RCA REVIEW

a technical journal

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

VOLUME XIX DECEMBER 1958 NO. 4

PRINTED IN U.S.A.

RCA REVIEW, published quarterly in March, June, September, and December by RCA Laboratories, Radio Corporation of America, Princeton, New Jersey. Entered as second class matter July 3, 1950 under the act of March 3, 1879. Second-class postage paid at Princeton, New Jersey, and at additional mailing offices. Subscription price in the United States and Canada; one year $2.00, two years $3.50, three years $4.50; in other countries: one year $2.40, two years $4.30, three years $5.70. Single copies in the United States, $.75; in other countries, $.85.

RCA REVIEW

a technical journal

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

Published quarterly by

RCA LABORATORIES

in cooperation with all subsidiaries and divisions of

RADIO CORPORATION OF AMERICA

VOLUME XIX | DECEMBER 1958 | NUMBER 4

CONTENTS

RCA REVIEW is regularly abstracted and indexed by *Industrial Arts Index Science Abstracts* (I.E.E.-Brit.), *Electronic Engineering Master Index, Chemical Abstracts, Proc. I.R.E.*, and *Electronic & Radio Engineer.*

RCA REVIEW

ON THE QUALITY OF COLOR-TELEVISION IMAGES AND THE PERCEPTION OF COLOR DETAIL*†

By

O. H. Schade, Sr.

RCA Electron Tube Division,
Harrison, N. J.

Summary—A theoretical and experimental study of the NTSC color system, supported by color photographs, shows that contrast range and color saturation obtained with commercial tricolor kinescopes provide a larger color space than provided by color motion pictures.

In fine detail more than 60 per cent of full color information is transmitted and reproduced by the NTSC system because the bandwidth restrictions of the electrical color signals (I,Q) do not affect definition in the vertical dimension and have a smaller effect on the reproduction of horizontal color detail than indicated by earlier evaluations which disregarded the two-dimensional nature of the image.

The detail color reproduction appears adequate to the eye because the color errors remaining, although perceptible, are small. This fact is significant because the spatial sine-wave response functions of the color discriminators of the visual system are found to be substantially independent of the color of light and similar to the spatial sine-wave luminance response function of the eye.

PART I—ELECTRICAL AND OPTICAL CHARACTERISTICS OF NTSC COLOR TELEVISION SYSTEM

General Characteristics of Color Systems

THE reproduction of color in a television or photographic system is based upon the trichromatic theory. The analysis of color in a television or photographic camera requires a discriminator mechanism having three different spectral sensitivities, resulting in three integrals or "primary signals" R, G, B** (or any linear transformation thereof, such as the NTSC values Y, I, Q) which are independent two-dimensional intensity functions of the object-point coordinates. In the synthesis of a colored image, luminance and color of all image points is restored by letting the primary signals control

* Manuscript received November 10, 1958.

† This paper was presented at the Convention of the Society of Motion Picture and Television Engineers of Detroit, Michigan, on October 22, 1958. It is also being published in the December, 1958 issue of the Journal of that Society.

** The spectral sensitivities or wavelength functions $f(\lambda)$ should be such that the primary signals are approximately proportional to the tristimulus values of the object, the tristimulus values being those of the receiver primaries.

three properly chosen "primary lights": the reproducing primaries red, green, and blue. The reproduction of color by a linear or a non-linear system must be independent of illumination intensity to conform with the requirements of vision. A stable color balance necessitates a constant ratio of the three signal functions, which requires:

(a) Matched transfer characteristics,
(b) Matched spatial frequency spectra of the system, and
(c) Matched noise-levels.

These specifications express a *registry requirement*. The *registry of three transfer characteristics* is an old problem in photography and a new problem in television, demanding high precision in maintaining relative and absolute gain and black-level stability in linear or non-linear amplifiers as well as in camera tubes and kinescopes, where sensitivity or response within the image area must be highly uniform. *Registry of the spatial frequency spectra* of three color images requires accurate time-delay and phase equalization in electrical channels and a closely matched image geometry in three scanning rasters. *Matched noise levels* become important when the signal-to-noise ratios are low, because unmatched levels result in a mismatch of transfer characteristics.

For the reason of "compatibility" in television, the color information, together with a luminance signal suitable for monochrome receivers, must be transmitted without increase of video bandwidth. The two additional color-transmission functions (I, Q) are, therefore, restricted to smaller passbands than the luminance function (Y).

The present analysis does not concern itself with faults of a temporary nature in the operation of a color system, but with the range and purity of colors obtainable, in particular with the unequal passbands of NTSC television signals and with commercially available color kinescopes. The computed colorimetric performance is checked by measurements, supported by visual observations (color photographs), and interpreted with reference to the performance of color photography and the visual system.

Color Television System for Measurements and Visual Tests

The color signals are generated by a light-spot slide scanner (see Figure 1) which is followed by three gamma-correction amplifiers and band-limiting filters feeding the cross mixing network or Matrix I (RCA "colorplexer") for translation of the input signals E_R, E_G, and E_B into E_Y, E_I, and E_Q signals. The matrix coefficients for the system are given in Table I. After band limitation by I and Q filters

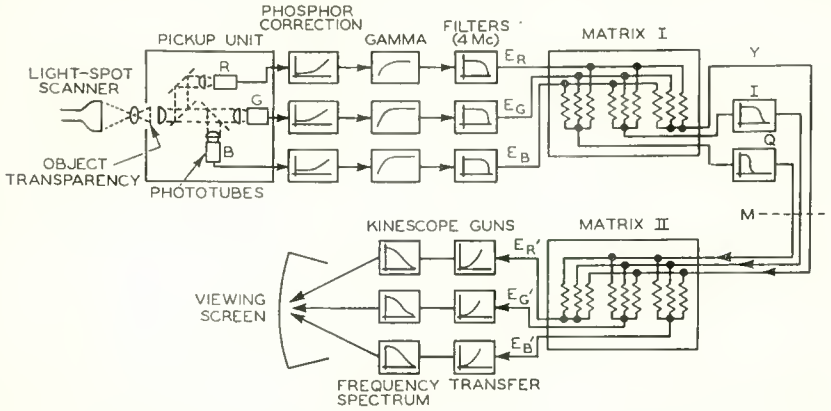


Fig. 1—Block diagram of color system. (Modulator and demodulator system (not shown) can be inserted ahead of Matrix II at point M.)

in the colorplexer, the signals are translated back to $E'_R E'_G E'_B$ in the inverse Matrix II. The primes indicate that each signal is now a frequency-dependent mixture of components from three unequal passbands as indicated in Table I and discussed later. These signals are then modified by the transfer characteristics and frequency characteristics of the electron guns of the color kinescope and superimposed on the kinescope screen, where they are converted to light energy. Modulation and demodulation circuits (not shown) can be switched in at point *M* to observe effects introduced by transmission of the multiplexed NTSC signal over a single channel.

For measurements of the dynamic transfer characteristic of the system, the light intensity in a fixed small area on the kinescope screen is observed by a multiplier phototube while a calibrated step tablet image is slowly drifting over this area in a vertical direction. The step signal is generated by a color transparency having the characteristics indicated in Figure 2. The slow vertical drift is generated

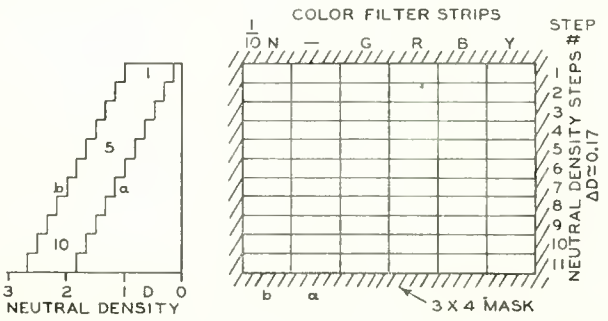


Fig. 2—Test pattern for measuring transfer characteristics.

Table I—Voltages in NTSC Color System

	Transmission Primaries, (Matrix I)				E' Components from Matrix II				Sum E' in Freq. Range Δf (mc)			Luminance Factor for Linear System		
Color	E	E_Y	E_I	E_Q	Gun	Y	I	Q	$(Y+I+Q)$ $\Delta f=0\rightarrow0.6$	$(Y+I)$ $0.6\rightarrow1.8$	Y $1.8\rightarrow4$	$l_{0.6}$	$l_{.6\ 1.8}$	$l_{1.8-4}$
R	1.0	.299	.596	.211	Red	.30	.57	.13	1.00	.87	.3	.30	.26	.09
	...				Green	.30	—.16	—.14	0	.14	.3	0	.08	.18
	...				Blue	.30	—.66	.36	0	—.36	.3	0	—.04	.03
												Σ .30	.30	.30
G	...	.587	—.274	—.523	Red	.59	—.26	—.33	0	.33	.59	0	.10	.18
	1.0				Green	.59	+.07	+.34	1.00	.66	.59	.59	.39	.35
	...				Blue	.59	+.30	—.89	0	.89	.59	.0	.10	.06
												Σ .59	.59	.59
B	...	.114	—.322	.312	Red	.11	—.31	+.20	0	—.20	.11	0	—.06	.033
	...				Green	.11	.09	—.20	0	.20	.11	0	.12	.065
	1.0				Blue	.11	.36	.53	1.00	.47	.11	.11	.05	.012
												Σ .11	.11	.11
W	1.0	1.	0	0	Red	1.	0	0	1.0	1.0	1.0	.30	.30	.30
	1.0				Green	1.	0	0	1.0	1.0	1.0	.59	.59	.59
	1.0				Blue	1.	0	0	1.0	1.0	1.0	.11	.11	.11
												Σ 1.0	1.0	1.0
Y	1.0	.886	.322	—.312	Red	.89	.31	—.20	1.0	1.20	.89	.30	.36	.67
	1.0				Green	.89	—.09	+.20	1.0	.80	.89	.59	.47	.53
	...				Blue	.89	—.36	—.53	0.	.53	.89	0.	.06	.09
												Σ .89	.89	.89
C_Y	...	.701	—.596	—.211	Red	.70	—.57	—.13	0	.13	.70	0.	.04	.213
	1.0				Green	.70	.16	.14	1.0	.86	.70	.59	.51	.415
	1.0				Blue	.70	.66	—.36	1.0	1.36	.70	.11	.15	.072
												Σ .70	.70	.70
M_G	1.0	.913	.274	.523	Red	.41	.26	.33	1.0	.67	.41	.30	.20	.123
	...				Green	.41	—.07	—.34	0.	.34	.41	0.	.20	.245
	1.0				Blue	.41	—.30	.89	1.0	.11	.41	.11	.01	.042
		1				2			3			Σ .41	.41	.41

by synchronizing the vertical oscillator of the light-spot scanner with a stable oscillator differing very slightly in frequency from the field frequency. The screen of the kinescope is thus illuminated continuously in its entire area by the drifting colored test pattern picture, while the monochrome step intensities are traced on a recorder. The optical step signal is generated in two sections. The main step tablet (a) covers a 70-to-1 range, and a 10-to-1 neutral filter strip (b) extends this range to 700. Measurement of the electrical signal steps at intermediate points of the system furnishes the various transfer characteristics shown in Figure 3.

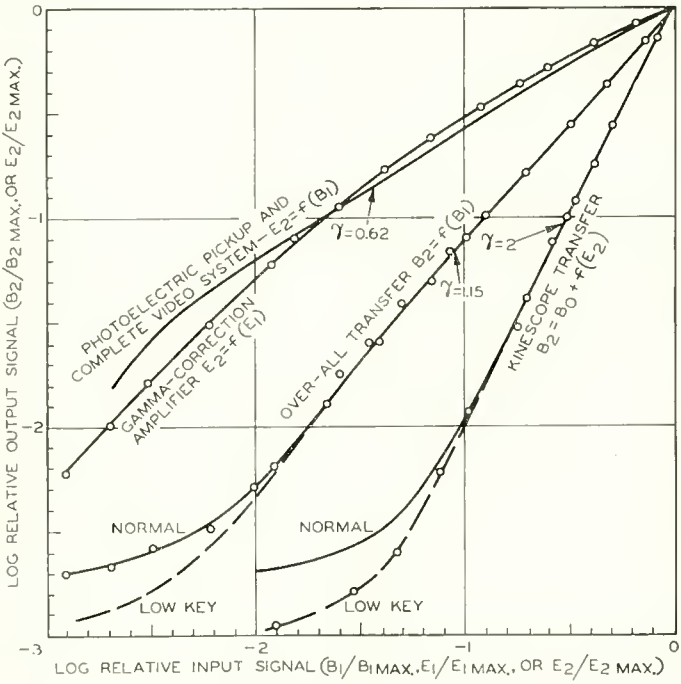


Fig. 3—Transfer characteristics of color system.

Picture-Tube Transfer Characteristics

The dynamic transfer characteristic shown in Figure 3 was measured on a 21AXP22 metal color kinescope. It follows a power law (close to a square law) with an additive constant B_0, which expresses the black-level illumination or light-bias on the screen resulting in a "toe" in the characteristic. The light bias is caused by diffuse electron excitation (secondaries), optical diffusion in the screen, and ambient light. It is variable, therefore, and has its lowest value for a dark

viewing room and for pictures having a high ratio of peak-to-average luminance. For normal picture material, the range of the transfer characteristic is similar to that measured with the colored test pattern (Figure 2), i.e., approximately 600 to 1 in a dark room.* This range is much higher than obtained in motion picture theaters (60 to 100 to 1) where the light bias, B_0, is caused by projection lens flare and ambient light; in fact it is even higher than the contrast range of the color film positive itself* which is in the order of 470 to 1.** The large range of the color kinescope cannot be photographed in its entirety on any Kodachrome or Ektachrome film and is further reduced by the color printing process in the various illustrations.

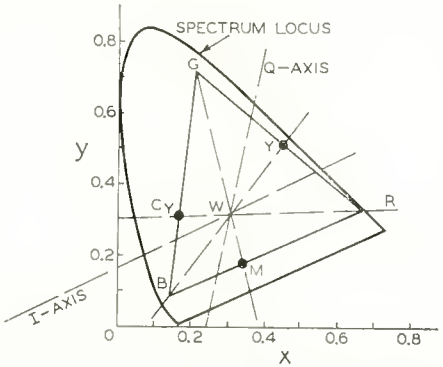


Fig. 4—CIE diagram showing location of television receiver primaries (R,G,B), I and Q axes, and color planes.

Excitation Purity (Color Saturation)

The degree of color saturation obtainable in a color image is determined by the reproducing primaries and the contrast range of the reproducer. The red, green, and blue primaries of the color kinescope are located in the CIE diagram as shown in Figure 4. A straight line drawn through the white point (illuminant C) intersects the spectrum locus at two points. The *excitation purity*, S, of a color located on such a line is, by definition,[1] the ratio of the distance from the color point to the white point, to the distance from the spectrum locus to the white point. The excitation purity is thus specified by a mixture

* The reader may compare these values with those given in Reference (2), showing transfer characteristics of early (1954) color kinescopes covering a range of hardly more than 20 to 1.

** Today's color kinescopes (21CYP22) have an even higher contrast because of reduced electron diffusion by secondary emission.

[1] Dean B. Judd, "Colorimetry," *National Bureau of Standards Circular 478*, March, 1950.

of a spectral light with white light. The fixed light bias, B_0, caused by ambient or scattered light will be assumed to have the neutral color of illuminant C. The addition of this white light moves all color points towards the white point and decreases the excitation purity of the color from the value S_c, computed for $B_0 = 0$ (see Table II), to a smaller value, S. The purity reduction by a fixed white light energy is obviously dependent on the excitation level of the color and its relative stimulus energy, defined as follows: The maximum luminance Y_{max} in a given color system is that of the white light ($Y_{max} = Y_{w\ max}$) obtained when the three primary lights (phosphors) are excited to selected maximum luminance values ($Y_{r\ max}$, $Y_{g\ max}$, $Y_{b\ max}$) giving the specified white light (Illuminant C for color television). The maximum

Table II—Tristimulus Energy Factors, w_c, Luminance Factors, l_c, and Excitation Purity, S_c, NTSC Standards

Color	w_c	l_c	Dominant λ (millimicrons)	S_c	Trichromatic Coefficients x	y	z
W	1.00	1.0	...	0	0.31	0.316	0.374
R	0.286	0.299	611	1.0	0.67	0.33	0.00
G	0.261	0.587	535	0.85	0.21	0.71	0.08
B	0.453	0.114	470	0.88	0.14	0.08	0.78
Y	0.547	0.886	573	0.88	0.45	0.512	0.038
M_G	0.739	0.413	535 compl.	0.625	0.345	0.177	0.48
C_Y	0.714	0.701	490.5	0.52	0.165	0.31	0.525

Munsell Lightness Value, V, and Luminance, Y

V	10	9	8	7	6	5
Y (%)	102.56	78.66	59.1	43.06	30.05	19.77
V	4	3	2	1	0.5	0.2
Y (%)	12.	6.56	3.126	1.21	0.581	0.237

luminance of a color, c, is therefore obtained when at least one of the primary lights reaches the excitation limit established for maximum white light excitation.

The excitation level of a color is hence specified by the excitation factor

$$k_c = Y_c / Y_{c\ max} \leqq 1. \tag{1}$$

The maximum luminance of a color relative to the maximum luminance is expressed by the luminance factor

$$l_c = Y_{c\ max} / Y_{w\ max}, \tag{2}$$

and the relative tristimulus energy of a color is given by the product

$$k_c w_c = k_c (X + Y + Z)_{c\ \max} / (X + Y + Z)_{w\ \max}. \tag{3}$$

The tristimulus energy factor, w_c, can be expressed in terms of trichromatic coefficients, y, and the luminance factor, l_c, by the simple relation

$$w_c / l_c = y_w / y_c. \tag{4}$$

The tristimulus energy factor, w_0, of the light bias is specified by the maximum contrast ratio C,

$$w_0 = w_w / C = 1/\mathrm{C}. \tag{5}$$

The mixture ratio $k_c w_c / w_0$ of the relative tristimulus energy of the color to the fixed relative tristimulus energy of the light bias is equal to the distance ratio $S/(S_c - S)$, which gives the desired purity relation

$$S/S_c = k_c w_c / (k_c w_c + w_0), \tag{6a}$$

and, with Equation (5),

$$S/S_c = k_c w_c / (k_c w_c + 1/C). \tag{6b}$$

The total relative luminance of the color and light bias energy is the sum

$$Y/Y_{\max} = k_c l_c + 1/C. \tag{7}$$

Inspection of Equation (6a) shows that only an ideal color reproducer having an absolute black level ($w_0 = 0$) provides a constant excitation purity for all colors, independent of excitation, k_c. A plot of excitation purity, S, as a function of relative luminance ($Y/Y_{\max}$) for all possible colors furnishes a *color space* in which arctan y/x is the vectorial direction of the color from the white point (see Figure 4), excitation purity is the vector length, and relative luminance is the elevation.[2] The axis of this space, erected over the white point, represents the gray scale, and vertical sections through this axis furnish *color planes* such as shown in Figures 5-7. The boundaries of these color planes include all possible colors and permit a comparison of different reproducers.

Given the NTSC primaries and constants (Table II) and the contrast range, C, of the picture tube, the lower boundary of television

---

[2] W. L. Brewer, J. H. Ladd and J. E. Pinney, "Brightness Modification Proposals for Televising Color Film," *Proc. I.R.E.*, Vol. 42, p. 174, January, 1954.

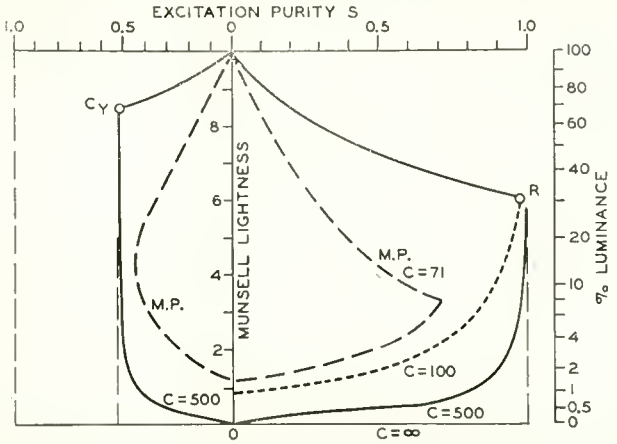


Fig. 5—Red-cyan color plane of additive television process (shadow-mask kinescope $C = 500$), solid lines, and subtractive M.P. process ($C = 71$) broken lines.[2]

color planes is constructed by selecting a color *on* the color triangle (Figure 4) and computing the coordinates S, (Y/Y_{max}) of points on the boundary curve with Equations (6) and (7) by assigning values between 0 and 1 to the parameter k_c.

The computed boundary extends up to the relative luminance $l_c + 1/C$ at full excitation ($k_c = 1$) of the primary color (or two-component mixture). Higher luminance values in the particular color plane can therefore be obtained only by adding light of complementary color which combines with a certain fraction k' of the color energy to a partial white excitation, having the relative stimulus energy $k'w_w$.

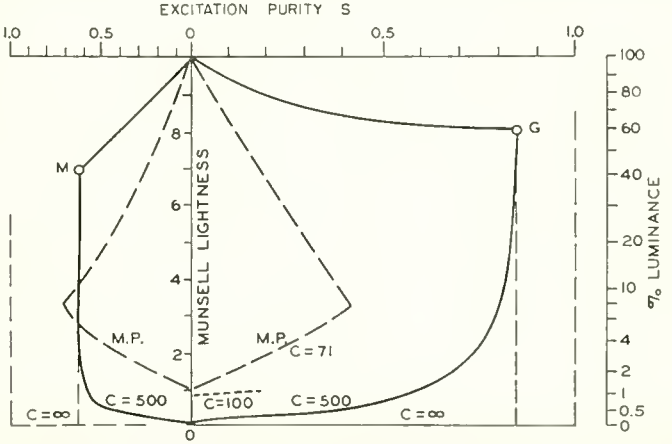


Fig. 6—Green magenta color plane of additive television process (shadow-mask kinescope $C = 500$), and subtractive M.P. process ($C = 71$) broken lines.[2]

The remaining relative stimulus energy of the color is $(1 - k')w_c$. The excitation purity is hence determined by the mixture of the remaining color amount $(1 - k')w_c$ with the amounts of white $k'w_w + w_0 = k' + 1/C$, which leads to the expression for the upper boundary of the color space;

$$S/S_c = (1 - k')w_c/[(1 - k')w_c + k' + 1/C], \tag{8}$$

where $k' \leqq 1$ (white excitation factor).

The relative luminance is given by

$$Y/Y_{max} = (1 - k')\, l_c + k' + 1/C. \tag{9}$$

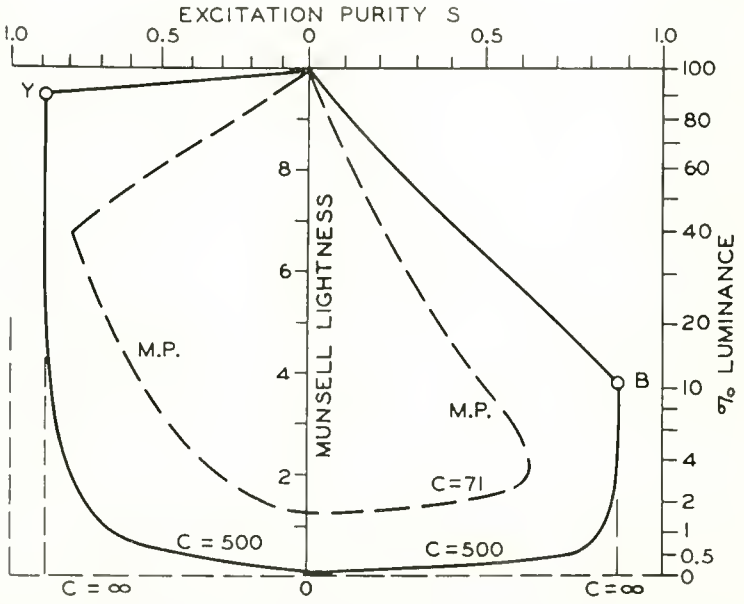


Fig. 7—Yellow-blue color plane of additive television process (shadow-mask kinescope $C = 500$), and subtractive M.P. process ($C = 71$) broken lines.[2]

The relative luminance scale in Figures 5-7 is distorted, because a better appreciation of the contrast range and color purity of a reproducer is obtained by using the psychophysical *Munsell lightness scale, V*, which divides the dynamic luminance range of the visual system into uniform lightness steps. (The conversion from per cent luminance to lightness, V, is given in Table II).

A "perfect" color reproduction requires an absolute black level ($C = \infty$) and reproduces all spectrum colors with 100 per cent purity;

i.e., the color triangle is replaced by the spectrum locus. The lower boundaries of "perfect" color planes are therefore rectangles, indicated by $C = \infty$ and $S = 1$ up to the relative luminance for the spectral color point as exemplified by the red television primary. The upper boundary of the "perfect" color plane is determined by Equations (8) and (9) with $1/C = 0$ and w_c equal to the tristimulus energy factor for the spectral color point. Because their form is not strongly dependent on the contrast, C, the upper boundaries of the television color planes R, G, B, and Y approach those of the "perfect" color planes for the corresponding spectral color points. It is seen from Figure 4 that the cyan and magenta regions of the color triangle lie approximately half way between the spectrum locus and the white point. The corresponding color planes in Figures 5 and 6 have, therefore, approximately one half the width of the corresponding "perfect" color planes. It is further evident that the 600-to-1 contrast range obtained with a color kinescope comes much closer to a "perfect" color reproduction than does a motion picture. The color space boundaries of a color motion picture (broken lines taken from Reference (2)) are, by comparison, much more restricted and illustrate the basic fact that a subtractive process rapidly loses saturation at higher luminance values, because an increase in film transmission reduces the dye concentration and its color filter action.

To approach the chroma obtainable by the addititive color television reproduction, color film could be modified to have a neutral density range of approximately 3.7, to be used with a minimum neutral highlight density near unity to retain a sufficiently high dye concentration. Unfortunately this modification leads to the requirement for ten times more light from the projector which results in heating and other film problems.

The difference in performance between normal color film and color television can be demonstrated strikingly by color photographs of television images having constant luminance (constant Y-signal) and various degrees of color saturation, obtained by a progressive increase of I and Q signals (chroma control). Photographs were taken with chroma signals 1, 2, and 3 times normal. The 2 times chroma value produced very high color saturation on the kinescope. The corresponding color photographs however taken with *normal exposures* show only minor increases in chroma as illustrated by plates I and II and expected from the color diagrams Figures 5-7, while the denser photographs reproduced in Plates III and IV taken with ½ normal exposure give a better reproduction of the kinescope chroma at the expense of a shorter distorted contrast range. Relative chroma values are fairly

well reproduced in the prints, although the color purity is lower than in the transparencies and considerably lower than in the original kinescope image.

Sine-Wave Spectra and Optical Passbands

The transmission of color requires three independent video signals as compared to one for a monochrome image. Equal definition in a color image requires thus, in theory, a transmission system having 3 times the information capacity of a monochrome system. An appraisal of the total information capacity of the NTSC color system can be obtained by comparing its information capacity with that of a color system having 3 equal independent channels, taking into account a number of non-ideal conditions arising in practical systems and deserving particular attention.

The electrical frequency spectrum of a *stationary* monochrome television picture signal is a line spectrum of discrete frequency components which are harmonics of the frame frequency (30 cycles). Because of this fact, a second line spectrum, the *color signals (I, Q)*, can be added to the monochrome or *luminance signal (Y)* by interleaving its frequency components with those of the *Y* spectrum. Interleaving is accomplished by modulation of a color carrier frequency (3.579545 megacycles) which is made an odd multiple of the half-line and half-frame frequencies. To permit separation of color signals by synchronous demodulators in the receiver, one color signal (*Q*) is transmitted with double sidebands. It is, therefore, limited to a 600-kilocycle bandwidth, and a 600-kilocycle filter is required after demodulation to eliminate all higher cross-product frequencies. The other signal, *I*, (also double-sideband up to 600 kilocycles) can have its bandwidth extended by single-sideband transmission. It is restricted to ½ the color carrier frequency, i.e., a bandwidth of 1.8 megacycles, and a 1.8-megacycle filter is required after demodulation to eliminate crosstalk.

A perfect separation of the interleaved luminance and chrominance signals can be achieved with interleaved comb filters when the image is stationary. The inexpensive continuous passband filters used in practical receivers, however, give rise to spurious color signals upon synchronous demodulation in the *I* and *Q* channels, caused by high-frequency *Y*-signal components, and the chrominance signal (modulated color carrier) produces periodic errors in the luminance signal. These errors change polarity in successive frames and would cancel out in a linear system when integrated (by the eye) over two frame periods

(1/15 second). Practical systems, however, are not linear (all kinescopes are rectifiers), and integration by the eye is incomplete, particularly for bright pictures. The errors, therefore, do not cancel completely even in stationary pictures. They become larger for moving objects, and attain full magnitude for random signals such as noise.

The high-frequency "crosstalk" from the common 2-4 megacycle region of a "flat" Y spectrum into the color demodulator band causes orange and blue color tinting of horizontal monochrome detail (coloring of resolution wedges in a monochrome test pattern), and normally fine-grained camera noise in this region of the Y channel is heterodyned

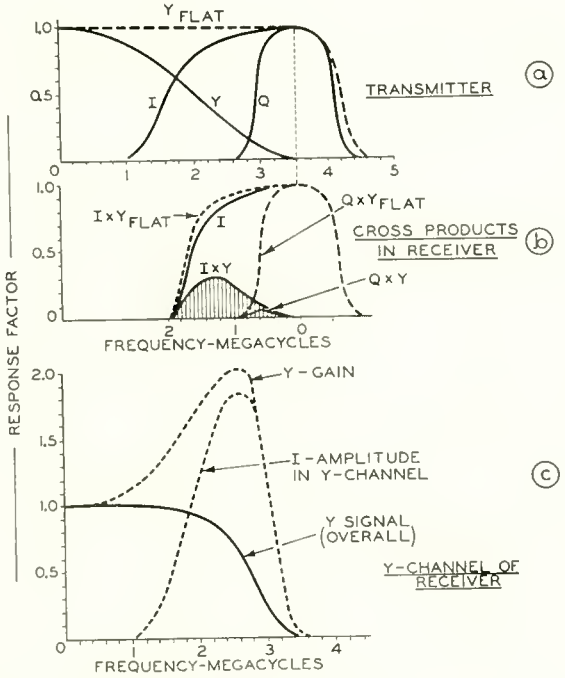


Fig. 8—Filter response and "crosstalk" in a color system.

into objectionable coarse color-noise (streaks) by the demolulators, as indicated by the large values of the broken-line cross product curves in Figure 8b.

These undesirable effects can be substantially eliminated by introducing a bandwidth limitation of 3.6 megacycles and a gradual "roll-off" into the Y channel *before synthesis* of the NTSC signal in the colorplexer of the transmitting station as shown by the solid-line curves in Figures 8a and 8b. Subsequent reemphasis (aperture correction) of the Y channel in the receiver (see Figure 8c) after removal of the

color carrier restores a good Y-signal response. The various degrees of color crosstalk are illustrated in Plates V and VI. (The bluish background color of the "white" test pattern photographs resulted from the accidental omission of the ultraviolet filter on the camera lens.) The striking reduction of noise obtained by the bandwidth limitation and "roll off" is not reproduced because of integration by the time exposure. Considerable improvement in noise crosstalk is obtained by the bandwidth limitation alone.

The complete elimination of the chrominance signal (modulated color carrier) from the Y signal (in the receiver) is not possible with continuous filters. It is therefore common practice to suppress the color carrier and its lowest sideband frequencies in the receiver by insertion of a filter (trap) which limits the Y-channel response to 3.6 megacycles (see Figure 8c). The carrier interference is thus completely eliminated in large areas, leaving only beat patterns of reduced amplitude (occurring near sharp vertical edges) from the remaining sideband components (I amplitude in Y channel Figure 8c) which generally contain little energy and permit aperture correction of the remaining Y channel at the receiver.*

There are three electrical passbands available for the transmission of color signals; these correspond to three two-dimensional optical passbands for the color image. Because of the rectilinear scanning process, the three optical passbands in the "vertical" (v) coordinate are alike. They are determined by the raster line number, and have the theoretical equivalent passbands:[3] $N_{e(Y)v} = N_{e(I)v} = N_{e(Q)v} =$ 490 lines. The equivalent passbands in the horizontal coordinate (h) are unequal and have the theoretical values $N_{e(Y)h} = 320$ for $\Delta f = 4$ megacycles, $N_{e(I)h} = 144$ for $\Delta f = 1.8$ megacycles, and $N_{e(Q)h} = 48$ lines for $\Delta f = 0.6$ megacycle.

The equivalent symmetrical passbands[3]

$$\overline{N}_e = \left(\frac{4}{\pi} N_{e(v)} N_{e(h)}\right)^{1/2} \tag{10}$$

are hence

$$\overline{N}_{e(Y)} = 446 \text{ lines},$$

$$\overline{N}_{e(I)} = 300 \text{ lines},$$

$$\overline{N}_{e(Q)} = 173 \text{ lines}.$$

* The aperture correction approximately doubles the chrominance signal amplitude in the Y channel.

[3] Otto H. Schade, "Image Gradation, Graininess and Sharpness in Television and Motion-Picture Systems, Part IV—Image Analysis in Photographic and Television Systems," *Jour. S.M.P.T.E.*, Vol. 64, p. 593, November, 1955.

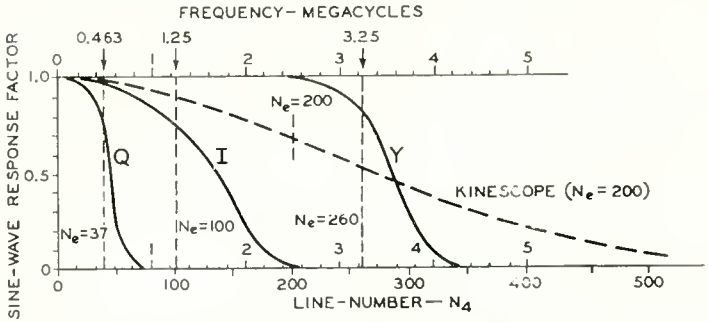
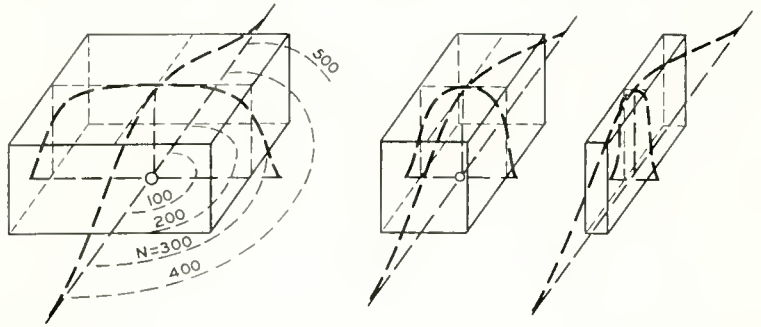


Fig. 9—Sine-wave spectra of horizontal (Y, I, and Q) transmission passbands used in the analysis.

Their sum $\Sigma \overline{N_e} = 919$ lines is hence 68 per cent of the sum of three equal passbands of 446 lines each. This simple appraisal of total information capacity in such a color system assumes theoretical rectangular frequency spectra having abrupt cutoff which are neither practical nor desirable for image transmission (monochrome or color) because of strong edge transients and spurious signals generated in the signal-separating process.

The sine-wave response characteristics of practical electrical passbands used in the following analysis are shown in Figure 9. Their equivalent spatial horizontal passbands are $N_{e(Y)h} = 260$,* $N_{e(I)h} = 100$ and $N_{e(Q)h} = 37$. The spatial vertical passband is determined by the cascaded value of the camera and kinescope sine-wave spectra and is in the order of $N_{e(v)} \simeq 200$ for each of the three signals. The equivalent spectrum spaces for the *three practical* channels are the rectangular solids illustrated in Figure 10. The equivalent symmetrical pass-



CHANNEL	Y	I	Q
N_e (v)	200	200	200
N_e (h)	260	100	37

Fig. 10—Two dimensional sine-wave spectra (Y, I, and Q passbands).

* This Y channel is somewhat wider than for receiver use (Figure 8) because it does not contain a trap circuit for the carrier frequency.

bands obtained with Equation (7) form cylindrical spectrum spaces having the radii

$$\overline{N}_{c(Y)} \simeq 257 \text{ lines},$$

$$\overline{N}_{c(I)} \simeq 160 \text{ lines},$$

$$\overline{N}_{c(Q)} \simeq 97 \text{ lines}.$$

Because of the coordinate transformation or cross-mixing processes and unequal passband limitations, the electrical sine-wave spectra ("frequency response") for different colors are not equal in this system and can be determined as follows. Referring to Figure 1, it is seen that the electrical frequency spectra of the color signals, E_R, E_G, E_B entering Matrix I are alike. The linear cross-mixing process to Y, I, and Q signals and back to $E'_R\ E'_G\ E'_B$ in the inverse Matrix II does not disturb the signal ratios in the range of 600 kilocycles where the frequency response in the Y, I, and Q channels is alike, i.e., $E'_R = E_R$, $E'_G = E_G$, $E'_B = E_B$. *This range includes the complete electrical frequency spectrum required for transmission of the vertical spatial passbands of the three color functions which remain, therefore, unaffected by the cross-mixing processes.* The *horizontal passbands* are normal in the 600-kilocycle range, i.e., a green signal E_G, for example, will result in a green signal E'_G on the green kinescope gun and in zero signals on the other two guns as indicated in Table I under the column $\Delta f = 0 \rightarrow 0.6$. In the range from 600 kilocycles to 1.8 megacycles, however, all Q coefficients in Matrix II are zero because of the Q-filter cutoff. The matrix is no longer the inverse of Matrix I, and causes signal voltages to appear on all three kinescope grids as shown in Table I. Beyond 1.8 megacycles, both Q and I coefficients are zero, with the result that all three kinescope guns receive equal signals. The last three columns of Table I show that the total luminance would remain constant in all sections of the total passband for a hypothetical *linear* kinescope, although the color or color mixture does not remain constant because of "spurious" signals. The complete video frequency spectra (at the control grids of the kinescope guns) have been plotted in Figures 11 and 12 for the six colors listed in the Table I. Note that all of them are different, and some of the spurious responses have negative lobes (negative signs of coefficients indicate a phase reversal).

Effect of Unequal Passbands on Color Detail

It is a rather widely accepted opinion that the NTSC color transmission provides a three-color presentation for large areas, a two-color

presentation for medium sized areas and a presentation of fine detail without color information. This cannot be true because full color information is transmitted in the *vertical* dimension up to the finest detail which is not affected by the inequality of the electrical passbands. Detail color information is also transmitted in the horizontal dimension, because only the true or "fundamental" color signal components have a normal frequency spectrum including a d-c component, while the frequency spectra of the "spurious" color components contain only horizontal a-c components and represent one-dimensional high-pass filters. It is improper to disregard the difference in d-c components and conclude from Figure 11, for example, that a vertical line group in a green sine-wave test pattern (for which the blue and red camera

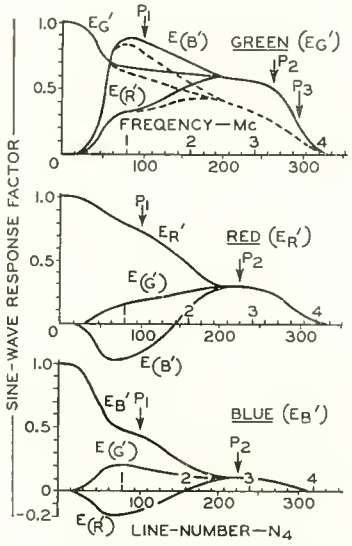


Fig. 11—Horizontal sine-wave spectrum components at kinescope grids for green, red, and blue camera signals.

signals are zero) would appear in black and white in the fine detail "mixed-high-frequency" region of line numbers $N_h > 200$.

Actually these lines appear green, because the green a-c signal only is raised completely above the black level by a d-c component, while the spurious color signals ($E'_{(R)}$ and $E'_{(B)}$) have an electrical zero-level axis, *are rectified by the kinescope,* and can only produce light of much lower intensity during positive half cycles. In computations of color mixtures for small areas it must be remembered that the normal Fourier relations between impulse forms, unit functions, and their spectra do not hold in systems containing nonlinear elements such as a kinescope.

To obtain the correct answer, it is necessary to convert frequencies to intensity functions (waveforms) at the input terminals of nonlinear elements such as the kinescope guns, project the waveforms over the nonlinear transfer characteristics, pass them through the low-pass filters representing the gun and optical performance, and then combine the light signals to color mixtures as indicated in Figure 1. The general *effect of the kinescope "spot" size* on fundamental color signals is a reduction of high-frequency components as indicated by broken line curves in Figure 11 for small signals. The spurious color signals,

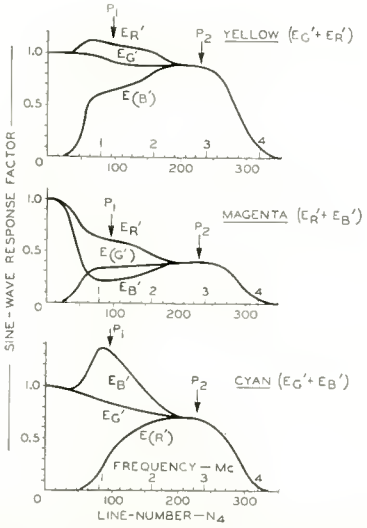


Fig. 12—Horizontal sine-wave spectrum components at kinescope grids for yellow, magenta, and cyan camera signals.

however, are rectified by the kinescope, so that each sine-wave frequency is replaced by a series of even-order harmonics which is attenuated more rapidly by the kinescope low-pass spectrum than indicated by the dotted lines, which lose their meaning as a sine-wave response.

The observed transfer characteristic of the actual kinescope is substantially a square-law characteristic. The intensity function of the fundamental green sine-wave pattern on the kinescope screen (neglecting kinescope spot size) can hence be calculated by squaring the instantaneous sine-wave grid-signal values measured from the

current cutoff point, as illustrated in Figure 13 (top) for two frequencies (points P_1 and P_2 in Figure 11). Because of the missing d-c component, the squared spurious red and blue a-c signals become very small and only their positive half-waves produce light, which adds to the green sine-wave light, on the kinescope screen, diluting the chroma as indicated. Similar conditions prevail for test patterns of other saturated colors (see Figure 13, bottom) as summarized in Table III.

The fine-line groups in the "mixed-high-frequency" region, therefore, retain substantially the color of the fundamental signal, although they acquire a spurious tint. Their relative amplitude and luminance is determined largely by the amplitude of the fundamental color component which is proportional to the Y component of the particular color and therefore lowest for blue (11.4 per cent). These findings are confirmed visually by turning off one or two of the three equal color signals generated by a "white" SMPTE test pattern ahead of Matrix I (see Figure 1) and comparing the color in the vertical line-wedge reproduction with the pure color reproduction in the horizontal line-wedges of the kinescope image.*

A similar result is obtaind by an *analysis of line images.* The images of fine horizontal lines are transmitted without color change (because of the matched vertical sine-wave response in all channels). The color of fine vertical lines, however, is affected by the unequal horizontal-frequency spectra and is determined by the sums of the three squared impulse components. The impulse forms computed for unit signals in the Y, I, and Q channels are shown in Figure 14, corresponding to the frequency spectra shown in Figure 9. For a given color-line signal the impulse forms at each of the three kinescope control grids are sums of these three impulses modified in amplitude by the amplitude factors for the particular color signal mixture (Table I, column group 2). The squared positive amplitudes of these impulse sums are transduced by the kinescope into light, i.e., into line-images, which are then added and analyzed for color mixture.

The impulse forms at the control grids of a color kinescope are shown in the top row of Figure 15a for a green line object. Note that only the fundamental color (green) has a positive average value (the spurious signal impulses red and blue contain no d-c term). A low-intensity green line object superimposed on a white background would represent a small signal on a d-c level. In this case the spurious as well as the fundamental signal components are transduced linearly and

* It is apparent that overemphasis of the electrical high-frequency response by excessive aperture correction increases the spurious color signals and chromatic errors.

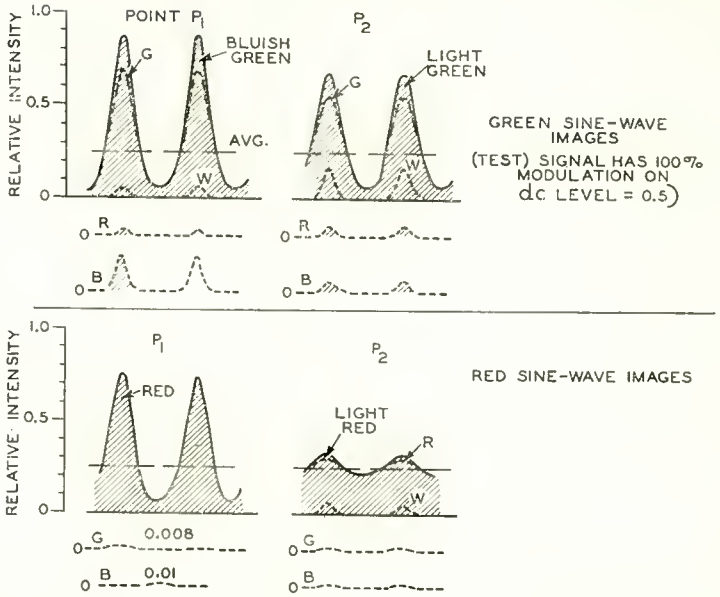
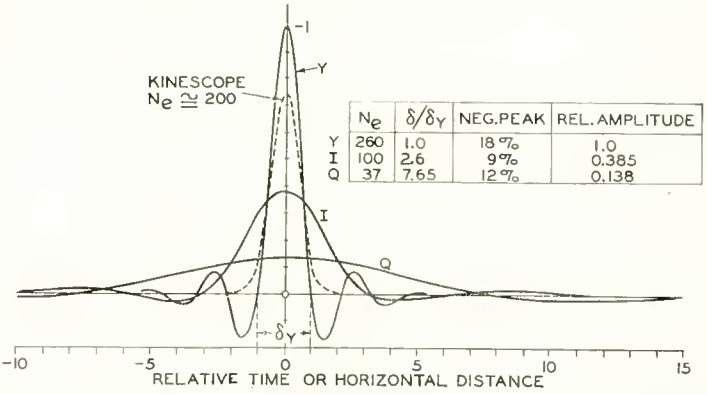


Fig. 13—Intensity function and components (broken lines) of green sine-wave images on the kinescope screen (top), and of red sine-wave images (bottom).

without rectification by the kinescope, resulting in the impulse sum shown at left in Figure 15(c). A high-intensity line object against a black background is a large signal on a zero d-c level, requiring separate squaring of the gun signals to obtain the intensity functions of the three colored line images on the kinescope screen shown in Figure 15b, which have no negative amplitudes (clipped by the kinescope). The



	Ne	δ/δY	NEG. PEAK	REL. AMPLITUDE
Y	260	1.0	18%	1.0
I	100	2.6	9%	0.385
Q	37	7.65	12%	0.138

Fig. 14—Unit-impulse forms of Y, I, and Q channel from Matrix I and of kinescope.

sum of these optical intensity functions is shown at the right side in Figure 15c. Since the transfer gain is adjusted so that equal electrical amplitudes result in "white," the amount of white contained in the impulse sum is given by multiplying the smallest one of three common components (R in this case) by three at any one distance from the line-image center. It is seen that the white dilution is lower in a squared large-signal line image than in a small signal, which by itself is not a pure color. This condition is also obtained for other colors as shown in Figures 16 and 17. It follows again that *vertical line images retain substantially their original hue* somewhat tinted, desaturated, or re-

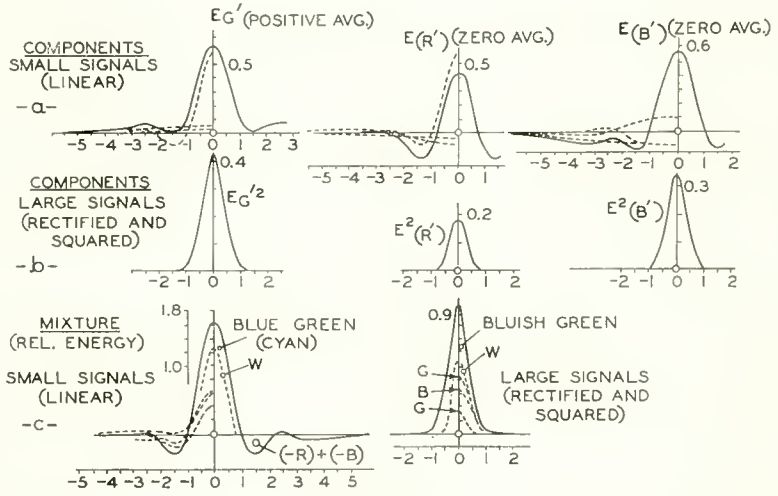


Fig. 15—Relative intensity of impulse components and sums forming the green line image.

duced in amplitude as summarizel in Table IV and substantiated visually in the color photographs Plate VII by comparison with the normal color reproduction of horizontal line objects. The illustration shows small sections of 5¼ x 7-inch picture reproductions. The faint blue tint in the vertical green lines is barely visible.

Color transitions at sharp edges can be synthesized in the same manner from the component step functions shown in Figure 18. The results are plotted in Figures 19-23. Inspection shows that both color and waveform distortions vary with color and are again relatively smaller in high-contrast transitions than in low-contrast transitions because of the square-law kinescope characteristics.

It should be pointed out that the effect of camera tube and kinescope beam sizes as well as the ½-power gamma correction for the kinescope

Table III—Color of Horizontal Sine-Wave or Periodic Line-Patterns

Object Color	P_1	P_2	P_3	(Fig. 13)
Green	bluish green (20% blue)	light green (26% white)	Green	
Red	red (1.5% green)	light red (17% white)		
Blue	blue (1.8% green)	light blue (22% white)		
Yellow	yellow (5% blue)	yellow (10% blue)		
Magenta	magenta (3% green)	magenta (4% green)		
Cyan.	cyan. (2% red)	cyan. (8% red)		

Table IV—Horizontal Line-Image Characteristics

Object Color	Incremental Impulse (Linear) Color*	Large-Signal Impulse (rectified, squared) Color*	Rel Brightness		Rel. Sharpness	
Green	light blue green (60% white)	bluish green (36% white)	normal	see Fig. 24	normal	see Plate VII
Red	red, orange tint (29% green)	red, slight or. tint (7% green)	reduced		fair	
Blue	greenish blue (17% green)	blue, green tint (7% green)	low		blurred	
Yellow	yellowish white (27% blue)	light yellow (21% blue)	normal		normal	
Magenta	reddish magenta (25% green)	reddish magenta (17% green)	reduced		fair (blue haze)	
Cyan.	bluish cyan. (20% red)	bluish cyan. (12% red)	good		normal	
White	white	white	normal		normal	

* The per cent color admixture refers to the area under the main lobe of the impulse or line. Incremental impulse color refers to the increment, not to the total which depends on the color of the set-up level.

have not been included in the line image and step-function evaluations. The gamma-correction amplifier itself would not change an ideal unit impulse or a step function, but it does round off the low-level corner of the actual S-shaped step function from the camera relative to its high-level corner, thus making the large-signal step functions more symmetrical and all step functions less steep with reduced transient ripples. To include the effect of the kinescope beam size, it is necessary to perform a convolution of its line image or its step-function response with the computed response forms. The finite kinescope beam size

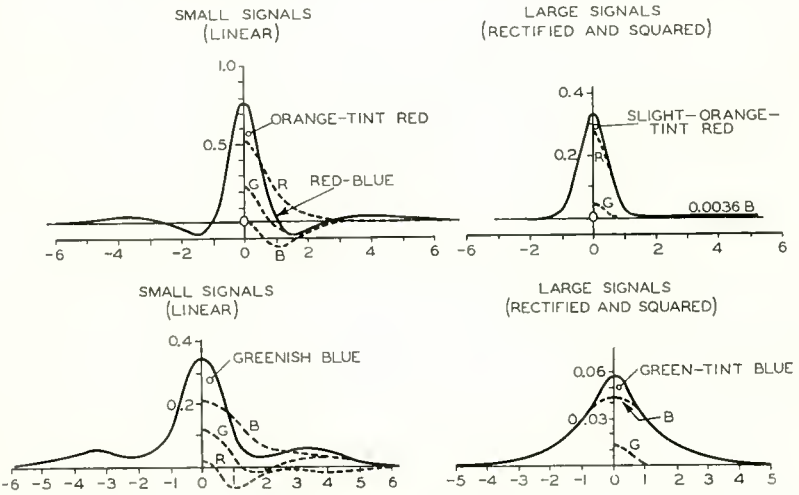


Fig. 16—Relative intensity of red and blue line images.

broadens the line images and transitions somewhat more, and further reduces high-frequency transients. The visibility of the luminance errors caused by the unequal passbands of the NTSC system can be assessed by computing the visual lightness of horizontal step functions and line images for red, green, and blue as shown in Figure 24. Constant-bandwidth vertical step functions and line images are shown for comparison, and indicate that the errors in the blue are readily visible as confirmed by observation (color plate VII).

It can thus be concluded that *considerable color information is transmitted and reproduced even in fine horizontal detail with NTSC color signals and that a high gamma (square-law) kinescope transfer characteristic is a definite asset because it reduces the color distortion introduced by unequal color-transmission channels.* Furthermore, the fact that relatively small color changes can be seen in the fine detail of color television images raises doubts that the eye's color mechanism

does not function when observing fine detail, and suggests a reappraisal of the capabilities of the visual system at normal luminance values.

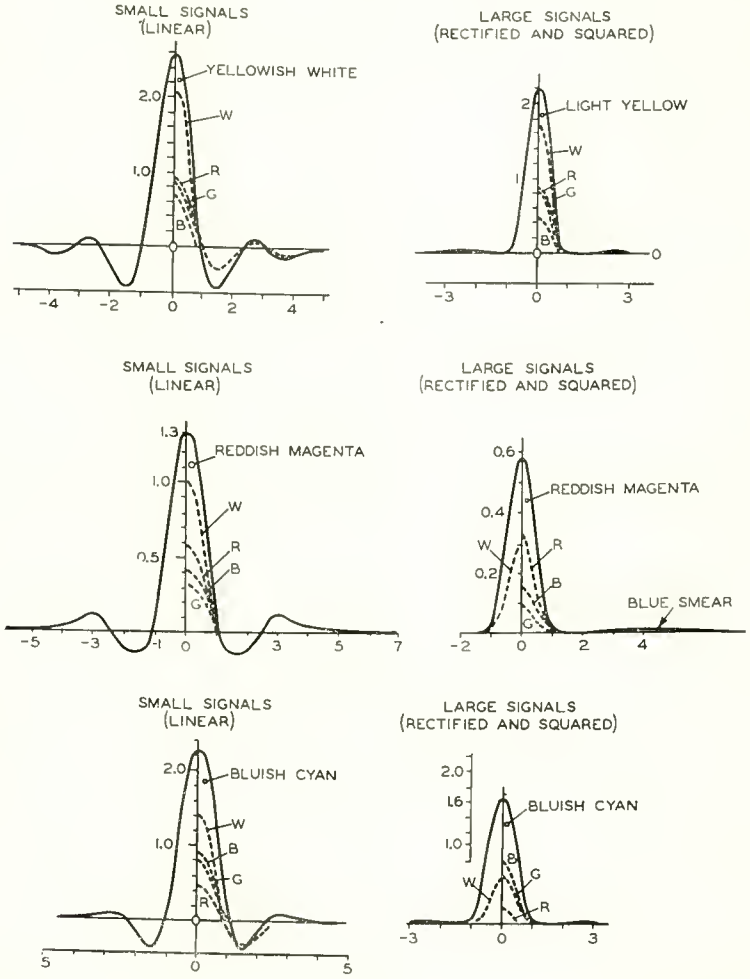


Fig. 17—Relative intensity of yellow, magenta, and cyan line images.

PART II—SOME CHARACTERISTICS OF VISUAL SYSTEM AND EFFECT ON OVER-ALL SYSTEM CONCEPT

Transfer Functions, Wavelength Functions, and Incremental Sensitivity for Luminance and Color

The visual system is a very complex system which terminates in a "computer" (the brain) having random connections, and very little is

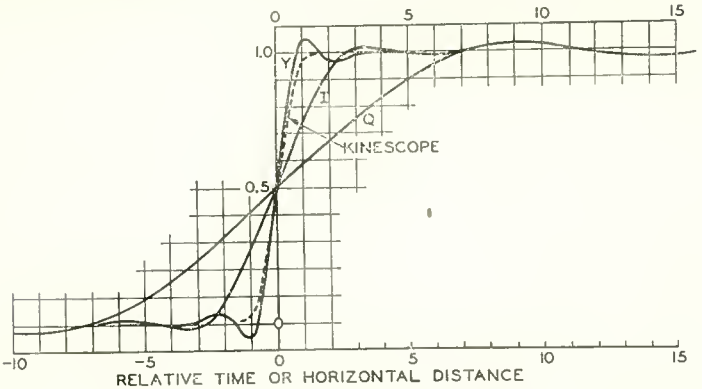


Fig. 18—Unit step-function components from Matrix I and of kinescope.

known about its memory and interpreting processes.* There is ample evidence that the system is nonlinear in many of its sections. It is well known that both hue and lightness of colored objects are functions of the surrounding background, although there is no spectral or colorimetric change. It is also known that the apparent "lightness" of a color mixture does not necessarily agree with the luminance value computed as the linear sum of component luminances. (A bluish-white

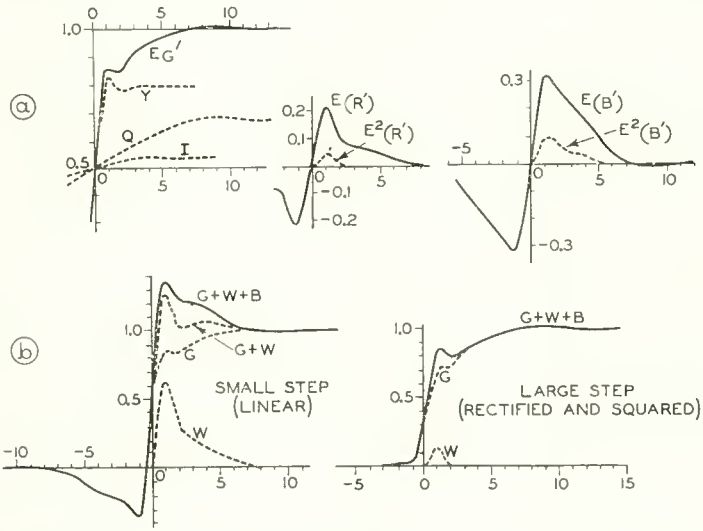


Fig. 19—Relative intensity of step-function components and sums forming the green step function image.

* An interesting theory on this subject is discussed in Reference (4).

[4] Frank Rosenblatt, "The Design of an Intelligent Automaton," *Research Trends*, Cornell Aeronautical Lab., Inc., Vol. VI, No. 2, Summer, 1958.

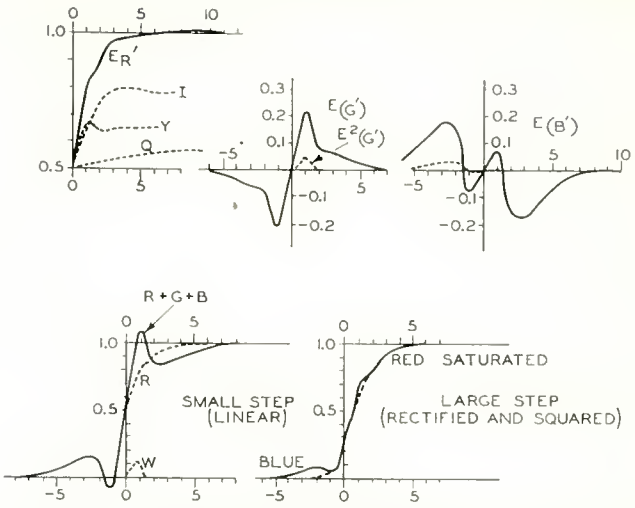


Fig. 20—Relative intensity of step-function components and sums forming the red step function image.

screen, for example, appears brighter than a screen of equal luminance illuminated by a low-color temperature light source.) Observations made by the author on equiluminous sine-wave mixtures indicate, likewise, that the fixed relation of lightness and luminance given in Table II does not necessarily apply to incremental amounts of the components in a mixture of primary colors, but that it is strongly dependent on the relative total amounts of the primaries.

With regards to chromatic sensitivity it is known that the just-

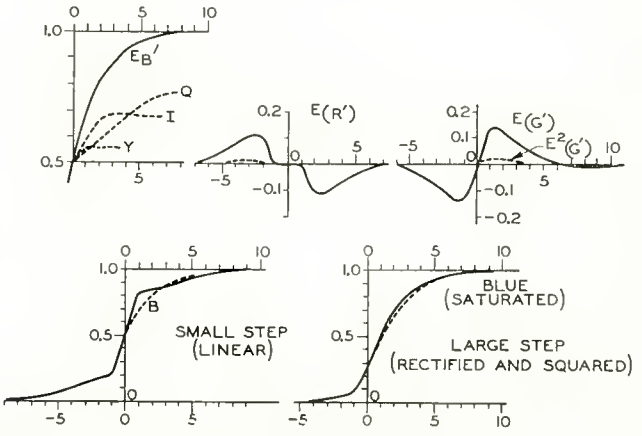


Fig. 21—Relative intensity of step-function components and sums forming the blue step function image.

perceptible amounts of pure color added to white is of the order of 1 per cent for red and blue and about 2 per cent for yellow and green.[5] The chromatic sensitivity $f/\Delta f$ or $\lambda/\Delta\lambda$ of the eye for spectral colors[6] has two peaks (less than 3:1) at yellow and cyan above a substantially uniform level from red (650 millimicrons) to blue (430 millimicrons). From the system analysis point of view, it does not appear likely that the wavelength functions of the eye's color discriminating system are

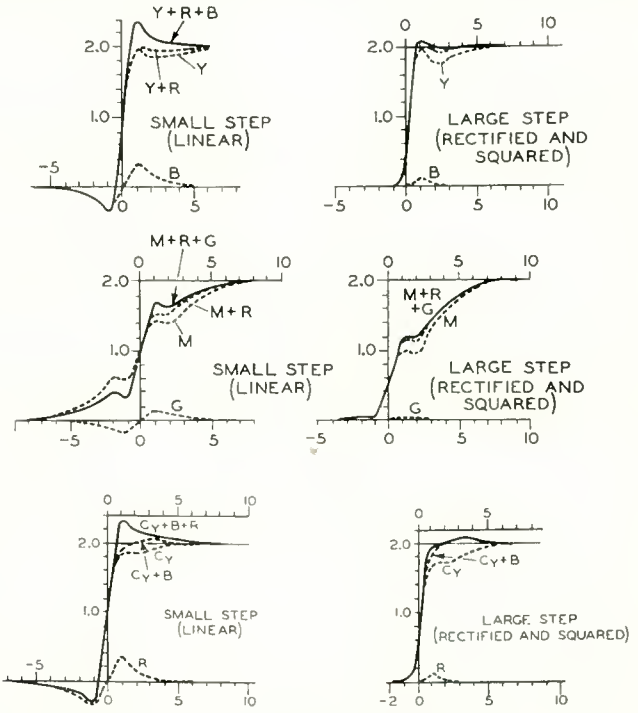


Fig. 22—Relative intensity of yellow, magenta and cyan step-function images.

as simply related to the wavelength function of its luminance channel as they are in a photographic or color television system, because the eye's luminance system continues to operate at light intensities far below the threshold for color vision.

It has been demonstrated that the eye can see all of the calculated defects in the color detail reproduction of the NTSC system, including those in low luminance colors, such as blue, for which the eye's "acuity"

[5] J. W. T. Walsh, *Photometry*, Constable, London, England, 1926, p. 67.
[6] L. A. Jones, "The Fundamental Scale of Pure Hue and Retinal Sensibility to Hue Differences," *Jour. Opt. Soc. Amer.*, Vol. 1, p. 63, March-May, 1917; W. D. Wright, *Researches on Normal and Defective Color Vision*, Kimpton, London, England, 1946.

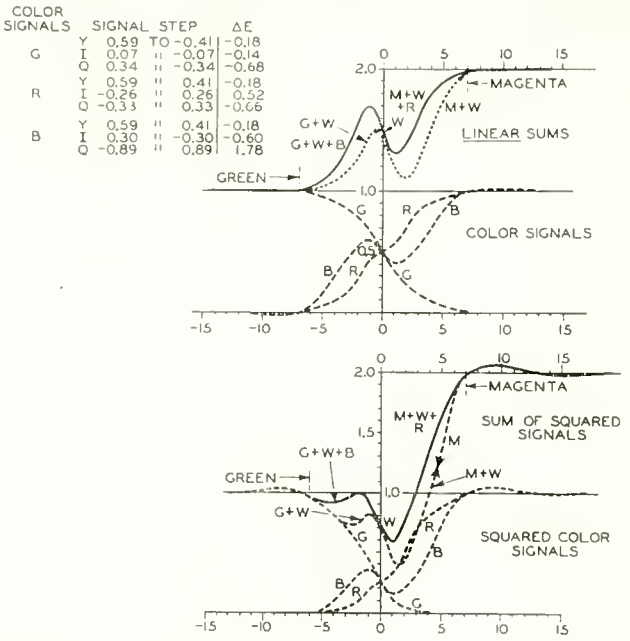


Fig. 23—Relative intensity of green to magenta step-function image.

is low. It is important to understand clearly that acuity and resolving power are not determined by the spatial sine-wave response alone, because they are strongly dependent on the noise level as well. The

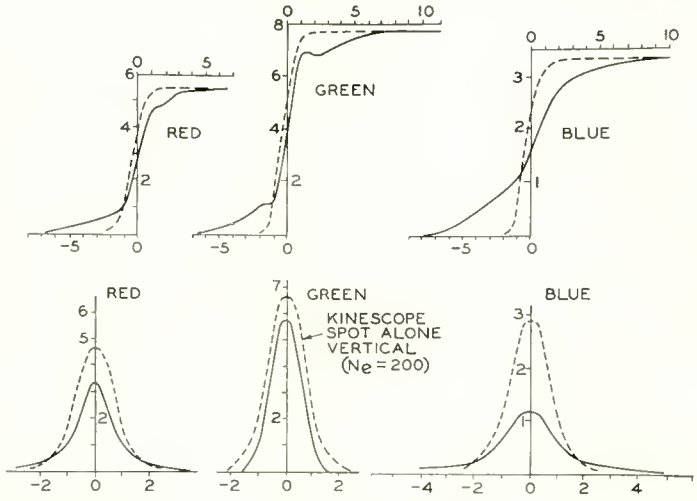


Fig. 24—Step functions and line images (NTSC) in munsell lightness values. Solid curves horizontal, broken line curves vertical coordinate.

remarkable visual constancy of color mixtures over a large part of the photopic range may be taken as an indication that the transfer characteristics, spatial sine-wave response characteristics, and noise levels of the eye's color discriminating system are well matched. Similar noise levels in turn indicate that the "discriminators" receive similar signal levels, or more likely, similar signal differences from the photo receptors in at least three principal regions of the visual spectrum. Their spatial sine-wave response appears to be well matched, because there is no evidence to the contrary in their performance (no color fringes, transients, or "smears").

It was therefore decided to determine the optical passbands of the visual color discriminators by sine-wave response measurements, following a procedure previously described for the luminance channel of the eye.[7]

Sine-Wave Response Functions

There have been general observations to the effect that resolution and sharpness of images (in microscopy, for example) do not change appreciably with blue, green, or red illumination, and that the readability of colored print of equal luminance is fairly independent of color with red slightly better than green or blue. These conclusions are not convincing, because they relate to an interpretation of signal combinations (depending on noise level and sine-wave response) from the luminance as well as the color mechanism. The same criticism can in part be applied to the single-color sine-wave response characteristics, Figure 25, measured by the author on an observer having normal vision. The response of a luminance system can, however, be eliminated by the use of *equi-luminous* sine-wave objects, which provide a sinusoidal variation of excitation purity at constant luminance. A red purity change, for example, is represented in the red color plane (Figure 5) by a sinusoidal modulation of the radius length (horizontal distance), which can be centered near the white axis (near-white background) or may be moved out towards a higher purity point. The generation of such sine-wave patterns on a color kinescope screen is relatively simple, as discussed in the Appendix.

Sine-wave patterns of constant luminance were set up by measuring the sine-wave luminance distribution with a luminance meter,* and

[7] Otto H. Schade, "Optical and Photoelectric Analog of the Eye," *Jour. Opt. Soc. Amer.*, Vol. 46, p. 721, September, 1956.

* Weston Foot-Lambert Meter, New Candle Model 931 with Viscor Filter, corrected for visual response. The meter is designed for kinescope measurements and gave correct luminance readings on the phosphor primaries and their mixtures. The readings agreed with comparative readings taken with a Macbeth Illuminometer.

adjusting the component intensity for a constant reading. It was observed immediately that luminance does not have a constant relation to visual brightness. The adjustment for constant luminance (by calculation or meter) differed from the incremental mixture values required for minimum flicker and constant visual brightness, particularly for equiluminous blue purity sine-waves in a near white background where the increment-component ratio was in error by several hundred per cent. The flicker test for equal brightness was made by reducing the temporal (electrical) sine-wave modulation frequency to 20 or 25 cycles and adjusting, for example, the blue component (see

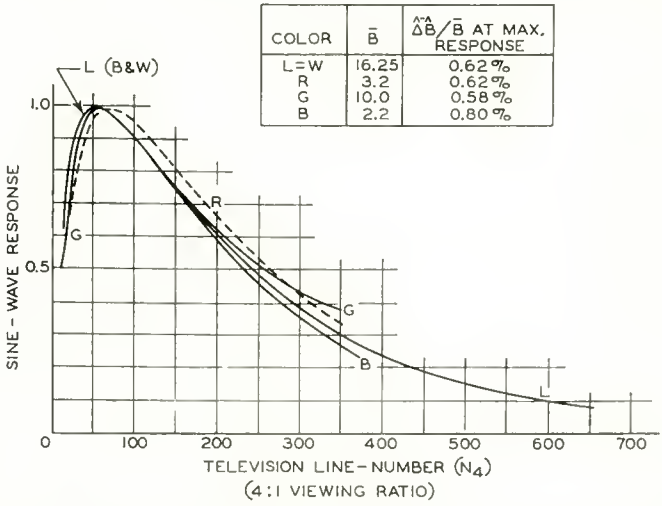


Fig. 25—Response of visual system to single-color sine-wave patterns.

Appendix) for minimum flicker. Because of the possibility of errors arising from the difference in phosphor decay time for different kinescope primaries, the brightness equality of the positive and negative purity half waves was judged visually at a higher modulation frequency producing a wavelength of about 2 centimeters on the kinescope screen. It was found to agree with the flicker test. When such a stationary wave pattern is observed while one component is varied, a sudden phase shift seems to occur in the sine-wave position when constant brightness is reached.

The contant-brightness condition near the white point requires more nearly* equal stimulus energy than equiluminance of the com-

* This might be expected from a discriminator near its neutral point. A more extended quantitative study of incremental luminance and brightness ratios in sine-wave objects would be of value. The author is aware of many publications (in the *Journal of the Optical Society of America* and elsewhere) discussing discrepancies in the total luminance of a color from standard sums.

ponents as most strikingly evident from blue purity modulation tests. It approaches equiluminance when the operating point (background color) is moved towards the spectrum locus, which is to be expected because the luminosity curve is determined from spectrum colors by brightness equality tests. This observation may well explain observed discrepancies in color noise visibility with the NTSC color system, because *equal luminance increments do not guarantee equal increments in visual brightness.*

At the time of this writing only one set of sine-wave response-measurements taken near the white axis ($B_W = 7$ foot-lamberts) is available. The designations red, green, blue, in Figure 26, refer to the color plane in which the purity modulation occurs, although the color in the areas of reduced "negative" purity (negative half cycles of sine wave) appears to the eye as a complementary color in com-

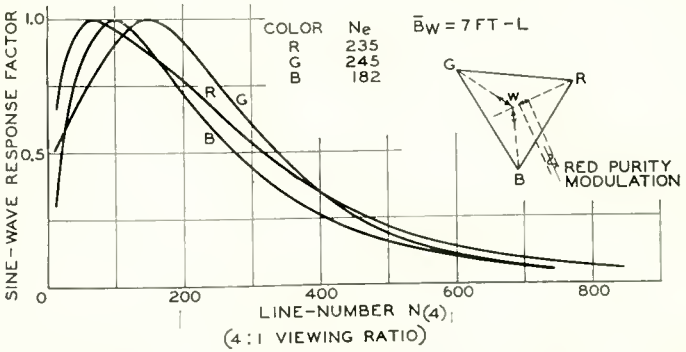


Fig. 26—Response of visual color discriminator system to constant brightness sine-wave patterns (purity modulation).

parison with the color of adjacent areas of higher "positive" purity, even at a larger distance from the white point. (The negative blue bar appears yellow, negative green appears magenta, etc.) Lacking a luminance change, the bars appear to have a constant purity as if produced by a square-wave rather than a sine-wave modulation.

The measured sine-wave response functions of the color discriminator elements in the visual system differ from one another by much smaller factors than the "standard" luminance values. All show a loss of low-frequency and d-c response as in the luminance channel. An analog of the visual system[7] contains, therefore, a high-pass filter or a negative frequency-limited feedback loop which superimposes a negative image of low definition on the normal image, "inhibiting" the normal low-pass filter response by addition of a negative response as reconstructed in Figure 27 (see Appendix). The feedback is least for red, and largest for green light, with blue intermediate. More data are

required to support a more quantitative analysis of the details of these results.

Equivalent Passbands and Number of Receptors in Effective Retinal Sampling Areas

The equivalent passbands, $N_{e(+)}$, of the main low-pass characteristics give information on the *effective discriminator areas in the retina* and appear to be in general agreement with Polyak's anatomically well-founded deductions[8] that the red end of the color spectrum is signaled

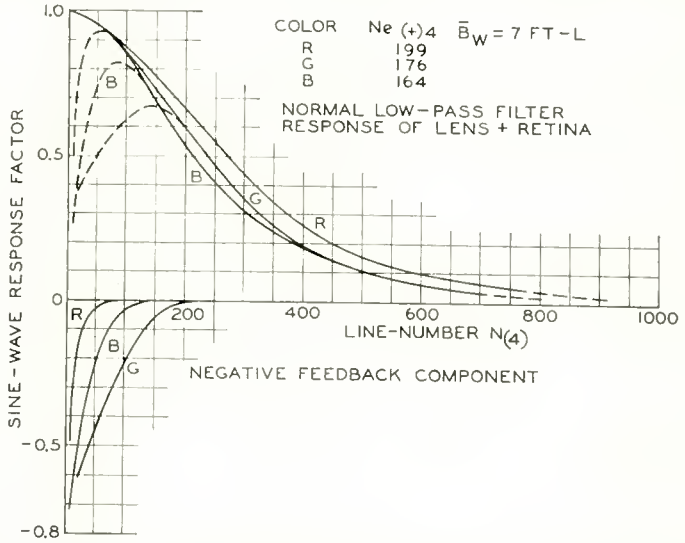


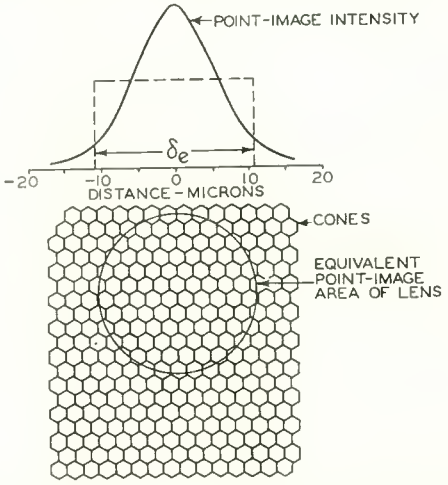
Fig. 27—Positive and negative sine-wave spectrum components of visual color discriminator system.

over direct "private line" connections between optic nerve fibers and cones and that it has the best detail response, while green and blue signals involve more diffuse matrix networks in the order stated.

The effective sampling area a (area of equivalent constant intensity point-image) can be computed from the equivalent passband by $a = 1.16/N_e^2$ (see Reference (3)). It has been shown in Reference (7) that at a luminance of 7 foot-lamberts the equivalent passband of the eye's lens is $N_{e(L)} = 50$ television lines per millimeter on the retina. The effective sampling area a_L of the optical point image produced by the lens of the eye or the retina includes therefore 88 2.2-micron cones at this illumination as illustrated in Figure 28. The normal positive

[8] S. L. Polyak, M.D., *The Retina*, University of Chicago Press, Chicago, Ill., 1941.

sine-wave response function (Figure 27) of the visual system is the product of the sine-wave response function of the lens and the sine-wave response function of the color discriminator neuron system. Since the equivalent area of the final point-image in a cascaded system is simply the sum of the equivalent areas of its cascaded components,



COLOR DISCRIM.	EQUIV. PASSBAND N_e, (L/MM) OF RETINAL AREA	DIA. δ_e OF EQUIV. AREA (μ)	NUMBER OF 2.2μ CONES
R	135	8.0	12
G	75	14.4	39
B	61	17.7	59
POINT-IMAGE OF LENS: 50 AT $\bar{B}$ = 7 FT-L		21.6	88

Fig. 28—Diameter δ_e (microns) of equivalent round sampling area of optical point-image and color discriminators of visual system relative to the diameter (2.2μ) of the photo-receptors (cones).

the equivalent areas, a_d, of the color discriminator system are obtained as the difference

$$a_d = a_m - a_L,$$

where a_d = effective area of discriminator neuron system, $a_m = 1.16/(N_{e(+)4}/4.25)^2$ = equivalent area of entire system,* and $a_L = 1.16/N_{eL}^2$ = 1.16/2500 square millimeters = equivalent sampling area of lens. The diameters obtained for equivalent circular sampling areas a_d of

* The equivalent passband $N_{e(+)4}$ of the over-all sine-wave function (Figure 27) is measured at an object distance equal to four length units (indicated by the subscript 4); since the effective distance to the retinal image is 17 millimeters (effective focal length of lens) this length unit corresponds to 17/4 = 4.25 millimeters on the retina, requiring division by 4.25 to obtain lines per millimeter on the retina.

the color discriminator neuron systems are listed in Figure 28.* The red discriminator receives signal contributions from approximately 12 cones, while the more diffuse green and blue discriminators collect signals from equivalent areas containing about 39 and 59 cones respectively. The intensity distribution in the actual point-image of the lens and the chromatic sampling areas is approximately gaussian as illustrated in Figure 28, and the actual areas include a larger number of cones with partial signal contributions. The relatively large number of cones associated with a single point-image area does give a logical explanation to the known fact that *the retina contains a much smaller number of matrix units (bipolar and ganglion cells) than cones, and that this ratio is adequate to derive color signals from the receptors, with only minor effects on the over-all passbands.*

System Concepts Unresolved

The complexity of the visual system, much of it unexplained, leaves room for much variety with regards to its possible functioning as a color system, and many different theories can be found in the literature.** The number of possible systems, however, is reduced by the high "quantum efficiency" of the eye[7,11] which does not seem to agree with the systems of area-sharing receptors, where each covers only a narrow region of the visible spectrum. Given supersensitive receptors, even slight dissimilarities in three spectral response characteristics similar to a luminosity function are quite sufficient to produce excellent color discriminators by signal subtraction and to retain full luminance sensitivity by addition. Like the eye, such a system will continue to "see" images in monochrome when the scene illumination becomes insufficient for operation of its discriminators.†

The nonlinearity and the matrix system of the retina allow for

* The signal integration computed from the measured sine-wave spectra agrees in order of magnitude with the number of synaptic connections from optic nerves over specialized types of ganglion and bipolar cells to the cones in Reference (8).

** Excellent discussions are given in References (9) and (10).

† At still lower illumination (scotopic range), cone vision ceases and only the rod system remains in operation.

[9] Gunnar Svaetichin and Rune Jonasson, I. A Technique for Oscillographic Recording of Spectral Response Curves; II. Spectral Response Curves from Single Cones; III. Receptor Mechanisms for Flicker and Fusion; IV. Notes on the ERG Analysis; V. The Cone Function Related to the Activity of Retinal Neurons; VI. Aspects on Human Photoreceptor Mechanisms," *Acta Physiologica Scandinavica*, Vol. 39, Supplementum 134, Stockholm, 1956.

[10] S. A. Talbot, "Recent Concepts of Retinal Color Mechanism," *Jour. Opt. Soc. Amer.*, Vol. 41, p. 895, December, 1951.

[11] A. Rose. "Quantum Effects in Human Vision," *Advances in Biological and Medical Physics*, Vol. V, Academic Press Inc., New York, 1957.

variation of amplitudes and weighting of respective gains (negative feedback) in combination with the luminance signal, so that the "computer" (the brain) assigns less importance to signals from dark-colored objects in a bright surrounding than it does to small-component signals in a color mixture, or when the general "lightness" of all colors in the viewing field is more balanced. A similar mechanism is effective for pure neutral-luminance signals, where lower-intensity signals are suppressed near strong ones, to enhance contrast by creating a subjective black-level. In such cases one could replace the low-intensity surrounding of an object by a black surrounding without introduction of error, but this does obviously not permit the omission of all low-intensity signals. For the same reason one cannot replace blue, or for that matter any dark-appearing color, by a dark neutral or black except under specific conditions.

Choice of Color Axes in External Systems Having Unequal Passbands

It follows from the discussion of the visual system that, given three unequal passbands for an external color-reproducing system, a preferred assignment to particular colors cannot be made on the basis of the eye's spatial frequency response, since there are apparently no major inequalities between color and luminance passbands in the visual system. One must look towards other characteristics, such as the chromatic aberration* of its lens, the lightness weighting, and the subjective black-level (feedback) mechanism, which, in a comparison, cause it to attach less importance to "dark" colors, thereby permitting errors in their reproduction to be noticed less frequently.

In the NTSC system the translation of the camera primaries R, G, B into the transmission primaries Y, I, Q containing a luminance channel (Y) was essential to establish compatibility with monochrome receivers. Matrix I also assigns a percentage of the full video passband to each color. In view of the above analysis it appears that a somewhat more panchromatic distribution of these percentages (higher blue content in Y) could be of some advantage.

Having assigned the normal wide passband to the luminance channel of the NTSC color system, there will be no inequality of optical passbands in the reproduction of black-and-white objects. The axes for the remaining two theoretically nonluminous** color signals (I, Q) must

* The axial chromatic aberration (*Jour. Opt. Soc. Amer.*, Vol. 47, No. 6, 1947) of the human eye is approximately 2 diopters from 400 to 550 millimicrons and less than 1 diopter from 550-700 millimicrons.

** Because of the nonlinearity in an actual system, the I and Q signals cause small luminance components,

pass through the white point, and thus affect pairs of complementary colors, such as red and blue-green, orange and cyan, yellow and blue, yellow-green and purple, green and magenta. A balanced psychological weighting condition might be indicated by letting the Q component be larger for high-luminance colors (which have a large wide-band Y component), but except for the red and blue-green color pair, all pairs contain one higher and one lower luminance color. The choice of I- and Q-axis directions is thus not very critical. The two color axes should include in general a large angle and the Q axis should not come too close to the low-luminance blue primary on one side, and point towards the high-luminance green region on the other side of the white point. These considerations place the I axis in the orange-cyan region, helping the medium-luminance red. The NTSC system axes have these locations (See Figure 4).

Conclusions

The above analysis as well as visual tests indicate that the bandwidth restriction in the I and Q channels causes small red-colored objects to lose some intensity and blur slightly (relative to white); green- and yellow-colored objects are sharp but lose some saturation, while small blue-colored objects have good saturation but lose intensity and sharpness. Because these effects are smaller in high-luminance colors, disappear towards white, and occur in one (horizontal) direction only, the degradation in a normal two-dimensional color picture transmitted by the NTSC system is relatively small in comparison with a color transmission over three equal independent channels as illustrated by plates VIII and IX. The relatively uncritical nature of the bandwidth assignments can be appreciated by deliberately disturbing the luminance weighting by an interchange of red, blue, or green signal connections at the input to Matrix I, making the corresponding change after Matrix II, and observing the results visually at the kinescope as illustrated by the photographs plates IX to XI. The reversal of luminance weightings for green and blue by interchange of matrix connections (plate X) can be detected only by the appearance of a transient (following the right side edge of the hat). The reversal of luminance weightings for red and green by interchange of matrix connections (plate XI) give no detectable effect for this subject as compared to the normal reproduction in plate IX. Critical observation of test pattern slides reveals that the defects caused by unequal passbands are readily visible and change to different colors when the above interchanges are made, and that there is a preference towards the NTSC choice. In a direct demonstration on a kinescope, most observers

(engineers) were unable to recognize the equal-bandwidth condition or the interchange of matrix connections when shown a variety of outdoor and indoor color pictures (SMPTE test series and many others).

The analysis has shown that commercial color kinescopes have a color contrast range resulting in a color space which is larger than that of commercial color motion pictures, that color reproduction errors caused by the unequal passbands of the system in fine detail are relatively small, and that they are considerably reduced by rectification in the nonlinear kinescope. It has also been shown that the spurious signals generated in the practical separation of chrominance and luminance signals from the composite (NTSC) signal can be minimized by proper choice of amplitude response in the various band-limiting filters at the transmitter and receiver, and result in a relatively small loss of bandwidth and picture sharpness as illustrated by plates XII and XIII.

The relative sharpness is well reproduced in the color plates. (Note the slight horizontal edge transient in the NTSC reproduction, plate XII.) The color saturation, however, is considerably reduced by the printing process.

Regarding the performance of commercially available color receivers, it can be stated that the contrast of the color tube is as good as that of the tube used in these tests. Pictures received from commercial color broadcasts can be, and on many occasions have been, as good as those observed in these tests. However, with color even more than with black-and-white reception, performance depends on proper adjustments of the receiver controls and, of course, the ambient light level. The fine detail monochrome performance should be close to the values given (again assuming correct adjustments), but the horizontal color detail is not quite as good, because it is present practice to use a common intermediate bandwidth for both I and Q channels in the receiver.

Appendix

Generation of Equi-Luminous and Constant-Brightness Sine Waves

A white background is set up first by adjustment of d-c components. The color gun (or guns) producing the desired hue, red for example, is then modulated with a sine-wave signal to generate a vertical sine-wave field, and the guns producing a complementary signal (white minus red = blue plus green) are simultaneously modulated by negative sine-

Plates I-IV—Ektachrome performance in recording constant luminance images. Plate I, normal chroma, normal exposure; Plate II, 2× chroma, normal exposure; Plate III, normal chroma at ½ exposure; Plate IV, 2× chroma at ½ exposure.

Plates V, VI—Monochrome test pattern reproductions by NTSC system. Plate V, left, 4-megacycle flat Y-channel input, demodulators off (no crosstalk); Plate V, right, 4-megacycle flat Y-channel input, demodulators on; Plate VI, left, 4-megacycle gradual roll-off Y channel, demodulators on, subsequent re-emphasis; Plate VI, right, 3-megacycle flat Y channel, demodulators on, normal Y response in receiver.

Plate VII—Sample of three line-object reproductions, 4-megacycle Y channel and standard NTSC I and Q channels; pure color signals obtained from a white pattern by turning off unwanted color channels in camera.

Plates VIII-XI—Effects of unequal passbands and luminance-weighting on kinescope image. Plate VIII, three independent 4-megacycle channels (R, G, B); Plate IX, $Y + I + Q$ signals, standard NTSC matrix and filters as in Figure 1; Plate X, $Y + I + Q$ signals and filters, but G and B reversed to Matrix I and from Matrix II; Plate XI, $Y + I + Q$ signals and filters, but R and G reversed to Matrix I and from Matrix II.

Plate XII—Complete NTSC signal with standard matrix and filters, including modulators, multiplexing, and demodulator circuits.

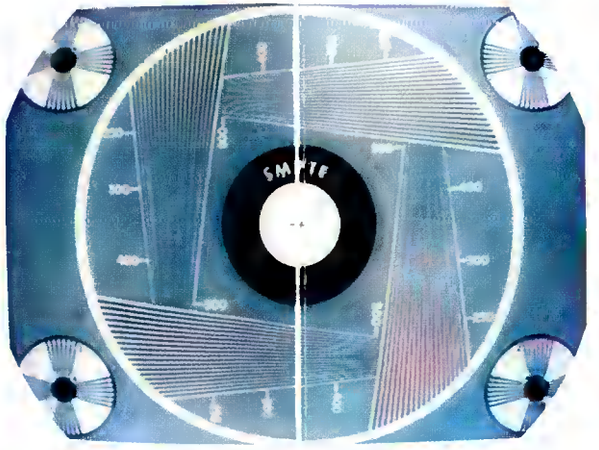
Plate XIII—Three independent 4-megacycle channels (R, G, B).

I

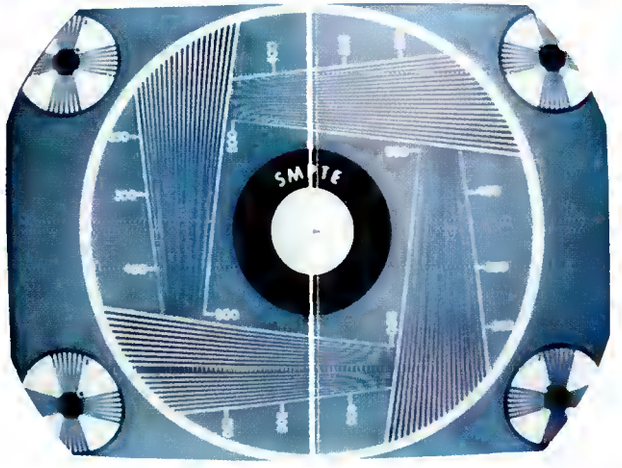
II

III

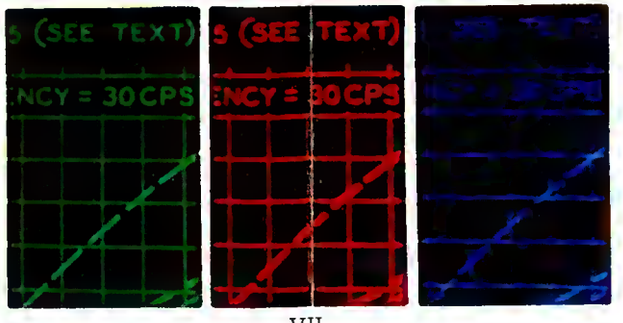
IV



V



VI



VII

VIII

IX

X

XI

XII

XIII

wave signals (opposite phase) of the same frequency.* To adjust the proper mixture, all guns are first modulated in phase to produce a "white" sine-wave, after which the red gun signal is reversed in phase and adjusted in amplitude for constant luminance or constant visual brightness of the pattern. It is observed that this adjustment varies with the background color (horizontal distance from the white-axis) which is, therefore, adjusted first by change of d-c components.

The adjustment for constant brightness (red sine-wave amplitude) must be made for the actual observer to make his luminance mechanism inoperative, because the values (Y) computed from the CIE "standard observer" curve are in error near the white axis as mentioned. This error is surprisingly large for blue (7 different observers).

Measurement of Sine-Wave Response Functions and Interpretation of Data

The observer views, from a fixed distance, a uniformly illuminated field on which an extended sine-wave bar pattern is faded in slowly by a master control ahead of the potential dividers for the component ...ls. The pattern becomes visible when the brain receives a just-...ptible signal. A plot of the required optical input signal modula-...ion is shown in Figure 29. The operating point in the color plane ...ear the white point and had a total luminance indicated in Figure ... $\bar{B}_W = 7$ foot-lamberts. Since the modulation is given in relative ... of the color under test and relative stimulus values are close to ...lues $w_R : w_G : w_B = 1{:}0.92{:}1.57$ of the white point, the minimum ...tible stimulus increments of the three colors differ by less than ...r of two from one another at any spatial frequency. The normal-...iprocal values of the sine-wave input signals plotted as a func-... spatial frequency, f_s, or line number, $N = 2f_s$, furnish the sine-...mplitude response of the visual system (Figure 26) provided ...eshold criterion is independent of frequency. Otherwise a cor-... is needed. Upon examination one finds that the visual threshold ...e limited and that detection of any object area requires a certain ... signal-to-noise ratio. Because the area of sine-wave bars of ... length and one half cycle width decreases in inverse proportion ... line number, signal-to-noise ratio, R_0, and sensitivity should ...ortional to $1/\sqrt{N}$, as illustrated by curve 1 in Figure 30. It is ...owever that the eye must vibrate (ocular tremor) in order to

* ...noninterlaced raster of 500 lines is used to obtain a direct propor-... of optical and electrical frequencies (see Reference (12)).

... to H. Schade, "A Method of Measuring the Optical Sine-Wave ... Spectrum of Television Image Display Devices," *Jour. S.M.P.T.E.*, ..., p. 561, September, 1958.

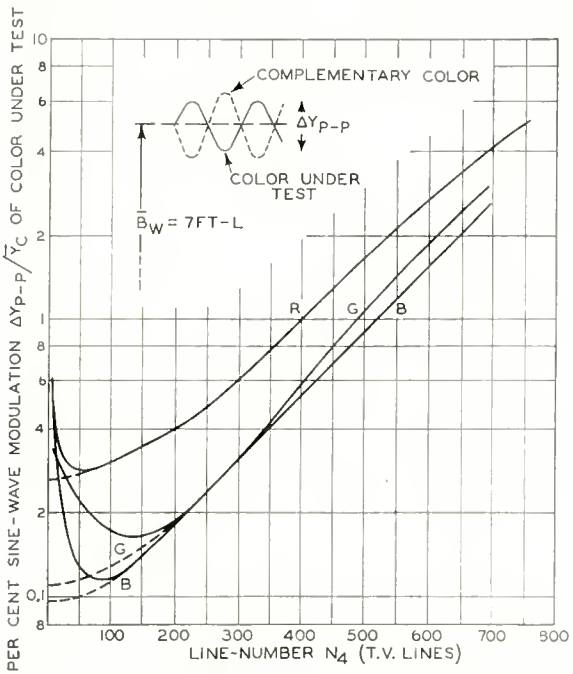


Fig. 29—Optical modulation intensities required for threshold vision (corrected for kinescope response).

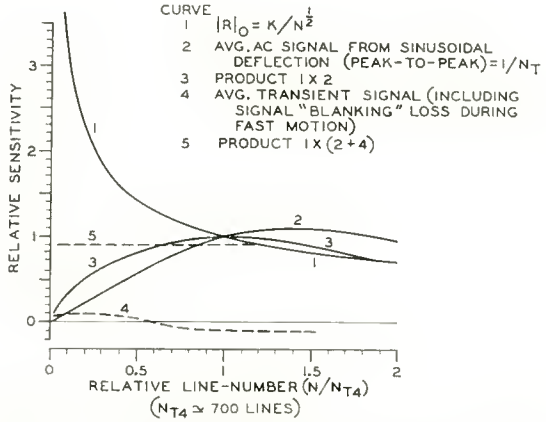


Fig. 30—Effect of signal-to-noise ratio, ocular tremor, and transient motion on threshold sensitivity.

see detail; i.e., it must generate a-c signals for transmission to the brain (its analog contains a blocking capacitor). The deflection amplitude of the tremor is quite small, 6 microns on the average,[13] which does not degrade the image, but is sufficient to generate strong a-c signals at line numbers where the half cycle length, $1/N$, of the retinal sine-wave image is in the order of the deflection amplitude. The a-c signal, however, decreases towards lower line numbers and goes to zero at $N = 0$. The average a-c signal developed at different points on the retina due to ocular tremor is, therefore, a function of line number, and calculations indicate the function shown by curve 2 in Figure 30. The deflection amplitude is specified by its reciprocal value, expressed as a television line number $N_{T(4)}$, for an object distance equal to four length units.*

The action of the tremor counteracts the sensitivity change due to the varying signal-to-noise ratio as shown by the product curve 3, but overcompensates at low frequencies. This deficiency is decreased by a partial d-c response[7] and by the continuous jerky motion of the eyeball when a larger image is observed. One becomes conscious of this motion and its effect in generating a-c signals when a barely visible large object or pattern is viewed, because the pattern fades out when fixed steadily. The signal generation in this case is a series of detached transients (a continuous large deflection would blur the image). A transient image appears after the sudden motion stops, just as in tests where ocular tremor and image motion are prevented artificially and transient vision is obtained after a change of image content. The dashed curves in Figure 30 show that a constant threshold requires only a relatively small increase in a-c signal at the low-frequency end of the response curve; this is obtainable by intermittent transient excitation.

In view of these facts, it appears justified to assume a constant threshold criterion in the upper portion of the sine-wave response characteristic of the eye which is obtained directly from the measured data. The lower portion is less well defined, but indicates that an analog system contains a high-pass filter in cascade with the normal low-pass filter representing the optic and retinal structure of the eye. The low-frequency section of the low-pass filter response can be extrapolated by matching the high-frequency section to a possible lens plus grain structure response characteristic (see Figure 27).

---

[13] Lorrin A. Riggs, "The Measurement of Normal Ocular Tremor by Corneal Reflection," *Jour. Opt. Soc. Amer.*, Vol. 42, p. 287, April, 1952; G. C. Higgins and K. F. Stults, "The Frequency and Amplitude of Ocular Tremor," *Jour. Opt. Soc. Amer.*, Vol. 42, p. 872, November, 1952.

* The reciprocal distance on the retina in television lines/mm is given by $N/mm = N_{(4)}/4.25$.

THE EFFECT OF RADIATION ON SILICON SOLAR-ENERGY CONVERTERS*†

By

J. J. Loferski and P. Rappaport

RCA Laboratories,
Princeton, N. J.

Summary—*The performance of silicon solar cells under simultaneous illumination and irradiation by various ionizing radiations has been observed for the purpose of estimating the useful life of such power sources in the environment of the I.G.Y. (International Geophysical Year) earth satellite. Experiments were conducted both in air and in vacuum for 2 mev electrons, 20 mev protons and 40 mev alpha particles. The maximum power, open-circuit voltage and short-circuit current experienced a decay. These results have been extrapolated to yield an estimated minimum of about 10^5 years until the output of such cells on a satellite drops to 75 per cent of the initial value.*

Possible modifications of these estimates based on tentative data acquired from satellite flights have been considered.

The decay of the cells is associated with changes in minority-carrier lifetime and deviations from expected behavior are discussed. No definite decay was observed during the limited exposure time of this study for the ultraviolet radiation between 2200Å and 3400Å or for x-rays from machines operated at 50-2,000 kilovolts.

Introduction

THE PURPOSE of this study was to investigate the effects of cosmic and solar radiations likely to be encountered in the upper atmosphere on the performance of silicon p-n junction solar energy converters. Specifically, experimental study was made of the effects on such cells of ultraviolet light, x-rays, gamma rays, electrons, protons and alpha particles both in air and in a vacuum. Radiation densities comparable to, or greater than, those expected outside the earth's atmosphere were used to ascertain at what levels they would produce detrimental effects on the electrical characteristics of the cells, such as short-circuit current (i_s), maximum voltage (V_{max}), maximum power (P_{max}), and maximum solar energy conversion efficiency (η_{max}). Experimental results are compared with theory.

Throughout these experiments the cells were simultaneously illuminated by a light source, whose power density approximated sunlight

* Manuscript received September 4, 1958.
† This work supported by the United States Army Signal Research and Development Laboratory Contract #DA36-039-SC-64643.

(100 milliwatts per square centimeter), and irradiated by the potentially detrimental radiation. The irradiated cells were kept within the temperature range of 25°C to 30°C because it is anticipated that satellite temperatures will have an equilibrium value within 15°C of the launching temperature.

TABLE I

COMPONENTS OF RADIATION IN THE UPPER ATMOSPHERE AND THE EXPERIMENTAL RADIATIONS SELECTED

UPPER ATMOSPHERE			EXPERIMENTAL	
RADIATION	ENERGY OR WAVELENGTH	AMOUNT	ENERGY OR WAVELENGTH	AMOUNT AND SOURCE
ULTRAVIOLET	75 - 2200 Å 2200 - 3400 Å	6×10^{-4} mw/cm^2 4.8 mw/cm^2	2200 - 3400 Å	SH 9.8×10^{-3} w/cm^2 S4 173×10^{-3} w/cm^2 B - H6 2.1 w/cm^2 UV LIGHT BURNERS
X-RAYS	7 - 60 Å	10^{-4} mw/cm^2	50 kev (0.24 Å) TO 250 kev 0.75 TO 2 mev	UP TO 10^4 ROENTGENS / HR X-RAY GENERATOR
ELECTRONS (AURORAL REGIONS)	10 - 500 kev	1%	2 mev	UP TO 100 μa/cm^2 VAN DE GRAAFF
PROTONS	≥ 1 Bev, AVERAGE IN NUCLEAR BURSTS 50 Mev	80% TOTAL FLUX 1 PARTICLE/cm^2/sec	10 - 20 mev	UP TO 100 μa/cm^2 CYCLOTRON
ALPHAS		19%	40 mev	UP TO 100 μa/cm^2 CYCLOTRON

Types of Radiation

Selection of the radiation used in this study was based upon the best available information on the nature of radiation components in the upper atmosphere.[1] It was, however, necessary to make compromises because it was not possible to duplicate these radiation conditions exactly. Table I summarizes the information given in Reference (1) and also lists the actual radiations available for the study.

As result of data gathered by the I.G.Y. Satellites 1958 Alpha and Gamma, it may be necessary to revise Table I.[2] The number of counts reported by the Geiger counters on the satellites was larger by a factor of about 3×10^4 than the values found by extrapolation from

[1] "The Environment of an Earth Satellite," *SCEL Technical Memorandum*, M-1747, March, 1956.

[2] State University of Iowa Report, SUI-58-5, May, 1958, to be published.

the high-altitude-balloon data on which Table I is based. The nature of the radiation is unknown, but it appears that it may be either high-energy electrons or protons of moderate energy, i.e., much lower in energy than the 10 bev protons of cosmic rays. The implication of these new data will be discussed.

In general, as broad a spectrum of electromagnetic radiation as possible was used so that predictions could be made concerning those areas where experimental sources were not available. For example, it is known that solar x-rays and ultraviolet radiation in the wavelength range between 3 Å and 200 Å exist in the upper atmosphere. Laboratory sources in that portion of the spectrum are not available, but radiation sources of wavelengths shorter than 0.24 Å and longer than 2200 Å are available so that the region in question can be straddled.

As for the particle radiations, cosmic ray energies are in excess of 10 bev. However, the highest energies available for this study were 40 mev for alpha particles and 17 mev for protons. The results of irradiations at these lower energies were extrapolated to predict the effect of higher energy particles. Sources of radiation used during this study are as follows.

Ultraviolet Radiation

Three different commercial ultraviolet lamps were used, namely the Hanovia Type S-H burner, GE Type BH-6 burner and a GE Type S-4 burner with glass envelope removed. The intensities of radiation in the spectral region 2200 Å to 3400 Å available from these sources, together with the amount of radiation from the sun in the same spectral region, are given in Table II. The high intensity available from the BH-6 permits the equivalent of one year's irradiation to be delivered to a cell in 20 hours.

X-Rays

The source of x-rays in the range between 50 and 250 kilovolts was the Westinghouse Quadrocondex X-ray machine of the Evans Signal Laboratories, while a 2 mev Van de Graaff Electron Accelerator (High Voltage Engineering Corp.) of those same laboratories was used for the 750 kv and 2,000 kv x-ray spectra. No filters were used to modify the continuous spectrum emitted by the target. Periods of irradiation were limited to a few hours at each machine energy.

High-Energy Electrons

The 2 mev Van de Graaff machine referred to above supplied electrons at an energy of 1.7 mev. The electrons were admitted into the air of the room after passing through a 0.005-inch aluminum window.

The electron current to the samples was determined by measuring the current to a metal plate of the same area as the samples, but thick enough to stop the electrons.

Protons and Alpha Particles

Cyclotrons at Princeton University and at the Brookhaven National Laboratories were used to supply 17 mev and 10 mev protons, respectively, while 40 mev alpha particles were supplied by the Brookhaven Cyclotron. These particles entered the air of the room by passing through a 0.001-inch Dural* window. The particle fluxes used were

Table II—Comparison between Different Criteria for Estimating the Effect of X-Radiation on Surfaces

Type of Radiation	Incident Power/cm² (μw)	Power Absorbed/cm² (μw)	No. of Photons/cm² sec. (N_{ph})
Atmospheric X-rays			(Approximate)
200 ev-1200 ev	0.1	0.1	3×10^9
665-1500	0.06	0.06	6×10^7
1000-1500	3×10^{-4}	3×10^{-4}	1×10^4
1200-1700	10^{-4}-10^{-5}	10^{-4}-10^{-5}	5×10^4-10^5
Machine Produced X-rays			
50 kv	4.6	1.95×10^{-3}	5.8×10^8
(550 r) 250 kv	532.0	0.20	1.3×10^{10}
(12,000 r) 750 kv	384.0	8.63	3.2×10^9
(8000 r) 2000 kv (5×10^4 r)	3850.	53.6	4.2×10^{10}

in the range of 10^8 particles per square centimeter per second. Flux measurements were made by observing the beam back-scattered from a foil according to the following scheme: The machine current was increased to values a few orders of magnitude above those to be used for the actual bombardment. This higher current, measured by means of a Faraday cup and a conventional microammeter, was compared to the number of counts registered by a photomultiplier arranged to receive some of the scattered particles. It was established that the photomultiplier counts were proportional to the current in the Faraday cup. Thereafter the number of counts registered by the photomultiplier were used to determine the actual flux of particles. To establish how the particles are distributed over the irradiated area, copper foils were substituted for the cells and activated by the particles. The foils were then cut into small pieces and the activity in the regions occupied by

* Registered trade mark.

the solar cells was determined as a fraction of the total activity contained in the copper.

The Solar Cells

The silicon solar cells were manufactured by the Hoffman Semiconductor Division, Evanston, Illinois and have been described in the literature.[3] The particular units used were 0.5 × 2 centimeters and about 0.05 centimeter thick. Figure 1 shows a sketch of the unit. The surface of the originally n-type silicon wafer had been converted to p-type by diffusion doping with boron. The diffused layer, which is about 1 micron thick, is removed from all surfaces except the top by etching. The active volume of the cell from which generated carriers can be collected is determined by the thickness of the diffused layer

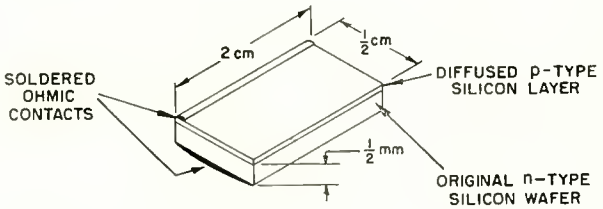


Fig. 1—Diagram of a silicon solar cell.

and the exposed area. Ohmic contacts are made to a thin top strip on the surface and the entire bottom surface by nickel plating and soldering.

At least six cells were used for each radiation type, three in air and three in a vacuum. The cells used had the highest values of η_{max} from among about 100 cells whose current–voltage characteristics had been measured before irradiation. For these cells η_{max} varied between 6 and 10 per cent. (η_{max} is defined as the ratio of maximum electrical power output, P_{max}, to solar power input as measured by an Epply pyrheliometer.)

Experimental Arrangement and Procedure

The samples were mounted on a water-cooled copper block, so oriented that the light beam (from a 150-watt projector spotlight) was normal to the beam of radiation, and both radiations were incident on the cells at an angle of 45° as shown in Figure 2. The intensity of

[3] M. B. Prince, "Silicon Solar Energy Converters," *Jour. Appl. Phys.*, Vol. 26, p. 534, May, 1955.

illumination was adjusted to deliver about 100 milliwatts per square centimeter of equivalent solar radiation to the samples. Whenever possible, constancy of the light source was assured by observing a signal from another silicon solar cell which was located well outside the radiation field. For the vacuum runs, a similar water-cooled block was suspended inside a Pyrex* vacuum chamber fitted with windows to admit both light and particles. The pressure in the vessel was maintained at about 10^{-5} mm Hg by means of a conventional oil-diffusion pump and liquid-air traps.

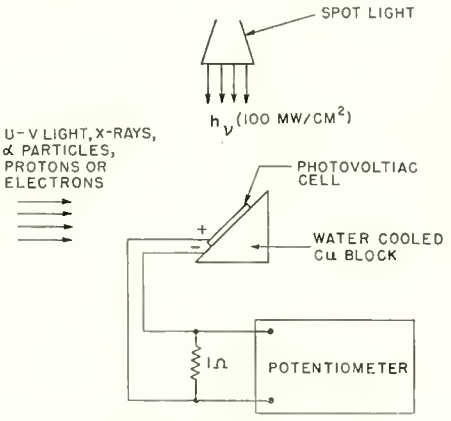


Fig. 2—Experimental arrangement and circuit for measurement.

Figure 2 also shows how electrical measurements were made on the cells. A current–voltage characteristic like that shown in Figure 3 was obtained by shunting in succession each of fifteen different precision resistors across the illuminated cell. The voltage developed across the shunt resistor was recorded on a Brown Electronik Recorder. From the resistance and voltage, V_{max}, i_s, P_{max} and η_{max} were computed. Each run took three minutes and was repeated periodically during irradiation. During some runs only three precision resistors were monitored to give V_{max}, i_s and P_{max}, thus allowing for more rapid observations during faster decays.

Experimental Results

Ultraviolet Light and X-Rays

Figures 4 and 5 show i_s, V_{max} and P_{max} as a function of irradiation time with the S-H Hanovia Lamp for three solar cells in air. It is

* Registered trade mark.

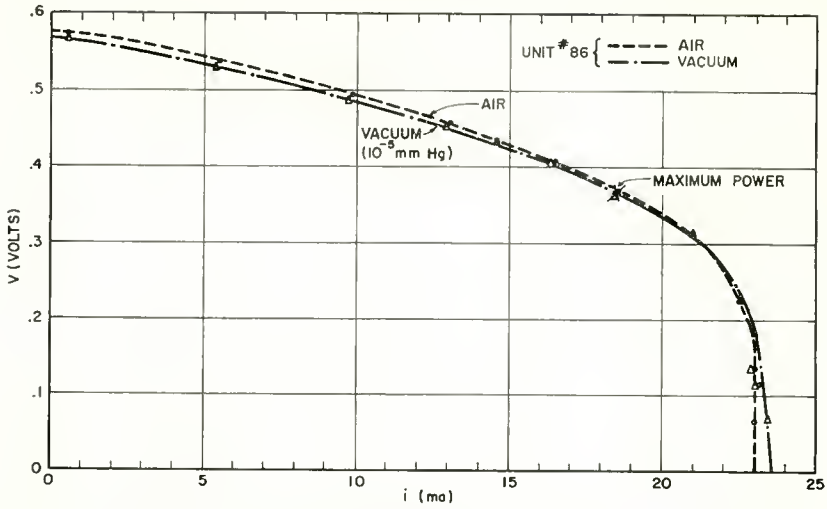


Fig. 3—Comparison of current–voltage (i–V) characteristics in air and in a 10^{-5} mm of Hg vacuum at a flux of 110 mw/cm^2.

evident from these figures that none of the three parameters changed during irradiation for a time nearly equivalent to one year in the upper atmosphere. Similar results were obtained using the S-4 burner

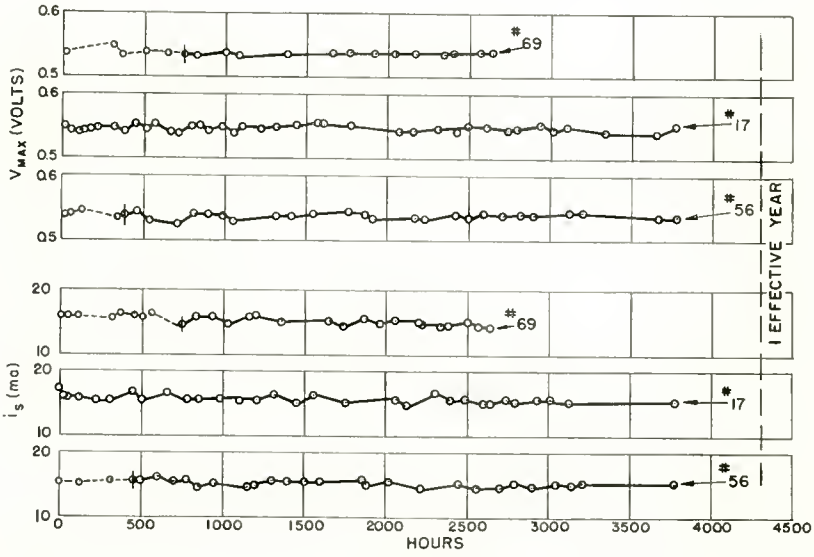


Fig. 4—i_s and V_{max} versus time for ultraviolet irradiation in air using type S-H Hanovia lamp.

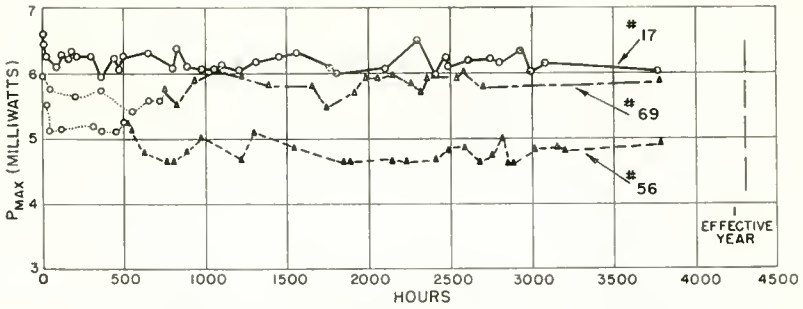


Fig. 5—P_{max} versus time for ultraviolet irradiation in air for the type S-H lamp.

on cells in vacuum; no changes were observed for periods equivalent to about 2.5 years in the satellite environment. The samples were irradiated by the BH-6 burner for a period equivalent to 50 years in the satellite environment for the spectral range where the burner overlaps the upper atmosphere ultraviolet spectrum. Figure 6 shows P_{max} versus time for this run, and indicates that a decrease equivalent to no more than 0.3 per cent per year may have occurred. The "decay" may, however, have been an experimental artifact traceable to the difficulty of maintaining constant illumination over the 1000-hour laboratory irradiation period. With this reservation in mind, assuming that the decay is real, one finds by extrapolation that 85 years would elapse before P_{max} would decay 25 per cent.

Results for irradiation at 50 and 250 kilovolts with x-rays are shown in Figures 7 and 8. These runs were made on successive days, some-

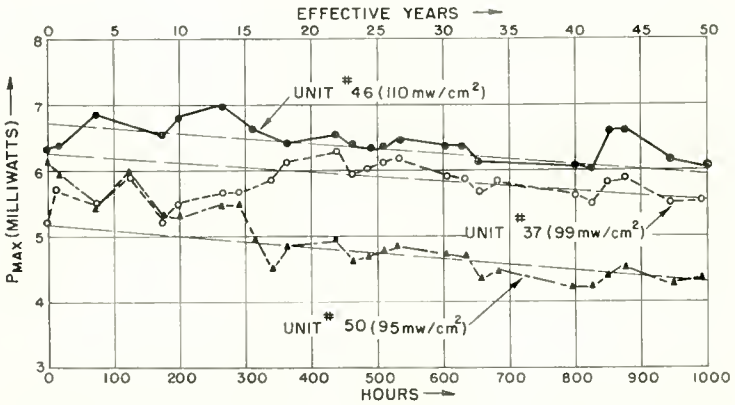


Fig. 6—P_{max} versus time for high intensity ultraviolet source.

times in vacuum and sometimes in air. As is evident from the figures, there was no monotonic change in any of the parameters; the scatter of the points in these figures is an indication of the limitations of the experimental method rather than of any changes occurring in the samples. Similar results were obtained with 750 and 2,000 kv x-rays. Thus, it is concluded that these x-rays do not affect the operation of solar cells. Table II shows both the total dose in roentgens (where applicable) and an estimated total number of incident photons, N_{ph}, received by the samples during the x-ray and ultraviolet irradiations. The quantity, N_{ph}, for the x-rays was computed by assuming that the average energy of a photon in the x-ray spectrum was ½ of the maximum photon energy.

There remains, however, the problem of inferring the life of these cells in the x-ray environment described in Table I in which the x-ray photon energy is much lower than that used for the irradiations described. Since even the 2,000 kv x-rays did not produce any discernible bulk damage it can be concluded that if the low-energy x-rays actually present in the upper atmosphere are going to produce a decay of the current–voltage characteristics of these cells, it will necessarily be the result of a surface effect. Such a surface effect could be proportional to one of the following parameters of the radiation: (a) the integrated number of photons, (b) the incident energy per square centimeter, (c) the energy absorbed near the surface per square centimeter. This latter quantity is determined by calculating the energy absorbed in the first micron of cell thickness.

Table II compares the values of these parameters for the x-radiations likely to be encountered above the atmosphere and the x-rays used in these experiments. All values are listed per unit time so that from the laboratory irradiation time one can determine directly the equivalent irradiation time in the satellite environment. As we have indicated above, P_{max} was constant during the laboratory x-irradiations within an experimental error of ± 5 per cent. To arrive at a lower limit on the life of the cells in the satellite environment, assume that they had experienced a decay of 5 per cent, which was not detected because of the limited precision of the experiments. Table III shows how much time will be required for the cells to lose 25 per cent of their initial P_{max} in the satellite environment based on these assumptions.

High-Energy Electrons

Three units were run individually at atmospheric pressure and three were run simultaneously in vacuum. The distribution of elec-

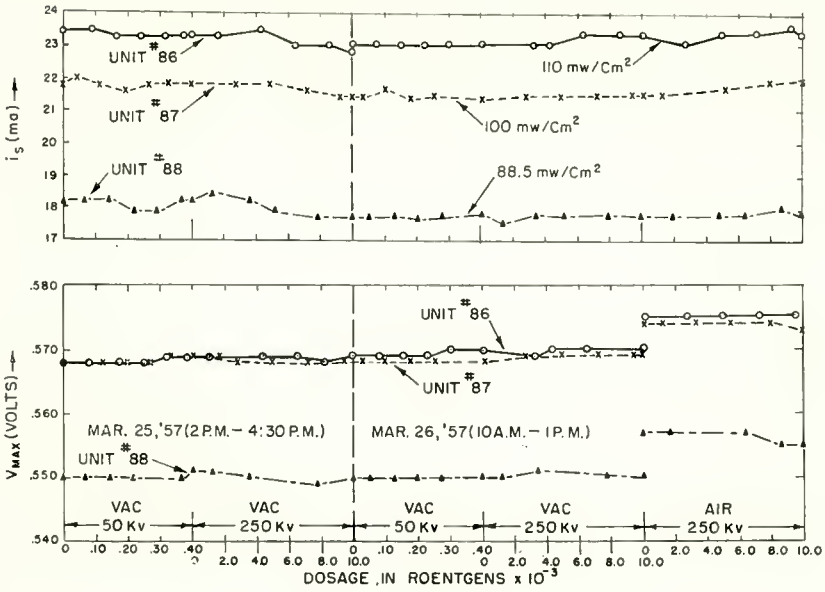


Fig. 7—i_s and V_{max} versus time for x-irradiation.

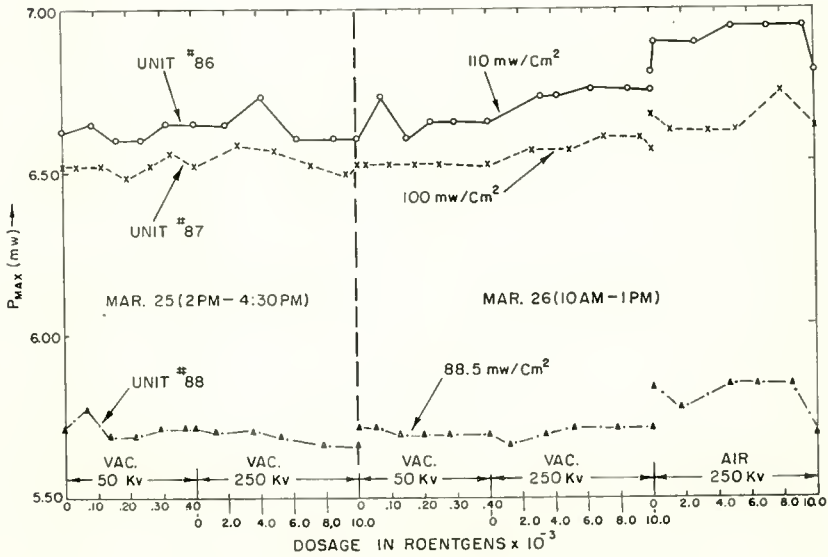


Fig. 8—P_{max} versus time for x-irradiation.

Table III—An Estimate of the Lower Limit on the Life of Cells Above the Atmosphere Based on Three Different Criteria (X-rays)

Type of Radiation	Incident Power	Power Absorbed	N_{ph}
50 kv and 250 kv	2.4×10^4 hrs.	9.0 hrs.	39 hrs.
750 kv and 2000 kv	3.8×10^5 hrs.	4.9×10^3 hrs.	70 hrs.
50 kv and 250 kv	8.1×10^4 hrs.	27 hrs.	103 hrs.

trons on the three samples in vacuum was not known accurately so that some of the difference in decay rates of these three units should be attributed to this nonuniform electron distribution. The units in air were bombarded individually hence no distribution problem existed. Figures 9 and 10 show V_{max}, i_s, and P_{max} plotted against the estimated total number of electrons incident on the sample. Table IV summarizes the results of the electron irradiation; the symbol ϕ_c represents the average flux at which the P_{max} falls to 75 per cent of its initial value. It is evident from the table that the value of ϕ_c is different for the samples irradiated in vacuum and those irradiated in air. The reason for this difference is uncertain. However, this does not affect estimates of the life of cells in the upper atmosphere, since the best available data on the electron flux indicates a value less than 10^{-2} per square centimeter per second. On this basis, P_{max} will decrease 25 per cent from its initial value after an irradiation time whose value lies between 10^{15} and 10^{17} seconds, i.e., between about 10^8 and 10^{10} years.

Protons

Two groups of three units were irradiated, the first with 10 mev protons in vacuum at Brookhaven, and the second with 17.6 mev protons in air at Princeton University. The results are shown in Figures 11 and 12 where i_s, V_{max} and P_{max} are plotted against the integrated flux, ϕ. Differences in the value of ϕ_c are partially due to the uncertainty about distribution of the flux. As with the electrons, these variations in the value of ϕ_c do not seriously affect an estimate of the elapsed time for P_{max} to be reduced by a predetermined amount in the upper atmosphere. Using the data from the run in vacuum at Princeton University, we find that on the average, about 3.5×10^{10} particles

Table IV—Flux of Particles Required to Reduce Solar Cell Power Output to 75% of Its Initial Value.

Particle Type	ϕ_c (Particles/cm^2)
1.7 mev electrons	Between 5×10^{13} (vacuum) and 1.5×10^{15} (air)
17.6 mev protons	3.5×10^{10}
40 mev α particles	4.4×10^9

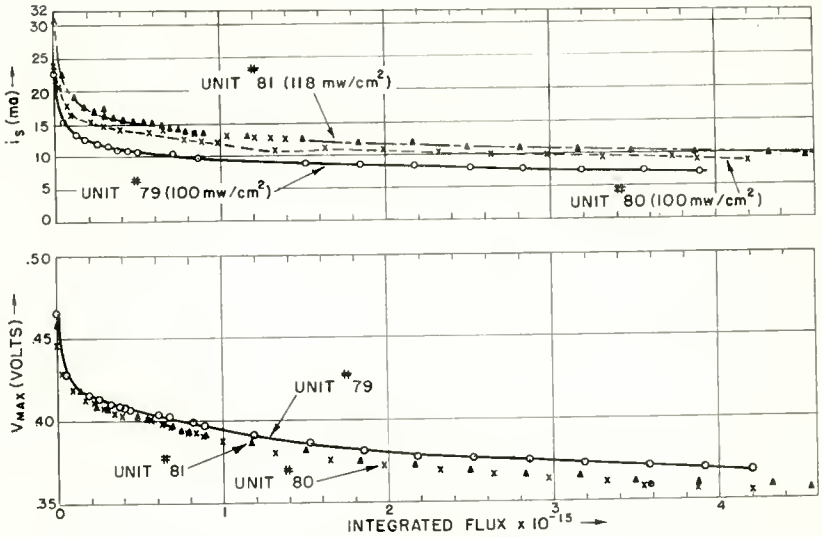


Fig. 9—i_s and V_{max} versus integrated flux for 1.7 mev electrons irradiated in vacuum.

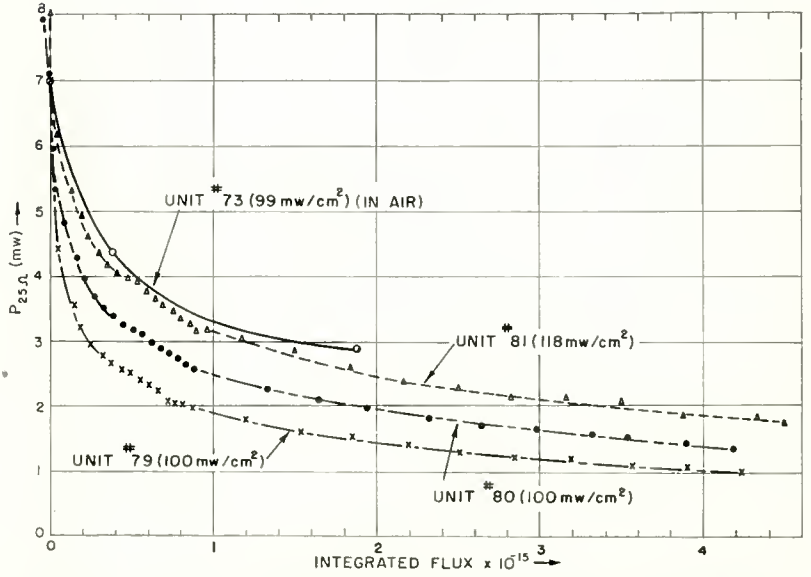


Fig. 10—P_{max} versus integrated flux for 1.7 mev electrons in vacuum.

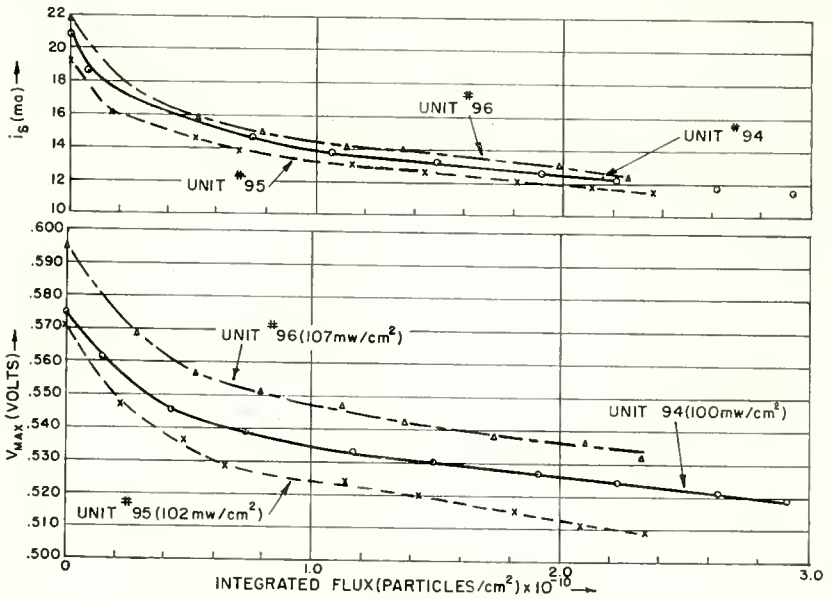


Fig. 11—i_s and V_{max} versus integrated flux of 17.6 mev protons in air.

per square centimeter are needed to produce a 25 per cent decrease in P_{max}, as recorded in Table IV.

Because of the difference in energy between the protons available

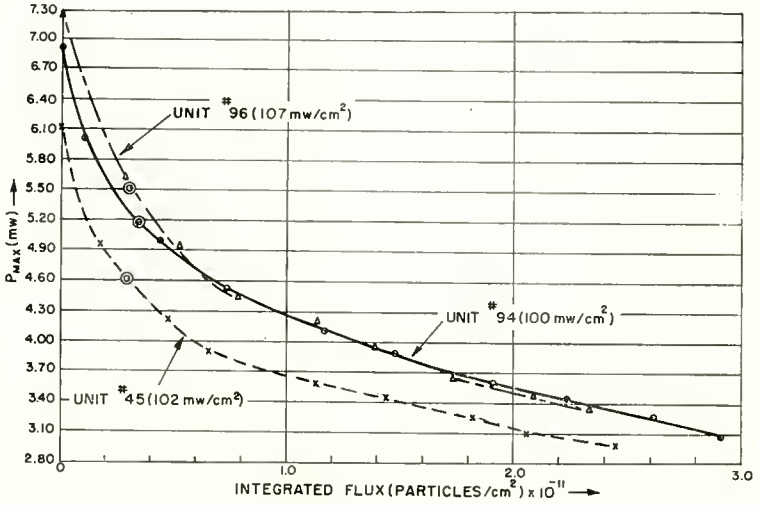


Fig. 12—P_{max} versus integrated flux of 17.6 mev protons in air.

from the cyclotrons and the high-energy cosmic ray protons, the relative numbers of defects produced by particles of different energies must be calculated. Following Seitz and Koehler,[4] note that the density of defects, N_D, produced by an integrated flux, ϕ, of high-energy particles penetrating a solid whose thickness is less than the range, i.e., a thin target, is given by

$$N_D = \bar{\nu} n \phi, \tag{1}$$

where n is the number of primary displacements per particle and $\bar{\nu}$ is the average number of secondary displacements per primary displacement. If two particles produce equal numbers of defects, it follows that

$$\bar{\nu}_1 n_1 \phi_1 = \bar{\nu}_2 n_2 \phi_2. \tag{2}$$

For a given species of particle of energy E,

$$n = \kappa/E \text{ so that}$$

$$\bar{\nu}_1 \frac{\kappa}{E_1} \phi_1 = \bar{\nu}_2 \frac{\kappa}{E_2} \phi_2 \tag{3}$$

and thus

$$\frac{\bar{\nu}_1}{E_1} \frac{E_2}{\bar{\nu}_2} = \frac{\phi_2}{\phi_1}. \tag{4}$$

Substitution into this equation shows that approximately 330 ten-bev protons must pass through a cell to produce as many defects as a single twenty-mev proton does. Since it is estimated that the proton flux in the satellite environment is about 1 particle per square centimeter per second, it follows that about 3.6×10^5 years are required to decrease $P_{\max}$ by 25 per cent as a result of bombardment by cosmic ray protons of the upper atmosphere.

Alpha Particles

Six cells were irradiated, three in air, three in vacuum. As before, some of the variation in values of ϕ_c from sample to sample is caused by uncertainties in determinations of the flux received by the sample. The average value of ϕ_c is shown in Table IV. To estimate the life of

[4] F. Seitz and J. S. Koehler, "Displacement of Atoms During Irradiation," *Solid State Physics*, Vol. 2, Academic Press, New York, N. Y., 1956, pp. 307-448.

such units in the alpha particle flux of the upper atmosphere, we combine the flux rate estimated in Table I, i.e., 0.2 alpha particles per square centimeter per second, with the average value of ϕ_c deduced from the experiments and take into account the difference in energy between the cosmic-ray alpha particles and those available during this study. When this is done, one finds that these cells should last 1.4×10^5 years before 25 per cent decay in P_{max} is reached as a result of the alpha-particle bombardment in the upper atmosphere.

Modified Estimates of Life Based On I.G.Y. Data

If the level of radiation in the satellite environment is as high as given by Van Allen based on data from the first satellites,[2] our estimates of the useful life of silicon solar cells must be revised downward. Assume, first of all, that the increased radiation is given by the lower limit deduced from the satellite, i.e., greater than previously estimated by a factor of 3×10^4. Then if the radiation consists of protons of about 20 mev, P_{max} from the cells would be reduced by 25 per cent after about 10 years in a field of this intensity. If the particles are electrons, then the high count rate is very likely to have been caused by bremstrahlung produced by these electrons bombarding the satellite shell and Geiger counter case. As has been shown, x-rays have essentially no effect on the cells. If we assume that electron current is given by the count rate, then such an arrival rate would cause a 25 per cent decrease in P_{max} after a minimum of 3×10^3 years, if nothing were interposed between the cells and the radiation field.

Correlation with Theory

Nuclear radiation can change three parameters which affect semiconductor devices—the conductivity, σ, the minority carrier lifetime, τ, and the surface recombination velocity, s. Changes in σ and τ have been attributed to a common cause, namely the production of excess vacancy-interstitial pairs associated with the displacement of atoms in the lattice by the bombarding particles. Such displacement can occur only if the energy transferred to a lattice atom exceeds a threshold energy, E_t, which has been found experimentally to be 13 ev for silicon.[5] To transfer such an amount of energy in a collision with a silicon atom, an electron must have an energy of 145 kev, while the heavier proton needs 98 ev, and the α particle 30 ev. By means of Compton scattering, x-ray photons can impart all of their energy to

[5] J. J. Loferski and P. Rappaport, "Radiation Damage in Ge and Si Detected by Carrier Lifetime Changes: Damage Thresholds," *Phys. Rev.*, Vol. 111, p. 432, July, 1958.

a single electron during an encounter, so that photons whose energy exceeds 175 kev can potentially displace silicon atoms. Earlier studies of radiation damage in electron-voltaic cells showed that changes in τ destroy the usefulness of a junction long before changes in σ become important. It was therefore expected the same would be true for these solar cells.

To provide a framework for discussion, the manner in which changes in τ would be reflected in the parameters of the photovoltaic effect will be reviewed. Figure 13 shows a cross-section of a p-n junction where the p-type region to the left of the junction, has a thickness x. Assume that there is no contribution to the photovoltaic effect from minority carrier generation in the n-type region to the right of the junction in the figure. This assumption is valid in this case, since the light which produces the effect is absorbed principally between the

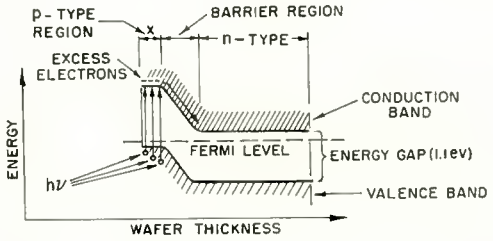


Fig. 13—Electron energy versus distance from front of cell.

surface of the material and the junction. A mathematical treatment of this model has been presented elsewhere.[6] Here it is sufficient to point out that the fraction of injected carriers which survive long enough to pass over the junction is proportional to $\tau^{1/2}$. Now τ is inversely proportional to the number of recombination centers, N_r, in the material. Such recombination centers are introduced into the forbidden energy gap by vacancy-interstitial pairs. For low irradiation fluxes it is reasonable to assume that the number of defects is proportional to the integrated flux, ϕ, so that

$$\frac{1}{\tau} \propto N_r \propto \phi = N_B t, \tag{5}$$

where N_B is the number of bombarding particles per square centimeter

[6] P. Rappaport, J. J. Loferski, and G. E. Linder, "The Electron-Voltaic Effect in Germanium and Silicon P-N Junctions," *RCA Review*, Vol. XVII, p. 100, March, 1956.

per second and t is the irradiation time. Thus, plots of τ^{-1} against t are linear for a constant flux. This relation has been verified by earlier work,[5] according to which the significant parameters of the photovoltaic effect depend on τ and therefore on ϕ in the following way:

$$i_s \propto \tau^{1/2} \propto \phi^{-1/2} \tag{6}$$

$$i_o \propto \tau^{-1/2} \propto \phi^{1/2} \tag{7}$$

$$\lambda V_{\max} = \ln \frac{i_s}{i_o} \propto \ln \tau \propto \ln \phi^{-1} \tag{8}$$

where $\lambda \equiv e/kT$ and i_o is the reverse current of the junction

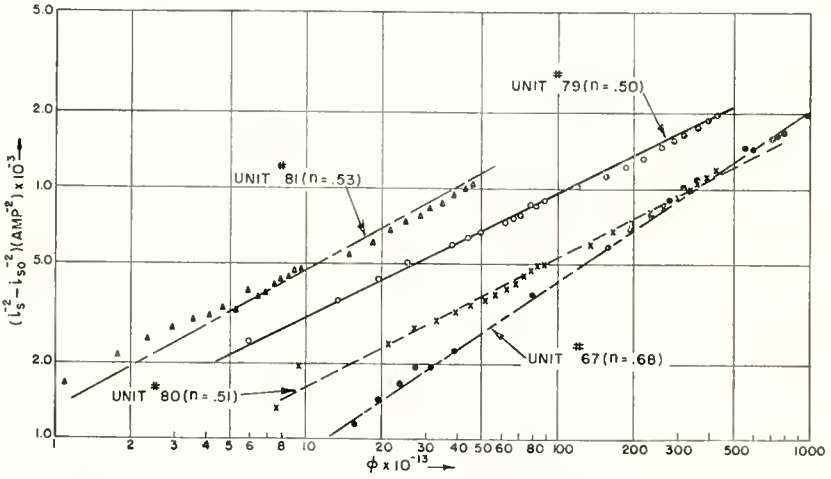


Fig. 14—Plots of ln $(i_x^{-2} - i_{so}^{-2})$ versus ln ϕ for three sample irradiated by 2 mev electrons in vacuum and one unit (No. 67) in air.

$$P_{\max} \approx i_s V_{\max} \propto \phi^{-1/2} \ln \phi^{-1}. \tag{9}$$

These relations will be true provided that only τ changes during irradiation. Change of σ will not affect i_s and will not produce much of a change in the other parameters.

The validity of the relations shown above for high-energy (> 145 kev) electron bombardment of silicon alloy diodes has been partially confirmed in previous experiments. The data from the present experiments deviates considerably from the predictions of Equations (6) through (9). For example, i_s^{-2} was found to be proportional to ϕ^n, where $1 > n > \frac{1}{2}$ instead of the expected value of $n = 1$. This is shown in Figure 14 where ln i_s plotted versus ln ϕ for four units

irradiated by 2 mev electrons. Further experiments are needed to explain these deviations.

Changes in surface recombination velocity could also affect i_s, V_{max} and P_{max}, although it is not expected that surface changes would follow the relatively simple form of Equations (6) through (9). Radiation which is composed of quanta not sufficiently energetic to produce bulk displacements could affect the efficiency of a photovoltaic call by changing the surface recombination velocity. Such changes might be caused by removing silicon or foreign atoms from the surface, or perhaps, by ionizing gas atoms in the vicinity of the surface and the subsequent adsorption of such ions onto the surface. If the latter were true, one would expect a different behavior of the cells during irradiation in vacuo and in air which has not been observed unequivocally.

Conclusions

On the basis of these experiments, it is concluded that silicon solar cells will not experience any serious deterioration for periods of the order of 10^5 years as a result of the radiations to which they will be exposed in the upper atmosphere assuming that the actual amount of radiation encountered corresponds to the best current estimates. Specifically, the only radiations which produced readily observable damage were 20 mev protons, 40 mev alpha particles and 2 mev electrons. This damage has been identified as a change in the bulk properties of silicon. From our data, it is possible to extrapolate to the time which shall have elapsed before the cells lose 25 per cent of their initial power output in the alpha particle and proton fluxes of the upper atmosphere. Such extrapolations indicate a minimum life of 3.6×10^5 years with protons, 1.4×10^5 years with alpha particles and 10^8 years with electrons. If minimum life is calculated on the basis of recent satellite data, these figures would be reduced by four orders of magnitude, or even more, provided that the radiation involved is corpuscular. Such a reduction in life is in the realm of speculation until further experiments with satellites yield more complete data on the nature of upper atmosphere radiation.

The electromagnetic radiations used did not produce any changes which could be labeled unequivocally as a decay in the power output. For ultraviolet radiation, it could be argued that P_{max} may decrease by 25 per cent after 85 years. As for x-rays, the best estimates of the type of electromagnetic radiation above the atmosphere indicate no x-rays of energy higher than 1,200 ev so that none of the electromagnetic radiation to be encountered could produce such bulk damage. The most pessimistic possible interpretation of the data, according to

which a 5 per cent decay during the irradiation was masked by experimental error, would lead one to estimate that a 25 per cent decrease in P_{max} could occur after 6.7 years in the x-irradiation of the upper atmosphere. It would have been desirable to have extended the x-ray radiation times so that these extrapolations could have approached the minimum life based on the particle irradiation, but this was impossible within the time limitations of this study. It should also be pointed out that none of these estimates of minimum life have included the effects of annealing which are likely to occur and which would increase the minimum life. Operation at a somewhat elevated temperature should also appreciably increase the minimum life. No significant differences between irradiation in vacuum and in air were noted.

It is important to point out that the decay of P_{max} is approximately inversely proportional to the flux so that a second decrease of 25 per cent would require 4/3 as much flux, i.e., time, as this first decrease, with correspondingly longer periods for further decay.

Acknowledgments

Ronald Hand of these Laboratories gave invaluable assistance in many phases of this work. B. Markow, L. McSherry, and R. Rast of the Evans Signal Laboratories gave generous assistance in operating their Van de Graaff accelerator and in supplying information concerning the total dosage received by the units. Thanks are also due to Professor R. Sherr and Dr. C. Brockman of Princeton University, and Dr. C. P. Baker of Brookhaven National Laboratories whose advice and help were indispensable for the Cyclotron experiments.

THE INFLUENCE OF DEFECT LEVELS ON PHOTOEMISSION*[1]

By

W. E. Spicer

RCA Laboratories,
Princeton, N. J.

Summary—The influence of defect levels on the characteristics of photoemitters is considered. These levels may affect a photoemitter through:

(1) The creation of photoemissive centers,

(2) The production of a bending of the bands at the semiconductor-vacuum interface and the semiconductor-backing interface, and

(3) The determination of the magnitude of thermionic emission from and the conductivity of the photoemitter.

It is shown that, if photoemission from the valence band is to be used, the photoemitter performance will be optimized by acceptor levels near the top of the valence band. If photoemission from defect levels is desired, these should lie near the bottom of the conduction band. However, quantum efficiency in excess of a few percent cannot be expected from defect levels.

Introduction

ALL PHOTOEMITTERS of practical interest such as Cs_3Sb, [Cs] $(NaK)_3Sb$, and Ag-O-Cs are semiconductors. In most cases[2] the photoemission of interest is due to the excitation of electrons from the valence band (intrinsic photoemission), whereas the position of the Fermi level and thus the conductivity and thermionic emission of the photoemitter are determined by defects.[3-6] It is the

* Manuscript received August 26, 1958.

[1] The term "defect" level covers any energy level introduced into a semiconductor by an imperfection in the lattice whether by a foreign atom (impurity level), a vacancy, or any other type of defect.

[2] The infrared response of Ag-O-Cs is probably not due to valence band excitation.

[3] A. H. Sommer, "N-Type and P-Type Conduction in Alkali Antimonide Photoemitters," *Jour. Appl. Phys.*, Vol. 29, p. 1568, November, 1958.

[4] J. Sakata, "Some Experimental Studies of the Conductivity and Thermoelectromotive Force of Cs_3Sb Photo-Cathodes," *Jour. Phys. Soc. Japan*, Vol. 8, p. 125, January-February, 1953; "Hall Effect in Cesium Antimonide With Audio-Frequency Currents," *Jour. Phys. Soc. Japan*, Vol. 9, p. 1030, November-December, 1954.

[5] J. E. Davey, "Thermionic and Semiconducting Properties of (Ag)-Cs_2O, Ag, Cs," *Jour. Appl. Phys.*, Vol. 28, p. 1031, September, 1957.

[6] W. E. Spicer, "Photoemissive, Photoconductivity, and Optical Absorption Studies of the Alkali-Antimony Compounds," *Phys. Rev.*, Vol. 112, p. 114, October, 1958.

of the order of 2×10^4 per centimeter are obtained. Thus, only about 10 per cent of the light is likely to be absorbed near enough to the S–V interface to produce photoelectrons. However, not all of the absorbed photons will produce photoelectrons. The maximum quantum efficiency obtained from photoemitters where all of the light was absorbed within 500 Å of the surface (intrinsic photoemitters) is about 30 per cent,[2] and there is no reason to assume that defect centers are any more effective in producing photoelectrons. Therefore, at best, about 30 per cent of the electrons excited from defect levels can be expected to escape from the photoemitter. Thus, under optimum con-

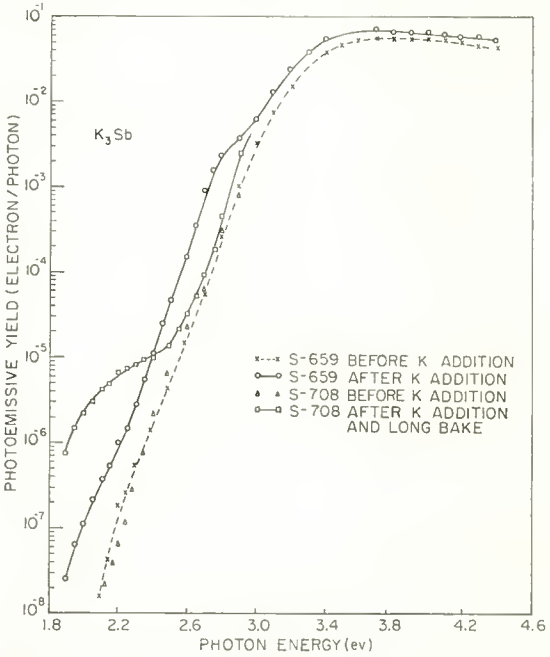


Fig. 2—Photoemission from K_3Sb. Defect photoemission at about 2.1 and 2.8 ev is produced by potassium addition. The response of S-708 is not followed to higher $h\nu$ since it was a thick layer in a tube of such a geometry that only illumination from the glass substrate was possible.

ditions, the maximum efficiencies to be expected from defect centers would seem to be of the order of a few per cent. It is possible that the infrared peak of the Ag-O-Cs photocathode whose yield is of this order is due to defect levels.[5] However, this has not been definitely established.

Band Bending

Because of the presence of surface states at the S–V interface and

of a second solid at the S–B interface, it is to be expected that internal electric fields will be set up in the photoemitter. This will result in a bending of the conduction and valence bands with respect to the Fermi level. Such bending may help or hinder photoemission.

First, let us examine the S–V interface (see Figure 3). Here either donor or acceptor surface states may be present. With n-type surface states which originally lay above the Fermi level, the surface states must be emptied of their electrons until all of the states which remain filled lie below the Fermi level. This will result in a positive charge layer at the surface. Thus the bands will bend downward at the surface. Conversely, it can be shown that acceptor states which lay below

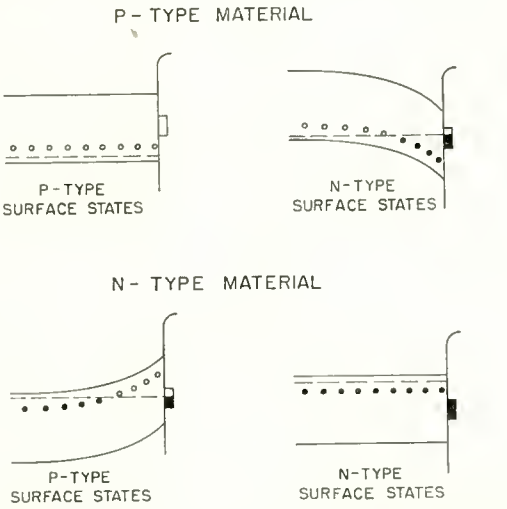


Fig. 3—The type of band bending which may be expected for surface states near the center of the forbidden gap. The Fermi level is indicated by the dashed line.

the Fermi level before the surface charge is redistributed will cause the bands to bend upward at the surface. These possibilities are illustrated in Figure 3 as are the cases where no bending will take place. Here we have assumed surface states near the center of the band and recognized that the Fermi level will lie in the upper half of the forbidden gap for n-type material and in the lower half for p-type material.

From these considerations one can make a simple generalization about the bending of the bands. With n-type bulk material, the bands will tend to bend upward, whereas with p-type material, the bands will tend to bend downward.

From Figure 3, the effect of the band bending on photoemission

may be deduced.[12] If the bands bend upward, an electron excited in the interior of the material must overcome a retarding potential in moving to the surface. Whereas, with bands bending downward, the electron will be effectively accelerated toward the surface. Thus bands which bend upward will hinder photoemission and those which bend downward will aid it.

It has been found[3,6] in the family of photoemitters, [Cs] $(NaK)_3Sb$, Cs_3Sb, Rb_3Sb, K_3Sb, and Na_3Sb, that the best emitters are p-type in bulk conductivity. How much this has to do with band bending and how much with the inherent properties of these materials has not been established. It is possible that the effect of oxygen on the Cs_3Sb cathode, i.e., the shift of the threshold toward longer wavelengths, is due to the introduction of new surface states which produce a favorable band bending.

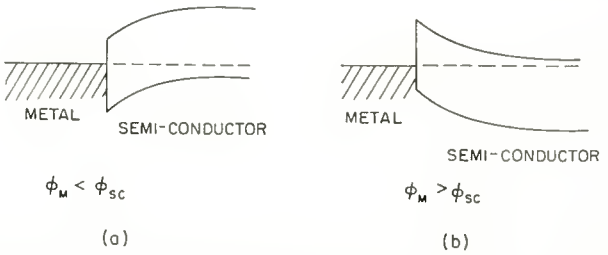


Fig. 4—Band bending at the interface between a metal and the photoemitter. ϕ_M is the work junction of the metal and ϕ_{SC} is that of the semiconductor.

Thus, there is, at present, little direct evidence of the effect of band bending on photoemission, probably because the effect is difficult to detect and because a concerted attempt has not been made to detect it. However, band bending has been so well established for semiconductors in general[13] that it is certain to take place also in the photoemitters.

In discussing the situation at the interface between the semiconductor and its supporting contact, we will assume, for simplicity, that the supporting material is a metal (similar arguments may be made for semiconducting substrates). If the semiconductor work function is greater than that of the metal, the situation shown in Figure 4(a) is obtained (an ohmic contact for electrons). Whereas if the semi-

[12] For the band bending to be of practical importance, there must be an appreciable potential drop within 500 Å of the surface. Due to the high impurity concentrations present in photoemitters[4,6], this condition is likely to be fulfilled.

[13] For example, see the review articles: R. H. Kingston, "Review of Germanium Surface Phenomena," *Jour. Appl. Phys.*, Vol. 27, p. 101, February, 1956.

conductor work function is smaller than that of the metal, the situation shown in Figure 4(b) is obtained (a rectifying contact for electrons). Remembering that the thickness of photoemitters may be comparable with the depth from which photoemission can occur, it is evident that such bending of the bands at the supporting surface–photoemitter interface can affect the photoemission. In particular, the case shown in Figure 4(b) will tend to accelerate electrons away from the supporting surface toward the vacuum and thus enhance the photoemissive efficiency, whereas that shown in Figure 4(a) may hinder photoemission.

Thermionic Emission and Electrical Conductivity

The thermionic emission from a photocathode at operating temperature determines the dark current in the tube and, thus, the ultimate noise level. For practical purposes, it is therefore desirable to reduce the thermionic emission as much as possible. Another important practical consideration is that the photocurrent which can be drawn from

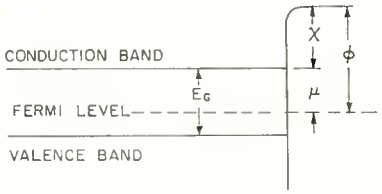


Fig. 5—Energy level diagram for a semiconductor without band bending. χ is the electron affinity, μ is the position of the Fermi level measured from the bottom of the conduction band and E_G is the band gap.

a cathode should not be limited by high cathode resistance. (Cathode resistivity is usually not a problem at room temperature but often becomes excessive at lower temperatures.) It will be shown here that both low thermionic emission and low resistance may be achieved simultaneously.

Thermionic emission, as well as conductivity, of a semiconductor is determined by the position of the Fermi level. The expression for the thermionic emission is

$$J = A_0 T^2 e^{-(\chi+\mu)/kT}, \tag{1}$$

where J is the thermionic current density, T is the absolute temperature, k is Boltzmann's constant, χ is the electron affinity, μ is the position of the Fermi level, and A_0 is a constant. $\chi + \mu = \phi$ is the thermionic work function. χ, μ, and ϕ are shown in Figure 5. μ is

measured downward from the bottom of the conduction band and is a positive number.

The expression for the conductivity is

$$\sigma = N_c \epsilon v \, e^{-\mu'/kT} \tag{2}$$

where N_c is the effective density of states in the band in which the conduction takes place, ϵ is the electronic charge, v is the carrier mobility, and μ' is the position of the Fermi level measured from the nearest band. Thus, for an n-type semiconductor, $\mu' = \mu$; for a p-type material, μ' is measured from the valence band and $\mu = E_G - \mu'$. For a simple system with only a single "defect" level, the value of μ will be given by

$$\mu_d = \frac{E_d}{2} + \frac{kT}{2} \ln \frac{N_c}{N_D} \tag{3}$$

for a donor, and

$$\mu_a = E_G - \frac{E_a}{2} - \frac{kT}{2} \ln \frac{N_c}{N_A} \tag{4}$$

for an acceptor level. Here E_d is the energy of the donor level measured from the bottom of the conduction band, N_D is the density of donors, E_G is the band gap, E_A is the acceptor activation energy measured from the top of the valence band, and N_A is the density of acceptor levels. For compensated materials, the Fermi level will roughly coincide with the dominant defect level. Thus, it is apparent, depending on the defects present, that the activation energy for thermionic emission from a given photoemitter may vary from χ to $E_G + \chi$; whereas the activation energy for conductivity may vary from zero to $E_G/2$. Both minimum thermionic emission and maximum conductivity may be obtained with an acceptor level located near the top of the valence band.

In experimental confirmation of this type of behavior, it is interesting to note that the multi-alkali [Cs](NaK)$_3$Sb cathode has a threshold of response extending to lower energy than the oxygen-treated Cs_3Sb (S-4 or S-11 surface)[14] while the dark current (thermionic emission) is roughly identical for these two cathodes.[14,15] Thus

---

[14] A. H. Sommer, "Multi-Alkali Photo Cathodes," *Trans. I.R.E. PGNS*, p. 8, November, 1956.

[15] R. W. Engstrom and W. Widmaier, private communication.

it appears that for the multi-alkali cathode, the Fermi level lies closer to the valence band than for the Cs_3Sb surface. This is confirmed by the fact that the Cs_3Sb surface becomes insulating at 77°K, whereas the multi-alkali surface does not.

Schaetti[16] has attempted to add a number of different impurities to Cs_3Sb and has investigated their effect on thermionic emission. He found a decrease by about a factor of ten in dark current with the addition of Tl, Mg, Ce, and Ba. Thus it would seem that the addition of these elements lowers the Fermi level in Cs_3Sb. It is not possible to go beyond this qualitative statement because the experimental conditions in Schaetti's work were too complex to allow more detailed conclusions.

Conclusions

It has been shown that for the minimum thermionic emission (and thus noise), maximum conductivity, and optimum effect of band bending, p-type defects near the top of the valence band are needed. However, if photoemission from defect levels (rather than the valence band) is desired, n-type levels near the bottom of the conduction band are desirable. It is estimated that the maximum quantum efficiency which can be expected from defect levels is about one per cent.

Acknowledgment

The author would like to express his appreciation to A. H. Sommer as well as to his other colleagues for many stimulating discussions and to A. H. Sommer and M. L. Kaiser for their aid in the experimental work.

[16] N. Schaetti, "Beeinflussung der Charakteristik uner Cs-Sb-Photo-Kathode durch Zusatz fremder Elemente," *Zeits. ang. Math. and Phys.*, Vol. 4, p. 450, 1953.

DRIVE FACTOR AND GAMMA OF CONVENTIONAL KINESCOPE GUNS*

By

R. D. Gold† and J. W. Schwartz‡

RCA Laboratories,
Princeton, N. J.

Summary—*A new and more accurate expression for the beam current–drive voltage relationship of conventional kinescope guns is presented. It is shown that, to a good approximation, the space-charge-free electric field at the cathode can be expressed as a quadratic function of radial distance from the cathode center. The expression is written in terms of the potential applied to grids 1 and 2 and four constants which depend only on gun geometry. The actual beam current is then related to the space-charge-free electric field. Drive factors [(maximum beam current)/(cutoff voltage)$^{3/2}$] are obtained for both the grid and cathode drive conditions.*

The values of gamma and drive factor obtained are in better agreement with experimental results than previous expressions. Considerable interest presently exists in guns which exhibit an increased cathode drive factor (and decreased grid drive factor) because of close grid 1–grid 2 spacing and low grid 2 voltage. The expressions presented here clearly indicate this behavior.

Introduction

THE problem of reduction in kinescope drive requirement is presently of interest for several reasons.[1-4] A decrease in the required drive voltage can lead to a significant cost reduction, since lower power video output stages can be used, or one stage may even be eliminated. A small drive voltage is important in the development of a-c/d-c type television sets, where relatively low B+ voltages make it inconvenient to obtain a large video output. In addition, low-drive kinescopes are particularly advantageous in transistorized television sets, where only small signal levels are normally available.

* Manuscript received July 24, 1958.

† RCA Laboratories, Princeton, N. J.

‡ Formerly, RCA Laboratories, Princeton, N. J.; now with Kaiser Aircraft and Electronics Corporation, Oakland, Cal.

[1] J. M. Sheehe, "Electron-Gun Design Permits Low E_{g2}," *Electronic Equipment*, p. 34, April, 1957.

[2] J. W. Schwartz, "The Annular Geometry Electron Gun: A New Device," *I.R.E. National Convention Record*, Vol. 6, Part 3, p. 13, 1958.

[3] J. W. Schwartz, "A New High Transconductance Electron Gun for Kinescopes," *RCA Review*, Vol. XIX, p. 232, June, 1958.

[4] D. Sillman, "Design Considerations in Transformerless Single Rectifier TV Receivers," *I.R.E. National Convention Record*, Vol. 6, Part 7, p. 133, 1958.

This paper discusses the drive and gamma characteristics of conventional triode-type kinescope guns. A theoretical analysis[5] is presented which leads to an expression for beam current in terms of the applied grid potentials and four geometric constants. Experimental results are in good agreement with the derived expressions.

Basic Equations

The beam current in a conventional kinescope is determined primarily by the geometry and applied potentials in the triode section. The masking apertures in succeeding electrodes normally intercept a small percentage of the transmitted triode current. Although the percentage may increase with increasing beam current, thus reducing the over-all gamma, this effect will be ignored.

The Langmuir–Child law may be used to express the current density, j, from a small segment of the cathode in terms of the space-charge-free electric field, E, normal to the segment. Thus,

$$j = kE^{3/2}, \tag{1}$$

where k is assumed constant. Under the Langmuir–Child assumptions for plane parallel electrodes and complete space-charge limitation, k contains a factor $1/\sqrt{d}$, where d is the electrode spacing. However, there are theoretical and experimental reasons for assuming k to be constant.

Kinescope control characteristics do not follow the familiar three-halves power law. This is because the active cathode emission area varies with drive voltage. Because of the axial symmetry of the electrode structure, the grid apertures, and the applied potentials (grid 1 negative, grid 2 positive), the field at the cathode decreases with radial distance, r, from the center of the cathode. At some critical radius, r_c, depending on the applied voltages, the field becomes zero. At greater distances ($r > r_c$) the field is negative. Therefore, essentially no emission takes place outside a circle of radius r_c. In order to find the cathode current, an expression must first be found for the functional dependence of field on radial distance from the cathode center.

From electrostatic theory, the potential at any spatial point can be expressed in terms of the electrode potentials and geometric coefficients of proportionality. Therefore, it follows that the electric field at the cathode can also be expressed in the same terms. Thus,

---

[5] See also K. Schlesinger, "Transfer Characteristics and Mu Factor of Picture Tubes," *Proc. I.R.E.*, Vol. 41, p. 528, April, 1953.

$$E(0, r) = K_1(r) V_1 + K_2(r) V_2, \tag{2}$$

where V_1 and V_2 are the grid 1 and grid 2 potentials, respectively. The coefficients $K_1(r)$ and $K_2(r)$ can each be expressed as a power series in r. Equation (2) is, therefore, rewritten as

$$E(0, r) = (A + Br^2 + \cdots) V_1 + (C - Dr^2 + \cdots) V_2, \tag{3}$$

where A, B, C, and D are arbitrary geometric constants. For practical purposes, the higher order terms can be neglected. This is justified as follows:

In axially symmetric fields, the potential $V(z, r)$ at any point can be expressed in terms of the axial potential ϕ and its derivatives.[6]

Thus,

$$V(z, r) = \phi - \frac{r^2}{4}\phi'' + \frac{r^4}{64}\phi^{(4)} - \frac{r^6}{2304}\phi^{(6)} + \cdots \tag{4}$$

Differentiating Equation (4) with respect to z,

$$E(0, r) = \phi' - \frac{r^2}{4}\phi''' + \frac{r^4}{64}\phi^{(5)} - \frac{r^6}{2304}\phi^{(7)} + \cdots \tag{5}$$

The terms in Equation (5) diminish rapidly at sufficiently small values of r. For the purposes of this derivation, only the first two terms will be retained. Figure 1 indicates that errors resulting from omission of the higher order terms are significant only at values of $r > r_c$. Thus, Equation (3) can be rewritten as

$$E(0, r) = (A + Br^2) V_1 + (C - Dr^2) V_2. \tag{6}$$

The value for the critical radius, r_c, is obtained by setting Equation (6) equal to zero. This gives

$$r_c = \sqrt{\frac{AV_1 + CV_2}{DV_2 - BV_1}}. \tag{7}$$

Combining Equations (6) and (7),

$$E(0, r) = (r_c^2 - r^2)(DV_2 - BV_1). \tag{8}$$

[6] I. G. Maloff and D. W. Epstein, *Electron Optics in Television*, McGraw-Hill Book Co., Inc., New York, N. Y., 1938, p. 77.

Equation (1) can be expressed in integral form to give the total cathode current, I;

$$I = 2\pi k \int_0^{r_c} E(0, r)^{3/2}\, r\, dr. \tag{9}$$

Upon integration, this becomes

$$I = M \frac{(AV_1 + CV_2)^{5/2}}{(DV_2 - BV_1)}, \tag{10}$$

where $M = \dfrac{2\pi}{5} k$.

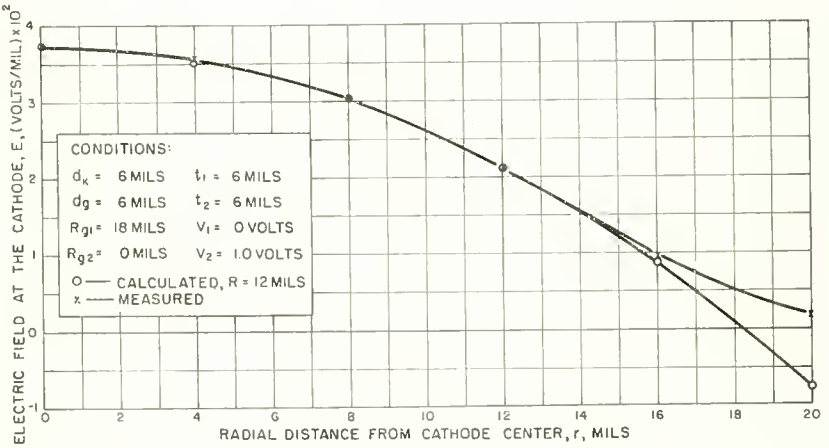


Fig. 1—Comparison of measured and calculated electric field, from resistor-board data.

In deriving this equation, it should be noted that space-charge effects were not rigorously taken into account. Although the Langmuir–Child law does include space-charge effects, Equation (2) was obtained from purely electrostatic considerations. It ignores the effects of the charge distribution on the potential near the cathode. Space-charge effects are accounted for in the selection of the constant M. As shown in a later section, the expressions derived using these approximations are in good agreement with experimental measurements.

Equation (10) is used to obtain expressions for a drive factor and gamma for both grid drive and cathode drive conditions. These expressions are in terms of the constants A, B, C, and D. Equation (6) can be used to obtain the values of these constants from data taken on

an electrostatic analog such as a resistor board.[7]

Drive Factor

For both grid and cathode drive, the drive factor is defined as the ratio of maximum beam current I_{max} (current at $V_1 = 0$) to the three-halves power of cutoff voltage V_{co}. It has the same units as perveance, is a function only of the tube geometry and the cathode emission properties,* and is independent of applied voltages. The drive factor can be expressed in terms of the constants previously defined, as follows: From Equation (10) at $V_1 = 0$,

$$I_{max} = M \frac{C^{5/2}}{D} V_2^{3/2}. \tag{11}$$

For grid drive, $I = 0$ at $V_1 = V_{co}$. Thus,

$$V_{co} = -\frac{C}{A} V_2. \qquad \text{(grid drive)} \tag{12}$$

Combining Equations (11) and (12),

$$I_{max} = M \frac{C}{D} A^{3/2} (-V_{co})^{3/2}. \qquad \text{(grid drive)} \tag{13}$$

The grid drive factor is therefore given by

$$P_{GD} \equiv \frac{I_{max}}{(-V_{co})^{3/2}} = M \frac{C}{D} A^{3/2}. \qquad \text{(grid drive)} \tag{14}$$

For cathode drive, the relationships

$$V_1 = -V_k, \tag{15a}$$

$$V_2 = V_2' - V_k \tag{15b}$$

must hold, where V_2' is the grid 2–ground voltage and V_k is the cathode–ground voltage. Substituting into Equation (10) and proceeding as for grid drive,

[7] G. Liebmann, "Solution of Partial Differential Equations with a Resistance Network Analogue," *British Journal of Applied Physics*, Vol. 1, p. 92, April, 1950.

* In practice, even under complete space-charge-limited conditions, cathode currents are not strictly independent of the cathode properties.

$$V_{co} = \frac{C}{A + C} V_2', \qquad \text{(cathode drive)} \quad (16)$$

and

$$P_{KD} \equiv \frac{I_{\max}}{(+V_{co})^{3/2}} = \frac{MC}{D} (A + C)^{3/2}. \text{ (cathode drive)} \quad (17)$$

For experimental evaluation purposes, the effects of variations in M for different tubes can be eliminated by considering the ratio P_{KD}/P_{GD}.

Thus,

$$\frac{P_{KD}}{P_{GD}} = \left(\frac{A + C}{A} \right)^{3/2}. \quad (18)$$

Expression for Gamma

The beam current in a kinescope is generally considered to be a power-law function of the signal or drive voltage, V_D. The exponent in this power law is called the tube gamma. It is usually determined experimentally as the slope of the logarithmic plot of current versus drive voltage;

$$\gamma \equiv \frac{\partial \log I}{\partial \log V_D} = \frac{V_D}{I} \frac{\partial I}{\partial V_D}. \quad (19)$$

Equation (10) can be rewritten to express the beam current in terms of drive voltage, from which an expresion for gamma is found. In doing this, grid drive operation and cathode drive operation are treated as separate cases.

Grid Drive

The grid drive voltage is given by

$$V_{GD} = V_1 - V_{co}. \quad (20)$$

Substituting Equations (12) and (20) into (10),

$$I = MA^{5/2} \frac{V_{GD}^{5/2}}{\left(D + \frac{BC}{A} \right) V_2 - BV_{GD}}. \quad (21)$$

Thus,

$$\gamma_{GD} = 2.5 + \frac{1}{\left[\left(\frac{D}{B} + \frac{C}{A}\right)\frac{V_2}{V_{GD}} - 1\right]}. \quad \text{(grid drive)} \quad (22)$$

Cathode Drive

The cathode drive voltage is given by

$$V_{KD} = V_{co} - V_K. \quad (23)$$

Since the grid is at a-c ground potential for cathode drive, from Equation (15a),

$$V_{KD} = V_{co} + V_1. \quad (24)$$

Substituting Equation (16),

$$V_1 = V_{KD} - \frac{C}{A + C} V_2'. \quad (25)$$

Combining Equations (15a), (15b), and (25),

$$V_2 = V_{KD} + \left(\frac{A}{A + C}\right) V_2'. \quad (26)$$

Inserting Equations (25) and (26) into (10),

$$I = M(A + C)^{5/2} \frac{V_{KD}^{5/2}}{\left(\frac{AD + BC}{A + C}\right) V_2' + (D - B) V_{KD}}. \quad (27)$$

Thus,

$$\gamma_{KD} = 2.5 + \frac{1}{\left[\frac{(AD + BC)}{(A + C)(B - D)}\frac{V_2'}{V_{KD}} - 1\right]}. \quad \text{(cathode drive)} \quad (28)$$

Equations (22) and (28) are of special interest because they show that gamma is not constant, but depends on tube geometry and the applied voltages. In the absence of drive voltage, $\gamma = 2.5$. As the

grid–cathode voltage is increased from cutoff to zero bias, gamma increases to a value on the order of 3 or 4. It is also noted that the denominator of the second term in these equations vanishes as the drive voltage is still further increased. It should be realized, however, that the results are valid only for $V_1 \leqslant 0$.

The range of values for gamma reported in the literature varies from 1.4 to over 3. This analysis provides a theoretical basis for the higher values of gamma. Measurements of gamma less than 2.5 may be due to an erroneously low measurement of cutoff voltage. The slope of the logarithmic current–drive-voltage curve (which is gamma) will then also be too low. Figure 2 demonstrates this qualitatively. Incorrect values for drive factor also result from errors in V_{co}. An accurate method for measuring V_{co} is described in the Appendix.

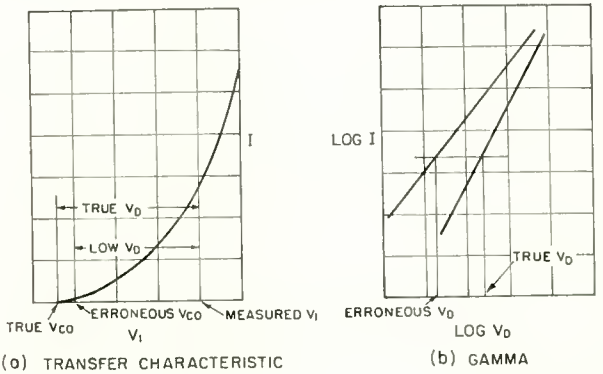


Fig. 2—Illustration of how errors in measuring V_{co} affect the apparent gamma.

Geometrical Dependence of Drive Factor

The constants A, B, C, and D have been obtained for various gun geometries with the aid of a resistor-board potential analog. Voltage measurements with two linearly independent sets of electrode voltages, at two values of r, are required for each geometry. The field at the cathode is obtained by taking the difference in potential between the cathode and the first measurable point in an axial direction from the cathode at the two different values of r. It has been found that the best quadratic approximation for the cathode field is obtained by choosing the radial positions $r = 0$ and $r \approx 2/3\, R_{g1}$, where R_{g1} is the radius of the grid 1 aperture. A schematic representation of the gun triode section is shown in Figure 3.

Figure 4a shows the values of A for variable grid 1–grid 2 spacing

and various values of grid 2 radius, R_{g2}. Figures 4b, c, and d show the values of B, C, and D, respectively, for the same conditions. Figures 5a, b, c, and d show measured values of the constants for variable cathode–grid 1 spacing with grid 1 thickness as parameter, for a grid 1–grid 2 spacing of 6 mils. Figures 6 and 7 show the values for the same conditions, but for grid 1–grid 2 spacing of 36 mils and 100 mils, respectively. Using Equations (14) and (17), numerical values of the normalized drive factors P_{GD}/M and P_{KD}/M are obtained from the constants A, B, C, and D for various geometrical configurations. The accuracy of the constants is in doubt for the 2-mil spacings because of the discrete nature of the resistor-board analog.

Figure 8 shows curves of P_{KD}/M and P_{GD}/M as functions of cathode–grid 1 spacing for a grid 1–grid 2 spacing of 36 mils and with grid 1 thickness, t_1, as a parameter. It is seen that the drive factor increases as the cathode–grid 1 spacing is made smaller. Also,

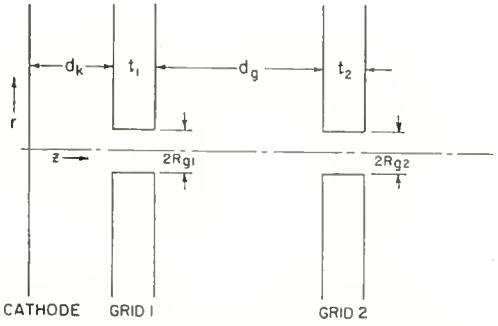
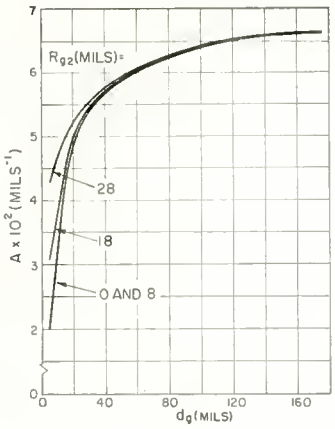


Fig. 3—Schematic of gun triode section.

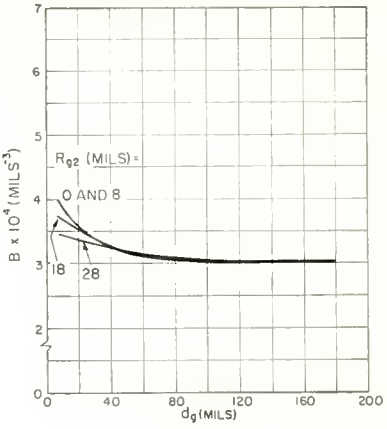
the drive factor is relatively insensitive to grid 1 thickness for large spacings, but increases with thinner grid 1 electrodes for small spacings. Figure 9 shows that the cathode drive factor can be greatly increased by decreasing the grid 1–grid 2 spacing. The grid drive factor, however, decreases in this case. The diameter of the grid 2 aperture has little effect on the drive characteristic except at close grid 1–grid 2 spacings.

Experimental Measurements of Drive Factor

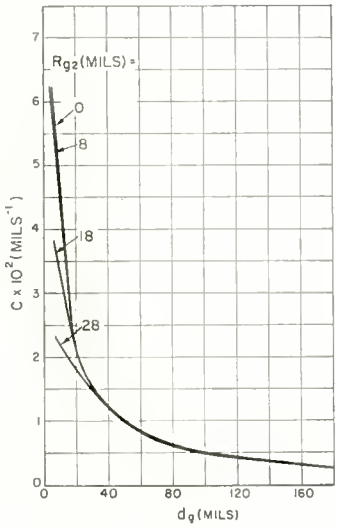
A series of electron-gun triode sections were made with different spacings, so that a comparison of the analytical predictions and experimental results could be made. Figure 10 shows the essential features of a typical gun. The cap at the end of G_2 insures that essentially all of the beam is collected and measured at that electrode.



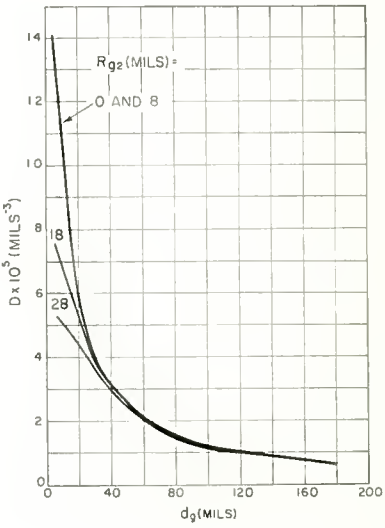
(a) GEOMETRIC CONSTANT A vs. SPACING d_g



(b) GEOMETRIC CONSTANT B vs. SPACING d_g



(c) GEOMETRIC CONSTANT C vs. SPACING d_g



(d) GEOMETRIC CONSTANT D vs. SPACING d_g

d_k = 4 MILS
t_1 = 4 MILS
R_{g1} = 18 MILS
t_2 = 6 MILS

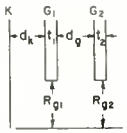
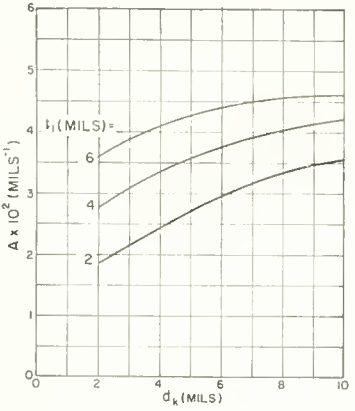
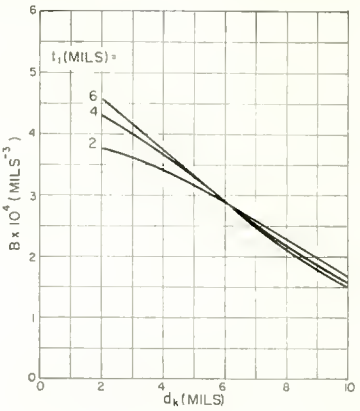


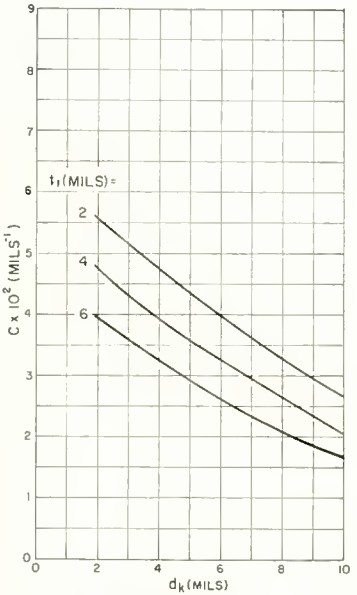
Fig. 4—Geometric constants versus grid 1–grid 2 spacing.



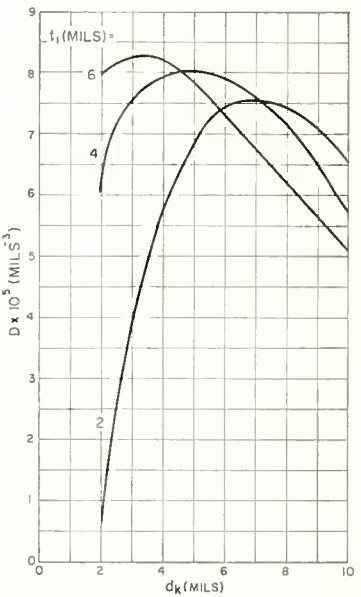
(a) GEOMETRIC CONSTANT A vs. SPACING d_k



(b) GEOMETRIC CONSTANT B vs SPACING d_k



(c) GEOMETRIC CONSTANT C vs. SPACING d_k



(d) GEOMETRIC CONSTANT D vs. SPACING d_k

d_g = 6 MILS
R_{g2} = 18 MILS
R_{g1} = 18 MILS
t_2 = 6 MILS

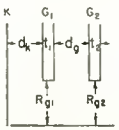


Fig. 5—Geometric constants versus cathode–grid 1 spacing; $d_g = 6$ mils.

The cathode–grid 1 spacing was varied from 2 to 10 mils (cold spacing), and the grid 1–grid 2 spacing from 2 to 140 mils. The beam current at zero bias and the cutoff voltage for both grid and cathode drive operation were measured, and the drive factor calculated.

Figure 11 shows measured values of P_{KD} for various grid 1–grid 2 spacings. The solid curve, obtained from Equation (17), is theoretical. In this equation, a value of $M = 0.75 \times 10^{-6}$ ampere/volt$^{3/2}$mil$^{1/2}$ was

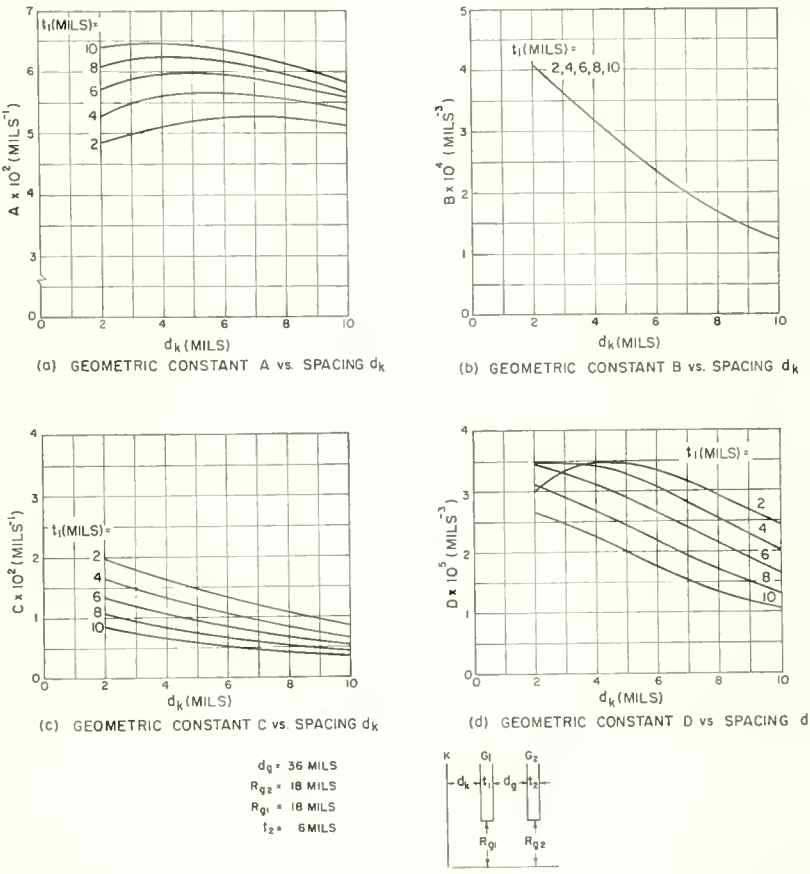


Fig. 6—Geometric constants versus cathode–grid 1 spacing; $d_g = 36$ mils.

chosen because it gave the best fit for the experimental points. The spread in the data is attributed to variations in cathode emission conditions.

Equation (18) indicates that the ratio P_{KD}/P_{GD} (which is always greater than unity) is independent of the emission constant, M. Figures 12 and 13 show the close correlation of experimental data with the analytical expressions for P_{KD}/P_{GD}.

Gamma

Figure 14 shows a logarithmic plot of beam current versus grid drive voltage for three different tubes with a grid 2 voltage of 200 volts. Figure 15 shows a similar plot for one of these tubes at $V_2' = 50$

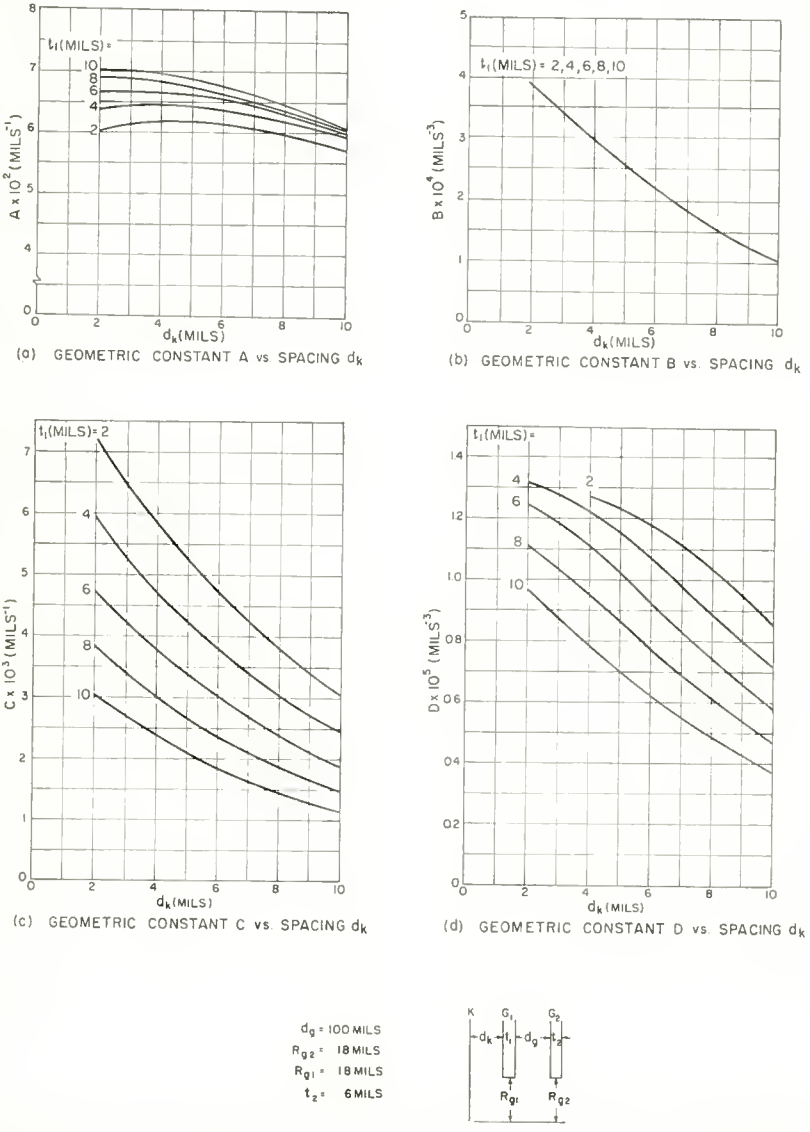


Fig. 7—Geometric constants versus cathode–grid 1 spacing; $d_g = 100$ mils.

volts and $V_2' = 400$ volts; grid drive and cathode drive curves are both shown. In each case gamma is greater than 2.50, and is seen to increase rapidly near zero bias. Figure 16 compares the experimental curves for grid drive with the theoretical curves obtained

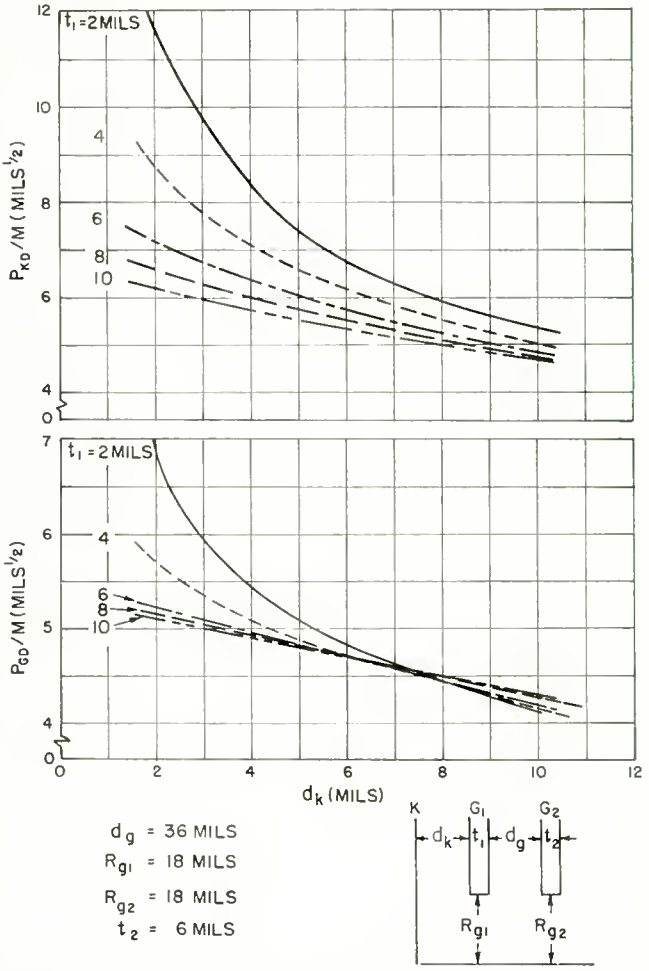


Fig. 8—Drive factor versus cathode–grid 1 spacing.

from Equation (21). In this equation, a value of $M = 0.62 \times 10^{-6}$ ampere/volt$^{3/2}$ mil$^{1/2}$ was chosen in order to permit a direct comparison for both values of V_2.

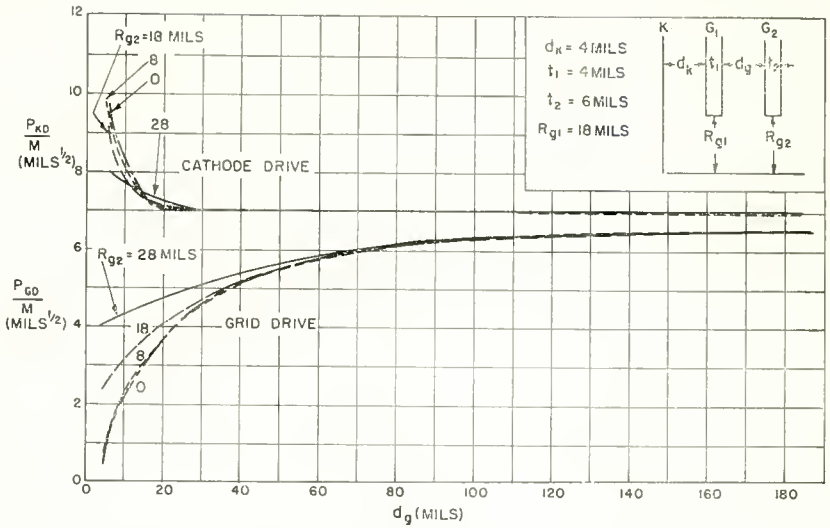


Fig. 9—Drive factor versus grid 1–grid 2 spacing.

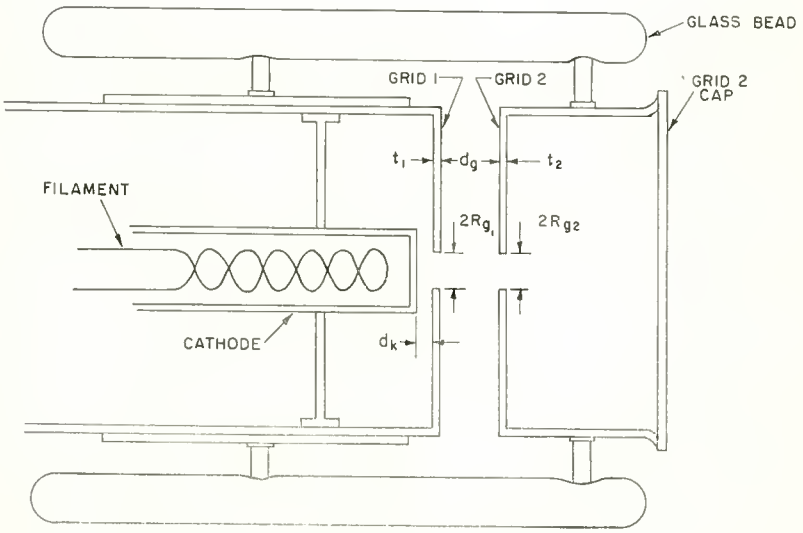


Fig. 10—Electron-gun triode section.

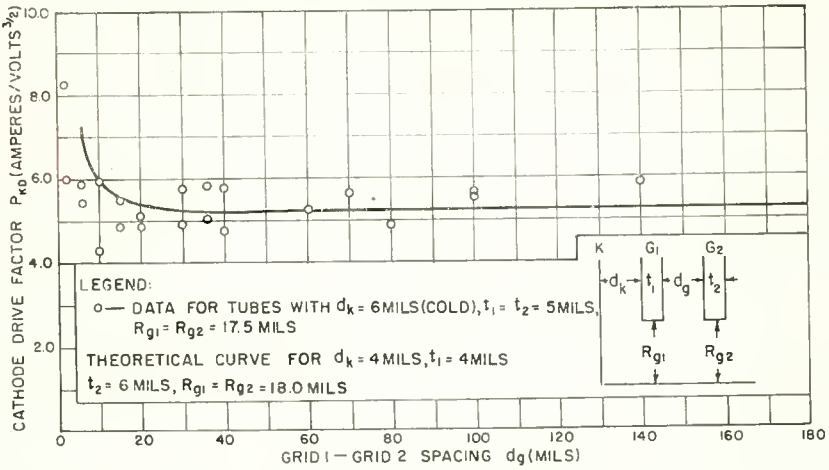


Fig. 11—Theoretical and experimental values of cathode drive factor versus grid 1–grid 2 spacing.

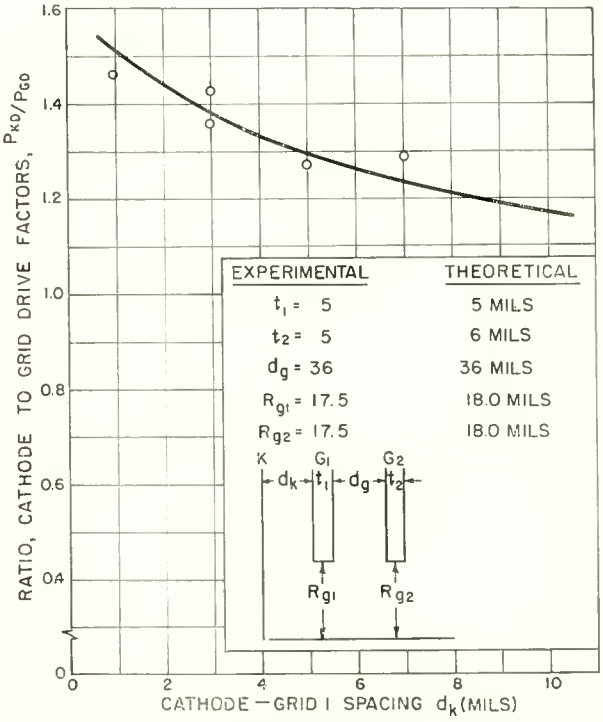


Fig. 12—Theoretical and experimental values of P_{KD}/P_{GD} versus cathode–grid 1 spacing.

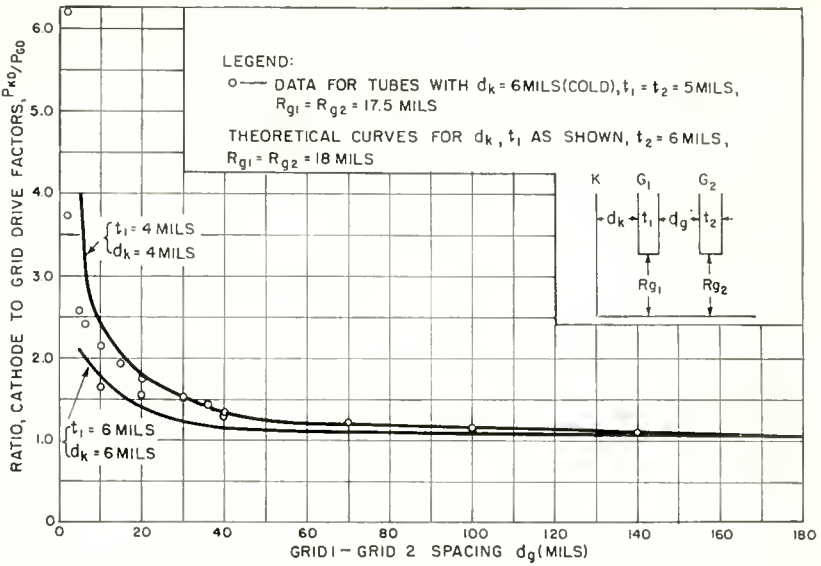


Fig. 13—Theoretical and experimental values of P_{KD}/P_{GD} versus grid 1–grid 2 spacing.

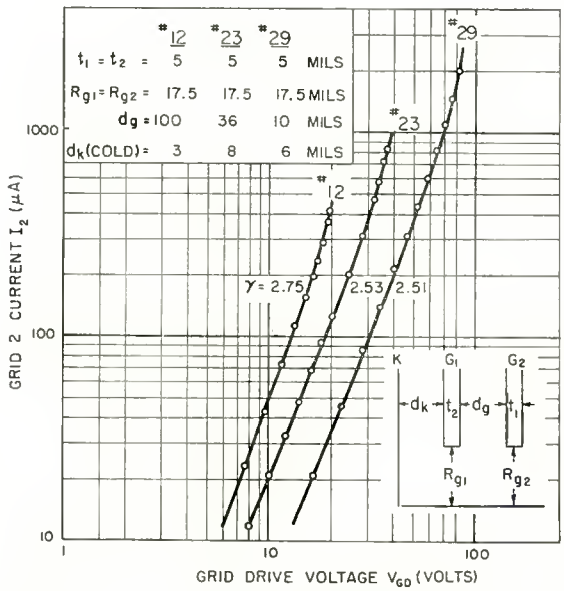


Fig. 14—Gamma for electron-gun triode section (three tubes, grid drive, $V_b = 200$ volts).

Conclusions

Equations have been derived which are in good agreement with experimental observations made by the authors and others. Unlike previously available expressions for gamma, experimentally observed values greater than three are accounted for. The cathode drive factor is shown to be always greater than the grid drive factor and, in some ranges, both are sensitively dependent on triode geometry. For close cathode–grid 1 and grid 1–grid 2 spacings, and for thin grid 1 discs, the cathode drive factor increases markedly. More than a two to one variation can be obtained from changes in triode geometry. For grid drive, the effect of a thin grid 1 disc and close cathode–grid 1 spacing increases the drive factor, but close grid 1–grid 2 spacing reduces the drive factor.

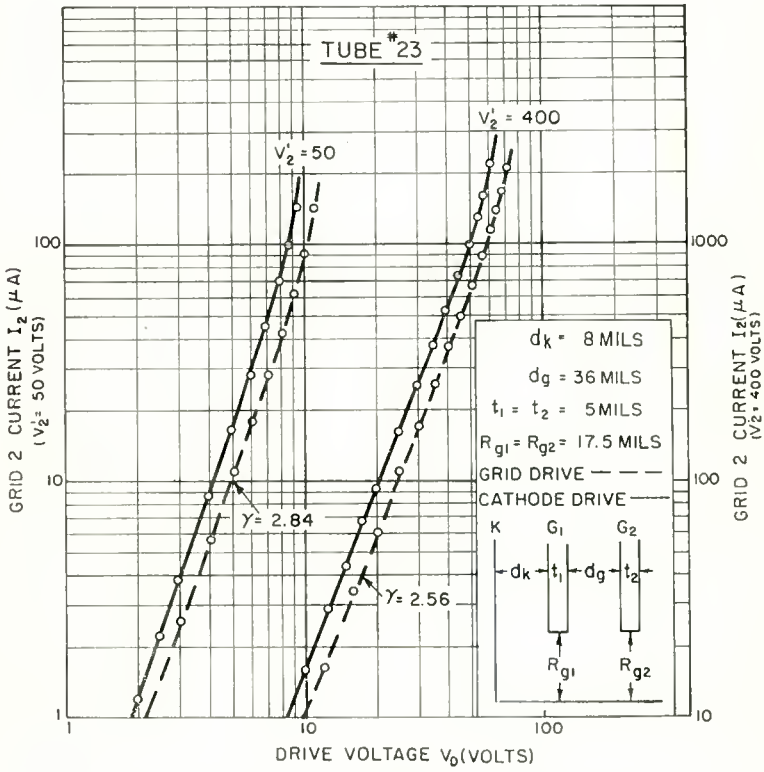


Fig. 15—Gamma for electron-gun triode section.

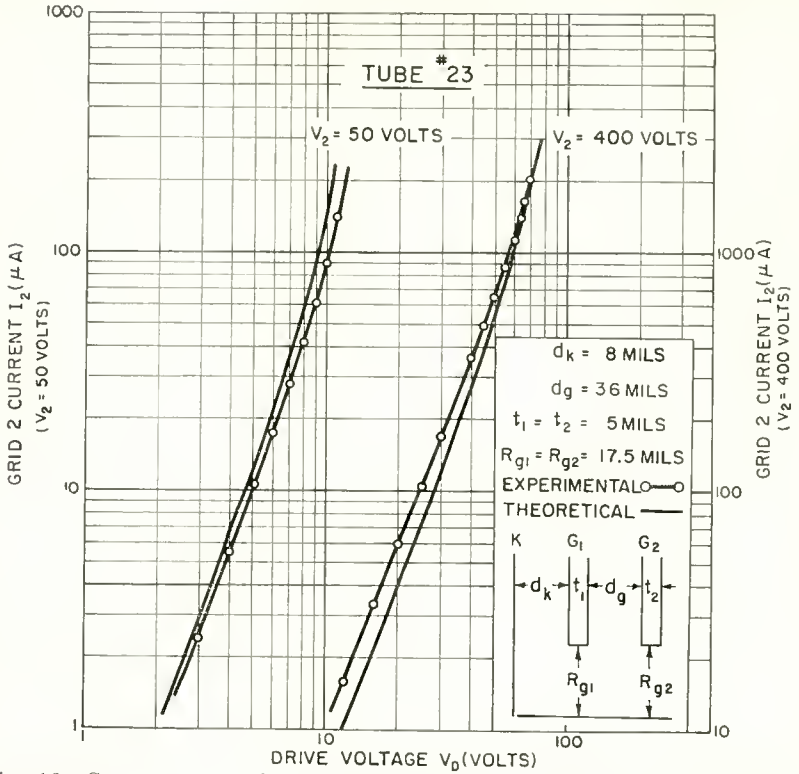


Fig. 16—Gamma comparison under grid drive conditions, theoretical and for gun triode section.

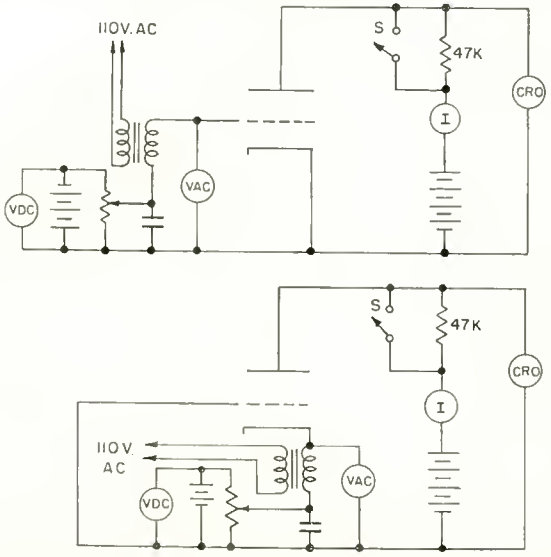


Fig. 17—Measuring circuit (top-grid drive, bottom-cathode drive).

Appendix

A circuit designed to determine V_{co} to within a few tenths of a volt is shown in Figure 17. A fixed a-c voltage is added to a variable d-c bias voltage obtained from a ten-turn precision potentiometer. The output waveform is viewed on a sensitive oscilloscope, and the potentiometer is adjusted so that the peak of the superimposed a-c voltage just starts to form a small pip in the output waveform (with switch S open). The magnitude of the cutoff voltage then equals the d-c voltage minus the peak a-c voltage. With switch S closed and both grid and cathode grounded, the maximum or zero bias current is measured. For the gamma measurements, switch S is closed and the a-c voltage source removed.

TWO BACKWARD-WAVE OSCILLATOR TUBES FOR THE 29,000-TO-74,000 MEGACYCLE FREQUENCY RANGE*†

By

Donald J. Blattner and Fred Sterzer

RCA Electron Tube Division,
Princeton, N. J.

Summary—*Two experimental tubes were built which demonstrated the feasibility of obtaining voltage-tunable oscillators having useful power output over very wide frequency ranges at millimeter wavelengths. The design and operating characteristics of these tubes are discussed. Their wide frequency range and milliwatt-level power outputs make them well suited for signal-generator and local-oscillator applications in the millimeter-wave spectrum.*

Extrapolation of the design indicates the possibility of tape-helix backward-wave oscillators delivering useful power at frequencies as high as 150,000 megacycles (2-millimeter wavelength).

Introduction

THIS paper describes the characteristics and performance of two experimental backward-wave oscillator tubes which were built to explore the possibilities of obtaining voltage-tunable oscillations over very wide frequency ranges in the millimeter-wavelength region. These tubes have been given developmental designations A-1073 and A-1074. The A-1073 oscillates between 29,000 and 49,000 megacycles, and the A-1074 oscillates between 48,000 and 74,000 megacycles. The frequency is varied by a single voltage adjustment. This frequency range is considerably larger than that obtained in previously described backward-wave oscillator tubes[1-6] operating at millimeter wavelengths.

---

* Manuscript received October 22, 1958.

† Much of the work described in this paper was performed under a contract sponsored by Wright Air Development Center.

[1] R. Kompfner, "Backward-Wave Oscillator," *Bell Lab. Record*, Vol. 31, p. 281, August, 1953.

[2] W. V. Christensen and D. A. Watkins, "Helix Millimeter-Wave Tube," *Proc. I.R.E.*, Vol. 43, p. 93, January, 1955.

[3] A. Karp, "Traveling-Wave Experiments at Millimeter Wavelength with a New, Easily Built, Spaceharmonic Circuit," *Proc. I.R.E.*, Vol. 43, p. 41, January, 1955.

[4] J. A. Noland and R. E. Lepic, "Backward-Wave Oscillators for the 17 to 41 KMC Band," *The Sylvania Technologist*, Vol. X, January, 1957.

Table I—Electrical and Mechanical Characteristics

	A-1073	A-1074
Frequency	29,000-49,000 megacycles	48,000-74,000 megacycles
Beam voltage	650-3,200 volts	750-3,100 volts
Beam current	3 milliamperes	2 milliamperes
Power output (calculated)	5-13 milliwatts	1.2-2.5 milliwatts
Helix inner diameter	0.031 inch	0.021 inch
Helix turns per inch	73	109
Cathode diameter	0.040 inch	0.030 inch
Cathode current density	1 ampere/centimeter^2	1 ampere/centimeter^2
Magnetic focusing field strength	1,200 gauss	1,200 gauss
Output waveguide	RG 97/U	RG 98/U

Both tubes use a tape-helix backward-wave circuit through the center of which a solid electron beam from a parallel-flow gun is focused by means of a uniform axial magnetic field. The r-f output signals are taken out of the tubes by means of standard waveguides (RG 97/U for the A-1073 and RG 98/U for the A-1074). A summary of the electrical and mechanical characteristics of both tubes is given in Table I.

Figure 1 is a photograph of the A-1074. Externally it is like the A-1073 except for the different size of the output waveguide and flange.

Description of Components

The major difficulty in designing tubes for the millimeter-wavelength range is that physical dimensions must be small, and therefore, the power-handling capacity and beam current of the tube are limited. Also, efficiency is low due to high r-f losses. The approach taken to minimize the problem of small size, with the resultant requirements on precision and tolerances in the mechanical design, was to make the

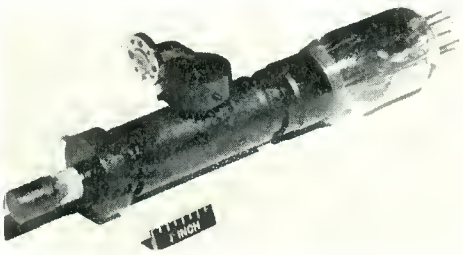


Fig. 1—A-1074 backward-wave oscillator tube.

[5] A. Karp, "Backward-Wave Oscillator Experiments at 100 to 200 Kilomegacycles," *Proc. I.R.E.*, Vol. 45, p. 496, April, 1957.

[6] R. W. Grow, D. A. Dunn, J. W. McLaughlin and R. P. Lagerstrom, "A 20 to 40 KMC Backward-Wave Oscillator," *Trans. I.R.E. PGED*, p. 152, July, 1958.

design of the tube as simple as possible. Figure 2 is a schematic diagram of the A-1073 showing the electron gun, the r-f output coupler, the helix, and the collector.

Tape-Helix Circuit

A tape helix was chosen as the slow-wave structure for the tube for both electrical and mechanical reasons. Electrically, the tape helix offers high backward-wave impedance over an octave frequency range, larger than that of any other known circuit, and coupling between r-f output and the helix over this frequency range is feasible. Moreover, the helix presents a comparatively large cross-sectional area over which the r-f field can interact with the electron beam.

Mechanically, the helix is a good circuit to use because conventional

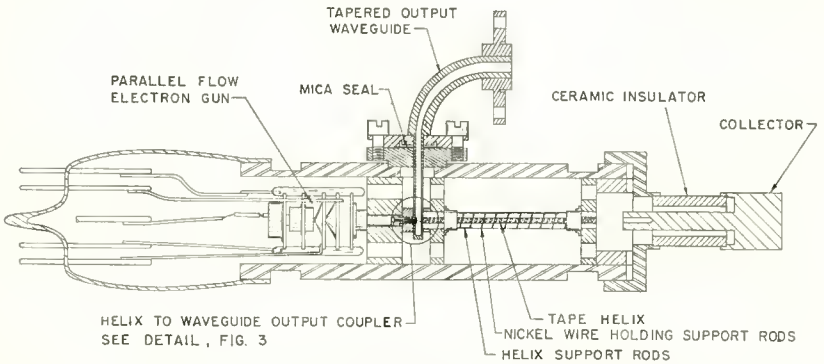


Fig. 2—Schematic diagram of tube assembly.

grid-winding techniques can be used to hold the very tight tolerances required in structures which are to be used at such extremely high frequencies. The helices are wound of tungsten tape 0.002 inch thick by 0.006 inch wide for the A-1073 and 0.002 inch thick by 0.0045 inch wide for the A-1074. The mean diameters of the helices in the two tubes are 0.0315 inch and 0.021 inch, respectively. These dimensions were chosen to assure that the ratio of helix circumference to free-space wavelength is less than 0.5 over the entire range of oscillation frequency. The helices are made long enough to be self-terminating. The tape-helix circuit is supported mechanically by three round quartz rods spaced equally around its outside diameter. The rods are held in contact with the helix by a fine nickel wire spiraling around their entire length, as shown in Figure 2.

Electron Gun and Collector

The Pierce-type parallel-flow electron gun[7] shown in Figure 2 was

[7] J. R. Pierce, *Theory and Design of Electron Beams*, D. Van Nostrand Co., Princeton, N. J., 1949.

designed for operation with the tape-helix circuit. Confined parallel flow was chosen as a simple and effective method of maintaining the electron beam close to the circuit. Electrons are emitted from the cathode in the presence of a confining axial magnetic field, and are accelerated to helix voltage by several electrodes whose spacing and voltage are so arranged as to keep the electrons moving in straight-line paths. The elements of the gun are: (1) an impregnated tungsten matrix cathode which produces a current density of 1 ampere per square centimeter, (2) a beam-forming electrode held at cathode potential, (3) an accelerating electrode, and (4) a drift tube, held at helix potential, which serves to collimate the electron beam before

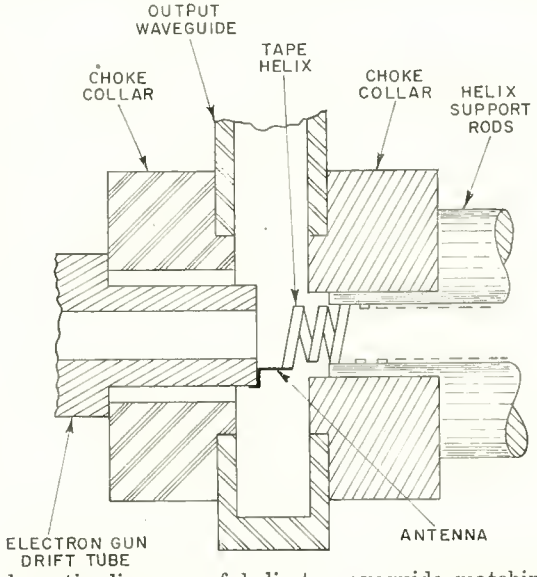


Fig. 3—Schematic diagram of helix-to-waveguide matching section.

it enters the helix. The inside diameter of the drift tube is 0.032 inch in the A-1073 and 0.021 inch in the A-1074.

After passing down through the tape-helix circuit, the electron beam is collected in the hollow cylinder of oxygen-free high-conductivity copper. This shape was chosen to reduce the number of secondary electrons reaching the helix.

R-F Output Circuit

The electromagnetic wave, which is excited in the tape helix by the electron beam, is radiated from an antenna at the gun end of the helix into the thin section of waveguide shown in Figure 3. The antenna connects to a post which is also the drift tube of the electron gun. Choke collars on both sides of the waveguide prevent radiation from

leaking out through the coupling apertures in the waveguide. The height of the section of waveguide at the antenna is 0.045 inch in the A-1073, and 0.030 inch in the A-1074. This dimension is then tapered up to standard waveguide size to that the actual output connections for the tubes are standard RG 97/U and RG 98/U waveguides and flanges. In most of the experimental tubes the output waveguide was sealed vacuum-tight by a "gasket-type" mica window.[8]

The transfer of energy from helix to waveguide is efficient over the entire frequency range of the tube. The position and penetration of the post to which the antenna is connected as well as the antenna length, were selected, on the basis of experiments on an X-band scale model, to provide optimum matching over the desired frequency range. The voltage standing-wave ratio of the scale model is shown in Figure 4 as a function of frequency. Similar data measured on an A-1073

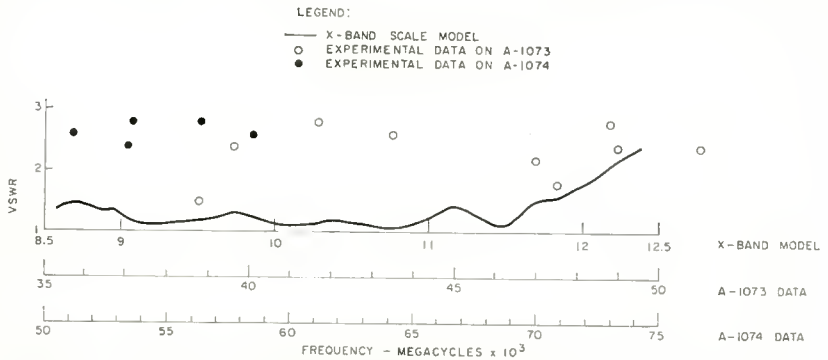


Fig. 4—Voltage standing-wave ratio versus frequency for the helix-to-waveguide matching section.

and an A-1074 are also shown. The voltage standing-wave ratio of the X-band model was below 2 throughout the frequency range of interest, and for the actual tubes the values were all below 3.

Tube Characteristics

Oscillation Frequency

The oscillation frequency of the A-1073 and A-1074 backward-wave oscillator tubes is varied by changing the helix potential. In Figure 5, the tuning curve for the A-1073, the frequency of oscillation is plotted as a function of helix voltage. As the helix voltage is varied from 650 to 3,200 volts, the frequency of the r-f output signal varies from 29,000 to 49,000 megacycles. The data given in Figure 5 for three different A-1073's indicates that the tuning curve is reproduced closely

[8] F. Sterzer, "Simple High-Temperature Vacuum Tight Mica Window," *Rev. Sci. Instr.*, Vol. 28, p. 208, March, 1957.

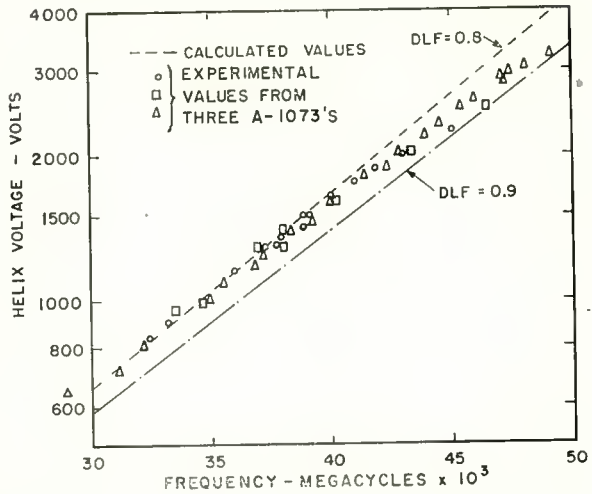


Fig. 5—Frequency of oscillation versus helix voltage for the A-1073.

from tube to tube. Calculated tuning curves, plotted as dashed lines in Figure 5,* are based on dielectric loading factors (DLF) of 0.8 and 0.9. A comparison between the experimental data and the calculated curves indicates that the DLF due to the quartz rods supporting the tape helix varies between the values of 0.8 and 0.9 and approaches the higher value at the upper end of the frequency scale.

Figure 6 shows the tuning curve of the A-1074 when the helix voltage is plotted against the corresponding oscillation frequency. The

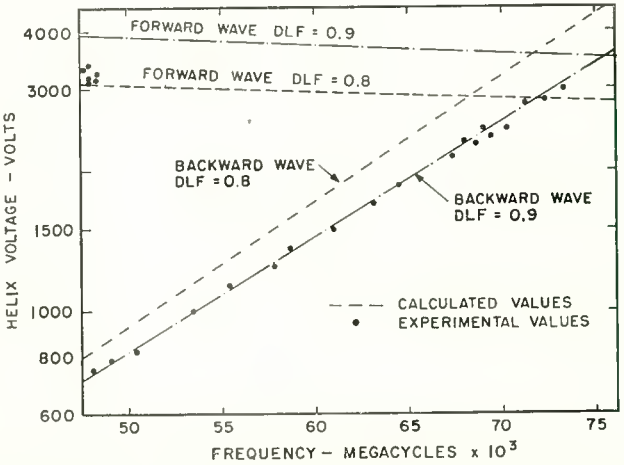


Fig. 6—Frequency of oscillation versus helix voltage for the A-1074.

* The Appendix outlines the calculations of the tuning curve, start-oscillation current, and power output.

dashed lines indicate the performance calculated on the basis of dielectric loading factors of 0.8 and 0.9.

Each of the calculated curves has two parts. The lower section, which has a positive slope and increases with frequency, is the voltage–frequency characteristic of the backward space harmonic wave on the helix with which interaction is desired. The upper branch, which has a nearly constant negative slope, is the voltage–frequency characteristic of the fundamental forward wave on the helix. The two curves intersect at that frequency at which the helix circumference is equal to one half of a wavelength in the presence of the dielectric. As long as the helix voltage is below the intersection value, the electron beam can interact only with the backward wave, but when the helix voltage is slightly greater than the intersection value, the electron beam can interact with either the backward wave or the forward wave. In the latter case the feedback necessary for oscillation is furnished by reflection from the collector end of the helix.

The experimental data in Figure 6 show that as the helix voltage is varied from 750 to 3,100 volts, all other parameters being held constant, the frequency of the r-f output signal varies from 48,000 to 74,000 megacycles. When the helix voltage is increased above 3,100 volts, the forward mode of oscillation takes over and the frequency jumps back to 48,000 megacycles. A comparison between the observed and calculated backward-wave oscillation curves indicates that the experimental DLF is 0.9; in the forward wave mode the DLF is closer to 0.8.

Start-Oscillation Current

The electron beam current required for the onset of oscillation in a backward-wave oscillator depends on the geometry of the beam and the length, loss, DLF of the r-f circuit, and the frequency. Figure 7 shows the start-oscillation current (measured on the collector) for the A-1073 plotted as a function of frequency. The dashed curve gives the calculated values of starting current, based on the experimentally measured DLF. Data taken on three different tubes are plotted. Improvement in the construction of successive tubes is reflected in their lower starting currents, and is probably due to better helix uniformity and beam alignment. In each case the starting currents follow the shape of the calculated curve, starting below 1 milliampere and rising slightly at the higher frequencies.

Figure 8 shows the start-oscillation current for the A-1074 plotted as a function of frequency. The dashed curve gives values of starting current calculated on the basis of the measured DLF, while the points

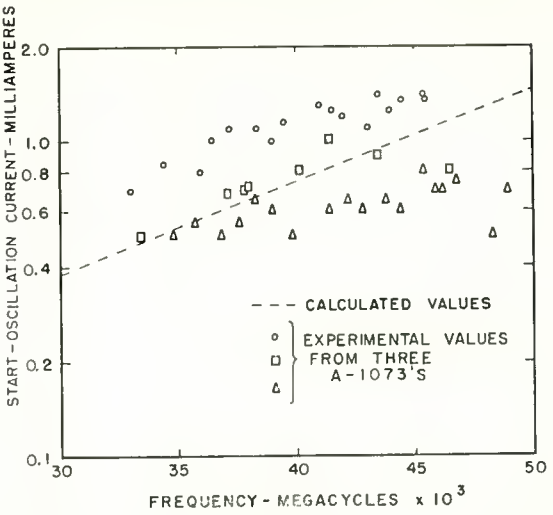


Fig. 7—Start-oscillation current versus frequency of oscillation for the A-1073.

show measured values. The agreement between calculation and experiment is not very good at the lower frequencies, but in every case the starting current lies well below 1 milliampere.

R-F Power Output

The r-f power output from a backward-wave oscillator tube is, of course, zero for electron beam currents below the start-oscillation value. As the beam current is increased beyond the start-oscillation

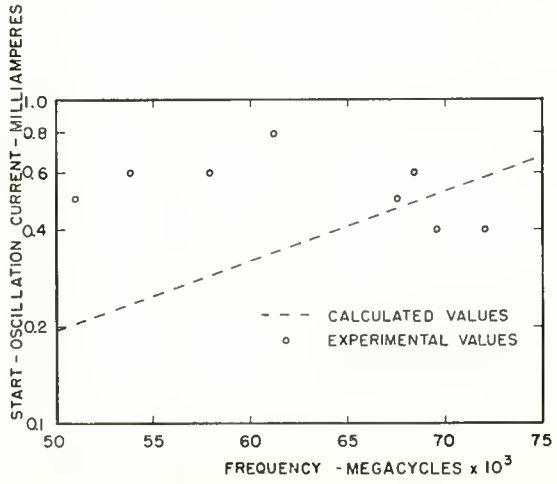


Fig. 8—Start-oscillation current versus frequency of oscillation for the A-1074.

value, the power output rises sharply at first and then levels off for currents above two or three times the starting current (see Figure 10, which is discussed below). The output power depends upon beam current and voltage, beam and helix geometry, and helix loss and dielectric loading. The output power to be expected from the A-1073 and A-1074 was calculated for beam currents of 3 milliamperes for the A-1073, and 2 milliamperes for the A-1074. The curves in Figure 9 show that the calculated power output of the A-1073 increases from 5 milliwatts at 30,000 megacycles to 13 milliwatts at 50,000 megacycles, and the calculated power output of the A-1074 rises from 1.2 milliwatts at 50,000 megacycles to 2.6 milliwatts at 74,000 megacycles.

Measurements of the power output of these tubes were made with a commercial bolometer. This instrument showed a power output of several milliwatts for the A-1073, and power of the order of a milliwatt for the A-1074. Although the power was by no means monotonic over the band, it was observed to be continuous. However, accurate power measurements at millimeter wavelengths are extremely difficult,[9] and commercially available bolometers are inefficient and are useful only to indicate relative power values.

Variation of R-F Power with Beam Current

Figure 10 shows the experimental values for the variation of r-f power output with beam current when all other parameters are held constant. Relative r-f power as measured with a bolometer is plotted as a function of the ratio of beam current to start-oscillation current. The power rises sharply to a point at which the ratio of beam current to starting current is between 2 and 3, and then levels off.

Variation of R-F Power with Focusing Magnetic Field

The output power from the A-1073 and A-1074 is sensitive to the intensity of the axial magnetic field applied to focus the electron beam. In Figure 11, the relative power output of the A-1073 is plotted against magnetic field strength. Relative beam transmission is shown in the same figure to make it clear that the loss of output power is not due to interception, but rather to the more subtle variation in the position of the electron beam. Evidently, the electrons flow closest to the helix over the greatest length for a magnetic field of 1,200 gauss. These data illustrate once again how essential it is that the electron beam be close to the circuit in millimeter wavelength tubes. Similar performance was observed in the A-1074, where, again, the optimum magnetic field was found to be 1,200 gauss. This value is remarkably low for

[9] W. M. Sharpless, "A Calorimeter for Power Measurements at Millimeter Wavelengths," *Trans. I.R.E. PGMTT*, p. 45, September, 1954.

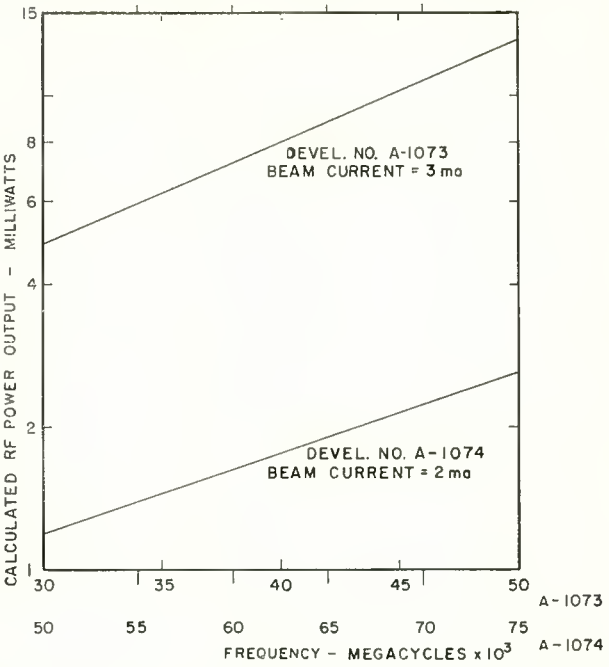


Fig. 9—Calculated r-f power output for the A-1073 and A-1074.

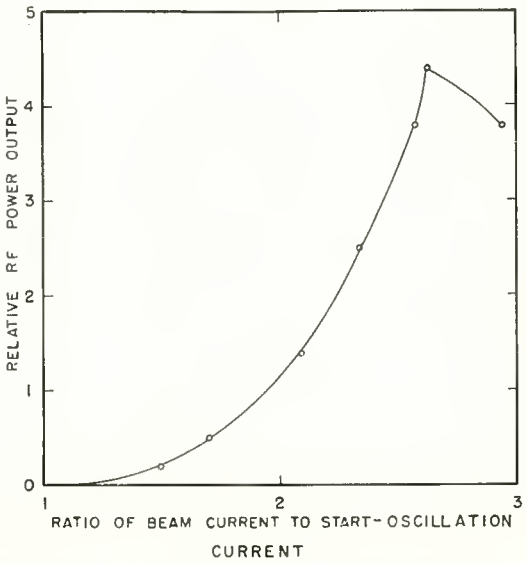


Fig. 10—Relative r-f power output versus the ratio of beam current to starting current for the A-1073 as measured on commercial bolometer.

tubes operating at millimeter wavelengths and is probably due to the simplicity of the mechanical features of the design.

Extrapolation to Higher Frequencies

Because the calculated and measured performance of the tubes were in fair agreement, the performance of tape-helix backward-wave oscillator tubes at higher frequencies can be predicted with some degree of confidence.

An extrapolation of the present design to a backward-wave tube which would operate in the frequency range between 90,000 and 150,000 megacycles results in a tube which would have a helix 0.6 inch long,

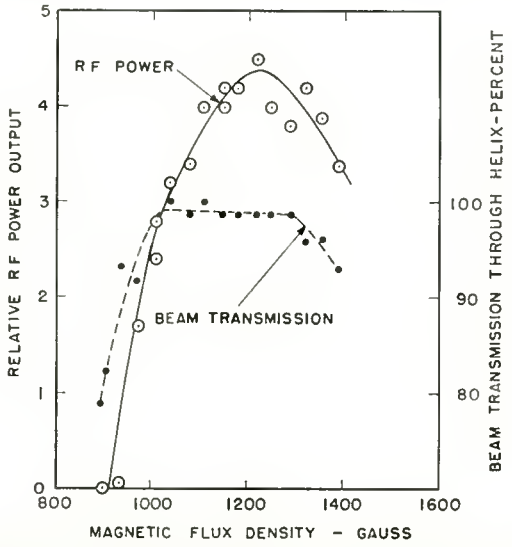


Fig. 11—Relative r-f power output and beam transmission through helix versus magnetic focusing field strength for the A-1073 as measured on commercial bolometer.

made of 0.001 by 0.0022 inch molybdenum tape wound at 218 turns per inch on a 0.0105 inch inside diameter.

The simplicity of the design makes this structure mechanically feasible. The tube should oscillate from 90,000 to 150,000 megacycles as the helix voltage is varied from 600 volts to 3,300 volts. Starting current is calculated to be 0.5 to 1.9 milliamperes over the frequency range. For a current density of 5 amperes per square centimeter (which is not impractical with impregnated cathodes), the collector current is 2.8 milliamperes and the expected power output is in the order of one milliwatt. It should be pointed out that while theory predicts certain power outputs, obtaining them over the entire operat-

ing range of an actual tube is another matter, and a great deal of developmental work must still be done before reliable performance and uniformity can be expected from this type of tube.

ACKNOWLEDGMENT

The authors are grateful to A. Cohen, F. E. Vaccaro, and H. J. Wolkstein who cooperated in the design and construction of these tubes.

APPENDIX—CALCULATION OF TAPE-HELIX BACKWARD-WAVE OSCILLATOR-TUBE PERFORMANCE

Calculation of helix voltage, starting current, and r-f power output as functions of frequency are described below. These quantities* can be determined reliably from the formulas if the physical properties of the tube, especially beam diameter and helix loss, are accurately known.

Helix Voltage versus Frequency

For oscillation, the velocity of the electron beam must be equal to $(v_{-1})_D$, the phase velocity of the first backward space-harmonic wave on the tape helix.

The subscript D denotes the presence of dielectric loading. Therefore, the beam voltage is given by

$$V_{osc} = \frac{1}{2\eta} (v_{-1})_D^2 \tag{1}$$

where η is the charge-to-mass ratio for the electron. The value of $(v_{-1})_D$ is found from the frequency, f, helix pitch, p, the dielectric loading factor, DLF, and the phase velocity of the fundamental wave on the *unloaded* helix, v_0, by the relation

$$\frac{2\pi f}{(v_{-1})_D} = \frac{2\pi f}{(DLF)\, v_0} - \frac{2\pi}{p} \tag{2}$$

or

$$|(v_{-1})_D| = \frac{(DLF)\, v_0}{\dfrac{(DLF)\, v_0}{pf} - 1}. \tag{3}$$

Therefore

* MKS units are used.

$$V_{osc} = 2.84 \times 10^{-12} \left(\frac{(DLF)\, v_0}{\frac{(DLF)\, v_0}{pf} - 1} \right)^2 \tag{4}$$

The value of v_0 is quite accurately given by

$$v_0 = \frac{cp}{2\pi a}, \tag{5}$$

where c is the velocity of light in free space, p is the helix pitch, and a is the mean radius of the helix. A correction to Equation (5) is given by Pierce.[10]

The helix voltage required for synchronism between the electron beam and the forward wave is

$$\begin{aligned} V_{\text{forward}} &= \frac{1}{2\eta} (DLF)\, v_0^2 \\ &= 2.84 \times 10^{-12} (DLF)\, v_0^2. \end{aligned} \tag{6}$$

Start-Oscillation Current

The current required to start oscillation depends on the length of the helix, the helix impedance, and the distributed loss on the helix. The frequency dependence of these parameters makes the starting current a function of frequency. Harman[11] gives the value of CN required for onset of oscillation as a function of distributed helix loss. N is the length, l, of the helix in backward wavelengths (λ_{-1}),

$$N = \frac{l}{\lambda_{-1}} = \frac{lf}{v_{-1}}, \tag{7}$$

and C is the helix gain parameter

$$C = \sqrt[3]{\frac{K_{-1} I}{4 V_{osc}}}. \tag{8}$$

In this expression, I is the beam current, V_{osc} is the helix voltage corresponding to the oscillation frequency involved, and K_{-1} is the

[10] J. R. Pierce, *Traveling Wave Tubes*, D. Van Nostrand Co., Princeton, N. J., 1950; Fig. 3.3, p. 25.

[11] W. A. Harman, "Backward Wave Interaction in Helix-type Tubes," *TR-13*, Electronic Research Lab., Stanford University, p. 3, April 26, 1954.

interaction impedance which the helix presents to the beam. Watkins and Ash[12] plot K_{-1} as a function of frequency, and show how to correct its value for dielectric loading effects.

$$(K_{-1})_D = (K_{-1})_{\text{UNLOADED}} \left[\frac{1 - \dfrac{2\pi fa}{c}}{(DLF)^2 - \dfrac{2\pi fa}{c}} \right]^2 \tag{9}$$

A further correction to K_{-1} to allow for the relationship between beam and helix diameters, is given by Harman[13]. Then

$$I_{\text{START } OSC} = \frac{4[(CN)_{\text{START}}]^3 V_{OSC} (\mathring{v}_{-1})_D^3}{K_{-1} l^3 f^3}. \tag{10}$$

Power Output

The power output is computed from the efficiency expression of Grow and Watkins[14]

$$\text{Efficiency} = (4/3)\ RC,$$

where R is an efficiency reduction factor (plotted as a function of helix loss by Grow and Watkins) and C is the helix gain parameter used above. R-F power output is therefore

$$P_{rf} = (4/3)\ RCIV_{OSC}. \tag{11}$$

This expression is valid at twice the start-oscillation current.

---

[12] D. A. Watkins and E. A. Ash, "The Helix as a Backward-Wave Circuit Structure," *Jour. Appl. Phys.*, Vol. 25, p. 782, June, 1954.

[13] W. A. Harman, *op. cit.*, p. 60.

[14] R. W. Grow and D. A. Watkins, "Backward-Wave Oscillator Efficiency," *Proc. I.R.E.*, Vol. 43, p. 848, July, 1955.

MEASUREMENT OF TRANSISTOR CHARACTERISTICS IN THE 3-250 MEGACYCLE FREQUENCY RANGE*

By

John H. O'Connell and Thomas M. Scott

RCA Laboratories,
Princeton, N. J.

Summary—The transistor characteristics which are of especial concern to the engineer applying transistors in HF-VHF circuits are: input, output, and unilateralization impedances; transconductance; and current gain.

Techniques and special apparatus which have been developed for the measurement of these characteristics, at frequencies up to 250 megacycles, are described in this paper. Commonly available commercial test instruments are used to the greatest practical extent.

Introduction

THE particular transistor characteristics which are of interest in the design of HF-VHF circuits, such as the r-f and i-f portions of a television or FM receiver, are defined as follows:

(1) *Input impedance components:* the shunt resistive component, R_{in}, and the shunt reactive component, X_{in}, of incremental input impedance, for either the common-emitter or common-base connection, with the collector-to-common voltage held constant (collector shorted to common at the signal frequency).

(2) *Output impedance components:* the shunt resistive component, R_{out}, and the shunt capacitive component, C_{out}, of incremental collector-to-common impedance, with the base-to-emitter voltage held constant. This impedance is the same for both the common-emitter and common-base connections.

(3) *Unilateralization impedance components:* the series resistive and capacitive components of the common-emitter unilateralization impedance. This impedance, when placed between the base terminal and a source of signal voltage equal and opposite to that at the collector, produces an input impedance at the base terminal which is

* Manuscript received July 16, 1958.

independent, over a restricted frequency range, of collector load impedance. Unilateralization also renders the output impedance independent of input source impedance. The input and output impedances become those discussed in (1) and (2) above, except for the incidental shunting effect of whatever practical unilateralization circuit components are used.

(4) *Transconductance:* the magnitude of the incremental ratio of collector current to base-to-emitter voltage, with either the collector-to-base or the collector-to-emitter voltage held constant. In the measurement procedures to be described the former connection is used (base common). While theoretically some difference could exist between measurements made in the two connections, for practical purposes it may be assumed that the common-emitter transconductance will be identical to the common-base transconductance.

(5) *Current gain:* the magnitude of the incremental ratio of collector current to emitter current with the collector-to-base voltage held constant. A knowledge of this ratio, or more particularly the manner in which it varies with frequency, is a prerequisite to analysis of the transistor as a device.

Items (1), (2), (3), and (4) bear directly on circuit design. The available unilateralized gain, in terms of these quantities, is given by:

$$\text{Power Gain} = \frac{(\text{Transconductance})^2 R_{\text{in}} R_{\text{out}}}{4}. \qquad (1)$$

The usable gain of a tuned amplifier is usually determined by considerations relating to stability and transistor interchangeability and, depending on the tolerance allowed in unilateralization, may be significantly lower than that obtained from the above expression.[1]

A figure-of-merit[2] for the transistor itself is the frequency at which the available common-emitter power gain falls to unity. This is, of course, the maximum frequency of oscillation in a loss-free circuit, and is given by

$$f_{\max} = \frac{1}{2}\sqrt{f_\alpha \frac{1}{2\pi R_{bb'} C_{b'c}}}. \qquad (2)$$

[1] T. O. Stanley and D. D. Holmes, "Stability Considerations in Transistor IF Amplifiers," *Transistors I*, RCA Laboratories, Princeton, N. J., 1956, p. 403.

[2] L. J. Giacoletto, "Study of PNP Alloy Junction Transistors from D-C Through Medium Frequencies," *RCA Review*, Vol. XV, p. 506, December, 1954.

The factor f_a, in the foregoing expression, is termed the alpha-cutoff frequency, and is the frequency at which the common-base current gain has declined to 0.707 of its low-frequency value (3 decibels down). The base lead resistance, $R_{bb'}$, which is not directly measurable, can be inferred from

$$R_{bb'} = \frac{f_a}{g_m f_g}, \tag{3}$$

where g_m is the intrinsic transconductance (38 millimhos per milliampere of collector current at 25°C in an ideal transistor) and f_g is the extrinsic transconductance-cutoff frequency, the frequency at which the transconductance, as measured at the transistor terminals, has declined to 0.707 of its low-frequency value. Alternatively, $R_{bb'}$ can be assumed to be the input resistance measured at a frequency so high that the intrinsic base-to-emitter capacitive reactance is low compared to the base-lead resistance; that is, at frequencies well above the intrinsic transconductance cutoff frequency. The collector-to-internal-base capacitance, $C_{b'c}$, appearing in Equation (2) can be assumed to be the difference between the transistor common-base output capacitance at zero emitter current, and the incidental shunt capacitance in the transistor mount. Typical values for the incidental capacitance range from 0.5 to 1.0 micromicrofarad.

Input and Output Impedance

Transistor input and output impedance components can be measured conveniently and accurately at frequencies up to 250 megacycles using a Boonton type 250-A RX Meter. Both the resistive and reactive components of output impedance are generally within the range of this meter and may be measured with full meter sensitivity (from 0.1 to 0.5 volt r-m-s across the transistor), provided the collector bias voltage is greater than about three volts.

For input impedance measurements, the voltage across the RX meter terminals is reduced so as not to overdrive the transistor. A simple modification of the RX meter, as recommended by the manufacturer, allows for adjustment of the signal voltage appearing at the meter terminals. To find the small-signal input impedance, the signal drive applied to the transistor is reduced by means of the added adjustment knob until the meter balance stops changing.

Adaptors (Figures 1 and 2) are used with the RX meter for input and output impedance measurements. These adaptors are arranged for common-emitter test. Short leads are used in signal-carrying

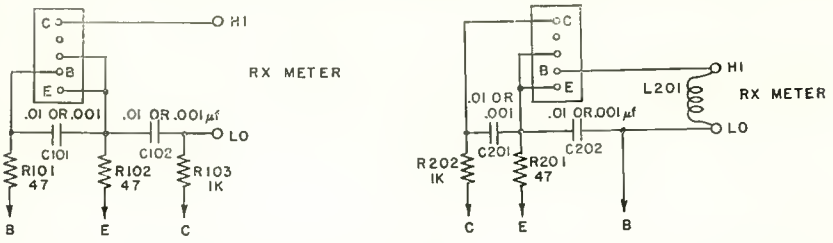


Fig. 1—(a) Output impedance adaptor; (b) input impedance adaptor.

circuits. Two different values of by-pass capacitance are used in the adaptors, as shown in the schematics. The higher value capacitors are suitable at frequencies from 2 to 20 megacycles, the lower value capacitors from 20 to 250 megacycles.

The range of reactive balance of the RX meter is +20 to —100 micromicrofarads. A transistor input capacitance in excess of 20 micromicrofarads necessitates an auxiliary reactance of opposite sign across the meter terminals, shown as L201 in Figure 1, to enable offsetting the meter reference balance from zero on the capacitance scale.

Unilateralization Impedance

The series resistive and capacitive components of the unilateralization impedance are measured using a specially designed test set in conjunction with a signal generator, a receiver, and an oscilloscope. The test set, shown in Figures 3 and 4, is also used in measurement of transconductance and current gain.

Fig. 2—Input and output impedance adaptors.

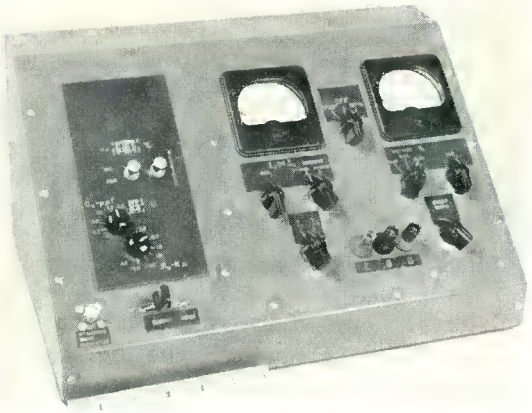


Fig. 3—Front view of transistor test set.

In the unilateralization measurement, a CW signal generator, set to about 100 millivolts output, is coupled via a relatively high impedance to the base of the transistor under test, as shown in Figure 5. The emitter is common. The collector load consists of a tightly coupled phase-reversing transformer tuned by the capacitance of a reverse-biased junction diode (the collector junction of a 2N139 transistor). The reversed-phase section of the transformer is coupled back to the transistor base through a calibrated, adjustable, series R-C unilateralizing impedance. The tuning of the collector circuit is varied back and forth past the fixed frequency of the input signal by varying the magnitude of the bias applied to the diode, and thus its capacitance,[3]

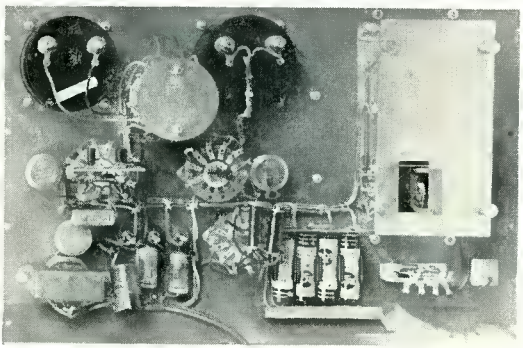


Fig. 4—Rear view of transistor test set.

[3] J. H. O'Connell and L. J. Giacoletto, "A Variable Capacitance Germanium Junction Diode for UHF," *RCA Review*, Vol. XVII, p. 68, March, 1956.

at a 60-cycle sweep rate. The resulting variation of collector load impedance produces a simultaneous variation in base input impedance if the circuit is imperfectly unilateralized. Such impedance variation results in a variation in signal voltage at the base. The base voltage is monitored with a communications-type receiver. The wave-form of the receiver a-f output is a replica of the detected signal envelope, and is observed on an oscilloscope, using a 60-cycle, variable phase, sine-wave sweep. Figure 6 shows waveforms corresponding to various degrees of unilateralization. In these oscillograms, the forward and reverse traces are not quite superimposed because of low frequency phase shift.

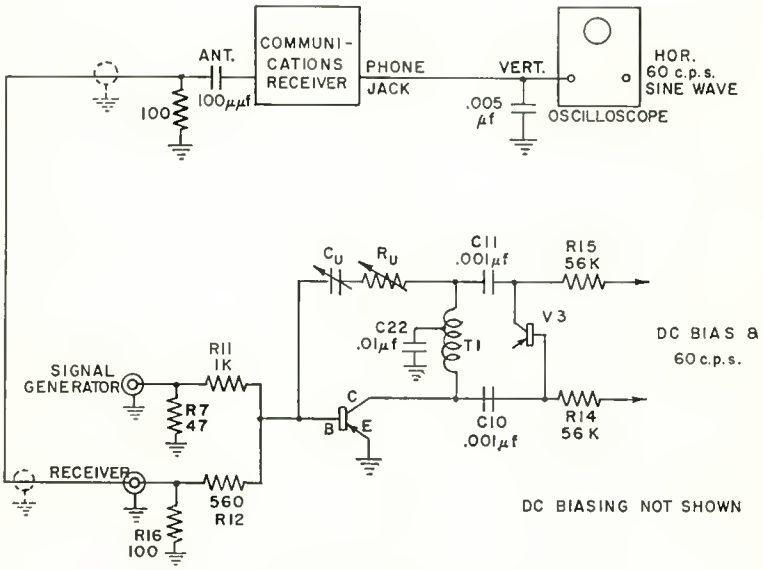


Fig. 5—Unilateralization test circuit.

Transconductance

Transconductance may be measured at a low signal frequency using the simple circuit shown in Figure 7. An adaptor incorporating this circuit is shown in Figure 8. In Figure 7, the input and output resistance values indicated have been chosen so that the transconductance, in mhos, is simply the ratio of v_2 to v_1. The input voltage, v_1, should not be so high as to overdrive the transistor; a maximum value of 0.1 volt is suggested. An audio signal generator set to about 2,000 cycles is considered a suitable signal source. Input and output voltages, v_1 and v_2, may be measured successively, using any of several commer-

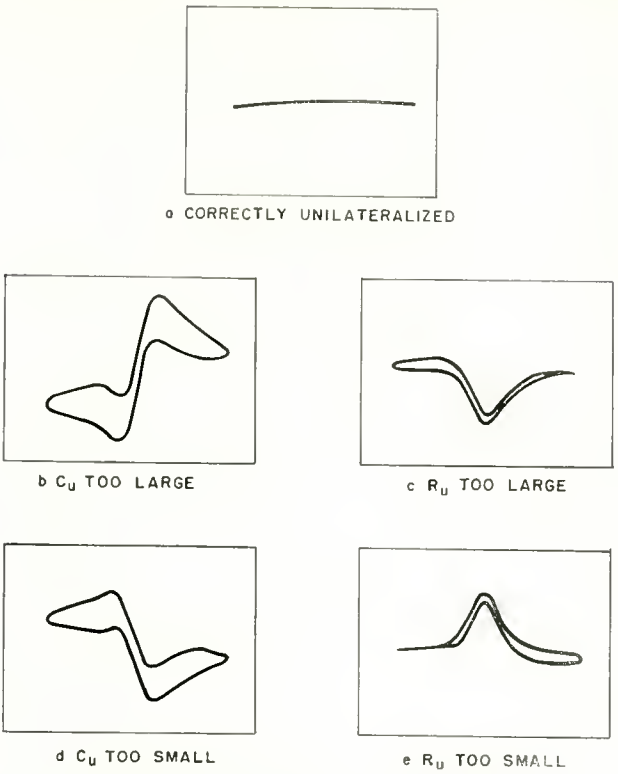


Fig. 6—Unilateralization test oscillograms.

cially available multi-range electronic voltmeters. Alternatively, the voltage at v_2 can be set to a suitable indicated value such as 0.04 volt, the voltmeter transferred to v_1, and the v_2 reading reproduced at v_1 by adjustment of a calibrated attenuator on the signal generator. The transconductance is then the voltage ratio indicated by the difference in attenuator settings, both of which should be at least 20 decibels below maximum attenuator output to avoid possibility of differential loading error.

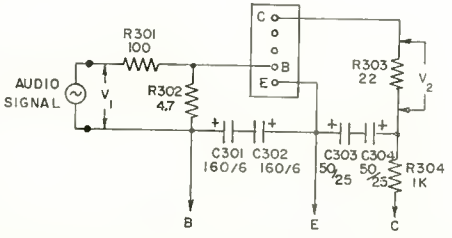


Fig. 7—Audio frequency transconductance test circuit.

Fig. 8—Audio frequency transconductance test adaptor.

At high frequencies, transconductance is measured on a relative basis as a function of frequency. The circuit used is shown in Figure 9. This circuit is incorporated in the test set of Figure 3.

A 30-millivolt, 90 per cent amplitude-modulated signal is connected to the input jack. The input is terminated by the bolometer, R6, and the voltage divider R7, R8, R9. The bolometer is an envelope detector, used to monitor the amplitude of the signal at the input to the voltage divider. The assumption is made that the input impedance of the transistor is high at all frequencies compared to the 6-ohm shunt section of the voltage divider, and that the voltage input to the transistor is, therefore, independent of frequency. (This is generally the case for high-frequency transistors operated at emitter currents less than a few milliamperes. Input impedance measurements, as previously described, will disclose any instance in which correction factors might be needed, and will provide the data for estimating such correction factors.)

A low-impedance path for collector current is provided by bolo-

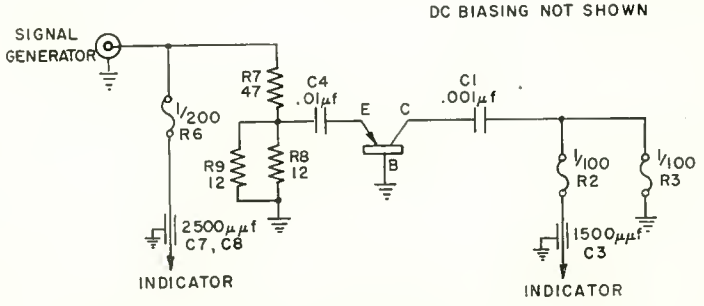


Fig. 9—HF-VHF transconductance test circuit.

meters R2 and R3, which are in parallel for signal frequencies, and which serve as an envelope detector for the collector current. The ratio of the signal modulation components available from the input and output bolometers is the measure of the relative transconductance at the frequency of measurement. Such ratios, obtained at frequency intervals of several megacycles, and plotted as decibels versus frequency, give a complete picture of the transistor transconductance characteristic. Such a curve is shown in Figure 10. The low-frequency, flat portion of the curve has the absolute transconductance value obtainable by measurement at a-f, as previously described.

Bias is applied sequentially to the input (R6) and output (R2, R3) bolometers from a Hewlett Packard 415B Standing-Wave Indicator,

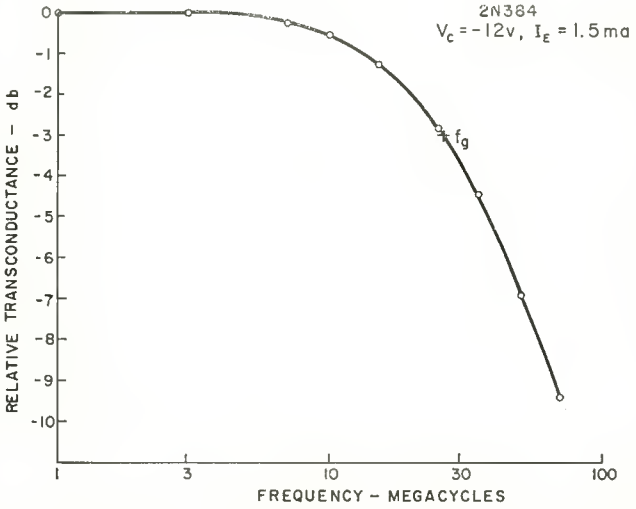


Fig. 10—Relative transconductance versus frequency plot.

which also provides narrow-band amplification and decibel-level indication of the detected signal. Input bolometer R6, a 1/200 ampere Littelfuse, is biased at approximately 4 milliamperes (the "high" current position on the HP 415B), at which current it exhibits approximately 500 ohms resistance. Output bolometers R2 and R3 are 1/100 ampere Buss fuses, and are series-biased at about 7 milliamperes (on the "high" position), at which current their parallel resistance is about 56 ohms. Collector signal current is coupled to the bolometers via capacitor C1, which is of sufficiently high audio impedance to divert any incidental a-f component of collector current from the bolometers and output indicator.

The signal generator used in the transconductance test should

preferably be one capable of being modulated at 80-90 per cent, so that operation well above the noise level is possible without using excessive signal drive on the transistor. Increasing the modulation from 30 to 90 per cent improves the signal-to-noise ratio by 4.7 decibels. A further improvement of about 2 decibels is obtainable by substituting a 400-cycle filter (available from the manufacturer) for the 1,000-cycle filter normally supplied in the standing-wave indicator, and using 400-cycle signal modulation. This latter improvement relates to the thermal time-constants of the bolometers.

Both the magnitude of transconductance and its frequency dependence change with emitter current, so care must be taken to maintain the emitter current at the desired value.

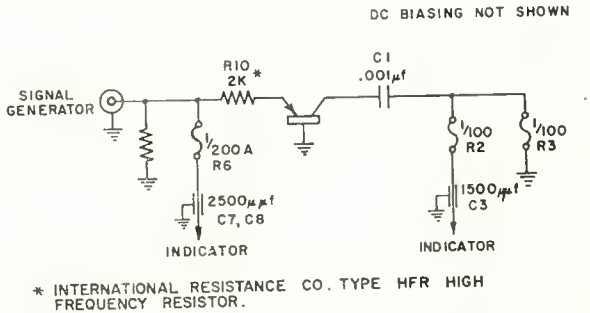


Fig. 11—HF-VHF current gain test circuit.

Current Gain

The method of measurement of relative common-base current gain as a function of frequency is quite similar to that described for transconductance measurement, and requires the same apparatus. The current gain test circuit in the test set is shown in Figure 11. This differs from the relative transconductance test circuit only in that a constant-current signal source is substituted for the constant-voltage source. In Figure 11, the 2000-ohm resistance, R10, is assumed to be invariant with frequency, and high compared to the transistor input impedance, so that with constant signal voltage applied, the emitter current is likewise constant. About 0.1 volt of 90 per cent modulated signal is used. As in the case of transconductance measurement, the transistor input is held constant by monitoring the signal detected by the input bolometer, R6. The required precision of measurement generally is not obtainable using the signal generator output indication for setting the transistor input at the successive frequencies.

The alpha-cutoff frequency can be found by a successive approximation approach, or by plotting the collector current as a function of frequency. Since the current gain is measured on a relative basis, factors such as the shunting effect of the resistances through which biases are supplied are of no concern as long as such factors are not frequency dependent. To obtain the current gain on an absolute basis, it may be assumed that the low-frequency current gain is unity. The plot of collector current versus frequency then transforms directly to a plot of common-base current gain versus frequency. Such a plot, obtained by measurement of a 2N384 drift transistor, is shown in Figure 12.

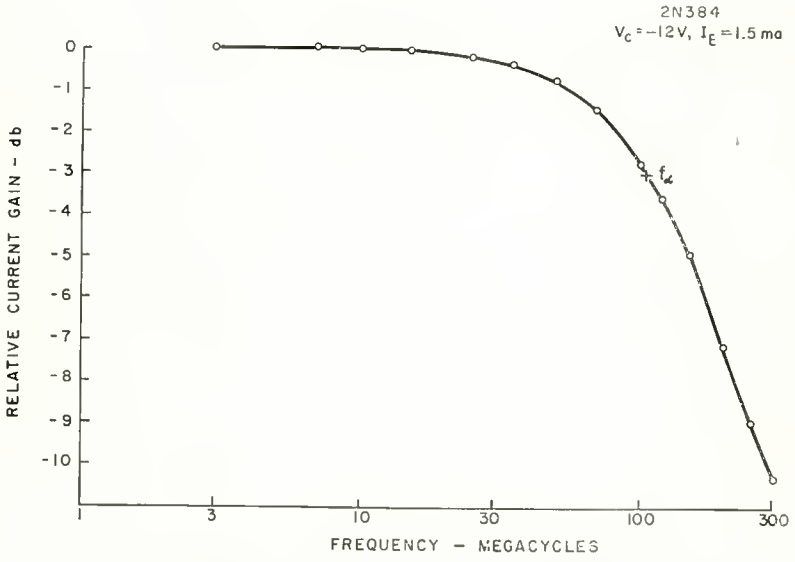


Fig. 12—Relative current gain versus frequency plot.

Correction Factor

When the measurement frequency, the collector load resistance, and the capacitances associated with the collector, have values such that an appreciable portion of the collector current is by-passed or fed back, a correction factor should be added to the indicated current gain and transconductance. It is shown in the appendix that in the case of grounded-base current gain measurement, the effect of feedback can be neglected. The numerical value of the first-order-approximation

current-gain correction factor, F, which accounts only for by-passing, is

$$F = \left| \frac{i_c}{i_{bol}} \right| = \sqrt{1 + (\omega R_L C)^2}, \tag{4}$$

where i_c is the collector signal current, i_{bol} is the signal current delivered to the output bolometer, ω is the angular frequency of measurement, R_L is the collector load resistance, and C is the collector-plus-incidental capacitance shunting the transistor output. This correction factor, expressed in decibels, is plotted in Figure 13 for the particular load resistance value, 56 ohms, which occurs in the circuit here discussed. The indicated number of decibels is added to the decibel-level

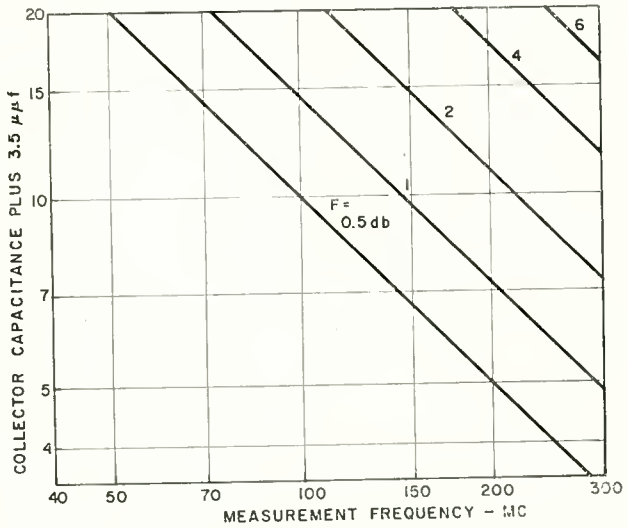


Fig. 13—Correction factor for excess collector capacitance.

indication of the standing-wave indicator. In entering the plot, the test-set incidental capacitance, 3.5 micromicrofarads, is added to the transistor collector capacitance.

It is shown in the appendix that in transconductance measurements, feedback to the transistor base lead resistance alters the apparent value of collector capacitance as a function of frequency. The equivalent shunt value of the feedback capacitance, $C_{b'c}$, is given by

$$C_{\text{equiv}} = C_{b'c} \left[1 + \frac{g_m R_{bb'}}{\sqrt{1 + \left(\frac{f}{f_g} \right)^2}} \right], \tag{5}$$

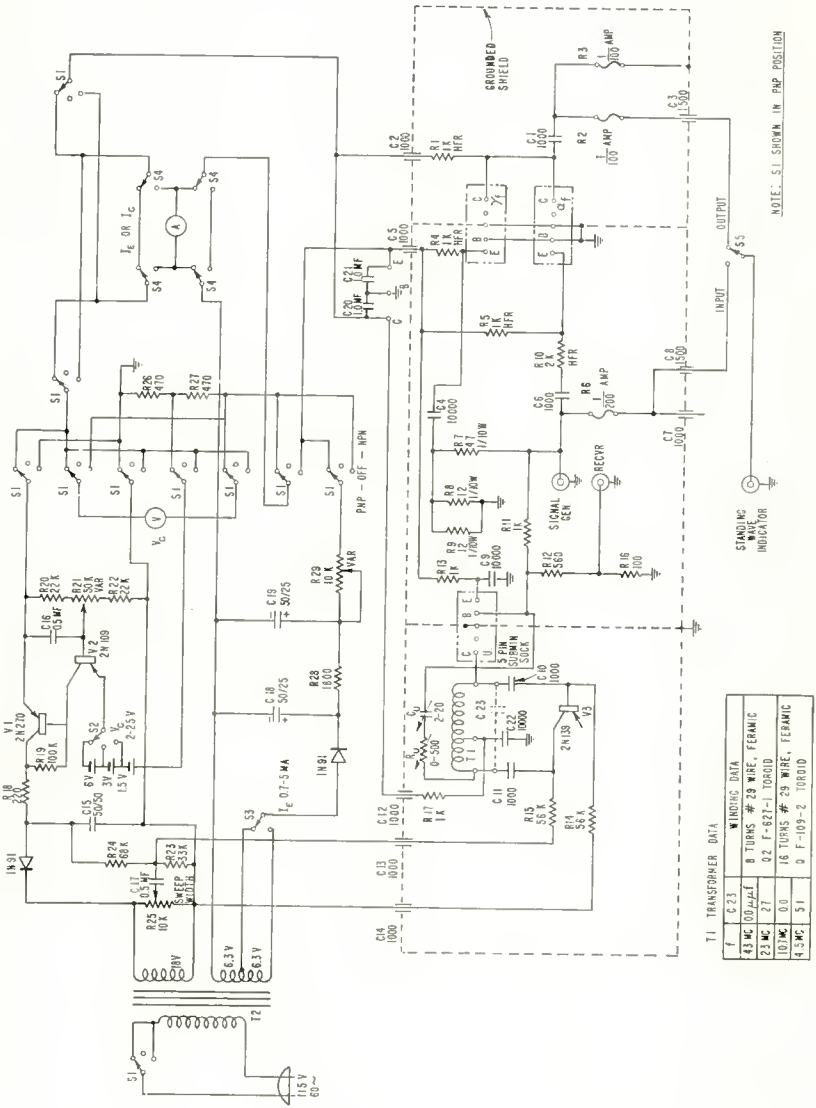


Fig. 14—Transistor test set circuit diagram.

in which f is the measurement frequency. C_{equiv} is added to the incidental transistor mount capacitance (typically 0.5 to 1.0 micromicrofarads) and the test set incidental capacitance (3.5 micromicrofarads) to find the ordinate of the correction factor in Figure 13. The factor f_g in Equation (5) is the true transconductance-cutoff frequency which cannot be known until after the correction factor is applied. However, the measured transconductance cutoff frequency can be substituted with small loss in precision.

Test Set

The test set consists of two principal sections: (1) an r-f compartment containing the unilateralization, transconductance, and current-gain test circuits, and (2) a power supply which provides

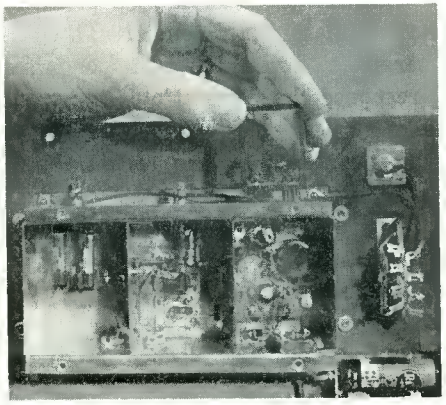


Fig. 15—Close-up view of r-f box.

biases for the transistor under test. The complete circuit diagram of the test set is shown in Figure 14.

The r-f compartment is shown in Figure 15. A separate transistor socket is used for each type of test. Partitions in the box bridge all sockets for input–output isolation, and provide direct low-impedance grounds for transistor shield terminals, and for base terminals of the transconductance and current-gain sockets.

The phase-reversing transformer (shown in Figure 15) is wound on a ferrite toroid core. For operation at the television intermediate frequency, 44 megacycles, the transformer is tuned by the collector, sweep diode, and incidental circuit capacitance. By plugging in a shunt capacitor (C23), this transformer can be tuned to 24 megacycles. Similar components provide for operation at the FM and television intercarrier i-f frequencies, 10.7 and 4.5 megacycles, respectively.

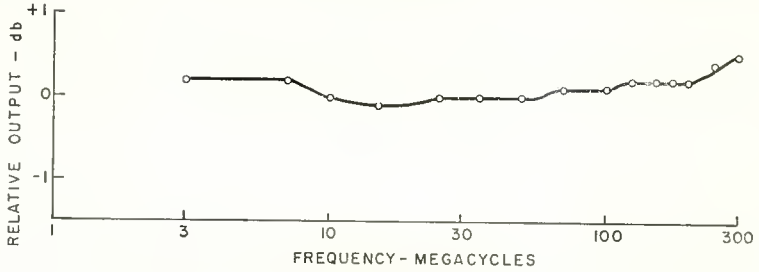


Fig. 16—Frequency response characteristic of the transconductance test circuit.

The phase reversing transformers have a collector-to-feedback section voltage ratio of seven to one. Correspondingly, the unilateralization adjusting controls, R_u and C_u, have scales indicating, respectively, seven times and one seventh the values at which they are set, so that when adjusted for the unilateralized condition, they indicate the values of the series unilateralization impedance components as defined earlier.

The frequency-response characteristic of the transconductance test circuit can be checked by plugging a minimum-lead-length 1,000-ohm high-frequency type resistor into the socket, between emitter and collector terminals. A typical characteristic so obtained is shown in Figure 16.

The current-gain test circuit was checked using a "short" between emitter and collector terminals of the socket. The resulting frequency-response curve is shown in Figure 17. The increased response at high frequencies results from a decrease in impedance of the emitter-current-limiting resistor, R10. This amount of deviation from the ideal flat response is not considered significant enough to warrant use of a correction factor.

Power Supply

The power supply operates from the a-c line. It provides two

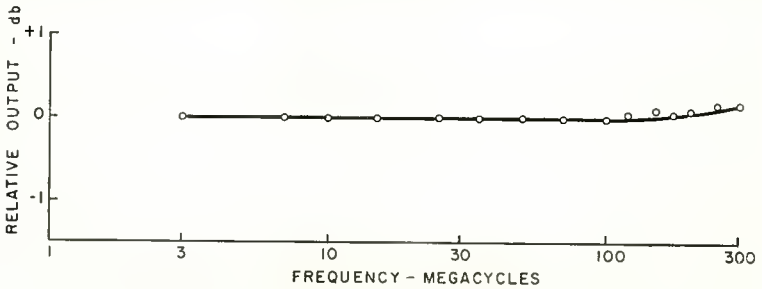


Fig. 17—Frequency response characteristic of the current gain test circuit.

independently adjustable outputs. The collector voltage supply is adjustable from 2 to 25 volts and is regulated. The emitter current supply, which is unregulated, is adjustable from 0.7 to 5.0 milliamperes. A PNP-NPN switch (S1) provides for polarity reversal. With switch S1 set at either PNP or NPN, the power supply output is available, at binding posts on the panel, for biasing a transistor on which input or output impedance measurements are being made. In the OFF position of this switch, an external bias source can be connected to the binding posts to take care of requirements beyond the ranges of the internal supplies.

The collector voltmeter, V_c, if returned to common (ground), would read in error by the amount of the IR drop in the 1,000-ohm collector r-f isolating resistor (R1 or R17). Since the emitter current closely approximates the collector current, a correction is obtainable by returning the voltmeter to a point similarly "above" common potential in the emitter current return path. R26 and R27 provide this compensation. In PNP tests, the voltage developed across R26 is placed in series with the regulator reference voltage. This produces a varia-

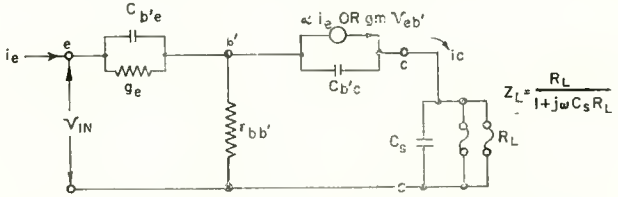


Fig. 18—Transistor common base equivalent circuit.

tion in output voltage of the collector bias supply, as a function of transistor current, of such direction and magnitude as to maintain the bias voltage constant *at the collector*. This convenient feature is not available in NPN tests.

Approximately 9 volts d-c reverse bias for the variable capacitance diode, V3, in the unilateralization test circuit, is obtained from the input of the collector bias supply at the tap between R23 and R24. The a-c sweep voltage for the diode is coupled in through C17 from a 10,000-ohm potentiometer (R25) across the 18-volt transformer secondary.

ACKNOWLEDGMENT

The authors would like to express their appreciation to A. A. Barco, T. O. Stanley, L. A. Freedman, and C. G. Seright for valuable discussions and suggestions concerning the problems and for editorial assistance in the composition of the paper.

Appendix

The equivalent shunt value of the feedback capacitance which is to be used in the correction factor of Equation (4) is obtained from an analysis of the transistor common-base equivalent circuit shown in Figure 18. The circuit elements of the equivalent circuit are defined as follows: $C_{b'e}$ is the sum of the emitter diffusion and transition capacitances, $C_{b'c}$ is the collector transition capacitance, C_s is the incidental capacitance which shunts the transistor output and includes transistor mount and socket capacitances and circuit capacitances, $g_e = (q/kT)I_e$ is the emitter junction conductance, $r_{bb'}$ is the base-lead resistance, α is the transistor common-base current gain, and R_L is the parallel resistance of the two bolometers used to measure collector current.

The two nodal equations below which are used to calculate transistor current gain and transconductance are obtained by equating the currents entering and leaving nodes b' and c.

$$i_e = V_{b'}\,(g_{bb'} + j\omega C_{b'c}) + \alpha i_e - j\omega C_{b'c}V_c\,, \tag{6}$$

$$\alpha i_e = -j\omega C_{b'c}V_{b'} + V_c(j\omega C_{b'c} + Y_L)\,. \tag{7}$$

Transistor current gain, which is defined as the ratio of collector current, i_c, to emitter current, i_e, can be found by simultaneous solution of Equations (6) and (7). If the substitution $V_c = i_c Z_L$ is made, then the ratio of i_c to i_e is

$$\frac{i_c}{i_e} = \frac{(\alpha + j\omega C_{b'c}r_{bb'})}{1 + j\omega C_{b'c}Z_L + j\omega C_{b'c}r_{bb'}}\,. \tag{8}$$

For most transistors, $j\omega C_{b'c}\, r_{bb'} \ll 1$ up to the alpha-cutoff frequency and can be ignored. Equation (8) can now be written

$$\frac{i_c}{i_e} = \frac{\alpha}{1 + j\omega C_{b'c}Z_L} = \frac{\alpha}{1 + \dfrac{j\omega C_{b'c}R_L}{1 + j\omega C_s R_L}}\,. \tag{9}$$

The relationship between the measured current, i_{bol}, and the collector current, i_c, is

$$i_c = i_{bol}\,(1 + j\omega C_s R_L)\,. \tag{10}$$

If Equation (10) is substituted into Equation (9),

$$\frac{i_{bol}}{i_c} = \frac{\alpha}{1 + j\omega R_L (C_{b'c} + C_s)}, \tag{11}$$

or

$$\left| \frac{i_{bol}}{i_c} \right| = \frac{\alpha}{\sqrt{1 + \omega^2 R_L^2 (C_s + C_{b'c})^2}} = \frac{\alpha}{F}. \tag{12}$$

It may be seen from Equation (12) that the ratio of bolometer current to input current multiplied by the correction factor, F, gives the magnitude of transistor current gain. The equivalent value of shunt capacitance to be used in the correction factor is the sum of collector capacitance plus incidental capacitance which shunts the transistor output.

Equations (6) and (7) can be simultaneously solved to give the ratio of output current, i_c, to the input voltage, V_{in}, which is defined as transistor transconductance. To eliminate the input current from Equations (6) and (7), the substitutions $i_e = Y_e(v_{in} - v_{b'})$ and $\alpha i_e = g_m (v_{in} - v_{b'})$ were made. The simultaneous solution of Equations (6) and (7) for transistor transconductance is

$$\frac{i_c}{v_{in}} = \frac{g_m \dfrac{j\omega C_{b'e}(g_m - j\omega C_{b'c})}{g_{bb'} + j\omega C_{b'e} + j\omega C_{b'c}}}{1 + j\omega C_{b'c} Z_L + \dfrac{j\omega C_{b'c} Z_L (g_m - j\omega C_{b'c})}{g_{bb'} + j\omega C_{b'e} + j\omega C_{b'c}}}. \tag{13}$$

Equation (13) can be simplified by assuming that $C_{b'c} \ll C_{b'e}$ and $g_m \gg j\omega C_{b'c}$ up to the transconductance cutoff frequency, f_g. Equation (13) simplified is then

$$\frac{i_c}{v_{in}} = \frac{\dfrac{g_m}{1 + j\omega C_{b'e} r_{bb'}}}{1 + j\omega C_{b'c} \dfrac{R_L}{1 + j\omega C_s R_L} \left(1 + \dfrac{g_m}{g_{bb'} + j\omega C_{b'e}} \right)}. \tag{14}$$

The numerator of Equation (14) is the intrinsic transconductance of the transistor and the denominator represents a measurement error caused by a finite value of load resistance which is to be corrected by Equation (4). The ratio of measured current, i_{bol}, to input voltage is

obtained by substituting the relationship of Equation (10) into Equation (14):

$$\frac{i_{bol}}{v_{in}} = \frac{\dfrac{g_m}{1 + j\omega C_{b'c} r_{bb'}}}{1 + j\omega R_L \left[C_s + C_{b'c} \left(1 + \dfrac{g_m r_{bb'}}{1 + j\omega C_{b'c} r_{bb'}} \right) \right]} . \tag{15}$$

$$\left| \frac{i_{bol}}{v_{in}} \right| = \frac{\dfrac{g_m}{\sqrt{1 + \left(\dfrac{f}{f_g} \right)^2}}}{\sqrt{1 + \omega^2 R_L{}^2 \left[C_s + C_{b'c} \left(1 + \dfrac{g_m r_{bb'}}{\sqrt{1 + \left(\dfrac{f}{f_g} \right)^2}} \right) \right]^2}} \tag{16}$$

$$= \frac{g_m}{F \sqrt{1 + \left(\dfrac{f}{f_g} \right)^2}}$$

From Equation (16) the value of shunt capacitance to be used in the correction factor is

$$C = C_s + C_{b'c} \left(1 + \frac{g_m r_{bb'}}{\sqrt{1 + \left(\dfrac{f}{f_g} \right)^2}} \right) , \tag{17}$$

and Equation (16) multiplied by the correction factor will give the ratio of measured output current to input voltage equal to the transistor transconductance.

GENERATION OF SECOND HARMONIC IN A VELOCITY-MODULATED ELECTRON BEAM OF FINITE DIAMETER*

By

F. Paschke

RCA Laboratories,
Princeton, N. J.

Summary—This second-order computation reveals that the amplitudes of the second-harmonic current and velocity do not in general vary periodically with distance. The maxima of second-harmonic current are closer to the input cavity than the corresponding maxima of the fundamental current. At very small beam diameters, the second harmonic exhibits an instability and both the current and velocity amplitudes rise with distance. This can be interpreted as a parametric oscillation where the two first-order Hahn-Ramo waves serve as low-frequency-pumping sources. The analysis agrees, at least qualitatively, with experimental results reported recently by Mihran.

Introduction

In klystron frequency multipliers and high-power klystron amplifiers a knowledge of harmonic generation is required. One-dimensional theories[1,2] of velocity-modulated electron beams yield, for small signals, the convection-current densities and velocities shown in Figure 3. The distributions are periodic for both the fundamental frequency and the second harmonic. The maximum convection-current density of the second harmonic is predicted to appear at the same distance from the input cavity as the maximum convection-current density of the fundamental. Mihran,[3] however, has shown experimentally and by machine computations on a disc-electron model that for beams of finite size

(1) the second-harmonic current does not vary periodically with distance;

(2) the maxima of the second harmonic current are closer to the

* Manuscript received September 22, 1958.

[1] D. R. Hamilton, J. K. Knipp, and J. B. H. Kuper, *Klystrons and Microwave Triodes*, McGraw-Hill Book Company, Inc., New York, N. Y., 1948.

[2] F. Paschke, "On the Nonlinear Behavior of Electron-Beam Devices," *RCA Review*, Vol. 18, p. 221, June, 1957.

[3] T. G. Mihran, "Measurement of Harmonic Currents in a Velocity-Modulated Electron Beam," 16th Annual Conference on Electron Tube Research, Quebec, Canada, June, 1958.

input cavity than the corresponding maxima of the fundamental;

(3) for thin beams, the second-harmonic current grows with distance. The growth resembles that occurring in a rippled-beam or velocity-jump amplifier.

Mihran attributed these phenomena to the presence of fringe fields of the beam. In an earlier paper[2] the nonlinear space-charge-wave equation has been developed by successive approximation into three simultaneous linear differential equations of second order. These have been solved[2] for the one-dimensional case, i.e., for negligible fringe fields. In the present paper the first two simultaneous equations will be solved with the fringe fields taken into account in order to demonstrate analytically the phenomena observed by Mihran.

Solution of the Nonlinear Space-Charge-Wave Equation

The following assumptions underlie the analysis:

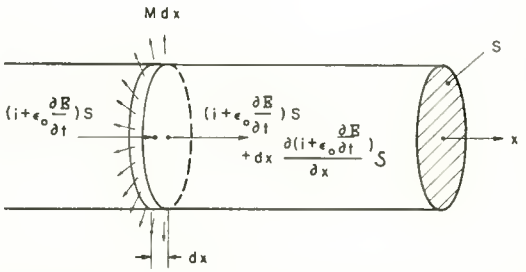


Fig. 1—Current flow through an infinitesimal beam element.

(1) The electron flow is confined by an infinitely strong magnetic field.

(2) The electron and phase velocities are small compared to the velocity of light.

(3) The electric r-f field E in the direction of the stream is approximately constant over the cross section of the beam. This is tantamount to assuming that the transverse electric field varies linearly across the beam.

(4) The effects of the potential depression caused by space charge across the beam and the thermal velocity distribution are negligible.

Figure 1 shows the current flow through an infinitesimal element of the beam; E is the axial electric field strength, i is the convection-current density, and S is the area of the beam cross section. The

displacement current per unit length M which flows away from the beam accounts for the fringe field. This field is particularly important in traveling-wave amplifiers, for it establishes the coupling between beam and delay line. According to the continuity equation of Maxwell's field theory, the total current flux into a given volume is zero. Thus for the infinitesimal beam element,

$$\frac{\partial i}{\partial x} + \epsilon_0 \frac{\partial^2 E}{\partial x \partial t} + \frac{M}{S} = 0. \tag{1}$$

As a variable the "displacement," a scalar quantity defined by

$$y = \epsilon_0 E + m \tag{2}$$

is used where m represents the fringe field and is given by

$$\frac{\partial^2 m}{\partial x \partial t} = \frac{M}{S}. \tag{3}$$

The fields are approximated by first- and second-order perturbations which are denoted by the subscripts 1 and 2. The subscript 0 refers to unperturbed quantities. The perturbations satisfy the following system of equations (taken from Reference 2):

$$\frac{\partial^2 y_1}{\partial x^2} + \frac{2}{v_0}\frac{\partial^2 y_1}{\partial x \partial t} + \frac{1}{v_0^2}\frac{\partial^2 y_1}{\partial t^2} + \beta_p^2 (y_1 - m_1) = 0, \tag{4a}$$

$$\frac{\partial^2 y_2}{\partial x^2} + \frac{2}{v_0}\frac{\partial^2 y_2}{\partial x \partial t} + \frac{1}{v_0^2}\frac{\partial^2 y_2}{\partial t^2} + \beta_p^2 (y_2 - m_2) \tag{4b}$$

$$= \frac{2}{i_0 v_0}\left(\frac{\partial y_1}{\partial t}\frac{\partial^2 y_1}{\partial x \partial t} - \frac{\partial y_1}{\partial x}\frac{\partial^2 y_1}{\partial t^2}\right) + \frac{2}{i_0}\left(\frac{\partial y_1}{\partial t}\frac{\partial^2 y_1}{\partial x^2} - \frac{\partial y_1}{\partial x}\frac{\partial^2 y_1}{\partial x \partial t}\right)$$

$$- \frac{3\eta}{\epsilon_0 v_0^2}\frac{\partial y_1}{\partial x}(y_1 - m_1).$$

The velocities are given by

$$v_1 = -\frac{1}{\rho_0}\left(\frac{\partial y_1}{\partial t} + v_0 \frac{\partial y_1}{\partial x}\right), \tag{5a}$$

$$v_2 = -\frac{1}{\rho_0}\left(\frac{\partial y_2}{\partial t} + v_0\frac{\partial y_2}{\partial x} + v_1\frac{\partial y_1}{\partial x}\right), \tag{5b}$$

and the current densities by

$$i_1 = -\frac{\partial y_1}{\partial t}, \tag{6a}$$

$$i_2 = -\frac{\partial y_2}{\partial t}. \tag{6b}$$

β_p is the plasma wave number of the infinitely extended beam, and ρ_0 the d-c space-charge density. The first-order perturbations determine the fundamental components and the second-order perturbations the second-harmonic components of the modulation. Although the theory is applicable only in cases where the transverse electric field strength varies approximately linearly across the beam, the analysis can be generalized as follows: The first-order exact solutions, the Hahn–Ramo waves, are well known. The propagation constants of these waves are given by

$$\beta = \beta_e \pm p(\omega)\ \beta_p \tag{7}$$

where β_e is the beam wave number and $p(\omega)$ the plasma-frequency-reduction factor. The latter is a function of the frequency and of the geometry.[4] The fringe-field parameter m_1 in Equation (4a) is now determined so that the solution of Equation (4a) coincides with the exact solution, Equation (7). Thus

$$y_1 - m_1 = p^2(\omega)\ y_1\,. \tag{8}$$

Similarly, in Equation (4b) for the second harmonic,

$$y_2 - m_2 = p^2(2\omega)\ y_2\,. \tag{9}$$

It is expected that this generalization predicts the dispersion of the second harmonic and therefore the field distribution along the axis satisfactorily. However, one can not calculate the field distribution across the beam by this method.

To obtain the proper solutions of Equations (4), it is assumed that

[4] G. M. Branch and T. G. Mihran, "Plasma Frequency Reduction Factors in Electron Beams," *Trans. I.R.E., PGED*, Vol. 2, p. 3, April, 1955.

the beam is modulated at $x = 0$ by a double layer with the small voltage $V_1 \sin \omega t$. If the further assumption is made that

$$\frac{\omega_p}{\omega} \ll 1, \tag{10}$$

one can neglect the excitation of harmonic velocities by the double layer.[2] Then the velocity modulation produced by the double layer at $x = 0$ is $(\eta\, V_1/v_0) \sin \omega t$. The proper solution of Equation (4a) is

$$y_1 = -\frac{i_0}{\omega_p p(\omega)} \frac{V_1}{2V_0} \sin \beta_p p(\omega)\, x \sin (\omega t - \beta_e x), \tag{11}$$

where V_0 is the d-c velocity of the beam in volts. With the aid of Equations (5a) and (6a)

$$v_1 = v_0 \frac{V_1}{2V_0} \cos \beta_p p(\omega) x \sin (\omega t - \beta_e x) \tag{12}$$

$$i_1 = i_0 \frac{\omega}{\omega_p p(\omega)} \frac{V_1}{2V_0} \sin \beta_p p(\omega) x \cos (\omega t - \beta_e x). \tag{13}$$

With the knowledge of the first-order perturbations Equation (4b) can be solved. The boundary condition of no second-harmonic modulation at $x = 0$ is met by the solution

$$y_2 = -i_0 \frac{\omega}{\omega_p^2 p^2(2\omega)} \left(\frac{V_1}{4V_0}\right)^2 \left[1 - \frac{3}{4\xi^2 - 1} \cos 2\beta_p p(\omega) x + \frac{4(1 - \xi^2)}{4\xi^2 - 1} \cos \beta_p p(2\omega) x \sin 2(\omega t - \beta_e x)\right]. \tag{14}$$

This solution is valid only under the Condition (10). The parameter ξ is defined by

$$\xi = \frac{p(\omega)}{p(2\omega)}. \tag{15}$$

ξ has been evaluated from the reduction factors given in Reference (4) for a pencil beam with radius b inside of a drift tube of radius a.

The result is shown in Figure 2. ξ approaches unity for extremely thick beams where practically no fringe fields are present and approaches ½ for very thin beams where the fringe fields are strong. From Figure 2 it can be seen that for practical beam diameters the assumption of one-dimensional fields ($\xi = 1$) is unrealistic. With the aid of Equations (5b) and (6b), the second-harmonic velocity and convection-current density are found to be

$$v_2 = v_0 \frac{\omega}{\omega_p p(2\omega)} \left(\frac{V_1}{4V_0}\right)^2 \left[\left(\frac{6\xi}{4\xi^2 - 1} - \frac{1}{\xi}\right) \sin 2\beta_p p(\omega) x - \frac{4(1-\xi^2)}{4\xi^2 - 1} \sin \beta_p p(2\omega) x\right] \sin 2(\omega t - \beta_e x). \tag{16}$$

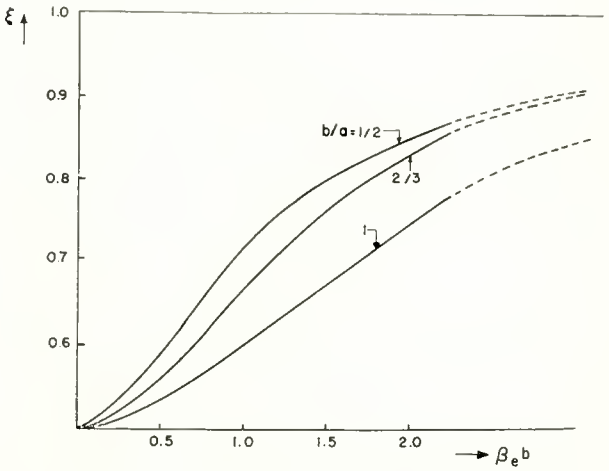


Fig. 2—Fringe-field parameter, ξ, versus normalized beam diameter.

$$i_2 = i_0 \frac{2\omega^2}{\omega_p^2 p^2(2\omega)} \left(\frac{V_1}{4V_0}\right)^2 \left[1 - \frac{3}{4\xi^2 - 1} \cos 2\beta_p p(\omega) x + \frac{4(1-\xi^2)}{4\xi^2 - 1} \cos \beta_p p(2\omega) x\right] \cos 2(\omega t - \beta_e x). \tag{17}$$

It can be seen from Equations (15), (16), and (17) that the velocity and current-density amplitudes are periodic functions of distance only if ξ happens to be a rational number. Even if the pattern is periodic, it is not as simple as that of the one-dimensional theory where $\xi = 1$.

For very thin beams $p(\omega)$ is approximately a linear function of frequency (unless $b/a = 0$) and $\xi \doteq \frac{1}{2}$. Then the inhomogeneous "driving" term on the right hand side of Equation (4b) contains a function which is a solution of the homogeneous differential equation. Consequently there is a resonance, and Equations (14), (16), and (17) become

$$y_2 = -i_0 \frac{\omega}{\omega_p^2 p^2(\omega)} \left(\frac{V_1}{8V_0}\right)^2 [1 - \cos 2\beta_p p(\omega) x + 3\beta_p p(\omega) x \sin 2\beta_p p(\omega) x] \sin 2(\omega t - \beta_e x), \tag{18}$$

$$v_2 = v_0 \frac{\omega}{\omega_p p(\omega)} \left(\frac{V_1}{8V_0}\right)^2 [\sin 2\beta_p p(\omega) x + 6\beta_p p(\omega) x \cos 2\beta_p p(\omega) x] \sin 2(\omega t - \beta_e x), \tag{19}$$

$$i_2 = i_0 \frac{2\omega^2}{\omega_p^2 p^2(\omega)} \left(\frac{V_1}{8V_0}\right)^2 [1 - \cos 2\beta_p p(\omega) x + 3\beta_p p(\omega) x \sin 2\beta_p p(\omega) x] \cos 2(\omega t - \beta_e x). \tag{20}$$

Thus for very thin beams where $\xi \doteq \frac{1}{2}$, the harmonic amplitudes increase with distance. The increase is predicted to be linear. However, perturbation terms of higher than second order in V_1/V_0 which have been neglected in the present analysis contain amplitude functions of higher than first power in x. Thus the series whose first two terms are given by Equations (11)-(13) and (18)-(20) converge satisfactorily only for distances relatively close to the excitation plane, $x = 0$. It may very well be that the actual increase in harmonic amplitudes is nonlinear. From the following, one can understand this instability physically. The second-order nonlinearity of the beam produces second-harmonic waves with twice the propagation constants of the first-order waves (Equation (7)):

$$\beta = 2\beta_e \pm 2p(\omega)\ \beta_p. \tag{21}$$

The waves described by Equations (7) and (21) have equal phase velocities and therefore travel in synchronism. In very thin beams where $\xi \doteq \frac{1}{2}$, the second-harmonic waves are as strongly supported by the beam as the first-order waves, because the propagation constants from Equation (21) are identical to those from Equation (7) for the frequency 2ω. Thus there is a situation where both the fundamental and second-harmonic frequencies are *supported* by the electron beam and where the waves associated with the two harmonic frequen-

cies travel in *synchronism*. Both conditions are required to be present if the instability is interpreted as *parametric oscillation* as suggested by Kompfner.[5] In an ordinary traveling-wave-type parametric amplifier or oscillator,[6] energy is injected at a pump frequency ω_{pp} and converted into energy at a signal frequency ω_s and an idling frequency ω_i. That the process can be reversed with energy injected at ω_s and ω_i and converted into energy at ω_{pp} is compatible with the conservation of energy.* The following relations must hold for the frequencies and associated propagation constants:[6]

$$\omega_{pp} = \omega_i + \omega_s, \tag{22}$$

$$\beta_{pp} = \beta_i + \beta_s. \tag{23}$$

These conditions are met by the waves described by Equations (7) and (21) if $\omega_s = \omega$ and $\omega_i = \omega$. Thus the instability of the second harmonic can indeed be interpreted as parametric oscillation at the frequency 2ω with low-frequency pumping at ω. The threshold of this oscillation is zero because the resonant structure, the electron beam, is lossless. It is expected that instabilities occur also for the higher harmonics as soon as

$$p(n\omega) \doteq np(\omega) \tag{24}$$

becomes a valid approximation.

Figures 3 to 5 illustrate how the fundamental and harmonic amplitudes vary along the beam for various values of ξ, that is, for various geometries. Figure 3 shows the familiar pattern of the one-dimensional theory, $\xi = 1$. Both the velocity and convection-current density are periodic functions of distance. However, for $\xi = 1/\sqrt{2}$, which corresponds to a beam with $\beta_e b \doteq 1.2$ and $b/a = 2/3$, the amplitudes of neither the velocity nor the convection-current density are periodic functions of distance. This is shown in Figure 4. It is noteworthy that the maxima of second-harmonic convection-current density are closer to the input cavity than the corresponding maxima of the fundamental convection-current density. Figure 5 shows the instability phenomenon which occurs in very thin beams ($\xi = \frac{1}{2}$). It should be mentioned that the amplitudes in Figures 3-5 are not drawn in correct

[5] R. Kompfner, Comments to Reference (3), 16th Annual Conference on Electron Tube Research, Quebec, Canada, June, 1958.

[6] P. K. Tien and H. Suhl, "A Traveling-Wave Ferromagnetic Amplifier," *Proc. I.R.E.*, Vol. 46, p. 700, April, 1958.

* The power relations cannot be given here since a fourth-order theory would be required.

relation to each other. In Figures 3 and 4 the amplitudes are shown in relation to their maximum possible value. The second-harmonic velocity amplitude shown in Figure 5a is the function in the parenthesis of Equation (19) divided by 50. The amplitude of the second-harmonic current density shown in Figure 5b is the function in the parenthesis of Equation (20) divided by 20.

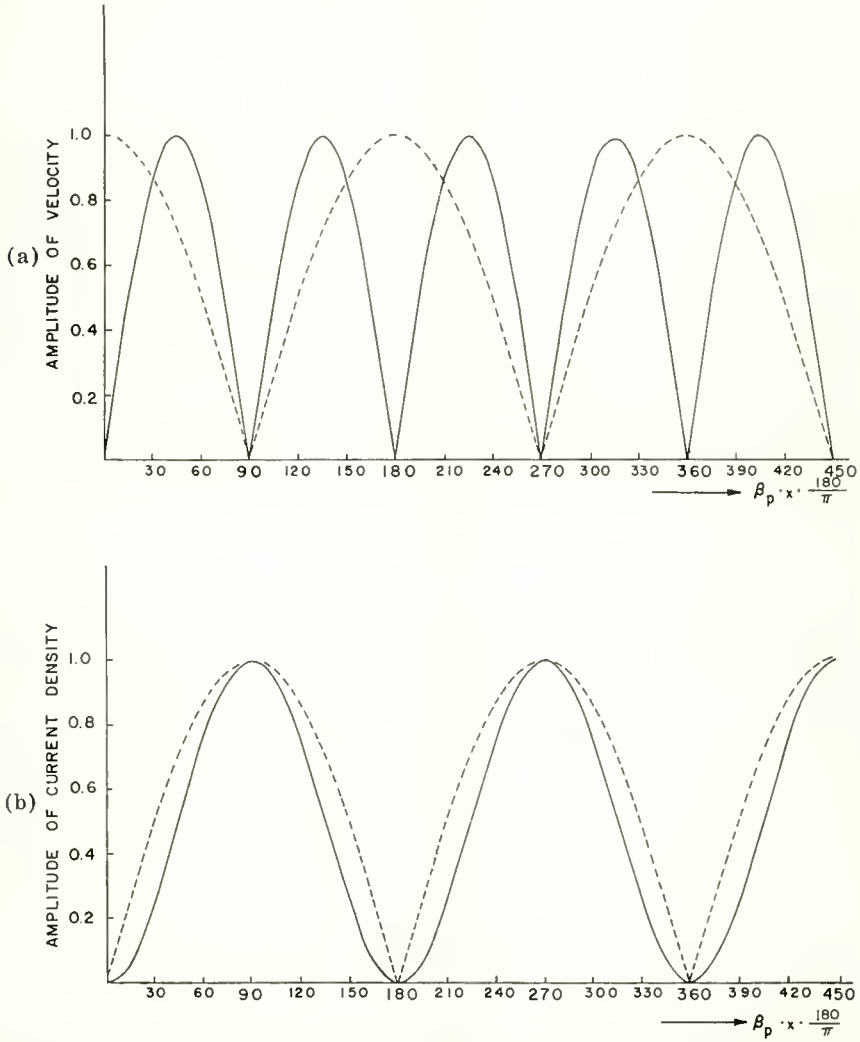


Fig. 3—Amplitude functions of the fundamental (dashed) and second-harmonic (solid) velocities (a) and convection-current densities (b) versus distance from the input cavity measured in plasma-phase angle for infinite beam diameter (one-dimensional theory, $\xi = 1$).

Conclusion

The generation of second harmonic in a velocity-modulated electron beam has been analyzed. The fringe field which is neglected in one-dimensional theories, has thereby been taken into account. The phenomena observed experimentally by Mihran[3] have been demonstrated analytically.

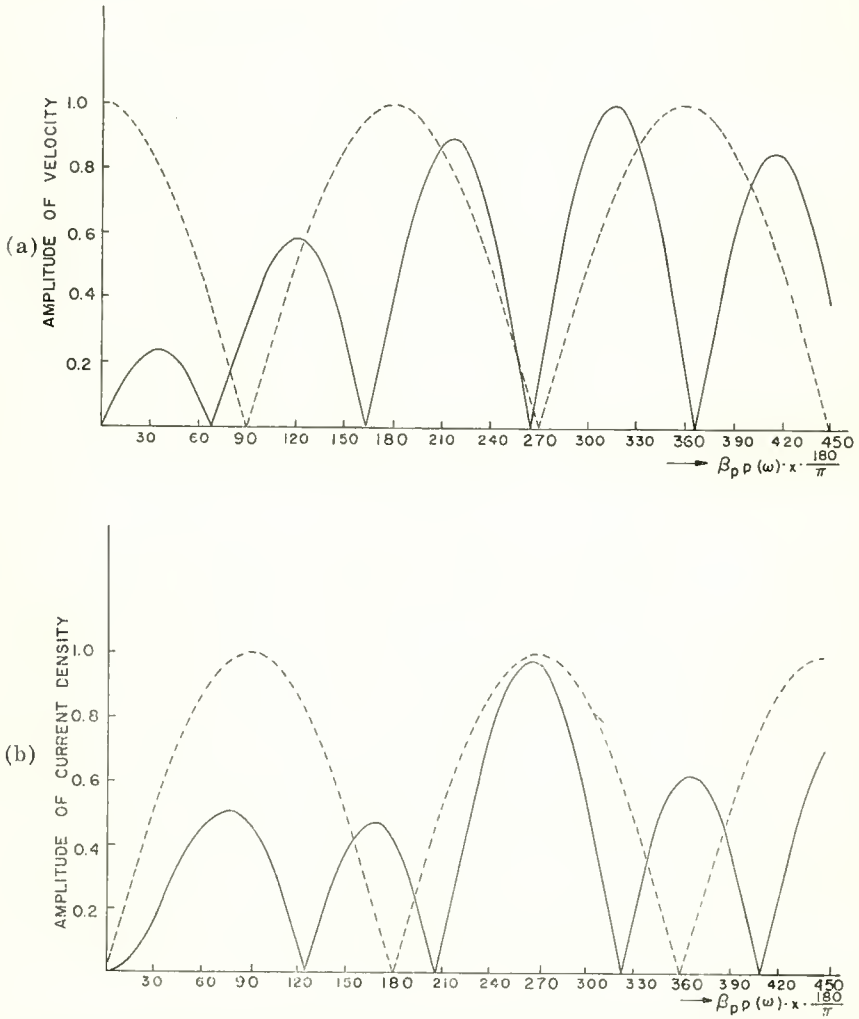


Fig. 4—Amplitude functions of the fundamental (dashed) and second-harmonic (solid) velocities (a) and convection-current densities (b) versus distance from the input cavity measured in plasma-phase angle for b/a = 2/3 and $\beta_{eb} = 1.2$ ($\xi = 1/\sqrt{2}$).

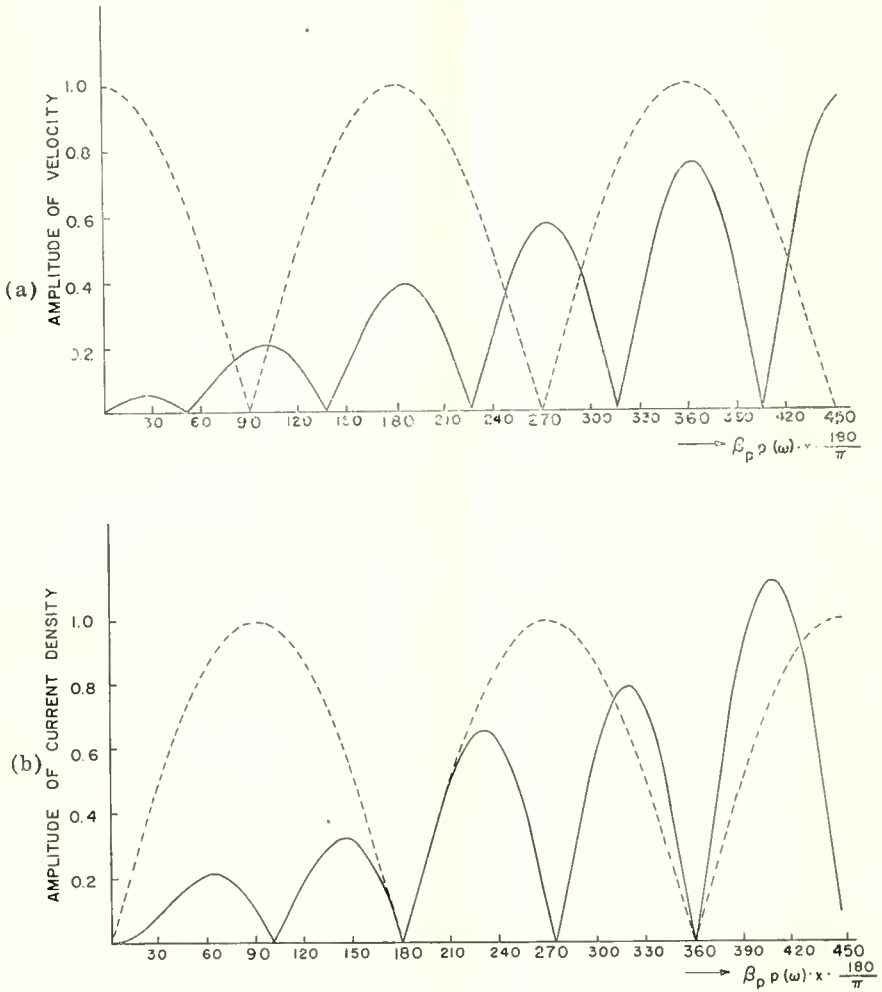


Fig. 5—Amplitude functions of the fundamental (dashed) and second-harmonic (solid) velocities (a) and convection-current densities (b) versus distance from the input cavity measured in plasma-phase angle for thin beams ($\xi = 1/2$).

Acknowledgment

The author is greatly indebted to L. S. Nergaard for reviewing the manuscript.

NOTES ON ERROR-CORRECTING TECHNIQUES I. EFFICIENCY OF SINGLE-ERROR-CORRECTING CODES WITH A CONSTANT BIT RATE OF TRANSMISSION*

By

Jacques Dutka†

RCA Defense Electronic Products,
New York, N. Y.

Summary—Under the stated assumptions, the efficiency of single-error-correcting codes for binary data compared with unprotected codes is investigated. In particular, the range of values of the bit error probability for which the former is advantageous, probability-wise, is determined as well as the values for which there is maximum relative efficiency. The relative power gain obtained by using single-error-correcting codes is also determined.

Introduction

In recent years, methods for constructing error-detecting and error-correcting binary codes have been developed by R. W. Hamming and others,[1,2] but comparatively little has been published on the efficiency of such codes compared with unprotected codes under a variety of conditions. Such an analysis would have to consider such comparisons as the bandwidth, the rate of information transfer, the relative power gain (or loss), the feasibility of mechanization, etc. This paper considers some limited questions in this area.

Assumptions

1. A Bernoulli distribution for bit errors is assumed. This means that the occurrence of an error in one bit is independent of the occurrence of an error in any other bit and that the probability of an error per bit is a constant, p.

2. Errors due to fading are not considered since there is little analytical information available on their occurrence or on the distribution of their amplitudes and durations.

* Manuscript received August 27, 1958.

† Dr. Dutka is also associated with Columbia University, New York, N. Y.

[1] M. J. E. Golay, "Notes on Digital Coding," *Proc. I.R.E.*, Vol. 37, p. 657, June, 1949.

[2] R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell Sys. Tech. Jour.*, Vol. 29, p. 147, April, 1950.

3. The time of transmission of a bit is assumed constant so that the time required to send a word is proportional to the length of the word. This insures that no additional bandwidth is required in transmitting a message in a protected code as compared with transmitting the message in an unprotected code. (A protected code is one in which redundant bits are added to an m-bit word before transmission so that errors which are in the received word may be detected or corrected.)

Efficiency of Single-Error-Correcting Codes

Let p denote the probability that a received bit is in error and let $q = 1 - p$. The probability that there is at least one error in a word m bits in length is $P_0 = 1 - q^m$, and this is the probability that a received word in an unprotected code is in error. In a single-error-correcting code, k redundant bits are added to the m information bits to produce an n-bit word. Here k is the smallest integer such that $1 + n \leq 2^k$. For an n-bit received word in a single-error-correcting code, the probability that there will be at least one error (after single-error correction is applied) is $P_1 = 1 - q^n - npq^{n-1}$.

It is now of interest to determine the range of values of the bit error probability for which a single-error-correcting code is advantageous as compared with an unprotected code. More precisely, it is desired to determine the values of p for which $P_1 \leq P_0$.

The answer may be summarized as follows:

$P_1 \leq P_0$ for $0 \leq p \leq 1 - \alpha$ where α depends on n and k, and $1/2 \leq \alpha < n\ (k - 1)/k\ (n - 1)$. Moreover $\alpha = 1/2$ if and only if $2^k = n + 1$. Single-error-correction coding is most efficient when $p = p' = (n - k)/k\ (n - 1)$.

Proof:

To show that $P_1 \leq P_0$ for $0 \leq p \leq 1 - \alpha$, it is sufficient to prove that the polynomial $f(q) = (1 - n)q^k + nq^{k-1} - 1$ is non-negative for $\alpha \leq q \leq 1$ where α is the smallest positive zero of $f(q)$. For if we have

$$(1 - n)\ q^k + nq^{k-1} - 1 \geq 0$$

for $\alpha \leq q \leq 1$, then on multiplying this inequality by $-q^m$ and rearranging,

$$(n - 1)\, q^n - nq^{n-1} \leq -q^m$$

$$1 - q^n - npq^{n-1} \leq 1 - q^m$$

or

$$P_1 \leq P_0.$$

It remains to show that $f(q) \geq 0$ for $\alpha \leq q \leq 1$. By Descartes' Rule of Signs, the polynomial $f(q)$ has k zeros of which two are positive (one of the zeros is $q = 1$), and the remaining zeros are complex except when k is odd, in which case there is one zero of $f(q)$ which is negative. An examination of the derivative $f'(q) = q^{k-2}\,[kq\,(1-n) + n\,(k-1)]$ shows that there are critical points when $q = 0$ and $q = q' = n\,(k-1)/k\,(n-1)$. For $q \geq 0$, $f(q)$ is an increasing function until $q = q'$ where $f(q)$ has a relative maximum, and $f(q)$ is a decreasing function for $q > q'$. Since $f(1) = 0$, it follows that $f(q')$ is positive. Also $f(1/2) = (1 + n - 2^k)2^{-k}$ is negative or zero according as $2^k > n + 1$ or $2^k = n + 1$. If $f(1/2 < 0$, then since $f(q') > 0$ and $f(q)$ is continuous, there exists an α, $1/2 < \alpha < q'$, such that α is a zero of $f(q)$. If $f(1/2) = 0$, then $2^k = n + 1$ and $\alpha = 1/2$. Since $f(q)$ has a maximum when $q = 1 - p = n\,(k-1)/k\,(n-1)$, it follows that single-error correction is most efficient (in the sense that the ratio $(1 - P_1)/(1 - P_0)$ is a maximum (when $p = (n - k)/k(n-1)$.

The statement is sometimes made with respect to single-error-correcting codes that they are useful only when the bit error probability is small and that at medium or high bit error probability rates, they offer no advantage over unprotected codes. In fact the claim has been made that under the latter conditions it is better to use unprotected codes. The validity of this statement will now be investigated.

It is sufficient to restrict ourselves to the case where $p < 1/2$. For if the probability of error of a received bit is known to exceed 1/2, the bit may be inverted (a "0" is replaced by a "1" and vice versa) and the bit error probability becomes less than 1/2. (Of course, if $p = 1/2$, neither protected nor unprotected codes will be of any value.) It is easily verified that

As $n \to \infty$, $p' = (n - k)/k\,(n-1) \to 0$. *The value of the bit error probability for which single-error-correction coding is most efficient tends to zero as the word length becomes infinite.*

For all n sufficiently large, $P_1 \leq P_0$, and thus a single-error-correcting code is advantageous compared with an unprotected code if only the word length is large. More precisely,

For $n \geq 13$, $P_1 \leq P_0$ *for* $0 \leq p \leq 1 - \alpha$ *where* $1 - \alpha > (1/2) + f(1/2)/f'(1/2)$ *and* $f(q) = (1 - n)q^k + nq^{k-1} - 1$. *As* $n \to \infty$, $(1 - \alpha) \to 1/2$. *Thus the upper bound of the bit error probability for which single-error correction is advantageous tends to 1/2 as the word length becomes infinite.*

Proof:

To show that $\alpha \to 1/2$ as $n \to \infty$ where α is the smallest positive

root of $f(q) = 0$, it suffices to find a number a_1 such that $a_1 > \alpha \geq 1/2$ and to verify that $a_1 \to 1/2$ as $n \to \infty$. A convenient choice for a_1 is the Newton–Raphson approximation

$$a_1 = \frac{1}{2} - \frac{f(1/2)}{f'(1/2)} = \frac{1}{2} - \frac{1 + n - 2^k}{2\,[n\,(k-2) + k]}.$$

Since $\log_2 (n+1) \leq k < 1 + \log_2 (n+1)$, $a_1 \to 1/2$ as $n \to \infty$. It remains to show that $a_1 > \alpha$ for $n \geq 13$. Subtracting α from a_1,

$$a_1 - \alpha = \frac{(1/2 - \alpha)\, f'(1/2) - f(1/2)}{f'(1/2)} \tag{1}$$

where $f'(1/2) > 0$ and it must be proved that the numerator on the right is positive. On expanding $f(\alpha)$ by Taylor's formula about the point $q = 1/2$,

$$f(\alpha) = f(1/2) + (\alpha - 1/2)\, f'(1/2) + \frac{(\alpha - 1/2)^2}{2!} f''(\xi)$$

where $1/2 < \xi < \alpha$. Since $f(\alpha) = 0$,

$$(1/2 - \alpha)\, f'(1/2) - f(1/2) = \frac{(\alpha - 1/2)^2}{2!} f''(\xi). \tag{2}$$

The left-hand side of Equation (2) is the same as the numerator of the right-hand side of Equation (1). The problem thus reduces to showing that $f''(q)$ is positive in the interval [1/2 to α] for $n \geq 13$. Now $q'' = n(k-2)/k(n-1)$ where $f''(q'') = 0$ and $f''(q) > 0$ for $0 < q < q''$. To show that $\alpha < q''$ (and thus that $f''(q) > 0$ in the interval [1/2 to α] it will now be proved that $f(q'') > 0$.

$$f(q'') = (q'')^k\,[(1-n) + n(q'')^{-1} - (q'')^{-k}]$$

$$= (q'')^k \left[\frac{2\,(n-1)}{k-2} - \left(\frac{k\,(n-1)}{n\,(k-2)} \right)^k \right],$$

$$f(q'') = \frac{2\,(q'')^k\,(n-1)}{n\,(k-2)} \left[n - \frac{k^2}{2\,(k-2)} \left(1 + \frac{2}{k-2} \right)^{k-2} \left(1 - \frac{1}{n} \right)^{k-1} \right]. \tag{3}$$

By direct substitution, it is found that the bracket on the right-hand

side of Equation (3) is positive for $k = 4$, with $13 \leq n \leq 15$ and also for $k = 5$ with $16 \leq n \leq 19$. Now since $(1 - 1/n)^{k-1} < 1$,

$$f(q'') > \frac{2\,(q'')^k\,(n-1)}{n\,(k-2)}\left[n - \frac{k^2}{2\,(k-2)}\left(1+\frac{2}{k-2}\right)^{k-2}\right]. \quad (4)$$

By substitution, it is found that the bracket on the right-hand side of Inequality (4) is positive for $20 \leq n < 64$ (and k defined by $\log_2(1+n) \leq k < 1 + \log_2(1+n)$). Finally, since $[1 + 2/(k-2)]^{k-2} < e^2$ for all $k > 2$,

$$f(q'') > \frac{2\,(q'')^k\,(n-1)}{n\,(k-2)}\left[n - \frac{k^2\,e^2}{2\,(k-2)}\right].$$

Now $2^{k-1} < n < 2^k$ for all k, (n cannot be an integral power of 2), and it is readily verified that for $k \geq 7$, $2^{k-1} - k^2\,e^2/2(k-2) > 0$. Hence, on summarizing the above, we see that $f(q'') > f(\alpha) = 0$ for $n \geq 13$ and since $f(q)$ is an increasing function for $0 < q < q'$, it follows that $q'' > \alpha$ and $f''(q) > 0$ in the interval $[1/2, \alpha]$. Thence we see that $a_1 > \alpha$ for $n \geq 13$, Q.E.D.

Table I—Single-Error-Correcting Code Compared with Unprotected Code

m	k	n	$p' = (n-k)/k(n-1)$	$1-\alpha$
10	4	14	.1923	.4845
20	5	25	.1667	.4643
40	6	46	.1482	.4595
60	7	67	.1300	.4309
80	7	87	.1329	.4606
100	7	107	.1348	.4827
2	3	5	.1667	.3596
5	4	9	.1563	.3664
12	5	17	.1500	.3906
27	6	33	.1406	.4115
58	7	65	.1295	.4273
121	8	129	.1182	.4390

Table I summarizes the results obtained above as applied to some words of convenient length. The first part of the table applies to words in which the number of information bits is $m = 10, 20, 40, 60, 80$ and 100. The second part of the table applies to words in which the length is $n = 5, 9, 17, 33, 65$ and 129, where n is of the form $2^{k-1} + 1$. This form represents the cases which are most unfavorable to single-error correction. For since $2^{k-1} < n < 2^k$, if we choose $n = 2^{k-1} + 1$, then m/n, the ratio of the number of information bits to the total number of bits, is a minimum. The column marked p' denotes the value of the bit

Table II — Probability of Error for Unprotected Word

p \ m	5	10	20	40	60	80	100
.0001	.00049	.00100	.00200	.00399	.00598	.00797	.00995
.0005	.00250	.00499	.00995	.01981	.02956	.03922	.04878
.001	.00499	.00996	.01981	.03923	.05826	.07692	.09521
.0025	.01244	.02472	.04883	.09528	.13945	.18147	.22144
.005	.02475	.04889	.09539	.18168	.25974	.33035	.39423
.007	.03451	.06784	.13107	.24496	.34392	.42991	.50464
.01	.04901	.09562	.18209	.33103	.45284	.55248	.63397

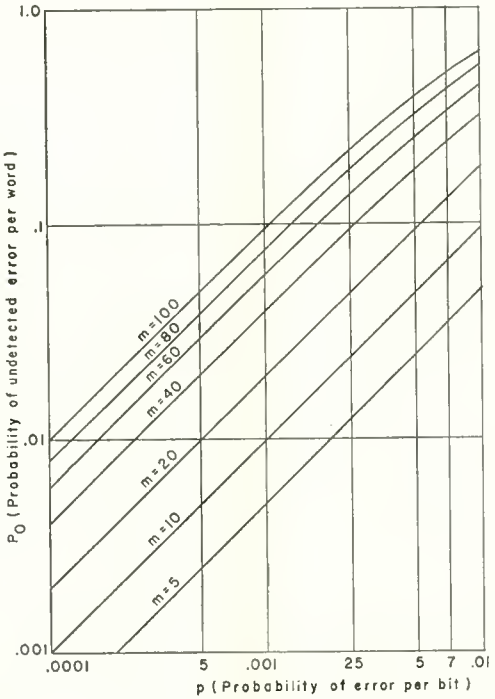


Fig. 1—Probability of undetected error per word versus probability of error per bit for unprotected code.

Table III — Probability of Error for Protected Word

p \ m	5	10	20	40	60	80	100
.0001	.00000	.00000	.00000	.00001	.00002	.00004	.00006
.0005	.00001	.00002	.00008	.00026	.00054	.00091	.00137
.001	.00003	.00009	.00030	.00101	.00212	.00354	.00529
.0025	.00022	.00056	.00180	.00601	.01241	.02032	.02981
.005	.00088	.00219	.00695	.02237	.04462	.07077	.10063
.007	.00171	.00422	.01321	.04139	.08040	.12442	.17269
.01	.00344	.00840	.02576	.07753	.14487	.21632	.29010

error probability for which single-error correction is most efficient, and the column marked $1 - \alpha$ denotes the value of the bit error probability which must not be exceeded for single-error correction to be advantageously compared with an unprotected code.

Table II gives the probability P_0 versus p and Table III gives the probability P_1 versus p for words containing $m = 5$, 10, 20, 40, 60, 80 and 100 information bits. Figures 1 and 2 denote the corresponding graphs. It will be noted that both sets of curves are approximately linear in form when drawn on log log graph paper. Table IV gives

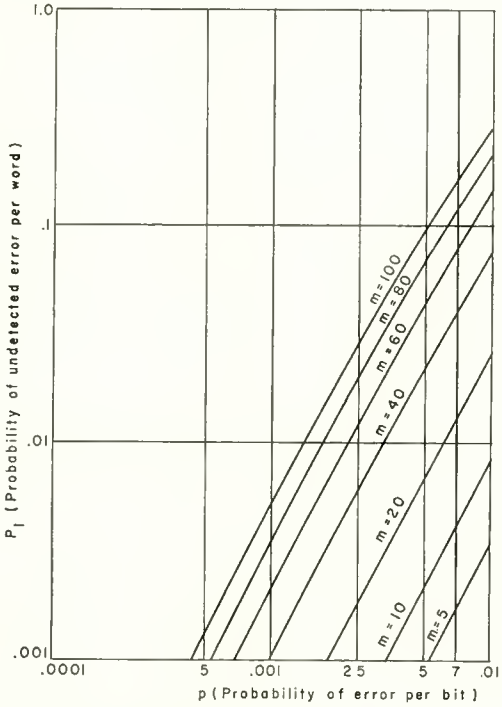


Fig. 2—Probability of undetected error per word versus probability of error per bit for protected code.

the number of information bits versus the total number of bits per word for $3 \leq n \leq 100$. Figure 3 shows a plot of k versus n and Figure 4 shows a plot of the ratio m/n versus n for the same range of n. Table V gives the probability of undetected error for a word when single-error correction is applied to a word of 52 information bits, to two subwords of 26 information bits each, and to four subwords of 13 information bits each. Figure 5 shows the corresponding graphs. It appears that comparatively little protection is gained by breaking up

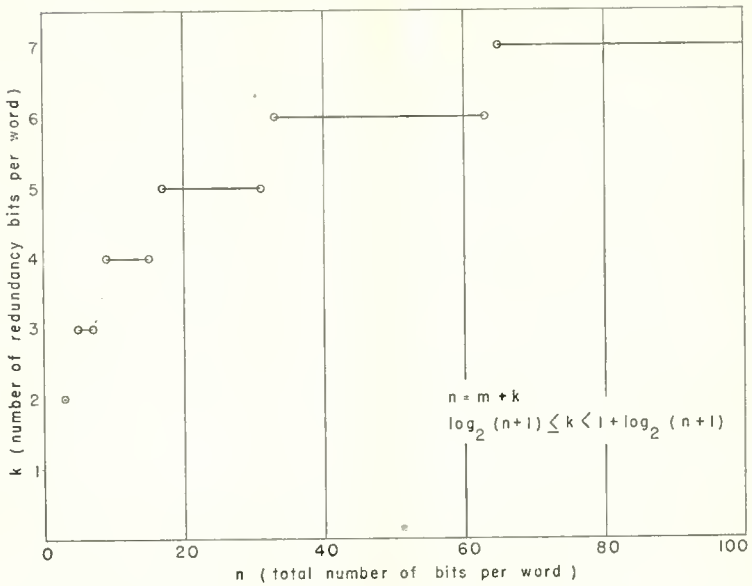


Fig. 3—Number of redundancy bits versus total number of bits per word.

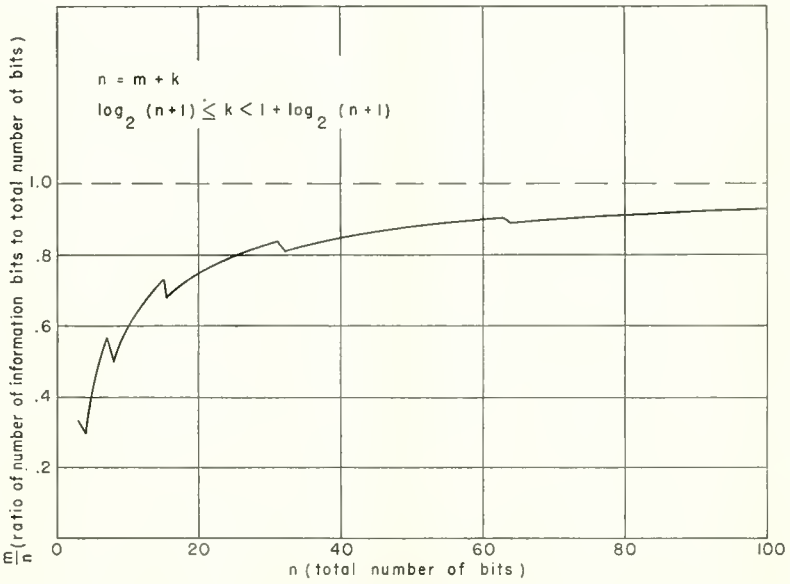


Fig. 4—Ratio of information bits to total bits per word versus total number of bits.

Table IV — Information Bits Versus Total Number of Bits Per Word

m	k	n	m/n
1	2	3	.3333
2-4	3	5-7	.4000 - .5714
5-11	4	9-15	.5556 - .7333
12-26	5	17-31	.7059 - .8387
27-57	6	33-63	.8182 - .9048
58-93	7	65-100	.8923 - .9300

$$n = m + k, \log_2 (n + 1) \leq k < 1 + \log_2 (n + 1)$$

Table V — Probability of Undetected Error for Protected Word

p	One 52-bit word with 6 redundancy bits	Two 26-bit words with 5 redundancy bits each	Four 13-bit words with 5 redundancy bits each
.0001	.00002	.00001	.00001
.0005	.00041	.00023	.00027
.001	.00159	.00091	.00061
.0025	.00941	.00553	.00372
.005	.03435	.02100	.01443
.007	.06260	.03944	.02755
.01	.11467	.07828	.05390

the 52-bit word into protected subwords when the number of extra redundancy bits required is taken into account.

Assume that a mark is represented by a d-c pulse of amplitude E volts and a space by a pulse of zero volts amplitude. It is also supposed that the transmitted signal is corrupted by additive Gaussian noise of zero mean and root mean square value v_r. An error in a received bit occurs when a mark (space) is sent and the noise is of negative (positive) polarity of magnitude $E/2$. If we suppose that the probability of transmitting a mark (space) is 1/2, then the probability of error in a received bit is

$$p = \frac{1}{2}\int_{-\infty}^{-E/2v_r} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt + \frac{1}{2}\int_{E/2v_r}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt = \int_{E/2v_r}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt.$$

Let p_1 denote the effective bit error probability for an m bit word defined by the relation $1 - (1 - p_1)^m = P_1$. That is, p_1 is what the bit error probability would have to be in order to yield the same word error probability for an m-bit word as the word error probability for an n-bit word which includes single-error correction and in which the bit error probability is p. Let $q_1 = 1 - p_1$ and let E_1, the effective pulse amplitude, be defined by

$$p_1 = \int_{E_1/2v_r}^{\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt.$$

The quantity 20 $\log_{10} E_1/E$ is a measure, in decibels, of the power gain (loss) obtained by using a single-error-correcting code (and more generally of a protected code). This quantity versus p and m is given in Table VI and the results are shown graphically in Figure 6.

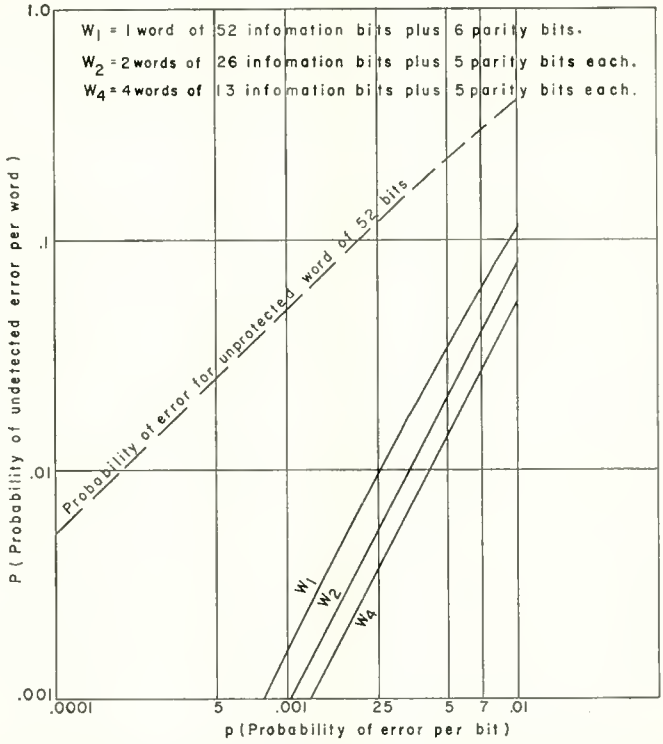


Fig. 5—Probability of undetected error per word versus probability of error per bit for Single Error-Correcting Code applied to subwords.

It would be of interest to obtain the quantity 20 $\log_{10} E_1/E$ as an explicit function of p_1 and p, and, as an initial step, E_1 as an explicit function of p_1. It does not appear likely that a usable exact form for E_1 as a function of p_1 can be obtained by inverting the above integral. Instead we shall obtain upper and lower bounds for E_1 as a function of p_1, and for the quantity 20 $\log_{10} E_1/E$. We have

$$2v_r \sqrt{-\frac{\pi}{2} \ln 4p_1q_1} \leq E_1 \leq 2v_r \sqrt{-2 \ln 4p_1q_1}\,, \tag{5}$$

and

$$\left| 20 \log_{10} \left(\frac{E_1}{E} \right) - 10 \log_{10} \left(\frac{\ln 4p_1q_1}{\ln 4pq} \right) \right| \leq 10 \log_{10} \frac{4}{\pi} = 1.049 \text{ db.} \quad (6)$$

Proof:

Consider the surface $z = (1/2\pi) \exp \{-(x^2 + y^2)/2\}$. The volume under this surface bounded by the right cylinder whose base is the square $-a \leq x, y \leq a$ is

$$\int_{-a}^{a} \int_{-a}^{a} z \, dxdy.$$

Table VI — Decibel Gain In Use of Protected Code

m	p	$E/2v_r$	$E_1/2v_r$	$20 \log_{10} E_1/E$
.5	.0001	3.7190	5.2626	3.0157
	.0005	3.2905	4.6339	2.974
	.001	3.0902	4.3387	2.947
	.005	2.5758	3.5759	2.850
	.01	2.3263	3.1991	2.767
10	.0001	3.7190	5.2148	2.936
	.0005	3.2905	4.5854	2.882
	.001	3.0902	4.2889	2.847
	.005	2.5758	3.5184	2.708
	.01	2.3263	3.1387	2.602
20	.0001	3.7190	5.1234	2.782
	.0005	3.2905	4.4662	2.654
	.001	3.0902	4.1736	2.611
	.005	2.5758	3.3922	2.392
	.01	2.3263	3.0092	2.236
40	.0001	3.7190	5.0204	2.606
	.0005	3.2905	4.3603	2.445
	.001	3.0902	4.0532	2.356
	.005	2.5758	3.2492	2.017
	.01	2.3263	2.8751	1.840
60	.0001	3.7190	4.9529	2.489
	.0005	3.2905	4.2889	2.302
	.001	3.0902	3.9738	2.184
	.005	2.5758	3.1714	1.807
	.01	2.3263	2.7944	1.592
80	.0001	3.7190	4.8980	2.392
	.0005	3.2905	4.2360	2.194
	.001	3.0902	3.9197	2.065
	.005	2.5758	3.1152	1.651
	.01	2.3263	2.7435	1.432
100	.0001	3.7190	4.8560	2 317
	.0005	3.2905	4.1942	2.107
	.001	3.0902	3.8762	1.969
	.005	2.5758	3.0732	1.534
	.01	2.3263	2.7048	1.309

The volume under the surface of a right cylinder whose base is the circle $x^2 + y^2 = a^2$ inscribed in the square is

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{a} e^{-r^2/2}\, r dr d\theta.$$

and the volume under the surface of a right cylinder whose base is the circle $x^2 + y^2 = 4a^2/\pi$, the area of the circle being equal to that of the square, is

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{2a/\sqrt{\pi}} e^{-r^2/2}\, r dr d\theta,$$

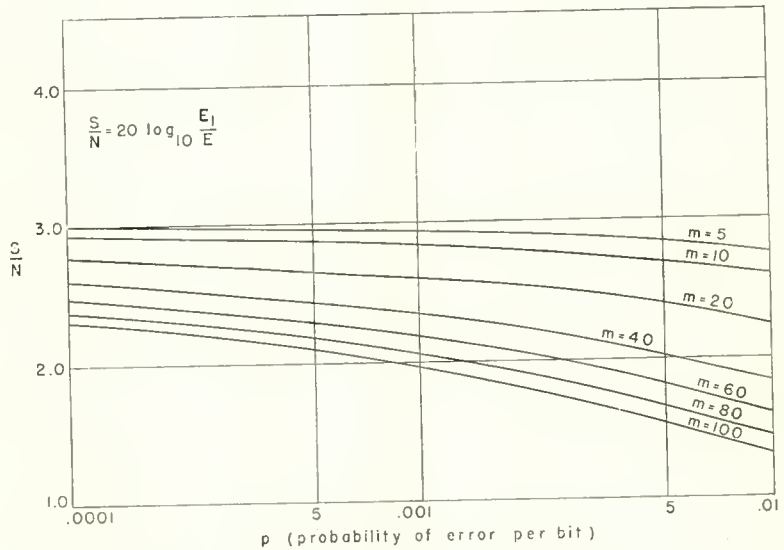


Fig. 6—Decibel gain in using protected code.

It may be readily shown that

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{a} e^{-r^2/2}\, r dr d\theta \leqq$$

$$\int_{-a}^{a} \int_{-a}^{a} z dx dy \leqq \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2a/\sqrt{\pi}} e^{-r^2/2}\, r dr d\theta$$

and on evaluating the integrals on the left and right,

$$1 - e^{-a^2/2} \leq \int_{-a}^{a} \int_{-a}^{a} z dx dy \leq 1 - e^{-2a^2/\pi}$$

Since

$$\int_{-a}^{a} \int_{-a}^{a} z dx dy = \left(\int_{-a}^{a} e^{-t^2/2}\, dt/\sqrt{2\pi} \right)^2,$$

it follows that

$$\sqrt{1 - e^{-a^2/2}} \leq \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{-t^2/2}\, dt \leq \sqrt{1 - e^{-2a^2/\pi}}.$$

(The relative error involved in approximating the middle term by the term on the right never exceeds 0.75 per cent.) Let $a = E_1/2v_r$. Then the middle term is equal to $2q_1 - 1$, and solving these inequalities for E_1 gives Equation (5).

From these inequalities we may write a similar set involving E and p and from these results we find

$$10 \log_{10} \left(\frac{\ln 4p_1q_1}{\ln 4pq} \right) - 10 \log_{10} \frac{4}{\pi} \leq 20 \log_{10} \frac{E_1}{E}$$

and

$$20 \log_{10} \frac{E_1}{E} \leq \left(\frac{\ln 4p_1q_1}{\ln 4pq} \right) + 10 \log_{10} \frac{4}{\pi}$$

which is equivalent to Equation (6).

The practical usability of these results in any particular application depends, of course, upon the magnitude of the error which is acceptable in the application.

Conclusions

At the cost of a reduction in the rate of information transmission by a factor m/n and the requirement of a temporary storage to compensate for the delay, the use of an error-correcting code has, under certain conditions, resulted in efficiency gains.

In this paper we have established:

(a) the ranges of bit error probability for which single-error

correction is advantageous and the dependence of these values on the word lengths;

(b) the word lengths for which the efficiency of single-error-correcting codes as compared to unprotected codes is maximum;

(c) the fact that for a typical case of a 52 information bit word, relatively little gain is obtained in applying single-error correction to subwords;

(d) the relative power gain in applying single-error correction for typical values of bit error probability and for typical word lengths;

(e) upper and lower bounds in analytical form for the relative power gain as functions of the actual and effective bit error probabilities.

AUTOMATIC CONTROL OF VIDEO TAPE EQUIPMENT AT NBC, BURBANK*

By

Robert W. Byloff

National Broadcasting Company, Inc.,
New York, N. Y.

Summary—*The presentation of television programs at the proper local time in the different time zones in the United States has created unusual program recording demands upon the network television broadcast industry. To solve this problem, the National Broadcasting Company has developed an integrated automatic system using video tape recorders and associated automatic control equipment.*

This paper describes the equipment and physical layout of this new facility, modification of equipment and methods employed to provide automatic cueing of video tape recorders, equipment that automatically maintains the television color signal at the required level, and sequencing devices that automatically switch the input and output circuits, automatically stop, start and rewind the video tape recorder machines all to provide one-, two-, three- or four-hour program delays or any combination thereof.

Introduction

THERE are four time zones across the United States — Eastern, Central, Mountain, and Pacific. In addition, during the summer months "daylight-saving" time is in effect in certain areas causing a fifth time zone to exist. Thus, there may be a four hour time differential between the east and west coasts.

Television programming is arranged with the time of day in mind; that is, there are morning, afternoon, and evening programs each designed to appeal to a particular audience group. It is important, therefore, that programs designed to go on the air at about eight o'clock in the evening, go on at that time in each part of the country. Unless the program is delayed by some means, a program going on the air at eight o'clock in the evening in New York would be on at four o'clock in the afternoon in Seattle, Washington, during summer months. In this case, the program broadcast in Seattle would not reach the audience for which it was intended.

Kinescope recording has been used for a number of years to delay programs in the far west, but until video tape recorders with their

* Manuscript received August 6, 1958.

ability to play back immediately, became available, it was impractical to delay programs for as little as one hour. Now it is possible to provide a network program service in which programs are aired in all sections of the country at the same local times.

Such a service was inaugurated by the National Broadcasting Company (NBC) in April, 1958. By means of video tape machines installed in Burbank, California, programs recorded from the eastern NBC network are played back to the central and Pacific coast networks one and three hours later, respectively. The map of Figure 1 indicates the time differential by the use of shaded areas. Clocks in these areas show a representative time in each zone. Thus when it is eight o'clock in the northeast, it is seven o'clock in areas on eastern standard and central daylight times, six o'clock in areas on central standard time, five o'clock in areas on mountain standard and Pacific daylight time, and four o'clock in a small area in the Pacific northwest which stays on Pacific standard time. The complete solution to this problem would be to provide services of one-, two-, three-, and four-hour delay to the respective areas. However, by providing only a one-hour and three-hour time delay, it is possible to provide the entire network with programs which are on, in terms of local times, at most only one hour different from the local time of origination in the east. In general this is being done with the entire NBC program schedule. Figure 1 shows the time schedule for a typical evening program across the country. In terms of local time; it is on at 8:00 PM in New York, 8:00 PM in Chicago, 7:00 PM in Omaha, 8:00 PM in Los Angeles, and 7:00 PM in Seattle.

Video tape recorders are presently arranged to record and play back reels of tape one hour in length. If a one-hour delay is required, it is impossible to record for a full hour, since time must be allowed to rewind the recorded reel and ready it for playback. To avoid discontinuities in program material, it was decided to record for half-hour periods. Thus, any switching operations could occur during "station break" periods. A recorder involved in the delay operation records for a half hour, rewinds and waits during the next half hour, plays back the one-hour delay during the third half hour, then rewinds and waits during the next one and a half hours so that during the seventh half hour, it can play back the three-hour delay. It then rewinds and waits during the eight half hour to be ready to record again during the ninth half hour. Other machines on a staggered basis fill out the schedule as shown in Figure 2. Note that it takes eight machines using this recording–playback format to do a continuous one- and three-hour delay. It is normal practice in network broadcasting to record the

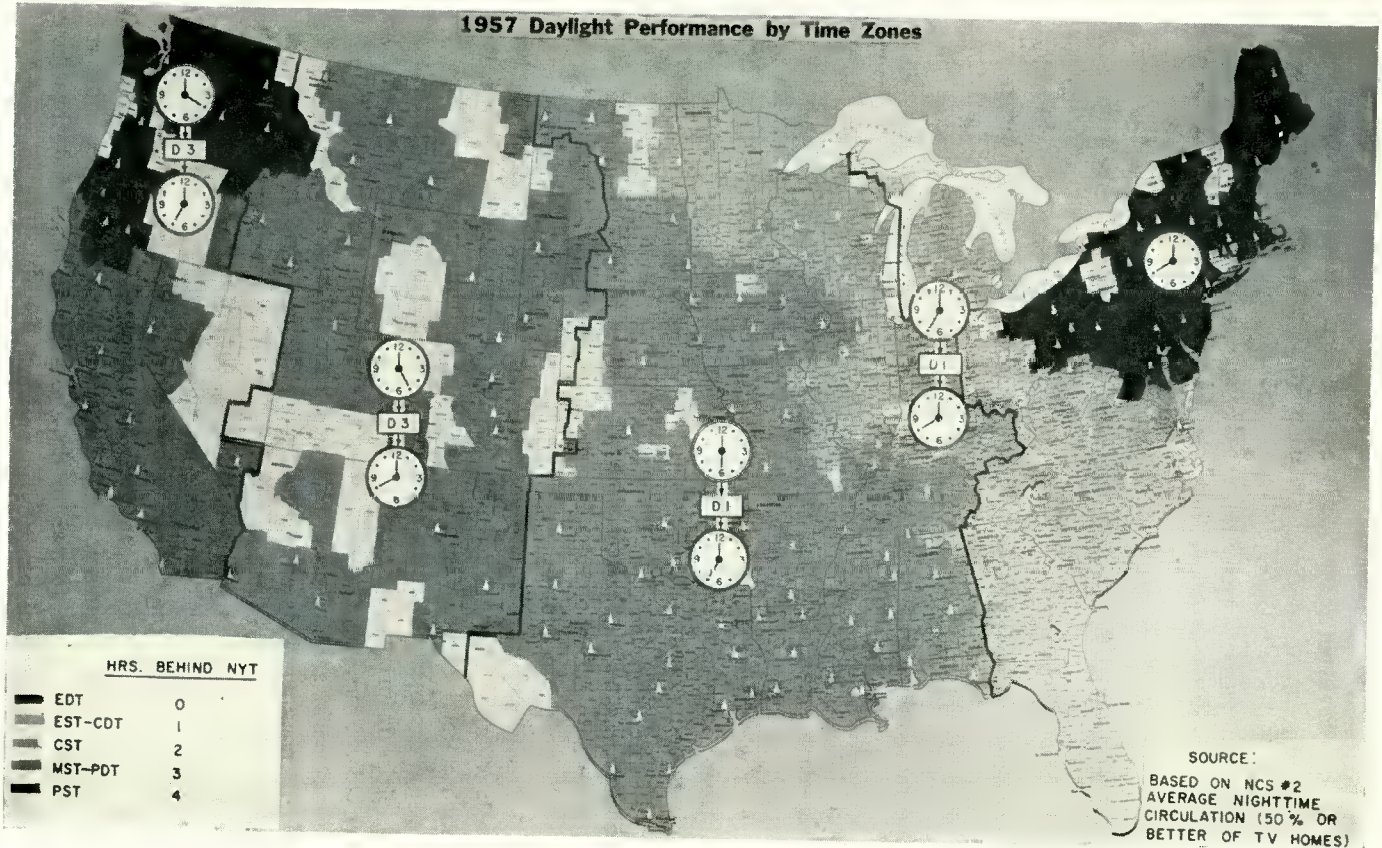


Fig. 1—NBC television coverage areas — full network of 189 stations excluding 19 satellites.

same program and play it back on two machines simultaneously to afford a "protection" copy recording. Thus, if an element in one chain of equipment should fail, the program can continue on the protection chain. They are normally termed "A" and "B" copies. For complete protection, therefore, sixteen machines are required. Fortunately, the program schedule is such that continuous recording and playback is not required for the entire day, and so it is possible to perform this operation with twelve machines which will allow recording and playback on a one- and three-hour delay basis for three hours. If an hour is then skipped, the cycle can be repeated.

Plan of Facility

The plan of the tape recording plant at Burbank, California, is shown in Figure 3. Since color programs occur at random times during

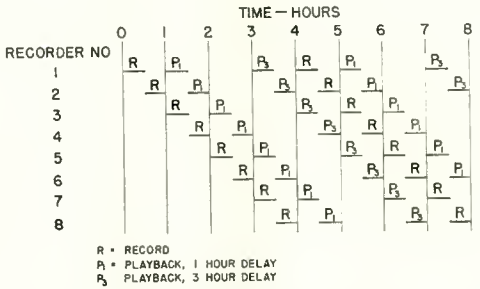


Fig. 2—Record-playback cycle for video tape recorders for one- and three-hour delayed broadcasting.

the schedule, and will increase in frequency, it was decided to "colorize" all the machines. The plant contains, at present, four RCA prototype color video tape recorders and eight RCA-colorized Ampex tape recorders. Each RCA recorder consists of six standard equipment racks and a monitor housing containing color and black-and-white monitors, a cathode-ray oscilloscope, VU meter, and a control panel. Each colorized Ampex machine consists of four video racks, a large console and a monitor. All of these equipments are installed in one space to gain maximum efficiency. It is possible to observe the operation of each machine from any spot in the room. The status panel at one end of the room shows the status of the machines in the automatic delayed broadcast system, and permits rearrangement of the operational sequence at any time. In the quality control room, the quality of the

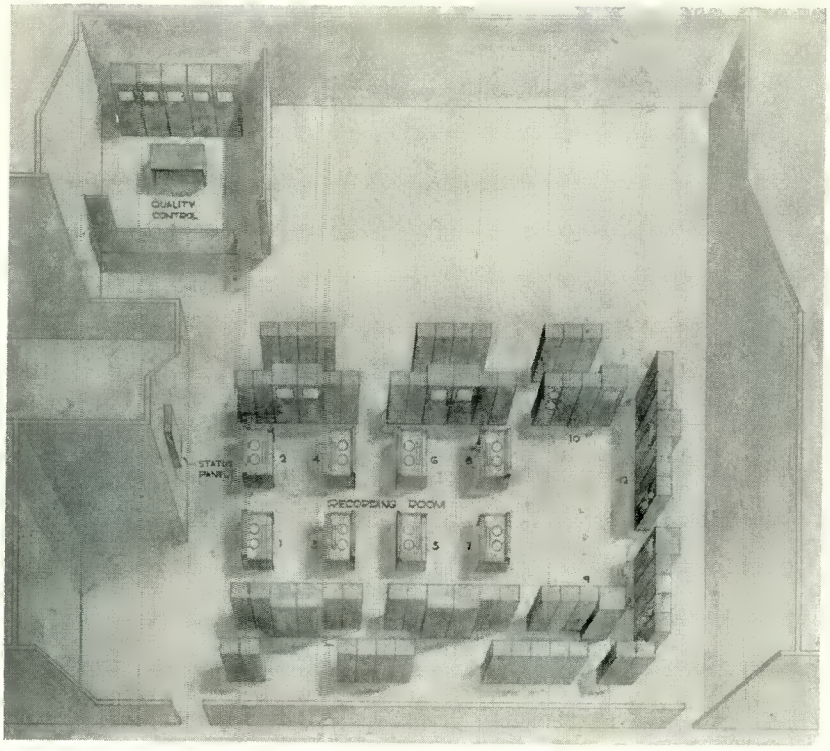


Fig. 3—Video tape equipment layout.

Fig. 4—Video tape recording room

Fig. 5—RCA color video tape recorders.

delayed playbacks is monitored in color and in black-and-white, and the choice of "A" or "B" copy is made.

The remaining space is used for maintenance, erasing facilities, parts storage, and for future expansion.

The tape room is shown in Figure 4; the RCA recorder facility is shown in Figure 5; the Ampex recorder facility is shown in Figure 6; and the quality control room is shown in Figure 7.

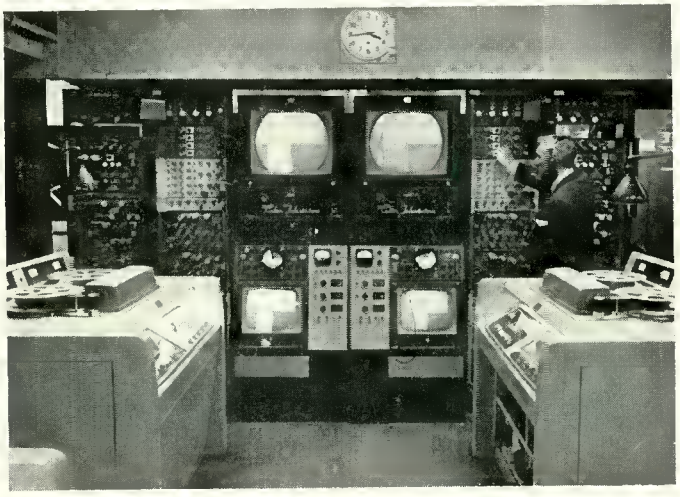


Fig. 6—RCA-colorized Ampex recorders.

Video and Audio System

The video and audio systems are identical in form; the two systems are shown by the block diagram of Figure 8 which also shows the control system. The twelve video tape recorders are connected between input and output switching systems.

The input switching system, which is used during the recording process, is fed by the incoming network line from the east, all the Burbank studio outputs, and test signals. These test signals are recorded at the beginning of each tape to act as a check on the operation of the machine before it is used on the air. Automatic-gain-control amplifiers, both in audio and video, immediately precede the video

Fig. 7—Tape-quality control room.

tape recorder input to insure that the recorded signal level is as high as permissible without overloading; this gives maximum signal-to-noise ratio. The a-g-c amplifiers also eliminate signal level variations.

Each video tape recorder is provided with a color monitor, a black-and-white monitor, a cathode-ray oscilloscope, a VU meter, and a loudspeaker. The monitors are switched automatically to the input or output of the machine according to whether the machine is recording or playing back.

The purpose of the output switching system, used during playback, is to connect the outputs of the twelve recorders to outgoing lines for delayed broadcasting and for playback to studios when integrating with other program material. The four lines provided for normal

delayed broadcast playback are nominally, the one-hour delay to the midwest, the three-hour delay to the Pacific coast, a two-hour delay, which is used for special purposes, and a four-hour delay for future use. Monitoring on these four lines is available in the quality control room, and control of output switching is provided to permit switching between "A" and "B" copy tapes.

Types of Control

The control system has several operating modes. Controls at each machine are provided to permit selection of the control mode desired as follows:

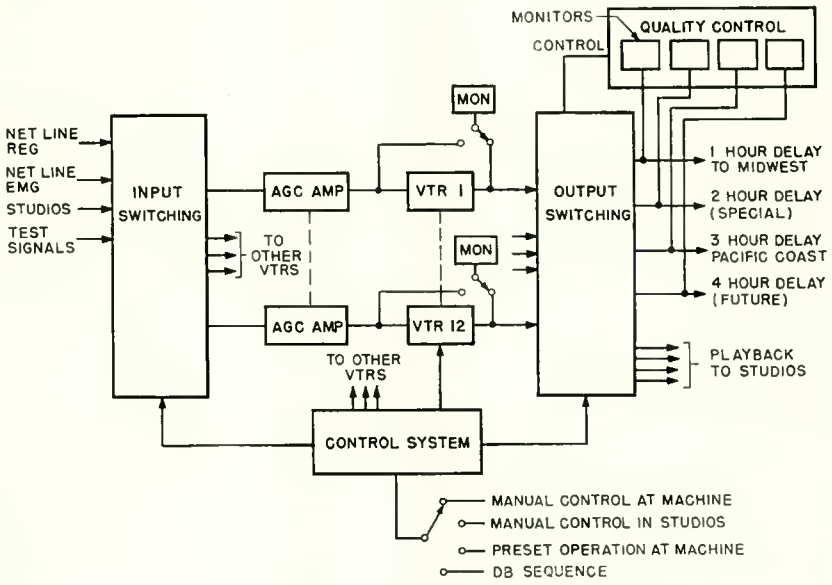


Fig. 8—Video and audio systems.

A. Manual Control at Machine — In this mode, the recorder operator can select the input line, the output line, and the recorder operation (i.e., record, playback, rewind, fast forward, or stop). This mode is used mainly for testing.

B. Manual Control in Studios — In this mode, the machine operation is controlled by any one of the Burbank studios. The studio to have control is selected at the machine.

C. Delayed Broadcast (D-B) Sequence — In this mode, all machine operations including input and output line selection and switching are assigned to the automatic D-B sequencer. The recorder enters the cyclic operation required for delay broadcasting as outlined above.

D. Preset Operation at Machines — In this mode, the operator may preset the machine to record or play, preselect the record or playback line, and set the time and duration of operation. The machine will automatically start, select the proper lines, record or play back for the specified period, and then automatically rewind. This mode is used for scheduled operations of a noncyclic type.

Almost all operating situations can be handled by control modes B, C, and D thereby eliminating manual switching operations in the tape room which was the goal in the design of the control system.

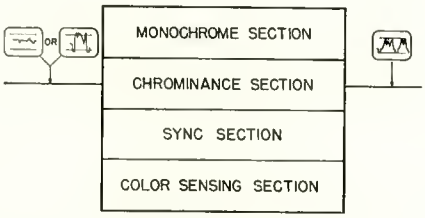


Fig. 9—Color video automatic-gain-control amplifier.

Automatic Control Devices

In order to simplify operation, certain devices were added to the tape recorders.

A. A-G-C Amplifiers — Automatic-gain-control amplifiers are used in the audio channel of the recorders. In addition, color video a-g-c amplifiers are used in the video paths. These amplifiers, which were developed recently by NBC, perform many functions as shown in Figure 9. The incoming video signal is passed through a filter which separates luminance and chrominance components. In the luminance channel sync is clipped, the base line cleaned up, and the signal controlled to a particular level. In the chrominance channel the burst is measured, and by this means chrominance is held to a predetermined level. A sync channel separates sync from the incoming video signal, clips and shapes it, and adds it to the luminance and the chrominance information at the output of the amplifier, through suitable delay networks. Thus the amplifier automatically sets and maintains separately

the levels of luminance, chrominance, and sync of a color signal. The color sensing section in the amplifier detects whether the incoming signal is color or monochrome, and if monochrome, automatically shuts off the chrominance channel, and widens the bandwidth of the luminance channel.

B. Automatic Cueing Equipment — The recorders are arranged so that they "cue up" automatically to a predetermined point, thus providing a selected length of playing time before the program material. The problem was to insure that after a recording was made for delayed broadcast operation, the machine would rewind and "cue up" to the proper position for playback.

Equipment for special cue tracks was not immediately available so it was necessary to improvise temporarily. The equipment was arranged to detect the absence of control track during rewinding and

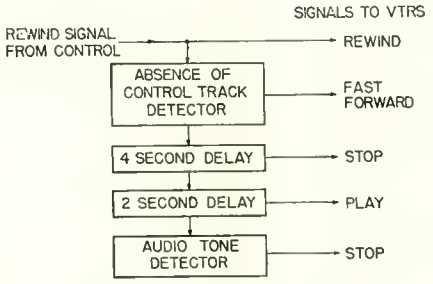


Fig. 10—Automatic cueing.

to detect the presence of 1000-cycle audio tone when playing to come up to cue. This tone is recorded at the beginning of each recording as a test. A diagram of the functioning of the cue system is shown in Figure 10. The rewind signal causes the machine to rewind and also to connect the "absence of control track detector." When the tape has rewound to the point where no control track is detected, a relay closes putting the machine in the fast forward mode. This is done to bring the machine to a stop without using the brakes and thereby reduce brake wear. After four seconds in the fast forward mode, the machine is given a stop signal, and two seconds later a play signal. The machine plays until it detects 1000-cycle audio tone on the audio track, then stops. The 1000-cycle signal is recorded automatically during the recording cycle exactly ninety seconds ahead of program material. When cue heads are available, the system will be changed to use specially recorded signals on this track for cueing.

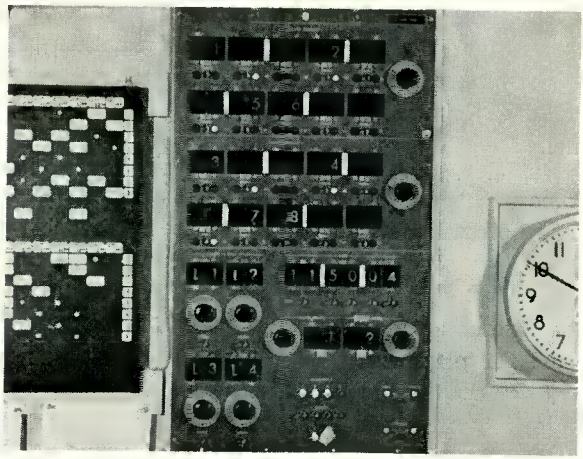


Fig. 11—Delayed-broadcast sequence status panel.

Delayed-Broadcast Sequencer

Automatic delayed broadcast operation is available through use of the D-B sequencer. This sequencer consists of a status panel and racks of control equipment. The status panel is shown in the photograph of Figure 11. The system has two similar parts: one for the "A" and another for the "B" copy tapes. Machines are assigned to

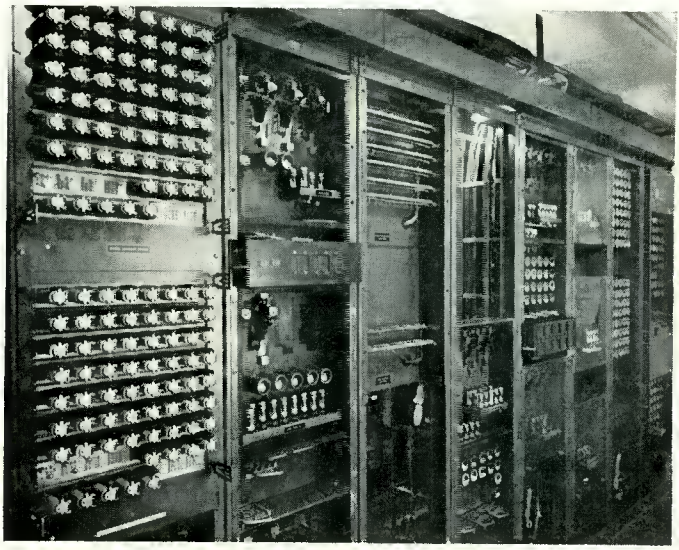


Fig. 12—Delayed-broadcast sequence control racks.

the "A" sequence at the top section of the panel. There are ten so-called "slots" for machine assignment, permitting ten machines to be assigned, so that continuous one-, two-, three-, and four-hour delays can be handled. Machines may be assigned in any order to these delay slots. Assignment is made by selecting the desired machine with the selector knob at the right side of the panel and then pressing the "enable" button under the appropriate slot. The read-outs indicate the machine assignments. A similar setup is provided for "B" copy tapes in the next section of the panel. Controls at the lower section of the panel provide input and output line assignments. These controls permit any input line to be used for recording and also permit assignment of any delay to any output line. The delay modes can also be selected at this location. Thus, the system can be set up for any combination of one-, two-, three-, and four-hour delays.

Figure 12 is a photograph of the control equipment racks associated with the D-B system.

Figure 13 shows the D-B sequence control system, in a very much simplified representation, as it would be if set up to record and play one- and three-hour delays, "A" copy only. Eight recorders are required to provide the complete delay. The heart of the system is the sequence stepping switch. This is a multideck-"rotary" switch mechanically ganged and electrically pulsed from clock impulses. The switch moves one step every half hour. The points around the switch correspond to machine positions. The drawing shows recorders 1-8 around the switch. Each deck represents a different machine operation. Thus, the first deck shown at the top of the drawing is for record. The drawing shows that this switch deck is in position to make VTR 1 record. The record impulse also sets up the input switching system to select the proper input line and connect it to VTR 1. One-half hour later, this switch will rotate one step and will cause VTR 2 to record, and the input switching system to switch the input line to VTR 2. This will continue on succeeding half hours to make all recorders record in sequence. The second deck is a rewind deck. It is shown in position so that VTR 2 is rewinding. Deck 3 is the three-hour delay play deck. Referring to Figure 2, note that while recorder 1 is recording, recorder 3 is playing the three-hour delay, and recorder 7 is playing the one-hour delay. This is the condition shown in Figure 13. Deck 3 also sends impulses to the output switching system to set up the proper output line to the recorder playing back the three-hour delay. A selector switch permits preselection of the line to be used for three-hour delay. The other decks in the sequence switch as shown in Figure 13 are: deck 4, rewind; deck 5, play two-hour delay; deck 6,

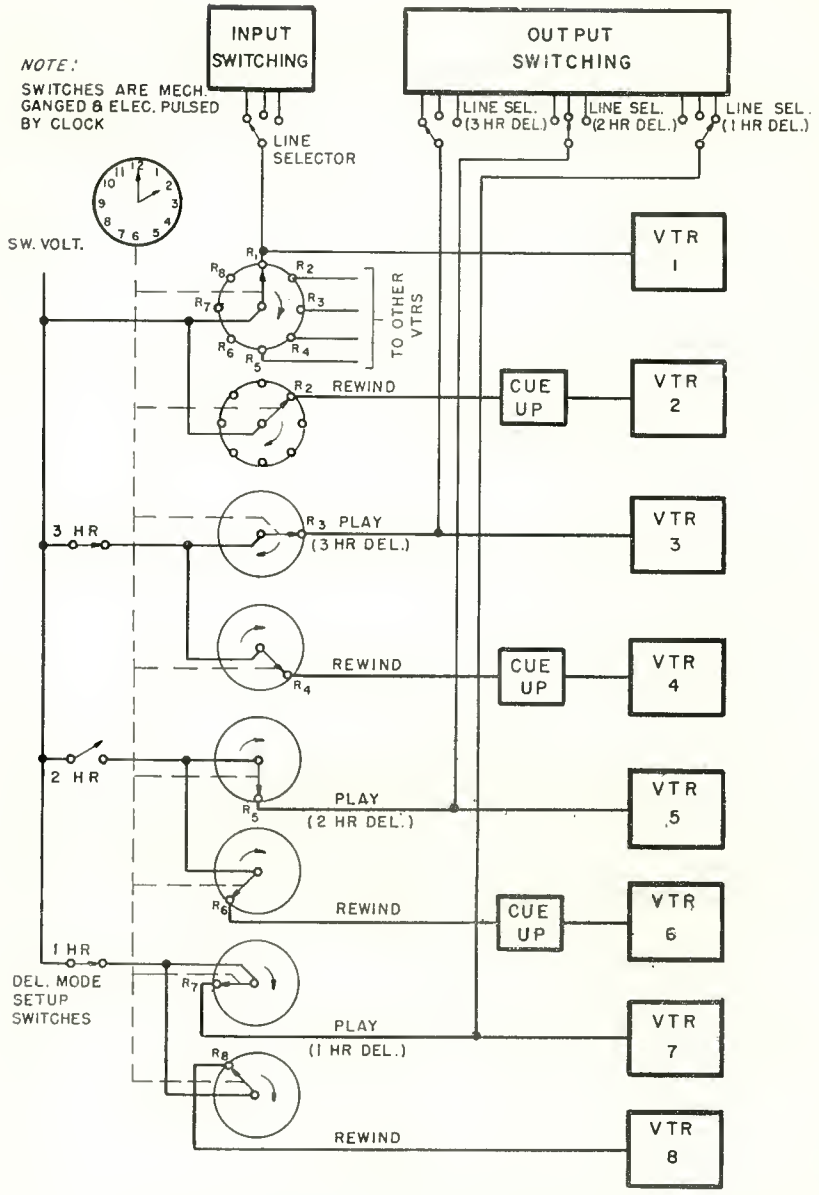


Fig. 13—Delayed-broadcast sequence system control.

rewind; deck 7, play one-hour delay; deck 8, rewind. Note that while deck 5 is calling for recorder 5 to play back the two-hour delay, the delay mode switch for two-hour delay is open, and therefore the playback does not occur. For the sake of simplicity, the system which allows any recorder to be assigned to any position in the sequence, and the complete system which permits four-hour delay are not shown.

Preset Operation

Figure 14 shows the time preset system with which each recorder is separately equipped. It enables an operator to preset recorders to record or play from any of the recording or playback lines at specified times and for specified intervals. When there is coincidence between the clock and the time set into a recorder, a pulse passes to the recorder which starts the recorder and switches the lines. At the same time

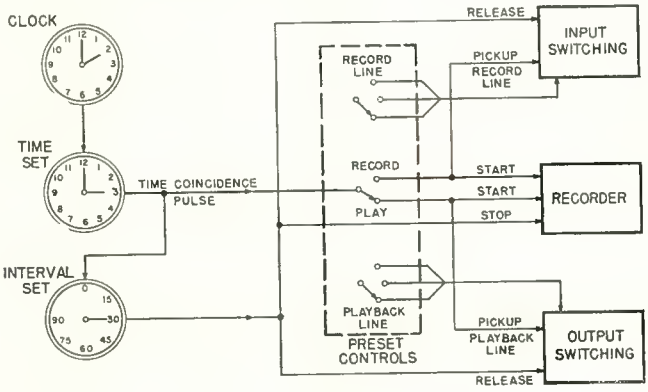


Fig. 14—Time preset system.

this pulse starts an interval timer, which stops the recorder and releases the lines when it has completed the interval.

This preset system supplements the D-B sequencer to provide essentially automatic operation of the tape equipment used for delayed broadcasting by NBC at Burbank.

Operating Experience

This equipment has been in operation since April, 1958. At the present time over four hundred machine hours per week are being logged for actual recording and playback. During this period the system has provided NBC with a reliable delayed broadcast operation with a minimum of confusion and errors. It has been very important in achieving efficiency in this operation.

POST-INSTALLATION PERFORMANCE TESTS OF UHF TELEVISION BROADCASTING ANTENNAS*

By

D. W. Peterson

RCA Laboratories,
Princeton, N. J.

Summary—*Determination of the power gain and radiation pattern of a modern television broadcasting antenna is difficult after it is installed. However, if ordinary coverage surveys do not give the expected results, tests on the installed transmitting antenna may be necessary to measure the performance of this part of the system. This paper undertakes a summary of the pertinent propagation information and outlines a procedure for field checking antennas that requires the minimum effort consistent with obtaining reliable results. Propagation effects due to terrain shadowing, clutter, and the earth's curvature are avoided as completely as possible.*

A check of antenna performance does not constitute even a partial coverage survey. Coverage surveying is quite another matter which includes *the effects of terrain shadowing, clutter, and the earth's curvature.*

Introduction

ANY post-installation check of a UHF television broadcasting antenna is generally recognized as much less accurate than factory tests made under relatively ideal conditions. Even at a properly designed factory test-site, antenna gain and pattern measurements unavoidably include earth reflection effects, although these effects are now usually negligible[1]. The philosophy adopted for post-installation antenna checking avoids a large part of the complication of propagation theory. There remains a simplified theory that must be applied. UHF broadcasting antennas need rarely be checked beyond a distance of ten miles. By thus restricting the distance and carefully choosing propagation situations, the theoretical requirements can be limited to simple free-space and plane-earth theories. Certain tests are required to know which theory to apply to a specific field measuring situation. If, in spite of the obvious superiority of factory test-site measurements, field testing of an antenna becomes necessary after installation, this may be done with reasonable reliability

* Manuscript received September 9, 1958.

[1] E. H. Shively, "Pattern-Testing the TF-24B UHF Antenna," *Broadcast News*, Vol. 69, p. 42, May-June, 1952, and E. H. Shively and L. D. Wetzel, "Measurements of RCA UHF TV Antennas," *Broadcast News*, Vol. 82, p. 14, February, 1955.

provided care is exercised and the limitations imposed by propagation effects are fully recognized.

Problems Associated with Field Measurement

In order to make a post-installation check of an antenna, one must define the size, location, and character of areas, known as first Fresnel (earth reflecting) zones, that may cause significant earth reflection effects. Also, the first Fresnel (shadowing) zone must be described quantitatively to know that significant terrain shadowing effects have been avoided. For the exceptionally smooth-earth situation there must be quantitative propagation laws for prediction, and a smooth-earth criterion is required to determine when to apply smooth-earth theory. Also, the near-field region must be determined. Means of establishing these factors are provided in this paper in the form of easily used curves and formulas. There is also a brief review of the applicable theory and recommended procedure for field measurement of antenna performance.

Since the object is to field-test the antenna, unpredictable propagation effects must be avoided as far as possible. This means that terrain shadowing, clutter effects, and shadowing from the earth's curvature should be completely avoided. In other words, all field measurements for post-installation antenna checking should be made in unmistakably uncluttered, line-of-sight locations.

However, one propagation effect, namely *earth reflection,* cannot be completely avoided. This effect can cause an increase above the free-space value of field strength (electric intensity) up to 2 to 1 and in rare cases even more. Also, earth reflection may reduce field strength 10 to 1 or more below the free-space value. Since the effect cannot be avoided, it must be taken into consideration. Examination of the field strength versus height characteristic often gives clues as to the existence and magnitude of earth reflection. However, these clues cannot always be relied upon since the limited height range examined may not reveal the entire character of the interference pattern resulting from earth reflection because of the complexity from more than one effective reflecting surface and the resulting multiple rays. It should be kept in mind that significant reflecting surfaces (portions of hills) may occur at any distance or may be situated away from the direct path.

Because of the unavoidable propagation effects mentioned, there is no positive assurance that even carefully selected measuring sites will yield the predicted field strengths. Nevertheless, field strengths in the

order of the free-space value are usually found over rough terrain provided the precautions discussed below are observed.

Some idea of the field strength to expect over smooth terrain can also be reached. However, a complication results from the fact that exceptionally smooth terrain may cause a lobed two-ray interference pattern. When this condition is suspected, the theoretical field strength versus height distribution must be calculated and measurements must be made to determine whether or not the condition exists. Sufficiently smooth earth will limit the distance of usable measuring sites to those sites where interference maxima and minima can be reached with feasible receiving-antenna heights.

UHF propagation is subject to severe shadowing effects from terrain obstacles. These effects can and should be avoided by maintaining at least first Fresnel (shadowing) zone clearance above terrain obstacles. Similarly, the effect of the local clutter of trees and buildings at the receiving site is to be avoided. Both the terrain shadowing and the clutter effects are highly frequency dependent. Since they can be avoided, these effects need not unduly complicate field performance checks of a broadcasting antenna.

Measurements of field intensity can be made too near as well as too far with respect to the transmitting antenna. Thus, the near-field region should be avoided because within this distance range the pattern has not fully formed. On the other hand, when measurements are attempted at too great a distance the field intensity will be time variable, another condition to avoid.

In cases where the azimuth pattern as measured at the factory test-site is accepted, only one radial line need be examined. The distance range of measurements along this one radial will usually include radiation from the entire useful part of the elevation pattern with sufficient sampling detail to establish the general character of the elevation pattern. Often only the maximum of the elevation pattern need be checked, preferably in several receiving sites. When all or part of the azimuth pattern must be checked, the process of examining the elevation pattern may be repeated along additional radial lines. A reasonable compromise to reduce the amount of measurement for a horizontal pattern check is to check pattern circularity by measurements on the elevation pattern maximum only.

When field strengths are below expected values, all relevant theoretical possibilities must carefully be considered in accounting for the condition. It is unfortunate but true that most of the propagation effects reduce rather than increase field strength with respect to the predictions of simple theory.

Free-Space Theory of Propagation

When propagation is presumed to occur in empty space, attenuation results from the divergence of energy from a point source. Antenna engineers like to think of an isotropically radiating source at the point; consequently, power flow through the space can be thought of as radial and with uniform power density flowing across any sphere concentric with the point source. An isotropic source has come into wide use as the reference for antenna gain. However, because of the ready availability of half-wave dipole antennas, this kind of reference is used instead for television antennas.

Transmitting-antenna power gain is commonly expressed as a power ratio, specifically the ratio of power radiated in free space on the elevation pattern maximum to the power maximum of a reference half-wave dipole, assuming that the power inputs are equal. The

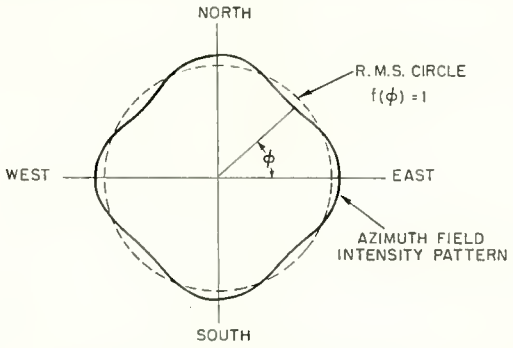


Fig. 1—Azimuth pattern.

azimuth field intensity patterns of most broadcasting antennas are not precisely circular although they usually closely approach circularity. The quoted antenna power gain, G, is the gain that an antenna with the same radiated power and the same elevation pattern would have if the azimuth field intensity pattern were precisely circular with unity magnitude. The circle is the root-mean-square circle of the azimuth pattern as shown in Figure 1. The true gain, G_{ϕ_1}, in a direction specified by ϕ_1 can be related to the quoted gain thus:

$$G_{\phi_1} = [f(\phi_1)]^2 G \tag{1}$$

where G_{ϕ_1} is the antenna power gain in the direction specified by ϕ_1,

$f(\phi_1)$ is the azimuth field intensity pattern evaluated for the direction ϕ_1,

G is the r-m-s antenna power gain.

Because of the use of a half-wave reference dipole for power gain, the free-space field intensity on the maximum of the dipole pattern will be useful.

$$E_0 = \frac{137.6\sqrt{P_d}}{r} \tag{2}$$

where E_0 is the free-space field intensity in millivolts/meter for a half-wave dipole,

P_d is the power input to the dipole in kilowatts,

r is the distance in miles between the antenna and a point in space.

In the direction specified by ϕ_1 the field intensity on the maximum of the elevation pattern from an antenna of power gain, G, is

$$E_{\phi_1} = \frac{137.6\sqrt{G\,P}}{r} f(\phi_1) \tag{3}$$

where E_{ϕ_1} is the field intensity in the ϕ_1 direction in millivolts/meter,

P is the power input of the transmitting antenna in kilowatts.

The theoretical free-space field intensity is expected under certain conditions when the smooth-earth criterion (discussed later) is *not* satisfied. Before this becomes a useful tool for checking an antenna in the field, methods of measurement and the necessary conditions must be understood. On the other hand, when the smooth-earth criterion *is* satisfied, plane-earth theory must be considered.

Ray Theory of Propagation Over Plane Earth

Propagation from a transmitting antenna located at P_2 to a point P_1 over smooth, reflecting earth is described in Figure 2. The earth is assumed to be a relatively smooth plane. The lobing which results from interference of the direct and earth-reflected rays is described and formulas are given for the calculation of field intensity over smooth, plane earth.

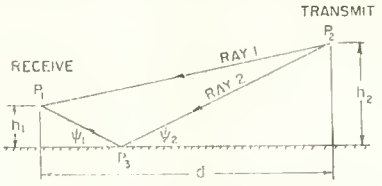


Fig. 2—Plane-earth reflection.

According to simple plane-earth ray theory, radiation from P_2 can reach P_1 by way of two rays, namely P_2P_1 (the direct ray) and $P_2P_3P_1$ (the earth-reflected ray). The incident grazing angle ψ_2 of the earth-reflected ray must equal the angle ψ_1 of the ray upon reflection. With horizontal polarization, as used for television broadcasting in the United States, there is a 180° phase reversal upon reflection at P_3, and there is complete reflection (magnitude of the reflection coefficient is unity) for smooth-earth conditions.

If the transmitting antenna consists of a half-wave dipole with a circular elevation pattern, the field intensity at P_1 will vary with height h_1 according to a simple periodic law, resulting from the alternately constructive and destructive interference of two rays, provided a simplifying restriction is made. The restriction is that the grazing angle ψ will be small (less than 10°) so that the reflection coefficient will remain close to unity and so that the inverse-distance law of free-space propagation will result in only small differences in the magnitudes of ray 1 and ray 2 field intensity contributions. The law of field intensity variation is, for horizontal polarization and $\psi < 10°$,

$$E_{h_1} = E_0 \sqrt{2 - 2 \cos \Delta}, \tag{4}$$

where E_{h_1} is the field intensity at height h_1 (h_1 is related to Δ in Equation (5)),

Δ is the difference in path length in electrical degrees between rays 1 and 2.

This formula, shown graphically in Figure 3, gives a limit on the

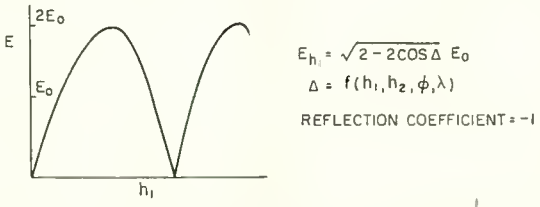


Fig. 3—Field-intensity variation with height.

maximum possible excursion of field intensity variation with height for any transmitting antenna as long as the receiving site is below the maximum of the elevation pattern. When the antenna has power gain by virtue of a narrowed elevation pattern, the variation of field intensity with height will be similar as long as both ray 1 and ray 2 are near the maximum of the elevation pattern. Under most other practical conditions, the excursions about the free-space field intensity will be less. With high-gain shaped-elevation-pattern antennas, the ray 1 and ray 2 contributions must be calculated separately if the magnitudes of the maxima and minima to be found with height are desired because portions of the pattern exhibit rapid changes with elevation angle. However, it usually suffices to approximately locate the maxima and minima with Equation (4) and then depend upon measurement in the field for the magnitudes. For a well-chosen measuring site, the average of these two values will be the free-space field.

The path difference Δ depends upon the geometry shown in Figure 2, and the wavelength and can be calculated from the formula[2]

$$\Delta = 0.136 \frac{h_1 h_2}{d\lambda} \tag{5}$$

where h_1 and h_2 are the measuring and transmitting antenna heights in feet,

d is the distance between P_1 and P_2 in miles,

λ is the wavelength in feet. $\lambda = \dfrac{980}{f}$, where f is in megacycles/sec.

Test Equipment for Field Measurement

A reliable field-intensity meter is, of course, indispensable. When reliable measurements are needed, it is advisable to check the instrument against laboratory generators or against a bolometer. This, of course, only checks the calibrating generator in the field-intensity meter. The entire instrument can be checked by establishing and measuring known field intensities.

An over-all check of the field intensity meter can be made in a large, flat, clear area (say a 10 acre field) by establishing a radiated field with a signal generator and a matched $\lambda/2$ dipole. The field-intensity-meter dipole height should be adjusted to pass through both interference maxima and minima at a distance great enough so the direct

[2] H. Reed and C. Russell, *UHF Propagation*, John Wiley & Sons, New York, N. Y., 1953, p. 110.

and earth-reflected rays are essentially the same length. The free-space field which has been established (actually the field strength which would have existed without the presence of the reflecting earth) can be calculated from Equation (2). Of course, transmission-line attenuation must be subtracted in the calculation of established field strength. The free-space field intensity from the measurement will be

$$E_0 = \frac{E_{\min} + E_{\max}}{2},$$

where $E_{\min}$ is the field intensity on the interference minimum, and
$E_{\max}$ is the field intensity on the interference maximum.

When measurements are made in hilly terrain or in the vicinity of possible reflectors (this cannot always be avoided), the standard dipole furnished with the field-intensity meter should be replaced with a corner reflector or paraboloid antenna. This can be calibrated against the dipole in a large, flat, clear area by direct substitution, taking care that both antennas are placed successively in the same field intensity. The field-intensity distribution in the substitution test area should first be checked for constancy by probing with the dipole.

Procedure for Field Measurement of Performance

Two kinds of useful measurements associated with two distinct types of propagation paths can be made in an effort to check antenna performance in the field.

The first kind consists of field strength versus height measurements with conditions such that two-ray (plane-earth) theory clearly applies. There will be no question about the existence of first Fresnel (shadowing) zone clearance since there will be no hills or other shadowing obstacles. It will be necessary to estimate the theoretical heights of maxima and minima and measure a field strength versus height distribution from which to calculate (by averaging a maximum and a minimum) the free-space field intensity. The area surrounding Miami, Florida, and the Illinois prairie are illustrative of places where the terrain is flat enough to apply smooth-earth theory. Paths over bodies of water can also be included in this category.

The second kind consists of field strength versus height measurements over terrain so rough that the two-ray (plane earth) theory as tested by the Rayleigh roughness criterion does *not* apply. Over rough terrain, there must be a test for first Fresnel (shadowing) zone clearance.

Both kinds of measurements must be made under line-of-sight conditions without the clutter of trees or buildings in the vicinity of the receiving antenna. They should be made as close as possible (outside the near-field region) to the transmitting antenna to avoid spherical earth effects. Other kinds of propagation paths will be encountered but, for the purpose of antenna checking, are to be avoided.

A straightforward procedure for checking television broadcasting antenna performance will now be outlined:

1. Measure the power fed to the broadcasting antenna after checking the VSWR of the antenna. An independent substitution wattmeter is preferable to simply relying on built-in power-monitoring devices. Subtract transmission line loss to learn antenna input power.
2. Select a number of measuring sites outside the near-field region (Equation (14)), but no further than necessary from the transmitting antenna, which are completely uncluttered with buildings or trees. From these sites, the transmitting antenna should be clearly seen to be well above all intervening terrain features. On-the-spot observation is necessary. If clutter at the receiving site is unavoidable, the receiving antenna should be kept well above the clutter.
3. Plot elevation profiles[3] for the intervening terrain for each potential test site.
4. Calculate the size and location of the first Fresnel (earth reflecting) zone, Equations (10), (11), and (12), and test this area by the smooth-earth criterion (Equation (6)). It must be recognized that terrain features outside the first Fresnel (earth reflecting) zone may alter the received field intensity.
5. If the smooth-earth criterion indicates smooth earth, calculate the theoretical location of field-intensity interference maxima and minima as a function of receiving antenna height (Equation (4)). If no theoretical maximum is accessible, the site should not be used.
6. If the smooth-earth criterion indicates rough earth, test the intervening terrain for first Fresnel (shadowing) zone clearance and abandon the site if the clearance test is not satisfied. The test should be applied to hills, buildings, and trees. (Figure 4)
7. Measure and plot field intensity versus height in the acceptable sites and compare data with appropriate theory.

The procedure outlined has been aimed at avoiding unpredictable

[3] If topographic maps are not available from either Federal or State sources, salient features can sometimes be measured with an altimeter. See H. Reed and C. Russell, *op. cit.*, p. 52.

propagation effects. This is not always possible; however, it will usually be possible to find sites which are satisfactory. The weakness of a measuring site will not always be apparent. Sometimes the field intensity versus height recordings will suggest likely shortcomings of a measuring site.

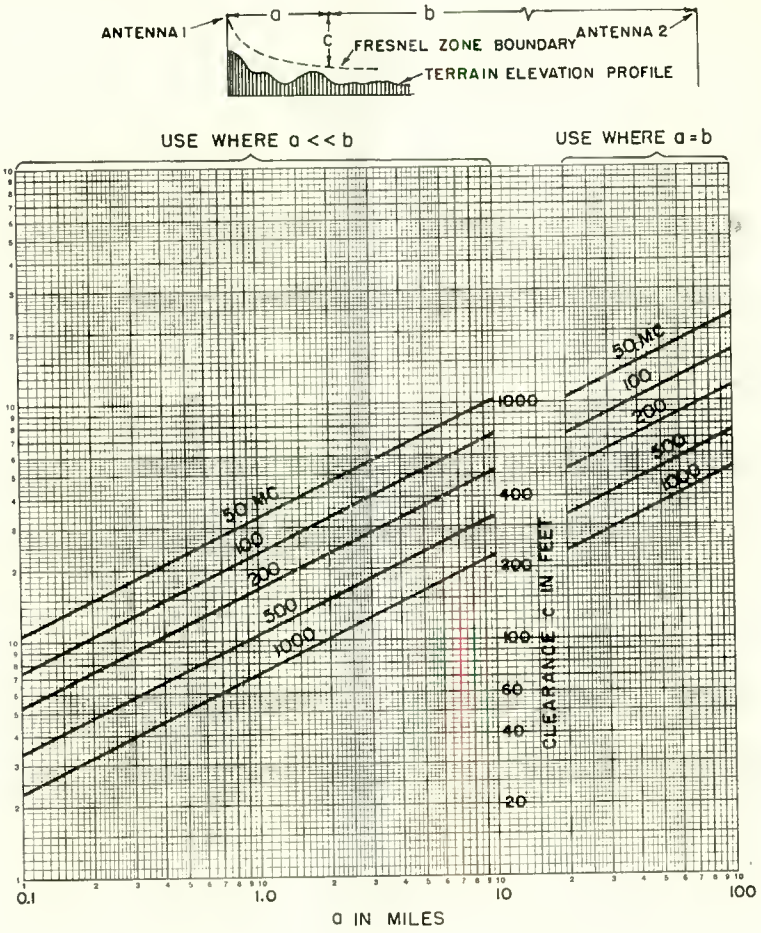


Fig. 4—First Fresnel (shadowing) zone clearance.

Expected Field Intensity

When the smooth-earth criterion indicates smooth earth in the first Fresnel (earth reflecting) zone, values ranging from the free-space field intensity to twice this value are to be expected at heights where interference maxima occur. Either the reflection coefficient can

be lower than unity or ray 1 and ray 2 of Figure 2 may differ significantly in length to account for this range of values. The field intensity versus height data should confirm the applicability of smooth-earth theory by the presence of interference maxima and minima. Several field intensity versus height distributions are desirable at each measuring site. Recordings at constant height of field strength versus distance are also valuable.

When the smooth-earth criterion indicates rough earth in the first Fresnel (earth reflecting) zone, free-space field intensities are to be expected. If the surroundings of the receiving site are hilly, it will be usual to observe complicated field intensity versus height distributions which will be a consequence of the immediate surroundings. Under these circumstances field intensities both above and below the free-space value are possible.

Another condition which may be encountered is intervening terrain

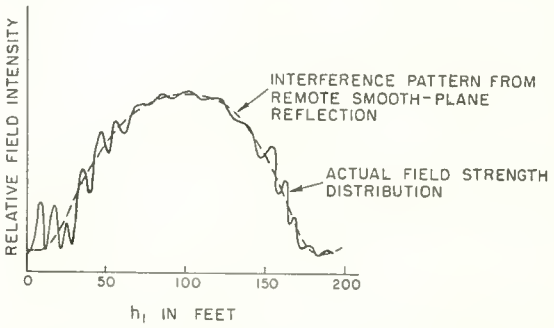


Fig. 5—Measured field strength versus height in hilly terrain.

in the first Fresnel (earth reflecting) zone which satisfies the smooth-earth criterion, accompanied by cluttered and/or hilly surroundings at the receiving site. Then there is the possibility that there will be an average reduction of field intensity below free-space values because of destructive interference throughout the measuring site. In other words, there may be superimposed on the observable maxima and minima an average trend of increasing (or decreasing) field intensity with height which is not fully observed because of the limited height range which is accessible. When this happens, there is no way (except exploration at unreasonable heights) of knowing the extent of the interference hidden in the measurements. Thus, in the typical field strength versus height distribution illustrated in Figure 5, examination of field strength up to 30 feet would fail to reveal that there had been a reduction in field strength at all heights up to 30 feet as a result of destructive interference. The pattern of destructive interference

would only have been apparent upon examination of received field strength up to 200 feet in height. The rapid variations with height are the result of reflecting surfaces on relatively nearby hills whereas the slower variation results from a more remote smooth plane in the first Fresnel (earth reflecting) zone.

"Before-and-After" Comparisons

There may be a requirement to test the improvement in antenna performance upon changing the transmitting antenna while keeping the same center of radiation; e.g., there may be an increase in gain, a change in pattern shape, or a shift in frequency. Presumably data both before and after the antenna change can be obtained. The problem here is basically easier than the absolute test of a single antenna because with due precaution the propagation effects can be virtually eliminated from the comparison. However, there are important exceptions.

When the first Fresnel (earth reflecting) zone is smooth according to the smooth-earth criterion, it will be necessary to proceed with care. It should be recognized that the direct and earth-reflected rays may come from significantly different parts of the elevation pattern. Thus, when there is a change of pattern shape, the contributions from the earth-reflected ray and direct ray both change in magnitude but not necessarily by the same factor. Unless this is understood there may be apparent inconsistencies in before-and-after comparisons even though the frequency and the height of the center of radiation remain the same and the antennas are functioning normally.

If the antenna change includes even a small change in location there will be drastic changes in the detailed field intensity distribution in rough terrain. This can easily be true even in areas of fringe reception at great distances. Furthermore, a shadowed fringe-area valley may receive median field intensities which are changed even though the *same* transmitting antenna is moved only a few miles transversely with respect to the propagation path.

When an antenna change is accompanied by a shift in frequency there will be propagation changes which must be given due regard in before-and-after comparisons. This will be true even though the elevation pattern and the center of radiation remain unchanged.

Another kind of precaution relates to the time dependence of received field intensity beyond the horizon of smooth earth or in the deep shadows of rough terrain. Because there is often substantial time variation, it is well in before-and-after comparisons to include only line-of-sight receiving sites. When necessary, co-channel interference

should be kept to a minimum by use of a unidirectional receiving antenna.

Comparisons can be made by, say, eight radial recordings of field strength measured at constant height as for a coverage survey. However, if the only object of the test is demonstration of the improvement or change of the transmitting antenna performance, it can be done much more easily by excluding cluttered surroundings and making 100 foot recording samples at line-of-sight locations chosen to measure on significant parts of the elevation pattern. Of course, the measurements will be made in exactly the same places before and after the antenna change.

Accuracy

The accuracy of field measurements as an indication of broadcasting antenna performance is determined by (1) the care exercised in choosing measuring sites and (2) the quality and amount of the data.

The matter of expected accuracy for absolute field-intensity measurements over commonly encountered kinds of terrain merits discussion. Field-intensity meters depend, for absolute accuracy, upon a bolometer or a crystal diode in the calibrating generator. High-quality UHF signal generators are claimed to have absolute voltage accuracies of $\pm$ 1.5 to 2 decibels. One would hardly expect better from a field-intensity meter calibrating generator. Furthermore, there will ordinarily be mismatch error in field intensity measurements unless the $\lambda/2$ dipole effective length factor incorporates mismatch correction. Thus, an accuracy of $\pm$ 2 decibels for a field-intensity meter would, by present practice, be good. More can be expected with respect to *relative* accuracy at a given time and frequency since the instruments depend upon accurate and reliable piston attenuators for relative values. Piston attenuator accuracies in the order of tenths of a decibel are easily attained.

Other accuracies which are involved in field-intensity measurements include the transmitting antenna azimuth and elevation patterns, transmitter power, transmission line attenuation, diplexer attenuation, and transmitting antenna gain. Any field evaluation of transmitting antenna performance should be based upon knowledge of all of the quoted factors and their probable accuracies.

Appendix

First Fresnel (Shadowing) Zone Clearance

If free-space propagation theory is to be applicable, there must be a certain minimum clearance between the terrain and a ray from the

transmitting antenna to the measuring site. The first Fresnel (shadowing) zone is bounded by an ellipsoid with the transmitting antenna and the measuring point at the foci. When there is first Fresnel (shadowing) zone clearance over terrain which is rough by the smooth-earth criterion, free-space propagation is expected. The required clearance can be determined from Figure 4.

Test to Determine Validity of Ray Theory Over Plane Earth

It has been convenient to treat earth reflection as though occurring at a point (ray theory). Actually, there is a significant reflecting area, elliptical in shape, known as the First Fresnel (earth reflecting) zone.[4] This area can be located and defined. The length of the ellipse is $2x_1$, the width is $2y_1$, as shown in Figure 6. If the first Fresnel (earth reflecting) zone surface is smooth enough to satisfy the Rayleigh

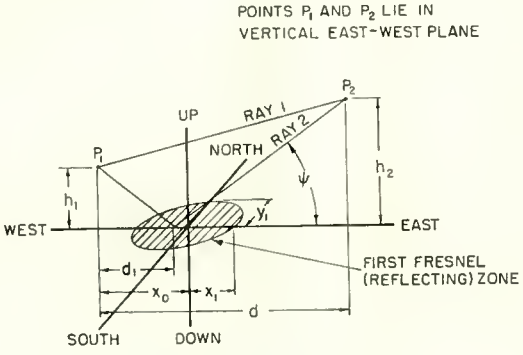


Fig. 6—Reflecting zone.

smooth-earth criterion, it is probable that the simple ray theory applies.

According to the smooth-earth criterion, a reflector behaves essentially like a mirror[5] if

$$H < \frac{\lambda}{16 \sin \psi} \tag{6}$$

where H is the height of a surface irregularity,

ψ is the angle of incidence.

According to this rule-of-thumb, reflection is specular or mirror-like

[4] K. A. Norton and A. C. Omberg, "The Maximum Range of a Radar Set," *Proc. I.R.E.*, Vol. 35, p. 4, January, 1947.
[5] H. Reed and C. Russell, *op. cit.*, pp. 236-238.

when the height, H, Figure 7, of surface irregularities satisfies the smooth-earth criterion.

When reflection can be assumed to be specular, based on the smooth-earth criterion, the reflection coefficient is assumed to be unity. Then the field-intensity contribution of either the direct or the earth-reflected ray can be computed from the free-space formula

$$E_0 = \frac{137.6\sqrt{P}}{r} \text{ millivolts per meter,}$$

where r is the length, in miles, of the rays. The interference maxima and minima will be, respectively, the sum and difference of the two contributions.

First Fresnel (Earth Reflecting) Zone

The formulas for computing the dimensions and location of the first Fresnel (earth reflecting) zone on plane earth follow. Refer to Figure 6 for the geometry and symbols.

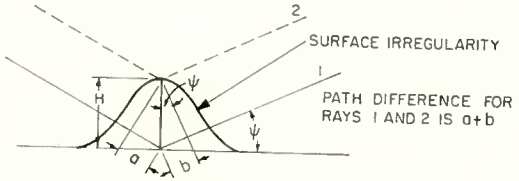


Fig. 7—Roughness criterion geometry.

wavelength, λ, is in feet

frequency, f, is in mc/sec.

heights h_1 and h_2 are in feet

distances x_0, x_1, y_1, d_1, d_2, and d are in miles.

The wavelength is

$$\lambda = \frac{980}{f}. \tag{7}$$

The grazing angle for the earth-reflected ray is

$$\psi = \tan^{-1}\left(\frac{h_1}{5280\, d_1}\right). \tag{8}$$

If $h_2 \gg h_1$, λ the distance to the earth-reflected-ray reflection point is

$$d_1 = \frac{h_1}{h_2} d. \tag{9}$$

The distance to the center of the first Fresnel (earth reflecting) zone ellipse is

$$x_0 \cong d_1 \left[1 + \frac{\lambda}{2 h_1 \sin \psi} \right]. \tag{10}$$

The length of the ellipse (major diameter) is

$$2 x_1 \cong 0.00378 \sqrt{\frac{h_1 \lambda}{\sin^3 \psi} \left[1 + \frac{\lambda}{4 h_1 \sin \psi} \right]}. \tag{11}$$

The width of the ellipse (minor diameter) is

$$2 y_1 \cong 2 x_1 \sin \psi. \tag{12}$$

While these formulas provide a useful guide, it should be recognized that significant reflections may occur outside the first Fresnel (earth reflecting) zone.

Transmitting Antenna Pattern and Power Gain

Since antenna gain is quoted in terms of r-m-s power gain, G, antenna azimuth and elevation patterns must be taken into account in the estimation of expected field intensity at a specific measuring site. Designating the azimuth pattern as $f(\phi)$ and the elevation pattern as $f(\theta)$, these patterns may be related to the coordinates of Figure 8. If the azimuth direction angle is ϕ_1 and the elevation angle θ_1 for a specific measuring site, the theoretical free-space field intensity at a point r_1 miles from the transmitting antenna will be

$$E_{\phi 1 \theta_1 r_1} = \frac{137.6 \sqrt{G P}}{r_1} f(\phi_1) f(\theta_1). \tag{13}$$

Avoidance of the Near-Field Region

The near-field region of the radiation pattern should be avoided since only far-field patterns are ordinarily published. Thus, measurements should not be made closer than

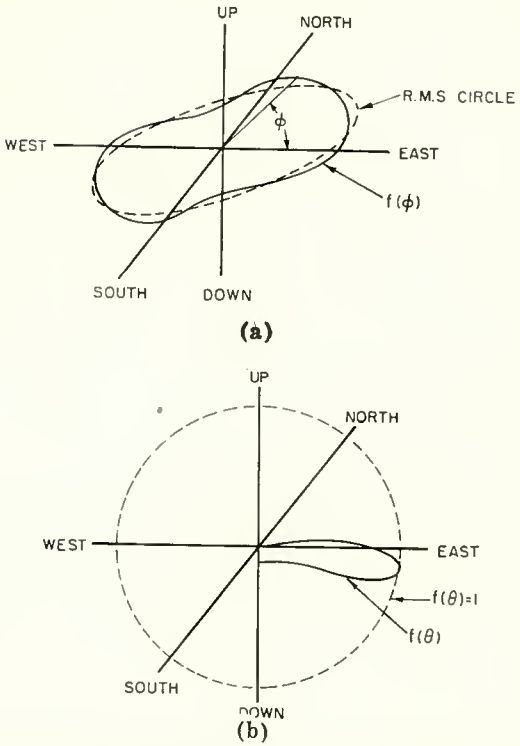


Fig. 8(a)—Azimuth pattern in horizontal NSEW plane; (b)—Elevation pattern in east plane.

$$d_{\min} = 2\frac{W^2}{\lambda}, \tag{14}$$

where $d_{\min}$ is the minimum measuring distance in feet,

W is the transmitting antenna aperture (length of radiating portion in feet),

λ is the wavelength in feet.

RCA TECHNICAL PAPERS†

Third Quarter, 1958

Any request for copies of papers listed herein should be addressed to the publication to which credited.

"Accuracy Control in a File Prccessor," J. C. Hammerton, *Electronic Engineering* (September) 1958

"Airborne Closed-Loop Television System," A. F. Flacco, *Jour. S.M.P.T.E.* (July) 1958

"Analysis of Low-Energy Sputtering," E. Langberg, *Phys. Rev.* (July 1) 1958

"The Application of Very Precise Frequency Control to Minimize Television Cochannel Interference," Part II, D. R. Mason, *Broadcast News* (August) 1958

"An Automatic Communications Switching System," R. E. Montijo, Jr., *Signal* (September) 1958

"Automatic Cuing of TP-6 Series Film Projectors," B. F. Melchionni, *Broadcast News* (August) 1958

"Batteries," C. K. Morehouse, R. Glicksman, and G. S. Lozier, *Proc. I.R.E.* (August) 1958

"Bilateral Conductivity in Power Transistors," I. G. Maloff, *Electronic Industries* (July) 1958

"Broadcast-Band Amplifier Circuits Using RCA-2N544 Drift Transistor," *RCA Application Note AN-176*, Semiconductor and Materials Division, Radio Corporation of America, Somerville, N. J. (September) 1958

"Broadcast-Band Frequency Converter Using RCA Transistors 2N140, 2N219, 2N411, or 2N412," *RCA Application Note AN-175*, Semiconductor and Materials Division, Radio Corporation of America, Somerville, N. J. (September) 1958

"Class B Complementary-Symmetry Audio Amplifiers," C. F. Wheatley, *Electronic Design* (August 6) 1958

"A Color 'Electrofax' Process," J. S. Rydz and S. W. Johnson, *RCA Review* (September) 1958

"Control Knobs for Military Electronic Equipment," T. G. Nessler, *Electrical Manufacturing* (September) 1958

"Design and Development of the 21CYP22 21-Inch Glass Color Picture Tube," C. P. Smith, A. M. Morrell, and R. C. Demmy, *RCA Review* (September) 1958

"Design Considerations for Direct-Coupled Transistor Amplifiers," J. E. Lindsay and H. J. Woll, *RCA Review* (September) 1958

"Design Methods to Improve the Stability of AM Directional Antenna Systems," G. H. Brown, *Broadcast News* (August) 1958

"Detection of Asymmetric Sideband Signals in the Presence of Noise," T. Murakami and R. W. Sonnenfeldt, *RCA Review* (September) 1958

"Diffraction by Smooth Cylindrical Mountains," H. E. J. Neugebauer and M. P. Bachynski, *Proc. I.R.E.* (September) 1958

"Drone Surveillance," J. L. Langevin, *Electronics* (July 4) (Letter to the Editor) 1958

"Effect of Crystal Growth Variables on Electrical and Structural Properties of Germanium," F. D. Rosi, *RCA Review* (September) 1958

"Effects of Radiation on Vidicon Performance," R. A. Davidson and B. H. Rosen, *Trans. I.R.E. PGNS* (August) 1958

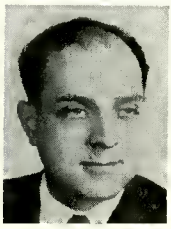
"Good Writing and Speech—Their Importance to the Engineer," A. N. Goldsmith, *Trans. I.R.E. PGEWS* (August) 1958
"Growth of Ferroelectric Crystals of the Type $(Glycine)_3H_2AB_1$," R. Nitsche, *Helvetica Physica Acta* (July) 1958
"Harmonic Generation with Nonlinear Reactances," K. K. N. Chang, *RCA Review* (September) 1958
"How to Get Best Performance from the TK-41 Color TV Camera," S. L. Bendell, H. N. Kozanowski, and T. J. Shipferling, *Broadcast News* (August) 1958
"How to Get Good Picture Quality from the TK-15 Vidicon Studio Camera," J. H. Roe, *Broadcast News* (August) 1958
"Large-Signal Surface Photovoltage Studies with Germanium," E. O. Johnson, *Phys. Rev.* (July 1) 1958
"Let's Use Our Heads," R. Samuel, *Trans. I.R.E. PGEWS* (August) 1958
"Loudspeakers and Negative Impedances," R. E. Werner, *Trans. I.R.E. PGA* (July-August) 1958
"Metallographic Aspects of Alloy Junctions," A. S. Rose, *RCA Review* (September) 1958
"A Method of Measuring the Optical Sine-Wave Spatial Spectrum of Television Image Display Devices," O. H. Schade, *Jour. S.M.P.T.E.* (September) 1958
"Minimize Local Oscillator Drift," Part 2, W. Y. Pan and D. J. Carlson, *Electronic Design* (August 6) 1958
"Mobility of Electrons in Germanium-Silicon Alloys," M. Glicksman, *Phys. Rev.* (July 1) 1958
"A New Arm for Vehicular Communications," J. R. Neubauer, *Trans. I.R.E. PGVC* (July) 1958
"New Broadband FM Antennas," W. First, *Broadcast News* (August) 1958
"A Note on the Dispersion of Interdigital Delay Lines," F. Paschke, *RCA Review* (September) 1958
"Optical Absorption in p-Type Gallium Arsenide," R. Braunstein Coauthor), *Phys. Rev.* (July 15) 1958
"A Parametric Amplifier Using Lower-Frequency Pumping," K. K. N. Chang and S. Bloom, *Proc. I.R.E.* (July) 1958
"Pickup Tube Performance With Slow Scanning Rates," C. T. Shelton and H. W. Stewart, *Jour. S.M.P.T.E.* (July) 1958
"Power Dissipation in Class B Circuitry," C. F. Wheatley, *Electronic Design* (September 17) 1958
"Proposal for Detection of Negative-Mass Carriers by Cyclotron Resonance," G. C. Dousmanis, *Phys. Rev. Letters* (July 15) ... 1958
"Radiation Damage in Ge and Si Detected by Carrier Lifetime Changes: Damage Thresholds," J. J. Loferski and P. Rappaport, *Phys. Rev.* (July 15) 1958
"Radiationless Recombination in Phosphors," L. Bess, *Phys. Rev.* (July 1) 1958
"The RCA Magnetic Disc Recorder," G. C. Weilenmann, *Broadcast News* (August) 1958
"RCA Video Tape Recorders in Operation at NBC Tape Central, Burbank, California," *Broadcast News* (August) 1958
"The Reaction of Sync Separators in Television Receivers to Impulse Noise," Dr. E. Luedicke, *Trans. I.R.E. PGBTR* (September) .. 1958
"The Relation of Utilization to the Shortage of Scientists," W. Milwitt and Co-authors, *Trans. I.R.E. PGEM* (September) 1958
"Reliability Control Based on Multiple Sequential Feedback," C. M. Ryerson, *Trans. I.R.E. PGRQC* (July) 1958
"Sealing a Calcium Fluoride Window to Glass," M. H. Greenblatt, *Rev. Sci. Instr.* (August) (Notes) 1958
"Simplified Theory of One-Carrier Currents with Field-Dependent Mobilities," M. A. Lampert, *Jour. Appl. Phys.* (July) 1958
"Solid-State Panel Amplifies X-Rays," B. Kazan, *Electronics* (September 12) 1958
"Some Aspects of Photoconductivity in Cadmium Selenide Crystals," R. H. Bube and L. A. Barton, *Jour. Chem. Phys.* (July) 1958

"Some Physical Properties of Tri-Glycina Sulfate and its Isomorths," E. Fatuzzo, *Helvetica Physica Acta* (July) 1958
"Sound Reproducing Systems—Monaural, Binaural, Monophonic, and Stereophonic," H. F. Olson, *Audio* (September) 1958
"Space-Charge-Limited Currents in ZnS Single Crystals," W. Ruppel, *Helvetica Physica Acta* (July) 1958
"Spectrographic Analysis of Silicon-Germanium Alloys," M. C. Gardels and H. Whitaker, *Analytical Chemistry* (September) 1958
"Stop-Go Scanning Saves Spectrum Space," H. E. Haynes and D. T. Hoger, *Electronics* (September 26) 1958
"Storage Capacity in Meteor-Burst Communication Systems," W. A. Helbig, *Proc. I.R.E.* (September) (Letter to the Editor) 1958
"Strategic Air Command Uses RCA Color Television to Expedite Vital Briefings," *Broadcast News* (August) 1958
"A Study of 468-Megacycle Tropospheric Scatter Propagation Over a 289-Mile Path," J. B. Atwood, G. B. MacKimmie, D. G. Shipley, and G. S. Wickizer, *RCA Review* (September) 1958
"A Study of the Molded Nickel Cathode," C. P. Hadley, W. G. Rudy, and A. J. Stoeckert, *Jour. Electrochem. Soc.* (July) 1958
"System Aspects," M. M. Tall, *Trans. I.R.E. PGRQC* (September) .. 1958
"Tension Stresses in Glass Coatings and in Glass-Metal Seals in the Annealing Range," J. C. Turnbull, *Jour. Amer. Cer. Soc.* (September) 1958
"Thermoelectric Heat Pumping," N. E. Lindenblad, *Elec. Eng.* (September) 1958
"Tube Noise Factor Chart," L. P. A. DeBacker, *Electronics* (July 18) (Electronics Reference Sheet) 1958
"TV Receiver Picture Area Losses," C. L. Townsend, *Trans. I.R.E. PGBTS* (September) 1958
"Unusual Tube Effects Cause Circuit Troubles," W. E. Babcock, *Electronics* (September 12) 1958
"Vacuum-Tube Requirements in Vertical-Deflection Circuits," K. W. Angel, *Trans. I.R.E. PGBTR* (September) 1958
"What Did They Say?" E. M. McElwee, *Trans. I.R.E. PGEWS* (August) 1958
"Which is What? A Review of Some Modern Amplitude-Modulation Systems," K. W. Uhler, *RCA Ham Tips* (September-October) 1958
"WINS Proves Reliability and Economy of 50-KW Ampliphase AM Transmitter," *Broadcast News* (August) 1958

AUTHORS

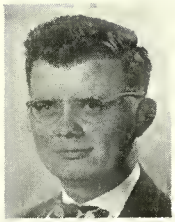
DONALD J. BLATTNER received the B.S. degree in Electrical Engineering in 1946 and the M.A. degree in Physics in 1949 from Columbia University. He joined the RCA Laboratories at Princeton, N.J. in 1953. He is now with the technical staff of RCA Microwave Advanced Development activity of the Electron Tube Division at Princeton, N.J. His work has been concerned with low-noise and traveling-wave-tube problems, backward-wave oscillators, parametric amplifiers, and microwave techniques for ultra-high-speed computers. Mr. Blattner is a member of Sigma Xi and a senior member of the Institute of Radio Engineers.

ROBERT W. BYLOFF received the B.S. degree in Electrical Engineering from the Massachusetts Institute of Technology in 1943. From 1943-46 he was in the U.S. Marine Corps serving overseas as a radar officer. After being discharged, he joined NBC as a junior project engineer. He has been a senior project engineer, supervisor of audio and video engineering design, and manager of project planning at NBC. In this latter position, he is responsible for specifying plans for all NBC technical facilities.



JACQUES DUTKA received the B.S. degree from the City College of New York, and the A.M. and Ph.D. degrees from Columbia University. From 1946 to 1953 he taught mathematics at Rutgers and Princeton Universities, and from 1953 to 1956 he was a mathematician with the Norden Laboratories. Since 1956 he has been a Senior Engineer with the Surface Communications Laboratory of RCA Defense Electronic Products in New York City. In addition he has, since 1954, been an Adjunct Associate Professor in the Electrical Engineering Department of Columbia University. Dr. Dutka is a member of Sigma Xi, the American Mathematical Society, and the Institute of Mathematical Statistics.

ROBERT D. GOLD received the B.E.E. degree from City College of New York in 1953. He spent three months with the RCA Victor Division as a Specialized Trainee before entering the U.S. Army. He worked on feasibility studies of radar counter-measure systems while in service. From 1955-1957 he attended Cornell University, where he received the M.S. degree in Electrical Engineering. He joined RCA Laboratories in 1957, where he has worked on television circuits and kinescopes, and on semiconductor devices. He also lectures in Electrical Engineering at City College of New York. Mr. Gold is a member of the Institute of Radio Engineers, Tau Beta Pi, and Eta Kappa Nu.



JOSEPH J. LOFERSKI received the B.S. degree from the University of Scranton in 1948, and the M.S. and Ph.D. degrees in Physics from the University of Pennsylvania in 1949 and 1953, respectively. He was a Research Associate at the University of Pennsylvania during 1952-53. He joined RCA Laboratories in 1952, where he has been engaged in research on semiconductor radiation converters, and radiation damage in semiconductors. Dr. Loferski is a member of Sigma Xi, American Physical Society, and the American Association for the Advancement of Science.

JOHN H. O'CONNELL received the B.S. degree in Electrical Engineering in 1954 from Fournier Institute of Technology, Lemont, Illinois. In the same year he joined RCA Laboratories and spent a year on a training program. Upon completion of his training program he joined the Industry Service Laboratory group in Princeton and he is now working on transistor circuit applications. Mr. O'Connell is an Associate Member of the Institute of Radio Engineers.

FRITZ PASCHKE (see *RCA Review*, Vol. XIX, No. 3, September 1958, p. 490.)

DONALD W. PETERSON (see *RCA Review*, Vol. XIX, No. 2, June 1958, p. 315.)

JAMES W. SCHWARTZ (see *RCA Review*, Vol. XIX, No. 2, June 1958, p. 316.)

PAUL RAPPAPORT studied chemistry at Temple University, Philadelphia, Pa. before entering the U. S. Navy as an electronic technician in 1944. As a civilian in 1946 he worked nine months at the Naval Air Experimental Station in Philadelphia in the capacity of Physicist. He received the B.S. and M.S. degrees in solid state physics in 1948 and 1949 at Carnegie Institute of Technology where he was a graduate teaching assistant from 1948 to 1949. He joined RCA Laboratories in 1949, where he has worked in the fields of secondary electron emission, direct conversion of radiation into electricity and radiation damage in semiconductors. Mr. Rappaport is a member of the American Physical Society, Pi Mu Epsilon, and Sigma Xi.

OTTO H. SCHADE was born and educated in Germany; he came to the United States in 1926. He joined the RCA Electron Tube Division at Harrison, N.J. in 1931. Since 1938 he has specialized in television circuits, camera tubes, and picture tubes. During the past few years he has been working on a unified method of image-system analysis including practical methods for measuring the aperture effect (sine-wave response function) of optical, photographic, and electronic system components. He has received numerous honors including the Modern Pioneers Award of the National Association of Manufacturers (1940), the Morris Liebmann Memorial Prize of the Institute of Radio Engineers (1950), the Fellow Award of the Institute of Radio Engineers (1951), and the Fellowship Award of the S.M.P.T.E. (1951). He is also the first recipient of the David Sarnoff Gold Medal Award of the Society of Motion Picture and Television Engineers (1951). Dr. Schade received the honorary degree of Doctor of Engineering from Rensselaer Polytechnic Institute in June, 1953.

T. M. SCOTT received the B.S. degree in Electrical Engineering from the University of Maryland in 1953, and the M.S. degree from the University of Wisconsin in 1958. He joined the technical staff of RCA Laboratories in 1953, and has been engaged in transistor measurements and circuit development. He is currently on leave of absence from RCA and is pursuing graduate studies at the University of Wisconsin, where he is also an instructor in electrical engineering. Mr. Scott is a member of Tau Beta Pi, Phi Kappa Phi, and an Associate Member of the Institute of Electrical Engineers.

WILLIAM E. SPICER received the B.S. degree from William and Mary College in 1949, the S.B. degree from M.I.T. in 1951, and the M.A. and Ph.D. degrees from the University of Missouri in 1953 and 1955, respectively. He joined the RCA Laboratories at Princeton in 1955 where he has been engaged in studying electron emission processes. Dr. Spicer is a member of Phi Beta Kappa, Sigma Xi, and the American Physical Society.

FRED STERZER received the B.S. degree in physics from the City College of New York in 1951, and the M.S. and Ph.D. degrees in physics from New York University in 1952 and 1955, respectively. He worked for the Allied Control Co. in New York City from 1952 to 1953. During the school year of 1953 to 1954 he was an instructor in physics at the Newark College of Engineering in Newark, N. J., and a research assistant at New York University working on microwave spectroscopy. He joined the RCA Tube Division in Harrison, N. J., in October, 1954, and transferred to Princeton in 1956. His work with the microwave tube advanced development activity is concerned primarily with traveling-wave tubes and backward-wave oscillators, parametric devices, and components for ultra-high-speed computers. Dr. Sterzer is a member of Phi Beta Kappa, Sigma Xi, the American Physical Society, and the Institute of Radio Engineers.

RCA REVIEW

a technical journal

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

INDEX

VOLUME XIX

TABLE OF CONTENTS

March

June

AUTHORS, VOLUME XIX