RCA REVIEW

a technical journal

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

VOLUME XVIII **JUNE 1957** **NO. 2**

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 at the Post Office at Princeton, New Jersey, under the act of March 3, 1879. 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 XVIII JUNE, 1957 NUMBER 2

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

DESIGN CONSIDERATIONS IN THE FIRST STAGE OF TRANSISTOR RECEIVERS

BY

LARRY A. FREEDMAN

RCA Laboratories,
Princeton, N. J.

Summary—*This paper presents a discussion of noise performance of transistor r-f stages utilizing capacitive antennas and of transistor mixer stages utilizing loop antennas. Examples of the noise performance to be expected with each type of antenna are included. Comparisons are drawn between transistor stages and corresponding tube stages. Consideration is given to design compromises between image rejection and insertion loss for an r-f stage employing both tuned input and interstage transformers. The procedure for transformer design for optimum insertion loss–image rejection performance is outlined.*

NOISE PERFORMANCE

Capacitive Antenna

A TYPICAL transistor r-f stage is shown in Figure 1a. A capacitive-type antenna is assumed (as for example the rod antenna of an automobile receiver). Figure 1b shows an equivalent circuit from the antenna to the r-f stage input where the antenna has been replaced by a voltage source, V_A, and an internal impedance consisting of a capacitance C_A. The transistor is represented by its input resistance, R_i, and the transformer tuned impedance by R_o. The signal power delivered to the r-f stage input by an unmodulated signal is given by

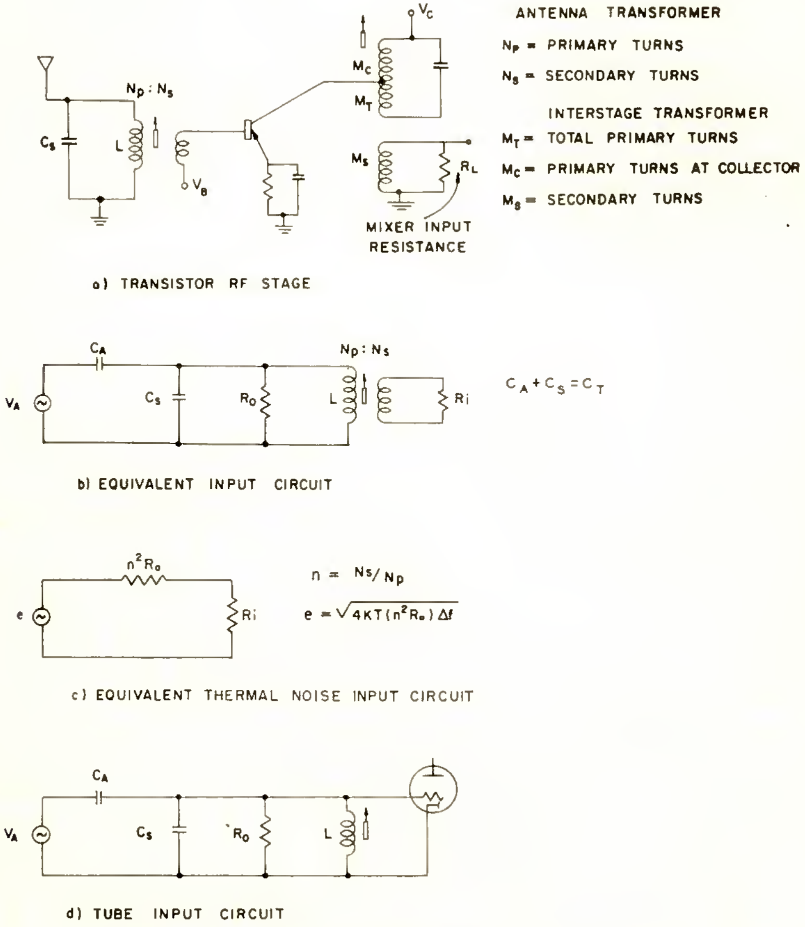
$$P_s = \left[n\, V_A \frac{C_A}{C_T} Q_{01} \frac{R_i}{n^2 R_o + R_i} \right]^2 \frac{1}{R_i}, \qquad (1)$$

where n is the transformer turns ratio, N_s/N_p, as indicated in Figure 1 (essentially unity coupling is assumed), C_T is the sum of the antenna capacitance and the total shunt capacitance across the antenna transformer, and Q_{01} is the unloaded Q of the antenna transformer. This expression is derived in Appendix I.

The equivalent thermal noise generator referred to the secondary of the equivalent circuit of Figure 1b is shown in Figure 1c. The antenna radiation resistance is negligible compared to R_o and is ignored. The thermal noise power delivered to the transistor is

$$P_{TH} = 4kT\, n^2\, R_o\, \Delta f \left[\frac{R_i}{n^2 R_o + R_i} \right]^2 \frac{1}{R_i}, \tag{2}$$

where k is Boltzmann's constant, T is absolute temperature, and Δf is the noise bandwidth. Since the r-f transistor noise factor, F, may be expressed as P_N/P_{TH}, where P_N is the total noise power referred to the transistor input,



a) TRANSISTOR RF STAGE

b) EQUIVALENT INPUT CIRCUIT

c) EQUIVALENT THERMAL NOISE INPUT CIRCUIT

d) TUBE INPUT CIRCUIT

Fig. 1—R-F stage utilizing capacitive antenna.

$$P_N = F P_{TH} \tag{3}$$

(assuming that noise from the second and succeeding stages is small and may be neglected). The signal-to-noise ratio is obtained from Equations (1), (2), and (3) and is given by

$$\frac{P_S}{P_N} = \frac{\left[V_A \dfrac{C_A}{C_T} Q_{01} \right]^2}{4kT\, R_o\, \Delta f\, F}. \tag{4}$$

For a receiver employing a linear detector the signal-to-noise ratio at the output of the audio amplifier, $(P_S/P_N)_A$, is related to the r-f signal-to-noise ratio by

$$\left(\frac{P_S}{P_N} \right)_A = m^2 \frac{P_S}{P_N} \frac{\Delta f}{\Delta f_A}, \tag{5}$$

where m is the per cent modulation and Δf_A is the over-all noise bandwidth of the receiver. (For the usual case where the receiver audio bandwidth is less than half of the bandwidth to the detector input, the over-all noise bandwidth is twice the audio noise bandwidth.) For a 30 per cent modulated signal and an output signal-to-noise ratio of 20 decibels, $(P_S/P_N)_A = 100$; the antenna voltage required for this signal-to-noise ratio is obtained by combining Equations (4) and (5) and is given by

$$V_A = \frac{10\sqrt{4kT\, R_o\, \Delta f_A F}}{0.3 \dfrac{C_A}{C_T} Q_{01}}. \tag{6}$$

This expression is retained in this form to facilitate comparison with the tube case discussed below (the mutual dependence of R_o, C_T, and Q_{01} should be noted, i.e., $Q_{01} \approx \omega\, C_T\, R_o$). Note that Equation (6) does not appear to be directly dependent on the turns ratio or operating Q of the antenna transformer. The transistor noise factor, however, is dependent on the turns ratio to the extent that the turns ratio determines the driving source impedance. This noise factor dependence is quite small in the vicinity of the matched source impedance. For example, a four-to-one mismatch in either direction causes an increase in F of about 2 decibels* corresponding to an increase of about 25 per cent in the signal required for 20 decibels signal-to-noise ratio.

A comparison of the noise performance of a tube and transistor r-f

* Determined experimentally.

stage may be made by comparing Equation (6) with a like expression for a tube stage. Figure 1d shows a typical antenna connection for a tube r-f stage where the primary of the antenna transformer of Figure 1b is connected directly to the grid. The tube noise factor, F_T, may be expressed as

$$F_T = \frac{R_o + R_{eq}}{R_o} \qquad (7)$$

where R_{eq} is the equivalent noise resistance of the tube. For typical remote-cutoff pentodes, the equivalent noise resistance is of the order of 3,000 to 5,000 ohms. For this type of connection at broadcast frequencies R_o is usually large compared to 5,000 ohms and F_T is very nearly unity. The signal required for 20 decibels signal-to-noise ratio is then

$$V_A = \frac{10\sqrt{4kT\, R_o\, \Delta f_A}}{0.3 \dfrac{C_A}{C_T} Q_{01}} . \qquad (8)$$

A comparison of Equations (6) and (8) shows that the tube r-f stage will be $\sqrt{F}$ times better than a transistor stage.

Example of Noise Performance Calculation

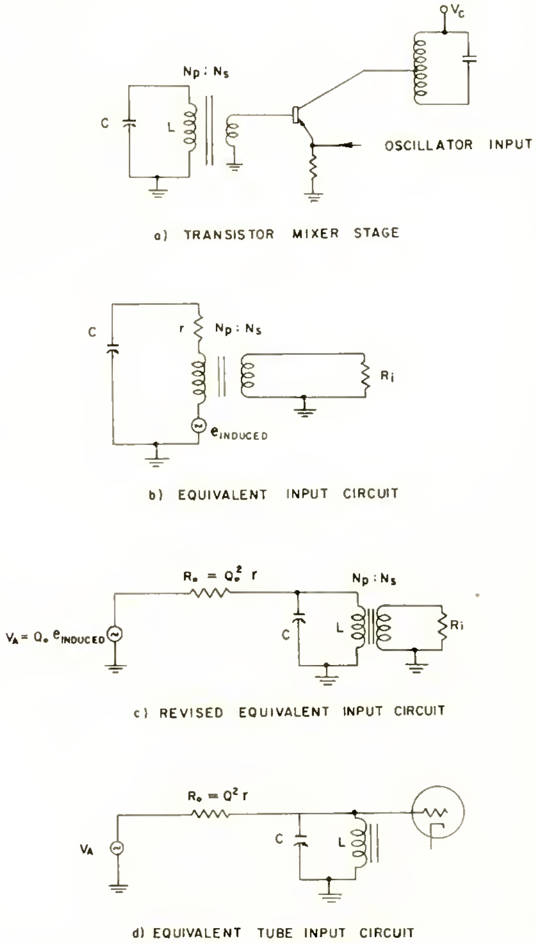
As an example, consider the r-f stage of a receiver for which the input circuit parameters are $C_A = 30$ micromicrofarads, $C_T = 100$ micromicrofarads, and $Q_{01} = 70$. The noise factor of the r-f transistor is 6 decibels and the over-all noise bandwidth is 3 kilocycles. Substituting the pertinent values in Equation (6),

$$V_A = \frac{10\sqrt{1.6 \times 10^{-20} \times 112 \times 10^3 \times 3 \times 10^3 \times 4}}{0.3 \times \dfrac{30}{100} \times 70} = 7.37 \text{ microvolts}$$

for a 20-decibel signal-to-noise ratio at 1 megacycle.

Loop Antenna

A mixer or converter first stage is frequently used in receivers employing a loop antenna. A transistor mixer first stage is shown in Figure 2a. The antenna circuit typically consists of a coil wound on a ferrite rod, L, a tuning capacitor, C, and a secondary winding for impedance transformation from antenna to transistor input. An input equivalent circuit of Figure 2a is shown in Figure 2b where e is the voltage induced by the signal field, r is the series resistance of the tuned transformer, and R_i is the transistor input resistance. Another equivalent circuit is shown in Figure 2c which is in the form of Figure



a) TRANSISTOR MIXER STAGE

b) EQUIVALENT INPUT CIRCUIT

c) REVISED EQUIVALENT INPUT CIRCUIT

d) EQUIVALENT TUBE INPUT CIRCUIT

Fig. 2—Mixer stage utilizing loop antenna.

1b. The voltage V_A, corresponding to a 20-decibel signal-to-noise ratio is given by

$$V_A = \frac{10\sqrt{4kT\,R_0 \Delta f_A F}}{0.3}, \tag{9}$$

which is derived in Appendix II. In terms of the induced voltage, Equation (9) may be written

$$e_{\text{induced}} = \frac{10\sqrt{4kT\,R_o \Delta f_A F}}{0.3\,Q_o}. \tag{10}$$

The signal input is usually described in terms of field strength so that making use of Equation (10),

$$\text{Field Strength} = \frac{e}{h} = \frac{10\sqrt{4kT\,R_o \Delta f_A F}}{0.3\,Q_o\,h} \tag{11}$$

for a 20-decibel signal-to-noise ratio, where h is the effective height of the antenna.

A comparison of the noise performance of a transistor mixer stage and a tube converter stage may be made by the use of Equation (11). An equivalent input circuit for a tube stage is shown in Figure 2d where the primary of the antenna transformer of Figure 2c is connected directly to the signal grid. The value of the tube noise factor to be used in Equation (11), depending on the input circuit parameters and the type of converter tube, may be determined from Equation (7). For typical converter tubes, R_{eq} is of the order of 150,000 to 300,000 ohms and the corresponding F_T is of the order of 1.5 to 6 decibels.

Examples of Noise-Performance Calculations for Loop Antennas

Noise-performance calculations will be made for two battery-operated receivers employing ferrite cored loop antennas, one receiver having a transistor mixer first stage and the other a pentagrid converter tube first stage. The following information applies for both receivers. The antenna core consists of a ferrite rod 0.25 inch in diameter by 7 inches long. The variable tuning capacitance is 8 to 172 micromicrofarads. The effective height of the antenna is about 0.004 meter at 1,000 kilocycles.* The noise bandwidth of the receivers is 3 kilocycles.

The receiver employing the transistor mixer has an unloaded antenna circuit Q of 200. Substituting the pertinent values in Equation (11),

* This value of effective height applies for a particular core material and inductance. It should be noted that for a given core the effective height increases with increasing inductance, but for a constant Q the ratio $\sqrt{R_o}/Q_o h$ appearing in Equation (11) is unchanged.

$$\text{Field Strength} = \frac{10\sqrt{1.6 \times 10^{-20} \times 612 \times 10^{3} \times 3 \times 10^{3}\,F}}{0.3 \times 200 \times 4 \times 10^{-3}} =$$

$$226\sqrt{F} \text{ microvolts per meter}$$

for 20 decibels signal-to-noise ratio at 1,000 kilocycles. For a transistor converter having a noise factor of 6 decibels, the field strength for a 20-decibel signal-to-noise ratio is 452 microvolts per meter.

The second receiver employs a 1R5, a popular converter tube in battery portables, having an equivalent noise resistance of about 200,000 ohms. The antenna unloaded Q is usually limited to about 100 by other circuit considerations. (Signal feedback, loss of high-frequency audio response, and loss of sensitivity due to tracking errors are all aggravated by increased Q.) Substituting the required terms in Equation (11),

$$\text{Field Strength} = \frac{10\sqrt{1.6 \times 10^{-20} \times 306 \times 10^{3} \times 3 \times 10^{3}\,F}}{0.3 \times 100 \times 4 \times 10^{-3}} =$$

$$320\sqrt{F} \text{ microvolts per meter}$$

for a 20-decibel signal-to-noise ratio at 1,000 kilocycles. The converter noise factor, as obtained from Equation (7) is 2.2 decibels at 1,000 kilocycles, resulting in a field strength of 412 microvolts per meter for a 20-decibel signal-to-noise ratio.

Insertion Loss and Image Rejection

Equivalent input and output circuits for the r-f stage of Figure 1a are shown in Figure 3. These circuits are convenient for the consideration of insertion loss and image rejection. In Figure 3a, the voltage source of Figure 1b is replaced by an equivalent current source, i, and the transistor input resistance is referred to the transformer primary and designated R_1. The transistor input power is

$$P_1 = i^2 R_o^{\,2} \frac{R_1}{(R_o + R_1)^2}, \qquad (12)$$

which maximizes at $i^2 R_1/4$ when $R_1 = R_o$. A curve of input power relative to maximum transistor input power (hereafter referred to as insertion loss) for varying R_1/R_o is shown in Figure 4.

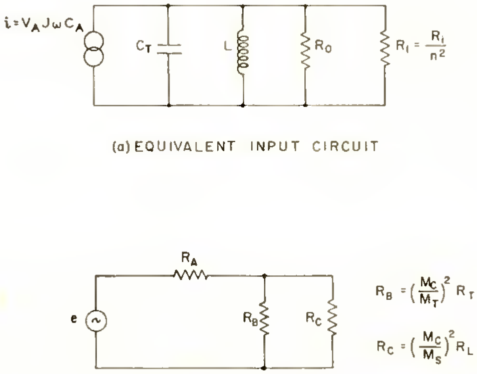
The operating Q of the antenna transformer, designated Q_1, is related to the unloaded Q by the relation

$$Q_1 = \frac{1}{1 + \dfrac{R_o}{R_1}} Q_{01}, \tag{13}$$

and the image rejection ratio, I, is very nearly

$$I = Q_1 \left(\frac{f}{f_o} - \frac{f_o}{f} \right), \tag{14}$$

where f_o is the frequency of the desired signal and f is the image frequency. Equation (12) is valid[1] for values of Q_1 greater than 10. Maximum image rejection is obtained when the transformer is unloaded $(Q_1 = Q_{01})$. The loss of image rejection relative to this maximum is also plotted in Figure 4.



(a) EQUIVALENT INPUT CIRCUIT

(b) EQUIVALENT INTERSTAGE CIRCUIT

Fig. 3—Equivalent circuits of Figure 1 r-f stage.

A significant improvement in image rejection at a small sacrifice in insertion loss results from operation at values of R_1/R_o greater than unity. For instance, a change from $R_1/R_o = 1$ to $R_1/R_o = 2.5$ results in an improvement of 3 decibels in image rejection at the expense of less than 1 decibel in insertion loss.

An output equivalent circuit which consists of a generator coupled to a load by means of a tuned transformer is shown in Figure 3b. The elements R_A, R_B, and R_C represent the output resistance of the r-f

[1] F. E. Terman, *Radio Engineers Handbook*, McGraw-Hill Book Company, Inc., New York, 1943, p. 144.

stage, the interstage transformer tuned resistance, and the load resistance, respectively. Maximum operating Q (maximum image rejection) for a given insertion loss obtains for the condition $R_C = R_A$. For this condition the power into the load resistance is

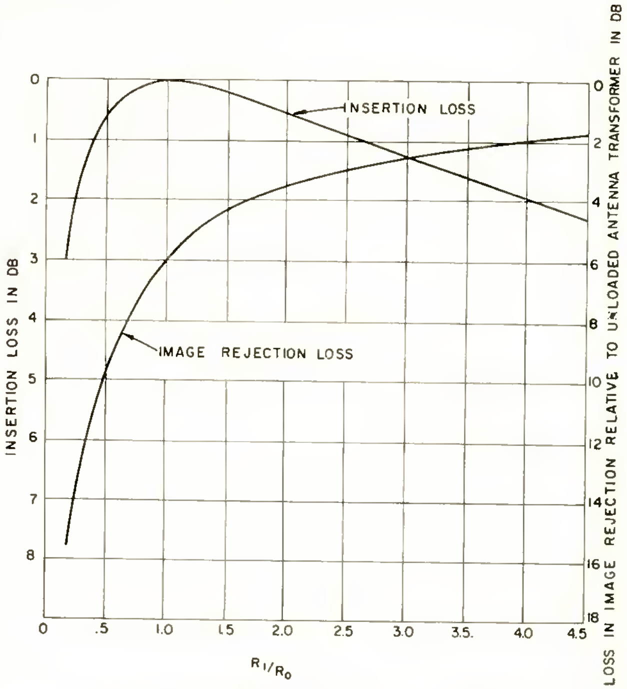


Fig. 4—Variation of insertion loss and image rejection with the ratio R_1/R_o for the antenna circuit of Figure 1.

$$P_2 = e^2 \frac{\left(\dfrac{R_B}{R_A}\right)^2}{R_A\left(2\dfrac{R_B}{R_A}+1\right)^2}, \tag{15}$$

and the operating Q of this transformer is

$$Q_2 = \frac{1}{1+2\dfrac{R_B}{R_A}} Q_{02}, \tag{16}$$

where Q_{02} is the unloaded Q.

The insertion loss, determined from Equation (15), is shown in Figure 5 as a function of R_B/R_A. The image rejection relative to an unloaded tuned transformer, also shown in Figure 5, is obtained from Equations (14) and (16) for varying R_B/R_A.

Example of R-F Stage Design

The design of an r-f stage which provides near optimum noise performance and 66 decibels image rejection in a receiver employing a 455-kilocycle i-f will be considered. The antenna and interstage transformers which were used each have an unloaded Q of 70.

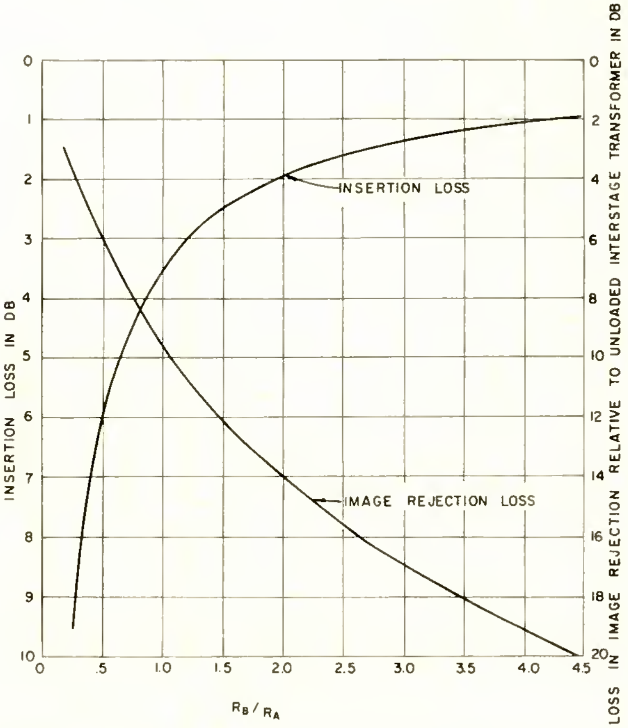


Fig. 5—Variation of insertion loss and image rejection with the ratio R_B/R_A for the interstage circuit of Figure 1.

Near optimum noise performance (determined experimentally) is obtained in the range of R_1/R_o from 0.4 to 2.5; for this example let $R_1/R_o = 1.5$. From Figure 4, the image rejection is 4.3 decibels below maximum. The maximum image rejection as calculated from Equation (14) would be 39.7 decibels at 1,000 kilocycles for $Q_o = 70$. Therefore, the antenna transformer provides 35.4 decibels of image rejection and an insertion loss of 0.2 decibel over that obtaining for the maximum input power condition. The remainder of the required image rejection, 30.6 decibels, is to be obtained from the interstage transformer. This allows for a tolerable loss of image rejection of 9.1 decibels, relative to an unloaded transformer. As determined from Figure 5, the ratio R_B/R_A is 0.92 and the consequent insertion loss is 3.8 decibels.

The r-f transistor has matched terminating resistances of 250 ohms input and 10,000 ohms output and 30 decibels matched gain; the mixer input resistance is 500 ohms. The capacitance required to tune the antenna transformer is 100 micromicrofarads; the capacitance required to tune the interstage transformer is 500 micromicrofarads. From the above information the turns ratios of both transformers may be determined. The tuned resistance of the antenna transformer at 1,000 kilocycles is

$$R_o = Q_o X_{C_T} = 70 \frac{1}{2\pi\, 10^6 \times 10^{-10}} = 112{,}000 \text{ ohms.}$$

For the turns ratio, n, of the antenna transformer,

$$1.5 = \frac{R_1}{R_o} = \frac{R_i}{n^2 R_o}, \quad \text{or} \quad n = \sqrt{\frac{R_i}{1.5\, R_o}}.$$

Inserting the values of R_i and R_o in this expression,

$$n = \sqrt{\frac{250}{1.5 \times 112{,}000}} = \frac{1}{26},$$

assuming unity coefficient of coupling. The turns ratios of the interstage transformer may be calculated in a similar manner. The tuned resistance at 1,000 kilocycles is given by

$$R_B = Q_{02}\, X_c = 70 \frac{1}{2\pi\, 10^6 \times 5 \times 10^{-10}} = 22{,}400 \text{ ohms.}$$

The turns ratio from the collector to the top of the tuned circuit is given by

$$\frac{M_C}{M_T} = \sqrt{\frac{0.92 \times 10{,}000}{22{,}400}} = \frac{1}{1.56},$$

and the collector-to-secondary turns ratio is given

$$\frac{M_C}{M_S} = \sqrt{\frac{10{,}000}{500}} = \frac{1}{0.224}.$$

The r-f stage gain from the base of the r-f transistor to the base of the mixer will be (30 — 3.8) or 26.2 decibels. The circuit with normalized turns ratios is shown in Figure 6.

In the above design procedure, consideration was limited to operation at 1,000 kilocycles. This procedure is satisfactory only if performance at the extremes of the band is not severely compromised. The relative change in image rejection obtained with two tuned circuits which are either unloaded or loaded by a constant amount (R_o/R_1 and R_B/R_A constant) is shown in Figure 7a, Curve A. If the transistor input and output resistances are constant over the band, the ratios R_o/R_1 and R_B/R_A will not be constant. Both the insertion loss and image rejection will differ, at the extremes of the band, from the mid-band value. The insertion loss will increase and the change in image rejection decrease, with respect to the constant-load conditions, as the frequency increases. The amount of change will depend on the design values of R_o/R_1 and R_B/R_A and may be obtained from Figures 4 and 5 with the aid of Equations (11) to (15).

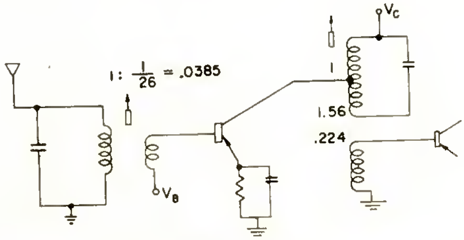
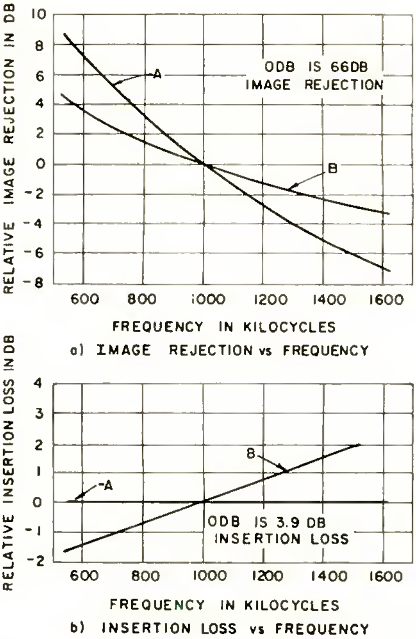


Fig. 6—Example of r-f stage input and interstage design.

The calculated variation in image rejection and insertion loss over the broadcast band for the circuit of Figure 6 is shown in Figure 7, Curve B, assuming constant transistor input and output resistances. Note in Figure 7b that while about ±2 decibels change in sensitivity results, the ratio of R_o/R_1 varies from 0.54 to 1.6 of the mid-band value, and is thus maintained within the previously defined limits for good noise performance. For a practical case with many present day transistors, the input and output resistances may decrease almost as rapidly as R_o and R_B. Curves A of Figure 7 will then be a closer approximation to the actual performance.

Minimum Insertion Loss for Prescribed Image Rejection

While the above design procedure usually yields satisfactory results, it is interesting to consider another approach. The total insertion loss, l, in the antenna and interstage transformers as obtained from Equations (11) and (14) is



a) IMAGE REJECTION vs FREQUENCY

b) INSERTION LOSS vs FREQUENCY

Fig. 7—Image rejection and insertion loss of circuit of Figure 6: (A) constant loading; (B) constant transistor input and output resistances.

$$l = \frac{\left(\dfrac{R_o}{R_1} + 1\right)^2 \left(2\dfrac{R_B}{R_A} + 1\right)^2}{4\dfrac{R_o}{R_1}\left(2\dfrac{R_B}{R_A}\right)^2} \tag{17}$$

and the total image rejection, obtained by combining Equations (13), (14), and (16), is

$$I = \frac{1}{\dfrac{R_o}{R_1} + 1} \times \frac{1}{\dfrac{2R_B}{R_A} + 1} Q_{01}\, Q_{02} \left(\frac{f}{f_o} - \frac{f_o}{f}\right)^2. \tag{18}$$

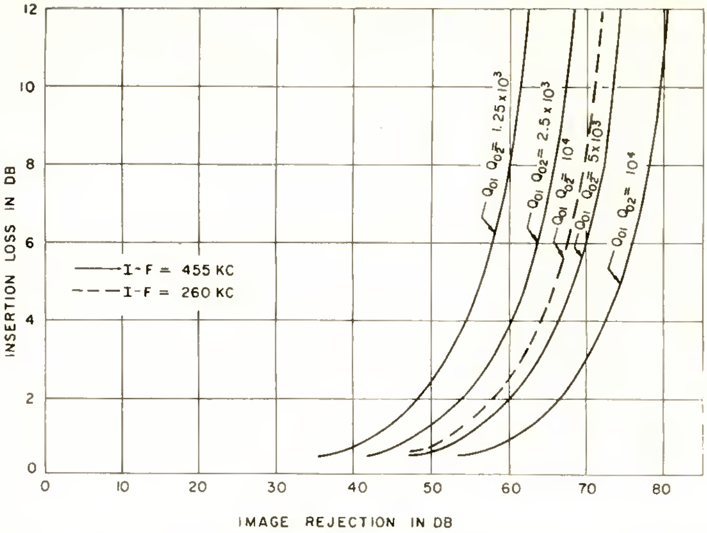


Fig. 8—Minimum insertion loss versus image rejection at 1,000 kilocycles for r-f stage of Figure 1.

The values of R_o/R_1 and R_B/R_A which give minimum loss for a given image rejection may be obtained from Equations (17) and (18) and are derived in Appendix III. A plot of minimum insertion loss versus image rejection, as obtained from Equation (17), for various values of $Q_{01}\ Q_{02}$ and an intermediate frequency of 455 kilocycles is shown in

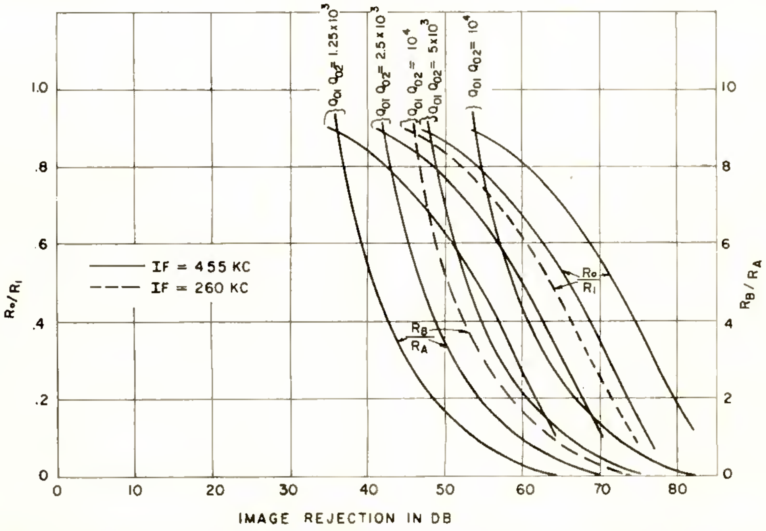


Fig. 9—R_o/R_1 and R_B/R_A versus image rejection (these values provide minimum insertion loss at 1,000 kilocycles for the r-f stage of Figure 1).

Figure 8. A change of 2-to-1 in the $Q_{01}\, Q_{02}$ product causes a 6-decibel change in image rejection for constant insertion loss. If, however, the image rejection is to be held constant, which is the usual case, the change in insertion loss for a change in $Q_{01}\, Q_{02}$ depends on both the value of image rejection and the intermediate frequency. A curve for an i-f of 260 kilocycles is plotted for comparison. Note that the 260-kilocycle curve indicates a loss of image rejection of 8.3 decibels (at 1,000 kilocycles) as compared to the 455-kilocycle curve for the same $Q_{01}\, Q_{02}$. Corresponding curves of R_o/R_1 and R_B/R_A, from which the transformer turns ratios may be determined, are shown in Figure 9.

Referring to the sample design in the previous section for a 455-kilocycle i-f and $Q_{01} = Q_{02} = 70$. Figure 8 indicates an insertion loss of 3.7 decibels for a 66-decibel image rejection. The values of R_o/R_1 and R_B/R_A, from Figure 9, are 0.505 and 1.05, respectively. Note that this value of R_o/R_1 is within the defined limits for good noise performance. The corresponding turns ratios, determined as in the above example, are shown on the circuit diagram of Figure 10. For this design a negligible decrease in insertion loss (0.2 decibel) is obtained; other values of $Q_{01}Q_{02}$ and image rejection may provide a more striking difference. This method produces near optimum noise performance only by chance. Therefore, its chief value lies in evaluating the results of the design compromises of the previous section.

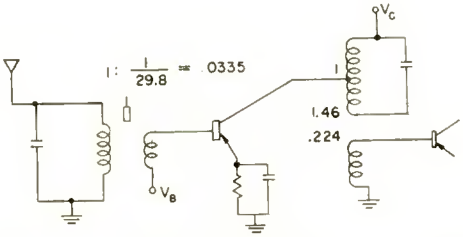


Fig. 10—Example of r-f stage input and interstage design for minimum insertion loss.

Appendix I—Derivation of Signal Power Delivered by a Capacitive Antenna

Referring to the equivalent circuit of Figure 1b and using the notation of the text, the signal voltage at the primary of the transformer is given by

$$V_S = V_A \frac{C_A}{C_T} Q_1 ,$$

where Q_1 is the loaded (operating) Q of the transformer. The signal voltage at the secondary winding is then $V_{S2} = nV_A\ (C_A/C_T)\ Q_1$, and the signal power delivered to the transistor is

$$P_S = \frac{(V_{S2})^2}{R_i} = \left(n\, V_A \frac{C_A}{C_T} Q_1 \right)^2 \frac{1}{R_i} .$$

Since the operating Q is related to the unloaded Q,

$$Q_1 = \frac{R_o R_i/n^2}{R_o + R_i/n^2} \times \frac{1}{\omega L} = Q_{01} \frac{R_i}{n^2\, R_o + R_i} .$$

The signal power may be expressed as

$$P_S = \left(n\, V_A \frac{C_A}{C_T} Q_{01} \frac{R_i}{n^2 R_o + R_i} \right)^2 \frac{1}{R_i} . \qquad (1)$$

Appendix II—Derivation of Signal Required for 20 Decibels Signal-to-Noise Ratio for a Loop Antenna

Referring to the equivalent circuit of Figure 2c and using the notation of the text, the signal voltage at the primary of the transformer is

$$V_S = V_A \frac{R_i/n^2}{R_o + R_i/n^2} = V_A \frac{Q}{Q_o} .$$

The signal voltage at the secondary winding is then $V_{S2} = n\ V_A\ Q/Q_o$, and the signal power delivered to the transistor is

$$P_S = \frac{(V_{S2})^2}{R_i} = (n\ V_A\ Q/Q_o)^2 \times \frac{1}{R_i} .$$

The noise power referred to the transistor input is

$$P_N = 4kT\ n^2\ R_o \Delta f \left[\frac{R_i}{n^2\ R_o + R_i} \right]^2 \frac{1}{R_i} F,$$

where F is the transistor noise factor.

For an output signal to noise ratio of 20 decibels and 30 per cent modulation,

$$\left(\frac{P_S}{P_N}\right)_A = 100 = \frac{(0.3n\, V_A\, Q/Q_o)^2 \dfrac{1}{R_i}}{4kT\, n^2\, R_o\, \Delta f_A \left[\dfrac{R_i}{n^2\, R_o + R_i}\right]^2 \dfrac{1}{R_i} F},$$

or,

$$V_A = \frac{10\sqrt{4kT\, R_o\, \Delta f_A\, F}}{0.3}. \qquad (9)$$

Appendix III—Derivation of Minimum Insertion Loss for a Specified Image Rejection

Equation (18) of the text is

$$I = \frac{1}{\dfrac{R_o}{R_1} + 1} \times \frac{1}{\dfrac{2R_B}{R_A} + 1}\, Q_{01}\, Q_{02} \left(\frac{f}{f_o} - \frac{f_o}{f}\right)^2, \qquad (18)$$

which may be written as

$$\frac{R_o}{R_1} + 1 = \frac{\left(\dfrac{f}{f_o} - \dfrac{f_o}{f}\right)^2}{I}\, Q_{01}\, Q_{02}\, \frac{1}{\dfrac{2R_B}{R_A} + 1}.$$

Introducing a constant,

$$C = \frac{\left(\dfrac{f}{f_o} - \dfrac{f_o}{f}\right)^2}{I}\, Q_{01}\, Q_{02},$$

which depends on the required image rejection, the intermediate frequency, and the unloaded Q's, Equation (18) becomes

$$\frac{R_o}{R_1} + 1 = \frac{C}{2\dfrac{R_B}{K_A} + 1}.$$

Substituting in Equation (17),

$$l = \frac{C^2}{4} \frac{1}{(R_o/R_1)} \frac{\left[\left(\frac{R_o}{R_1}\right) + 1\right]^2}{\left[C - \left(\frac{R_o}{R_1}\right) - 1\right]^2}. \tag{17a}$$

Differentiating Equation (17a) with respect to R_o/R_1,

$$\frac{\partial l}{\partial\left(\frac{R_o}{R_1}\right)} = \frac{C^2}{4}\left[\left\{\left(\frac{R_o}{R_1}\right)^3 + 2\left(\frac{R_o}{R_1}\right)^2(1-C) + \left(\frac{R_o}{R_1}\right)(1-C)^2\right\} 2\left(\frac{R_o}{R_1}+1\right) - \left(\frac{R_o}{R_1}+1\right)^2\left\{3\left(\frac{R_o}{R_1}\right)^2 + 4\left(\frac{R_o}{R_1}\right)(1-C) + (1-C)^2\right\}\right]\left[\left(\frac{R_o}{R_1}\right)^3 + 2\left(\frac{R_o}{R_1}\right)^2(1-C) + \frac{R_o}{R_1}(1-C)^2\right]^{-2}$$

Equating this expression to 0 to determine the value of R_o/R_1 for minimum l,

$$\frac{R_o}{R_1} = \frac{-(C+2) + \sqrt{C^2+8C}}{2},$$

and the corresponding value of R_B/R_A is

$$\frac{R_B}{R_A} = \frac{C}{-(C+2) + \sqrt{C^2+8C}} - \frac{1}{2}.$$

The minimum loss for a given C is then given by Equation (17a):

$$l = \frac{C^2}{2} \frac{\left[-C + \sqrt{C^2+8C}\right]^2}{\left[-(C+2) + \sqrt{C^2+8C}\right]\left[3C - \sqrt{C^2+8C}\right]^2}.$$

OPTIMIZING THE DYNAMIC PARAMETERS OF A TRACK-WHILE-SCAN SYSTEM

By

JACK SKLANSKY

RCA Laboratories,
Princeton, N. J.

Summary—*Systems such as search radar, in which the input data arrives intermittently, frequently require a device for continuously estimating the "present" value of the input. In radar terminology, this device is called a "track-while-scan system."*

A common type of track-while-scan system is characterized by two dynamic parameters, one parameter correcting the position error, and the other correcting the velocity error. Such a system is analyzed here, and a scheme for optimizing the two dynamic parameters is suggested. Tools are derived for the optimization scheme in the form of charts and formulas describing the stability, transient response, noise, and maneuver error as functions of the dynamic parameters.

One interesting result is a formula for the mean square response of the system to white noise.

A numerical example illustrates a suggested optimization procedure using the derived charts and formulas.

INTRODUCTION

THE efficient use of a search radar in which one or more targets appear on the screen intermittently usually demands a device for tracking the targets automatically. Such a device, called a "track-while-scan system," must make an estimate of each target's instantaneous position from the sampled-data information provided by the radar. In a more general sense, the term "track-while-scan system" may denote any system which estimates the "present" value of a signal from the "past" sampled values of the signal, the sampling taking place at regular intervals.

In this paper, a common type of track-while-scan system is analyzed, and graphs and formulas for optimizing the system's dynamic parameters (or "smoothing constants") are derived.

In the particular system to be analyzed, the target motion is restricted to a plane surface, and the estimated course is a sequence of straight-line constant-velocity paths, successive corrections in the computed position and velocity of the target being made proportional to the deviations of the computed from the measured positions at sampling instants.[1] It is assumed that the data fed to the track-while-scan system (the "input") is in cartesian coordinates; the information is ordinarily obtained from a coordinate converter operating on the raw polar-coordinate data. An illustrative input wave and the resulting form of the output are shown in Figure 1.

The constants of proportionality, α and β, used in correcting the position and velocity, respectively, of the estimated target course completely characterize the performance of this particular track-while-scan system. These constants are the dynamic parameters or so-called "smoothing constants" of the system. The purpose of this paper is to develop graphical and analytical tools for optimizing these smoothing constants with respect to the following four aspects of performance: stability, transient response, noise, and maneuver error. Much of the analysis makes use of "z-transforms," a tool especially suited to linear sampled-data systems. One interesting result is a formula for the mean-square response of the system to white noise.

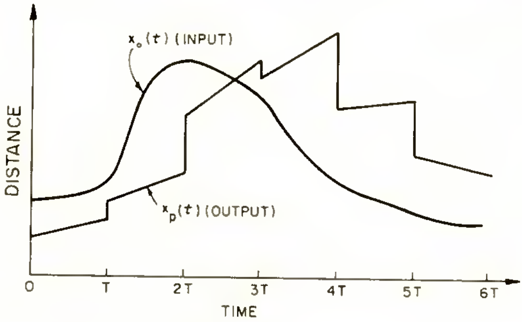


Fig. 1—A track-while-scan input and the resulting output. For illustrative purposes the input here is varying much more rapidly within each sampling period than would occur in practice.

A numerical example serves to illustrate how the charts and formulas may be used.

Preliminary Analysis

In the system under consideration, it is assumed that the target path lies in a plane, or that one is interested only in the projection of the target onto some flat surface such as the horizontal plane. It is

[1] A less general analysis of such a system has been described by E. A. Mechler, J. B. Russell, and M. G. Preston in "The Basis for the Optimum Aided-Tracking Time Constant," *Jour. Frank. Inst.*, Vol. 248, p. 327, October, 1949.

also assumed that the input data is received in the form of cartesian coordinates $x_o(nT)$ and $y_o(nT)$, or is converted to this form by some appropriate coordinate converter. The track-while-scan system is assumed to operate on each coordinate separately, and to combine the outputs vectorially.

Successive corrections to the position and velocity of a stored target coordinate $x_p(t)$ (henceforth referred to as the *output*) are made proportional to the differences between the observed coordinate $x_o(t)$ (henceforth referred to as the *input*) and the output $x_p(t)$ at the sampling instants. As time progresses from the interval $(n-1)T < t \leqslant nT$ to the next interval, $nT < t \leqslant (n+1)T$, the x_p-axis intercept of $x_p(t)$ is raised by the amount $\alpha[x_o(nT) - x_p(nT)]$, and the slope of $x_p(t)$ is increased by the amount $\beta[x_o(nT) - x_p(nT)]$. α and β are the "position smoothing constant" and "velocity smoothing constant," respectively, of the system.

If neither α nor β is zero, a constant-velocity target will eventually be followed with zero error, assuming there is no noise. However, some noise in the input is unavoidable, and the target will not always follow a constant-velocity course. Obviously, the system varies in its ability to smooth noise and reduce maneuver error for different values of α and β. Techniques are described for determining values of α and β which will minimize noise response and maneuver error while maintaining acceptable transient response and stability.

The track-while-scan system described above is equivalent to a sampled-data feedback system in which: (a) the error, defined by $e_x(t) \equiv x_o(t) - x_p(t)$, is converted to a train of impulses,

$$e_x^*(t) \equiv \sum_n e_x(nT)\,\delta(t - nT), \tag{1}$$

where $\delta(t)$ is the Dirac delta function; and (b) this train is applied to a single- and double-integrator combination whose impulsive response, $g(t)$, is the sum of a step function of height α and a ramp of slope β/T. The block diagram is shown in Figure 2. The single- and double-integrator combination is labeled by its Laplace transform, $\mathcal{L}g(t) = \alpha/s + \beta/Ts^2 \equiv G(s)$, and the transformation of the continuous error $e_x(t)$ into the impulse train $e_x^*(t)$, defined in Equation (1), is indicated by the sampling switch.

Because $x_p(t)$ ordinarily contains discontinuities at the sampling instants, it is important to remember that the sampling switch samples $e_x(t)$ at instants just *preceding* the discontinuities in $x_p(t)$. In other words the amount of correction which takes place in each new sampling period is proportional to the error at the end of the previous sampling period, and not to the error at the beginning of the new period. This is because in any physical implementation $g(t)$ will not have a true discontinuity at the origin, but will rise with a finite, positive slope. Account is taken of this fact in the derivation given below of the system transfer function (specifically, at Equation (4)).

With the aid of the block diagram, the input and output signals can be related by Laplace transforms, and the difference-equation technique often used in the analysis of sampled-data systems can thereby be avoided.

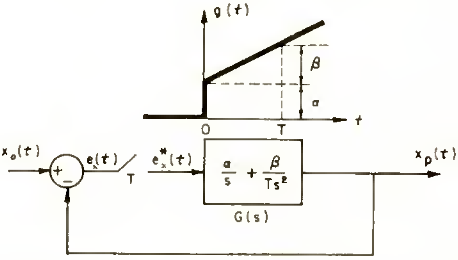


Fig. 2—The sampled-data feedback system equivalent to the particular track-while-scan system under analysis.

Using results available in the literature,[2] one can relate the output and input immediately:

$$X_p(s) = \frac{X_o^*(z)\ G(s)}{1 + G^*(z)}, \tag{2}$$

where $X_p(s)$ is the Laplace transform of $x_p(t)$, $X_o^*(z)$ is the Laplace transform of $\sum_n x_o(nT)\,\delta(t - nT)$, with the variable z replacing ϵ^{sT}. Quantities such as $G(s)$ and $G^*(z)$ in Equation (2), and others appearing elsewhere, are defined similarly.

Now, since

$$G(s) = \frac{\alpha}{s} + \frac{\beta}{Ts^2}, \tag{3}$$

it follows that

$$G^*(z) = \sum_{n=0}^{\infty} [(\alpha + \beta n)\ z^{-n}] - \alpha. \tag{4}$$

[2] J. R. Ragazzini and L. A. Zadeh, "Analysis of Sampled-Data Systems," *Trans. A.I.E.E.*, Vol. 71, p. 225, November, 1952. This paper includes derivations of the basic Laplace and "z-transform" techniques for the analysis of sampled-data systems.

The subtraction of α is required here because, as explained above, sampling at the discontinuity of $g(t)$ takes place at an instant just *preceding* the discontinuity.

It can be shown that Equation (4) sums to the following closed form:

$$G^*(z) = \frac{(\alpha + \beta)\, z^{-1} - az^{-2}}{(1 - z^{-1})^2}. \tag{5}$$

Putting Equations (5) and (3) into (2) yields

$$X_p(s) = X_o{}^*(z) \frac{(1 - z^{-1})^2 \left(\dfrac{\alpha Ts + \beta}{Ts^2} \right)}{1 - (2 - \alpha - \beta) z^{-1} + (1 - \alpha) z^{-2}}. \tag{6}$$

This is the desired expression—it completely describes the output $x_p(t)$ in terms of the sampled input data $x_o(nT)$.* This expression will be the starting point in the remaining analysis of all four properties of interest: stability, transient response, noise, and maneuver error.

STABILITY

From Equation (2) it follows that a necessary and sufficient condition for stability is that the zeros of $1 + G^*(\epsilon^{sT})$ lie within the left half of the s plane, or, equivalently, that the zeros of $1 + G^*(z)$ lie inside the unit circle of the z plane. From Equations (5) or (6), it is seen that the zeros of $1 + G^*(z)$ are the roots of the following equation:

$$z^2 - (2 - \alpha - \beta) z + (1 - \alpha) = 0. \tag{7}$$

By transforming the interior of the unit circle of the z plane into the left half of a new y plane via the transformation $y = (z - 1)/(z + 1)$, one can apply the Routh–Hurwitz conditions for stability to the coefficients of Equation (7). The resulting requirements for stability are $\alpha > 0$, $\beta > 0$, and $(2\alpha + \beta) < 4$. An additional stable condition exists when $\beta = 0$, since in that case the zero of the denominator of Equation (6) at $z = 1$ is cancelled by one of the zeros of the numerator.

* When $\alpha = \beta = 1$, Equation (6) reduces to a transfer-function description of the "first-order hold" discussed in Reference (2). Hence the track-while-scan system of this paper may be regarded as a generalized first-order hold.

The resulting necessary and sufficient conditions for the stability of the track-while-scan system are $\alpha > 0$, $\beta \geqslant 0$, and $(2\alpha + \beta) > 4$. These inequalities determine a "stability triangle" in the α–β plane, for which all internal points and all points on the base $(\beta = 0)$ in the interval $0 < \alpha < 2$ correspond to a stable system. This triangle is shown in Figure 3.

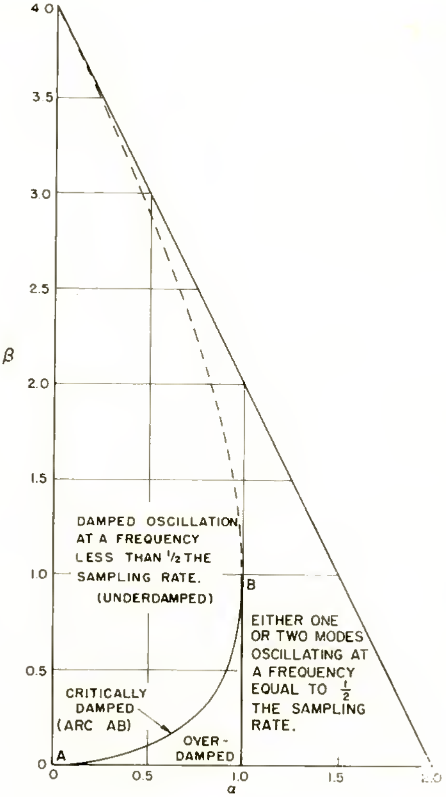


Fig. 3—The stability triangle.

The conditions for underdamped, critically damped, and overdamped transient response are found by inspecting the sign of the discriminant of Equation (7) if the roots are complex, or the signs of the roots themselves if the roots are real.* The resulting conditions are

* If the roots are real, a positive root corresponds to an overdamped or critically damped natural mode, while a negative root corresponds to a damped natural mode oscillating at one half the sampling frequency.

$(\alpha+\beta)^2 < 4\beta$	$<==>$	underdamped,
$\beta \leqslant 1,\ (\alpha+\beta)^2 = 4\beta$	$<==>$	critically damped,
$\beta < 1,\ \alpha \leqslant 1,\ (\alpha+\beta)^2 > 4\beta$	$<==>$	overdamped.
All other values of (α, β) inside the stability triangle and on the base $(\beta = 0)$ in the interval $0 < \alpha < 2$	$<==>$	The transient response contains at least one damped oscillatory natural mode with a rate of oscillation equal to one half the sampling frequency.

The corresponding regions in the stability triangle are indicated in Figure 3.

It should be noted that the condition $(\alpha+\beta)^2 = 4\beta$ alone does not describe critical damping. When, under this condition, $\beta > 1$, the transient will have an oscillatory component, although the envelope of the oscillatory component has the appearance of a critically damped wave. For the region satisfying both $(\alpha+\beta)^2 > 4\beta$, and α or $\beta > 1$, the transient has a similar characteristic, but with an overdamped envelope.

The relations between the stability triangle and the roots of the characteristic equation are described in more detail in Appendix I.

Transient Response

The transient response of the system becomes significant when the input is switched from one time-shared target to another or when the target makes a sharp maneuver. In this section, some curves describing the transient response as a function of α and β are derived.

It is convenient to evaluate the transient performance of this system by considering the position error in response to a step of velocity. The reason for this is that a system which is nominally critically damped or overdamped in accordance with the labeling of Figure 3 will still yield a position response having an overshoot when the input is either a single pulse or a step of position. No position overshoot will appear, however, when the input is a step of velocity or acceleration.

Using Equation (5) and some elementary manipulations of z-transforms,[2] one relates the z-transforms of $e_x(t)$ and $x_o(t)$:

$$E_x^*(z) = \frac{X_o^*(z)}{1+G^*(z)} = \frac{(1-z^{-1})^2 X_o^*(z)}{1-(2-\alpha-\beta)z^{-1}+(1-\alpha)z^{-2}}. \qquad (8)$$

If the input is a unit step of velocity, then $x_o(t) \equiv 0$ or t when $t < 0$ or $t \geqslant 0$, respectively. Hence $X_o^*(z) = Tz^{-1}/(1-z^{-1})^2$, and

$$E_x^*(z) = \frac{Tz^{-1}}{1-(2-\alpha-\beta)z^{-1}+(1-\alpha)z^{-2}}. \tag{9}$$

For the underdamped region of the stability triangle, the above expression is the z-transform of a time function of the form $\epsilon^{-ut}\sin vt$, while for the critically damped and overdamped regions the corresponding time functions have no oscillations.

Usually, a desirable choice of α and β will yield a slightly underdamped transient response to a step of velocity, i.e., an error response whose natural mode is of the form $\epsilon^{-ut}\sin vt$. An indication of the violence of oscillation of this function is given by the ratio of a peak value of $|\epsilon^{-ut}\sin vt|$ to the peak value immediately preceding. This ratio, whose definition is illustrated in Figure 4, will be called the *decrement ratio,* and will be denoted by r. This ratio has an additional significance—it is roughly equal to the relative overshoot of $e_x(t)$ in

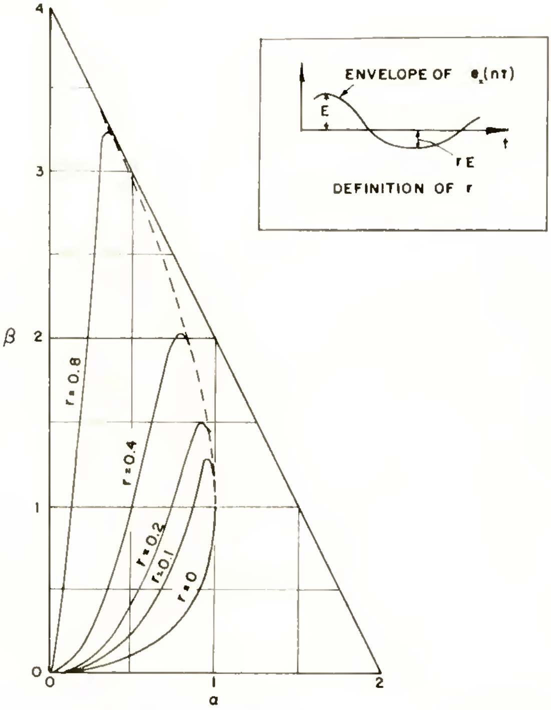


Fig. 4—Contours of constant decrement ratio, r.

response to a step of acceleration, provided that the frequency of oscillation of the natural mode is several times smaller than one-half the sampling frequency.

Corresponding to any given pair (α,β), the value of r can be obtained by relating (α,β) to u/v by means of the denominator of Equation (9). It can be shown that the ratio u/v is, in turn, related to r by the formula $r = \exp(-\pi u/v)$. In this way, several contours of constant r were computed; they are plotted in Figure 4. These contours are useful for both analysis and synthesis, i.e., for either evaluating the transient response, or for determining regions in the α–β plane corresponding to acceptable transient performance.

Noise

The input coordinate data is subject to corruption by random noise, such as (a) range jitter and (b) wander of the target's apparent center of reflection. It is desirable to know how the choice of α and β affects the degree to which the noise is smoothed or exaggerated by the system.

The following figure of demerit was chosen to describe the inefficacy of the system in smoothing the input noise:

$$\rho \equiv \text{noise ratio} \equiv \sqrt{\frac{\overline{x_p^2(t)}}{\overline{x_o^2(t)}}}, \tag{10}$$

assuming $x_o(t)$ is white noise. In words, the noise ratio, ρ, is defined as the ratio of the r-m-s value of the output to the r-m-s value of the input. The bandwidth of the input noise is assumed to be finite, but much greater than all natural frequencies of the system.

We shall now derive an expression for ρ^2 in terms of α and β. From Equation (2) it is seen that $x_p(t)$ may be considered to consist of a sequence of responses of a system whose transfer function is $G(s)/[1+G^*(z)]$ to the impulse train

$$x_o^*(t) \equiv \sum_n x_o(nT)\ \delta\ (t-nT). \tag{11}$$

Since $x_o(t)$ is effectively white noise, the successive samples $x_o(nT)$ are completely uncorrelated. Hence the successive responses are uncorrelated, and the mean square of their sum equals the mean of the integrated squares of the individual responses to each impulse. The problem thus reduces to finding the ensemble average of the integrated square of

$$x_o(nT)\ \mathcal{L}^{-1}\left[\frac{G(s)}{1+G^*(z)}\right], \tag{12}$$

and dividing the result by the ensemble average of the integrated square of the section of $x_o(t)$ in the interval $nT < t < (n+1)T$, where n is chosen arbitrarily. Since the interval is T time units long, the integral of $x_o^2(t)$ over the interval is *on the average* equal to T times the ensemble average of $x_o^2(nT)$.

Hence

$$\rho^2 = \frac{1}{T}\int_{-\infty}^{\infty}\left[\mathcal{L}^{-1}\frac{G(s)}{1+G^*(z)}\right]^2 dt. \tag{13}$$

Now let

$$C(s) \equiv \frac{G(s)}{1+G^*(z)}$$

$$=\left[(1-z^{-1})^2\frac{\alpha Ts+\beta}{Ts^2}\right]\left[\frac{1}{1-(2-\alpha-\beta)\,z^{-1}+(1-\alpha)\,z^{-2}}\right]. \tag{14}$$

Let the expressions in the first and second pairs of brackets be denoted by $A(s)$ and $B^*(z)$, respectively. It is convenient to express the time function $\mathcal{L}^{-1}C(s)$ in terms of $\mathcal{L}^{-1}A(s)$ and the coefficients appearing in the series expansion of $B^*(z)$. Thus,

$$c(t) = \mathcal{L}^{-1}\sum_n b_n\,A(s)\,z^{-n} = \sum_n b_n\,\mathrm{a}(t-nT), \tag{15}$$

where

$$c(t) \equiv \mathcal{L}^{-1}\,C(s),$$

$$a(t) \equiv \mathcal{L}^{-1}\,A(s),$$

and

$$\sum_n b_n\,z^{-n} \equiv B^*(z).$$

Fortunately, $a(t)$ is bounded in the time domain, thereby affording a simplification in the evaluation of $\int_0^\infty c^2(t)\,dt$. The function $a(t)$ is the sum of the doublet and the triangular waveforms shown in Figure 5. Stated in analytical terms,

$$a(t) = \begin{cases} \alpha + \dfrac{\beta}{T} t, & \text{for } 0 < t \leqslant T, \\ -\alpha + 2\beta - \dfrac{\beta}{T} t, & \text{for } T < t \leqslant 2T. \end{cases} \tag{16}$$

Let $c_N(t)$ for $N = 0, 1, 2 \cdots$ be defined by

$$c_N(t) \equiv c(t) \{1[t - NT] - 1[t - (N+1)T]\}, \tag{17}$$

where $1(t) \equiv$ unit step function $\equiv \begin{cases} 1 & \text{for } t \geqslant 0, \\ 0 & \text{for } t < 0. \end{cases}$

Then, by Equations (15), (16), and (17),

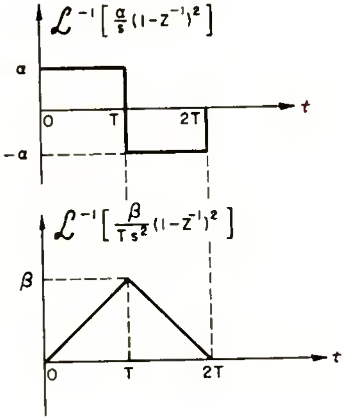


Fig. 5—The two waveforms summing to $a(t)$.

$$c_N(t) = b_{N-1}\, a(t_N + T) + b_N\, a(t_N), \tag{18}$$

where $t_N \equiv t - NT$. Applying Equation (16) to Equation (18) for the interval $0 < t_N \leqslant T$, we obtain

$$c_N(t_N) = b_{N-1}\left(-\alpha + \beta - \frac{\beta}{T} t_N\right) + b_N\left(\alpha + \frac{\beta}{T} t_N\right), \tag{19}$$
$$0 < t_N \leqslant T.$$

Then,

Maneuver Error

Another important factor governing the choice of the smoothing constants is the error incurred when the target follows a curved course. This "maneuver error" is an additional indication of the poorness of performance of the system.

For convenience of analysis, it is assumed that the target is following a circular arc in the x–y plane at a constant angular velocity ω. The maneuver error, $e(t)$, will be resolved into $e_x(t)$ and $e_y(t)$.

The results derived in this section describe the steady-state values of the maneuver error. For this reason, the symbols $e(t)$, $e_x(t)$ and $e_y(t)$ in this section denote steady-state values.

From the geometry of $e_x(t)$, $e_y(t)$, and $e(t)$, it follows that

$$e(t) = \sqrt{e_x{}^2(t) + e_y{}^2(t)}, \tag{27}$$

and that the x and y coordinates of the true target position, namely $x_o(t)$ and $y_o(t)$, will both be sinusoidal time functions, separated in phase by 90 degrees, since the target is assumed to follow a circular course at constant speed. Thus, we write

$$x_o(t) = R\epsilon^{j\omega t},$$

and

$$y_o(t) = jR\epsilon^{j\omega t}, \tag{28}$$

where $R \equiv$ radius of the path. In the above equations, $x_o(t)$ has been arbitrarily set at zero, and $y_o(t)$ has been arbitrarily assumed to lead rather than lag $x_o(t)$. (These assumptions do not incur any loss of generality.) Setting $z = \epsilon^{j\omega T}$ in Equation (8), one obtains an expression for the ratio of $e_x(nT)$ to $x_o(nT)$ in the steady state, i.e., one obtains the ratio $e_x(nT)/(R\epsilon^{j\omega nT})$ for positive integral values of n, the values of n being taken large enough to make the transient component of $e_x(nT)$ negligible. Thus,

$$e_x(nT) = \left[\frac{R}{1 + G^*(\epsilon^{j\omega T})}\right]\epsilon^{j\omega nT}. \tag{29}$$

Since $y_o(t)$ leads $x_o(t)$ by 90 degrees it follows that $e_y(nT)$ must lead $e_x(nT)$ by 90 degrees. Furthermore, the vectors $\mathbf{e}_x(t)$ and $\mathbf{e}_y(t)$ are perpendicular in space. Hence their vector sum, $e(t)$, is circularly polarized: it is constant in magnitude at all sampling instants, and it rotates in equal angular increments, $\Delta\theta$, from sample to sample, the rate of rotation, $\Delta\theta/T$, being equal to the angular target velocity, ω.

The magnitude of $\boldsymbol{e}(nT)$ and its angular displacement from the radius vector joining the center of the circular course to the target position is given by the magnitude and phase of the bracketed quantity in Equation (29). The magnitude of the total error, $e(nT)$, is

$$e = \left| \frac{R}{1 + G^{*}(\epsilon^{j\omega T})} \right|, \tag{30}$$

where the argument of $e(nT)$ has been omitted, because $e(nT)$ is constant for all integral n. Setting $z = \epsilon^{j\omega T}$ in Equation (5) and substituting the resulting expression for $G^{*}(\epsilon^{j\omega T})$ in Equation (30) gives an exact expression for e:

$$e = \frac{\dfrac{4a}{\omega^2} \sin^2 \left(\dfrac{\omega T}{2} \right)}{\sqrt{[(2-\alpha) \cos \omega T - (2-\alpha-\beta)]^2 + \alpha^2 \sin^2 \omega T}}, \tag{31}$$

where R has been replaced by a/ω^2, a being the centripetal acceleration of the target.

Ordinarily the sampling rate is such that $\omega T \ll \pi$, in which case $\sin \omega T \approx \omega T$ and $\cos \omega t \approx 1 - \omega^2 T^2/2$. Then Equation (31) reduces to

$$e \approx \frac{aT^2}{\beta \sqrt{1 + \left[\dfrac{\alpha^2 - 2\beta + \alpha\beta}{\beta^2} \right] \omega^2 T^2}}, \tag{32}$$

when $\omega T \ll \pi$.

If, in addition, the second term under the radical is much less than unity, then the expression becomes still simpler:

$$e \approx \frac{aT^2}{\beta}, \tag{33}$$

when $\qquad \omega T \ll \pi, \quad \dfrac{\alpha^2 - 2\beta + \alpha\beta}{\beta^2} \omega^2 T^2 \ll 1.$

Equations (31), (32), and (33) are the desired results.

It is noted that the error in all three equations is exactly proportional to the target's acceleration. Furthermore, if the sampling rate is high enough for Equation (32) to hold, then the magnitude of the error is almost independent of the target velocity and is proportional to the square of the sampling period. The phase lag of the error becomes negligible when the second term under the radical of Equation (32) is sufficiently small. These properties are in conformity with physical considerations.

From Equation (33) it is seen that if the number of samples per revolution of the target in its circular course is large, and if the second term under the radical of Equation (32) is negligible, then the steady-state maneuver error is virtually independent of α. The mechanism underlying this property is demonstrated by the situation in Figure 7. In the steady state the successive values of the error are constant in magnitude, and the straight-line portions of the computed course must therefore take on equal angular displacements in progressing from one sampling period to the next. Each of these angular displacements is equal to $\Delta\theta$, the angular displacement per sampling

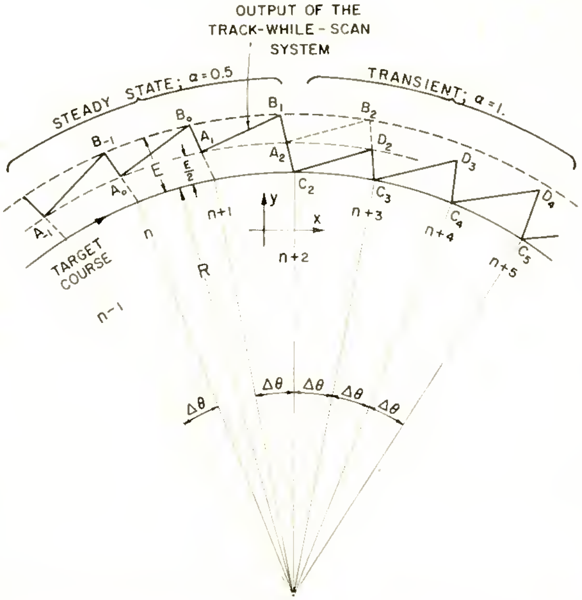


Fig. 7—Diagram illustrating mechanism by which the steady-state error remains essentially unaltered under a change in the position constant, α.

period of the radius vector of the target from the center of curvature of the course.

The first three periods of the figure illustrate a steady-state condition for the case $\alpha = 0.5$; the angle between $A_{-1}B_{-1}$ and A_0B_0 equals the angle between A_0B_0 and A_1B_1. At the sampling instant numbered $n + 2$, the value of α is suddenly raised to 1. This brings the error in the computed position down to zero at this instant. The angular orientation of C_2D_2 is equal to that which A_2B_2 would have had if α had not been changed, because the error E at the end of the preceding sampling period equals the steady-state value attained when α was 0.5. Therefore, D_2C_3 is nearly equal to $E/2$. This means that the velocity correction (practically all of which for this figure is y-oriented) for the interval beginning with the $(n + 3)^{\text{th}}$ instant is only half that required to maintain a steady-state condition. Therefore the angle of C_3D_3 with respect to the target course is about 50 per cent larger than that of C_2D_2. The error D_3C_4 is thus roughly equal to $3E/4$. Hence the velocity correction for the following interval is again too small for the steady-state condition, and D_4C_5 is roughly equal to $7E/8$. In this way the succeeding values of error at the end of each interval approach E monotonically until the steady-state value of E is attained. (More specifically, the error is roughly $(1—2^{-k})E$ when k sampling periods following the change in α have elapsed.) Thus, under the assumptions of high sampling rate and negligible value of the second term under the radical of Equation (32), which are implied by the geometry in Figure 7, it is seen that a change in α has practically no effect on the steady-state value of e. These assumptions often are realized in practice, especially when the targets are relatively slow.

It is instructive to deduce Equation (33) directly from the simplified situation in Figure 7. In the steady state, the difference in angle between a straight-line path such as A_1B_1 and the path A_0B_0 of the preceding sampling period is equal to the angular displacement $\Delta\theta$ of the target's radius vector during a full sampling period. This angular displacement is also nearly equal to the correction in slope, namely, βe divided by the arc length per period, $R\Delta\theta$. Thus:

$$\Delta\theta = \frac{\beta e}{R\Delta\theta},$$

and
$$e = \frac{R\Delta\theta^2}{\beta} = \frac{R\omega^2 T^2}{\beta}. \tag{34}$$

Since $R\omega^2 = a$, this result is equivalent to Equation (33).

Procedure for Optimizing the Smoothing Constants

The selected values of α and β should as nearly as possible result in a system which minimizes both noise and maneuver error, and provides dynamic stability and good transient response. Usually a three-way compromise will need to be made among noise, maneuver error, and transient response.

In the following numerical example, a procedure for choosing α and β from the derived results will be illustrated. For special requirements, variations of the procedure should suggest themselves.

Numerical Example

Suppose the target moves in a circular path in the x-y plane with a velocity of 400 miles per hour and a centripetal acceleration of 1g. Suppose in addition that the random noise in the observed value of target position has an r-m-s value of 50 feet, that successive samples of the noise are completely uncorrelated, and that the sampling period is 3 seconds.

Problem: Find values of α and β which will provide dynamic stability and good transient response, and will yield optimally small noise and maneuver error in the computed position of the target.

Solution: Let the optimization criterion be in the form of a figure of demerit, F, defined as the sum of the square of the maneuver error and the mean square of the output noise. Thus,

$$F \equiv e^2 + \rho^2 n_i^2, \tag{35}$$

where n_i is the r-m-s value of the input noise. The constants α and β will now be chosen so as to yield a minimum value of F inside the stability triangle.

First we compute the value of ωT to determine whether Equations (32) or (33) are applicable:

$$\omega T = \frac{aT}{v} = \frac{(32.2)\,(3600)\,(3)}{(400)\,(5280)} = 0.1647 \text{ radian.} \tag{36}$$

Since ωT satisfies the condition $\omega T \ll \pi$, Equation (32) should be a satisfactory approximation. For convenience it shall be assumed that the values of α and β to be chosen will be such that Equation (33) will also be acceptable. We then are able to write e^2 in terms of β:

$$e^2 = \left(\frac{aT^2}{\beta}\right)^2 = \frac{37{,}300}{\beta^2}\text{ ft}^2. \tag{37}$$

In the α–β plane the contours of constant e^2 coincide with contours of constant β; thus, in Figure 6 they are horizontal straight lines. It is therefore evident that the minimum value of F must lie at a point of tangency between one of the constant-e^2 lines and one of the constant-ρ^2 contours. To determine this point, the values of F for several points of tangency are computed, and the point of minimum F determined by interpolation.

A typical computation of F runs as follows: Consider the contour of $\rho^2 = 3$. Its point of tangency with a constant-β line occurs at $\beta = 1.63$. Hence

$$e^2 = \frac{37{,}300}{(1.63)^2} = 14{,}030 \text{ ft}^2,$$

$$\rho^2 n_i^2 = 3(50)^2 = 7{,}500 \text{ ft}^2,$$

$$F = e^2 + \rho^2 n_i^2 = 21{,}530 \text{ ft}^2.$$

By repeating a few such computations, a minimum value of $F = 20{,}100$ ft^2 is found to occur roughly at $(\alpha,\beta) = (0.5,1.93)$, corresponding to a noise ratio of $\rho^2 = 4$. Since both inequalities of Equation (33) are satisfied here, the approximation assumed in Equation (37) is justified.

Unfortunately, the fact that the smoothing constants just determined will give optimum performance for the steady-state tracking of a circular course does not guarantee good *transient* performance. The latter is important, because the target often will not follow a circular course for more than a 90-degree arc. It is still more important when the system switches from one target to another on a time-sharing basis. Satisfactory transient performance can be obtained by confining one's choice of (α,β) to the region where position error in the response to a step of velocity will be critically damped or will be underdamped with a decrement ratio no greater than, say, 10 per cent. This is indicated by the shaded region in Figure 8.

When the figure of demerit, F, of the numerical example is minimized within this shaded region, the optimum smoothing constants occur when β is a maximum, namely, $(\alpha,\beta) = (0.95, 1.28)$, corresponding to which $\rho^2 = 3.3$ and $F = 31{,}050$ ft^2. The reason is that the noise component of F is much smaller than the maneuver-error component everywhere within the shaded region.

Thus, in situations where the maneuver is the major source of error inside the shaded region of Figure 8, the pair $(\alpha,\beta) = (0.95, 1.28)$ will be an optimum. Such situations often arise in practice.

If desired, the permissible values of (α,β) may be confined to the contour of critical damping. In that case, if the noise component in

F is smaller than the maneuver error along this contour—as it is in the numerical example—the resulting optimum parameters will be $(\alpha,\beta) = (1,1)$. This corresponds to $\rho^2 = 2.67$ and, for the numerical example, $F = 44{,}000$ ft^2.

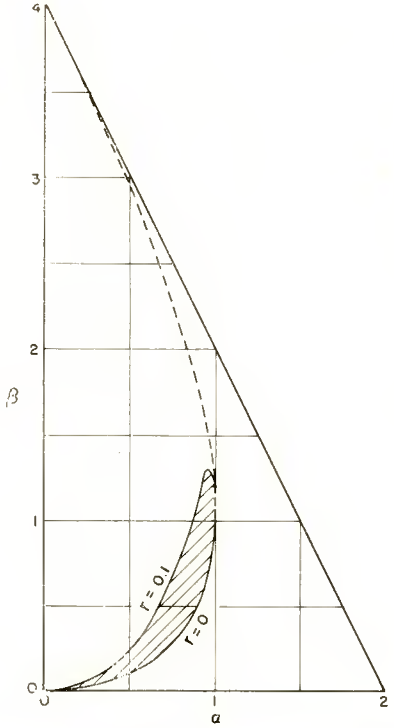


Fig. 8—The shaded region indicates the values of (α, β) for which the decrement ratio of the position error in response to a step of velocity lies between 0 and 0.1. In the illustrative example, this region is assumed to correspond to a satisfactory transient response.

Nonmaneuvering Targets

Although the preceding example involved a maneuvering target, the optimization procedure illustrated in it could be applied to a nonmaneuvering target, i.e., a target flying a straight-line course at constant speed. It is intuitively clear that the optimum values of α and β for such a target should be significantly different than for a maneuvering target, since greater smoothing will usually be desirable.

In choosing the smoothing constants for this case, it is important to take account of a small unintentional maneuver error or "drift" in the target motion. When this is done, the optimum value of α will be relatively small, usually below 0.5.

Conclusions

With the aid of the graphs in Figures 4 and 6 and the formulas given by Equations (25), (31), (32), and (33), optimizing the smoothing constants becomes a reasonably straightforward matter.

The results of this paper are restricted to two-dimensional tracking with no more than two independent dynamic parameters, although an extension to the three-dimensional two-parameter case may not be too difficult, especially if the target can be assumed to move two-dimensionally over several sampling periods.

Acknowledgment

The author is grateful for many valuable discussions with J. R. Ford, E. C. Hutter, and V. D. Landon of RCA Defense Electronic Products, Princeton.

Appendix I

The regions of the stability triangle labeled in Figure 9 are described below in terms of the roots in the z-plane of the characteristic equation (Equation (7)) and the frequency of oscillation of the natural modes.

These results were obtained from the following general expressions for the roots of Equation (7):

$$z \equiv \epsilon^{sT} = \frac{1}{2}\left[2\text{-}\alpha\text{-}\beta \pm \sqrt{(2\text{-}\alpha\text{-}\beta)^2 - 4(1-\alpha)}\right] \qquad (38)$$

$$\equiv \frac{1}{2}\left[2\text{-}\alpha\text{-}\beta \pm \sqrt{(\alpha+\beta)^2 - 4\beta}\right]. \qquad (39)$$

Region I: Complex roots with positive real parts, corresponding to an underdamped mode oscillating at a frequency below $1/(4T)$.

Region II: Complex roots with negative real parts, corresponding to an underdamped mode whose frequency of oscillation, f, lies in the interval $1/(4T) < f < 1/(2T)$.

Region III: Two negative real roots, corresponding to two damped modes both oscillating at the frequency $f = 1/(2T)$.

Region IV: Two real roots of opposite sign, the negative root corresponding to a damped mode oscillating at the frequency $f = 1/(2T)$, and the positive root corresponding to an overdamped mode.

Region V: Two unequal, positive real roots, corresponding to two overdamped modes.

Arc AB: Two equal, positive real roots, corresponding to critical damping.

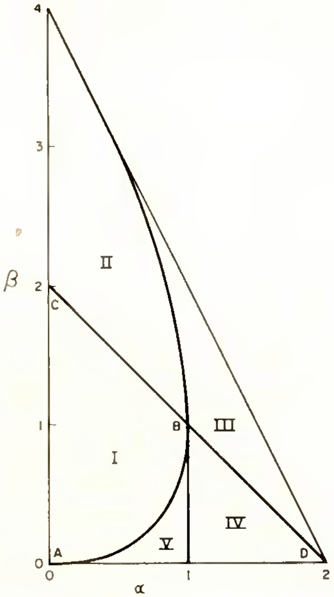


Fig. 9—The stability triangle broken down into regions corresponding to the forms of the roots of the characteristic equation.

Line segment BC: Two conjugate, pure imaginary roots, corresponding to an underdamped mode oscillating at $f = 1/(4T)$.

Appendix II

Proof of Equation (22), an important relation used in the derivation of the formula for the noise ratio, is given below.

It was stated that if $B^*(z)$ is a rational function of z, and if all of its poles lie inside the unit circle, then

$$\sum_{n=0} b_{N-m} b_N = \begin{array}{l}\text{Residue}\\ \text{inside}\\ \text{unit}\\ \text{circle}\end{array} \left[B^*(z) B^*(z^{-1}) z^{m-1} \right]. \tag{22}$$

Proof: $B^*(z) B^*(z^{-1}) = \sum_{i=0}^{\infty} b_i z^{-i} \sum_{j=0}^{\infty} b_j z^j$

$$= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} b_i b_j z^{j-i} \tag{40}$$

Let $m = i - j$. Then

$$B^*(z) B^*(z^{-1}) = \sum_{i=0}^{\infty} \sum_{m=-\infty}^{\infty} b_{i-m} b_i z^{-m}$$

$$= \sum_{m=-\infty}^{\infty} r_m z^{-m}, \tag{41}$$

where $r_m \equiv \sum_{i=0}^{\infty} b_{i-m} b_i$.

By Laurent's theorem,

$$r_m = \frac{1}{2\pi j} \oint B^*(z) B^*(z^{-1}) z^{m-1} \, dz, \tag{42}$$

where the path of integration is a closed curve separating the poles of $B^*(z)$ and of $B^*(z^{-1})$. From this integral, Equation (22) follows immediately.

CONCENTRIC-SHEAR-MODE 455-KILOCYCLE ELECTROMECHANICAL FILTER

BY

R. W. GEORGE

RCA Laboratories,
Princeton, N. J.

Summary—*This paper describes an experimental 455-kilocycle electromechanical filter of simplified design. The filter consists of four magnetostrictive ferrite disk resonators operating in the concentric shear mode. The entire filter is only 1 inch long and ⅝ inch in diameter. Broad-band electrical terminations and low insertion loss make this filter ideal for use in transistorized equipment.*

INTRODUCTION

ELECTROMECHANICAL filters are known to be particularly suitable for use in single-sideband communications equipment. These filters meet the precision performance requirements and are smaller, lighter, lower in cost and in some cases more rugged than the equivalent L-C or crystal-type filters. There is a much larger potential use of electromechanical filters in relatively low-cost equipment such as automobile and broadcast receivers. The performance requirements here are much less critical, but the cost is of paramount importance. The work described here was carried out with these factors in mind. The result is a small filter which is well-adapted to use with transistors. While it can be made much more cheaply than currently available electromechanical filters, further cost reduction is necessary before the filter can be used in low-cost broadcast receivers.

The filter is an assembly of magnetostrictive ferrite disks resonant in the concentric shear mode[1] and quarter-wave torsional couplers. A complete filter assembly removed from the case is shown in Figure 1 and the parts are shown in more detail in Figure 2. Separate couplers are cemented between each disk. Magnetic bias is obtained from residual circular magnetism obtained by passing a current through a hole in the center of the disk. Excitation is by application of the r-f field radially to the side of the disk. This makes it possible to measure the resonant frequency of each disk as well as to use the end disks

[1] W. van B. Roberts, "Some Applications of Permanently Magnetized Ferrite Magnetostrictive Resonators," *RCA Review*, Vol. 14, p. 3, March, 1953.

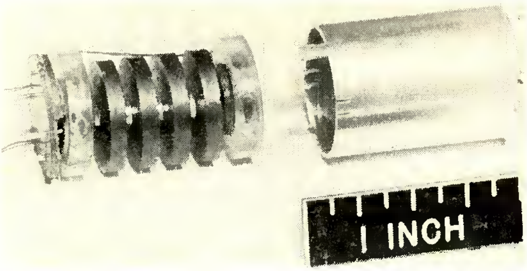


Fig. 1—Filter assembly removed from case.

of the filter for efficient electromechanical transducers. Uniform circular bias and exciting field give negligible excitation of spurious modes of vibration.

The transducers and the mechanical system have very low loss so that the important power losses occur in the coils which are coupled to the transducers. The coils are enclosed in low-loss ferrite cores, Figure 3. These ferrite cores minimize eddy-current loss in the shield can and concentrate the r-f field in the transducer disk to obtain maximum electromechanical coupling to the energized portion of the transducer. Broad-band electrical termination is possible when the electromechanical coupling coefficient is relatively large, and the magnetostrictively energized portion of the end resonator, also a transducer in this case, is relatively small.[2] It will be apparent that electrical damping should be in the external terminal load resistance as much as possible so as to minimize electrical transmission loss.

An axial pivot is fixed to each end of the disk-and-coupler assembly. These pivots fit into cup-shaped bearings in the two coil cores. The cores are in turn supported by ring spacers and the shield can. The pivots should not only space the core from the disk but should have low electrical loss as well. Dural pivots were found to lower the Q of

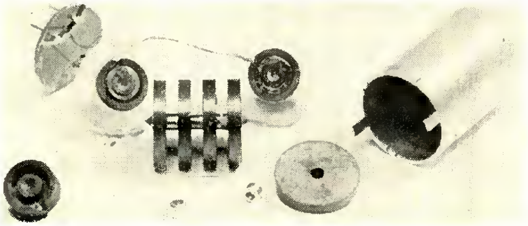
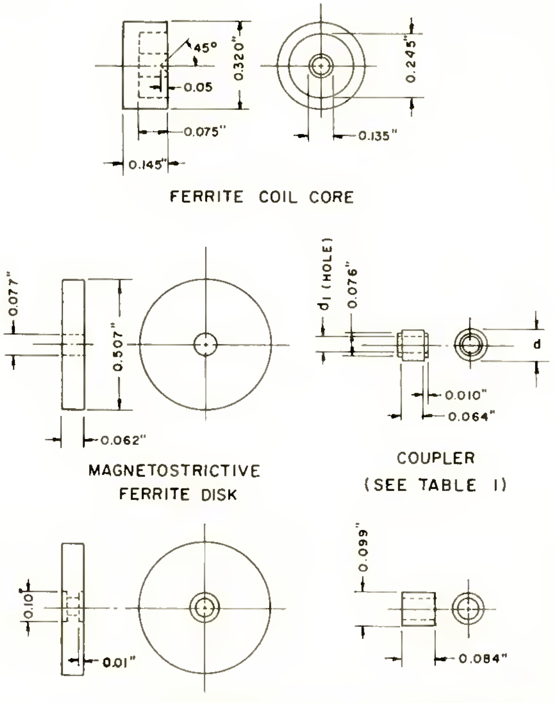


Fig. 2—Filter subassemblies and parts.

[2] R. W. George, "Electromechanical Filter for 100 KC Carrier and Sideband Selection," *Proc. I.R.E.*, Vol. 44, p. 14, January, 1956.

the coil appreciably. Care must be taken to ensure that supporting and spacing parts do not contribute appreciable mechanical damping to the end resonator disks. Difficulty was encountered from this source when pressure was put on rounded lucite pivots; this resulted in a 5 to 10 decibel increase in the insertion loss. Mechanical damping has been satisfactorily reduced by making the pivot bearings hard and smooth with a coating of A-2 cement[†] and using pointed lucite pivots. Spacing between the ferrite core and the disk should generally be as small as practical, on the order of 0.01 to 0.02 inch.



FERRITE COIL CORE

MAGNETOSTRICTIVE FERRITE DISK

COUPLER (SEE TABLE I)

FERRITE DISK AND COUPLER VARIATION

Fig. 3—Mechanical filter details.

Design

It is possible to calculate the filter dimensions to give the required bandwidth on the basis of the kinetic energy of the resonators and the couplers. However, it is more practical to use empirical methods.

An attempt was made to use coupling coefficients of the Campbell

[†] Product of the Armstrong Products Co., Warsaw, Ind.

type.[3] This was tried by using identical couplers with end disks ½ the thickness of the interior disks. Usable filters were obtained; however, the attenuation on the low-frequency side of the passband was sloppy and included unexpected spurious responses. Filters made using identical disks with somewhat larger end couplers eliminated these difficulties. The spurious responses of these filters were generally at least 60 decibels down over the frequency range of 160 to 1,600 kilocycles. One unexplained response about 50 decibels down has been found in some filters around 700 kilocycles.

In the narrower-band filters, the couplers are more delicate and the assembly with cement requires more care. For a bandwidth of somewhat less than 2 kilocycles, it might be better to use thicker resonator disks to increase the required mass of the couplers. The thickness should probably not exceed about 0.075 inch which is approximately a quarter wavelength in the torsional mode.

The input and output damping or load resistance values can be varied over a fairly wide range with various values of coupling-coil inductance and high-Q coils. Low transmission loss is also obtained by having high Q, L-C circuits.

Magnetostrictive Ferrite Resonator Disks

The disk resonators were made of nickel–iron ferrite containing very little cobalt impurity.[4] The resonant frequency of the disk is principally a function of diameter. The hole in the center which is necessary in order to magnetize the disk contributes a minor modification of the resonant frequency. Nonuniformity in the thickness of a disk also results in variations of resonant frequency. Experience with small lots of 25 to 50 disks pressed by hand indicates that the 455-kilocycle disks might be expected to have resonant frequency variations of less than ±500 cycles. In the filters described, it was found that the resonant frequency of the completed filter was some 1,100 to 1,300 cycles lower than the resonant frequency of the disks themselves because of the method of assembly of couplers and disks.

The disks used in the experimental filter had a diameter of about 0.507 to 0.509 inch and a hole diameter of 0.077 inch. Nominal thickness was about 0.062 inch. They were selected or tuned to be matched in frequency within about ±50 cycles for each filter. It is difficult to

[3] W. van B. Roberts and L. L. Burns, "Mechanical Filters for Radio Frequencies," *RCA Review*, Vol. 10, p. 348, September, 1949.

[4] R. L. Harvey, "Ferrites and Their Properties at Radio Frequencies," *Proceedings of the National Electronics Conference*, Vol. 9, p. 287, February, 1954.

lower the resonant frequency of a disk in the concentric shear mode because this requires a removal of material to make the disk thinner in the region of the nodal ring zone which has a diameter of about 0.7 times the disk diameter. If the disks are to be tuned, it is more practical to make them a little low in frequency and then raise the frequency by grinding the outside edge or corners.

Physical faults such as voids and cracks not only reduce the electromechanical coupling coefficient and residual magnetic bias, but also result in spurious resonant frequency responses; the main resonant frequency is usually altered considerably and is accompanied by lower amplitude at resonance.

The temperature–frequency coefficient of these filters is of the order of 40 parts per million per degree centigrade. This coefficient can be decreased by improving the characteristics of the ferrite.

The assembled mechanical portion of the filter was re-magnetized by the standard current to insure against loss of bias due to shock or exposure to strong magnetic fields while handling and assembling the disks. It may also be desirable to age the filter at an elevated temperature to stabilize the residual bias if operating temperatures are expected to be higher than usual.

Data on Experimental Filters

The experimental filters were first tested in a completely closed shield with leads at each end. Stray input–output coupling was very small, being more than 60 decibels down. This coupling is minimized by the use of small coupling coils and cores which have an outside diameter considerably less than the ferrite disk diameter. The stray coupling obtained with the shield can shown in Figure 1, having all leads at the open end of the can, was somewhat more than 40 decibels down. A material improvement results from closing the open end of the can. The ferrite coil cores may be increased in size somewhat to increase the coupling which might be required for very-wide-band filters, or to make more space for more efficient or higher impedance coils.

The circuit shown in Figure 4 was used in testing these filters. Some variation in coil-core spacing to the filter resulted in unequal values of R_1 and R_2 and C_1 and C_2. Tuning capacitance was on the order of 500 micromicrofarads when using 75-turn coils. Tests were made with an output voltage, E_2, of about 0.1 volt; however, the filters operate satisfactorily with output voltages of up to 10 volts or more. Response curves for 4 filters are shown in Figures 5 to 9. The shape factor is on the order of 3 to 3.5.

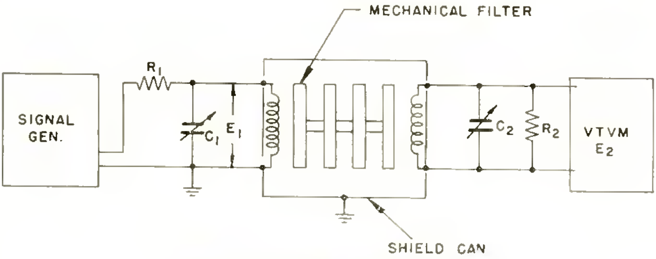


Fig. 4—Filter test circuit.

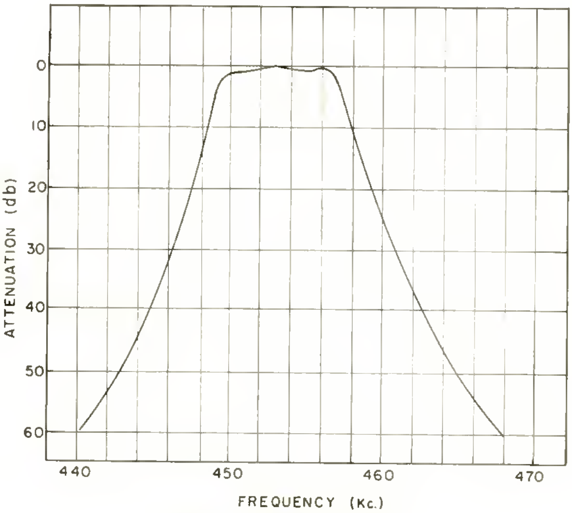


Fig. 5—Filter No. 8.

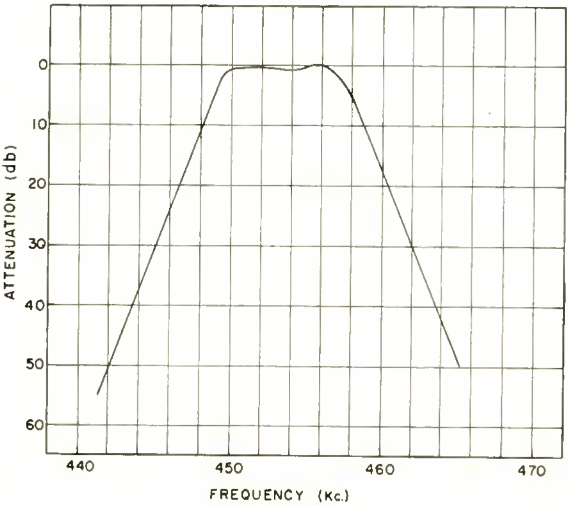


Fig. 6—Filter No. 9.

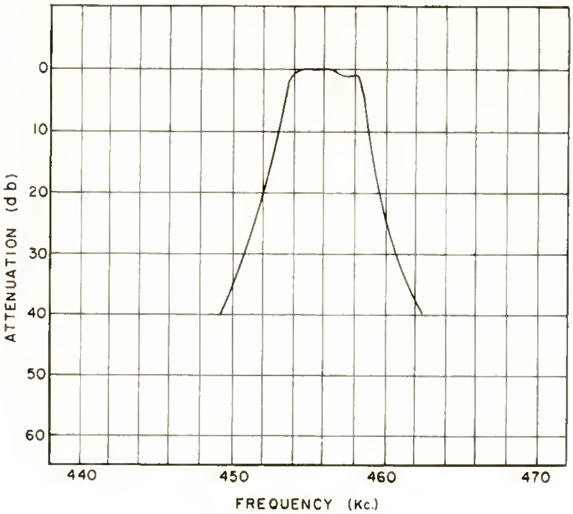


Fig. 7—Filter No. 10.

Electrical insertion loss is estimated to be less than 6 decibels. This was based on the power input to the filter at the terminals where E_1 is measured, and the output power in R_2. E_1 varies considerably with frequency in the pass-band with the result that this method of measurement is not very satisfactory. The coils were layer or random wound with about 75 turns of No. 36 enamel wire. Lower insertion loss could be achieved by use of higher-Q coils.

Mechanically perfect joints between resonators and couplers are

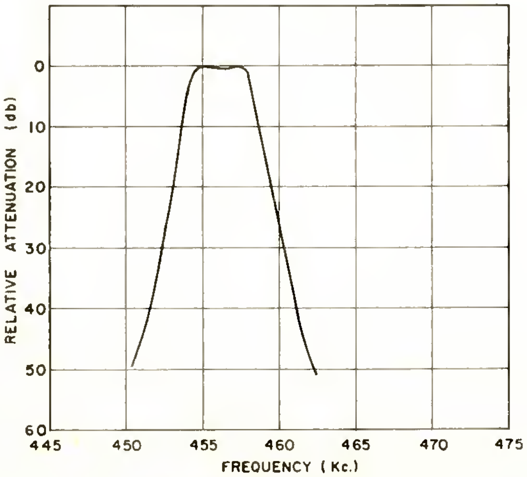


Fig. 8—Filter No. 11.

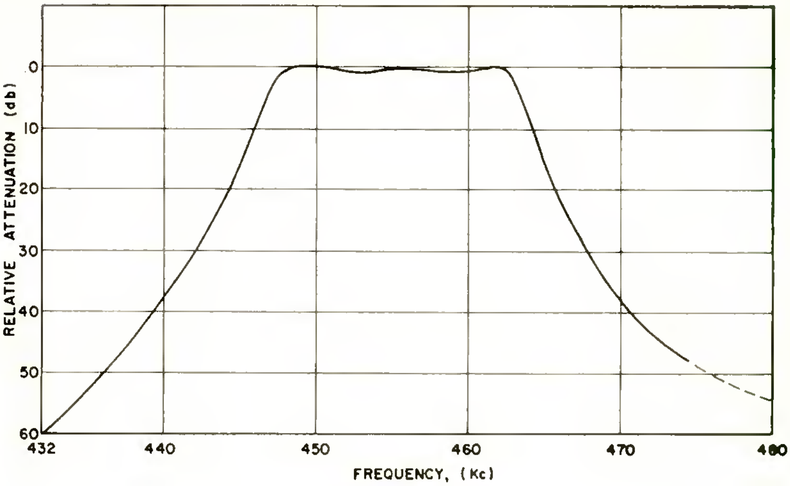


Fig. 9—Filter No. 12.

desired. Satisfactory results were obtained with A-2 cement.

The use of quarter-wave couplers permits the use of the largest possible diameter which will give the desired coupling coefficient. The couplers should be made of material having low mechanical losses. This material should also have a reasonably low frequency–temperature coefficient if the ferrite resonator disks have a very low frequency–temperature coefficient. Bandwidths obtained with various parts dimensions are shown in Table I and Figure 3.

Pivots or other supporting-spacing means which are in the flux path of the coils should have low electrical loss as well as good mechanical characteristics. Pivots should be either loose or cemented in the filter element in such a way as to minimize mechanical losses. Cemented

Table I—Data On Experimental Filters

		Couplers					
		Interior		End			
Filter Number	Bandwidth (cps)	d (inches)	d_1 (inches)	d (inches)	d_1 (inches)	R_1 (ohms)	R_2 (ohms)
8	8260	0.095	0.062	0.100	0.062	9000	18000
9	9500	0.095	0.058	0.100	0.058	18000	18000
10	5340	0.090	0.058	0.095	0.058	18000	9000
11	4000	0.086	0.066	0.090	0.066	50000	4000*
12	17000	0.107	0.058	0.115	0.058	40000	18000

* Across a tap on the output coil. Or this tap can be connected directly to a transistor circuit.

pivots may appreciably detune the end resonator and therefore should have small mass. Another mounting arrangement which was tried was a loose-fitting bakelite shaft through the filter. Small washers at each end spaced the filter with respect to the coil cores. This method of mounting is satisfactory if very little or no pressure is put on the spacers so that the filter element is positioned with substantially no binding.

A variation in disk and coupler shapes is shown at the bottom of Figure 3. A recess on each side of the disk centers the coupler and pivot-bearing assembly. The couplers are simply cut-off pieces of tubing with a permissible length variation of perhaps 2 or 3 per cent. The outer diameter is kept constant to fit the recessed hole in the ferrite disk, while the inner diameter is chosen to give the desired coupling coefficient. The couplers may be made of a variety of materials including possibly some ceramics. The dimensions given here apply only to the use of dural.

ACKNOWLEDGMENT

The author is indebted to R. L. Harvey of RCA Laboratories, who made all of the ferrite parts used in the experimental filters.

VAPOR PRESSURE DATA FOR THE MORE COMMON ELEMENTS

By

Richard E. Honig

RCA Laboratories,
Princeton, N. J.

Summary—*The most recent data available concerning vapor pressures, melting and boiling points, and heats of sublimation have been selected, tabulated, and plotted for 57 elements, many of which are of special interest to workers in the fields of electronics and high-vacuum technique. It has been found convenient to present vapor pressure data graphically as plots of log p (mm Hg) versus log T (°K), and also to tabulate the absolute temperatures for fixed pressures. This collection contains data published or available before March 1, 1957.*

Introduction

In 1948, R. R. Law[1] published vapor pressure data for some 30 substances, presenting all information graphically on a single sheet as curves of log p versus log T. Law's collection is based entirely on prewar data. A review article published by Brewer[2] in 1950 contains a substantial amount of information obtained during the war years.

In the intervening years, a large volume of new information has come into existence, and the need for an up-to-date collection of vapor pressure data has become more and more pronounced. This review, begun in December 1956, was initially based largely on the 1955 edition of "Metallurgical Thermochemistry" by Kubaschewski and Evans[3] as well as a comprehensive literature search carried out by the author. Upon completion of the first version, two unpublished compilations of data, by Stull and Sinke[4] and by Hultgren,[5] became available which

[1] R. R. Law, "Vapor Pressure Data for Various Substances (A Graphical Presentation)," *Rev. Sci. Instr.*, Vol. 19, p. 920, December, 1948.

[2] L. Brewer, *The Chemistry and Metallurgy of Miscellaneous Materials: Thermodynamics*, McGraw-Hill, New York, N. Y., 1950, p. 13.

[3] O. Kubaschewski and E. L. Evans, *Metallurgical Thermochemistry*, John Wiley & Sons, Inc., New York, N. Y., 1956, 2nd Ed.

[4] D. R. Stull and G. C. Sinke, "Thermodynamic Properties of the Elements," *Advances in Chemistry*, Series 18, American Chemical Society, Washington, D. C., 1956.

[5] R. R. Hultgren, "Selected Values for the Thermodynamic Properties of Metals and Alloys," Report of the Minerals Research Laboratory, University of California, Berkeley, California, 1956.

contained critical evaluations of data up to and including 1956. The present review is largely based on these three collections.

DATA

Presentation

Vapor pressure data may be presented very succinctly with the help of the equation

$$\log_{10} p = A\,T^{-1} + B\log_{10} T + C\,T + D \qquad (1)$$

where

p = pressure, expressed in this paper in millimeters of mercury,

T = absolute temperature in °K,

A, B, C, D = coefficients characteristic of the element.

In this system, employed by Kubaschewski and Evans,[3] the coefficients are readily tabulated in a minimum of space, but the evaluation of pressures for a given set of temperatures is quite cumbersome. The converse process—finding temperatures corresponding to given pressures—can only be done by trial and error and usually requires the service of a computing machine.

A more useful way of presenting vapor pressure data is to tabulate pressures corresponding to fixed temperatures, as was done by Stull and Sinke,[4] or to quote temperatures for given pressures. Both systems of tabulation were employed by Hultgren.[5] In general, we are interested in a fixed pressure range ($10^{-8} < p < 760$ mm Hg) which is the same for all elements, while the corresponding temperatures may lie anywhere between 0 and about 5,000°K. If curves are to be drawn from such tabulations, it is found that the points will be suitably spaced if pressure decades are employed. For this reason, temperatures have been tabulated at fixed pressures in the present paper.

Vapor pressure curves have been presented in the past usually as plots of log p versus $1/T$, largely because the resulting curves are straight lines over a limited temperature range and have slopes which yield heats of vaporization directly. Unfortunately, this type of temperature scale compresses the high temperature end so much that it is not feasible to present many vapor pressure curves on the same sheet over a wide temperature range. From a practical standpoint, the log p versus log T plots first introduced by Law[1] are by far the most suitable and are employed in this paper.

Selection and Treatment

A compilation of this type is based on the data of many workers who measured vapor pressures by different techniques under widely differing experimental conditions. Thus it is often very difficult to choose between two apparently equivalent sets of results, and at times the preference expressed may be subjective and arbitrary. Still, every effort has been made to include all data, based on reliable experimental measurements, for those elements that are of some practical importance. The only exceptions are iridium, palladium, and rhodium for which there exist to date only estimates.

The vapor pressure data available from the three major collections[3-5] were compared graphically by plotting log p versus log T. Kubaschewski and Evans' data,[3] given as the coefficients of Equation (1), were first converted into a tabulation of temperatures corresponding to fixed pressures with the help of a computing machine. They were then entered on the log p versus log T plots, and smooth curves drawn through the points. In the very complete collection by Stull and Sinke[4] vapor pressures are tabulated as $\log_{10} p$ (atm) at fixed temperatures. These could be put directly on the log p versus log T plots with the help of a special scale. Hultgren's data[5] for selected metals, tabulated as temperatures at fixed pressures (in atm), could be plotted directly.

Data based on Kubaschewski and Evans' collection[3] and on a literature search made by the author had already been plotted when the other two compilations became available. The curves were left unchanged wherever the pressures agreed with the other two sources to within about 20 per cent, which is the accuracy to be expected from vapor pressure measurements. Where there were larger discrepancies, preference was usually given to the evaluations of Hultgren[5] and Stull and Sinke.[4] It was felt that Hultgren's evaluation of a few selected metals was likely to be somewhat more thorough than Stull and Sinke's collection covering most of the elements, and that either was preferable to the extensive but less up-to-date collection by Kubaschewski and Evans who treated most of the elements and many compounds. From the curves mentioned above, the temperatures corresponding to fixed pressures were read off and tabulated.

For those elements having two or more important gaseous species of known concentrations, total vapor pressure curves were obtained by adding up the individual curves graphically. These are identified by a $\sum$ preceding the chemical symbol. In those cases where the atomic species is known to be the predominant contributor to the total vapor pressure, contributions from molecular species were neglected and the

Table I—Vapor Pressure Data and Melting and Boiling Points for the More Common Elements

SYMBOL	ELEMENT	DATA TEMP. RANGE °K	M.P. °K	TEMPERATURES (°K) FOR VAPOR PRESSURES (mm Hg)											B.P. °K	REF.	CURVE SHEET
				10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10^{1}	10^{2}			
Ag	SILVER	994-1273	1234	852	903	961	1030	1105	1195	1305	1440	1610	1830	2120	2435	3,4,5	A
Al	ALUMINUM	1410-1468	932	950	1010	1080	1155	1245	1355	1480	1620	1820	2050	2350	2720	3,4	A
As_4	ARSENIC		1090	377	400	423	447	477	510	550	590	645	710	790	886	4	A
Au	GOLD	1000-1260	1336	1045	1115	1180	1260	1355	1470	1605	1780	1980	2240	2580	2982	4,5	A
B	BORON		2300	1650	1735	1850	1960	2100	2250	2430	2650	2930	3300	3730	4200	4	A
Ba	BARIUM	1333-1411	983	560	600	640	690	740	810	900	1000	1140	1310	1570	1910	4	B
Be	BERYLLIUM	1172-1552	1556	972	1030	1100	1175	1260	1365	1485	1640	1840	2060	2370	2750	3,4	B
ΣBi	BISMUTH	682-770	544	590	629	672	723	781	851	934	1035	1165	1330	1555	1832	3,4	A
Br_2	BROMINE		266	113	120	128	137	147	159	174	192	214	243	282	331	3,4	A
ΣC	CARBON			1950	2030	2140	2250	2380	2520	2700	2900	3140	3420	3800	(4200)	4	B
Ca	CALCIUM	800-920	1123	555	590	630	675	725	790	865	960	1090	1240	1480	1765	4	B
Cd	CADMIUM	473-533	594	346	368	393	422	455	494	540	594	665	759	883	1038	3,4,5	B
Cl_2	CHLORINE		172					103	112	123	136	153	172	201	239	4	A
Co	COBALT	1363-1522	1768	1200	1265	1345	1435	1535	1650	1790	1970	2180	2440	2770	3150	3,4	A
Cr	CHROMIUM	1162-1561	2176	1125	1180	1250	1335	1435	1540	1665	1830	2010	2240	2550	2938	4,5	B
Cs	CESIUM		303	256	274	295	319	348	383	425	479	550	646	786	958	3,4	A
Cu	COPPER	1242-1879	1357	1005	1065	1135	1215	1305	1415	1545	1700	1895	2140	2460	2851	3,4,5	A
Fe	IRON	1356-1519	1812	1150	1220	1290	1380	1480	1595	1740	1910	2120	2370	2710	3130	5	A
Ga	GALLIUM	1230-1518	310	845	899	961	1030	1115	1210	1330	1470	1645	1870	2170	2510	3,4	B
Ge	GERMANIUM	1510-1885	1210	1085	1150	1225	1310	1415	1535	1680	1855	2070	2350	2710	3100	3,4	A
Hg	MERCURY		234	199	213	228	245	265	289	318	354	398	456	534	630	3,4	A
I_2	IODINE		387	178	188	199	212	226	242	262	285	312	345	388	456	3,4	A
In	INDIUM	1000-1348	429	770	820	877	943	1020	1110	1220	1350	1515	1730	2010	2364	3, 5	A
Ir	IRIDIUM	EST.	2727	1720	1840	1950	2070	2220	2380	2580	2800	3100	3440	3900	(4400)	4	B
K	POTASSIUM	373-1033	337	294	314	337	364	396	435	481	539	614	715	857	1039	3,4	A
La	LANTHANUM	1600-1900	1193	1260	1330	1430	1535	1650	1800	1970	2170	2420	2730	3180	3640	4	A
Li	LITHIUM	732-1353	454	505	539	577	621	672	733	806	896	1010	1155	1355	1604	3,4	A
Mg	MAGNESIUM	1009-1293	923	462	490	524	560	603	655	715	790	885	1000	1175	1377	4,5	A

SYMBOL	ELEMENT	DATA TEMP. RANGE °K	M.P. °K	TEMPERATURES (°K) FOR VAPOR PRESSURES (mm Hg)											B.P. °K	REF.	CURVE SHEET
				10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10^1	10^2			
Mn	MANGANESE	EST.	15b7	807	855	910	970	1040	1125	1220	1340	1500	1700	1975	2324	4,5	B
Mo	MOLYBDENUM	2070-2504	2850	1855	1970	2110	2260	2440	2650	2900	3200	3570	4040		5100	3,4	A
Na	SODIUM	537-1201	371	350	373	400	431	468	511	563	628	710	818	967	1163	3,4	A
Nd	NEODYMIUM		1297	1005	1070	1150	1230	1335	1465	1615	1810	2050	2370	2800	3360	4	B
Ni	NICKEL	1307-1583	1725	1185	1260	1330	1415	1520	1630	1770	1940	2150	2400	2750	3112	4,5	B
P_4	PHOSPHORUS		870	320	335	355	380	403	430	460	495	535	580	640	704	4	A
Pb	LEAD		601	617	658	705	760	824	900	992	1105	1250	1435	1695	2024	3,4	B
Pd	PALLADIUM	EST.	1823	1180	1255	1330	1430	1535	1660	1820	2000	2240	2560	2940	3400	4	A
ΣPo	POLONIUM	711-1018	527	387	408	432	460	493	536	587	655	745	860	1025	1235	4	A
Pt	PLATINUM		2043	1560	1650	1755	1875	2015	2180	2350	2590	2860	3210	3630	4100	4	B
Rb	RUBIDIUM		312	270	288	312	337	368	406	449	501	573	665	800	974	4	B
Re	RHENIUM		3453	2200	2330	2480	2640	2830	3060	3330	3670				5900	4	A
Rh	RHODIUM	EST.	2239	1550	1640	1745	1860	1980	2130	2300	2520	2800	3120	3530	(4000)	4	A
ΣS	SULFUR		392	241	256	272	291	313	339	370	408	456	519	606	717	3,4	B
ΣSb	ANTIMONY		903	550	580	615	655	700	750	815	890	1030	1250	1570	1910	4	A
ΣSe	SELENIUM		490	357	375	394	417	440	470	505	550	620	702	820	958	4	B
ΣSi	SILICON		1688	1200	1270	1355	1450	1555	1680	1820	1990	2200	2430	2740	3060	11	B
Sn	TIN	1424-1688	505	937	999	1070	1155	1250	1365	1500	1670	1885	2160	2540	2952	3,4,5	B
Sr	STRONTIUM		1043	499	533	571	615	667	729	804	896	1015	1170	1380	1640	3,4	B
Ta	TANTALUM	2000-3270	3270	2230	2360	2510	2670	2860	3080	3340	3645	4010			5700	3,4	B
Te_2	TELLURIUM		723	451	476	503	534	569	609	656	711	793	906	1065	1260	3,4	B
Ti	TITANIUM	1587-1764	1945	1330	1415	1500	1600	1715	1850	2000	2200	2450	2750	3130	3559	5	B
Tl	THALLIUM		577	558	595	637	685	741	808	888	986	1110	1270	1480	1740	3,4	A
U	URANIUM	1630-1970	1403	1405	1495	1600	1715	1855	2010	2200	2430	2720	3070	3540	4200	3,4	A
V	VANADIUM	1662-1882	2130	1428	1510	1600	1705	1824	1960	2120	2310	2560	2840	3220	3650	3,4	B
W	TUNGSTEN		(3650)	2340	2480	2640	2820	3030	3280	3570	3915				5800	3,4	B
Xe	XENON		161									105	120	140	165	4	B
Zn	ZINC	512-650	693	396	421	449	481	519	563	615	678	758	864	1005	1179	3,4,5	B
Zr	ZIRCONIUM	1949-2054	2125	1745	1850	1975	2110	2275	2460	2670	2920	3250	3620		4688	5	A

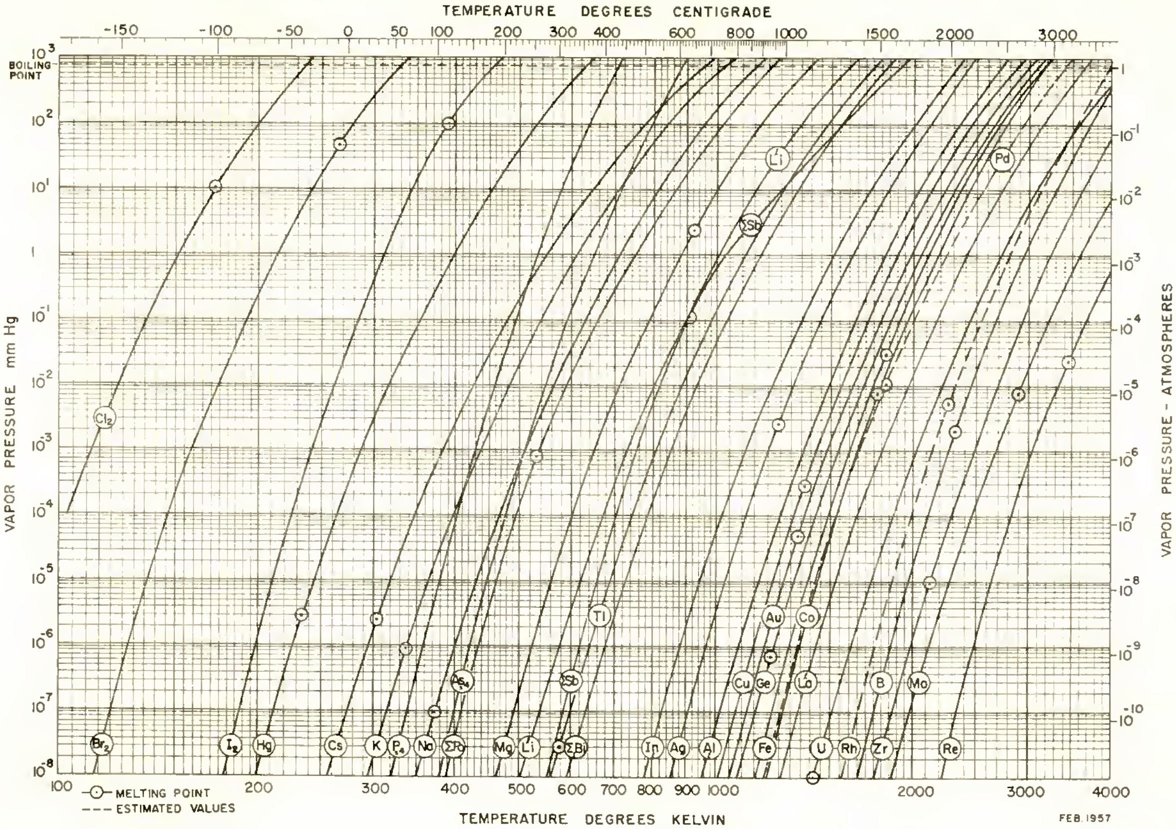


Fig. 1A—Vapor pressure curves for the more common elements.

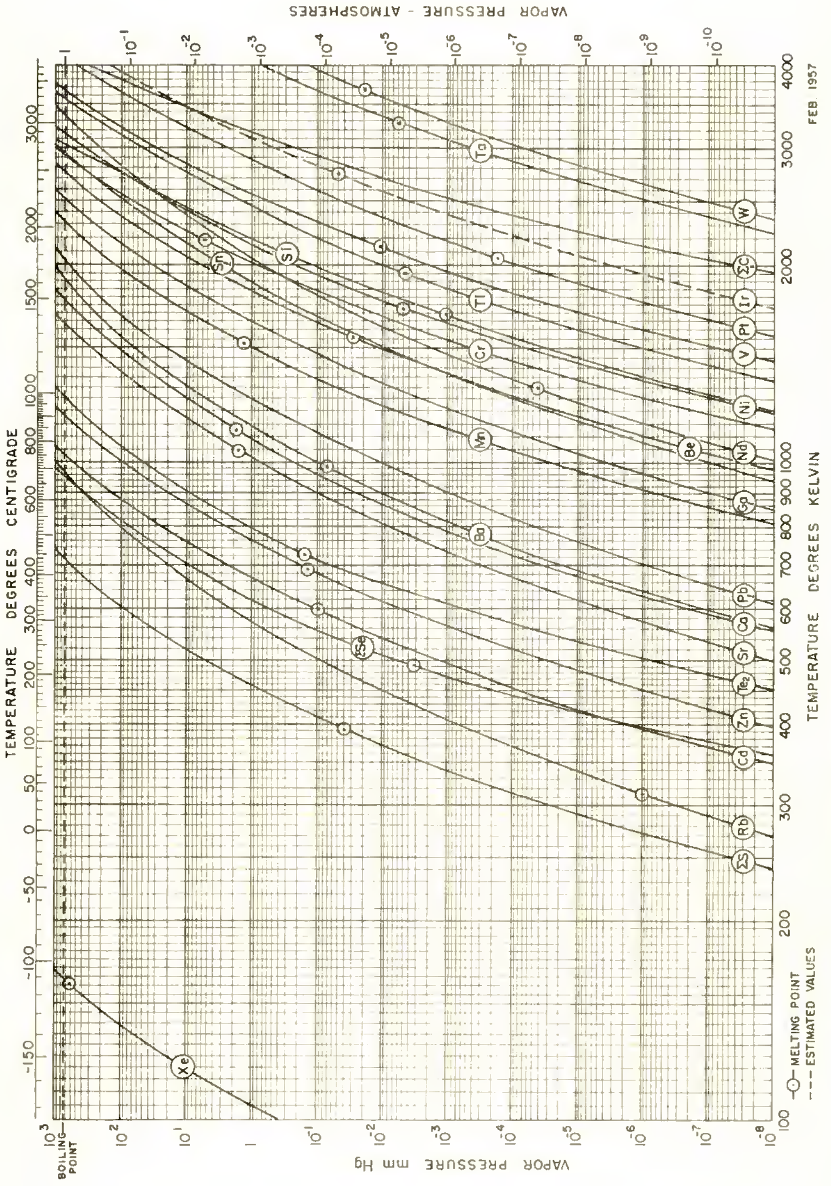


Fig. 1B—Vapor pressure curves for the more common elements.

symbol given without any subscript. For the few elements that consist largely of one molecular species, the appropriate subscript has been added.

Results

Table I contains the desired vapor pressure data, as well as the melting and boiling points, literature references, and in many instances the temperature range over which the original experimental data were obtained. The uncertainty in pressure considered tolerable, ±20 per cent, corresponds to a temperature uncertainty of between 1 and 2 per cent. To simplify matters, temperatures have been tabulated to ±1° between 100° and 1,000°K, to ±5° between 1,000° and 2,000°K, and to ±10° above 2,000°K, corresponding to the smallest increments that can be read on the graph. Thus it should be remembered that the last figure quoted is not always significant. In the reference column, the three collections rather than the original articles have been quoted, thereby avoiding the repetition of a very extensive bibliography.

Figures 1A and 1B present the vapor pressure data in graphical form.* The curves were placed on two separate sheets in such a way as to minimize interference. The circled point shown on most curves is the melting point. Where the melting point falls outside the pressure range of the graph, the letters "s" (solid) or "l" (liquid) have been appended to the chemical symbol. The last column of Table I indicates on which sheet a given element will be found.

Table II gives the most recent values available for the heats of sublimation, ΔH_{298}, quoted for the individual elemental species at 298°K. These heats are of considerable practical interest because they are a measure of lattice energies. The accuracy of the values quoted will lie somewhere between ±0.01 and 10 kcal/mole, is often difficult to assess objectively, and is not available for all entries. For this reason, estimate errors have not been tabulated, but indications of accuracy may be had from the number of significant figures given.

Discussion

Every effort has been made to include in this collection of vapor pressure data all information available as of March 1, 1957.

Aside from the three entries where data are based on estimates (iridium, palladium, and rhodium), there are three elements which

* A limited number of "wall-size" reproductions of these charts are available; requests for copies should be directed to RCA Laboratories, Princeton, N. J.

Table II—Heats of Sublimation of Elemental Species at 298°K

SPECIES	ΔH_{298} kcal/mole	REF.	SPECIES	ΔH_{298} kcal/mole	REF.	SPECIES	ΔH_{298} kcal/mole	REF.
Ag_1	68.2	5	Ge_1	92	7	Re_1	187	6
Ag_2	99	8	Ge_2	119	9	Rh_1	(133)	4
Al_1	77.5	4,7	Hg_1	14.65	4,7	S_1	57.5	7
As_1	>69	4	Hg_2	27.4	7	S_2	30.8	4,7
As_2	>48	4	I_2	14.88	4,7	S_4	28	7
As_4	34.5	4,7	In_1	57	4,7	S_6	28	7
Au_1	84.7	5	Ir_1	(150)	4	S_8	24.4	4,7
Au_2	119	8	K_1	21.42	4	Sb_1	63	7
B_1	141	4,7	K_2	30.6	4	Sb_2	56	7
B_2	213	7	La_1	100	4,7	Sb_4	49	7
Ba_1	41.7	4	Li_1	38.44	4	Se_1	49.4	4
Be_1	77.9	4	Li_2	50.5	4	Se_2	34.1	4
Bi_1	47.5	4	Mg_1	35.4	5	Se_6	35.4	4
Bi_2	55.3	4	Mn_1	67.1	5	Si_1	105	7
Br_2	7.45	4	Mo_1	157.5	4	Si_2	138	9
C_1	170.9	7	Na_1	25.85	7	Sn_1	72.0	5
C_2	200	9	Na_2	33.4	7	Sn_2	96	9
C_3	190	10	Nd_1	76.8	4	Sr_1	39.1	4
Ca_1	42.3	4,7	Ni_1	101.2	5	Ta_1	186.8	4
Ca_2	79	7	P_1	80	4	Te_1	46.5	4
Cd_1	26.78	5	P_2	42.7	4	Te_2	39.6	4
Cd_2	50.8	7	P_4	30.8	4	Ti_1	112.8	5
Cl_2	4.878 (@ 239°K)	4	Pb_1	46.8	4	Tl_1	43	4
Co_1	101.6	4	Pb_2	71	9	U_1	117.2	4
Cr_1	94.9	5	Pd_1	(94)	4	V_1	122.7	4,7
Cs_1	18.7	4	Po_1	34.5	4	W_1	200	4
Cs_2	26.6	4	Po_2	32.9	4	Xe_1	3.02 (@ 165°K)	4
Cu_1	81.0	5	Pt_1	134.8	4	Zn_1	31.22	5
Cu_2	115	8	Rb_1	19.6	4	Zn_2	56	7
Fe_1	99	7	Rb_2	27.6	4	Zr_1	146	4,5
Ga_1	65	4						

[6] L. Brewer, private communication.

[7] L. Brewer, "Heats of Sublimation of the Elements," University of California Radiation Laboratory Report 2854 (revised), February, 1955.

[8] J. Drowart and R. E. Honig, "Mass Spectrometric Study of Copper, Silver, and Gold," *Jour. Chem. Phys.*, Vol. 25, p. 581, September, 1956.

[9] R. E. Honig and J. Drowart, "Vaporization of Compounds and Alloys at High Temperature," Technical Note 1, University of Brussels, Belgium, 1956.

[10] R. E. Honig, "Mass Spectrometric Study of the Molecular Sublimation of Graphite," *Jour. Chem. Phys.*, Vol. 22, p. 126, January, 1954.

require discussion: silicon, aluminium and uranium. The vapor pressure data presented for silicon were obtained by the writer by an indirect method.[11] Recent, direct measurements of SiC by Searcy seem to have yielded[6,12] a silicon pressure of 1.6×10^{-6} mm Hg at 1,500°K (about 10 × lower than the present value) and $\Delta H_{298} = 113$ kcal/mole, but to date no further experimental data or details have been received. The data for aluminium are based on a study by Brewer and Searcy[13] who used BeO and TaC crucibles and observed considerable scattering of points. If the TaC results may be discarded, and if the effects of BeO–Al interaction are negligible, as suggested by Brewer,[6] then their data are much more consistent and will yield $\Delta H_{298} = 78.4$ kcal/mole. The vapor pressures given for uranium in Table I are consistent with $\Delta H_{298} = 117.2$ kcal/mole, as quoted in Table II. The heat of sublimation preferred by Brewer[6] is $\Delta H_{298} = 122.7$ kcal/mole.

Acknowledgments

It is a pleasure to acknowledge the courtesy of Drs. D. R. Stull and G. C. Sinke, Dow Chemical Company, and of Prof. R. R. Hultgren, University of California, who made available to the author prepublication copies of their collections of thermochemical data. The author is greatly indebted to Prof. L. Brewer, University of California, for bringing to his attention the two collections mentioned above and for supplying him with additional recent information. Thanks are due to R. Serrell and S. Skillman of the Computing Laboratory at RCA Laboratories for help in converting some of the data into suitable tabular form.

[11] R. E. Honig, "Sublimation Studies of Silicon in the Mass Spectrometer," *Jour. Chem. Phys.*, Vol. 22, p. 1610, September, 1954. These silicon vapor pressures are based on a comparison with germanium. The values given in Table I of the present report have been adjusted to take into account the most recent germanium vapor pressure data (A. W. Searcy and R. D. Freeman, "Measurement of the Molecular Weights of Vapors at High Temperature," II "The Vapor Pressure of Germanium and the Molecular Weight of Germanium Vapor," *Jour. Chem. Phys.*, Vol. 23, p. 88, January, 1955).

[12] A. W. Searcy, private communication.

[13] L. Brewer and A. W. Searcy, "The Gaseous Species of the Al-Al_2O_3 System," *Jour. Amer. Chem. Soc.*, Vol. 73, p. 5308, November, 1951.

ALLOYING PROPERTIES OF GERMANIUM FREE OF EDGE DISLOCATIONS*

By

C. W. Mueller

RCA Laboratories,
Princeton, N. J.

Summary—*Germanium that is substantially free of edge dislocations has alloying properties which are considerably different from those of germanium with many edge dislocations. In particular, surface spreading in equilibrium processes is difficult to control. Experiments are described showing the effect of surface tension and germanium crystallographic plane on indium alloy dot diameter. Alloying with controlled spreading may be accomplished by arranging for initial dissolution of the germanium by an unsaturated solution. An alloying method is described which consists of wetting an indium dot on germanium at 300°C in hydrogen and then using a very rapid heating in nitrogen to the alloy temperature of 550°C. The resulting junctions are extremely flat and uniform.*

Introduction

THE advances in the techniques of growing semiconductor single crystals have made materials available with a greatly improved physical structure. Germanium single crystals that are essentially free of edge dislocations can now be readily grown. The alloying or dissolution properties of indium on germanium that is free from edge dislocations is different than on germanium with many edge dislocations.

In the formation of p-n junctions for transistors, it is necessary that the area of the junction and its location be accurately controlled. The (111) crystal plane is a natural barrier in the crystal that forms the desirable planar junctions.[1] If, however, a crystal has no edge dislocations, resistance to penetration by the indium from a saturated solution is so strong that excessive and uncontrolled horizontal spreading of the molten dot occurs.[2]

This paper discusses the important factors in alloying on disloca-

* Presented at The Electrochemical Society Meeting, Cleveland, Ohio, October 3, 1956.

[1] C. W. Mueller and N. H. Ditrick, "Uniform Planar Alloy Junctions for Germanium Transistors," *RCA Review*, Vol. XVII, p. 46, March, 1956.

[2] H. V. Kettering and J. I. Pankove, "Crystal Quality as a Factor in Controlling the Alloying of Indium onto Germanium Surfaces," presented at the I.R.E.-A.I.E.E. conference on Semiconductor Research, Lafayette, Indiana, June 26, 1956.

tion-free germanium and shows that controlled alloying on dislocation-free material may be achieved by a process in which the dissolution occurs mainly by an unsaturated indium–germanium solution.

Physical Model

The forces concerned with the spreading of a molten dot can best be discussed with the help of Figure 1. The temperature cycle of an alloy junction is usually adjusted so that a saturated solution of germanium in indium occurs, thereby making the process essentially independent of time. Hence, the total volume of germanium taken in solution is the same for all units. The forces that determine the shape of this volume and how these forces can be controlled so that the desired shape is always attained will be discussed.

At the present time an examination of these forces in a fundamental manner is difficult because of the sketchy state of knowledge about the wetting and flow of one metal on another,[3] especially for indium and germanium. Since the shape of the solid–liquid interface is the result of the combined actions of the several forces, the rates of liquid motion in directions "A," "B," and "C" (Figure 1) are important. The rates are not definitely known, but they can be qualitatively discussed and their relative magnitudes illustrated by experiment.

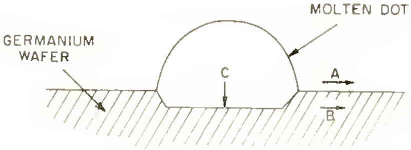


Fig. 1—Factors influencing spreading of a molten dot.

The motion of the molten indium can be resolved into horizontal and vertical components with respect to the surface of the wafer, which is nearly a (111) plane. From examination of junctions we know that the fastest horizontal rate is toward the corners of the triangle formed by the (111) planes that intersect the surface. Since maximum spreading must be controlled, the fastest rate must be contained, and this is the one considered.

The "A" rate (Figure 1) is a horizontal surface rate that depends on the surface tension of the molten dot and the wettability of the

[3] G. L. J. Bailey and H. C. Watkins, "The Flow of Liquid Metals on Solid Metal Surfaces and its Relation to Soldering, Brazing, and Hot-Dip Coating," *Jour. Inst. of Metals*, Vol. 80, p. 57, 1952.

surface by the dot material. It is influenced by surface conditions, temperature, and dot material.

The "B" and "C" rates are not influenced by the surface; they are mainly determined by the dissolving of the germanium by the indium. They are solution or alloying forces and depend upon the temperature and the bulk properties of the germanium, i.e., crystallographic direction and perfection of the crystal. If the molten dot is uniformly mixed, the "B" and "C" rates will be the same when the crystal bulk properties are isotropic. Excessive spreading can be prevented by decreasing the rates "A" and "B" and increasing rate "C." Experiments will be described which illustrate how these rates can be greatly changed by the choice of experimental conditions.

In all the experiments, germanium areas free from edge dislocations were used. All the germanium wafers were etched for 10 minutes in CP4 and then examined microscopically for etch pits. Regions free of etch pits associated with edge dislocations[4] were then selected. This was not difficult in material that had an average count of 100-300 pits per square centimeter.

Indium–zinc balls (0.5 per cent zinc) 0.014 inch in diameter cleaned in 1 per cent HCl were used for dot material unless otherwise stated. These dots spread slightly less than pure indium dots. Uniform wetting was reproducibly attained using hydrogen as the flux at temperatures of about 300° C. No liquid fluxes were used. The large undercutting of the surface on dislocation-free germanium by the molten dot aids in removing any initially unwet areas.

Surface Tension

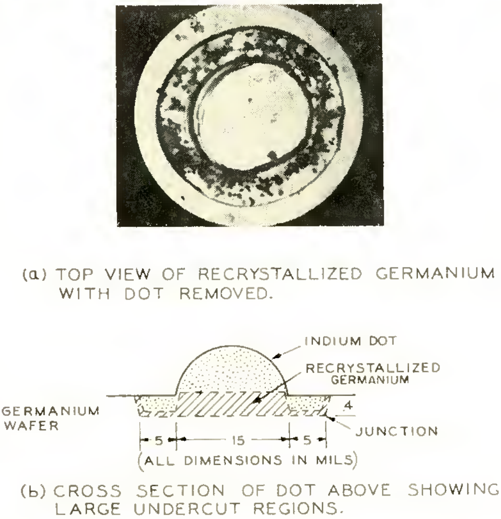
Several methods are available to reduce rate "A" which is closely connected with surface tension:

1. The use of a slightly oxidizing atmosphere.
2. Mechanical constraint by a jig.
3. Use of a dot material such as lead alloys having a higher surface tension.

In this paper only method 1 will be discussed. Method 2 has the disadvantage of mechanical wear at high temperatures and usually does not sufficiently limit junction area because the molten liquid can undercut the surface. Method 3 was tried and it was found that the use of lead inhibits excessive spreading only when lead-rich alloys ($>$50 per cent) are used. However, in this case the alloy penetration is very shallow.

[4] S. G. Ellis, "Dislocations in Germanium," *Jour. Appl. Phys.*, Vol. 26, p. 1140, September, 1955.

The use of a slightly oxidizing atmosphere during alloying is probably the simplest way to change rate "A." An experiment that eliminates surface tension gives an insight into the nature of the bulk alloying mechanisms involved in forming a junction. The indium–zinc dot is first soldered or wetted down in a pure hydrogen atmosphere at 300° C. The alloying step, heating to 550° C, is then done in a gas from which all the oxidizing gases have not been removed—line nitrogen, for example. Traces of oxygen and water vapor attack the indium very slightly and form a thin oxide film. This oxide film forms a membrane over the dot and restrains the dot in a hemispherical form. The diameter of the hemisphere, as determined by a machinist's microscope or shadowgraph, does not grow during the alloying step.


(a) TOP VIEW OF RECRYSTALLIZED GERMANIUM WITH DOT REMOVED.

(b) CROSS SECTION OF DOT ABOVE SHOWING LARGE UNDERCUT REGIONS.

Fig. 2—"Hat brim" type of structure from use of an oxidizing atmosphere (not drawn to scale).

The use of an oxidizing atmosphere thus provides good control of the surface forces (rate "A" of Figure 1). However, on close examination of the alloy dot, one finds that the germanium surface has been undercut, i.e., rate "B" has not been reduced and a "hat brim" type of structure has been created as shown in Figure 2. The surface of the crystal of Figure 2 is parallel to the (111) plane. Heating and cooling was done at a rate of 20° C per minute between 400° and 550° C. Typical dimensions of the outer brim, as shown in Figure 2, can be large, thus causing high electrical capacitance and large collector-to-emitter spacing.

Note that the horizontal spread from the original dot is more than 10 times the vertical penetration. The recrystallized germanium is very thin in the outer brim region because as the dot cools, mixing is restricted and the amount of germanium available for recrystallization is small. One sees that merely restricting the surface rate of spread does not by itself restrict the actual junction because of the very slow rate at which the indium penetrates the horizontal (111) plane. Consequently the bulk crystal properties must also be considered as a means of coping with the problem of lateral undercutting.

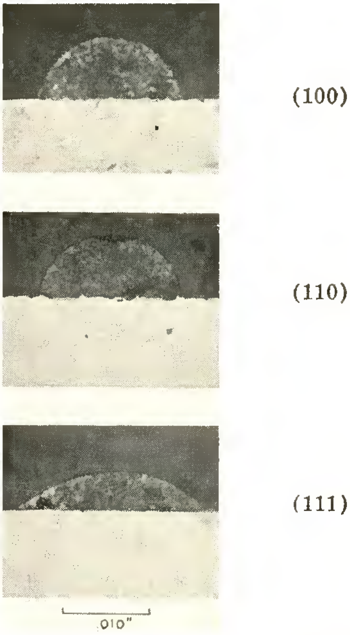


Fig. 3—Cross section of junctions produced on different crystal planes.

Crystal Axis

One way of changing directional properties of the alloying process is by cutting the germanium crystal so that a plane different from the (111) is parallel to the surface. Figure 3 shows a series of cross section photographs illustrating the results of alloying on the different crystal planes when all are processed in the same manner at the same time: namely, wetting down of the 0.014-inch spheres in hydrogen at 300° C and then alloying to 550° C in hydrogen with a heating and cooling rate of 20° C per minute between 400° and 550° C. Micrometer measurements and the pictures clearly show that there is much less spreading on the (100) and (110) planes. Close examination of the junctions of Figure 3 as well as the examination of others from which the indium was removed, shows that the wetting on the (100) and (110) planes was considerably poorer than on the (111) plane. On the (111) plane the large undercutting removes unwet areas. This experiment does not mean that it is impossible to use the (100) or the (110) planes of germanium, but that the wetting on these planes is more difficult. In other words, a wetting problem is substituted for the spreading problem. Consequently, it is desirable to have a method of utilizing the (111) crystal plane.

Rate of Dissolution

If indium in contact with germanium is heated slowly under conditions approaching equilibrium, then germanium atoms are continually leaving, and returning to, the germanium wafer. In this case the differences in binding forces acting on the atoms in different crystallographic locations have a chance to be effective. On the other hand, if the solution rate is rapid (as in unsaturated indium) the crystal forces are overcome by the more violent dissolving forces and the differences of binding energy for variously located atoms have less effect. Under a rapid heating, nonsaturated condition, the solution of germanium by indium is that of a liquid drop acting at nearly the same speed in all directions. Consequently while the liquid is in a highly unsaturated condition, the (111) plane does not hinder penetration as markedly. Experiments on the solution rate of oriented germanium wafers in molten indium are consistent with this hypothesis.[5]

At the end of the heating period while the temperature is constant, the solution becomes nearly saturated. Hence crystal forces on the atoms are comparatively stronger and can level the bottom of the junction.

A very rapid temperature rise promotes germanium solution by unsaturated indium. In Table I the large effect of changing the rate of temperature rise is indicated. The indium–zinc dots (0.5 per cent zinc) on all the units of the table were first soldered in hydrogen on dislocation-free germanium at 300° C. The diameter of the dots was then measured in two perpendicular directions. An atmosphere of line nitrogen was used for the alloying. As shown in the table, the units that were heated at 10° C per minute spread an average of 18.7 mils. while the rapidly heated (1,500° C per minute) units spread only

[5] B. Goldstein, "The Dissolution of Germanium by Molten Indium," *RCA Review*, Vol. XVIII, p. 213, June, 1957.

TABLE 1

GROUP	HEATING RATE 300°C. TO 550°C.	MINUTES HELD AT 550°C.	COOLING RATE 550°C. TO 300°C.	FINAL DIAM. (MILS)	INCREASE IN DIAM. (MILS)
A	10°C/MIN.	3	10°C/MIN.	35.8	18.7
B	1500°C/MIN.*	3	"	19.6	2.3
C	1500°C/MIN.*	30	"	19.7	2.4
D	125°C/MIN.	3	100°C/MIN.	30.6	13.4

* NOT CORRECTED FOR HEAT LOSS DUE TO THERMOCOUPLE LEADS

2.3 mils. The sidewise alloying spread or undercutting was thus reduced by a factor of about eight.

In order to determine if the solution reached saturation with the very rapid heating rate the top temperature of 550° C was held in group C (Table I) for 30 minutes. The data shows that additional spreading is negligible. The depth of penetration is about the same. Although the rate of rise must be rapid, time at the top temperature is not critical. Equilibrium is reached at the top temperature and the process is no longer a dynamic process. This is desirable because a dynamic process is generally more difficult to control. The decrease of temperature from the maximum can be independently controlled to give the desired type of recrystallization (see Figure 4).

In order to secure the fast rate of rise shown, a jig with a large heat mass cannot be used because of its thermal inertia. In the experi-

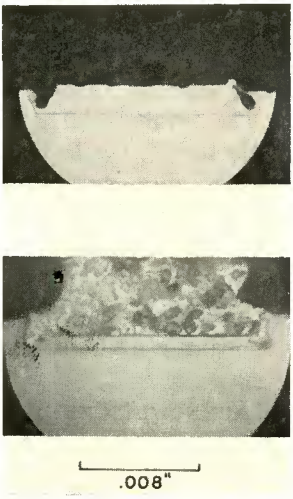


Fig. 4—Cross section of junctions made by using very rapid temperature rise: top view, slow cooling; bottom view, rapid cooling.

ments described, the germanium pellets were placed on a very thin sheet of mica and pushed into a furnace which had been brought to the top temperature of 550° C. Laying germanium wafers on a graphite boat and pushing this into a furnace at 550° C produces the results shown in group D of the table. Although the spreading is somewhat less than for group A, it is not nearly as small as that shown by the units of groups B and C with the extremely rapid temperature rise.

A good alloying process should also produce junctions that are crystallographically flat. The cross sections of junctions made with a very rapid temperature rise as described above are shown in Figure 4 which shows that the flatness is determined by a crystal plane. These units made on the dislocation-free germanium not only show flat bottoms but also show sides that are determined by crystal planes. If 5 junctions are made in a line on a wafer and a cross section of the entire group is made, one finds that they are all perfectly formed. This degree of uniformity of the junctions made as described above is very striking, and shows that, for this type of crystal, a slow rate of heating is not as important a parameter as heretofore supposed.

Conclusions

By arranging the germanium alloy cycle so that initial dissolution by an unsaturated indium solution occurs, the restrictive action of the (111) crystal plane may be minimized. This action gives a simple method of making junctions without excessive spreading on germanium free of edge dislocations. As the final atom layers are removed the solution becomes saturated and the restrictive action of the (111) plane again becomes very strong leveling the bottom of the junction. A two-step process is used. The dot material is wetted to the germanium at 300° C in hydrogen. No liquid flux is used. The alloying is then done in line nitrogen with an extremely rapid temperature rise to the alloying temperature of 550° C. Equilibrium is reached in less than three minutes and the temperature decrease can be governed by the type of recrystallized region desired.

The junctions formed by the above process are extremely flat and uniform when examined optically at high magnification.

Acknowledgment

The skill and patience of Mrs. E. Moonan in preparing and sectioning many samples is greatly appreciated.

THE DISSOLUTION OF GERMANIUM BY MOLTEN INDIUM*

BY

B. GOLDSTEIN

RCA Laboratories,
Princeton, N. J.

Summary—*Experiments have been performed on the dissolution of germanium by molten indium as functions of crystal axis, temperature, and germanium concentration of the solvent. Plots of germanium dissolution versus time are generally characterized by two regions indicating an initial rate and a final rate. For the initial rates the [111] crystal direction is slowest and the [110] crystal direction is fastest. As the dissolution proceeds, other crystal faces become exposed so that the final rates, being less a function of any one crystal direction, are more nearly equal. When the dissolution rates are varied by temperature or by solvent constitution, it is found that as the over-all dissolution rate increases, the rates in the three major crystal directions become more nearly equal. There are similarities between the dissolution of germanium by molten indium and the etching of germanium by an acid etch.*

INTRODUCTION

MOLTEN indium pellets alloyed in the customary manner[1] onto germanium surfaces with low edge dislocation density tend to spread excessively.[2] Since bulk dissolution properties may be expected to play an important role in alloying, one of the approaches to coping with the problem of excessive spreading lay in modifying that part of the alloying procedure in which the germanium is dissolved by the molten indium. Accordingly, experiments were performed to answer certain specific questions relating to such dissolution. These experiments include a study of dissolution as a function of crystal axis orientation, temperature, and amount of germanium already in the molten indium solvent.

* Presented at The Electrochemical Society Meeting, Cleveland, Ohio, October 3, 1956.

[1] C. W. Mueller and N. H. Ditrick, "Uniform Planar Alloy Junctions for Germanium Transistors," *RCA Review*, Vol. XVII, p. 46, March, 1956.

[2] H. V. Kettering and J. I. Pankove, "Crystal Quality as a Factor in Controlling the Alloying of Indium onto Germanium Surfaces," presented at the I.R.E.-A.I.E.E. conference on Semiconductor Research, Lafayette, Indiana, June 26, 1956.

EXPERIMENTAL RESULTS

Thin wafers of low-dislocation-density germanium about one half inch long, one eighth inch wide and about 40-60 mils thick were cut and carefully oriented so that the large-area faces were perpendicular to a major crystallographic axis. The wafers were surface ground and then etched in CP4 to remove about three mils of worked surface. (This was later increased to five mils; however, it was found that the results were independent of the amount removed by the CP4 beyond the original three mils.) The wafers were coated with a flux* to insure good uniform wetting and then partially dipped into the solvent (pure indium or a germanium–indium solution). Constant, nonviolent agitation was supplied to prevent the germanium already in solution from interfering with the dissolution process. After a certain time the wafers were removed, and the excess indium dissolved by HCl. No trace of recrystallized germanium could be detected.

The dipped portion of the wafer was then compared to the undipped portion under a microscope. In the initial stages of the dissolution (5-10 seconds) a pattern is formed on the wafer surfaces consisting of plateaus of germanium which had been only slowly attacked by the indium, surrounded by regions of more rapidly attacked germanium. These plateaus vary in size from one micron to several mils. They are mostly circular with some few of the smaller ones triangular and hexagonal. With time the plateaus get smaller, finally disappear, and an over-all decrease in wafer thickness is observed. The surface at this point is not flat, and the decrease in thickness can be recorded only as a gross, averaged quantity. The data is plotted as the amount of germanium dissolved (in terms of the decrease in thickness) versus time. The measurements of decrease in thickness are accurate to about $\pm 10^{-3}$ centimeter.

Figure 1 shows the rates of dissolution of germanium by pure indium at 347°C as functions of the primary crystal axes. The rates are given in Figure 1 and also in Table I. Each point on a curve represents a separate test wafer. These, and subsequent curves are characterized by two regions indicating an "initial" and "final" dissolution rate. To compare initial and final rates, straight line segments have been used together with a "knee" to indicate the gradual transition from the initial to the final region.

A possible reason for the two regions in the curves may be thought at first to be due to an insufficient CP4 etch leaving some worked surface to be dissolved by the indium before the undisturbed bulk of the

* Similar to the flux described by Mueller and Ditrick in Reference (1).

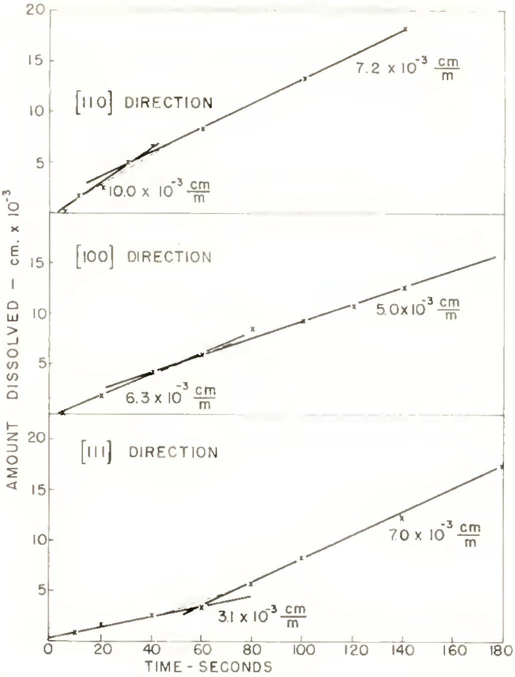


Fig. 1—Dissolution of germanium by molten indium, $T = 347°C$.

crystal was reached. However, when 5×10^{-3} centimeter, of germanium (the "break point" in the curves of Figure 1) was etched in CP4 in addition to the amount ordinarily etched, essentially the same curves are produced. This ruled out the possibility mentioned above. It is now thought that the two regions generally observed in the dissolution rates are due to the fact that although the test wafers have flat oriented surfaces initially, as the dissolution proceeds the surfaces become more and more roughened. This means that ultimately the direction of the indium's attack is distributed among several crystal directions and not solely in any one direction. Thus, the initial regions

Table I—Rates of Dissolution of Germanium by Molten Indium

	$T = 347°C$ Rate, in cm $\times 10^{-3}$/min.	
Crystal Orientation	Initial	Final
(110)	10.0	7.2
(100)	6.3	5.0
(111)	3.1	7.0

in the curves of Figure 1 are much more the property of crystal axis than the final region. It is found that, consistent with this, differences in the dissolution rates observed for the final regions are much less than differences in rates for the initial regions (see, for example, Table I). Note that in the initial rates direct experimental evidence is found for something which had heretofore been only strongly suspected—namely that the formation of flat alloy junctions on a (111) crystal plane is due to the fact that indium dissolves germanium slowest in the [111] direction.

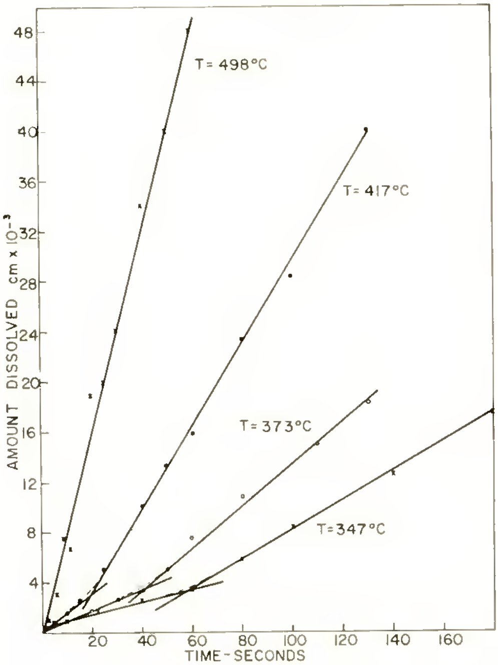


Fig. 2—Dissolution of germanium by molten indium, [111] direction.

The final rates given in Table I are little different from one another, the greatest difference being between the [111] and the [100] directions, a factor of 1.4. However, the initial [110] rate is 3.2 times as fast as the initial [111] rate, and later on it is shown that when the actual conditions of junction fabrication are more closely approximated, the ratio of the (110) alloying rate to the (111) alloying rate is even greater.

The dissolution rate of the (111) oriented test wafers was then determined at 373°C, 417°C, and 498°C. This data is shown in Figure 2. At 373°C and 417°C initial and final regions are again present, with increased rates for both. At 498°C only one rate appears although an initial rate may very well be present but not detectable by our methods because of the rapidity of the indium's attack. The dissolution rate increases sevenfold as we go from 347°C to 498°C. When the available data of Figure 2 is plotted as dissolution rate versus inverse absolute temperature, the results are as shown in Figure 3. Here, we see that the dissolution rate depends exponentially on the inverse absolute temperature. If we assume that the temperature dependence of the rate-determining step in the dissolution process is given by a Boltzmann factor, then the activation energy associated with the dissolution process is .53 electron volt. Interestingly enough, this is an activation energy associated with the etching of germanium by No. 2 acid etch found by P. Camp.[3] Furthermore, on comparing the etch data of Camp's Table II with our own data (Table I) it is found that the dependence of the dissolution rate on crystal axis is the same

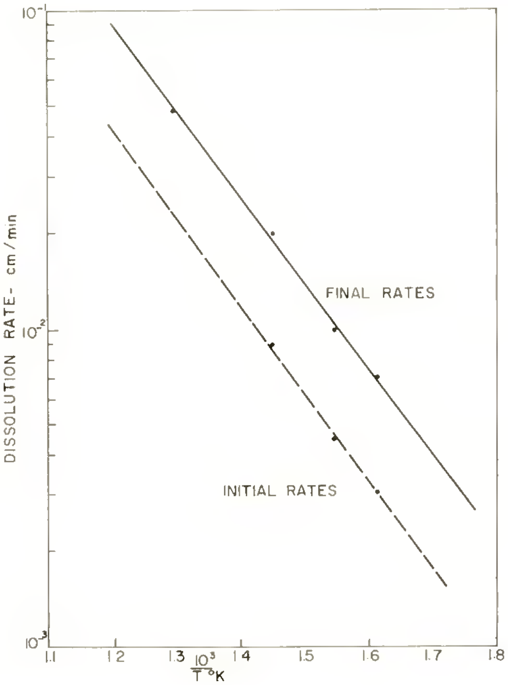


Fig. 3—Temperature dependence of dissolution rate of germanium by molten indium, [111] direction.

[3] P. J. Camp, "A Study of the Etching Rate of Single-Crystal Germanium," *Jour. Electro. Chem. Soc.*, Vol. 102, p. 586, October, 1955.

(qualitatively) for both No. 2 etch and molten indium. This behavior is true for both the initial and the final rates.* Thus, even though acid etching involves the formation of chemical compounds while dissolution by indium does not, identical activation energies and identical dependence of dissolution rate on crystal axis suggest that both acid etching and dissolution by molten indium may be very similar in mechanism.

Increasing the temperature of the molten indium increased the rate of dissolution as expected. An experiment will now be described in which a slower dissolution rate was produced not by decreasing the temperature, but rather by adding germanium to the molten indium. In this way the conditions present during the fabrication of an alloy junction are more nearly approximated, since in the actual fabrication, which customarily involves very slow heating, the germanium is dissolved by an indium–germanium solution. Accordingly, a mixture of germanium and indium was prepared which produced, at the temperature of the experiment, a solution about 75 per cent saturated with germanium. The dissolution rates were determined and the results are given in Figure 4. Essentially the same general characteristics as those found in Figure 1 appear. Two rates, an initial and a final one, are apparent in each of the curves. The final rates are less than the initial ones for the [110] and [100] directions, while the opposite holds for the [111] direction. Again we note that differences among the final rates are less than those among the initial rates.

Table II—Rates of Etching of Germanium by No. 2 Etch

Crystal Orientation	Rate, in cm. × 10^{-4}/min. Initial	Final
(110)	9.6	6.0
(100)	8.4	3.8
(111)	8.0	4.3

There are some important differences, however, between these curves and those in Figure 1. The general dissolution rates are about fifteen times slower than the rates using pure indium (note the abscissa scale change in Figure 4 from seconds to minutes). It is seen that with these much slower rates, the difference between the (110) rate and the (111) rate is given by a factor of five. Elaborating this par-

* The identity for the initial rates of the two processes may not be significant. The initial rate with No. 2 acid etch may correspond to the removal of worked surface. In the case of dissolution by molten indium, the worked surface had already been removed.

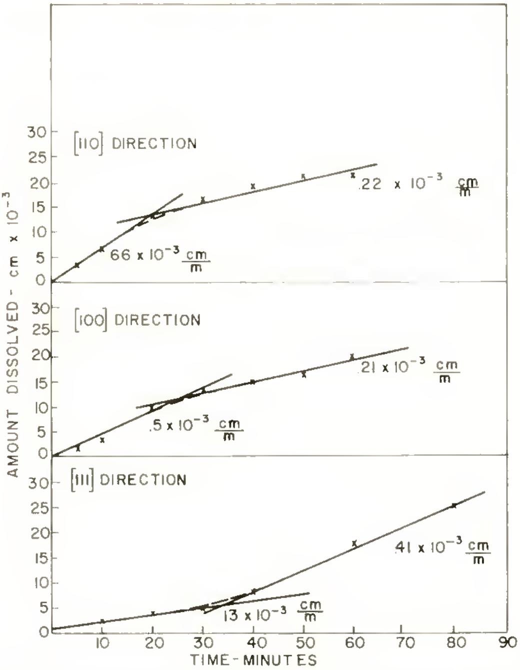


Fig. 4—Dissolution of germanium by molten indium–germanium solution (75 per cent saturation). $T = 347°\text{C}$.

ticular point, the initial rates have been listed in Table III for those conditions where the over-all rates are markedly different. The bottom line in this table gives the ratios of the (110) rate to the (111) rate. From this it is concluded that the slower the dissolution rate, whether it be due to temperature or to solvent composition, the greater the difference in initial dissolution rates, and consequently the greater the ratio of lateral dissolution speed to vertical dissolution speed on a (111) oriented surface. These are shown as the bulk dissolution

Table III—Dissolution Rate of Germanium

	Rate, in cm × 10^{-3}/min.			
Crystal Orientation	347°C In—75% Ge	347°C Indium	417°C Indium	500°C Indium
(110)	.66	10.0	17	~ 50
(100)	.50	6.3	14	~ 50
(111)	.13	3.1	9	~ 50
$\frac{(110)\ \text{rate}}{(111)\ \text{rate}}$	5.0	3.2	1.9	~ 1

processes B and C, Figure 5, in which the factors relating to the spreading of a pellet of molten indium during alloying are illustrated schematically. Process A, Figure 5, relates to wetting and surface tension forces.

Discussion

It has been found that surface undercutting, process B, Figure 5, is the major cause for excessive spreading of indium during alloying.[4] One way to minimize this effect would be to make processes B and C more nearly equal. The data in Table III indicates that this may be accomplished by increasing the over-all dissolution rate. This has, in fact, been found to be the case.[4] When slow heating of the alloy units is used (as is customarily done) the germanium is dissolved by a solution already essentially saturated with germanium. This produces a very slow dissolution rate and a great disparity between the rates of B and C, Figure 5. When a very fast heating rate is employed, causing most of the alloying to take place at high temperature and with the dissolving solution far from germanium saturation, much faster dissolution rates and more nearly equal B and C rates are obtained. It is under these latter conditions that one finds no excessive spreading. It is to be emphasized that the fast heating rate *per se* does not produce the favorable alloying conditions, but rather the high temperature and nonsaturated condition of the solvent. We may remark in passing that even with the rapid heating, one can still produce the required flat junctions for which it was thought the slow heating was necessary.

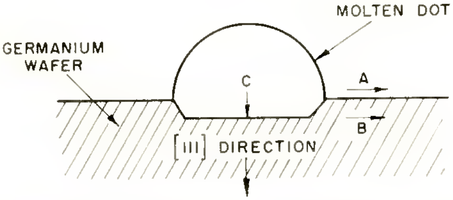


Fig. 5—Factors influencing spreading of a molten dot. Process (A) relates to wetting and surface tension forces. Processes (B) and (C) relate to bulk dissolution processes.

[4] C. W. Mueller, "Alloying Properties of Germanium Free of Edge Dislocations," *RCA Review*, Vol. XVIII, p. 205, June, 1957.

ON THE NONLINEAR BEHAVIOR OF ELECTRON-BEAM DEVICES

By

F. Paschke

RCA Laboratories,
Princeton, N. J.

Abstract

Summary—*The nonlinear space-charge-wave equation is derived. By third-order successive approximation this complicated equation is split into three simultaneous linear differential equations, which are solved for the case of a velocity-modulated electron beam (klystron). When excited at the fundamental frequency ω, the beam produces all harmonics with the frequency $n\omega$. At fairly low levels the maximum current amplitudes are found to be proportional to the $(1-n/2)^{th}$ power of the direct current. Equations for the gain and efficiency of a two-cavity klystron are given. The saturation efficiency is computed to be about 55 per cent. Thus the space-charge, or "longitudinal debunching," does not reduce the efficiency of a klystron appreciably below the well-known ballistic theoretic value of 58.2 per cent. Furthermore, it is shown that a computation of kinetic power flow, using the linear solutions, is not quite justifiable. With the simultaneous linear differential equations given, it is possible to treat nonlinear phenomena in traveling wave-tubes.*

Introduction

THE theoretical evaluation of efficiency, harmonic frequency generation, mixing, etc. in electron-beam tubes requires an understanding of the nonlinear behavior of the device. This problem has been the subject of many papers, each dealing with one special type of tube, e.g., klystron and traveling-wave tube. Webster[1,2] investigated a velocity-modulated electron beam (klystron) under conditions of arbitrary signals, but took into account only the linear space-charge term. Hence, this theory is, for large signals, a purely ballistic one. Gittins and Sullivan,[3] however, showed experimentally that with any appreciable current the neglect of space charge in a

[1] D. L. Webster, "Cathode-Ray Bunching," *Jour. Appl. Phys.*, Vol. 10, p. 501, July 1939.

[2] 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.

[3] J. F. Gittins and A. B. J. Sullivan, "On Electron Bunching by Traveling Waves," *S.E.R.L. Tech. Jour.*, Vol. 5, p. 127, December 1955.

velocity-modulated beam is never justifiable. Nordsieck,[4] in his traveling-wave-tube theory, followed a typical set of electrons and, again neglecting the space charge, computed their velocities and positions by numerical integration. Tien, Walker, and Wolontis[5] and Rowe[6] included space-charge effects by superimposing the space-charge force and the force of a divergence-less electric-circuit field. The space-charge field was calculated by replacing the delay line (helix) by a perfectly conducting drift tube which may lead, according to Reference (7), to appreciable errors. In addition, certain assumptions were made for the distribution of space charge along the axis. Tien, Walker, and Wolontis computed the space-charge force from a disk-charge model. Rowe presumed a sinusoidal space-charge variation along the axis. Neither presumption holds in actual practice.

A straightforward way of calculating the nonlinear behavior of electron-beam devices is to take the electron stream not as a certain limited number of electrons but as a "fluid" where the discrete charges are thought to be "smeared out" and to integrate the nonlinear space-charge-wave equation for the given boundary conditions. Here no superposition of a space-charge field and a circuit field is possible because of the nonlinearity of the field equation. In Maxwell's field theory there is only *one* definite total electrical field strength for which it is valid to say that

$$\operatorname{div} E = \rho/\epsilon_0$$

inside of the beam, and

$$\operatorname{div} E = 0$$

outside of the beam. This "Eulerian" approach, however, is valid only until electron overtaking occurs which means that the velocity becomes a multi-valued function of space. As will be seen later, the theory always yields (for a given position) definite velocities. This means that in our fluid model overtaking does not occur.

The nonlinear space-charge-wave equation is very complicated and

[4] A. Nordsieck, "Theory of the Large Signal Behavior of Traveling-Wave Amplifiers," *Proc. I.R.E.*, Vol. 41, p. 630, May 1953.

[5] P. K. Tien, L. R. Walker, and V. M. Wolontis, "A Large Signal Theory of Traveling-Wave Amplifiers," *Proc. I.R.E.*, Vol. 43, p. 260, March 1955.

[6] J. E. Rowe, "A Large Signal Analysis of the Traveling-Wave Amplifier," *Trans. I.R.E. PGED*, Vol. 3, p. 39, January 1956.

[7] F. Paschke, "Die Wechselseitigkeit der Kopplung in Wanderfeldröhren," *Arch. Elektr. Uebertr.*, Vol. 11, p. 137, April, 1957.

extremely difficult to integrate. On the other hand we do not need a complete solution, correct for all power levels. In order to predict saturation and the generation of the first two harmonics, a third-order theory is sufficient. The purpose of this paper is to derive the nonlinear space-charge-wave equation and to solve it by third-order successive approximation for the special case of a drifting velocity-modulated beam (klystron).

The Space-Charge-Wave Equation

To simplify the analysis we make the following assumptions:

1. One-dimensional confined electron flow.
2. The electric r-f field E in the direction of the stream is approximately constant over the cross section of the beam.
3. The electron velocity is small compared to the velocity of light.
4. The effects of the potential depression caused by space-charge across the beam and the thermal velocity distribution are negligible.

The convection-current density

$$\boldsymbol{i}_0 + \boldsymbol{i} = (\rho_0 + \rho)(v_0 + v), \tag{1}$$

$$\boldsymbol{i}_0 = \rho_0 v_0$$

appears in the continuity equation

$$-\frac{\partial i}{\partial x} = \frac{\partial \rho}{\partial t}. \tag{2}$$

The subscript 0 refers to the unperturbed d-c quantities at the input plane. The perturbations i, ρ, and v are caused by the r-f field and are functions of time and distance from the input plane. According to the assumptions listed above, the electrons are influenced only by the axial electric field E. Thus the force equation is

$$\frac{\partial v}{\partial t} + (v_0 + v)\frac{\partial v}{\partial x} = \eta E. \tag{3}$$

The coupling between the beam and the structure is produced by the transverse electric field; hence, a displacement current per unit length, M, flows from the beam (Figure 1). This coupling current enters into the continuity equation for the total current flowing through an infinitesimal beam element;

$$\frac{\partial i}{\partial x} + \epsilon_0 \frac{\partial^2 E}{\partial x \partial t} + \frac{M}{S} = 0. \tag{4}$$

We introduce now a new variable, the "displacement"

$$y = \epsilon_0 E + m, \tag{5}$$

where m represents the transverse coupling field and is given by

$$\frac{\partial^2 m}{\partial x \partial t} = \frac{M}{S}. \tag{6}$$

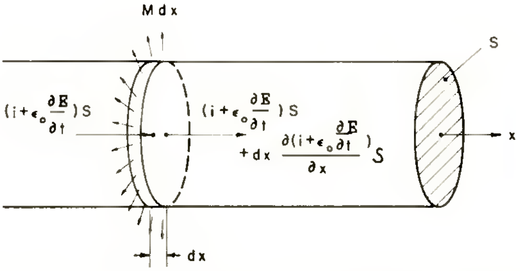

M d x
S
$(i+\epsilon_0 \frac{\partial E}{\partial t})S$
$(i+\epsilon_0 \frac{\partial E}{\partial t})S$
$+dx \frac{\partial(i+\epsilon_0 \frac{\partial E}{\partial t})S}{\partial x}$
x
d x

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

Equations (1)-(6) yield the following nonlinear differential equation for the displacement y:

$$\left(\frac{\partial^2 y}{\partial t^2} + v_0 \frac{\partial^2 y}{\partial x \partial t}\right)\left(\rho_0 + \frac{\partial y}{\partial x}\right)^2 - \frac{\partial^2 y}{\partial x \partial t}\left(\frac{\partial y}{\partial t} + v_0 \frac{\partial y}{\partial x}\right)\left(\rho_0 + \frac{\partial y}{\partial x}\right)$$
$$+ \left(i_0 - \frac{\partial y}{\partial t}\right)\left(\frac{\partial^2 y}{\partial x \partial t} + v_0 \frac{\partial^2 y}{\partial x^2}\right)\left(\rho_0 + \frac{\partial y}{\partial x}\right)$$
$$- \left(i_0 - \frac{\partial y}{\partial t}\right)\left(\frac{\partial y}{\partial t} + v_0 \frac{\partial y}{\partial x}\right)\frac{\partial^2 y}{\partial x^2} + \frac{\eta}{\epsilon_0}(y - m)\left(\rho_0 + \frac{\partial y}{\partial x}\right)^3 = 0. \tag{7}$$

The velocity v, i.e., the total velocity minus the initial d-c value, is related to the solution of Equation (7) by

$$v = -\frac{\dfrac{\partial y}{\partial t} + v_0 \dfrac{\partial y}{\partial x}}{\rho_0 + \dfrac{\partial y}{\partial x}}. \tag{8}$$

For calculations of efficiency only a third-order theory is required. Hence, terms of fourth and higher order in Equations (7) and (8) may be neglected without making appreciable errors if the discussion is restricted to powers below and up to saturation level. Under this assumption Equation (7) yields

$$\begin{aligned}
&\frac{\partial^2 y}{\partial x^2} + \frac{2}{v_0}\frac{\partial^2 y}{\partial x \partial t} + \frac{1}{v_0^2}\frac{\partial^2 y}{\partial t^2} + \beta_p{}^2\,(y - m) \\
&\quad + \frac{2}{i_0 v_0}\left(\frac{\partial y}{\partial x}\frac{\partial^2 y}{\partial t^2} - \frac{\partial y}{\partial t}\frac{\partial^2 y}{\partial x \partial t}\right) + \frac{2}{i_0}\left(\frac{\partial y}{\partial x}\frac{\partial^2 y}{\partial x \partial t} - \frac{\partial y}{\partial t}\frac{\partial^2 y}{\partial x^2}\right) \\
&\quad + \frac{3y}{\epsilon_0 v_0{}^2}\frac{\partial y}{\partial x}(y - m) \\
&\quad + \frac{1}{i_0{}^2}\left(\frac{\partial^2 y}{\partial x^2}\left(\frac{\partial y}{\partial t}\right)^2 - 2\frac{\partial^2 y}{\partial x \partial t}\frac{\partial y}{\partial x}\frac{\partial y}{\partial t} + \frac{\partial^2 y}{\partial t^2}\left(\frac{\partial y}{\partial x}\right)^2\right) \\
&\quad + \frac{3\eta}{\epsilon_0 i_0 v_0}\left(\frac{\partial y}{\partial x}\right)^2 (y - m) = 0.
\end{aligned} \tag{9}$$

Here $\beta_p = \omega_p/v_0$ is the plasma-wave number. For very low power levels we can even neglect second-order terms. Under such a condition all terms in Equation (9) except the first four can be omitted; this leads to the well known space-charge-wave equation for small signals.[8,9] We, however, try a third-order approximative solution

$$\begin{aligned}
y &= y_1 + y_2 + y_3 + \cdots \\
m &= m_1 + m_2 + m_3 + \cdots \\
v &= v_1 + v_2 + v_3 + \cdots .
\end{aligned} \tag{10}$$

Again omitting terms of fourth and higher order we get

[8] W. C. Hahn, "Small Signal Theory of Velocity-Modulated Electron Beams," *General Electric Review*, Vol. 42, p. 258, June 1939.

[9] S. Ramo, "Space-Charge and Field Waves in an Electron Beam," *Phys. Rev.*, Vol. 56, p. 276, August 1939.

$$f_1 + f_2 + f_3 = 0, \tag{11}$$

where

$$f_1 = \frac{\partial^2 y_1}{\partial x^2} + \frac{2}{v_0}\frac{\partial^2 y_1}{\partial x \partial t} + \frac{1}{v_0^2}\frac{\partial^2 y_1}{\partial t^2} + \beta_p^2 (y_1 - m_1),$$

$$\begin{aligned} f_2 = {} & \frac{\partial^2 y_2}{\partial x^2} + \frac{2}{v_0}\frac{\partial^2 y_2}{\partial x \partial t} + \frac{1}{v_0^2}\frac{\partial^2 y_2}{\partial t^2} + \beta_p^2 (y_2 - m_2) \\ & + \frac{2}{i_0 v_0}\left(\frac{\partial y_1}{\partial x}\frac{\partial^2 y_1}{\partial t^2} - \frac{\partial y_1}{\partial t}\frac{\partial^2 y_1}{\partial x \partial t}\right) + \frac{2}{i_0}\left(\frac{\partial y_1}{\partial x}\frac{\partial^2 y_1}{\partial x \partial t} - \right. \\ & \left. \frac{\partial y_1}{\partial t}\frac{\partial^2 y_1}{\partial x^2}\right) + \frac{3\eta}{\epsilon_0 v_0^2}\frac{\partial y_1}{\partial x}(y_1 - m_1), \end{aligned}$$

$$\begin{aligned} f_3 = {} & \frac{\partial^2 y_3}{\partial x^2} + \frac{2}{v_0}\frac{\partial^2 y_3}{\partial x \partial t} + \frac{1}{v_0^2}\frac{\partial^2 y_3}{\partial t^2} + \beta_p^2 (y_3 - m_3) \\ & + \frac{2}{i_0 v_0}\left(\frac{\partial y_1}{\partial x}\frac{\partial^2 y_2}{\partial t^2} + \frac{\partial y_2}{\partial x}\frac{\partial^2 y_1}{\partial t^2} - \frac{\partial y_2}{\partial t}\frac{\partial^2 y_1}{\partial x \partial t} - \frac{\partial y_1}{\partial t}\frac{\partial^2 y_2}{\partial x \partial t}\right) \\ & + \frac{2}{i_0}\left(\frac{\partial y_1}{\partial x}\frac{\partial^2 y_2}{\partial x \partial t} + \frac{\partial y_2}{\partial x}\frac{\partial^2 y_1}{\partial x \partial t} - \frac{\partial y_1}{\partial t}\frac{\partial^2 y_2}{\partial x^2} - \frac{\partial y_2}{\partial t}\frac{\partial^2 y_1}{\partial x^2}\right) \\ & + \frac{1}{i_0^2}\left(\frac{\partial^2 y_1}{\partial x^2}\left(\frac{\partial y_1}{\partial t}\right)^2 - 2\frac{\partial y_1}{\partial t}\frac{\partial y_1}{\partial x}\frac{\partial^2 y_1}{\partial x \partial t} + \frac{\partial^2 y_1}{\partial t^2}\left(\frac{\partial y_1}{\partial x}\right)^2\right) \\ & + \frac{3\eta}{\epsilon_0 v_0^2}\left(\frac{\partial y_2}{\partial x}(y_1 - m_1) + y_2\frac{\partial y_1}{\partial x}\right) + \frac{3\eta}{\epsilon_0 v_0 i_0}(y_1 - m_1)\left(\frac{\partial y_1}{\partial x}\right)^2 . \end{aligned}$$

To simplify the analysis considerably we set the terms of equal order equal to zero. This procedure, of course, eliminates some of the possible solutions of Equation (11). However, as will be shown later, the remaining solutions are sufficient to satisfy the boundary conditions in practical cases. The equations are

$$f_1 = 0, \tag{12a}$$

$$f_2 = 0, \tag{12b}$$

$$f_3 = 0. \tag{12c}$$

Similarly we find from Equations (8) and (10) for the velocity terms v_1, v_2, and v_3 the relations

$$-\rho_0 v_1 = \frac{\partial y_1}{\partial t} + v_0 \frac{\partial y_1}{\partial x} \tag{13a}$$

$$-\rho_0 v_2 = \frac{\partial y_2}{\partial t} + v_0 \frac{\partial y_2}{\partial x} + v_1 \frac{\partial y_1}{\partial x} \tag{13b}$$

$$-\rho_0 v_3 = \frac{\partial y_3}{\partial t} + v_0 \frac{\partial y_3}{\partial x} + v_2 \frac{\partial y_1}{\partial x} + v_1 \frac{\partial y_2}{\partial x}. \tag{13c}$$

Equations (12) represent a system of *linear differential equations* and the procedure of computing is now clear; we must first solve Equation (12a), put the solution into (12b), solve this Equation and put the solutions of (12a) and (12b) into Equation (12c) and solve it. The total solutions $y_1 + y_2 + y_3$ and $v_1 + v_2 + v_3$ must satisfy the boundary conditions. This approach to the solution of a nonlinear differential equation is known as the successive-approximation method.

Velocity Modulation by a Sinusoidal R-F Voltage

Let us assume a cavity with very narrow gap at $x = x_0$ modulating a beam of velocity v_{00}. The voltage across the gap is given by

$$V = -V_1 \sin(\omega t + \phi). \tag{14}$$

From the conservation of energy we find, for the total velocity of the beam leaving the gap,

$$\left(\frac{dx}{dt}\right)_{x_0} = v_{00} \sqrt{1 - \frac{2\eta V_1}{{v_{00}}^2} \sin(\omega t + \phi)}. \tag{15}$$

Developing Equation (15) for the case

$$\left|\frac{2\eta V_1}{{v_{00}}^2}\right| \ll 1$$

into a series and omitting again terms of higher than third order, we get

$$\left(\frac{dx}{dt}\right)_{x_0} = v_{00}\left[1 - \left(\frac{\eta V_1}{2v_{00}^2}\right)^2 - \frac{\eta V_1}{v_{00}^2}\left(1 + \frac{3}{2}\left(\frac{\eta V_1}{2v_{00}^2}\right)^2\right)\sin\,(\omega t + \phi)\right.$$

$$\left. + \left(\frac{\eta V_1}{2v_{00}^2}\right)^2 \cos 2\,(\omega t + \phi) + \left(\frac{\eta V_1}{2v_{00}^2}\right)^3 \sin 3\,(\omega t + \phi)\right]. \qquad (16)$$

Currents and Velocities in a Modulated Beam

We assume a velocity-modulating cavity with a very narrow gap at

$$x = x_0 = -\frac{\pi}{2\beta_p}. \qquad (17)$$

Equation (12) must be solved for $m = 0$, i.e., no coupling to the outside and a velocity at $x = x_0$ given by Equation (16). After basic but rather involved algebraic manipulations, the solution for the displacement y is found to be

$$\begin{aligned}
\frac{\omega_{p0}}{i_0}\,y = {} & \frac{\eta V_1}{v_{00}^2}\cos\beta_p x\,\sin\,(\omega t - \beta_e x) \\
& - \frac{1}{4}\left(\frac{\eta V_1}{v_{00}^2}\right)^2 \frac{\omega}{\omega_{p0}}\left[(1 + \cos 2\,\beta_p x)\,\sin 2\,(\omega t - \beta_e x)\right. \\
& \quad + \frac{\omega_{p0}}{\omega}\cos\beta_p x\,\cos 2\,(\omega t - \beta_e x) \\
& \quad \left. + \frac{\omega_{p0}}{\omega}\sin 2\,\beta_p x\,(1 - \cos 2\,(\omega t - \beta_e x))\right] \\
& - \frac{1}{16}\left(\frac{\eta V_1}{v_{00}^2}\right)^3 \left(\frac{\omega}{\omega_{p0}}\right)^2 \left[\frac{1 + 9\left(\dfrac{\omega_{p0}}{\omega}\right)^2}{2}\cos 3\,\beta_p x\,\sin\,(\omega t - \beta_e x)\right. \\
& \quad + \frac{3 - 13\left(\dfrac{\omega_{p0}}{\omega}\right)^2}{2}\cos\beta_p x\,\sin\,(\omega t - \beta_e x) \\
& \quad + 2\left(\frac{\omega_{p0}}{\omega}\right)^2 \sin 2\,\beta_p x\,\sin\,(\omega t - \beta_e x)
\end{aligned}$$

$$+\frac{\omega_{p0}}{\omega}(1+\cos 2\,\beta_p x)\cos(\omega t-\beta_e x)$$

$$\left.-3\frac{\omega_{p0}}{\omega}(\sin 3\,\beta_p x+\sin\beta_p x)\cos(\omega t-\beta_e x)\right]$$

$$+\frac{3}{16}\left(\frac{\eta V_1}{v_{00}{}^2}\right)^3\left(\frac{\omega}{\omega_{p0}}\right)^2\left[\frac{1+\left(\dfrac{\omega_{p0}}{\omega}\right)^2}{2}(\cos 3\,\beta_p x+3\cos\beta_p x)\sin 3\,(\omega t-\beta_e x)\right.$$

$$-\frac{\omega_{p0}}{\omega}(\sin 3\,\beta_p x+\sin\beta_p x)\cos 3\,(\omega t-\beta_e x)$$

$$+\frac{\omega_{p0}}{\omega}(1+\cos 2\,\beta_p x)\cos 3\,(\omega t-\beta_e x)$$

$$+\frac{2}{3}\left(\frac{\omega}{\omega_{p0}}\right)^2\sin 2\,\beta_p x\sin 3\,(\omega t-\beta_e x)$$

$$\left.-2\left(\frac{\omega_{p0}}{\omega}\right)^2\cos\beta_p x\sin 3\,(\omega t-\beta_e x)\right]. \tag{18}$$

The velocity, as derived from Equations (13), is given by

$$\frac{v}{v_{00}}=-\frac{1}{4}\left(\frac{\eta V_1}{v_{00}{}^2}\right)^2+\frac{\eta V_1}{v_{00}{}^2}\sin\beta_p x\sin(\omega t-\beta_e x)$$

$$+\frac{1}{4}\left(\frac{\eta V_1}{v_{00}{}^2}\right)^2\frac{\omega}{\omega_{p0}}\left[-\sin 2\,\beta_p x\sin 2\,(\omega t-\beta_e x)\right.$$

$$-\frac{\omega_{p0}}{\omega}\sin\beta_p x\cos 2\,(\omega t-\beta_e x)$$

$$\left.+\frac{\omega_{p0}}{\omega}(1+\cos 2\,\beta_p x)(1-\cos 2\,(\omega t-\beta_e x))\right]$$

$$-\frac{1}{16}\left(\frac{\eta V_1}{v_{00}{}^2}\right)^3\left(\frac{\omega}{\omega_{p0}}\right)^2\left[\frac{5}{2}\left(1+\left(\frac{\omega_{p0}}{\omega}\right)^2\right)\sin 3\,\beta_p x\sin(\omega t-\beta_e x)\right.$$

$$+\frac{1}{2}\left(5-7\left(\frac{\omega_{p0}}{\omega}\right)^2\right)\sin\beta_p x\sin(\omega t-\beta_e x)$$

$$+\left(3\frac{\omega_{p0}}{\omega}-2\left(\frac{\omega_{p0}}{\omega}\right)^{2}\right)\sin 2\,\beta_p x\cos\,(\omega t-\beta_e x)-2\left(\frac{\omega_{p0}}{\omega}\right)^{2}(1+\cos 2\,\beta_p x)\,\sin\,(\omega t-\beta_e x)$$

$$+\left(5\frac{\omega_{p0}}{\omega}-2\left(\frac{\omega_{p0}}{\omega}\right)^{3}\right)\cos 3\,\beta_p x\cos\,(\omega t-\beta_e x)$$

$$+\left(3\frac{\omega_{p0}}{\omega}+2\left(\frac{\omega_{p0}}{\omega}\right)^{2}\right)\cos\beta_p x\cos\,(\omega t-\beta_e x)$$

$$-\frac{1}{2}\left(7-\left(\frac{\omega_{p0}}{\omega}\right)^{2}\right)\sin 3\,\beta_p x\sin 3\,(\omega t-\beta_e x)-\frac{1}{2}\left(7-5\,\Big)\frac{\omega_{p0}}{\omega}\Big)^{2}\right)\sin\beta_p x\sin 3\,(\omega t-\beta_e x)$$

$$-\left(5\frac{\omega_{p0}}{\omega}-2\left(\frac{\omega_{p0}}{\omega}\right)^{3}\right)(\cos 3\,\beta_p x+\sin 2\,\beta_p x)\cos 3\,(\omega t-\beta_e x)$$

$$-\left(3\frac{\omega_{p0}}{\omega}+2\left(\frac{\omega_{p0}}{\omega}\right)^{3}\right)\cos\beta_p x\cos 3\,(\omega t-\beta_e x)+2\left(\frac{\omega_{p0}}{\omega}\right)^{2}(1+\cos 2\,\beta_p x)\sin 3\,(\omega t-\beta_e x)\Bigg]. \tag{19}$$

The boundary conditions at the excitation plane are met if ϕ in Equations (14)-(16) is set equal to

$$\phi=\frac{\pi}{2}\frac{\omega}{\omega_p}. \tag{20}$$

The plasma-frequency ω_{p0} in the amplitude terms of Equations (18) and (19) is the value before the beam enters the modulating cavity. Behind the cavity the plasma frequency is slightly changed and depends upon the amplitude because of the relation between d-c velocity and amplitude given by Equation (16).

Beside the well known low-level waves[8,9] with the propagation constants

$$\gamma=\frac{\omega\pm\omega_p}{v_0}, \tag{21}$$

we have at large signals additional waves with the propagation constants

$$\gamma = \frac{\omega}{v_0}, \quad \frac{\omega \pm 2\omega_p}{v_0}, \quad \frac{\omega \pm 3\omega_p}{v_0} \tag{22}$$

for the fundamental frequency;

$$\gamma = \frac{2\omega}{v_0}, \quad \frac{2\omega \pm \omega_p}{v_0}, \quad \frac{2\omega \pm 2\omega_p}{v_0} \tag{23}$$

for the first harmonic; and

$$\gamma = \frac{3\omega}{v_0}, \quad \frac{3\omega \pm \omega_p}{v_0}, \quad \frac{3\omega \pm 2\omega_p}{v_0}, \quad \frac{3\omega \pm 3\omega_p}{v_0} \tag{24}$$

for the second harmonic.

In most practical cases we may neglect high-order-space-charge terms in Equations (18) and (19) since

$$\frac{\omega_{p0}}{\omega} \ll 1, \tag{25}$$

and

$$\left(\frac{\eta V_1}{v_{00}^2}\right)^2 \ll 1. \tag{26}$$

Under these assumptions we now compute the current density which is simply the negative time derivative of the displacement y:

$$\begin{aligned}
\frac{i}{i_0} = & -\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}} \cos \beta_p x \cos (\omega t - \beta_e x) \\
& + \frac{1}{2}\left(\frac{\eta V_1}{v_{00}^2}\right)^2 \left(\frac{\omega}{\omega_{p0}}\right)^2 (1 + \cos 2\,\beta_p x) \cos 2\,(\omega t - \beta_e x) \\
& + \frac{1}{32}\left(\frac{\eta V_1}{v_{00}^2}\right)^3 \left(\frac{\omega}{\omega_{p0}}\right)^3 (\cos 3\,\beta_p x + 3 \cos \beta_p x) \cos (\omega t - \beta_e x) \\
& - \frac{9}{32}\left(\frac{\eta V_1}{v_{00}^2}\right)^3 \left(\frac{\omega}{\omega_{p0}}\right)^3 (\cos 3\,\beta_p x + 3 \cos \beta_p x) \cos 3\,(\omega t - \beta_e x).
\end{aligned} \tag{27}$$

By taking into account Expressions (25) and (26), Equation (19) yields the following simplified equation for the velocity:

$$\frac{v}{v_{00}} = \frac{\eta V_1}{{v_{00}}^2} \sin \beta_p x \sin (\omega t - \beta_e x)$$
$$- \frac{1}{4} \left(\frac{\eta V_1}{{v_{00}}^2} \right)^2 \frac{\omega}{\omega_{p0}} \sin 2\, \beta_p x \sin 2\, (\omega t - \beta_e x)$$
$$- \frac{5}{32} \left(\frac{\eta V_1}{{v_{00}}^2} \right)^3 \left(\frac{\omega}{\omega_{p0}} \right)^2 (\sin 3\, \beta_p x + \sin \beta_p x) \sin (\omega t - \beta_e x)$$
$$+ \frac{7}{32} \left(\frac{\eta V_1}{{v_{00}}^2} \right)^3 \left(\frac{\omega}{\omega_{p0}} \right)^2 (\sin 3\, \beta_p x + \sin \beta_p x) \sin 3\, (\omega t - \beta_e x). \qquad (28)$$

Figure 2 is a plot of the absolute values of current density and velocity of the fundamental frequency versus distance for different power levels. The absolute maximum of current density (saturation) appears at $x = 0$ for a power level given by

$$\frac{\eta V_{1\text{sat}}}{{v_{00}}^2} \frac{\omega}{\omega_{p0}} = \sqrt{\frac{8}{3}}. \qquad (29)$$

The first set of curves corresponds to small signals, the third to saturation level. The fourth curve shows the amplitude distribution for higher than saturation level; the validity of this curve is doubtful because of the neglect of higher than third-order terms. Figure 3 shows the amplitudes of the first two harmonics versus distance. Again the limitation to third- and lower-order term does not allow us to predict the saturation behavior of the harmonics.

The question arises as to how large the power levels can be and still have Equations (27) and (28) represent good approximations of the exact solutions. To extrapolate to higher harmonics let us consider the "quasi-linear" case, which means neglect of saturation effects but not of the harmonic generation. The maximum current density (Equation (27)) appears at $x = 0$ and for quasi-linearity is given by

$$\frac{i}{i_0} = - \frac{\eta V_1}{{v_{00}}^2} \frac{\omega}{\omega_{p0}} \cos \omega t + \left(\frac{\eta V_1}{{v_{00}}^2} \frac{\omega}{\omega_{p0}} \right)^2 \cos 2\, \omega t - \frac{9}{8} \left(\frac{\eta V_1}{{v_{00}}^2} \frac{\omega}{\omega_{p0}} \right)^3 \cos 3\, \omega t. \qquad (30)$$

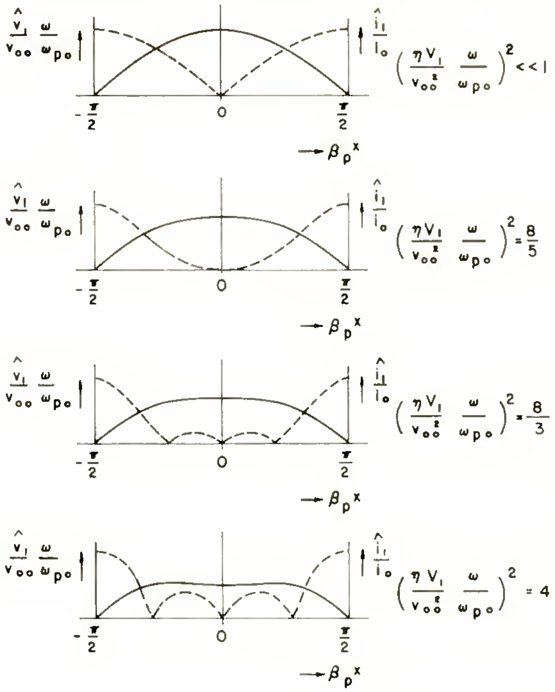


Fig. 2—Amplitudes of the fundamental current density (solid) and velocity (dashed) for different power levels versus distance.

This equation seems to be the first part of the series

$$\frac{i}{i_0} = \sum_{n=1}^{\infty} 2 \frac{\left(-\frac{n}{2}\right)^n}{n!} \left(\frac{\eta V_1}{v_{00}^2} \frac{\omega}{\omega_{p0}}\right)^n \cos n\omega t. \tag{31}$$

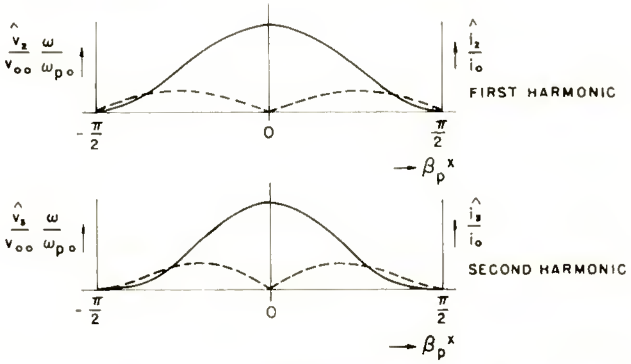


Fig. 3—Amplitudes of the harmonic current densities (solid) and velocities (dashed) versus distance.

Figure 4 shows this spectrum for different power levels. The series is convergent only if

$$\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}} < \frac{2}{e} \doteq 0.736. \tag{32}$$

Thus we may conclude that Equation (27) is a valid approximation for the *total* current density only for power levels well below the limiting value given by Equation (32). This does not necessarily mean that the first terms of the power series for the *fundamental* current at $x = 0$ which can be obtained from Equation (27),

$$\left(\frac{i_1}{i_0}\right)_{x=0} = -\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}}\left(1 - \frac{1}{8}\left(\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}}\right)^2 + \cdots\right)\cos\omega t, \tag{33}$$

are applicable only for power levels given by Equation (32). This series seems to converge much more rapidly than Equation (31). At saturation level (Equation (29)), for example, the second term is 1/3 of the first term. It should be noted that Equation (33) consists of the first two terms of the Bessel function of first order. This vague extrapolation

$$\left(\frac{i_1}{i_0}\right)_{x=0} = -2J_1\left(\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}}\right)\cos\omega t = -\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}} \tag{34}$$
$$\left[1 - \frac{1}{8}\left(\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}}\right)^2 + \frac{1}{192}\left(\frac{\eta V_1}{v_{00}^2}\frac{\omega}{\omega_{p0}}\right)^4 - \cdots\right]\cos\omega t$$

leads to a relation which is *formally* the same as the well-known equation of the ballistic theory,[1]

$$\frac{i_1(z)}{i_0} = 2J_1\left(\frac{\eta V_1}{v_{00}^2}\frac{\omega z}{v_{00}}\right)\cos\omega t, \tag{35}$$

where z is the distance from the plane of excitation;

$$z = x + \frac{\pi}{2\beta_p}. \tag{36}$$

This equivalence explains the good agreement between some experiments and ballistic theory.[2]

To recapitulate, the *total current density or velocity* can be computed from Equations (18) and (19) but only for levels well below the limit given by Equation (32). We can, however, predict the behavior of the current and velocity of the *fundamental frequency* up to saturation level.

Comparison with the Ballistic Theory

Webster's ballistic theory[1] predicts convection currents in a velocity-modulated beam which are given by

$$\frac{i}{i_0} = 2 \sum_{n=1}^{\infty} J_n \left(\frac{n\eta V_1}{v_0^2} \beta_e z \right) \cos n\,(\omega t + \phi). \qquad (37)$$

This result is valid for negligible space charge and our theory should agree in this case with Webster's. Shifting our coordinate system by the transformation of Equation (36) we get from Equation (27)

$$\begin{aligned} \frac{i}{i_0} = {} & \frac{\eta V_1}{v_0^2} \frac{\omega}{\omega_p} \sin \beta_p z \cos\,(\omega t - \beta_e x) \\ & + \left(\frac{\eta V_1}{v_0^2} \frac{\omega}{\omega_p} \right)^2 \sin^2 \beta_p z \cos 2\,(\omega t - \beta_e x) \\ & + \frac{1}{8} \left(\frac{\eta V_1}{v_0^2} \frac{\omega}{\omega_p} \right)^3 \sin^3 \beta_p z \cos\,(\omega t - \beta_e x) \\ & - \frac{9}{8} \left(\frac{\eta V_1}{v_0^2} \frac{\omega}{\omega_p} \right)^3 \sin^3 \beta_p z \cos 3\,(\omega t - \beta_e x). \end{aligned} \qquad (38)$$

In Equation (38) the plasma frequencies ω_p and ω_{p0} were considered equal because Equation (37) is based on the assumption of a pure sinusoidal velocity modulation with the d-c velocity unchanged. For

$$\beta_p z \ll \frac{\pi}{2}, \qquad (39)$$

the space-charge force certainly is very small since there is no bunch formed yet. In this case we get from Equation (38)

$$\frac{i}{i_0} = -\frac{\eta V_1}{v_0^2}\beta_e z\left(1-\frac{1}{8}\left(\frac{\eta V_1}{v_0^2}\beta_e z\right)^2\right)\cos\,(\omega t-\beta_e x)$$
$$+\left(\frac{\eta V_1}{v_0^2}\beta_e z\right)^2\cos 2\,(\omega t-\beta_e x)-\frac{9}{8}\left(\frac{\eta V_1}{v_0^2}\beta_e z\right)^3\cos 3\,(\omega t-\beta_e x). \tag{40}$$

The amplitudes of the currents of frequencies $n\omega$ are given by

$$n=1:\quad \left|\frac{\hat{\imath}_1}{i_0}\right| = \frac{\eta V_1}{v_0^2}\beta_e z\left(1-\frac{1}{8}\left(\frac{\eta V_1}{v_0^2}\beta_e z\right)^2\right), \tag{41a}$$

$$n=2:\quad \left|\frac{\hat{\imath}_2}{i_0}\right| = \left(\frac{\eta V_1}{v_0^2}\beta_e z\right)^2, \tag{41b}$$

$$n=3:\quad \left|\frac{\hat{\imath}_3}{i_0}\right| = \frac{9}{8}\left(\frac{\eta V_1}{v_0^2}\beta_e z\right)^3. \tag{41c}$$

From Equation (37) on the other hand, by developing the Bessel-functions into a power series for small arguments,

$$\left|\frac{\hat{\imath}_n}{i_0}\right| = 2\,\frac{\left(\dfrac{n\eta V_1}{2v_0^2}\beta_e z\right)^n}{n!}\left(1-\frac{\left(\dfrac{n\eta V_1}{2v_0^2}\beta_e z\right)^2}{n+1}+\cdots\right). \tag{42}$$

As expected, the Equations (41) and (42) are in agreement. According to measurements of Gittins and Sullivan[3] the neglect of space charge with any appreciable beam current is not justifiable except for positions close to the excitation plane where Equation (39) holds.

Two-Cavity Klystron

We assume a narrow-gap cavity at $x=-\pi/(2\beta_p)$ which produces a velocity modulation given by Equation (16). At the output cavity, which is assumed to be at $x=0$, the current density of the fundamental frequency is given by Equation (33). We now introduce the beam voltage

$$V_0=\frac{v_{00}^2}{2\eta}. \tag{43}$$

Saturation of the current density is reached at

$$\frac{V_{1\,\mathrm{sat}}}{2V_0}\frac{\omega}{\omega_{p0}}=\sqrt{\frac{8}{3}},\tag{44}$$

where

$$|\hat{\imath}_{1\,\mathrm{sat}}|=2i_0\left(\frac{2}{3}\right)^{3/2}.\tag{45}$$

In order to prevent reflection of electrons, the Q of the second cavity must be adjusted so that the r-f gap-voltage does not exceed the d-c beam voltage. For maximum output the gap voltage equals the d-c voltage, and the saturation power delivered from the beam to the cavity is given by

$$P_{2\,\mathrm{sat}}=\frac{1}{2}\,|\hat{\imath}_{1\,\mathrm{sat}}|\,SV_0.\tag{46}$$

Here S is the area of the cross section of the beam. The electronic saturation efficiency is given by

$$\eta_{\mathrm{el}}=\frac{P_{2\,\mathrm{sat}}}{i_0\,SV_0}=\frac{1}{2}\,\frac{|\hat{\imath}_{1\,\mathrm{sat}}|}{i_0}.$$

From Equation (45),

$$\eta_{\mathrm{el}}=\left(\frac{2}{3}\right)^{3/2}\doteq 54.4\%.\tag{47}$$

If we use the vaguely extrapolated Equation (34), the calculated saturation efficiency agrees with the result of the ballistic theory, i.e.,

$$\eta_{\mathrm{el}}=58.2\%.\tag{48}$$

In any case we can conclude that the space charge or the so called "longitudinal debunching" has no appreciable influence on the electronic efficiency. This is in agreement with experimental results.[3] For other data, however, such as velocity- and current-modulation or gain, the ballistic theory only holds near the plane of excitation as previously discussed.

If the equivalent gap resistances R_1 and R_2 of the input and output cavities are known, the gain of the klystron can be calculated. The input power, P_1, is given by

$$P_1 = \frac{V_1^2}{2R_1}, \tag{49}$$

and the output power by

$$P_2 = \frac{|\hat{i}_1|^2 S^2}{2} R_2 . \tag{50}$$

Equations (33), (43), (49), and (50) yield an expression for the gain,

$$G = \frac{P_2}{P_1} = \frac{R_1 R_2}{4\left(\dfrac{V_0}{I_0}\right)^2} \left(\frac{\omega}{\omega_{p0}}\right)^2 \left[1 - \frac{1}{8}\left(\frac{V_1}{2V_0} \frac{\omega}{\omega_{p0}}\right)^2\right]^2, \tag{51}$$

where I_0 is the d-c beam current,

$$I_0 = i_0 S. \tag{52}$$

Thus the small-signal gain is given by

$$G_{ss} = \frac{1}{4} \frac{R_1 R_4}{\left(\dfrac{V_0}{I_0}\right)^2} \left(\frac{\omega}{\omega_{p0}}\right)^2 \tag{53}$$

and the saturation gain (from Equation (44)) by

$$G_{sat} = \frac{1}{9} \frac{R_1 R_2}{\left(\dfrac{V_0}{I_0}\right)^2} \left(\frac{\omega}{\omega_{p0}}\right)^2 . \tag{54}$$

From Equations (53) and (54) one can see that the ratio of saturation gain to low-signal gain is 4/9, which corresponds to a difference of 3.52 decibels.

Because of the limitation of Equation (32), the harmonic output can be predicted only for levels far below saturation. Figure 4 shows the spectrum for three different power levels which is given by Equation (31). While in the ballistic theory all the harmonic-current amplitudes depend linearly upon the direct current (Equation (37)), in the theory presented here a relation

$$|\hat{i}_n|_{x=0} \propto i_0^{\,1-\frac{n}{2}} \tag{55}$$

is predicted. The fundamental current is proportional to the square root of the direct current, as is well known. The first harmonic output is predicted to be independent of the direct current whereas the higher harmonic amplitudes should *decrease* with increasing direct current. This is understandable because the smaller the d-c beam current for a given r-f power level the more nonlinear the system becomes. The electron stream acts like a linear system if the r-f current is very much smaller than the direct current.

Kinetic Power Flow

The third-order solution makes it possible to clarify the question of kinetic-power flow in electron streams. This concept was the subject of a number of papers.[10-15] According to References (14) and (12), it is a consequence of the conservation of energy that

$$\operatorname{div} \overline{P_K} = \overline{iE}, \tag{56}$$

where $\overline{P_K}$ is the time average of the kinetic-power flow P_K which, in the most general form,[14] is given by

$$P_K = \frac{1}{2\eta} (i_0 + i) \ (v_0 + v)^2. \tag{57}$$

Using Equations (18), (19), (43), and (57),

[10] H. W. König, "Selbsterregung von Triodenschaltungen im Ultrakurzwellengebiet," *Wissensch. Veröff. aus den Siemens-Werken*, Vol. 20, p. 10, 1941.

[11] L. J. Chu, "A Kinetic Power Theorem," 1951 I.R.E. Conference on Electron Devices, Durham, N. H.

[12] W. H. Louisell and J. R. Pierce, "Power Flow in Electron Beam Devices," *Proc. I.R.E.*, Vol. 43, p. 425, April 1955.

[13] L. R. Walker, "Power Flow in Electron Beams," *Jour. Appl. Phys.*, Vol. 26, p. 1031, August 1955.

[14] H. Bauer, "Der elektrokinetische Leistungsfluss und die Eigenschaften passiver Achtpole," *Die Telefunken-Röhre*, Vol. 33, p. A79, October, 1956.

[15] H. W. König, "Kinetische Energiedichte und kinetischer Leistungsfluss in Elektronenströmungen," *Oesterreichisches Ingenieur Archiv.*, Vol. 10, p. 221, 1956.

$$\overline{P}_K = \frac{i_0 V_0}{8} \left(\frac{V_1}{V_0} \right)^2 \cos^2 \beta_p x. \tag{58}$$

In this case the d-c term

$$P_{K0} = i_0 V_0$$

is of no interest and is therefore omitted. Equation (56) can be checked by computing $\overline{iE}$ from Equations (4) and (5) $(m = 0, M = 0)$;

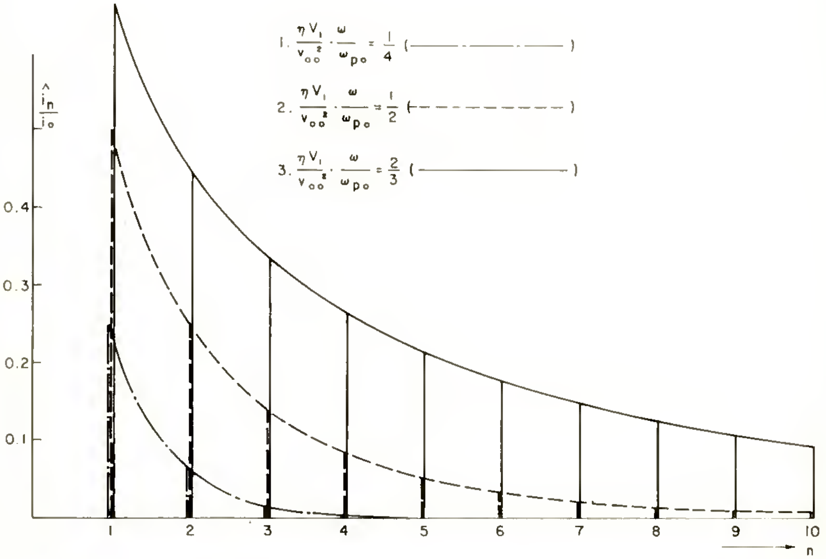


Fig. 4—Frequency spectrum of the convection-current density in a quasi-linear electron beam.

$$\overline{iE} = \frac{i_0}{\epsilon_0} \overline{y} - \frac{1}{2\epsilon_0} \frac{\overline{\partial (y^2)}}{\partial t} = \frac{i_0}{\epsilon_0} \overline{y}. \tag{59}$$

This equation yields a nonzero value for $\overline{iE}$ only if there is a d-c term of the displacement, hence, a d-c term of the electric field. Indeed, we find from Equations (18) and (59)

$$\overline{iE} = -\frac{i_0 V_0}{8} \left(\frac{V_1}{V_0} \right)^2 \beta_p \sin 2\beta_p x. \tag{60}$$

From Equation (58),

$$\operatorname{div} \overline{P}_k = \frac{\partial \overline{P}_k}{\partial x} = -\frac{i_0 V_0}{8} \left(\frac{V_1}{V_0} \right)^2 \beta_p \sin 2\beta_p x, \tag{61}$$

which agrees with Equations (56) and (60). The time-average value of *iE* in Equation (60) is caused by space charge, and originates from a *second-order* term of the displacement *y*. Thus the calculation of kinetic power flow by use of only the linear solution is *not* justifiable, even at arbitrarily low levels. For ion neutralization a d-c term of the electric field cannot exist. In such a case we have to omit the d-c term in Equation (18) and the kinetic power flow is *zero*. These results are in agreement with those of Walker.[13]

The physical explanation for the maximum of kinetic-power flow

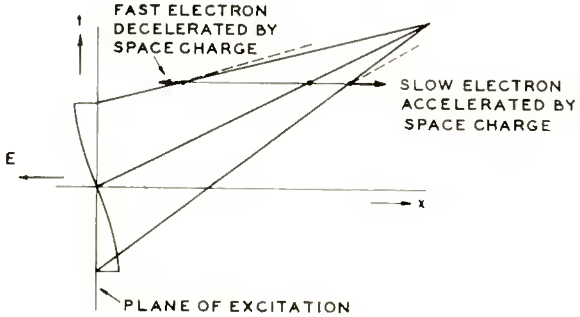


Fig. 5—Electron-time table.

at $x = 0$ is the increased average velocity of the beam at that point due to space-charge effects. In the "electron-time table" (Figure 5) the ballistic paths of three representative electrons are plotted, one passing the excitation gap at maximum r-f field (fast electron), one at zero, and one at minimum r-f field (slow electron). Because of the higher charge density, the space-charge forces are obviously much more efficient for the slow electrons. Hence, the slow electrons are accelerated by the space charge more than the fast electrons are decelerated. The net effect in the equilibrium state of the plasma oscillation at $x = 0$ must be an increase in average velocity. Indeed, Equation (19) yields the average velocity

$$\frac{v_{00} + \overline{v}}{v_{00}} = 1 + \frac{1}{4} \left(\frac{\eta V_1}{v_{00}{}^2} \right)^2 \cos 2\beta_p x, \tag{62}$$

which has a maximum at $x = 0$. The value at excitation plane ($\beta_p x = -\pi/2$) agrees, of course, with the boundary condition given by Equation (16).

Note on the Nonlinear Behavior of Traveling-Wave Tubes

The system of linear differential Equations (12) allows us to calculate also the nonlinear behavior of traveling-wave tubes.

If we make the simplifying assumption that at the harmonic frequencies there is no coupling between beam and delay line, it is obvious from Equations (12) that the nonlinear behavior of traveling-wave tubes can be described by the small-signal parameters, since the small-signal solution y_1 determines also the higher-order solutions y_2 and y_3.

This seems to be in contradiction to the theory of Tien, Walker, and Wolontis[5] who found that, at least for high levels, the plasma frequency *reduction factor* R must be added as a new parameter. It has, however, been shown that this parameter also appears at low levels because it determines the coupling of the space-charge wave to the delay line.[7]

Acknowledgment

The author is greatly indebted to S. Bloom for reviewing the manuscript of this paper.

A SURVEY OF METHODS USED TO DETERMINE CONTACT POTENTIALS IN RECEIVING TUBES

BY

EDWARD R. SCHRADER

RCA Electron Tube Division,
Harrison, N. J.

Summary—*This paper discusses the nature of contact potential and the effects of changes in grid–cathode contact potential on "bias shift" in receiving tubes. It describes the method of determining true grid–cathode contact potential by extrapolation of portions of the grid-characteristic curve, as well as two direct methods of measurement which are generally used in preference to the extrapolation method because of their relative simplicity. Although these direct methods do not indicate true contact potential, they can, in many cases, provide information of considerable value in bias-shift studies. The relationships among the three methods and the application for which each is best suited are discussed.*

INTRODUCTION

A PROBLEM which frequently confronts manufacturers and large-scale users of receiving tubes is a variation in the grid-bias requirements of an amplifier type. This variation (called bias shift) may occur from lot to lot, among tubes in a particular lot, or during the life of an individual tube. Although normally of small magnitude, this variation can cause a substantial change in a major characteristic, particularly when it occurs in a high-mu triode or a high-transconductance type. It is particularly objectionable to the equipment manufacturer because it may necessitate frequent modification of circuit design to change the value of an applied bias, or the use of a compromise value of bias which limits the performance obtainable from the equipment.

It is more or less customary to attribute bias shifts to changes in contact potential (the difference in potential developed between dissimilar metals which are in direct mechanical or electrical contact). This difference of potential is created by, and equal to, the difference in the work functions of the metals, and its polarity is such that the metal having the higher work function is negative with respect to the other. In the case of a tube, the electrode having the lowest work function is usually the cathode. Consequently, the contact potential between the cathode and grid of a triode amplifier tube is equivalent to an applied negative grid bias, and changes in the work function of either the cathode or the grid will cause corresponding changes in the effective grid bias. Such changes in work function may result from manufacturing variations in the composition of the grid and cathode materials, or from contamination of the grid by the cathode during the tube manufacturing process or the life of the tube.

The evaluation of true contact potentials in tubes is complicated by the fact that the resulting bias voltages can be affected by the presence of negative grid currents due to gas, grid emission, and leakage. Because there is no method of direct measurement which will eliminate the effects of all of these contributory factors, true contact potential cannot be measured in terms of the equivalent bias voltage. Nevertheless, a number of methods for the determination of "contact potential" (so-called) have been developed and are used throughout the industry. Some of these methods are based on direct measurement of grid-bias voltages or grid currents and, therefore, include the effects not only of true contact potential but also of potentials due to other factors. This paper describes the various methods now in use and their advantages, shortcomings, and proper applications. In the interest of standardization, it also suggests a system of nomenclature and symbolization which indicates the character of the results obtained by each of the methods described and the conditions under which the results are obtained.

WORK FUNCTION

Before proceeding with a discussion of contact potential, it is desirable to review the concept of work function. An electron a very short distance from the surface of a conductor in a vacuum is subjected to a combination of image and quantum forces which tend to prevent its escape from the parent metal.[1a,2a,3a] Minor quantum effects (such as "tunnel" processes) which permit passage of a few electrons through the work-function barrier will be neglected here. The electron may pass through the surface of the metal, but it will return unless it possesses enough energy to overcome the restraining forces. This behavior is comparable to that of a ball which is propelled up an incline, and which will roll back if it has not been pushed hard enough to make it go over the top. When a metal is heated, the internal electrons partake of the thermal energy in varying amounts, some getting

[1] G. Herrmann and S. Wagener, *The Oxide-Coated Cathode*, Vol. II, Chapman & Hall Ltd., London, England, 1951: (a) p. 13; (b) p. 168; (c) p. 192; (d) pp. 32, 188; (e) p. 190; (f) p. 34.

[2] E. L. Chaffee, *Theory of Thermionic Vacuum Tubes*, McGraw-Hill, N. Y., 1933: (a) p. 56; (b) p. 80.

[3] W. B. Dow, *Fundamentals of Engineering Electronics*, John Wiley Inc., N. Y., 1952: (a) p. 218; (b) p. 203.

enough to permit them to escape into the vacuum. The work function of a metal (the equivalent of the height of the incline) is the total energy which an electron at the Fermi level must possess to escape completely and is usually expressed in electron volts. A plot of the distance out from the surface to which electrons having given amounts of initial energy will go is called an "energy profile" diagram. Such a diagram is shown in Figure 1.

The energy profile representing the work function of an emitter is altered by the presence of an applied electric field between the emitter and an adjacent electrode. A difference in work function alone between an emitter and an electrode with which it is in electrical contact is equivalent to an applied field. Under such conditions the energy profile may not have the form shown in Figure 1, but climbs or dips in accordance with the sign of the applied field. These effects are illustrated in Figure 2. In Figure 2a, curve I shows the retarding field produced when the emitter is in electrical contact with a collector electrode having a larger work function. Curves II and III indicate the effect of progressively increasing anode currents which superimpose a space-charge field on the retarding field. In curve II the additional effect of the space-charge field is not sufficient to interfere with the flow of electrons to the collector. In curve III, however, the current is assumed high enough to produce a space-charge field sufficient to limit the electron flow. Figures 2b and 2c show the effects on the energy profile of applied potentials having the polarities indi-

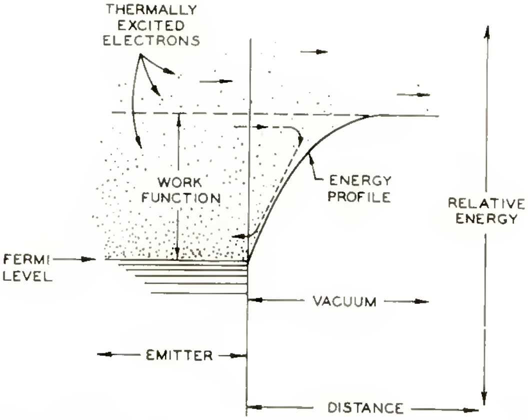


Fig. 1—Energy-profile diagram of a thermionic emitter in vacuum, in the absence of an applied field. Only electrons having energy levels higher than that indicated by the dashed line can escape; all others return to the emitter.

cated. Curves II and III in both figures show the effect of the presence of an increasing number of electrons in the space between the electrodes, curve III, in each case, showing space-charge limitation. Note that if the collector is sufficiently positive with respect to the emitter so that the current is limited only by the emitter work function, the resulting field is termed a "saturating" or "temperature-limiting" field (Figure 2c, curves I and II).

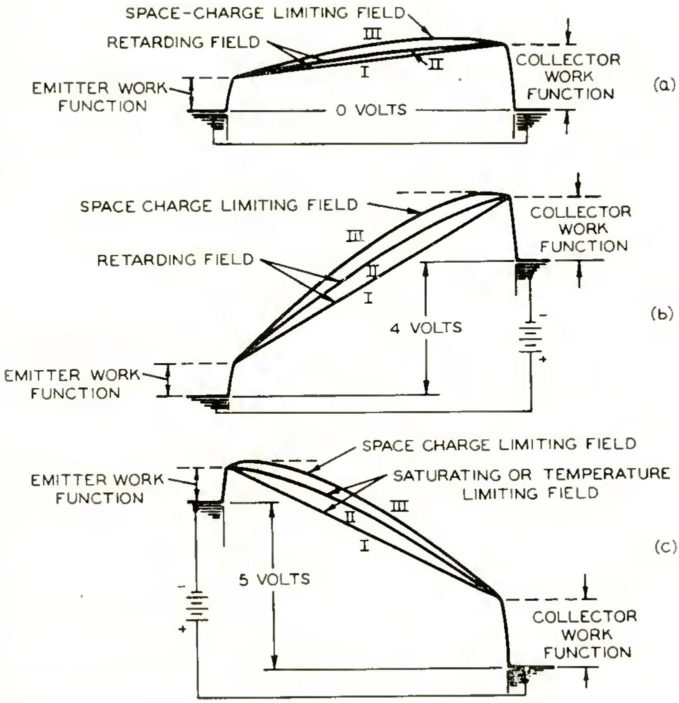


Fig. 2—Typical energy-profile diagrams of a thermionic cathode in the presence of a collector electrode having a higher work function, showing the effects of various applied voltages.

What has been said thus far applies to pure metal emitters but not necessarily to oxide-coated cathodes which are the types most commonly used in receiving tubes. The mechanism of emission from the oxide cathode is more complex than that from a simple metal and is not yet fully understood.[1b] The work function of an oxide is a function of current[4a] and is believed to depend on temperature as well.[1c,4b] For most measurements on commercial electron tubes, however, the cur-

[4] L. S. Nergaard, "Studies of the Oxide Cathode," *RCA Review*, Vol. 13, December, 1952: (a) pp. 512, 520; (b) p. 496.

rent and temperature effects can be considered second-order effects. In this discussion of contact potential it is assumed that the simplified picture of work function applies. It is also assumed that the contact potential of interest is that between the cathode and the adjacent electrode.

Methods of Measurement

The methods used to determine contact potential fall in three general categories. In one, a potential is determined graphically by extrapolation of two well defined regions of the grid-current–grid-voltage characteristic curve. This procedure yields what will be considered true contact potential. In the second method, a potential is determined from the applied voltage required to produce a chosen value of collector current. In the third, a potential is determined from the reading on a high-impedance voltmeter connected between grid and cathode.

The oxide cathode, being a semiconductor, exhibits a thermovoltaic effect which modifies the contact potential and must be corrected for in exacting work. The magnitude of this correction is of the order of 0.003 volt per degree of temperature differential across the coating.[1d] There is also a small effective voltage due to the initial velocity of the electrons.[3b]

The three methods and their significance are described and an appropriate title and symbol suggested for each.

Method 1 — Intersection Method

In this method the cathode is heated to a temperature in the neighborhood of 600°-700°K by operation of the heater at 24 per cent of its rated voltage (for example, 1.5 volts for a 6.3-volt heater). To minimize the effect of the plate (assuming a triode), the plate is tied to the cathode. The logarithm of the grid current is then plotted as a function of the applied grid voltage. The resulting grid-characteristic curve usually contains three clearly defined regions representing the effects of the retarding, space-charge, and saturating fields.

The technique by which such a curve may be used to determine true contact potential is illustrated in Figure 3. Assume that the cathode has a work function of 1 volt. Figure 3a gives the energy-versus-distance configuration corresponding to an applied grid voltage of −0.25 volt, resulting in an assumed grid current of 0.001 microampere. Even though the grid current is quite small the electrons form a space-charge field, with the result that the potential distribution between the grid and cathode is not linear. The space-charge field, however, does not form a barrier higher than that of the grid work function, so that the latter plus the applied field are the only factors limiting the current. Only electrons having thermal energies greater than 1.25 electron volts can reach the grid to make up the 0.001 microampere of current.

Figure 3b shows the profile for zero applied voltage, i.e., the Fermi levels of the cathode and grid have the same energy value. Under these conditions the only retarding field affecting the electrons is that produced by the difference in work function, i.e., by the true contact potential. The grid, therefore, accepts some lower-energy electrons, with the result that its current increases. When the grid is made positive with respect to the cathode (Figure 3c), the profile approaches the horizontal, and at a certain point the space-charge field begins to become important in limiting the current. Up to this point the grid characteristic plot is a straight line, because in the retarding-potential region the collector current from an oxide cathode, like that from a pure metal emitter, is generally a logarithmic function of applied voltage as predicted by theory.[1d] The interposition of the space-charge barrier at this point, however, tends to reduce the current, with the result that the curve starts to round off (Figure 3d).

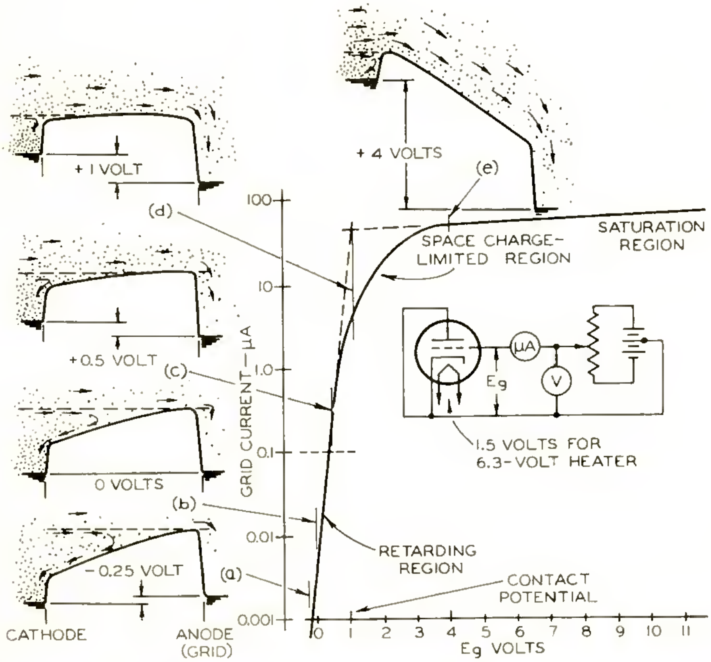


Fig. 3—Diagram showing the principal regions of a grid-characteristic curve, and determination of true contact potential by the intersection method.

Figure 3e shows that if the grid is made sufficiently positive, both the grid work function and the space-charge barrier no longer have a limiting effect on the current and all of the electrons emitted are collected. However, the current in this saturation or temperature-limited region of the grid characteristic curve does not remain constant, but increases slightly with applied voltage due to a slight change of the cathode work function by the applied field.

If no space-charge barrier were present, the retarding-field region of the curve would change abruptly into the saturation region at the applied grid voltage which sets the upper rims of the cathode and grid work functions at the same level (Figure 3d). This applied voltage is numerically equal to the difference in work functions and, therefore, is the true contact potential which we are trying to determine. Because there is always a rounding of the curve due to space charge, this voltage, E_{CP}, is given by the intersection of the extrapolations of the retarding and saturation regions of the curve. In circuit analysis E_{CP} must be preceded by a minus sign because it acts in the same manner and direction as a negative grid bias.

Although the intersection method is the best one available for determining true contact potential in commercial receiving tubes, it has several shortcomings. (1) To avoid damage by saturation currents the curve must be taken at low cathode temperature. Because of the dependence of the cathode work function on temperature, it is to be expected that the contact potential would be different at operating temperature. (2) Resistive layers on the cathode base metal and on the grid may affect the distribution of the applied voltage, thereby distorting the grid-current curve.[5] Consequently, even the retarding-field region of the curve may be rounded, making extrapolation difficult and leading to erroneous results. The curve can also be rounded by nonuniformity of the cathode work function, as evidenced by variations in emission density over the surface of the oxide coating.[1e] Some authorities believe that the shape of the curve may also be influenced by the porosity and roughness of the oxide surface. Fortunately, these adverse factors are not significant in most tubes and the values obtained when precautions are taken to assure clean electrode surfaces show good agreement with the expected contact potentials. (3) The intersection method yields true contact potential only for plane-parallel electrodes since the slope of the retarding portion of the curve varies with the electrode geometry.[1f, 2b] Fortunately, most tubes approach

[5] G. C. Dalman, "Effects of Cathode and Anode Resistance on the Retarding Potential Characteristics of Diodes," *Jour. Appl. Phys.*, Vol. 25, p. 1263, October, 1954.

the plane-parallel configuration closely enough so that, to a first approximation, this consideration may be ignored.

Method 2—Potential Producing an Arbitrary Reference Current

An applied potential which produces a certain chosen reference current in the retarding-field region of the grid-characteristic curve is called a retarding potential, and is symbolized E_{RC}, the subscript C indicating a relatively cool cathode (one operated at 24 per cent of the rated heater voltage). The reference current chosen should be low enough to assure that the corresponding applied potential is truly within the retarding-field region. A suitable value is 0.1 microampere. From Figure 3 it can be seen that the retarding potential which gives +0.1 microampere is determined only by the magnitude of the grid work function, the cathode work function becoming significant only in the saturation region. Differences in values of E_{RC}, therefore, reflect differences in grid work functions, and are equal to differences in true contact potential only if the cathode work function can be assumed to be constant.

If the +0.1 microampere reference current is outside the true retarding-field region because of poor cathode activation or large voltage drops across electrode resistances, differences in E_{RC} may not truly indicate differences in grid work functions. Consequently, the smaller the reference current that can reliably be measured (assuming no leakage currents are present), the more accurately will changes in retarding potential indicate changes or variations in grid work function. It is especially important that measurements of retarding potential made for comparative purposes be made at the same cathode temperature. Because increases in cathode temperature raise all portions of the grid-current–grid-voltage curve along the current axis, they have little effect on the contact potential but substantial effects on the voltage necessary to produce a given grid current. An increase in cathode temperature, therefore, results in a reduction of the value of E_{RC}, i.e., E_{RC} becomes less positive, or more negative.

Because retarding potentials lie well below the saturation region, they may be measured at normal cathode temperatures. Under these conditions it is necessary to apply a negative voltage to the grid to retard the majority of the available electrons. Figure 4 shows the location of the retarding potential point on a conventional grid-current–grid-voltage curve, the subscript H indicating a "hot cathode" (normal temperature). E_{RH} will not only vary (like E_{RC}) with changes in grid work function but will also shift slightly with changes in grid emission and leakage currents, which increase with cathode temperature. It is, therefore, not as accurate as E_{RC} as an indicator of contact potential, but it is, nevertheless, useful as a means of showing gross shifts in grid work function.

A grid-current–grid-voltage curve taken with voltage applied to the plate differs from one taken with the plate floating or grounded because of negative currents resulting from gas ions and increased leakage currents. Retarding potentials measured under these conditions provide a general indication of bias shifts due to all causes, and are, therefore, only partially dependent upon contact potential. To avoid confusion with the retarding potentials E_{RC} and E_{RH}, which are measured with the plate floating or grounded, a retarding potential measured with the plate positive is termed "limiting potential" and is symbolized as $E_{L(X)}$, the subscript X representing the reference current in microamperes. Contact potential is sometimes defined as the potential required to limit the grid current to zero microamperes, i.e., $E_{L(0)}$. Another value frequently used to indicate or determine bias shifts is $E_{L(0.1)}$ which has practically the same significance as $E_{L(0)}$. If negative grid currents are used for reference, the limiting potential may be difficult to determine because of the instability of the negative component due to gas. It is undesirable, therefore, to reference the limiting potential to currents in the negative region.

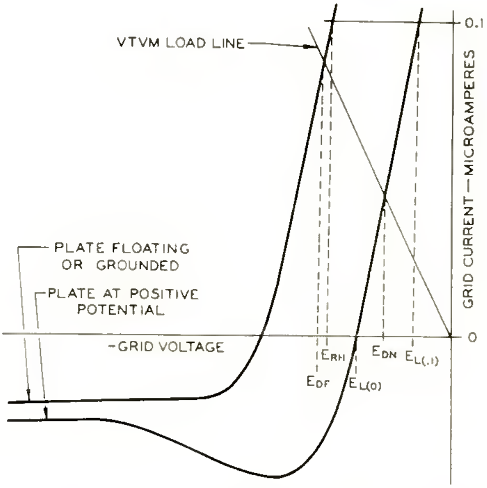


Fig. 4—Curves showing differences in limiting and developed potentials when tube is operated with plate floating or grounded (subscript F), and with plate at an applied positive potential with respect to the cathode (subscript N).

Method 3—Voltmeter Method

If a d-c voltmeter having a resistance of at least 10 megohms (such as a vacuum-tube voltmeter) is connected between the grid and cathode, and no other connection is made to the grid, a voltage drop will be developed across the meter by the grid current of the tube. Graphically, this voltage drop (called "developed potential") is determined by the intersection of the load line with the grid-current curve and, therefore, will have one value when measured with the plate at a positive potential, and another when measured with the plate floating or grounded. These two developed potentials are symbolized, respectively, as E_{DN} (N indicating normal plate voltage) and E_{DF} (F indicating a floating or grounded plate). It can be seen in Figure 4 that E_{DF} and E_{DN} differ very little from retarding and limiting potentials measured under the same conditions and, consequently, provide similar information. The developed potentials, however, are more easily measured than the others.

Relationship Between Methods

Because all three methods described above, with variations, are or have been used to determine contact potential, it is desirable to establish as precisely as possible the relationships among them. The writer has participated in conferences of tube-industry scientific and engineering personnel at which at least three of the various potentials defined above have been repeatedly but inaccurately referred to as "contact potential". A typical example of this misuse occurred in a discussion of a manufacturing variation which resulted in a decrease in the grid work function, and a corresponding increase in developed bias. This increase was erroneously described as an increase in contact potential. The error of this description can be seen by reference to Figure 3, which shows that a decrease in grid work function moves the entire grid-characteristic curve to the left, and, therefore, *reduces* the contact potential.

A decrease in cathode work function raises the saturation region of the curve, but does not shift the retarding-field line. An increase in cathode temperature enlarges the entire curve, particularly the rounded portion. Theoretically, cathode temperature also determines the slopes of the retarding and saturation regions, but in practice these slopes may also depend upon other factors.[5]

The curves of Figure 4 show true cathode-to-grid retarding-field currents combined with negative grid currents. A decrease in grid work function shifts the curves to higher negative voltages and increases the absolute value of the developed, limiting, and retarding potentials. An increase in ion currents due to an increase in tube gas content, however, makes $E_{L(X)}$ and E_{DN} less negative but does not affect the other measurements. Shifts in cathode work function have little effect on the values shown in Figure 4 since, at normal cathode operating temperatures, the primary grid current consists of electrons having energies higher than the energy of the cathode work-function barrier.

Table I gives the results of measurements made on an experimental twin triode similar to type 6CG7 by each of the methods described above. The grids of the two triode units in the experimental tube were made of different materials, but were equally contaminated from the cathode coating. The differences, not only in magnitude but also in the direction of change obtained with the various methods, emphasize the necessity for care in the choice of method used to determine bias shifts.

Table I—Measurements on an Experimental Twin Triode

Potential	Triode Unit I (Magno-Nickel Grid) (volts)	Triode Unit 2 (Silver-Plated Magno-Nickel Grid) (volts)	Difference (Unit 2-Unit 1) (volts)
E_{CP}	1.27	0.50	—0.77
E_{RC}	0.92	0.16	—0.76
E_{RH}	—0.96	—1.35	—0.39
$E_{L(0.1)}$	—0.91	—1.13	—0.22
$E_{L(0)}$	—1.00	—1.25	—0.25
E_{DF}	—0.94	—1.31	—0.37
E_{DN}	—0.85	—1.13	—0.28

Conclusions

Each of the methods described above has a specific field of application in bias-shift studies. The intersection method is the only one which will indicate shifts in true contact potential, i.e., those resulting from changes in the work functions of the cathode, the grid, or both. The retarding-potential method is best suited for measurements of shifts due to changes in grid characteristics—E_{RC} showing the effects of grid contamination, and E_{RH} the combined effects of grid contamination, leakage, and grid emission. Limiting potential and developed potential show shifts due to all causes. To avoid confusion, the method used for obtaining bias-shift data should always be clearly identified.

Acknowledgment

The writer wishes to express his sincere thanks to R. W. Lambdin and E. G. Widell for contributing their time and effort in this survey and to Mrs. D. A. Yanchusk for correcting and proofreading the text.

General References

I. F. Patai and M. A. Pomerantz, "Contact Potential Differences," *Jour. Frank. Inst.*, Vol. 252, p. 239, September, 1951.

S. Friedman and L. N. Heynick, "Circuit for Determination of Contact Potentials and Electron Temperatures from Retarding-Field Characteristics," *Rev. Sci. Instr.*, Vol. 26, pp. 17-19, January, 1955.

R. M. Bowie, "This Matter of Contact Potential," *Proc. I.R.E.*, Vol. 24, p. 1501, November, 1936.

INFLUENCE OF SURFACE OXIDATION ON ALPHA$_{cb}$ OF GERMANIUM P-N-P TRANSISTORS

BY

J. TORKEL WALLMARK

RCA Laboratories,
Princeton, N. J.

Summary—*Oxidation of the germanium surface is shown to be a major factor influencing surface recombination and thereby the current transfer ratio, α_{cb}, of germanium p-n-p transistors. When the transistor is etched electrolytically in KOH, the surface becomes covered by a hydrated germanium monoxide–dioxide layer giving low surface recombination and high α_{cb}. During the subsequent life of the transistor this layer grows through further oxidation, the higher the temperature the more rapid the growth, causing a reduction in α_{cb}. As the oxidation is rapid initially and then slows down with time, it has proven advantageous in manufacturing to subject new transistors to heat treatment. This creates a certain thickness of oxide which during the subsequent life of the transistor does not change rapidly even at elevated temperatures.*

INTRODUCTION

DRIFT of the electrical characteristics of transistors, especially of the current-transfer ratio, α_{cb}, has long been a major problem in transistor applications. It has been known for some time that this drift is associated with changes in surface conditions.[1]

It is shown here that oxidation of the germanium surface is a major factor in the slow decay of α_{cb} in p-n-p transistors. The rate of decay increases with temperature, and α_{cb} may change to such an extent that transistor circuits designed around the initial value of α_{cb} fail to operate properly after a period of time. Our knowledge of the oxidation of germanium is still rather incomplete, but with the help of recent publications, an outline of the processes responsible for this behavior can now be sketched.

Evidence is presented to support the picture that n-type germanium electrolytically etched in KOH has a low surface recombination velocity, a high α_{cb}, and is covered by a thin GeO layer. The surface layer oxidizes further at temperatures between 25° and 125°C according to the so-called "cubic" law, as is the case with copper. This oxidation is in the form of a GeO_2 layer whose growth is responsible for an increase in the surface recombination velocity and thereby the decay in the current-transfer ratio during life.

[1] R. H. Kingston, "Review of Germanium Surface Phenomena," *Jour. Appl. Phys.*, Vol. 27, p. 101, February, 1956.

Earlier Work on Atomically Clean Surfaces

On a freshly cleaved germanium surface, Green and Kafalas[2] have shown that at temperatures between $-68°$ and $+25°C$, chemisorption in less than five seconds covers the surface with about three oxygen atoms per surface germanium atom, thus building up a very thin initial oxide layer. From then on the further chemisorption follows the usual logarithmic law,[3]

$$N = C_1 \log 60t + C_2, \qquad (1)$$

where N is the number of oxygen atoms absorbed per cm^2,

t is time in hours,

$C_1 = 1.4 \times 10^{14}$,

$C_2 = 8.4 \times 10^{14}$.

A logarithmic law was found for both p-type and n-type germanium, and could be followed for 22 days until the rate was too small to be observed.

It has been shown by Law and Meigs[4] that at 500°C oxidation follows the parabolic law:

$$N = 60C_3t^{1/2}, \qquad (2)$$

where $C_3 = 9 \times 10^{13} \times p^{1/2}$, p is the oxygen pressure, and t is the time in hours. This is typical of a process where diffusion through the formed oxide layer is the rate-determining factor. At still higher temperatures, Law and Meigs[4] and Bernstein and Cubicciotti[5] report a linear law in which the rate-determining factor is the evaporation of germanium monoxide from the surface:

$$N = C_4t, \qquad (3)$$

where C_4 is a pressure-dependent factor.

[2] M. Green and J. A. Kafalas, "Oxidation of Clean Surfaces of Germanium Below 25°C," *Phys. Rev.*, Vol. 98, p. 1566, 1955.

[3] M. Green, J. A. Kafalas, and P. H. Robinson, "Interaction of Oxygen with Clean Germanium Surfaces: Part I Experiment," *Semiconductor Surface Physics*, University of Pa., 1957, p. 349.

[4] J. T. Law and P. S. Meigs, "The High Temperature Oxidation of Germanium," *Semiconductor Surface Physics*, University of Pa., 1957, p. 383.

[5] R. B. Bernstein and D. Cubicciotti, "The Kinetics of the Reaction of Germanium and Oxygen," *Jour. Amer. Chem. Soc.*, Vol. 73, p. 4112, September, 1951.

Earlier Work on "Real" Surfaces

While the above results are applicable to surfaces in contact with oxygen alone, it is obvious that etches commonly used to clean the germanium surface will lead to chemisorption of etch constituents so that we are dealing with a contaminated surface. On such a surface the contamination may significantly change the rate of further oxidation in a manner well known from oxidation studies[6] once the germanium has been removed from the etch. Furthermore, as most germanium etches are oxidizing, they will leave an initial thickness of oxide on the surface. It has been found by Green and Smythe[7] that an etched surface may be covered by three monolayers of oxide five minutes after etching, increasing to six monolayers after three days' exposure to room air. Contaminated surfaces are presently less understood than clean surfaces, and the considerable amount of data that exists on surface traps has not yet been interpreted and correlated with the chemical and physical processes that are responsible for the traps. The present article constitutes a step in this direction.

Two Basic Surface Treatments

The results of two well-investigated surface treatments for germanium are taken as a starting point. These results are used to form a hypothesis about the oxidation of germanium which is then checked against the behavior of germanium transistors on life test. The two surface treatments referred to are the so-called No. 5 etch* and anodic oxidation in glacial acetic acid. The No. 5 etch is known to leave germanium monoxide, GeO, on the germanium surface, thereby making the surface conductivity strongly n-type.[8] This gives p-n-p transistors a low surface recombination velocity and consequently a high α_{cb} and a low saturation current.

The anodic oxidation treatment in acetic acid has been worked out by Zwerdling and Green[9] and is known to leave a layer of germanium dioxide, GeO_2, on the surface. This dioxide layer influences the surface conductivity from n-type towards less n-type as is shown in the follow-

[6] R. K. Hart, "The Oxidation of Aluminium in Dry and Humid Oxygen Atmosphere," *Proc. Roy. Soc. A*, Vol. 236, p. 68, July, 1956.

[7] M. Green and R. Smythe, as reported in Reference (1).

* The No. 5 etch, developed by S. G. Ellis of these Laboratories, consists of 40 parts 48% HF, 6 parts 30% H_2O_2, 24 parts H_2O.

[8] S. G. Ellis, private communication.

[9] S. Zwerdling and M. Green, "Anodic Oxide Films on Germanium," *Jour. Electrochem. Soc.*, Vol. 103, p. 61, March, 1956 (abstract #60). Also Lincoln Lab. M.I.T., Quarterly Progress Report, August 1, 1956.

ing experiment. Four groups of p-n-p transistors were etched electrolytically in KOH and subsequently two of the groups were given the anodic oxidation treatment. Thereafter all groups were put on 85°C shelf life test. Figure 1 is a plot of α_{cb} versus time. Each experimental point is an average of three units. As can be seen in Figure 1 the untreated control groups behave quite normally, the experimental points falling around the theoretical curve. The oxidized groups first show very low α_{cb} which, however, climbs rapidly (due to removal of bound etch constituents) and approaches a curve similar to the solid curve but shifted to the left. In other words, the oxidation treatment has applied an oxide layer of the same thickness as would be applied by heating to 85° for some 30 hours.

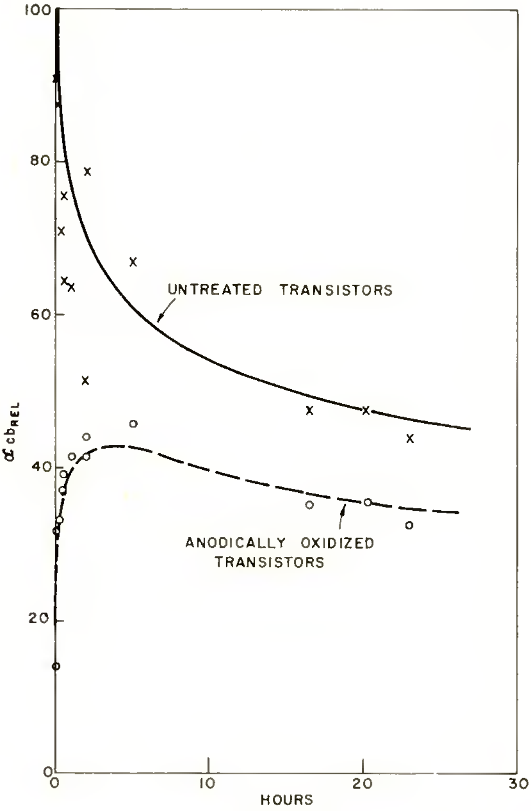


Fig. 1—Influence on transistors of anodic oxidation in glacial acetic acid plus sodium acetate. In the ordinate scale, 100 represents the value of α_{cb} immediately after the electrolytic etch. The solid curve represents Equation (10).

The terms GeO and GeO_2 have been used loosely as it is known that these oxides under the prevalent conditions take up a rather indefinite amount of H_2O forming various hydrated forms such as di- or metagermanic acids, etc.[10] For silicon, which behaves in a similar manner, these processes have been studied, but for germanium the details are not as well known. Therefore the terms monoxide (GeO) and dioxide (GeO_2) have been used for simplicity without specifying amount of hydration or atomic arrangement.

Without offering further proof at this time, it will be assumed that similar oxidation processes take place during the conventional processing of transistors. In particular, it will be assumed that the high α_{cb} of p-n-p transistors found after an electrolytic etch in KOH is a result of a layer of GeO on the surface of the germanium, while the subsequent decay of α_{cb} may be interpreted as due to a growing layer of GeO_2 on top of the GeO layer. In line with this it has been found by E. O. Johnson and G. C. Dousmanis of these Laboratories that a germanium surface freshly etched electrolytically in KOH is n-type, but in the course of many hours gradually goes over more and more towards less n-type.

That oxidation of the surface should influence the electrical characteristics is entirely in line with the present view of oxidation processes.[11,12] In the adsorption of gas atoms on the surface, strong electric fields are set up, as witnessed by the well-known changes in surface recombination velocity, surface conductivity, etc., caused by exposure to different gas species.[1] The strong electric field is fundamental in the early growth of an oxide layer by aiding the transport of ions through the layer. As the layer grows, the electric field at the metal–oxide interface necessarily decreases, and it is this gradual reduction of the electric field that is responsible for the change in surface recombination velocity and thereby also α_{cb} of transistors.

Experimental Results

The experimental work consists of a number of life tests on transistors, comprising measurements of α_{cb} versus time under various conditions.

[10] O. H. Johnson, "Germanium and its Inorganic Compounds," *Chem. Rev.*, Vol. 51, p. 431, December, 1952.

[11] K. Hauffe, *Reaktionen in und an Festen Stoffen*, Springer Verlag, Berlin, Germany, 1955, p. 411.

[12] W. E. Garner, *Chemistry of the Solid State*, Butterworths Scientific Publications, London, England, 1955, p. 336.

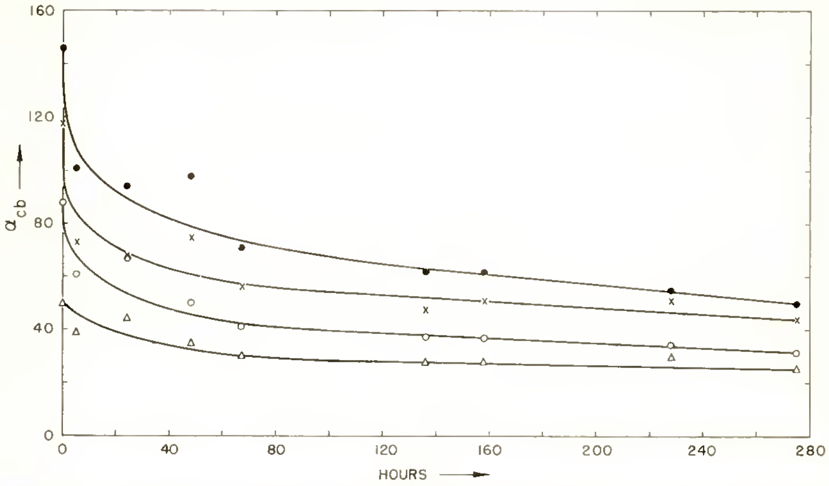


Fig. 2—Life test results for encapsulated silicone-covered p-n-p germanium transistors. Shelf life at 70° C. (F. L. Hunter)

It may be pointed out here that this data should not be interpreted as the performance of commercially available transistors. Factory procedures generally include an "aging" step, e.g., heating to an elevated temperature for a period of time, which reduces α_{cb} to approximately 35 per cent of its original value and after which moderate temperatures within the rating of the transistor do not cause signifi-

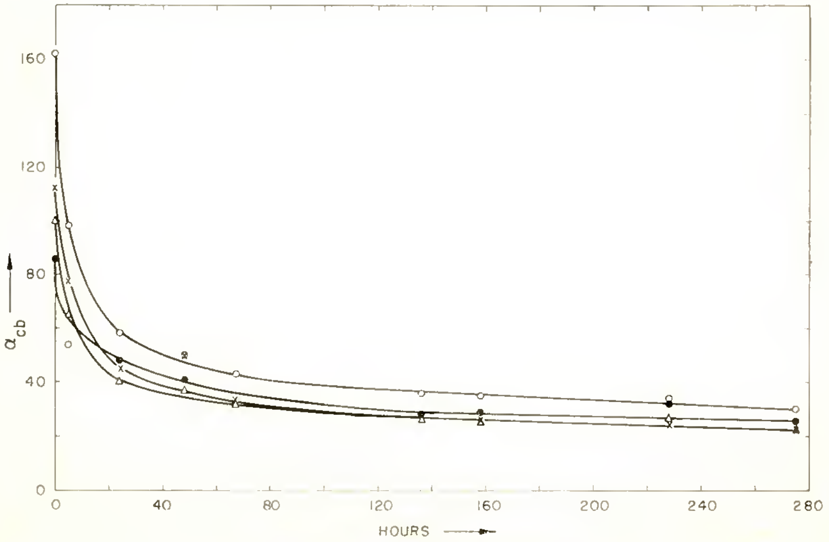


Fig. 3—Shelf life at 85° C.

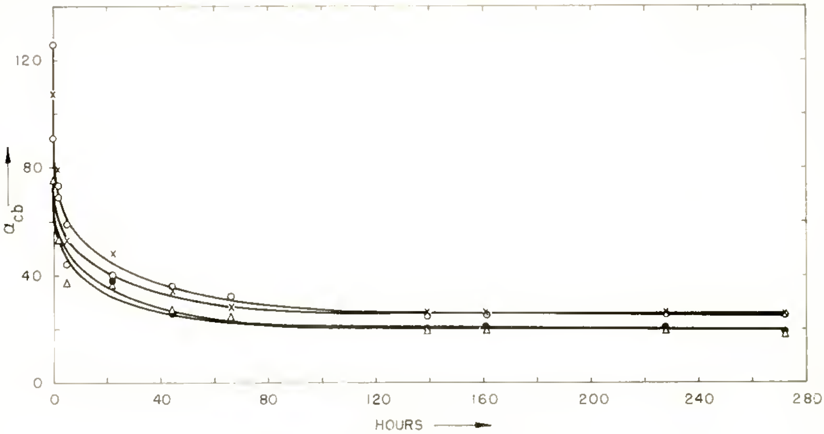


Fig. 4—Shelf life at 105°C.

cant further changes. Thus commercial processes employ in some degree an oxidation process to achieve stability.

Figures 2-5 present life test data for four groups of p-n-p germanium transistors at four different temperatures. In each group there is one transistor for which α_{cb} seems to decay somewhat differently from the other three, perhaps because of a slight error in the initial α_{cb} reading. Therefore, an average has been taken of the three similar units, disregarding the fourth. In order to let all transistors contribute equally to the average, relative values have been used, setting the initial $\alpha_{cb} = 100$. The results are shown in Figure 8. For each point, the

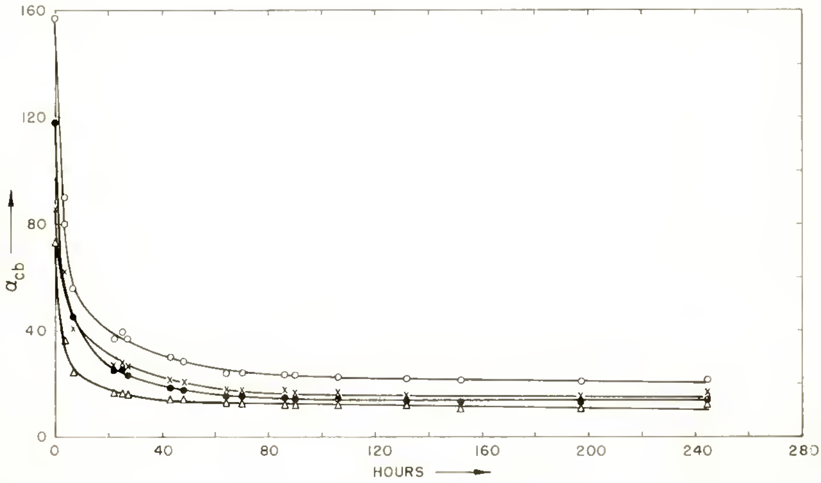


Fig. 5—Shelf life at 125°C.

spread between the three values is shown. The transistors had been electrolytically etched in KOH and hermetically encapsulated with a silicone resin covering the germanium surface.

Other runs were made in which it was established that the silicone was inconsequential and that units in silicone or air showed identical behavior within the accuracy of the measurements. Vacuum-sealed units showed a reduced decay as expected if the decay is caused by an oxidation process. Figure 6 represents such a run in which the units were sealed at 10^{-4} mm Hg and life tested at 100°C, 70°C, and room temperature. Each point is an average of three units. Relative values are used as explained above. The theoretical curves are identical to Equation (10) except for a factor 0.09 representing the influence of reduced oxygen pressure.

In order to test the theory describing these results, especially the temperature dependence of the oxide growth, a room-temperature life test was run; this is shown in Figure 7. These transistors were sealed in hermetic cans in dry air without silicone. The results are also shown in Figure 9, where each point represents an average of five transistors, and again the spread is shown. In analyzing this test, the three transistors with low α_{cb} have been discarded.

Outline of Method of Analysis

The thickness of the growing GeO_2 layer cannot be very large at the temperatures (25° to 125°C) and times involved. It is shown later that this growth cannot exceed one hundred monolayers. It is reasonable, therefore, to assume that the increase of the surface recombintion velocity, s, is directly proportional to the increase of the thickness of the layer. For $\Delta w \ll w$,

$$\frac{ds}{dt} = p \frac{dw}{dt}, \tag{4}$$

where w is the thickness of the GeO_2 layer and p is a proportionality factor. Upon integration,

$$s = s_o + pw. \tag{5}$$

α_{cb} is directly related to the surface recombination velocity by the equation[13]

[13] W. M. Webster, "On the Variation of Junction-Transistor Current-Amplification Factor with Emitter Current," *Proc. I.R.E.*, Vol. 42, p. 914, June, 1954.

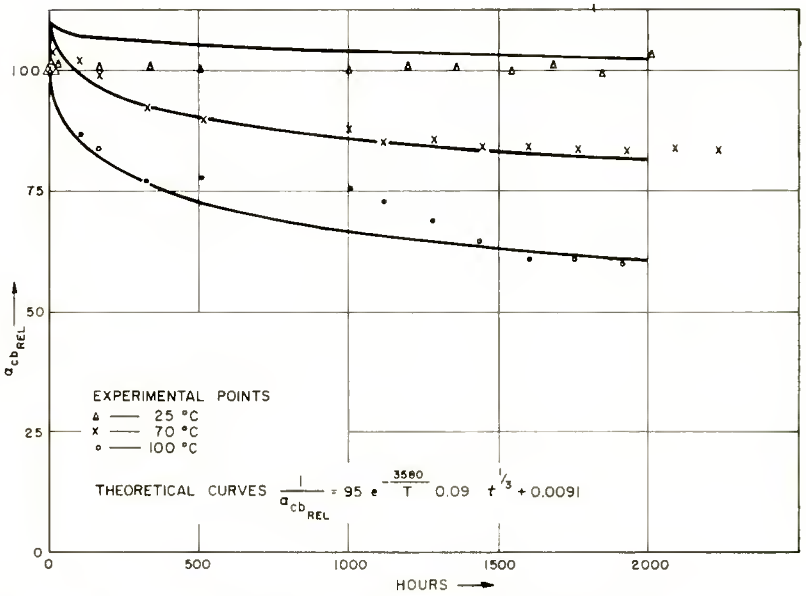


Fig. 6—Life test results for vacuum-sealed (10^{-4} mm Hg) transistors. Shelf life at 105°C. (R. R. Johnson)

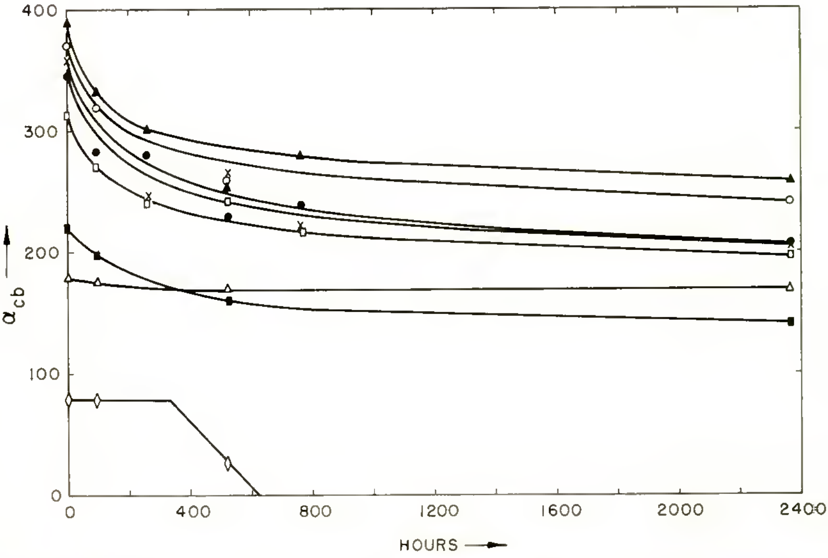


Fig. 7—Life test results for transistors sealed in dry air, no silicone. Shelf life at room temperature. (E. L. Jordan)

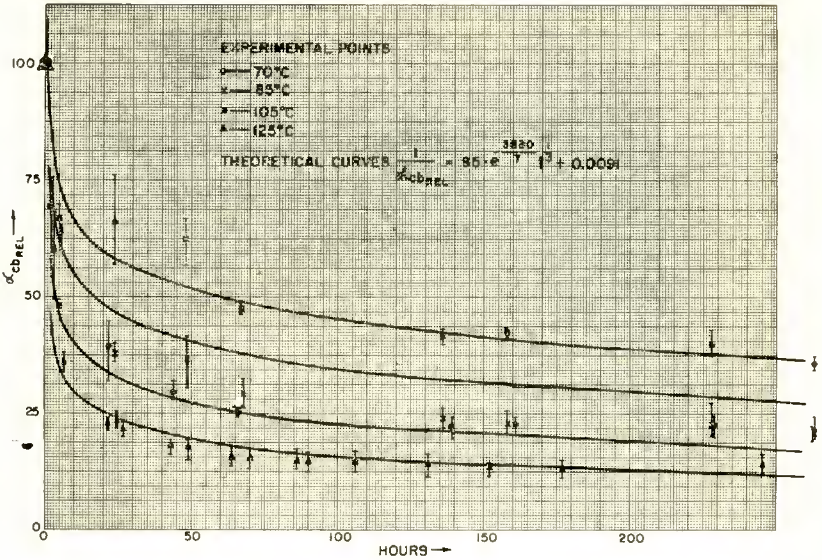


Fig. 8—Figures 2-5 replotted using average of relative values of α_{cb}. Range within which values fall is indicated.

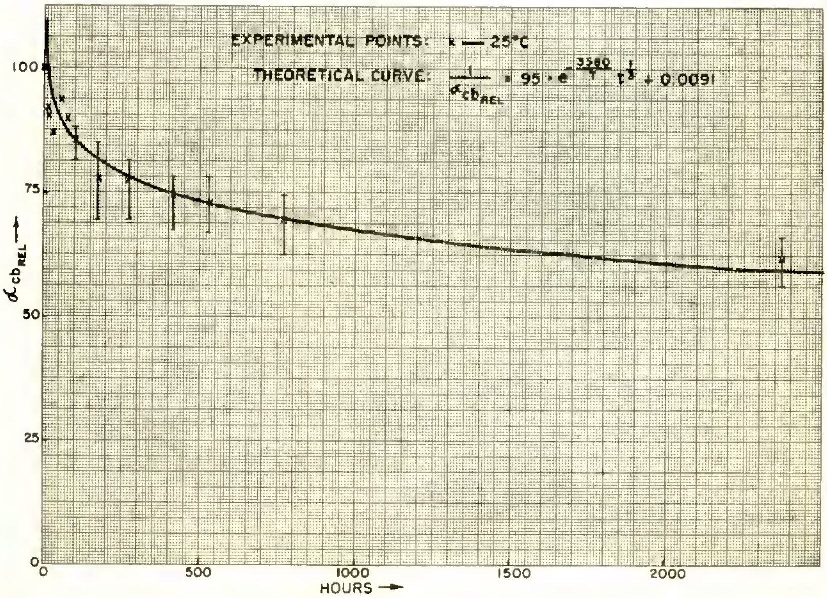


Fig. 9—Figure 7 replotted using average of relative values of α_{cb}. Range within which values fall is indicated.

$$\frac{1}{\alpha_{cb}} = C_5\, s + C_6\,, \tag{6}$$

where C_5 and C_6 are given by the geometry, material, etc. of the transistor. Combination of Equations (5) and (6) then gives

$$\frac{1}{\alpha_{cb}} = \frac{1}{\alpha_{cb_o}} + C_5 pw, \tag{7}$$

where

$$\frac{1}{\alpha_{cb_o}} = C_5 s_o + C_6.$$

Now by insertion of experimental values of α_{cb} versus time in Equation (7), the growth rate of the oxide layer can be obtained from the resulting relation of w versus time.

Analysis of the Data

It is known that within a certain temperature range the oxidation process follows one of a small number of possible rate equations.[11,12] Using Equation (7) above, it is now possible to fit the rate equations to the experimental data. It is then found that the so-called "cubic" rate equation,

$$d = C_7 t^n + C_8, \tag{8}$$

where $n = 1/3$, fits the data very well while none of the other rate equations can account for the initial drop. In Figures 10 and 11 experimental points from Figure 8 have been redrawn and a number of possible rate equations have been fitted at two points. As can be seen, good fit can only be claimed for the "cubic."

Equation (8) applies to the case of oxidation of a metal with a thin oxide layer. The validity of the assumption of a thin oxide layer (< 500 Å) will be demonstrated by the direct oxidation experiment described below.

In Equation (8), the coefficient C_7 contains an exponential factor characterizing the activation energy for oxidation. Thus Equation (8) can be written as

$$\frac{1}{\alpha_{cb}} = C_9 e^{\frac{-C_{10}}{T}}\, t^{1/3} + C_6. \tag{9}$$

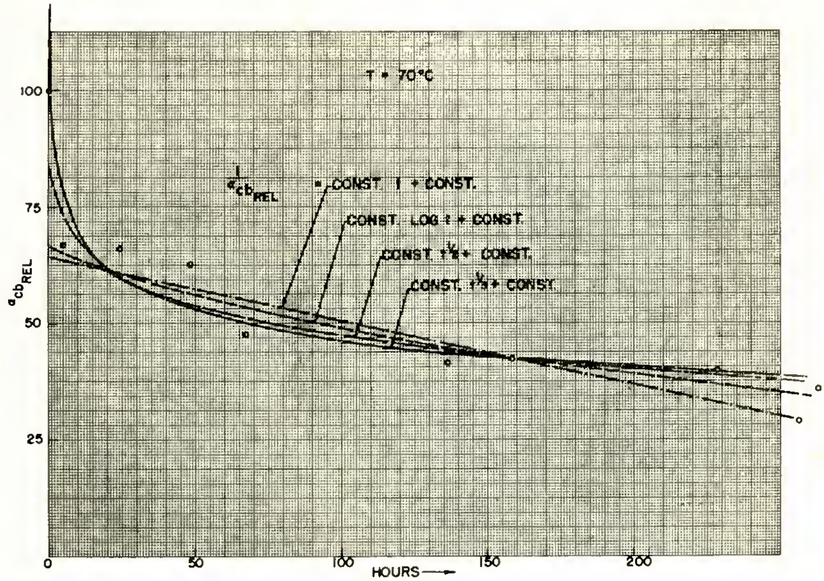


Fig. 10—Some known oxidation equations fitted to the 70°C curve of Figure 8.

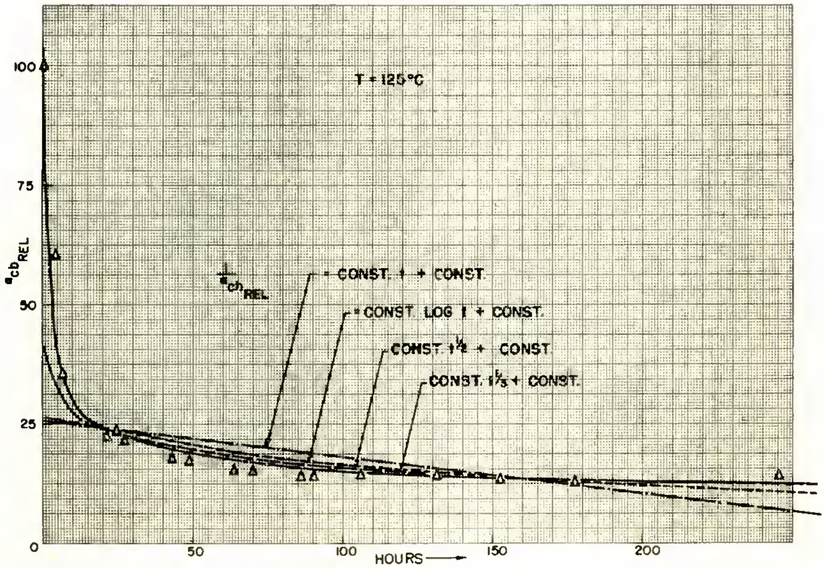


Fig. 11—Some known oxidation equations fitted to the 105°C curve of Figure 8.

Equation (9) has been fitted to the experimental points in Figure 8 using coincidence at two arbitrary points, namely 20 hours and 160 hours. The result is the equation

$$\frac{1}{\alpha_{cb}} = 95\, e^{-\frac{3600}{T}}\, t^{1/3} + 0.0091, \tag{10}$$

which is represented by the curves of Figure 8. As can be seen in Figure 9, this equation fits the room-temperature curve also.

We can press the fit a little further. At a first glance, it is surprising that the theoretical curves, Equation (10), do not go through the point (0,100), in other words, that the constant in Equation (10) is not 0.01. However, it has been found by R. R. Johnson of the RCA Semiconductor Division, that when transistors are taken off a high-temperature life test at, say, 85°C and brought to room temperature for the purpose of measurements, α_{cb} shows a decline for about 48 hours as shown in Figure 12. This was not known at the time of the life test described above and therefore no time interval was fixed between removal from life test and measurement. The procedure was to remove the units in the morning and to measure them at any time during that day. It may therefore be assumed that an average of four hours elapsed between removal and measurement corresponding to a decline in α_{cb} of about 9 per cent as shown in Figure 12. (This fact at the same time may explain some of the spread of the measured α_{cb} values.) However, the initial α_{cb} was measured before any heat treatment and therefore would correspond to the end value in Figure 12. It may be expected then that all the points except the initial reading are about 9 per cent high as indeed seems to be borne out by Equation (10). When the measurements are corrected for this effect they obey Equation (11):

$$\frac{1}{\alpha_{cb}} = 95e^{\frac{-3600}{T}}\, t^{1/3} + 0.01. \tag{11}$$

The value of the activation energy in Equation (10), (7200 cal/mol) at first seems unusually small. However, recent work on the oxidation of copper,[14] which according to the reasoning outlined above follows a similar oxidation process, also suggests a small activation energy, about 9,000 cal/mol.

[14] F. W. Young, J. V. Cathcart, and A. T. Gwathmey, "The Rates of Oxidation of Several Faces of a Single Crystal of Copper as Determined with Elliptically Polarized Light," *Acta Metallurgica*, Vol. 4, p. 145, No. 2, 1956.

Direct Oxidation Experiment

In order to provide a direct proof of the oxidation hypothesis set forth above to explain slow decay of α_{cb} of p-n-p transistors, the following experiment was performed. Samples of n-type germanium were heated in an oven at 85°C for various periods of time. Subsequently the samples were washed in an NaOH solution to remove any oxide formed. The amount of germanium in the solution was then determined quantitatively by a spectrographic method. Figure 13 shows the results. The experimental points are in reasonable agreement with the cube-root-of-time law suggested earlier. In order to retain the oxide in easily soluble form it was found necessary to keep the samples at 100 per cent relative humidity during the heat treatment.

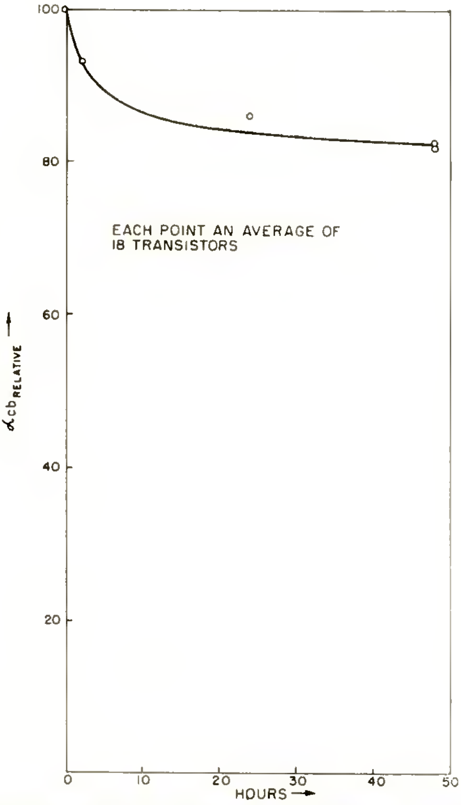


Fig. 12—Off-life decline of α_{cb} ("48-hour effect"). (R. R. Johnson)

The results in Figure 13 make it possible to deduce a quantitative expression for the oxidation rate. Assuming that at the low relative humidity present in encapsulated transistors the oxidation rate is not too different from the rate at 100 per cent relative humidity, the following expression gives an upper limit for the oxide layer thickness in Å at 85°C:

$$w \leqq 21\ (t + 1.3)^{1/3}. \tag{12}$$

It should be pointed out that this oxidation rate is valid for an etched, and consequently hydrated, surface and therefore may be higher than for a clean germanium surface.

Influence of a Sequence of Temperatures

From the picture outlined above some predictions may be made concerning the influence of oxidation on units that are subjected to a sequence of temperatures. As an example, consider transistors that are first aged at room temperature for 24 hours. Following this they are heated for 24 hours at 105°C. With the help of Equation (10) we may now follow the change in α_{cb} due to oxidation. After the first 24 hours of room temperature aging, α_{cb} is down to 85 per cent of its initial value. After the next 24 hours at 105°C, α_{cb} is down further to 30 per cent of the initial value. From then on at a temperature of 85°C, $\alpha_{cb_{rel}}$ can be predicted from the following time dependent expression:

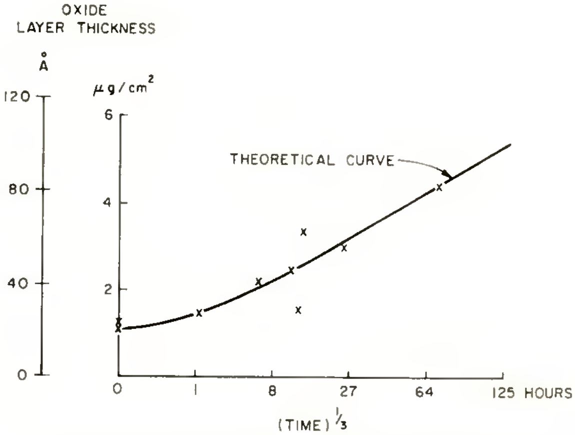


Fig. 13—Thickness of oxide layer versus time.

$$\frac{1}{\alpha_{cb_{rel}}} = 0.00427\ (t + 160)^{1/3} + 0.033. \tag{13}$$

Alternatively, assuming that the value of α_{cb} at the start of the 85°C shelf life test is 100 per cent, and that all measurements (except the initial one) are made an average of 4 hours after removal from the oven, we have

$$\frac{1}{\alpha_{cb_{rel}}} = \frac{1}{3.3 \times 1.09} [0.00427 (t + 160)^{1/3} + 0.033]. \qquad (14)$$

This relation is plotted in Figure 14. In a similar manner the results

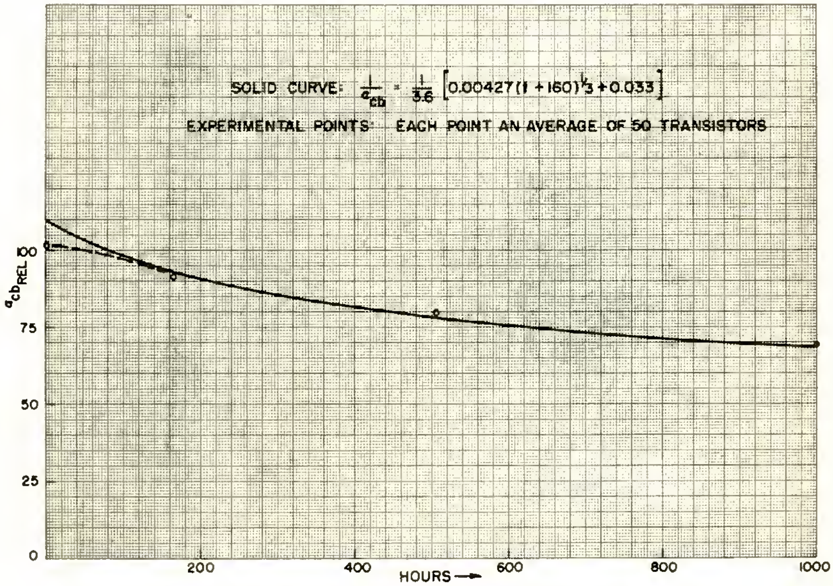


Fig. 14—Theoretical curve and experimental values of α_{cb} for shelf life at 85°C.

of any combination of temperature and aging time can be predicted. Figure 14 also shows some experimental points from life tests of units that have been subjected to the above-mentioned sequence of temperatures.

The first constant in Equation (13) will be different for different types of transistors and depends on the ratio C_5/C_6 in Equation (6).

It should be pointed out once again that factors other than oxidation also influence the change of α_{cb} during life. However, interpretation of other effects is difficult unless the influence of oxidation is first singled out.

Conclusions

In view of the evidence presented, it appears likely that in germanium p-n-p transistors which have been electrolytically etched in KOH and subsequently hermetically sealed, α_{cb} is influenced by oxidation as follows:

1. The conventional electrolytic etch in 40 per cent KOH leaves a composite hydrated oxide layer, GeO — GeO_2 on the surface.

2. The high α_{cb} resulting from this etch is mainly a consequence of the GeO layer.

3. The GeO_2 layer will grow with time and the decay of α_{cb} is a consequence of this growth.

4. The value of α_{cb} after storage of a p-n-p germanium transistor at a temperature T°K can be predicted from

$$\frac{\alpha_{cbo}}{\alpha_{cb}} - 1 = C_1 e^{-\frac{3600}{T}} (t + t_o)^{1/3},$$

where α_{cbo} is the initial value after electrolytic etch, and C_1 is a constant which depends on the geometry of the transistor. For a 2N77 transistor, $C_1 = 95$. t_o refers to the oxide thickness at the start of the storage period and is the time necessary to grow this thickness at the temperature T. t_o may be obtained from

$$\frac{\alpha_{cbo}}{\alpha_{cb1}} - 1 = C_1 e^{-\frac{3600}{T}} t_o^{1/3},$$

where α_{cb1} is the value at the start of the storage period.

Acknowledgment

The author has benefited greatly from discussions with R. H. Plumlee. Most of the experimental data has been made available by F. L. Hunter, R. R. Johnson, E. L. Jordan, A. Kerken, A. L. Kestenbaum and S. C. Simons of the RCA Semiconductor Division. The assistance by M. C. Gardels and H. H. Whitaker with the spectrographic analysis is gratefully acknowledged.

PULSE-FIRING TIME AND RECOVERY TIME OF THE 2D21 THYRATRON*

By

John A. Olmstead and Melvin Roth

RCA Electron Tube Division,
Harrison, N. J.

Summary—*The time-dependent or dynamic characteristics of the 2D21 are of considerable importance to the application engineer. In many cases, these characteristics determine the ability of the associated circuit to function properly. Three of the most important dynamic characteristics of the 2D21 are pulse-firing time, ionization time, and grid-recovery time.*

Pulse-firing time is the time required after application of the grid pulse for the ionization in the grid region to reach a value such that the grid loses control of the breakdown process. In the 2D21 this time is essentially equal to the ionization time (time required for anode conduction). Pulse firing time is a function of anode voltage, and grid–cathode voltage, the significant parameter being the grid–cathode voltage. When this voltage is positive, the grid draws appreciable current and proper circuit analysis must be applied to determine pulse source requirements.

Recovery time is the period required for the grid to regain control after anode current flow has been interrupted by removal of anode voltage. It is primarily a function of anode current prior to extinction, anode voltage, and the grid–cathode voltage. The actual grid–cathode voltage must be determined by proper circuit analysis, with particular attention given to the ion grid current flowing during deionization.

With proper circuitry the 2D21 can be ionized in 1 microsecond, and grid control can be recovered within 100 microseconds. The data presented should enable the application engineer to design the necessary circuitry with a minimum of trial and error. Although the data is applicable only to the 2D21, the general concepts apply to other thyratrons.

Introduction

THE 2D21 thyratron has found wide application in the industrial, military, and computer fields because of its small size and desirable characteristics as a low-impedance switch. A great many of these applications involve the use of the 2D21 in pulse-type circuits. Because adequate data on the dynamic characteristics of the 2D21 has not been available, designers have not been able to predict the end performance of such circuits, and, consequently, have been compelled to design them largely by trial-and-error methods.

* Portions of this paper are based on work done for a thesis submitted to Newark College of Engineering by J. A. Olmstead in June, 1957 as part of the requirements for the degree of Master of Science in Electrical Engineering.

This paper discusses and presents quantitative data on the two most important dynamic characteristics of the 2D21—pulse-firing time and grid-recovery time.

Pulse-firing time is the time required for the ionization in the grid region to reach a value which causes the grid to lose control of the breakdown process.

Grid-recovery time is the time required after removal of the anode voltage for the grid to regain a degree of control sufficient to prevent conduction when the anode voltage is reapplied. Although this interval is sometimes called the deionization time, the term recovery time is more precise because the grid can regain control even when a substantial amount of ionization is present.

Pulse Firing

"Pulse firing" of the 2D21 takes place in the following manner: Electrons emitted by the cathode are accelerated by the applied anode potential and/or grid potentials, and acquire sufficient energy to ionize the included gas. However, the region of the tube in which the electrons acquire ionization energies and the number which reach this region are determined primarily by the grid–cathode potential. At low grid–cathode potentials, electrons can acquire ionization energies only in the anode region, and few electrons reach this region. When the grid is made highly positive with respect to the cathode (for example, by application of a grid pulse) large numbers of electrons can acquire ionization energies in both the grid and anode regions.[1]

If the grid pulse voltage produces sufficient positive ions in the grid region, these ions will surround the grid and neutralize the effects of the applied grid potential. As a result, the grid loses its ability to control the anode current and breakdown eventually occurs.

The interval required for the grid to lose control after application of the firing pulse is called the pulse-firing time. The somewhat longer interval required for the tube to become completely ionized is called the ionization time. The pulse-firing time and ionization time of the 2D21 are essentially equal and will be treated as such.

Measurement of Pulse-Firing Time

A block diagram of the circuit used to measure the pulse-firing time of 2D21, and the associated anode-voltage and grid-voltage waveforms are shown in Figure 1. The anode voltage is a 60-cycle square

[1] J. D. Cobine, *Gaseous Conductors,* p. 456, 457, McGraw-Hill Book Company, Inc., New York, 1941.

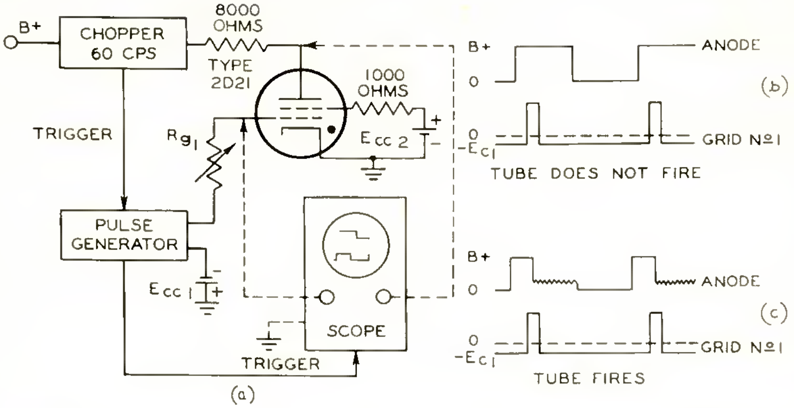


Fig. 1—(a) Circuit used to measure pulse-firing time of the 2D21; (b) and (c) anode and grid-voltage waveforms.

wave. The applied grid voltage is a positive-going pulse delayed so that it occurs during the positive portion of this square wave. The adjustable series resistance R_{g1} in the grid circuit is provided to facilitate grid-current measurements. Figure 1(b) shows the grid and anode wave forms under conditions such that breakdown does not occur. Figure 1(c) shows the wave forms under breakdown conditions.

Figure 2 is a time expansion of the breakdown waveforms in Figure 1(c) showing the important parameters, namely peak grid–cathode voltage, E_{gk}, critical pulse width, and ionization time, T_i.

The shaded portion of the grid pulse is unnecessary for firing the

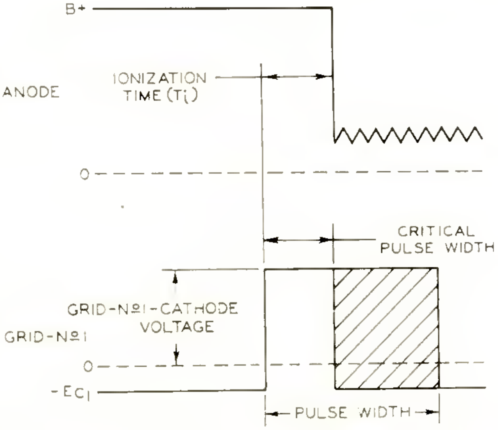


Fig. 2—Expanded presentation of waveforms shown in Figure 1c, showing relation between ionization (pulse-firing) time and critical pulse width. Shaded portion of grid-voltage pulse is superfluous.

tube. In fact, the tube will fire even if the grid-pulse duration is slightly less than the ionization time. In the 2D21, however, the difference between the critical pulse duration and ionization time is so small that it can be neglected. Figure 3 shows the maximum required grid–cathode voltage versus anode voltage for various ionization (pulse-firing) times.

The curves in Figure 3(a) are for a shield-grid voltage of 0 volts and those in Figure 3(b) for a shield-grid voltage of +9 volts. These curves can be used to determine the maximum grid–cathode voltage required for any desired ionization time. For example, with 200 volts on the anode, and 0 volts on the shield grid, the grid–cathode voltage required to ionize the tube *within* 5 microseconds is +17 volts.

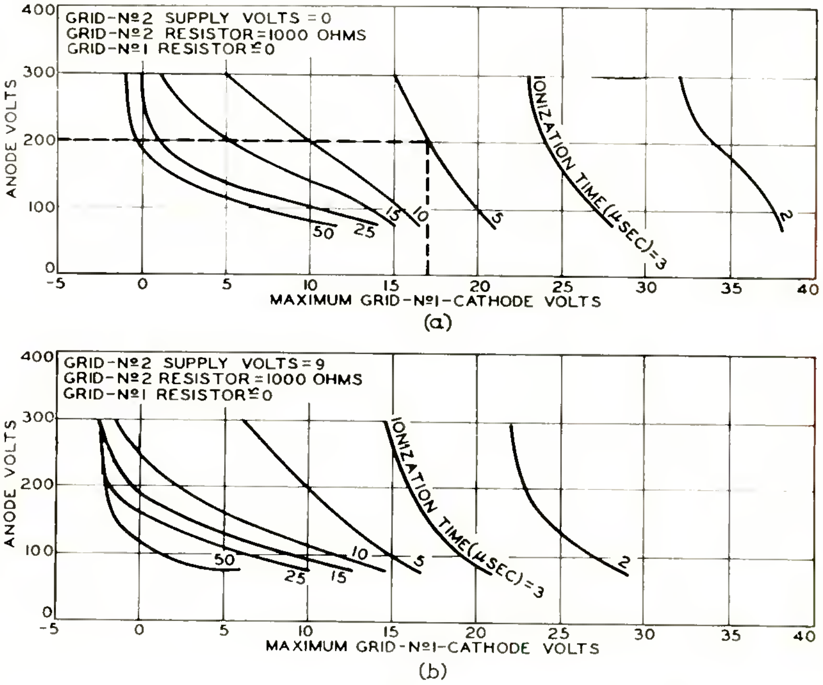


Fig. 3—Maximum grid–cathode voltage versus anode voltage with ionization time, E_{cc2}, and R_{g2} as parameters.

Note that the required grid–cathode voltage varies inversely with ionization time and that the rate of variation is nonlinear. For example, the difference between the grid–cathode voltages required for ionization times of 2 and 3 microseconds is substantially greater than the difference between those required for ionization times of 25 and 50 microseconds.

Absolute time jitter due to variations in the grid-pulse amplitude will, therefore, be less at shorter ionization times.

The grid-bias voltage may have any value within tube ratings; *the important parameter is the actual grid–cathode voltage.*

Figure 4 shows curves of maximum grid current versus grid–cathode voltage, with shield voltage as a parameter. These curves show the grid current flowing prior to firing. Any series impedance between the pulse source and the grid, therefore, becomes an important consideration in determining the actual grid–cathode voltage, because the applied pulse will be attenuated by the resulting voltage drop. The actual grid–cathode voltage for a given grid-circuit impedance and pulse-source voltage can be determined by circuit analysis.

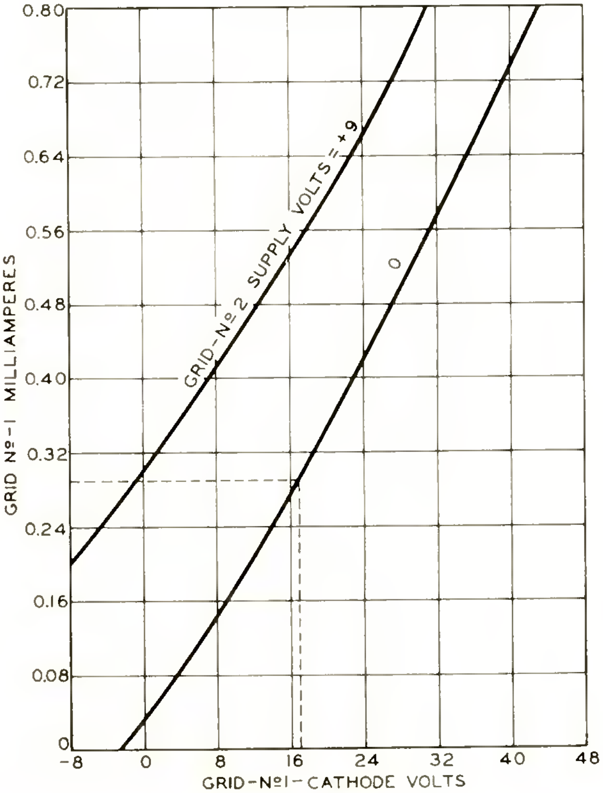


Fig. 4—Maximum grid-current versus grid–cathode voltage, with E_{cc2} as a parameter.

Figure 5 shows the minimum grid–cathode voltages required for various pulse-firing times and anode voltages. This information is particularly valuable when the 2D21 is to be used in coincidence-type circuits because the curves can be used to determine the maximnm pulse width and/or amplitude that can be applied without firing the tube. For example, with 200 volts on the anode and 0 volts on the shield-grid, a pulse of 5 microseconds duration may fire the tube if the resulting grid–cathode voltage exceeds 8 volts.

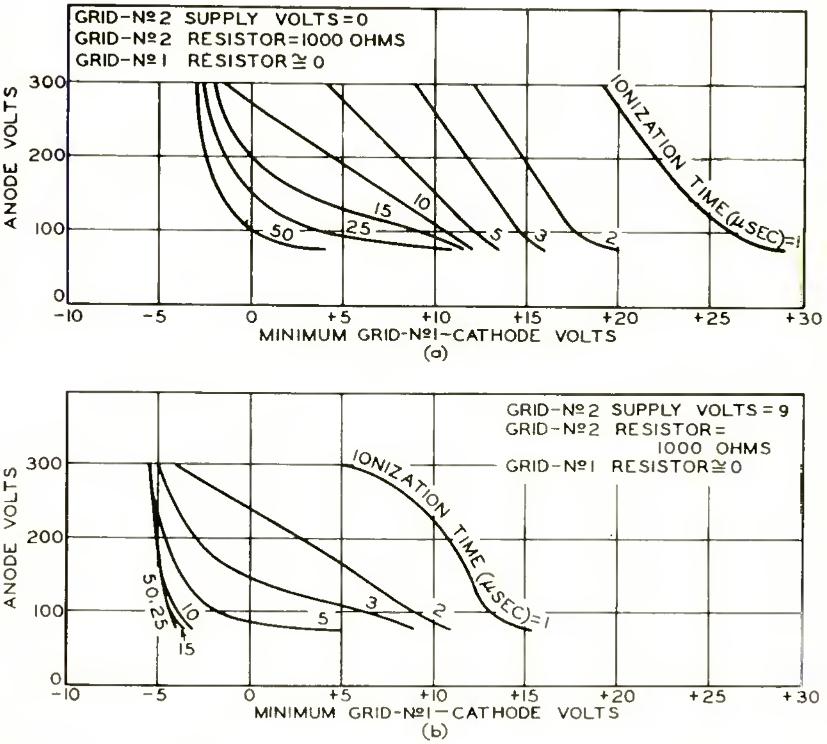


Fig. 5—Minimum grid–cathode voltage required for various ionization times, with anode voltage, E_{cc2}, and R_{g2} as parameters.

Note that the data shown was obtained with the use of a resistance of 1,000 ohms in series with the shield grid. The use of this resistance reduces ion bombardment of the shield and thereby improves tube life. The value of 1,000 ohms was found satisfactory for use in all circuits.

Recovery of Grid Control in the 2D21

Once conduction has been established in the 2D21, flow of anode current cannot be interrupted or controlled by grid action. The grid loses control because it is surrounded by a sheath of positive ions which neutralize the effects of any applied grid voltage. Although this sheath is relatively opaque to electrons, it does not fill the entire grid aperture, and, therefore, does not obstruct the flow of electrons to the anode. Increasing the negative grid voltage would normally expand this sheath and thereby permit the grid to control the anode current. During conduction, however, an increase in negative grid potential shifts the region of ion production in such a manner as to increase the ion density in the grid region rather than to allow sheath expansion.

Ion production can be stopped only when the anode voltage is reduced below the extinction value (approximately zero volts). The resulting elimination of ion production permits the ion sheath surrounding the grid to expand. When the sheath expands sufficiently to close the grid aperture the grid begins to reassert control. If the grid potential at this point is sufficiently negative, subsequent reapplication of anode voltage will not refire the tube. As deionization progresses, the grid acquires greater degrees of control, until it reaches its prefiring condition. Recovery time, therefore, is a function of the grid–cathode voltage at the instant anode potential is reapplied. *This actual grid–cathode voltage is not necessarily the bias voltage.*

To determine the actual grid–cathode voltage during the deionization period, it is necessary to know the magnitude of the ion current flowing in the grid circuit throughout this period. This ion current (called the "after-glow grid current") depends upon the ion density within the tube and consequently, on the anode current prior to extinction, and is, therefore, substantially independent of the external grid-circuit parameters. The after-glow grid current decays approximately exponentially from its initial value. If the voltage drop across the bias impedance due to the after-glow current tends to exceed the bias voltage, electrons as well as positive ions flow to the grid, cancelling a portion of the after-glow current and thereby preventing the grid from becoming positive with respect to the cathode. When the after-glow current decays to a value such that the resulting voltage drop across the bias impedance is equal to the bias voltage, the grid potential begins to return to the bias value. The resulting delay in the return of the actual grid–cathode voltage to the bias value is responsible for the increase in recovery time observed when a grid-bias impedance is present.[2]

[2] L. Malter and E. O. Johnson, "Studies of Thyratron Behavior," *RCA Review*, Vol. XI, pp. 165-189, June, 1950.

Measurement of Recovery Time

A block diagram of the setup used to measure the recovery time of the 2D21, and the associated waveforms are shown in Figure 6. The test method shown is similar to one now under consideration by the Joint A.I.E.E.–I.R.E. Subcommittee on Gas Tubes. A master oscillator places repetitive positive pulses simultaneously on the control grid and anode. The tube conducts for a period sufficiently long to establish steady-state conditions. Recovery time is then measured by means of a "probe" pulse which may be applied to the anode at any desired time after cessation of the anode-current pulse. The delay of the probe pulse with respect to the end of the anode-current pulse is set at the desired recovery time. With the anode-conduction current and the probe-pulse voltage at the desired values, the grid–cathode voltage is then adjusted until the tube just fails to conduct during the probe pulse.

Measurement of the after-glow current is facilitated by resistor R_{g1} in the grid circuit.

Figure 7 shows the critical grid–cathode voltages required for various recovery times. Figure 7a is for a peak anode current of 12 amperes and 7b for a peak anode current of 5 amperes. Note that in Figure 7a after 120 microseconds of deionization, a grid–cathode voltage of—68 volts is required to prevent refiring at an anode potential of 200 volts. In Figure 7b, for the same anode voltage and recovery time, the grid–cathode voltage required to prevent refiring is now only —38 volts.

Figure 8 shows similar curves for peak anode currents of 1 ampere and 0.5 ampere. Note that as peak anode current decreases, the required grid–cathode voltage for a given recovery time also decreases. It has been found that anode currents of less than 0.5 ampere have little effect on critical grid–cathode voltage. The curve for 0.5 ampere may, therefore, be used for lower currents.

Figure 9 shows the decay of after-glow current with time. These curves make it possible to determine the instantaneous ion grid current at any time during deionization. The actual grid–cathode voltage may be determined by circuit analysis. This analysis applies even though the resultant grid–cathode voltage sometimes appears to go positive.

The magnitude of the positive-ion current flowing to the grid during conduction is the magnitude at zero deionization time. This current can be used to determine the grid–cathode voltage during the conduction period; it should not exceed the —10-volt rating for the 2D21. An example illustrating the use of the data presented is given below.

The circuit under consideration is shown in Figure 10. The known parameters are: anode-supply voltage $(E_{bb}) = 200$ volts, grid-supply

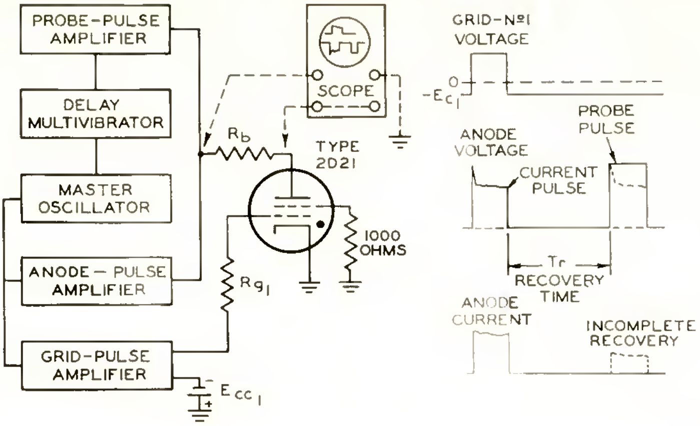


Fig. 6—(a) Circuit used to measure grid-recovery time of the 2D21; (b) and (c) anode and grid-pulse waveforms.

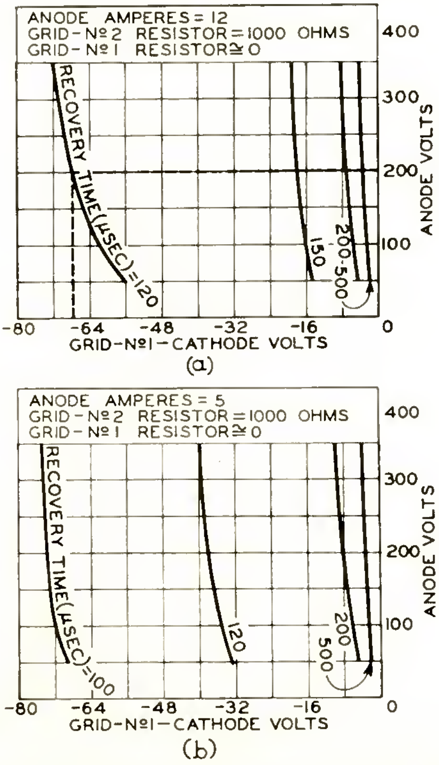


Fig. 7—Critical grid–cathode voltage as a function of recovery time, anode voltage, and R_{g2}, for peak anode currents of 12 and 5 amperes.

voltage (E_{cc1}) = −75 volts, peak anode current (i_b) = 12 amperes, ionization time (T_i) = 5 microseconds max., grid-recovery time (T_r) = 120 microseconds max.

The problem is to determine the maximum value of R_{g1}, and the minimum pulse-source voltage E_s required to fire the tube within 5 microseconds. The value of R_{g1} is determined from the data given in Figures 7a and 9. In Figure 7a, the intersection of the curve for $T_r = 120$ microseconds with the coordinate $E_{bb} = 200$ volts shows that a grid–cathode voltage of −68 volts at the instant E_{bb} is reapplied will prevent firing. In Figure 9, the intersection of the curve for $i_b = 12$ amperes with the 120-microsecond coordinate shows that the after-glow current flowing in the grid circuit 120 microseconds after conduction ceases is 0.15 milliampere.

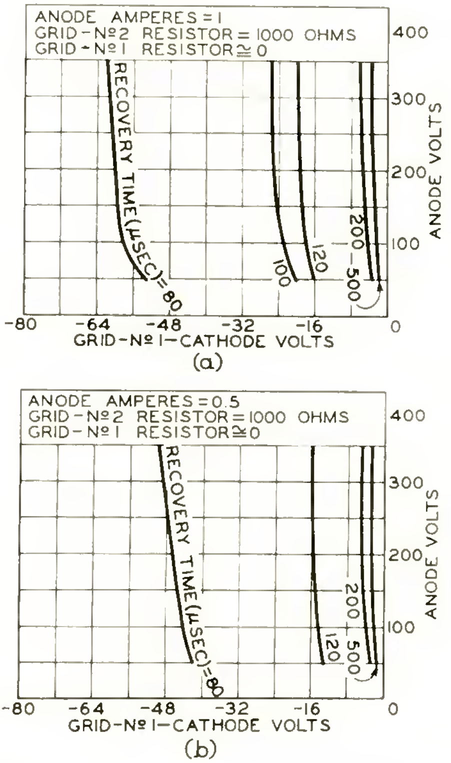


Fig. 8—Critical grid–cathode voltage as a function of recovery time, anode voltage, and R_{g2}, for peak anode currents of 1 and 0.5 ampere.

From these values ($E_{g1k} = -68$ volts, and $i_{g1} = 0.15$ milliampere) the maximum value for R_{g1} is determined as follows:

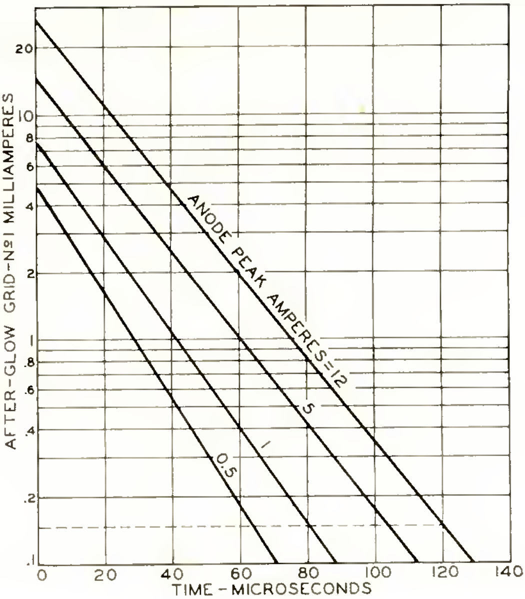


Fig. 9—"After-glow" grid-current as a function of recovery time and peak anode current.

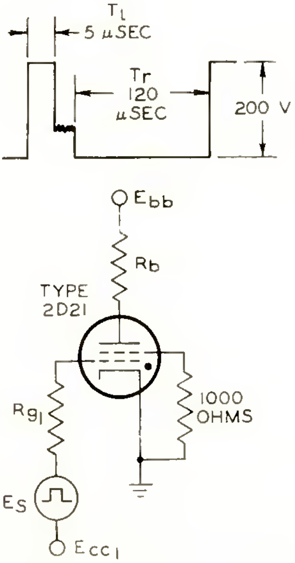


Fig. 10—Basic circuit showing parameters involved in sample calculation of R_{g1} and E_{s1}.

$$E_{g1k} = E_{cc1} - i_{g1}R_{g1},$$

$$R_{g1} = \frac{E_{cc1} - E_{g1k}}{i_{g1}}$$

$$= \frac{75 - 68}{0.15 \times 10^{-3}} \approx 46{,}500 \text{ ohms.}$$

The next unknown to be determined is the minimum pulse-source voltage E_s required to fire the tube within 5 microseconds. This voltage is found from the data shown in Figures 3a and 4. In Figure 3a, the intersection of the curve for $T_i = 5$ microseconds with the $E_{bb} = 200$ volts coordinate shows that a grid–cathode voltage of +17 volts is required to fire the tube. In Figure 4, the intersection of the curve for $E_{g2} = 0$ volts with the +17-volt coordinate shows that the prefiring grid current is 0.29 milliampere.

From these values ($E_{g1k} = +17$ volts, $i_{g1} = 0.29$ milliampere, $R_{g1} = 46{,}500$ ohms, and $E_{cc1} = -75$ volts) the pulse-source voltage E_s is found as follows:

$$E_s = E_{g1k} + i_{g1}R_{g1} + E_{cc1}$$

$$= 17 + (0.29 \times 10^{-3} \times 46{,}500) + 75 = 106 \text{ volts.}$$

Conclusion

When the 2D21 is to be used in applications requiring short ionization and recovery times, the external circuit parameters must be carefully chosen. The important parameter is the grid–cathode voltage. To determine this voltage, knowledge of the grid currents, and careful analysis of the external grid circuits are necessary. Generally grid impedances should be kept at a minimum. Large grid signals and bias voltage may be required.

Ionization times as short as 1 microsecond may be achieved in the 2D21 by pulsing the grid sufficiently positive. Because appreciable grid current flows under these conditions, the effective impedance of the signal source becomes a very important consideration. This impedance attenuates the pulse and should, therefore, be kept as low as possible, preferably below 50,000 ohms.

Grid control in the 2D21 can be recovered in about 100 microseconds if the circuit is properly designed. The important parameters affecting recovery time are anode current prior to extinction, anode voltage, and the grid–cathode voltage. The grid–cathode voltage is determined by the after-glow grid current, grid-circuit impedance, and bias voltage. A grid-circuit impedance below 50,000 ohms and large bias supply voltages are recommended.

The data presented covers the ranges of operating conditions normally employed for the 2D21, and should enable the design engineer to predict the end performance of his circuit with reasonable accuracy.

RCA TECHNICAL PAPERS†

First Quarter, 1957

Any request for copies of papers listed herein should be addressed to the publication to which credited.

"Analysis and Synthesis of Transitional Butterworth-Thomson Filters and Bandpass Amplifiers," Y. Peless and T. Murakami, *RCA Review* (March)	1957
"Automatic Conelrad Alarm Provides Constant Guard for Conelrad 'Alert'," G. D. Hanchett, *RCA Ham Tips* (February)	1957
"Automatic Gain Control For Monochrome Video," A. H. Turner, *Broadcast News* (February)	1957
"Automatic Level Control for Film Systems," W. L. Hurford, *Trans. I.R.E. PGBTS* (February)	1957
"Automation Re-Examined," J. J. Graham, *Trans. I.R.E. PGIE* (March)	1957
"Basic Logic Circuits for Computer Applications," G. W. Booth and T. P. Bothwell, *Electronics* (March)	1957
"Calculation of the Parameters of Ridge Waveguides," Tsung-Shan Chen, *Trans. I.R.E.* **PGMTT** (January)	1957
"Calculations of the Risk of Component Applications in Electronic Systems," J. A. Connor, *Trans. I.R.E. PGRQC* (January)	1957
"CdS-Type Photoconductivity in ZnTe Crystals," R. H. Bube and E. L. Lind, *Phys. Rev.* (March 15)	1957
"Color-TV Video Generator for Convergence," R. Samuel, *Service* (January)	1957
"Confined Electron Flow in Periodic Electrostatic Fields of Very Short Periods," K. K. N. Chang, *Proc. I.R.E.* (January)	1957
"Constant Current Generator," H. S. Sommers, Jr., *Rev. Sci. Instr.* (Notes) (March)	1957
"Current Steering in Magnetic Circuits," J. A. Rajchman and H. D. Crane, *Trans. I.R.E. PGEC* (March)	1957
"Design, Construction, and High-Frequency Performance of Drift Transistors," A. L. Kestenbaum and N. H. Ditrick, *RCA Review* (March)	1957
"The Design of Periodic Permanent Magnets for Focusing of Electron Beams," F. Sterzer and W. W. Siekanowicz, *RCA Review* (March)	1957
"Ear-Insert Microphone," R. D. Black, *Jour. Acous. Soc. Amer.* (February)	1957
"Effect of a Negative Impedance Source on Loudspeaker Performance," R. E. Werner, *Jour. Acous. Soc. Amer.* (March)	1957
"Effect of Annealing in Various Gases on the Bulk Lifetime of Germanium," K. Weiser, *Jour. Appl. Phys.* (February)	1957
"The Effect Of Electrical Loading On Microphone Response," R. E. Werner, *Broadcast News* (February)	1957
"Extension of Babinet's Principle to Absorbing and Transparent Materials, and Approximate Theory of Backscattering by Plane, Absorbing Disks," H. E. J. Neugebauer, *Jour. Appl. Phys.* (March)	1957

† Report all corrections or additions to *RCA Review*, RCA Laboratories, Princeton, N. J.

"Ferrite Apertured Plate for Random Access Memory," J. A. Rajchman, *Proc. I.R.E.* (March) 1957

"'Immittance' New Tool For Network Analysis," H. M. Wasson, *Electronic Industries* (February) 1957

"A Linear-Logarithmic Amplifier for Ultra-Short Pulses," H. Kihn and W. E. Barnette, *RCA Review* (March) 1957

"Low-Field Electroluminescence in Insulating Crystals of Cadmium Sulfide," R. W. Smith, *Phys. Rev.* (February 1) 1957

"Miniature ITV Camera Uses Drift Transistors," L. E. Flory, G. W. Gray, J. M. Morgan, and W. S. Pike, *Electronics* (January) 1957

"A Mobile Control Room For Color TV," H. H. Klerx, *Broadcast News* (February) 1957

"Molded Expanded Polystyrene," R. Kuhlman, *Modern Packaging* (February) 1957

"Molecular Amplification and Generation of Microwaves," J. P. Wittke, *Proc. I.R.E.* (March) 1957

"One-Dimensional Treatment of the Effective Mass in Semiconductors," I. Adawi, *Phys. Rev.* (February 1) 1957

"Operations Research as a Managerial Instrument of Advice and Decision," Z. Prihar, *Trans. I.R.E. PGEM* (March) 1957

"Operations Research: The Inventory Problem," Z. Prihar, *Trans. I.R.E. PGEM* (March) 1957

"Portable TV Station for Remote Pickups," L. E. Flory, G. W. Gray, J. M. Morgan, and W. S. Pike, *Electronics* (February) 1957

"Prediction of Tube Failure Rate Variations," M. P. Feyerherm, *Trans. I.R.E. PGRQC* (January) 1957

"Production Testing in the Automatic Factory," H. S. Dordick, *Trans. I.R.E. PGIE* (March) 1957

"Properties of Some Low-Temperature Solder Glasses," J. Gallup and A. G. F. Dingwall, *The American Ceramic Society Bulletin* (February 15) 1957

"Reduction of Co-Channel Television Interference by Precise Frequency Control of Television Picture Carriers," W. L. Behrend, *Trans. I.R.E. PGBTS* (February) 1957

"Reflection Point and Angle Nomographs," Z. Prihar, *Electronics* (Reference Sheet) (January) 1957

"Simple High-Temperature Vacuum-Tight Mica Window," F. Sterzer, *Rev. Sci. Instr.* (Notes) (March) 1957

"Situations that Affect the Productivity of Engineers," M. C. Batsel, *Trans. I.R.E. PGEM* (March) 1957

"Some Practical Aspects of High-Fidelity Design," Roy S. Fine, *Trans. I.R.E. PGA* (March-April) 1957

"Space-Charge Limitation on the Focus of Electron Beams," J. W. Schwartz, *RCA Review* (March) 1957

"Systems Engineering—A Growing Concept," E. W. Engstrom, *Elec. Eng.* (February) 1957

"Test Chart Transparencies Speed Matching Of Live TV Cameras," E. P. Bertero, *Broadcast News* (February) 1957

"Thermionic Current in a Parallel-Plane Diode," L. J. Giacoletto, *Trans. I.R.E. PGED* (January) 1957

"Thermistors Compensate Transistor Amplifiers," A. J. Wheeler, *Electronics* (January) 1957

"Validity of Traveling-Wave-Tube Noise Theory," W. R. Beam and R. C. Knechtli, *RCA Review* (March) 1957

"Walter Reed Army Medical Center Uses RCA Compatible Color TV System For Medical Education," *Broadcast News* (February) 1957

Correction:

In the paper entitled "Validity of Traveling-Wave-Tube Noise Theory," by W. R. Beam and R. C. Knechtli, which appeared on pages 24-38 of the March 1957 issue, the following information was inadvertently omitted:

The authors wish to thank R. W. Peter for his guidance.

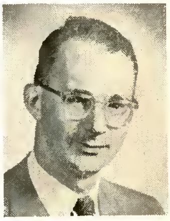
Part of the work described in this paper was performed under a Signal Corps Contract.

AUTHORS

LARRY A. FREEDMAN received the B.S. degree in Electrical Engineering from Drexel Institute of Technology in 1948. From 1948 to 1950 he was a research assistant in the Electrical Engineering Department of Rutgers University. He received the M.S. degree from Rutgers University in 1950. In 1950 he joined the RCA Laboratories in Princeton, N. J., where he is working on transistor circuit applications. Mr. Freedman is a member of the Institute of Radio Engineers, Phi Kappa Phi, Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.

R. W. GEORGE received the B.S. degree in Electrical Engineering from Kansas State College in 1928. On graduation he joined the Radio Corporation of America at Riverhead, N. Y., where he worked in the fields of high frequency propagation, measuring equipment radio relays, and other phases of communications. He transferred to the David Sarnoff Research Center of RCA Laboratories at Princeton, N. J. in 1951, where he has been concerned mostly with the development of electromechanical filters. Mr. George is an affiliate of Sigma Xi and a Senior Member of the Institute of Radio Engineers.

B. GOLDSTEIN received the B.S. degree from Brooklyn College in 1949. He then attended the Polytechnic Institute of Brooklyn where he held teaching and research fellowships, and from which he received an M.S. in 1951 and a Ph.D. in 1955. His research while at Brooklyn Polytechnic was on the infrared properties of ZnS phosphors. He joined the semiconductor physics group at the RCA Laboratories in Princeton in 1954, where he has been making basic studies on p-n junction and atomic diffusion in semiconductors. Dr. Goldstein is a member of the American Physical Society and Sigma Xi.



RICHARD E. HONIG received the B.S. degree in Electrical Engineering from Robert College, Istanbul, Turkey, in 1938; the M.S. degree in 1939, and the Ph.D. degree in Physics in 1944, both from the Massachusetts Institute of Technology. In 1940-1941, he was instructor of Physics at Bluffton College, Ohio. He taught and did research work at M.I.T. as Research Assistant from 1941 to 1944, and as Research Associate from 1944 to 1946. He was at the Socony Vacuum Research Laboratories, Paulsboro, N. J., from 1946 to 1950, working mainly in the field of mass spectrometry. Since 1950, he has been at RCA Laboratories, Princeton, N. J., engaged in fundamental research in solid-state physics. In 1955/6, on a year's leave of absence from RCA Laboratories, he studied vaporization phenomena at the University of Brussels, Belgium. Dr. Honig is a member of Sigma Xi and the American Physical Society.

Charles W. Mueller received the B.S. degree in Electrical Engineering from the University of Notre Dame in 1934, the S.M. degree in Electrical Engineering in 1936 from the Massachusetts Institute of Technology, and the degree of Sc.D. in Physics from M.I.T. in 1942. From 1936 to 1938 he was associated with the Raytheon Production Corporation, first in the engineering supervision of factory production of receiving tubes, and then in the development of gas-tube voltage regulators and cold-cathode thyratrons. From 1938 to 1942 he worked at M.I.T. on the development of gas-filled special-purpose tubes for counting operations. Since 1942 he has been a member of the technical staff of RCA Laboratories in Princeton, N. J. where he has been engaged in research on high-frequency receiving tubes, secondary electron emission phenomena, and solid-state devises. Dr. Mueller is a member of the American Physical Society and Sigma Xi and a Senior Member of the Institute of Radio Engineers.

John A. Olmstead received the B.S. degree in Electrical Engineering from the University of Buffalo in 1952 and the M.S. degree in Electrical Engineering from Newark College of Engineering in 1957. He joined the Radio Corporation of America as a specialized trainee in 1952. He has since been engaged in the design and development of special purpose tubes, particularly gas filled types, in the receiving tube design activity at Harrison, New Jersey.

Fritz Paschke received the degree of Diplom-Ingenieur in 1953 and the Dr. techn. sc. in 1955 from the Technical University Vienna. Between 1953 and 1955 he was a research assistant at the Institute of High-Frequency Techniques in Vienna. Since April 1956, Dr. Paschke has been a member of the technical staff at RCA Laboratories, Princeton, N. J.

Melvin Roth received the B.S. degree in Electrical Engineering from Carnegie Institute of Technology in 1955. He then joined the Radio Corporation of America as a specialized trainee. Since November 1955 he has been a member of the Industrial Receiving Tube Application Laboratory in Harrison, New Jersey, specializing in gas filled devices.

Edward R. Schrader received the B.A. degree in Physics from Columbia University in 1950 and the M.S. degree in Physics in 1953 from the Polytechnic Institute of Brooklyn. He is now working toward a Ph.D. degree. He was engaged in oceanographic research for the American Museum of Natural History from 1950 to 1951. As an engineer for the Sperry Gyroscope Company he worked on aircraft instrument research and development from 1951 to 1954. Since 1954 he has been with the Receiving-Tube Chemical and Physical Laboratory of the RCA Electron-Tube Division at Harrison. His work is chiefly concerned with the investigation of cathode and grid problems. Mr. Schrader is a member of the American Physical Society and of the Society of the Sigma Xi.

JACK SKLANSKY received the B.E.E. degree from the City College of New York in 1950, the M.S.E.E. degree from Purdue University in 1952, and the Doctor of Engineering Science degree from Columbia University in 1955. While at Purdue he served as a teaching assistant. At Columbia he was a Eugene Higgins fellow from 1952 to 1954 and a research assistant at the Electronics Research Laboratory during two summers and 1954-55. During previous summers he was associated with the U. S. Bureau of Reclamation, the U. S. Naval Ordnance Laboratory, and the Bell Telephone Laboratories. Since 1955, he has been a member of the technical staff of the David Sarnoff Research Center, where he has been investigating problems in satellite instrumentation, missile dynamics, and radar-data processing. Dr. Sklansky is a member of the Institute of Radio Engineers, the American Institute of Electrical Engineers, Tau Beta Pi, Sigma Xi, and Eta Kappa Nu.

J. TORKEL WALLMARK received the Civilingenjor degree in Electrical Engineering from Royal Institute of Technology, Stockholm, Sweden, in 1944, the Teknologie Licentiat degree in 1947 and the Teknologie Doktor degree in 1953. He was a research assistant with the Institute from 1945-1953 and also held positions with various other organizations active in research in electron tube and semiconductor problems. From 1947 to 1948 Dr. Wallmark was a fellow of the American-Scandinavian Foundation at RCA Laboratories, to which he returned in 1953; he is presently working on semiconductor surface problems.