Space Programs Summary No. 37-29, Volume IV

for the period August 1, 1964 to September 30, 1964

Supporting Research and Advanced Development

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

October 31, 1964

Space Programs Summary No. 37-29, Volume IV

for the period August 1, 1964 to September 30, 1964

Supporting Research and Advanced Development

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

October 31, 1964

Preface

The *Space Programs Summary* is a six volume, bimonthly publication designed to report on JPL space exploration programs, and related supporting research and advanced development projects. The subtitles of all volumes of the *Space Programs Summary* are:

Vol. I. The Lunar Program (Confidential)

Vol. II. The Planetary–Interplanetary Program (Confidential)

Vol. III. The Deep Space Network (Unclassified)

Vol. IV. Supporting Research and Advanced Development (Unclassified)

Vol. V. Supporting Research and Advanced Development (Confidential)

Vol. VI. Space Exploration Programs and Space Sciences (Unclassified)

The *Space Programs Summary,* Volume VI consists of an unclassified digest of appropriate material from Volumes I, II, and III and a reprint of the space science instrumentation studies of Volumes I and II.

W. H. Pickering, *Director*
Jet Propulsion Laboratory

Space Programs Summary No. 37-29, Volume IV

Prepared under Contract No. NAS 7-100, National Aeronautics & Space Administration

Contents

SYSTEMS DIVISION

GUIDANCE AND CONTROL DIVISION

ENGINEERING MECHANICS DIVISION

Contents (Cont'd)

Contents (Cont'd)

Contents (Cont'd)

Contents (Cont'd)

TELECOMMUNICATIONS DIVISION

Contents (Cont'd)

SYSTEMS DIVISION

I. Systems Analysis

A. Equations of Motion for a Double-Precision Trajectory Program

F. M. Sturms, Jr.

As part of a general effort to develop double-precision trajectory computer programs, the basic subroutines of existing programs are being examined to determine their suitability for coding in double precision. This article, presenting one of the typical investigations, is illustrative of the type of modification that must be incorporated in order to achieve consistent double-precision accuracy. Articles will be appearing in future issues of the *SPS*, Vol. IV, giving important results in the development of double-precision trajectory programs.

In Ref. 1, Cowell equations of motion of a probe are derived in coordinates with respect to any one of n bodies constituting an n-body gravity field. In the transformation from an inertial origin to the selected central body, terms are added giving the accelerations on the central body due to the gravitational attractions of the non-central bodies. The Earth and Moon are considered to be non-spherical when the direct accelerations on the probe are determined; however, only the point mass effect is included in the accelerations on the central body due to the other bodies.

It is felt that the attraction between the Earth and Moon should include the first-order oblateness terms for trajectories computed in double precision. This article presents the derivation of equations of motion including these terms, based on relationships from Chapter III of Ref. 2.

1. Basic Equations of Motion

Define an inertial Cartesian coordinate system $R(X, Y, Z)$ in which the axes are parallel to the 1950.0 mean equinox and equator coordinate system. Let X_0, Y_0, Z_0 be the coordinates of the probe of mass M_0, and let X_i, Y_i, Z_i $(i = 1, \cdots, N)$ be the coordinates of N bodies of mass M_i. Further specify that M_1 be the Earth and M_2 the Moon. Let the force potential between any two bodies be denoted by U_{ij} such that the components of force on M_i due to M_j are given by

$$F_{r_{ij}} = \frac{\partial U_{ij}}{\partial X_i} \tag{1}$$

and similar equations in Y and Z. Then, according to Newton's second law,

$$\ddot{X}_i = \frac{1}{M_i} \sum_{\substack{j=0 \\ i \neq j}}^{N} \frac{\partial U_{ij}}{\partial X_i}, \qquad i = 0, 1, \cdots, N, \tag{2}$$

are the equations of motion of the $N + 1$ bodies in the inertial coordinate system.

Let a parallel coordinate system $r(x, y, z)$ be defined as being centered in one of the N bodies, M_c. Then,

$$\bar{r}_j = \bar{r}_{cj} = \bar{R}_j - \bar{R}_c \tag{3}$$

represents the radius vector from the central body to the jth body. For the probe, we then have (neglecting the acceleration of the probe on the central body):

$$\begin{aligned} \ddot{x}_0 &= \ddot{X}_0 - \ddot{X}_c \qquad (4) \\ &= \frac{1}{M_0} \sum_{j=1}^{N} \frac{\partial U_{0j}}{\partial X_0} - \frac{1}{M_c} \sum_{\substack{j=0 \\ j \neq c}}^{N} \frac{\partial U_{cj}}{\partial X_c} \\ &= \frac{1}{M_0} \frac{\partial U_{0c}}{\partial X_0} + \sum_{\substack{j=1 \\ j \neq c}}^{N} \left(\frac{1}{M_0} \frac{\partial U_{0j}}{\partial X_0} - \frac{1}{M_c} \frac{\partial U_{cj}}{\partial X_c} \right) \\ &= \frac{1}{M_0} \frac{\partial U_{0c}}{\partial X_0} + \sum_{\substack{j=1 \\ j \neq c}}^{N} \left(\frac{1}{M_0} \frac{\partial U_{0j}}{\partial X_0} + \frac{1}{M_c} \frac{\partial U_{cj}}{\partial X_j} \right) \end{aligned}$$

and similarly for y_0 and z_0.

Now, let the potential be expressed as the point mass term plus the non-spherical term:

$$U_{ij} = \frac{GM_iM_j}{R_{ij}} + U'_{ij}. \tag{5}$$

The non-spherical term of interest in this article appears in the Earth–Moon potential:

$$U_{12} = \frac{GM_1M_2}{R_{12}} + U'_{12}. \tag{6}$$

In practice, the coordinates x_i, y_i, z_i are used rather than X_i, Y_i, Z_i. From the relations:

$$X_j - X_i = x_j - x_i, \tag{7}$$

$$\frac{\partial U_{ij}}{\partial X_i} = \frac{\partial U_{ij}}{\partial x_i}, \tag{8}$$

$$R_{ij} = [(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2]^{1/2}, \tag{9}$$

Eq. (4) becomes

$$\begin{aligned} \ddot{x}_0 = &- GM_c \frac{x_0}{r_0^3} - \sum_{\substack{j=1 \\ j \neq c}}^{N} GM_j \left(\frac{x_0 - x_j}{r_{0j}^3} + \frac{x_j}{r_j^3} \right) \\ &+ \frac{1}{M_0} \frac{\partial U'_{0c}}{\partial x_0} \\ &+ \sum_{\substack{j=1 \\ j \neq c}}^{N} \frac{1}{M_0} \frac{\partial U'_{0j}}{\partial x_0} \qquad \text{(for all non-zero } U'_{0j}\text{)} \\ &+ \frac{1}{M_c} \frac{\partial U'_{12}}{\partial x_i}, \end{aligned} \tag{10}$$

where

$i = 1$ if the central body is the Moon

$= 2$ if the central body is the Earth

and the last term in Eq. (10) is zero if the central body is neither the Earth nor Moon. Eq. (10) above may be compared to Eq. (3) of Ref. 1. The first line is identical to the form of Eq. (3) of Ref. 1, and the next two lines are equivalent to oblate perturbation terms discussed on p. 5 of Ref. 1. The important difference between the two equations is the last term above, which is due to the mutual attraction of the Earth and Moon.

2. Derivation of U'_{12}

Let dM_1 and dM_2 be differential elements of the mass of the Earth and Moon, respectively. Define parallel coordinate systems (ξ, η, ζ) and (ξ', η', ζ') centered in the Earth and Moon, respectively, such that the center of mass of the Moon is located on the ξ-axis at a distance r_{12}. The distance between the mass elements is then given by

$$\begin{aligned} \Delta^2 &= (r_{12} + \xi' - \xi)^2 + (\eta' - \eta)^2 + (\zeta' - \zeta)^2 \\ &= r_{12}^2 \left[1 - \frac{2(\xi - \xi')}{r_{12}} \right. \\ &\quad \left. + \frac{(\xi - \xi')^2 + (\eta - \eta')^2 + (\zeta - \zeta')^2}{r_{12}^2} \right]. \end{aligned} \tag{11}$$

Letting

$$q = \frac{\xi - \xi'}{r_{12}\alpha}, \tag{12}$$

$$\alpha^2 = \frac{(\xi - \xi')^2 + (\eta - \eta')^2 + (\zeta - \zeta')^2}{r_{12}^2}, \tag{13}$$

Eq. (11) becomes

$$\Delta^2 = r_{12}^2 (1 - 2q\alpha + \alpha^2). \tag{14}$$

The element of force potential, by Newton's law of gravity, is

$$dU_{12} = \frac{GdM_1dM_2}{\Delta}$$

$$= \frac{GdM_1dM_2}{r_{12}} (1 - 2q\alpha + \alpha^2)^{-1/2}. \tag{15}$$

The quantity raised to the half power may be recognized as the generating function of a power series in α, with coefficients consisting of Legendre polynomials in q.

$$dU_{12} = \frac{GdM_1dM_2}{r_{12}} [1 + P_1(q)\alpha + P_2(q)\alpha^2 + \cdots]. \tag{16}$$

Then, since $\alpha < 1$ and the series is convergent,

$$U_{12} = \int_{M_2}\int_{M_1} \frac{GdM_1dM_2}{r_{12}} [1 + P_1(q)\alpha + P_2(q)\alpha^2 + \cdots]$$

$$= U_{12}^{(0)} + U_{12}^{(1)} + U_{12}^{(2)} + \cdots. \tag{17}$$

Evaluating the Legendre polynomials, we have

$$U_{12}^{(0)} = \int_{M_2}\int_{M_1} \frac{GdM_1dM_2}{r_{12}} = \frac{GM_1M_2}{r_{12}}, \tag{18}$$

which is the point mass term (see Eq. 6):

$$U_{12}^{(1)} = \int_{M_2}\int_{M_1} \frac{G\alpha q}{r_{12}} dM_1dM_2$$

$$= \iint G\frac{\xi - \xi'}{r_{12}^2} dM_1dM_2 = 0 \tag{19}$$

since ξ and ξ' are measured from the centers of mass.

$$U_{12}^{(2)} = \int_{M_2}\int_{M_1} \frac{G}{r_{12}} \frac{1}{2} (3q^2 - 1)\alpha^2 dM_1dM_2$$

$$= \int_{M_2}\int_{M_1} \frac{G}{r_{12}^3} \left\{ \frac{3}{2}(\xi - \xi')^2 - \frac{1}{2}[(\xi - \xi')^2 + (\eta - \eta')^2 + (\zeta - \zeta')^2] \right\} dM_1dM_2$$

$$= \frac{G}{r_{12}^3} \iint \left[\left(\xi^2 - \frac{1}{2}\eta^2 - \frac{1}{2}\zeta^2 \right) + \left(\xi'^2 - \frac{1}{2}\eta'^2 - \frac{1}{2}\zeta'^2 \right) - 2\xi\xi' + \eta\eta' + \zeta\zeta' \right] dM_1dM_2. \tag{20}$$

In Eq. (20), the product terms $(-2\xi\xi' + \eta\eta' + \zeta\zeta')$ yield nothing due to the fact that the coordinate systems are at the centers of mass. The remaining terms may be regrouped as follows:

$$U_{12}^{(2)} = \frac{GM_2}{r_{12}^3} \int_{M_1} \left[\xi^2 + \eta^2 + \zeta^2 - \frac{3}{2}(\eta^2 + \zeta^2) \right] dM_1 + \frac{GM_1}{r_{12}^3} \int_{M_2} \left[\xi'^2 + \eta'^2 + \zeta'^2 - \frac{3}{2}(\eta'^2 + \zeta'^2) \right] dM_2. \tag{21}$$

Then,

$$\left.\begin{aligned} \int (\xi^2 + \eta^2 + \zeta^2)\, dM_1 &= \frac{1}{2}(A + B + C), \\ \int (\xi'^2 + \eta'^2 + \zeta'^2)\, dM_2 &= \frac{1}{2}(A' + B' + C'), \\ \int (\eta^2 + \zeta^2)\, dM_1 &= I, \\ \int (\eta'^2 + \zeta'^2)\, dM_2 &= I', \end{aligned}\right\} \tag{22}$$

where A, B, and C are the moments of inertia of the Earth about the principal axes and I is the moment of inertia of the Earth about the ξ-axis, and similarly for the Moon using the primed notation. Then, to first order,

$$U'_{12} = U^{(2)}_{12} = \frac{GM_2}{r^3_{12}} \frac{1}{2} (A + B + C - 3I)$$

$$+ \frac{GM_1}{r^3_{12}} \frac{1}{2} (A' + B' + C' - 3I'). \quad (23)$$

The quantity I is given in terms of the principal moments of inertia by

$$I = Aa^2 + Bb^2 + Cc^2, \quad (24)$$

where a, b, and c are the direction cosines of the ξ-axis with respect to the principal axes. A similar expression is written for I'. The coordinates along the principal axes of the Earth may be taken as the true equinox and equator of date coordinates, denoted by $\bar{x}$, $\bar{y}$, $\bar{z}$. Then,

$$r_{12} = (\bar{x}^2_2 + \bar{y}^2_2 + \bar{z}^2_2)^{1/2}, \quad (25)$$

$$\left.\begin{aligned} a &= \frac{\bar{x}_2}{r_{12}}, \\ b &= \frac{\bar{y}_2}{r_{12}}, \\ c &= \frac{\bar{z}_2}{r_{12}}. \end{aligned}\right\} \quad (26)$$

However, r_{12}, a, b, and c must be expressed in terms of the 1950.0 coordinate system in order to obtain the partial derivatives indicated in Eq. (10). The relationship between the two systems is given by

$$\begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = NA \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad (27)$$

where the elements of the rotation matrix NA (denoted by a_{ij}) are obtained from the precession and nutation, as described on pp. 66–70 of Ref. 1. Then,

$$\left.\begin{aligned} a &= \frac{a_{11} x_2 + a_{12} y_2 + a_{13} z_2}{r_{12}}, \\ b &= \frac{a_{21} x_2 + a_{22} y_2 + a_{23} z_2}{r_{12}}, \\ c &= \frac{a_{31} x_2 + a_{32} y_2 + a_{33} z_2}{r_{12}}, \end{aligned}\right\} \quad (28)$$

$$r_{12} = (x^2_2 + y^2_2 + z^2_2)^{1/2}. \quad (29)$$

A similar analysis can be performed for the Moon, yielding

$$a' = -\frac{\bar{x}'_1}{r_{12}}$$

$$= \frac{m_{11} x_2 + m_{12} y_2 + m_{13} z_2}{r_{12}}, \quad (30)$$

$$b' = -\frac{\bar{y}'_1}{r_{12}}$$

$$= \frac{m_{21} x_2 + m_{22} y_2 + m_{23} z_2}{r_{12}}, \quad (31)$$

$$c' = -\frac{\bar{z}'_1}{r_{12}}$$

$$= \frac{m_{31} x_2 + m_{32} y_2 + m_{33} z_2}{r_{12}}, \quad (32)$$

where $(\bar{x}', \bar{y}', \bar{z}')$ are Moon-centered coordinates along the Moon's principal axes (true equator). The m_{ij} are elements of the rotation matrix from the 1950.0 system to the Moon's true equator of date:

$$\begin{pmatrix} \bar{x}' \\ \bar{y}' \\ \bar{z}' \end{pmatrix} = M \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix}, \quad (33)$$

as described on pp. 78 and 79 of Ref. 1. The sign reversals arise from the relations

$$\left.\begin{aligned} x'_1 &= -x_2, \\ y'_1 &= -y_2, \\ z'_1 &= -z_2. \end{aligned}\right\} \quad (34)$$

U'_{12} may be written as a function of x'_1 (for the Moon as the central body) or x_2 (for the Earth as the central body) by using Eqs. (34) in Eqs. (28) through (32). Equivalently, in Eq. (10) one may utilize the relationship:

$$\frac{\partial U'_{12}}{\partial x_1} = -\frac{\partial U'_{12}}{\partial x_2}. \quad (35)$$

3. Partial Derivatives of U'_{12}

The partial derivatives of U'_{12} with respect to x, y, and z as indicated in Eq. (10) may now be obtained explicitly by means of Eqs. (23) through (25), (28) through (32), and (34). For the Earth as the central body, we have for the last term of Eq. (10):

$$\frac{1}{M_1}\frac{\partial U'_{12}}{\partial x_2} = -\frac{3}{2}\frac{GM_2}{r_{12}^4}\left[\left(\frac{A+B+C-3I}{M_1}\right)\frac{\partial r_{12}}{\partial x_2} + \left(\frac{A'+B'+C'-3I'}{M_2}\right)\frac{\partial r_{12}}{\partial x_2} + \frac{1}{M_1}\frac{\partial I}{\partial x_2}r_{12} + \frac{1}{M_2}\frac{\partial I'}{\partial x_2}r_{12}\right]. \quad (36)$$

Now, from Eq. (29),

$$\frac{\partial r_{12}}{\partial x_2} = \frac{x_2}{r_{12}}. \quad (37)$$

From Eqs. (24), (28), and (37),

$$\frac{\partial I}{\partial x_2} = 2Aa\left[\frac{a_{11}}{r_{12}} - \frac{ax_2}{r_{12}^2}\right] + 2Bb\left[\frac{a_{21}}{r_{12}} - \frac{bx_2}{r_{12}^2}\right] + 2Cc\left[\frac{a_{31}}{r_{12}} - \frac{cx_2}{r_{12}^2}\right], \quad (38)$$

and similarly

$$\frac{\partial I'}{\partial x_2} = 2A'a'\left[\frac{m_{11}}{r_{12}} - \frac{a'x_2}{r_{12}}\right] + 2B'b'\left[\frac{m_{21}}{r_{12}} - \frac{b'x_2}{r_{12}}\right] + 2C'c'\left[\frac{m_{31}}{r_{12}} - \frac{c'x_2}{r_{12}}\right]. \quad (39)$$

For obtaining the partials with respect to y_2 and z_2, the following changes may be made in Eqs. (36) through (39):

$$x_2 \rightarrow y_2, z_2,$$

$$a_{j1} \rightarrow a_{j2}, a_{j3},$$

$$m_{j1} \rightarrow m_{j2}, m_{j3}.$$

If the Moon is the central body, the last term of Eq. (10) is obtained from Eq. (35) as

$$\frac{1}{M_2}\frac{\partial U'_{12}}{\partial x'_1} = \frac{3}{2}\frac{GM_1}{r_{12}^4}\left[\quad\right], \quad (40)$$

where the terms in the brackets are the same as those obtained in Eq. (36). It should be noted that, in evaluating the bracketed terms, x_2, y_2, and z_2 are the coordinates of the Moon with respect to the Earth, regardless of whether Eq. (36) or Eq. (40) is being evaluated.

For obtaining the attractions between a non-spherical body and the probe, which is considered a point mass, the potential is expressed in terms of spherical harmonics.

$$U'_{0j} = \frac{GM_0M_j}{r_{0j}}\left[-\sum_{n=2}^{N} J_n\left(\frac{R_m}{r_{0j}}\right)^n P_n(\sin\phi) + \sum_{n=2}^{N}\sum_{m=1}^{n}\left(\frac{R_m}{r_{0j}}\right)^n P_n^m(\sin\phi) \times (C_{n,m}\cos m\lambda + S_{n,m}\sin m\lambda)\right]. \quad (41)$$

The moments of inertia about the principal axes of a tri-axial ellipsoid may be obtained in terms of the spherical harmonic coefficients. From equations in Ref. 3:

$$\left.\begin{aligned} A &= \frac{1}{5}MR_m^2\left(2 - \frac{5}{3}J_2 - 10C_{2,2}\right),\\ B &= \frac{1}{5}MR_m^2\left(2 - \frac{5}{3}J_2 + 10C_{2,2}\right),\\ C &= \frac{1}{5}MR_m^2\left(2 + \frac{10}{3}J_2\right), \end{aligned}\right\} \quad (42)$$

where R_m is the mean radius, which, for the purpose of deriving Eqs. (42), was taken as the root-mean-square of the three principal semi-axes of the ellipsoid.

In order to be consistent with the previous assumptions about the orientation of the principal axes of the Earth, A should be equal to B, and therefore $C_{2,2}$ is zero for the Earth. For the Moon, $C_{2,2}$ may be non-zero. If $C_{2,2}$ for the Earth is taken as non-zero, then the a_{ij} matrix of Eq. (27) must be replaced by the matrix corresponding to a new $(\bar{x}, \bar{y}, \bar{z})$ coordinate system fixed in the rotating Earth, with the $\bar{x}\bar{y}$-plane the Earth's true equatorial plane, and the $\bar{x}$-axis along the major axis of the elliptical equatorial cross section analogous to the principal axes for the tri-axial Moon. The new transformation is given by:

$$\begin{pmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{pmatrix} = \begin{pmatrix} \cos(\theta + \lambda_m) & \sin(\theta + \lambda_m) & 0 \\ -\sin(\theta + \lambda_m) & \cos(\theta + \lambda_m) & 0 \\ 0 & 0 & 1 \end{pmatrix} \times NA \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \tag{43}$$

where θ is the apparent Greenwich sidereal time and λ_m is the East longitude of the major axis of the equatorial ellipse, given by

$$\tan 2\lambda_m = \frac{S_{2,2}}{C_{2,2}}. \tag{44}$$

References

1. Holdridge, D. B., *Space Trajectories Program for the IBM 7090 Computer*, Technical Report No. 32-223, Jet Propulsion Laboratory, Pasadena, California, March 2, 1962.

2. Brouwer, D., and Clemence, G. M., *Methods of Celestial Mechanics*, Academic Press, New York, 1961.

3. Clarke, V. C., Jr., *Constants and Related Data for Use in Trajectory Calculations (As Adopted by the Ad Hoc NASA Standard Constants Committee)*, Technical Report No. 32-604, Jet Propulsion Laboratory, Pasadena, California, March 6, 1964.

II. Program Engineering

A. Development of a Biological Sterility Indicator for Dry Heat Sterilization

A. Irons

One NASA sterilization requirement states that all spacecraft with a possibility of planetary impact must be sterilized with dry heat. The sterilization cycle for flight equipment is currently specified as 135 +4, −0°C (275°F) for 24 hr in an inert atmosphere such as dry nitrogen.

Biological sterilization requires the use of a test procedure which will indicate, with a high degree of reliability, whether or not a particular sterilization procedure has been successful in destroying all of the biological life associated with the materials being processed. The ultimate test procedure should be one of testing for survival of organisms by means of biological assay.

The uncontrollability of variables associated with direct biological assay of some components makes it difficult to directly examine any given spacecraft parts for the presence of viable organisms. A biological sterility indicator, then, should be one that can be subjected to the sterilization cycle in close proximity to the components while remaining divorced from the influence of these variables. This will permit control of the environment surrounding the test organism, which is mandatory for a consistent, reliable, and predictable biological sterility indicator.

The successful sterilization of a biological indicator is an indirect test and will not, in itself, prove that sterility of the spacecraft has in fact been achieved. It is merely an indicator that a particular process was reliably applied. The process applied to produce sterility must, therefore, have been previously shown to be reliable and predictable as to the probability of attaining sterility.

Certain bacterial spores, particularly in the dry state, are known to be highly resistant to adverse conditions. If a process is selected which will successfully sterilize a system embodying such resistant spore forms, and if it has been reliably shown that these spore forms are not sterilized by anything less than the stipulated process, then it can be stated that the sterilization procedure was applied with a high degree of reliability.

The Wilmot Castle Company, under contract to JPL, is attempting to develop a biological sterility indicator for a dry heat sterilization cycle of 135°C for 24 hr in an atmosphere of dry nitrogen. The timing of the cycle shall begin after all parts have reached equilibrium at 135°C. The indicator will probably be in the form of a sealed tube or ampule (carrier) which will contain the test organism plus an insulating substrate in an atmosphere of dry nitrogen.

The spacecraft design has not been established; therefore, the time for the entire spacecraft to reach sterilization temperature, as well as the time to cool to ambient after sterilization, is not known. Because of the unknown time factors, one or more of the following methods may have to be employed:

(1) Establish a set of organism kill curves (time and temperature required to kill a given microbial population) for various come-up and come-down times, for example:

Time to reach 135°C, hr	8	16	24
Time at 135°C, hr	24	24	24
Time to reach ambient, hr	4	8	12
Total	36	48	60

(2) Establish the practicability of insulating the test system so that it can be made to conform to the spacecraft component with the slowest come-up time.

(3) Use a substance to act as a thermal barrier to encase the tube or ampule containing the test organism: one which can be altered to produce various come-up times. It should have provisions for the installation of thermocouples to the internal and external portion of the test system, so that the temperature within the thermal barrier and within the test system may be determined as desired. The proposed configuration of the test system makes it mandatory to seal the thermocouples into the tube or ampule in a manner which will prevent the escape of the dry nitrogen atmosphere contained within the test system.

(4) Utilize various substrates which may be used as thermal barriers within the test tube or ampule, rather than external to it.

(5) Achieve the desired time–temperature relationship by using different combinations and variations in the number of organisms, physical makeup of the substrate, and/or physical makeup of the carrier.

In summary, this task calls for the development of a biological sterility test system for dry heat in accordance with certain specified constraints. In general, this will include use of an organism sufficiently resistant to dry heat, determination of the proper substrate and carrier, the establishment of sufficient survival–kill data to indicate and guarantee an efficient sterilization cycle at the specified time and temperature, and the establishment of a reasonable form of the system for practical use. Sufficient testing must be done to demonstrate the adequacy, stability, and reproducibility of the test system.

B. The Microbiological Profile of Clean Rooms

J. J. McDade

The probability of obtaining a sterile spacecraft by any sterilization procedure is enhanced by keeping the initial microbial contamination to a minimum level. Reduced levels of microbial contamination may be obtained through proper environmental control of the spacecraft assembly area during assembly and checkout. Such an environment can also afford the secondary benefit of improved reliability.

Decreased levels of particulate contamination have been obtained in industrial clean rooms. Each clean room is classified into one of several classes (Class II, III, or IV, as defined in the Air Force Technical Order 00-25-203, or Class 100, 10,000, or 100,000, as defined in Federal Standard No. 209) according to set limits of tolerable particulates of specific sizes. No attempt is made to differentiate this total particulate contamination into microbial-bearing or non-microbial particulate matter. A few preliminary reports (Refs. 1–3) have been made on the microbiology of clean rooms. Yet, at present, data on the levels of microbiological contamination that may exist within industrial clean rooms are scanty or wholly lacking. Therefore, a microbiological sampling study was initiated to determine the levels of viable particulate contamination that exist within the different classes of clean rooms during various conditions of operation and activity. The results obtained

in this study, when considered with those obtained during other spacecraft sterilization studies, will aid in establishing the class of clean room required to effect adequate microbiological control during assembly and checkout of spacecraft required to be sterile.

1. Experimental

a. Phase 1. Phase 1 is designed to standardize the procedures to be used in the study. A variety of culture media will be screened to determine the best medium or media for the maximum recovery of thermophiles and aerobic and anaerobic spore-forming and asporogenous mesophiles from the intramural air and surfaces of clean rooms. Culture media for the best recovery of fungi will also be included. Air sampling times and sites and surface sampling areas will be determined during this phase. At the end of Phase 1, microbiological sampling techniques will be standardized, and a rough approximation of the levels of microbiological contamination within clean rooms should be available.

b. Phases 2 and 3. Phase 2 (the actual testing phase) and Phase 3 (data recording and evaluation) will be conducted simultaneously. In Phase 2, the air and surfaces in each of three clean rooms (one each of Classes II, III, and IV) will be sampled daily for six sampling periods of 1 wk each over a 6-mo period. The particular sampling week for a given room was randomly selected; i.e., the Class II clean room will be sampled daily during the 3rd, 8th, 9th, 11th, 16th, and 19th weeks of the 6-mo period. Similarly, the Class IV clean room will be sampled during the 5th, 7th, 10th, 12th, 14th, and 20th weeks. Sampling periods for the Class III clean room were also randomized over this period. Thus, each room will be sampled microbiologically for six time periods, or a total of 30 days. As far as practical, at least six sampling days will be times when the room is not in use (normal idle days, weekends, etc.) to determine the background levels of microbiological contamination. The remaining 24 sampling days will be times when the room is fully staffed and is being operated under full working conditions. For comparison, microbiological samples will be collected from at least one specified area outside each clean room being sampled.

To assay the number of airborne viable particles, the all-glass impinger (AGI, Fig. 1), the Andersen sampler (Fig. 2), several types of slit samplers (Figs. 3 and 4), and

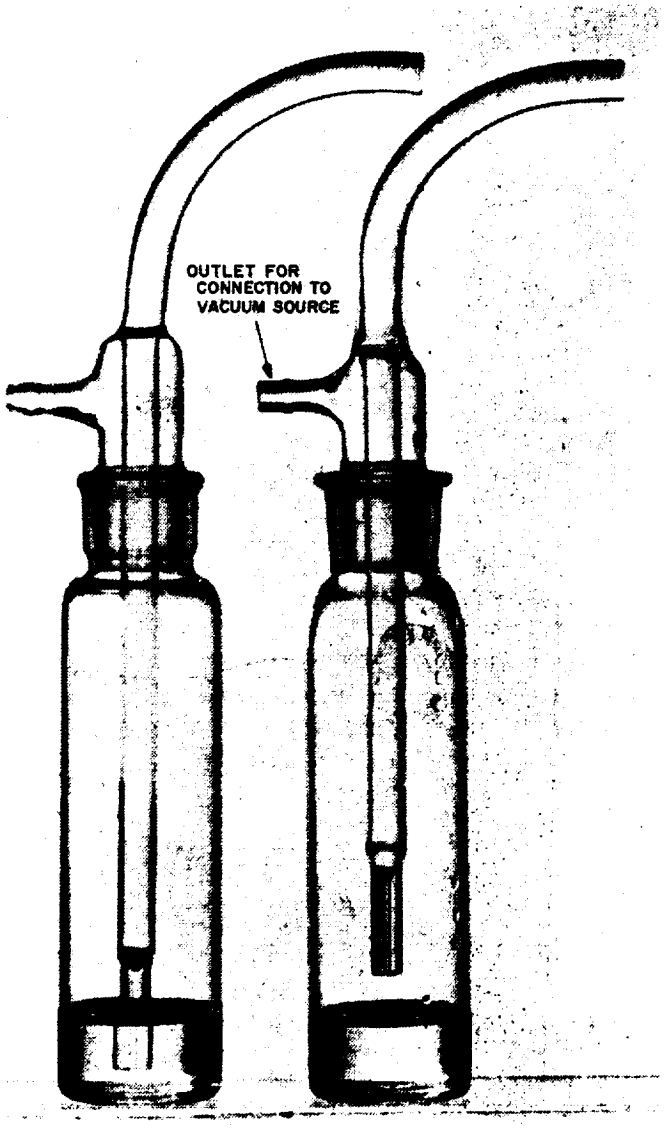


Fig. 1. All-glass impinger

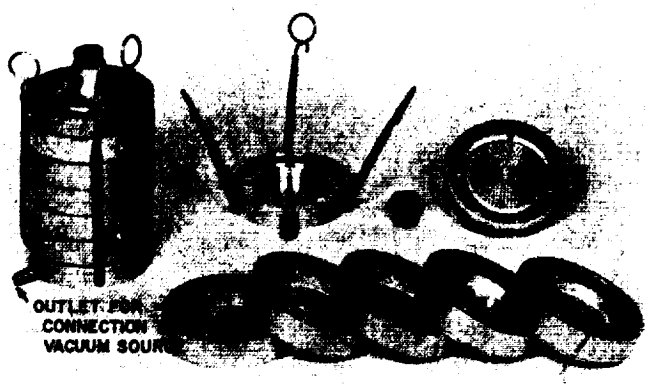


Fig. 2. Andersen sampler

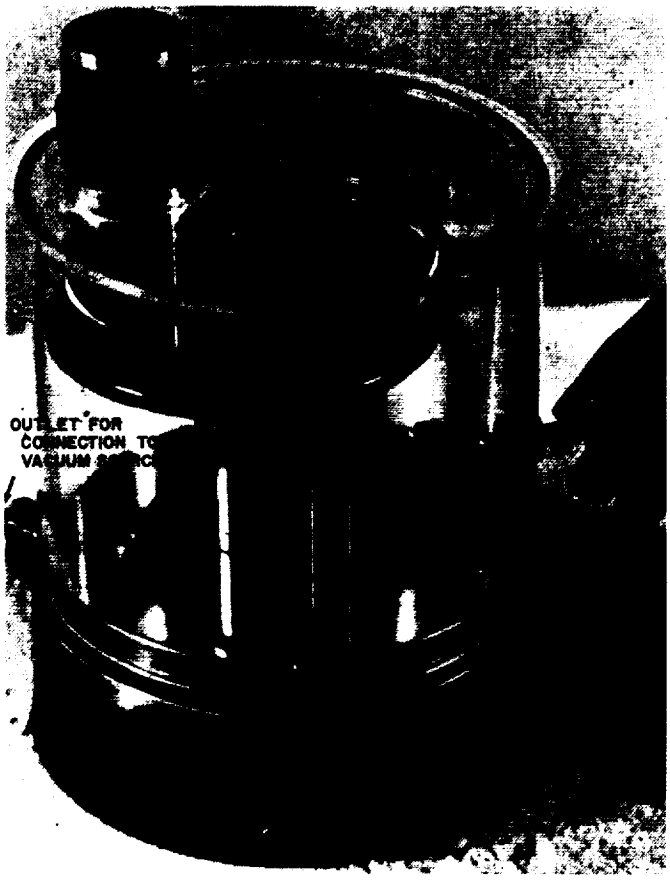


Fig. 3. Elliott slit sampler

Fig. 4. Pot-type slit sampler

agar settling plates (Fig. 5) will be used. Periodically, an estimate of the total particulate matter shall be made with a light-scattering device. Also, air samples will be collected with a membrane filter and examined microscopically for total particulate counts. The degree of surface contamination will be assessed with the Rodac

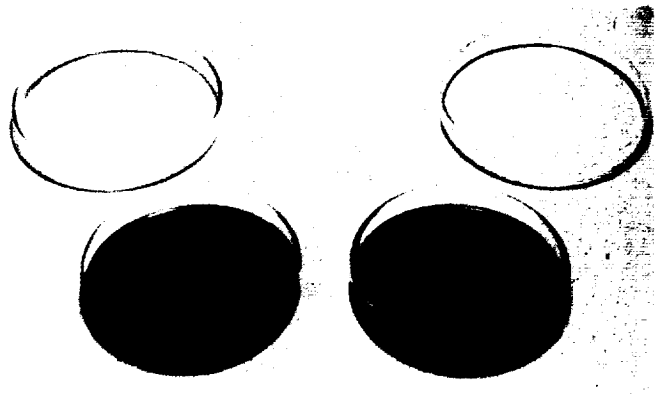


Fig. 5. Agar settling plate

plate[1] (Fig. 6a) and by the use of sterile 1- × 3-in. settling strips of polished stainless steel, lucite, and glass.

In addition to the sampling procedures described above, sterile strips of polished stainless steel, lucite, and glass will be handled by clean room personnel. This procedure is included to obtain some estimate of the amount and type(s) of microbiological contamination that may occur on spacecraft parts as a result of human handling.

2. Results and Discussion

The entire study was designed to obtain a measure of the number of viable aerobic and anaerobic microorganisms present within the intramural environment of clean rooms. Other aspects of the study include:

(1) An estimation of the aerial fallout of particles containing viable aerobic and/or anaerobic microorganisms. From such data, it should be possible to determine the degree of microbial contamination that accumulates on polished stainless steel, lucite, and glass surfaces exposed to clean room environments for a 6-mo period.

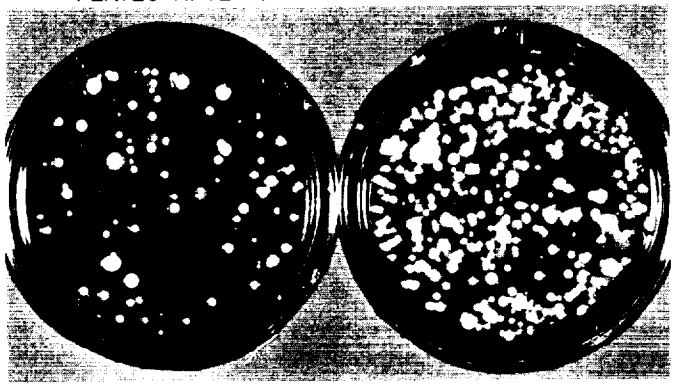
(2) A determination of whether or not a microbial spore population accumulates on exposed surfaces. These data should provide some information on the type of microbial flora accumulating on exposed surfaces over the study period.

(3) A determination of the degree of contamination that occurs on initially sterile surfaces after handling by clean room personnel. Human handling

[1]Hall, L. B., and Hartnett, M. J., "Measurement of Bacterial Contamination on Surfaces in Hospitals," *Public Health Reports* (To be published).

(a) APPLICATION OF THE RODAC PLATE

(b) MICROBIOLOGICAL COLONIES DEVELOPING ON RODAC PLATES AFTER SURFACE SAMPLING AND INCUBATION



TYPICAL RECOVERY FROM LESS HEAVILY CONTAMINATED SURFACES

TYPICAL RECOVERY FROM HEAVILY CONTAMINATED SURFACES

Fig. 6. Use of the Rodac plate in surface sampling for microorganisms

may result in excessive microbiological contamination of spacecraft parts or the vehicle itself. Some estimate of the microbial contamination that may result from clean room personnel handling is essential to the over-all process of spacecraft sterilization.

Airborne particulates occur as aerosols, i.e., solid or liquid particles suspended in air. Such particles may be a single microorganism surrounded by a film of dried organic matter or a clump of several to many cells of the same or differing microbial species. Viable microorganisms may also be attached to nonviable particles such as dust, lint, etc. Airborne viable particles may be disseminated for considerable distances. Airborne particles may also settle onto surfaces and later be re-aerosolized by violent actions that agitate the contaminated surfaces. The air sampling devices were selected to discriminate, qualitatively and quantitatively, the airborne viable particles within clean rooms into thermophiles, aerobic and anaerobic spore-forming and asporogenous mesophiles, and fungi.

Liquid impingement samplers, such as the AGI (Fig. 1), bubble air through a collecting fluid. Samplers of this type break up clumps of microorganisms and also wash microorganisms off dust and other nonviable particles on which some are carried. After sampling a measured volume of air (usual flow rate, 12.5 liters/min; usual sampling time, 15.0 min), portions of the collecting fluid will be transferred to plates of solid culture medium. To determine the number of airborne microbial spores, other portions of the same collecting fluid will be heat-shocked (80°C/15 min) and plated on solid culture medium. Incubation conditions and temperatures will be chosen to allow assay of the aerobic and anaerobic sporogenous and asporogenous mesophiles, thermophiles, and fungi from each air sample collected with the AGI sampler.

Solid media impaction samplers, such as the slit sampler (Figs. 3 and 4) and the cascaded sieve sampler (Fig. 2), deposit particles suspended in air directly onto the surface of a solid culture medium. Upon incubation, the deposited particles containing viable microorganisms will, in the proper physiological environment, grow and divide repeatedly, resulting in the formation of colonies on the culture medium.

In the operation of slit samplers, a given volume of air (flow rate, 1.0 ft^3/min) to be sampled is drawn through a narrow inlet slit.[2] Passage through the slit increases the velocity of the air stream being sampled. The high-speed jet of air is then forced to make a 90-deg turn at the face of a solid culture medium contained in a Petri plate located immediately below the slit (Figs. 3 and 4). Many of the particles suspended in the air stream cannot make this sharp turn and continue on in a straight line, to be impacted onto the culture medium. Furthermore, in a slit sampler, the Petri plate is mechanically rotated under the slit. Such movement causes the impacted particles to be distributed over the portion of the medium that moves under the slit. Thus, concentrations of colonies in one or more sections of the culture medium can be related to the time of their aerosolization as viable particles.

The Andersen six-stage cascaded sieve sampler (Fig. 2) operates in a manner similar to that of slit samplers. A

[2]Sampling time for the Elliott slit sampler (Fig. 3) is limited to 15 min; that for pot-type slit sampler (Fig. 4) is up to 60 min. New models (Reyniers) sample up to 120 min.

given volume of air (flow rate, 1.0 ft^3/min; usual sampling time, 15.0 min) is drawn through a fixed number of small, evenly spaced holes in perforated disks. A stationary Petri plate, containing solid culture medium, is located beneath each of the six perforated disks. The holes in the top stage (perforated disk) are comparatively large (1.81 mm D). In each succeeding stage, the hole size decreases, with the smallest holes (0.25 mm D) located in the bottom stage (perforated disk). Thus, the velocity of the air passing through the sampler increases with the passage through each stage, reaching near-sonic speed as it passes through the holes of the bottom stage. The size of the particles being impacted depends considerably upon the velocity of the air stream through the sampler. Thus, the largest particles tend to be deposited onto the culture medium in the Petri plate beneath the top stage, where the air-stream velocity is lowest. Furthermore, the smallest particles tend to be impacted onto the medium beneath the bottom stage, with the intermediate particle sizes being distributed onto the plates of culture medium beneath the other four stages.

Through the use of the slit and cascaded sieve air sampling devices, it will be possible to measure, in broad terms, the size distribution of airborne particulates and also obtain a time–concentration relationship for the occurrence of microbiological aerosols or "showers." It may be possible to trace movements of aerosolized microorganisms and to detect clean room objects, personnel, and/or activities that generate showers or clouds of airborne viable particles.

The Rodac plate is a recent variation of the contact principle of surface sampling (Fig. 6a). For use, the bottom half of the plate is filled with a culture medium. Upon solidification, a convex surface of standard dimension ($\cong$ 4.0 $in.^2$) results. This convex area is impressed onto the surface to be sampled. Microorganisms existing as surface contaminants are pressed onto the surface of a suitable culture medium. After contact, a top is placed on the plate and the entire unit is incubated. Following incubation, the colonies that developed are counted and the microbial contamination per unit surface area may be estimated (Fig. 6b). The Rodac plate is easily prepared, is readily usable, and gives reproducible results.

Agar settling plates (Fig. 5) have been included to obtain information on the aerial fallout of viable particles for short (1- to 6-hr) exposure periods. The settling plate, exposed for short intervals of one to several hours, tends to provide an indication of the predominant type(s) of viable airborne microorganisms in the environment.

For long-term exposures such as the 6-mo study period, and for determining the type(s) of microorganisms that accumulate on surfaces within clean rooms, the polished stainless steel, lucite, and glass settling strips will be used. The exposed strips will be sampled periodically by shaking in a collection fluid. After shaking, the collecting fluid will be treated according to the procedure described for the AGI sampler. Thus, it will be possible to determine the presence of thermophiles, mesophilic spores and vegetative cells, and fungi from each settling strip. Finally, with this procedure, it should be possible to determine if a spore population might be expected to accumulate on exposed surfaces within an environmentally controlled area. Sterile polished stainless steel, lucite, and glass strips that have been handled by clean room personnel will also be shaken in broth and assayed for viable organisms by the technique described above for settling strips.

Earlier studies by the author (Refs. 4–6; also [3,4] below) and other investigators (Refs. 7–10) have demonstrated that relative humidities in the 50.0 to 60.0% range are quite lethal for surface-exposed vegetative cells. However, preliminary studies (unpublished results of the author) would indicate that this range of relative humidity has little, if any, effect on spores of a strain of *Bacillus subtilis* var. *niger*. The accumulation of a spore population on exposed surfaces would present a considerably more difficult challenge to any sterilization procedure than the contamination of similar surfaces with a vegetative cell population.

It is essential to know the degree of contamination and the type(s) of contaminating organism(s) that may result on spacecraft parts from the hands of clean room personnel. The primary source of microorganisms within clean rooms is the particles disseminated from the clean room personnel (Refs. 1–3 and Ref. 11). A single touch with a heavily contaminated hand or glove might seriously contaminate a spacecraft. Such an action might negate all earlier precautions and sterilization procedures, thereby defeating the purpose of spacecraft assembly in an environmentally controlled area. The human contamination aspect of the present study should add to our knowledge of contamination sources within clean rooms.

[3]McDade, J. J., and Hall, L. B., "Survival of *Staphylococcus aureus* in the Environment. II. Effect of Elevated Temperature on Surface-Exposed Organisms," *American Journal of Hygiene* (To be published).

[4]McDade, J. J., and Hall, L. B., "Survival of Gram-Negative Bacteria in the Environment. I. Effect of Relative Humidity," *American Journal of Hygiene* (To be published).

Taken collectively, the results obtained from the different sampling devices and procedures should be valuable and needed additions to the knowledge of clean room microbiology. This study should produce a comprehensive summary of the levels of microbiological contamination that exist in the different classes of industrial clean rooms under conditions of normal activity and operation. Such data are essential to any plan to define the environmental conditions for the assembly of spacecraft required to be sterile.

C. An Experimental Study of Sterile Assembly Techniques

J. J. McDade

The NASA requirement that all spacecraft and probes which might impact an extraterrestrial body be sterile has been widely promulgated. The use of dry heat appears to be the method of choice for the terminal sterilization of a planetary landing spacecraft. However, at present, certain spacecraft items cannot withstand the dry heat cycle without serious degradation in reliability. Therefore, if an entirely heat-stable spacecraft is not possible, reliable sterile assembly techniques must be developed and be ready for use to incorporate heat-labile components into sterile spacecraft.

The present study[5] is being conducted to evaluate the potential of obtaining a sterile electronic unit through use of a procedure in which each assembly step takes place in a glove box containing an ethylene oxide (ETO) atmosphere. Later in the study, sterile parts will be assembled into an electronic unit in a glove box containing sterile air or sterile nitrogen. Each step in the sterile assembly procedure will be checked through positive control procedures, sterility checks, and systematic tests. Performance tests will be conducted on the units assembled in the glove box. The assembled units will be operated in glove box atmospheres of ETO, sterile nitrogen, or sterile air, and also outside the glove box. The performance and reliability of the intramurally and extramurally tested units will be compared.

[5]This study is being conducted at the Lockheed Missiles and Space Company, Sunnyvale, California, under JPL contract.

1. Experimental

Essentially, this is a feasibility study to determine the following factors:

(1) The feasibility of assembly of a small electronic unit within a glove box through the use of glove box techniques.

(2) The efficacy of the sterilant gas conditions and time cycle to surface-decontaminate electronic parts, using a known concentration of bacterial spores contained on typical electronic parts used in the test circuit.

(3) The effect, if any, of assembly in ETO, sterile nitrogen, or sterile air on sample soldering, bonding, staking, and nut/bolt assembly procedures.

(4) The effect of assembly in ETO, sterile nitrogen, or sterile air on a complete unit assembly and on the performance and/or reliability of the test unit.

(5) The effect of operating the test unit in ETO, sterile nitrogen, or sterile air on the performance and/or reliability of the test unit.

(6) The efficacy of a sterile sealing system for the test circuit.

The basic units employed in the study are shown in Fig. 7. The main glove box for assembly, a 3- × 3- × 6-ft unit, has been modified by the addition of a gas lock and a smaller glove box on one end of the unit. The other major units include a gas generator and purge system (sterilant gas generator) and the gas concentration and relative humidity continuous-readout system (monitoring apparatus). The general test program is as follows:

All electronic parts to be used will be heated to 135°C for 24 hr. Then, the heat-treated parts will be inoculated with *Bacillus subtilis* var. *niger* spores. The inoculated parts will be transferred to the main glove box and exposed to ETO gas (12.0% ethylene oxide/88.0% freon-12 gas mixture). The exact exposure time necessary to decontaminate the inoculated parts will be determined during this phase. Sterility checks will be conducted by placing the part to be tested in sterile broth. All sterility test broths will be incubated at 37°C. After 24 and 48 hr of incubation, the broths will be examined visually for turbidity. Incubation will be continued for 7 days. If no growth is visually detectable after 7 days, portions of the broth will be plated on trypticase soy agar. The remaining portion of the previously incubated sterility broth will be inoculated with *Bacillus subtilis* var. *niger* spores and incubated. If bacteriostasis did not occur during the

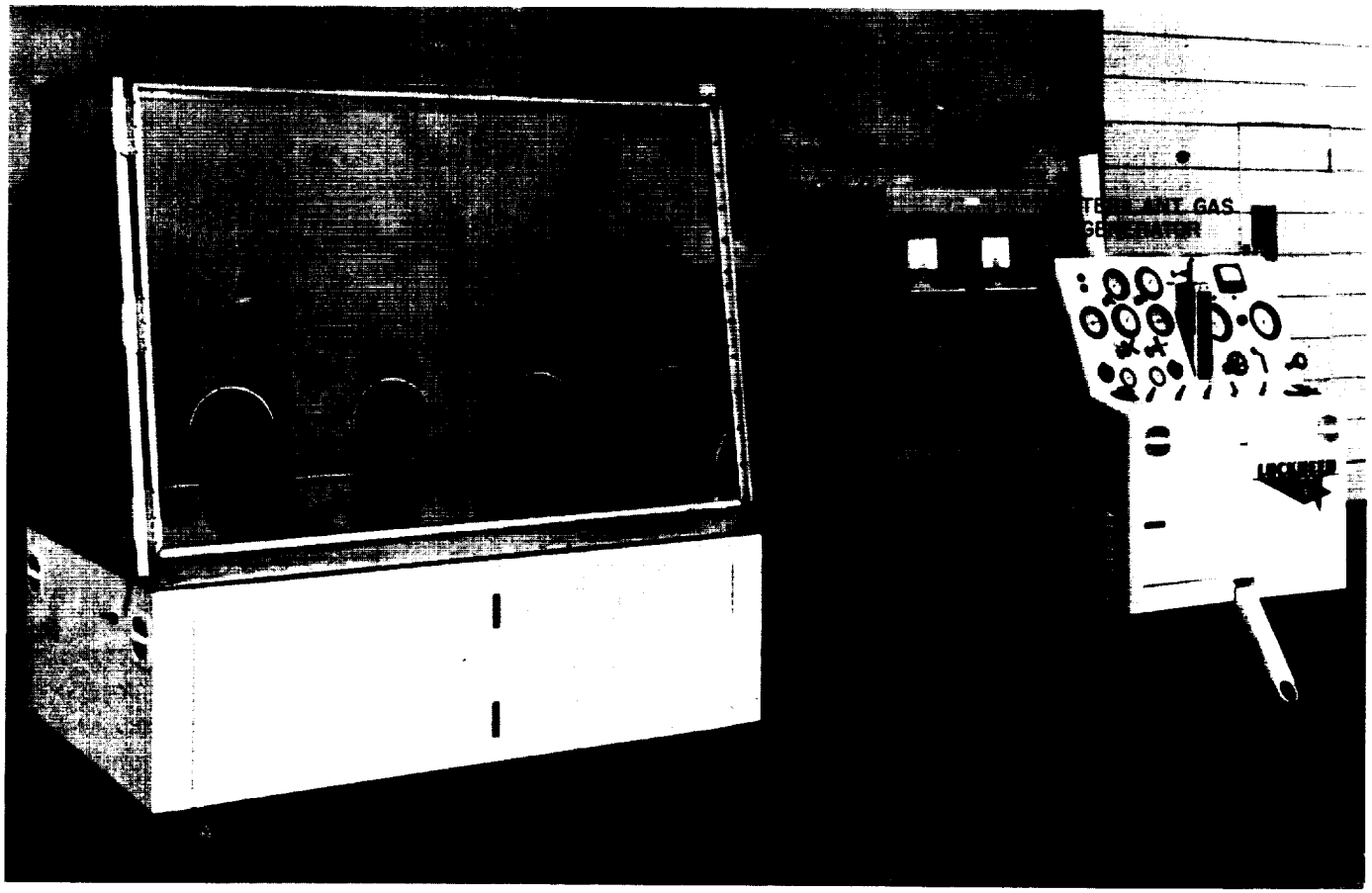


Fig. 7. Basic units employed in study of sterile assembly techniques

sterility check, the freshly inoculated spores should germinate and proliferate in the broth.

When the ETO exposure cycle has been determined, a breadboard of the full circuit will be assembled outside the main glove box. The breadboard assembly will be tested to determine its performance and operational characteristics. A paper-tape-controlled comparator system will be used to guarantee accuracy and repeatability of the assembled breadboard tests. Upon completion of these checks, the breadboarded circuit will be disassembled and all the parts will be placed in the main glove box. Prior to the introduction of these parts, the interior of the main glove box will be exposed to ETO.

ETO will be released into the main glove box, and the breadboard circuit will be reassembled in an ETO atmosphere. The same personnel will reassemble the breadboard using the same techniques. A checkout identical to the previous breadboard tests will be performed with the tape checkout system. The performances of both assemblies will be compared.

Upon completion of the above (i.e., the assembly of a breadboard circuit, successfully tested without degradation in ETO), printed circuit assemblies will be made. Some of the printed circuit boards will be performance-tested in the main glove box. After this testing, these circuit boards will be sterility-tested. Other glove-box-assembled circuit boards will be removed and performance-tested in air. Finally, the remaining glove-box-assembled printed circuit boards will be mounted in sealed containers having an internal ETO environment. Some of the sealed containers will be moved to the small glove box containing sterile air. Then, the containers will be opened, and the internal circuit boards, as well

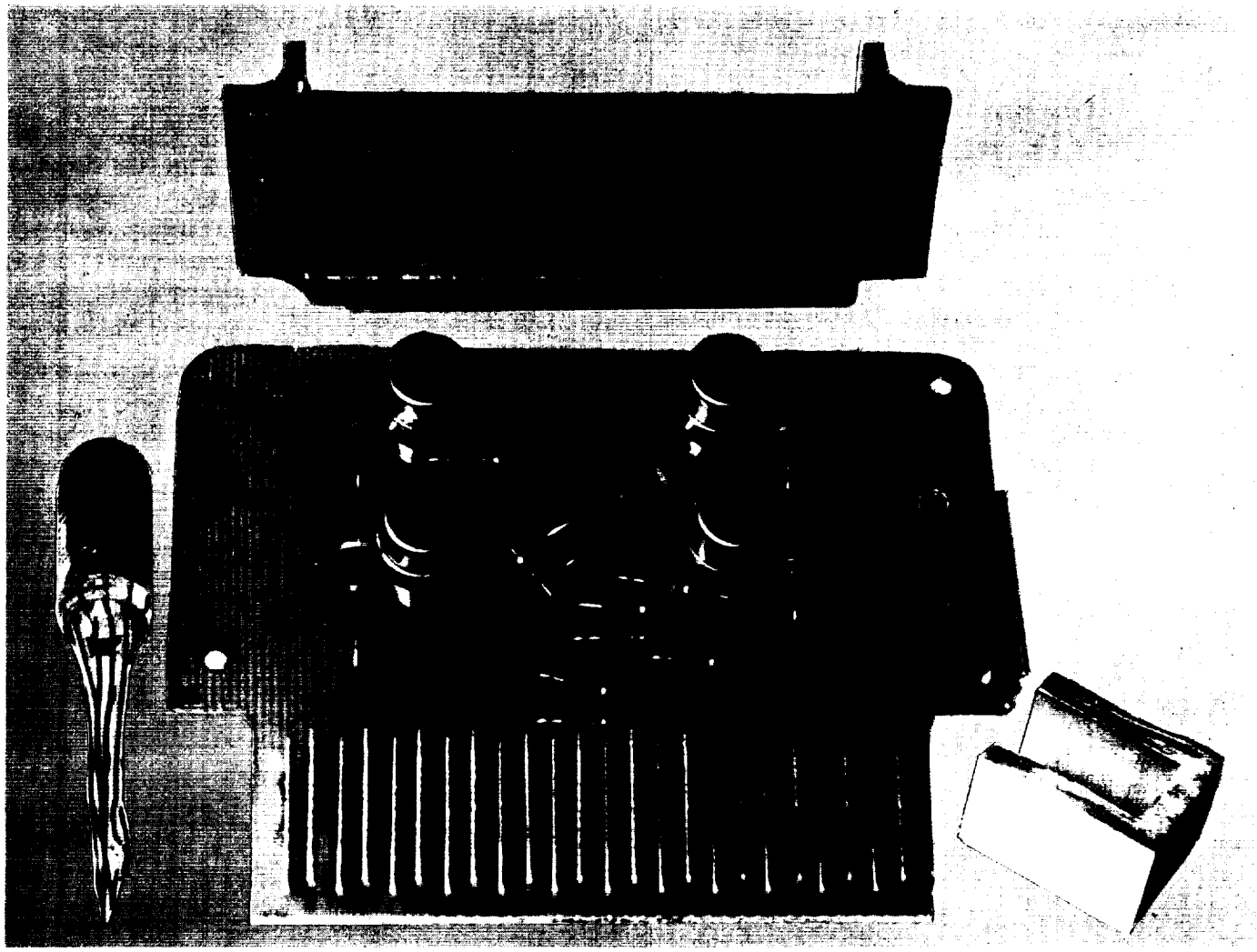


Fig. 8. Typical electronic assembly to be tested

as the containers, will be checked for sterility. The remaining sealed circuit boards will be removed from the glove box, and these "canned" circuits will be subjected to a series of performance tests, including handling and drop tests. These units will then be returned to the main glove box, and the containers will be exposed to ETO. Following this, the containers will be opened and some of the circuit boards will be tested for sterility.

After this series of tests has been completed, the sequence beginning with the introduction of heat-treated parts into the main glove box through breadboard assembly and checkout to printed circuit boards will be repeated in a sterile nitrogen atmosphere and then in a sterile air atmosphere. Appropriate performance tests and sterility checks will be included. A series of assemblies with routine parts (not sterilized) will be used.

2. Results and Discussion

The first phase of work (1-mo period) was concerned with the modification of equipment and the performance of tests to standardize procedures. The test circuit (Fig. 8), as well as several alternate circuits, was breadboarded and checked out with the tape-controlled checkout system. The circuits were checked for sensitivity to small variations in component values. The original circuit, with the addition of a capacitor and with slight component-value changes, was found to be the optimum circuit. The circuit design and printed circuit card design were completed.

Experiments were conducted to determine the best method for surface-inoculating electronic parts with spores of *Bacillus subtilis* var. *niger*. Immersing the part

in a suspension of spores in distilled water, followed by drying, appeared to be a promising method.

The results obtained from this study should be most interesting. Use of the paper-tape-controlled comparator system will ensure the accuracy and repeatability of tests on the breadboarded test circuit. The test tape developed with the breadboarded circuit will be used to determine and maintain circuit performance in the transition from the breadboard to printed circuit card. Since the test circuits will be assembled in an ETO atmosphere within the glove box, it will be most important to see what effect, if any, ETO gas has on the operation of dip soldering, hand soldering, staking, epoxy bonding, potting, and even nut/bolt connecting. Backup operations include repeating the entire assembly procedure in glove box atmospheres of sterile nitrogen and sterile air.

D. Evaluation of Microbiological Filters for Liquids and Gases

A. Irons

The production of sterilizable spacecraft, their sterilization prior to launch, and the maintenance of sterility after launch and until planetary impact will almost certainly require the use of microbiological filters for gases and will probably require the use of filters for liquids. Certain spacecraft assembly areas and glove box systems, as well as controlled environment areas housing sterilized spacecraft prior to launch, will have sterile air requirements. Air delivered to these areas will no doubt be filtered first through a positive filter to remove all viable particles. Gases carried aboard spacecraft, such as those used in some attitude-control devices, will also have to be sterilized, probably in the same manner. If it becomes necessary to use heat-labile liquids in the production of spacecraft (for example, liquid propellants and battery electrolyte) or in extraterrestrial microbiological experiments, it will be essential to utilize a microbiological filter capable of sterilizing liquids.

Reports from laboratories engaged in work involving sterile filtration of liquids and gases indicate that claims made for the efficiency of filters are sometimes exaggerated. This factor, a cause of great concern, has led to the present study, which is being conducted by the Wilmot Castle Company. This study is designed to test the efficiency and evaluate the reliability of commercially available filters in terms of their ability to sterilize liquids and gases. The results of this study should provide a basis for the selection of filters required to produce the required conditions of sterility. Table 1 lists the microbiological filters which will be evaluated in the present study. A minimum of 10 each of these filters will be evaluated.

In all cases, for both liquid and gas filters, the manufacturers' recommendations of pressure differential across the face of the filter will be followed. In general, the procedures to be used in evaluating the filters will be as follows:

a. Liquids. An 18- to 24-hr broth culture of a bacterial strain whose size is small enough to present a maximum challenge to the filter will be utilized. A minimum of 10^3 ml of an aqueous suspension of the organism, containing approximately 10^2 viable organisms/ml, will be drawn through the filter being challenged. An assay of the cell suspension being filtered will be conducted simultaneously with the filtration test to determine the number of viable organisms/ml of suspension.

The filtrate will be incubated at optimum temperature for a period of 1 wk, but will be examined at 24-hr intervals for evidence of microbial growth. Adequate positive and negative controls of both media and organism, both before and after filtration, will be maintained to ensure validity of the results.

If growth appears in the filtrate, the contaminating organism(s) will be isolated and identified. This step is necessary so that the possible cause(s) of filter failure may be determined. Filter failure may result if:

(1) The pore size of the filter is too large (degrading filter efficiency).

(2) The integrity of the filter is broken (degrading reliability).

(3) An inadequate sterilization cycle has been applied to the test system prior to use, as evidenced by the presence of organisms in the filtrate other than those used to challenge the filter (contamination).

(4) Faulty technique is employed.

If some other cause is suspected, it will also be reported.

Table 1. Types of microbiological filters to be evaluated

Manufacturer	Description
American Air Filter Co., Inc.	Type F A11G6R2 (glass) Type F, 8- × 8- × 5⅞-in. (ceramic asbestos)
Cambridge Filter Corp.	1 A 50 filter 1 D 50 filter, moisture resistant
Cox Instruments	Filter elements, 0.5 μ
Cuno Engineering Corp.	6D-O asbestos–cellulose disk 6D-DO asbestos–cellulose disk
Flanders Filter, Inc.	Size "B" 7-A-10a (glass) Size "B" 5-F-40d (ceramic)
Gelman Instruments Co.	47-mm gelatin, 0.5 μ GM-6 cellulose ester GM-8 cellulose ester Glass Type "A" Glass Type "E" with binder GA8, triacetate, 0.2 μ Epoxy membrane, Versapor 6424
Herman, F. R. & Co.	Pads, asbestos D-6 Pads, asbestos D-10
Millipore Filter Corp.	Standard filter, HAWP 04700, Size 0.45 μ Standard filter, GSWP 04700, Size 0.22 μ Standard filter, WHWP 04700, Size 0.45 μ Standard filter, PHWP 04700, Size 0.30 μ Microtube cartridges, CA12-440-00 Cellulose ester, 8.0 μ, 142 mm, SCWP 14250
Mine Safety Appliances Co.	Glass fiber, CU-72926 Ultra filter cartridge, 15-85759 Oxygen liquid cartridge, 15-85759 Oxygen liquid cartridge, 15-83915 Glass, cylindrical, 15-86695 Glass, cylindrical, 15-85202
Pall Corporation	Filter elements, Ultipor ACF4463 UW (0.15 μ)
Schleicher and Schuell Co.	B-6, 0.4 Bac-T-Flex, 47 mm B-11, 0.25 Bac-T-Flex, 47 mm MC4, 0.25 Bac-T-Flex, 47 mm with built-in prefilter 0.2, 0.25 Organic Solvent Resistant, 47 mm
Selas Flotronics	10 Micro-porous porcelain FPS-56 01 Micro-porous porcelain FPS-56 015 Micro-porous porcelain FPS-56 02 Micro-porous porcelain FPS-56 03 Micro-porous porcelain FPS-56
Will Corporation	Berkefeld candles Seitz asbestos pads, clarifying grade Seitz sterilization disk, coarse Seitz sterilization disk, fine Pyrex, sintered glass filters, ultrafine (0.9 to 1.4 μ)

All observations will be recorded and explanations offered for filter failure whenever possible. All tests will be repeated as many times as required to achieve a high confidence level, with a minimum of 10 tests for each type of filter being evaluated.

b. Gases. Procedures to evaluate the efficiency of gas filters are more complex than those outlined for liquid filters. Numerous methods can be used to evaluate the efficiency of filters to remove solid particles such as dust or smoke (DOP; dioctyl-phthalate test) from air. Other tests utilize various dyes in the form of dust particles to challenge filters. The ultimate test of the efficiency of a filter to remove viable microorganisms is to challenge the filter with viable microorganisms. In the case of a gas filter, this means viable microorganisms suspended in a stream of gas moving through the test filter at its rated capacity.

A general procedure which can be employed is as follows: An aerosol of heat-shocked (to destroy vegetative cells) *Bacillus subtilis* spores is introduced into an air duct upstream from the filter being challenged. Air sampling probes are inserted upstream and downstream from the filter. These probes can be fastened to Andersen air samplers located outside of the duct. Particles picked up by the probes will be impinged on solid culture media located in the samplers. If these plates are incubated at the proper temperature (34 to 37°C) for at least 72 hr, the resulting colonies of microorganisms can be counted. If the number of organisms introduced into the duct upstream of the filter is known, and if the volume of gas passing through the filter is also known, then we can calculate how many organisms/ft^3 are challenging the filter. If we know the volume of gas being sampled by the probe(s) and we have established the efficiency of the sampling system, we can determine the efficiency of the filter being challenged.

The use of probes upstream and downstream simplifies the procedure. Obviously, if the upstream and downstream samplers are sampling the same volume of gas, we can determine the efficiency of the filter by the formula:

$$\text{Removal efficiency, } \% = \frac{\left(\begin{matrix}\text{number of particles}\\ \text{upstream}\end{matrix}\right) - \left(\begin{matrix}\text{number of particles}\\ \text{downstream}\end{matrix}\right)}{(\text{number of particles upstream})} \times 100.$$

This procedure is illustrative of one method of determining the relative efficiency of a given filter.

References

1. Michaelson, G., and Vesley, D., "Industrial White Rooms vs Hospital Operating Rooms," *Air Engineering,* Vol. 5, pp. 24–29, 1963.

2. *The Level of Microbial Contamination in a Clean Room During an Eleven Week Test Period,* Protection Branch Report of Test No. 10-64, Physical Defense Division, Fort Detrick, Frederick, Maryland, February 3, 1964.

3. *Microbial Contamination in a Clean Room When Occupied by Operating Personnel,* Protection Branch Report of Test No. 1-65, Physical Defense Division, Fort Detrick, Frederick, Maryland, August 12, 1964.

4. McDade, J. J., and Hall, L. B., "An Experimental Method to Measure the Influence of Environmental Factors on the Viability and the Pathogenicity of *Staphylococcus aureus* in the Environment," *American Journal of Hygiene,* Vol. 77, pp. 98–108, 1963.

5. McDade, J. J., and Hall, L. B., "Survival of *Staphylococcus aureus* in the Environment. I. Exposure on Surfaces," *American Journal of Hygiene,* Vol. 78, pp. 330–337, 1963.

6. McDade, J. J., and Hall, L. B., "Effect of the Environment on Virulence," National Conference on Institutionally Acquired Infections, *U. S. Public Health Monograph No. 1188,* pp. 91–97, 1963.

7. Bateman, J. B., McCaffrey, P. A., O'Connor, R. J., and Monk, G. W., "Relative Humidity and the Killing of Bacteria. The Survival of Damp *Serratia marcescens* in Air," *Applied Microbiology,* Vol. 9, pp. 567–571, 1961.

8. Dunklin, E. W., and Puck, T. T., "The Lethal Effect of Relative Humidity on Airborne Bacteria," *Journal of Experimental Medicine,* Vol. 87, pp. 87–101, 1948.

9. Kethley, T. W., Fincher, E. L., and Cown, W. B., "The Effect of Sampling Method Upon the Apparent Response of Airborne Bacteria to Temperature and Relative Humidity," *Journal of Infectious Diseases,* Vol. 100, pp. 97–102, 1957.

10. Lidwell, O. M., and Lowbury, E. J., "The Survival of Bacteria in Dust. II. The Effect of Atmospheric Humidity on the Survival of Bacteria in Dust," *Journal of Hygiene,* Vol. 48, pp. 21–27, 1950.

11. McDade, J. J., "Sources of Microbiological Contamination Within Clean Rooms," *Proceedings of the American Association for Contamination Control,* Session XI, Paper 3, 1964.

GUIDANCE AND CONTROL DIVISION

III. Spacecraft Electrical Power

A. Power Sources

D. W. Ritchie and S. S. Luebbers

1. Photovoltaic Solar Power Systems, D. W. Ritchie

Two approaches will extend the usefulness of the solar photovoltaic power systems for space utilization: A decrease in weight, and an increase in the useful deployed surface. Since present launch vehicle shroud constraints have limited the available surface area of solar arrays, we are planning to examine unfurlable systems. These concurrent developments should result in more flexible power systems that could permit a greater variety of experiments on future space missions.

a. Thin solar cell development. A shipment of 1000 thin silicon solar cells has been received at JPL. These thin solar cells (Fig. 1) are single-crystal silicon, phosphorus diffused, with nominal physical dimensions of 0.788 × 0.788 × 0.008 in. Fig. 2 presents the current-voltage characteristics of solderless, normal, and thin types of n/p solar cells. The lower short-circuit current and the corresponding lower maximum available power output of the thinner cell are due to the lower collection efficiency at long wavelengths.

The tests performed to date on the thin solar cells have been to determine the sunlight I-V characteristics and to compare these to the present p/n systems. Two small solar cell matrixes were fabricated using p/n and thin n/p type solar cells (Figs. 3, 4). The two systems were fabricated using techniques designed for *Mariner* Mars '64. The solar cell matrixes were mounted on 0.004-in. glass-fiber board. Tests on matrixes of the small cells show that the n/p thin cell system, even with an area increase to compensate for the lower cell efficiency, can exhibit a weight reduction of nearly 30%. These weight savings exclude substrate weight considerations, but they include the glass-fiber board mentioned earlier. Development of lighter weight and nonmagnetic interconnection materials could further reduce the weight per unit area.

In conjunction with the development of the thin, lightweight solar cells, various possible designs of lightweight deployable solar panel structures are being studied.

b. Large-area lightweight, photovoltaic arrays with feasible deployment systems. The total fixed area for solar panel structures now available under a *Surveyor-Centaur* type of shroud is around 100 to 150 ft^2. The *Ranger-*

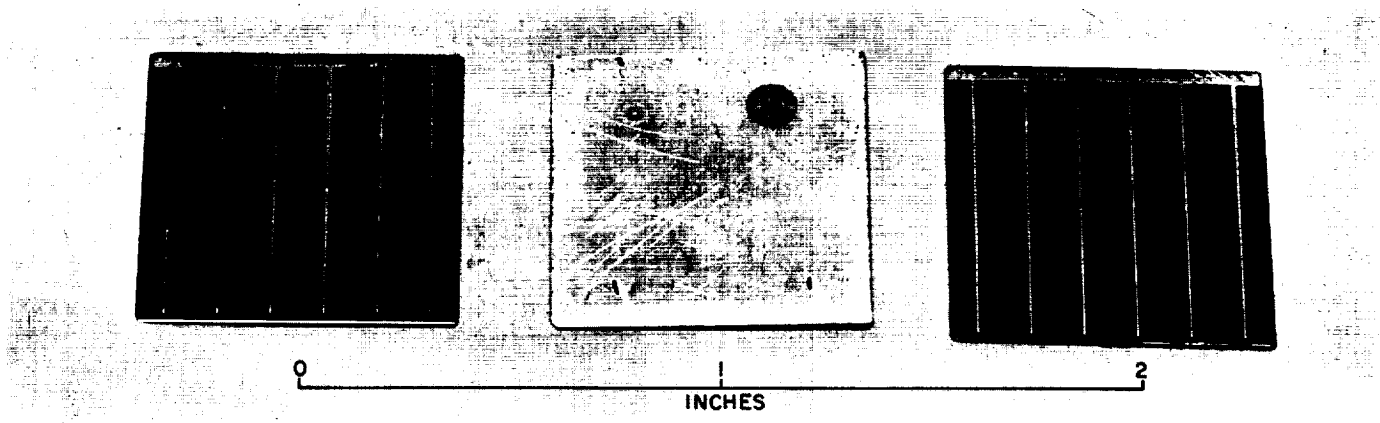


Fig. 1. Thin *n/p* type solar cells

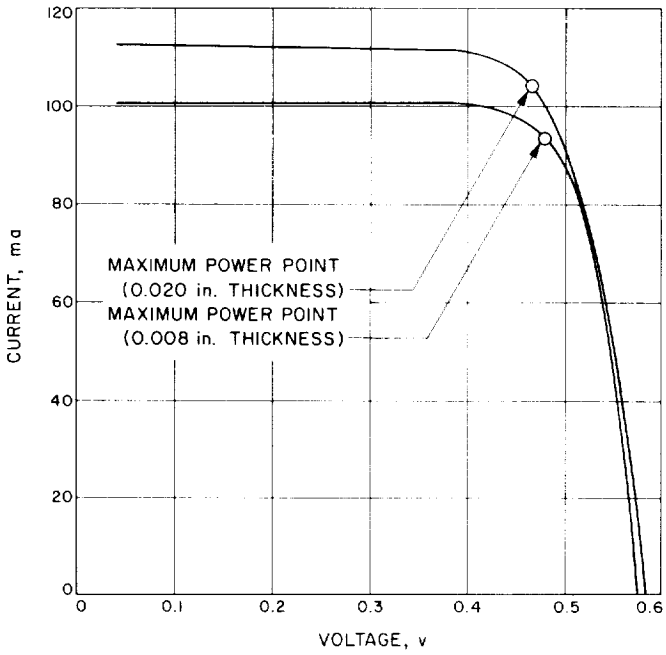


Fig. 2. Solderless solar cell *I-V* characteristics

Mariner shroud can house a maximum of 75 ft². Present photovoltaic power sources can supply little over 4 watts/ft² at Mars encounter. Thus the maximum power available using the largest area of the *Surveyor-Centaur* type shroud is limited to 600 watts at Mars. Future interplanetary spacecraft might require as much as 600 ft² of solar panel area. Presently considered goals for solar panel sizes are between 200 and 500 ft², and goals for total array weight are 0.5 to 0.6 lb/ft², including cells and wiring.

The end result of this development will be the fabrication of a representative prototype test element demonstrating the soundness and reliability of the design concepts and a mastery of the technology involved. The design and development of lightweight photovoltaic solar cell arrays associated with lightweight deployable structures should make this type of power system competitive with other systems now under consideration. Present limitations of the photovoltaic systems to approximately 1 kw could then be extended to much higher levels.

2. *Electrode Work Function in Cesium Thermionic Converters*, S. S. Luebbers

a. Introduction. Since the operation of a thermionic converter is intimately connected with the electrode work functions, a procedure for determining these functions has been developed. The motive diagram (Ref. 1) for a thermionic converter, operating in the arc mode with ion rich sheaths, is illustrated in Fig. 4. From this diagram, we may write an expression for the output voltage.

$$V_o = \phi_E - \phi_c - V_p \qquad (1)$$

where V_o is the output voltage, ϕ_E the emitter work function, ϕ_c the collector work function, and V_p the plasma sustaining voltage. The output voltage is dependent upon both the emitter and collector work functions. The voltage V_p represents the voltage loss incurred in the converter due to the combination sheath and plasma drops.

As part of the testing performed on thermionic converters, a close examination of collector work function is included. In many cases, a degradation of converter performance may be linked directly to an increase in collector work function. With such an increase, the output voltage necessarily decreases. The mechanisms by which the collector work function increases are not

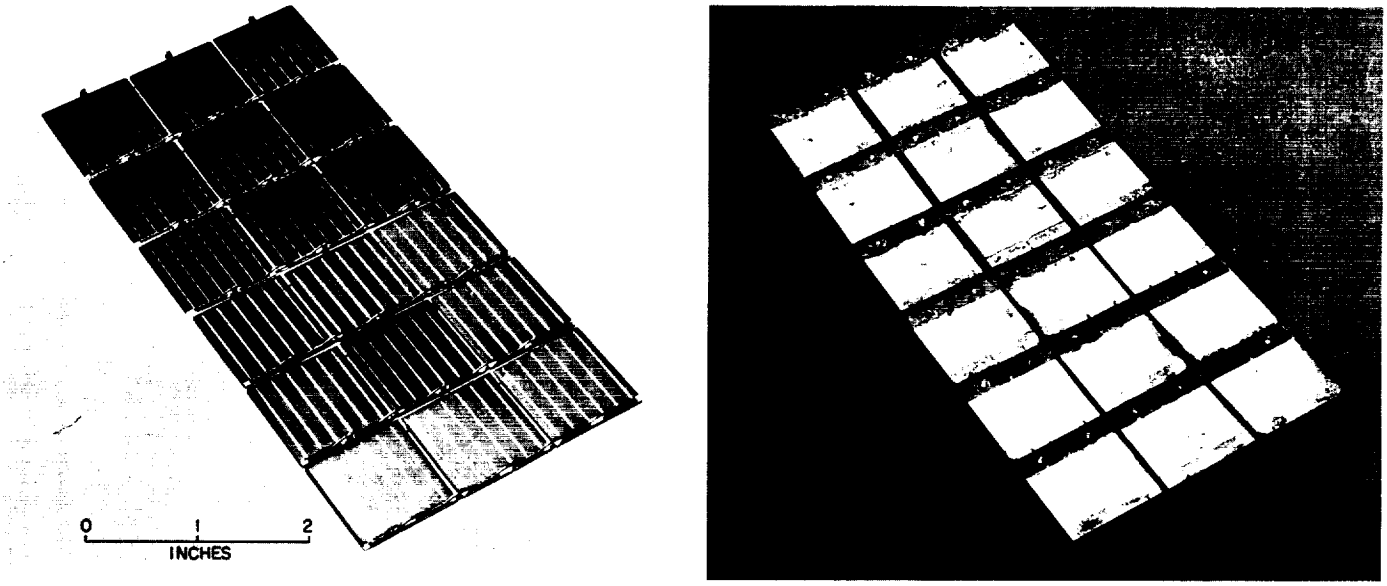


Fig. 3. Matrix of thin *n/p* solar cells; back view shows solder connections

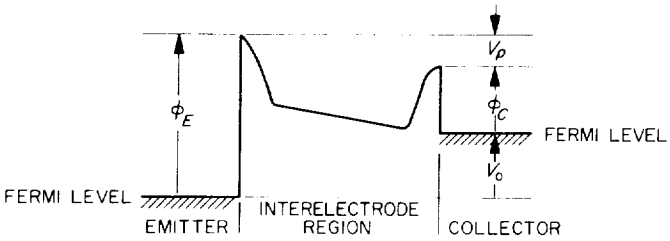


Fig. 4. Motive diagram for a thermionic converter

clearly understood. However, in most cases a collector contamination is observed. A photograph of a contaminated collector is presented in Fig. 5, and the resulting decrease in power output may be noted in the volt-ampere characteristics given in Fig. 6. Spectrographic analysis of the collector deposit has been performed, and typical results are summarized in Table 1. The remainder of this discussion covers laboratory techniques for determining collector work function. The method utilized was developed by Thermo Electron Engineering Corporation, Waltham, Massachusetts, and reported in Ref. 2.

b. Theoretical considerations. To determine the collector work function, two conditions must be satisfied. These conditions are: (1) Emitted electrons experience no collisions with cesium ions within the interelectrode space, i.e., no volume ionization is present. (2) The electrons leaving the emitter surface see no maxima in the motive diagram within the interelectrode region, i.e., complete space-charge neutralization. Under the preceding conditions, in contrast to the arc mode illustrated in Fig. 4, the motive diagram assumes the form illustrated in Fig. 7. Fig. 7 demonstrates the fact that the converter must be operated in the retarded region to obtain work function measurements.

Referring to Fig. 7, the emitting surface would tend to emit saturation current j_s; however, a retarding potential

Table 1. Semiquantative x-ray spectrographic analysis of collector deposit; molybdenum substatum

Metal	Collector I	Collector II	Collector III
Tantalum	84	24	62
Silicon	0.056	0.35	9.6
Magnesium	0.0093	0.0004	0.036
Manganese	0.0082	0.0035	—
Molybdenum	0.008	0.43	1.6
Iron	0.12	0.59	0.20
Copper	0.004	—	0.088
Tin	0.043	—	—
Titanium	0.015	0.001	0.64
Silver	0.073	0.007	—
Cobalt	0.012	0.34	—
Aluminum	0.009	0.01	—
Calcium	0.0082	0.0005	2.8
Chromium	0.0025	0.0024	tr
Columbium	—	—	—
Nickel	0.043	0.12	—
Zirconium	—	—	—

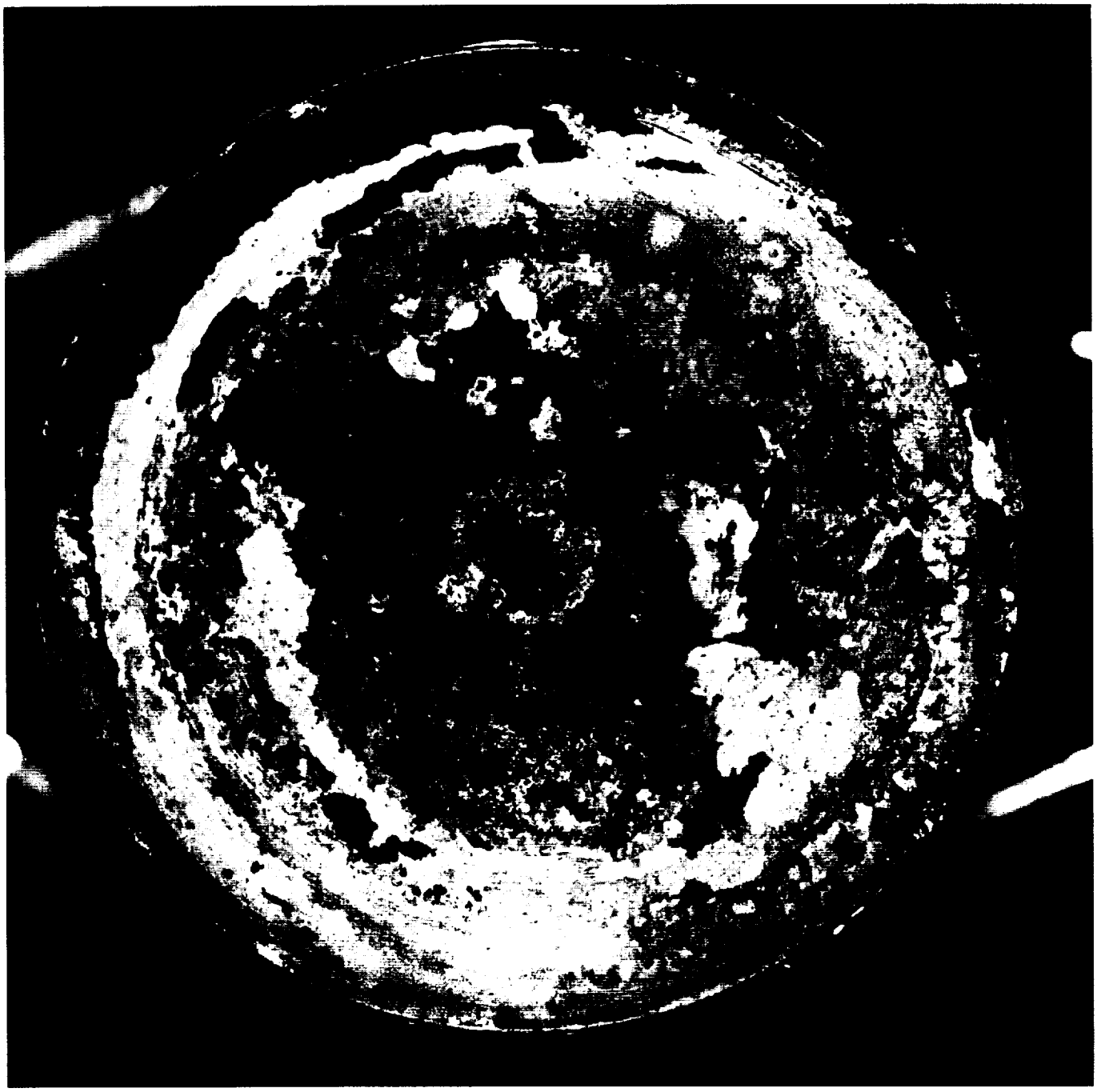


Fig. 5. Contaminated collector surface of thermionic converter

V_d is experienced by the emitted electrons. The collected current density may be expressed as:

$$j = j_s \exp(-V_d/KT) \tag{2}$$

where j is the collected current density, j_s the emitter saturation current density, V_d the retarding potential, k the Boltzmann constant in ev/deg, and T the temperature, °K.

Writing the Richardson equation for emitter saturation current j_s results in

$$j = AT^2 \exp -(\phi_E/kT + V_d/kT) \tag{3}$$

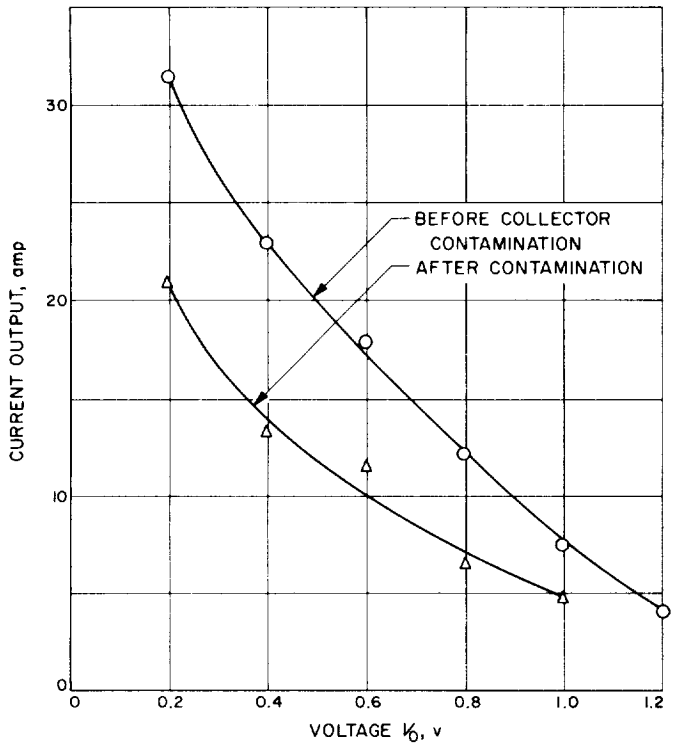


Fig. 6. Thermionic converter volt-ampere characteristics

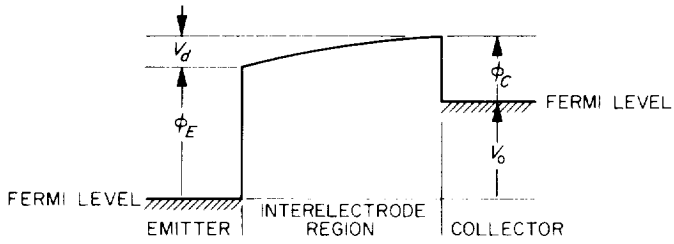


Fig. 7. Motive diagram in the retarding region

where A is the constant, 120.1 amp/cm²°K.

Substituting

$$\phi_E + V_d = \phi_c + V_0 \tag{4}$$

results in

$$j = AT^2 \exp -(\phi_c/kT + V_0/kT) \tag{5}$$

Taking the logarithm and solving for the collector work function yields

$$\phi_c = kT \ln (AT^2/j) - V_0 \tag{6}$$

Utilizing Eq. (6), we may now calculate the collector work function under the conditions previously described. To make a meaningful calculation, we must make the following fundamental assumption:

> The work function of a material, under certain temperature conditions, is determined only by the ratio of surface temperature to that of cesium reservoir temperature (or cesium pressure).

Experimental evidence indicates that this assumption is reasonably valid. There is a tendency for the work function to increase slightly with increasing collector temperature; however, extremely accurate measuring techniques are required to detect the difference. With the acceptance of the preceding hypothesis, we may determine the collector work function in the extinguished mode and conclude that the ignited mode has an almost identical work function.

c. Laboratory experimentation. The laboratory apparatus employed to determine collector work functions is schematically represented in Fig. 8. The required retarding potential is obtained from the power supply S_1. The voltages supplied to the x-y recorder will plot the volt-ampere characteristics of the converter as the R_L and S_1 are varied. A typical volt-ampere curve, obtained in the low current region is illustrated in Fig. 9.

Examination of Fig. 9 should reveal an exponential curve in the region where the previously outlined conditions are satisfied. A translation of the linear plot in Fig. 9 to a logarithmic plot is given in Fig. 10, where the curve is seen to possess a straight line portion.

Rewriting Eq. (5) yields:

$$\ln j = \ln AT^2 - \phi_c/kT - V_0/kT \tag{7}$$

This equation describes the linear portion of the curve shown in Fig. 10. If Eq. (7) is to describe the logarithmic plot, the slope of the straight line portion should be determined by $1/kT$. This is indeed the situation illustrated in Fig. 10. A straight line is drawn for the given emitter temperature, $T = 1600$°C, and is shown in Fig. 10.

Examination of Fig. 9 reveals that the converter delivers reverse current in the retarded region (>2 volts). This reverse current is a combination effect of collector back-emission and converter leakage resistance. The converter leakage resistance is caused by conduction along the cesium coated insulators of the converter. This resistance may be determined by the slope of the volt-ampere curve in saturated reverse region. Calculating the leakage resistance gives:

$$R = V/I = 1.47/0.038 = 38.7\Omega$$

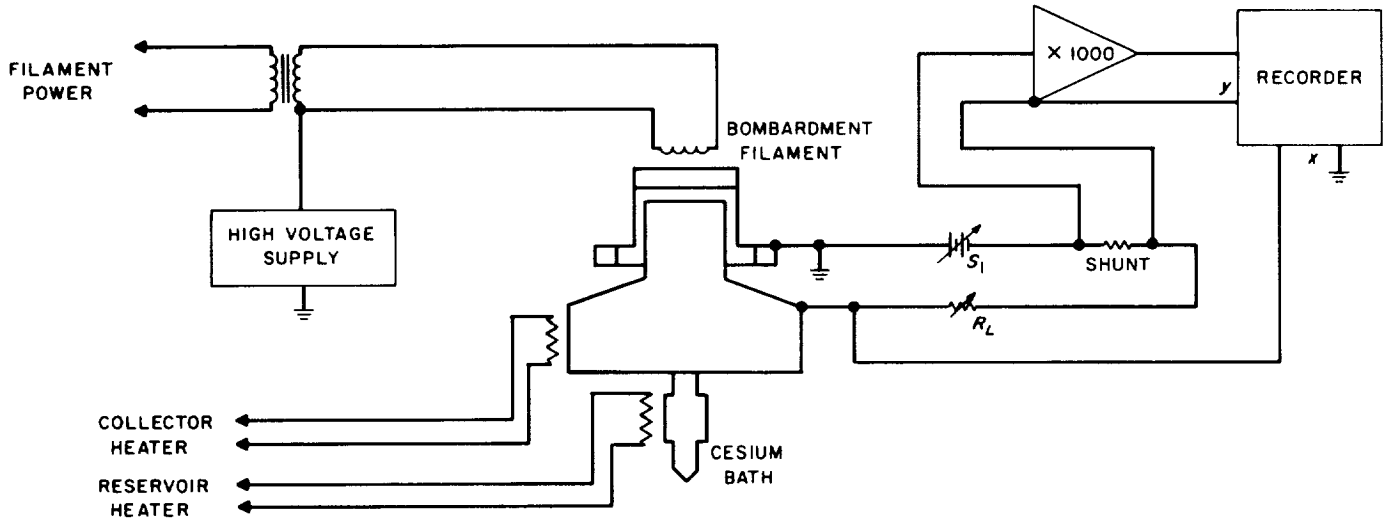


Fig. 8. Circuit diagram for work function measurement

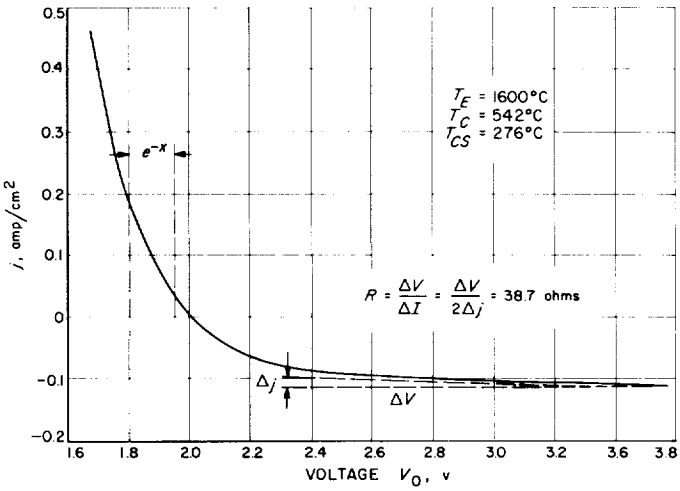


Fig. 9. Converter voltage versus amperes per square centimeter

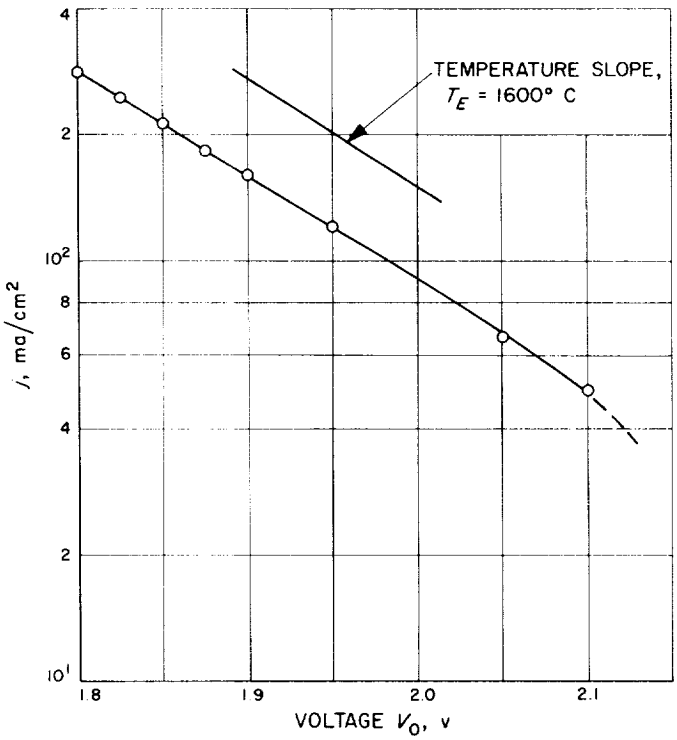


Fig. 10. Logarithmic plot of Fig. 9

Since the operational internal impedance of the converter is in the milliohm region, this shunting resistance has little effect on the converter performance. A knowledge of the resistance does, however, allow determining of the collector back emission. By extending the linear portion of the saturated reverse current curve to the zero volt intercept, the collector back emission may be read directly from the current axis intercept. The reverse current is added to the forward emission since the meter readings account only for net current. This has been done in Fig. 10.

Having established the previous facts, the work function calculation proceeds readily as follows:

(1) Select the point on Fig. 10
$V_0 = 1.9$ volts

(2) At this point
$j = 0.161$ amp/cm²

(3) $\phi_c = kT \ln (AT^2/j) - V_0$
$\phi_c = 1.61$ ev

d. Conclusions. The preceding calculation of work function has two shortcomings. The calculation does not account for edge or patch effects encountered in hardware converters. The calculated work function, therefore, represents an average over the collector surface including all edge effects. The real advantage of this measurement is that it gives a comparison by which the performance degradation, sometime observed in hardware converters, may be correlated to a change in collector work function without sectioning the devices.

B. Electrical Conversion

W. K. Shubert and T. J. Williams

1. Switched Mode Voltage Regulators, W. K. Shubert

a. Introduction. Voltage regulation can be accomplished in several ways. Normal dissipative series regulators are the most common [Fig. 11(a)]. If the input voltage varies appreciably and if the desired output voltage is not close to the input voltage, low efficiency and high heat dissipation result. For spacecraft applications, efficient operation of a regulator with large input voltage variation is very important. Hence, switched mode regulation is required [Fig. 11(b)]. On both the *Ranger* and the *Mariner* spacecraft, switched mode booster regulators have been used with excellent performance.

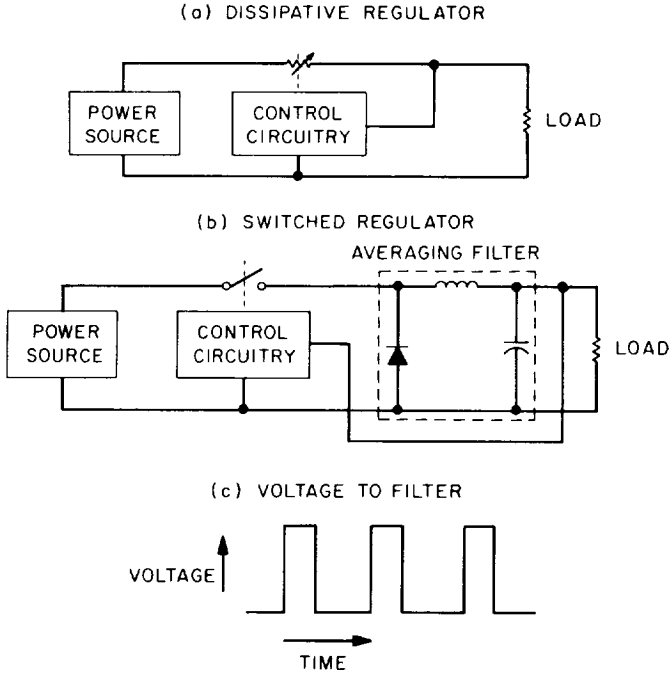


Fig. 11. Switched regulator compared with dissipative regulator

All fast-switching transistors introduced recently have collector-emitter breakdown voltages (BV_{CEO}) on the order of 80 to 100 v. *Booster* regulators to supply a 50-v input require transistors with a BV_{CEO} of twice the maximum input voltage plus derating, or about 150 v. *Switched* regulator transistors, however, would require a BV_{CEO} of only the maximum input voltage plus a safety factor. Therefore, operation could be practical with a 75-v input, using a 100-v BV_{CEO} transistor. With the use of these transistors, the switched regulator would be more efficient and capable of higher power operation than the booster regulator. Fig. 11 shows the switched regulator circuit in comparison with the series dissipative type. With switched operation of a transistor, losses are due only to leakage, switching, saturation, and base drive.

b. Specifications. Preliminary specifications for the switched regulator are given in Table 2. Generally, these specifications are derived from those which would be applicable to *Mariner C*. With a booster regulator, input voltage could not rise above the output voltage. Therefore, with a 52-v output, the solar panel voltage had to be shunt-regulated with zener diodes to 50 v. Higher maximum input voltage will allow the removal of these shunt regulators. Low output voltage with the same output power necessitates more current and heavier wire, which are both undesirable, but the output voltage of a down regulator must be less than the minimum input voltage. Wide load variation was specified to allow for a situation similar to *Mariner C* in which there was a back-up regulator. This regulator was used for light loads during portions of the flight. The *L-C* output filter

Table 2. Design specifications for 400-watt regulator

Input voltage, v	25 to 70
Output voltage, v	20
Load, w	40 to 400
Operating temperature, °C	−10 to +75
Regulation, %	±1
Ripple, v	< 0.2 peak-to-peak
Turn-on transient, v	< 2 overshoot
Line and load transients, v	±2 for 50 msec
Efficiency, %	85 for load > 200 w

requires a specification on the turn-on transient. Overshoot can be higher than input voltage if not considered in the design. Overall load and line transient response is difficult to calculate, but the specified value seems possible and, with design modifications, could be improved. The efficiency specification is based on calculated estimates. This requirement is much higher than *Mariner C* (74%) because of the faster switching and the exclusion of system losses which were included in the *Mariner C* value. A block diagram of the switched regulator designed to meet these requirements is shown in Fig. 12.

c. Circuit operation. The main switch is two 2N1937 transistors manufactured by Texas Instruments. This transistor is not as new or fast as others available. However, this device is readily available, is almost as fast as other transistors, has excellent saturation characteristics, and appears more reliable than later transistors.

Open-loop input regulation is a prediction of the pulse width required to keep the output voltage constant with varying input voltage. Adding this control allows the feedback loop gain to be smaller. The input voltage control of pulse width is a bias winding on the magnetic amplifier. Current through this winding increases linearly with increasing input voltage. Although a hyperbolic function would be ideal (Fig. 13), a linear function was generated because it required fewer parts and fewer initial adjustments. Open loop voltage control should

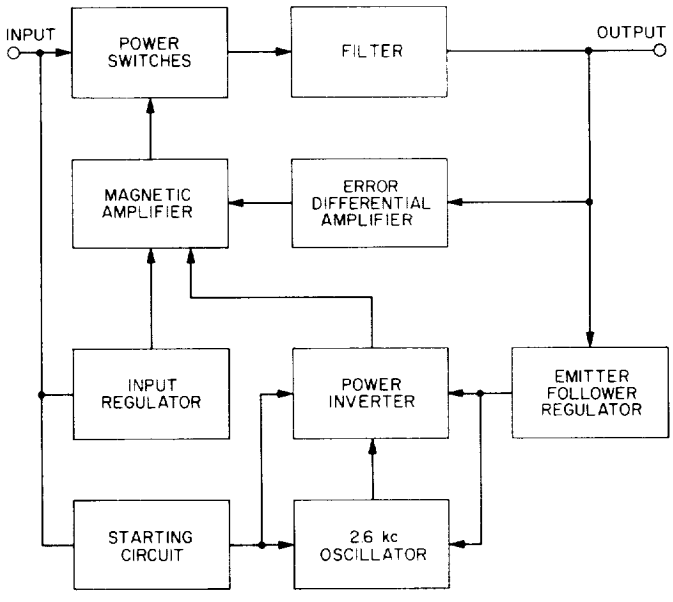


Fig. 12. Diagram of 400-w switched regulator

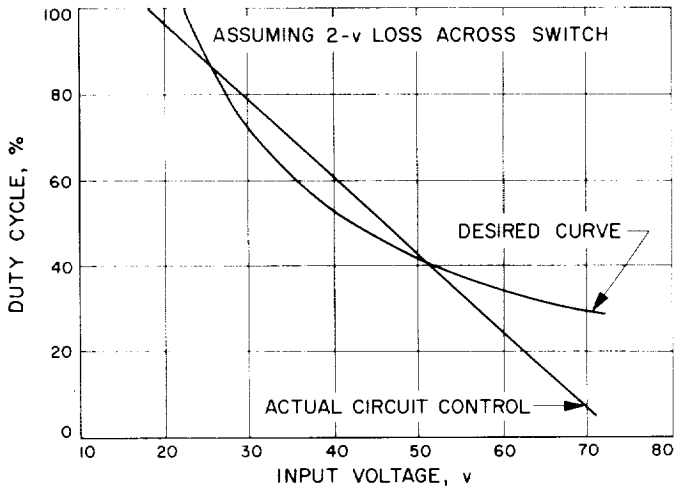


Fig. 13. Duty cycle versus input voltage of 400-w switched regulator

also reduce output transients due to line changes because a prediction of pulse width is made as the line voltage changes.

A two-stage oscillator was used to attempt to keep the frequency of operation constant with a synchronizing source. Since synchronization is not necessarily required, the small oscillator could be discarded. Originally, the oscillator and power inverter were operated directly off the output voltage. Unfortunately, the magnetic amplifier pulse width is sensitive to drive voltage, and a positive feedback loop occurred. This loop was eliminated by adding a stage of regulation to isolate the magnetic amplifier drive from the output.

The operating frequency of 2.6 kc was selected as a compromise between switching losses, filter size, and transient response time. Increased frequency would mean a lower-weight filter, better transient response, and higher switching losses in the main transistors. Even with this low frequency, the storage time of the 2N1937 transistors has created a loop problem for the light loads (40 w) and high input voltages ($>$ 60 v). Under these conditions, a small variation in base drive creates a large change in collector pulse width, (Fig. 14). Therefore, the loop has an artificially high gain which makes the regulator unstable. This condition must be avoided. The simplest method is to replace the transistors with a type of transistor that has a lower storage time. Another approach is to redesign the transistor turn-off circuitry to reduce storage time. An automatic reduction in storage time would occur if the base drive current was proportional to the output current. As yet no modification

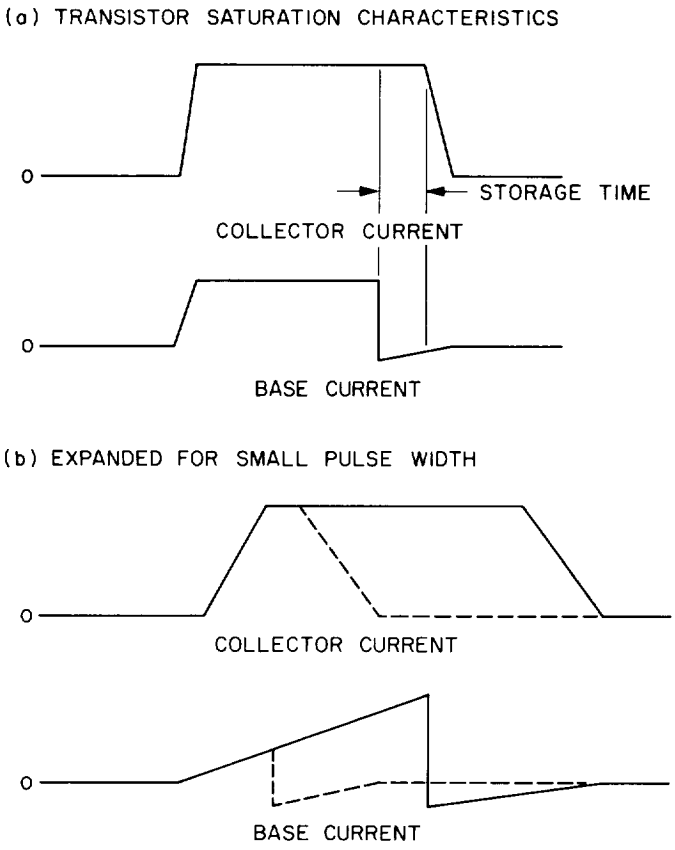


Fig. 14. Effect of base drive variation on collector pulse width

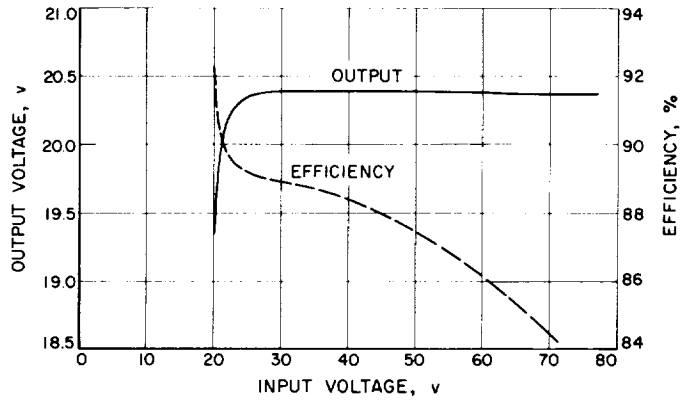


Fig. 15. Output voltage and efficiency versus input voltage for a nominal 235-w load

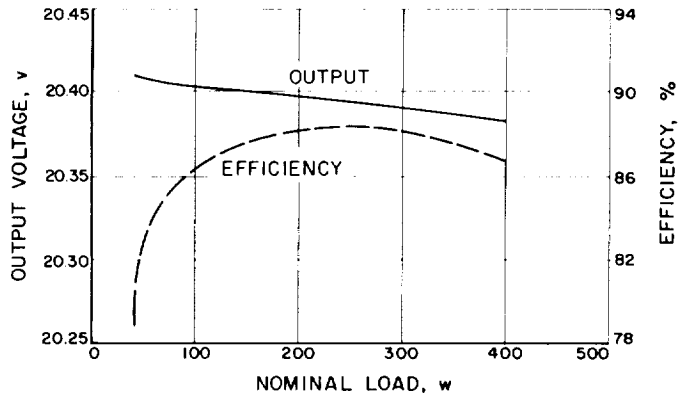


Fig. 16. Output voltage and efficiency versus nominal load for a nominal 35-v input

has been made. Since this instability does not affect operation at other input voltages, efficiency and regulation tests were made with the present configuration. The results are shown in Figs. 15 and 16.

d. Future work. Because of the higher speed of lower voltage transistors, the incorporation of the switched-mode down regulators into spacecraft power systems would improve the efficiency with solar panel-battery sources. Transient tests will, therefore, be run with the present configuration to determine what the response would be. Tests will also be made with different switch frequencies and transistor types to determine what effect these parameters have on transient response and efficiency. Circuits will be selected that can meet the requirements of future space projects.

2. Base Modulated Low-Voltage Converter,

T. J. Williams

a. Introduction. The use of the thermionic diode for supplying electrical power to spacecraft has made necessary the development of new techniques for the efficient utilization of the low voltages made available by this device. Because of the low voltages, the electrical current levels necessary for typical power requirements are often very high. Therefore the low-voltage, high-current power must be converted to the higher voltages but lower current levels that are normally required in the spacecraft. This must be accomplished efficiently and reliably without exceeding weight limitations.

Minneapolis-Honeywell, under contract to JPL, has developed two low-voltage converters. One operates from a 0.7-v source at up to 150 amp with an average efficiency of 80%. The other operates from a 3.5-v source at up to 50 amp with an average efficiency of 90%. Although the efficiencies of both units are good, the weight (which was not a design restriction) is excessive, being almost 30 lb each. Also, the units are not regulated. Regulation with load is approximately 7% for the 0.7-v unit and 4% for the 3.5-v unit for load changes of 75% of full load. The regulation for input

voltage change is proportional to the input voltage change. Since a spacecraft normally requires regulated power, several companies are attempting to develop a regulated voltage, high-current source.

b. Design approach. Probably the most straightforward approach to the problem of obtaining regulated power from a low-voltage source is to convert the low voltage to a higher voltage, as with the M-H converter. The higher voltage can then be regulated with a switched regulator, as is done in present systems. This approach, however, has the inherent disadvantages of poor efficiency and excessive weight. In addition, reliability is not as good as desired because of the large number of parts involved. The efficiency is low and the weight is high because of the double switching process. If we can assume an 80% efficiency for the low-voltage converter and an 80% efficiency for the switched regulator, then the overall efficiency is only 64%.

The other approach, as is proposed here, is to regulate as well as up-convert in the same switching process. It should be possible to maintain efficiencies only slightly less than that of the basic low-voltage converter unit. In addition, the weight would be substantially less than in the previous approach. The unit will, however, be heavier than a converter of an equivalent power rating which utilizes standard input voltages because of the large conductors and bus wire required to carry the heavy currents.

The technique to be used is to pulse-width modulate the base drive to the switching transistors in the converter. An error signal which is obtained by comparing the DC output of the converter to a reference voltage would cause the duty cycle of the drive to the switching transistors to be varied so as to maintain a constant output voltage. Either a magnetic amplifier or solid state techniques may be used for this purpose.

c. Problem areas. The greatest single problem is the limitations of the switching transistor. The input voltage to be switched may be as low as 0.5 v or as high as 4 v, depending upon the number and the electrical connection of the thermionic diodes. If the power level to be switched is, for example, 200 w, then the switched currents may be anywhere from 50 to 400 amp. This calls for a transistor with extremely low saturation resistance and high collector current capability. Fortunately, there are a few such devices available. Germanium transistors exhibit saturation characteristics that are superior to those made of silicon. At the lower input voltages (<2 v), it is imperative that germanium units be used. The MHT 2205 made by Minneapolis-Honeywell has a collector-to-emitter saturation voltage ($V_{CE(SAT)}$) of 0.1 v at 50 amp. Another unit, the MHT 2101, has approximately the same $V_{CE(SAT)}$ at 150 amp. Both units have a collector breakdown voltage of 5 v, which limits their use in converter applications to maximum input voltages of approximately 2 v. Higher voltage germanium units are available, but their saturation resistance is higher (0.45 v at 65 amp).

The germanium units have two disadvantages: First, they are slow switching (t_r = 11 μsec, t_s = 30 μsec, t_f = 17 μsec); this limits the upper frequency at which they can be used due to excessive switching losses. Second, the maximum junction temperature is 110°C, which limits the maximum power dissipation; this depends upon the thermal resistance from the junction to ambient.

An MHT 2205 operating from a 2-v source and switching 25 amp (1-kc rate) will dissipate approximately 5 w, of which 3 w are switching losses, 1.25 w are saturation losses, and 0.75 w is drive loss. The maximum amount it can dissipate at an ambient temperature (T_A) of 75°C is 12 w if a total thermal resistance from junction to ambient (θ_{JA}) of 3°C/w is achieved.

At input voltages greater than 2 v, it becomes practical to use certain silicon transistors such as the MHT 8301 for switching. This transistor has $V_{CE(SAT)}$ = 0.6 v at 10 amp. This would result in a 3-w saturation loss for 50% duty cycle, but it is offset by the greatly reduced switching losses. Also, the maximum junction temperature is increased to 200°C, allowing for greater heat dissipation.

Another problem that may be critical with germanium transistors is that of obtaining a low θ_{JA} in order to dissipate the power lost without exceeding the maximum allowable junction temperature. One possible solution to this problem is to use a modified common emitter connection for the switching transistors whereby it is possible to electrically connect all the collectors. Since the collectors of the transistors are electrically connected to the case, this approach permits the mounting of all transistors on a common heat sink without insulators, thereby decreasing the thermal resistance from case to sink, which is one of the three thermal resistances that make up θ_{JA}. Such an approach may, however, cause electrical problems in the drive circuit due to interwinding capacitance of the drive transformers. This problem will be investigated.

Other problems include matching transistors to be used in parallel and eliminating spikes and other circuit disturbances due to stray circuit inductance and large switched currents.

C. Energy Storage

W. L. Long

1. Sealed Silver-Zinc Battery Development,

W. L. Long

Improvement in spacecraft batteries is a continuing need. Higher watt-hours per pound, greater temperature range, longer life, and better reliability are areas where improvement can be utilized. In addition, secondary batteries require greater cycle life and, in some cases, the capability of accepting a continuous trickle charge.

Contracts were entered into with Yardney Electric Corporation and Power Sources Division of Whittaker Corporation for each company to develop, independently, sealed secondary silver-zinc battery cells with approximately the following characteristics:

Capacity, 43-amp-hr above, 1.43 v at 10 amp

Temperature range, 30 to 140°F

Weight, 610 g maximum

Cycle life, 3 or more

Trickle charge acceptance, 0.020 amp into fully charged cell for 15 days.

After the cells have been designed, fabricated, and successfully tested, 18 cells will be fabricated and assembled into a battery for delivery to JPL. Here each battery will be subjected to rigorous tests, including electrical performance and environmental and life testing. Power Sources has built twenty-five 4 amp-hr cells with variations in the following parameters:

$$(1)\ \text{HgO content} = \frac{\text{weight HgO}}{\text{weight (HgO + ZnO)}} \times 100$$

$$(2)\ \text{Cell fit} = \frac{\text{total dry free space + total dry membrane thickness}}{\text{total dry membrane thickness}}$$

$$(3)\ \text{Material ratio} = \frac{\text{total weight ZnO}}{\text{total weight Ag}}$$

$$(4)\ \text{Electrolyte ratio} = \frac{\text{volume of electrolyte}}{\text{plate area} \times \text{dry thickness of cell pack}}$$

$$(5)\ \text{State-of-charge} = \frac{\text{initial capacity of Zn}}{\text{theoretical capacity of Ag}} \times 100$$

Each cell contained 6 silver and 7 zinc plates; plate area was 29.4 in.^{2}. Separator materials were Viscon 3001 on the negative, and 4 layers of Permion 600 on the positive. Each cell was fitted with a pressure gage and sealed under vacuum before the first charge. Charging current was 0.20 amp to 2.00 v, then 0.10 amp to 2.00 v. Cell pressures remained below 5 psi and returned to 10 to 25-in. Hg vacuum upon stand or discharge. Discharge current was 2.0 amp on first and third cycles, and 1.0 amp on the second cycle. Cell capacities and plateau voltages were recorded and tabulated.

Initial conclusions from this work are: (1) An electrolyte ratio of less than 0.45 materially reduces capacity. (2) Presence of 3 or 4% mercuric oxide in the positive plate increases cell capacity. (3) Increasing cell fit ratio from 2.0 to 2.5 increases cell internal impedance. (4) Variations in material ratio and state-of-charge has little or no effect on voltage or capacity.

All of this work was at room temperature. Selected cells will be cycled at 30°F and 140°F; some of them will be trickle-charged. The results of this work will be incorporated into 43 amp-hr cells for further evaluation and testing.

References

1. Nottingham, W. B., *Thermionic Emission*, Technical Report 321, p. 28, Research Laboratory of Electronics, M.I.T., December 10, 1956.

2. Kitrilakis, S. S., and Weinstein, J. H., *Second Annual Technical Summary Report for the Thermionic Emitter Materials Research Program*, Report 27-64, Thermo-Electron Engineering Corporation, Waltham, Mass., September 30, 1963.

IV. Guidance and Control Research

A. Magnetics Research

F. B. Humphrey and K. Kuwahara[1]

1. Magnetoresistance Investigation of RIS Films

a. Introduction. The anomalous magnetic anisotropy found in vacuum-evaporated Ni-Fe films is generally a well-behaved uniaxial anistropy (*RS 36-11*, pp. 16–20). Some time ago, Wilts (*RS 36-12*, Vol. I, pp. 40–43) showed that the anistropy would assume a considerably different character when the film was subjected to controlled oxidation and reduction. It seemed as if the anisotrophy axis could be changed by relatively weak magnetic fields, even less than those reported by Prosen (Ref. 1). In summarizing all of the then-known anomalous effects in thin films, Cohen (Ref. 2) named this effect "Rotatable Initial Susceptibility" (henceforth called RIS). Two types were observed by Cohen, called RIS-I and RIS-II. Wilts had fabricated RIS-II and, subsequently (*SPS 37-22*, Vol. IV, p. 26), he fabricated RIS-I by his simple oxidation-reduction procedure.[2] Most of the investigations have been made using a hysteresis loop tracer although some preliminary studies using a torque magnetometer have also been made (*SPS 37-22*, Vol. IV, p. 29). Recently Wilts (Ref. 3) finished a comprehensive study of both types, but again the mode of investigation was limited to hysteresis and torque curve studies. Since the detailed mechanism of this strange behavior is still not understood, it seemed reasonable to use the magnetoresistance effect (*SPS 37-26*, Vol. IV, p. 47) to continue the investigation.

This investigation will be mainly concerned with RIS-II films. At reasonable drives, say 10 Oe, the hysteresis loop is essentially the same in all directions and, for low drives, the anisotropy field, H'_k, is smallest along the direction of a previously applied AC field (Ref. 3). For very low fields, the torque curve is proportional to sin α; and for higher fields above a threshold, H_m, a large constant torque appears. The value of the rotational hysteresis integral has a value of 3.5 independent of thickness (Ref. 3). The value of H_m decreases as the thickness increases.

b. Experimental method. To avoid ambiguity of the initial condition, before each measurement a field of 260 Oe (highest field available with the present apparatus) was applied parallel to the plane of the film and then rotated in the plane of the film more than one revolution in the negative direction, and stopped at a desired angle. If the field is then reduced through zero,

[1] Formerly Resident Research Appointee to JPL, now returned to the University of Hiroshima, Faculty of Engineering, Hiroshima, Japan.

[2] It should be noted that the terms I and II have been reversed from that used in *RS 36-12* and *SPS 37-22* (Vol. IV, pp. 26, 29) to make the nomenclature consistent with Ref. 2.

to large negative fields, magnetoresistance measured as a function of the field is a measurement that can be conveniently compared to a hysteresis loop measurement. If, however, the field is decreased to a certain value and then rotated, the magnetoresistance measured as a function of the rotation angle can be conveniently compared to a torque measurement.

Experimental data, which will be described hereafter, is generally for one typical film. It is 1-cm square, 2000 Å thick, vacuum-evaporated from a melt of 80% Ni-20% Fe at a pressure of 10^{-5} torr. The glass substrate temperature during evaporation was 300°C. It was then heated to about 450°C for 5 min in dilute oxygen, then reduced at 450°C for 15 min in pure hydrogen. Vacuum-evaporated gold contacts were then placed along two sides such that the current was along the original easy axis.

c. Results. The resistance of the RIS-II films is a function of the rotation angle and the field magnitude. The results of one experiment are shown in Fig. 1. A field of 260 Oe is rotated in the negative direction, and the resistance is measured as a function of the direction of applied field α. Occasionally, the rotation is stopped and the field reduced at angles indicated. Each vertical line in Fig. 1 corresponds to the change in resistance when the field is removed, with the final resistance for zero field shown as a horizontal bar. Note that the bar is not always at the end of the vertical line. The $\cos^2 \alpha$ part of this curve is as would be expected for H_0 where the magnetization would be essentially aligned along the field direction. As the field is removed, the magnetization relaxes back in the direction from which it came. As was shown in detail in Eq. (2) of *SPS 37-26*, Vol. IV, p. 47, the magnitude of the final resistance indicates the average resultant angle between the magnetization and the current. The results for the data of Fig. 1 are shown in Fig. 2, where the predicted angle through which M has relaxed is shown as a function of applied field. The average angle is about −35 deg, where the negative value indicates that M relaxes back to the direction from which it came.

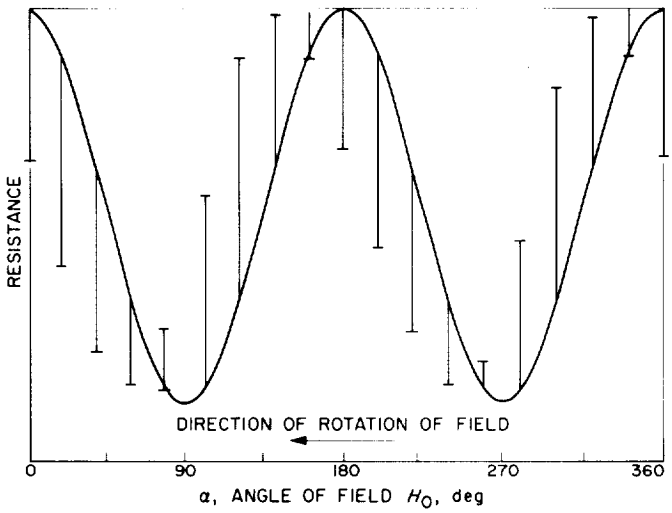


Fig. 1. Resistance change of RIS-II films with field directions

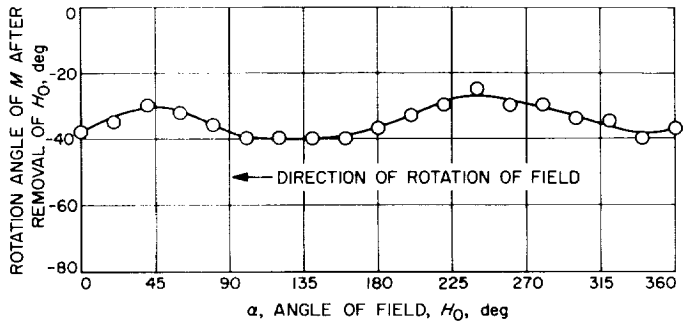


Fig. 2. Relaxation angle upon removal of field as a function of angle of applied field

For many experiments, this procedure to establish an "easy" axis by rotating a large field (H_0) and then stopping at some angle, α, will be followed. In order to avoid ambiguity, the angle of the field during such a procedure will be referred to as α_0. Details of the resistance change for three angles α_0 can be seen in Fig. 3, where the resistance as a function of field is shown for decreasing fields with angle α_0 as a parameter. The resistance change corresponding to a decrease in field from high positive values to zero corresponds to the data of Fig. 1. The field is further changed to negative values in a manner similar to a loop tracer as shown. The extreme values of resistance, corresponding to the expected values when M rotates coherently, are indicated as $R_\parallel$ and $R_\perp$. Since discontinuous changes in these curves are not observed, it can be concluded that, during this change of the field, rotation of the magnetization takes place in a nearly coherent manner without discontinuities.

The rotation angle of M as a function of the applied field can be calculated for each curve in Fig. 3, as was done previously. The results for the three curves are plotted in Fig. 4. The ordinate is the calculated rotation angle of M measured from the direction of the field. This curve is essentially the same as the usual curve of M as a function of H.

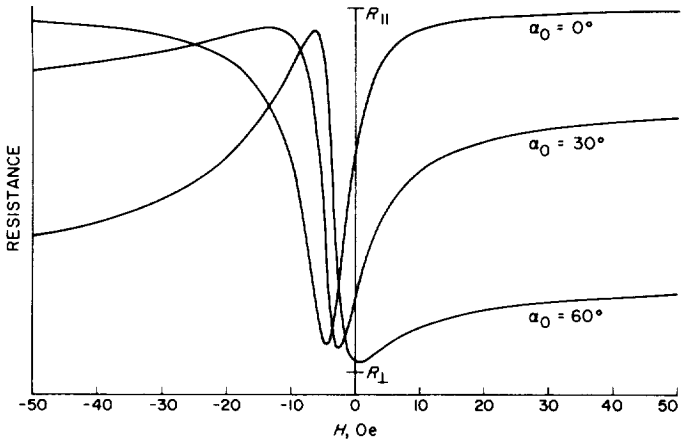


Fig. 3. Magnetoresistance as a function of field with angle α_0 as a parameter

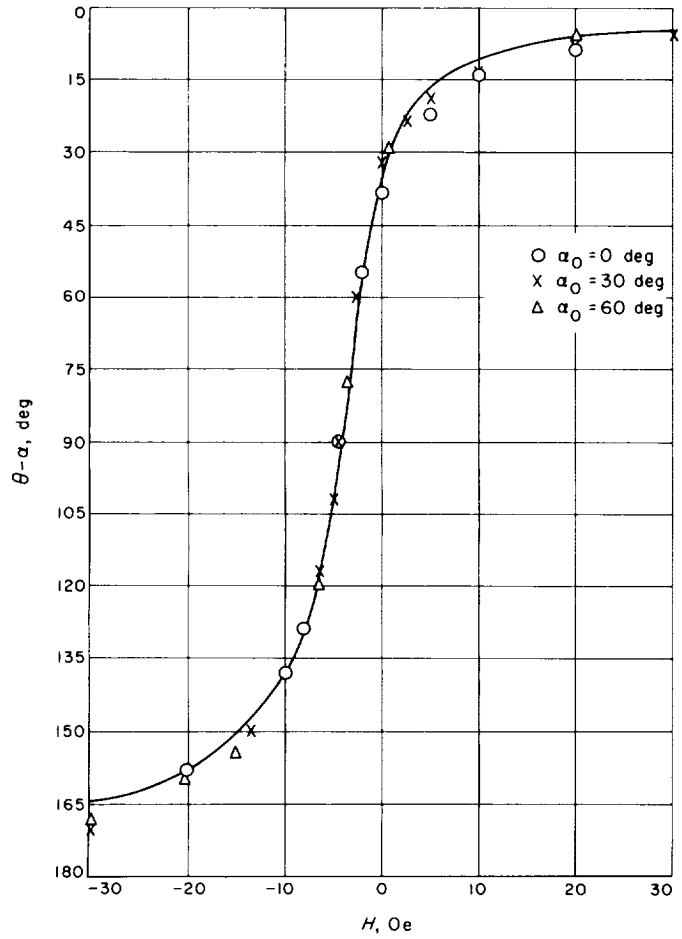
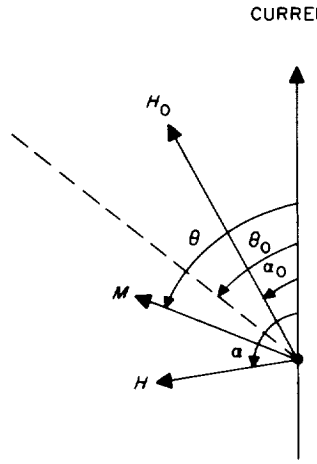


Fig. 4. Rotation angle, $\theta - \alpha_0$, as a function of field for an applied field derived from the data of Fig. 3

A phenomenological fit to the data of Fig. 4 can be obtained by the expression:

$$K \sin \frac{\theta - \theta_0}{2} = M H \sin (\alpha - \theta), \qquad (1)$$

where, as can be seen in Fig. 5, H is applied at some angle α to the direction of the current and, for this experiment $\alpha = \alpha_0$, M is in equilibrium at some angle θ from the current, and θ_0 is the angle of magnetization when $H = 0$. The solid curve in Fig. 4 is obtained by choosing 35 deg and 8.7 Oe as $(\theta_0 - \alpha_0)$ and K/M, respectively.



DEFINITIONS OF SYMBOLS

H	MAGNETIC FIELD
H_0	LARGE MAGNETIC FIELD USED TO CHANGE "EASY AXIS"
α	ANGLE BETWEEN CURRENT AND APPLIED FIELD H
α_0	ANGLE α WHEN "EASY AXIS" IS ESTABLISHED
θ	ANGLE BETWEEN THE AVERAGE MAGNETIZATION AND THE CURRENT
θ_0	ANGLE BETWEEN THE "EASY AXIS" AND THE CURRENT
H_K	ANISOTROPY FIELD FOR NORMAL FILM
H_K'	ANISOTROPY FIELD FOR RIS FILM
H_m	FIELD FOR MAXIMUM HYSTERESIS LOSS (WILTS, REF. 3)
H_{crit}	FIELD FOR MAXIMUM HYSTERESIS LOSS FROM MODEL
R_{II}	RESISTANCE OF FILM WHEN A LARGE FIELD IS APPLIED PARALLEL TO CURRENT FLOW
$R_\perp$	RESISTANCE OF FILM WHEN A LARGE FIELD IS APPLIED PERPENDICULAR TO CURRENT FLOW

Fig. 5. Definition of symbols of various angles; the reference direction is the direction of the current

d. Torque curves. As mentioned previously, when the field is rotated in the plane of the film while the magnetoresistance is measured, data similar to torque curves can be obtained. Such curves can now be calculated from Eq. (1). Fig. 6 compares the calculated curve with experimental points for a field of 1 Oe. This curve is typical for $H < 0.5\,(K/M)$ (in this case 4.4 Oe). For $0.5\,(K/M) < H < (K/M)$, the calculated curve shown

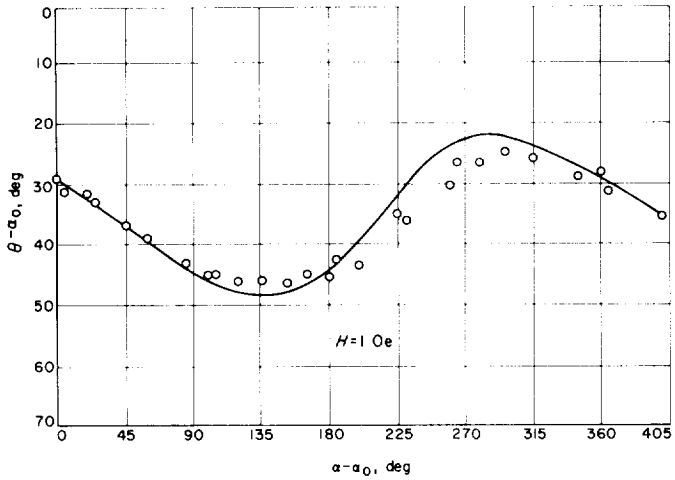


Fig. 6. Relation between the field direction ($\alpha - \alpha_0$) and *M* direction ($\theta - \alpha_0$) when $H = 1$ Oe is rotated in the film plane; the circles are obtained by magnetoresistance measurement on the same sample used for Fig. 1; the curve is calculated from Eq. (1)

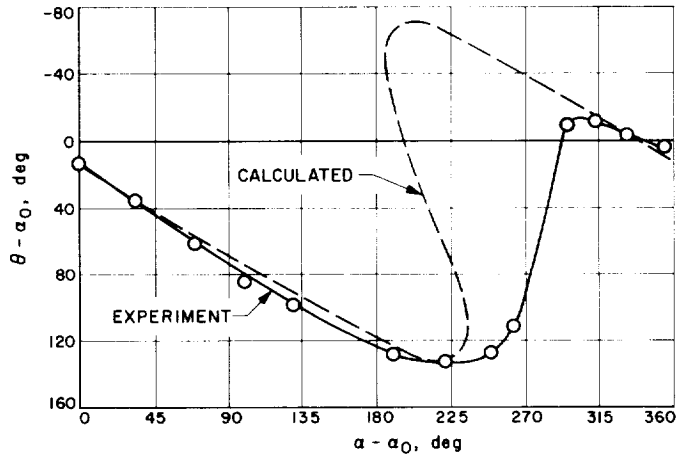


Fig. 7. The same relation as that in Fig. 6, when $H = 7$ Oe, the same sample as used for Figs. 1 and 6

in Fig. 7 ($H = 7$ Oe) has a range of angles where triple values of θ are predicted. As can be seen, a rapid variation (a definite discontinuous change is seen in the original magnetoresistance curve) is observed from $\alpha = 260$ deg to $\alpha = 290$ deg, presumably caused by wall motion that can be considered as a transition between two stable states in the calculated curve, indicated by the dashed line. At fields higher than K/M, experimental points show the nearly constant torque as indicated in Fig. 8 ($H = 10$ Oe). The experimental data are somewhat at variance with the calculated curve, but the general shape is still expressed by the equation.

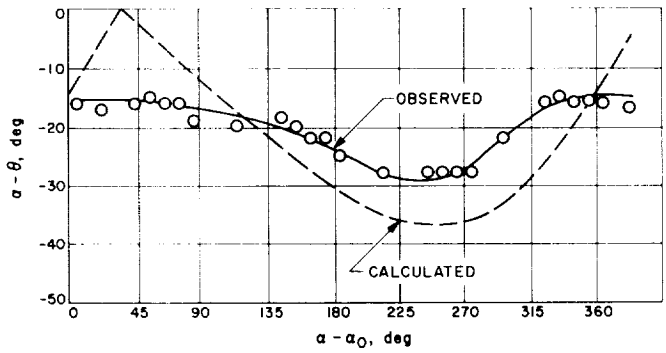


Fig. 8. Relation between ($\alpha - \theta$) and ($\alpha - \alpha_0$) when $H = 10$ Oe; the circles are obtained by magnetoresistance measurement; the curve is calculated from Eq. (1), the same sample as for Figs. 1 and 6

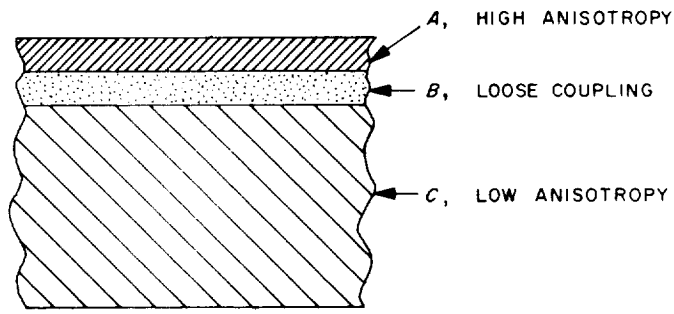


Fig. 9. A schematic cross section of a proposed model for RIS-II films; the surface layer *A* has high anisotropy fields distributed at random in direction and magnitude; the bottom layer C has low anisotropy fields; the middle layer *B* is a transition region, and a loose coupling is assumed

e. A proposed model for RIS-II films. A schematic cross section of a film, illustrating a model consistent with Eq. (1), is illustrated in Fig. 9. The film is considered as composed of three layers. The bottom layer, *C*, has a usual uniaxial anisotropy which is randomly distributed in magnitude as well as in direction with H_k values usually greater than those of layer *C*. The transition layer between *A* and *C* is indicated as B. It is assumed that the thicknesses of *A* and *B* are small compared with the total thickness of the film; therefore, the observable magnetic characteristics of the sample are due to the layer *C*. Exchange interactions act to align

the magnetization within the layers A and C. Also, between the layers A and C interaction is through the layer B in which a loose coupling is assumed.

According to such a model, when a field, greater than H_K for layer C, but less than H_K for layer A, is applied and rotated in the film plane, the average magnetization direction in A should be behind the field direction. Meanwhile the magnetization in the layer C should be almost in the same direction as that of the applied field. When the field is removed, the torque exerted on C by the field vanishes, and hence C becomes aligned to A by way of the exchange through B. This relaxation is the origin of $(\theta_0 - \alpha_0)$.

The exchange coupling between the magnetization in the layers A and C through B is plausible as the origin of $(\theta_0 - \alpha_0)$, but difficult to handle quantitatively. An estimation of the exchange interaction can be obtained as follows. The boundary, B, has the same configuration of magnetic spins as that presumed in a Bloch wall, when the magnetization direction is different for the layers A and C since the magnetization rotates in planes parallel to the wall. Assuming that each successive rotation angle between adjacent spins is the same, exchange energy, E_{ex}, stored in a unit area of B is expressed, similar to the Bloch wall energy (Ref. 4), as

$$E_{ex} = \frac{A}{T_B}(\Delta\theta)^2 \tag{2}$$

where A is a constant related to the exchange coefficient, T_B is the thickness of the layer B, and $\Delta\theta$ is the total angle between the magnetizations of A and C. The torque L_{ex} exerted on C by A is

$$L_{ex} = \frac{\partial E_{ex}}{\partial(\Delta\theta)} = 2A(\Delta\theta)/T_B . \tag{3}$$

On the other hand, the torque L per unit area derived from Eq. (1) for small $\Delta\theta\left(\sin\Delta\theta/2 - \frac{\Delta\theta}{2}\right)$ and for thickness T_C is

$$L = K T_C \frac{\Delta\theta}{2} . \tag{4}$$

Here, T_C is the thickness of the layer C.

From Eqs. (3) and (4),

$$K T_C = 4A/T_B . \tag{5}$$

Putting experimental values into K and T_C and assuming $T_B \sim 200$ Å, A is estimated as

$$A \sim 6 \times 10^{-8} \text{ erg/cm} .$$

This value of A is less by one order of magnitude than the known value of A for usual ferromagnets under ordinary conditions (Ref. 4). In the present model, then, B is considered as a loose coupling layer with the loose coupling attributed to vacancies or some other defects in the layer.

f. Dependence of threshold upon thickness. The existence of a threshold was mentioned earlier in Part *d*, where Eq. (1) was solved as is done for torque curves. It was noted that characteristic solutions were obtained for fields less than $K/2M$, greater than K/M and between the two values. The thickness dependence can be obtained in a similar way but after rewriting Eq. (1) as a balance of torques per unit area:

$$K T_C \sin\frac{\theta - \theta_0}{2} = M H T_C \sin(\alpha - \theta) ,$$

As mentioned in Eq. (5), $K T_C$ is a constant which can be called K'. Then

$$K' \sin\frac{\theta - \theta_0}{2} = M H T_C \sin(\alpha - \theta),$$

and the critical field is

$$H_{crit} = \frac{K'}{M T_C} .$$

It is noted that, as known from Eq. (1), a field higher than this H_{crit} can pull M around the film. As will be mentioned later, the calculated rotational hysteresis loss is constant above H_{crit} and zero below H_{crit}. The field corresponds to H_m of Wilts (Ref. 3). Assuming that the exchange constant A is the same even for significant changes in composition, and that the depth of oxidation (color before reduction) is the same, the thickness of layers A and B is probably the same for all films, and the critical field should be:

$$H_{crit} \propto \frac{1}{M\left[T - (T_A + T_B)\right]} \tag{6}$$

where T is the total film thickness. Fig. 10 shows the data of Wilts for $M H_m$ as a function of total film thickness for three compositions. The solid line is the proportionality of Eq. (6) with a match such that $T_A + T_B = 300$ Å.

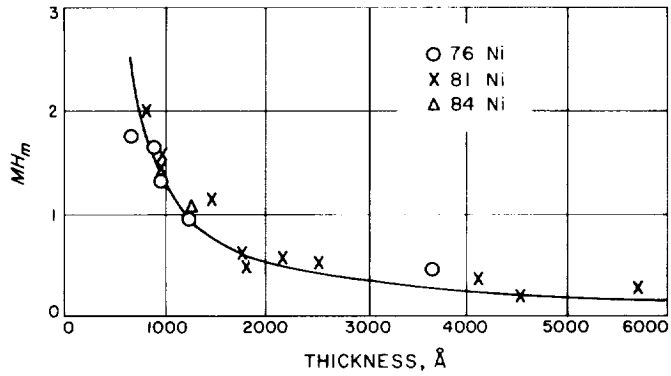


Fig. 10. Experimental points of maximum rotational field, H_m, of RIS-II films as a function of thickness (Wilts); the curve is derived from the proposed model

g. Relation between initial susceptibility and H_M. The anisotropy field, H'_K, can be predicted from Eq. (1) by considering the usual experiment where a small field is applied perpendicular to θ_0 and the change in flux parallel to the field is observed. The initial susceptibility, χ_i, can be calculated as:

$$\chi_i = \frac{dM}{dH} = \frac{d\,(M_s \sin\theta)}{dH} = M_s \cos\theta \frac{d\theta}{dH}$$

from Eq. (1) assuming $\theta_0 = 0$, $\alpha = 90$ deg and a single domain,

$$K/2 \cos(\theta/2) \frac{d\theta}{dH} = M_s \cos\theta - M_s H \sin\theta \frac{d\theta}{dH}.$$

Therefore

$$\chi_i = \frac{M_s^2 \cos^2\theta}{(K/2)\cos(\theta/2) + M_s H \sin\theta} \qquad (7)$$

which, for small H and small θ gives:

$$\chi_i \equiv \frac{M_s}{H'_K} = \frac{2M_s^2}{K}.$$

Hence,

$$H'_K = \frac{K}{2M_s} = \frac{H_{crit}}{2}. \qquad (8)$$

The measured magnetoresistance was converted into fraction of the film saturated. This fraction is plotted in Fig. 11 as M/M_s versus H. The field H'_K. where the extrapolation of the initial linear portion of the curve crosses the value for $M/M_s = 1$, will be $H_{crit}/2$ as can be seen from Eq. (8). In this case the relation seems to be satisfactory since H_{crit} is 8.7 Oe and H'_K is 4.8 Oe.

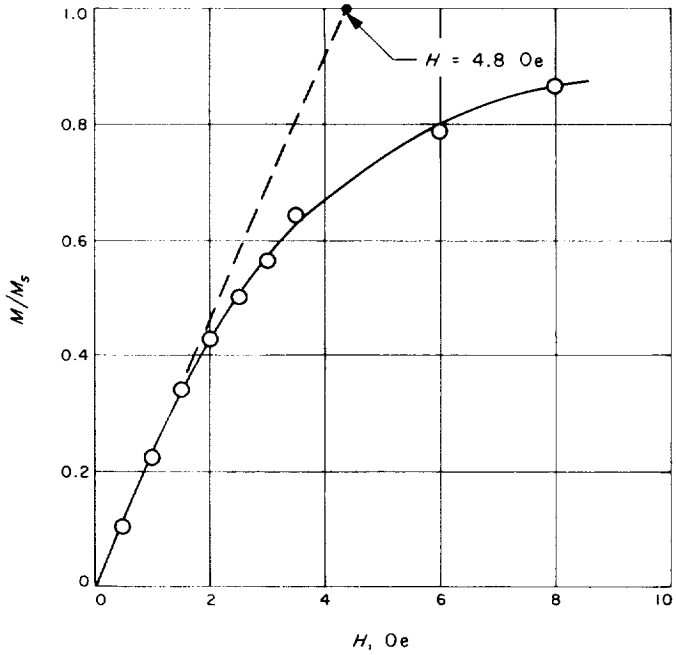


Fig. 11. Experimental increase of M/M_s obtained by magnetoresistance effect as a function of field; the field H is applied at right angles to M, and M/M_s is measured to the same direction

An investigation of H'_K as a function of thickness was performed by Wilts with the use of a loop tracer. His data are plotted in Fig. 12, with the curve calculated by Eqs. (6) and (8), choosing $T_c = 300$ Å and a proportional constant to fit the data. As can be seen, the fit is fairly good except for thinner films. It seems, however, that the relation of Eq. (8) is not always valid through the wide range of thickness shown in Fig. 12, and especially in thinner films since H'_K is larger than H_{crit}, as can be seen by comparing Figs. 10 and 12.

h. Rotational hysteresis integral. The rotational hysteresis integral has been used to indicate the mechanism of rotational hysteresis loss (*SPS* 37-22, Vol. IV, p. 29). The rotational hysteresis loss (per unit volume) can be calculated as the integral of the torque per cycle:

$$W_r = \int_0^{2\pi} M_s H \sin(\alpha - \theta)\, d\alpha \qquad (9)$$

and is shown in Fig. 13. The torque can be calculated from Eq. (1). It is assumed that hysteresis loss occurs in the region where the solution of Eq. (1) is triple-valued. Where the field is greater than K/M, the hysteresis is

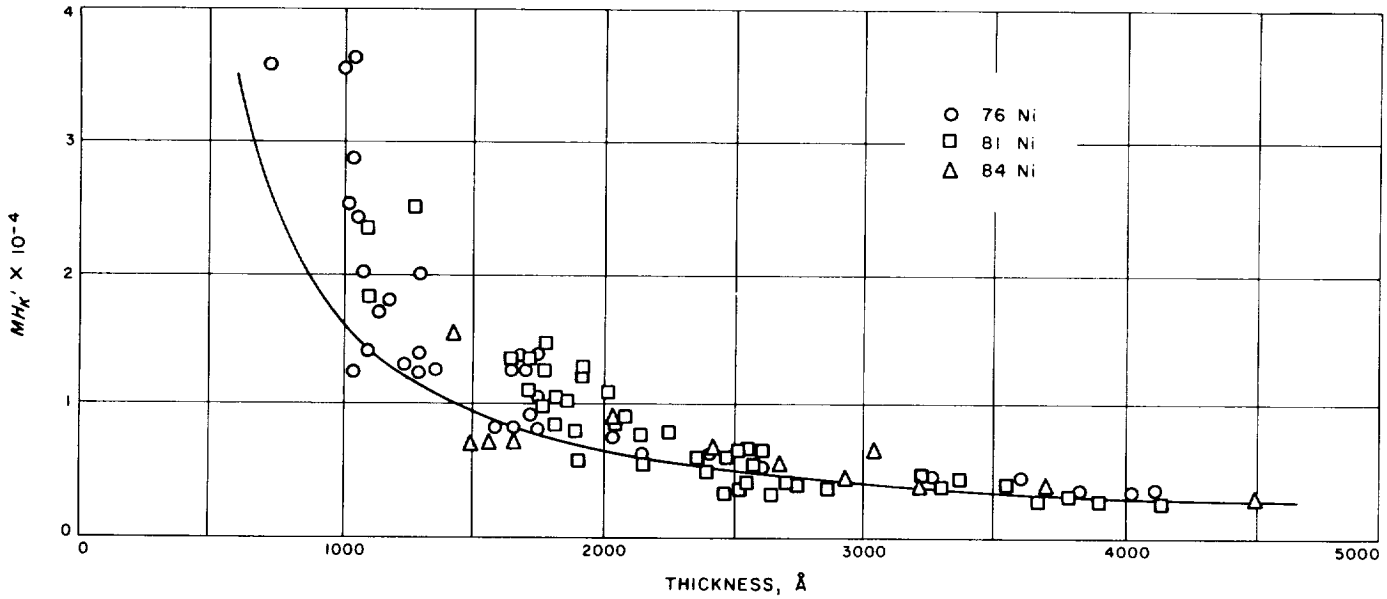


Fig. 12. Experimental points of minimum initial susceptibility field, H_K', of RIS-II films as a function of thickness; the curve is derived from the proposed model

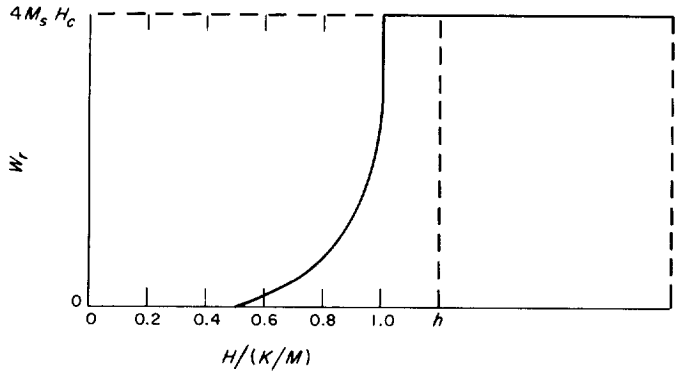


Fig. 13. Calculated rotational hysteresis loss, W_r, as a function of H (solid curve); the dashed curve is expected in taking account of wall motion and dispersion

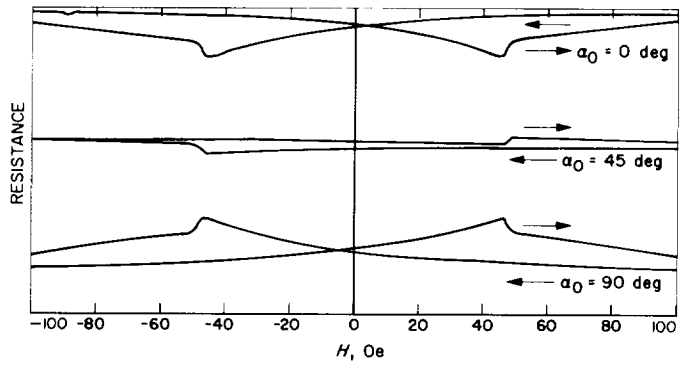


Fig. 14. Magnetoresistance hysteresis loop of an RIS-I film

constant with a value of $4\,H_cM_s$. The integral of rotation hysteresis loss I_H is calculated as

$$I_H = \int_0^\infty \frac{W_r}{M_s}\, d\left(\frac{1}{H}\right) = 4.5. \qquad (10)$$

This value of I_H as well as the behavior of W_r are in fair agreement with those observed in RIS-II films by Wilts.

i. RIS-I film. Typical magnetoresistance hysteresis loops for $\alpha_0 = 0$, 45 and 90 deg are shown in Fig. 14 for an RIS-I film. In contrast with the case of RIS-II films, the total resistance change is small and is sometimes quite sudden as is characteristic with wall motion. This wall motion takes place at the same field for all α_0's, as can be seen in the figure. The magnetization reversal process seems to be composed of: first, dispersion of magnetization around H_0; second, wall motion; and third, decreasing dispersion around $-H$.

Although wall motion was not generally found in the RIS-II films, the wall motion just mentioned in the RIS-I type can also be observed in the DIS-II type if the experiment is performed after the RIS-II film is magnetized along the direction of Oe. A typical magnetoresistance hysteresis loop for an RIS-I film which has been subjected to such a procedure is shown in Fig. 15, in the

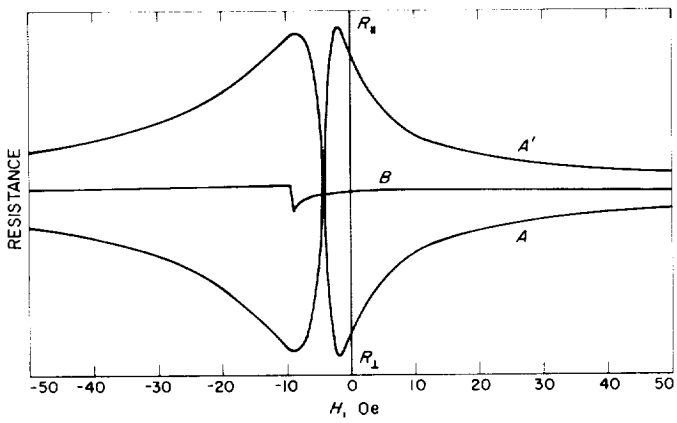


Fig. 15. Magnetoresistance hysteresis loop of RIS-II film (the same one as used to get Figs. 1 through 8), for different preceding treatments

case of $\alpha_0 = 45$ deg. The curves A and A' were obtained for reference after rotating H_0 in negative and positive angle directions, respectively. The curve B shows no large amount of coherent rotation but definite wall motion. In this case it is confirmed that rotation preceding the wall motion is composed of almost the same amount of rotations in the positive and negative directions.

j. Experiment of etching off the surface of the film. To obtain some evidence of the validity of the proposed model, an experiment was performed to etch off the surface layer of RIS-II films. The etchant adopted was "Mirrofe" (Ref. 1). In Fig. 16 are shown hysteresis loops observed before (A) and after (B) the etching for a typical film. The etching time was 260 sec using 90× diluted Mirrofe solution. The decrease in thickness is estimated as about 300 A. A drastic change was produced by taking off the thin top layer, and the RIS-II film was converted into an RIS-I film.

k. Discussion. The poor agreement with loop tracer experiments and torquemeter experiments at higher fields indicates that the formulation discussed above may not be valid for high field strengths. Also the single domain nature of the formulation is for a situation where the single domain character is far from assured. The inability to clearly state the situation observed in demagnetization, plus the problem encountered in the formulation when θ_0 is not fixed, leaves more to be done. In spite of these limitations, the formulation of this model for RIS films has been quite fruitful in predicting many of the observed effects and in pinpointing the areas where understanding is lacking.

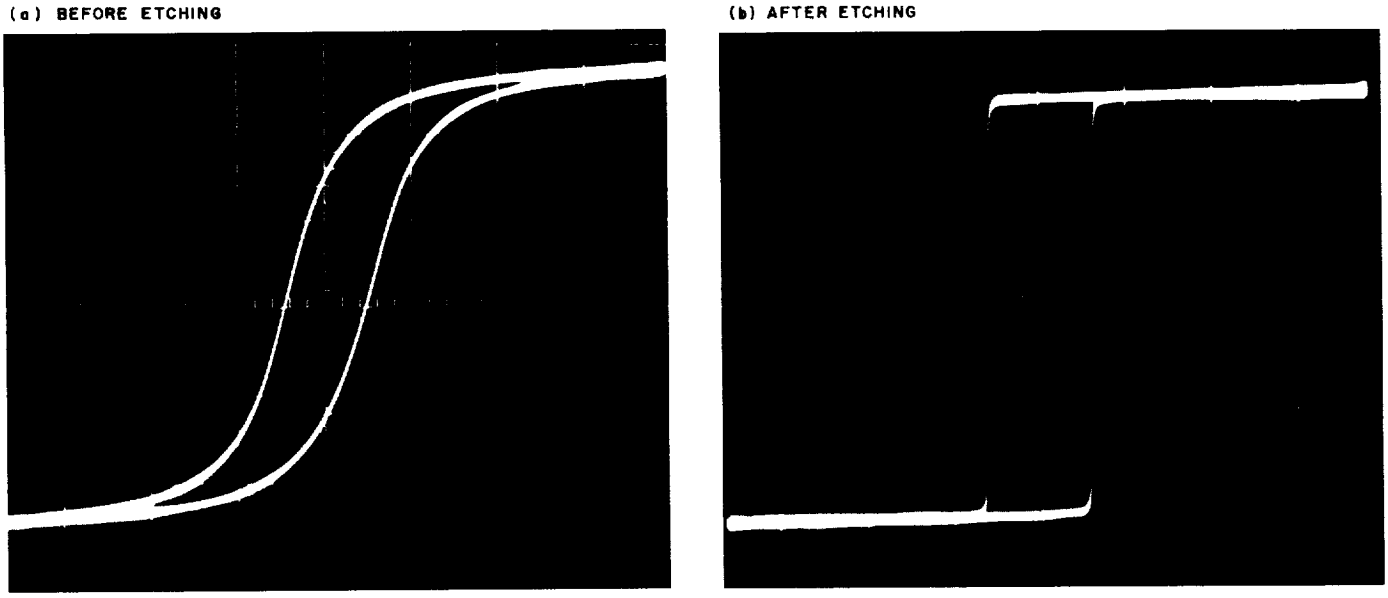


Fig. 16. Hysteresis loops before (a) and after (b) etching

References

1. Prosen, R. J., Holmes, J. O., and Green, B. E., *Journal of Applied Physics,* 32: 915, 1961.

2. Cohen, M. S., *Journal of Applied Physics,* 33: 2964, 1962.

3. Wilts, C. H., "Quasi-static Properties of RIS Films of Oxidized Permalloy," *Journal of Applied Physics,* 35: 2097, 1964.

4. Kittel, C., *Introduction to Solid State Physics,* Second Edition, p. 433, John Wiley & Sons, New York, 1956.

ENGINEERING MECHANICS DIVISION

V. Materials Research

A. Electronic Materials

I. Weinberg

1. Thermoelectric Power in Metals and Alloys

The thermoelectric power is perhaps the most sensitive electronic transport property of a metal (Ref. 1). This being the case, it can serve as a sensitive probe into such solid-state phenomena as electron–electron, electron–impurity, and electron-phonon interactions. Furthermore, with respect to ultrahigh purity, the thermoelectric power is highly susceptible to the presence of trace elements in solid solution. Accordingly, a measurement program has been instituted on a series of copper alloys, and an apparatus has been constructed covering the temperature range from 4.2 to 320°K.

Since alloying effects are studied in relation to pure copper, initial efforts have been concentrated on this material. Thermal potentials are measured using a Rubicon thermofree potentiometer. Temperatures from 4.2 to 77°K are determined by germanium resistance thermometers, with copper–constantan thermocouples in use above 77°K. The thermoelectric power is obtained by differentiating the curve of voltage versus temperature, using a computer program to obtain a least-squares fit to the experimental data. The present results (solid curve in Fig. 1) exhibit two pronounced peaks at 8 and 75°K. Experimentally, the low-temperature peak is attributed to the presence of trace amounts of iron (Ref. 2). Recent theoretical work (Ref. 3) attributes this peak to spin-dependent scattering of conduction electrons by trace amounts of magnetic ions coupled via the Overhauser giant spin density waves. The high-temperature peak results from a competition between Umklapp and normal phonon-drag processes (Ref. 2). The magnitude of this peak is susceptible to the presence of impurities having significant mass difference with respect to the host material.

The present data are compared with the recent results of Henry and Schroeder (Ref. 4) in Fig. 1. In Ref. 2, it has been shown that the magnitude of the low-temperature peak can differ markedly between different batches of 99.999% stated-purity copper obtained from the same manufacturer (American Smelting and Refining Company). The spread in low-temperature results is thus

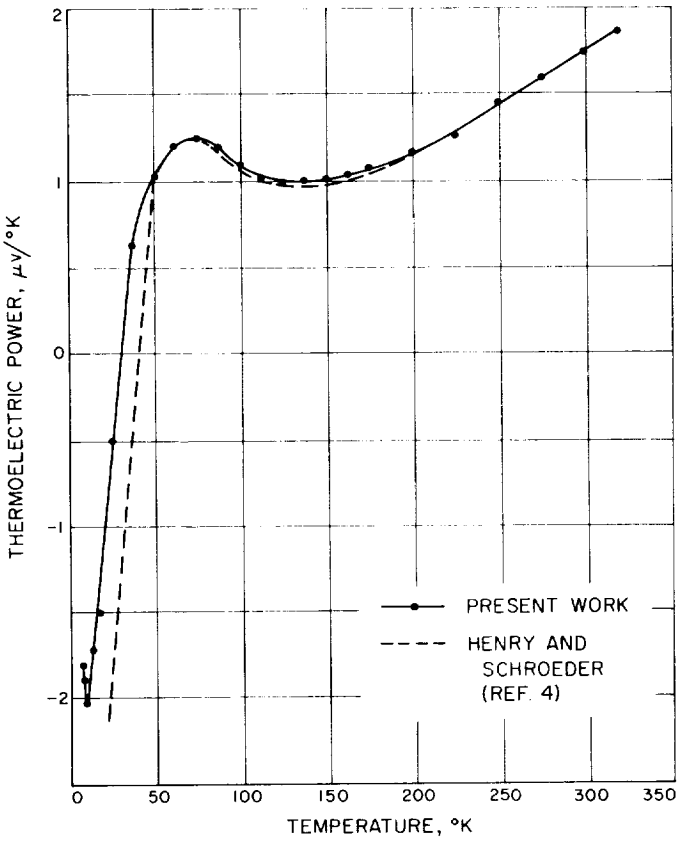


Fig. 1. Thermoelectric power of copper from 4.2 to 320°K

attributable to small differences in trace impurities. The agreement above 50°K indicates an insignificant impurity difference with respect to foreign atoms of mass significantly different than that of copper. At the higher temperatures, the results are intrinsic to the host material and agreement is excellent. The present results are thus considered adequate for use in a study of alloying effects on the electron–phonon interaction in copper. This program is now under way.

B. Crystal Growth and Perfection

P. Shlichta

1. High-Strength Filamentary Crystals

a. Use of whiskers. When metal "whiskers" were first discovered, their filamentary form and ultrahigh strength were attributed to the lack of non-axial dislocations (Ref. 5). Since screw-dislocation growth spirals were presumably present only at the tip of a whisker, the whisker would be incapable of lateral growth. In tensile or bend tests, no dislocations would be under stress; thus, the whisker would exhibit the strength predicted for perfect crystals. Subsequent observations, however, complicated this simple interpretation, so that there has been considerable debate as to whether the occasionally observed ultrahigh strength of whiskers resulted from the comparative absence of dislocations, from the perfection of the whisker surfaces, or merely from their small size.

The most notable contradiction of the simple absence-of-dislocation model was the often-cited observation that smaller specimens had higher strengths. This was reported for whiskers of copper and iron (Ref. 6), cadmium and zinc (Ref. 7), alkali halides (Refs. 8 and 9), and silicon (Ref. 10). Moreover, large imperfect crystals, when reduced to whisker-like dimensions, exhibited a similar strength increase with decreasing size. This was shown for silicon rods (Ref. 10), copper microwires (Refs. 11 and 12), and even particles ground off from bulk metal (Ref. 13).

A critical reappraisal of the above data, however, suggested that this "size effect" might be illusory. All of the materials tested—whiskers, fine wires, or Taylor wires —were usually weak, the strong ones being more the exception than the rule. Moreover, all of the techniques used for preparation and testing—selecting and detaching whiskers from a combustion boat of reduced metal halide, chemically polishing wires to order-of-magnitude reductions in cross section, and mounting specimens in a tensile test apparatus—incurred an appreciable mortality rate; i.e., the weaker specimens tended to be destroyed before they could be tested. Needless to say, the thinnest weak specimens were the ones most easily destroyed. Since data for these "unfit" specimens were generally not included in the reported test data, a "law of natural selection" obtains which is reflected in an apparent size effect over and above any real effect.

One of the goals of research on filamentary crystals at JPL is the determination of the cause of the ultrahigh strength of whiskers by the unambiguous determination of the effect of size on the strength of perfect and imperfect filamentary crystals. To do so, it is necessary to develop materials and techniques that do not involve the element of selection described above. The specimens used would have to be grown by a process which gives reproducible yields, i.e., with approximately the same size and strength distribution each time. *All* of the specimens in each batch would have to be tested. The thinning-

down process would have to be free of failures. The possibility of an oxide or phosphate skin, such as that which usually remains after chemical polishing, would have to be eliminated. The mechanical testing procedures would have to be simple and rapid enough to permit testing of a large number of specimens so as to make a statistical treatment valid.

After several years of effort, most of the above techniques have been refined to a satisfactory degree, but the problem of obtaining reproducible specimen material is still unsolved. Attempts were made to standardize the preparation of copper whiskers by the reduction of cuprous bromide in a hydrogen–argon atmosphere at elevated temperature. The purity and physical condition of the halide, the composition and flow rate of the gas, the temperature and geometry of the reaction chamber, and the surface condition and dimensions of the copper combustion boats were all carefully controlled. Nonetheless, after more than 100 experimental runs, the whiskers were still erratic in appearance, and only a small percentage had high strengths. It was therefore decided to abandon whiskers entirely and concentrate on the production of filamentary dislocation-free metal crystals by other means.

b. Use of Taylor wires. The most promising starting material for this effort was Taylor wires. These glass-clad metal filaments are made by inserting a metal wire into glass tubing, heating the composite until the metal melts and the glass softens, and then drawing the composite down to a fiber. This was originally done by hand (Ref. 14), but a conventional glass-fiber drawing machine can be modified for drawing large quantities of Taylor wires (Fig. 2).

As ordinarily produced, a 1-cm test-specimen length of Taylor wire is either polycrystalline or a highly distorted single crystal. Therefore, the first efforts were concentrated on the recrystallization of Taylor wires into single crystals containing few or no dislocations. Efforts were made to secure, as starting material, large quantities of high-purity copper and/or gold Taylor wire, from 1 to 50 μ in diameter, enclosed in a fused-silica fiber. Recrystallization was attempted by: (1) dissolving away the glass shell with hydrofluoric acid and then recrystallizing by strain-annealing the metal wire in a vacuum, (2) passing a molten zone down the Taylor wire by means of a miniaturized zone melter, and (3) sealing off the ends of a 6-in. length of Taylor wire and remelting and slowly solidifying the metal core by heating the entire specimen in a furnace with a small temperature gradient.

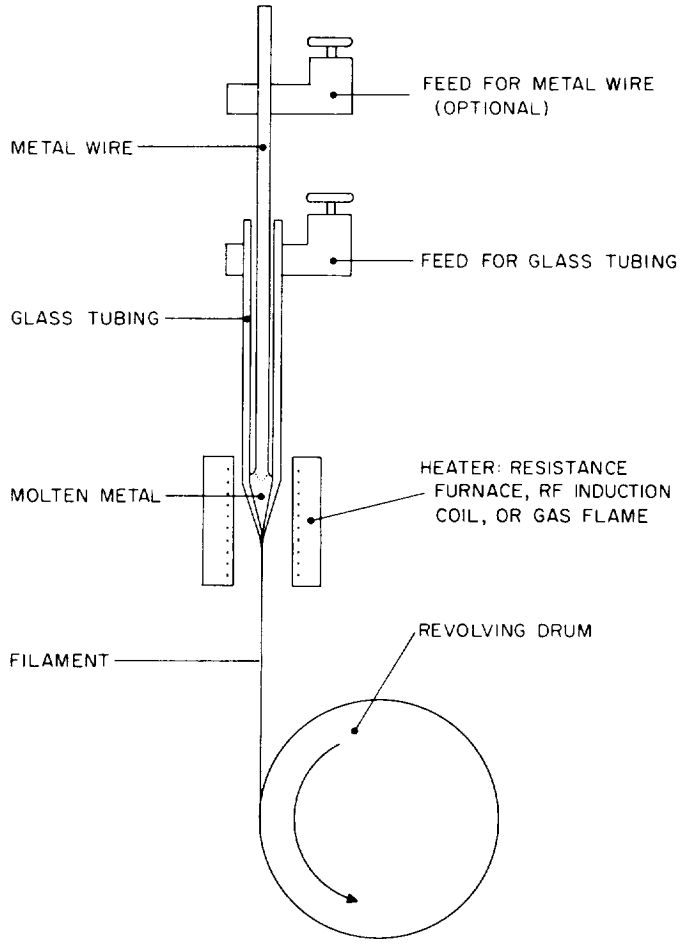


Fig. 2. Machine for drawing Taylor wires

Unfortunately, during this phase of the program, there was no reliable source of Taylor wires with the desired quality, and only occasional specimens with widely varying properties were available. Therefore, preliminary attempts at each of these techniques were unsuccessful, mainly because of defects in the starting material. Dissolving the glass shell with hydrofluoric acid often resulted in severe etching of the copper core. Moreover, numerous breaks and indentations were revealed in the stripped core; these were presumably due to bubbles of gas in the original molten metal. Attempts at strain-annealing these wires disclosed numerous experimental difficulties, such as thermal grooving at grain boundaries. Other laboratories engaged in similar research have reported similar difficulties.[1] Zone melting of Taylor wires resulted in isolated metal droplets due to evolution of gas upon melting. Moreover, there was considerable sagging and distortion of the glass shell in copper-in-silica

[1]Hughes, E. J., Materials Research Corporation (Private communication).

and tin-in-Pyrex wires, indicating that this technique may be suitable only for such combinations as tin-in-silica. There was also reason to believe, as will be shown below, that even optimum zone melting would result in crystals as distorted as those in the original wire. Heating sealed-off specimens above the melting point of the metal caused large evolutions of gas which distended the glass shell into a chain of bubbles, within each of which the metal solidified as a small sphere.

It became obvious that, if any of these techniques were to be successful, a far better quality of starting material would be needed. This would entail drawing the Taylor wire as part of the research effort. This, in turn, suggested the possibility that the Taylor-wire drawing process might ultimately be improved to the point where dislocation-free crystals could be drawn directly, so that a subsequent recrystallization process would be unnecessary. Accordingly, a study was undertaken of the causes of defects in Taylor wires.

The evolution of gas, both in the original drawing process and during subsequent remelting of the metal, was presumed to result from dissolved gas in the metal itself. Accordingly, three specimens of copper wire—commercial grade, OFHC, and ASARCO 99.999+%-pure—were heated in a vacuum. It was first noted that very little gas was evolved until the metal was actually melted; hence, vacuum annealing of the metal used for drawing Taylor wire would probably be of little value. It was also noted that, although the commercial wire evolved far more gas than did the OFHC and ASARCO wires, even the latter specimens gave off enough gas to cause considerable trouble in the resultant Taylor wire. It seemed necessary, therefore, to vacuum-melt and recast the metal used for drawing Taylor wires; preparations for this process are now being made. An alternate procedure—the use of metal wire of considerably smaller diameter than the bore of the glass tubing so as to provide clearance for the gas to escape—has been used elsewhere with some success, but this procedure is not foolproof and can lead to other difficulties, such as the necessity for separate feeds for the metal and the glass tubing.

Another defect, one which has greatly hindered the application of Taylor wires to microelectronics, is the occurrence of breaks or discontinuities in the metal core. Since this defect often occurs when the heater is at too low a temperature, it has sometimes been attributed to the stretching of the solidified metal filament until it breaks. In most cases, however, this defect can be traced to gas evolution. Observations on the necked-down portion of the composite have, on several occasions, shown gas bubbles, trapped in the metal, being stretched by the drawing process until they form breaks in the resultant filament. Thus, this defect can presumably also be eliminated by outgassing the metal before use.

X-ray diffraction was used to determine the crystallographic defects present within the metal core. Schulz topographs were taken using the scanning camera described previously for the study of ruby laser rods (*SPS 37-21*, Vol. IV, pp. 37–39). These topographs reveal distortions in the crystal lattice in the same way that the shape of the reflection of sunlight from a mirror onto a wall reveals irregularities and distortions in the mirror's surface. Topographs of Taylor wires from several different sources are shown in Fig. 3. Note that: (1) hand-drawn wires are usually polycrystalline, and (2) nearly all the machine-drawn wires are deformed single crystals with a severe, finely spaced zig-zag distortion. This might be interpreted as indicating that the metal solidifies as an undistorted single crystal which is subsequently deformed, thereby producing numerous slip bands.[2] Several plausible explanations can be advanced for the occurrence of these slip bands:

(1) They might, in some way, result from irregularities in the drawing speed and, hence, in the rate of solidification. This idea is somewhat corroborated by the polycrystalline structure of hand-drawn wires, which are presumably drawn in an even more irregular manner, and also by the erratic movement of the luminous "tail" of the filament, which is observed just below the heater during drawing. However, the exact mechanism for the formation of slip bands in this manner is somewhat obscure; moreover, it is hard to believe that the feed mechanisms of the different machines from which these specimens were drawn all had the same magnitude and frequency of irregularity. Nonetheless, to eliminate this possible source of defects, improved feed mechanisms and heater controls are being designed.

(2) Slip bands could result from stretching of the viscous glass fibers after the metal had solidified. This is somewhat confirmed by the observation of birefringence in the glass shells of most Taylor wires. This suggests that some deformation took place well below the softening point, which for most glasses is below the melting point of the metal. On this basis, one would

[2]Lockwood, P., Owens–Corning Fiberglas Research Center (Private communication).

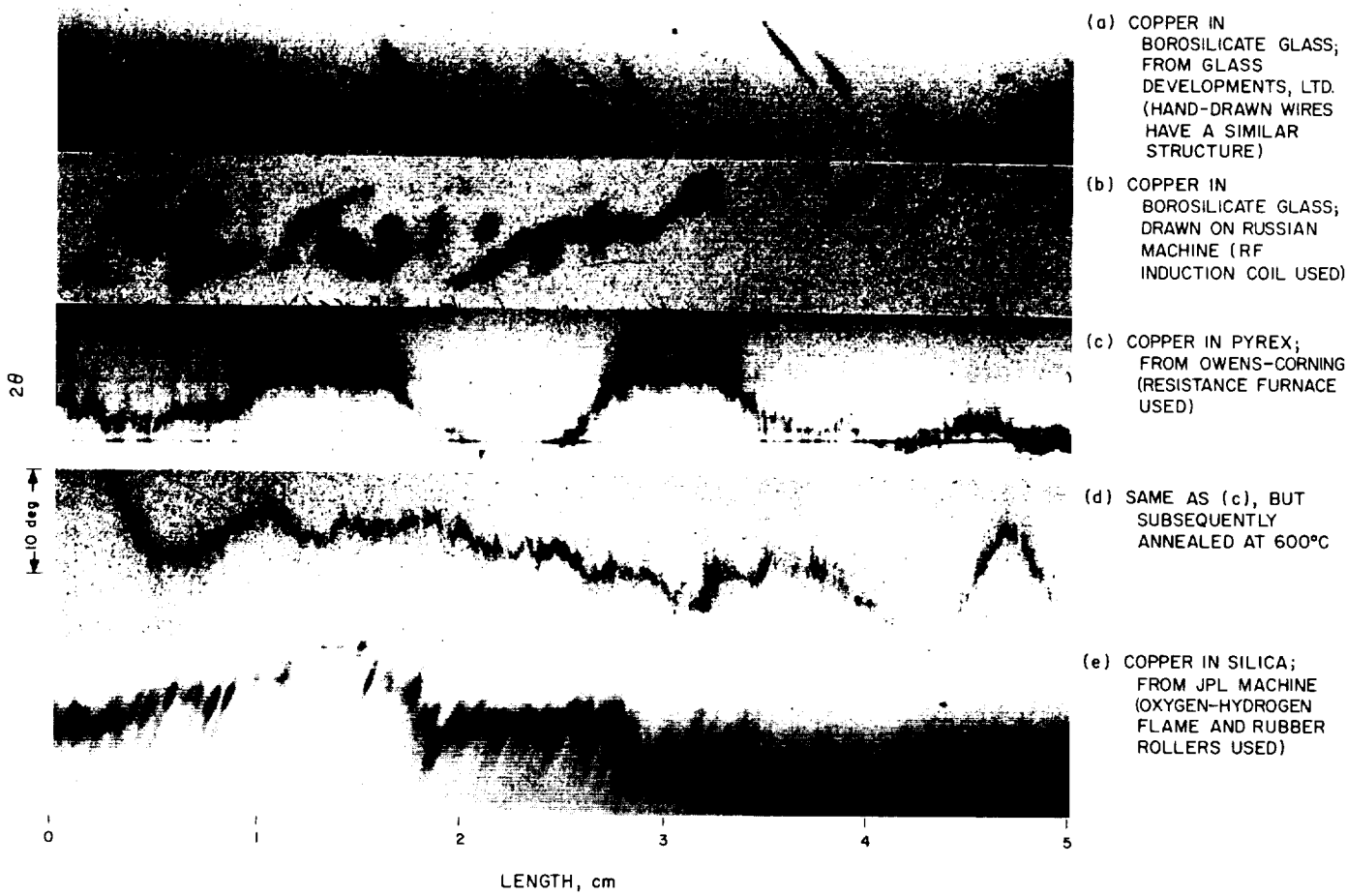


Fig. 3. X-ray diffraction topographs of Taylor wires

presume that the best glass would be the one with the highest possible softening point. Unfortunately, however, these distortions are qualitatively similar for copper-in-Pyrex and copper-in-silica wires, although the latter glass should be quite rigid at the melting point of copper. The following explanation is thus suggested:

(3) The metal core could be deformed by stretching as a result of differential thermal contraction as the composite filament cools to room temperature. The glasses used thus far have coefficients of thermal expansion ranging from 4 to 25% of those of copper or gold; therefore, the resultant strain (1 to 3%) should be sufficient to cause considerable deformation of the metal. The obvious remedy for this source of slip bands would be to use a glass having a coefficient of thermal expansion matching that of the metal. Unfortunately, however, glasses with high coefficients of thermal expansion tend to have lower softening points, thereby increasing the likelihood of the deformation mechanism described in (2) above. Indeed, the assumption that both mechanisms contribute to the deformation of the metal may satisfactorily account for the observation that copper-in-silica and copper-in-Pyrex wires have approximately the same type and amount of deformation.

At present, attempts are being made to solve the two problems simultaneously by choosing the glass which has the best combination of softening point and coefficient of thermal expansion. If this fails, preliminary trials suggest the feasibility of composite glass tubing, consisting of an outer shell of high-softening-point glass and an inner liner of low-melting-point glass having a coefficient of thermal expansion matching that of the metal core; this would be, in effect, a radial graded seal. Still another approach would be to insert an air blast immediately below the heater in order to chill (and thereby harden) the glass shell as quickly as possible. Finally, if all other

techniques fail, it may be necessary to resort to low-melting-point metals, such as indium, where the temperature change from melting point to room temperature is small enough to make thermal strains negligible.

(4) Slip bands might be caused by strains resulting from the high thermal gradients of the rapidly cooling wire. This mechanism is not, at present, thought to be important inasmuch as the extremely small diameter of the wires should ensure the absence of curved thermal gradients and, therefore, of thermal strains. If necessary, however, this effect might be remedied by inserting a long low-gradient annealing furnace immediately below the heater.

(5) The substructure might not result from slip bands, but rather from lineage or mosaic crystallites caused by the extremely high linear rate of solidification. This is somewhat substantiated by the observation that the substructure remains qualitatively unchanged by prolonged annealing at 600°C (Fig. 3d); one would expect slip bands to recrystallize under these conditions. Certainly, the rates of solidification encountered in the present drawing techniques (e.g., 3×10^3 to 15×10^3 cm/min) would cause severe mosaic structure in macroscopic crystals. Moreover, Taylor (Ref. 14), who drew his wires at vastly slower speeds than are now employed, remarked at the unusual strength and elasticity of his wires. This strongly suggests the desirability of studying the effect of drawing speed on the perfection of Taylor wires. Such a program is planned for the near future. It is hoped that rates as low as 1 cm/min can ultimately be achieved.

(6) Some deformation might result from the mechanics of drawing the filament by winding it around a revolving drum. Studies of similarly designed glass-fiber machines indicate the possibility of an oscillatory "whipping" of the filament about the axis of drawing. To check this possibility, a drawing mechanism consisting of two revolving rubber rollers was constructed. This device draws the filament without causing any such whipping; nonetheless, the substructure of wires drawn in this manner was the same as that observed in filaments drawn by a revolving drum. Thus, this effect is apparently negligible.

(7) Another potential source of trouble lies in the localized areas of oxidation frequently encountered in copper Taylor wires. These spots may provide nuclei for parasitic crystallites; alternatively, since they cause adhesion of the metal to the glass, they may provide points of stress concentration and consequent initiation of plastic deformation. Accordingly, it seems desirable to eliminate all sources of adhesion of metal to glass. This can be accomplished by carefully cleaning and drying the glass and metal and by either excluding air from the copper composites or using gold.

In view of the gross deformations observed in Taylor wires thus far, it may seem premature to consider the sources of individual dislocations. These individual dislocations are, however, fairly easy to prevent. The various causes and remedies for these are as follows:

(a) Dislocations can be caused by very slight thermal or mechanical stresses; the prevention of these conditions has already been discussed.

(b) Some authors believe that dislocations can result from the coagulation and collapse of vacancy clusters. This can be avoided by allowing the excess vacancies time to diffuse out to the surface of the metal, i.e., by drawing Taylor wires of very small diameters (1 to 10 μ) at very slow speeds ($\simeq$ 1 to 10 cm/min).

(c) Dislocations can also be caused by strains resulting from the clustering or precipitation of impurities. This suggests the advisability of using the purest possible metal as starting material.

(d) Another cause of dislocations, as pointed out by Jackson (Ref. 15), is the presence of solid particles in the metal during crystallization; these can cause dislocations either by disturbing the arrangement of the atoms as the melt solidifies or by the thermal strains set up around the incorporated dirt particles as the crystal cools. This indicates the necessity not only of using ultrapure metal, but also of having a scrupulously clean handling procedure.

(e) Dislocations parallel to the filament axis might, in theory, propagate indefinitely. It seems unlikely that these would not quickly wander out to the surface of the metal and there terminate, but, even if they remained, such axial dislocations would not impair the strength of the resultant whiskers.

Thus, we see that, although the possible causes of defects in Taylor wires are numerous indeed, there appears to be a practicable means of preventing each of them. Therefore, the goal of the current effort in this field—the production, in quantity, of nearly perfect single crystals by the Taylor-wire process—appears to be an attainable, albeit difficult, one.

C. Ceramics

M. H. Leipold and T. Nielsen

1. Pure Carbide Ceramic Research

It was reported in *SPS 37-26*, Vol. IV, p. 69, that TaC powder of satisfactory purity was available for initial investigation of mechanical properties. The vacuum hot press, described in the above *SPS*, pp. 65–69, has been employed to densify this powder into useful specimen blanks. A series of pressings has been conducted in the temperature range from 2000 to 2500°C on a special low-O_2, high-purity grade of TaC (obtained from Cerac, Inc., Butler, Wisconsin). The particle size of this material was approximately 5 μ. Densities of approximately 11 g/cm^3 (theoretical X-ray density = 14.53 g/cm^3) were obtained at 2200°C and pressures of approximately 6000 psi. When attempts were made to increase the density by increasing the temperature, the graphite punches began to deform, increasing the friction between the punches and the sides of the die case and reducing the effective pressure on the material to be compacted. Consequently, at temperatures as high as 2500°C and applied pressures of 7000 psi, densities of only 12.1 to 13 g/cm^3 were obtained. Attempts to reduce this friction by contouring the punches were moderately successful. Other grades of graphite are being obtained which should exhibit greater resistance to deformation at these temperatures and loads.

Since densification of powders during hot pressing may be enhanced by reducing the particle size of the starting material, the TaC powder was ground by means of a single pass through a fluid energy mill (Jet Trost Model TX). This powder was then pressed for 1 hr at 2200°C, 6500 psi, under a vacuum of 10^{-4} torr. The density obtained was 14.3 g/cm^3. The microstructure is shown in Fig. 4(a). The etch pits appeared during etching to define the grain boundaries. This microstructure is typical of the central portion of the specimen, while the areas near the surface showed grains approximately $\frac{1}{10}$ this size along with some free carbon. It thus appears that considerable grain growth occurred in the center of the specimen, while grain growth at the edges was inhibited by the presence of free carbon from the die case.

The effect of temperature and pressure on the density and grain size of compacts produced from this reground TaC will be determined before routine production of test specimens begins. Any effect of this grinding on the purity of the material and the particle size of the reground powder will be determined.

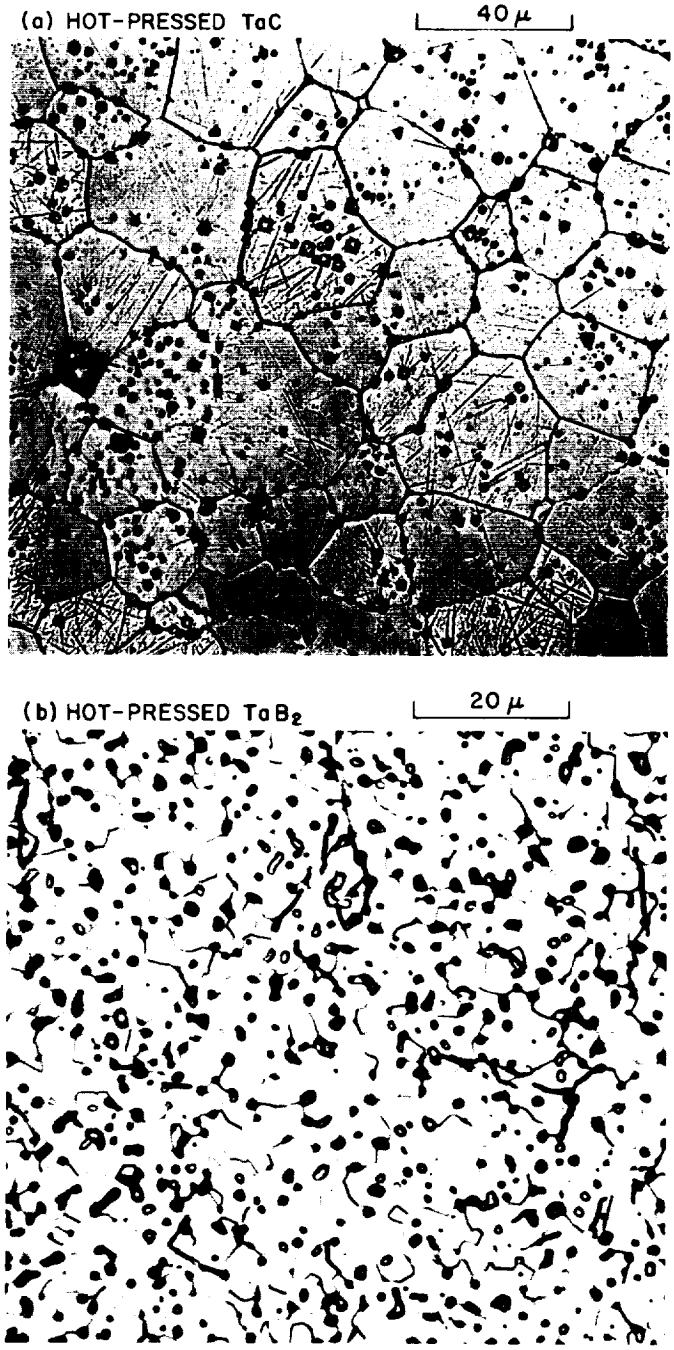


Fig. 4. Microstructures of hot-pressed TaC and TaB_2 (mechanically polished and etched with 20% HNO_3, 20% HF, 60% lactic acid)

The capability to produce dense specimens of other refractory materials was demonstrated with the fabrication of two TaB_2 compacts with a density of 11.4 g/cm^3 (theoretical density = 12.3 g/cm^3) at 2100°C, 6000 psi, in a vacuum of 10^{-4} torr for 1 hr. The microstructure is

shown in Fig. 4(b). Note what appears to be a second phase (long needle-like crystals, probably free boron). In this case, the theoretical density noted would not be correct with this composition.

2. *Pure Oxide Ceramic Research*

A chemical technique for the production of high-purity MgO has been evaluated (Ref. 16). Eleven batches of material have been produced, and the process has been reduced to a routine procedure. Further automation of the process is contemplated in order to reduce the amount of handling required.

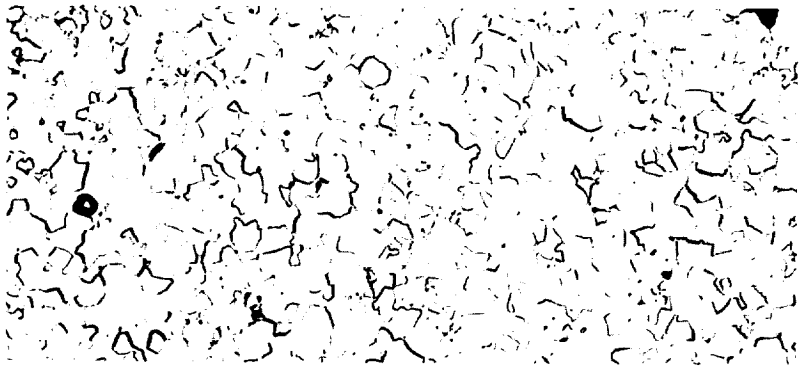
The material obtained from the chemical precipitation process is a basic magnesium carbonate known by the mineral name hydromagnesite ($4MgO \cdot 3CO_2 \cdot 4H_2O$). Thermogravimetric analysis of this material during calcining in a vacuum of 10^{-1} to 10^{-6} torr has indicated a two-step process of decomposition. The first step is the loss of the combined H_2O in the hydromagnesite structure. This loss begins when the vacuum is applied to the material at room temperature and occurs very rapidly at a temperature of the order of 100°C. The X-ray structure indicates that the material remaining after this loss is poorly crystalline hydromagnesite. The second step in the decomposition process is the loss of the combined CO_2 to periclase (MgO). This second loss begins at approximately 300°C and occurs quite rapidly at 350°C.

The particle size of the MgO powder produced by this decomposition was determined by a standard X-ray diffraction technique (Ref. 17). An average particle diameter of 107 A was determined from material calcined at 400°C, ranging up to 185 A at 750°C. At 850°C, the rate of particle growth increases rapidly to produce an average particle size of 1100 A. The bulk of the JPL material has been calcined at 650°C. This temperature seems to offer a satisfactory compromise of speed and completeness of decomposition.

The determination of hot-pressing parameters in refractory dies for this high-purity MgO and for commercial types of MgO has continued. The problem of die failure has not been entirely solved. Commercial Al_2O_3 has proved generally satisfactory for punches, but not for die cases. The use of molybdenum support rings for die cases (*SPS 37-26*, Vol. IV, pp. 65–69) is limited to vacuum or inert-gas operation and does not eliminate failure. Molybdenum die cases are satisfactory under the same atmosphere limitations. They must, however, be used with Al_2O_3 punches to eliminate galling (differential thermal expansion must be provided for in sizing of the parts). Also, the soft inner surface of the molybdenum die case has shown a tendency to be scratched and roughened during pressing, thus limiting the usefulness of the case. The possibility of plating the molybdenum with molybdenum disilicide for both atmosphere protection and surface hardness is being investigated. Commercial SiC dies (Carborundum Co., Grade KT) have also been used for pressings with satisfactory results. However, their failure seems comparable to that of JPL Al_2O_3 dies. The JPL Al_2O_3 dies produced by hot pressing (*SPS 37-26*, Vol. IV, pp. 65–69) are still the most satisfactory. The life of these dies has been extended by eliminating hot ejection (Ref. 18) of the pressed part and thus reducing thermal shock.

The results determined from examination of the hot-pressed MgO have begun to demonstrate the differences in behavior at high temperature of various levels of purity in starting material. Fig. 5 shows the differences in the structures of as-pressed and reheated materials as exhibited by a good commercial MgO (Fisher Scientific Co., Grade M-300) and the high-purity material produced at JPL (described briefly in the above *SPS*). Both specimens were pressed in air at 1100°C, 15,000 psi, for 0.5 hr in Al_2O_3 dies, and were reheated in air to 2200°C. The grain size of the JPL MgO as-pressed specimen is approximately two orders-of-magnitude greater, suggesting that grain growth in the commercial material has been inhibited by impurities. The grain size for the JPL MgO after reheating to 2200°C, shown in Fig. 5, is not entirely representative. Several grains were visible which were as large as the entire figure, while some were smaller. The smaller grains shown appeared to have been limited in growth by the presence of macrocracks and the edges of the specimen. Note that the JPL MgO in general shows clean, regularly curved, grain boundaries with very little evidence of second phase, while the commercial material shows a second phase, very irregular grain boundaries, and considerable reheat porosity. This reheat porosity, first discovered during thermal expansion measurements and typical of refractory oxides (Ref. 19), is believed to be the result of volatilization of small quantities of impurities.

Attempts to analyze the pure JPL MgO for the presence of impurities have necessitated the development of improved analytical techniques. Some success was obtained by using spark-source mass spectrography; however, development of a new specimen preparation procedure was necessary. MgO is not an electrical conductor; therefore, the conventional technique of sparking



AS-PRESSED SPECIMENS

250 μ

JPL MgO: MECHANICALLY POLISHED, ETCHED WITH 50% HNO_3 AND OPTICAL MICROGRAPH

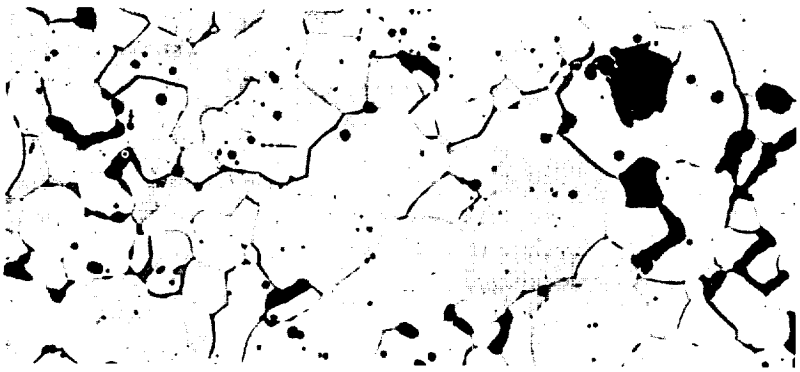
1 μ

COMMERCIAL MgO: MECHANICALLY POLISHED, ETCHED WITH STEAM AND ELECTRON MICROGRAPH BY REPLICA TECHNIQUE

SPECIMENS REHEATED TO 2200°C

250 μ

JPL MgO: MECHANICALLY POLISHED, ETCHED WITH 50% HNO_3 AND OPTICAL MICROGRAPH



250 μ

COMMERCIAL MgO: MECHANICALLY POLISHED, ETCHED WITH 50% HNO_3 AND OPTICAL MICROGRAPH

Fig. 5. Microstructures of hot-pressed MgO

between two chips of the sample to produce volatile material for analysis in the mass spectrograph is not applicable. Previous techniques which have necessitated powdering of a dense piece and mixing with a metal to form a conducting matrix are too likely to introduce impurities to be used with these highly pure materials. The technique developed may be applied to dense specimens or compacted powders which have been densified by hydrostatic pressing. These chips of dense material or compacts are then sparked against a conducting probe of a suitable material. (High-purity silicon has proved very effective.) Analysis of the sample may be made for all elements, except that of the probe. This technique is now being used to study the purity of these various MgO specimens.

3. Thermal Expansion

The measurement of thermal expansion of pure well-characterized oxides (*SPS 37-23*, Vol. IV, pp. 52, 53; *SPS 37-24*, Vol. IV, pp. 48–51; Ref. 18) has continued with a study of NiO. The starting material was analytical reagent-grade nickel sulphate which was calcined 20 hr at 900°C in air. The resulting submicron NiO powder may be hot-pressed, as described in Ref. 20, into dense polycrystalline bodies. However, attempts to hot-press a 0.75-in. D × 3-in.-long blank required for a thermal expansion specimen were unsuccessful. The blank produced was 90+% of theoretical density at the ends, but only 80% in the middle. The failure to attain density is believed to be a result of poor pressure uniformity, caused by friction between the walls of the die and the long blank. By limiting the blank length to 2 in., uniform pieces were obtained which were greater than 95% of theoretical density. These shorter pieces were then welded together by means of an additional hot-pressing, employing a small quantity of NiO powder between the two parts. The joint could be detected macroscopically by variations in color, but it was not detected microscopically.

The pressing parameters used for these three hot-pressing operations were 10,000 psi at 1000°C (inside die temperature) for 90 min in O_2. The final density of the blank was 6.44 g/cm^3. Since all of our previous results have shown that grain size does not affect thermal expansion (Ref. 19), the specimen was thermally stabilized with respect to grain growth and density by heating to 1800°C for 1 hr in a 95 wt % O_2–5 wt % N_2 atmosphere. After this heat treatment, the weld joint could not be detected microscopically or macroscopically (Fig. 6).

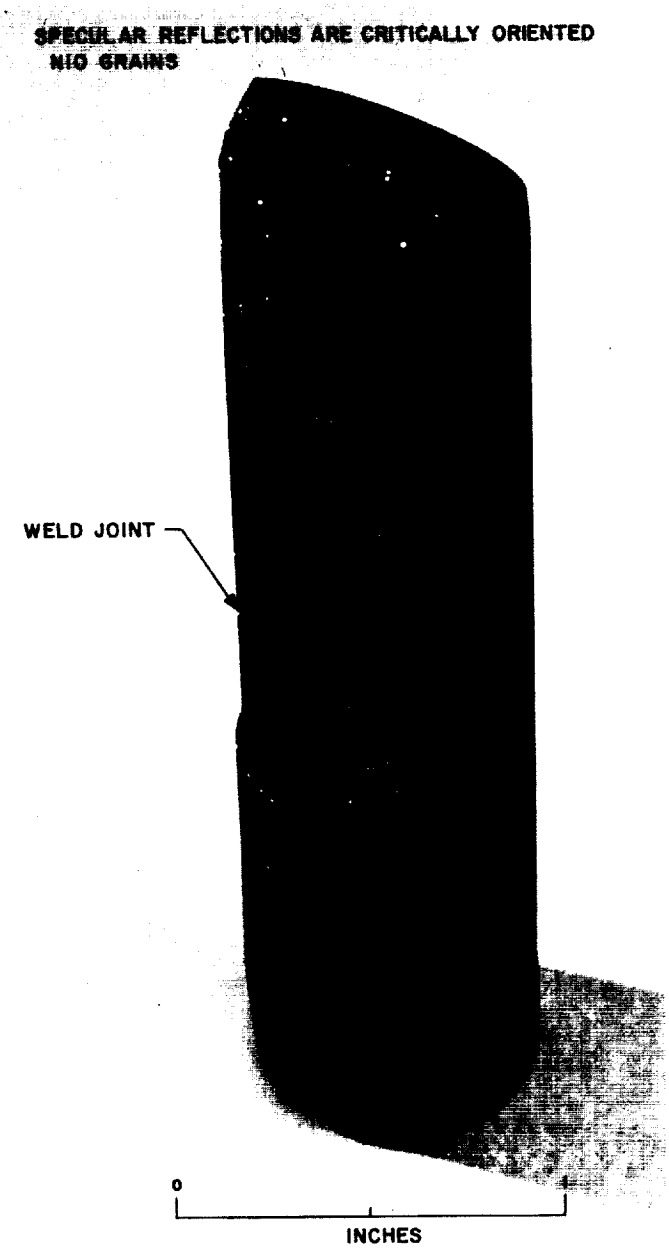


Fig. 6. Hot-pressed and hot-press-welded NiO specimen showing weld area

Two thermal expansion tests were conducted on this specimen to 1800 and 1930°C, respectively; the composite data are shown in Fig. 7. During the second test, surface melting occurred on the specimen surface at a local temperature of 1935°C, causing fusion of the specimen to the support. The specimen characteristics, as determined by techniques previously described (Ref. 19), are listed in Table 1.

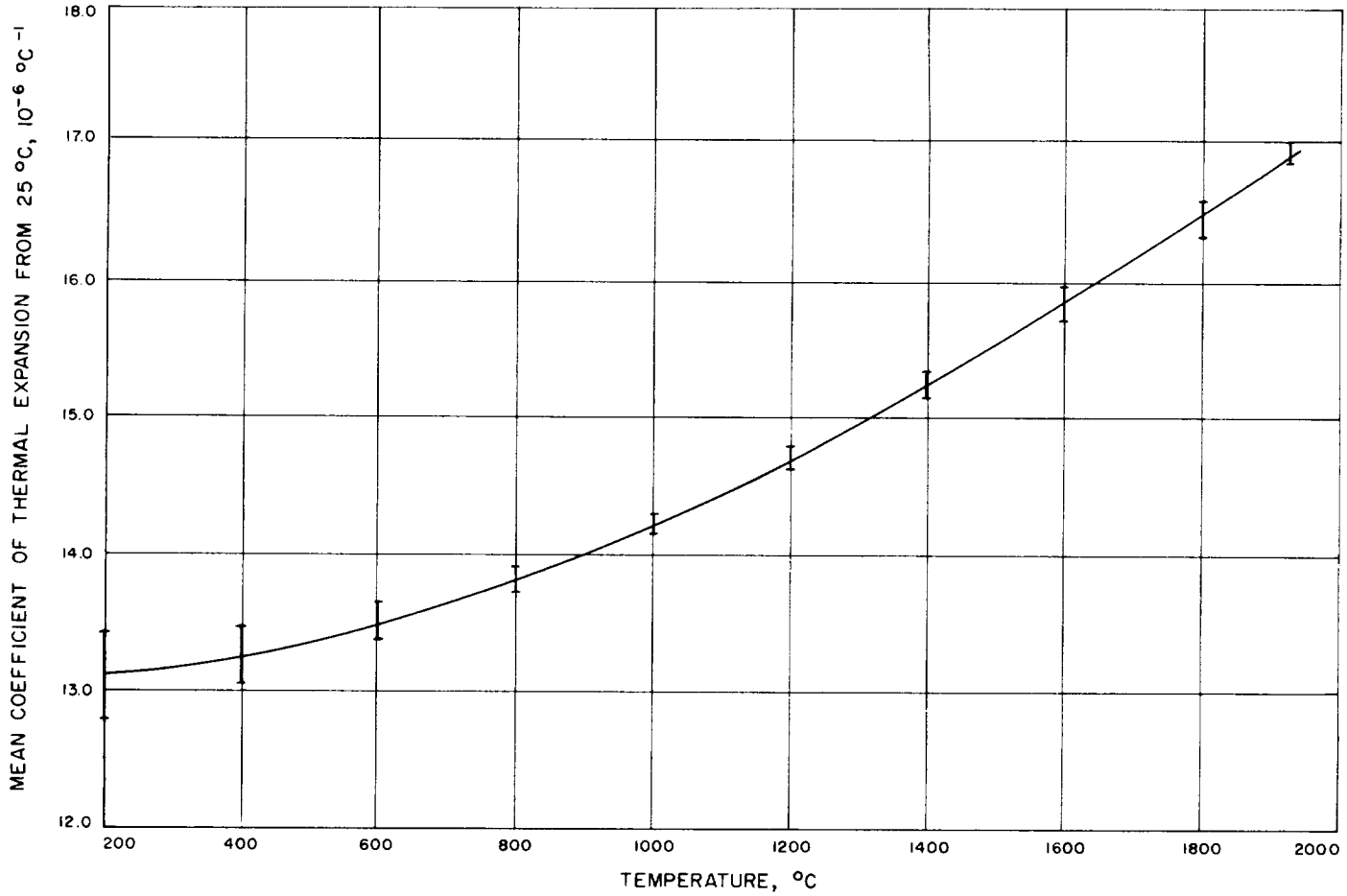


Fig. 7. Mean coefficient of thermal expansion vs temperature for NiO

An attempt was made to determine the thermal expansion of CeO_2 during the same period. The starting material, reported 99.9%-pure by the Research Chemical Corporation, had a particle size of approximately 5 μ. This material, when hot-pressed for 45 min at 10,000 psi and 1150°C (inside die temperature), was only 75% dense. A fluid energy mill (Jet Trost Model TX) was used to reduce the particle size to less than 1 μ. This finer CeO_2, when pressed under the same parameters, resulted in a specimen having a density of 96.5%. As with the NiO, the technique of producing two short specimens and subsequently welding them together was employed. Fig. 8 shows the weld joint. It can be seen that the CeO_2 did not weld as satisfactorily as did the NiO.

Table 1. Specimen characteristics

Maximum test temperature, °C	Grain size, μ	Lattice parameter, A	Density, g/cm^3	Test atmosphere
As fabricated	≅1.5	4.1782	6.44	—
1800, heat treat	140	4.1779	6.57	95 wt % O_2–5 wt % N_2
1800	145	4.1781	6.56	95 wt % O_2–5 wt % N_2
1930	185	4.1752	—[a]	95 wt % O_2–5 wt % N_2

[a] Specimen reacted with setter.

This CeO_2 specimen fractured during its initial heat treatment, apparently due to a sudden increase in temperature. Data obtained up to 1000°C in the dilatometer (Ref. 19) are shown in Fig. 9. Because of the limited interest in CeO_2, additional fabrication attempts are not contemplated.

Fig. 8. Hot-pressed and hot-press-welded CeO_2 specimen showing weld area

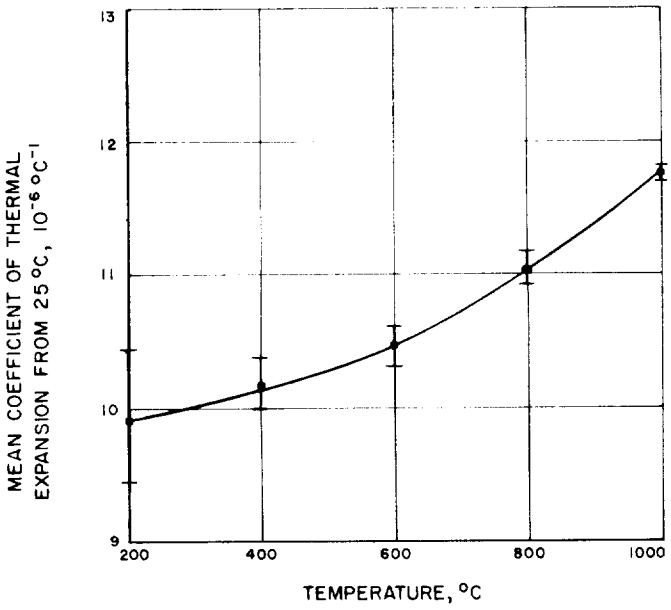


Fig. 9. Mean coefficient of thermal expansion vs temperature for CeO_2

D. Parachute Sterilization and Vacuum Compatibility

R. G. Nagler

Interest in deceleration devices for Mars entry led to a questioning of the sterilization and long-vacuum-exposure compatibility of available parachute materials. A contract was let with Cook Electric Company, Tech-Center Division, Morton Grove, Illinois, to determine the resistance of available parachute materials to thermal and chemical sterilization followed by long vacuum exposure (Ref. 21).

1. Test Plan

Specimens of silk, Nylon 66, Dacron, and Nomex, along with representative pyrotechnic materials, were exposed to the test sequence shown in Fig. 10. Seventy samples were removed after each test to measure strength degradation due to exposure to the test conditions. These samples were composed of ribbon, fabric, and cord in flat, folded and compacted, and twisted and compacted configurations. The ribbon was both sewed and unsewed. Table 2 shows the strength measurements made on each configuration: tensile tests for the ribbon and cord, and burst and permeability tests for the fabric. Extra specimens of ribbon, cord, and fabric from each material were exposed to the entire environmental sequence. These were then subjected, while still in the vacuum, to a sudden applied load similar to that expected during Mars entry. The environmental test conditions are shown in Table 3.

Table 2. Parachute material test matrix

Weave form	Material property test	Sample configuration			Samples subjected to sudden applied load in vacuum
		Flat	Folded and compacted	Twisted and compacted	
Ribbon	Tensile	5 Unsewed	5 Unsewed	5 Unsewed	3 Unsewed
		5 Sewed	5 Sewed	5 Sewed	3 Sewed
Fabric	Burst	5 Unsewed	5 Unsewed	5 Unsewed	1 Unsewed
	Permeability	5 Unsewed	5 Unsewed	—	—
Cord	Tensile	5 Unsewed	5 Unsewed	5 Unsewed	3 Unsewed

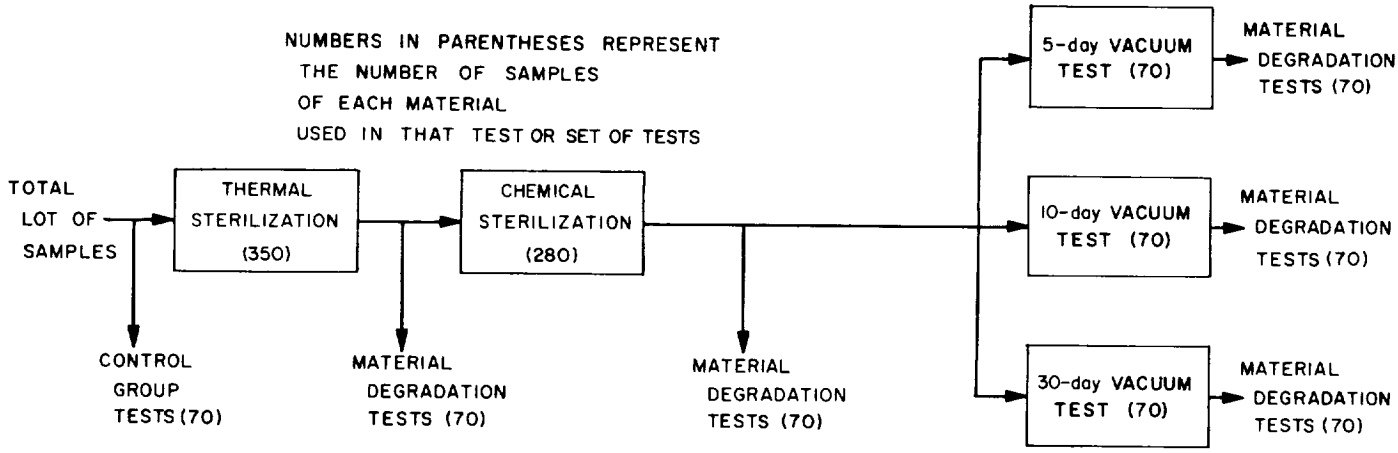


Fig. 10. Flow diagram of environmental and material degradation test sequence

Table 3. Environmental test conditions

Test	Conditions
Thermal sterilization	Three cycles at 294 ± 3°F for 36 hr/cycle in a dry nitrogen atmosphere, with return to approximately 80°F between cycles, per JPL Specification X50-30275-TST-A
Chemical sterilization	Exposure to mixture of 12% ethylene oxide–88% freon-12 for two temperature cycles of 75 ± 5°F and 104 ± 5°F for 24 hr each, with an ethylene oxide concentration of 550 ± 50 mg/liter and a relative humidity of 40 to 50%
5-day Vacuum exposure	Pressure, 1.0 to 3.9 × 10^{-6} torr Temperature, ~160°F
10-day Vacuum exposure	Pressure, 0.8 to 4.0 × 10^{-6} torr Temperature, ~160°F
30-day Vacuum exposure	Pressure, 0.5 to 4.0 × 10^{-6} torr Temperature, ~160°F

Table 4. Effects of sterilization and vacuum exposure on strength of parachute cloths

Cloth	Preliminary test	Sterilization		5-, 10-, and 30-day vacuum exposure	Sudden-shock-opening test
		Thermal	Chemical		
Silk	Failed	—	—	—	—
Nylon 66	Border	Fabric and cord lost 80%, ribbon lost 20%	—	—	Failed
Dacron	OK	½ lost 20%, ½ varied ± 10%	Gained ½ of loss	No effect	OK
Nomex	OK	Varied ± 5 to 10%	No further loss	No effect	OK

2. Results

a. Cloth materials. The results obtained for the various parachute cloth materials are summarized in Table 4. Silk failed "catastrophically" in preliminary thermal sterilization tests. Nylon 66 fabric and cord lost 80% of their strength after thermal sterilization, whereas the ribbon lost only 20%. The materials were identical, except that the processing technique used for the ribbon material produces a more oriented or linearized fiber. The difference in behavior can be attributed to the fact that the test temperature is near the knee in the strength–temperature curve (Fig. 11). Therefore, a 5°F variation in temperature may cause a similar 80% drop in strength in the ribbon. The ribbon also failed in a sudden-shock-opening test in vacuum at a load of 50% of the design strength. This seems to confirm reports of water-loss embrittlement reported elsewhere.

Dacron and Nomex both appear to be acceptable materials, with material variations being less in Nomex. Fig. 12 shows the averages and maximum variations of the five-sample groups used in each test situation (see Table 2). Dacron seems to have a somewhat consistent gain of strength (and weight) due to chemical sterilization. The gradual loss of this added strength (and weight) during vacuum tests, back to the original value prior to

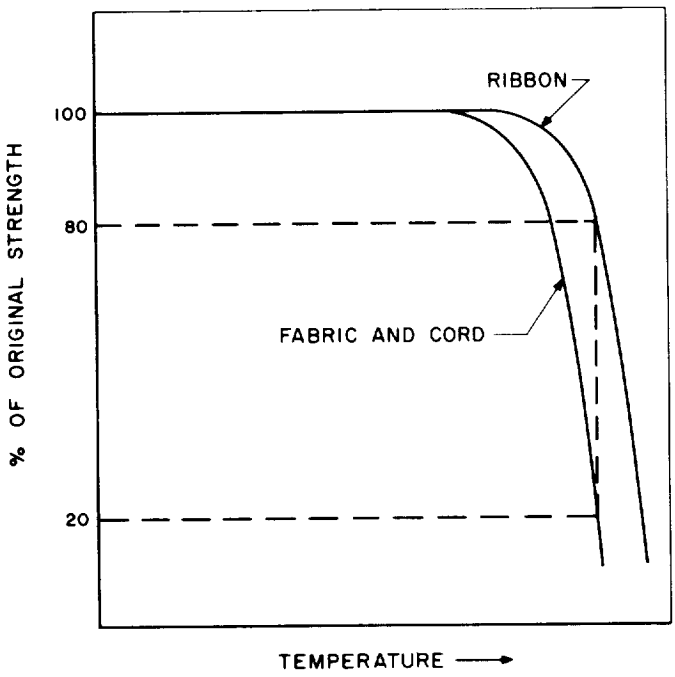


Fig. 11. Effect of temperature on the strength of parachute cloths

chemical sterilization, may indicate absorbed water effects from the 40 to 50% relative humidity of the chemical sterilization.

Folded or twisted and compacted specimens of all of the materials showed no measurable variation from the flat specimens. Sewed specimens, on the other hand, lost 10% of their strength, probably due to fiber damage. The permeability of Dacron increased slightly (maximum 8%), whereas that of Nomex decreased slightly (maximum 8%). Neither Nomex nor Dacron failed when subjected to sudden shock loads in vacuum at 55% of listed tensile strength. (Normal design strength is 50% of listed strength.)

b. Pyrotechnic devices. Ten suppliers of pyrotechnic devices and materials were contacted: five supplied pressure-generating devices, and two supplied reefing cutters. All devices operated after exposure to all standard environments. No problems were expected with the pressure-generating devices, but it was noticed that thermal sterilization caused some sensitizing, so that they reached higher peak pressures faster. The seals on some of the pressure-generating devices were purposely punctured before vacuum exposure. Some of these samples failed, indicating that vacuum exposure of sealed devices for much longer periods than those in this test (i.e., 7 to 9 mo of travel to Mars) may be detrimental. The success of the reefing cutters was a surprise, but the high-energy primers used may not be suitable for other constraints in the parachute and vehicle design for Mars entry.

3. Conclusions

In conclusion, both Nomex and Dacron appear to be suitable for sterilizable parachute or deceleration-device cloths. All that remains to be done is to perhaps better define their strength limits to allow lower weight designs than the safety factor of 2 used in normal parachute designs allows. Pressure-generating devices are probably satisfactory, but work is needed on the interrelations between the time of vacuum exposure and the quality of various seals. Reefing cutters are perhaps less of a problem than was expected before this study, but much work is necessary before a reliable device suitable for a Mars entry retardation system will be ready.

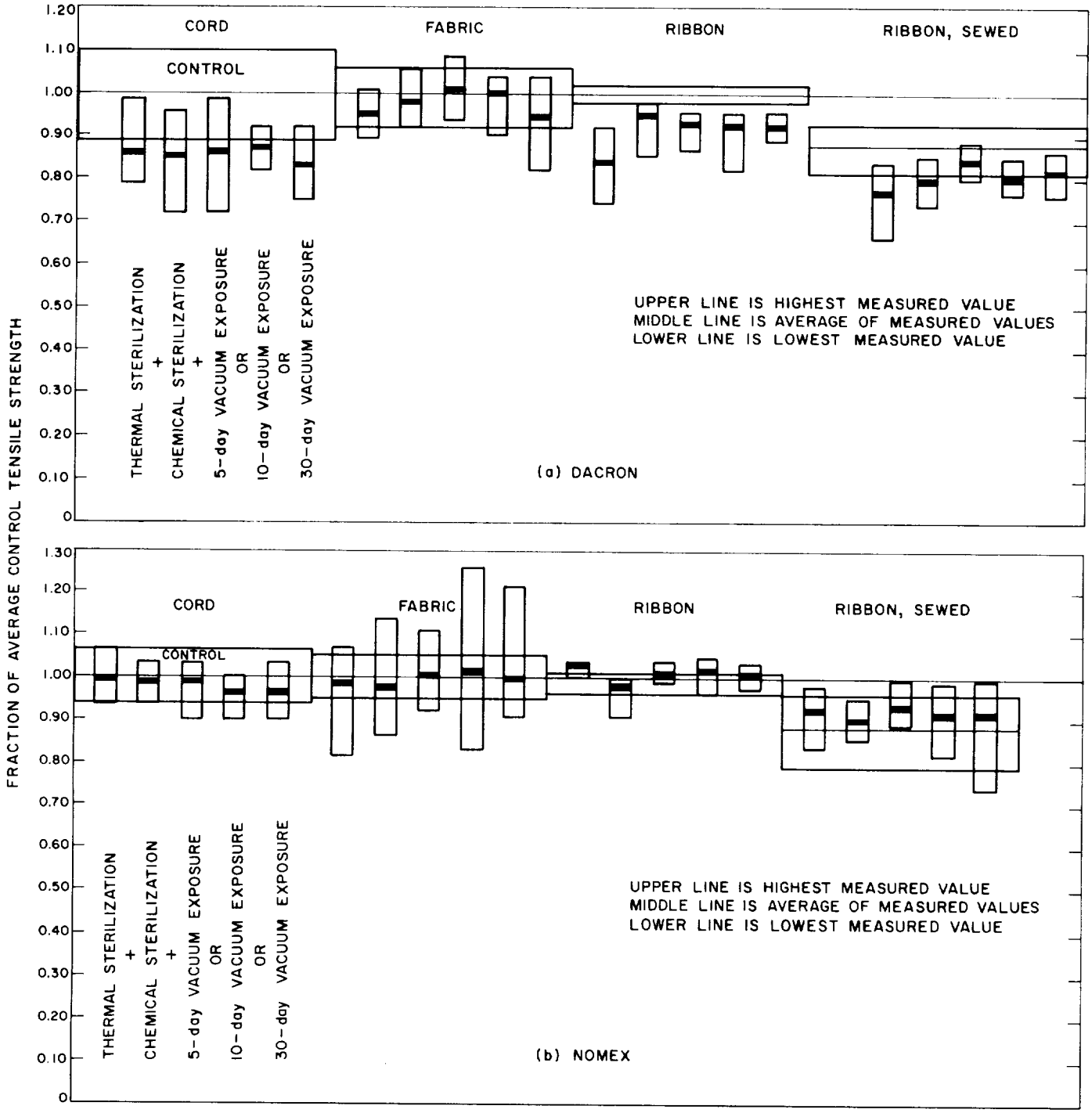


Fig. 12. Tensile-strength losses in Dacron and Nomex due to thermal and chemical sterilization and vacuum exposure

References

1. Ziman, J. M., *Electrons and Phonons*, p. 396, Oxford University Press, New York, 1960.

2. Gold, A. V., MacDonald, D. K. C., Pearson, W. B., and Templeton, I. M., "The Thermoelectric Power of Pure Copper," *Philosophical Magazine*, Vol. 5, p. 765, 1960.

3. Van Zandt, L. L., *Some Effects of Static Spin Density Waves on Electrical Transport*, PhD Thesis, Department of Physics, Harvard University, Cambridge, Massachusetts, February 1964.

4. Henry, W. G., and Schroeder, P. S., "The Low Temperature Resistivities and Thermopowers of α-Phase Copper–Zinc Alloys," *Canadian Journal of Physics*, Vol. 41, p. 1076, 1963.

5. Herring, C., and Galt, J. K., "Elastic and Plastic Properties of Very Small Metal Specimens," *Physical Review*, Vol. 85, pp. 1060, 1061, 1952.

6. Brenner, S. S., "Tensile Strength of Whiskers," *Journal of Applied Physics*, Vol. 27, pp. 1484–1491, 1956.

7. Coleman, R. V., Price, P. B., and Cabrera, W., "Zinc and Cadmium Whiskers," *Journal of Applied Physics*, Vol. 28, pp. 1360–1362, 1957.

8. Guylai, Z., "Festigkeits-und Plastizitätseigenschaften von NaCl-Nadelkristallen," *Zeitschrift für Physik*, Vol. 138, pp. 317–321, 1954.

9. Gordon, J. E., "Current Research on the Strength of Whiskers and Thin Films at Tube Investments Laboratory," *Growth and Perfection of Crystals*, pp. 219–229, R. H. Doremus, B. W. Roberts, and D. Turnbull, Editors, Wiley & Sons, Inc., New York, 1958.

10. Pearson, G. L., Read, W. T., and Feldman, W. L., "Deformation and Fracture of Small Silicon Crystals," *Acta Metallurgica*, Vol. 5, pp. 181–191, 1957.

11. Shlichta, P. J., "Comparison of Copper Whiskers with Ultrafine Wires," *Growth and Perfection of Crystals*, pp. 214–218, R. H. Doremus, B. W. Roberts, and D. Turnbull, Editors, Wiley & Sons, Inc., New York, 1958.

12. Barton, P. W., Hughes, E. J., and Johnson, A. A., "Yielding in Copper Single Crystals Having Diameters in the Range to μ to 150μ," *Journal of the Physical Society of Japan*, Vol. 19, pp. 407, 408, 1964.

13. Backer, W. R., Marshall, E. R., and Shaw, M. C., "The Size Effect in Metal Cutting," *Transactions of the AIME*, Vol. 74, pp. 61–72, 1952.

14. Taylor, G. F., "A Method of Drawing Metallic Filaments and a Discussion of Their Properties and Uses," *Physical Review*, Vol. 23, pp. 655–660, 1924.

15. Jackson, K. A., "On the Origin of Dislocations," *Philosophical Magazine*, Vol. 7, pp. 1615, 1616, 1962.

16. Quincey, R. B., *Preparation of High Purity Magnesium Oxide*, Report ORNL-TM-302, Oak Ridge National Laboratory, Tennessee.

17. Klug, H. P., and Alexander, L. E., *X-Ray Diffraction Procedures*, p. 491, John Wiley & Sons, Inc., New York, 1959.

References (Cont'd)

18. Spriggs, R. M., Brissette, L. A., Rosetti, M., and Vasilos, T., "Hot Pressing Ceramics in Alumina Dies," *American Ceramic Society Bulletin,* Vol. 42, No. 9, pp. 477–479, 1963.

19. Nielsen, T. H., and Leipold, M. H., "Thermal Expansion in Air of Ceramic Oxides to 2200°C," *Journal of the American Ceramic Society,* Vol. 43, No. 8, p. 572, 1964.

20. Spriggs, R. M., Brissette, L. A., and Vasilos, T., "Pressure Sintering Nickel Oxide," *American Ceramic Society Bulletin,* Vol. 43, No. 8, p. 572, 1964.

21. Anderson, Arthur, *Effect of Biological Sterilization and Vacuum on Certain Parachute Retardation System Components,* Final Report on JPL PO-BE4-229753, Cook Electric Company, Tech-Center Division, Morton Grove, Illinois, August 1964.

VI. Applied Mechanics

A. Feasibility Study of Parachutes for Use in the Mars Atmosphere

J. Brayshaw

Mars landing vehicles or atmospheric probes may require a decelerator system, auxiliary to the basic entry vehicle, to furnish low impact velocity or extension of descent time, respectively. The staged parachute system is a promising candidate decelerator technique because of its high drag-force-to-weight ratio and extensive experience with Earth recovery systems. Preliminary results of a contract with Cook Electric Company, Tech-Center Division, Morton Grove, Illinois, indicate that such a system is feasible for the possible range of Mars atmospheric conditions presently thought to exist, although certain development areas (discussed below) require early attention.

1. Conditions

Nominal design conditions assumed were: (1) a JPL Mars 10-mbar surface pressure and lowest density atmospheric model (no wind), (2) an initial entry velocity of 26,000 ft/sec at an altitude of 800,000 ft, (3) entry path angles of -20 to -90 deg (below horizon), (4) an entry vehicle ballistic coefficient of 0.17 slug/ft^2, and (5) a vehicle Earth weight of 350 lb. Changes in decelerator system performance resulting from variations of the above constraints were also considered. All weights are expressed in Earth units.

2. Recommended System

The two-stage decelerator system described below with its related performance is the recommended design, for the above nominal conditions, resulting from this study.

(1) A mortared 12-ft D meter hyperflo parachute is deployed from the entry vehicle at Mach 3.0 and a minimum altitude of 32,000 ft. Total weight is 350 lb.

(2) A 59-ft D extended-skirt reefed parachute is extracted by the first stage at Mach 0.85 and a minimum altitude of 24,000 ft. During this deployment, a 40-lb aft vehicle cover is removed. Total weight is 310 lb.

(3) The extended-skirt parachute is disreefed and fully inflated at an equilibrium descent rate of approxi-

mately 30 ft/sec and a minimum altitude of 22,000 ft. During the disreefing process, the 210-lb front heat shield is jettisoned. Total weight is now 100 lb.

(4) After a 12.5-min minimum descent time in condition (3) above, a payload weighing 75 Earth lb is deposited on the Mars surface.

(5) Total weight attributable to the decelerator system is 54 lb (65% parachute, 16.5% explosive disconnects, and 18.5% sensors, mortar, and accessories).

3. Design Trends

The study showed that first-stage deployment at higher Mach numbers and, consequently, at higher vehicle decelerations (if selected to occur at a higher altitude to circumvent atmospheric and topographic uncertainties) incurred considerable first-stage weight penalties because of increased aerodynamic heating and inertial loads. Furthermore, this change did not significantly increase second-stage deployment altitude. Thus, Mach 3.0 deployment (for which some successful flight-test experience has been accumulated) was selected.

4. Work Remaining

Major efforts still to be accomplished in the development of this decelerator system are:

(1) The design of a sequencing system to sense maximum tolerable deployment conditions in order to initiate stage deployments at maximum practicable (design) altitudes. Several likely methods proposed must be analyzed and tested.

(2) A flight-test program conducted at Earth altitudes greater than 100,000 ft in order to simulate second-stage parachute deployment mechanics and drag and stability performance on Mars. This test program would check deployment times, transient loads, and equilibrium performance extrapolated in the present study from current Earth experience gained in a much greater density environment at low altitude.

(3) A wind tunnel test program to check the sensitivity of a high-speed first-stage parachute in the wake conditions behind blunt entry vehicle shapes. High-speed parachutes have been successfully flown behind rocket-launched test vehicles at the contemplated Mars density and Mach number, but only with a pointed slender vehicle shape.

(4) Long-term (hundreds of days) materials tests to determine the effects of in-transit vacuum exposure after sterilization. Recent Cook tests over periods up to 30 days on the proposed Nomex and Dacron materials have demonstrated insignificant fabric strength loss, but have been inconclusive in establishing a rate of strength loss versus time. Pyrotechnic devices also need such a check.

ENGINEERING FACILITIES DIVISION

VII. Computer Applications Data Systems

A. On a Consistent Ephemeris

B. G. Marsden

Astronomers are often criticized for their continued use of values of astronomical constants which now appear to be considerably in error. Disapproval has also been expressed that some of the values recommended for adoption at the recent General Assembly of the International Astronomical Union are not the best available; in fact, the Union has positively rejected any plan to change certain constants, such as the constant of general precession in longitude and the mass ratios of the planets to the Sun, when the adopted values have been in use for more than sixty years and are known in some cases to be far from the truth. It has been supposed that lunar and planetary ephemerides based on these constants will be entirely useless for the space experiments which JPL and other organizations have planned.

The point is, many of these quantities depend upon each other in an extremely intricate manner, and a change in the value of one can have a profound effect on the others. If the precession constant is altered, the proper motions of all the stars are altered, too. Frequent changes of this constant in the past have made it an extremely tiresome process to put all the observations of Mars, for instance, on a basis suitable for comparison with theory. If the astronomers of two centuries ago had been as foresighted as those who flourished at the beginning of this century, a formidable amount of unnecessary calculation would have been avoided. The process of comparison has been further complicated by inadequacies in the theory of Mars that is currently in use; empirical terms having been introduced in an attempt to remedy the situation. There has been no improvement with the publication (Ref. 1) of a numerical integration of the motion of Mars, for the reduction of Earth-bound observations necessarily requires knowledge of the position of the Earth. In all the

available ephemerides of the Earth it is supposed that the Venus–Sun mass ratio is exactly 1:408000; on the other hand, in the integration of the motion of Mars a ratio of 0.00000246:1 was used, which differs from the other value in the third significant figure. Since the perturbations by Venus on both Earth and Mars are quite large, a comparison of the observations with these theories is impossible.

What is required is a solution of the fifty-fourth order system of equations that Newton's law of gravitation provides (relativistic corrections being applied in some prescribed manner), the constants of integration being selected so that, for instance, the squares of the residuals between the observed and computed positions can be neglected. At JPL, numerical integrations of the orbits of the nine principal planets have been fitted to so-called source data, and considerable effort has gone into the business of making as close a fit as possible. The source data have consisted of Newcomb's (Ref. 2) theories of Mercury, Venus and the Earth–Moon barycenter, augmented by the corrections obtained by Clemence (Ref. 3) and Duncombe (Ref. 4), Clemence's (Refs. 5 and 6) theory of Mars, and the numerical integration (commonly called the SSEC) by Eckert, Brouwer and Clemence (Ref. 7) of the motions of the five outer planets, plus the corrections (Ref. 8) to take into account the effects of the four inner ones.

There are good reasons why the integrations and the source data differ in something like the sixth figure. Newcomb's theories of the inner planets are only of the first order in the masses of the disturbing planets; consequently, numerous periodic terms, the precise number of which is hard to estimate, are missing from them. This is also true of the corrections to allow for the effects of the inner planets on the outer ones. Clemence's third-order theory of Mars is the best hitherto devised for any planet, but there is no guarantee that neglected higher-order terms are not significant. In any case, residuals of a periodic nature should not cause much difficulty. More serious is the fact that the source data contain secular terms which are inconsistent with the Newtonian equations.

Newcomb's calculations of the secular motions of the elements are based only on first-order theories, but, with one exception to be noted later, they are in tolerable agreement with the observations. However, Newcomb used the value $p_{1900} = 5024''.93$ for the general precession in longitude per Julian century at 1900, the currently accepted value—that used for Clemence's theory of Mars and for the integrations of the outer planets—being $0''.82$ larger. Thus, corrections of $+0''.82T$ (T is measured in centuries from 1900.0) should be applied to Newcomb's values of the longitudes of perihelia ($\tilde{\omega}$) of the orbits of Mercury, Venus and the Earth, and to the longitudes of the ascending nodes (Ω) of the orbits of Mercury and Venus. Further, the empirical terms applied by Newcomb to the longitudes of perihelia ($+43''.37T$ for Mercury, $+16''.98T$ for Venus and $+10''.45T$ for the Earth) and that applied to the perihelion of the Earth–Moon barycenter ($+7''.68T$) to allow for the nonsphericity of the system, should be *removed* before the integration is fitted to the theory.

Further difficulties arise if the corrections found by Clemence and Duncombe (Refs. 3 and 4) are applied to the orbital elements of these planets, because Clemence assumed that $p_{1900} = 5026''.50$ and Duncombe that $p_{1900} = 5026''.41$. Although the empirical terms were replaced by relativistic corrections, these latter must still be removed. In addition, Duncombe's secular variation corrections were derived from different values of the masses of the planets. If the mean longitude and motion of the Earth are changed, there are difficulties with the definition of ephemeris time. In the case of Mars, the relativistic motion of the perihelion should be removed from Clemence's theory. For the five outer planets the comparison with observations was made *before* the corrections for the effects of the inner planets were applied. Since these corrections contain secular terms, it would seem desirable to fit the nine-planet integration to the SSEC directly, rather than to the SSEC plus inner planet corrections. Actually, the difference between the two procedures is negligible. The recommended source data for all the planets are summarized in Table 1; the correction to the orbit of Mars must be applied differentially to the ephemeris provided by the U. S. Naval Observatory.

Table 1. Source data

Mercury	Newcomb with $\Delta\tilde{\omega} = -42''.55T$, $\Delta\Omega = +0''.82T$
Venus	Newcomb with $\Delta\tilde{\omega} = -16''.16T$, $\Delta\Omega = +0''.82T$
Earth–Moon	Newcomb with $\Delta\tilde{\omega} = -17''.31T$
Mars	Clemence with $\Delta\tilde{\omega} = -1''.35T$
Jupiter–Pluto	SSEC

After the numerical integrations of the Newtonian equations have been fitted, corrections for relativity must be applied. This is best done by moving the perihelion an amount

$$\frac{12\pi^2 a^2}{c^2 P^2 (1 - e^2)}$$

per revolution period P, where a and e are the semimajor axis and eccentricity of the orbit. A consistent and convenient way of applying the correction is to calculate at each step (from the position and velocity given by the integration) the osculating elements, adjust the argument of perihelion by the relativistic correction and the mean anomaly by the correction with sign reversed, and then recalculate the position and velocity. Values of the corrections for all the planets are given in Table 2. At some future date it may be desirable to integrate instead equations which take into account general relativity more rigorously, such as those given by Brouwer and Clemence (Ref. 9). In this case the fit should be made to source data increased by the motions of the perihelia given in Table 2.

Table 2. Relativistic motions

Mercury	$\Delta\bar{\omega} = +42''.981T$
Venus	$\Delta\bar{\omega} = +8''.625T$
Earth–Moon	$\Delta\bar{\omega} = +3''.839T$
Mars	$\Delta\bar{\omega} = +1''.351T$
Jupiter	$\Delta\bar{\omega} = +0''.062T$
Saturn	$\Delta\bar{\omega} = +0''.014T$
Uranus	$\Delta\bar{\omega} = +0''.002T$
Neptune	$\Delta\bar{\omega} = +0''.001T$
Pluto	$\Delta\bar{\omega} = 0''.000T$

It also is necessary to apply corrections to allow for the departures from sphericity of the planetary systems. The longitude of perihelion of the Earth–Moon system should be increased by $+7''.694T$, that of Jupiter by $+0''.006T$ and that of Saturn by $+0''.001T$.

A consistent numerical integration of the equations of motion of the nine principal planets, produced according to the precepts described here, will be a worthwhile result. If it covers the period 1700–2000, it will permit observations, both optical and radar, to be compared with it in a satisfactory manner; then, and only then, one can think seriously about changing the constants of integration and planetary masses and produce a new set of more accurate ephemerides. This integration has the further advantage of enabling one to investigate the expression for the obliquity of the ecliptic, for this is the element mentioned earlier where the observed and computed motions disagree. Since the ecliptic is defined to be the *mean* path of the Earth (or the Earth–Moon system) some methods of numerical analysis must be applied to the integration of the motion of the Earth–Moon system in order to remove the periodic terms. The resulting data can then be analyzed in the manner given by Newcomb (Ref. 2) or, better, by Andoyer (Ref. 10) and the theory of planetary precession revised.

In discussing the observations of the planets an ephemeris of the Earth, rather than of the Earth–Moon system, is needed. Consequently, an ephemeris of the Moon is required. It has been suggested (Ref. 11) that the terms in Brown's lunar theory containing the factor

$$\alpha_1 = \frac{1}{0.9990931420}\frac{\bar{a}}{A}\frac{E-M}{E+M},$$

$\bar{a}$ and A being the mean distance of the Moon and the astronomical unit, and E and M the masses of the Earth and the Moon, be adjusted to agree with more recent values of these quantities. If we take $GE = 398603$ km^3 sec^{-2} (G being the constant of gravitation) $A = 149{,}600{,}000$ km, $\mu^{-1} = E/M = 81.30$, we must add the terms given in Table 3. The resulting expression for sine parallax must then be multiplied by 0.999927379 (rather than by 0.999953253) in order to be consistent with an Earth equatorial radius of 6378.160 km. Finally, the heliocentric position of the Earth ($\mathbf{r}_E$) is obtained from that of the Earth–Moon barycenter ($\mathbf{r}_B$) and the geocentric position of the Moon ($\boldsymbol{\rho}$) by

$$\mathbf{r}_E = \mathbf{r}_B - \frac{\mu}{1+\mu}\boldsymbol{\rho}.$$

Table 3. Corrections to the lunar ephemeris

To the longitude:	$+0''.168 \sin D$
	$-0''.001 \sin 3D$
	$+0''.011 \sin (l + D)$
	$-0''.025 \sin (l - D)$
	$-0''.004 \sin (l - 3D)$
	$-0''.024 \sin (l' + D)$
	$-0''.001 \sin (l' - D)$
	$+0''.001 \sin (2l + D)$
	$-0''.002 \sin (2l - D)$
	$-0''.002 \sin (2l - 3D)$
	$-0''.002 \sin (l + l' + D)$
	$+0''.001 \sin (l - l' - D)$
	$-0''.001 \sin (2F - D)$
To the latitude:	$+0''.014 \sin D$
	$+0''.002 \sin (l + D)$
	$-0''.001 \sin (l - 3D)$
	$-0''.002 \sin (l' + D)$
To sine parallax:	$+0''.0013 \cos D$
	$+0''.0001 \cos (l + D)$
	$+0''.0001 \cos (l - 3D)$
	$-0''.0002 \cos (l' + D)$

l and l' are the mean anomalies of the Moon and Sun, F the argument of latitude of the Moon, and D the synodic angle $\lambda - \lambda'$.

In the past, two different values of the Earth–Moon mass ratio have been in official use: $\mu^{-1} = 81.53$ for the lunar theory and $\mu^{-1} = 81.45$ for the ephemeris of the Earth. It is not difficult, but very effective, to replace both these values by $\mu^{-1} = 81.30$.

There are undoubtedly errors of a periodic nature in Brown's lunar theory, and the question naturally arises as to how useful it would be to integrate numerically the equations of motion of the Moon. This could be desirable for aiding the reduction of observations of a satellite orbiting the Moon. It will be necessary either to remove the secular effects of the figures of the Earth and the Moon ($\Delta\tilde{\omega} = +6''.44T$, $\Delta\Omega = -6''.14T$) or to allow for them by including the harmonic terms in the equations of motion. If they are removed, they must be replaced after the fit, and one should include relativistic motions (Ref. 12) of $\Delta\tilde{\omega} = +1''.97T$, $\Delta\Omega = +1''.91T$. One should also remove the tidal term in the mean longitude $\Delta\lambda = -11''.22T^2$ and replace it afterwards.

The set of ephemerides would not be complete without the nutation and aberration day numbers. For the former, Woolard's theory (Ref. 13) should be used. For the latter, there should be introduced the value of the constant of aberration consistent with the astronomical unit and the velocity of light ($c = 173.1422$ AU per day), namely, $\kappa = 20''.4958$. They are calculated from

$$C = \frac{\dot{y}_\oplus}{c} - \kappa e \cos \tilde{\omega} \cos \varepsilon$$

$$D = \frac{\dot{x}_\oplus}{c} - \kappa e \sin \tilde{\omega},$$

ε being the obliquity of the ecliptic, the elliptic terms being necessary since they are conventionally included in the star positions.

A set of ephemerides based on the above precepts will be of great use to astronomy and will serve as the beginning of any program to improve further any astronomical constants, the assumed values of which are those recommended by the International Astronomical Union. The only serious inconsistency remaining is that between the astronomical unit and the ratio of the mass of the Sun to that of the Earth plus the Moon (which should be 328912, rather than 329390).

References

1. Brady, J. L., and Vienop, E., *Heliocentric Coordinates of Mars, 1800–2000*, Lawrence Radiation Laboratory, Livermore, California, 1962.
2. Newcomb, S., *A Compendium of Spherical Astronomy*, The Macmillan Co., New York, New York, 1906.
3. Clemence, G. M., *Astronomical Papers of the American Ephemeris*, Vol. XI, Part I, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1943.
4. Duncombe, R. L., *Astronomical Papers of the American Ephemeris*, Vol. XVI, Part I, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1958.
5. Clemence, G. M., *Astronomical Papers of the American Ephemeris*, Vol. XI, Part II, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1949.
6. Clemence, G. M., *Astronomical Papers of the American Ephemeris*, Vol. XVI, Part II, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1961.

References (Cont'd)

7. Eckert, W. J., Brouwer, D., and Clemence, G. M., *Astronomical Papers of the American Ephemeris,* Vol. XII, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1951.

8. Clemence, G. M., *Astronomical Papers of the American Ephemeris,* Vol. XIII, Part V, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1954.

9. Brouwer, D., and Clemence, G. M., *The Solar System,* Vol. III, p. 43, G. P. Kuiper and B. M. Middlehurst, Ed., University of Chicago Press, Chicago, Illinois, 1961.

10. Andoyer, H., *Bulletin Astronomique de l'Observatoire de Paris,* Vol. 28, p. 67, 1911.

11. Clemence, G. M., *Astronomical Journal,* Vol. 53, p. 169, 1948.

12. deSitter, W., *Monthly Notices of the Royal Astronomical Society,* Vol. 77, p. 155, 1917.

13. Woolard, E. W., *Astronomical Papers of the American Ephemeris,* Vol. XV, Part I, Nautical Almanac Office, U. S. Naval Observatory, U. S. Government Printing Office, Washington, D. C., 1953.

VIII. Aerodynamic Facilities

A. Wind Tunnels

J. Minich, V. Johns, R. Hiller, M. Argoud, and B. Dayman, Jr.

1. Wind Tunnel Test 20-566, Mauler,

J. Minich and V. Johns

Wind Tunnel Test 20-566 was of the General Dynamics Corp. (Pomona, California) 0.2-scale *Mauler* model. The test was made to determine the aerodynamic cross-control derivatives and the effect of simulating the rocket exhaust plume at the model base. The approximate aerodynamic parameters for the test were Mach numbers 1.65, 2.41, and 3.26, with corresponding Reynolds numbers/in. of 0.338×10^6, 0.386×10^6, and 0.303×10^6. The test variables and ranges were angles of attack from -10 to $+20$ deg, roll angles from 0 to -90 deg, and tail-deflection angles from 0 to -20 deg.

The model configuration comprised a cylindrical body with a modified tangent ogive nose and a 6-deg boattail. The configurations had four rectangular-planform, uncambered tail surfaces. Forces and moments were obtained for the complete configuration.

2. Base Pressure Measurements on a Two-Dimensional Wedge,

R. Hiller

Wind Tunnel Test 20-572 was an investigation of the effects that the aspect ratio, model width (b)/base height (h), has upon the base pressure of a two-dimensional wedge.

The models were 6-deg half-angle wedges (which spanned the wind tunnel) with base heights of 0.5, 1.0, 2.0, and 4.0 in., giving a geometric aspect ratio of 36, 18, 19, and 4.5, respectively. At Mach number 4.54, the corresponding aspect ratios based upon the distance between the Mach line and model base intersection are 32, 15, 6, and 2, respectively. The base of the model was instrumented with six pressure taps located on one-half of the span centerline.

Data were taken at Reynolds numbers/in. of 0.03×10^6 through 0.33×10^6, but only in a limited Reynolds number/in. range of 0.03×10^6 through 0.08×10^6 is the Reynolds number based upon a model length (in this instance the base height) constant for the four models.

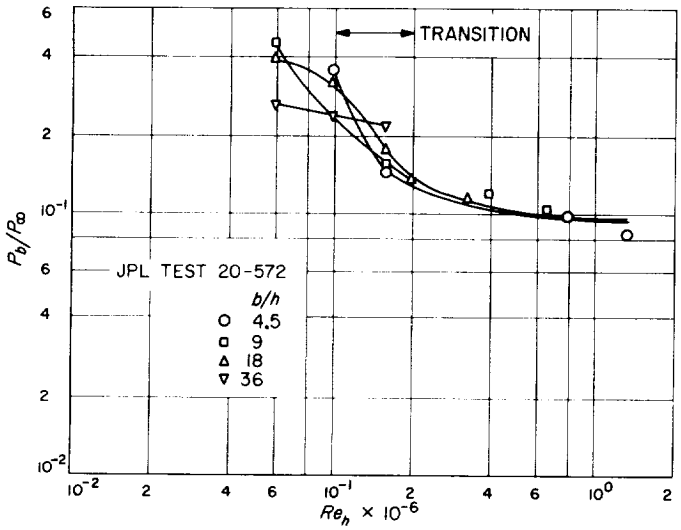


Fig. 1. Base pressure versus Re_h for different aspect ratios

Fig. 1 shows the base pressure plotted versus Re_h for the different aspect ratios. The area indicated as transition was determined from shadowgraph and spark schlieren photographs. The data show the pressure ratio at Re_h of 0.06×10^6 is lowest and at Re_h of 0.16×10^6 is highest on the smallest model. This crossover occurs in the transition range, and it would appear that one of the more significant effects of varying b/h is to delay or alleviate transition.

Future investigations will include additional models with higher aspect ratios to help define the effect of aspect ratio on base pressure in the laminar region.

3. Wind Tunnel Test 21-168, Nike-Iroquois,

M. Argoud and V. Johns

Wind Tunnel Test 21-168 was of the Space General Corp. (El Monte, California) 0.06-scale *Nike-Iroquois* model. The test was performed to investigate the effects of interactions between stabilizing fin geometry and the body-shed vortex system in producing vehicle rolling moments.

The approximate aerodynamic parameters for the test were Mach numbers 5 to 8 and corresponding Reynolds numbers/in. of 0.12×10^6 and 0.29×10^6.

The test variables and ranges were angles of attack from -4 to $+16$ deg, angles of roll from 0 to 60 deg, and five different afterbodies differing in the planform geometry.

The model configuration was made up of a 29.4-in.-long centerbody with a 20-deg included-angle conical nose and five different afterbodies differing in the planform geometry of the attached stabilizing fins; in one case, the afterbody had only one fin.

Forces and moments were obtained for the complete configuration.

4. Dynamic Stability Studies, B. Dayman, Jr.

The effect of several variables (cone angle, nose bluntness, base configuration and Mach number) on the dynamic stability of free-flight cones was investigated during Wind Tunnel Test 21-171 in the hypersonic tunnel and 20-599 in the supersonic tunnel. In all cases the wire-release (Ref. 1) technique was used to launch the models. The nomenclature used for these tests is given in Table 1.

The models (0.5 and 1.0 in. diameter) used were constructed of thin plastic or metal (aluminum or magnesium) shells which were ballasted with lead spherical

Table 1. Nomenclature

A	model reference area, $\pi d^2/4$, in.^{2}
C_D	drag coefficient, C_D = total drag/qA
C_m	pitching moment coefficient, C_m = pitching moment/qAd
C_{m_α}	effective linear pitching moment slope per radian
$C_{m_q} + C_{m_{\dot\alpha}}$	dynamic damping-in-pitch coefficient per radian, $C_{m_q} + C_{m_{\dot\alpha}} = \left[\frac{\partial C_m}{\partial\left(\frac{\dot\theta d}{V}\right)}\right] + \left[\frac{\partial C_m}{\partial\left(\frac{\dot\alpha d}{V}\right)}\right]$ (assumed constant over cycle)
d	model base diameter (reference length), in.
L	model length (for blunted models, length taken as that for sharp-nose model), in.
M	freestream Mach number
q	freestream dynamic pressure, lb/in.^{2}
r_B	model base radius, in.
r_N	radius of blunted nose, in.
V	freestream velocity, in./sec
$X_{c.g.}$	distance of center of gravity aft of nose (for blunted models from sharp nose before blunting), in.
α	angle of attack
$\bar\alpha_{env}$	effective (average) oscillation amplitude over a complete trajectory, $\bar\alpha^2_{env} = 2\delta^2$, deg
δ	root-mean-squared angle of attack during a complete trajectory, deg
σ	cone semi-vertex angle, deg
$\dot\theta$	model angular velocity, deg/sec

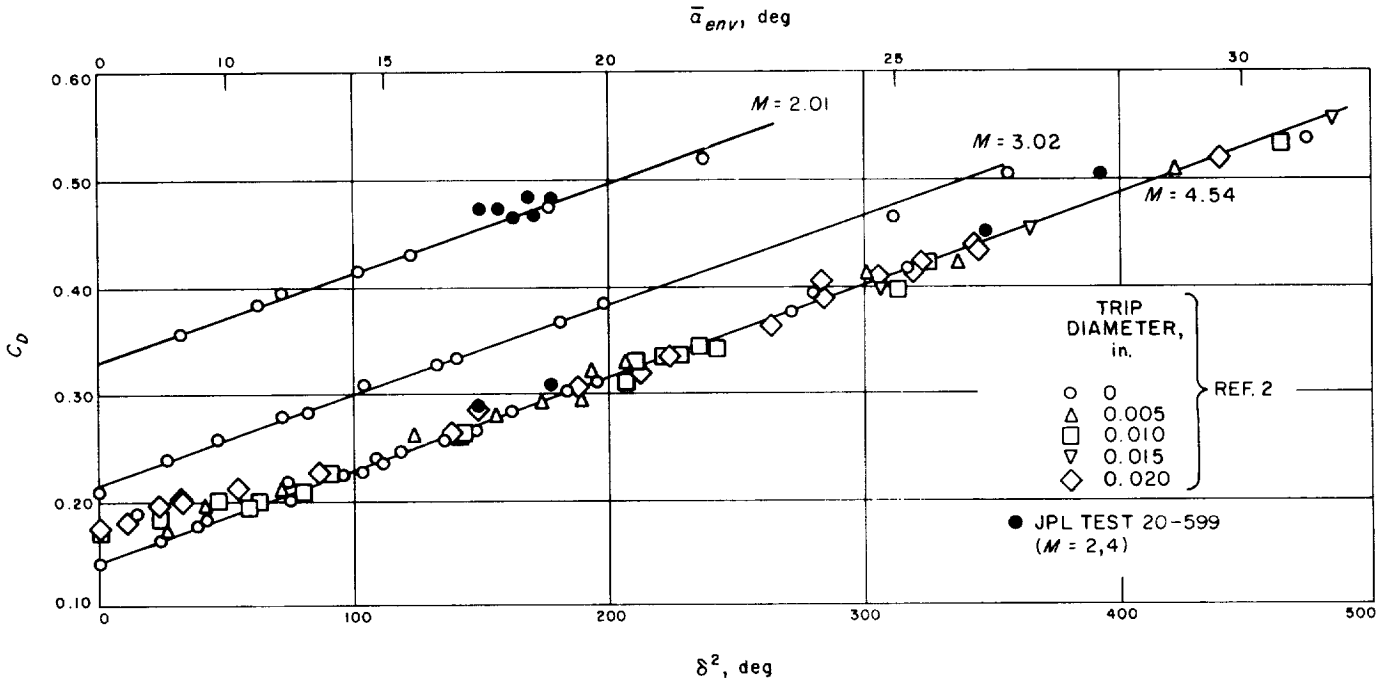


Fig. 2 Effect of oscillation amplitude on effective cone drag

cores. They were released into free-flight trajectories at angles of attack from 20 to 40 deg. The model motion was recorded on high-speed (2000–5000 frames/sec) half-frame 35-mm motion picture film using conventional high-speed cameras. Back-light (either silhouette or schlieren) was used to outline the model. Multi-flash (500–1000 flashes), short-duration (2 μsec) strobe light, synchronized to the camera, was used in order to eliminate model-image motion during film exposure (60 μsec or longer). Although the model oscillatory motion was normally confined to the vertical plane, a second camera was used to record the motion in the horizontal plane.

Although the purpose of these tests was measurement of model dynamic stability, pitching moment slope and drag were obtained during the data reduction. The effect of oscillation amplitude on drag is shown in Fig. 2. Data from these tests are compared to the extensive data of Ref. 2 in which gun-launched tests were run during April 1964. The comparison is satisfactory for both flat and hemispherical afterbodies on the sharp-nose, 10-deg half-angle cones. Fig. 3 shows the cone pitching moment slope as a function of Mach number. Contrary to the drag comparison, the effect of the hemispherical afterbody affects the pitching moment appreciably. Fig. 4 presents cone damping through a Mach number range ($2 < M < 6$) as a function of oscillation amplitude.

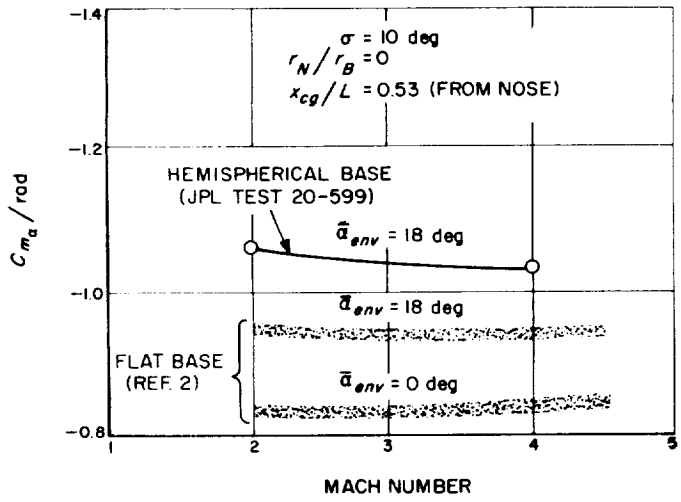


Fig. 3. Cone effective pitching moment slope

Dynamic stability data from Ref. 2 are transferred to the conditions of Wind Tunnel Test 20-599 for presentation in Fig. 5. Here again the recent data compare favorably with the previous data. The hemispherical afterbody on the cone models does not affect the damping, with or without boundary layer trip. The trip has been shown (Ref. 2) to give turbulent cone boundary layer and wake at zero angles of attack.

Limited tests were performed at $M = 6$ in order to compare the damping of cones with different apex angles. In Fig. 6 the comparison of trends with Newtonian theory is shown to be quite good. At $M = 2$ and 4, several flights were made with flat-based cones blunted to a nose radius to base radius ratio of 0.2. This amount of blunting decreased the damping significantly more than predicted by Newtonian theory (see Fig. 7). The location of the model center of gravity, in respect to the base, was the same for the blunted and sharp cones.

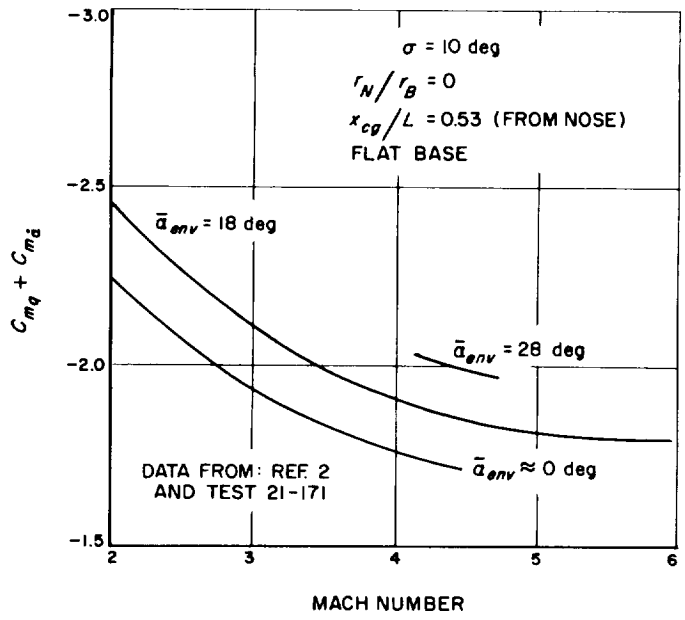


Fig. 4. Effect of Mach number on cone dynamic stability

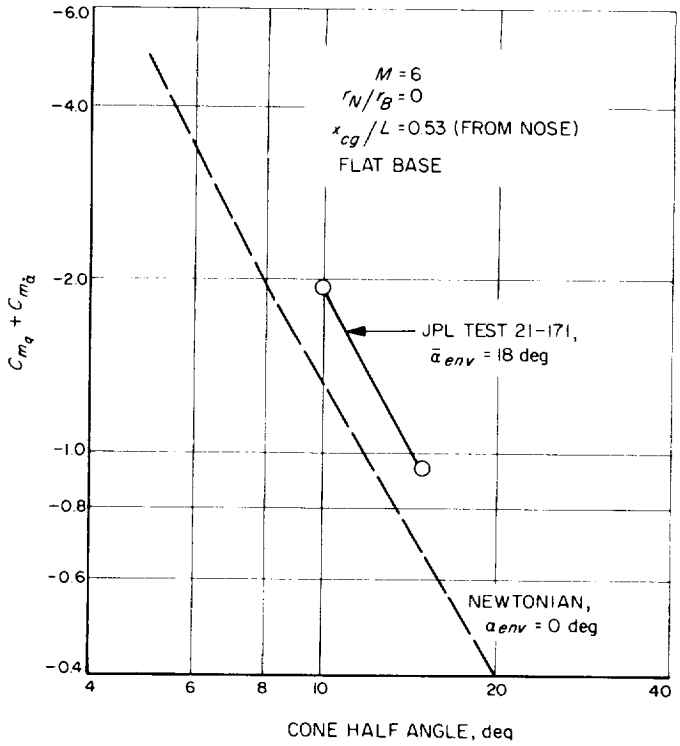


Fig. 6. Effect of apex angle on cone dynamic stability

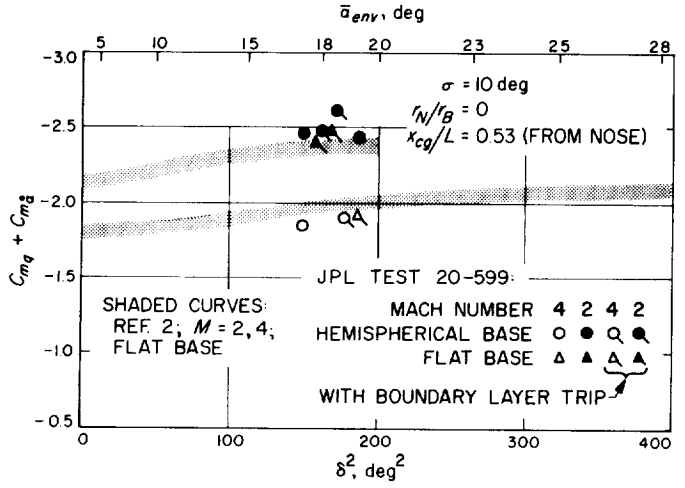


Fig. 5. Effect of oscillation amplitude on cone dynamic stability

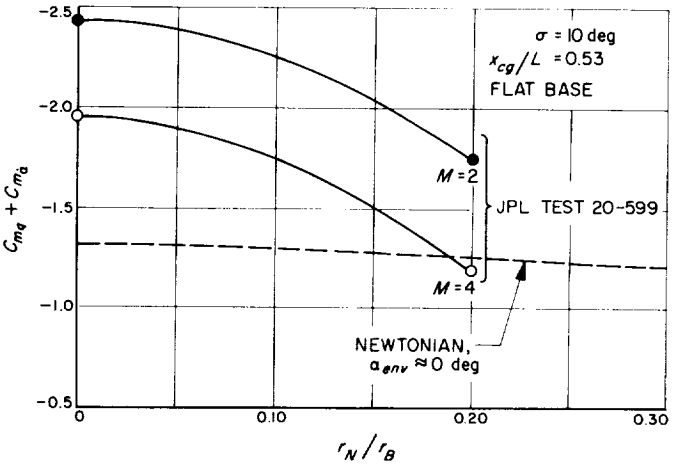


Fig. 7. Effect of nose bluntness on cone dynamic stability

B. Hypervelocity Laboratory

F. R. Livingston

1. Shock Tunnel Design and Performance

The Mach-12 shock tunnel has been designed for operation from the reflected region of a 6760-ft/sec incident shock propagated into 12.5 cm Hg of air in a 3-in.-diameter shock tube. Unheated hydogen is used as the driver gas in the shock tube. Conditions behind the reflected shock at the shock tunnel entrance are: temperature = 6370°R; pressure = 108,000 lb/ft^2; and enthalpy = 2000 Btu/lb.

The nozzle inviscid contour is axisymmetrically shaped with 12-deg half-angle source flow and contour computed by Cresci (Ref. 3) for frozen flow. The throat radius has been adjusted to 0.030833 ft to allow a Mach number 12 test section condition for the nonequilibrium flow in the nozzle. The nozzle has been shortened by not including the downstream 40% portion of the theoretical length. Theoretical exit diameter is 2.5 ft. The inviscid contour has been corrected for boundary layer displacement thickness by the method of Enkenhus and Maher (Ref. 4) as applied by R. McKenzie[1] at the NASA Ames Research Center. The corrected nozzle is 12.6 ft long and has an exit diameter of 3.6 ft.

[1]Personal communication, July 1964.

Using the vibrational freezing criterion of Stollery and Park (Ref. 5) and applying the recently measured nitrogen vibration relaxation times of Hurle, Russo and Hall (Ref. 6) in an expanding flow, the vibrational temperature at freezing was estimated to be 3350°R at the design condition. Knowing the freezing point, other test core parameters have been calculated assuming equilibrium flow to the freezing point and frozen expansion to the test core.

Pertinent shock tunnel parameters are:

Translational-rotational temperature, °R	276
Static pressure, atm	3.00×10^{-4}
Velocity, ft/sec	9460
Density, slug/ft^3	1.29×10^{-6}
Mach number	12.0

The shock tunnel uses an existing 3-in.-diameter shock tube for the hot gas supply. Most other components of the shock tunnel have been designed and are now being built. The nozzle and test section are being made of resin impregnated fiberglass by Tolo, Inc., of Fullerton, California.

References

1. Dayman, B., Jr., "Simplified Free-Flight Testing in a Conventional Wind Tunnel," *Technical Report No. 32-346*, Jet Propulsion Laboratory, Pasadena, California, October 1962.
2. Jaffe, P., and Prislin, R. H., "Effect of Boundary Layer Transition on Dynamic Stability over Large Amplitudes of Oscillation," *AIAA Preprint 64-427*, (Presented at the First AIAA Annual Meeting, Washington, D. C., June–July 1964).
3. Cresci, R. J., "Tabulation of Coordinates for Hypersonic Axisymmetric Nozzles," Wright Air Development Center, Technical Note 58-300, Part 1, October 1958.
4. Enkenhus, K. R., and Maher, E. F., "The Design of Axisymmetric Nozzles for High Temperature Air," *Advances in Hypervelocity Techniques*, Plenum Press, New York, March, 1962.

References (Cont'd)

5. Stollery, J. L., and Park, C., "Computer Solutions to the Problem of Vibrational Relaxation in Hypersonic Nozzle Flows," *Journal of Fluid Mechanics*, Vol. 19, Part 1, May, 1964.

6. Hurle, S. R., Russo, A. L., and Hall, H. G., "Experimental Studies of Vibrational and Dissociative Nonequilibrium in Expanded Gas Flows," AIAA Conference on Physics of Entry into Planetary Atmospheres, AIAA Paper 63-439, Massachusetts Institute of Technology, August, 1963.

PROPULSION DIVISION

IX. Solid Propellant Engineering

A. Heat-Sterilization Propellants

L. C. Montgomery

Since JPL's adoption of the dry-heat sterilization approach to producing a sterile landing capsule, the Propulsion Division has taken three approaches to the problem. The prime approach is to find an "off-the-shelf-propellant" that would satisfy the immediate needs of planetary landing programs initially reported in *SPS 37-26*, Vol. IV. The second approach is to develop a "backup" propellant from known heat-resistant ingredients as reported in *SPS 37-27*, Vol. IV. The third is to develop a higher energy heat-sterilizable system to upgrade the propulsion system for future requirements.

The heat-sterilization requirements state that each component must reliably operate after being subjected to sterilization temperatures of 145°C for a period of 36 hr. Three cycles at this temperature are required to accommodate component changes and subsequent sterilization that may be required at the launch site.

1. Propellants Tested

In response to inquiries from JPL, seven companies indicated they had candidate propellants capable of meeting the sterilization requirements indicated in *SPS 37-26*, Vol. IV. These companies supplied JPL with samples of propellants for sterilization evaluation. The propellants are identified as follows: (1) Aerojet-General, Sacramento propellant 583 AF; (2) Atlantic Research Corporation propellant 413A; (3) Naval Propellant Plant propellant PVC-A; (4) Rocketdyne, McGregor propellant RDS-510-2A; (5) Thiokol Alpha propellant TP-H-8162; (6) Thiokol Elkton propellant TP-H-3105; and (7) United Technology Corporation propellant UTX 5113. The propellants TPH-3105, TPH-8162, RDS-510-2A and UTX-5113A are based on a polybutadiene binder system; the

propellants 413A and PVC-A are plasticized polyvinylchloride systems, and the propellant AN 583 AF is a polyester styrene binder system.

The planned test program consisted of initial screening of the propellants by slump testing of cantilevered specimens. The propellants surviving this test were then subjected to a more comprehensive physical evaluation test program consisting of heating and testing of tensile specimens, and 3- × 3- × 6-in. block specimens used for tensile tests. All the data is referenced to similar data taken from nonheated specimens.

2. Equipment

The equipment items used in these tests were disposable ovens and an Instron tester. Sterilization heating of specimens was done in both air and nitrogen atmospheres. Heating under an atmosphere of nitrogen was accomplished by a flow of nitrogen gas through the oven. The exhaust from the oven was bubbled through water to assure a slight positive pressure inside the oven. The disposable ovens were constructed using a cylindrical commercial metal can wrapped with electrical heating tape and then well insulated. The temperature of the oven was controlled by use of a manually adjusted Variac until precise temperature controllers were obtained.

3. Testing and Results

The initial screening evaluation tests consisted of heating cantilevered ½- × ½- × 4-in. specimens suspended with a 3-in. overhang. The specimens were heated to 145°C, maintained at that temperature for 36 hr and then returned to ambient. Those specimens surviving the first heating cycle were then subjected to a second cycle at this temperature and then a third cycle. The tests were performed in atmospheres of both nitrogen and air.

Four propellants from these initial tests qualified for further evaluation in this program. These were the polyester styrene-based propellant (AN 583AF) and three polybutadiene based propellants (RDS-510-2A, TP-H-3105 and TP-H-8162). The AN-583 AF propellant showed no change due to heating during any of the temperature cycles except for a slight discoloration. The RDS-510-2A, TP-H-3105 and TP-H-8162 all softened on the first heating and showed a drop of approximately 1 in. at the free end of the bar. Subsequent cycles showed no further change except for further darkening of the samples. Identical tests, except for a nitrogen atmosphere, were made on identical samples. The results of the nitrogen atmosphere tests could not be distinguished from those of the air atmosphere tests.

A carboxyl-terminated polybutadiene propellant (UTX-5113A) and both plasticized polyvinyl chloride propellants (413A and PVC-2) did not survive the initial heating cycle of the slump tests in air or nitrogen gas. In air the polybutadiene softened to the point that it dropped off the holder into the bottom of the oven; in nitrogen it "wilted" and hung straight down. The polyvinylchloride propellant broke off at the point of maximum stress. The break indicated that the propellant had started to slump; then a "brittle" fracture occurred.

The next series of tests were for determining more precisely the effect of the sterilization heat cycles on the physical properties of the four propellants which survived the slump test. In these tests precut JANAF tensile specimens, torsion samples, and 3- × 3- × 6-in. blocks of propellants were heat cycled and tested.

The tensile samples were heated in atmospheres of both air and nitrogen. Typical results of the tests of the propellants are shown in Figs. 1 through 4.

Although the amount of data taken is minimal, some interesting trends have been noted. In the TP-H-3105 propellant the ultimate tensile strength increases to a maximum with the first heat cycle in air and shows no further change in subsequent heating. However, a slightly higher maximum is reached on the second heat cycle in the N_2 atmosphere, and then the ultimate tensile strength drops off. The percent elongation at ultimate strength is approximately the same in both cases.

In the case of the RDS-510-2A propellant there appears to be no differences in the ultimate tensile strength if the propellant is heated in air or N_2, and little change occurs. Even though there is little change in the ultimate tensile strength, a great change occurs in the elongation in the two tests. The elongation drops from the 20 to 30% bracket to around 7.5% on the first heating cycle and continues a downward trend to under 5% on the third cycle. Little deviation from these trends occur with the samples heated in N_2.

The ultimate tensile strength data for propellant TP-N-8162 heated in atmospheres of air and N_2 showed

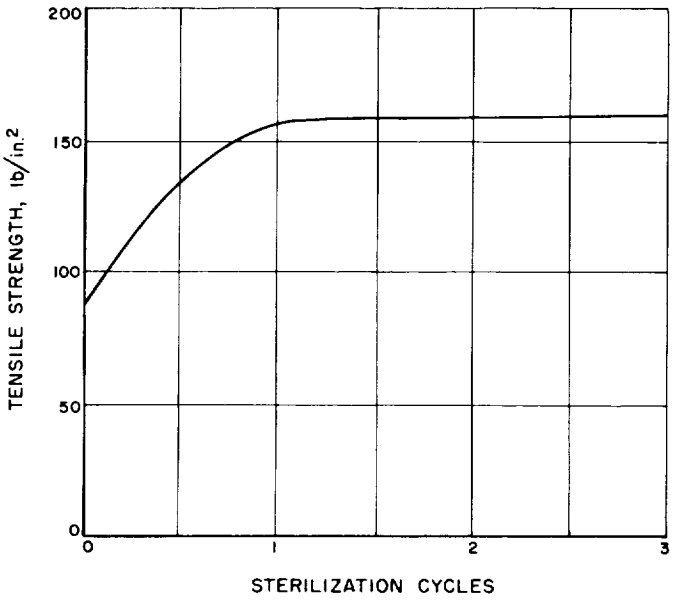


Fig. 1. Tensile strength of propellant TP-H-3105 versus number of sterilization cycles at 294 ± 2°F for 36 hr in air

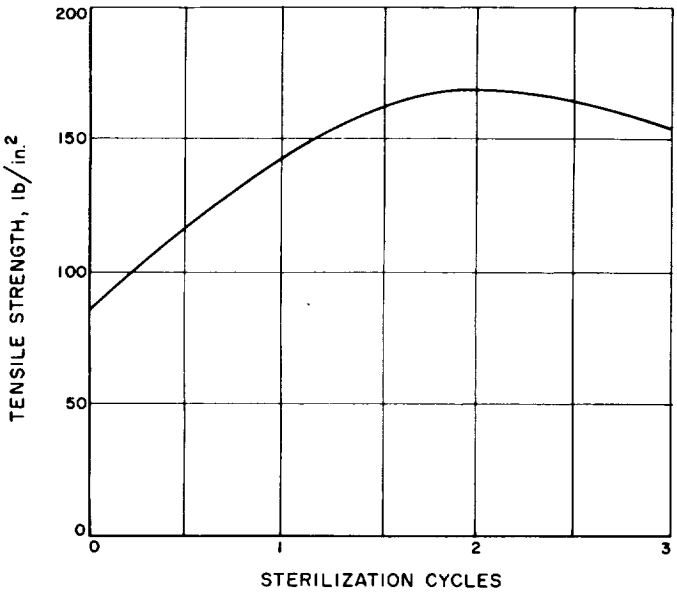


Fig. 2. Tensile strength of propellant TP-H-3105 versus number of sterilization cycles at 294 ± 2°F for 36 hr in N_2

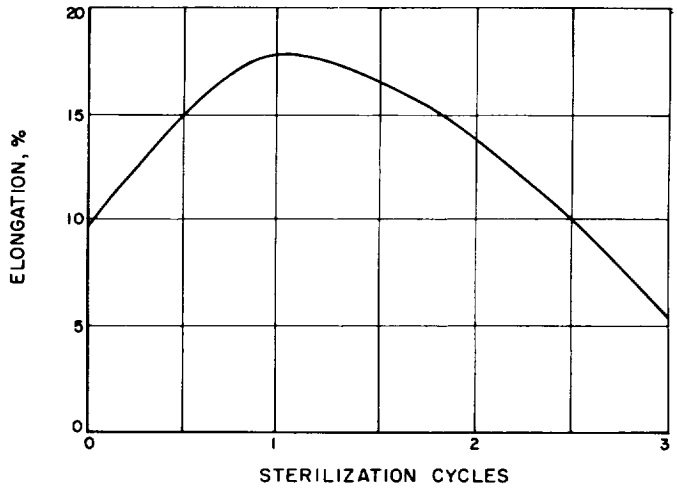


Fig. 3. Elongation percent of propellant TP-H-3105 versus number of sterilization cycles at 294 ± 2°F for 36 hr in air

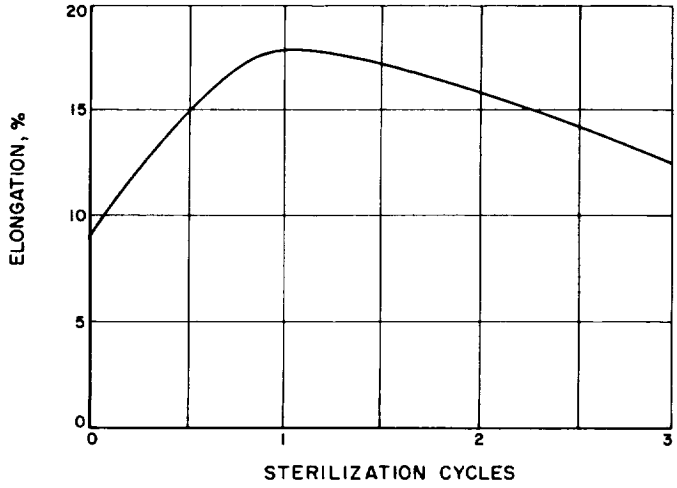


Fig. 4. Elongation percent of propellant TP-H-3105 versus number of sterilization cycles at 294 ± 2°F for 36 hr in N_2

a slight trend down on the first cycle, an increase to a maximum tensile strength on the second cycle and a drop to a minimum just under the tensile strength for the nonheated specimens on the third cycle. However, the propellant elongation acted similarly to the RDS 510-2A propellant and dropped drastically from around 20% elongation to 5% on the first sterilization cycles in both atmospheres and continued down with each successive heating.

The AN-583-AF propellant indicated even more unexpected trends. Both the ultimate tensile strength and the elongation increased. The elongation at ultimate tensile strength was improved very slightly by sterilization heating in both atmospheres. The ultimate tensile strength of the propellant when heated in air increased

from 250 to 430 psi, to 540 and to 600 psi on the successive heatings in air. The nitrogen atmosphere caused the ultimate tensile strength to level off at just about 500 psi on the second two heating cycles.

Since the tensile bars just discussed were precut, a large surface area was exposed to the surrounding environment. Therefore, to get a better picture of the heating effect on the internal propellant, blocks of propellant 3 × 3 × 6 in. were subjected to sterilization heating in an atmosphere, then cut into JANAF tensile specimens and tested. Typical results of these tests are shown in Figs. 5 and 6. The charts at the bottom indicate the position each tensile specimen had occupied in the 3- × 3- × 6-in. block. The surface crust was left on specimens numbered 1, 4, 5, and 8. Specimens numbered 2, 3, 6 and 7 were out from internal sections of the block with the test section being no closer than ¼ in. from the outside edge of the block. Specimens 9 and 10 were also supposed to be internal specimens. However in some cases not enough material remained to allow for the heated crust to be taken off the specimen. Therefore data from these specimens are shown on the plots but are not used in the analysis. It should be noted that the specimens 1, 4, 5 and 8 have only one side of crust, and therefore the data from these specimens will not necessarily agree with the data from specimens heated in the form of precut JANAF specimens. However, trends from the specimens containing crust should be similar. Hereafter in this discussion preheated JANAF tensile bars will be referred to as the "heated specimens" and the term "block specimens" refers to those tensile specimens cut from 3- × 3- × 6-in. block after heating.

The ultimate tensile strength data from propellant TP-H-3105 block specimens show the same trends as the heated JANAF specimens, but lower values. Surprisingly the specimens having a crust show a slightly lower ultimate tensile strength than do the center cut specimens. The elongation shown by the block specimens having crust agree with the heated specimens in that a maximum is reached on the first cycle and then elongation drops off with successive heating cycles. The elongation of the internal block specimens continues to increase.

The ultimate tensile strength from the block specimens of RDS 510-2A propellant does not agree with that from the heated specimens on the first sterilization heat cycle, but tests of specimens from the last two heat cycles do show agreement. This anomaly will be further investigated. The internal block specimens show a trend to decrease in ultimate tensile strength on the first heat cycle and an increase on subsequent cycles, while ultimate tensile strength of the crust samples decrease with

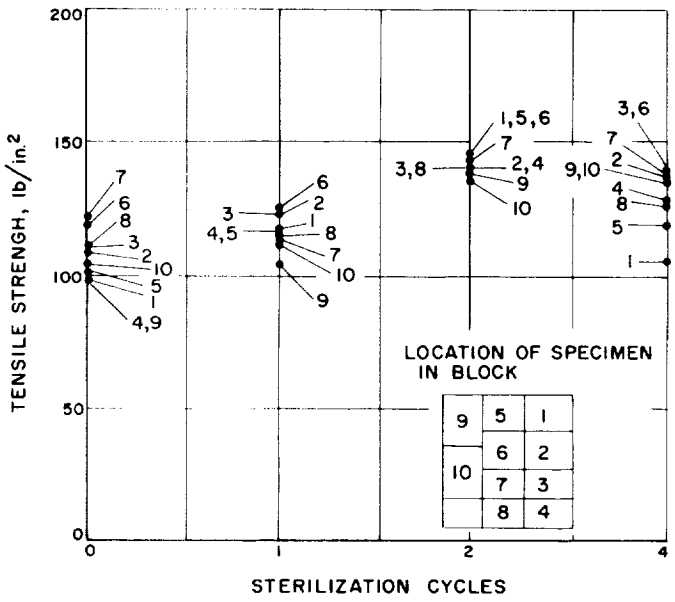


Fig. 5. Tensile strength of propellant TP-H-3105 versus number of sterilization cycles at 294 ± 2°F for 36 hr when heat sterilized in block form

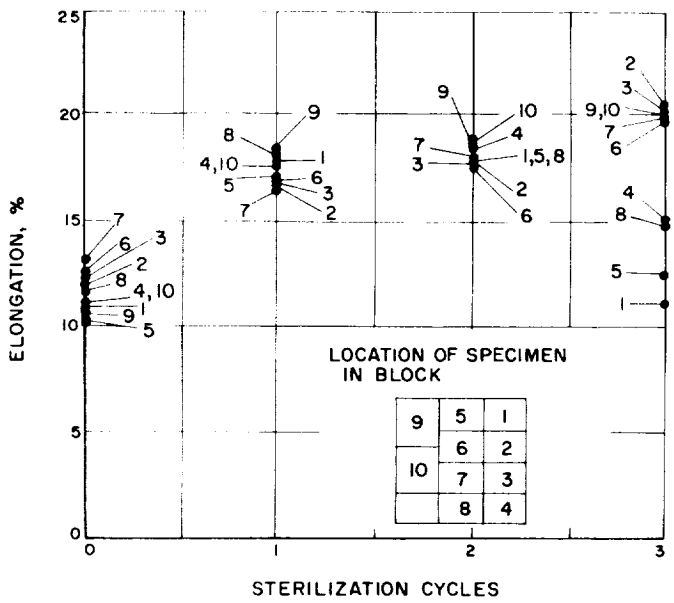


Fig. 6. Elongation percent of propellant TP-H-3105 versus number of sterilization cycles at 294 ± 2°F for 36 hr when heat sterilized in block form

each heat cycle. The downward trend in elongation of the external block specimens agrees with that of the heated specimens, but the internal block specimens show an increase in elongation on the first heat cycle and then drop to just below the initial elongation condition on cycles 2 and 3.

For the propellant AN-583-AF, the trends indicated by the heated specimens hold true in the outside block specimens as well. The ultimate tensile strength at the internal part of the block shows a slight decrease on the first two cycles but drops from about 290 to 230 psi on the third cycle. Again the elongation shows a very slight tendency to increase, which is a very important factor for this study.

The block of propellant TP-H-8162 cracked internally when heated, as shown in Fig. 7. No external evidence gave indication that this was happening although it was measured, weighed, and examined after each heat cycle. However when it was cut for fabrication of JANAF bars it was found to be unusable.

Torsion tests were run on each of the four propellants. These data have been compiled and are now being evaluated. If this information is found to contribute significantly to the aims of this study, it will be included in the final report.

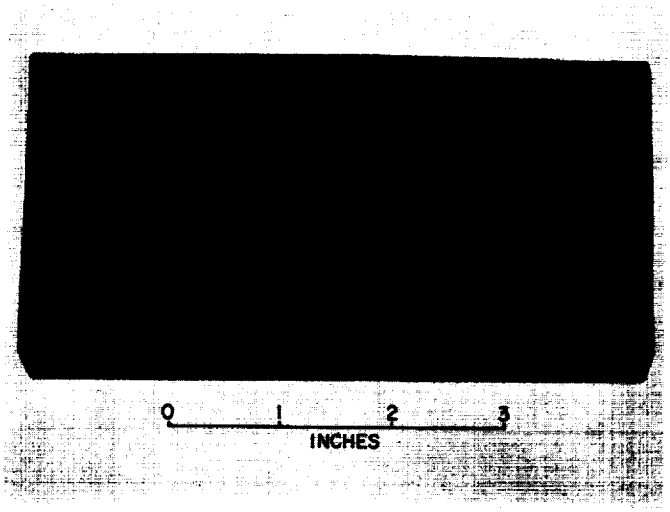


Fig. 7. Cut surface of TP-H-8162 propellant after being heated in block form for three 36-hr sterilization cycles at 295°F, showing internal cracking

4. Summary

The trends indicated in this report are gleaned from data on propellants submitted by propellant manufacturers as those which can withstand heat sterilization. The propellants were heat sterilized in small and large pieces; in aid and N_2. Slump, tensile, and torsion tests were run.

It was found that the polyvinyl chloride propellants acted alike and broke in the slump test when heat sterilized.

Small changes in binder formulation have a profound effect on the way in which the polybutadiene binder responds to sterilization heat cycling. Of the two carboxyl-terminated polybutadiene propellants, one "melted" under the sterilization treatment while the other showed little change in tensile strength and a drastic decrease in elongation.

Of the two polybutadiene acrylic acid binder propellants one initially increased in tensile strength and elongation, followed by a decrease in elongation in the second and third cycles, while the elongation of the second one was drastically reduced on the first cycle while showing little change in tensile strength. This second propellant was found to crack internally when heated in a large block. These data emphasize the need for basic studies in the effects of formulation on propellant characteristics in regard to this problem and the need for basic studies in understanding the internal stresses of a grain under various heating conditions.

The polyester styrene propellant showed a large increase in tensile strength from heat sterilization with a slight tendency to increase in percent elongation.

For this program the propellants which show the least decay in physical properties both in the internal grain and the surface are those which can be used in a heat sterilizable propulsion system. The two propellants which appear the most promising are TP-H-3105 and AN 583 AF.

The propellants will be cartridge loaded in the initial designs to avoid the problem of differential expansion of case and liner and/or case and propellant. The required design will allow for expansion of the propellant without interference with the case.

B. Low-Pressure Combustion Studies

Leon Strand

1. *Influence of Aluminum Coarseness on Completeness of Low-Pressure Combustion*

In *SPS 37-27*, Vol. IV, pp. 49–50, it was reported that the substitution of coarser aluminum than that usually used in JPL polyurethane-ammonium perchlorate propellants resulted in propellant extinction characteristics which approached those of polyurethane-ammonium perchlorates containing little or no aluminum in their formulations. Incomplete low-pressure combustion of the coarser aluminum in the modified propellant, resulting in propellant burning characteristics that approach those of a nonaluminized propellant, was postulated as a possible explanation for these results.

The three propellants reported on have since been test fired in the JPL 3-in.-D, 6-in.-long flanged test motor in the Edwards Test Station vacuum test system. The motor chamber lengths and nozzle throat diameters were similar to those used in transparent motor firings with these propellants. The purpose of these tests was to obtain comparative low-pressure c^* efficiencies for the three 16% aluminum propellants used in the low-pressure unstable combustion investigation.

Table 1 gives some of the important results for the four test firings. The $\bar{c}^*$ values reported are each the averaged value for the two motor pressure gage digital records. The $\bar{c}^*$ value for Run Number 1 has to be discounted as erroneous, because of its unreasonably low value. The remaining test results are as expected, the $\bar{c}^*$ efficiencies progressively worsening with increased aluminum coarseness. The 31-μ aluminized propellant $\bar{c}^*$ efficiency value is 10% less than the value obtained for JPL-540 propellant, with its 7-μ aluminum particle size. The value of the $\bar{c}^*$ efficiency for the 15 to 17-μ aluminized propellant is 8% less.

It is not possible from these few tests to draw definite conclusions concerning the influence of incomplete combustion on low-pressure unstable combustion. However, the differences in completeness of combustion (c^* efficiency) between the unmodified and modified propellants do appear large enough in magnitude to at least partially account for the different low-pressure extinction characteristics.

The remaining area of possible influence of aluminum particle size on low-pressure propellant combustion lies within the propellant solid phase itself. In a private communication Dr. Ralph Anderson of the United Technology Center suggested that the influence of the propellant macroscopic structure, i.e., packing density of various sized oxidizer and aluminum granules, on propellant subsurface reactions could be an important consideration in attempting to explain the different propellant low-pressure extinction characteristics. Additional tests are being planned in order to pursue these investigations further.

Table 1. Test motor firing results

Run No.	Propellant	Al, %	Average Al particle size, μ	Initial propellant		Throat diameter, in.	Run time, sec	P_{ce}, psia	$\bar{c}^*$, ft/sec
				Diameter, in.	Length, in.				
200	JPL-540	16	7	2.501	3.99	0.751	1.12	53.8	986
201	JPL-540	16	7	2.505	3.99	0.654	2.96	48.9	2539
206	JPL-540 Mod. 1	16	31	2.400	4.450	0.679	2.87	51.5	2278
222	JPL-540 Mod. 2	16	15–17	2.503	3.989	0.654	3.20	53.0	2338

*Subscript: extinction

X. Polymer Research

A. Poly(Propylene Oxide), the Correct Name for Polypropylene Glycol

R. F. Landel and J. D. Ingham

Inconsistencies have appeared in the naming of polymers (particularly with respect to polyethers) in previous publications and reports from the Propulsion Division. This originated in part because these were relatively new materials when initially employed for propellant binders and technically incorrect names were adopted. These have been widely used in the scientific and technological literature, followed by more correct naming in some JPL reports.

Specifically, the polymer obtained from propylene oxide has been most often called polypropylene glycol and abbreviated as PPG. This terminology persisted until about 1960, when the name polyoxypropylene glycol (POPG) began to appear. This is the correct designation if the emphasis is to be on the fact that the material is a glycol with a polyether backbone rather than a hydrocarbon backbone. However this nomenclature is not widely used nor is it consistent with the general method applied to most polymers. By convention, an addition polymer should be named by using the prefix "poly" followed by the name of the monomer unit in parentheses, e.g., poly(styrene), poly(vinyl acetate), poly(methyl methacrylate). Thus the polymer based on propylene oxide or ethylene oxide should be called poly(propylene oxide) and poly(ethylene oxide), which may be abbreviated as PPO or PEO. Interestingly enough, the standard method is normally used for PEO. Unless stated otherwise, it is ordinarily assumed that PPO or PEO are hydroxy-terminated. If these terminal groups have been modified, for example to carboxylic acid groups, then the material should be designated as carboxy-terminated poly (propylene oxide). Dimers and trimers should be called di- and tri-propylene oxide instead of di- and tripropylene glycol as in the past.

Since it is formally correct and relatively concise, this naming method has been adopted by the Propulsion Division and will be used in all succeeding publications whenever possible.

B. Hydrolysis of Higher Molecular Weight Alkoxypropionitriles

S. H. Kalfayan

1. Introduction

Carboxy-terminated polyalkylene oxides can be prepared by the hydrolysis of cyanoethylated polyalkylene oxides, e.g., PPO-1000, PEO-600,[1] etc. The cyano group can be changed to the carboxyl group either by acid or alkaline hydrolysis. However, for cyanoethylated polyalkylene oxides, alkaline hydrolysis is ruled out because, in the presence of alkalis, such alkoxypropionitriles are decomposed into the original reactants, acrylonitrile and glycol, particularly at higher temperatures (Refs. 1, 2). Hydrolysis with concentrated hydrochloric acid was used in all the experiments reported below; 40% sulfuric acid was also tried but extensive discoloration took place in the reaction mixture, and hydrolysis with this acid was discontinued.

Analyses of hydrolyzed products indicated less carboxy-terminated material than expected. At first this was thought to be due to incomplete hydrolysis of the nitrile to carboxy acid. However, analysis indicated the presence of products other than unconverted nitrile or amide which would be an intermediate in the hydrolysis of nitrile to acids. A "model" hydrolysis reaction with β-2-butoxypropionitrile also showed other products formed in appreciable amounts after hydrolysis with concentrated HCl. The isolation of pure dicarboxylated polyalkylene oxide from the hydrolysis mixture proved to be difficult. Several methods of purification were tried, of which precipitation as an insoluble salt showed the most promise.

2. Experimental Results and Discussion

a. Hydrolysis of cyanoethylated PPO-1000. Since carboxy-terminated PPO-1000 is the compound of greatest interest, most of the hydrolysis experiments were carried out with cyanoethylated PPO-1000. The hydrolysis procedure consisted of heating the cyanoethylated glycol with twice the amount of concentrated HCl, either in sealed and agitated ampoules, or in flasks where the mixture was mechanically stirred. After removing the ammonium chloride formed by filtration, the reaction mixture was taken up in ether and washed free of chloride ion. The product was analyzed for acidity and OH content after removal of solvent and residual moisture. It was found that the temperature and time recommended in the literature (Ref. 2) for the hydrolysis of low molecular weight alkoxypropionitriles (75 to 80°C for 4 hr and 95 to 100°C for ½ hr) was not sufficient to complete the hydrolysis of the high molecular-weight alkoxypropionitrile. IR spectra still showed the presence of CN groups. Therefore, several combinations of time and temperature were attempted through the range of 75 to 100°C and 4- to 8-hr reaction time. Longer heating time produced more carboxy acid, but also led to cleavage of the alkoxy compound as evidenced from the increase of OH content (Table 1, Experiments 2 and 3). Some of the hydrolyzed products were also analyzed for chlorine, with positive results. Highest conversion to acid was obtained when the cyanoethylated PPO-1000 was heated with hydrochloric acid for 6 hr at 70 to 75°C and 2 hr at 95 to 100°C (Experiment 4).

b. Hydrolysis of β-2-butoxypropionitrile. In order to check for side reactions in the hydrolysis reaction, β-2-butoxypropionitrile was employed as a model compound, since low molecular weight side products would be more easily identifiable than those from cyanoethylated PPO-1000. The butoxypropionitrile was prepared according to methods previously described (Ref. 3) and showed a single peak when analyzed by gas chromatography. Like cyanoethylated PPO-1000 it is a secondary alkoxypropionitrile.

The hydrolysis of β-2-butoxypropionitrile was performed as recommended (Ref. 2), except that the reactions were carried out in sealed ampoules to avoid the loss of volatile side products. Not all of the side products from the hydrolysis of β-2-butoxypropionitrile were identified, but those that were shed some light on the hydrolysis of secondary alkoxypropionitriles.

After heating with HCl 4 hr at 70 to 75°C and ½ hr at 95 to 100°C, the composition of the hydrolysis products, as determined from weights of distillation fractions

[1]PPO = poly(propylene oxide); PEO = poly(ethylene oxide)

Table 1. Hydrolysis of cyanoethylated PPO-1000

Experiment No.	OH content cyanoethlyated PPO-1000, meq/g	Cyanoethylated glycol in mixture, %	Acidity of hydrolized product, meq/g	OH content of hydrolized product, meq/g	Carboxylic acid in hydrolized mixture[a], %	Hydrolysis temperature, °C	Total time, hr	Purified product		
								Acidity, meq/g	OH content, meq/g	Carboxylic acid,[a] %
1	0.52	73.5	0.74	—	43	76–80 (4 hr) 95–100 (½ hr)	4½			
2	0.17	91	1.17	0.24	68.4	95–100	5			
3	0.17	91	1.21	0.40	71	95–100	6½			
4	0.26	87	1.36	0.27	79.5	70–75 (6 hr) 95–100 (2 hr)	8			
5	0.49	75	1.02	—	60	85–90	4	1.07[b]	0.26	62.5
6	0.40	79.6	1.11	0.52	67	70–75 (5 hr) 95–100 (1 hr)	6	1.15[c]	0.30	70.7
7	0.52	73.5	0.94	—	55	95–100	6	1.21[d]	0.28	71
8	0.26	87	1.23	0.28	72	70–75 (5 hr) 95–100 (1 hr)	6	1.26[d]	0.21	73.7
9	0.26	87	1.23	0.28	72	70–75 (5 hr) 95–100 (1 hr)	6	1.43[e]	—	83.6

[a]Obtained from the theoretical acidity of 1.71 meq/g and the experimental acidity, Col. 4
[b]Method 1 for purification
[c]Method 2 for purification
[d]Method 3, aluminum salt precipitated
[e]Method 3, silver salt precipitated

and gas chromatographic peak areas, whenever these were obtainable, was roughly as follows:

β-2-butoxypropionic acid	60–61%
β-2-butoxypropionitrile	3–4%
2-chlorobutane	15–16%
2-butanol	4-5%
Unidentified	14–15%

For identification, chemical as well as gas chromatographic techniques were used. Fractions distilling above 65°C/0.4 mm of Hg could not be detected with a 20% Carbowax 20M column on Chromosorb P operated at a column temperature of 145°C. This included the β-2-butoxypropionic acid which distilled at 85–87°C/0.4 mm of Hg. Two unidentified peaks appeared with the fraction containing β-2-butoxypropionitrile (40–65°C/0.4 mm of Hg). The fraction distilling at 90–147°C/0.4 mm (but mostly at 140–147°C/0.4 mm) reacted vigorously with 10% $NaHCO_3$. There are, thus, at least four other unidentified compounds, two of which are, most probably, polymeric acids.

Presence of 2-butanol and 2-chlorobutane among the hydrolysis products, indicates the cleavage of the ether link in the alkoxypropionitrile. The cleavage might have been brought about thermally, or by the action of the mineral acid.

Christian and Hixon (Ref. 2) carried out the hydrolysis of 1,4-di-(2-cyanoethoxy)-pentane, I, with concentrated HCl and reported that about ⅓ of the hydrolyzed product was 2-chloroamyloxypropionic acid, II.

$$CH_3\,CH(OCH_2\,CH_2\,CN)\,CH_2\,CH_2\,CH_2O\,CH_2\,CH_2CN$$

I

$$CH_3\,CH(Cl)\,CH_2\,CH_2\,CH_2O\,CH_2CH_2\,COOH$$

II

The same authors reported yields, as much as 20% lower, for β-secondary alkoxypropionic acids than for β-primary alkoxypropionic acids. For example, β-1-propoxypropionic

acid was obtained in 80% yield, whereas β-2-propoxypropionic acid was obtained in 60% yield; β-2-methyl-1-butoxypropionic acid could be obtained in 69% yield and β-2-amyloxypropionic acid was obtained in only 49% yield.

These findings indicate that secondary alkoxypropionitriles are more susceptible to ether cleavage.

c. Purification of dicarboxy polypropylene oxide. Isolation of the desired dicarboxy polyalkylene oxide from the other components of the reaction mixture was attempted by using the following methods: (1) column chromatography, (2) extraction of an ethereal solution with alkali, and (3) precipitation of the carboxylic acid as an insoluble salt. In method (1) a solution of hydrolyzed material in hexane was passed through a column of Amberlite CG-400 (an anion exchange resin) and further eluted with hexane. Elution was then continued with CH_3OH/HCl. With method (2) an ethereal solution of the hydrolyzed product was extracted with 5% NaOH, the extract acidified with HCl and re-extracted with ether. After drying and removal of solvent, the acidity and OH content of the residue were determined. Method (3) consisted of precipitating the carboxylated material as an insoluble salt from an aqueous medium, washing the dried precipitate with petroleum ether or hexane and reconverting the salt to the carboxy acid by treatment with a mineral acid. Viscous oils resulted when the precipitation of cupric, magnesium, manganese, barium and zinc salts were tried. The silver salt was semisolid, and the aluminum salt had the consistency of soap. The silver and aluminum salts could be "washed" by Sohxlet extraction. The others swelled extensively in the extracting liquid.

Gains were marginal when purification methods (1) and (2) were employed to obtain dicarboxylated PPO's (Experiments 4 and 6). The acidities increased only by 2–4%. Using method (3) the acidity could be increased by as much as 16% (Experiment 7). However, the gain was still marginal when the acidity of the unpurified product was relatively high to begin with (Experiment 8). Highest acidity obtained by purification was by precipitating the silver salt (Experiment 9).

It is concluded that the hydrolysis of alkoxypropionitriles is complicated by side reactions, and in the case of high molecular weight alkoxypropionitriles, such as cyanoethylated PPO-1000, the side products are not readily removable. A purification method worth further investigation is column chromatography.

C. A Commentary on the Gibbs-DiMarzio Theory of the Glass Transition

J. Moacanin and R. Simha[2]

1. Introduction

The transition from liquid to glass is manifested by marked changes, within a narrow interval about the glass transition temperature T_g, in the first derivatives of the extensive thermodynamic quantities, such as heat capacity, compressibility, and thermal expansivity, as well as transport coefficients such as viscosity. Gibbs and DiMarzio (Ref. 4, 5, 6) used a statistical thermodynamic argument to predict for amorphous polymers the existence of a second-order transition, as defined by Ehrenfest, at a uniquely defined temperature T_2, at which the system has only one degree of freedom. This result was obtained by using the quasicrystalline lattice model to calculate the number of degrees of freedom for a mixture of holes and polymer molecules as a function of temperature. The experimentally observed T_G varies with molecular weight, diluents, copolymer composition, etc., in the same manner as the theory predicts for T_2. But the experimental T_G is time dependent, and is usually 30 to 50°K higher than T_2, (Ref. 7).

Recently, Simha and Boyer (Ref. 8) proposed for polymers the following relationship between the expansivities for the liquid α_L and glassy α_G states and T_G:

$$(\alpha_L - \alpha_G)\, T_G = K_1 \tag{1}$$

This equation with K_1 being a constant may be derived from the postulate that at T_G, the free volume fraction defined as

$$V - V_{0,L}\,(1 + \alpha_G\, T)\,/V \simeq V - V_{0,L}\,(1 + \alpha_G\, T)\,/V_{0,L}$$

is the same for all polymers; V is the total volume at temperature T, and $V_{0,L}$ the liquid volume extrapolated to 0°K. This concept of T_G as an iso-free volume state was first proposed by Fox and Flory (Ref. 9). An examination of experimental results showed that T_G for a number of polymers varying widely in structure, intermolecular forces, chain flexibility and geometry, ranged between 140 and about 420°K; whereas K_1, although not constant, varied only between 0.08 to 0.13°K.

[2] Consultant, University of Southern California, Los Angeles, Calif.

The purpose of this paper is twofold. First, we wish to explore the relation between the Gibbs–DiMarzio theory and the Simha–Boyer result. Our second purpose is to examine the relative importance of the chain flexibility and intermolecular energy in determining T_G. These two quantities, although not independent of each other, must enter any theory of T_G (or T_2). Qualitatively it is clear that the intermolecular energy plays but a minor role in the Gibbs–DiMarzio development.

2. Recapitulation of the Gibbs-DiMarzio Theory

The basic postulate of the theory, that the system at T_2 has only one degree of freedom, requires that the configurational entropy S_{conf} vanishes at this temperature. This condition for a polymer phase consisting of n_x chains of size x and n_0 holes may be expressed in the form (Ref. 4, Eq. 22):

$$\frac{S_{conf}(T_2)}{n_x kT_2} = 0 = \phi\left(\frac{E_0}{kT_2}\right) + \lambda\left(\frac{\epsilon}{kT_2}\right) + (1/x)\ln \left\{\left[(z-2)x+2\right](z-1)/2\right\} \quad (2)$$

where

$$\phi\left(\frac{E_0}{kT}\right) \equiv \ln\left(\frac{V_0}{S_0}\right)^{z/2-1} + \frac{V_0}{V_x}\ln\left(\frac{V_0}{S_0}\right) \quad (3)$$

and

$$\lambda\left(\frac{\epsilon}{kT}\right) \equiv \frac{x-3}{x}\left\{\ln\left[1+(z-2)\exp\left[-\epsilon/(kT)\right]\right] + \frac{\epsilon}{kT}\frac{(z-2)\exp\left[-\epsilon/(kT)\right]}{1+(z-2)\exp\left[-\epsilon/(kT)\right]}\right\} \quad (4)$$

E_0 is the energy of interaction for a pair of occupied nearest-neighbor segments, and gives the energy change when a contact between two filled sites is broken by introducing a vacant site. Its magnitude should be of the order of van der Waal's interactions. The fraction of sites which are unoccupied is defined as $V_0 = n_0/(n_0 + xn_x)$ and was shown to be related to E_0 by (Ref. 4, Eq. 11):

$$\ln\frac{V_0^{z/2-1}}{S_0^{z/2}} = \frac{z E_0 S_x^2}{2kT} \qquad T \geqq T_2 \quad (5)$$

with

$$S_0 = \frac{zn_0}{\left[(z-2)x+2\right]n_x + zn_0} = \frac{V_0}{1 - \frac{V_x}{z/2}\left(1-\frac{2}{x}\right)}$$

and

$$S_x = 1 - S_0;\ V_x = 1 - V_0;$$

z is the coordination number.

The parameter ϵ is the difference between the energy ϵ_2 for the "flexed" bond orientation and ϵ_1, the energy of the remaining $(z-2)$ possible orientations.

To express the thermal expansivities we use the following relations for the total volume:

$$V_{T>T_2} = C(T)\left[xn_x + n_0(T)\right] \quad (6)$$

$$V_{T<T_2} = C(T)\left[xn_x + n_0(T_2)\right]$$

where $C(T)$ is the volume of a lattice site at temperature T. Above T_2, n_0 is temperature dependent, whereas for $T < T_2$ the number of holes remains fixed at the value reached at T_2. It follows therefrom that the difference between the expansivity for the liquid state α_L and that for the glassy state α_G (i.e., for $T > T_2$ and $T < T_2$, respectively), at the limit $T \to T_2$ is given by:

$$(\alpha_L - \alpha_G)_{T=T_2} = \frac{1}{C}\frac{\partial C}{\partial T} + \frac{1}{xn_x + n_0}\frac{\partial n_0}{\partial T} - \frac{1}{C}\frac{\partial c}{\partial t} = \frac{1}{xn_x + n_0}\frac{\partial n_0}{\partial T}$$

By implicit differentiation of Eq. (5) and rearrangement of terms one finally arrives at the equation given by Gibbs and DiMarzio (Ref. 4, Eq. 24):

$$\alpha_L - \alpha_G = \frac{\frac{z E_0}{2kT_2^2} V_0 S_x^2}{\frac{z S_x}{2} - \frac{z E_0 S_0 S_x^2}{kT_2} - (z/2-1)V_x} \quad (7)$$

3. Discussion

For our purpose the expression for $(\alpha_L - \alpha_G)$ can be recast into a form analogous to Eq. (1) by multiplying Eq. (7) by T_2, and expressing terms in $zE_0(kT_2)$ by means of Eq. (5):

$$(\alpha_L - \alpha_G)T_2 = \frac{V_0 \ln\frac{V_0^{z/2-1}}{S_0^{z/2}}}{\frac{zS_x}{2} - 2S_0\ln\frac{V_0^{z/2-1}}{S_0^{z/2}} - (z/2-1)V_x} \quad (8)$$

This product, which is to be compared with K_1, is a function of V_0 only at fixed z and x, since the other parameters appearing on the right hand side of the equation can be expressed in terms of V_0. Moreover, V_0 is then uniquely determined at T_2, by the ratio ϵ/E_0 by virtue of Eqs. (2) to (5), as will be shown. Hence K_1, will ultimately be related to the intermolecular energy and chain flexibility.

In order to relate V_0 values to the experimental range of K_1 we show in Fig. 1 plots of the product $(\alpha_L - \alpha_G)\, T_2$ versus V_0, Eq. (7), for x from 10 to infinity, and $z = 4$, 8 and 12; the values selected for z should adequately represent lattice geometries of interest. Furthermore, if T_2 lies approximately 30 to 50°K below the experimental T_G (Ref. 7), then $T_2 \approx 4/5\ T_G$, and the experimental K_1 range of 0.08 to 0.13 (Ref. 7) should correspond to that of 0.064 to 0.104 for the theoretical values of the product. From Fig. 1, it can be seen that to these "corrected" K_1 values there correspond ranges for V_0 from 0.017 to 0.029, and 0.018 to 0.032 for $x = 10$ and ∞, respectively. For all practical purposes the asymptotic values are already reached for $x = 100$, and hence we shall set this discussion $x = \infty$. Interestingly enough, the relation between K_1 and the range of "experimental" V_0 is but slightly dependent on chainlength, and virtually independent of z. Also, the mean value for V_0 for large x is 0.025, which incidentally, is identical to the average fractional free volume of the WLF equation (Ref. 10).

The relationship between V_0 at T_2 and the intermolecular energy E_0 is determined by Eq. (5) and is illustrated in Fig. 2. Although these plots are strongly dependent on z they nearly superpose if the abscissa is multiplied by z. This is not surprising, since E_0 refers to the energy of interaction for a nearest-neighbors pair, whereas the total energy of interaction depends on the total number of pairs which is proportional to z. Thus we find that at the transition $zE_0/(kT_2)$ values which correspond to the range of "experimental" V_0 are insensitive to z and cover the fairly narrow range between about 5 and 6, implying near proportionality between E_0 and T_2.

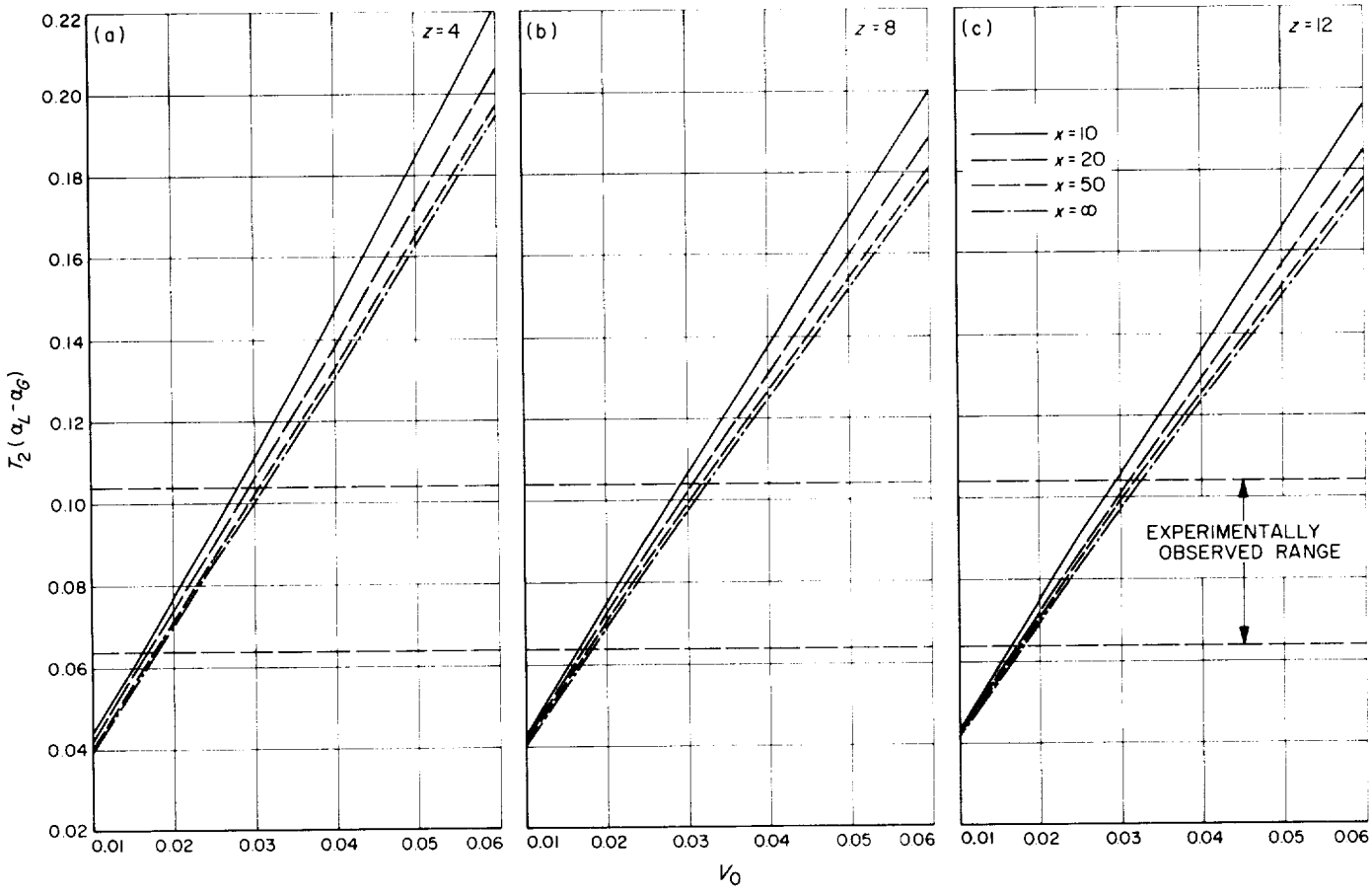


Fig. 1. $T_2(\alpha_L - \alpha_G)$ versus V_0

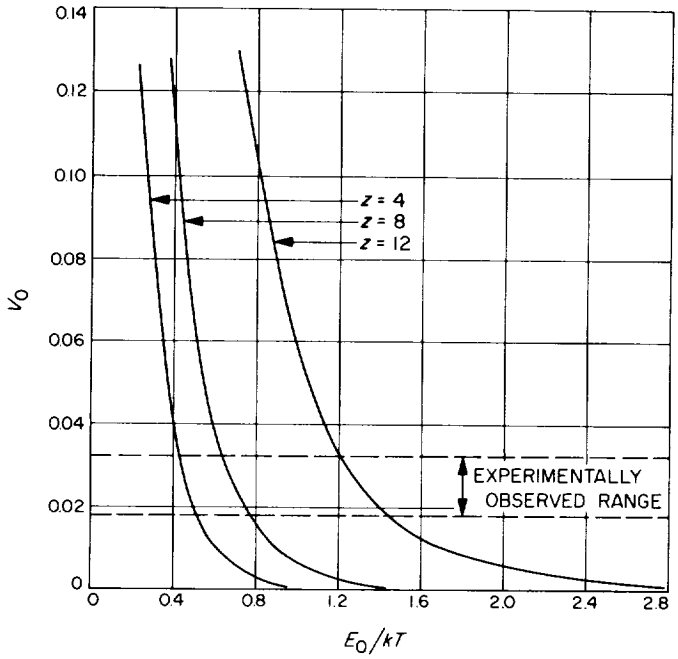


Fig. 2. Dependence of the unoccupied volume V_0 on the intermolecular energy E_0 (x = ∞)

So far we have not considered chain stiffness as measured by ϵ. But it is this quantity together with E_0 that determine T_2 through Eq. (2). The third term in Eq. (2) becomes negligible for reasonably large x. Thus the condition to be satisfied at $T = T_2$ becomes simply $-\phi = \lambda$. Figs. 3 and 4 show plots of the functions $-\phi$ versus $E_0/(kT)$ from Eq. (3) and λ versus $\epsilon/(kT)$ from Eq. (4). The values for both $E_0/(kT)$ and $\epsilon/(kT)$ which satisfy the desired condition for ϕ and λ are readily obtained by inspection of Figs. 3 and 4. Plots of $\epsilon/(kT_2)$ versus $E_0/(kT_2)$ obtained in this manner are given in Fig. 5. From the "experimental" $E_0/(kT_2)$, as computed from V_0, we get for $\epsilon/(kT_2)$ about 3.9 to 4.0, for $z = 12$, 3.3 to 3.4, for $z = 8$, and 2.1 to 2.2, for $z = 4$. Here z is an important factor, because by assumption only one of the $(z - 1)$ possible bond orientations is in a different energy state. Thus its weighting factor on the chain stiffness is strongly dependent on the geometry of the lattice.

From inspection of Fig. 5, the following relations between T_2 and the parameters ϵ and E_0 become apparent. In general, T_2 is a monotonically increasing function of ϵ at fixed E_0 seen as follows. Let $T_2'' > T_2'$, then $E_0/(kT_2'') < E_0/(kT_2')$, and from Fig. 5 it follows that $\epsilon''/(kT_2'') > \epsilon'/(kT_2')$, hence $\epsilon'' > \epsilon'$. Also, we note that for $\epsilon \to 0$, the transition temperature vanishes because the ratio $\epsilon/(kT_2)$ has to remain finite. Turning our attention to the effect of the intermolecular energy we note

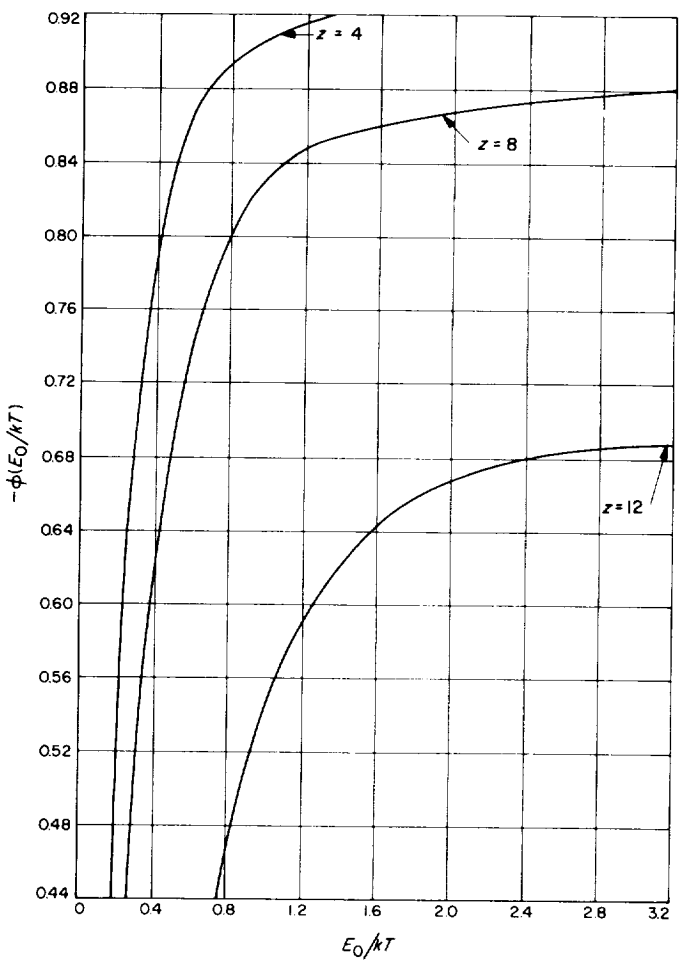


Fig. 3. The intermolecular energy dependent term of the configurational entropy versus $E_0/(kT)$ (x = ∞)

that for $E_0 \to \infty$ (i.e., $V_0 \to 0$ at all temperatures), T_2 is finite and is determined by the asymptotic $\epsilon/(kT_2)$ value. Furthermore, at the limit of $V_0 \to 0$ the change in slope at T_2 in the volume versus temperature relation is given from Eq. (8).

$$\lim_{V_0 \to 0} (\alpha_L - \alpha_G)\, T_2 = \lim_{V_0 \to 0} (-V_0 \ln V_0) = 0$$

Since T_2 remains finite as long as ϵ is finite, it follows that at the limit $(\alpha_L - \alpha_G)$ must be zero, showing that for this case there is no discontinuity in the volume expansivities. More generally $(\alpha_L - \alpha_G)$ will become negligible whenever $\epsilon < E_0$, since $V_0\ (T_2)$ rapidly approaches zero whenever $\epsilon/E_0 < 1$, since then $V_0\ (T_2)$ will virtually vanish (Fig. 6).

Although this is a surprising result, inasmuch as we are accustomed to associate T_G (and by inference T_2)

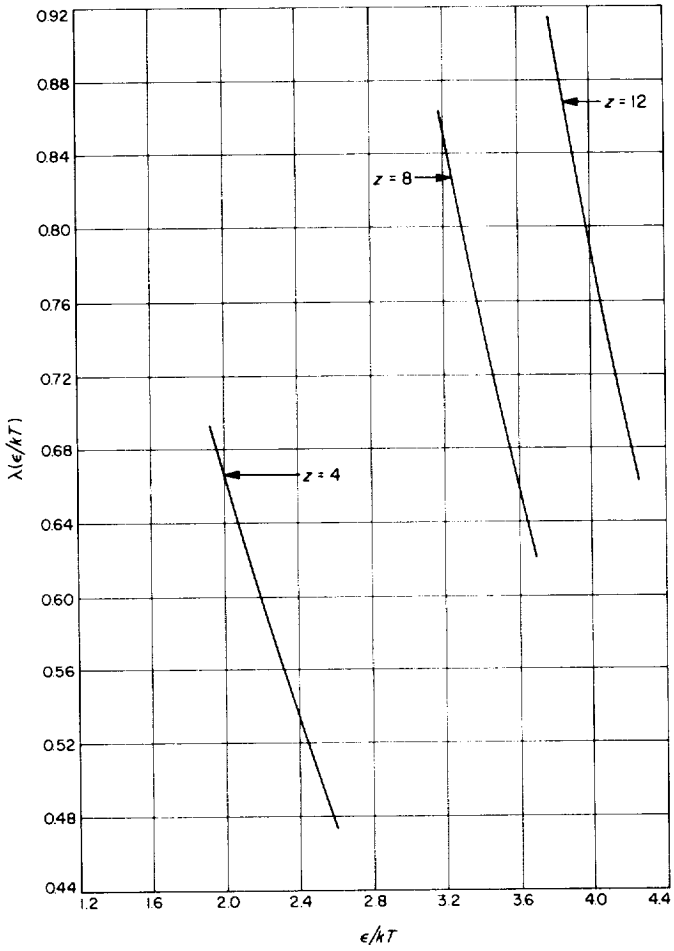


Fig. 4. The flex energy dependent term of the configurational entropy versus ε/(kT) (x = ∞)

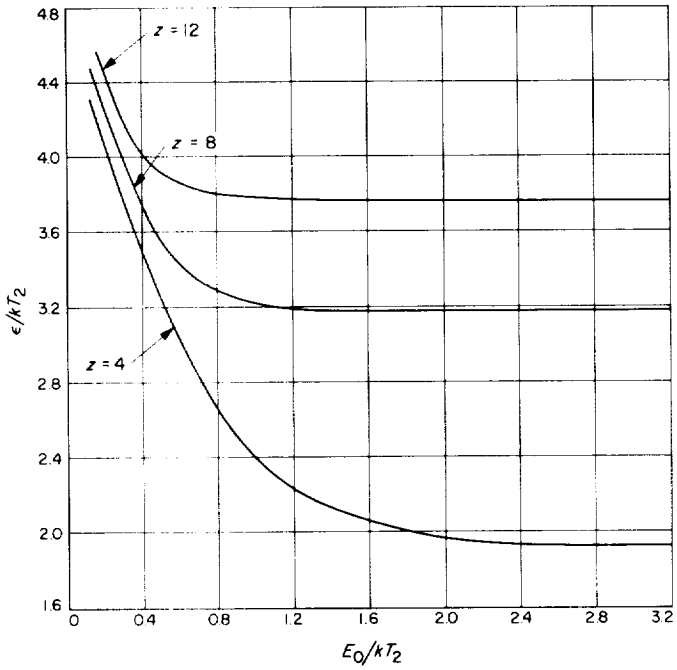


Fig. 5. Relationships between the transition temperature T_2 and the energy parameters ε and E_0 (x = ∞)

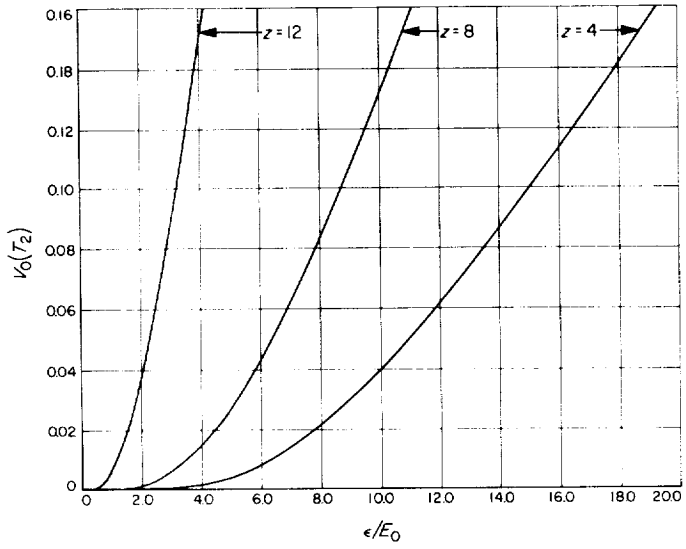


Fig. 6. The unoccupied volume at T_2 versus ε/E_0 (x = ∞)

with a change in volumetric behavior, there is no thermodynamic inconsistency in having a discontinuity in $(\partial S/\partial T)_P$ which is not paralleled by one in $(\partial V/\partial T)_P$, as may be seen by considering the relation

$$T\,(\partial S/\partial T)_P = (\partial E/\partial T)_P + P\,(\partial V/\partial T)_P$$

In general, as long as ϵ is finite, there will be always a discontinuity of T_2 in the derivative of the internal energy because of freezing-in of bonds in a fixed configuration. As a consequence, V_0 reaches a constant value, i.e., $V_0\ (T < T_2) = V\ (T_2)$, and a discontinuity in the volume derivative ensues. If, however, $V_0 = 0$ for all temperatures, the discontinuity in the volume term will disappear, but without removing the one in the energy term, and thus the discontinuity in the entropy derivative is preserved.

Numerical values for ϵ and E_0 can be estimated from experimental volume-temperature data in the following manner. The experimental value of K_1 is used along with Eq. (8) or Fig. 1 to calculate $V_0\ (T_G)$, which in turn is related to $E_0/(kT_G)$ through Eq. (5), and finally $\epsilon/(kT_G)$, or strictly $\epsilon/(kT_2)$ is obtained from Fig. 5. Thus Gibbs and DiMarzio used polystyrene data to calculate $E_0/(kT_G) = 1.19$ and $\epsilon/(kT_G) = 2.25$ (for $z = 4$). It is difficult to assess the physical significance of these numbers, since results obtained from independent methods are not

available. In principle the intermolecular parameter E_0 should be related to the cohesive energy density *CED*. By assuming that $(CED)^{1/2}$ (i.e., heat of vaporization) is given by the product of the number of contacts and E_0, one finds that the only way to obtain agreement with experiment is to assign to a lattice site a volume much smaller than that of the smallest natural chain segment. This of course violates the lattice model, but is a well known result of Eyring (Ref. 11), who deduced that the volume of a "hole" in a liquid should be about one sixth that of a molecule or segment. There are data available on the rotational potentials for gas molecules and these should be related to ϵ, but one cannot state *a priori* how the potential is affected by going to the liquid state, or by the presence of intersegmental bonds in a polymer.

We have attempted, however, to examine the internal consistency of the data by comparing dilatometric data on polystyrene and polydimethylsiloxane. Both polymers have identical K_1 values (Ref. 8), and consequently equal V_0 (T_G), $E_0/(kT_G)$ and $\epsilon/(kT_G)$. It follows therefore that the proportionality constant relating E_0 and ϵ values for polystyrene to those for polydimethylsiloxane is 2.49, namely the ratio of T_G's for the two polymers, respectively. We find, however, that the ratio for the $(CED)^{1/2}$ values (Ref. 12) gives 1.18 (i.e., 8.6/7.3), in disagreement with the assumption of proportionality between E_0 and the heat of vaporization. The value of ϵ for polydimethylsiloxane should be about zero according to force-temperature coefficient measurements (Ref. 13), whereas polystyrene should have a finite ϵ, again in disagreement with the T_G ratios (see, however, Ref. 14, p. 1364, for discussion of ϵ for polystyrene).

In conclusion, we have derived an expression for the Simha–Boyer parameter K_1 using the Gibbs–DiMarzio theory of the glass transition. This expression is a function only of the fractional unoccupied volume at T_2 $(K_1 \approx -V_0 \ln V_0)$. The analysis of the theory shows that T_2 is proportional to the chain stiffness parameter ϵ, and is zero for a chain of zero stiffness. Wide variations in the intermolecular energy parameter E_0 effect relatively minor changes in T_2, and hence T_G. But for values of E_0 that are large relative to ϵ the theory predicts the absence of the discontinuity in the volume expansivities, which is normally observed at T_G. Yet polydimethylsiloxane, a polymer of zero chain stiffness, exhibits the usual change in expansivities about a T_G which is considerably above absolute zero. It is apparent that the theoretical deductions are inconsistent with observations for cases where intermolecular forces are significant in comparison to chain stiffness.

D. Structure and Electrical Properties of Poly-9-Vinylanthracene

A. Rembaum and A. Henry

1. Introduction

In 1962 Inoue et al. (Ref. 15) reported in Japan that poly-9-vinylanthracene (PVAn) forms charge transfer complexes with iodine, at room temperature, which exhibit low electrical resistivity. Simultaneous investigations of the same system in our laboratory have established that PVAn is characterized by a new type of structure substantially different from the one assumed by the Japanese workers. Also, the electrical properties of the PVAn-iodine complexes can be understood in the light of our results, which were recently confirmed by Michel and Baker (Ref. 16). Furthermore we have found that the PVAn-iodine complexes prepared at 100 to 200°C showed a higher concentration of free spins and considerably lower resistivity than those formed at room temperature. These experimental facts were attributed (Ref. 17) to a hydrogen abstraction reaction taking place mainly at elevated temperatures and leading to a conjugated double bond system, which remains complexed with iodine. In order to confirm this hypothesis the dehydrogenation of the polymer by means of sulfur and selenium was studied and the electrical properties of the dehydrogenated products were examined.

2. Experimental Technique and Results

a. Dehydrogenation. The dehydrogenation runs were carried out in an evacuated system using 50-50 mixtures of PVAn and sulfur or PVAn and amorphous selenium. The apparatus consisted of two glass ampoules connected via a ground glass joint and vacuum stopcock. The mixture was contained in one of the ampoules, the other serving as a liquid nitrogen trap. The whole system was evacuated and the same brought to the desired temperature by means of a small, electrically heated sand furnace. A mercury manometer measured the buildup of H_2S or H_2Se pressure in a section of the apparatus, the volume of which was calibrated beforehand, and the pressure measurements were recorded at 25°C. Since it is possible that some degradation might take place, the purity of the evolved gas at 150 and 200°C was checked by mass spectrometry and was found to be 96% pure. The dehydrogenated solid product was dissolved in carbon disulfide, precipitated with methanol and dried in a vacuum oven at 40°C. Refluxing equal weights of PVAn and sulfur or selenium in dimethylformamide at 153°C for 2 hr yielded dark colored products similar to those

obtained by dehydrogenation with sulfur. Sublimed iodine was added at room temperature to the dehydrogenated samples dissolved in benzene. After precipitation with methanol and drying, the iodine content of the complexes was of the order of 25%.

b. Electrical measurements. Most resistivity determinations were performed as a function of pressure on dry samples by means of an apparatus previously described (Ref. 18). In order to avoid a build-up of static charge the vanadium steel anvils serving as electrodes were shorted before every measurement during the pressurization process. Since an inherent residual voltage of the order of 50 to 200 mv could be observed on samples of high resistivity, when this voltage exceeded 50 mv, the resistance was evaluated from a current versus voltage plot. In this case, the circuit consisted of dry batteries shielded by means of an aluminum box and connected in series with a Keithley 600A electrometer, the whole being securely grounded.

The Seebeck coefficient was obtained using a Cambridge System 112B Thermoelectric probe and a Keithley 610A electrometer. The power probe and the electrometer provided the temperature and voltage change respectively.

c. Physical and chemical analysis. Absorption spectra were determined on a Cary Model 14 and a 421 Perkin-Elmer spectrophotometer. Elemental analysis was carried out by the Elek Analytical Laboratory and the number-average molecular weight was estimated by means of a Mechrolab osmometer.

d. Results. The rate of gas evolution during the dehydrogenation experiments is shown in Fig. 7, in which the amount of hydrogen sulfide is plotted against time of heating. The absorptions in the IR, UV, and visible regions of the spectrum are compared in Figs. 8, 9, and 10, respectively. The resistivity of PVAn dehydrogenated in the presence of sulfur or selenium was independent of pressure and is recorded in Table 2. The resistivity of sulfur-dehydrogenated PVAn complexed with iodine at room temperature was found to be of the same order of magnitude as of PVAn reacted directly with iodine at 200°C (compare Table 3 with Fig. 11). The resistivity decreased smoothly and reversibly as the pressure was increased.

Fig. 11 shows a plot of voltage versus current across a PVAn-iodine complex. The slope of the curve changes

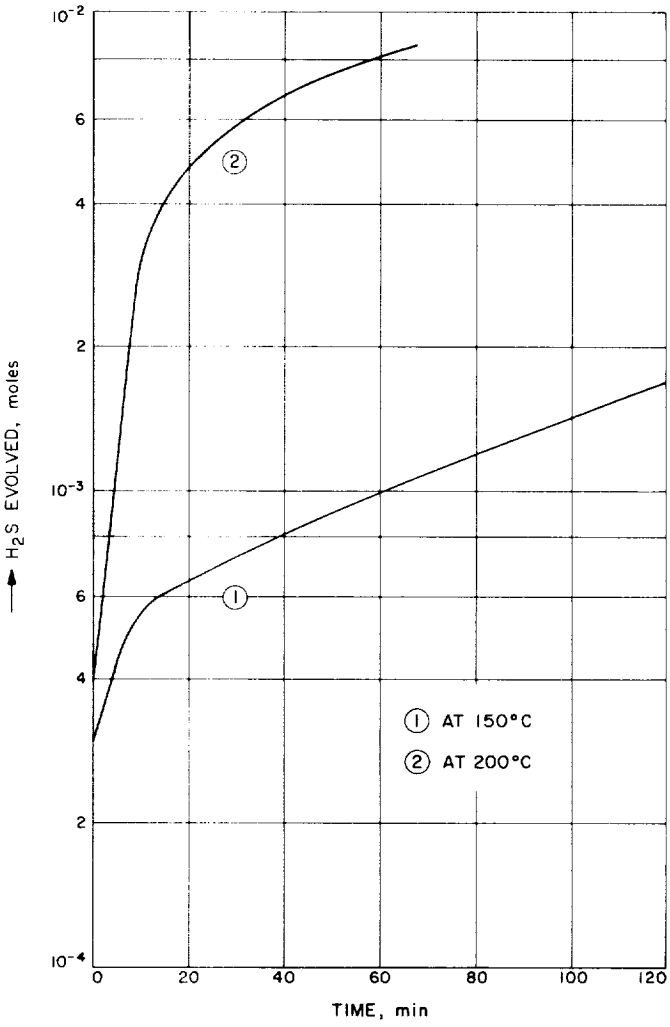


Fig. 7. Rate of dehydrogenation of PVAn (2.5×10^{-3} moles) by means of sulfur

quite markedly, the sample being non-ohmic, since the resistivity drops by a factor of 6 as the voltage is increased. This effect is not predominant in the pressure versus resistivity runs. As the potential never exceeded the value of 1 v, these measurements were well within the ohmic region of the material. The increase of voltage was extended to the point where the sample underwent an irreversible change at about a field strength of 5000 v/cm, yielding a material of very low resistivity (130 ohm-cm).

The Seebeck coefficient determined on a pressed pellet of PVAn-iodine complex was found to be of the order of 4.0 mv/°C. This is significantly larger than the value observed for some commercial inorganic semiconductors, e.g., doped silicon, 0.408 mv/°C. From this value we

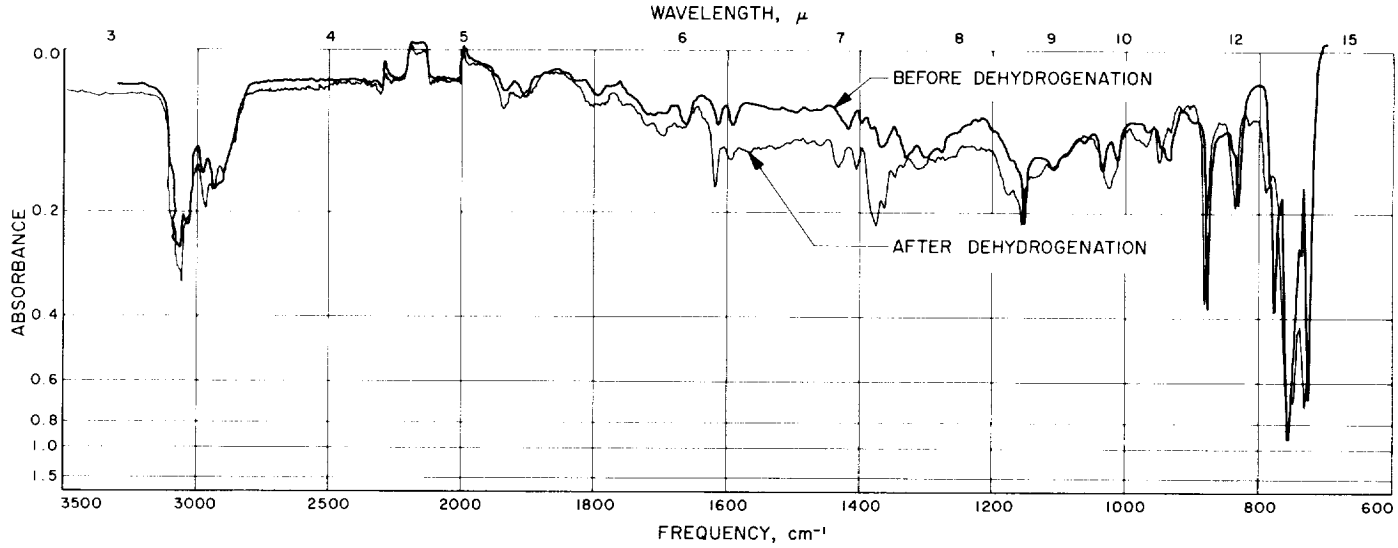


Fig. 8. IR spectrum of PVAn in CS_2 before and after dehydrogenation at 200°C

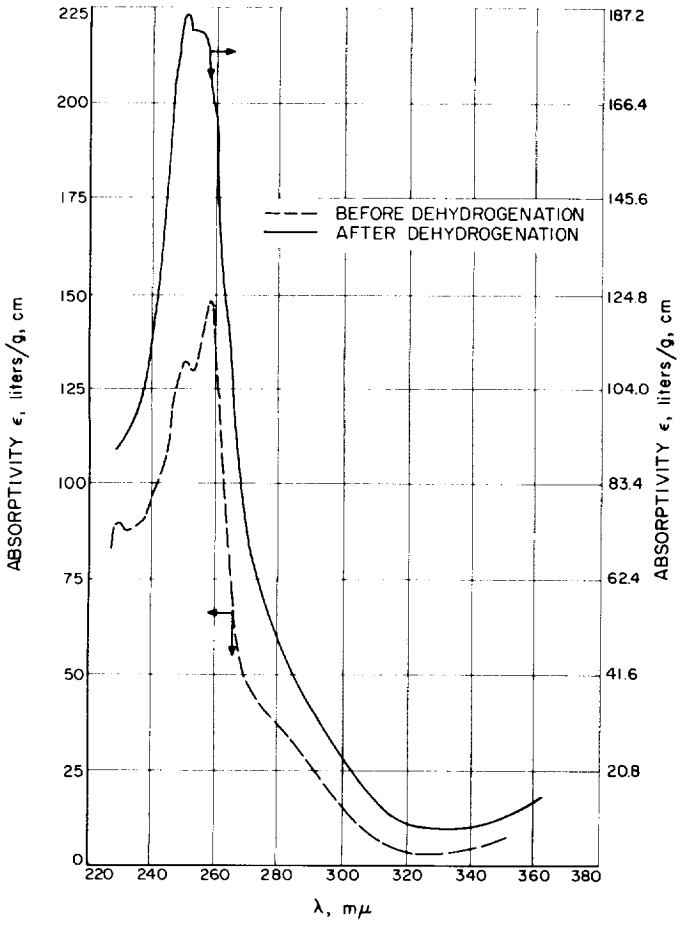


Fig. 9. UV absorption spectrum of PVAn

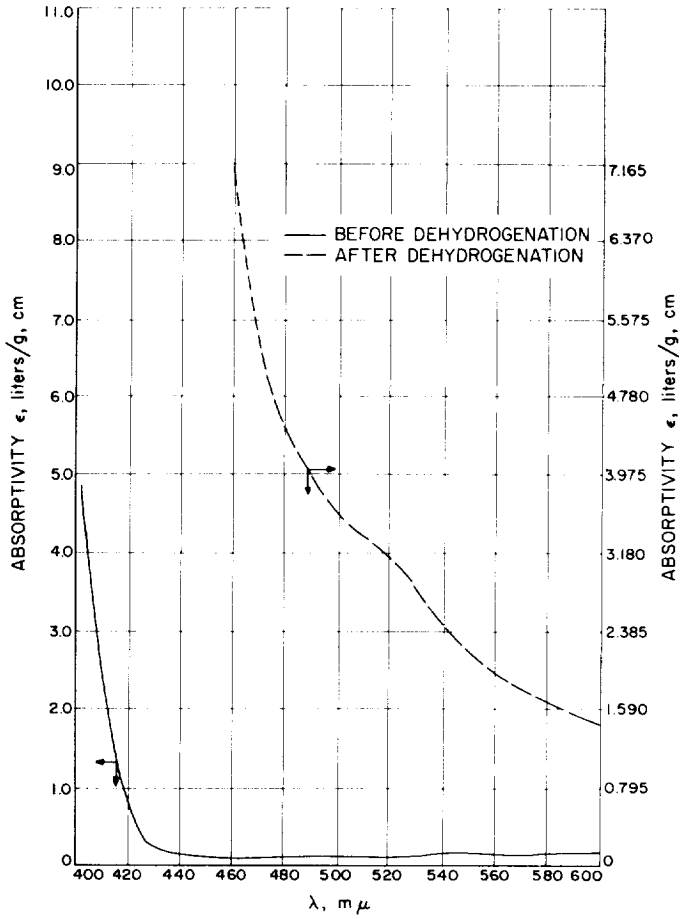


Fig. 10. Absorption spectrum of PVAn in the visible range

Table 2. Electrical properties of dehydrogenated PVAn

Dehydrogenating agent	Reaction temperature, °C	Time, hr	Color	Number-average molecular weight $M\bar{n}$	Calculated moles H_2S or H_2Se/mole of PVAn	Resistivity,[a] ρ(ohm-cm) $\times 10^{-10}$
S	200 (vac)	2	Brown	1000	2.87	5.3
S	150 (vac)	2	Brown	1200	0.64	7.3
Se	220 (vac)	2	Brown	1500	0.32	3.8
S	153 (in dimethyl formamide)	2	Brown	1000	—	2.4
Se	153 (in dimethyl formamide)	2	Tan	1400	—	1.5
S	128 (vac)	2	Yellow	1420	—	6.6

[a] The resistivity of unreacted PVAn is $> 10^{15}$ ohm-cm and its $M\bar{n} \cong 1800$.
The resistivity of PVAn complexed with iodine at room temperature was found to be of the order of 10^{11} ohm-cm.

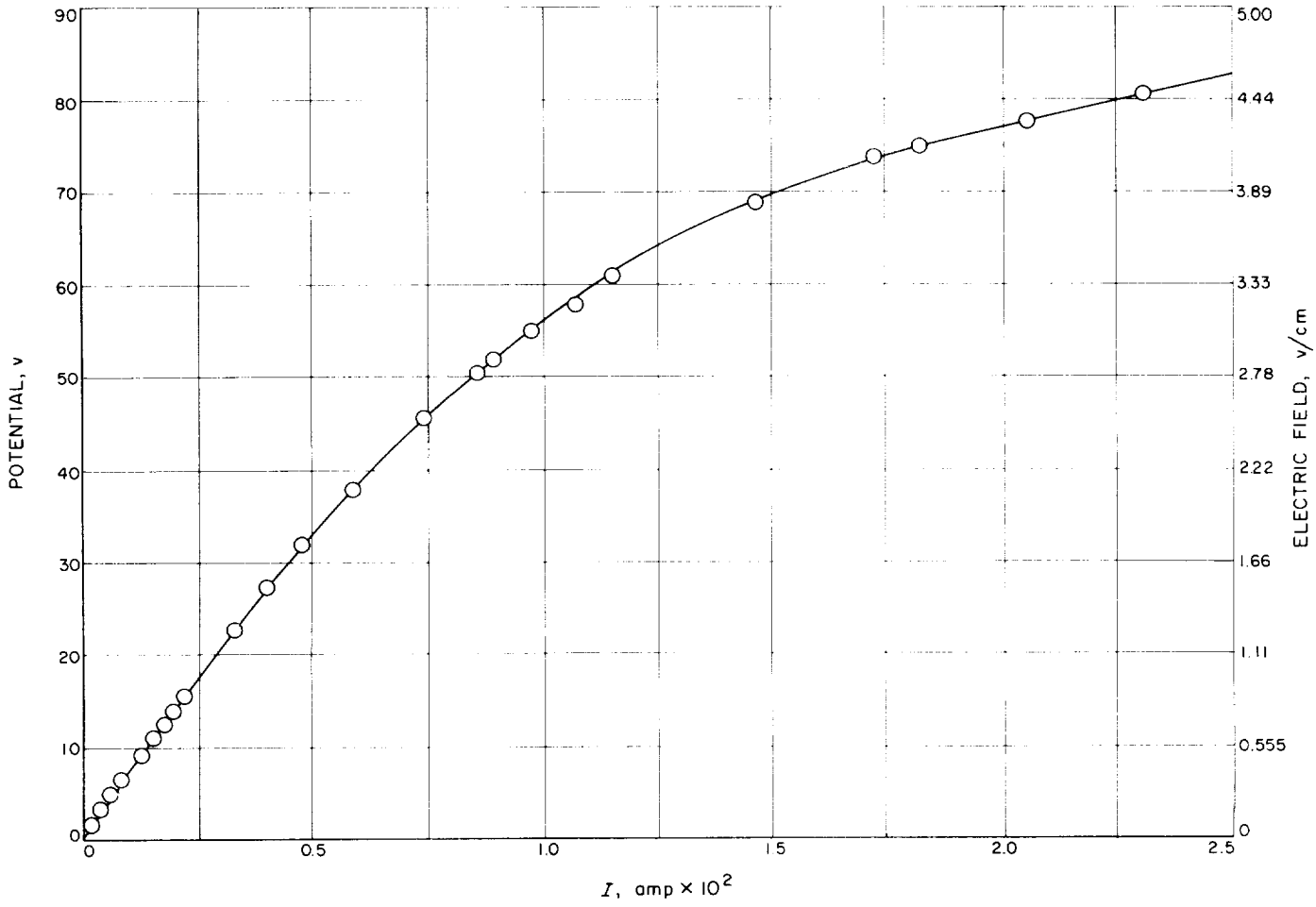


Fig. 11. Voltage versus current for a PVAn-iodine complex

Table 3. Resistivity of PVAn heated with sulfur and subsequently complexed with iodine at room temperature

Dehydrogenation temperature, °C	Iodine content, %	Resistivity at 14000 atm
150	25	1.85×10^7
200	28	1.3×10^7

calculate the figure of merit Z defined by

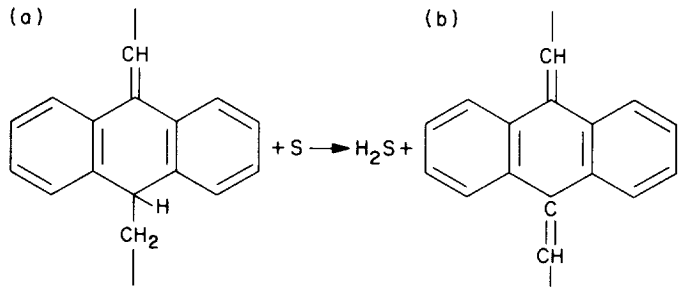
$$Z = S^2 \sigma / K \quad (1)$$

where S is the Seebeck coefficient, and σ and K the electrical and thermal conductivity, respectively. Thus Z of the PVAn-iodine complex is found to be of the order of 2×10^{-8} (with $K = 10^{-3}$ w $cm^{-1}deg^{-1}$), i.e., a factor of 20 better than Z for doped silicon (Ref. 19).

3. Conclusions

On the basis of these results we may conclude that PVAn undergoes structural changes during the dehydrogenation process which yields a product of lower resistivity than the starting material.

These changes became evident on examination of the absorption spectra in the IR and UV range. Fig. 8 shows that a PVAn sample dehydrogenated by means of S at 200°C exhibits new absorption peaks at 3.5, 6.18, 7.12 and 7.28 μ, all of which may be attributed to formation of (b) from (a).



This is also consistent with the UV spectrum and the visual observation of color. The amount of H_2S evolved at 200°C is greater than stochiometry requires (Fig. 7 and Table 2). The excess of gas may be due to the occurrence of some degradation, and the decrease of the molecular weight (Table 2) is consistent with this conclusion. There is also a possibility of formation of allene structures contributing to the formation of H_2S in greater amounts than theory predicts, since an allene compound was recently synthesized from materials similar to PVAn (Ref. 20).

The results recorded in Table 2 show that similar restrictivity values are obtained irrespective of the amount of hydrogen sulfide evolved and color of the sample. Although there is little doubt that dehydrogenation decreases the resistivity considerably (Table 3) the extent of dehydrogenation seems to have little effect on the resistivity values and further study is required to elucidate this finding.

References

1. Kalfayan, S. H., and Havlik, A. J., "Polyoxyalkylenes Terminated With Different Functional Groups: Rate Studies of Cyanoethylation and Hydrolysis of Cyanoethylated Materials," *SPS 37-17*, Vol. IV, October 20, 1962, pp. 178–180.

2. Christian, R. V., and Hixon, R. M., *Journal of the American Chemical Society*, Vol. 70, p. 1333, 1948.

3. Kalfayan, S. H., "Kinetics and Mechanism of the Cyanoethylation of 1- and 2-Propanols," *SPS 37-27*, Vol. IV, June 30, 1964, pp. 61–63.

References (Cont'd)

4. Gibbs, J. H., and DiMarzio, E. A., *Journal of Chemical Physics,* Vol. 28, pp. 373–383, 1958.

5. DiMarzio, E. A., and Gibbs, J. H., *Journal of Chemical Physics,* Vol. 28, pp 807–813, 1958.

6. Gibbs, J. H., "Nature of the Glass Transition and the Vitreous State," in *Modern Aspects of the Vitreous State,* Vol. 1, J. D. Mackenzie, Editor, Butterworths, London, 1960, p. 152.

7. Gibbs, J. H., "The Glass Transition. Thermodynamic and Statistical Considerations," in *Unsolved Problems in Polymer Science,* National Academy of Sciences, National Research Council, Publication 995, Washington, D. C., 1962, p. 107.

8. Simha, R., and Boyer, R. F., *Journal of Chemical Physics,* Vol. 37, pp. 1003–1007, 1962.

9. Fox, Jr., T. G., and Flory, P. J., *Journal of Applied Physics,* Vol. 21, p. 581, 1950.

10. Williams, M. L., Landel, R. F., and Ferry, J. D., *Journal American Chemical Society,* Vol. 77, p. 3701, 1955.

11. Hirai, N., and Eyring, H. J., *Journal of Applied Physics,* Vol. 29, pp. 810–816, 1958.

12. Tobolsky, A. V., *Properties and Structures of Polymers,* John Wiley and Sons, Inc., New York, N.Y., 1960, p. 64.

13. Flory, P. J., Hoeve, C. A. J., and Cifferi, A., *Journal of Polymer Science,* Vol. 34, p. 337, 1959.

14. Boyer, R. F., *Rubber Reviews,* Vol. 36, pp. 1303–1421, 1963.

15. Inoue, H., Noda, K., Takiuchi, S., and Imoto, E., *Journal of the Chemical Society* (Japan), Vol. 65, p. 128, 1962.

16. Michel, R. H., and Baker, W. P., *Journal of Polymer Science,* Part B, Vol. 2, p. 163, 1964.

17. Rembaum, A., Henry, A., Waits, H., "Anionic Polymerization of 9-Vinylanthracene and Semiconductivity of Poly-9-Vinylanthracene-Iodine Complexes," *American Chemical Society, Division of Polymer Chemistry Preprints,* p. 109, New York Meeting, September 1963.

18. Rembaum, A., Moacanin, J., Cuddihy, E., *Research Summary,* Vol. 36-14, Jet Propulsion Laboratory, Pasadena, California, May 1, 1963, p. 152.

19. Hannay, N. B., *Semiconductors,* pp. 313–388, Reinhold Publishing Corp., New York, N.Y., 1959.

20. Gustav, C., "Reactions of Exo-Cyclic Vinyl Halides With Phenyl Lithium," Thesis, University of Illinois, 1963.

Errata

The following corrections should be noted for *SPS 37-27*, Vol. IV:

(1) Page 60, the 5th line following Eq. (1) should read:
defined by $d\tau = k_p A\, dt$. The condition for first order

(2) Page 60, Eq. (4) should read:

$$(P_N - 4)^{-1}$$

(3) Page 59, Equations under *Propagation* should read:

$$\bar{A}_j^- + A \xrightarrow{kp} \bar{A}_j^- + 1$$

$$A_j^- + A \xrightarrow{kp} A_j^- + 1$$

XI. Propulsion Research

A. Ignition and Combustion in Nitrogen

R. A. Rhein

1. Introduction

Nitrogen is regarded as a major component in the atmosphere of Mars and a minor but important constituent of the Venusian atmosphere. In order to utilize these atmospheres for chemical propulsion, a series of experiments were conducted to find the ignition temperature in nitrogen of various powdered metals which might be used as possible fuels.

2. Experimental

Attached to a conventional high-vacuum system is the nitrogen ignition apparatus, Fig. 1, and the sample holder, Fig. 2. The ignition apparatus and sample holder are evacuated. The sample holder crucible is filled with the powdered metal, lowered below the thermocouple, and then raised until the thermocouple is in the mass of powdered metal. The system is filled with nitrogen

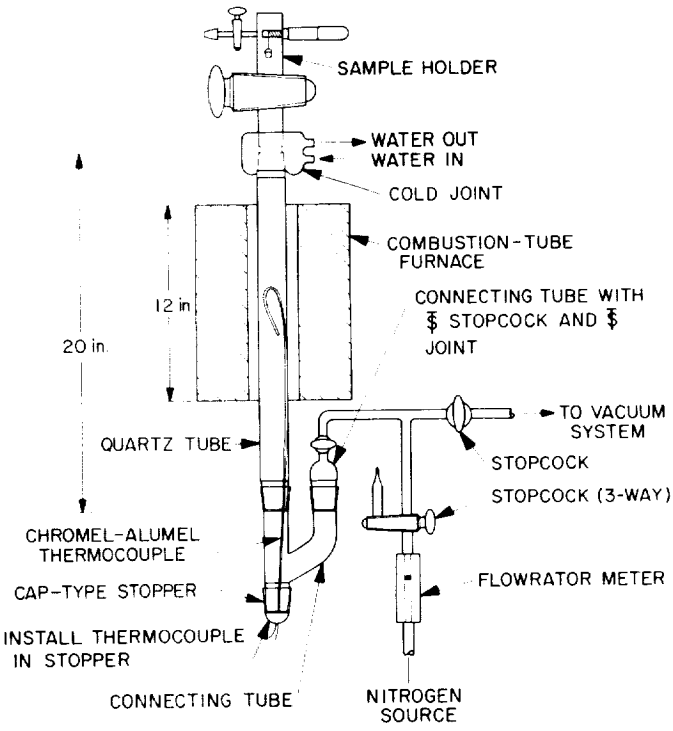


Fig. 1. Nitrogen ignition apparatus

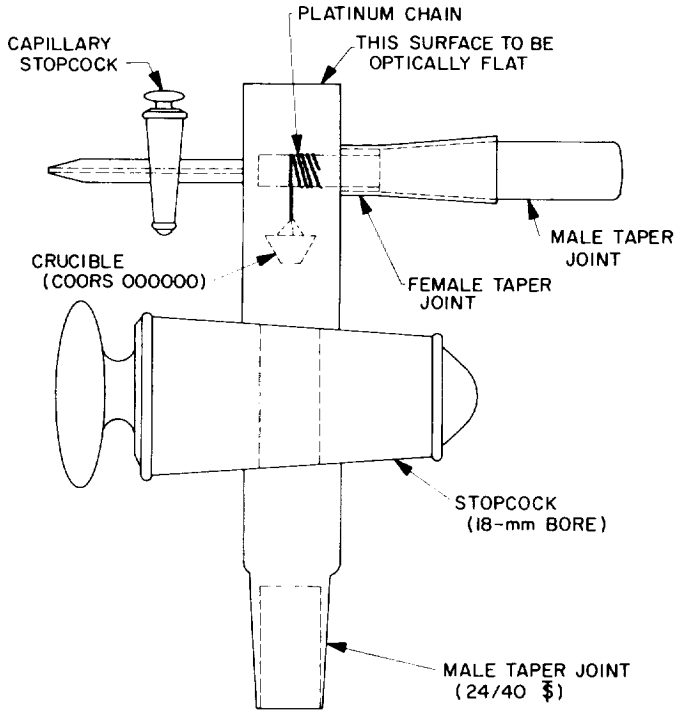


Fig. 2. Sample holder

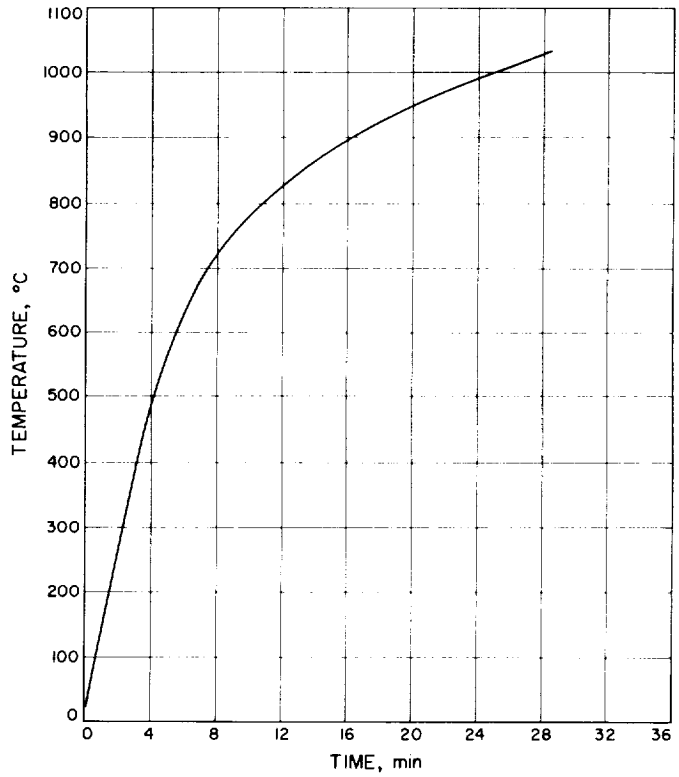


Fig. 3. Furnace temperature as a function of the time interval following switching on

(Linde Co., extra-dry high-purity grade, 99.995% pure, point $< -73°C$, $O_2 < 30$ ppm) and then flow rate of 100 ml/min is established. The furnace is switched on, and ignition is observed both visually and by an essentially discontinuous temperature increase registered by the thermocouple. A typical relationship between the thermocouple temperature in the furnace and the time interval is shown on Fig. 3.

3. Results

a. Lithium. The ignition temperature of lithium in nitrogen has been variously reported at 170°C (Ref. 1), 450°C (Ref. 2), and at dull red heat (Ref. 3). The literature reports for small pieces or chunks of Li. Here, powdered lithium (Foote Mineral Co., New Johnsonville Operations, $\leq$ 100-μ particle size, Lot 401-03) was found in two experiments to ignite at 410 and 388°C, and was seen to burn vigorously.

b. Beryllium. Beryllium powder is reported to react at moderate speed with N_2 at 1100°C to form Be_3N_2. However, it does not react more readily at 1300°C (Refs. 4, 5). Here, when a very finely divided Be powder (Ultra Fine Beryllium Powder; National Research Corp., <0.1-μ particle size) was heated with nitrogen in the reaction tube shown in Fig. 4, ignition occurred at 504 and 527°C in two experiments.

c. Magnesium. The literature is somewhat contradictory regarding the reaction of magnesium in nitrogen. Magnesium powder (100% 100 mesh, 80% 270 mesh) ignited at 530°C (Refs. 6, 7) and magnesium reacted readily when heated (Refs. 7, 8). Elsewhere it is reported that there is no reaction under 600°C (Ref. 9), the reaction begins at 670°C (Ref. 10), and that a steady reaction occurs in the region 700 to 800°C.

Here magnesium powder (Reade Mfg. Co., Inc., (−325 mesh, rated 99.9% pure) was heated to 954°C and there was no indication of ignition. There was a yellowish powder present after cooling; however, this is presumably Mg_3N_2. Some magnesium powder was heated rapidly to 1071°C in the apparatus in Fig. 4, and no ignition in nitrogen occurred. However, the yellow material, presumably Mg_3N_2, was noted after the tube cooled down.

d. Calcium. Ignition of Ca in N_2 occurs at dull red heat (Refs. 11–14). Certain alloys of Ca can react faster

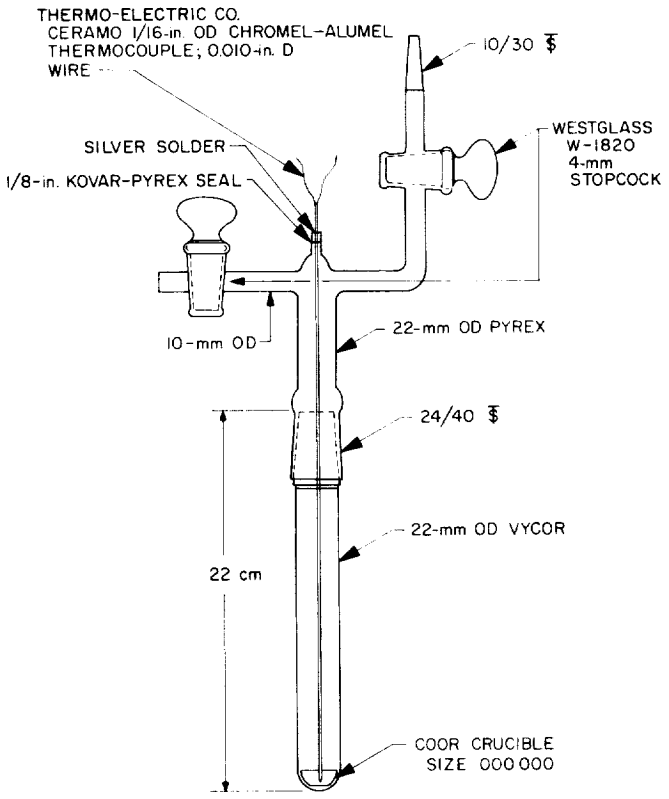


Fig. 4. Reaction tube

or slower with N_2 (Ref. 15). Here, −325 mesh calcium powder (99.9% pure, from the Research Chemicals Div. of Nuclear Corp. of America) was found in two experiments to ignite at 327 and 360°C and observed to burn vigorously. After cooling and treating the solid product with water, an NH_3 odor was observed, indicating the presence of Ca_3N_2 in the solid.

e. Boron. Amorphous boron does not react with N_2 at 900°C; the reaction begins at 1230°C (Refs. 16–19). Here, ultrafine boron powder (National Research Corp., 0.02 to 0.06-μ particle size) was heated in N_2 to 893°C, and there was no evidence of ignition. A chemical analysis indicated 0.8% N in the product.

f. Aluminum. In the literature, it was reported that powdered Al ignites at 720°C (Ref. 20), or 820°C (Refs. 21, 22). Others state that the reaction is vigorous, but not self-sustaining at 700–750°C (Ref. 6), a reaction occurs above 800°C (Refs. 23, 24), at 850–875°C, and that the best temperature for AlN preparation is at 900°C (Ref. 25). It was here found that neither powdered Al (Reynolds Aluminum 1-131 Atomized Powder, 99.3% pure, average particle size 8 to 9 μ) nor ultrafine aluminum powder (National Research Corp. 0.03-μ average particle size, 93% pure, with oxide as the impurity) ignited in N_2 up to 1080°C.

g. Cerium. It is found in the literature that cerium ignited and burned in nitrogen at 780°C (Ref. 26) and cerium wire will ignite in nitrogen at 850°C (Ref. 27). Here, cerium powder (VARLACOID Chemical Co., New York; order BH4-288601: −325 mesh powder under kerosene) after having its kerosene removed by extraction with hexane, was seen to ignite at 216°C in N_2 to produce intense combustion. Treatment of the solid product with water after it had cooled down produced NH_3, indicating the nitride was, indeed, formed.

h. Mischmetall. Here, powdered Mischmetall (VARLACOID Chemical Co., New York, order BH4-288601: −325 mesh, packed under kerosene), was treated with hexane to remove the kerosene it was packed in, and was found to ignite in nitrogen at 209°C.

i. Titanium. From the literature, molten titanium burns in nitrogen (Ref. 28); highly divided titanium burst into flame in nitrogen at 800°C (Refs. 29, 30), and 10.5-μ powder ignited in commercial N_2 at 760°C (Ref. 31). Here, powdered Ti (A. D. Mackay Co., New York, 1 to 5-μ particle size) ignited in nitrogen at 830°C when the initial nitrogen temperature was 540°C. It was found to be necessary to begin the experiment with initially hot nitrogen because when the experiment was begun with the nitrogen at ambient temperature, there was no ignition observed.

j. Zirconium. The ignition temperature for powdered Zr in N_2 is reported from the literature as 530°C for −325 mesh (Ref. 6), 790°C for 3.3-μ sized particles (Ref. 31), and no ignition at 820°C for 17.9-μ sized particles (Ref. 31). Here, for 3-μ Zr powder (Charles Hardy, Inc., zirconium powder 120-A grade, Lot 103-2, 94–95% pure, oxide impurity, order BH4-288629) no ignition was observed in the tube furnace apparatus on heating to 970°C, but in the reaction tube (Fig. 4) vigorous ignition was seen at 490 and 525°C in two experiments. In each case, it took approximately 2 min to heat the zirconium from room temperature to ignition temperature.

k. Thorium. Reportedly, 7.2-μ sized thorium powder ignited at 500°C in commercial nitrogen (Ref. 31). Here, −325 mesh Th powder (Charles Hardy, Inc., order BH4-288629) was seen to ignite at 620°C.

l. Uranium. From the literature it was found that 10.8-μ size uranium powder ignited in commercial nitrogen at 410°C (Ref. 31). Here, −200 mesh U^{238} (depleted uranium) coated with 2% Viton (The Great Southern Mfg. and Sales Co.) was found to ignite in nitrogen in two experiments at 354°C in the reaction tube and 360°C in the tube furnace apparatus.

m. Chromium. Although it was reported that pyrophoric chromium, prepared by distilling its amalgam, ignites in N_2 when warmed (Refs. 28, 32), other references report a slow reaction on heating (Refs. 14, 16, 33). Here, −325 mesh Cr powder, 99.85% pure (VARLACOID Chemical Co., New York, order BH4-288601) was heated to 1170°C in the reaction tube, Fig. 4, and there was no evidence of ignition.

n. Manganese. Although it was reported that finely divided Mn reacted with N_2 when heated (Refs. 14, 17, 33, 34), there was no mention of ignition. Here, −325 mesh Mn powder, 99% pure (Charles Hardy, Inc., order BH4-288629) was heated to 1316°C in the reaction tube, and there was no indication of ignition. The results are summarized in Table 1.

Table 1. Ignition temperatures in nitrogen of powdered metals

Metal	Condition	Ignition temperature, °C
Lithium	100 μ	388, 410
Beryllium	0.1 μ	504, 527
Magnesium	−325 mesh	No ignition to 1071[a]
Calcium	−325 mesh	327, 360
Boron	0.02 to 0.06 μ	No ignition to 893
Aluminum	8–9 μ size	No ignition to 1080
	0.03 μ size	
Cerium	−325 mesh	216
Mischmetall	−325 mesh	209
Titanium	1–5-μ size	830
Zirconium	3-μ size	490, 525[a]
Thorium	−325 mesh	620
Uranium	−200 mesh, Viton coated	360
Chromium	−325 mesh	No ignition to 1170[a]
Manganese	−325 mesh	No ignition to 1316[a]

[a]Heated in the reaction tube.

References

1. Markowitz, M. M., and Boryta, D. A., "Lithium Metal—Gas Reactions," *Journal of Chemical and Engineering Data*, Vol. 7, October 1962, pp. 586–591.
2. Dafert, F. W., and Miklauz, R., "New Compounds of Nitrogen and Hydrogen With Lithium," *Monatshefte für Chemie und verwandte Teile anderer Wissenschaften*, Vol. 31, 1910, pp. 981–996.
3. Ouvrard, L., "On a Nitride of Lithium," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences*, Vol. 114, 1892, p. 120.
4. Fichter, F., and Brunner, E., "Beryllium Nitride," *Zeitschrift für anorganische und allgemeine Chemie*, Vol. 93, 1915, p. 84.
5. von Antropoff, A., and Kruger, K. H., "Kinetics of the Reaction of the Alkaline Earth Metals With Nitrogen and Other Gases. II, III, IV," *Zeitschrift für physikalische Chemie*, Vol. A167, 1933, pp. 49–70.
6. Hartmann, I., Nagy, J., and Brown, H. R., "Inflammability and Explosibility of Metal Powders," Report on Investigations, U.S. Bureau of Mines, No. 3722, 1943.
7. Weintraub, A. A., "Flaming Metal," *Safety Engineering*, Vol. 99, No. 3, 1950, pp. 34–35, 47–48.

References (Cont'd)

8. Merz, V., "Remarks on Magnesium," *Berichte der deutschen chemischen Gesellschaft,* Vol. 24, p. 3940.

9. Remy-Gennete, P., "Researches on the Alkaline Earth Hydrides," *Annales de Chimie,* Vol. 19, 1933, pp. 263–426.

10. Matignon, C., and Lassieur, "Action of Nitrogen and Oxygen on Magnesium," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences,* Vol. 154, 1912, pp. 63–65.

11. Moissan, H., "Preparation and Properties of Calcium Nitride," *Bulletin de la Société Chimique de France,* Series 3, Vol. 21, 1899, p. 881.

12. Moissan, H., "Preparation and Properties of Calcium Nitride," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences,* Vol. 127, 1898, p. 498.

13. Maquenne, L., "On the Fixation of Nitrogen by the Alkaline-Earth Metals," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences,* Vol. 121, 1895, p. 1147.

14. Tammaun, G., "Velocity of Formation of Some Nitrides," *Zeitschrift für anorganische und allgemeine Chemie,* Vol. 124, 1922, pp. 25–35.

15. Ruff, O., and Lauterbach, H., "The Absorption of Nitrogen by Calcium and Its Alloys," *Zeitschrift für anorganische und allgemeine Chemie,* Vol. 121, 1922, p. 167.

16. Mellor, J. W., "A Comprehensive Treatise on Inorganic and Theoretical Chemistry," Longmans, Green and Co., Ltd., London, 1953, pp. 75–144.

17. Moissan, H., "Studies of the Properties of Amorphous Boron," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences,* Vol. 114, 1892, p. 617.

18. Moissan, H., "Studies of Amorphous Boron," *Annales de Chimie et de Physique,* Series 7, Vol. 6, 1895, pp. 296, 313.

19. Rizzo, H. F., "Oxidation of Boron at Temperatures Between 400° and 1300° in Air," *Boron, Synthesis, Structure, and Properties,* Plenum Press, Inc., New York N.Y., 1960, pp. 175–189.

20. Fichter, F., "Aluminum Nitride," *Zeitschrift für anorganische und allgemaine Chemie,* Vol. 54, 1907, p. 322.

21. Neumann, B., and Kroger, C., and Haebler, H., "The Heat of Formation of Nitrides. II. The Heat of Nitriding of Lithium, Aluminum, Beryllium, and Magnesium," *Zeitschrift für anorganische und allgemeine Chemie,* Vol. 204, 1932, pp. 81–96.

22. Wolf, J., "Synthesis of Aluminum Nitride From the Elements," *Zeitschrift für anorganische und allgemeine Chemie,* Vol. 83, 1913, pp. 159–162.

23. Russ, F., "An Experiment Demonstrating the Formation of Aluminum Nitride," *Zeitschrift für Elektrochemie,* Vol. 19, 1919, pp. 923–925.

24. Sieverts, A., and Krumbhaar, W., "Solubility of Gases in Metals," *Berichte der deutschen chemischen Gesellschaft,* Vol. 43, 1910, pp. 893–900.

25. Marre, F., "The Fixation of Atmospheric Nitrogen by the Use of Aluminum Nitride," *Genie Civil,* Vol. 61, 1912, pp. 30–33.

References (Cont'd)

26. Dafert, F. W., and Miklauz, R., "Compounds of Cerium With Nitrogen and Hydrogen," *Monatschefte für Chemie und verwandte Teile anderer Wissenschaften,* Vol. 33, 1912, p. 911.

27. Muthmann, W., and Kraft, K., "Investigations of Cerium and Lanthanum," *Liebigs Annalen der Chemie,* Vol. 325, 1902, pp. 261–278.

28. Hardtung, H., "Beitrage zur Kenntnis der Metallnitride," Hanover, 1912.

29. Weiss, L., and Kaiser, H., "Über Metallisches Titan," *Zeitschrift für anorganische und allgemeine Chemie,* Vol. 64, 1910, p. 352.

30. Barksdale, Jelks, *Titanium, Its Occurrence, Chemistry, and Technology,* Ronald Press, Inc., New York, N.Y., 1949, p. 49.

31. Hartmann, I., Nagy, J., and Jacobsen, M., "Explosive Characteristics of Ti, Zr, Th, U, and Their Hydrides," Reports on Investigations, U.S. Bureau of Mines, No. 4835, 1951.

32. Valensi, G., "The Dissociation of Chromium Azide," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences,* Vol. 187, 1928, pp. 293–296.

33. Neumann, B., Kroger, C., and Haebler, H., "The Heat of Formation of the Nitrides. I. Manganese Nitride and Chromium Nitride," *Zeitschrift für anorganische und allgemeine Chemie,* Vol. 196, 1931, pp. 65–78.

34. Valensi, G., "The Action of Nitrogen on Manganese," *Comptes Rendus Hebdomadaires des Séances de l'Academie des Sciences,* Vol. 187, 1928, pp. 376–378.

XII. Liquid Propulsion

A. Advanced Liquid Propulsion Systems

R. N. Porter, B. H. Johnson, W. H. Tyler, and H. B. Stanford

1. Introduction, R. N. Porter

The Advanced Liquid Propulsion Systems (ALPS) program is investigating selected problems generated by spacecraft operational requirements for propulsion systems capable of high inherent reliability, long-term storage in space, multiple start in free fall (zero gravity), and engine throttling. The solutions proposed to satisfy these requirements have been incorporated in the ALPS system.

Recent accomplishments include the testing of several additives to hydrazine and some individual injector elements. A 315-sec firing test of a boron/pyrolytic thrust chamber has been made. Some data on the long-term permeability of Teflon bladders has been obtained.

2. Injector Development, B. H. Johnson

a. An investigation of reaction inhibitors. The nitrogen tetroxide-hydrazine (N_2O_4-N_2H_4) propellant combination exhibits an extremely rapid liquid phase reaction rate. This factor requires the utilization of special techniques to accomplish liquid phase mixing. In conjunction with the attempts to physically increase the mixing of a doublet element (*SPS 37-22, 24,* and *28,* Vol. IV), a program was contracted to the Dynamic Science Corporation to find a chemical inhibitor for the N_2O_4-N_2H_4 reaction (Ref. 1). Such an inhibitor presumably would allow mixing to be accomplished before sufficient energy could be released to disrupt the impingement process. The ignition delay time was chosen as a convenient parameter to be measured in testing a variety of chemicals for their effectiveness as inhibitors.

The apparatus used to measure ignition delay was quite similar to that developed by Kilpatrick and Baker (Ref. 2). It consists of a closed bomb of about 450 cm^3 volume into which the propellants are injected as high-velocity jets. A high-response pressure transducer (flush-mounted Kistler) and a photocell were used to detect ignition. The ignition delay times were recorded by photographing the screen of an oscilloscope. In operation, a fast-opening solenoid valve (90% open in 2 msec) allows 1500 psig nitrogen gas to act against a large driving piston. This piston first actuates a trigger mechanism for the oscilloscope and then moves the smaller fuel and oxidizer pistons, forcing the propellants to break their respective Teflon retaining discs and enter the bomb simultaneously through

short orifices. The jets thus formed have a diameter of 0.060 in. and impinge tangentially in a swirl cup at a 90 deg angle to each other. The propellant pistons were sized so as to give a mixture ratio (oxidizer/fuel) of 1.2. Before each test the bomb was flushed with nitrogen gas to insure an inert atmosphere.

A number of chemical additives to the N_2H_4 were tested in this apparatus with the results shown in Fig. 1. The ignition delay time was increased to 4.0 msec and decreased to 1.2 msec from the 3.0 msec measured for the neat propellants. Several of these more promising additives were tested in a rocket engine with a single element doublet injector having an impingement angle of 60 deg and orifice diameters of 0.236 in. The chamber used had a characteristic length L^* of 170 in. The additives tested were, by weight, 1 and 2% ethylenediamine tetraacetic acid, 1 and 2% ethyl bromide, 1% fluorobenzene, 1% triethylborate, and 1, 2, 9.1, 13.0, and 16.7% water. None of these additives, except high percentages of water, increased c^* at all, or decreased c^* by more than 3% from the nominal efficiency of 66%. Fig. 2 shows the effect of water concentration on the performance of this injector. Peak performance measured was 76% of theoretical c^*, using 16.7% water in the fuel. Theoretical c^* is assumed to be that of N_2O_4-N_2H_4 at a mixture ratio of 1.2 (5810 ft/sec). An increase in combustion roughness was noted as the water content of the fuel was increased.

Another method tested for using water to study the combustion effects on the mixing process was to add a water orifice to the doublet element used above. The water orifice was located in the plane of the propellant orifices and midway between them. The three orifices thus impinge at a common point. The 0.080-in. D for the water orifice was chosen to satisfy, at a water flow rate of 0.48 lbm/sec, the criterion presented in Ref. 3 for optimizing mixing in a triplet element using nonreactive fluids.

Peak performance with this tripropellant element was at the design water flow rate of 0.48 lbm/sec where it reached 4150 ft/sec, or 72% of the theoretical c^* for N_2O_4-N_2H_4 at a mixture ratio of 1.2. Decreasing the water flow rate to 0.41 lbm/sec dropped performance to 71%

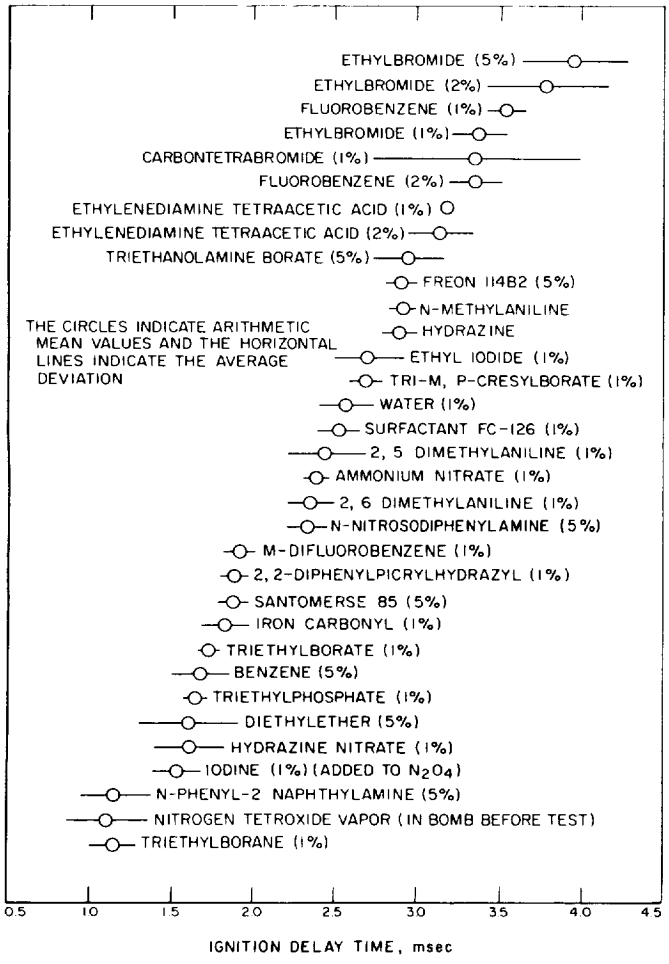


Fig. 1. Ignition delay times of N_2O_4-N_2H_4 with various fuel additives

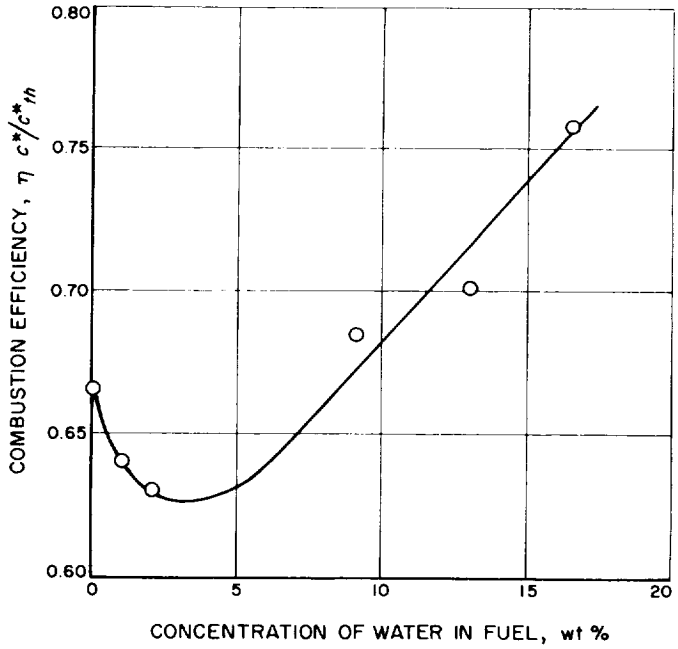


Fig. 2. Variation of combustion efficiency with water concentration in fuel

and increasing the water flow rate to 0.53 lbm/sec dropped performance to 67%.

It is encouraging that the addition of water to the fuel can significantly increase performance. This technique will not be practical, however, until an additive is found which will be more effective than water and will be required in much lower concentrations. Two different approaches to this problem are being studied at Dynamic Science Corporation under a new contract.

b. Multielement injector investigation. The primary emphasis in the ALPS injector development program has been on the development of a single-element injector at the 2000-lb thrust level using the N_2O_4-N_2H_4 propellant combination. Results from testing a variety of such elements have shown none to date which seem to offer high performance. Because of this, much of the ALPS injector effort has been redirected towards multielement injectors. The type of element chosen for study was an impinging-sheet doublet. Each sheet is formed by directing a jet against a suitable solid deflector (Fig. 3). This element

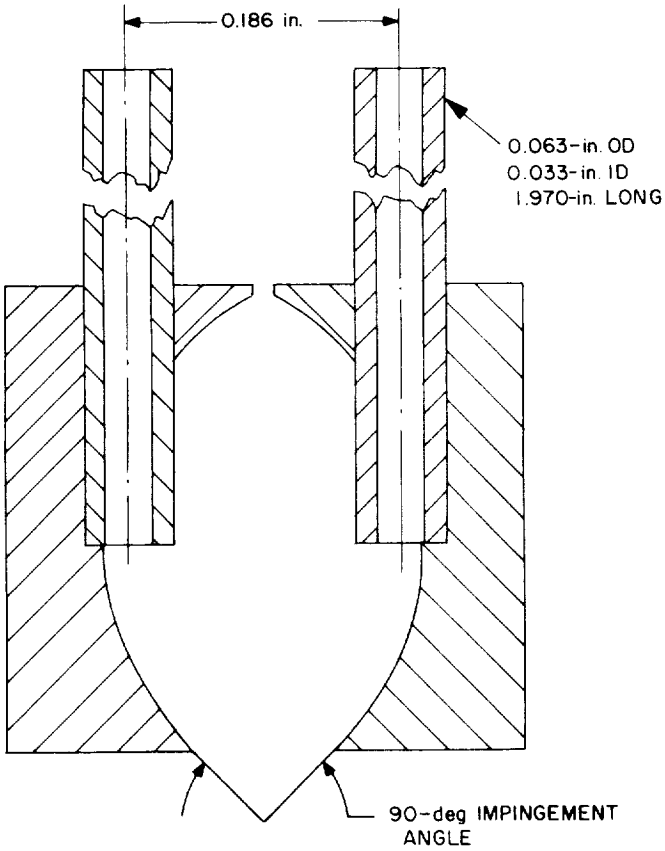


Fig. 3. Typical impinging sheets element

was chosen over impinging jets in order to avoid problems of jet misimpingement due to manufacturing tolerances and to provide a degree of film cooling to the injection face.

In order to optimize the impinging-sheets doublet element for use in multielement injectors, a single-element injector has been designed and fabricated to be tested in a small, uncooled thrust chamber at a flow rate of 0.0835 lbm/sec, a mixture ratio of 1.2, and a chamber pressure of 150 psia. This 25-lb-thrust chamber has a contraction ratio of 7.11 and an L^* of 19 in.

The injector was designed such that the orifices are parallel and axial and can be located at four discrete spacings (0.186, 0.372, 0.558, and 0.744 in.). Preliminary tests have shown that the peak performance was always measured at the closest (0.186 in.) spacing. In this position the edges of the deflectors were spaced approximately 0.08 in. apart. Several different deflector geometries have been tested and c^* rather consistently remained at about 5500 ft/sec, or 95% of theoretical (uncorrected for heat losses), with a 90-deg impingement angle and orifice diameters of 0.33 in. (L/D = 60), corresponding to an injector pressure drop, ΔP_{inj}, of 290 psi. Decreasing the impingement angle to 60 deg dropped c^* to 90% and increasing the angle to 120 deg did not seem to affect performance. Increasing the orifice diameters to 0.043 in. (L/D = 45, ΔP_{inj} = 165 psi) caused unstable combustion at a frequency of about 230 cps and performance dropped to 89%. The instability was eliminated by increasing the flow rate to about 0.100 lbm/sec (ΔP_{inj} = 220 psi).

Tests were also made with impinging jets doublet elements to get comparative data at the same operating conditions using the same thrust chamber. The injector was accurately fabricated such that the orifice centerlines impinged within 0.001 in. The peak performance (95% of theoretical c^*) using 0.033-in. D jets was at an impingement angle of 60 deg. The performance dropped to 86% at a 45-deg impingement angle and to 84% at a 90-deg impingement angle. While the preliminary tests thus indicate that peak performance was equal for these two types of elements, combustion was considerably smoother with the impinging-sheets type.

Further tests will be made to optimize the impinging-sheets element and this optimized element will then be tested in various multielement injectors.

3. *Thrust Chamber Development,* W. H. Tyler

The ALPS program has conducted successful demonstration test firings of radiation-cooled, free-standing pyrolytic graphite (PG) thrust chamber assemblies at the 100-lb and 200-lb thrust level.

Two long-duration tests have been completed with a new 100-lb thrust injector (Mod IV) and two boron/pyrolytic graphite alloy (BP) thrust chamber assemblies. One chamber assembly fractured after 310 sec (*SPS 37-28*, Vol. IV) and the other after 315 sec of continuous firing. The exact cause of these chamber failures is unknown. In each case chamber and throat erosion appeared minimal until failure with only a 4-psi decrease of chamber pressure for the 315-sec test. The nominal chamber pressure was 150 psia at a mixture ratio of 1.2. Characteristic velocity c^*, was approximately 5640 ft/sec. The new 100-lb thrust injector design (designated Mod IV) has demonstrated reduced average erosion rates compared to previous injector designs (*SPS 37-28*, Vol. IV).

During the preparation for the long-duration BP alloy chamber tests, two 30-sec checkout runs were made with a cracked PG chamber. This chamber had developed a localized cracked, delaminated area on the periphery during the first test but, since limited test hardware was available for checkout runs, it was decided to use this chamber for an additional test. Fig. 4 is a post-run photograph of the chamber. The second test was completed successfully, although some additional wall cracking occurred. These tests demonstrated encouraging evidence of durability of the pyrolytic graphite material.

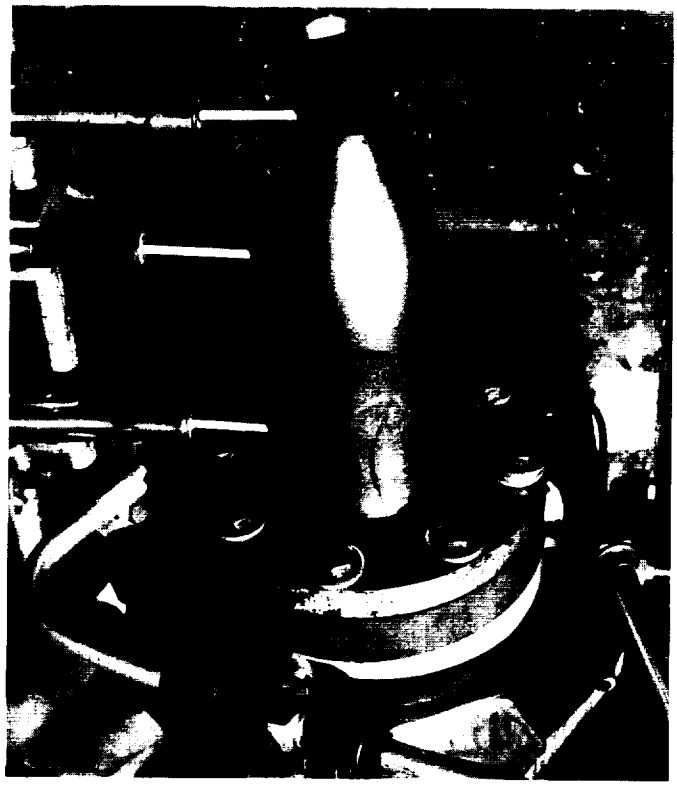


Fig. 4. Post-test view of thrust chamber showing cracks and delamination on outer periphery

4. *Bladder Development,* H. B. Stanford

A basic concept of the ALPS design specifies that the oxidizer (N_2O_4) and fuel (N_2H_4) be stored in bladders in a single tank for periods in excess of one year. Partial expulsion may occur at any time throughout the storage period. During this time, the propellant vapors which may have diffused through the bladder walls will be free to migrate throughout the ullage and to mix together. The results may not be catastrophic (*SPS 37-15*, Vol. IV).

To study the problems posed by the above-mentioned conditions a test facility was constructed at Edwards Test Station (*SPS 37-20*, Vol. IV). Here the candidate propellants may be stored in 18-in.-D hemispherical bladders placed in stainless steel tanks in combinations as follows: (1) fuel or oxidizer in one bladder with water in the other, (2) fuel in both bladders, (3) oxidizer in both bladders, and (4) fuel in one bladder and oxidizer in the other. The bladders used to date have been manufactured by Dilectrix, Inc. of Farmingdale, Long Island, N. Y. These are of current state-of-the-art seamless construction made from tetrafluoroethylene-fluorinated ethylene propylene (TFE-FEP) sprayed and sintered Teflon with a nominal wall thickness of 0.010 in. When installed in the stainless steel test tank the bladders are separated by a stainless steel partition which is a physical support but is not gas tight.

The tests are made at ambient temperature which ranges from +15 to +115°F. The test tank is maintained at an internal pressure of 250 psi with nitrogen gas and is protected from overpressurization by a 365-psi burst diaphragm. In this environment the gases permeating through either bladder are free to migrate throughout the ullage space within the tank.

The purpose of a recent test was to determine the amount of N_2O_4 that would permeate through the oxidizer bladder and into the fuel bladder (which in this case was filled with water) under the above-mentioned

conditions. The amount of N_2O_4 crossover from oxidizer to fuel bladder was measured by remotely monitoring the *p*H factor of the water in the fuel bladder and by analyzing samples of the water for NO_2 by titration.

After 12 weeks the indication of NO_2 accrual in the water bladder, based on *p*H reading, had dropped below 1; readings were discontinued at that time due to the decreased significance of the readings below that point.

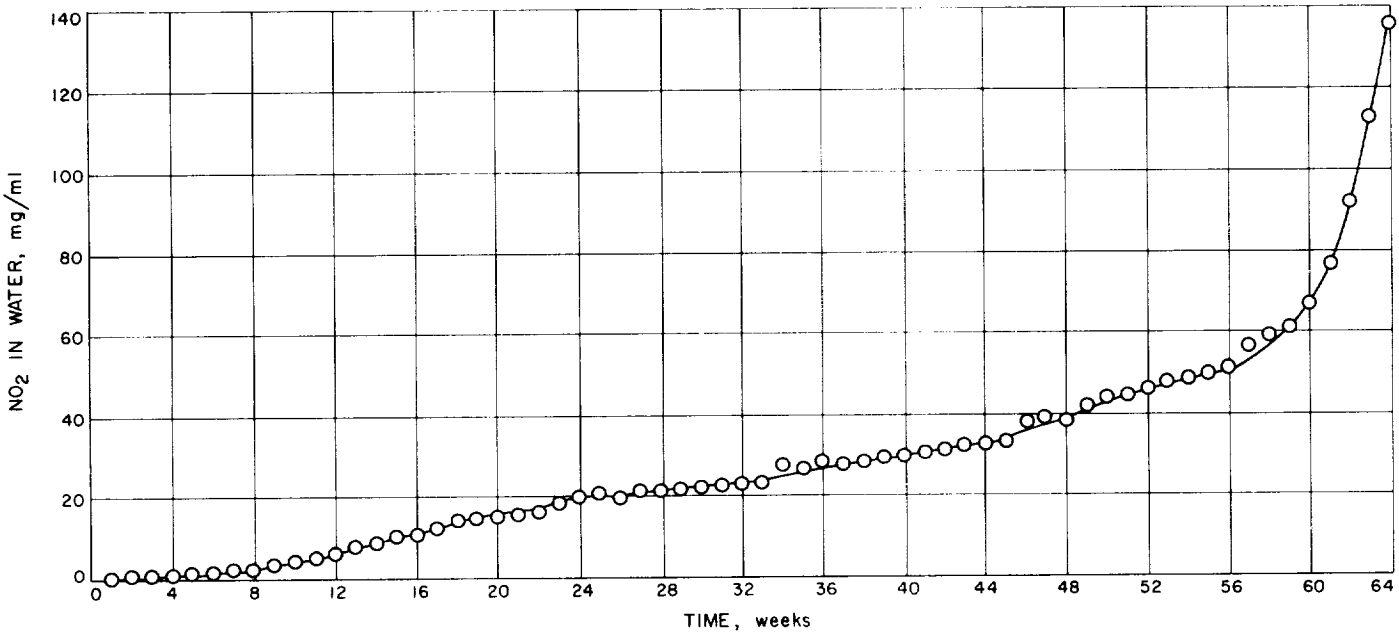


Fig. 5. NO_2 accrual in water-filled Teflon bladder due to permeation through N_2O_4 bladder

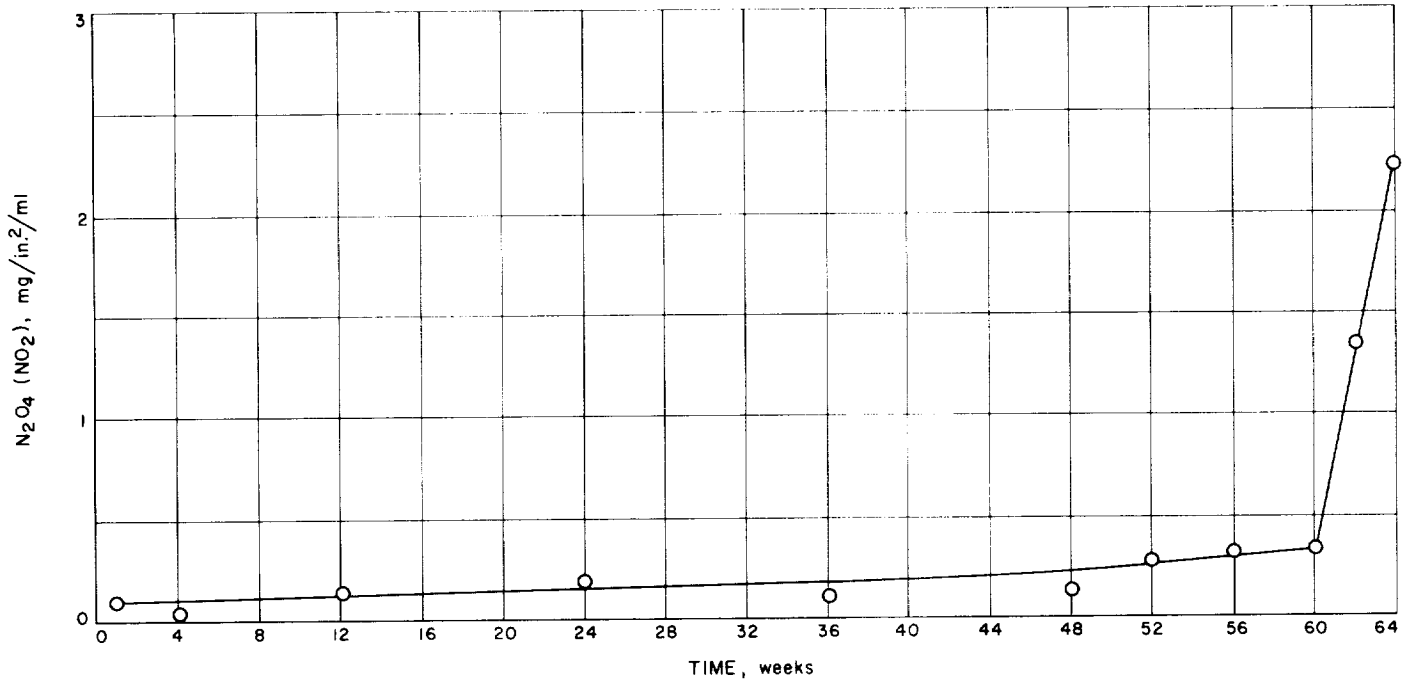


Fig. 6. Calculated permeation rate of NO_2 from N_2O_4 bladder into water-filled bladder

Measurement of NO_2 accrual by weekly sampling and titration was continued throughout the test. The NO_2 content in the water bladder increased at a nearly constant rate until the 60th week as indicated by the NO_2 accumulation curve (Fig. 5) and the permeation rate curve (Fig. 6). At that time a sharp increase in rate was detected, probably due to damage to one of the bladders. Although it is not possible to ascertain definitely which one failed, there are indications that the water bladder may have been cut due to displacement of the divider panel far enough off center to pinch the water bladder between the edge of the panel and the tank wall. This would have been possible because sampling depleted by some 7200 ml the initial quantity of 25400 ml of water. Thus the partially emptied water bladder would not counterbalance the pressure from the NO_2 bladder.

Whatever the cause, it is evident from the data that the NO_2 content in the water bladder increased rapidly at this time, and the permeation rate increased to that expected for a single layer of 0.010 in. thick, sprayed and sintered TFE-FEP Teflon material. Both of these conditions would have been apparent with either bladder broken. However, during the 64th week the burst diaphragm rated at 365 psi failed at 250 psi. Saturation with N_2O_4 alone had not caused this diaphragm to burst during the 60 weeks when both bladders were intact. Therefore, rupture of the water bladder apparently admitted water to the diaphragm area, and this, together with the N_2O_4 already present, formed nitric acid (HNO_3) which attacked the burst diaphragm and caused it to fail after 4 weeks' exposure.

The conclusions that can be drawn from this experiment are:

(1) N_2O_4 will permeate into the fuel bladder and certainly saturate the entire ullage in a system of this kind.

(2) The permeation rate of N_2O_4 through two bladders walls is less by a factor of 10 than has been measured through a single membrane of similar material, and remains at a reasonably constant level as long as the system remains intact.

More tests will be made in the future.

References

1. Weiss, H. G., Johnson, B., Fisher, H. D., Gerstein, M., "Modification of the Hydrazine-Nitrogen Tetroxide Ignition Delay," *AIAA Journal*, (to be published).

2. Kilpatrick, M., and Baker, L. L., Jr., Fifth Symposium on Combustion, Reinhold Publishing Corp., New York, N. Y., 1955, Paper No. 11, p. 196.

3. Elverum, G. W., Jr., and Morey, T. F., "Criteria for Optimum Mixture Ratio Distribution Using Several Types of Impinging Stream Elements," Memorandum No. 30-5, Jet Propulsion Laboratory, Pasadena, California, February 25, 1959.

XIII. Advanced Propulsion Engineering

A. Liquid MHD Power Conversion

D. Elliott, D. Cerini, R. Eddington

The long lifetimes required of electric-propulsion powerplants make nonrotating cycles attractive. A nonrotating powerplant under investigation at JPL is the liquid magnetohydrodynamic system shown schematically in Fig. 1. In this cycle a fluid, such as cesium, circulates in the vapor loop and causes a liquid metal, such as lithium, to circulate through an MHD generator in the liquid loop. The cesium leaves the radiator as condensate, flows through an EM pump to the mixer, vaporizes on contact with the lithium, atomizes and accelerates the lithium in the nozzle, separates from the lithium in the separator, and returns to the radiator. The lithium leaves the separator at high velocity (typically 500 ft/sec), decelerates through the production of electric power in the MHD generator, and leaves the generator with sufficient velocity (typically 300 ft/sec) to return through a diffuser to the reactor where the lithium is reheated.

Results of previous tests were reported in Refs. 1 and 2.

1. Diffuser Performance

In the two-phase diffuser tests it was found that diffuser efficiencies decreased rapidly with increasing gas-to-liquid volume ratio. The explanation lies in the supersonic nature of the inlet flow, as first pointed out in Ref. 3.

The low sonic velocity, and possibility of supersonic phenomena, in two-phase flows has long been known (Refs. 4, 5). The sonic velocity depends on the pressure p and density ρ through the well-known relation

$$c = \left(\frac{dp}{d\rho}\right)^{1/2} \quad (1)$$

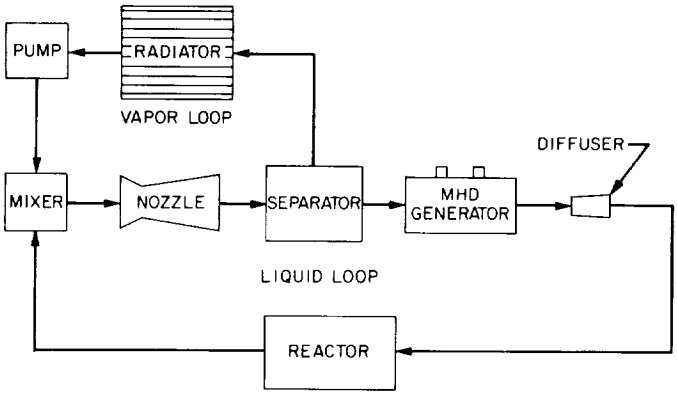


Fig. 1. Liquid MHD power conversion cycle

For a two-phase mixture with gas-to-liquid mass ratio r_m and volume ratio r_v the density is

$$\rho = \rho_l \frac{1 + r_m}{1 + r_v} \tag{2}$$

where ρ_l is the liquid density.

For the volume ratios of interest here ($r_v < 10$), temperature changes are negligible so that

$$_v p = \text{const.} \tag{3}$$

Substituting Eqs. (2) and (3) into Eq. (1), the sonic velocity is

$$c = (1 + r_v)\left[\frac{p}{\rho_l r_v(1 + r_m)}\right]^{1/2} \tag{4}$$

For water at atmospheric pressure and volume ratio of 1.0 (with any gas, since r_m can be neglected) the sonic velocity is only 66 ft/sec.

At velocity V the Mach number is

$$M = \frac{V}{c} \tag{5}$$

Since the mixture entering the separator capture slot and flowing to the generator and diffuser in a liquid MHD conversion system has volume ratios of 1.0–2.0 and velocities of 300–500 ft/sec, the flow is supersonic and shock waves can theoretically occur. In particular, a normal shock wave is possible for which the theoretical pressure rise is obtained as follows:

The continuity equation across the normal shock is

$$\rho_1 V_1 = \rho_2 V_2 \tag{6}$$

The momentum equation is

$$\rho_1 V_1^2 - \rho_2 V_2^2 = p_2 - p_1 \tag{7}$$

Combining Eqs. (6) and (7), and employing Eqs. (2) and (3), the pressure downstream of the shock is

$$p_2 = \frac{\rho_l V_1^2 r_{v_1}(1 + r_m)}{(1 + r_{v_1})^2} \tag{8}$$

Comparing Eqs. (8) and (4), the pressure ratio is simply

$$\frac{p_2}{p_1} = M_1^2 \tag{9}$$

For water at 19 psia, volume ratio of 1.5, and velocity of 250 ft/sec, the Mach number is 3.26 and the normal shock downstream pressure is $p_2 = 202$ psia. The downstream volume ratio and velocity are 0.141 and 114 ft/sec, respectively.

If the mixture can decelerate over a finite distance, with area change, then the restriction of Eq. (6) does not apply and the momentum equation at each point is

$$\rho V dV = -dp \tag{10}$$

Integrating this equation with the aid of Eqs. (2) and (3), the pressure attainable with shock-free, isentropic, flow is

$$p_2 = p_1 + \tfrac{1}{2}\rho_l (1 + r_m)(V_1^2 - V_2^2) - r_{v_1} p_1 \ln \frac{p_2}{p_1} \tag{11}$$

Setting $V_2 = 0$, Eq. (11) gives the isentropic stagnation pressure. For water at 19 psia, volume ratio of 1.5, and velocity of 250 ft/sec, the isentropic stagnation pressure is 356 psia. If Eq. (11) is applied to the shock downstream conditions calculated previously, the additional recovery obtainable is found to be 79 psi. A subsonic diffuser, however, is only about 86% efficient so that the actual recovery downstream of the shock would be 68 psi. The pressure obtainable with a normal shock is, therefore 202 + 68 = 270 psi. The corresponding efficiency is 270/356 = 76%.

Fig. 2 presents efficiencies calculated in this way for $p_1 = 14$ psia and for $v_1 = 200$ and 300 ft/sec. It is seen

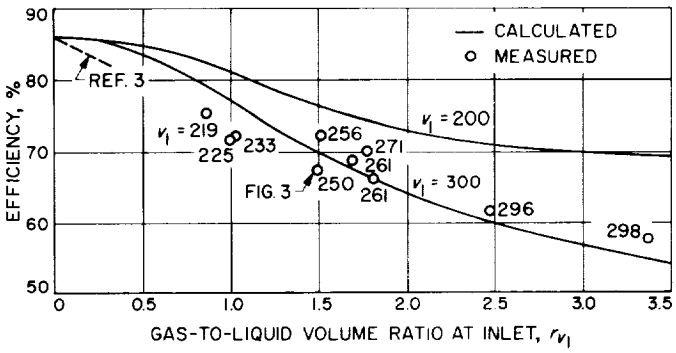


Fig. 2. Comparison of measured annular diffuser efficiencies with values calculated for inlet normal shock plus 86% recovery of downstream stagnation pressure

that the calculated efficiencies follow the observed downward trend with volume ratio.

The annular diffuser, however, did not have the inlet normal shock assumed in the curves. It had a convergent inlet of 0.77 area ratio and a slightly divergent throat of 12:1 length-to-gap ratio in which the normal shock occurred at a reduced Mach number. Table 1 presents performances predicted for this geometry. With no friction in the convergent section and throat, the exit pressure would be 310 psia, giving an efficiency of 87%. With friction included, however, the exit pressure is calculated to be only 248 psia and the efficiency 70%, in good agreement with the measured exit pressures which ranged

Table 1. Diffuser performance with V_1 = 250 ft/sec, p_1 = 19 psia, r_{r1} = 1.5, and M_1 = 3.26

Type of recovery	Exit pressure, psia	Efficiency, %
Isentropic	356	100
Inlet normal shock + 86% recovery of downstream stagnation pressure	270	76
Frictionless supersonic flow to throat with 0.77 contraction ratio + normal shock + 86% recovery of downstream stagnation pressure	310	87
Same as above with friction added to supersonic and normal shock section corresponding to skin fraction coefficient C_f = 0.005 and to wall area of annular diffuser	248	70
Measured	240–260	67–73

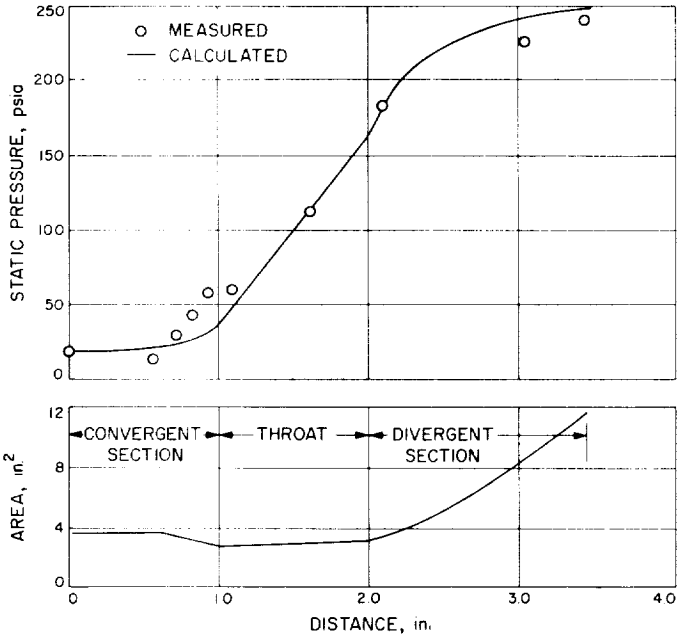


Fig. 3. Comparison of calculated and measured static pressure profiles in annular diffuser

from 240 psia (the highest exit pressure attainable with no effect on inlet pressure) to 260 psia (the highest pressure attainable before occurrence of a normal shock at the inlet).

Fig. 3 compares the measured static pressures in the annular diffuser with those calculated. The agreement is good except for a premature pressure rise in the convergent section which could be due either to oblique shocks or to backward slip of the gas.

It appears, then, that the annular diffuser failed to achieve better than inlet-shock performance because the length of inlet and throat employed to reduce the Mach number and achieve the normal-shock rise added friction losses which exceeded the reduction in shock loss.

2. *Supersonic Two-Phase Tunnel*[1]

To gain a more detailed understanding of supersonic phenomena in two-phase flow, a two-phase tunnel has been constructed. Figs. 4 and 5 are photographs of the tunnel, and Table 2 gives dimensions and operating conditions.

[1]This work is being performed as a C.I.T. graduate research project by Capt. R. B. Eddington, USAF.

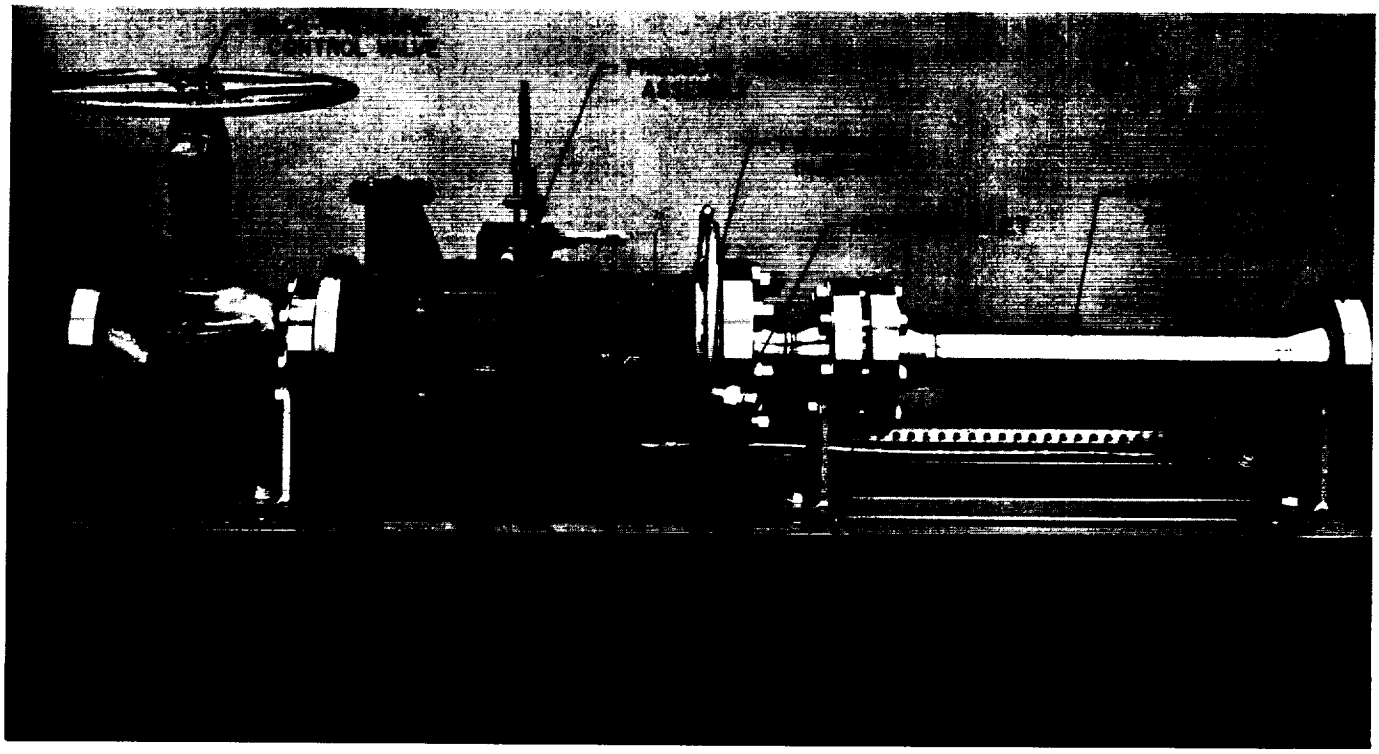


Fig. 4. Supersonic two-phase tunnel

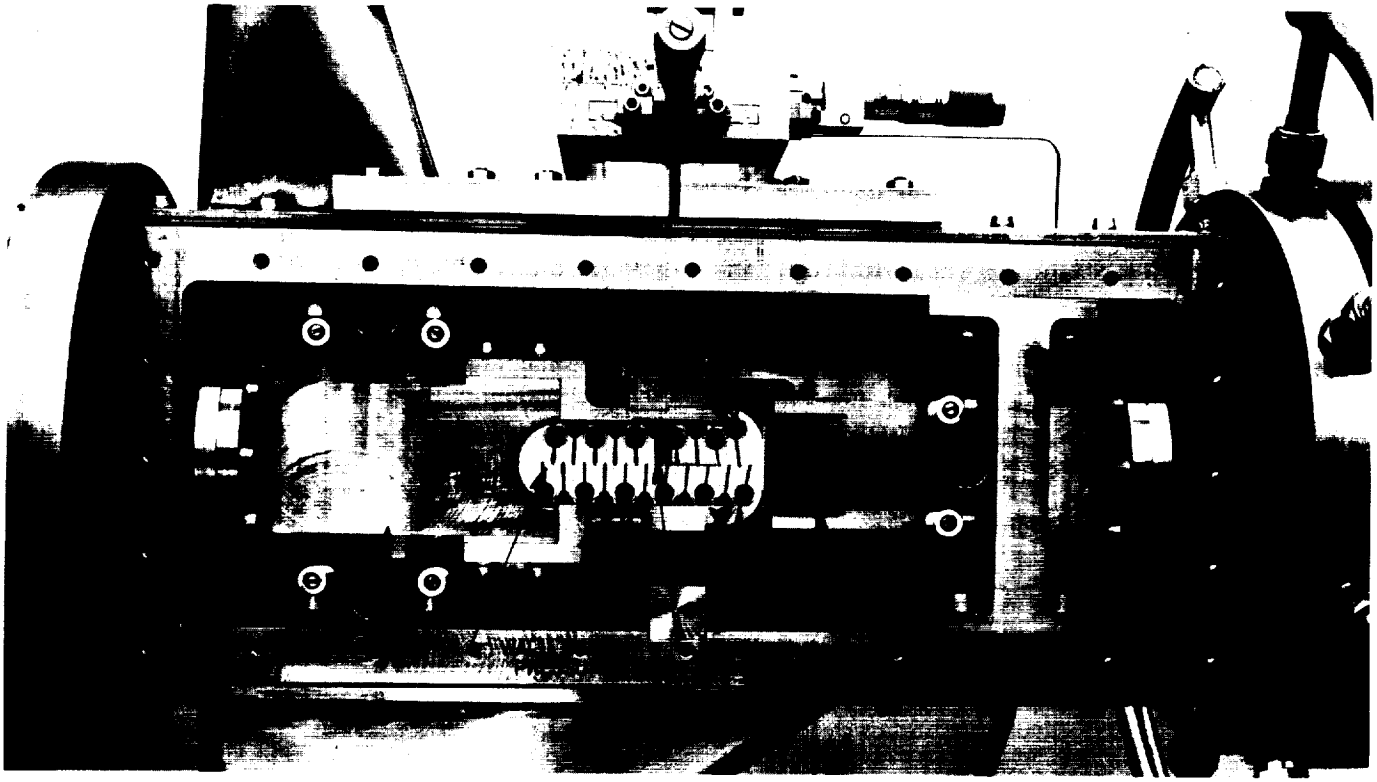


Fig. 5. Tunnel test section

Table 2. Two-phase tunnel dimensions and operating conditions

Upstream height	1.956 in.
Upstream width	0.756 in.
Upstream cross-section	1.479 in.^{2}
Water injection area	0.7396 in.^{2}
Volume ratio r_v	1.0
Test section height	1.600 in.
Test section width	0.381 to 0.756 in.
Test section length	4.00 in.
Water injection pressure	0–1000 psi
Water flow rate	0–100 lb/sec
Gas injection pressure	0–500 psi
Gas flow rate	0–300 ft^3/min
Mixture velocity	0–334 ft/sec
Mach number at 1.0 atm	0–5.14
Reynolds number (maximum, based on tunnel height)	2×10^6
Power density	443 hp/in.^{2}

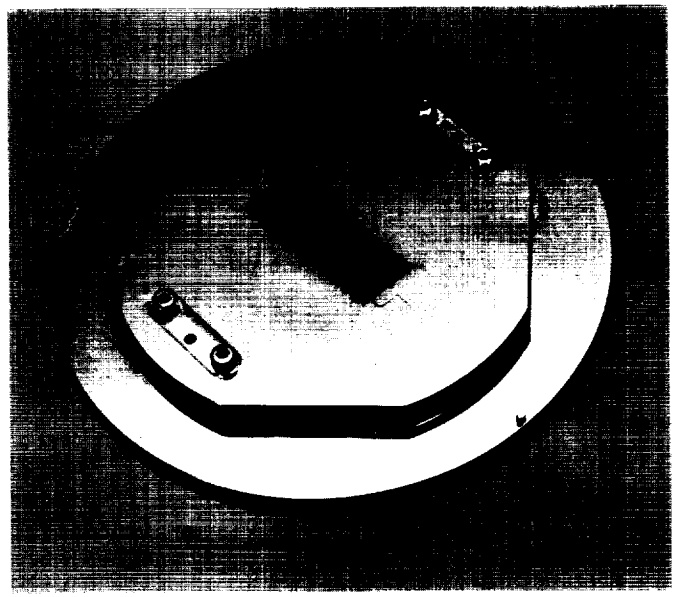


Fig. 6. Two-phase injector

The tunnel utilizes an 800-hp water pump having a maximum capacity of 720 gpm at 1100 psi. The tunnel consists of a rectangular upstream duct which directs a two-phase jet of $r_v = 1.0$ across a short gap to an adjustable-width downstream duct with bleed-off of the incoming boundary-layer flow. Transparent side walls in moveable carriages allow the test section width to be changed for removal of the side boundary layer, and fixed top and bottom knife edges remove the boundary layers from the top and bottom surfaces and serve as a holding point for normal shocks. A probe housing is attached to the top wall for three-dimensional positioning of a probe in the flow at pressures up to 1000 psi. One view block is fitted with static pressure taps along its horizontal centerline.

The two-phase injector (Fig. 6) is a bundle of 192 0.093-in. OD × 0.070-in. ID tubes which injects water through the tube holes and nitrogen through the cusp areas between adjacent tubes.

Preliminary tunnel operation showed clearly visible oblique and normal shocks of pressure ratios reaching 40:1. Shock intensity and the positioning of normal shocks were readily variable through control of water velocity, gas flow (changing upstream tunnel pressure at constant $r_v = 1.0$) and downstream throttling.

Figs. 7, 8, and 9 show flows at 275 ft/sec. In Fig. 7, shocks from the upper knife, probe body, and probe tip can be seen where they intersect the transparent side wall. Fig. 8 shows a normal shock positioned at the knife edges. The static pressure to the right (upstream) of the shock is near atmospheric and to the left is approximately 300 psig. Fig. 9 shows a normal shock positioned just downstream of the probe tip. The intersection of the normal shock and the conical shock from the probe tip is visible.

Fig. 10 is a ½-μsec flash photograph at the same conditions as Fig. 8. The detailed flow structure appears to be that of ligaments of water elongated in the flow direction giving a striated appearance. In motion pictures at 25,000 frames/sec, these ligaments could be followed moving through the shock waves.

Total and static pressure surveys have been taken across normal shocks at velocities from 155 to 325 ft/sec with good repeatability. Fig. 11 shows agreement within 8% between experimental and theoretical [Eq. (8)] values for the downstream static pressure of a normal shock. The upstream pressure is near atmospheric for the experimental points shown. This data is preliminary pending refinement of tunnel velocity values.

Flow around wedges and cones has also been investigated in the tunnel. Fig. 12 shows a 20-deg double wedge mounted in the test section, and Fig. 13 shows the shock pattern at a velocity of 300 ft/sec. Fig. 14 shows the same flow at an exposure of ½-μsec. The flow shows a double shock pattern which is currently being investigated with pressure surveys.

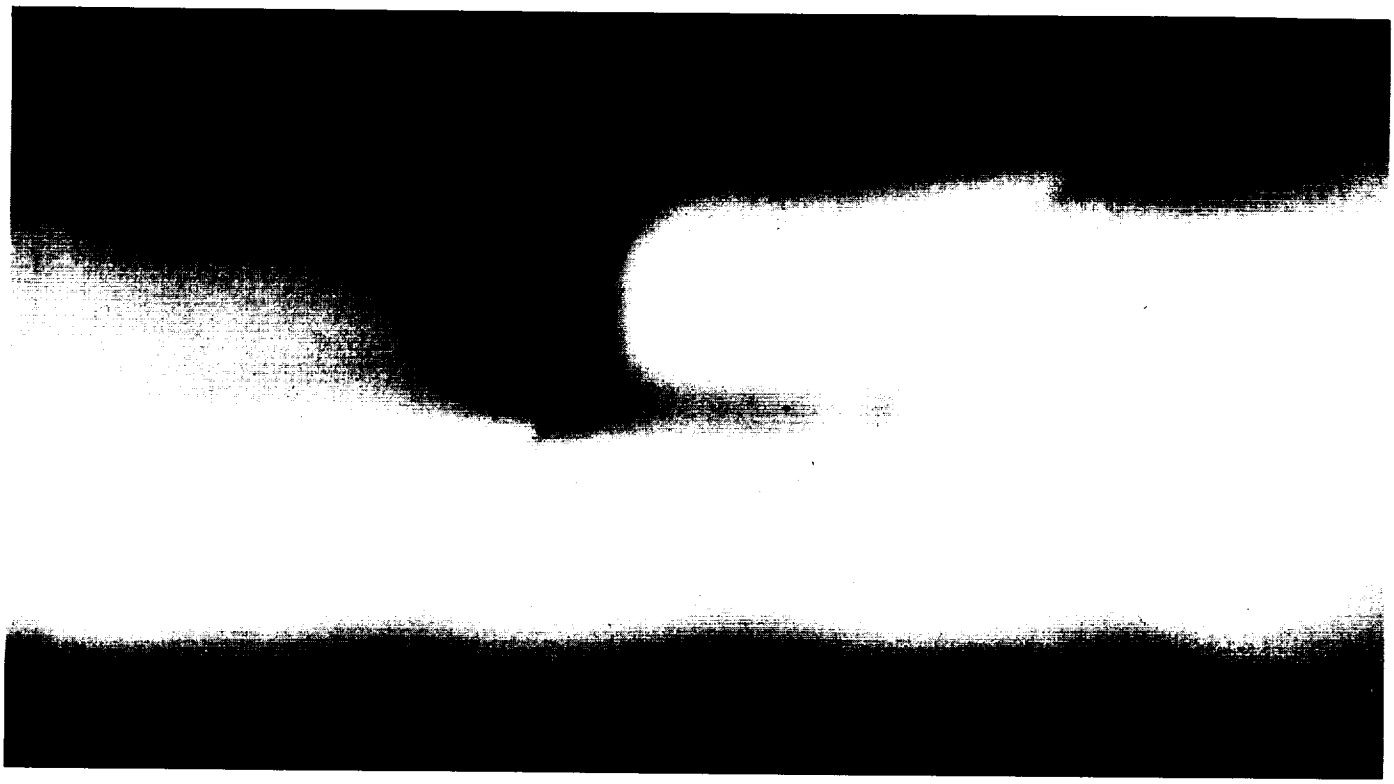


Fig. 7. Oblique shocks from inlet and probe at 275 ft/sec ($M = 4.2$)

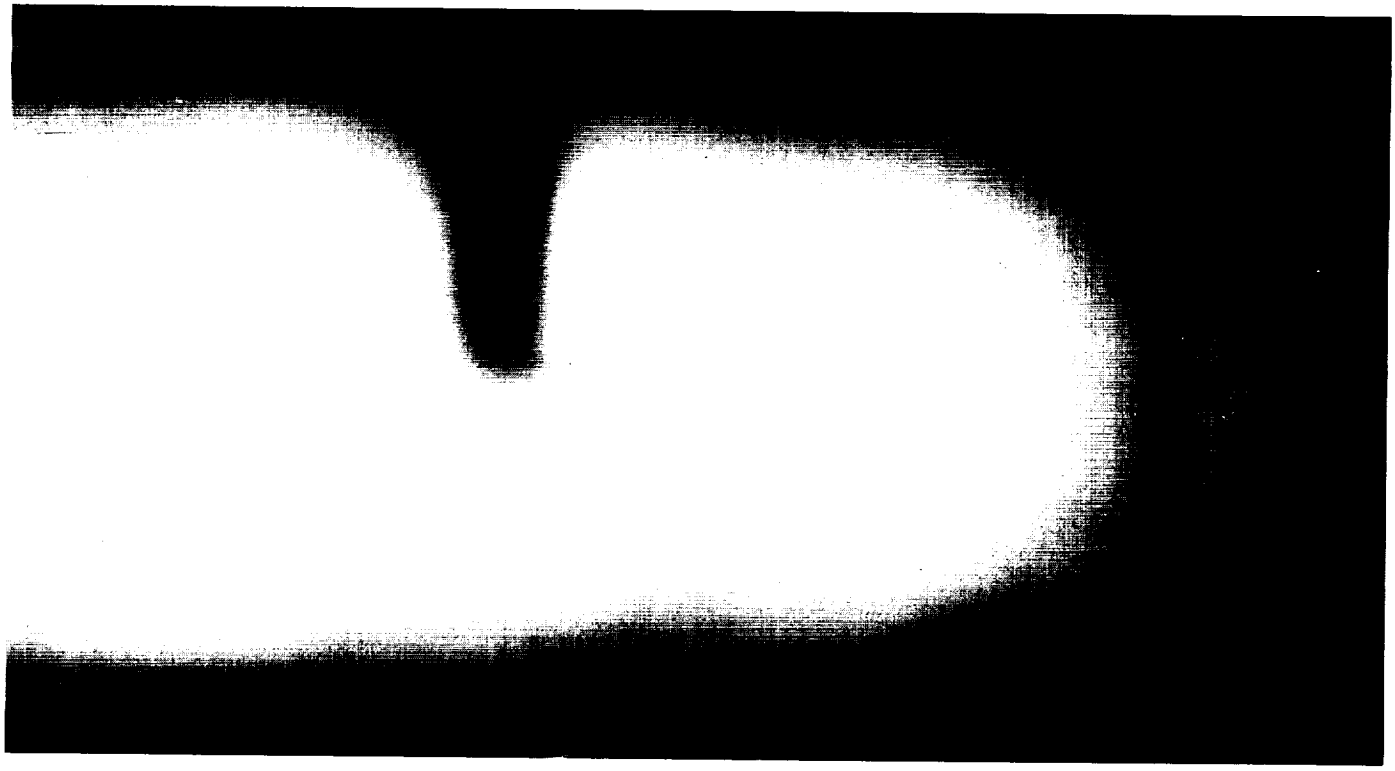


Fig. 8. Normal shock at inlet, 275 ft/sec

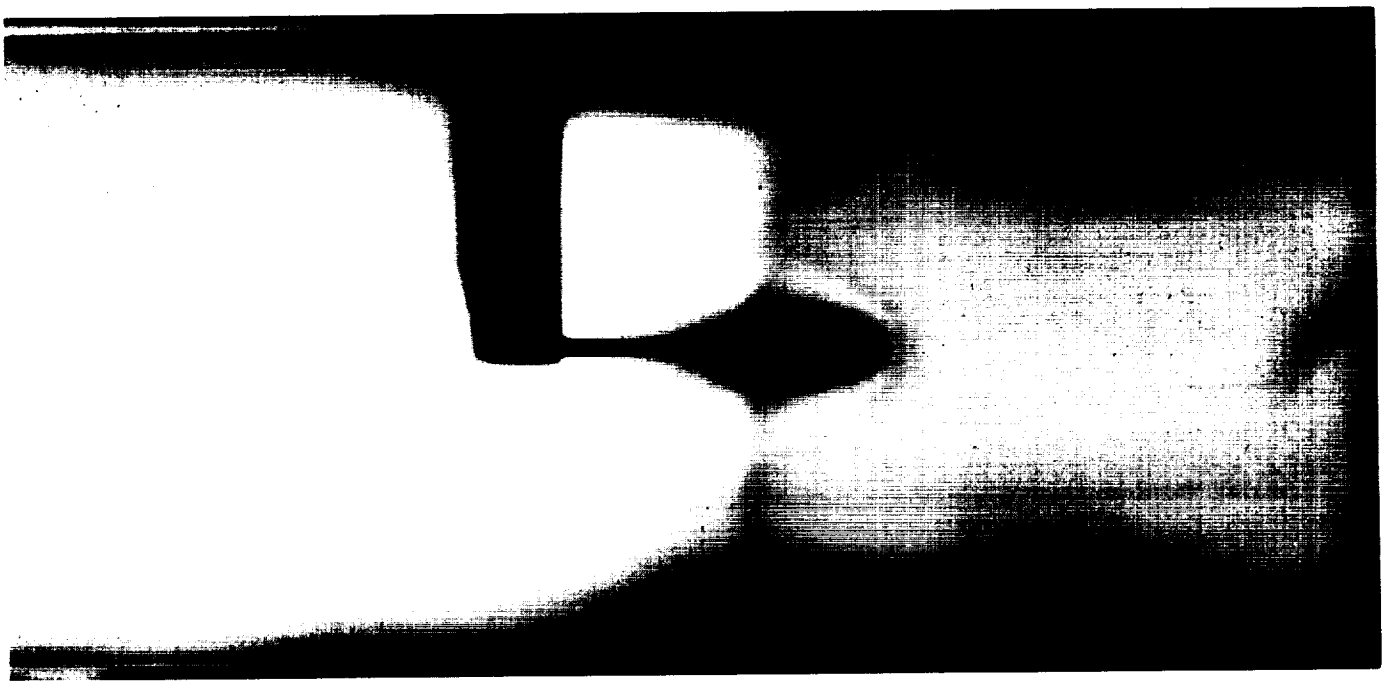


Fig. 9. Intersection of normal and oblique shocks at probe, 275 ft/sec

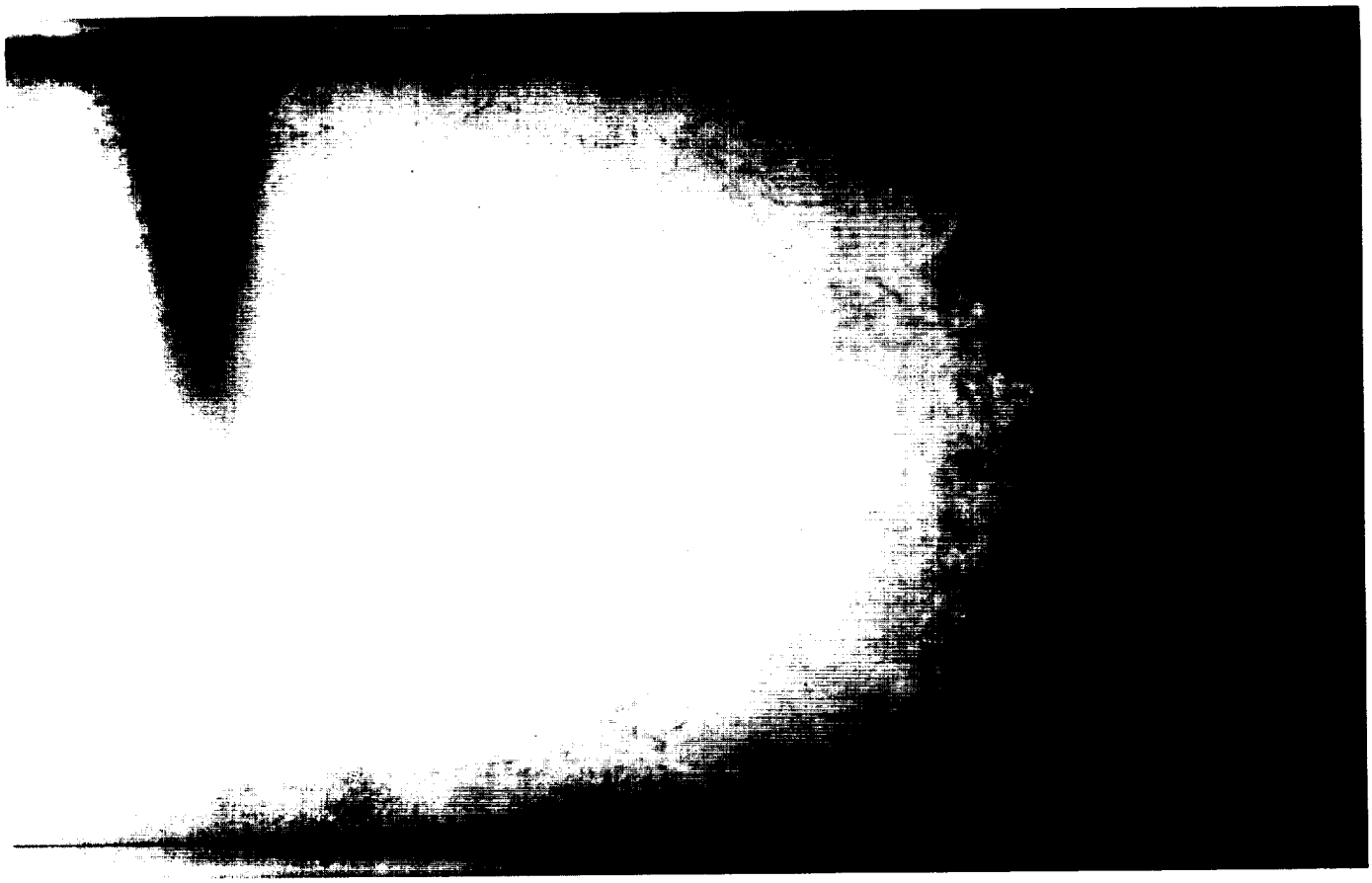


Fig. 10. Half-microsecond exposure of normal shock at 275 ft/sec

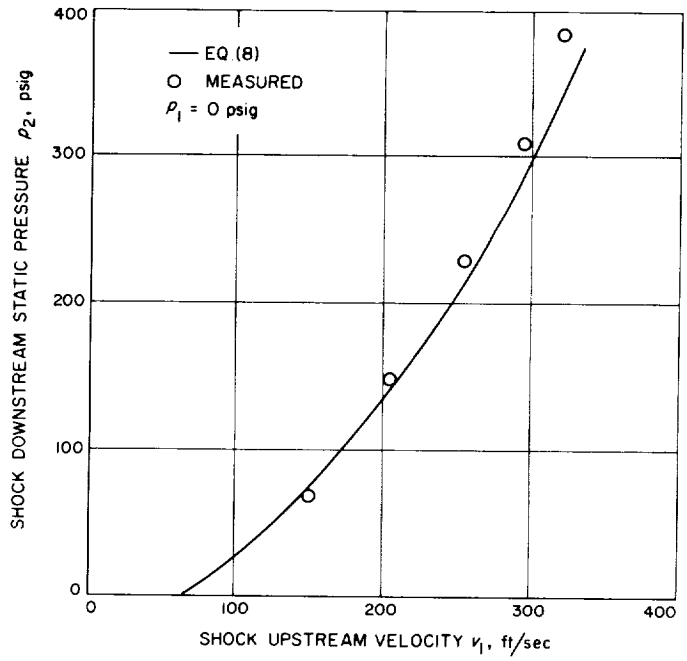


Fig. 11. Comparison of theoretical and experimental shock downstream static pressure in two-phase tunnel

Fig. 12. 20-deg double wedge model in test section (pressure taps visible in rear transparent wall)

Fig. 13. Flow around wedge at 300 ft/sec

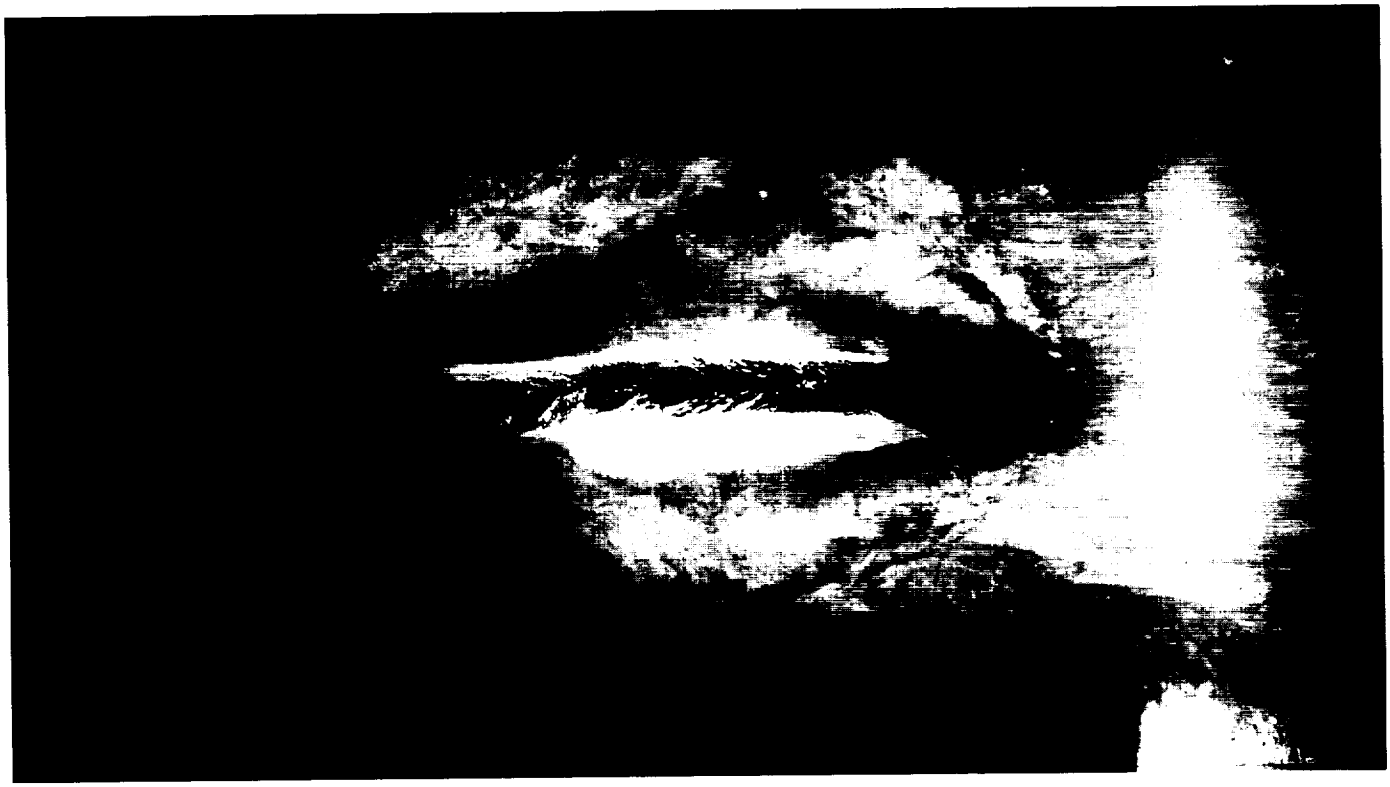


Fig. 14. Half-microsecond exposure of flow around wedge at 300 ft/sec

References

1. Elliott, D. G., Cerini, D. J., and Weinberg, E., *Investigation of Liquid MHD Power Conversion,* AIAA Paper No. 64-760. Third Biennial Aerospace Power Systems Conference, Philadelphia, Pa., September 1–4, 1964.

2. Elliott, D. G., Cerini, D. J., Hays, L., O'Connor, D., and Weinberg, E., "Liquid MHD Power Conversion," *SPS 37-28,* Vol. IV, pp. 86–92, Jet Propulsion Laboratory, Pasadena, California, August 31, 1964.

3. Schneiter, G. R., *Diffuser Studies With Single and Two-Phase Flows,* Report No. TM-62-1, Jet Propulsion Center, School of Mechanical Engineering, Purdue University, Lafayette, Indiana, April 1962.

4. Tangren, R. F., Dodge, C. H., and Seifert, H. S., "Compressibility Effects in Two-Phase Flow," *Journal of Applied Physics,* Vol. 20, No. 7, July 1949, pp. 637–645.

5. Campbell, I. J., and Pitcher, A. S., *Shock Waves in a Liquid Containing Gas Bubbles,* Admiralty Research Laboratory, Report A.R.L. Rl/G/H7/17/0, Teddington, Middlesex, England, August 1957.

SPACE SCIENCES DIVISION

XIV. Lunar and Planetary Instruments

A. A Conductivity Instrument as Part of a Life Detection System

J. R. Clark

Work continues in the Exobiology Instruments Group on the development of basic tools for the measurement of fluorescence, Tyndall effect (light scattering by living organisms), hydrogen ion concentration, and optical polarization.

Conductivity measurement is another area currently under investigation. The measurement applies to exobiology in two ways: (1) to determine the physical conductivity of a planetary surface so as to yield some information about the moisture content of the soil, and (2) to monitor conductivity of a solution while specific chemical reactions take place. From the conductivity changes the presence or absence of such elements as carbon, hydrogen, and possibly oxygen and nitrogen can be detected.

By applying proper techniques (Refs. 1–4) it is possible to determine the amount of carbon in a 1-mg sample to an accuracy of $\pm 0.29\ \mu g$ (Ref. 1). Hydrogen content can be determined to an accuracy of $\pm 0.19\ \mu g$ (Ref. 2). It is also possible to determine quantities of carbon and hydrogen in a given sample simultaneously in a composite experiment (Ref. 3). If these elements are found to exist in the soil, then the probability that life exists is increased. This is basically an experiment to test for a suitable environment for life as we know it.

The basic procedure of the experiment is to bubble the products of combustion from a sample (soil) through a potassium hydroxide solution. Carbon or hydrogen is detected by noting the conductivity changes throughout the absorption process. The amount of carbon or hydrogen in a given sample is directly proportional to the conductivity change.

In order to implement the above technique, a conductivity meter to detect the required small changes was

developed. In principle, the resistance of solution forms the resistive arms of a modified Wien bridge oscillator, as shown in Fig. 1. The two resistive elements of the conventional Wien bridge oscillator are shunted by a third resistance which has a value about four times the other two. This has the effect of requiring that the amplifier gain be increased from the usual value 3 to 13/4. The resonant frequency as an added effect is also changed from $\omega_0 = 1/RC$ to $\omega_0 = 1.5/RC$. The oscillator contains a dc feedback loop for bias stability and an ac feedback loop for amplitude stabilization. The ac feedback loop (or AGC loop) contains the peak detector, dc amplifier, and two field-effect transistors to operate as voltage-controlled resistors. Two such transistors are used in parallel to provide the required dynamic range of control in amplitude. One difficulty in measuring conductivity by this method is the large load imposed on the forward amplifier when measuring solutions of very high conductivity.

The oscillator possesses a frequency stability with respect to temperature of +0.03%/°C over a frequency

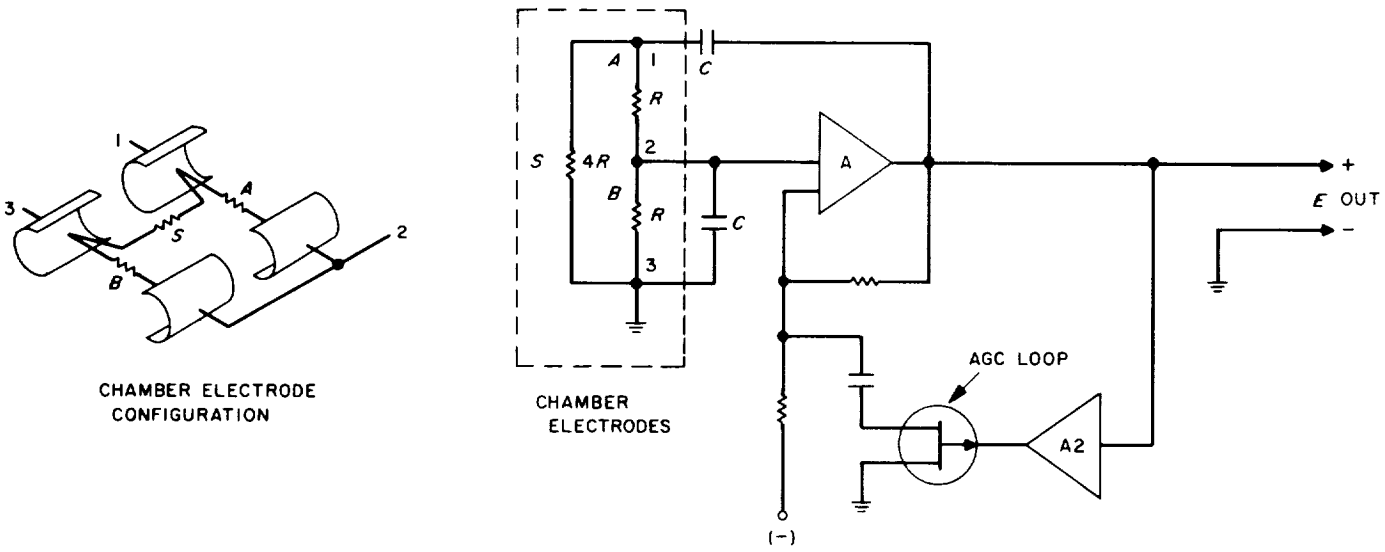


Fig. 1. Conductivity instrument block diagram

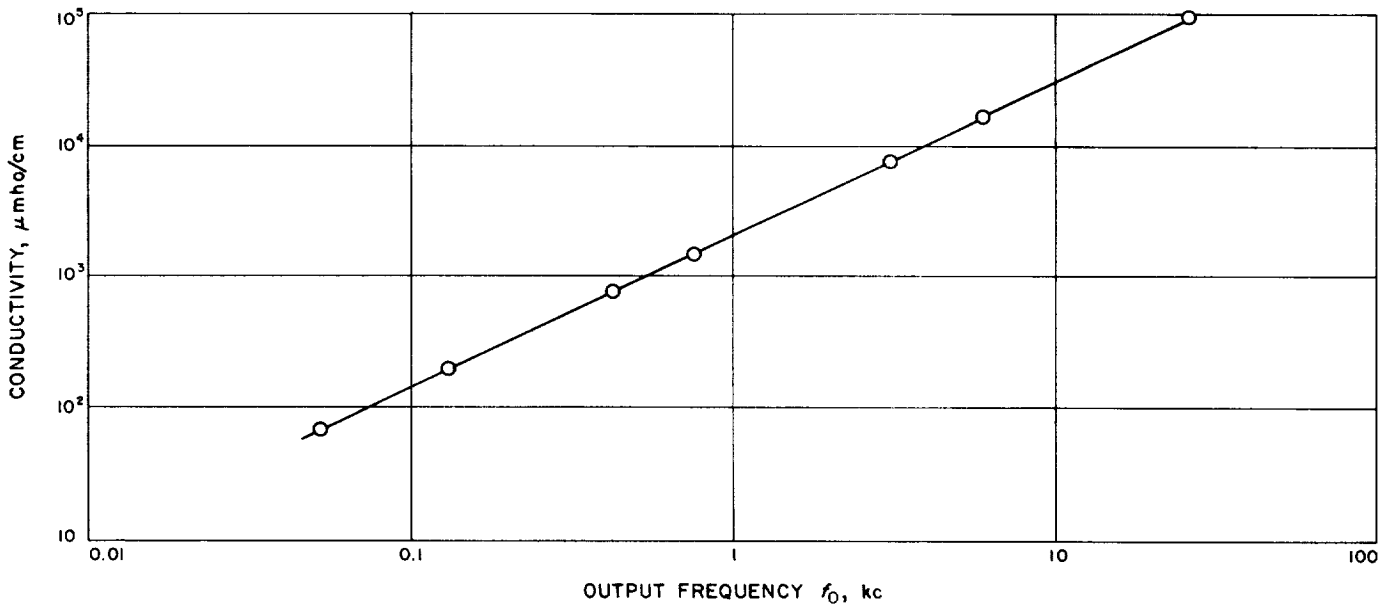


Fig. 2. Conductivity preliminary calibration curve

of 0.1 to 100 kc. The oscillator also possesses amplitude stability with respect to frequency of less than ±0.5 db over the same frequency range. These measurements were made with fixed resistors substituted for the resistances of solutions. In Fig. 2, however, a preliminary calibration curve is shown using standard solutions of known conductivity.

B. Sample Collection by Aerosol Method

S. B. Tuttle

The aerosol collection of particulate matter offers unique characteristics particularly for an exobiological life detection experiment. The system does not require gravity orientation providing the spacecraft is within 15 deg of vertical. It permits collection of particulates from multiple points. It is discriminatory of particle size and both the maximum and minimum particle can be adjusted by design. In the absence of particulate matter, it will deliver atmospheric gases with their entrained micro-organisms. The only controls needed are a pyrotechnic valve to eject the aspirators and start the process, and a photo-electric diverter to stop delivery to any experiment.

A previous feasibility test indicated that the scheme was workable. During recent months, effort was directed toward optimum performance. There are two components which present unique problems, i.e., an aspirator and a separator. The aspirator must accelerate particulate matter at ground level to a velocity in excess of its terminal free-fall velocity using high-pressure gas. The separator must exhaust a large percentage of the aerosol gas and deliver an enriched mixture to the experiment.

This article presents the development of an aspirator suitable for operation at standard atmospheric pressure.

1. Design Factors Affecting Performance of Aspirator

The functions of the aspirator are divided into four processes:

(1) Acceleration of particles at ground level.

(2) Entrainment of particles and surrounding gas by viscous friction at the periphery of the jet.

(3) Over-expansion of the jet to a pressure below that of the induced gases.

(4) Reconversion of energy in the combined streams to increased pressure (and reduced velocity), as imposed by the aerosol transport and recovery system.

The important aspects in the design of the various parts of an aspirator are discussed in the following paragraphs. The cross section of a typical aspirator is shown in Fig. 3.

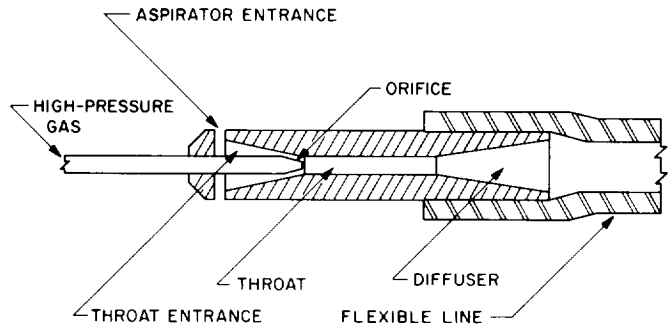


Fig. 3. Aspirator

a. Aspirator entrance. The entrance shall be as close to the ground as possible. It should have a limited width such that it cannot clog with large, low density particles. An annular slot 0.02 in. wide × 0.40 in. long has been adopted.

b. Orifice. The primary jet orifice is a choked nozzle of 0.01 in. diameter. The flow rate at 50 psig is 2000 to 2500 cm^3/min, measured at standard temperature and pressure. A cartridge of liquid CO_2 (12.5 g CO_2) gives a sampling period of about 2 min for a single aspirator.

c. Throat entrance. The entrance passage is a conical section with an included angle of 25 deg with edges bell-mouthed for maximum efficiency.

d. Throat. The mixing chamber may be either a constant-pressure or a constant-area type. Various experimental works indicate that the constant-area type works well at subsonic velocities over a wide range of conditions. The throat area is a critical dimension and a small change in throat area makes a large change in performance. The throat diameter and length are dimensions to be optimized.

e. Diffuser. A divergent section between the throat and delivery line is necessary. This will be made conical with an included angle of 8 deg and a final diameter of 0.12 in.

2. Method of Development of Aspirator

The small dimensions of these components introduce size factors for which there is little technical or experimental information. The critical dimensions involved as design parameters are:

(1) Minimum diameter of hose, at least 10 ft in length, connecting aspirator and separator.

(2) Minimum flow rate and storage tank capacity for high-pressure gas.

(3) Dimensions defining size and configuration of aspirator.

The criteria used to judge the performance capability of an aspirator are as follows:

(1) The ratio of total to primary volume flow shall be as high as possible.

(2) This volume ratio shall not be adversely affected by a large change in pressure ratio.

(3) The volume ratio shall be approximately linear in relation to head pressure.

(4) The stall pressure shall be at least 200% of the normal operating back pressure.

Time and facility has not permitted an exhaustive study of all parameters affecting performance. Consequently, only critical parameters were varied over the effective ranges. Other less critical parameters were arbitrarily frozen based on linear proportionality or based on performance factors not affected by size.

Tests at atmospheric pressure indicate that optimum performance is produced at a throat-to-orifice diameter ratio of approximately 2. As the pressure ratio diminishes, the volume ratio increases slightly for any diameter ratio less than 6. The results of these tests are shown graphically in Fig. 4.

3. Collection Rate

With a copious supply of particulars, this aspirator will aerosolize and deliver at least 1 g/min of particulate having a specific gravity of 2. The Minivator experiment requires a sample of 0.1 g.

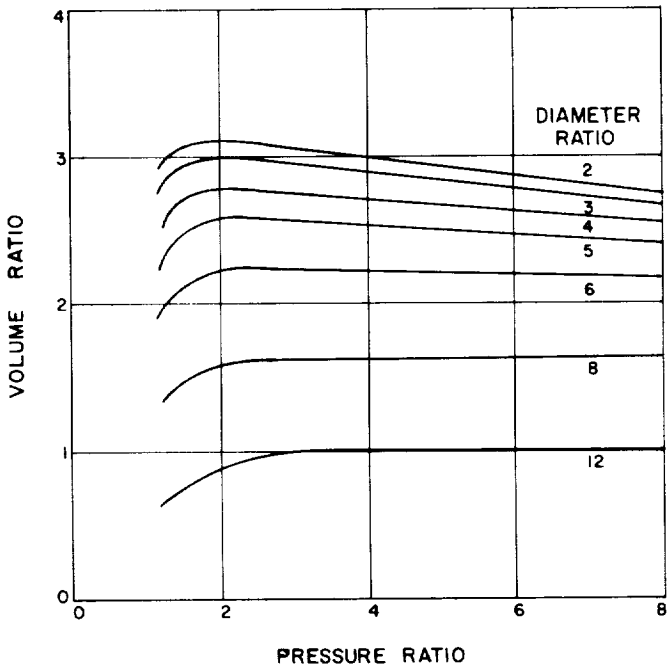


Fig. 4. Aspirator characteristic

4. Future Test Program

A continuing series of tests is in progress wherein the aspirator will be optimized for similar characteristics at reduced pressure. The pressure range simulating the Martian atmosphere ranges from 5 to 100 mm Hg. The test to be performed introduces a variable leak rate into the chamber which complicates the pressure control. A test set-up which performs in an acceptable manner is shown in Fig. 5.

The ensuing program involves the following:

(1) Investigate design factors affecting performance of the aspirator at reduced pressure.

(2) Investigate the parameters of cyclone separator design affecting optimum performance on particulates of 5 to 100 μ.

(3) Develop ejection equipment for multiple aerosol sampling.

(4) Develop pressure transducers for use in this pressure range.

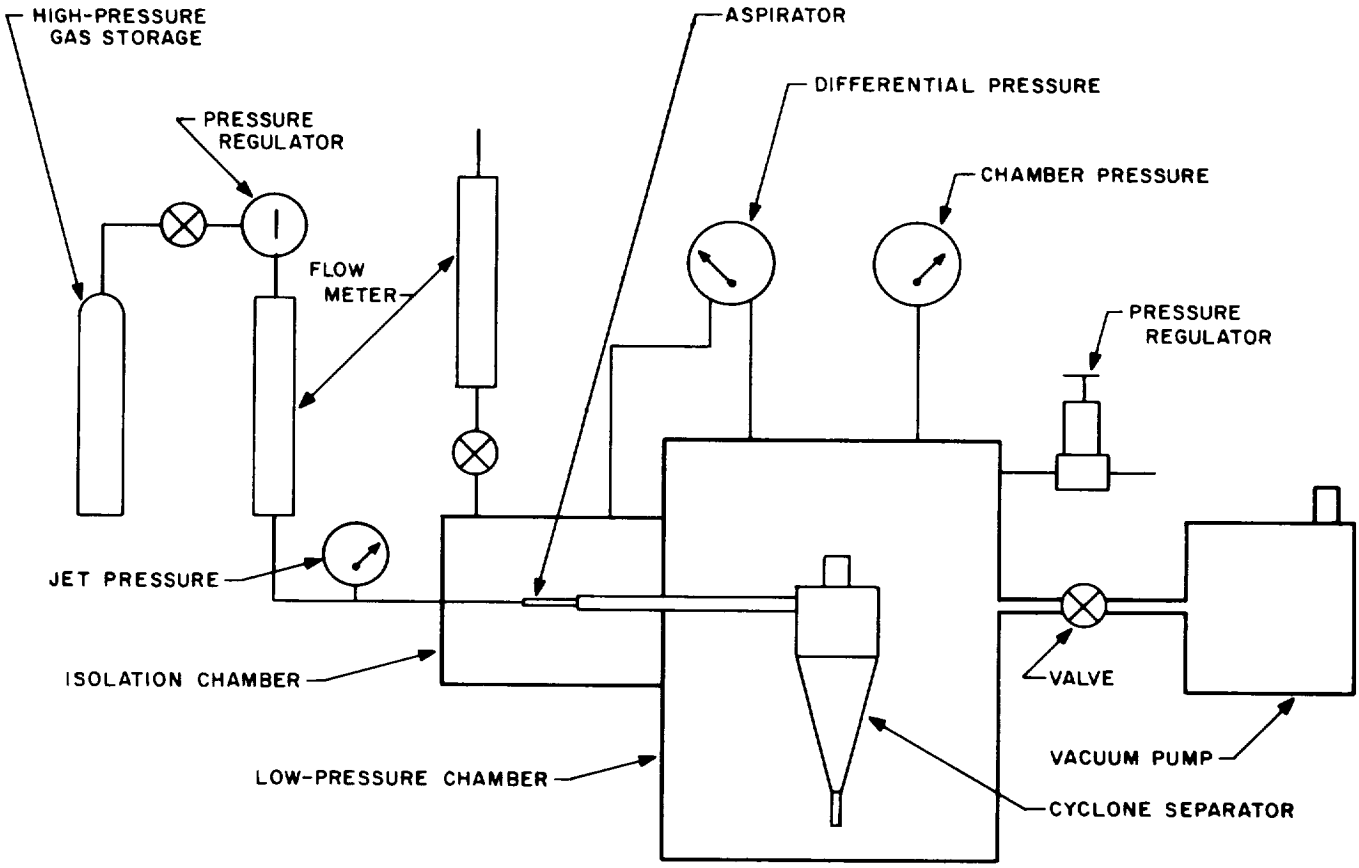


Fig. 5. Low-pressure test set-up

References

1. Greenfield, S., "A Conductimetric Micro Method for Determining Carbon in Organic Compounds," *The Analyst*, Vol. 85, July 1960.

2. Greenfield, S., and Smith, R. A. D., "A Conductimetric Micro Method for Determining Hydrogen in Organic Compounds," *The Analyst*, Vol. 87, November 1962.

3. Greenfield, S., and Smith, R. A. D., "A Conductimetric Micro Method for the Simultaneous Determination of Carbon and Hydrogen in Organic Compounds," *The Analyst*, Vol. 88, November 1963.

4. ASTM Designation D 1125-59 T, *Tentative Methods of Test for Electrical Conductivity of Industrial Water and Industrial Waste Water*, 1950, Revised, 1959.

XV. Space Instruments

A. Antenna Pattern Measurements in the 13-mm Band on the Goldstone 30-ft Antenna

M. L. Kellner and F. T. Barath

1. Introduction

The Goldstone 30-ft precision antenna was used in June and July 1964 for radiometric observations of Venus in the 13-mm microwave band. Since the antenna had not been previously used at this short wavelength, it was felt desirable to obtain simple elevation and azimuth patterns to evaluate the feed-antenna combination utilized for beamwidth, symmetry, and sidelobe structure information. In addition, the sidelobe structure is needed if absolute measurements are to be made on an extended target such as the Moon. This article describes the equipment and procedures used, and the data obtained.

2. Implementation

The patterns of the 30-ft antenna were obtained by utilizing a transmitter located at the Tiefort Mountain collimation site and the radiometer used for the 13-mm band Venus observations as the receiver on the antenna. The far field for the antenna in the 13.5-mm (22-Gc) region is approximately 7.6 mi; therefore, the 13-mi distance between the Venus site and Tiefort was more than adequate for the measurements. The free-space loss over this path is 146 db at 22 Gc, neglecting atmospheric attenuation.

Fig. 1 shows the boresighting bench used as the transmitter. A klystron, with a maximum output power of

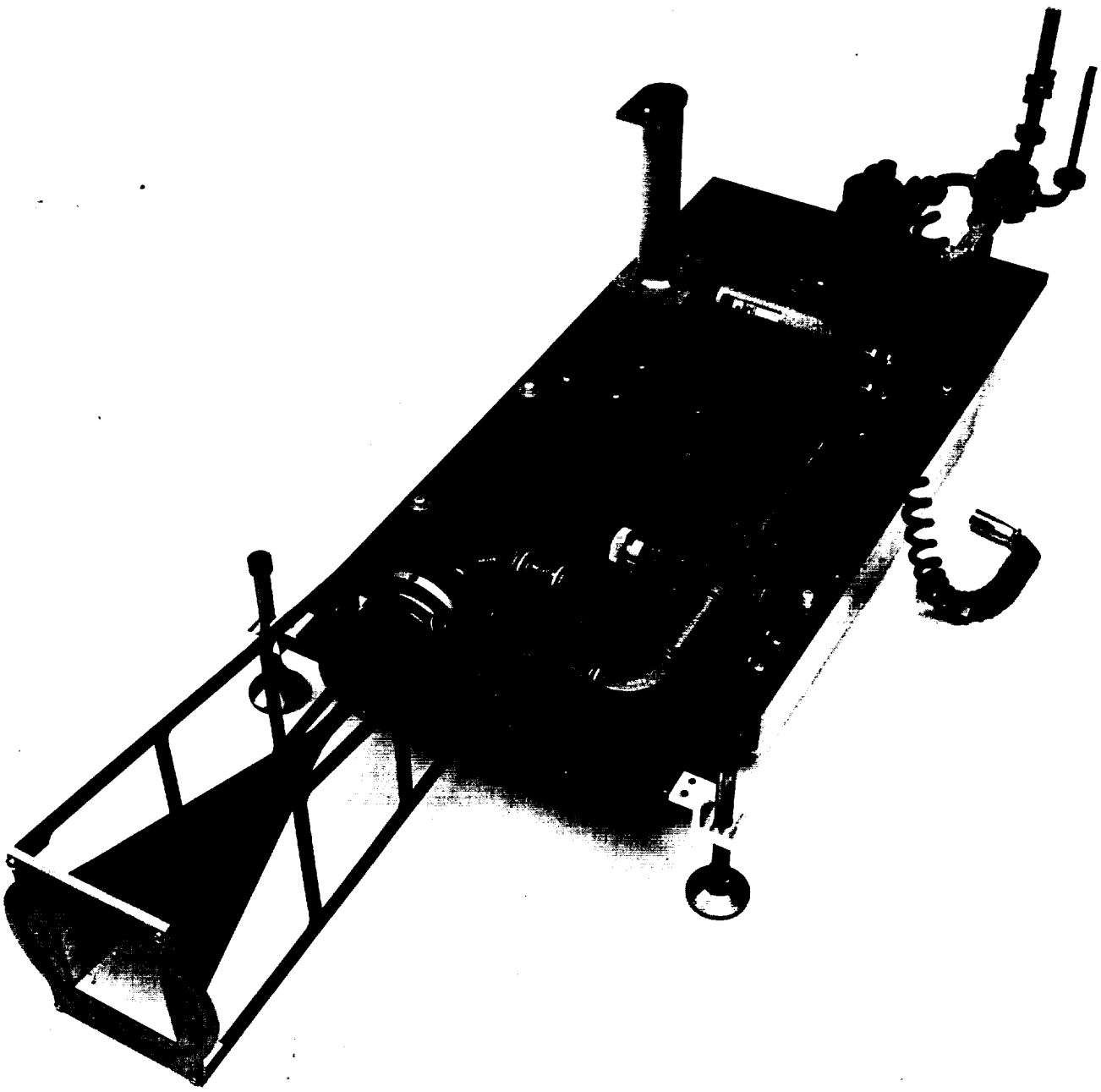


Fig. 1. Boresighting bench (over-all view)

approximately 20 mw, was the RF signal source. It was followed by an isolator, a level-setting attenuator, a waveguide switch that permits the power to route into a load when desired, a directional coupler that feeds a precision wavemeter and temperature-compensating bolometer for frequency and power monitoring, two attenuators, and a standard gain horn. One of the two attenuators before the horn was a precision attenuator which permits accurate output power level changes for linearity calibration purposes. A rotating step-twist was incorporated into the standard-gain-horn mounting structure to allow rapid polarization changes of the transmitted signal. Two

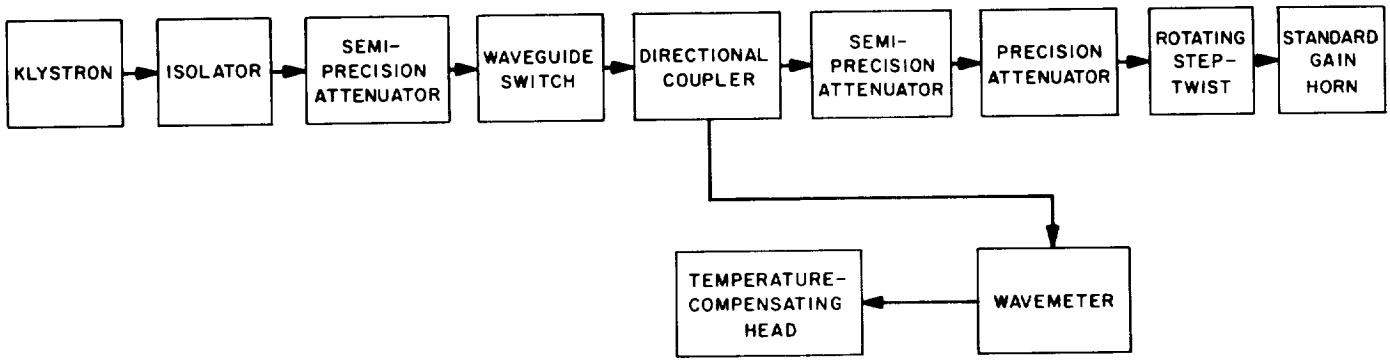


Fig. 2. Boresighting bench block diagram

klystrons were required to cover the 20- to 24-Gc band over which antenna patterns were run. Fig. 2 shows the block diagram of the system.

All the components, with the exception of the klystron power supply and power meter, and an electronic prime power regulator were mounted on a ½-in. aluminum plate for rigidity. Three leveling screw-jacks on the plate permitted fine adjustment of the output beam position. A telescope could also be mounted on the plate for visual collimation of the output beam.

The system performed its function adequately and had sufficient stability for accurate pattern measurements, the power output fluctuations not exceeding 0.02 db during several hours. High leakage of RF power from the klystron was observed, however, and all tests had to be run with maximum output power to insure that the relative level of leakage was negligible.

3. Procedure and Results

The first step in the measurements was to properly focus the feed horn on the antenna. This was accomplished at 21.9 Gc by observing the on-axis feed position yielding maximum output from the radiometer when the antenna was pointed at the boresight bench at Tiefort. All power levels and the radiometer performance were carefully monitored during this and subsequent measurements. The optimum point was found to be ⅜ in. nearer to the antenna than the theoretical position.

With the feed locked in proper position, the antenna was scanned both in elevation and azimuth at a rate of 0.004 deg/sec. The radiometer-receiver time constant was set at 1 sec and the output recorded on a 10-in. chart recorder. Each 0.01-deg displacement of the antenna was called out by the antenna operator and marked on the chart.

Two types of patterns were run in this manner. The first type was obtained by setting the radiometer gain so that the peak output on-axis was several db below compression. This allowed the main antenna lobe to be recorded in detail, but the dynamic range was insufficient to see the sidelobe structure. The second type pattern was run by allowing the radiometer to become saturated on the main lobe by increasing the gain. The sidelobes were thus recorded with fair detail.

Fig. 3 shows patterns obtained at 23.0 Gc. The following table summarizes the main characteristics at this frequency:

Scan plane	3-db beamwidth, deg	10-db beamwidth, deg	1st sidelobe, db
Azimuth	0.094	0.148	−15.0
Elevation	0.128	0.285	−18.5

Patterns were obtained at 21.5, 21.9, 22.5, 23.0, 23.5, and 23.9 Gc using the same technique. The patterns shown in Fig. 3 are representative of all the frequencies, with the exception that the sidelobes are somewhat lower near 21.9 Gc, at which frequency the feedhorn was focused, and also the azimuth and elevation patterns are more nearly alike.

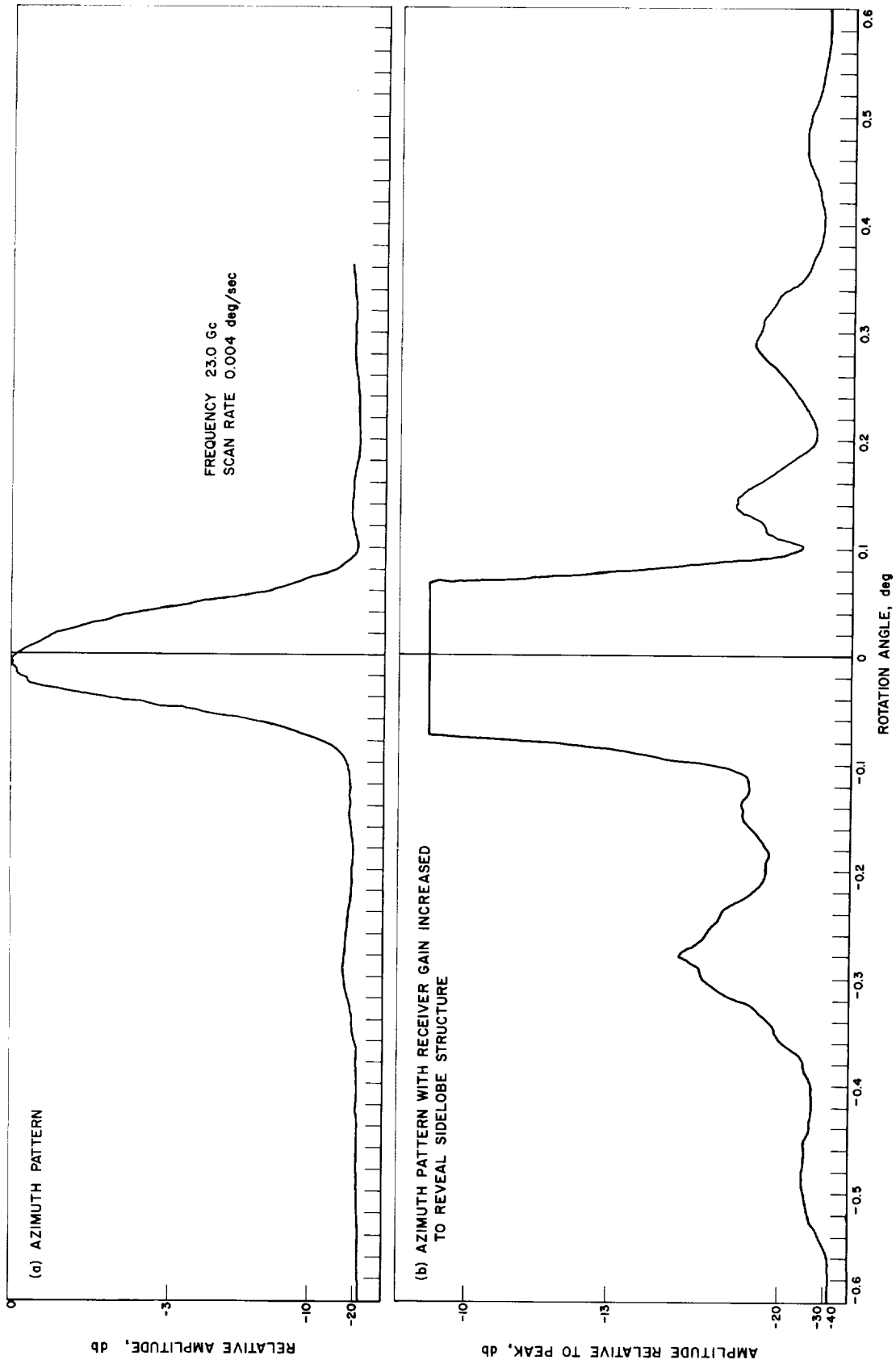


Fig. 3. Antenna pattern measurements in the 13-mm band on the Goldstone 30-ft antenna

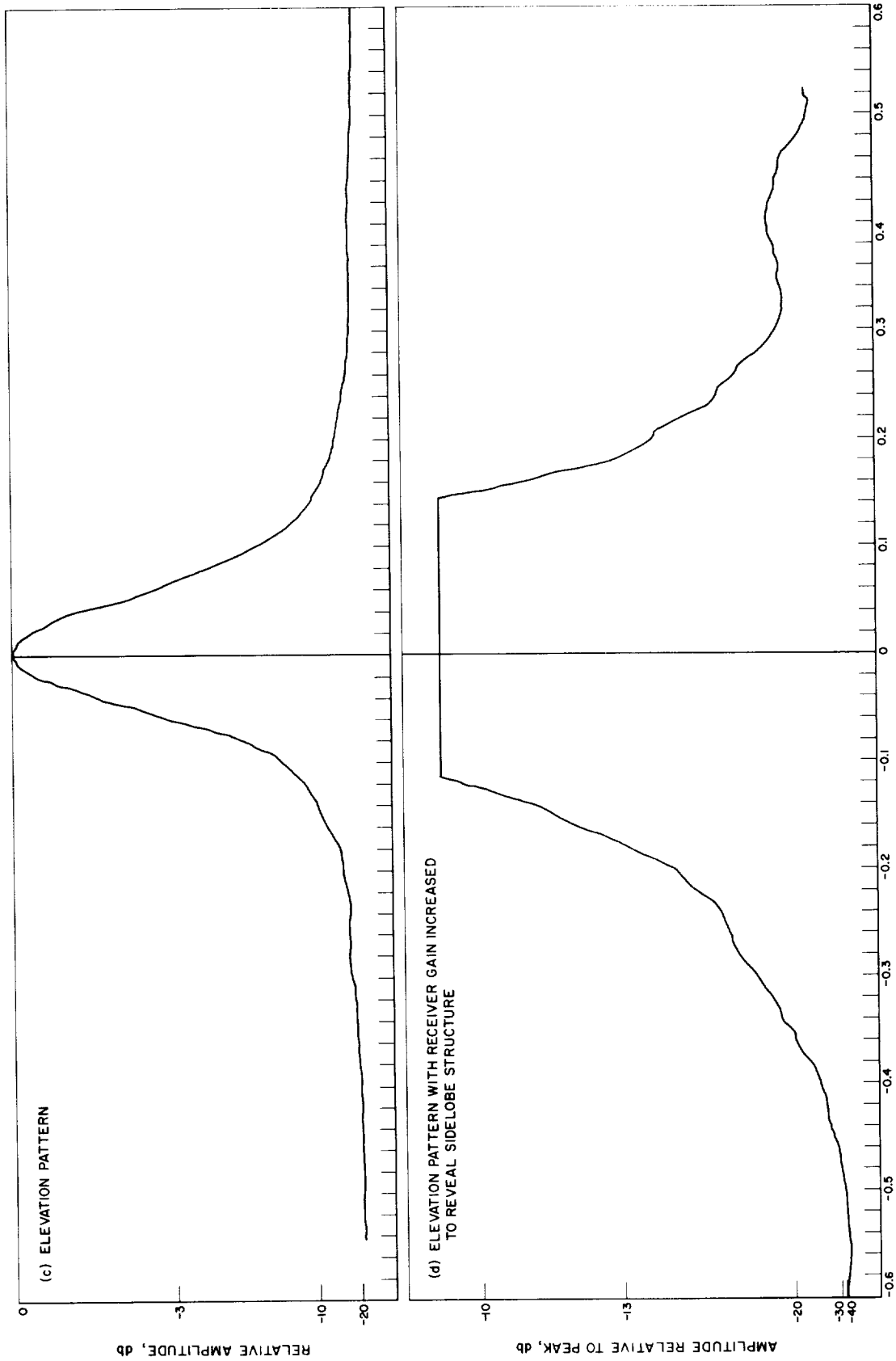


Fig. 3. Antenna pattern measurements in the 13-mm band on the Goldstone 30-ft antenna (cont'd)

B. A 13-mm Band Microwave Radiometer System for Earth-Based Venus Observations

E. J. Johnston and F. T. Barath

1. Introduction

The findings of the *Mariner 2* microwave radiometer experiment and of Earth-based radio astronomical measurements (Ref. 1) on Venus at 13.5 mm generated considerable scientific interest in further observations in this wavelength region. Thus, in early 1964 it was decided that an Earth-based program of high scientific value could be implemented in time to take advantage of the Venus near-conjunction period around June 18, 1964.

The program was to utilize a variable frequency radiometer to accurately measure the integrated Venusian brightness temperature at the water vapor line (13.5 mm) and to determine any spectral variations in the brightness temperature in a region as broad as possible around this line. A secondary purpose of the experiment was to establish the parameters of future experiments of this type, both Earth-based and spacecraft-borne, as well as to gain experience applicable to such instruments.

The instrumentation was designed, built, and tested in-house by the Radioscience Group, Space Instruments Section, and used highly successfully from June 29 to July 20, 1964, in conjunction with the Goldstone 30-ft antenna. Installation, calibration, and operation were done jointly with personnel from the Lunar and Planetary Sciences Section, the Communications Element Research Section, and the Communications Engineering and Operations Section.

This article describes in detail the basic radiometer system design and performance. A Technical Report, incorporating this and other aspects of the experiment, is in preparation.

2. Radiometer System

a. Basic system design and performance. The radiometer system used for the 13-mm band Venus measurements was a conventional superheterodyne receiver in a switching or Dicke configuration. The switching was performed between two primary antenna feed horns, one "signal" horn positioned at the focal point of the 30-ft antenna, and one "reference" horn positioned in the focal plane 3 in. below the focal point. In this manner, the sky temperature approximately 1.2 deg above the main beam was used as a reference source temperature. Since the signal contribution of Venus is only a few degrees, the requirement of nearly equal magnitudes between the signal and the reference sources was met. The use of two horns had the additional advantage of effectively cancelling contributions from varying atmospheric attenuation in as much as any such variations were observed by both horns and the radiometer measured only the difference between them.

Radiometers of this type have a basic sensitivity given by the equation:

$$\Delta T = k\frac{(FL-1)\,T_0}{(\beta\tau)^{1/2}}$$

where

ΔT = rms temperature fluctuation, °K

k = modulation factor, usually between 3 and 5

F = receiver noise factor

L = RF loss preceding mixer

T_0 = ambient temperature, °K

β = receiver predetection bandwidth, cps

τ = integration time, sec

Substitution of the radiometer's parameters in the above formula indicated a theoretical sensitivity of 0.67°K with a modulation factor of 4 and 10 sec of integration time. Subsequent calibrations yielded measured values of 0.49°K at the best and 0.96°K at the worst operating frequency, with the same integration time.

The radiometer was narrow band (10 Mc), but had the capability of being electronically and continuously tuned from 20.6 to 24.0 Gc. During the observation program, 10 discrete frequencies were chosen in the above band, and daily measurements were performed at as many of these 10 as possible.

Throughout the observations, calibration signals from a gas discharge noise source were injected into the system. The calibration signal magnitude at each frequency

was carefully determined in the laboratory by comparison with a heated standard load.

The output of the radiometer was recorded in both digital and analog form. The digital recordings were used for computer data reduction, while the analog strip-chart recording was used in real-time for determining proper antenna pointing and data quality. The analog recording was integrated for 1 sec in an R-C network; the digital integration was 45 sec.

During the observation period of June 29 to July 20, 1964, the radiometer system performed without any problems for over 500 hr. Prior and subsequent testing has increased the total number of operational hours to 814. It is felt that, with routine maintenance, several thousand hours of reliable operation is possible with this system and others of the same general design.

b. Receiver. Fig. 4 shows a simplified block diagram of the radiometer system. All the RF components of the receiver were mounted in a cylindrical shielded enclosure at the apex of the 30-ft antenna, as can be seen in Fig. 5. This arrangement kept lossy waveguide lengths at a minimum and aided in effective shielding against nearby RF sources. Provisions were made to remotely control a motor-driven wavemeter within the enclosure for frequency determination. Also remotely controlled were the calibration noise source and local-oscillator backward-wave-oscillator (BWO) frequency. Signal transmission to the main receiver console was at 60 Mc IF frequency. Power and control functions were transmitted to the RF enclosure via a fully-shielded multiconductor cable. The enclosure could be temperature-controlled to within 1°C with a transistorized proportional-control heating system.

The main receiver console, shown in the right center of Fig. 6, contained signal processing components from

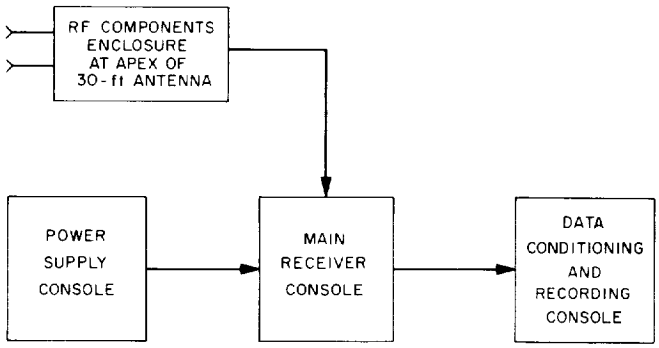


Fig. 4. Radiometer system simplified block diagram

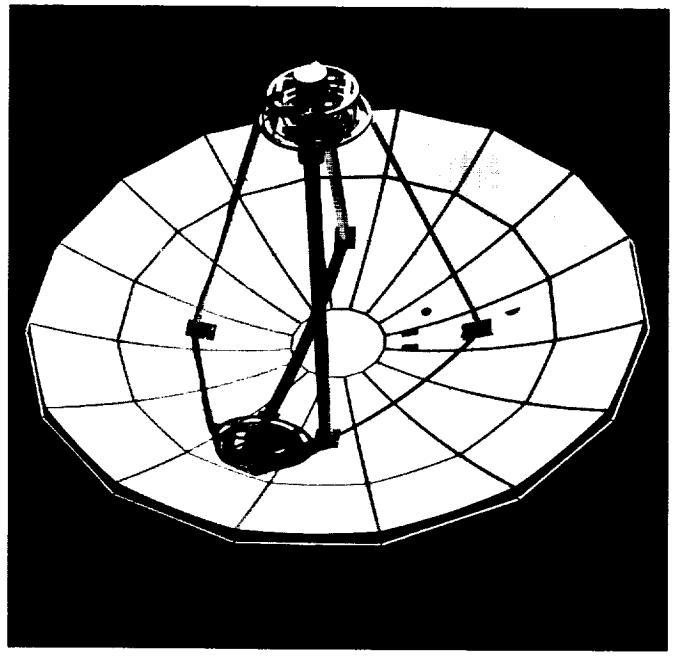


Fig. 5. RF components enclosure at apex of 30-ft antenna

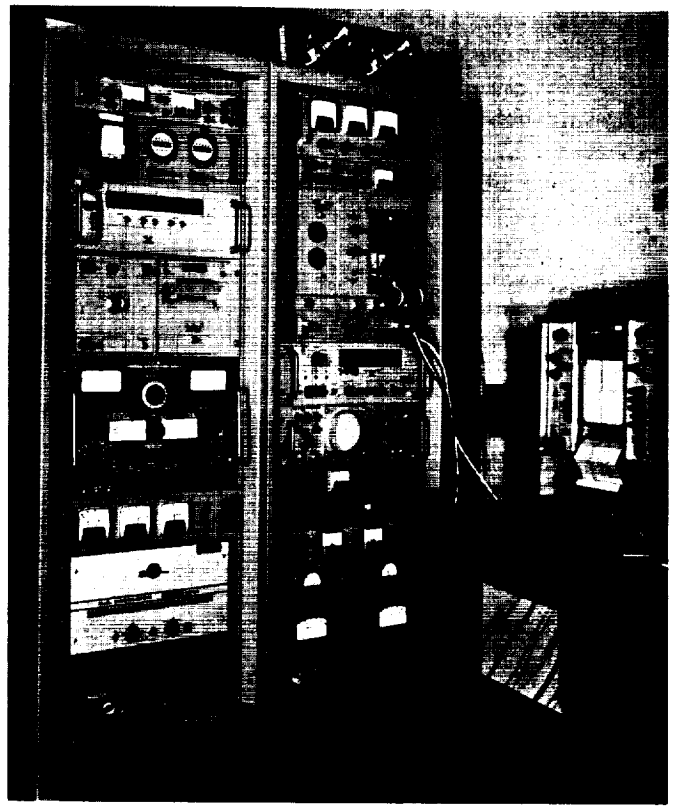


Fig. 6. Main receiver console, power supply console, and analog recorder

IF amplifiers through phase detection and dc amplification. Also included in this console were the remote controls for calibration noise-source actuation and wavemeter frequency adjustment. A digital voltmeter and oscilloscope were also available in this rack to permit monitoring various voltages and waveforms.

The power supply console (at the left in Fig. 6) contained slow and fast line regulators, power monitoring meters, the BWO power supply, digital printer and clock, running time meters, and an RF enclosure temperature monitoring recorder. The digital clock and recorder in this console were not used as primary data handling equipment but were included for system testing and back-up purposes. To the far right in Fig. 6 may be seen the analog recorder used for direct observation of data. Close examination will reveal a series of four noise-source calibration pulses near the center of the chart.

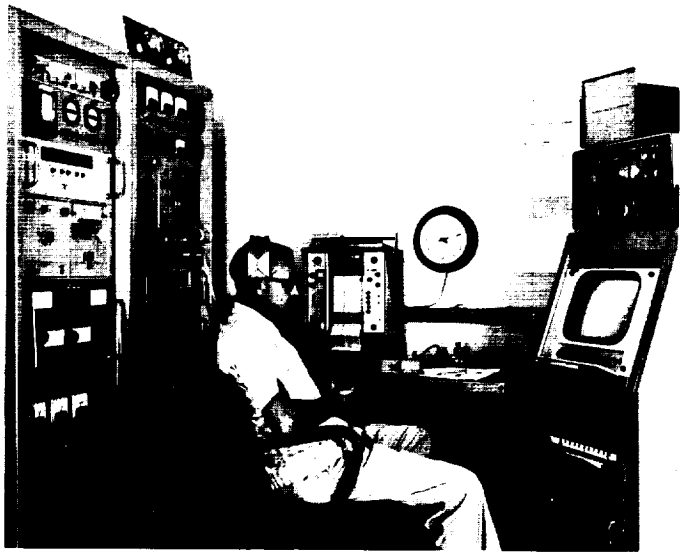


Fig. 7. Radiometer system operating position (over-all view)

An over-all view of the radiometer operating position is shown in Fig. 7. From left to right are the power supply console, main receiver console, analog recorder, WWV time standard receiver, TV monitor, and intercom. Not shown, in an adjoining room, was the data conditioning and recording console. This latter console was placed remote from the operating position because of the high noise level inherent in the paper tape punch used for digital data recording. The closed-circuit TV monitor displayed the information from a camera collimated with the antenna axis. Venus could be easily observed on the screen and antenna pointing was considerably eased. Thereby, correlation between the radio and optical axes allowed predictable pointing offsets to be computed. Typically, the radio and optical axes were less than 0.1 deg apart.

A detailed block diagram of the radiometer is shown in Fig. 8. The arrows interconnecting the blocks show signal path routes. Two hundred ft of signal and power cable connected the RF components enclosure and the main receiver console. The data conditioning and recording rack was separated from the receiver console by 40 ft of cable.

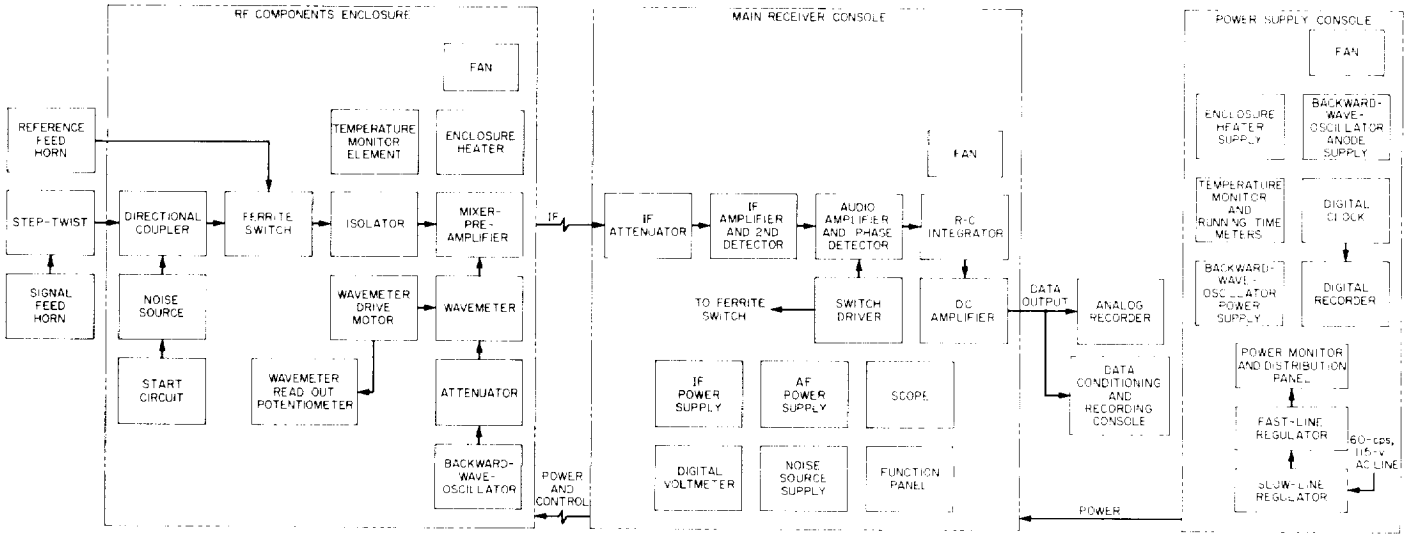


Fig. 8. Radiometer system detailed block diagram

The over-all measured radiometer performance characteristics were:

RF frequency range, Gc	20.6 to 24
Predetection bandwidth, Mc	10
Noise figure (single sideband), db	10.5
R-C integration times available, sec	0.1, 0.3, 1, 3, 10, 30, 100
Sensitivity with 10-sec integration, Gc at °K rms	20.6 at 0.71 21.0 at 0.49 21.9 at 0.60 22.5 at 0.96 24.0 at 0.71
Short-term stability, °K/hr	~1.0
Available system gain, db	170
Switching rate, cps	1100

The RF tuning range was primarily limited by the bandwidth of the ferrite switch used for switching between the signal and the reference feed horns (Fig. 8). A somewhat greater range might have been used; however, increasing insertion loss in the switch would have severely limited the radiometer sensitivity. Predetection bandwidth was determined by the post-IF-amplifier bandwidth. This amplifier was specially designed for radio astronomy purposes, exhibiting excellent gain stability characteristics.

System noise temperature was primarily determined by the noise figure of the mixer-preamplifier. Additional losses in the ferrite switch, isolator step-twist, and waveguides also contributed to the system noise temperature. Because of varying noise figure as a function of frequency, sensitivity calibrations were required throughout the band of interest. System sensitivity characteristics were actually measured every 500 Mc from 20.6 to 24.0 Gc.

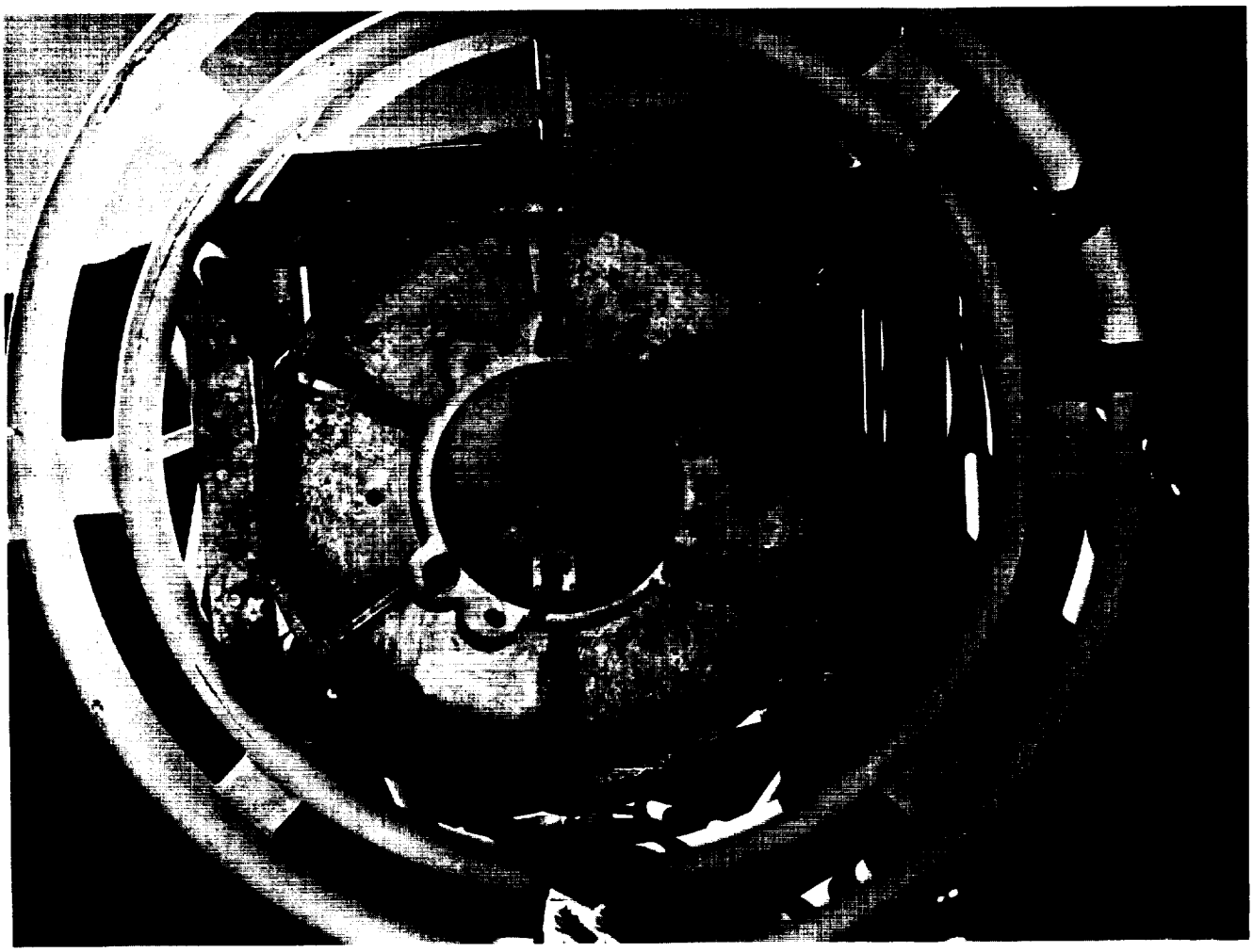


Fig. 9. Close-up of antenna apex structure showing the signal feed horn with the step-twist and the reference horn

A rather novel arrangement was used for rotating the polarization of the signal feed horn. A step-twist, consisting of a series of rotatable quarter-wave sections, allowed 0 to 90 deg polarization change. The device can be seen at the center of Fig. 9, which is a close-up view of the apex structure.

In addition to providing a convenient means for adjusting the vertical and horizontal polarization position of the feed horns, the rotation capability could also be used for polarization studies. These studies were not carried out during the course of this program, however, due to lack of time.

Following post-amplification and second detection, a tuned audio amplifier was used. This amplifier was capable of having its center frequency, gain, and bandwidth adjusted. In this case, the center frequency was set to the 1100-cps switching frequency, while the bandwidth was adjusted for 20 cps. The audio amplifier was followed by a phase detector in a balanced configuration for low offset voltage and minimum drift characteristics. The balanced arrangement was carried on through the integrator and dc amplifier, and yielded excellent over-all stability.

The low output impedance of the dc amplifier facilitated driving the various data recording instruments. A digital voltmeter, an analog chart recorder, and a voltage-to-frequency converter were used simultaneously on the dc amplifier output for data conditioning and recording purposes.

C. Film Scanner Using GaAs Light Source

A. Spitzak and J. D. Allen

In this article the characteristics of a particular type of film scanner system are investigated. The scanner's light source is GaAs light-emitting P-N junction, and the light sensor is a silicon planar photosensor.

The GaAs source is a Texas Instrument SNX-100 which is contained in a TO-18 can. The source radiates in the near infrared at 890 mμ. The sensor, a Texas Instrument LS 400, has its maximum response at about 980 mμ, and has about a 95% response at 890 mμ. A typical output of the light source is 2×10^{14} photons/sec when driven by a 100-ma current source. An expected current out of the sensor for this amount of light is about 100 ma.

The advantages to be obtained by utilizing the above devices in a film scanner are many. The source is extremely efficient and requires no high voltage to drive it. It is a solid-state device, extremely rugged, stable, and reliable. The sensor is also solid state, has a high quantum yield, and also requires no high voltages as do photomultiplier tubes. Both devices are very small and would be ideal choices for spacecraft-borne systems. The two devices are shown in Fig. 10.

It was necessary to determine what was the smallest scanning spot size obtainable at the film emulsion which could be utilized by the sensor and how a conventional film scanned by infrared compares with the results obtained by scanning with a light source in the visible range. Also, some of the characteristics of the sensor were investigated—such as its linearity and frequency characteristics at low currents.

A table that had been constructed for other scan tests was used in these tests. This table has two adjustable slits located at right angles to each other which are illuminated by a light source. This illuminated rectangle is imaged by a microscope objective (12.5/0.3) in space (Fig. 11). The minification of the system is approximately 18:1. The particular film pattern emulsion to be investigated is adjusted relative to this imaged spot until the two are together. In earlier work when the light was visible this coincidence could be observed by another microscope (Fig. 12). How this was done when the spot was not visible will be discussed later. The microscope views the emulsion and spot at 90 deg to the normal light path by the use of a mirror. Nikon microscope attachments allow a photograph to be taken of the emulsion; Fig. 13 shows such a photograph. The film is Ansco Super Hyscan, and the spot is approximately 0.003 in. square at the emulsion. The light source in this case is a Sylvania arc lamp. After set-up the mirror is removed from the light path, and the light transmitted by the emulsion is read by a photomultiplier tube. The film can be moved at a controlled rate perpendicular to the light path. This is done by driving an extremely accurate lead screw table by a hysteresis-synchronous motor. The speed of the motor can be controlled over a 10:1 range by varying the input frequency to the motor.

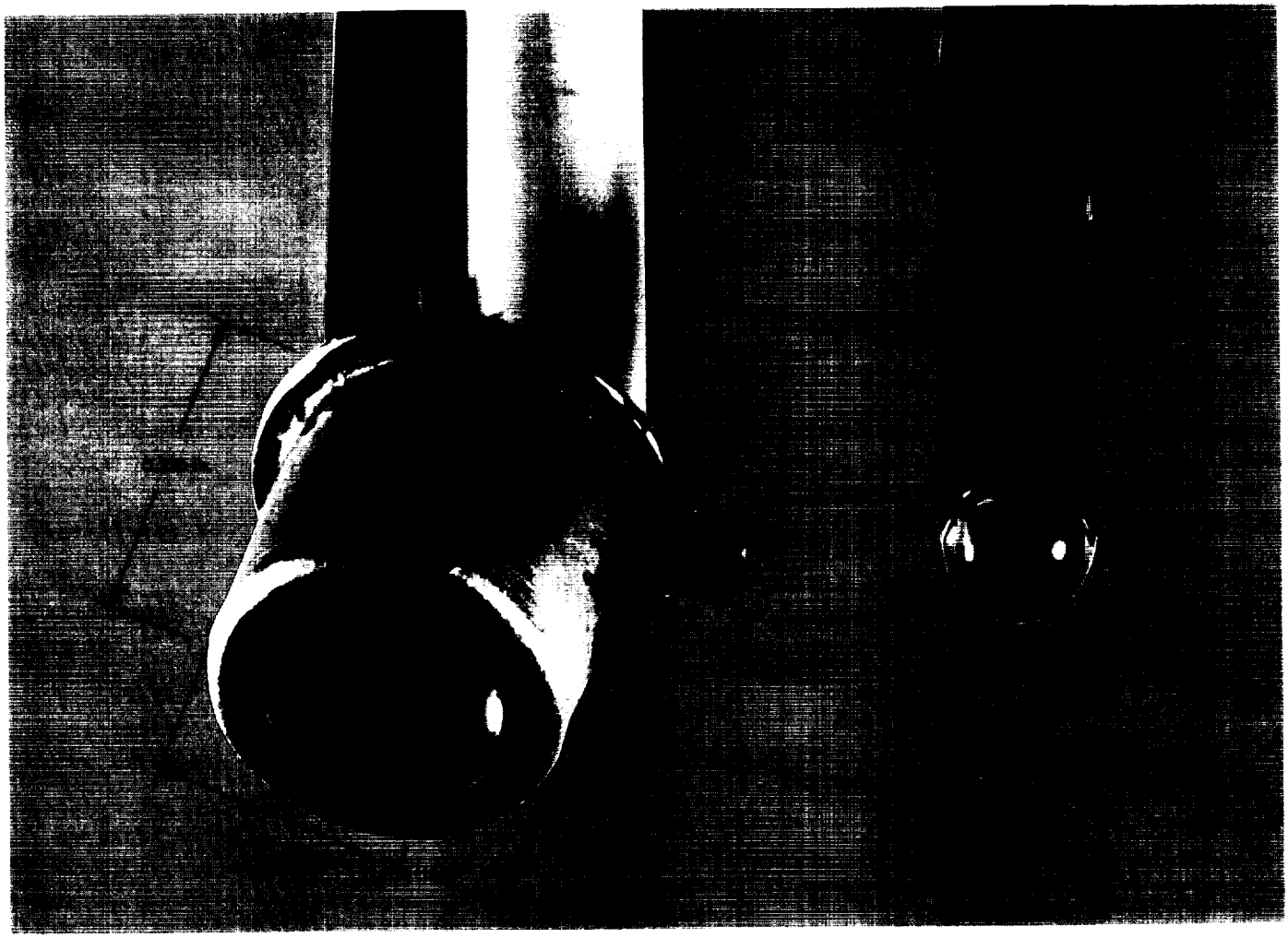


Fig. 10. Film scanner light source and sensor

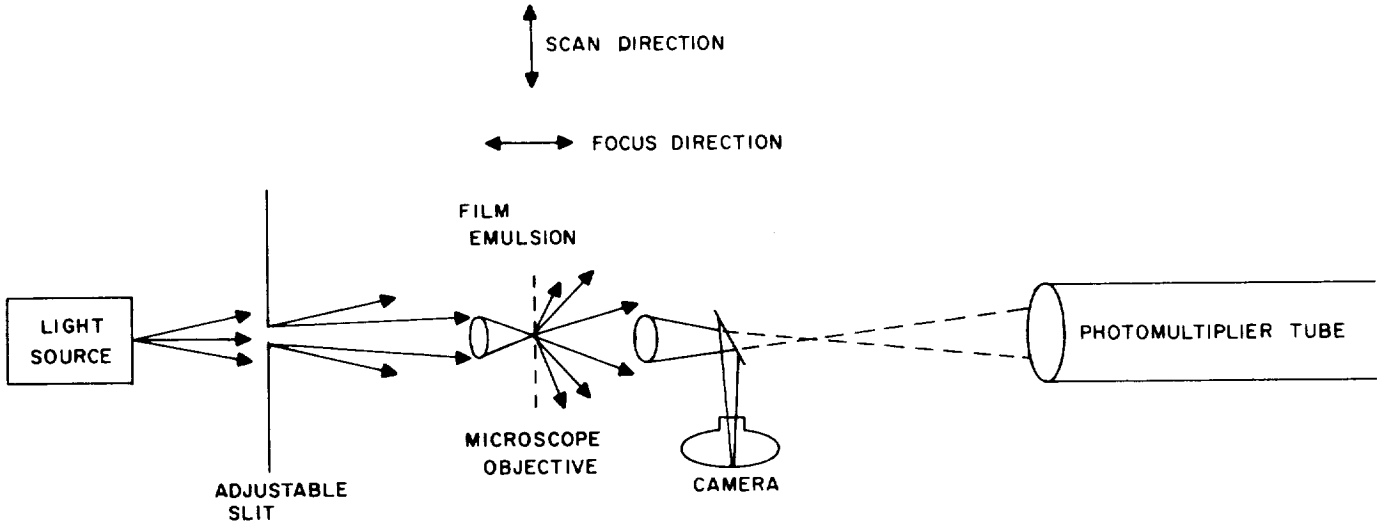


Fig. 11. Optical schematic

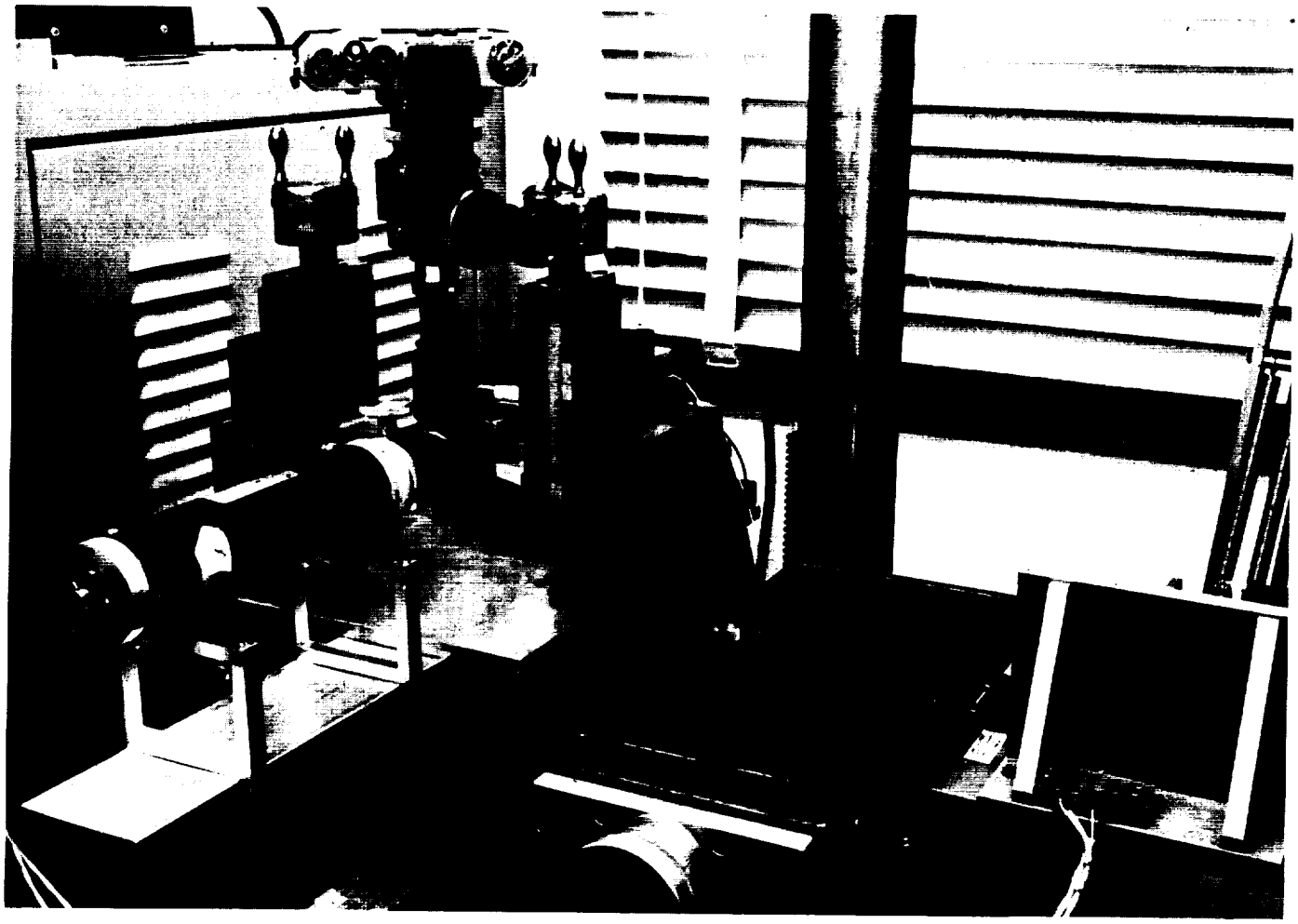


Fig. 12. Microscope setup for observing coincidence of emulsion and imaged spot

The two slits were calibrated for the given optics used in the system by measuring the distance travelled across a sharp density transition from full-on and full-off of the light transmitted. This was done with a very low-grain commercial Air Force test slide (AFTS). The calibration figure was about 0.0006 in./10 divisions of the slits. The arc lamp and the photomultiplier were then replaced by the GaAs source and the photosensor. This arrangement is shown in Fig. 14.

For these tests the sensor current was read by a Keithly 610A electrometer which was also used as an amplifier for oscilloscope display. Before display a low-pass filter was inserted in the system to minimize the high frequency pick-up problem at low currents. Earlier, without this filter and using sinusoidal GaAs currents (the GaAs source is capable of light modulation in the megacycle range), the frequency response of the sensor was found to be down 3 db at 8 kc when the amplifier input impedence was 2 KΩ. During the course of this investigation, some work was also done using a modulated light source which may make the design of an amplifier for the sensor easier. Future work will concern sensors having higher frequency response.

The sensor current versus light incident curve (Fig. 15) is obtained at low currents for two different biasing voltages (1 and 9 v). As the same slope and sensitivity are obtained in both cases, the lower biasing voltage is used in all later tests. The calibration is done using the inverse square law relation of light-incident-to-distance from the light source. In this test only the illuminated slit is used. About a 0.010-in.^{2} slit is illuminated by the GaAs source with 140 ma through it. This source current is chosen to give the initial sensor current at the first point, which is ½-in. from the slit. The maximum output is obtained at each distance from the slit by sensor alignment.

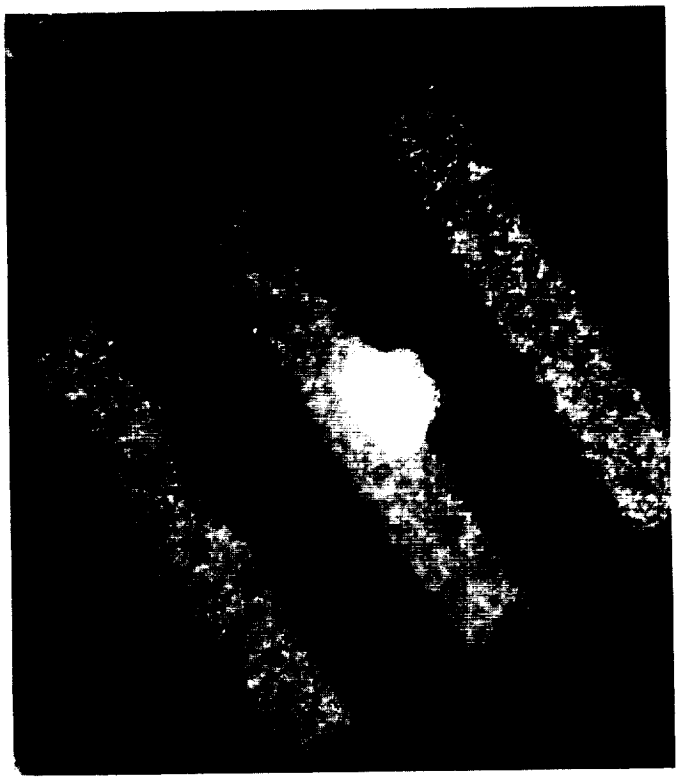


Fig. 13. Photograph of emulsion taken using Nikon microscope attachments

The sensor calibration for the 1-v bias is shown in Fig. 16, and a γ of 1.18 is indicated [current out $= C_1$ (light incident)$^\gamma$]. A further check on this value was made by using the known linear relationship between light output versus current of the GaAs source; the number obtained was similar.

Some AFTS are reimaged on Kodak-type 4400 aerial film with various background and bar densities. A typical bar pattern chosen for investigation is the series of six three-bar patterns shown in Fig. 17 which range in size from about 0.0025 to 0.0014 in. (Group 3). The negative diffuse densities for this print are 1.4 and 0.4 as measured by a MacBeth densitometer. To focus the infrared spot on the emulsion, the following technique is used. The emulsion is placed against a glass slide. A razor edge is also brought against the glass slide. The imaged spot is focused until the light-on-to-light-out transfer is made in the shortest distance. This number also experimentally reaffirms the earlier slit-calibration figure. The depth of focus for the particular optics used turned out to be about 0.001 in. This should be approximately the emulsion thickness, but the spot is also moved about the obtained figure to maximize the response from the actual emulsion.

When the sensor is used to read the light transmitted through the film, it reads only a narrow cone of the transmitted light. Since only a small portion of the light actually transmitted by the film activates the sensor, the measurement is specular. The MacBeth instrument which originally measured the density of the films during processing is a diffuse density-measuring instrument, i.e., it measures all the transmitted light. To verify that reading the film with the infrared source and a visible source is essentially the same, a number of different densities are read with both types of light, keeping all other components and parameters the same.

The readings as measured by the two systems are plotted against each other (Fig. 18). Similar readings would give a straight line at 45 deg. This data is not offered as being conclusive since the arc-lamp experiment was not as well controlled as the one which used the GaAs source. However, the lack of any great discrepancy indicates essentially the same reading between both systems. A large proportion of the energy from the arc lamp was in the infrared region, and later tests will have this component eliminated.

The square-wave modulation of the output signal is shown in Fig. 19. These are obtained using an AFTS printed on Type 4400 film whose output signal gives an effective contrast of 50:1. These contrasts are essentially specular measurements by the sensor. The spot sizes used are on the order of 0.00075 and 0.0015 in.^{2} The current through the GaAs is adjusted to obtain approximately the same change in output current in the sensor. These currents are 400 ma for the 0.00075-in.^{2} spot and 115 ma for the 0.0015-in.^{2} spot. The current output from the sensor went between 6×10^{-10} and 3×10^{-8} amp. A Polaroid of the 0.00075-in.^{2} scan between the square-wave line-pair (lp) numbers of 10.1 to 14.3 lp/mm is shown in Fig. 20.

In Fig. 21 a lower contrast (30:1) AFTS of the same group of line pairs (8.08 to 14.3 lp/mm) scanned by a 0.0015-in.^{2} spot is shown. To be more complete a large bar pattern (bars 0.010 wide) scanned by a 0.0015-in.^{2} spot is shown in Fig. 22; film grain noise is evident here.

With the knowledge gained about the potential of this system for film scanning, an automated system is being designed. In conjunction with the scanner, a data reproduction system is also being constructed. This is necessary because of the extremely low data rates that such a system will use; 20 cps will probably be the maximum system bandwidth expected.

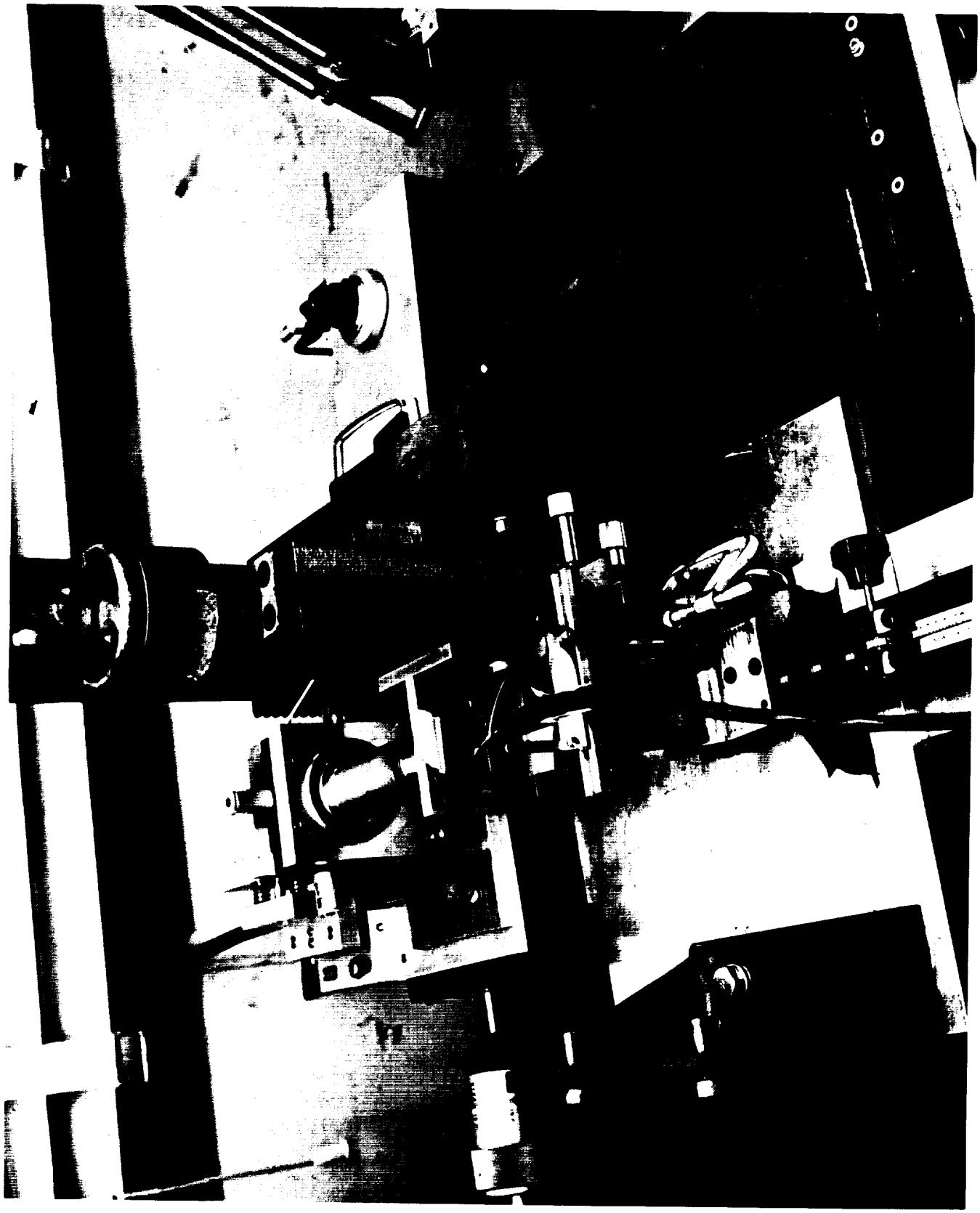


Fig. 14. Test arrangement for slit calibration

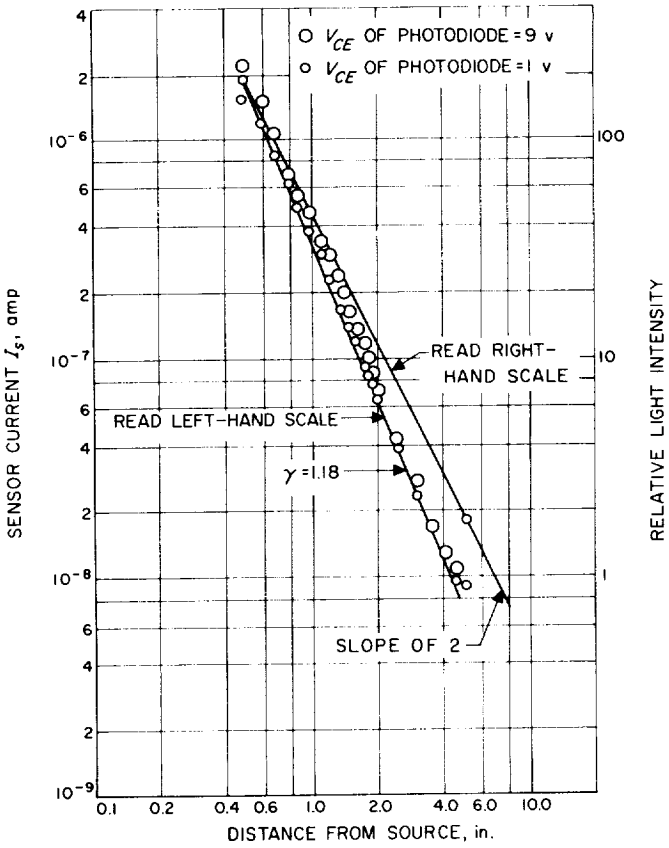


Fig. 15. Sensor current vs distance from source

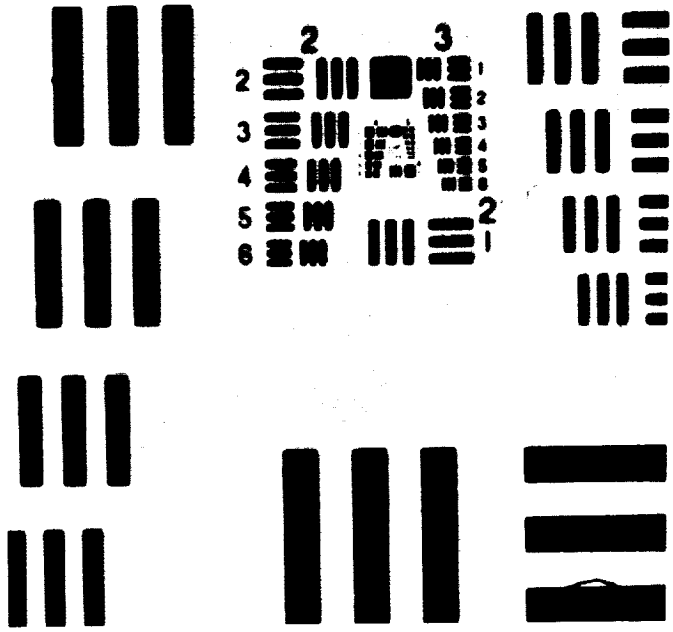


Fig. 17. Typical bar pattern

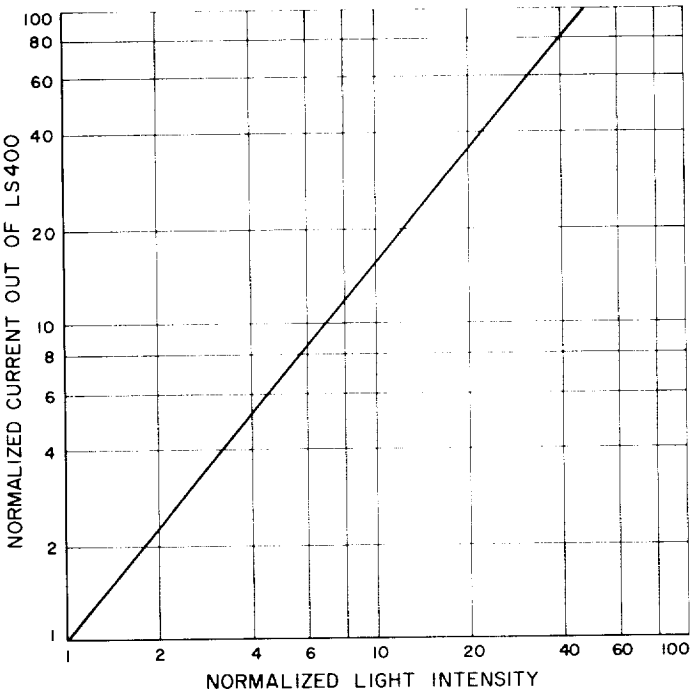


Fig. 16. Relative sensor output vs light incident

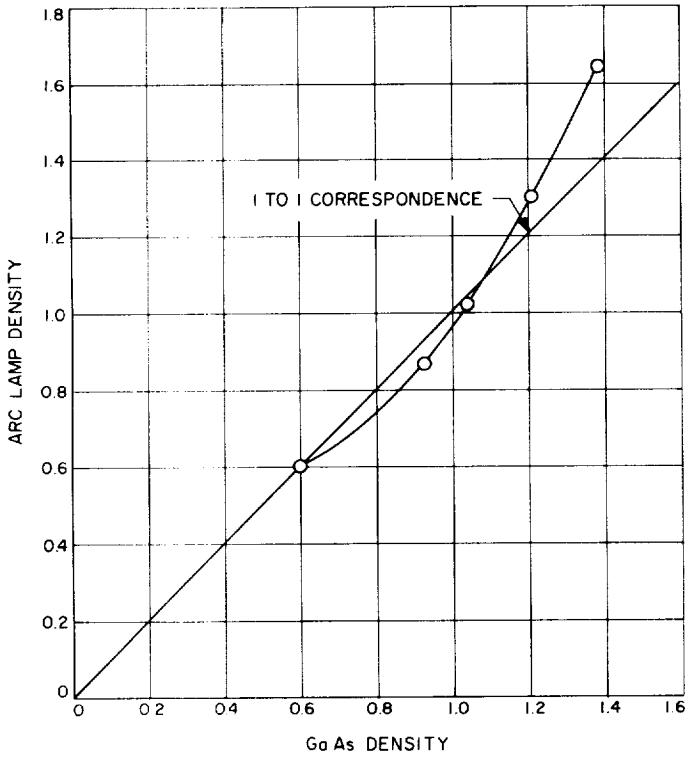


Fig. 18. Equivalent diffuse density irradiated by arc lamp to give same signal output when irradiated by GaAs source

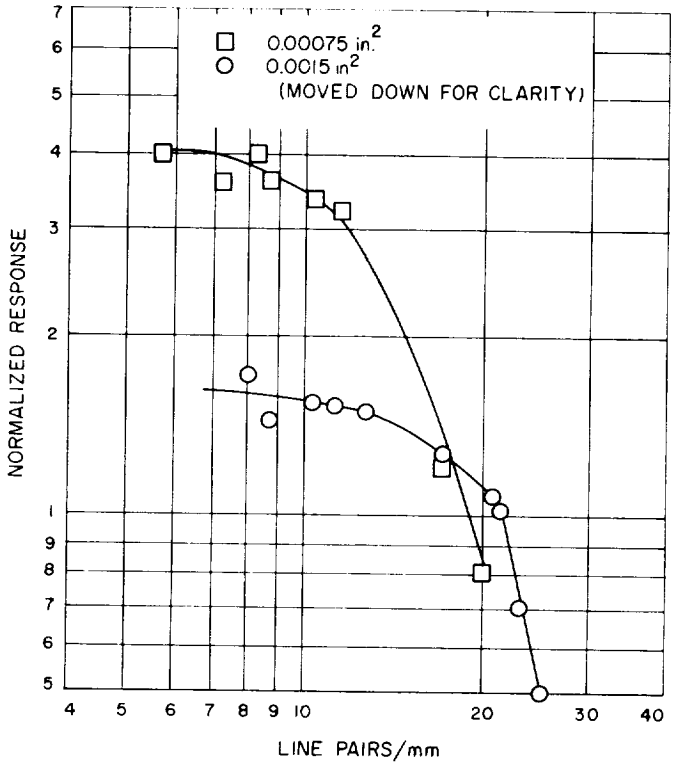


Fig. 19. Square wave modulation

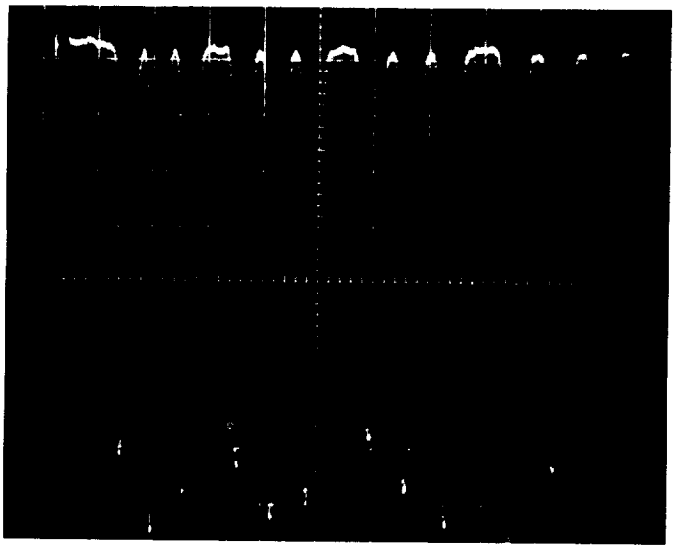


Fig. 20. Polaroid photograph of 0.00075-in.² scan between 10.1 and 14.3 line pairs/mm

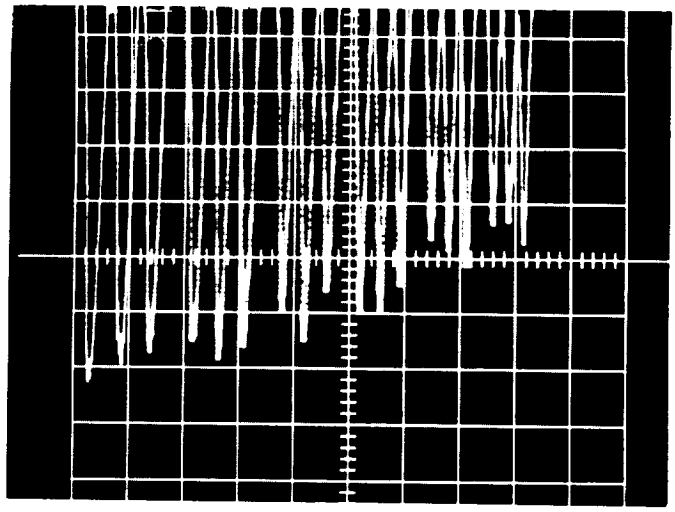


Fig. 21. Polaroid photograph of 0.0015-in.² scan between 8.08 and 14.3 line pairs/mm (30:1 contrast)

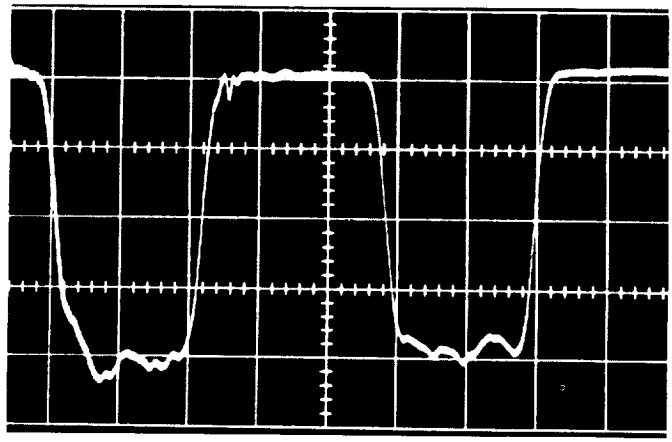


Fig. 22. Polaroid photograph of 0.0015-in.² scan of a large bar pattern

The most immediate requirement is to improve the light-gathering capability of the sensor. The constant γ of the sensor measured over a light-incident range of 100:1 and the expected continuation of this constant γ for two or three more decades indicates that this system could have a very large dynamic range. However, unless the maximum amount of light transmitted by the film is gathered, higher currents will have to be used in the GaAs source; and the dynamic range of the electronic system will have to be large. Better sensors are available to do this job and will be investigated. These sensors also have a higher frequency response, and modulation techniques will be studied at the same time. In addition, more efficient GaAs sources will be investigated.

D. Grain Noise Measurements of Pan X 4400

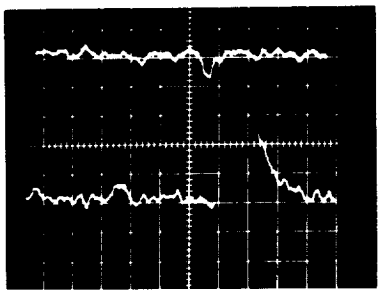
A. Spitzak and J. D. Allen

The use of Pan X 4400 for space reconnaissance is being considered. Some preliminary work was done using the optical setup described in the preceding article; the grain noise of the film was measured by the transmitted infrared light.

A grey scale is exposed on a piece of film along with an Air Force test slide for alignment purposes. The development is in D-19 for 5 min. The diffuse density measurement of the grey scale as measured by a MacBeth densitometer is between 0.07 and 1.68. The grey scale is then used in the specular measuring setup. Two scanning spot sizes used are square and of sides 0.00075 and 0.0015 in. The focus is adjusted to obtain maximum response from the system.

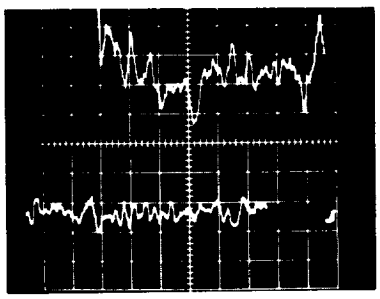
The resultant system output is shown in the accompanying set of photographs (Figs. 23 and 24). At the present time preparations are being made to record these signals on tape and to determine the frequency spectrum and amplitude distribution. This data will be used to pick the best scanning spot size for highest signal-to-noise ratio from the film scanner system and to determine the allowable variation of this spot size.

Using the available data, the value $\Delta I/I$ versus the previously measured diffuse density was plotted with much smoothing for both spot sizes (Fig. 25). (ΔI was obtained at this time with no consideration of the ampli-



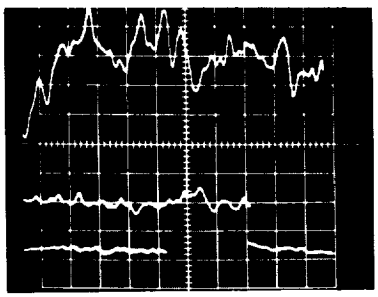
D	I, amp
0.07	7.0×10^{-8}
0.32	2.1×10^{-8}

$V = 1.0 \times 10^{-8}$ amp/cm



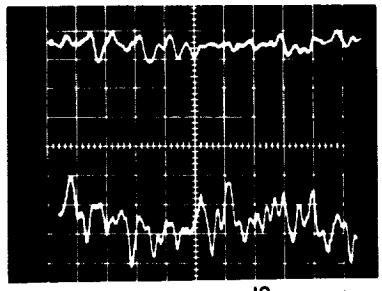
D	I, amp
0.32	2.1×10^{-8}
0.54	7.5×10^{-9}

$V = 2.5 \times 10^{-9}$ amp/cm



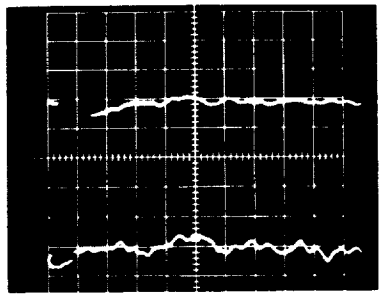
D	I, amp
0.54	7.5×10^{-9}
0.82	2.4×10^{-9}

$V = 1.05 \times 10^{-9}$ amp/cm



D	I, amp
0.82	2.4×10^{-9}
1.09	8.0×10^{-10}

$V = 2.63 \times 10^{-10}$ amp/cm

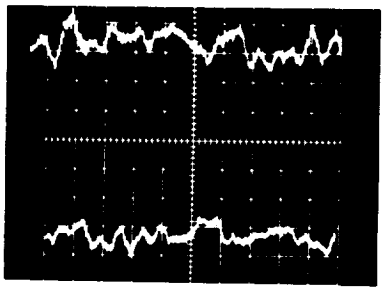


D	I, amp
1.09	8.0×10^{-10}
1.35	3.7×10^{-10}

$V = 8.7 \times 10^{-11}$ amp/cm

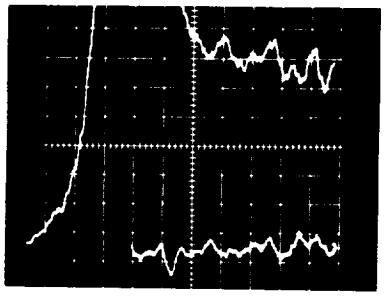
D = DENSITY
I = CURRENT
V = VERTICAL SENSITIVITY
BANDWIDTH = 35 cps
TABLE SPEED (HORIZONTAL SWEEP ON SCOPE) = 0.005 in./sec

Fig. 23. Pan X 4400 grain measurements (0.00075-in.^{2} spot size system)



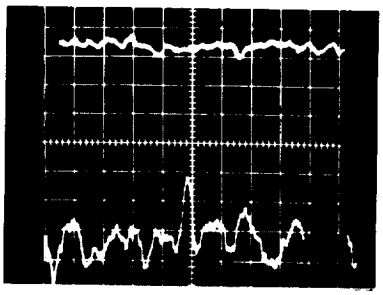
D	I, amp
0.07	3.2×10^{-7}
0.32	9.4×10^{-8}

$V = 1.6 \times 10^{-8}$ amp/cm



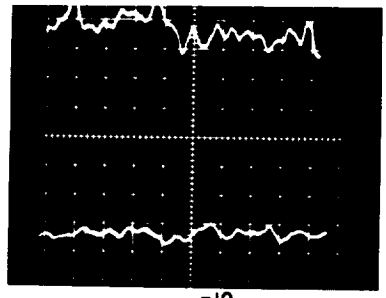
D	I, amp
0.32	9.4×10^{-8}
0.54	3.5×10^{-8}

$V = 8.2 \times 10^{-9}$ amp/cm



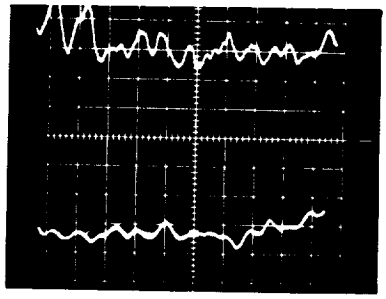
D	I, amp
0.54	2.8×10^{-8}
0.82	6.5×10^{-9}

$V = 2.5 \times 10^{-9}$ amp/cm



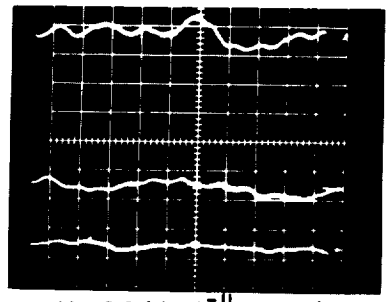
D	I, amp
0.82	6.5×10^{-9}
1.09	2.1×10^{-9}

$V = 6.6 \times 10^{-10}$ amp/cm



D	I, amp
1.09	2.1×10^{-9}
1.35	8.5×10^{-10}

$V = 2.0 \times 10^{-10}$ amp/cm



D	I, amp
1.35	8.5×10^{-10}
1.55	5.0×10^{-10}
1.68	3.3×10^{-10}

$V = 6.6 \times 10^{-11}$ amp/cm

D = DENSITY
I = CURRENT
V = VERTICAL SENSITIVITY

BANDWIDTH = 35 cps
TABLE SPEED (HORIZONTAL SWEEP ON SCOPE) = 0.005 in./sec

Fig. 24. Pan X 4400 grain measurements (0.0015-in.^{2} spot size system)

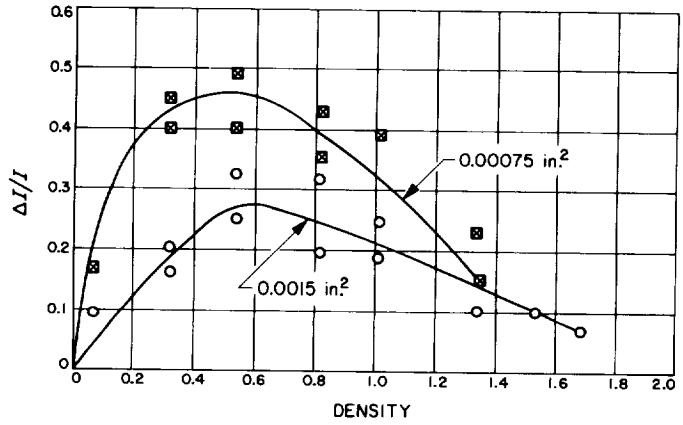


Fig. 25. $\Delta I/I$ vs diffuse density measurement for square scanning spots 0.00075 and 0.0015 in.^{2}

tude distribution by measuring the maximum excursion limits of the signal.) What is interesting to note is that the peak of the signal in both cases occurs at a diffuse density of about 0.6. If it is considered in this preliminary discussion that the transmitted light gathered specularly differs only by a constant ratio from that gathered diffusely (i.e., is not a function of density), then with

$$T = K_1 E^{-\gamma},$$

and since

$$\Delta T_{signal} \simeq \frac{dT}{dE} \Delta E,$$

where

$$\frac{dT}{dE} = \frac{\partial T}{\partial E} + \frac{\partial T}{\partial \gamma} \frac{d\gamma}{dE},$$

and considering that in the range under investigation at any time γ is a constant, then

$$\Delta T_{signal} \simeq -\gamma K_1 E^{-(\gamma+1)} \Delta E,$$

$$\frac{\Delta T_{signal}}{T} = \frac{-\gamma \Delta E}{E}.$$

Generally, it is desired that $\Delta E/E = K_2$ over the usable exposure range of the film. To be able to measure the signal so obtained, it must exceed the film noise contribution to the output signal. The film noise $\Delta T/T$ is greatest at a density of 0.6 and it is advisable that the highest film γ occur at this density so that the contribution $\Delta T_{signal}/T$ due to the exposure be the largest where that due to the film is the largest. Suitably lower γ's above and below the density of 0.6 can be used to obtain the desired exposure range.

Reference

1. Barath, F. T., *Symposium on Radar and Radiometric Observations of Venus during the 1962 Conjunction*, Technical Report No. 32-533, Jet Propulsion Laboratory, Pasadena, California, February 1964.

XVI. Chemistry

A. A Soldering Flux Which Leaves a Protective Film

A. J. Bauman

Electronic components to be exposed to high temperatures, as in spacecraft sterilization procedures, should not contain corrosive soldering flux residues. Commercial fluxes based on inorganic acid salts of hydrazine leave no residues. It occurred to us that such fluxes might be improved to leave a protective film if Schonhorn's method (Ref. 1) were applicable.

Schonhorn has described a method for bonding polyethylene to aluminum by means of an "oriented monolayer adhesive." The adhesive is a condensed monolayer of stearic acid which bonds chemically to the specially cleaned surface during application by the Blodgett (Ref. 2) technique.

It seemed likely that hydrazine salts of long-chain organic acids might heat-decompose to leave a coat of bound organic acid on the hydrazine reduced metal surface, and thus be good soldering fluxes.

We therefore made the diacid hydrazine salts (Ref. 3) of stearic and perfluoro myristic and perfluoro octanoic acids[1] (Ref. 4). All the salts were excellent fluxes for lead–tin solder on copper; however, the stearic acid salt left no film. The perfluoro acids left a residue which was contrastingly colored against the discolored oxide film of the copper and which resisted the action of 6N HCl vapor for about 12 hr. The contact angle of water on this "protective surface" was about 18 deg, measured by the tilting slide method. This suggests that the perfluoro acid coating is not an oriented monolayer and is quite porous. It may be worthwhile to repeat Schonhorn's work using a film balance and a perfluoro myristic acid monolayer on hydrazine to form an "oriented monolayer adhesive" for a Teflon–FEP protective coating. The new fluxes appear to be promising for space applications.

[1]We gratefully acknowledge gifts of perfluoro octanoic and perfluoro myristic acids from Dr. George Van Dyke Tiers of Minnesota Mining and Manufacturing Company.

B. The Microwave Spectrum, Structure, and Dipole Moment of 2,4-Dicarbaheptaborane (7)

R. A. Beaudet and R. L. Poynter

1. Abstract

The microwave spectra of all mono-substituted carbon and boron isotopic forms, and one di-substituted boron isotopic form of 2,4-dicarbaheptaborane (7) have been analyzed. The skeletal boron and carbon atomic coordinates determined from the rotational constants show that the molecule is a pentagonal bipyramid. The two carbon atoms lie in the pentagonal base plane and are separated by one boron atom. The two apex boron atoms appear to lie on, or very near the *c*-axis which is perpendicular to the pentagonal base. A molecular dipole moment of 1.32 $\pm$0.03 D is oriented along the planar symmetry axis.

2. Introduction

A number of carboranes have recently been prepared (Ref. 5). The structures of these compounds are poorly known since no detailed analyses have been made of their spectra at this time. While the structure of most of the carborane compounds will probably be determined by electron diffraction methods, a few of these compounds appeared to have rotational spectra amenable to microwave analysis. The microwave spectra and final structure of 2,4-dicarbaheptaborane (7), which belongs to the latter group, is reported here (Ref. 6). A preliminary communication of this molecular structure was recently given (Ref. 7).

Very little preliminary structure information was available. A molecular weight determination provided an empirical formula of $C_2B_5H_7$ (Ref. 5). Infrared and NMR measurements (Ref. 5) indicated that no bridge hydrogen atoms occurred and that there were two pairs of non-equivalent boron atoms in this molecule. A cage structure was proposed, but the carbon atom positions were unknown. Three possible structures can occur as shown in Fig. 1. Lipscomb and Hoffman (Ref. 8) used an extended Hückel MO theory to predict electron densities for these configurations. Their order-of-magnitude results slightly favored a model in which both carbon atoms were located in the pentagonal base and were separated by a boron atom. Each of the three possible structures would have a dipole moment but, depending upon the exact molecular configuration, the moment could be oriented along any of the three molecular principal axes or skewed with respect to them. Thus, the dipole selection rules could consist of one or more types which could not be ascertained beforehand. (See Fig. 1.) The bond lengths and angles could only be estimated by analogy with similar bonds in unrelated molecules. Initial spectral predictions were therefore uncertain.[2]

3. Experimental Methods

A 10-mg sample of 2,4-dicarbaheptaborane (7) was obtained on loan from Professor T. Onak. After each use the sample was carefully distilled back into the sample bulb in order to avoid sample loss. Professor Onak kindly prepared a similar quantity of the ^{13}C-substituted $C_2B_5H_7$. These samples had minor impurities present, primarily

[2]Lipscomb's study of dihydrocarborane had not been published when this work was started.

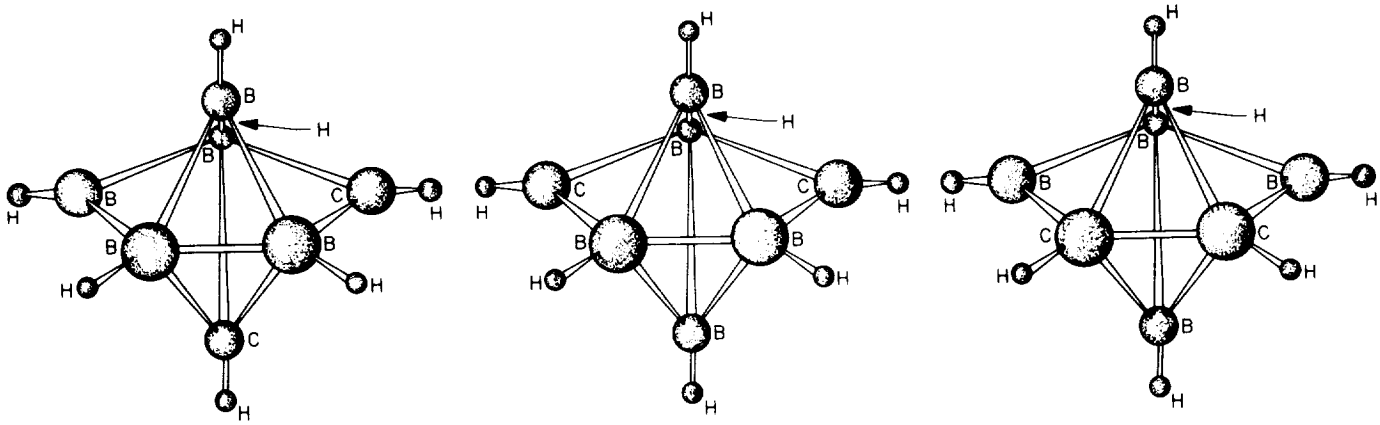


Fig. 1. Structures of carborane

dihydrocarborane, but were sufficiently pure (95%) that no serious trouble was encountered in their study.

The microwave spectrometer, which was of the standard 100-kc Stark modulation type, has been described previously (Ref. 9). The spectra was studied at dry ice temperatures, and at pressures of about 10–30 μ. Microwave frequencies were measured using interpolation receiver methods. The microwave frequency markers were generated by multiplication from a 1-Mc quartz-crystal-controlled oscillator, HP Model 104 AR, with a stability of ± 5 parts in 10^{10} and accuracy of the same order, as determined by comparison with WWV standard signals.

4. Spectra

Considerable trouble was expected in predicting and interpreting the observed spectra. A cursory examination of the spectrum showed an extremely large number of both strong and weak lines throughout the 8- to 40-Gc region. Spectra were predicted for each of the three possible structures. The spectra for each of these molecular models could have different types of spectral patterns, depending upon the selection rules. (See Fig. 1.) Furthermore, small changes in the assumed structures could change the rotor from a prolate to an oblate asymmetric top. Therefore, the predicted spectra could only provide a crude guide in the initial search.

Boron in natural occurrence consists of two isotopes, ^{10}B and ^{11}B with relative abundance of 19 and 81%, respectively. Thus, a large variety of isotopic species of $B_5C_2H_7$ will exist in sufficient quantity to be spectroscopically visible. The normal species, all ^{11}B atoms, accounts for 35% of the molecules. The B (1) and B (5) singly-substituted species each accounts for 16% of the molecules, while the B (3) singly-substituted species accounts for only 8%. Because of the large number of permutations, the B (1)–B (5) doubly-substituted species also makes up 8% of the molecules. The abundances of all other doubly- and more highly-substituted species are below 2%. Each of the five boron isotopic species mentioned above and one ^{13}C species was studied.

In addition, both boron isotopes have small nuclear quadrupole moments. As a consequence, the lower J-transitions will be broadened by an amount which depends upon the magnitude of the quadrupolar coupling in this molecule. Although such a broadening was observed, it was not large enough to give trouble except with the $J = 2 \leftarrow 1$ transitions whose line widths were 1 Mc. The $J = 3 \leftarrow 2$ transitions were used to calculate the molecular rotational constants.

Bearing in mind the above problems, the spectrum was scanned for strong low J lines with resolvable Stark effects. A number of these lines were located and tentative J assignments made. Sets of these lines were used to calculate rotational constants, which were then used to predict the low J transition frequencies. This process was repeated until an assignment was obtained which accounted for all transitions. The final assignment was made by selecting the three $3 \leftarrow 2$ transitions which were most sensitive to the three moments of inertia. This assignment indicated that the molecule was an oblate rotor ($\kappa = +0.54$) with *a*-type selection rules. A zero *b*-dipole was confirmed by Stark-effect measurements. Based on this structure and crude estimates of the bond distances and angles, prediction of the moments and frequencies of the single-substituted ^{10}B species could be obtained. The transitions of the isotopically-substituted species were located from these estimates and by comparing their Stark effects to those of the normal species. The rotational constants and moments of inertia determined from these assignments are given in Tables 1 and 2.

Table 1. 2,4-dicarbaheptaborane (7) rotational constants

Species	Rotational constants, Mc/sec		
	A	B	C
"Normal" (all ^{11}B)	4820.610	4586.817	3799.808
B (5)	4860.584	4632.607	3856.411
B (3)	4820.685	4670.139	3856.725
B (1)	4884.066	4644.363	3799.609
C (2)	4760.124	4578.088	3756.122
B (1)–B (5)	4925.179	4691.303	3856.241

Table 2. 2,4-dicarbaheptaborane (7)

Species	Moments of inertia, amu-Å^2		
	I_A	I_B	I_C
"Normal" (all ^{11}B)	104.8687	110.2139	133.0412
B (5)	104.0062	109.1245	131.0885
B (3)	104.8671	108.2475	131.0778
B (1)	103.5062	108.8483	133.0482
C (2)	106.2012	110.4241	134.5886
B (1)–B (5)	102.6422	107.7592	131.0942

One doubly-substituted ^{10}B species was also studied and assigned. This molecule, the most abundant of the doubly-substituted species, has transitions which are about the same intensity as the weakest singly-substituted species. Stark effects were resolvable for this species although they were quite weak. It was not practical to

study the other doubly-substituted species which have much weaker intensities.

The assignment was verified by predicting the $J = 4 \leftarrow 3$ transitions. A number of other moderately strong transitions were observed in addition to those belonging to this carborane. It is believed that these transitions are due to either another carborane, dihydrocarborane, present as an impurity, or to high J Q-branch transitions. No attempt was made to analyze these lines. Further purification was impractical without risking possible loss of the already small sample.

5. Structure

The spectra of 6 isotopic carborane molecules have been assigned. By using Kraitchman's equations (Ref. 10), as recommended by Costain (Ref. 11), the coordinates of each substituted atom were determined. These results are given in Table 3. The bond distances and angles derived from these coordinates are given in Table 4.

Table 3. Atomic coordinates in $C_2B_5H_7$

Atom		Coordinates, Å		
No.	Type	a	b	c
1	B	0	0	1.1647
2	C	0.4009	1.1828	0
3	B	1.3956	0	0
5	B	1.1204	0.8254	0

Table 4. Bond lengths and angles in $C_2B_5H_7$

Bond	Length, Å	Angles
C (2)–B (6)	1.5627	C (2)–B (3)–C (4) = 99°52′ ±30′
C (2)–B (3)	1.5455	B (3)–C (4)–B (5) = 116°51′
C (2)–B (1)	1.7077	C (4)–B (5)–B (6) = 103°13′
B (5)–B (6)	1.6508	B (1)–B (3)–B (7) = 79°42′
B (1)–B (5)	1.8146	
B (1)–B (3)	1.8177	

The structure corresponds to a pentagonal bipyramid, with the two carbon atoms separated by a boron atom and located in the pentagonal base plane. The skeletal structure is shown in Fig. 2, where the principal axes are denoted by arrows. As can be seen from Table 3, the B (1) and B (7) atoms appear to lie on the *c*-axis. From the symmetry of this molecule, there is no *a priori* reason for these atoms to be located on this axis. In earlier results (Ref. 7), they were reported as being off axis by ±0.1346 Å.

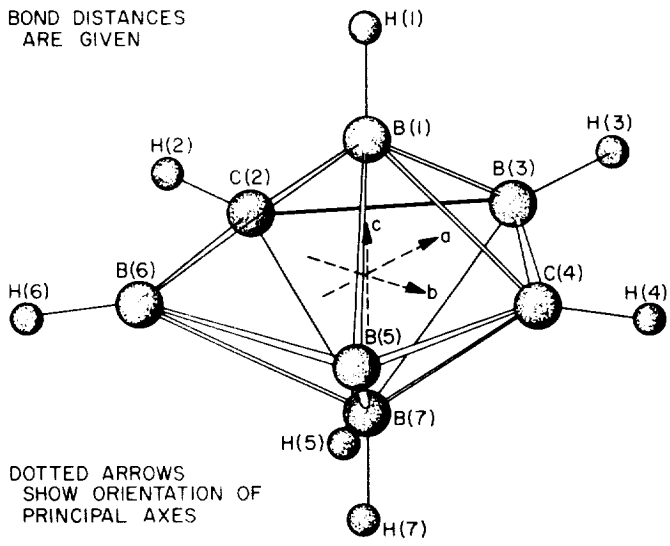


Fig. 2. Molecular structure of 2,4-dicarbaheptaborane, $C_2B_5H_7$

The doubly-substituted ^{10}B–^{10}B species was studied and assigned in the hope that Pierce's (Refs. 12 and 13) second difference method could be used to give a more accurate value for the B (1) coordinates. Unfortunately, the moment second differences were too small to give an improved set of coordinates: the magnitudes of the differences were of the same order as the experimental errors. A careful check of the rotational constants of all species showed that the assignment procedure had differed slightly for this isotopic molecular species. The use of an *a*-sensitive transition had been omitted. Re-examination of the spectrum showed that this was indeed an important consideration, for the discrepancy between the B (1) and the B (1)–B (5) isomeric carborane rotational constants and their predicted values for the *c*-coordinate of the B (1) atom could be removed. Using the new set of rotational constants for this species, the B (1) atom *c*-coordinate calculated from either of these isomers was 1.1647 Å. In addition, the *a*-coordinate for this atom was now reduced to practically zero, as calculated from either isomer. To the accuracy of our results, then, the B (1) and B (7) atoms are within ±0.01 Å of the *c*-axis.

6. Dipole Moment

The molecular dipole moment was determined by observing the Stark effect of several $J = 3 \leftarrow 2$ transitions. The only effect of the weak boron nuclear quadrupole moments was to produce a slight broadening of the low J transitions, an effect which rapidly decreased with increasing J, and was only barely noticeable for the $J = 3 \leftarrow 2$ transitions. The linearity of the plots of ν vs ε^2

for the Stark effect confirmed that the quadrupole coupling effect was negligible.

All measurements were made at dry ice temperatures. The absorption cell was calibrated with OCS immediately before and after the Stark measurements were made on the carborane. No *b*- or *c*-dipole component was observed, thus the dipole moment appears to be directed along the *a*-axis. These measurements are believed to be accurate within ±3%. The results are given in Table 5.

Table 5. Stark effect and dipole moment

Transition	Frequency, Mc/sec	M	$\Delta\nu/\Delta^2$	μ (Debyes)
$3_{22} \leftarrow 2_{21}$	25159.93	0	2.40×10^{-6}	1.30
$3_{12} \leftarrow 2_{11}$	25647.34	1	5.44×10^{-6}	1.34
$3_{02} \leftarrow 2_{02}$	23716.31	2	9.30×10^{-6}	1.33
			mean	1.32 ±0.03

7. Discussion

This is the largest molecule for which a skeletal structure has been determined by microwave spectroscopy. While heavier molecules have been studied by microwave spectroscopy, very few structures have been determined for these molecules. Preliminary bond distances and angles which were determined by other methods were available to assist in the spectral analysis of these other molecules. For dicarbaheptaborane, such preliminary information was not available, a fact which, while it caused considerable experimental trouble, provided a challenge to test the utility and power of microwave spectroscopy for the study of relatively large unknown molecules.

C. Induced Infrared Absorption of Solutions of H_2 and D_2 in Liquid Neon

G. E. Ewing[3] and S. Trajmar

1. Introduction

The induced infrared absorption spectra of liquid hydrogen (Ref. 14) and its solutions (Refs. 15–17) provide a means for studying molecular rotational energy levels and to a certain extent translational energy in the liquid state. A knowledge of these energy levels is important for the determination of molecular motion and intermolecular potential functions. It is also helpful to the understanding of quantum effects which have been shown to influence the thermodynamic properties of these systems (Refs. 18–22).

A study of H_2 and D_2 in solution with liquid neon provides a particularly interesting system for a test of the cell model interpretation of vibration-translation absorption features. The Lennard-Jones and Devonshire (hereafter called LJD) model of the liquid state which was originally derived for pure liquids (Ref. 23) can be applied, to a good approximation, to the binary H_2–neon (or D_2–neon) solutions because of the similarity in the solute and solvent Lennard-Jones 6–12 interaction parameters (Ref. 20). A direct comparison of the LJD potential with the spectroscopically-determined solute-solvent interaction potential can therefore be made. Furthermore, since the large quantum effects discussed for H_2–neon and D_2–neon solutions (Refs. 20 and 21) are related to the solute translational and rotational quantum levels, a spectroscopic determination of these quantities is therefore of interest.

2. Experimental

The 0.5-m low temperature absorption cell was similar to the one previously described (Ref. 17). In cases where phase separations occurred in the neon solution the use of a partition and mask eliminated possible error due to the superposition of spectra of two different phases. A number of holes were drilled in the partition so that equilibrium between the separated portions could be more easily achieved. In thermal contact with the sample tube was a compartment filled with liquid neon (27.2°K), which served as the primary coolant. A copper radiation shield cooled by liquid nitrogen surrounded the coolant compartment. The dewar vacuum and sample filling lines were flexible so that the cell could be rocked to hasten mixing of the H_2–neon (or D_2–neon) solution.

Spectra were obtained with a Perkin-Elmer Model 210 monochromator equipped with a PbS detector and a 640 line/mm grating for the H_2–neon experiments, and a thermocouple and a 240 line/mm grating for the D_2–neon experiments. Both gratings were used in first order and interference filters were used to eliminate unwanted orders. The light source, a tungsten lamp, was collimated through the absorption cell and focused onto the monochromator entrance slit. The chopper was placed just in front of the entrance window of the cell.

[3]Department of Chemistry, Indiana University, Bloomington, Indiana.

The region of the hydrogen absorption where spectral slit widths were about 1 cm^{-1} was calibrated using the overtone of CO, a mercury emission line (Ref. 24), and the absorption features of 1,2,4-trichlorobenzene (Ref. 25). The ν_3 fundamental of CH_4 and the ν_1 fundamental of HCN were used to calibrate the D_2 region (Ref. 24) where spectral slits were about 5 cm^{-1}. Frequency accuracy was limited by the uncertainty in determining peak maxima of the H_2 and D_2 features.

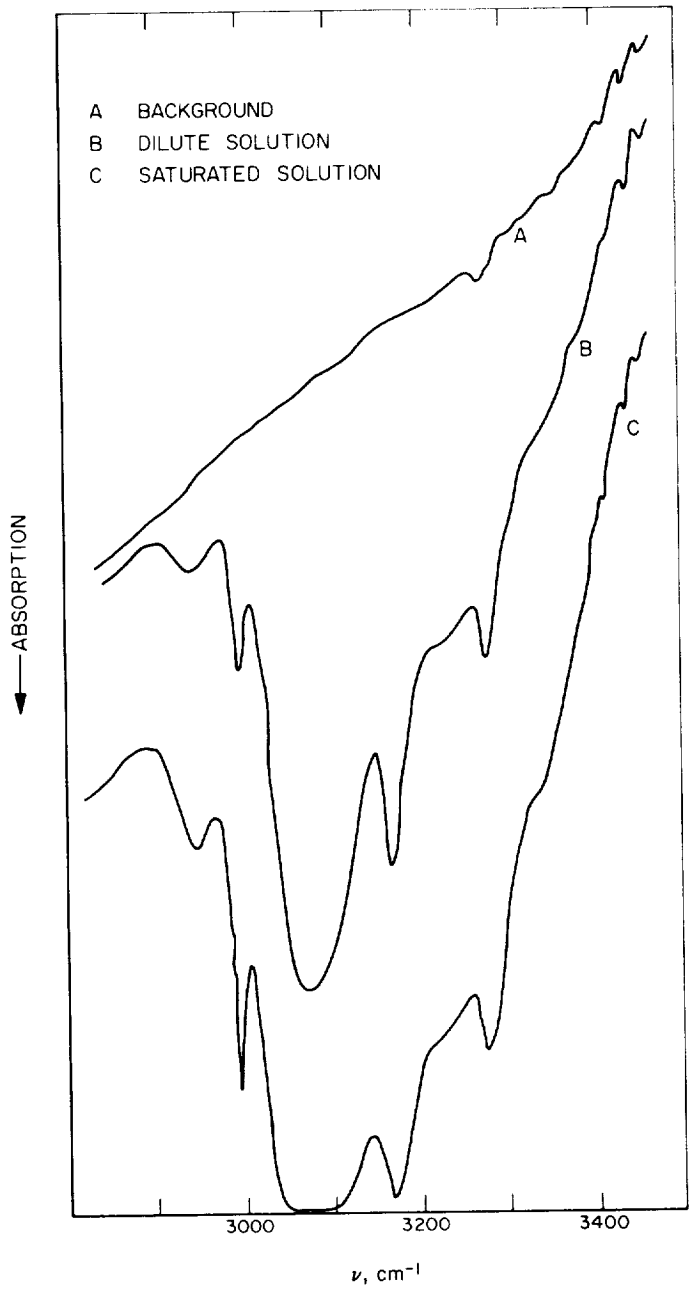


Fig. 3. Absorption spectrum of D_2 in liquid neon

3. Results and Observations

a. D_2 in neon. Neon gas was condensed into the cold absorption cell until the liquid just covered the metal partition. The background spectrum (Fig. 3, Curve A) revealed no absorptions in the neon. A weak feature at 3280 cm^{-1} is due to the absorption of the sapphire cell windows, and the features in the 3300–3500 cm^{-1} region are attributed to atmospheric water vapor absorptions. Deuterium was admitted under a pressure of about 3 atm until a small quantity of liquid D_2 had condensed over the neon. After an hour an interface was observed in the neon below the partition. This interface, which moved very slowly down into the neon, presumably represents the boundary between neon and neon saturated with D_2. A similar interface has recently been observed for H_2–argon and H_2–nitrogen solutions.[4] After rocking the cell a number of times to achieve better mixing, the interface and all Schlieren patterns disappeared. The spectrum of this saturated solution is shown in Fig. 3, Curve C. A dilute solution of D_2–neon which was obtained by pumping out some D_2 is shown in Fig. 3, Curve B. With an increase in solute concentration there is a general increase in intensity of all absorptions and a broadening of the 3278.6 cm^{-1} feature. The frequency error in determining the sharp absorptions, ($\nu_{1/2} \sim 10$ cm^{-1}) denoted S (0), S (1), and Q_Q, is ± 2 cm^{-1}. The weak broad feature labeled Q_P is determined to within ± 5 cm^{-1}, and strong broad absorption Q_R is uncertain to about ± 10 cm^{-1}. The poorly defined $S_R(0)$ and $S_R(1)$ absorptions are identified only as shoulders. These frequencies are listed in Table 6.

b. H_2 in neon. The background spectrum of pure liquid neon which was used for H_2 solution experiments is shown in Fig. 4, Curve A. Hydrogen under a pressure of 3 atm

[4]Holleman, G., unpublished results.

Table 6. Infrared absorption of D_2 in liquid neon[a]

D_2 in neon	D_2-gaseous[b]	Assignment
3340		$S_R(1)$
3278.6	3278.4	$S_1(1)$
3230		$S_R(0)$
3167.3	3166.2	$S_1(0)$
3078		Q_R
2992.8	2993.5	$Q_1(0)$ Q_4
	2991.4	$Q_1(1)$
2948		Q_P

[a]All frequencies are in cm^{-1}.
[b]Calculated from Raman data on low-pressure gaseous D_2, Stoicheff, B., *Canadian Journal of Physics*, Vol. 35, p. 730, 1957.

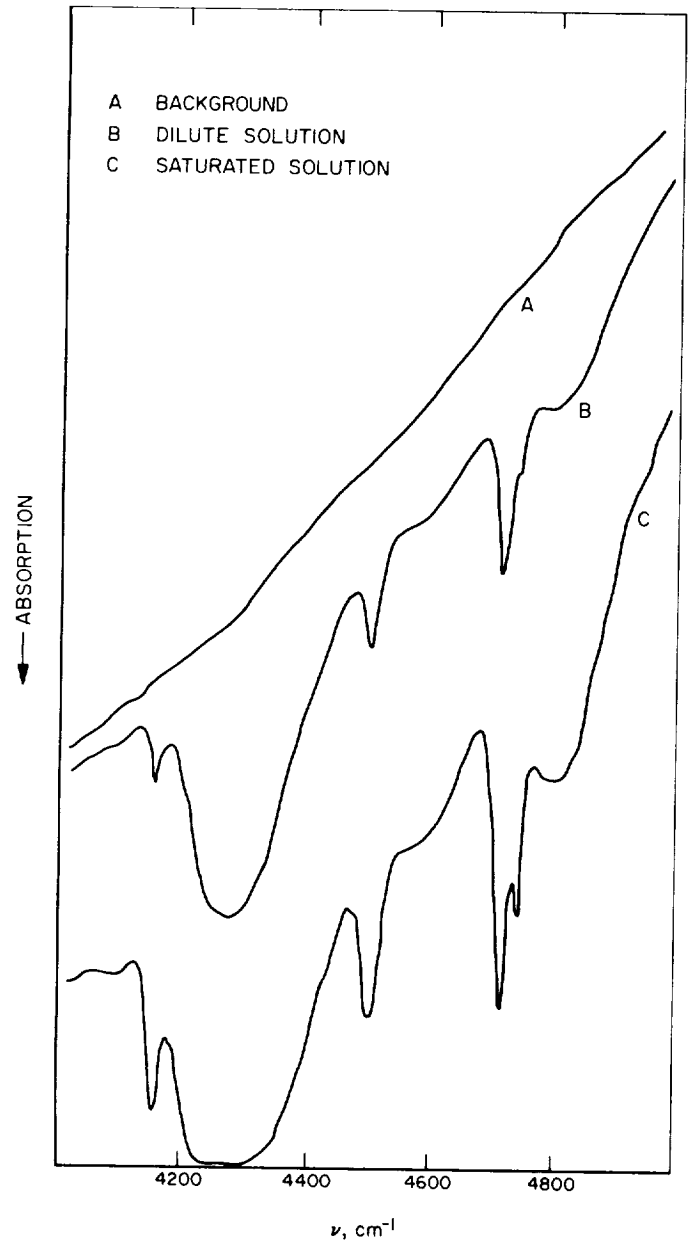


Fig. 4. Absorption spectrum of H_2 in liquid neon

was admitted to the pure liquid neon. Although Schlieren patterns indicated that the solution was not at equilibrium, no interface was observed. After rocking the cell a number of times, the spectrum shown in Fig. 4, Curve C was recorded. A more dilute solution of H_2 in neon was obtained by pumping out some of the dissolved H_2. This is shown in Fig. 4, Curve B. Two features, 4746.8 cm^{-1} and 4507.4 cm^{-1}, are very concentration dependent. In the dilute solution 4746.8 cm^{-1} is a shoulder of 4719.1 cm^{-1}. In Curve C 4746.8 cm^{-1} is well resolved, and 4507.4 cm^{-1} can be distinguished. A spectrum taken of a neon solution saturated with H_2 at 5 atm shows 4746.8 cm^{-1} and 4719.1 cm^{-1} of about equal intensity. A listing of all the observed absorptions is collected in Table 7. All features except the broad absorptions, $S_R(1)$, $S_R(0)$, Q_R, and Q_P, are estimated to be accurate to ± 2 cm. The weak but fairly well resolved $S_R(1)$ and Q_P features can be determined to ± 5 cm^{-1}. Determination of peak maxima of the weak $S_R(0)$ and very strong Q_R is difficult and the uncertainty is probably ± 10 cm^{-1}.

Table 7. Infrared absorption of H_2 in liquid neon[a]

H_2 in neon	H_2-gaseous[b]	Assignment
4823		$S_R(1)$
4746.8	4748.2	$Q_1(0) + S_0(1)$
	4742.3	$Q_1(1) + S_0(1)$
4719.1	4712.7	$S_1(1)$
4598		$S_R(0)$
4507.4	4515.5	$Q_1(0) + S_0(0)$
	4509.6	$Q_1(1) + S_0(0)$
4499.2	4498.7	$S_1(0)$
4280		Q_R
4160.2	4161.1	$Q_1(0)$
4154.8	4155.2	$Q_1(1)$
4107		Q_P

[a] All frequencies are in cm^{-1}.

[b] Calculated from Raman data on low-pressure gaseous H_2, Stoicheff, B., *Canadian Journal of Physics*, Vol. 35, p. 730, 1957.

c. Rotational energy levels. Sharp absorption features ($\nu_{1/2} \sim 10$ cm^{-1}) similar to the ones observed in H_2–argon (and D_2–argon) solution studies (Ref. 17) can be assigned to vibration-rotation transitions. A comparison of the solution phase absorptions with the calculated transitions based on gas-phase vibration-rotation energy levels and Raman selection rules ($\Delta J = 0, +2$) shows differences of 2 cm^{-1} or less (see Tables 6 and 7). In H_2–neon solutions even the $Q_1(0)$ and $Q_1(1)$ transitions are distinguished. Thus, within experimental error the rotatory energy levels for neon solution and gas-phase H_2 and D_2 are the same.

The two concentration dependent absorptions in the H_2–neon solution spectra are assigned to double transitions. One type of transition which occurs in the $S(1)$ region is $Q_1(0) + S_0(1)$. The suffix added to the usual spectroscopic notation indicates the change in vibrational level, Δv, of the molecule. The double transition $Q_1(1) + S_0(1)$ overlaps the $Q_1(0) + S_0(1)$ absorption. The assignment of simultaneous transitions in the $S(0)$ region is shown in Table 7.

No structure is observed in the S (0) and S (1) regions for D_2–neon solutions which could be assigned to double transitions. A broadening of S (1) with increasing D_2 concentration suggests, however, that this absorption is a blend of S_1 (1) and Q_1 (0) + S_0 (1) and Q_1 (1) + S_0 (1) transitions. The difference in frequency between single and double transitions in the S (0) region for D_2 is smaller than in the S (1) region, and thus they are more difficult to distinguish.

d. Translational energy levels. It is convenient to discuss the vibration-translation changes in H_2–neon (or D_2–neon) in terms of a simple model of the liquid state. The cell model which has been used recently to interpret the infrared spectra of dilute H_2–argon (and D_2–argon) solutions (Ref. 17) will now be applied to the neon solutions. The motion of the light solute molecule which is confined to a cell of molecular dimensions is best considered quantized, and its translational energy levels will be identified by the quantum number n. The second type of motion involves translation of the relatively heavy solvent molecules which make up the walls of the cell.

As H_2 (or D_2) moves toward the walls of its cell, overlap interactions cause a small dipole moment to be set up in the system. Motion of the solvent which results in an aspherical cell also produces a dipole moment. The combination of the fundamental molecular vibration of H_2 with the changing dipole caused by its translation or the motion of the solvent (cell) is therefore infrared-active. On the basis of these liquid-state interactions broad features observed in H_2–neon may be assigned to vibration-translation transitions involving changes in translation energy of either the solute or solvent.

Since previous studies of H_2-solution systems (Refs. 17 and 19) indicate that translational energy level spacing of the solute is of the order of 100 cm^{-1}, the feature Q_R will be assigned to the vibration-translation transition ($v' = 1$, $n' = 1 \leftarrow v'' = 0$, $n'' = 0$). Thus, the H_2 molecule on absorption of a photon increases its internal vibration and also increases its translational quantum level within its cell. An estimate of the translational energy level spacing is provided by subtracting the vibrational fundamental from the vibration-translation combination band: Q_R ($v' = 1$, $n' = 1 \leftarrow v'' = 0$, $n'' = 0$) − Q_Q ($v' = 1 \leftarrow v'' = 0$) = $E_1 - E_0$, where E_1 and E_0 are the energy of the first excited and the zero translational energy level, respectively. This value, $E_1 - E_0$, is found to be 122 cm^{-1}, which is in order of magnitude agreement with previous determinations of translational energy level spacing of the solute. The selection rules for a vibration-translation combination: $\Delta v = +1$, $\Delta n = +1$, $\Delta J = 0$ are consistent with this assignment (Ref. 17). Weaker absorptions which may be assigned to transitions involving both translational and rotational changes of the solute are also observed in H_2- and D_2–neon solutions: S_R (0), where ($v' = 1$, $J' = 2$, $n' = 1 \leftarrow v'' = 0$, $J'' = 0$, $n'' = 0$); and S_R (1), where ($v' = 1$, $J' = 3$, $n' = 1 \leftarrow v'' = 0$, $J'' = 1$, $n'' = 0$). To the approximation that the solute-solvent overlap interactions are independent of the angular orientation of H_2 about its center of mass, these transitions are forbidden (Ref. 17). However, it has been shown theoretically (Ref. 26) that there is a small angular dependent contribution to the induced dipole moment from overlap interactions, and this relaxes the selection rule for the translating molecule. A summary of translational energy level spacing is collected in Table 8.

Table 8. Translational energy levels of H_2 and D_2 in neon[a]

	H_2 in neon	D_2 in neon
$Q_R - Q_Q$	122	85
$Q_Q - Q_P$	50	45
S_R (0) − S (0)	99	60
S_R (1) − S (1)	104	60

[a] All frequencies are in cm^{-1}.

There is a large change of $Q_R - Q_Q$ from 122 cm^{-1} for H_2–neon to 85 cm^{-1} for D_2–neon. There is also a shift of S_R (0) and S_R (1), but these features are so poorly defined in D_2–neon that a quantitative evaluation is not possible. These shifts are further evidence that the translational levels observed spectroscopically as the broad combination bands Q_R, S_R (0), and S_R (1) are related chiefly to the motion of the solute. The magnitude of the Q_R isotope shift makes possible an examination of suitable potential functions which could describe the solute-solvent interaction. A number of intermolecular interaction potentials have been suggested for the liquid state. The simple harmonic oscillator and hard sphere box potentials have been used (Ref. 27), as well as more detailed descriptions like those of Lennard-Jones and Devonshire (Ref. 23).

The harmonic oscillator potential, $V = \frac{1}{2}kr^2$, where r is the distance of the solute from the center of its cell and k is the interaction force constant, appears to describe the solute-solvent interaction satisfactorily. The energy of the nth translational level is given by (Ref. 28)

$$E_n = \left(n + \frac{3}{2}\right)\frac{1}{2\pi c}\left(\frac{k}{M}\right)^{1/2}$$

The translational energy level spacing for D_2 in neon is predicted by this potential to be $(\frac{1}{2})^{\frac{1}{2}}$ less than that of H_2 in neon. The value of $(Q_R - Q_Q)_{D_2}$, which is calculated to be $(\frac{1}{2})^{\frac{1}{2}}(Q_R - Q_Q)_{H_2} = (\frac{1}{2})^{\frac{1}{2}}(122) = 86\ cm^{-1}$, is in good agreement with the experimentally determined quantity: $(Q_R - Q_Q)_{D_2} = 85\ cm^{-1}$.

It is interesting to compare the harmonic oscillator potential which was found consistent with the spectroscopic data with the potential derived from the LJD model of the liquid state. This model was devised for the determination of thermodynamic properties of pure liquids. It is assumed that each molecule moves in a cell which consists of z nearest neighbors. The erratic positions of these neighbors is accounted for by assuming that they are smeared out over a sphere. The radius of the sphere a is determined from the average distance between nearest neighbors in the liquid. This potential is given as a function of the displacement r of the molecule from the center of the sphere:

$$V(r) = \tfrac{1}{2} z\,|\varepsilon^*| \left\{ \frac{-r^{*6}}{2a^5 r}\left[\left(1 - \frac{r}{a}\right)^{4} - \left(1 + \frac{r}{a}\right)^{-4}\right] + \frac{r^{*12}}{10a^{11} r}\left[\left(1 - \frac{r}{a}\right)^{10} - \left(1 + \frac{r}{a}\right)^{10}\right]\right\}$$

The quantities ε^* and r^* are the interaction parameters which appear in the 6–12 potential of the gaseous molecules. The parameters for the H_2–Ne interaction were calculated in the usual way by taking the geometric mean of the pure gas quantities for ε^* and the arithmetic mean for r^*. The values for H_2–Ne seem close enough to those for Ne–Ne to expect the H_2 molecules to fit into liquid neon cells without great distortion. It is reasonable therefore to apply the LJD model to the binary neon solutions.

It is customary to choose the number of nearest neighbors in the liquid to be the same as in the crystal; for face-centered cubic neon this is $z = 12$. The nearest neighbor distance is given by $a = (2)^{1/6} v^{1/3}$, where v is the molecular volume calculated from the density of the liquid, and $(2)^{1/6}$ is the geometric factor consistent with assuming the packing arrangement in the liquid is that of the solid. The value of a for neon is calculated to be 3.40 Å. The LJD potential determined using these values of ε^*, r^*, and a is presented in Fig. 5. For the purposes of this calculation, the interaction force constant k for the harmonic oscillator potential is determined from

$$Q_R - Q_Q = E_1 - E_0 = 122\ cm^{-1} = \frac{1}{2\pi c}\left(\frac{k}{M}\right)^{\frac{1}{2}}$$

and yields for H_2–Ne, $k = 1.8 \times 10^3$ dynes/cm.

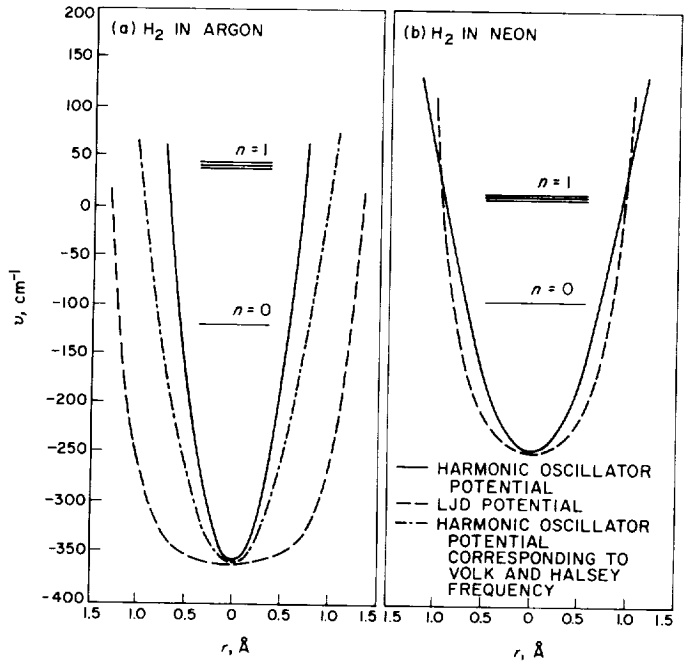


Fig. 5. Comparison of harmonic oscillator and Lennard-Jones-Devonshire potentials for H_2 in solution

This potential is also presented in Fig. 4. There is order of magnitude agreement between the two potentials which lends consistency to the interpretation of the translational energy levels which are observed in the infrared absorption band Q_R as arising from oscillations of H_2 in a cavity in the liquid.

A comparison of the spectroscopically-determined H_2–Ar quadratic potential with the LJD potential can be made. The value of the interaction force constant was found to be $k = 3.0 \times 10^3$ dynes/cm (Ref. 17). Using the value $a = 4.06$Å, the LJD potential is evaluated and compared to the spectroscopically-determined potential in Fig. 5. There is a large discrepancy. There is a considerable difference between the H_2–Ar and Ar–Ar interaction parameters. This suggests that there may be a change in the structure of the liquid when an H_2 is exchanged for an Ar. In order to understand the solubility difference between H_2–Ar and D_2–Ar in terms of a translational quantum effect, Volk and Halsey (Ref. 19) also found that the LJD potential was a poor approximation. They found that the quadratic potential shown by the dot-dash curve in Fig. 5 was consistent with their data. The approximate agreement between the Volk-Halsey potential and the one derived from infrared studies has already been mentioned (Ref. 17).

Since the features $S_R(1)$ and $S_R(0)$ are well defined in the H_2–neon system, they provide additional sources for the determination of the solute-solvent interaction potential. The translational energy level spacings for H_2 in neon are calculated to be $S_R(1) - S_1(1) = 100 \pm 5$ cm^{-1}, $S_R(0) - S_1(0) = 99 \pm 10$ cm^{-1}, and $Q_R - Q_Q = 122 \pm 10$ cm^{-1}. Within experimental error, the spacings determined from $S_R(1)$ and $S_R(0)$ are less than those determined from Q_R. This has also been observed in the spectrum of solid H_2 (1). From theoretical studies of the induced infrared absorption of compressed H_2, $|\langle u \rangle|^2$ has been shown to have different dependence on cluster configuration for $\Delta J = 0$ and $\Delta J = 2$ transitions (Ref. 27). In the liquid phase this implies that the peak maximum of $N_0|\langle u \rangle|^2 P(k)$ will therefore be different for Q_R and S_R absorptions. Unfortunately, the difference in liquid-state interaction potential calculated from S_R or Q_R absorptions is significant, since k depends on the square of the translational energy level spacing. The interaction force constant determined from $S_R(1)$ is $k = 1.2 \times 10^3$ dynes/cm, and from Q_R it is $k = 1.8 \times 10^3$ dynes/cm. The quadratic potential function calculated from $S_R(1)$ is less steep than the one in Fig. 5, and is in better quantitative agreement with the LJD potential.

The weak absorption features designated by Q_P have been assigned to "hot bands," and they involve transitions of the type ($v' = 1$, $n' = 0 \leftarrow v'' = 0$, $n'' = 1$). A translationally-excited H_2 (or D_2) molecule therefore loses a quantum of translational energy on absorption of a photon.

4. Conclusions

There is little doubt that H_2 and D_2 rotate freely in liquid neon solution. The near coincidence of 9 H_2– and D_2–neon-induced infrared absorptions with vibration-rotation transitions calculated for the gas-phase molecules indicate that rotatory levels $J = 0, 1, 2, 3$ in solution phase are essentially the same as for the gas-phase molecules. The rotational partition function for H_2 or D_2 in neon solution is therefore to a good approximation the same as in the gas phase.

The assignment of broad absorption features as vibration-translation combination bands is consistent with the concept that solute translational energy levels are best considered quantized. The $(2)^{1/2}$ decrease in translational energy levels of the solute when D_2–neon is compared to H_2–neon makes differences in the thermodynamic properties of these solutions seem reasonable. It also suggests that a description of solute-solvent interaction as a hard sphere box potential is a poor approximation. A harmonic oscillator potential which is consistent with the solute isotope shift can be calculated from the spectroscopically-determined translational levels and is in reasonable agreement with the Lennard-Jones and Devonshire cell model of the liquid state. Unfortunately, the unique determination of translational energy level spacing from the broad absorption features is complicated by lack of information on transition probability of the H_2–solvent cell system. If this information becomes available, the construction of translational partition functions for these solution systems will be greatly aided by spectroscopic data.

D. Two-Center Coulomb Integrals

M. Geller

A difficulty that arises in the application of non-relativistic quantum mechanics to molecular systems is the evaluation of the integrals originating from the use of trial wavefunctions. For wavefunctions composed of Slater-type atomic orbitals, one needs to evaluate one- and two-electron integrals associated with orbitals on one, two, three, and four different atomic centers. Although the one-center integrals can be evaluated rather easily, the evaluation of the two-electron, two-center integrals is a difficult task. The present investigation is concerned with a derivation and general expression for the two-electron, two-center Coulomb integrals based on the Fourier-convolution method introduced by Prosser and Blanchard (Ref. 29) for one-electron, two-center integrals and used by the author for one-electron, two-center integrals over solid spherical harmonics (Ref. 30) and later extended to two-electron, one and two-center integrals (Ref. 31 and 32).

The Coulomb integral

$$C_{NLM}^{N'L'M'}(p_a, p_b; R) = \int [NLM]_{a1} \frac{1}{r_{12}} [N'L'M']_{b2}\, dT_1\, dT_2$$

$$= [NLM_a \,|\, N'L'M'_b], \qquad (1)$$

where $[NLM]$ is the basic charge distribution defined by Roothaan (Ref. 33),

$$[NLM] = \left(\frac{2L+1}{4\pi}\right)^{1/2} \frac{2^L p^{N+2}}{(N+L+1)!} \times r^{N-1} \exp(-pr)\, S_{L,M}(\theta, \phi), \quad (2)$$

and

$$S_{L,0}(\theta, \phi) = \left(\frac{2L+1}{4\pi}\right)^{1/2} P_L(\cos\theta),$$

$$S_{L,\pm|M|}(\theta, \phi) = \left(\frac{2L+1}{2\pi} \cdot \frac{(L-|M|)!}{(L+|M|)!}\right)^{1/2} \times P_L^{|M|}(\cos\theta) \begin{Bmatrix} \cos|M|\phi \\ \sin|M|\phi \end{Bmatrix} \quad (3)$$

is equivalent, by the convolution theorem, to

$$(2\pi)^{-3} \int [NLM]^T_{a1} \left(\frac{1}{r_{12}}\right)^T [N'L'M']^T_{b2}\, e^{-i\vec{k}\cdot\vec{R}}\, d\vec{k}, \quad (4)$$

where the superscript refers to the Fourier transform, i.e.,

$$f(r)^T = \int e^{+i\vec{k}\cdot\vec{r}} f(\vec{r}) d\vec{r}. \quad (5)$$

The transform of the basic charge distribution, $[NLM]^T$, has been given by the author (Ref. 30) and the transform of r_{12}^{-1} is simply given by $4\pi k^{-2}$. Integrating over the angular coordinates of $\vec{k}$ and simplifying the resulting expression, we have for the general Coulomb integral

$$C^{N'L'M'}_{NLM}(p_a, p_b; R) = (-1)^M \delta_{M,M'} Q^{N'L'}_{NL} \sum_{s=0}^{[1/2(N-L)]} \sum_{t=0}^{[1/2(N'-L')]} (-1)^{s+t} S_{N,L}(s) T_{N',L'}(t) \times \sum_{r=0}^{L<} (-1)^r (2L + 2L' - 4r + 1)\, C^{L+L'-2r}(LM; L'M) \times W^{N+1,N'+1}_{L+L'-2r,\, s+t+r}(p_a, p_b; R), \quad (6)$$

where

$$Q^{N',L'}_{N,L} = 2^{2L+2L'+1} p_a^{-L} p_b^{-L'} \times [(2L+1)(2L'+1)]^{1/2} / (2L+1)!\,(2L'+1)!,$$

$$S_{N,L}(s) = \frac{(N-L-2s+1)_{2s}}{(2L+2)_{2s}} \frac{(s+1)_L}{p_a^{2s}},$$

$$T_{N',L'}(t) = \frac{(N'-L'-2t+1)_{2t}}{(2L'+2)_{2t}} \frac{(t+1)_{L'}}{p_b^{2t}}, \quad (7)$$

and

$[X]$ means the largest integer in X,

$(X+1)_n = (X+n)!/X!$,

$L_<$ is the lesser of L and L',

$C'(LM; L'M)$ are the Condon-Shortley coefficients (Ref. 34) (8)

and

$$W^{p,q}_{m,n}(p_a, p_b; R) = \frac{p_a^{2p} p_b^{2q}}{\pi} \int_0^\infty \frac{k^{m+2n} j_m(kR)\, dk}{(k^2+p_a^2)^p (k^2+p_b^2)^q}, \quad (9)$$

where $j_m(kR)$ are the spherical Bessel functions (Ref. 35).

The expression for the general Coulomb integral [Eq. (6)] involves a triple summation which is over a limited number of terms; as for example, for $N = N' = 5$, the maximum number of terms arising is 18 (when $L = L' = 1$). Often, the number of terms can be further reduced by the use of the recurrence relations for the spherical Bessel functions. We also note that the Coulomb integral vanishes if M and M' are different and further, that the integral is independent *of* M (Ref. 33).

The final difficulty is the one-dimensional infinite integral over k [Eq. (9)]. Although this integral can be evaluated analytically, the result is rather cumbersome; in fact, the integration can be carried out simply and rapidly numerically. Moreover, retaining the integral and evaluating it numerically allows one to either let the charges be equal ($p_a = p_b$) or the distance R go to zero or to do both without additional complications.

A JPL technical report will shortly be available with complete details as to the derivation of Eq. (6), recurrence relations for the W, and tables of the transforms of the $[NLM]$ and of the 83 Coulomb integrals (through $N = 4$) in terms of the auxiliary function W.

E. Sigma-Bonded Alkyl Compounds of Niobium and Tantalum, and Applications to Vapor Phase Plating

G. L. Juvinall

1. Chemistry

Recent developments in the organo-metallic chemistry of the elements of Group VA have been confined mainly to arene complexes. To this date, there have been no reports of sigma-bonded alkyl compounds of these metals, although many attempts to prepare them have been described (Ref. 36). Accordingly, we wish to report the first successful syntheses of alkyl derivatives of niobium and tantalum. Trimethyldichloroniobium, $(CH_3)_3NbCl_2$, and trimethyldichlorotantalum, $(CH_3)_3TaCl_2$, have been prepared by the low-temperature exchange of methyl groups and chlorine between dimethylzinc and the pentachlorides of niobium and tantalum, respectively.

In a typical experiment, 7.45 millimoles of $NbCl_5$ were sublimed *in vacuo* and treated with 18.26 millimoles of $(CH_3)_2Zn$ in 15 ml of pentane. The vessel containing the $NbCl_5$ and pentane was cooled to $-78°C$ and the $(CH_3)_2Zn$ was admitted in small portions because of the exothermic character of the reaction. A precipitate (probably $ZnCl_2$) was observed immediately following addition of the first portion of $(CH_3)_2Zn$. After each addition, the reaction mixture was warmed nearly to room temperature and agitated. After the final addition of $(CH_3)_2Zn$, all volatiles were removed at room temperature; the $(CH_3)_3NbCl_2$ was trapped at $-36°C$. The yield was 12.7% based on $NbCl_5$. Trimethyldichlorotantalum was prepared in an identical manner. All manipulations were carried out in the high-vacuum system.

Trimethyldichloroniobium forms golden yellow crystals which sublime readily in vacuum at room temperature. However, the compound will darken and release methane when left at room temperature for several hours. It appears to be stable indefinitely at $-78°C$.

Trimethyldichlorotantalum forms pale yellow crystals of similar volatility. This compound, however, appears to be much more thermally unstable than the niobium analogue. The thermal stability of both compounds was found to be adversely affected by small quantities of impurities. Both compounds are highly reactive toward water and air.

The formulae of the new compounds were established by hydrolyzing freshly prepared samples *in vacuo* with aqueous KOH, measuring the resultant CH_4 directly by means of a Sprengel pump, and determining niobium and tantalum gravimetrically as the pentoxides. The chloride was also determined gravimetrically. The CH_4 was subsequently shown to be quantitatively pure by means of infrared and mass spectroscopy. For each compound, all analytical data were determined independently on the same weighed sample.

$(CH_3)_3NbCl_2$	
Calculated	CH_3, 21.59; Nb, 44.47; Cl, 33.94
Found	CH_3, 21.6; Nb, 44.6; Cl, 34.0
$(CH_3)_3TaCl_2$	
Calculated	CH_3, 15.19; Ta, 60.93; Cl, 23.88
Found	CH_3, 14.8; Ta, 63.0; Cl, 23.6

In addition, samples of the new compounds were subjected to slow hydrolysis by exposure to the air, and the residues were analyzed spectrographically. No significant quantities of zinc or other extraneous metals were found to be present.

Attempts to obtain the molecular weights by gas density and vapor pressure depression measurements have not thus far been successful because of the instability of these compounds under the experimental conditions used.

The mass spectrum of the niobium compound is indicative of $(CH_3)_3NbCl_2$ monomer, although parent peaks (mass 208, 210, 212) were not observable at an ionizing voltage of 70 ev. Major fragments are the $[(CH_3)_2NbCl_2]^+$ ions at masses 193, 195, and 197 (relative abundances of 57, 37, and 6%, respectively).

Proton nuclear magnetic resonance spectra were obtained for $(CH_3)_3NbCl_2$ at $-10°C$ in CCl_4 containing a trace of $(CH_3)_4Si$. All spectra were obtained with a Varian A-60 spectrometer. Trimethyldichloroniobium exhibits a peak (line width 0.6 cps) 29.8 cps upfield from $(CH_3)_4Si$ which is characteristic for protons on a carbon atom directly bonded to a metal atom. For purposes of comparison, $(CH_3)_2Zn$ gives a peak (line width 0.6 cps) 28.6 cps upfield from $(CH_3)_4Si$. A sample containing both $(CH_3)_3NbCl_2$ and $(CH_3)_2Zn$, and $(CH_3)_4Si$ in CCl_4 was also examined. Only one peak (line width 0.9 cps) 29.6 cps upfield from the internal standard was observed. The merging of the two peaks and the slight broadening are suggestive of methyl group exchange between $(CH_3)_2Zn$

and $(CH_3)_3NbCl_2$. No nuclear magnetic resonance spectra were obtained for $(CH_3)_3TaCl_2$ because of the rapid decomposition of this material in CCl_4 at $-10°C$.

2. Metal Film Deposition

Attempts to produce metallic films by the thermal decomposition of $(CH_3)_3NbCl_2$ and $(CH_3)_3TaCl_2$ were successful. Under controlled conditions, it was possible to deposit metallic films evenly in the hot zone of the apparatus. The films shown were produced by passing $(CH_3)_3NbCl_2$ and $(CH_3)_3TaCl_2$ through quartz tubes with an annular area heated to 700°C. Pressures did not exceed 1 mm; similar experiments at higher pressures yielded films which did not adhere well to the quartz wall. Volatile pyrolysis products include methane and a presently unidentified halogenated hydrocarbon.

Fig. 6. Niobium film deposited in quartz tube

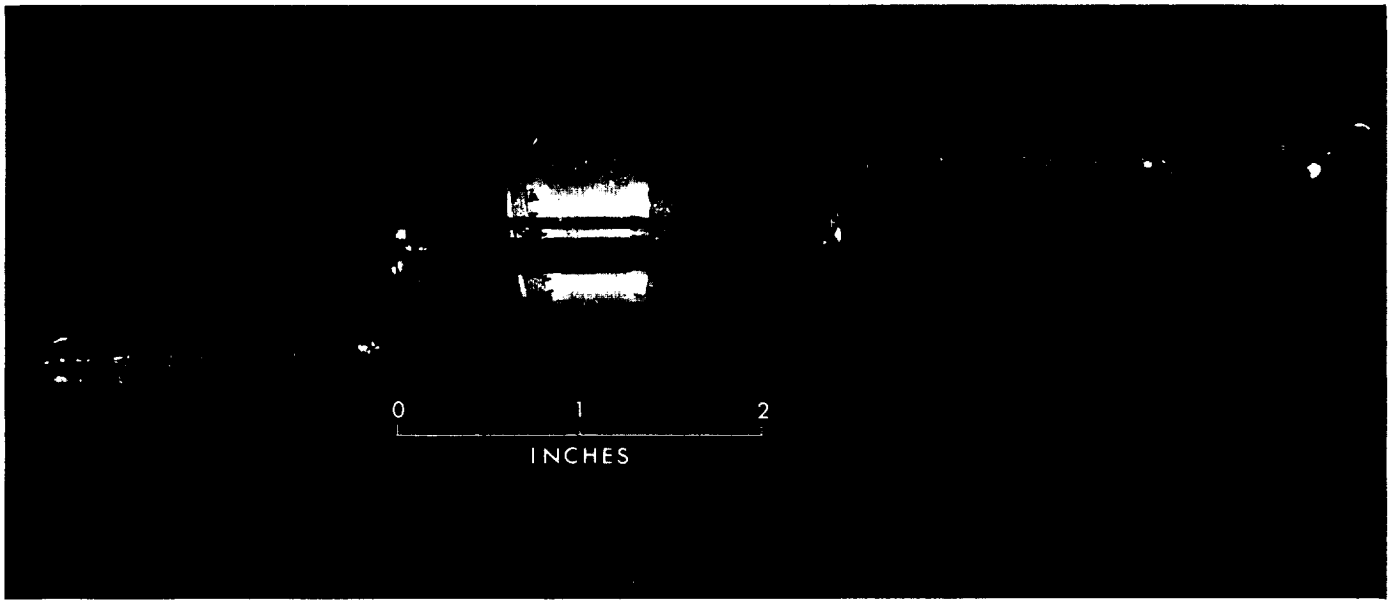


Fig. 7. Tantalum film deposited in quartz tube

At present, only the niobium film, shown in Fig. 6, has been found to be a superconductor. The tantalum film is shown in Fig. 7.

Pyrolysis of these new organometallic compounds appears to have great promise as a technique for vapor phase deposition of niobium and tantalum films under milder conditions than required by other processes. Studies of these compounds are continuing, as well as the search for other new organometallic compounds of the transition metals.

F. The Near Ultraviolet Bands of MgO; Analysis of the $D^1\Delta$–$A^1\pi$ and $C^1\Sigma^-$–$A^1\pi$ Systems

S. Trajmar and G. E. Ewing[5]

1. Introduction

The near ultraviolet spectrum of magnesium oxide (3600–3900 Å) has been extensively studied (Refs. 37–42) but only the (0-0) and part of the (1-1) band of the $C^1\Sigma$–$A^1\pi$ system have been analyzed (Ref. 42).

A reinvestigation of the isotope shift and the vibrational and rotational analysis of the bands in the 3766–3830 Å region are discussed here. A preliminary report on the assignment of these bands has been published recently (Ref. 43).

2. Experimental

The controlled atmosphere arc spectrum of $Mg^{16}O$ and $Mg^{18}O$ has been photographed in the 3766–3830 Å region with a 21-ft concave grating spectrograph in the second order. The line and band head frequencies were measured on a semiautomatic measuring machine of the Tomkins and Fred type. A thorium microwave discharge tube supplied the wavelength standards. The spectral features are designated by wavelength in air or by vacuum wave numbers. The $Mg^{16}O$ frequencies were taken from the plate used and described in connection with the analysis of the $C^1\Sigma$–$A^1\pi$ system (Ref. 42), since the linear dispersion on the plate was about four times larger than on the recently-taken second order plates.

3. The $D^1\Delta$–$A^1\pi$ System

a. Rotational analysis. A series of intense Q band heads and the (0-0) R branch are easily recognized in the 3798–3830 Å region and are characteristic of singlet electronic transitions with $\Delta\Lambda = 1$. The broadening and splitting of the lines at high J values, the double band head formation in the R branches, and the observation that the R (0) line is missing but the R (1) line is present show that the lower electronic state is $^1\pi$ and the upper one is $^1\Delta$.

The preliminary analysis followed the usual procedure (Ref. 44). It has been found that the second combination differences for the lower electronic state are within the limits of experimental error identical with the $A^1\pi$ state, which is known from the analysis of two other systems (Refs. 42 and 45). This fortunate situation confirmed the assignment of the electronic states and served for further cross checking in the analysis.

The final values of the rotational constants and band origins have been obtained from curve fitting of the measured line frequencies to a fourth-order polynomial by the methods of least squares; the values are given in Table 9. The values of α_e and B_e have been obtained by linear extrapolation from B_0 and B_1. The constants represent the average values of the Λ-doublet components.

Table 9. The rotational and vibrational constants for the $A^1\pi$, $D^1\Delta$, and $C^1\Sigma^-$ states of $Mg^{16}O$ (in cm^{-1})

	$A^1\pi$	$D^1\Delta$	$C^1\Sigma^-$
B_0	0.5030	0.4990	0.4984
D_0	1.17×10^{-6}	1.26×10^{-6}	1.27×10^{-6}
H_0	0.59×10^{-11}	–	0.48×10^{-11}
B_1	0.4990	0.4942	0.4936
D_1	1.18×10^{-6}	1.29×10^{-6}	1.30×10^{-6}
B_e	0.5050	0.5014	–
α_e	0.0040	0.0048	–
ω_e	664.4	632.5	632.4
$\omega_e X_e$	3.9	5.3	5.2
(0–0)	–	26,272.04	26,500.94
(1–1)	–	26,237.41	26,466.26

[5]Department of Chemistry, Indiana University, Bloomington, Indiana.

Lines corresponding to $J > 30$ are broad and at $J = 45$ the Λ-doubling is resolved. Double R band head formation is clearly observed in the (0-0) and (1-1) bands; unfortunately the former, which is free of spectral overlap, is obscured in large part by the plate holder, and the lines in the head region and the head itself cannot be measured with the same accuracy as the rest of the system. From the measured splitting δ one obtains the values of $q = B_0^c - B_0^d = 6 \pm 2 \times 10^{-5}$ and $p = D_0^c - D_0^d = 8 \pm 3 \times 10^{-9}$ cm^{-1}, where q and p are coefficients in the equation

$$\delta = qJ(J+1) + p\,[J(J+1)]^2$$

Here the assumption is made that the Λ-doubling is negligible in the $D^1\Delta$ state compared to the one in the $A^1\pi$ state.

b. Vibrational analysis and isotope shift. The small separation of the consecutive Q band heads (40 cm^{-1}) and the smooth increase of the isotope shift with the vibrational quantum number indicate that the bands form a sequence. The identification of the $A^1\pi$ state as the lower electronic state leaves no doubt that the bands analyzed are the (0-0) and (1-1) bands. The vibrational frequency and anharmonicity constant for the $D^1\Delta$ state were obtained from the (0-0), (1-1), and (2-2) band origins and from the known vibrational frequency and anharmonicity of the $A^1\pi$ state (Ref. 45).

It has been found that the band origins obtained from the curve fitting differ sometimes as much as 0.2 cm^{-1} from the measured ones. This is understandable considering the difficulty of measuring intense band heads accurately, and considering that the measurement is done at the maximum density point, which is not at $Q(0)$. In certain cases interference from an overlapping strong line can contribute to this discrepancy. In calculating the vibrational constants the values of the band origins obtained from the curve fitting were used except for the (2-2) band where no extrapolated value was available.

Table 10 summarizes the calculated and measured isotope shifts for the different band heads. The shifts have been calculated from the equations

$$\Delta\nu_v = \nu_v - \nu_v^i = (1-\rho)(\omega_e' - \omega_e'')(v+\tfrac{1}{2}) - (1-\rho^2) \times (\omega_e' X_e' - \omega_e'' X_e'')(v+\tfrac{1}{2})^2$$

$$\Delta\nu_r = \nu_r - \nu_r^i = (1-\rho^2)(\nu_r - \nu_0)$$

Table 10. Isotope shifts for the $D^1\Delta$–$A^1\pi$ system (in cm^{-1})

Band head	Calculated			Measured total
	Vibrational	Rotational	Total	
Q(0–0)	−0.56	–	−0.56	26,271.9 − 26,272.3 = −0.4
Q(1–1)	−1.83	–	−1.83	26,237.2 − 26,238.8 = −1.6
Q(2–2)	−3.28	–	−3.28	26,200.1 − 26,203.0 = −2.9
Q(3–3)	−4.91	–	−4.91	26,160.1 − 26,164.8 = −4.7
R^d(0–0)	−0.56	3.25	2.69	26,320.6 − 26,318.2 = 2.4
R^d(1–1)	−1.83	2.84	1.01	26,279.7 − 26,278.7 = 1.0

In obtaining the measured isotope shifts the actually-measured not the extrapolated head positions were used.

4. The $C^1\Sigma^-$–$A^1\pi$ System

The 3766–3800 Å region has previously been assigned to the $C^1\Sigma^+$–$A^1\pi$ transition (Ref. 42). The branches of the (0-0) band have been identified but the complete assignment of the (1-1) band was not achieved because of the difficulties caused by the serious overlap of the rotational structure. With the help of the $Mg^{18}O$ spectrum and the well established $A^1\pi$ state constants a reinvestigation of this system was undertaken.

In assigning the (1-1) band lines it has been found that the symmetry of the $C^1\Sigma$ state was minus and not plus as it had previously been believed. Unambiguous symmetry assignment is not possible from the (0-0) band alone. By accident the wave number difference between consecutive Q branch lines in the (0-0) band is approximately equal to the corresponding combination defects defined by

$$\epsilon(J) = [R(J) - Q(J+1)] - [Q(J) - P(J+1)]$$

therefore, lowering of the J numbering of the Q lines by one will just change the sign of the combination defect and the symmetry of the $C^1\Sigma$ state. The J numbering of the R and P lines is not affected.

It has been found that an excellent reproduction of the measured rotational lines and band head positions is possible with constants obtained from least squares fitting of the experimental data to sixth-order polynomials. For the R and P branches,

$$\nu(m) = \nu_0 + (B_v' + B_v'')m + (B_v' - B_v'' - D_v' + D_v'')m^2 + (-2D_v' - 2D_v'' + H_v' + H_v'')m^3 - (D_v' - D_v'' - 3H_v' + 3H_v'')m^4 + 3(H_v' - H_v'')m^5 + (H_v' - H_v'')m^6$$

and for the Q branches,

$$\nu(m) = \nu_0 + (B'_v - B''_v) J(J+1) - (D'_v - D''_v) \times J^2(J+1)^2 + (H'_v - H''_v) J^3(J+1)^3$$

The inclusion of the $HJ^3(J+1)^3$ term in the rotational term value is needed to fit the high J value lines well. A similar treatment of the (1-1) band or the bands of the $D^1\Delta$–$A^1\pi$ system would not be meaningful because the accuracy of the line measurements in those cases is inferior to the (0-0) band of the $C^1\Sigma^-$–$A^1\pi$ system.

All three branches of the (1-1) band have been identified as well as the R and Q band heads of the (0-0), (1-1), (2-2), and (3-3) bands. The calculated and experimental isotope shifts are in good agreement. The band head positions and isotope shifts are listed in Table 11, and the constants of the $C^1\Sigma^-$–$A^1\pi$ system are given in Table 9.

Table 11. Isotope shifts for the $C^1\Sigma^-$–$A^1\pi$ system (in cm^{-1})

Band head	Calculated			Measured total
	Vibrational	Rotational	Total	
Q(0–0)	−0.55	–	−0.55	26,500.7 − 26,501.1 = −0.4
Q(1–1)	−1.83	–	−1.83	26,466.2 − 26,467.7 = −1.5
Q(2–2)	−3.26	–	−3.26	26,429.0 − 26,431.8 = −2.8
Q(3–3)	−4.89	–	−4.89	26,388.5 − 26,393.4 = −4.9
R(0–0)	−0.55	2.95	2.40	26,545.1 − 26,543.0 = 2.1
R(1–1)	−1.83	2.57	0.74	26,504.8 − 26,504.4 = 0.4
R(2–2)	−3.26	2.29	−0.97	26,463.3 − 26,464.1 = −0.8
R(3–3)	−4.89	1.99	−2.90	26,418.2 − 26,421.5 = −3.3

5. Discussion

It seems difficult to extend the rotational analysis into the (2-2), (3-3), etc., band regions but one can conclude that the spectral features in the 3766–3830 Å region are accounted for by the $C^1\Sigma^-$–$A^1\pi$ and $D^1\Delta$–$A^1\pi$ systems. The irregularity in the isotope shift observed earlier was caused by the lack of sufficient resolution in the $Mg^{18}O$ spectrum. The complexity of the spectrum is the consequence of excessive overlapping of the rotational structure of the highly excited MgO emitter.

The vibrational and rotational constants of the $C^1\Sigma^-$ and $D^1\Delta$ states are disturbingly similar. One could consider that we are dealing here with Hund's case c coupling where $C^1\Sigma^-$ and $D^1\Delta$ correspond to the 0^- and 2 components of a Hund's case a (or b) π^3 state. (The 1 component has not been observed.) Hund's coupling case c or transition from case a or b to case c occurs for small intermolecular distances (hydrides) or for the less stable states of molecules containing heavy atoms (Ref. 44). It would be quite unusual, therefore, to expect strong spin orbit coupling and large multiplet splitting for MgO, which is a light molecule, and the electronic states in question are quite stable ($D_e \simeq (\omega_e)^2/4\,\omega_e X_e = 19{,}000$ cm^{-1}).

No perturbation has been found in the spectrum.

G. A Purine Derivitive of Aminoethylcellulose and Its Possible Use for the Detection, Isolation, and Study of Desoxyribonucleic Acid

H. H. Weetall and N. Weliky

Nucleic acids being universal constituents of terrestrial life, antibodies to these substances are of particular interest as specific reagents for detecting desoxyribonucleic acids or ribonucleic acids on Mars.

There have been many attempts to induce formation of antibodies specific to desoxyribonucleic acid (Refs. 46 and 47), but in only one instance have antibodies toward purified desoxyribonucleic acid been demonstrated (Ref. 48). In this case, the desoxyribonucleic acid was obtained from T4 bacteriophage and was partially denatured. This virus nucleic acid uniquely contains a glucosylated pyrimidine, and the antibodies produced were found to be specific to this particular portion of the molecule.

Substances which have the ability to react with desoxyribonucleic acid are frequently found in the sera of patients with systemic *lupus erythematosus* (Refs. 49 and 50). These substances are produced spontaneously; however, the stimulus for their production is unknown.

Since the recent synthesis of 6-halomethylpurines and pyrimidines (Ref. 51), it has become a simple matter to couple these bases to carrier proteins and obtain antibodies specific to these haptenic groups.

Using antisera containing anti-purine antibodies, Butler, Beiser, Erlanger, Tanenbaum, Cohen, and Bendich (Ref. 52) have found that these antibodies will react with desoxyribonucleic acids.

Because the anti-purine antibodies are capable of reacting with nucleic acids, it appears feasible to use them for the detection of nucleic acids.

We have developed a simple technique for isolating anti-purine antibody through the use of 6-trihalomethylpurines and related compounds. Aminoethylcellulose was coupled to 6-trichloromethylpurine in tetrahydrofuran-water, maintaining the pH alkaline. The quantity of purine coupled was dependent upon the percentage of tetrahydrofuran; the less water present, the greater the coupling (Table 12).

Table 12. Coupling of 6-trichloromethylpurine to aminoethylcellulose

Column designation	Ratio of THF* to water, v/v	Purine coupled, mg	Free aminoethyl groups coupled, %	Non-specific protein released, mg
A	4:96	1.5	1.6	0.21
B	50:50	12.2	12.8	0.14
C	75:25	54.6	57.4	0.10

*Tetrahydrofuran

The anti-purine antisera was prepared by first coupling freshly prepared 6-trichloromethylpurine to keyhole limpet hemocyanin (Ref. 52). Six New Zealand white rabbits received weekly injections of this antigen in complete Freund's adjuvant for three consecutive weeks. The rabbits were bled six days after the last injection.

A column containing 1 g of the purine-cellulose derivitive was prepared and tested first for the retention and release of nonspecific protein. Two ml of normal rabbit sera were passed through the column at neutral pH, and the column was washed with 1% NaCl until no absorption could be detected spectrophotometrically at 220 mμ, in a flow-through cell. The column was then eluted with 0.1 molar phosphate buffered at pH 2.3. As increased coupling occurred, the non-specific protein released by the columns at pH 2.3 decreased (Table 12). By further increasing the quantity of antigen coupled, the non-specific protein may be lowered even more.

For isolating the antibody, the column releasing the least nonspecific protein (Column C) was chosen. Since the antiserum used for the experiment contained 0.52 mg of anti-purine antibody per ml of serum, the highest theoretical purity obtainable was 80% as determined by precipitability with an antigen made by coupling the purine to bovine serum albumin.

Two ml of anti-purine sera were passed through the column and the column washed as described above. For determining purity (Ref. 53), the antibody was eluted with 1% NaCl adjusted to pH 2.3 with hydrochloric acid. Of the 1.04 mg of specific antibody passed through the column, 95% was retained. The isolated protein amounted to 1.24 mg of which 74% was precipitable with the purine coupled to bovine serum albumin. The data show conclusively that anti-purine antibody can be purified by this technique. By using column materials which release less non-specific protein, the purity of the isolated antibody should be increased further.

The isolation of anti-purine and anti-pyrimidine antibodies should also enable studies of their physical and chemical properties to be made and comparisons made both with anti-desoxyribonucleic acid antibodies produced against T4 bacteriophage and autoantibodies found in *lupus erythematosus* sera. These studies could lead not only to a significant increase in the understanding of antibody synthesis but to contributions to the knowledge of the etiology of *lupus erythematosus*.

References

1. Schonhorn, H. "Generalized Approach to Adhesion via the Interfacial Deposition of Amphipathic Molecules. I. Adhesion of Polyethylene to Aluminum," *Journal of Polymer Science*, Part A, 1, pp. 2343–2359, 1963.
2. Blodgett, K. B., "Films Built by Depositing Successive Monomolecular Layers on a Solid Surface," *Journal of the American Chemical Society*, Vol. 57, p. 1007, 1935.
3. Audrieth, L. F., and Ackerson, B. "The Chemistry of Hydrazine," Chapter 8, John Wiley & Sons, Inc., New York, 1951.
4. Kauck, E. A., and Diesslin, A. R., "Some Properties of Perfluorocarboxylic Acids," *Industrial and Engineering Chemistry*, Vol. 43, p. 2332, 1951.
5. Onak, T., Gerhart, F. J., and Williams, R. E., *Journal of the American Chemical Society*, Vol. 85, p. 3378, 1963.
6. The proposed nomenclature for cage molecules has been given by R. Adams, *Inorganic Chemistry*, Vol. 2, p. 1087, 1963.
7. Beaudet, R. A., and Poynter, R. L., *Journal of the American Chemical Society*, Vol. 86, p. 1258, 1964.
8. Hoffmann, R., and Lipscomb, W. N., *The Journal of Chemical Physics*, Vol. 36, p. 3489, 1962.
9. Poynter, R., *The Journal of Chemical Physics*, Vol. 39, p. 1962, 1963.
10. Kraitchman, J., *American Journal of Physics*, Vol. 21, p. 17, 1953.
11. Costain, C. C., *The Journal of Chemical Physics*, Vol. 29, p. 864, 1958.
12. Pierce, L., *Journal of Molecular Spectroscopy*, Vol. 3, p. 575, 1959.
13. Krisher, L. C., and Pierce, L., *The Journal of Chemical Physics*, Vol. 32, p. 1619, 1960.
14. Gush, H., Hare, W., Allin, E., and Welsh, H., *Canadian Journal of Physics*, Vol. 38, p. 176, 1960.
15. Bulanin, M., and Orlova, N., *Optika i Spektroskopia*, Vol. 11, p. 551, 1961. English translation: *Optics Spectroscopy (USSR)*, Vol. 11, p. 298, 1961.
16. Vu, H., Atwood, R., and Staude, E., *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, Vol. 257, p. 1771, 1963.
17. Ewing, G., and Trajmar, S., *The Journal of Chemical Physics*, Vol. 41, p. 814, 1964.
18. Bigeleisen, J., *The Journal of Chemical Physics*, Vol. 61, p. 87, 1964.
19. Volk, H., and Halsey, G., *The Journal of Chemical Physics*, Vol. 33, p. 1132, 1960.
20. Simon, M., *Physica*, Vol. 29, p. 1079, 1963.
21. Simon, M., *Physical Review Letters*, Vol. 5, p. 319, 1963.
22. Hare, W., and Welsh, H., *Canadian Journal of Physics*, Vol. 36, p. 88, 1958.

References (cont'd)

23. Lennard-Jones and Devonshire, *Proceedings of the Royal Society,* (London), Vol. 163A, p. 53, 1937; Vol. 165A, p. 1, 1938.

24. *Tables of Wave Numbers for the Calibration of Infra-Red Spectrometers,* Butterworth, Inc., Washington, D. C., 1961.

25. Acquista, N., and Plyler, E., *Journal of Research, National Bureau of Standards (U.S.),* Vol. 49, p. 13, 1952.

26. Van Kranendonk, J., *Physica,* Vol. 23, p. 825, 1957; Vol. 24, p. 347, 1958; Vol. 25, p. 337, 1959.

27. Hirschfelder, J., Curtiss, J., and Bird, R., *Molecular Theory of Gases and Liquids,* John Wiley & Sons, New York, 1954.

28. Powell, J., and Craseman, B., *Quantum Mechanics,* Addison-Wesley, Reading, Massachusetts, 1961.

29. Prosser, F. P., and Blanchard, C. H., "On the Evaluation of Two-Center Integrals," *The Journal of Chemical Physics,* Vol. 36, p. 1112, 1962.

30. Geller, M., "Two-Center Integrals Over Solid Spherical Harmonics," *The Journal of Chemical Physics,* Vol. 39, p. 84, 1963.

31. Geller, M., "Two-Electron, One- and Two-Center Integrals," *The Journal of Chemical Physics,* Vol. 39, p. 853, 1963.

32. Geller, M., and Griffith, R. W., "Zero-Field Splitting, One- and Two-Center Coulomb Integrals," *The Journal of Chemical Physics,* Vol. 40, p. 2309, 1964.

33. Roothaan, C. C. J., "A Study of Two-Center Integrals," *The Journal of Chemical Physics,* Vol. 19, p. 1445, 1951.

34. Slater, J. C., "Quantum Theory of Atomic Structure," Vol. I, p. 309, McGraw-Hill Book Co., Inc., New York, 1960.

35. Erdelyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G., "Higher Transcendental Functions," Vols. I, II, McGraw-Hill Book Co., Inc., New York, 1953.

36. Cotton, F. A., *Chemical Reviews,* Vol. 55, p. 551, 1955.

37. Brewer, L., and Porter, R., *The Journal of Chemical Physics,* Vol. 22, p. 1867, 1954.

38. Gaydon, A. G., and Pesic, D., *Proceedings of the Physical Society* (London), Vol. A73, p. 244, 1959.

39. Bulewicz, E. M., and Sugden, T. M., *Transactions of the Faraday Society,* Vol. 55, p. 720, 1959.

40. Pesic, D., *Proceedings of the Physical Society* (London), Vol. A76, p. 844, 1960.

41. Brewer, L., and Trajmar, S., *The Journal of Chemical Physics,* Vol. 36, p. 1585, 1962.

42. Brewer, L., Trajmar, S., and Berg, R., *Astrophysical Journal,* Vol. 135, p. 955, 1962.

References (Cont'd)

43. Trajmar, S., and Ewing, G., *The Journal of Chemical Physics,* Vol. 40, p. 1170 1964.

44. Herzberg, G., *Molecular Spectra and Molecular Structure,* Vol. I, D. Van Nostrand Co., Inc., New York.

45. Lagerqvist, A., and Uhler, U.,*Arkiv fur Fysik,* Vol. 1, p 459, 1949.

46. Blix, U., Iland, C. N., and Stacy, M., *British Journal of Experimental Pathology,* Vol. 35, p. 241, 1954.

47. Lackman, D., Mudd, S., Sevag, M. G., Smolens, J., and Wiener, M., *Journal of Immunology,* Vol. 40, p. 1, 1941.

48. Levine, L., Murakami, W. T., Van Vunakis, H., and Grossman, L., *Proceedings of the National Academy of Sciences,* Vol. 46, p. 1038, 1960.

49. Deicher, H. R. G., Holman, H. R., and Kunkel, H. G., *Journal of Experimental Medicine,* Vol. 109, p. 97, 1959.

50. Stollar, Levine, D. L., Lehrer, H. I., and Van Vanakis, H., *Proceedings of the National Academy of Sciences,* Vol. 48, p. 874, 1962.

51. Cohen, S., Thom, E., and Bendich, A., *Journal of Organic Chemistry,* Vol. 27, p. 3545, 1962.

52. Butler, V. P., Beiser, S. M., Erlanger, B. F., Tanenbaum, S. W., Cohen, S., and Bendich, A., *Proceedings of the National Academy of Sciences,* Vol. 48, p. 1597, 1962.

53. Weliky, N., Weetall, H. H., Gilden, R. V., and Campbell, D. H., *Immunochemistry,* in press.

XVII. Fluid Physics

A. A Proposal Concerning Viscous Flow Past Bluff Obstacles

S. Childress

We have previously (*SPS 37-23*, Vol. IV) described the branching of solutions of Euler's equations representing flow past a finite symmetric eddy, with respect to the vorticity (assumed constant) over the interior of the eddy. In the present note this branching is made the basis for a new proposal concerning stationary solutions of the Navier-Stokes equations which represent uniform flows past a finite two-dimensional obstacle, e.g., a circular cylinder.

Experience has shown that the stationary viscous flow past a circular cylinder cannot be maintained up to arbitrarily large Reynolds number Re. Regardless of how the experiment is performed, fluctuations which are periodic or almost periodic in time are eventually observed. In the range of Re where these are not present, the evolution of the flow with Re is marked by the growth of a pair of attached, elongated rotational eddies downstream of the obstacle, with the result that the wetted surface of the obstacle upstream of the eddy becomes a progressively smaller fraction of the bounding zero streamline. (Some recent observations of this evolution are given in Ref. 1.) Mathematically, it is known that the flow is uniquely determined by the usual conditions if Re is sufficiently small.

Because of these facts, we are led to the conjecture that the stability of the system under consideration is a function of Re. This is, of course, true of most viscous flows (a notable exception being Poiseuille flow). The stability which we refer to here may be decided in principle as follows: We deal with a set F of twice continuously differentiable solenoidal vector fields $\boldsymbol{f}$ which satisfy the boundary conditions on velocity in the stationary problem. We associate with any $\boldsymbol{f}$ a norm $N(\boldsymbol{f})$, and consider a non-stationary process with initial values from F. A given stationary solution $\boldsymbol{q}$ is then said to be stable if we can choose an ϵ so that the non-stationary solution tends for large time to $\boldsymbol{q}$, provided that $N(\boldsymbol{q} - \boldsymbol{f}) \leqslant \epsilon$. Otherwise $\boldsymbol{q}$ is unstable. It is this kind of stability that one hopes to settle by linearization of the equations about $\boldsymbol{q}$.

In the theory of stationary solutions of the Navier-Stokes equations, instability reappears in the phenomenon of *branching*. We here refer to branching with respect to a positive parameter, namely *Re*. If the number of solutions changes at a branch point Re_c, we call this a *bifurcation point*. In linear stability analyses bifurcation is associated with the neutral disturbances of the stability boundary. An important case is where the analytic continuation of the stable flow with respect to the parameter past the point $Re = Re_c$ is unstable, with new stable stationary flows branching away from Re_c. An example of this kind of branching occurs at the Taylor boundary in circular Couette flow. The point to bear in mind is that the real stable solution in $Re < Re_c$ is here continued analytically in *Re* through Re_c to a new real solution, and the fact that the continuation is unstable is of no importance to the theory of stationary solutions.

Not all bifurcation points at Re_c need be of the above type. The analytic continuation of the real stable solution may not be real, in which case for some $Re < Re_c$ there is a second, unstable branch. We can diagram these two possibilities by assigning to any stationary solution a number j, and assume for simplicity that j is not the same for any two of the solutions considered. The first and second possibilities may then be represented as shown in Figs. 1(a) and 1(b), respectively, the solid lines denoting stable and the dotted unstable solutions.

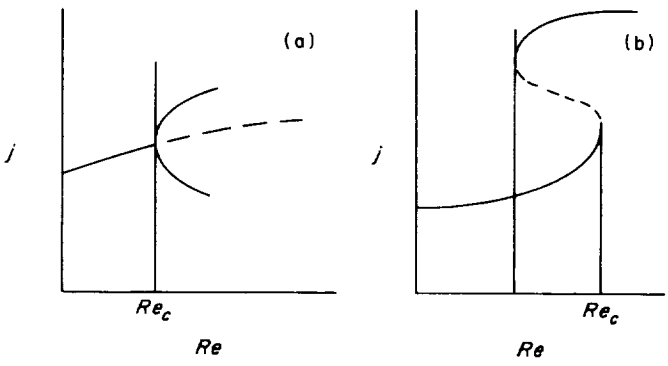


Fig. 1. Branching of stationary solutions

The conjecture we now adopt is that branching of the second kind occurs in the problem of flow past an obstacle. Some immediate consequences are now summarized:

(1) Since the solution is known to be unique for *Re* sufficiently small, there is a second critical point marking the termination of the unstable branch which, if not a singular point, is necessarily a bifurcation point, so that a second stable branch exists [see Fig. 1(b)].

(2) At $Re = Re_c$ the flow is neutrally stable according to the linear stability theory. However, for $Re > Re_c$ the analytic continuation, which is not real, can never be observed. Thus, if $Re > Re_c$, we cannot expect a stationary solution to be "near" (in the sense of j) to the solution at Re_c.

(3) For some $Re < Re_c$ there is a branch of stationary solutions which, because they are unstable, cannot in practice be maintained.

(4) Interpreted dynamically, the presence of an unstable solution near to the stable one when *Re* is only slightly less than Re_c means that small perturbations of the system could lead to non-stationary motions at values of the Reynolds number less than critical. Thus, in practice a calculated limit Re_c could never be reached.

In connection with (2) we should remark that the idea that the flow "goes unstable" at $Re = Re_c$ is here rejected, since this terminology carries with it the implication that real flows, stable or unstable, exist on either side of Re_c, which then leads up to the first type of branching and not the second. At the same time, therefore, we reject the idea that careful experimentation can in any way postpone the bifurcation, except by increasing Re_c through a change of the problem itself. Because of this, our conjecture carries with it a note of caution concerning interpretation of an *inviscid limit* in the Navier-Stokes theory. That is, an observed stationary flow, irrespective of the Reynolds number, cannot be close to the limit unless the two flows lie on the same branch. We return to the question of the inviscid limit below.

To sum up our position, it is proposed that the continuation of a given stationary flow past a circular cylinder at a Reynolds number where it is uniquely defined by a boundary value problem for the Navier-Stokes equations, to arbitrarily large Reynolds numbers, is not an analytic one in the expansion parameter. In fact, the continuation is thought to consist of a finite sequence of continuous branches, and in particular there may occur points of bifurcation where two branches terminate. If two solutions of the same problem are continuations of one another, lie on the same branch, and occur at the same Reynolds number, they are identical. A bifurcation point is believed to mark the theoretical

limit for stationary experimental continuations of observed flows.

One can attempt to test the above conjectures in two ways. First, if we start from the alternative proposition, namely that only one branch of solutions need be considered, then we are led to the problem of finding large asymptotic approximations for *Re* (the problem is a singular one) guided by experience with the solutions in the Reynolds number range where the solutions are stable. In doing so, one looks for approximations which satisfy certain broad constraints of consistency, for example, conservation of mass, momentum, and angular momentum. However, as yet we are aware of no asymptotic solution (other than the one we mention below) which is free of inconsistency in the conservation laws. In particular, the existence of an elongated wake bubble leads to the conjecture that the length of this bubble grows indefinitely as *Re* is increased. In fact, in order that the limiting drag be finite, it is necessary that the length grow like *Re*. However, it can be shown that this wake-bubble model leads to a paradox which is as yet unresolved.

Several other possible limits have been examined with a similar result, so that it now appears to be difficult to find a limit representing flow past an obstacle which has finite positive drag. Batchelor (Ref. 2) has suggested that the limiting drag is necessarily zero, and has proposed a cusped eddy model which is free of the inconsistency of the wake-bubble model. However, as yet it is not known whether or not this model is in fact ever a limit in the sense intended. Our proposal states, of course, that *any* consistent model, regardless of its relation to observed flows, must be considered as a candidate for the limit, and any asymptotic solution of the Navier-Stokes equations for large *Re* could be close to, and on the same branch as, the limit.

A second test is based on the introduction of mathematical models for the Navier-Stokes solutions over a finite range of *Re* which includes the first bifurcation, in the hope that models which are physically close would display similar behavior. One procedure is to model the kinematics of distributed vorticity, while neglecting its diffusion, through the use of rotational solutions of Euler's equations. Some representative solutions are described in a report to be published.[1] The symmetric eddies which we use here were described in *SPS 37-23*, Vol. IV. If the parameter is taken to be the constant vorticity ω in the eddy, then it can be shown that the proposed bifurcation occurs when

$$\frac{1}{16}\left(\frac{\omega L}{U}\right)^2\left(\frac{L_e}{L}\right)^{3/2}\left(\frac{C_D}{\pi}\right)^{1/2} = r_0 \doteq 0.71 \qquad (1)$$

In this expression L is a reference length, L_e is the length of the eddy, and C_D is the drag coefficient. If L_e/L, C_D, and $\omega L/U$ are taken as functions of *Re* which are such that the left-hand side of Eq. (1) attains the value r_0, then branching of the proposed type will occur in the model and Re_c can be calculated.

A numerical computation of flow around a circular cylinder at $Re = 40$ (usually the highest value that can be maintained in experiments) was reported in Ref. 3, and can be used to check the order of magnitude of Re_c as predicted by Eq. (1). If values of L_e/L and C_D are estimated from this computation, Eq. (1) gives $\omega L/U \doteq 1.2$. The numerical results show a variation of $\omega L/U$ (L being the radius of the cylinder) from 0.25 to about 2 over the eddy region, and a good average value appropriate to the model might be 0.75. Note that since Re_c would necessarily be greater than 40, our comparison must overestimate $\omega L/U$. Thus, in view of the simplicity of the model, the agreement is surprisingly close in order of magnitude. A direct computation of Re_c should be possible if diffusion of vorticity can be modeled in a sufficiently simple way, a refinement which is now being studied.

B. Boundary-Layer Tripping in Super- and Hyper-sonic Flows

F. R. Hama

Artificial tripping of a laminar boundary layer to produce a turbulent one is often desired in wind-tunnel model tests in order to simulate full-scale conditions. The tripping, however, becomes increasingly difficult as the Mach number increases. The difficulty is twofold. First, the shear layer, which is formed behind a tripping device, becomes more stable at high Mach numbers. Therefore, it becomes difficult to create the discrete vortices which lead the flow to the inception of turbulence, as the recent

[1] Childress, S., "Solutions of Euler's Equations Illustrating Effects of Finite Eddies."

experimental investigations on the detailed process of transition have revealed (Ref. 4). Second, as the Mach number increases, it becomes increasingly difficult in wind tunnels to obtain a sufficiently large roughness Reynolds number $Re_k = u_k\, k/\nu$ (u_k is the unperturbed velocity at the top of the roughness element, k the roughness height, and ν the kinematic viscosity defined at a certain height in the boundary layer), because of the following reasons:

(1) The boundary-layer thickness increases so that u_k/U_∞ for a given roughness height decreases, U_∞ being the free-stream velocity.

(2) The temperature ratio increases and hence the kinematic viscosity increases.

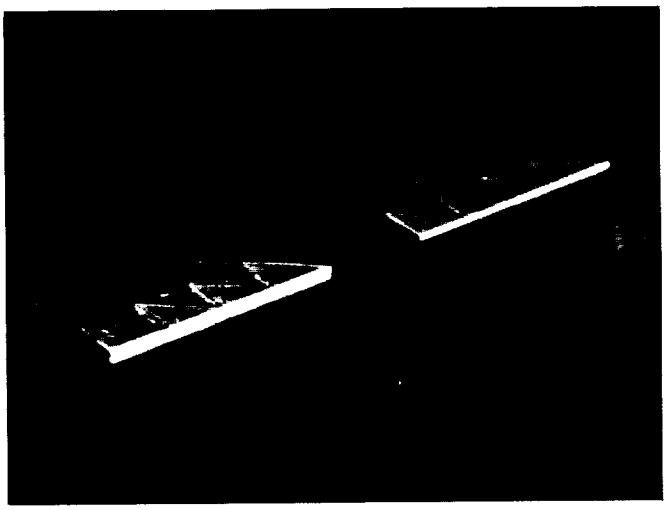


Fig. 2. Triangular-patch stimulators as mounted at 2 in. from the leading edge of a flat plate

(3) An attempt to increase Re_k by increasing the roughness height k results in the formation of a shock wave in front of the roughness element, thus decreasing the effective u_k/ν.

All of these difficulties are of fundamental nature and do not appear to be easy to avoid. The present investigation is not an attempt to resolve the difficulties but is an attempt to test whether a particular tripping device, which has proved to be considerably more efficient than conventional ones in low-speed flows, is equally efficient in supersonic flows.

The type of tripping device investigated is seen in Fig. 2, mounted on a flat plate surface. Notice that the flow direction is from left to right and that, because of the surface reflection, the thickness of the tripping device appears to be twice the actual thickness, which was 0.04 in. for the element on the left and 0.02 in. for the one on the right. The same type of device as mounted on a cone is shown in Fig. 3.

The use of this device was originally suggested by a preliminary observation that a hairpin-shaped, highly-concentrated vortex loop is responsible for the breakdown into turbulence (Ref. 5). This tripping device is apparently quite efficient in deforming the vorticity field in the boundary layer into a hairpin shape. Although the detailed process of the formation of the hairpin vortex in natural transition was later found to be more complex than initially supposed, the important role played by the hairpin vortex has been completely verified (Refs. 4 and 6).

The tests were conducted on a flat plate (Fig. 2) at Mach number M = 2, 3, 4, and 5 in the 20-in. supersonic wind tunnel and on an 8-deg (included) cone (Fig. 3) at

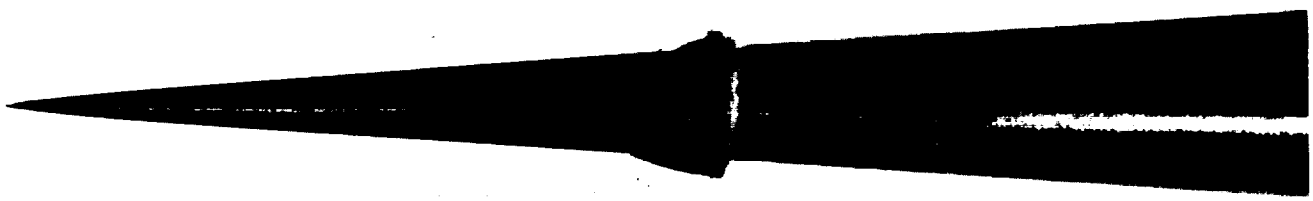


Fig. 3. Triangular-patch stimulator (0.04-in. thick) as mounted at 2 in. from the tip of a cone

$M = 6$ in the 21-in. hypersonic tunnel. The tripping elements were in the shape of equilateral triangles and their thickness was either 0.02 or 0.04 in. For the supersonic tests, the elements were 0.5 in. on a side, and were glued to the flat plate at 2 or 6 in. from the leading edge. For the hypersonic tests, the elements were approximately 0.2 in. on a side, and were fitted to the cone at the location 2 in. from the tip. The supersonic tests were made by the use of a hot-wire anemometer; the transition point x_t was taken to be the location of maximum mean-square output when the hot wire was traversed in the flow direction at a fixed height (approximately 0.04 in.). Owing to the lack of a hot-wire traversing mechanism, the transition point in the hypersonic tests was determined from spark-Schlieren pictures, which clearly revealed the state of the boundary layer. Cross examination with a fixed hot wire but with varying total pressure of the wind tunnel, however, indicated close agreement between the results of the two methods of observation (see also Ref. 7).

All of the raw data obtained in the present tests are plotted in Figs. 4–7. Included also in Fig. 5 is the transition point obtained behind a conventional trip wire of 0.02-in. diameter glued at 6 in. from the leading edge of the flat plate. In order to demonstrate the advantage of the present tripping device in comparison with spherical roughness elements, in particular, the method of analysis utilized by van Driest and Blumer (Ref. 8) was adopted. The variation of the transition location generally follows

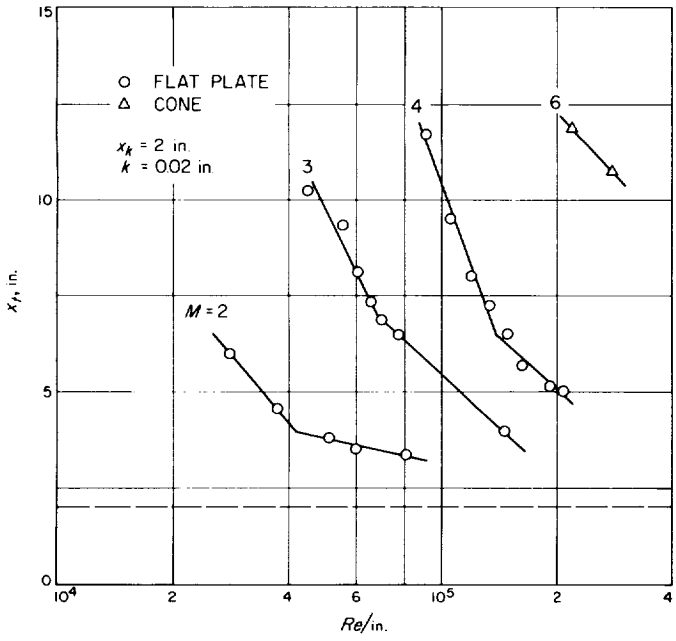


Fig. 4. Transition point x_t vs flow Reynolds number Re/in. (x_k = 2 in., k = 0.02 in.)

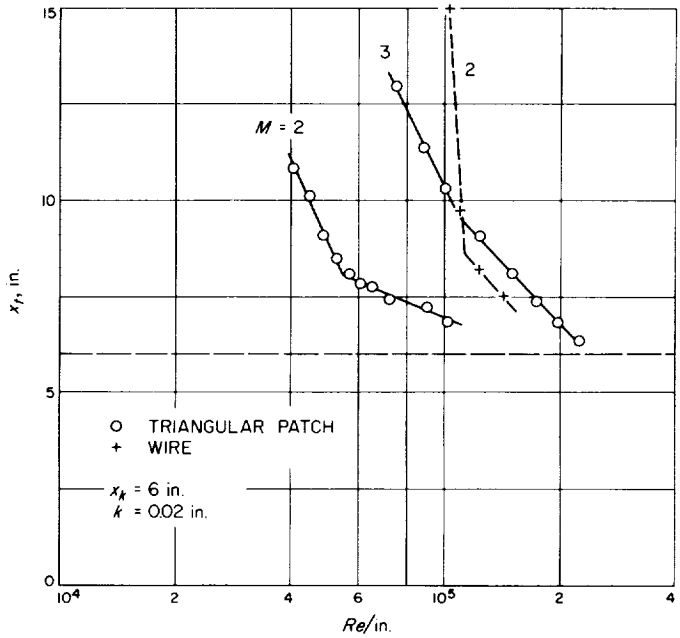


Fig. 5. Transition point x_t vs flow Reynolds number Re/in. (x_k = 6 in., k = 0.02 in.)

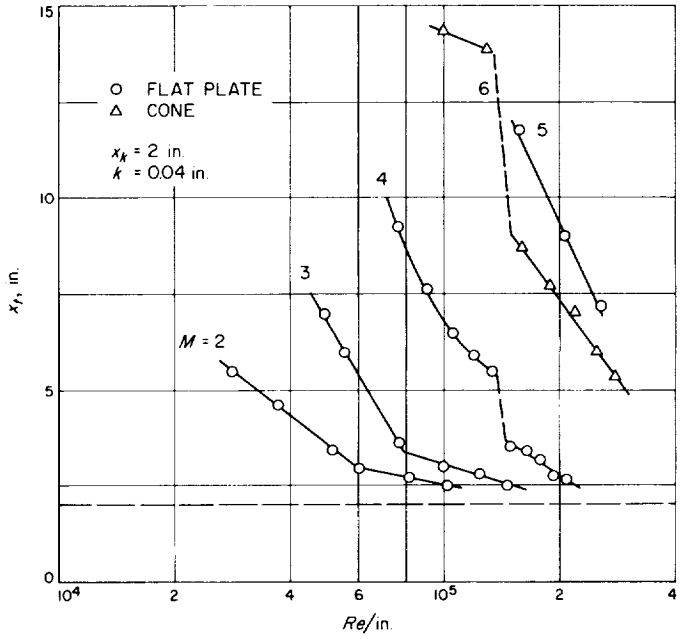


Fig. 6. Transition point x_t vs flow Reynolds number Re/in. (x_k = 2 in., k = 0.04 in.)

a pattern such as sketched in Fig. 8. As the flow Reynolds number ($Re/\text{in.} = U_\infty/\nu_\infty$) is increased from a very low value, the transition point first moves forward quickly from the natural transition point. However, after the transition point reaches the so-called *effective* point, its

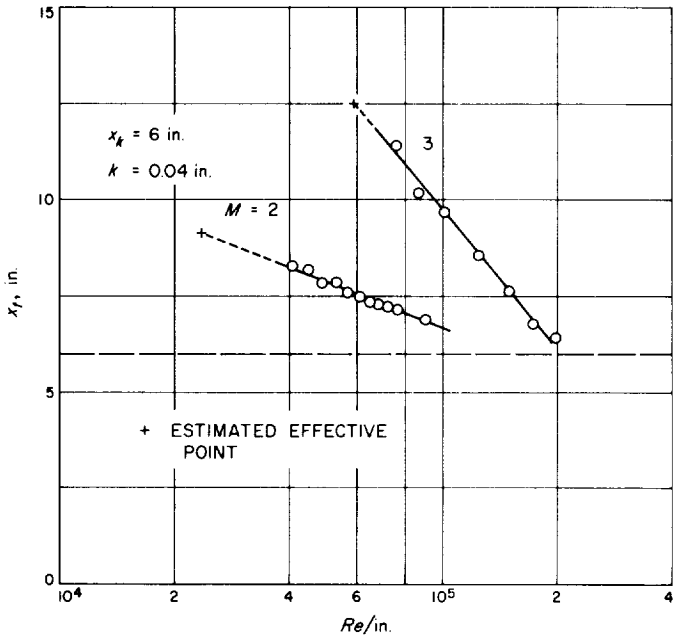


Fig. 7. Transition point x_t vs flow Reynolds number Re/in. (x_k = 6 in., k = 0.04 in.)

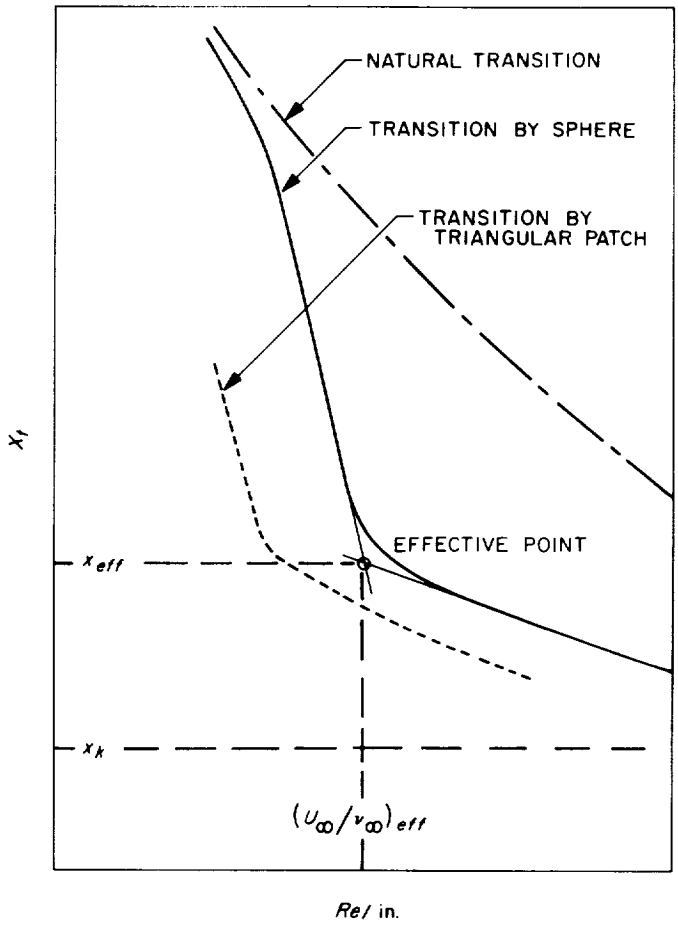


Fig. 8. General relationship between transition point and flow Reynolds number, and definition of the effective point

further forward advancement is slow. Since the transition process is in fact a gradual evolution rather than a spontaneous explosion and hence takes time and distance, the transition point will never really reach the trip position no matter how large the Reynolds number may be, unless overridden by natural transition. In practice, it suffices to have a transition point a short distance downstream from the tripping device.

The effective point was well-defined in all of the results of Ref. 8. The same was true for most of the present data, but in some cases it could only be defined somewhat artificially and in others it was not evident at all, as in Fig. 7. When the effective point x_{eff} can be defined, the following two Reynolds numbers may be defined:

$$Re_{x_k} = (U_\infty/\nu_\infty)_{eff}\, x_k \quad , \quad Re_{t_{eff}} = (U_\infty/\nu_\infty)_{eff}\, x_{eff}$$

In general, the transition Reynolds number $U_\infty\, x_t/\nu_\infty$ assumes a minimum value $Re_{t_{eff}}$ at the effective point.

From experimental results obtained behind spherical roughness elements, van Driest and Blumer found that Re_{x_k} is a function of the roughness number x_k/k and the free-stream Mach number M such that, for a flat-plate experiment,

$$Re_{x_k}^{3/4} = 43.2\left(1 + \frac{\gamma - 1}{2} M^2\right)\frac{x_k}{k} \qquad (1)$$

A similar correlation is obtained for the present tripping device. The equation

$$Re_{x_k}^{3/4} = 27\left(1 + \frac{\gamma - 1}{2} M^2\right)\frac{x_k}{k} \qquad (2)$$

shown as a solid line in Fig. 9, represents the open-symbol data reasonably well. Direct comparison with the spherical roughness elements may be found in Fig. 10 for x_k/k = 100. It is clear that a boundary layer can be tripped by the triangular-patch stimulator at a substantially lower Reynolds number than by the spherical roughness. The above equation fails to correlate data for x_k/k as small as 50, because the basic assumption that the roughness height is small compared with the boundary-layer thickness is no longer applicable. Moreover, it is seen that the boundary-layer tripping becomes less effective as well.

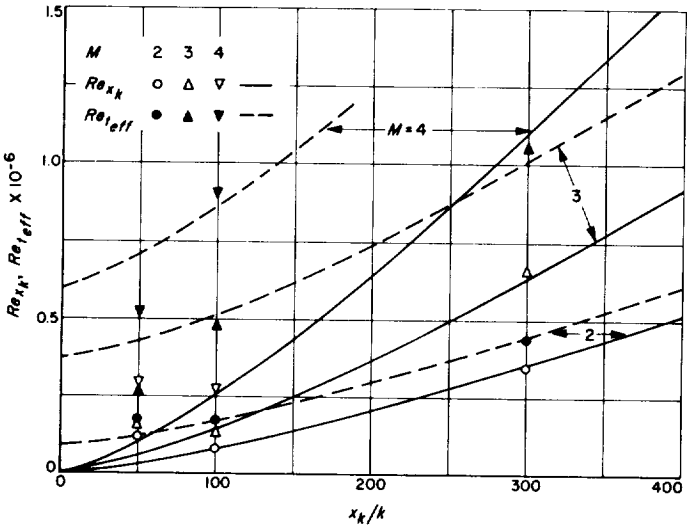


Fig. 9. Transition Reynolds numbers Re_{x_k} and $Re_{t_{eff}}$ defined at the effective point vs roughness number x_k/k

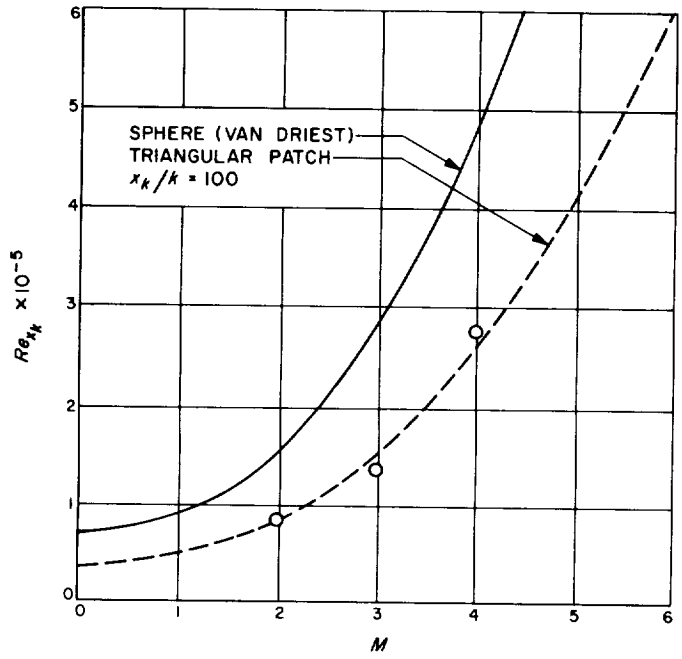


Fig. 10. Transition Reynolds number Re_{x_k} vs Mach number M (x_k/k = 100)

The second correlation obtained by van Driest and Blumer is that a Reynolds number based upon the distance between the trip location and the transition point at the effective-point condition (Fig. 8) is a unique function of the free-stream Mach number and is independent of x_k/k, except for the extreme values of x_k/k.

Within the limited amount of experimental data obtained in the present investigation, a similar correlation has also been established. This Reynolds number, which may be called the delayed-transition Reynolds number, is plotted in Fig. 11 in comparison with van Driest's curve for the spherical roughness elements. Also shown is one result for the conventional trip wire obtained during the present investigation (Fig. 5). The comparison shows that the triangular-patch stimulator induces transition not only at a lower free-stream Reynolds number but also at a location closer to the trip than other devices, as schematically shown by the dotted line in Fig. 8.

From these correlations, we may now roughly estimate the possible effective points for those conditions in which no effective points were evident, such as in Fig. 7. We can see that the lowest flow Reynolds numbers used during the tests were not small enough to reach the effective points.

We may further predict, by extrapolation, the effective-point transition Reynolds number at, say, $M = 6$ on a flat plate. In order to take advantage of the effectiveness of the trip, x_k/k is chosen to be 100. From Fig. 10, $Re_{x_k} = 0.6 \times 10^6$, and from Fig. 11, $Re_{t_{eff}} - Re_{x_k} = 1.25 \times 10^6$; hence $Re_{t_{eff}} = 1.85 \times 10^6$, which will be the smallest transition Reynolds number attainable with the present scheme of correlation. There still remains a choice as to how to attain this minimum transition Reynolds number. If the purpose is to obtain the smallest

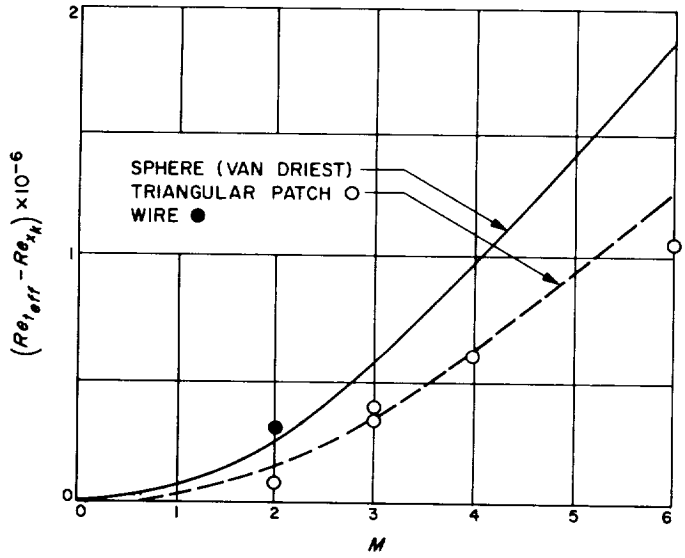


Fig. 11. Delayed-transition Reynolds number ($Re_{t_{eff}} - Re_{x_k}$) vs Mach number M

transition distance from the leading edge, the minimum transition Reynolds number should be attained at the maximum flow Reynolds number available in the wind tunnel. For example, the maximum flow Reynolds number in the 21-in. hypersonic wind tunnel at the Jet Propulsion Laboratory is about 0.3×10^6/in. at $M = 6$. Therefore, $x_{eff} = 6$ in., $x_k = 2$ in., and $k = 0.02$ in. On the other hand, if the purpose is to maintain transition within a certain range of the flow Reynolds number, say, above 0.1×10^6/in., then it would be advantageous to attain the effective transition Reynolds number at this flow Reynolds number. In this case, we have $x_{eff} = 18.5$ in., $x_k = 6$ in., and $k = 0.06$ in.

It is noted here that the constant factor in Eq. (2) should probably read about 20 when the triangular-patch stimulator is applied on a cone. Under conditions similar to the above example, we obtain $Re_{t_{eff}} = 1.6 \times 10^6$

It was already noted that the correlation equation did not hold for very small values of the roughness number x_k/k. For the experimental results obtained behind the spherical roughness elements at $M = 2.7$, Re_{x_k} and $Re_{t_{eff}}$ were found to take the forms sketched in Fig. 12(a). Whereas Re_{x_k} follows the kind of correlation given by Eq. (1), $Re_{t_{eff}}$ levels off below a certain critical value of the roughness number $(x_k/k)_c$, which was about 200 for the spherical roughness elements on a cone. Therefore, there is no advantage in applying a large roughness near the tip, such that the roughness number becomes less than 200.

With the triangular-patch stimulator, on the contrary, Re_{x_k} tends to level off or even slightly increase below a certain value of the roughness number, but the delayed-transition Reynolds number $(Re_{t_{eff}} - Re_{x_k})$ vastly decreases, resulting in appreciable reductions in $Re_{t_{eff}}$ [see Fig. 9 and Fig. 12(b)]. Such reductions seem to become larger as the Mach number increases. Therefore, we may expect a reduction in the transition Reynolds number by applying a relatively large triangular-patch stimulator at a location near the leading edge. For example, the minimum transition Reynolds number obtained at $M = 6$ with the 0.04-in. triangular patch on the cone was about 1.3×10^6, which may be compared with the natural transition Reynolds number at least 5×10^6.

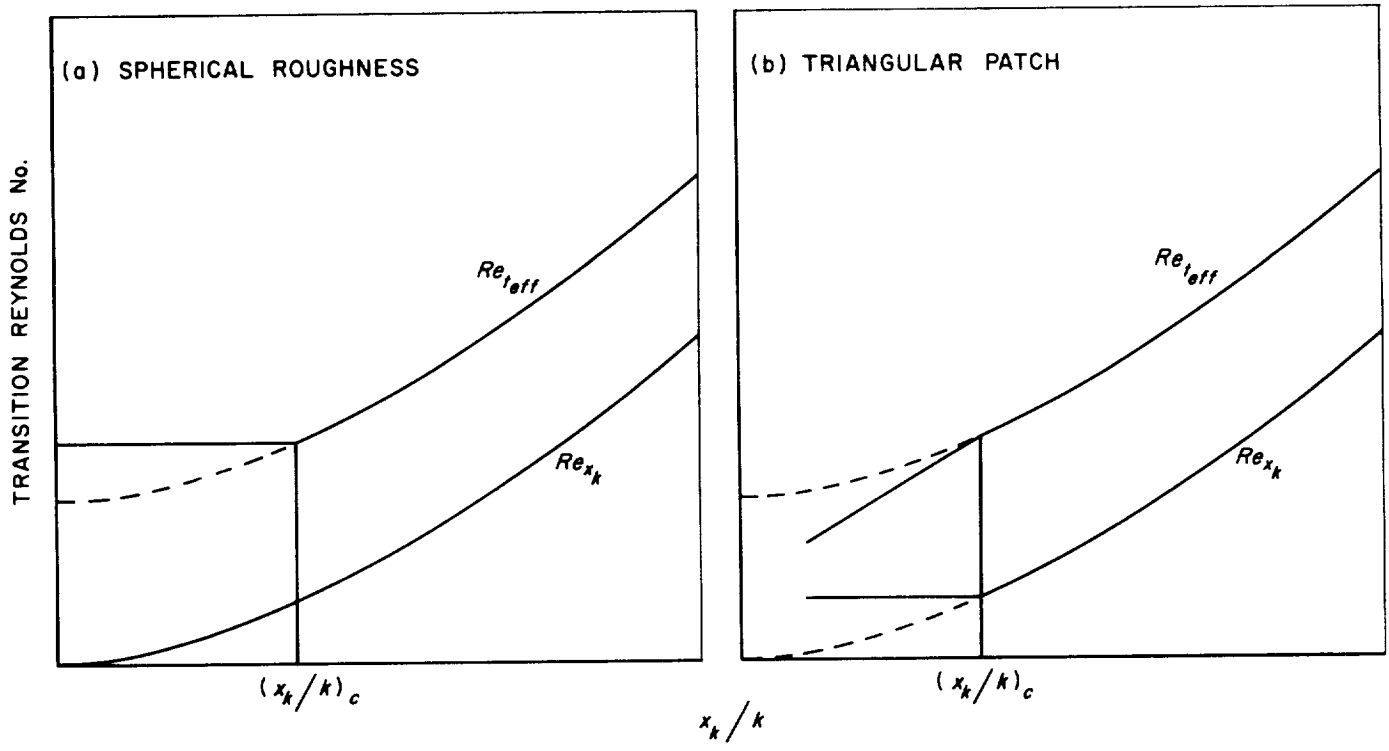


Fig. 12. Trends of transition Reynolds numbers Re_{x_k} and $Re_{t_{eff}}$ for (a) spherical roughness and (b) triangular patch at small values of roughness number x_k/k

C. A Linearized Boundary-Layer Solution for a Finite-Radius Rotating Disk

L. M. Mack

The problem usually associated with a finite-radius disk in a rotating flow, with the flow rotating faster than the disk, is that of a boundary layer which starts at the edge of the disk with zero thickness. However, when β^*, the angular velocity of the disk, is equal to $\omega^*(r^*)$, the angular velocity of the outer flow, at $r^* = r_1^*$, and $\beta^* < \omega^*$ for $r^* < r_1^*$, the boundary layer starts at r_1^* with a non-zero thickness. These two different starting conditions correspond to the two found with the conventional boundary layer. On a flat plate, the boundary layer starts at the leading edge with zero thickness; on a blunt body it starts at the stagnation point with non-zero thickness. For the rotating boundary layer, Stewartson (Ref. 9) has given the similarity solution for the zero-thickness starting condition. With r_1^* the radius of the disk, this type of solution applies whenever $\beta^* < \omega^*(r_1^*)$. In Ref. 10, the present author gave the similarity starting solution for the second case. In this solution, the point r_1^* acts very much as a stagnation point, with the radial velocity a linear function of a distance from r_1^*, provided $d(\omega^* r^*)/dr^* \neq \beta^*$ at r_1^*. The latter situation, where the flow in the vicinity of r_1^* is locally solid body with angular velocity β^*, is not considered further here.

With the stagnation point starting condition, r_1^* is the radius at which the angular velocities are equal and not necessarily the radius of the disk. However, if the radius of the disk is larger than r_1^*, the flow for $r^* > r_1^*$ is independent of the flow for $r^* < r_1^*$ in the sense that there are no streamlines connecting the two regions. The flow for $r^* > r_1^*$ will be of the von Kármán type with the secondary flow outward. If for some $r^* = r_2^*$ the angular velocities are again equal, the similarity solution of Ref. 10 also applies near r_2^*, but the flow will be in the opposite direction from the flow near r_1^*. Therefore, r_2^* is not comparable to a stagnation point, but to a reverse stagnation point, or to the point $r^* = 0$ in the Rogers and Lance family of rotating-disk similarity solutions (Ref. 11), with the outer flow rotating faster than the disk.

In Ref. 10 mention was made of a momentum-integral solution which serves to carry the boundary-layer solution from the region near r_1^* to $r^* = 0$. Near $r^* = 0$, this momentum-integral solution is identical to the infinite-radius solution for the ratio of angular velocities $\omega^*(0)/\beta^*$ for both starting conditions. This result indicates that the boundary-layer solution on a disk of finite radius in a rotating flow, unlike the usual boundary layer, has a predictable solution at its end point, $r^* = 0$. When the disk is at rest, the solution at the end point is the Bödewadt solution; when the disk is rotating, the end-point solution is one of the Rogers and Lance solutions.

In this note an analytic solution is presented for the finite-radius disk which reduces to the infinite-radius solution as $r^* \to 0$. The solution is a linearized one, and the linearization is only possible with the stagnation-point starting condition. A further necessary condition for the linearization is that near r_1^* the derivative of v_∞^*, the tangential velocity of the outer flow, with respect to r^* must differ from β^* by only a small quantity. It was shown in Ref. 10 that when this condition is satisfied, the similarity solution for the stagnation-point starting condition reduces to a linearized solution. The linearized solution for an infinite rotating disk in an infinite rotating flow was given by Squire (Ref. 12). Both of these linearized solutions are of the Ekman type, since they are linearizations about a solid-body rotation. The solution to be given here is valid for all r^* and includes the solutions of Squire and Ref. 10.

The dimensionless axisymmetric boundary-layer equations in cylindrical coordinates are

$$u\frac{\partial u}{\partial r} + w\frac{\partial u}{\partial z} - \frac{v^2}{r} = -\frac{v_\infty^2}{r} + \frac{\partial^2 u}{\partial z^2} \tag{1}$$

$$u\frac{\partial v}{\partial r} + w\frac{\partial v}{\partial z} + \frac{uv}{r} = \frac{\partial^2 v}{\partial z^2} \tag{2}$$

where r and z are the radial and axial coordinates, and u, v, w are the radial, tangential, and axial velocities. The definitions of the dimensionless quantities are

$$r = \frac{r^*}{r_1^*} \qquad z = \frac{z^*}{r_1^*}(Re_t)^{1/2}$$

$$u = \frac{u^*}{v_1^*} \qquad v = \frac{v^*}{v_1^*} \qquad w = \frac{w^*}{v_1^*}(Re_t)^{1/2} \tag{3}$$

In these definitions, r_1^* is the radius where $v_\infty^* = \beta^* r^*$, v_1^* is the value of v_∞^* at r_1^*, and

$$Re_t = \frac{v_1^* r_1^*}{\nu^*} \tag{4}$$

is the tangential Reynolds number at r_1^*. Since the reference velocity v_1^* is equal to $\beta^* r_1^*$, the dimensionless tangential velocity of the disk is equal to r.

With $v - r$ regarded as a small quantity, the linearized equations for u and v are the well-known Ekman equations

$$2(v_\infty - v) = \frac{\partial^2 u}{\partial z^2} \tag{5}$$

$$2u = \frac{\partial^2 v}{\partial z^2} \tag{6}$$

The boundary conditions for u and v are

$$\begin{aligned} &\text{at } z = 0: && u = 0, && v = r \\ &\text{at } z \to \infty: && u \to 0, && v \to v_\infty \\ &\text{at } r = 1: && && v = 1 \end{aligned} \tag{7}$$

By separation of variables, the solutions of Eqs. (5) and (6) appropriate to the boundary conditions are easily found to be

$$u = -(v_\infty - r)\, e^{-z} \sin z \tag{8}$$

$$v = r + (v_\infty - r)(1 - e^{-z} \cos z) \tag{9}$$

The axial velocity is determined from the continuity equation

$$\frac{\partial}{\partial r}(ru) + \frac{\partial}{\partial z}(rw) = 0 \tag{10}$$

to be

$$w = \frac{1}{2}\left[\left(\frac{dv_\infty}{dr} - 1\right) + \left(\frac{v_\infty - r}{r}\right)\right](1 - e^{-z} \cos z - e^{-z} \sin z) \tag{11}$$

This linearized solution is a special case of the more general solution obtained by Lewellen (Ref. 13) in a comprehensive study of linearized vortex flows.

Near the radius $r = 1$, the velocity components are

$$u = -(1 - v_1')(1 - r)\, e^{-z} \sin z \tag{12}$$

$$v = 1 + (1 - v_1')(1 - r)(1 - e^{-z} \cos z) \tag{13}$$

$$w = -\frac{1}{2}\left[(1 - v_1') - (1 - v_1' - v_1'')(1 - r)\right] \times (1 - e^{-z} \cos z - e^{-z} \sin z) \tag{14}$$

where v_1' and v_1'' are dv_∞/dr and d^2v_∞/dr^2, respectively, at $r = 1$. Eqs. (12) to (14), except for the $1 - r$ term in Eq. (14), are the linearized similarity solutions obtained in Ref. 10. With $v_1' < 1$, the radial velocity increases linearly away from $r = 1$, just as for the ordinary two-dimensional or axisymmetric stagnation point, and the axial velocity is negative and directed into the boundary layer. However, dw/dr is not, in general, equal to zero at $r = 1$ as at an ordinary stagnation point because of the absence of symmetry in the present case.

Since the outer flow must be a solid-body flow near $r = 0$, it is possible to write

$$v_\infty = \omega r \quad \text{for} \quad r \to 0 \tag{15}$$

where ω (> 1) is the ratio of $\omega^*(0)$, the angular velocity of the outer flow at $r = 0$, to the angular velocity of the disk. The three velocity components near $r = 0$ are, from Eqs. (8), (9), (11), and (15),

$$u = -(\omega - 1)\, re^{-z} \sin z \tag{16}$$

$$v = r + (\omega - 1)\, r\,(1 - e^{-z} \cos z) \tag{17}$$

$$w = (\omega - 1)(1 - e^{-z} \cos z - e^{-z} \sin z) \tag{18}$$

The above solution is identical to the solution of Squire for the rotating disk of infinite radius in a solid-body flow of infinite extent. This agreement of the finite-radius and Squire solutions demonstrates that within the framework of the linearized theory the infinite-radius similarity solution is attained by the boundary layer on a finite-radius disk as $r \to 0$.

It can be observed that when ω is a function of r, Eqs. (16) and (17) are still valid at all r, but at any r except near $r = 0$ the axial velocity must be computed from the more general Eq. (11) instead of Eq. (18). Consequently, the two velocity components u and v, as well as the radial mass flow in the boundary layer, are the same at any radius as the corresponding infinite-radius solution for the local value of ω. However, the axial velocity differs from the axial velocity of the infinite-radius solution, and must be computed from Eq. (11) which follows directly from the continuity equation once u has been obtained. The use of the infinite-radius solution to establish the relation between the radial mass flow and the local angular velocity, with the axial velocity related to the mass flow by the continuity equation, has been proposed by Turner (Ref. 14). This procedure is seen to be exact in the limit of the linearized solution.

When the outer flow is such that the linearization does not apply, even though $\beta^* = \omega^*(r_1^*)$, a Blasius type series expansion about $r = 1$ can be used to solve the boundary-

layer equations, just as the Stewartson series was used for the flow with $\beta^* < \omega^*(r_1^*)$ (Ref. 15). In the latter case, no matter what the rotation of the disk the boundary layer starts at a singular point, and the series method has difficulty in following the solution as it adjusts to the solid-body outer flow. Consequently, it is not possible to settle in a definitive manner the question of whether or not the infinite-radius solution is attained as $r \to 0$. In the present problem, the series method can be examined first for an outer flow where the linearization is valid and the correct solution is known. Then the outer flow can be gradually moved away from the linear range to find if the infinite-radius solution is still attained as $r \to 0$. The series method may be more successful in handling this problem than it was in the Stewartson case.

D. The Inviscid Stability of the Cooled Laminar Boundary Layer

L. M. Mack

The material presented in previous volumes of *Space Programs Summaries* concerning the stability of the laminar boundary layer deals exclusively with the boundary layer on an insulated surface. In the present contribution, the inviscid stability of the cooled boundary layer is considered. It was first brought out by Lees (Ref. 16) that cooling has a stabilizing effect on the boundary layer, and that over a certain range of Mach numbers two-dimensional disturbances can be completely stabilized by cooling. Much effort has been devoted to the computation of the cooling, or surface temperature, required for complete stability. The stability criterion used in these investigations is derived from the asymptotic theory. In view of the unreliability of the asymptotic method for computing neutral-stability curves at Mach numbers above 2, as well as the indirect nature of the criterion for complete stabilization, and in view of the recently demonstrated existence of additional instability regions associated with multiple inviscid solutions (Ref. 17), it is of considerable interest to investigate the effect of cooling by means of direct numerical methods. The inviscid stability characteristics of cooled boundary layers are treated in this article, and the stability characteristics at finite Reynolds numbers will be taken up in a future *Space Programs Summary*.

The situation regarding the effect of cooling on the inviscid stability before the discovery of the multiple inviscid solutions can be summarized as follows: Amplified solutions were regarded as near neighbors of neutral solutions, and in particular of the neutral solution with the eigenvalues α_s, c_s, where c_s is the dimensionless phase velocity equal to the mean velocity at η_s, the point in the boundary layer where $v_0 \sim (u'/T)'$ is zero (u and T are the mean velocity and temperature; the primes refer to derivatives with respect to η), and α_s is the corresponding dimensionless wave number. The locus of amplified solutions in an eigenvalue diagram goes from the neutral subsonic (relative to the free stream) solution (α_s, c_s) to the neutral sonic solution (0, $1 - 1/M_1$). Supersonic disturbances were ruled out. As the boundary layer is cooled, η_s, which for an insulated surface is greater than η_1, the point at which $u = 1 - 1/M_1$, moves closer to and eventually below η_1 [see Fig. 13 for the distribution through the boundary layer of v_0 for several values of the boundary-layer cooling parameter defined by Eq. (1)

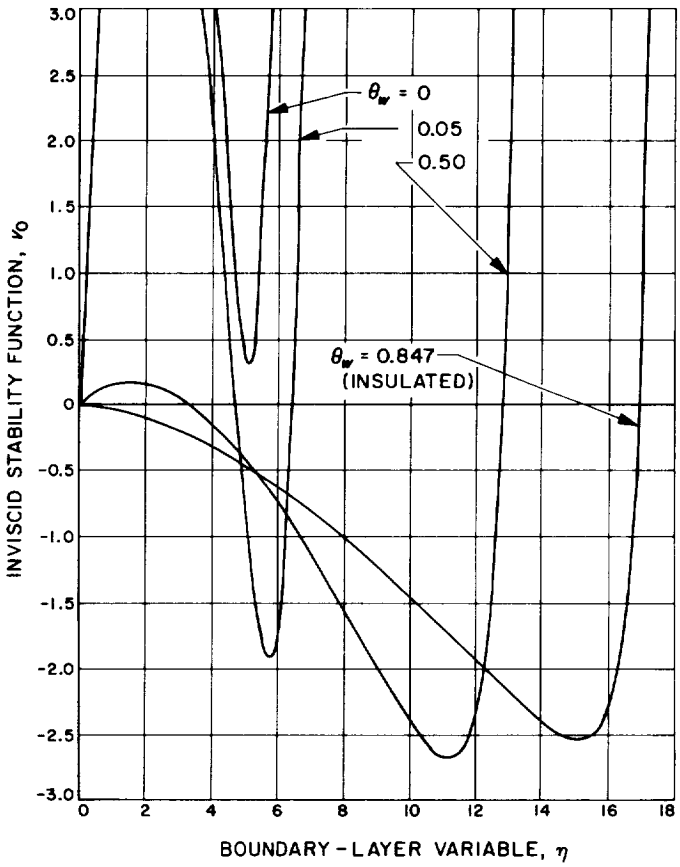


Fig. 13. Distribution through boundary layer of stability function v_0 for four values of θ_w at $M_1 = 5.8$

below]. Consequently, the combination of the requirements that the disturbances be subsonic relative to the free stream, and that the neutral solution (α_s, c_s) exists, resulted in the criterion for inviscid instability that $v_0 = 0$ and that $\eta_s > \eta_1$.

When the multiple inviscid solutions are taken into account, it is immediately seen that this reasoning applies only to the first mode, because the first mode is the only one for which the eigenvalues of the amplified solutions lie between α_s, c_s and 0, $1 - 1/M_1$. The eigenvalues of the second and higher modes go from the neutral subsonic solutions (α_{sn}, c_s), where the subscript n is the mode number, to the singular neutral solutions (α_{1n}, 1). The important feature of the singular neutral solutions is that they are present whether the boundary layer is cooled or not. The only requirement is that the mean flow relative to $c_r = 1$ must have a supersonic region. For an insulated surface, the relative Mach number at the surface is first equal to one at about $M_1 = 1.1$. By sufficient cooling, it is always possible to create a supersonic relative flow region even at much smaller Mach numbers. For example, with T_w^*/T_1^*, the ratio of the wall to the free-stream temperature, equal to 0.5, the surface Mach number relative to a wave moving with free-stream velocity is unity or larger for $M_1 \geq 0.707$. Consequently, since a neutral inviscid solution is always available with sufficient cooling, amplified solutions can also be expected.

A detailed study of the effect of cooling has been made at $M_1 = 5.8$. The parameter for the cooled boundary-layer solutions is θ_w, which is defined by

$$\theta_w = \frac{i_w^* - i_1^*}{i_0^* - i_1^*} \tag{1}$$

where i is the enthalpy, and the subscripts w, 1, and 0 refer to wall, free-stream, and stagnation conditions, respectively. Fig. 14 gives three of the more commonly used temperature ratios as functions of θ_w at $M_1 = 5.8$. The temperature T_r^* is the temperature of the insulated wall, or recovery temperature. The numerical results are presented in the form of eigenvalue diagrams in Figs. 15, 16, and 17. Fig. 15 gives α as a function of c_i, the imaginary part of the wave velocity, for an insulated surface and four values of θ_w. Figs. 16 and 17 give the corresponding c_r as a function of c_i. Only neutral and amplified solutions are considered. For $\theta_w = 0.50$ and 0.10, η_s is greater than η_1. For $\theta_w = 0.05$, η_s is almost coincident with η_1, and for $\theta_w = -0.10$, v_0 has no zero.

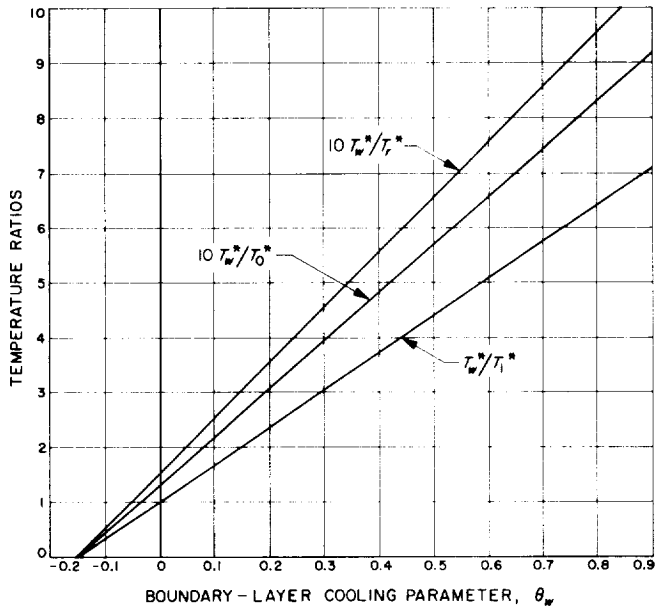


Fig. 14. Commonly-used temperature ratios as functions of the boundary-layer cooling parameter θ_w at $M_1 = 5.8$

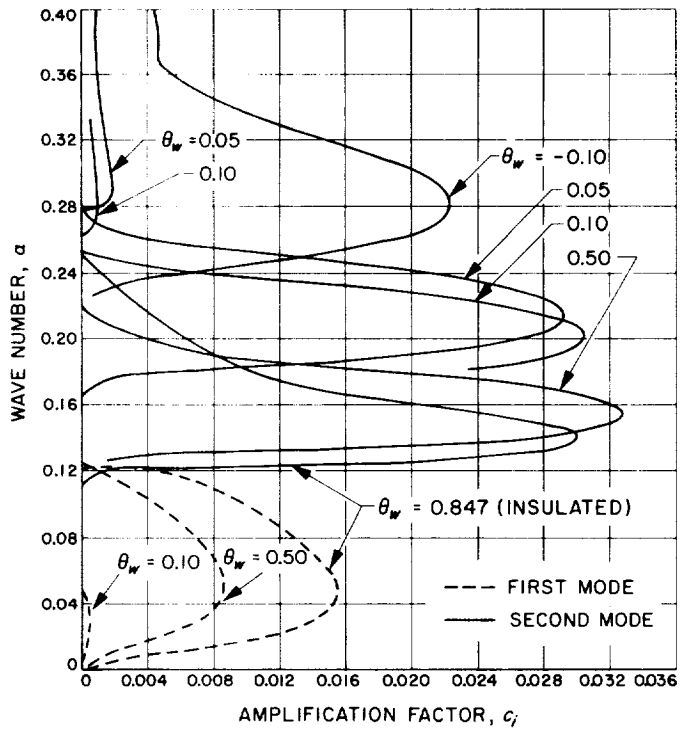


Fig. 15. Eigenvalue diagram α vs c_i for $\theta_w = 0.847$ (insulated), 0.50, 0.10, 0.05, and −0.10 at $M_1 = 5.8$

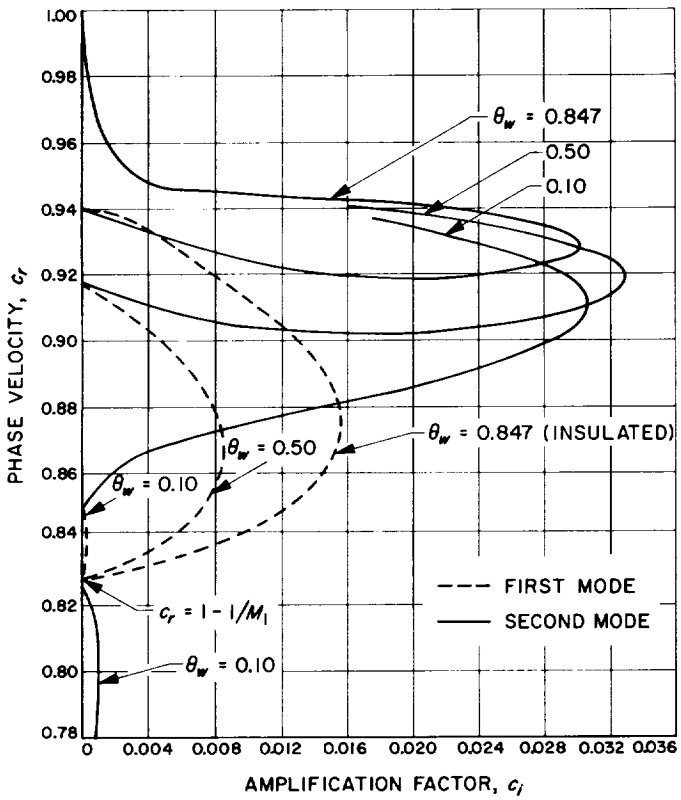


Fig. 16. Eigenvalue diagram c_r vs c_i for $\theta_w = 0.847$ (insulated), 0.50, and 0.10 at $M_1 = 5.8$

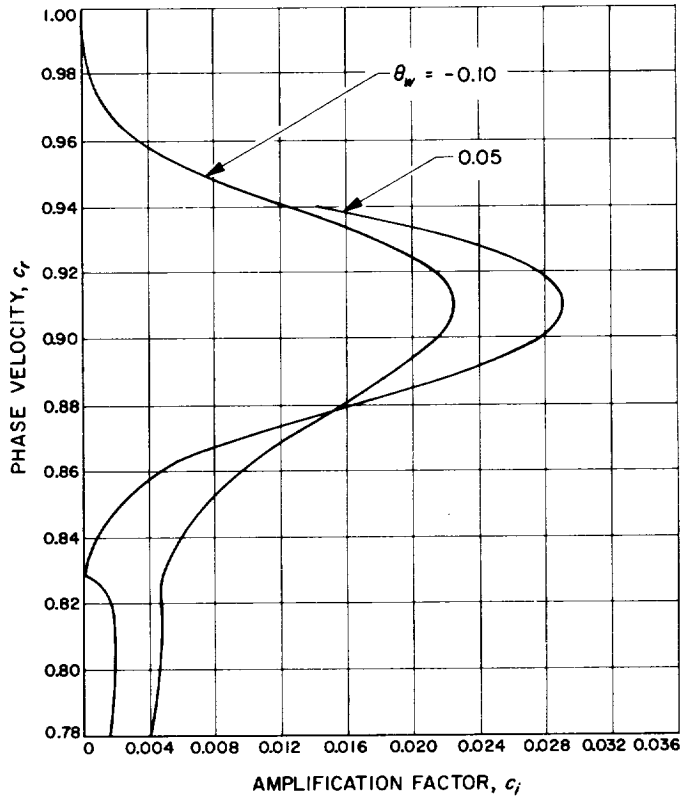


Fig. 17. Eigenvalue diagram c_r vs c_i for $\theta_w = 0.05$ and -0.10 at $M_1 = 5.8$

The effect of cooling on the first mode is as previously described. As $\eta_s \rightarrow \eta_1$, the maximum time rate of amplification, $(\alpha c_i)_{max}$, decreases sharply, and when $\eta_s = \eta_1$ ($\theta_w = 0.05$), the first mode disappears completely. Only viscous instability could produce unstable first-mode disturbances at finite Reynolds numbers for $\theta_w \leq 0.05$, but since the action of viscosity has been found to be only stabilizing for $M_1 > 3$ for an insulated surface, such a possibility is not considered likely.

The effect of cooling on the second mode is seen to be quite different. The wave number of the singular neutral solution, α_{12}, increases with increasing cooling, and the wave number α_{s2} associated with $v_0 = 0$ first decreases but then increases. The maximum value of c_i, in direct contrast to the first mode, is affected only a small amount by cooling. Of particular interest is what happens when η_s is either below η_1 or does not exist at all. For $\theta_w = 0.05$, when the two points almost coincide, there is a neutral solution (α_{s2}, $1-1/M_1$) which is an end point of the locus of amplified solutions starting at the singular neutral solution. However, there is in addition another locus of amplified solutions for which (α_{s2}, $1 - 1/M_1$) is an end point. This family of solutions is characterized by phase velocities which are supersonic relative to the free stream. Even though these solutions are supersonic, they still satisfy the boundary condition of zero disturbance amplitude at infinity since $c_i \neq 0$ and the computer program selects the square root of $1 - M_1^2(1 - c)^2$ with a positive real part (the solutions for $\eta > \eta_\delta$ are proportional to $\exp\{-\alpha\left[1-M_1^2(1-c)^2\right]^{1/2}\eta\}$). For $\theta_w < 0.05$, the locus of amplified solutions that starts at the singular neutral solution joins up with the supersonic family of solutions at some $c_i > 0$ without passing through a second neutral point. The junction of the two solutions moves to larger c_i and α with increasing cooling. The supersonic amplified solutions extend to values of α at least as large as unity and to values of c_r as small as 0.4. Apparently another singular neutral point exists near $c_r = 0.4$.

The existence of the supersonic family of amplified solutions for $\theta_w = 0.05$ which joins up with the subsonic family suggests that these solutions might also exist for $\theta_w > 0.05$, in which case they will be completely separate from the subsonic solutions. This expectation proved to be correct for $\theta_w = 0.10$ and 0.15, and eigenvalues for $\theta_w = 0.10$ are shown in Figs. 15 and 16. The interesting feature of this group of solutions is that the end

point at $c_i = 0$ is a supersonic neutral solution. Since the program selects $+ i$ as the square root of -1, this solution is an undamped outgoing wave. The propagation of energy to infinity is balanced by a positive Reynolds stress in the boundary layer which transfers energy from the mean flow to the disturbance as needed for a neutral solution. Supersonic amplified or neutral solutions were also searched for with $\theta_w = 0.25$, but none could be found.

Fig. 18 gives the maximum time rate of amplification of the subsonic solutions as a function of the ratio of the wall temperature to the recovery temperature. The stabilizing effect of cooling on the first mode, and the destabilizing effect on the second mode, are clearly shown. It is only for $T_w/T_r \rightarrow 0$ that a slight destabilizing effect on the second mode is noted. The effect of cooling on the third mode is also destabilizing, but even though the third mode is more unstable than the first mode over most of the range of T_w/T_r, it is the second mode which is dominant at all wall temperatures. Only actual calculations can determine whether the second mode is also destabilized at finite Reynolds numbers by cooling.

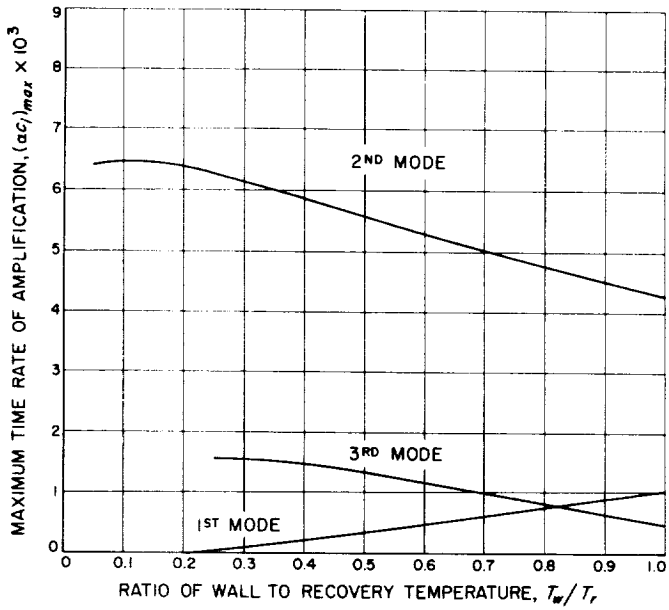


Fig. 18. Maximum time rate of amplification of first three modes at $M_1 = 5.8$ as function of ratio of wall to recovery temperature

In view of the fact that cooling can create a supersonic region of relative flow in the boundary layer at any Mach number, or enlarge an already existing small supersonic region, it is of interest to consider the effect of cooling at Mach numbers where for an insulated surface the instability is dominated by the first mode. The appearance, or the increased importance, of the second mode as a result of cooling could possibly be related to such things as the transition reversal phenomenon or the early transition of highly cooled, low Mach number boundary layers. However, at low Mach numbers, the inviscid theory is less helpful than at high Mach numbers because viscous instability is still important. For instance, at $M_1 = 2.2$ and with an insulated surface, the maximum viscous amplification rate is about 10 times the maximum inviscid amplification rate.

The effect of cooling on the inviscid stability at $M_1 = 2.2$ is to stabilize the first mode and destabilize the second mode, just as at $M_1 = 5.8$. At $T_w/T_r = 0.282$, the second mode $(\alpha c_i)_{max}$ is 0.00125, or about 60 times the first mode $(\alpha c_i)_{max}$ for the uncooled boundary layer, and about 10 times the uncooled second-mode value. The corresponding α is 1.12, compared to 0.897 for the uncooled second mode and 0.032 for the uncooled first mode. Although cooling results in a definite increase in the inviscid instability of the second mode, the high wave number of the most amplified disturbance decreases the chance of the second mode being an important source of instability at finite Reynolds numbers because of the large viscous damping associated with high wave numbers. Unfortunately, these same high wave numbers will make the program for the solution of the complete stability equations inadequate to investigate this point because at $M_1 = 2.2$ the maximum αR at which the program can operate is about 250.

References

1. Grove, A. S., et al, "An Experimental Investigation of the Steady Separated Flow Past a Circular Cylinder," *Journal of Fluid Mechanics*, Vol. 19, Part 1, pp. 60–80, 1964.

2. Batchelor, G. K., "A Proposal Concerning Laminar Wakes Behind Bluff Bodies at Large Reynolds Number," *Journal of Fluid Mechanics*, Vol. 1, pp. 388–398, 1956.

References (Cont'd)

3. Kawaguti, M., "Numerical Solution of the Navier-Stokes Equations for Flow Around a Circular Cylinder at Reynolds Number 40," *Journal of the Physical Society of Japan*, Vol. 8, No. 6, pp. 747–753, 1953.

4. Hama, F. R., and Nutant, J., "Detailed Flow-Field Observations in the Transition Process in a Thick Boundary Layer," *Proceedings of the 1963 Heat Transfer and Fluid Mechanics Institute*, p. 77, 1963.

5. Hama, F. R., "An Efficient Tripping Device," *Journal of the Aeronautical Sciences*, Vol. 24, p. 236, 1957.

6. Klebanoff, P. S., Tidstrom, K. D., and Sargent, L. M., "The Three-Dimensional Nature of Boundary-Layer Instability," *Journal of Fluid Mechanics*, Vol. 12, p. 1, 1962.

7. Potter, J. L., and Whitfield, J. D., "Effect of Slight Nose Bluntness and Roughness on Boundary-Layer Transition in Supersonic Flows," *Journal of Fluid Mechanics*, Vol. 12, p. 501, 1962.

8. van Driest, E. R., and Blumer, C. B., "Boundary-Layer Transition at Supersonic Speeds—Three-Dimensional Roughness Effects (Spheres)," *Journal of the Aerospace Sciences*, Vol. 29, p. 909, 1962.

9. Stewartson, K., "On Rotating Laminar Boundary Layers," *Boundary Layer Research*, Symposium Freiburg (1957), pp. 59–71, Springer-Verlag, Berlin, 1958.

10. Mack, L. M., "The Laminar Boundary Layer on a Rotating Disk of Finite Radius in a Rotating Flow," *SPS 37-18*, Vol. IV, pp. 43–46, Jet Propulsion Laboratory, Pasadena, California, 1962.

11. Rogers, M. H., and Lance, G. N., "The Rotationally Symmetric Flow of a Viscous Fluid in the Presence of an Infinite Rotating Disk," *Journal of Fluid Mechanics*, Vol. 7, Part 4, pp. 617–631, 1960.

12. Squire, H. B., *British Aeronautical Research Committee* 16,021, 1953.

13. Lewellen, W. S., "Linearized Vortex Flows," Report No. ATN-64 (9227)-4, Aerospace Corp., El Segundo, California, 1964.

14. Turner, J. S., "The Constraints Imposed by a Class of Strong Vortices by an Axial Boundary," Woods Hole Oceanographic Institute (Unpublished).

15. Mack, L. M., "The Laminar Boundary Layer on a Disk of Finite Radius in a Rotating Flow," *Research Summary* No. 36-14, pp. 103–105, Jet Propulsion Laboratory, Pasadena, California, 1962.

16. Lees, L., "The Stability of the Laminar Boundary Layer in a Compressible Fluid," NACA Report No. 876, 1947.

17. Mack, L. M., "The Stability of the Compressible Laminar Boundary Layer According to a Direct Numerical Solution. Part IV," *SPS 37-24*, Vol. IV, pp. 271–274, Jet Propulsion Laboratory, Pasadena, California, December 31, 1963.

where $\varepsilon^{\mu\nu\rho\sigma}$ is the four-dimensional Levi-Civita tensor, antisymmetric on all indices with

$$\varepsilon^{0123} = 1 \tag{10}$$

In addition to F and G, one can diagonalize $\underset{\sim}{M}^2 = j(j+1)$ and $M_3 = m$, thus obtaining states $|fgjm>$ or $|k\nu jm>$.

It is well known that the most general Lorentz transformation may be written in terms of two rotations and one translation

$$\Lambda = RZR' \tag{11}$$

where we have taken the translation in the z direction. In terms of the Euler angles the unitary operator representing Λ is given by

$$U(\Lambda) = U(R)\, U(Z)\, U(R') \tag{12}$$

where

$$U(R) = e^{-i\alpha M_3} e^{-i\beta M_2} e^{-i\gamma M_3} \tag{13}$$

$$U(Z) = e^{-i\varepsilon N_3} \tag{14}$$

$$U(R') = e^{-i\beta' M_2} e^{-i\gamma' M_3} \tag{15}$$

$$0 \leq \alpha, \gamma, \gamma' \leq 2\pi,\ 0 \leq \beta, \beta' \leq \pi,\ 0 < \varepsilon < \infty \tag{16}$$

and

$$\tanh \varepsilon = v \tag{17}$$

(with the velocity of light set equal to unity).

With these operators the spherical functions are defined by

$$\left\langle k\nu jm \mid U(R)\, U(Z)\, U(R') \mid k'\nu' j' m' \right\rangle = \delta_{kk'}\, \delta(\nu - \nu') \sum_{j'' m''} \mathcal{D}^{j}_{mm''}(\alpha, \beta, \gamma)\, Z^{k\nu m''}_{jj'}(\varepsilon) \times \mathcal{D}^{j'}_{m'', m'}(0, \beta', \gamma') \tag{18}$$

where the $\mathcal{D}$'s are the well-known matrix elements of finite rotations and

$$Z^{k\nu m''}_{jj'}(\varepsilon) = \left\langle k\nu jm'' \mid e^{-i\varepsilon N_3} \mid k\nu j' m'' \right\rangle \tag{19}$$

Although the matrix elements of N_3 are well known (Ref. 3), the corresponding matrix elements for finite translations have not been obtained. Thus, the problem of determining the relativistic spherical functions reduces to an evaluation of $Z^{k\nu m''}_{jj'}$.

Our derivation of $Z^{k\nu m''}_{jj'}$ is based on the representation theory of L as given by Naimark (Ref. 4). Let H be the Hilbert space of all (non-analytic) functions $f(z)$, $z = x + iy$, for which

$$\| f \|^2 = \iint dxdy\, |f(z)|^2 < \infty \tag{20}$$

The scalar product of two functions f, g in H is

$$(f, g) = \iint dxdy\, \overline{f(z)} \cdot g(z) \tag{21}$$

where the bar denotes complex conjugation.

Now, consider the set of all complex 2×2 unimodular matrices

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \tag{22}$$

that is, the group $SL(2,C)$. Corresponding to an α in $SL(2,C)$ we have a unitary transformation on H such that

$$U(\alpha)f(z) = \left[\overline{(\alpha_{12}z + \alpha_{22})}\right]^{i\nu - k - 1} (\alpha_{12}z + \alpha_{22})^{i\nu + k - 1} f\left(\frac{\alpha_{11}z + \alpha_{21}}{\alpha_{12}z + \alpha_{22}}\right) \tag{23}$$

This mapping is a unitary representation of $SL(2,C)$ on H. The set of all such representations for $2k$ an integer and ν a real number is called the principal series of representations of $SL(2,C)$.[1] For each fixed k and ν these representations are irreducible. However, it is well known (Ref. 5) that the Lorentz group L is homomorphic to $SL(2,C)$. That is, for an arbitrary Lorentz transformation Λ one can find an α such that

$$\alpha = N\left[\Lambda^{\mu}_{\mu} + \sum_{j=1}^{3} (\Lambda^{0}_{j} + \Lambda^{j}_{0} - i\varepsilon^{0j\mu}_{\tau}\, \Lambda^{\tau}_{\mu})\sigma_j\right] \tag{24}$$

[1]The spherical functions for the supplementary series will be given elsewhere.

References (Cont'd)

3. Kawaguti, M., "Numerical Solution of the Navier-Stokes Equations for Flow Around a Circular Cylinder at Reynolds Number 40," *Journal of the Physical Society of Japan,* Vol. 8, No. 6, pp. 747–753, 1953.

4. Hama, F. R., and Nutant, J., "Detailed Flow-Field Observations in the Transition Process in a Thick Boundary Layer," *Proceedings of the 1963 Heat Transfer and Fluid Mechanics Institute,* p. 77, 1963.

5. Hama, F. R., "An Efficient Tripping Device," *Journal of the Aeronautical Sciences,* Vol. 24, p. 236, 1957.

6. Klebanoff, P. S., Tidstrom, K. D., and Sargent, L. M., "The Three-Dimensional Nature of Boundary-Layer Instability," *Journal of Fluid Mechanics,* Vol. 12, p. 1, 1962.

7. Potter, J. L., and Whitfield, J. D., "Effect of Slight Nose Bluntness and Roughness on Boundary-Layer Transition in Supersonic Flows," *Journal of Fluid Mechanics,* Vol. 12, p. 501, 1962.

8. van Driest, E. R., and Blumer, C. B., "Boundary-Layer Transition at Supersonic Speeds—Three-Dimensional Roughness Effects (Spheres)," *Journal of the Aerospace Sciences,* Vol. 29, p. 909, 1962.

9. Stewartson, K., "On Rotating Laminar Boundary Layers," *Boundary Layer Research,* Symposium Freiburg (1957), pp. 59–71, Springer-Verlag, Berlin, 1958.

10. Mack, L. M., "The Laminar Boundary Layer on a Rotating Disk of Finite Radius in a Rotating Flow," *SPS 37-18,* Vol. IV, pp. 43–46, Jet Propulsion Laboratory, Pasadena, California, 1962.

11. Rogers, M. H., and Lance, G. N., "The Rotationally Symmetric Flow of a Viscous Fluid in the Presence of an Infinite Rotating Disk," *Journal of Fluid Mechanics,* Vol. 7, Part 4, pp. 617–631, 1960.

12. Squire, H. B., *British Aeronautical Research Committee* 16,021, 1953.

13. Lewellen, W. S., "Linearized Vortex Flows," Report No. ATN-64 (9227)-4, Aerospace Corp., El Segundo, California, 1964.

14. Turner, J. S., "The Constraints Imposed by a Class of Strong Vortices by an Axial Boundary," Woods Hole Oceanographic Institute (Unpublished).

15. Mack, L. M., "The Laminar Boundary Layer on a Disk of Finite Radius in a Rotating Flow," *Research Summary No. 36-14,* pp. 103–105, Jet Propulsion Laboratory, Pasadena, California, 1962.

16. Lees, L., "The Stability of the Laminar Boundary Layer in a Compressible Fluid," NACA Report No. 876, 1947.

17. Mack, L. M., "The Stability of the Compressible Laminar Boundary Layer According to a Direct Numerical Solution. Part IV," *SPS 37-24,* Vol. IV, pp. 271–274, Jet Propulsion Laboratory, Pasadena, California, December 31, 1963.

Erratum

The following correction should be noted for *SPS 37-28*, Vol. IV: The first equation appearing on p. 124 should read:

$$\frac{U_0 - U}{U_0 - U_c} = \exp\left(-\alpha^2 \eta^2\right) J_0\left(i\eta\right)$$

XVIII. Physics

A. Effective Resonance Integrals in Heterogeneous Arrays

H. Ludewig

It is to be expected that flux depressions in heterogeneous reactor elements may markedly affect neutron absorption at resonance energies.[1]

To begin an investigation of this phenomenon, the fractional change in resonance integral is being computed for a simple one-dimensional system made up of a slab of resonance absorber embedded in a neutron source region. Breit-Wigner resonance line shapes are assumed for the cross section, and the resonance integral is defined in the usual way (Ref. 1).

A preliminary calculation with an assumed quadratic spatial flux distribution in the element demonstrated significant variation in resonance integral from that normally calculated with flat flux distribution, over certain ranges of slab size. This preliminary calculation did not include the effect of a flux depression on the collision density in the slab.

[1]Cohen, E. R., Lecture Notes, A. M. 201, California Institute of Technology, Pasadena, California.

In view of this result, a more accurate model has been formulated in which the flux variation is contained in the collision density equation:

$$\left[\begin{array}{l}\text{coll. density at } du \text{ about } u,\\ d\bar{r} \text{ about } \bar{r}, \text{ and } d\Omega \text{ about } \Omega\end{array}\right]$$

$$= \int_{\bar{u}'}\int_{\bar{r}'}\int_{\bar{\Omega}'} \left[\begin{array}{l}\text{scat. coll. density at } du'\\ \text{about } u', d\bar{r}' \text{ about } \bar{r}',\\ \text{and } d\bar{\Omega}' \text{ about } \bar{\Omega}'\end{array}\right]$$

$$\times \left[\begin{array}{l}\text{prob. of scat. from}\\ u'\bar{\Omega}' \text{ to } u\bar{\Omega}\end{array}\right]$$

$$\times \left[\begin{array}{l}\text{prob. of traveling the distance}\\ |\bar{r}' - \bar{r}| \text{ and having the first}\\ \text{coll. in } d\bar{r} \text{ about } \bar{r}, du\\ \text{about } u, \text{and } d\bar{\Omega} \text{ about } \bar{\Omega}'\end{array}\right] d\bar{\Omega}' d\bar{r}'\, du'$$

$$+ \left[\begin{array}{l}\text{coll. density due to source of neutron at}\\ du \text{ about } u, dr \text{ about } \bar{r}, \text{ and } d\bar{\Omega} \text{ about } \bar{\Omega}\end{array}\right]$$

where $u_2 = \ln E^*/E$ = lethargy; E^* = reference energy.

Equations of the above type have been previously treated (Ref. 2); however, in that treatment the assumption was made that the neutrons lost no energy when colliding with an absorber nucleus—which is a good approximation in the case of heavy absorbers. An assumption of this

nature eliminates the lethargy integral in the above equation; however, by retaining the integral it is possible to treat problems in which absorbing material is mixed with a moderator, i.e., uranium oxide, uranium carbide, or uranium mixed with zirconium hydride, etc. In these cases the neutrons will certainly slow down in the absorbing element and the integral should be included.

The collision density has been expanded in two series of Legendre polynomials; one for angular variations, and one for spatial variations. This leads to an integral equation which determines the lethargy dependent expansion coefficient. At the present time an attempt is being made to solve this integral equation by an iteration method; once this is done and the coefficients are known, the resonance integral can be computed, and hence the fractional change may be determined.

B. Zero-Order, Degenerate, Time-Dependent Perturbation Theory—Modes in Resonantly Driven Quantum Systems

M. M. Saffren

Von Roos, in his study of the interaction of laser beams with matter, has examined the interaction of a laser beam with a hydrogen atom when the beam has a frequency equal to a level difference in the atom. This necessitates degenerate time-dependent perturbation theory. Apparently there has been no careful study of such perturbations except perhaps for magnetic resonance phenomena. In this article we give a general treatment and find zero-order wave functions that resemble stationary eigenfunctions and that are independent of initial conditions. We then compare time-dependent degenerate perturbation theory with time-independent degenerate perturbation theory.

We start with a Hamiltonian H that has a time-independent part, H_0, whose eigenfunctions ψ_n and eigenvalues $\hbar\omega_n$ we assume to be known, and a time dependent part:

$$H - H_0 = \sum_\alpha H_\alpha \exp(i\omega_\alpha t) + \sum H_\alpha^+ \exp(-i\omega_\alpha t) \quad (1)$$

(H_α^+ denotes adjoint of H_α.) The wave function of the system, ψ, has the equation of motion

$$\left(H_0 - \frac{\hbar}{i}\frac{\partial}{\partial t}\right)\psi = (H_0 - H)\,\psi \quad (2)$$

In the usual way we expand ψ as

$$\psi = \sum a_n(t)\,\psi_n(r)\exp(i\omega_n t) \quad (3)$$

and find

$$\frac{\hbar}{i}\dot{a}_s(t) = \sum_{\alpha,m}(H_\alpha)_{sm}\exp\left(i(\omega_\alpha - \omega_s + \omega_m)t\right)a_m(t) + \sum_{\alpha,n}(H_\alpha^+)_{sn}\exp\left(i(-\omega_\alpha - \omega_s + \omega_n)t\right)a_n(t) \quad (4)$$

We now split up the two sums into "secular" terms and ordinary terms. The secular terms are characterized by the resonance conditions $\omega_\alpha - \omega_s + \omega_m = 0$, or $-\omega_\alpha - \omega_s + \omega_n = 0$. We have

$$\begin{aligned}\frac{\hbar}{i}\dot{a}_s(t) = \sum_{\alpha,m}\{&(H_\alpha)_{sm}\,\delta(\omega_\alpha - \omega_s + \omega_m) \\ &+ (H_\alpha^+)_{sm}\,\delta(\omega_\alpha - \omega_m + \omega_s)\}\,a_m(t) \\ &+ \sum_{\alpha,m}\{(H_\alpha)_{sm}\left[1 - \delta(\omega_\alpha - \omega_s + \omega_m)\right] \\ &\times \exp(i\omega_\alpha t) + (H_\alpha^+)_{sm} \\ &\times \left[1 - \delta(\omega_\alpha - \omega_m + \omega_s)\right]\exp(-i\omega_\alpha t)\} \\ &\times \exp\left(i(\omega_m - \omega_s)t\right)a_m(t)\end{aligned} \quad (5)$$

Here $\delta(\omega_\alpha - \omega_s + \omega_m)$ is a delta function which we hereafter write as $\delta_{\alpha sm}$.

We can eliminate the secular terms by writing for the $a_s(t)$ involved in a resonance,

$$a_{s'}(t) = \sum_j a_{s'j}\,\alpha_j(t)\exp(i\Omega_j t) \quad (6)$$

Here Ω_j and $a_{s'j}$ are constants, and if there are N states in resonance the sum runs to N. The a_{nj} are solutions of

$$\sum \mathcal{H}_{sn}\,a_{nj} = \hbar\,\Omega_j\,a_{nj} \quad (7)$$

Evidently the Ω_j are the eigenvalues of the matrix

$$\mathcal{H}_{sn} = \sum_\alpha\left[(H_\alpha)_{sn}\,\delta_{\alpha sn} + (H_\alpha^+)_{sn}\,\delta_{\alpha ns}\right] \quad (8)$$

Since the matrix $\{a_{nj}\}$ diagonalizes $\mathcal{H}$, the matrix is unitary if we can show $\mathcal{H}$ is hermitian. But clearly $(H_\alpha^+)_{sn} = (H_\alpha)_{ns}^+$, so $\mathcal{H}_{sn} = \mathcal{H}_{ns}^+$, and $\mathcal{H}$ is indeed hermitian. Thus the relation $\sum a_{sk}^+ a_{s'j} = \delta_{jk}$ expresses the unitarity of the matrix a'_{sk}. (We note, using the invariance of the trace,

that since $\mathcal{H}_{nn} = 0$, $\sum \Omega_j = 0$.) The $\alpha_j(t)$ correspond to amplitudes of modes ϕ_j,

$$\phi_j = \sum a_{s'j}\, \psi_{s'}(r) \exp(i\omega_{s'}t) \tag{9}$$

which appear in the new expansion of ψ:

$$\psi = \sum_{m \neq m'} a_m(t)\, \psi_m \exp(i\omega_m t) + \sum_j \alpha_j(t)\, \phi_j(r,t) \exp(i\Omega_j t) \tag{10}$$

Now by expressing a_s in terms of α_j we obtain

$$\frac{\hbar}{i} \sum_j a_{s'j}\, \dot{\alpha}_j(t) \exp(i\Omega_j t) = \sum_{m \neq m'} h^{(0)}_{s'm}(t) \exp(i(\omega_m - \omega_{s'})t)\, a_m(t) + \sum_{m',r} h^{(0)}_{s'm'}(t)\, a_{m'r} \exp(i(\omega_{(m'r)} - \omega_{s'})t)\, \alpha_r(t) \tag{11}$$

Here

$$h^{(0)}_{s'm}(t) = \sum_\alpha \{(H_\alpha)_{s'm}[1 - \delta_{\alpha s'm}] \exp(i\omega_\alpha t) + (H_\alpha)_{ms'}[1 - \delta_{\alpha m s'}] \exp(-i\omega_\alpha t)\} \tag{12}$$

and

$$\omega_{(m'r)} = \omega_{m'} + \Omega_r \tag{13}$$

Using the unitarity of a_{sj}, we find that

$$\frac{\hbar}{i} \dot{\alpha}_k = \sum_{s',\, m \neq m'} h^{(0)}_{s'm}(t)\, a^+_{s'k} \exp\left(i[\omega_m - \omega_{(s'k)}]t\right) a_m(t) + \sum_{s',\, m',\, r} h^{(0)}_{s'm'}(t)\, a_{m'r}\, a^+_{s'k} \times \exp(i(\omega_{(m'r)} - \omega_{(s'k)})t)\, \alpha_r(t) \tag{14}$$

If $s \neq s'$, we also have

$$\frac{\hbar}{i} \dot{a}_s = \sum_{m \neq m'} h^{(0)}_{sm}(t) \exp(i(\omega_m - \omega_s)t)\, a_m(t) + \sum_{r,\, m'} \{h^{(0)}_{sm'}(t)\, a_{m'r}\} \exp(i(\omega_{(m'r)} - \omega_s)t)\, \alpha_r(t) \tag{15}$$

The resonance conditions now become

$$\omega_{(m'r)} - \omega_s = \pm\, \omega_\alpha \tag{16}$$

and

$$\omega_{(m'r)} - \omega_{(js')} = \pm\, \omega_\alpha \tag{17}$$

Letting s'' denote a state resonant with a mode, and $\tilde{r}$ denote a mode resonant either with a state or another mode, we introduce amplitudes $\beta_j(t)$ and set

$$a_{s''}(t) = \sum_j \beta_j(t) \exp(i\bar{\Omega}_j t)\, b_{s''j} \tag{18}$$

$$\alpha_{\tilde{r}}(t) = \sum_j \beta_j(t) \exp(i\bar{\Omega}_j t)\, b_{\tilde{r}j} \tag{19}$$

The $b_{s''j}$ and b_{rj} are determined from

$$\sum_{\tilde{r}} \bar{\mathcal{H}}_{s'',\tilde{r}}\, b_{\tilde{r}j} = \bar{\Omega}_j\, b_{s''j} \tag{20}$$

$$\sum_{s''} \mathcal{H}_{\tilde{r},s''}\, b_{s''j} + \sum_{\tilde{\chi}} \bar{\mathcal{H}}_{\tilde{r}\tilde{\chi}}\, b_{\tilde{k}j} = \bar{\Omega}_j\, b_{\tilde{r}j} \tag{21}$$

Here

$$\bar{\mathcal{H}}_{s''\tilde{r}} = \sum_{\alpha,\, m'} [(H_\alpha)_{s''m'}(1 - \delta_{\alpha s''m'})\, \delta_{\alpha s''(m'\tilde{r})} + (H_\alpha)^+_{m's''}(1 - \delta_{\alpha m's''})\, \delta_{\alpha(m'\tilde{r})s''}]\, a_{m'\tilde{r}} \tag{22}$$

$$\bar{\mathcal{H}}_{\tilde{r}\tilde{k}} = \sum_{s',\, m'} [(H_\alpha)_{s'm'}(1 - \delta_{\alpha s''m'})\, \delta_{\alpha(m'\tilde{k})\,(s',\tilde{r})} + (H_\alpha)^+_{m's'}(1 - \delta_{\alpha m's'})\, \delta_{\alpha(s'\tilde{r})\,(m'\tilde{k})}]\, a^+_{s'\tilde{r}}\, a_{m'\tilde{k}} \tag{23}$$

This secular equation introduces the new modes

$$\vartheta_j = \sum_{\tilde{r}} b_{\tilde{r}j}\, \phi_{\tilde{r}} \exp(i\Omega_{\tilde{r}}t) + \sum_{s''} b_{s''j}\, \psi_{s''} \exp(i\omega_{s''}t) \tag{24}$$

and a corresponding modified expansion for ψ,

$$\psi = \sum_{m \neq m',\, m''} a_m(t)\, \psi_m \exp(i\omega_m t) + \sum_{\tilde{r} \neq \tilde{r}} \alpha_r(t)\, \phi_r \exp(i\Omega_r t) + \sum_j \beta_j\, \vartheta_j \exp(i\bar{\Omega}_j t) \tag{25}$$

$$\vartheta_j = \sum_{\tilde{r},\, s} b_{\tilde{r}j}\, a_{s\tilde{r}}\, \psi_s \exp(i\omega_s t) \exp(i\Omega_{\tilde{r}}t) \tag{26}$$

The equations of motion of the amplitudes a_m, α_r, β_j are rather complicated and will not be given here.

We return now to the case where the β modes are absent, and indicate the significance of the α modes. We now suppose that the system at $t = 0$ is in a state, s', that appears in one of the modes. Then to zero order $\alpha_r(t) = \alpha_r(0)$, and $\alpha_r(0) = a^+_{s'r}$. To zero order the wave function, ψ, is then

$$\sum_{r,\, m'} A_{m'r}(s')\, \psi_{m'} \exp(i\omega_{m'}t) \exp(i\Omega_r t),\ A_{m'r}(s') = a^+_{s'r}\, a_{m'r} \tag{27}$$

Physical quantities, Q, associated with the system (Q = dipole moment, or charge, or current, etc.) have the expectation value

$$\langle\psi|Q|\psi\rangle = \sum_{m',n',r,k} A^{+}_{m'r}\, A_{n'k}\, Q_{m'n'} \times \exp\left(i[(\omega_{n'} + \Omega_k) - (\omega_{m'} + \Omega_r)]t\right) \quad (28)$$

Thus, to zero order the resonant perturbation causes the system to act as if it now had $N_{m'}$ stationary levels $\hbar\,[\omega_{m'} + \Omega_r]$, $r = 1, \cdots, N_{m'}$ ($N_{m'}$ is the number of states made degenerate with m' by resonance) where it had only one, $\hbar\,\omega_{m'}$, when no resonant perturbation was applied. It is not hard to see that these new levels persist to all higher orders.

As an application of this equation let Q be the projection $Q_{m'}$ on to the state m'. Then we see that the probability of finding the system in the state m' is

$$\langle\psi|Q_{m'}|\psi\rangle = |\langle\psi_{m'}|\psi\rangle|^2 = \sum_{r,k} A^{+}_{m'r}\, A_{m'k} \exp\left(i[\Omega_k - \Omega_r]t\right) = \sum_{r,k} A^{+}_{m'r}\, A_{m'k} \cos\left(\Omega_k - \Omega_r\right) t \quad (29)$$

Since at $t = 0$ this is $\delta_{m's'}$, we see that we have obtained in zero order the probability of finding the system in the state m', knowing that at $t = 0$ it was in the state s'.

Connection with degenerate perturbation theory for stationary systems. We know that the explicit time dependence of a Hamiltonian reflects the fact that it describes a subsystem interacting with the larger system of which it is a part. The Hamiltonian of the entire system is time independent, however. Thus all time-dependent Hamiltonians are approximate, and so the modes found above are also approximate. In this section we obtain a time-dependent Hamiltonian from a time-independent one, and compare the modes derived from each. We find the modes to be different unless certain care is exercised in deriving the time-dependent Hamiltonian from the time-independent one.

We proceed now to derive the time-dependent Hamiltonian. Let $H = H_0 + H_0' + H_I$ denote a time-independent Hamiltonian, where H_0 and H_0' commute with one another but not with H_I. Now $H\bar{\psi} = \hbar/i(\partial\bar{\psi}/\partial t)$ is the equation of motion of the wave function $\bar{\psi}$, and we expand $\bar{\psi}$ in eigenfunctions of H_0' as

$$\sum_n \bar{\psi}_n(t)\, \tilde{\psi}_n \exp\left(i\tilde{\omega}_n' t\right) \quad (30)$$

where $\tilde{\psi}_n$ and $\hbar\,\tilde{\omega}_n$ are the eigenfunctions and eigenvalues of H_0', and $\bar{\psi}_n(t)$ is a function of the variables of H_0. We then have as the equation of motion

$$\sum_n H_0\, \bar{\psi}_n(t)\, \tilde{\psi}_n \exp\left(\tilde{\omega}_n t\right) + \sum_n H_I\, \bar{\psi}_n(t)\, \tilde{\psi}_n \exp\left(i\tilde{\omega}_n t\right) = \frac{\hbar}{i} \sum \frac{\partial\psi_n(t)}{\partial t}\, \tilde{\psi}_n \exp\left(i\tilde{\omega}_n t\right) \quad (31)$$

which becomes

$$H_0\, \bar{\psi}_n(t) + \sum_{m'} (H_I)_{mn} \exp\left(i[\omega_n - \omega_m]t\right) \bar{\psi}_m = \frac{\hbar}{i} \frac{\partial\bar{\psi}_n}{\partial t} \quad (32)$$

We now turn H into a time-dependent Hamiltonian for the untilded variable by assuming that

$$\bar{\psi} = \psi(t) \sum A_n^{+}\, \tilde{\psi}_n \exp\left(i\tilde{\omega}_n t\right) \quad (33)$$

Physically this ansatz assumes that the part of the system described by the tilded variables drives the subsystem described by the untilded variables, there being no reaction back on the driver by the driven system. The wave function ψ then satisfies

$$H(t)\,\psi \equiv H_0\,\psi + \frac{\left[\sum A_m^{+} A_n\, H_{mn}\, e^{i[\tilde{\omega}_n - \tilde{\omega}_m]t}\right]}{\sum |A_m|^2}\,\psi = \frac{\hbar}{i}\frac{\partial\psi}{\partial t} \quad (34)$$

By ignoring the reaction we are regarding the tilded subsystem as classical. If we denote the sum,

$$\sum_{mn} \delta\left(\omega_\alpha + \tilde{\omega}_m - \tilde{\omega}_n\right) A_m^{+}\, A_n\, H_{nn} \quad (35)$$

by H_α, we see $H(t)$ to be the Hamiltonian for ψ that we have been considering all along (we assume also that $H_\alpha = 0$ when $\alpha = 0$). If we now write

$$\psi = \sum a_n(t) \exp\left(i\omega_n t\right) \psi_n \quad (36)$$

we have in fact

$$\frac{\hbar}{i}\,\dot{a}_m(t) = \sum a_n(t)\, A_b^{+}\, A_c\, (H_{bc})_{m,n} \times \exp\left(i[(\tilde{\omega}_c + \omega_n) - (\tilde{\omega}_b + \omega_m)]t\right) \quad (37)$$

$$\sum |A_b|^2 = 1 \quad (38)$$

This makes a connection between the time-independent Hamiltonian H and the time-dependent Hamiltonian $H(t)$, which we considered before.

Had we not tried to make this connection but instead written down the precise equation of motion of $\underline{\bar{\psi}}$, we would have written, to begin with,

$$\underline{\bar{\psi}} = \sum B_{bm}(t)\, \tilde{\bar{\psi}}_b\, \psi_m \exp\left(i[\tilde{\omega}_b + \omega_m]t\right) \tag{39}$$

and we would have obtained at this point,

$$\frac{\hbar}{i}\dot{B}_{bm}(t) = \sum B_{cn}(t)\,(H_{bc})_{m,n} \times \exp\left(i\left[(\tilde{\omega}_c + \omega_n) - (\tilde{\omega}_b + \omega_m)\right]t\right) \tag{40}$$

instead of the expression obtained for $\dot{a}_m(t)$.

The secular equation would have been

$$\sum_{bm,\,cn} B_{bm,j}\,(H_{bc})_{mn}\,\delta\left[(\tilde{\omega}_c + \omega_n) - (\tilde{\omega}_b + \omega_m)\right] = \hbar W_j\, B_{cn;j} \tag{41}$$

compared to

$$\sum a_{mj}\, A_b^+\,(H_{bc})_{mn}\,\delta\left[(\tilde{\omega}_0 + \omega_n) - (\tilde{\omega}_b + \omega_n)\right] = \hbar\,\Omega_j\, A_c^+\, A_{nj} \tag{42}$$

The corresponding modes are

$$\phi_j = \sum_{bm} B_{bm,j}\, \tilde{\bar{\psi}}_b\, \psi_m \exp\left(i[\omega_0 + \omega_m]t\right) \tag{43}$$

in one case, and

$$\bar{\phi}_j = \left(\sum_m a_{mj}\, \psi_m\, e^{i\omega_m t}\right)\left(\sum_b A_b^+\, \tilde{\bar{\psi}}_b\, e^{i\tilde{\omega}_b t}\right) \tag{44}$$

in the other; there is no apparent relation between them. However, the a_{mj} are functions of the A_b, and the A_b can presumably be chosen in some self-consistent fashion, so as to minimize the difference between the modes. One such fashion might be to require that

$$\delta_{A_b} \sum_j |W_j - \Omega_j|^2 = 0 \tag{45}$$

where δ_{A_b} signifies variation with respect to A_b. If, with optimum values of A_b, $\sum_j |W_j - \Omega_j|^2$ is indeed small, the two sets of modes ϕ_j, $\bar{\phi}_j$ are interchangeable, at least in zero-order calculations.

In any case we see that some care must be used in replacing H by $H(t)$, and that the modes derived from $H(t)$ may not correspond to the modes derived from H unless $H(t)$ is chosen properly. So although energy shifts can be calculated either way, it would always seem preferable to calculate them from the time-independent Hamiltonian H rather than from $H(t)$.

C. Spherical Functions of the Lorentz Group

P. Burt and J. S. Zmuidzinas

The determination of spherical functions for the homogeneous Lorentz group is of interest in connection with studies in elementary particle physics. However, the only cases treated have been for zero spin. In this article we wish to give a derivation of the relativistic spherical functions for arbitrary spin. This generality is necessary for the study of physically realized elementary particle states.

As the basis element of the Lie algebra of the Lorentz group L we take

$$M_{\mu\nu} = (\underset{\sim}{M}, \underset{\sim}{N}) \tag{1}$$

$$\underset{\sim}{M} = (M_{23}, M_{31}, M_{12}) \tag{2}$$

$$\underset{\sim}{N} = (M_{01}, M_{02}, M_{03}) \tag{3}$$

$$M_{\mu\nu} = -M_{\nu\mu} \tag{4}$$

The commutation relations of $M_{\mu\nu}$ are

$$[M_{\mu\nu}, M_{\rho\sigma}] = i\left[g_{\rho\nu} M_{\mu\sigma} - g_{\sigma\nu} M_{\mu\rho} + g_{\mu\sigma} M_{\nu\rho} - g_{\mu\rho} M_{\nu\sigma}\right] \tag{5}$$

where Greek indices run from 0 to 3 and

$$g_{00} = -g_{11} = -g_{22} = -g_{33} = 1 \tag{6}$$

$$g_{\mu\nu} = 0,\ \mu \neq \nu \tag{7}$$

Irreducible unitary representations of L are labeled by the eigenvalues of the two Casimir operators

$$F = -\frac{1}{2} M_{\mu\nu} M^{\mu\nu} = \underset{\sim}{N}^2 - \underset{\sim}{M}^2 = f = 1 + \nu^2 - k^2 \tag{8}$$

$$G = \frac{1}{4}\varepsilon^{\mu\nu\rho\sigma} M_{\mu\nu} M_{\rho\sigma} = 2\,\underset{\sim}{M}\cdot\underset{\sim}{N} = g = 2k\nu \tag{9}$$

where $\epsilon^{\mu\nu\rho\sigma}$ is the four-dimensional Levi-Civita tensor, antisymmetric on all indices with

$$\epsilon^{0123} = 1 \tag{10}$$

In addition to F and G, one can diagonalize $\underset{\sim}{M}^2 = j(j+1)$ and $M_3 = m$, thus obtaining states $|fgjm>$ or $|k\nu jm>$.

It is well known that the most general Lorentz transformation may be written in terms of two rotations and one translation

$$\Lambda = RZR' \tag{11}$$

where we have taken the translation in the z direction. In terms of the Euler angles the unitary operator representing Λ is given by

$$U(\Lambda) = U(R)\, U(Z)\, U(R') \tag{12}$$

where

$$U(R) = e^{-i\alpha M_3}\, e^{-i\beta M_2}\, e^{-i\gamma M_3} \tag{13}$$

$$U(Z) = e^{-i\varepsilon N_3} \tag{14}$$

$$U(R') = e^{-i\beta' M_2}\, e^{-i\gamma' M_3} \tag{15}$$

$$0 \leq \alpha, \gamma, \gamma' \leq 2\pi,\ 0 \leq \beta, \beta' \leq \pi,\ 0 < \varepsilon < \infty \tag{16}$$

and

$$\tanh \varepsilon = v \tag{17}$$

(with the velocity of light set equal to unity).

With these operators the spherical functions are defined by

$$\left\langle k\nu jm \mid U(R)\, U(Z)\, U(R') \mid k'\nu' j' m' \right\rangle$$

$$= \delta_{kk'}\, \delta(\nu - \nu') \sum_{j'' m''} \mathcal{D}^{j}_{mm''}(\alpha, \beta, \gamma)\, Z^{k\nu m''}_{jj'}(\varepsilon)$$

$$\times\ \mathcal{D}^{j'}_{m'', m'}(0, \beta', \gamma') \tag{18}$$

where the $\mathcal{D}$'s are the well-known matrix elements of finite rotations and

$$Z^{k\nu m''}_{jj'}(\varepsilon) = \left\langle k\nu jm'' \mid e^{-i\varepsilon N_3} \mid k\nu j' m'' \right\rangle \tag{19}$$

Although the matrix elements of N_3 are well known (Ref. 3), the corresponding matrix elements for finite translations have not been obtained. Thus, the problem of determining the relativistic spherical functions reduces to an evaluation of $Z^{k\nu m''}_{jj'}$.

Our derivation of $Z^{k\nu m''}_{jj'}$ is based on the representation theory of L as given by Naimark (Ref. 4). Let H be the Hilbert space of all (non-analytic) functions $f(z)$, $z = x + iy$, for which

$$\| f \|^2 = \iint dxdy\, |f(z)|^2 < \infty \tag{20}$$

The scalar product of two functions f, g in H is

$$(f, g) = \iint dxdy\, \overline{f(z)} \cdot g(z) \tag{21}$$

where the bar denotes complex conjugation.

Now, consider the set of all complex 2×2 unimodular matrices

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \tag{22}$$

that is, the group $SL(2,C)$. Corresponding to an α in $SL(2,C)$ we have a unitary transformation on H such that

$$U(\alpha)f(z) =$$

$$\left[\overline{(\alpha_{12}z + \alpha_{22})}\right]^{i\nu - k - 1} (\alpha_{12}z + \alpha_{22})^{i\nu + k - 1} f\left(\frac{\alpha_{11}z + \alpha_{21}}{\alpha_{12}z + \alpha_{22}}\right) \tag{23}$$

This mapping is a unitary representation of $SL(2,C)$ on H. The set of all such representations for $2k$ an integer and ν a real number is called the principal series of representations of $SL(2,C)$.[1] For each fixed k and ν these representations are irreducible. However, it is well known (Ref. 5) that the Lorentz group L is homomorphic to $SL(2,C)$. That is, for an arbitrary Lorentz transformation Λ one can find an α such that

$$\alpha = N\left[\Lambda^{\mu}_{\mu} + \sum_{j=1}^{3} (\Lambda^{0}_{j} + \Lambda^{j}_{0} - i\varepsilon^{0j\mu}_{\tau}\, \Lambda^{\tau}_{\mu})\sigma_j\right] \tag{24}$$

[1] The spherical functions for the supplementary series will be given elsewhere.

where

$$N = \left[4 + (\Lambda_\nu^\nu)^2 - \Lambda_\nu^\mu \Lambda_\mu^\nu + i\varepsilon^{\mu\nu\kappa\lambda} \Lambda_{\mu\nu} \Lambda_{\kappa\lambda}\right]^{-\frac{1}{2}} \tag{25}$$

and the σ_j are the Pauli spin matrices. Thus, for our purposes, the problem of determining the representation of L reduces to that of finding the principal series of representations of $SL(2C)$.[2] In particular, for fixed ν and k we look for functions $F_{jm}^{k\nu}(z)$ such that

$$M_3 F_{jm}^{k\nu} = m F_{jm}^{k\nu} \tag{26}$$

$$\underset{\sim}{M}^2 F_{jm}^{k\nu} = j(j+1) F_{jm}^{k\nu} \tag{27}$$

$$F F_{jm}^{k\nu} = (\underset{\sim}{N}^2 - \underset{\sim}{M}^2) F_{jm}^{k\nu} = (1 + \nu^2 - k^2) F_{jm}^{k\nu} \tag{28}$$

$$G F_{jm}^{k\nu} = 2\underset{\sim}{M} \cdot \underset{\sim}{N} F_{jm}^{k\nu} = 2k\nu F_{jm}^{k\nu} \tag{29}$$

These functions may be found by writing the infinitesimal generators as

$$M_+ = M_1 + iM_2 = -\partial - \bar{z}^2 \bar{\partial} + (i\nu - 1 - k)\bar{z} \tag{30}$$

$$M_- = M_1 - iM_2 = z^2\partial + \bar{\partial} - (i\nu - 1 + k) z \tag{31}$$

$$M_3 = -z\partial + \bar{z}\bar{\partial} + k \tag{32}$$

$$N_+ = N_1 + iN_2 = i\left[\partial - \bar{z}^2 \bar{\partial} + (i\nu - 1 - k)\bar{z}\right] \tag{33}$$

$$N_- = N_1 - iN_2 = i\left[-z^2\partial + \bar{\partial} + (i\nu - 1 + k) z\right] \tag{34}$$

$$N_3 = i\left[z\partial + \bar{z}\bar{\partial} - i\nu + 1\right] \tag{35}$$

with

$$\partial = \frac{\partial}{\partial z} \tag{36}$$

$$\bar{\partial} = \frac{\partial}{\partial \bar{z}} \tag{37}$$

and z and $\bar{z}$ are treated as independent variables. With these generators one can easily verify that Eqs. (28) and (29) are satisfied identically. Furthermore, letting $z = re^{i\theta}$, Eqs. (26) and (27) are solved by

$$F_{jm}^{k\nu}(r, \theta) = N_{jm}^{k\nu} e^{i(k-m)\theta} r^{k-m}(1 + r^2)^{i\nu+m-1} \times P_{j+m}^{k-m,-k-m}\left(\frac{1 - r^2}{1 + r^2}\right) \tag{38}$$

where

$$N_{jm}^{k\nu} = e^{i\eta} \pi^{-\frac{1}{2}} \left[\frac{(2j+1)(j+m)!(j-m)!}{(j+k)!(j-k)!}\right]^{\frac{1}{2}} \tag{39}$$

and $\eta = \eta(j, \nu)$ is a real number chosen so that the $F_{jm}^{k\nu}$ satisfy the conditions of a canonical basis (Ref. 3). $P_n^{\alpha\beta}$ is a Jacobi polynomial defined by

$$P_n^{\alpha\beta}(x) = \frac{(-)^n}{2^n n!}(1 - x)^{-\alpha}(1 + x)^{-\beta} \frac{d^n}{dx^n} \times \left[(1 - x)^{n+\alpha}(1 + x)^{n+\beta}\right] \tag{40}$$

The orthogonality relation for the F's is

$$\int_0^{2\pi} d\theta \int_0^\infty r dr\, \overline{F_{j'm'}^{k\nu}(r, \theta)}\, F_{jm}^{k\nu}(r, \theta) = \delta_{jj'} \delta_{mm'} \tag{41}$$

To compute the matrix element Z, we note that the unitary operator $\exp(-i\varepsilon N_3)$ in Eq. (19) is given by $U(\alpha)$,

$$\alpha = \begin{pmatrix} e^{-\varepsilon/2} & 0 \\ 0 & e^{\varepsilon/2} \end{pmatrix} \tag{42}$$

in the representation given by Eq. (23). Thus, one has

$$Z_{jj'}^{k\nu m} = e^{(i\nu-1)\varepsilon} \int_0^{2\pi} d\theta \int_0^\infty r dr\, \overline{F_{jm}^{k\nu}(r, \theta)}\, F_{j'm}^{k\nu}(e^{-\varepsilon/2} r, \theta) \tag{43}$$

The evaluation of the integrals in Eq. (43) is straightforward. One obtains finally

$$Z_{jj'}^{k\nu m} = \beta(jj';k\nu m;\varepsilon) \sum_{p=0}^{j+m} \sum_{q=0}^{j'+m} C(jj';km;p,q) \times F(q + 1 - m - i\nu, k - m + 1; p + q + 2 - 2m; 1 - e^{-2\varepsilon}) \tag{44}$$

[2] This problem has been considered by Joos (Ref. 3); however, his results are incorrect.

T_1 F-40X OUTPUT R_1 1 Ω R_2 750 Ω R_3 500 Ω R_4 2 kΩ R_5 10 kΩ

JPL SPACE PROGRAMS SUMMARY NO. 37-29, VOL. IV

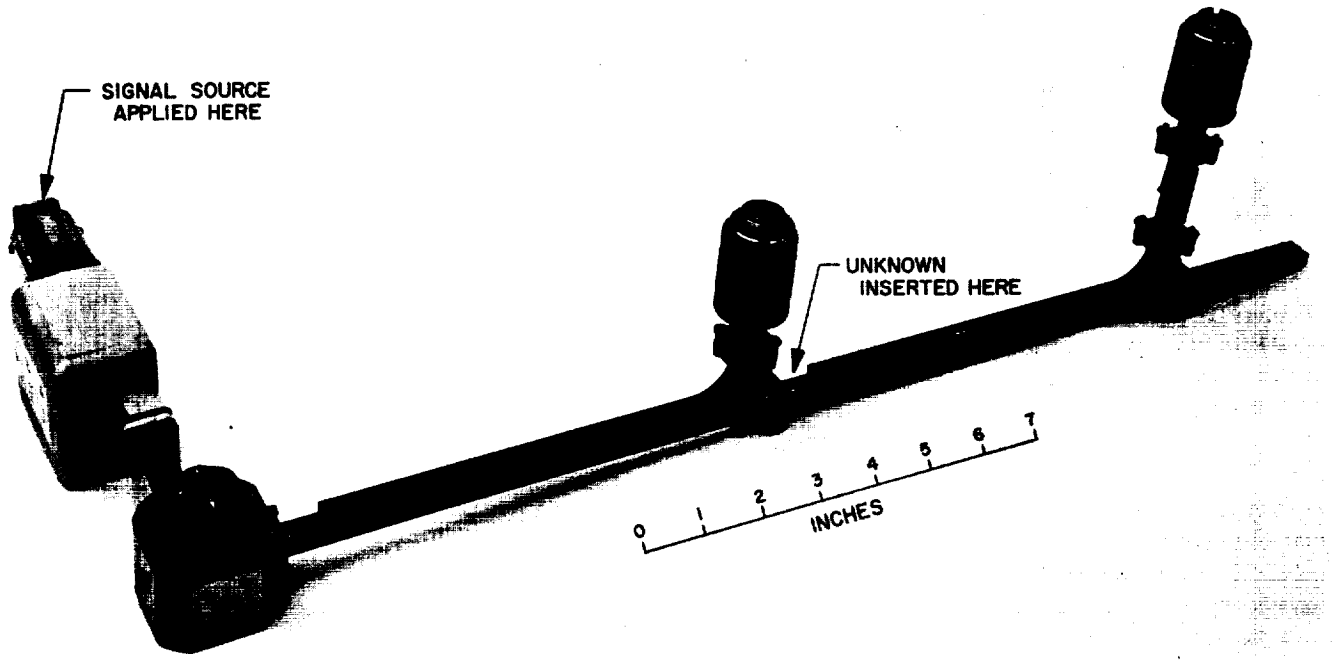


Fig. 19. K-band waveguide components for dc comparator insertion loss test set

B. Optical Communications Components

W. H. Wells

1. Lasers: Variable Q Optical Resonator Having Tilted Mirrors for Far Infrared Laser

a. Summary. We previously showed that there is a strong possibility of laser action in rotational transitions of polar molecules. To design a suitable resonator for such a laser, we have extended the theoretical work of Fox and Li (Refs. 5, 7) by studying the normal modes of an optical resonator consisting of two tilted plane mirrors. We considered larger tilt angles in order to use the spill-over light at the edge of the smaller mirror as the output coupling of the resonator. This scheme is not suitable near the visible spectrum, but is especially suited for

determines the mode shape and its threshold gain. We solved for the lowest TEM_0 modes, and found that the single-pass gain required to oscillate increases almost linearly from 3% at $\beta = 1.0 \times 10^{-2}$ to 28% at $\beta = 8.0 \times 10^{-2}$.

b. Recent work. Soon after the invention of lasers, a number of workers studied the modes of optical resonators that are formed by two mirrors facing one another (Refs. 4, 5, 6, and many others). Fox and Li (Ref. 7) described the modes of a pair of tilted mirrors, but they considered only a small amount of tilt that might occur from imperfect mirror alignment. We are designing a resonator that employs greater tilt as a means of extracting the output from a far infrared laser (*SPS 37-24*, Vol. IV, pp. 140–148; *SPS 37-25*, Vol. IV, pp. 121-124).

To understand the modes of a pair of tilted mirrors, first consider the geometric optics approximation. As shown in Fig. 20, a ray tends to "walk out" the more open side even though the initial direction of the ray was inward. The geometric approximation breaks down at the sharp edge of the mirror, and wave effects must be con-

where

$$N = \left[4 + (\Lambda_\nu^\nu)^2 - \Lambda_\nu^\mu \Lambda_\mu^\nu + i\varepsilon^{\mu\nu\kappa\lambda} \Lambda_{\mu\nu} \Lambda_{\kappa\lambda}\right]^{-\frac{1}{4}} \tag{25}$$

and the σ_j are the Pauli spin matrices. Thus, for our purposes, the problem of determining the representation of L reduces to that of finding the principal series of representations of $SL(2C)$.[2] In particular, for fixed ν and k we look for functions $F_{jm}^{k\nu}(z)$ such that

$$M_3 F_{jm}^{k\nu} = m F_{jm}^{k\nu} \tag{26}$$

$$\underset{\sim}{M}^2 F_{jm}^{k\nu} = j(j+1) F_{jm}^{k\nu} \tag{27}$$

$$F F_{jm}^{k\nu} = (\underset{\sim}{N}^2 - \underset{\sim}{M}^2) F_{jm}^{k\nu} = (1 + \nu^2 - k^2) F_{jm}^{k\nu} \tag{28}$$

$$G F_{jm}^{k\nu} = 2\underset{\sim}{M} \cdot \underset{\sim}{N} F_{jm}^{k\nu} = 2k\nu F_{jm}^{k\nu} \tag{29}$$

These functions may be found by writing the infinitesimal generators as

$$M_+ = M_1 + iM_2 = -\partial - \bar{z}^2 \bar{\partial} + (i\nu - 1 - k)\bar{z} \tag{30}$$

$$M_- = M_1 - iM_2 = z^2\partial + \bar{\partial} - (i\nu - 1 + k) z \tag{31}$$

$$M_3 = -z\partial + \bar{z}\bar{\partial} + k \tag{32}$$

$$N_+ = N_1 + iN_2 = i\left[\partial - \bar{z}^2 \bar{\partial} + (i\nu - 1 - k)\bar{z}\right] \tag{33}$$

$$N_- = N_1 - iN_2 = i\left[-z^2\partial + \bar{\partial} + (i\nu - 1 + k) z\right] \tag{34}$$

$$N_3 = i\left[z\partial + \bar{z}\bar{\partial} - i\nu + 1\right] \tag{35}$$

with

$$\partial = \frac{\partial}{\partial z} \tag{36}$$

$$\bar{\partial} = \frac{\partial}{\partial \bar{z}} \tag{37}$$

and z and $\bar{z}$ are treated as independent variables. With these generators one can easily verify that Eqs. (28) and (29) are satisfied identically. Furthermore, letting $z = re^{i\theta}$, Eqs. (26) and (27) are solved by

$$F_{jm}^{k\nu}(r, \theta) = N_{jm}^{k\nu} e^{i(k-m)\theta} r^{k-m}(1 + r^2)^{i\nu+m-1} \times P_{j+m}^{k-m,-k-m}\left(\frac{1 - r^2}{1 + r^2}\right) \tag{38}$$

where

$$N_{jm}^{k\nu} = e^{i\eta} \pi^{-\frac{1}{2}} \left[\frac{(2j+1)(j+m)!(j-m)!}{(j+k)!(j-k)!}\right]^{\frac{1}{2}} \tag{39}$$

and $\eta = \eta(j, \nu)$ is a real number chosen so that the $F_{jm}^{k\nu}$ satisfy the conditions of a canonical basis (Ref. 3). $P_n^{\alpha\beta}$ is a Jacobi polynomial defined by

$$P_n^{\alpha\beta}(x) = \frac{(-)^n}{2^n n!}(1 - x)^{-\alpha}(1 + x)^{-\beta}\frac{d^n}{dx^n} \times \left[(1 - x)^{n+\alpha}(1 + x)^{n+\beta}\right] \tag{40}$$

The orthogonality relation for the F's is

$$\int_0^{2\pi} d\theta \int_0^\infty r dr\, \overline{F_{j'm'}^{k\nu}(r, \theta)}\, F_{jm}^{k\nu}(r, \theta) = \delta_{jj'}\delta_{mm'} \tag{41}$$

To compute the matrix element Z, we note that the unitary operator $\exp(-i\varepsilon N_3)$ in Eq. (19) is given by $U(\alpha)$,

$$\alpha = \begin{pmatrix} e^{-\varepsilon/2} & 0 \\ 0 & e^{\varepsilon/2} \end{pmatrix} \tag{42}$$

in the representation given by Eq. (23). Thus, one has

$$Z_{jj'}^{k\nu m} = e^{(i\nu-1)\varepsilon} \int_0^{2\pi} d\theta \int_0^\infty r dr\, \overline{F_{jm}^{k\nu}(r, \theta)}\, F_{j'm}^{k\nu}(e^{-\varepsilon/2} r, \theta) \tag{43}$$

The evaluation of the integrals in Eq. (43) is straightforward. One obtains finally

$$Z_{jj'}^{k\nu m} = \beta(jj';k\nu m;\varepsilon) \sum_{p=0}^{j+m} \sum_{q=0}^{j'+m} C(jj';km;p,q) \times F(q + 1 - m - i\nu, k - m + 1; p + q + 2 - 2m; 1 - e^{-2\varepsilon}) \tag{44}$$

[2] This problem has been considered by Joos (Ref. 3); however, his results are incorrect.

where

$$B = \pi \overline{N_{jm}^{k\nu}} N_{j'm}^{k\nu} e^{(i\nu+m-k-1)\varepsilon} 2^{m-k} \frac{(j+k)!(j'+k)!(k-m-2)!}{(j-m)!(j'-m)!} \tag{45}$$

$$C = [(-)^{j+j'-2m-p-q}] \frac{(j-m+p)!(j'-m+q)!(p+q-m-k)!}{p!(j+m-p)!q!(j'+m-q)!(p-k-m)!(q-k-m)!(p+q-2m+1)!} \tag{46}$$

F is a hypergeometric function defined by

$$F(a,b;c;x) = \sum_{n=0}^{\infty} \frac{(a+n)!}{a!} \frac{(b+n)!}{b!} \frac{c!}{(c+n)!} x^n \tag{47}$$

Strictly speaking Eq. (44) is valid only for $m + k < 1$. However, if this condition is not satisfied, one can use certain symmetry properties of $F_{jm}^{k\nu}$ to obtain analogous formulas. A more complete discussion will be given elsewhere.

References

1. Meghreblian, R. V., and Holmes, D. K., "Reactor Analysis," McGraw-Hill, New York, 1960.
2. Corngold, N., Brookhaven National Laboratory, Report No. 445 (T-93), 1958.
3. Joos, H., *Fortschritte der Physik*, Vol. 10, p. 65, 1962.
4. Naimark, M. A., *American Mathematical Society Translations*, Ser. 2, Vol. 6, 1957.
5. Macfarlane, A. J., *Journal of Mathematical Physics*, Vol. 3, p. 1116, 1962.

XIX. Applied Science

A. A Gas Chromatograph for the Analysis of the Martian Atmosphere

W. F. Wilhite

1. Introduction

A gas chromatograph was proposed (Ref. 1) for a possible *Mariner* 1966 landing capsule mission to provide an analysis of the Martian atmosphere during descent of the capsule. The gas chromatography instrumentation proposed is not limited to the *Mariner B* 1966 mission but can be used to provide a descent analysis of the Martian atmosphere on any Mars landing mission.

A gas chromatograph consists of a column-detector system, which separates and detects the various components in a gaseous mixture and is therefore the heart of the instrument, the signal processing and associated electronics, the sampling valve for injecting a sample of the gaseous mixture into the chromatograph, and the carrier gas system. The development of the signal processing and associated electronics was described in a previous *SPS* (Ref. 2). The present article describes the development of a column-detector system which demonstrates the scientific feasibility of performing the mission described in the referenced proposal. The proposal assumed several limitations which affect the development of the column-detector system. These are:

(1) An analysis of no greater than 50 sec with a desired time of 30 sec.

(2) The major components to be analyzed were carbon dioxide, argon, and nitrogen; and the minor components were oxygen and water.

It has been decided to consider the analysis of the major components of primary importance so that engineering data to design future survivable capsules may be obtained.

2. Description of Column-Detector System

Carbon dioxide, argon, and nitrogen, although only three gaseous components, represent a rather difficult separation in gas chromatography, a separation which at the present time cannot be adequately performed on one column. It was decided to utilize a series arrangement column-detector system so that the inaccuracy of stream-splitting, necessary for parallel arrangement, would not be a problem. The column-detector system would then be arranged in the following order:

(1) A column for separating CO_2 from the composite (argon-nitrogen peak).

(2) A detector for detecting the composite and CO_2 peaks.

(3) A column for separating argon and nitrogen.

(4) A detector for detecting argon and nitrogen.

The problem of separating the composite from CO_2 was attacked first. Separation of CO_2 from an air (composite) peak is a rather fundamental separation in gas chromatography; however, to perform this separation very rapidly requires high resolution and very sharp, narrow peaks. These necessary characteristics are not provided by the fundamental separation methods. The achievement of highly resolved narrow peaks was attempted using various liquid substrates in saturated concentrations on acid-base washed and silanized diatomaceous earth supports. Some of the substrates tested included di(2-ethylhexyl)sebacate, Dow Corning DC 550 silicon oil, didecylphthalate, and dioctylphthalate. These substrates were tested with column loadings ranging from 30 to 35%. The results were not desirable due to incomplete resolution of the composite and CO_2 peak.

Resolution of the composite and CO_2 peak can be readily achieved using an activated silica gel; however, the CO_2 peak becomes extremely broad and therefore rapid analysis is difficult. It was decided to try an unactivated silica gel packing in order to achieve a narrow CO_2 peak without losing the desired resolution. This proved to be successful, but over a period of time the silica gel column would slowly activate and the CO_2 peak would broaden, slowing the analysis time. Success was finally achieved by activating the silica gel in the normal way and then deactivating the silica gel with a low loading of a polar substrate. The best results were achieved with 0.1% diglycerol on silica gel that had been activated at 250°C for 1 hr. A further refinement in this column development for reducing the required time of resolving the composite-CO_2 peaks was the utilization of very short columns of the above prepared packing material utilizing a silica gel particle size of 160 to 170 mesh. With these characteristics, a column length of only 3 in. was required for excellent separation and peak shape of the composite and CO_2.

The next area of development was a column to separate argon and nitrogen. This separation was achieved using molecular sieve 5A activated for 1 hr at 400°C. As in the case with the silica gel column, in order to provide as rapid as possible an analysis, a 160- to 170-mesh particle size molecular sieve 5A was used. A column of only 9 in. in length of this material was required to separate argon and nitrogen.

The detectors utilized in the system were provided by Dr. J. E. Lovelock, an advisor in gas chromatography to JPL. The detector is a dual-cell cross-section ionization detector, the dimensions and materials of construction of which are shown in Fig. 1. The details of the performance characteristics of this detector were given by Dr. Lovelock in a recent paper (Ref. 3). The outlet of a gas chromatograph column containing carrier gas and the sample is fed into one of the cells, and carrier gas only is fed into the other cell. The output signal, since one electrode of both cells is common, is the differential equivalent to the presence of the sample components. This mode of detector use cancels effects caused by fluctuating pressure and temperature.

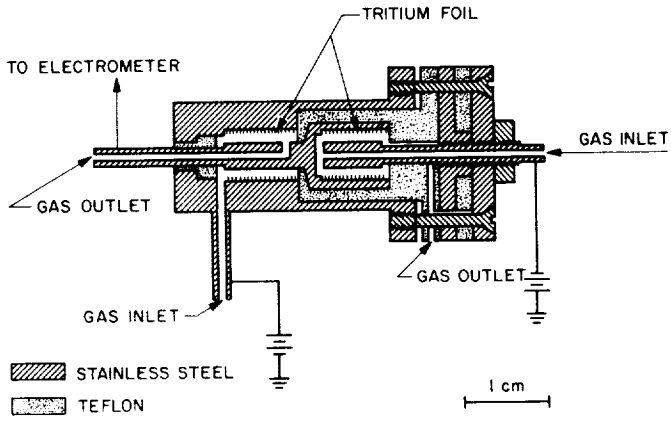


Fig. 1. Cross-sectional view of Lovelock dual-cell cross-section detector

The detector that will finally be used in this particular system is one that was developed by the JPL Space Instrument Development Section using Dr. Lovelock's detector as a design guide. A comparison of the two detectors is shown in Fig. 2. The performance of the detector developed by the JPL Space Instrument Development Section will be described in a future *SPS* article. The internal volume of the newer detector is only 60 microliters as compared with 250 microliters in Dr. Lovelock's dual-cell detector. This reduction in volume improved the performance characteristics of the column detector system significantly, as will be discussed later.

The system was assembled as shown in Fig. 3. The order of the components were as shown: the silica gel column, first detector, molecular sieve column, and second detector. Dummy columns approximating the pressure drop through the actual columns were placed in series with the reference side of each detector so that the first detector operated with the same pressure on both cells. This was found to be fairly critical because of the sensitivity to pressure fluctuations that the cross-section detector exhibits.

The 15-ft section of 0.050-in.-ID empty tubing was necessary between the first detector and the second column in order to delay the detection of the argon peak by the second detector until the CO_2 peak has been detected by the first detector.

The resolution achieved with the Lovelock dual-cell cross-section detector in the system is shown by chromatogram in Fig. 4a. The operating parameters for this chromatogram are as follows:

Inlet pressure	70 psi
Flow rate	135 cm^3/min
Analysis time	Approximately 30 sec
Components	Composite, carbon dioxide, argon, and nitrogen
Operating temperature	Room temperature

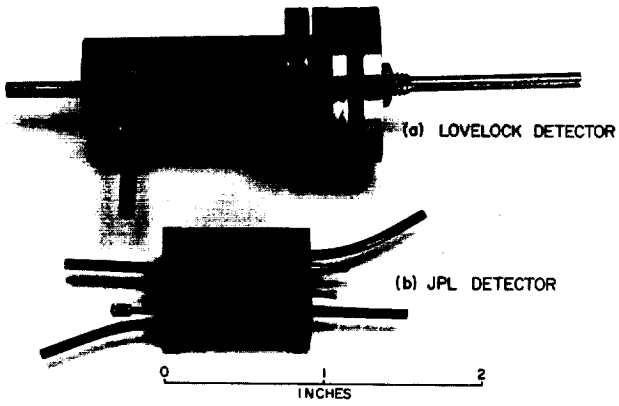


Fig. 2. Comparison of Lovelock detector with JPL detector

By using the detectors developed by the JPL Space Instrument Development Section with one-fourth the volume of the Lovelock detector, the chromatogram in Fig. 4b was obtained. Some of the system operating parameters were changed because of the superior resolu-

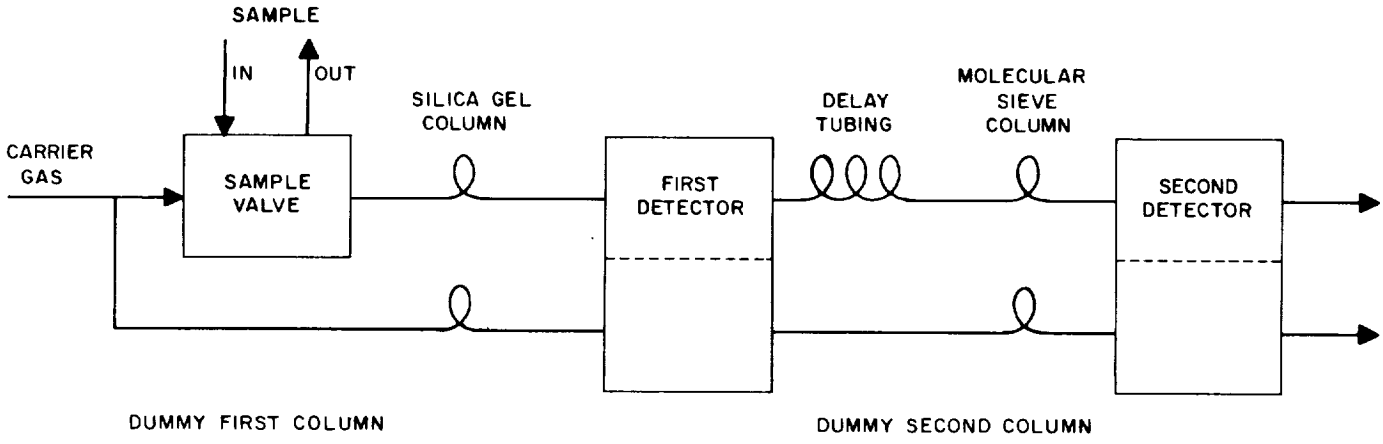


Fig. 3. Schematic of column-detector system

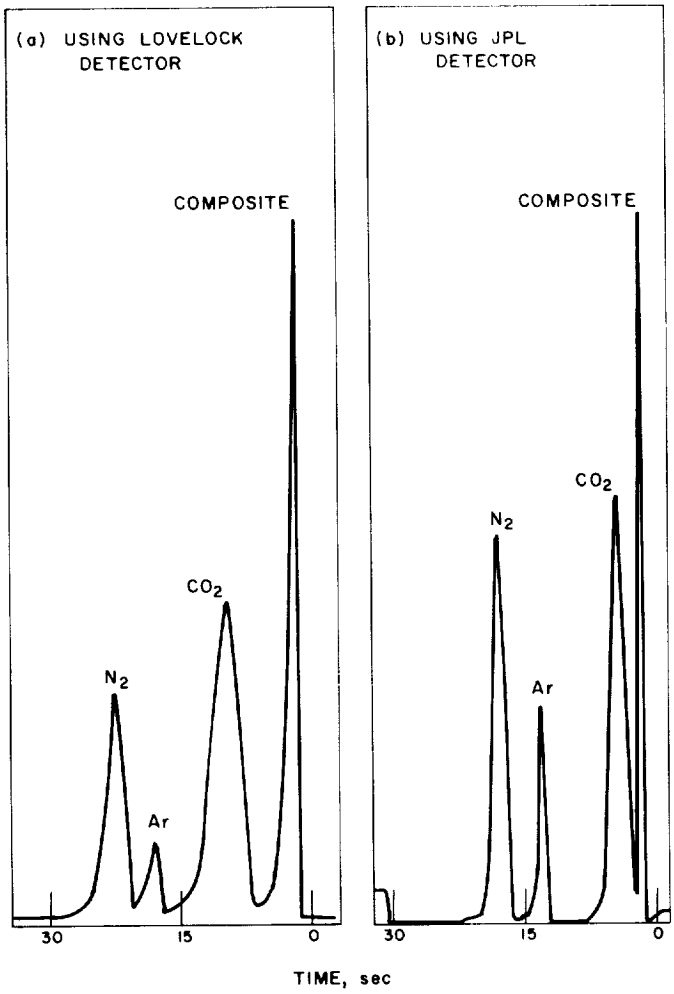


Fig. 4. Chromatograms from (a) Lovelock detector and (b) JPL detector

tion that these detectors afford. Also, the delay tubing between the first detector and second column was reduced from 15 to 8 ft in length. The operating parameters for this system are as follows:

Inlet pressure	70 psi
Flow rate	98 cm^3/min
Analysis time	20 sec
Components	Composite, carbon dioxide, argon, and nitrogen
Operating temperature	Room temperature

The improvement in performance between the two systems is attributable to less band spreading of the sample components as they enter the detector in the smaller detector. The latter system is now being mated to the signal processing and associated electronics. The overall performance of this gas chromatograph system will be studied extensively in the next few months.

Development is now proceeding on column-detector systems which should provide analysis in one-tenth the time of the system described in this report with a flow rate of carrier gas one-hundredth of the amount needed for the present system. Resolution and sensitivity will also be greatly improved with the newer system. The new column-detector will weigh far less than the present system.

References

1. Lipsky, S. R., Bentley, K., Bowman, L., Josias, C., Mertz, H., Wilhite, F., "*Mariner B* 1966 Mission: Analysis of the Atmosphere of Mars by Gas Chromatography Instrumentation," proposal submitted to JPL Space Sciences Division by S. R. Lipsky, Yale University Medical School, New Haven, Conn., November 15, 1963.

2. Josias, C., Bowman, L., Mertz, H., "A Gas Chromatograph for the Analysis of the Martian Atmosphere," *SPS 37-27*, Vol. IV, pp. 85–91, Jet Propulsion Laboratory, Pasadena, California, June 30, 1964.

3. Lovelock, J. E., Shoemake, G. R., Zlatkis, A., "Improved Ionization Cross-Section Detectors," *Analytical Chemistry*, Vol. 36, p. 1410, 1964.

TELECOMMUNICATIONS DIVISION

XX. Communications Elements Research

A. Low-Noise Amplifiers

C. T. Stelzried, W. V. T. Rusch, R. Brantner, and S. Petty

1. 90-Gc Millimeter Wave Work,

C. T. Stelzried and W. V. T. Rusch

a. Introduction. The objective of millimeter wave work is to investigate millimeter wave components and techniques for application to antenna and microwave engineering scale model work applicable to conventional DSIF frequencies, and to ascertain the future applicability of this frequency range to space communications and tracking. This involves the development of instrumentation for accurate determination of insertion loss, VSWR, power and equivalent noise temperature of passive elements, and gain and bandwidth of active elements at millimeter wavelengths.

In order to bring together state-of-the-art millimeter wave circuit elements and evaluate their use in a simple system, a radio telescope consisting of a 60-in. antenna and a superheterodyne radiometer has been built. The radio telescope was used to observe the 90-Gc temperature of the Moon during the December 30, 1963 eclipse (*SPS 37-25*, Vol. IV, pp. 113–117 and *SPS 37-26*, Vol. IV, pp. 181–189). This experiment was a joint effort by personnel from JPL and the Electrical Engineering Department of the University of Southern California.

b. Recent work. The 90-Gc radiometer is currently undergoing improvements and modification (Fig. 1). The 30-Mc IF amplifier is being replaced with a wide-band (150-Mc bandwidth) transistor commercial amplifier (Hewlett-Packard Model 461A). The noise figure of the standard amplifier is about 12 db at 30 Mc fed from a 50-Ω source.

Removing the 50-Ω input matching network lowers the noise figure at 30 Mc to 6 db from a 100-Ω source. A wide-band transformer (North Hills Model 1501) is being modified and adapted to match the push–pull output from the Raytheon mixer to the single-ended amplifier input. The following analysis is given to estimate the effect of the change in the IF amplifier on

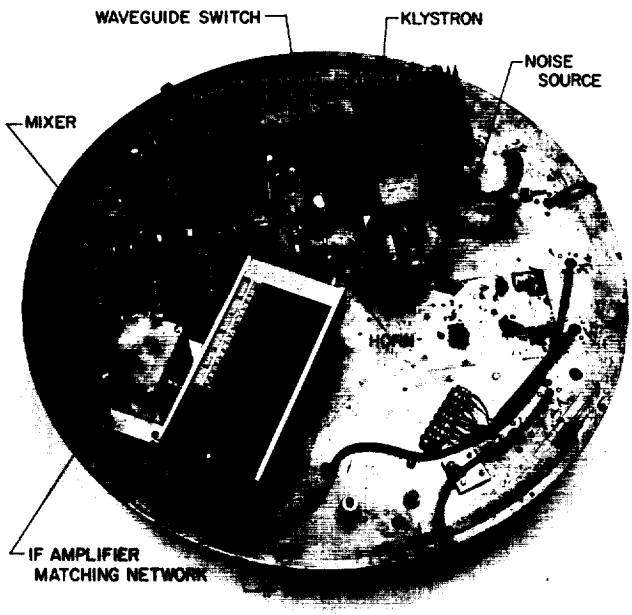


Fig. 1. Components of 90-Gc radiometer

system performance. The noise figure F of the mixer and IF amplifier combination shown in Fig. 2 is (Ref. 1):

$$F = \left[\left(\frac{S}{N}\right)\text{input} \Big/ \left(\frac{S}{N}\right)\text{output}\right]^{\text{with source at temperature } T_0 = 290^\circ\text{K}} \quad (1)$$

where

$\frac{S}{N}$ input = signal-to-noise power ratio at the amplifier output

$\frac{S}{N}$ output = signal-to-noise power ratio at the amplifier input

Define:

L_M = mixer conversion loss ratio

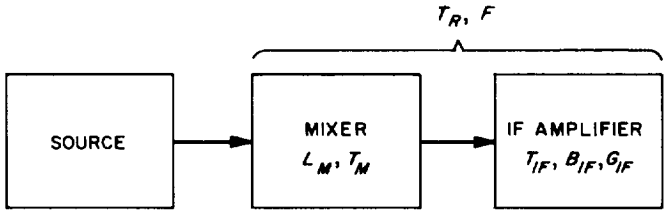


Fig. 2. Block diagram of mixer and IF amplifier used in noise temperature derivation

T_M = mixer equivalent output noise temperature at the IF amplifier frequency

T_{IF} = equivalent input noise temperature of IF amplifier

G_{IF} = available power gain of IF amplifier at midband

G_f = gain at frequency f of IF amplifier

$B_{IF} = \frac{1}{G_{IF}}\int G_f df$ = bandwidth of IF amplifier

k = Boltzmann's constant = 1.38×10^{-23} joules/°K

F = noise figure ratio of mixer and IF amplifier combination

T_R = equivalent noise temperature of mixer and IF amplifier combination

With a relatively narrow-band IF amplifier and the signal frequency $>>$IF frequency the conversion loss of the mixer is equal at the signal and image frequencies. For a radiometer with an input signal expressed as a temperature change ΔT_S, the available input and output signal and noise powers are:

$$\begin{aligned} S_{input} &= 2k(\Delta T_S)B_{IF} \\ S_{output} &= 2k(\Delta T_S)B_{IF}\frac{G_{IF}}{L_M} \\ N_{input} &= 2kT_0B_{IF} \\ N_{output} &= \frac{2kT_0B_{IF}G_{IF}}{L_M} + (T_M + T_{IF})G_{IF}kB_{IF} \end{aligned} \quad (2)$$

The factor of 2 appears due to the mixer response at both the image and signal frequencies.

Substituting into Eq. (1) and reducing,

$$F = 1 + \frac{L_M}{2T_0}(T_M + T_{IF}) \quad (3)$$

Converting to equivalent noise temperature,

$$T_R = \frac{L_M}{2}(T_M + T_{IF}) \quad (4)$$

The equivalent noise temperature of the millimeter radiometer using the Raytheon balanced mixer and 30-Mc JPL IF amplifier combination has been approximately 20,000°K. The conversion loss and noise temper-

ature of the Raytheon diodes are not known. Assuming $T_{IF} \simeq 300°K$ and $T_R = 20{,}000°K$, then

$$L_M \simeq \frac{40{,}000}{T_M + 300} \tag{5}$$

If the reasonable assumption is made that T_M lies between 100 to 500°K, then L_M is between 100 and 50. With this assumption there can be determined the effect on system performance of going from an IF amplifier with 10-Mc bandwidth and 300°K equivalent noise temperature to an amplifier with 150-Mc bandwidth and 1000°K ($F \approx 6$ db) equivalent noise temperature. For the wideband IF amplifier T_R is between:

$$T_{R_1} \simeq \frac{100}{2}(100 + 1000) = 55{,}000°K$$

and

$$T_{R_2} = \frac{50}{2}(500 + 1000) = 75{,}000°K$$

The thermal jitter ΔT of a Dicke radiometer with a perfect integrator is given by (Ref. 2)

$$\Delta T \approx \frac{2T_S}{(\tau B_{IF})^{1/2}} \tag{6}$$

where

$T_S = T_R + T_A$ = total system equivalent noise temperature

T_A = equivalent noise temperature of antenna

τ = integrating time

Since $T_R >> T_A$,

$$\Delta T \simeq \frac{2T_R}{(\tau B_{IF})^{1/2}} \tag{7}$$

The thermal jitter ΔT for the wide-band IF amplifier is between 2.9 to 3.9°K for a 10-sec integrating time. For the original 10-Mc IF amplifier the thermal jitter was 4.0°K. This estimate indicates that the system sensitivity with the wide-band IF amplifier will be about the same as with the narrow-band amplifier. If a wide-band low-noise front end for the IF amplifier can be obtained, then the performance can be substantially improved. If a low-noise wide-band IF amplifier of 300°K equivalent noise temperature were obtained, we have, using the same assumptions for T_M, a range for T_R of 20,000 to 40,000°K. The thermal jitter would then be between 1 to 2°K.

A significant advantage of the wide-band IF amplifier should be found in taking RF antenna patterns. Frequency instability in the RF transmitter and receiver local oscillator has caused the signal to wander out of the passband of the IF amplifier. Raising the IF frequency from 10 to 150 Mc will reduce this effect.

RG-99/u waveguide calibrated thermal terminations are under construction (Fig. 3). These terminations have improvements over previously constructed units. The waveguide stainless steel sections have been gold plated with an electroless plating technique (*SPS 37-28*, Vol. IV, pp. 147–150) to lower the insertion loss and increase the calibration accuracy. The heat sink surrounding the termination is anodized aluminum. Anodizing provides electrical insulation for the nichrome heater wires. These terminations will be controlled at 50 and 100°C with mercury thermostats.

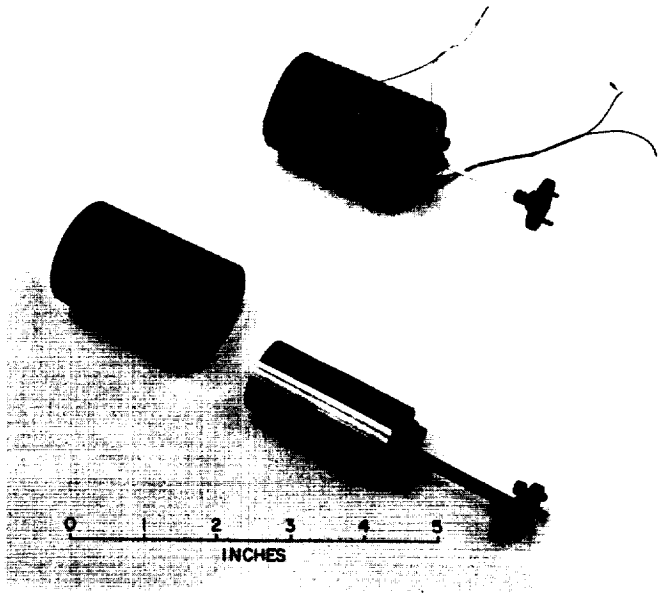


Fig. 3. RG-99/u waveguide calibrated thermal termination assemblies

A new 90-Gc transmitter facility has been developed as an aid to making antenna patterns and general system evaluation. A large heat sink for the klystron will enable operation at a temperature low enough to ensure stable operation of the transmitter without an air blower. Fig. 4 is a block diagram of the transmitter, and Fig. 5 shows the transmitter along with the associated power

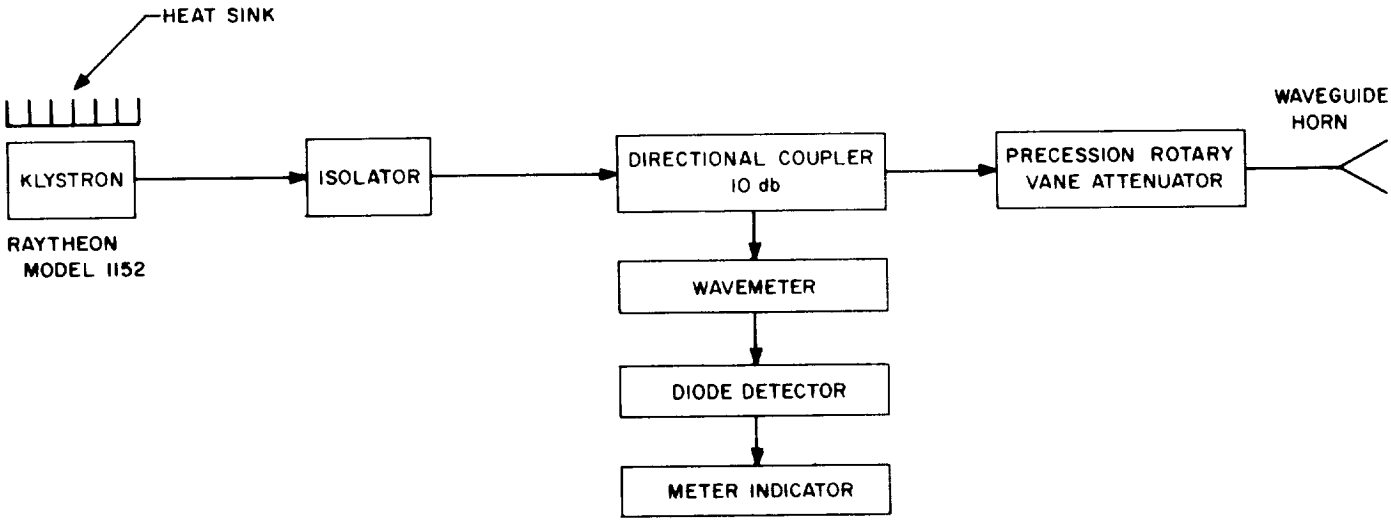


Fig. 4. Block diagram of 90-Gc transmitter

Fig. 5. The 90-Gc transmitter waveguide components and electronics

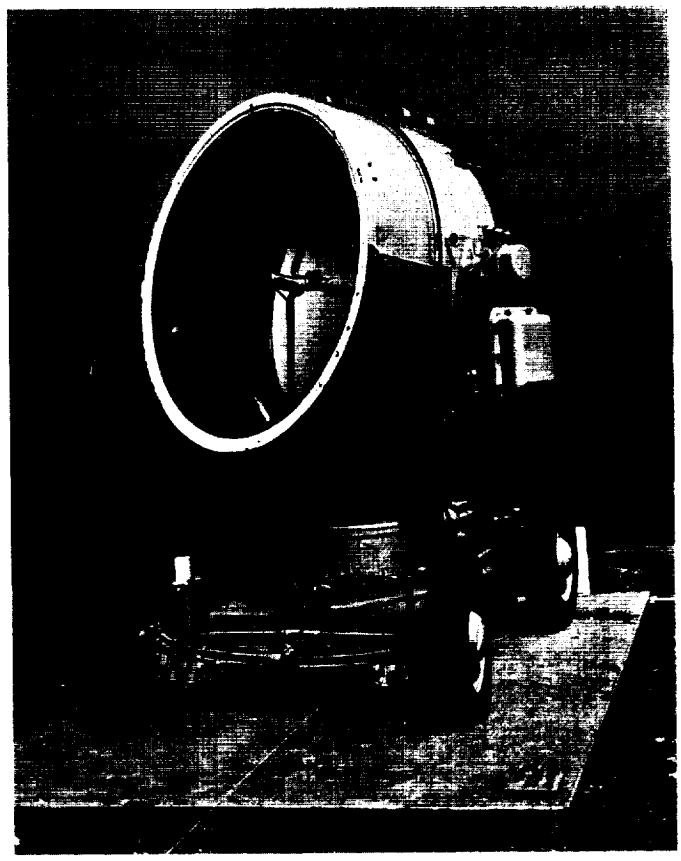


Fig. 6. The 90-Gc searchlight antenna on mounting pad

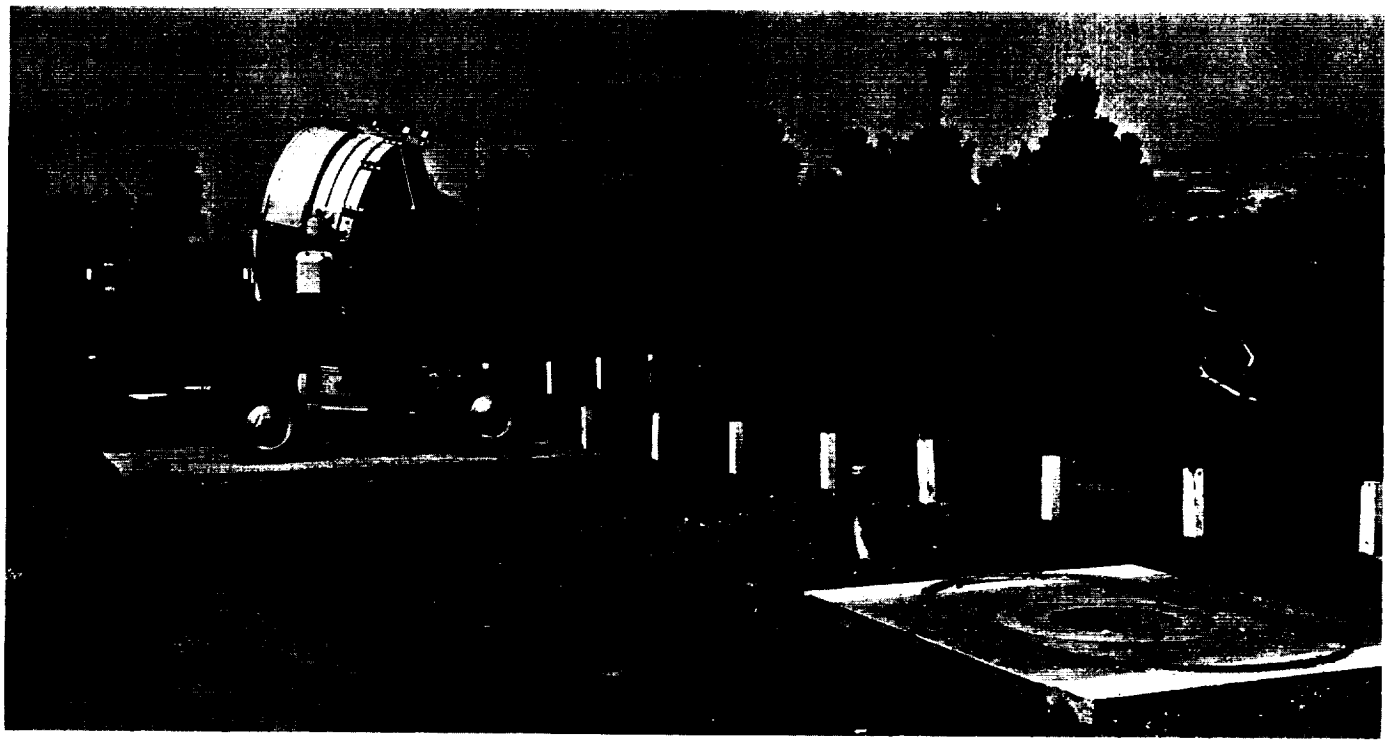


Fig. 7. The 90-Gc searchlight antenna and Contraves phototheodolite pad

supplies and auxiliary equipment. By combining both the electronic and waveguide components in a single rack, a portable facility which can be set up easily under varying conditions of terrain and antenna location has been developed.

A concrete mounting pad has been fabricated on the Mesa Antenna Range where the 90-Gc antenna will be located for field experiments; it is adjacent to a Contraves phototheodolite installation (Figs. 6 and 7). The phototheodolite will be used to support experiments with the 90-Gc radiometer as well as possible future optical frequency component experiments.

2. *Microwave Radiometer Bias Supplies,* R. Brantner

In microwave radiometry and related space communication system instrumentation, strip chart recorders are used in both laboratory tests and in field operations. In these operations, dc bias sources are used in conjunction with the recorders. A typical bias source consists of a mercury cell supplying current to a resistor network. A disadvantage of this system is the possibility of the mercury cell failing during a test or field operation, and because the cell is out of sight, it is easy to forget to replace it periodically. For these reasons an ac-powered bias source was constructed and compared to a battery-powered bias source similar to those now in use. Schematic diagrams of the test bias sources are shown in Fig. 8. In practice, the 100-Ω resistors, R_6, are replaced by helipots or Fluke decade potentiometers. The output voltage is taken from the common to the potentiometer tap so as to be variable from 0 to 100 mv. In the tests, the current through D_6 was set at 7.50 ma, and the output voltage was set at 100.0 mv using a digital voltmeter. Results of tests in an environmental test oven were as follows:

Temperature, °C	Output voltage, mv	
	Ac-powered	Battery-powered
−10	100.0	100.5
+25	100.0	100.0
+60	100.2	100.3

A precaution to be observed in assembling the ac-powered bias source is that R_5 and R_6 should have similar temperature coefficients of resistance. It is expected that ac-powered bias sources will be used in future radiometry operations.

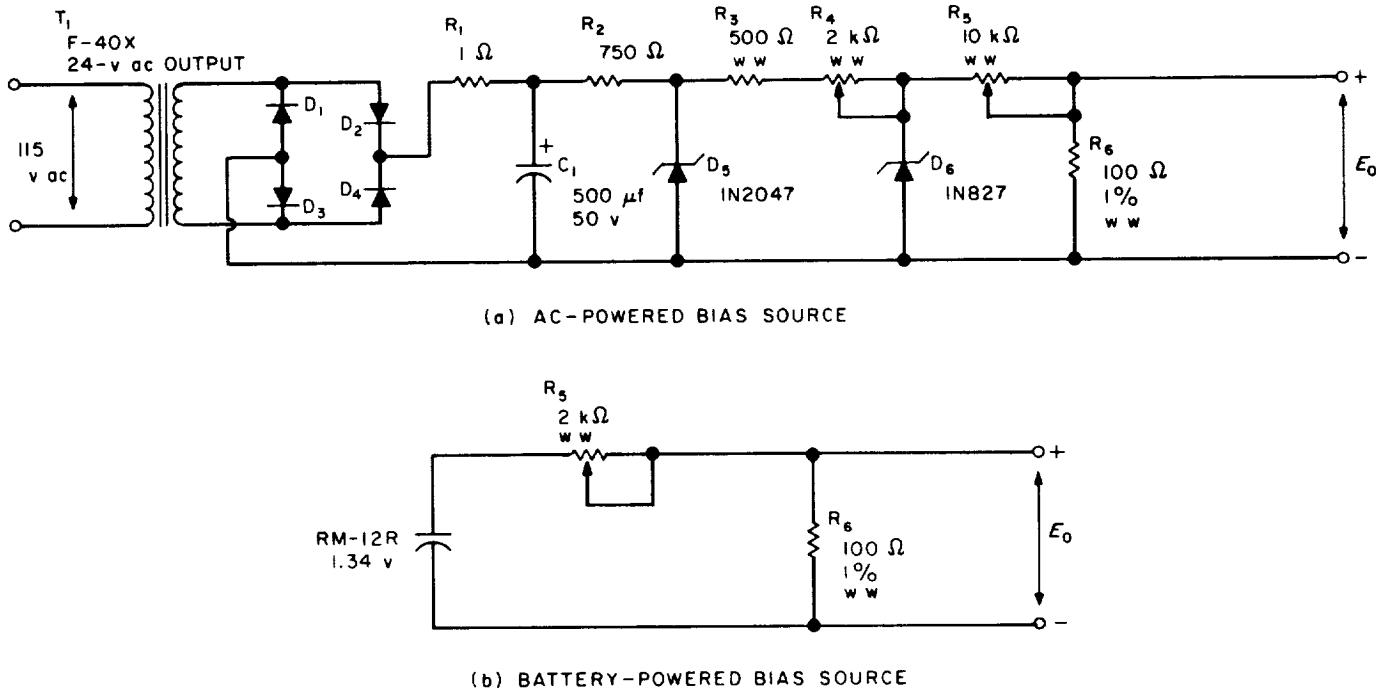


Fig. 8. Ac-powered and battery-powered bias sources

3. Solid State Circuits, R. Brantner

a. Summary. An intermediate frequency amplifier using field effect transistors (FET) is being studied for possible application as a low-noise amplifier where space and power requirements preclude use of vacuum tubes. Experimental models of a 10-Mc amplifier have been built and tested.

b. Recent work. A series of tests was made to determine the necessity of neutralizing the various stages of the IF amplifier. Results indicated that, with all stages operating in the common source mode, neutralization of the input stage was essential for stability, but that all remaining stages could be unneutralized with little or no degradation of performance. With the amplifier operating in this manner, various methods of matching the signal source to the first stage were tried. Most successful was an *L*-network, that gave a bandwidth of about 2.1 Mc and a NF of 5.2 db, using a 200- to 300-Ω source. For all the above tests, type TA-2330 FET's were used exclusively.

Further tests using FET types TA-2330 (RCA) and FE-400 (Amelco) resulted in an amplifier consisting of a common-source input stage and a common-gate second stage, both using FE-400's, a common-source third stage using a TA-2330, and an output impedance-matching stage. None of the stages was neutralized. With this configuration, power gain = 38 to 39 db, bandwidth = 3.3 Mc, and NF = 2.8 to 2.9 db, using a source impedance of 200 to 400 Ω. A lower noise and higher gain version of the FE-400, the type U-1166, is now available from Amelco. Experiments will continue, using these and other newly-released types of high-transconductance FET's. Work is also proceeding on modifications of the output impedance-matching stage. At present, this stage is quite stable but rather complex. Various modifications are being investigated in an attempt to simplify the stage without degrading its stability.

4. Microwave Noise Source, C. T. Stelzried and S. Petty

a. Introduction. Primary requirements on a gas tube noise source for calibrating a microwave receiving system are: low insertion loss, good short- and long-term stability of the injected noise, long life, and ease of operation. Present techniques used at JPL utilize a noise tube on the secondary arm of a directional coupler to inject a noise pulse of the desired amplitude. The excess noise added to the system by the unfired source adds a substantial amount of noise (from 0.75 to 3°K, typically); to keep the added noise from seriously degrading the per-

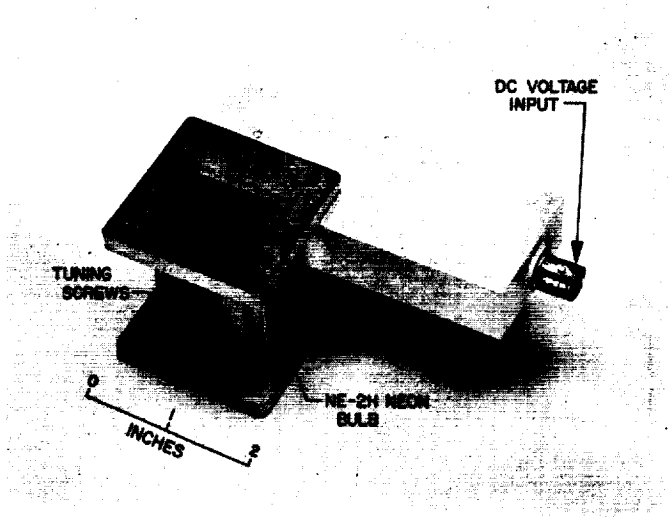


Fig. 9. H-band (RG-51/u) experimental microwave noise source

formance of very low noise systems, very large decoupling factors, up to 26 db are used. The net result is a noise pulse which is sometimes too small for proper calibration. Short-term instability of the excess noise from commercial microwave noise sources has been measured (*SPS 37-18*, Vol. IV, pp. 188–191). Short-term instability can be traced to changes in environmental temperature, mechanical vibration, and voltage supply. Electrical and gas impurity and previous history such as electrical overloads contribute to long-term instability. In an operational configuration, long cable runs between the noise source and its power supply are sometimes necessary. As a result, difficulty has been experienced with the high voltage pulsing required to excite the noise source. Some noise sources have had a very short life under operational conditions. A novel approach to solve some of these problems is described here.

b. Recent work. A commercial neon bulb Type NE-2H has been used to fabricate an experimental H-band waveguide (RG-51/u) noise source (*SPS 37-18*, Vol. IV, pp. 188–191). This tube with its close electrode spacing does not require a high voltage firing pulse. The tube is inserted directly in the waveguide (Figs. 9 and 10) so that a directional coupler is not required. The two electrodes extend approximately 0.010 and 0.014 in., respectively, below the waveguide surface. Tuning screws are used to provide an impedance match VSWR of 1.02 at 8448 Mc, the operating frequency. The series dropping carbon resistors with their associated stray capacity are used to provide RF shielding of the DC voltage lead to

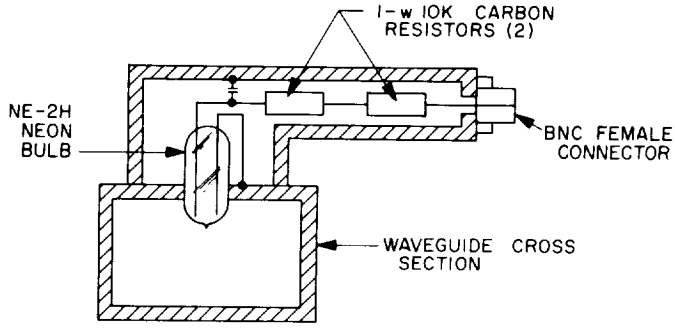


Fig. 10. Experimental RG-51/u waveguide neon bulb noise source installation

eliminate pickup at the operating frequency. The insertion loss of the noise source guide without the bulb installed was 0.005 db. With the bulb installed the insertion loss is 0.016 db so that the increase in insertion loss due to the bulb is 0.011 db. This adds about 0.8°K to the system temperature. The increase in insertion loss when the tube is fired as a function of tube current is shown in Fig. 11 and in *SPS 37-27*, Vol. IV, pp. 150, 151.

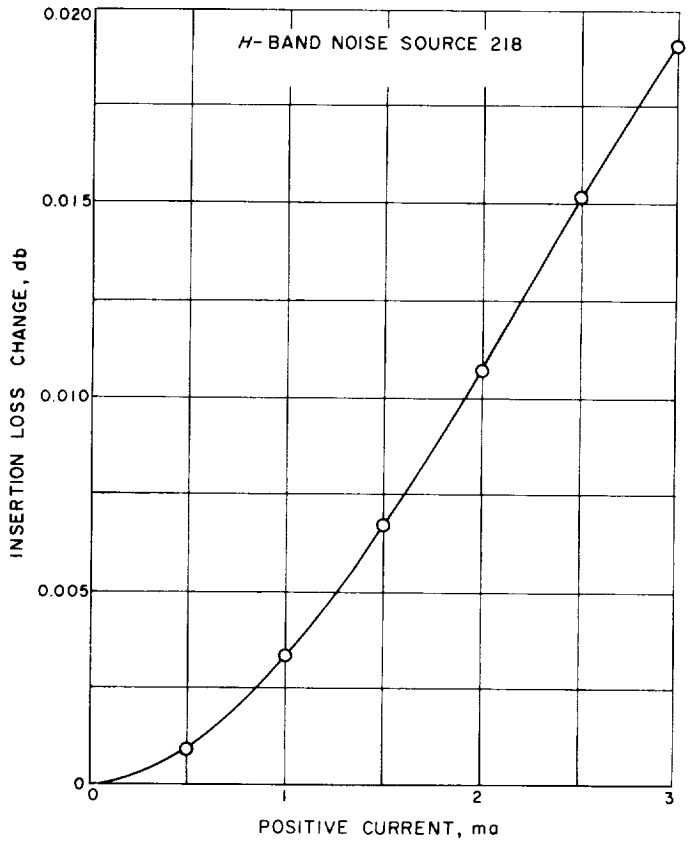


Fig. 11. RG-51/u waveguide noise source insertion loss change with tube fired versus tube current

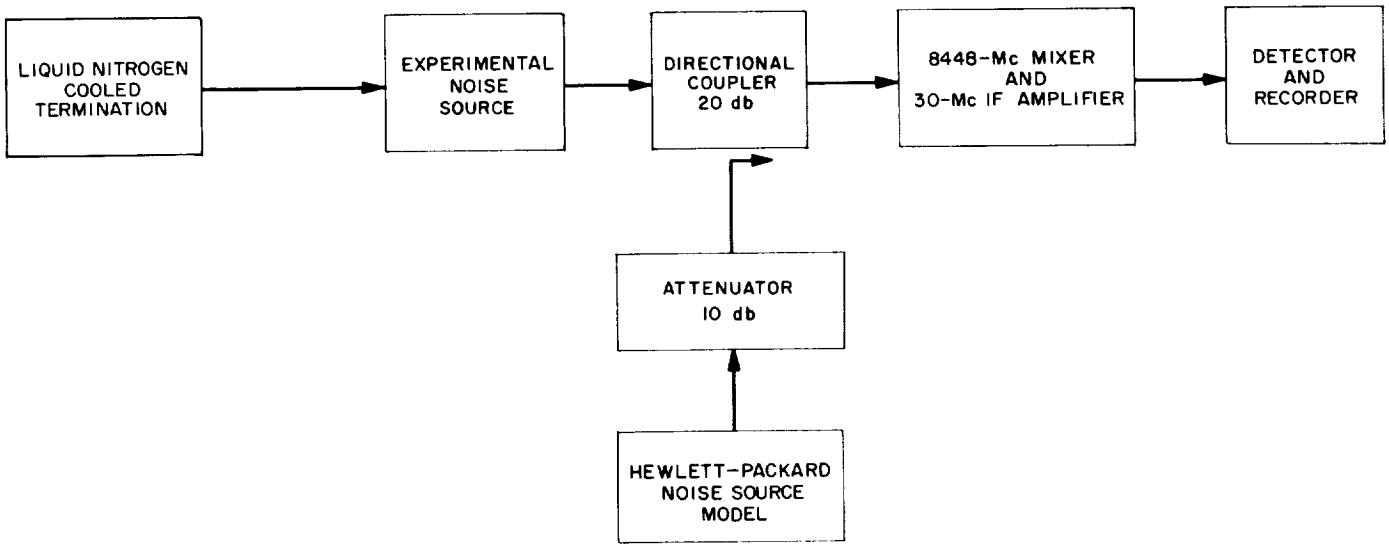


Fig. 12. Block diagram of 8448-Mc total power radiometer used for evaluation of experimental noise source

The excess noise of the experimental noise source was evaluated with a "total power" radiometer (Fig. 12) using the H-band maser monitor receiver (Ref. 3). The liquid nitrogen cooled input termination lowers the system temperature over that obtained with an ambient termination increasing the measurement accuracy. Fig. 13 shows the radiometer recording for the noise source excess noise calibration versus positive tube current. The operating parameters are:

System temperature	650°K
Bandwidth	1 Mc
Time constant	1 sec
Noise source temperature	24.6°K

One unit on the relative power scale is approximately 7.4°K as obtained from calibration with the Hewlett-Packard noise source. The excess noise versus tube current is shown in Fig. 14 for positive, negative and ac current. The electrode excited by the positive current

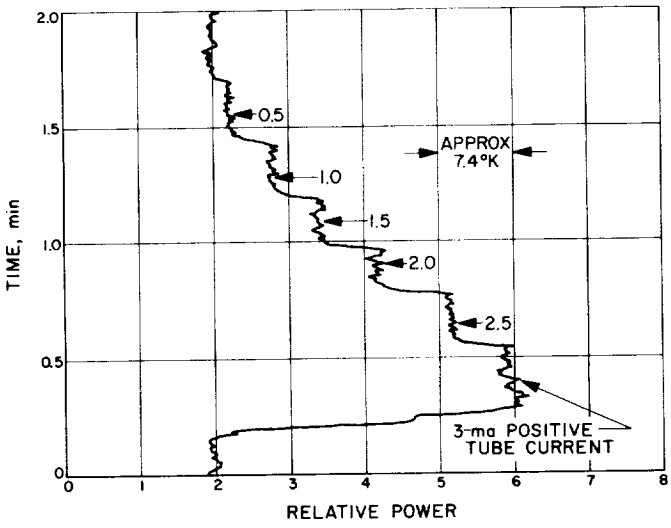


Fig. 13. 8448-Mc total power radiometer recording of experimental noise source calibration

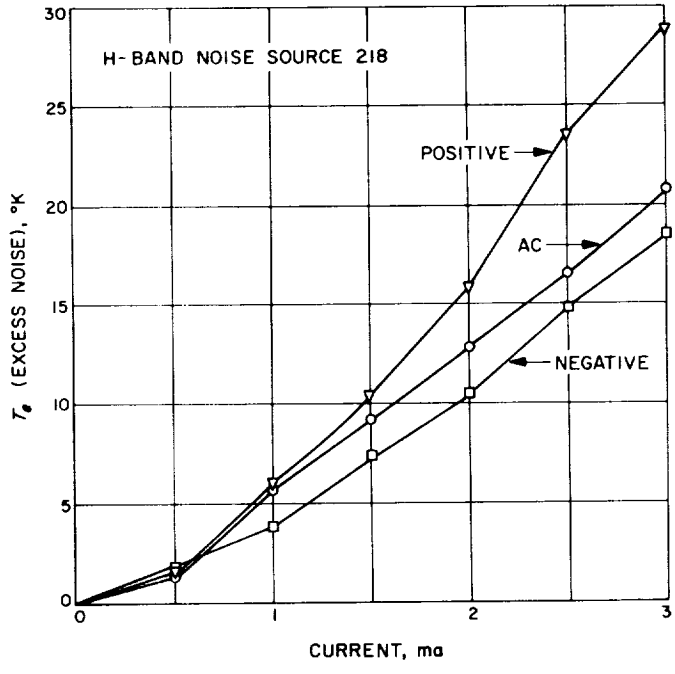


Fig. 14. Experimental noise source excess noise versus tube current for positive, negative and ac current

projects approximately 0.004 in. lower into the waveguide resulting in a higher excess noise. With 3-ma tube current, 1% current change results in approximately 1% excess noise change.

Two stainless steel waveguide sections were fabricated and used to thermally insulate the noise source. Fig. 15 shows the excess noise for various positive currents versus ambient temperature. With 3-ma tube current the excess noise sensitivity to ambient temperature is approximately 0.15°K/°C.

The insertion loss L through a gas relates the equivalent noise temperature T_g of the gas to the equivalent excess noise temperature T_e by

$$T_e = T_g (1 - L) \simeq T_g \frac{L\,(\text{db})}{4.343} \qquad (1)$$

With 3-ma tube current the insertion loss and excess noise were measured to be 0.0191 db and 28.7°K, respectively. This indicates an approximate noise temperature for the gas in the NE-2H tube with 3-ma tube current of

$$T_g \simeq 28.7 \frac{4.343}{0.0191} = 6520°\text{K}$$

Eq. (1) can be used to calculate the excess noise from the measured gas temperature and the insertion change with the tube fired. Fig. 16 shows the measured and calculated excess noise versus positive tube current. The small difference is within measurement accuracies.

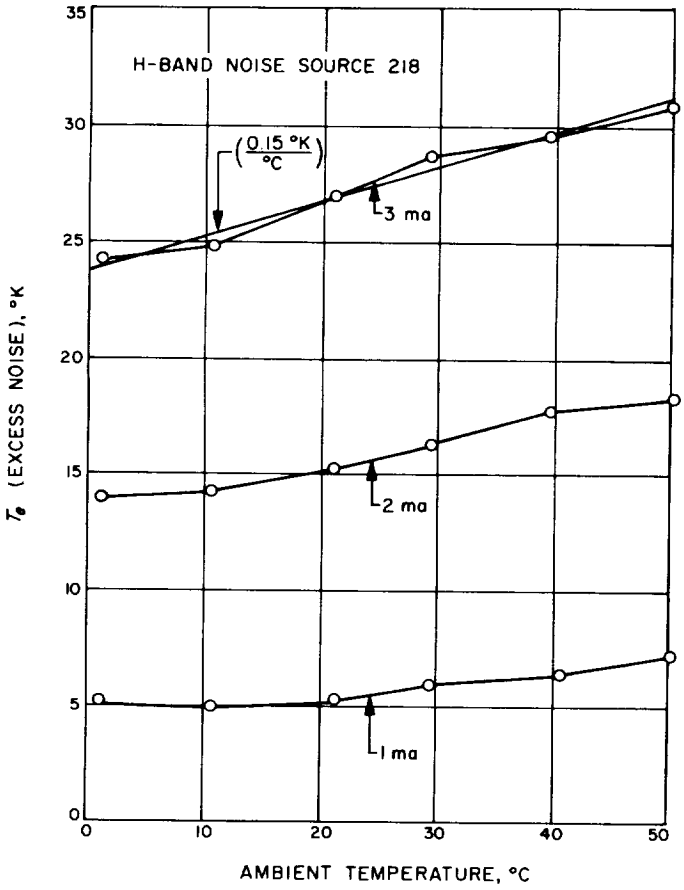


Fig. 15. Experimental noise source versus ambient temperature for various tube currents

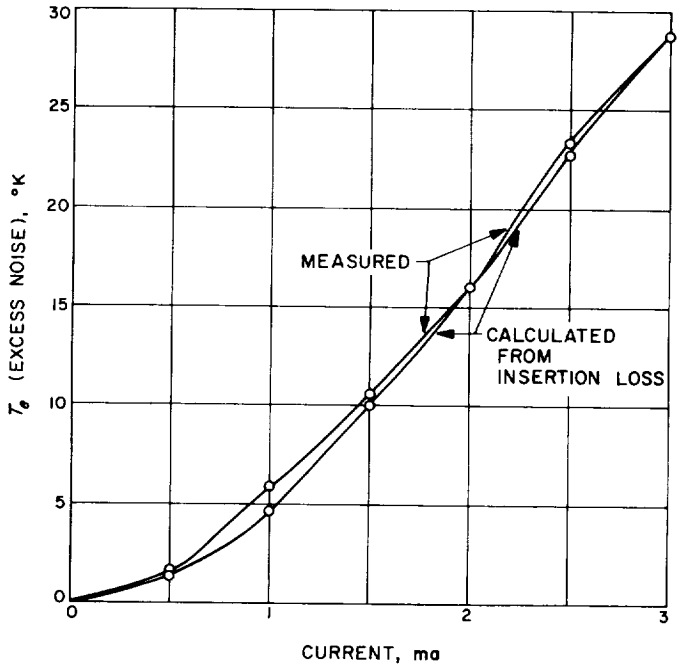


Fig. 16. Experimental noise source measured and calculated excess noise versus tube current

5. The 13.5-mm Thermal Termination, C. T. Stelzried

a. Introduction. A gas tube noise source was used for daily noise calibrations of a 13.5-mm radiometer installed on the 30-ft precision antenna at the Goldstone Venus Site during the period June 29 to July 30, 1964. The experiment in cooperation with the Lunar and Planetary Sciences Section of the Space Science Division of JPL was performed to: (1) experimentally determine the very short wavelength performance of the 30-ft antenna for comparison with calculated performance, and (2) measure the water content in the atmosphere of Venus with radio astronomy techniques during the June Venus conjunction. The measurements were performed at discrete frequencies, spanning the water vapor line: 20.6, 21.0,

21.5, 21.9, 22.2, 22.5, 23.0, 23.5, and 24.0 Gc. Techniques were required to accurately calibrate the gas tube noise source at these frequencies.

b. Recent work. Calibrated thermal terminations have been constructed for use in calibrating the radiometer noise source used with the 13.5-mm radiometer. The excess noise coupled into the radiometer from the noise source was approximately 15°K. The difference in temperature between a liquid nitrogen cooled and an ambient termination is approximately 210°K. For best accuracy in comparison calibration techniques it is desirable to have a difference temperature between the calibration terminations of approximately the same magnitude as the temperature to be calibrated. Calibration errors are due principally to termination calibration accuracy and the effect of nonlinearity and resolution in the radiometer system.

The two termination temperatures chosen were ambient and 50°C. The "ambient" load was obtained with a commercial K-band termination thermally insulated with a section of stainless steel transmission line. The "hot" load was specially fabricated and temperature controlled (Figs. 17 and 18). A copper block surrounds and heat sinks the termination. The termination is insulated with foam plastic and gold plated stainless steel transmission line. The mercury thermostat[1] has a rated absolute accuracy and control resolution of 0.05°C. This was found to be consistent with the measured thermocouple voltage. The microwave portion (Fig. 19) of the dc comparator microwave insertion loss test set (Ref. 3) was assembled and used to measure accurately the loss of the outer copper and inner stainless steel waveguide sections. The loss L_1 of the plated steel waveguide section was 0.0537 db. The loss L_2 of the input copper waveguide section was 0.0411 db. These losses were measured at a frequency f_0 of 21.9 Gc. The calibrated difference temperature $(T' - T_0)$ between the ambient and hot load at a frequency f is

[1] Philadelphia Scientific Glass Co., Inc., 9th and Ridge Ave., Perkasie, Penn.

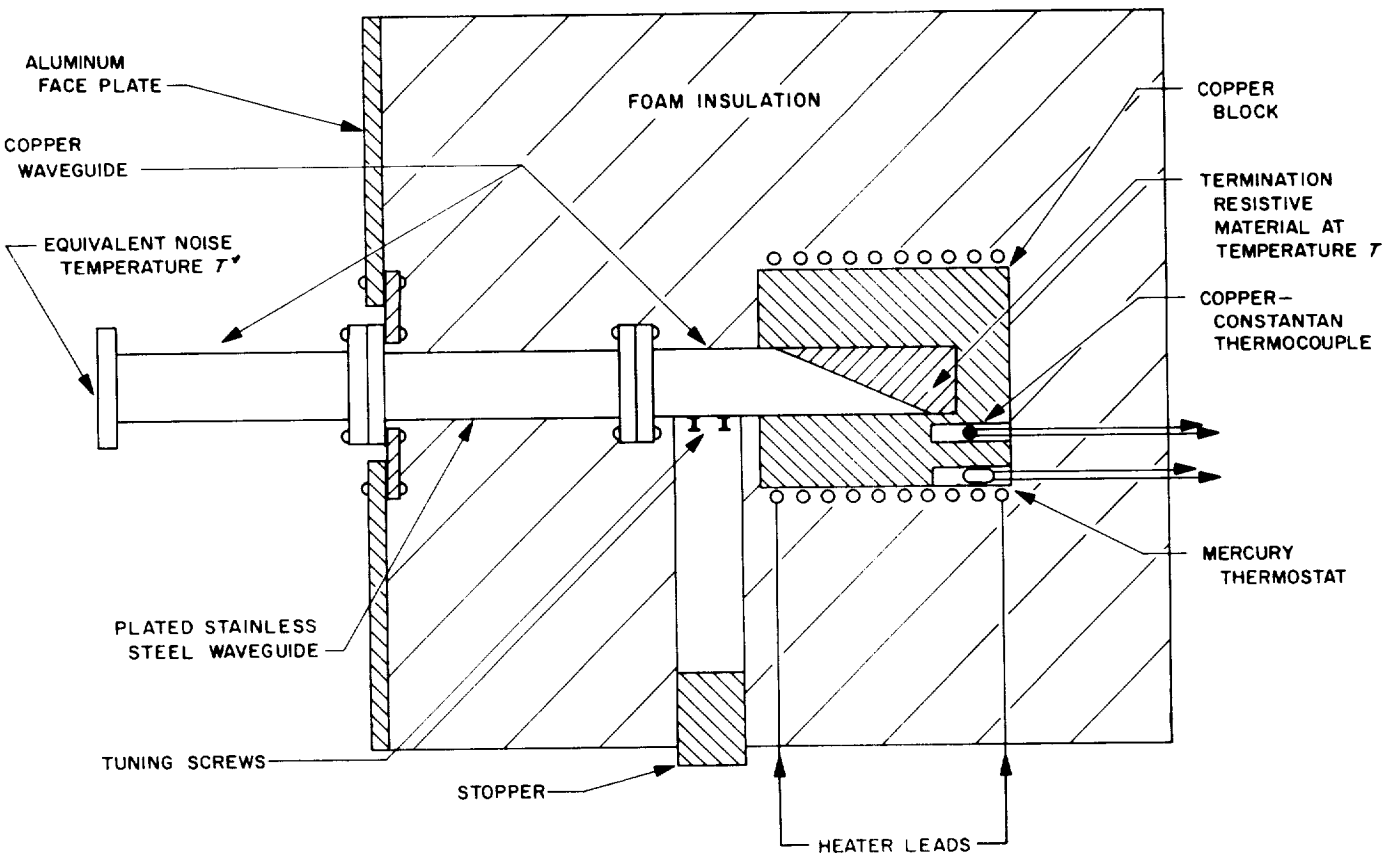


Fig. 17. K-band waveguide heat-regulated calibrated thermal termination

related to the actual termination temperature difference $(T - T_0)$ by (*SPS 37-28*, Vol. IV, p. 189)

$$(T' - T_0) \simeq (T - T_0)\left[1 - .2303\left(\frac{L_1}{2} + L_2\right)\left(\frac{f}{f_0}\right)^{\frac{1}{2}}\right] \quad (1)$$

Table 1 presents the tabulated difference temperatures for the discrete frequencies used in the experiment. The consequence of the lossy waveguide is to decrease the effect of the actual termination temperature difference by about 1.6%. A conservative estimate of 10% accuracy in the calibration indicates an accuracy of approximately 0.16% for the microwave calibration. This represents about 0.04°K accuracy for a 25°K temperature difference.

Tuning screws were fabricated with soldered nuts so that the screws bottom out at the required depth for a good microwave match (VSWR less than 1.02). Separate screws are used for each discrete frequency. In this way the termination can be matched in use at any of the required frequencies by changing the matching screws. Any loss associated with the matching need not be accounted for in the calibration since the loss occurs at the termination temperature T.

Table 1. Equivalent temperature difference ratios for K-band calibrated thermal termination

f, Gc	$\frac{T' - T_0}{T - T_0}$
20.6	0.9844
21.0	0.9843
21.5	0.9841
21.9	0.9840
22.5	0.9837
23.0	0.9836
23.5	0.9834
24.0	0.9832

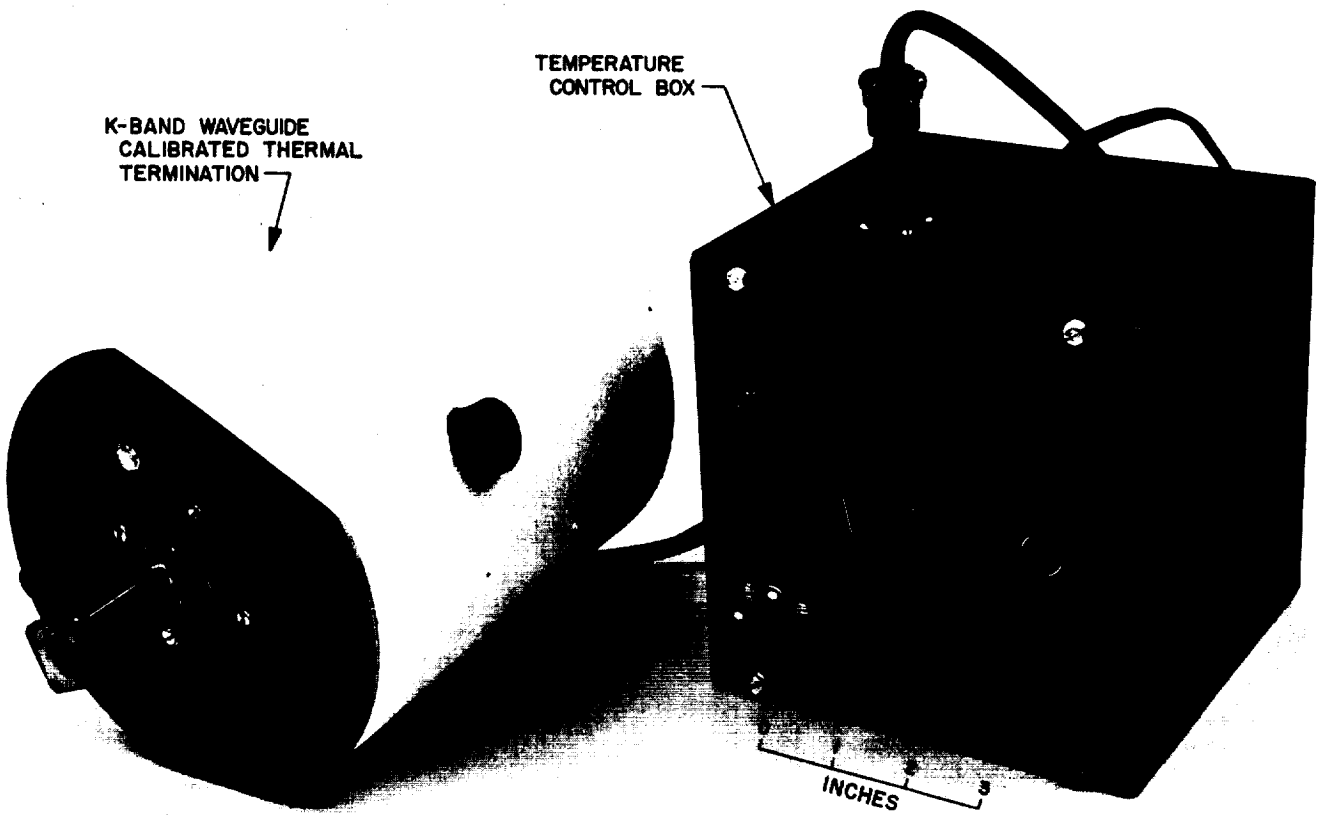


Fig. 18. K-band waveguide temperature-regulated thermal termination and control box

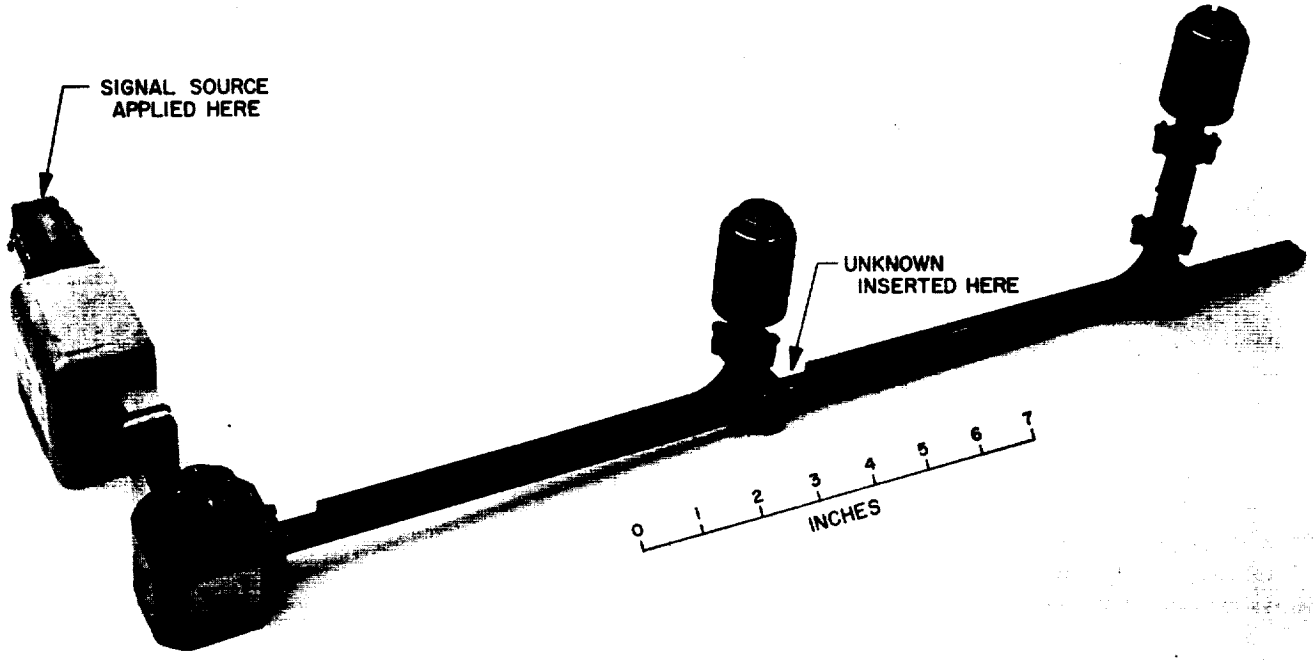


Fig. 19. K-band waveguide components for dc comparator insertion loss test set

B. Optical Communications Components

W. H. Wells

1. Lasers: Variable Q Optical Resonator Having Tilted Mirrors for Far Infrared Laser

a. Summary. We previously showed that there is a strong possibility of laser action in rotational transitions of polar molecules. To design a suitable resonator for such a laser, we have extended the theoretical work of Fox and Li (Refs. 5, 7) by studying the normal modes of an optical resonator consisting of two tilted plane mirrors. We considered larger tilt angles in order to use the spill-over light at the edge of the smaller mirror as the output coupling of the resonator. This scheme is not suitable near the visible spectrum, but is especially suited for intercepting a molecular beam having gain in the far infrared. Under reasonable conditions a single parameter,

$$\beta = (\text{tilt angle})\,(\text{number of wavelengths between mirrors})^{1/2},$$

determines the mode shape and its threshold gain. We solved for the lowest TEM_0 modes, and found that the single-pass gain required to oscillate increases almost linearly from 3% at $\beta = 1.0 \times 10^{-2}$ to 28% at $\beta = 8.0 \times 10^{-2}$.

b. Recent work. Soon after the invention of lasers, a number of workers studied the modes of optical resonators that are formed by two mirrors facing one another (Refs. 4, 5, 6, and many others). Fox and Li (Ref. 7) described the modes of a pair of tilted mirrors, but they considered only a small amount of tilt that might occur from imperfect mirror alignment. We are designing a resonator that employs greater tilt as a means of extracting the output from a far infrared laser (*SPS 37-24*, Vol. IV, pp. 140–148; *SPS 37-25*, Vol. IV, pp. 121-124).

To understand the modes of a pair of tilted mirrors, first consider the geometric optics approximation. As shown in Fig. 20, a ray tends to "walk out" the more open side even though the initial direction of the ray was inward. The geometric approximation breaks down at the sharp edge of the mirror, and wave effects must be considered. The diffraction pattern from the edge returns some light to positions back inside the resonator, from which it proceeds to walk out again. In the case of a laser oscillation, the medium between mirrors has gain,

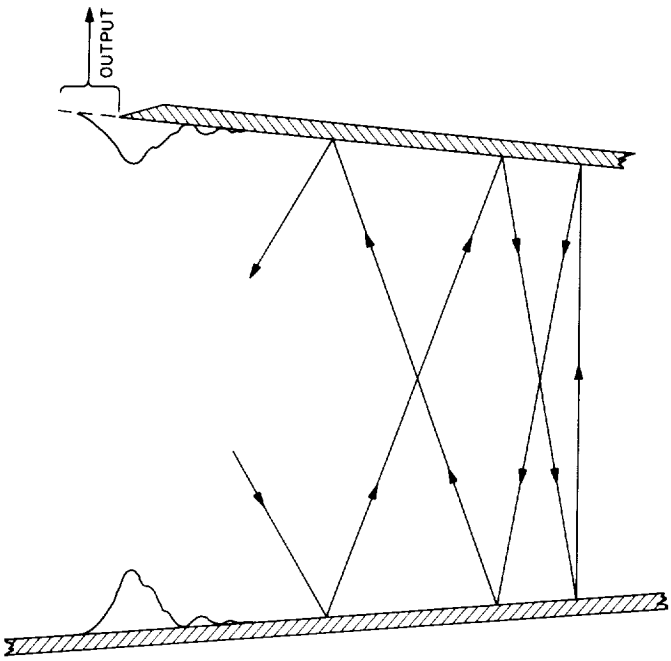


Fig. 20. Ray optics, normal mode, and output of tilted optical resonator

and the diffraction pattern provides sufficient positive feedback to maintain oscillation. The portion of the light pattern that spills over the edge of the shorter mirror provides the output coupling; the more the tilt, the greater the coupling, or the lower the Q of the resonator. When one mirror has adjustable tilt, this scheme allows one to optimize the output coupling for maximum power output, the optimum Q depending on the gain of the medium.

To our knowledge, a tilt output has not been employed before, certainly not at the short wavelengths of conventional lasers. At these wavelengths the angles would be very small, and would make adjustment highly critical. Moreover, it is easy to obtain output coupling at short wavelength by using partially transparent mirrors and eliminating "walk out" altogether by the use of concave mirrors.

For far infrared the situation changes. Partially transparent mirrors would be too lossy, so holes in mirrors have been used to extract the output. But spilling the radiation over the edge of a mirror is effectively the same as spilling through a hole, and the geometry of a straight-edge offers special advantages for extracting power from a molecular beam. The beam would enter the resonator from the more open side.

To intercept a molecular beam, the mode configuration should be broad in the directions that are transverse to the beam, and as thin as possible in the direction along the beam. Transverse broadness is desired merely to intercept a wide beam for maximum power. Longitudinal thinness is desired to minimize the doppler effect of various transverse velocity components that tend to take molecules out of resonance with the mode of oscillation. If a molecule is traveling with a velocity such that it remains in exact resonance with a perturbating oscillation, then its emission or absorption of power is the same whether the mode is thin and intense, or wide and weak, so long as the time integral of oscillation amplitude is constant for the duration of the interaction. But for molecules slightly out of resonance, a thick mode allows time for a phase difference to accumulate between the internal motion of the molecule and the oscillation of the field. Beyond 90 deg of phase difference, the molecule begins to reabsorb the energy it was emitting at the start of the interaction (Ref. 8).

The above argument for a thin mode configuration was presented in the conventional manner as a doppler effect. However, in this special case there is a fundamentally simpler description of the same effect. Consider the nodal planes of zero electric field in a resonator that consists of two nearly parallel mirrors. Now suppose the molecules that traverse the oscillation without crossing a nodal plane are in perfect resonance and finish delivering their power just as they pass out of the oscillation region. It follows that a molecule traveling at an angle sufficient to cross nodal planes will experience a reversal of the electric field at each crossing. These reversals change photon emissions into reabsorptions and vice versa to greatly reduce the efficiency of laser action. The useful fraction of the beam is on the order of the angular spread of trajectories that do not cross nodal planes (except outside the half-power limits of the mode) divided by the total angular spread of molecular trajectories in the beam. The numerator in this fraction equals a half-wavelength over the longitudinal (half-power) thickness of the mode, so that the latter is to be minimized for maximum efficiency.

Tilted rectangular mirrors meet the mode shape requirements nicely. The transverse breadth is merely the length of the rectangle. The lowest order mode, i.e., the one which first breaks into oscillation as the gain of the medium increases, is a single thin bright fringe (with weak satellite fringes) on the spillover (forward) edge. When a molecular beam enters the resonator at the forward edge, still further mode narrowing results

from saturation effects. The tail of the mode that extends back from the forward edge is suppressed by disappearing gain of the medium beyond the point where the beam has delivered most of its available power.

The remainder of this report is the mathematical solution for the lowest order mode configurations and threshold gains in resonators having greater tilt than those considered by Fox and Li. Interest in these modes extends somewhat beyond our special applications. The tilt occurs accidentally in some solid-state lasers that have relatively poor optical homogeneity, especially the tiny injection lasers. We shall restrict the problem to two dimensions, i.e., infinite strip mirrors. In our case, this corresponds to neglect of edge effects at the extremities of long rectangular mirrors (where beam density may be below threshold anyway). In certain cases (Refs. 4 to 7) it is valid to solve a square mirror by combining the solution for two strips which intersect to form the square.

For reasonable values of threshold gain, say 3 to 40%, the mirrors are nearly parallel, and the normal modes almost exactly plane polarized TEM. Therefore, scalar diffraction theory suffices, and we need only discuss wave amplitude without specifying which component of the E or H field is meant.

Obviously one strip mirror should be considerably wider than the other so that the power output will not be divided, but will all appear as spillover at the narrow mirror. Then we may as well take the wide mirror to be an infinite plane at distance b from the output edge of the narrow mirror. In this case the narrow mirror sees an image of itself at the apparent distance of $2b$. Therefore, let us solve the equivalent problem of identical mirrors at distance $2b$, since mathematical simplicity results from having the equation for propagating light to the left identical to the one for propagation to the right.

The basic equation for $2D$ diffraction is

$$u_2(x_2) = \frac{G}{(2b\lambda)^{1/2}} \int_0^w \exp\left[j\frac{\pi}{4} + j\frac{2\pi}{\lambda}\rho'_{12}(x_1, x_2) \right] \times u_1(x_1)\, dx_1, \tag{1}$$

where x_1, x_2 are the coordinates along the two mirrors, u_1 and u_2 the amplitude distributions on the mirrors, ρ' is the distance between points (lines) x_1 and x_2, the light propagates from 1 to 2, G is a phenomenological constant to account for amplitude gain in one pass, and the width of the mirror extends from zero to w. The geometry of Fig. 21 is used to express ρ' explicitly, and Eq. (1) becomes

$$\int_0^w \exp j\left[\frac{2\pi}{\lambda}\rho_{12}(x_1, x_2)\right] u(x_1)\, dx_1 = \frac{(2b\lambda)^{1/2}}{Ge^{j\delta}} u(x_2) \tag{2}$$

where

$$\rho_{12} = \left\{ \left[2b - \alpha(x_1 + x_2) \right]^2 + \left[x_1 - x_2 \right]^2 \right\}^{1/2} - 2b \tag{3}$$

and

$$\delta = \frac{\pi}{4} + \frac{4\pi b}{\lambda}. \tag{4}$$

In Eq. (2) the subscripts on u were omitted, because we wish to find normal modes which (by definition) reproduce the same amplitude distribution with each bounce, i.e., $u_2(x) = u_1(x) = u(x)$. Now Eq. (2) has the form of an integral eigenvalue equation where $u(x)$ is the eigenfunction and $(2b\lambda)^{1/2} G^{-1} e^{-j\delta}$ is the complex eigenvalue.

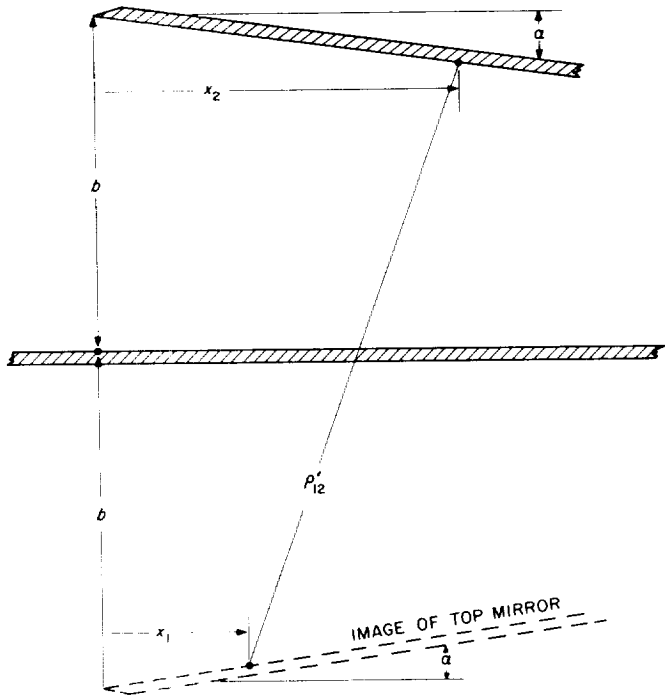


Fig. 21. Notation for diffraction integral

The solution of Eq. (2) as it stands would be a three parameter family of curves for various values of α, b/λ, and w/λ. Through good fortune we are able to reduce the solution to a one parameter family. First we note that the greatest tilt Fox and Li studied gives very little light power on the back edge of the strip mirror, and intuitively we expect this situation to hold or become more pronounced as tilt increases. Hence, we let $w \to \infty$ and expect little error. In practice we let w be as large as possible, consistent with a 100 point representation of the function for computer solution. More points require a more costly and elaborate computer program. For two solutions we varied w and confirmed that the results were insensitive to mirror width, provided the mirror is wide compared to the main fringe of light. As a further precaution, we faded out the upper limit by multiplying the last four points by 0.8, 0.6, 0.4, and 0.2. This eliminates knife edge diffraction fringes, which would otherwise be the main error from a short integration range.

After eliminating w, the other parameter reduction results from expanding ρ in power series to second order in α and $x_{1,2}/b$. Then α and b/λ occur only in the combination

$$\beta = \alpha(b/\lambda)^{1/3} \tag{5}$$

which becomes the single parameter. The expansion is valid whenever

$$X^4 \ll 10b^3\lambda \text{ and } X^3 \ll \frac{\lambda b^2}{\alpha}, \tag{6}$$

where X is the maximum value of x on the (practical) integration range. Since we do not know X at the outset, we merely used the expansion first, then examined the results and picked a value for X where the eigenfunction had decayed to a small value. Then the inequalities (6) were examined to find the range of b and λ for which the results were valid. Fortunately, over the range of β we treated, the lowest order mode TEM_0 is valid for $b/\lambda > 100$, which includes all optical and far infrared experiments with reasonable laboratory dimensions.

The simplified eigenvalue problem (Eq. 2) is expressed in the final form

$$\int_0^\infty \exp j\left[-(y+s) + \frac{(y-s)^2}{8\pi\beta^2}\right]\Psi(s)ds = \frac{8^{1/2}\pi\beta}{Ge^{j\delta}}\Psi(y) \tag{7}$$

where

$$y = \frac{2\pi\alpha x}{\lambda} \text{ and } \Psi(y) = u\left(\frac{\lambda y}{2\pi\alpha}\right) = u(x)$$

This form was solved in the usual way by iteration. One makes a best guess at $\Psi(s)$, substitutes it in the integrand, and integrates to find an improved function. Continuing, one puts $\Psi^{(n-1)}$ into the integrand, then integrates to find $\Psi^{(n)}$, $G^{(n)}$, and $\delta^{(n)}$, until the process converges to the highest eigenvalue (lowest G). The iteration corresponds to the physical process of triggering an oscillation with a burst of noise when the gain is set where only one mode can oscillate. Then we allow the other modes to decay in many bounces (iterations) to find the shape of the lowest order mode. As a convergence criterion, we required that the real and imaginary parts of the eigenvalue ($8.9\beta G^{-1}\ e^{-j\delta}$) not change more than 0.001 in one iteration, about $1.1 \times 10^{-2}\beta^{-1}\%$ accuracy. This was relaxed a factor of three for the higher values of β. The full range of β was from 10^{-2} to 10^{-1}.

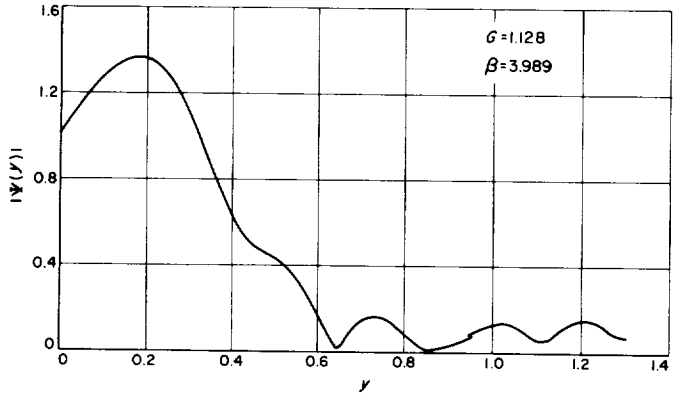


Fig. 22. Amplitude distribution of TEM_0 mode with 13% gain per pass at threshold of oscillation

Fig. 22 shows the magnitude of the eigenfunction, i.e., $|\Psi(y)|$ when $\beta = 3.989 \times 10^{-2}$. This is one of the cases that was tested for dependence upon the upper limit of integration. The part of the curve from $y = 0$ to 0.95 was obtained with the upper limit $Y = 0.95$. The remainder was obtained with $Y = 1.3$. The curves meet with a barely perceptible discontinuity, and would coincide similarly over the range from 0.0 to 0.95. Eigenvalues also match closely:

Y	G	δ, rad	(G-1) 100, %
0.95	1.1281	−0.1012	12.81
1.30	1.131	−0.1039	13.1

Although G was defined as single pass amplitude gain at the threshold of oscillation through the medium of the equivalent problem (distance $2b$), it also equals

power gain for a single pass (distance b) in the original problem, since power $\sim |\Psi|^2$, and the equivalent resonator is twice as long as the original, which squares the gain.

Over the range of β that we studied, G varied almost linearly from 3% at $\beta = 10^{-2}$ to 28% at $\beta = 8 \times 10^{-2}$. Fig. 23 gives gain in percent as a function of β. To use this graph in resonator design, one would estimate the gain of the laser medium under optimum oscillation conditions. The graph then tells what value of β gives this gain. The wavelength of the spectral line and β then give the combinations of resonator length and tilt angle that will produce the desired output according to the formula (from Eq. 5)

$$\alpha b^{1/2} = \beta \lambda^{1/2}$$

We checked the computer program by repeating the most tilted case treated by Fox and Li (Ref. 7). In their notation this is $N = 2.5$ and $\delta = \lambda/36$, which corresponds to our $\beta = 0.1242$. Agreement was satisfactory even though our program was not fully convergent at the time. We found 3.8% gain compared to 3.4. The eigenfunction magnitude $|\Psi(y)|$ matched to 3%. The phase of Ψ differed up to 2 deg in the main fringe, and up to 8 deg in satellite fringes in the tail of Ψ. These differences may be caused by our fading out the upper integration limit.

In all, we solved 7 values of β ranging from 1.0×10^{-2} to 7.98×10^{-2}. The shape of $|\Psi(y)|$ for $\beta = 0.040$, Fig. 22, is representative, so instead of plotting all 7 functions, let us extract certain features and plot them versus β. Fig. 24 gives the eigenphase δ and gain in a form that graphs more accurately than Fig. 23. The characteristic features of the eigenfunctions occur at y values that are nearly proportional to $\beta^{1/2}$. The ratios of certain of these y values to $\beta^{1/2}$ are plotted in Figs. 25 and 26.
The quantities are:

(1) c, the position of maximum $|\Psi|$.

(2) a, distance (in y units) from c to e^{-1} intensity point, i.e., $|\Psi(c)/\Psi(c + a)|^2 = e$.

(3) Intercept, the place where a straightedge intercepts the y-axis if adjusted to the nearly linear trailing edge of the graph of $|\Psi(y)|$.

Beyond the intercept, each eigenfunction has little fringes that lead to an average relative intensity $|\Psi/\Psi(c)|^2$ of about 0.4% scattered back into the resonator. Finally, the relative amplitude at the spillover edge $|\Psi(0)/\Psi(c)|$ is plotted in Fig. 26.

Fox and Li reported convergence difficulty with the iteration process because gains of the first two modes were not sufficiently different to cause the second to

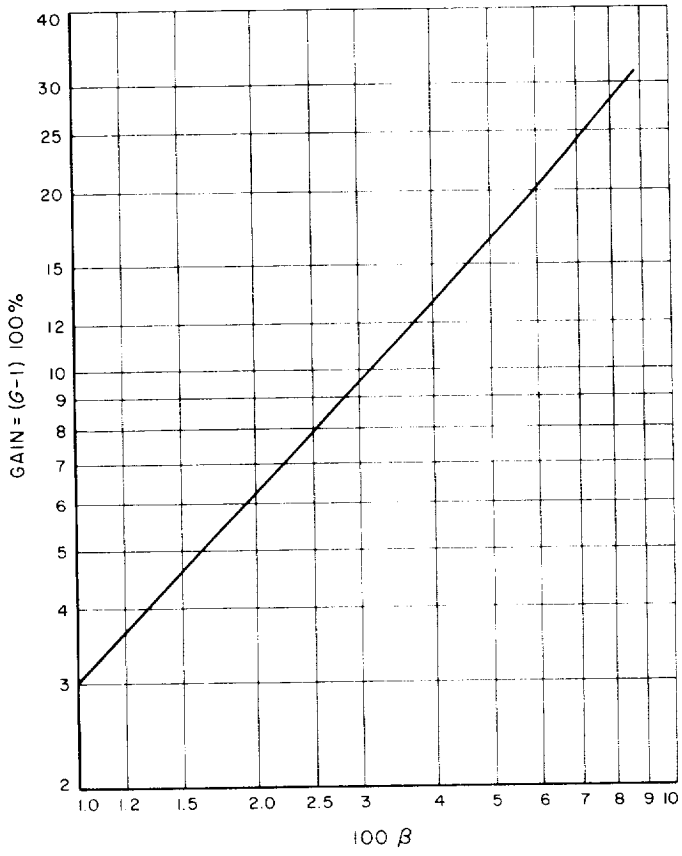


Fig. 23. Single-pass power gain at oscillation threshold

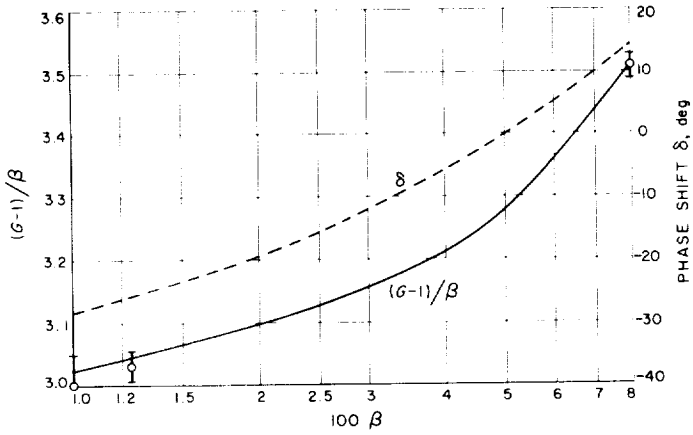


Fig. 24. Gain and phase shift

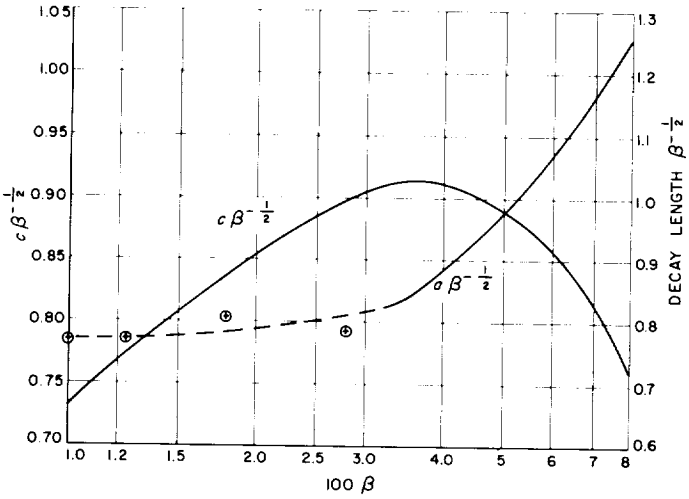


Fig. 25. Position of maximum and decay length

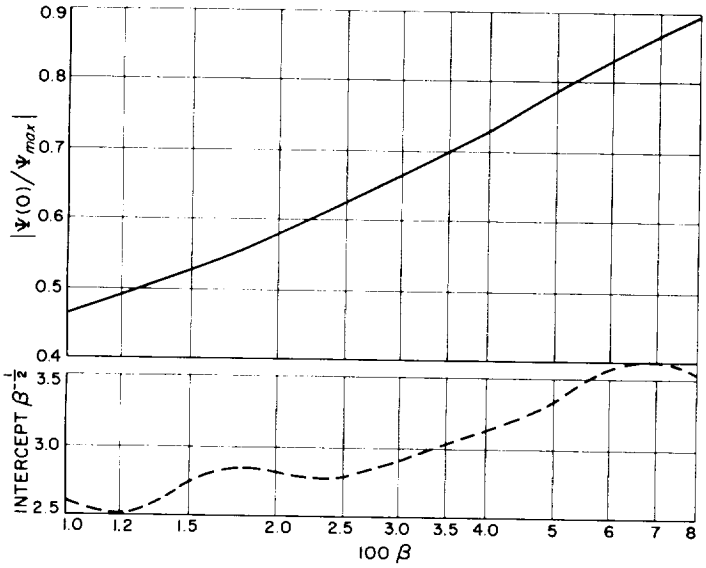


Fig. 26. Relative amplitude at output edge and intercept of trailing edge of $|\Psi(y)|$

decay significantly faster than the first. They disentangled the two modes from the beats they exhibit in successive iterations. We were able to isolate the lowest order mode by a simpler technique, which is the mathematical analog of a familiar physical process, introduction of additional loss that discriminates against unwanted modes. A loss in a mirror corresponds to painting part of it black, or in the mathematics of diffraction, leaving out part of the integration range. Of course, the integration range must be restored during the iteration, but this can be done gradually so as not to re-excite the suppressed modes with a large transient. For the lowest order mode, we started with the range min $(0, c - a) < y < c + a$ and let it slide out to $0 < y < Y$.

We are attempting to isolate the next mode TEM_1 by omitting a part of the integration range near $y = c$, which discriminates strongly against the TEM_0 mode, since the latter's maximum occurs at that point. This is analogous to touching a harp string in the center to damp the fundamental and hear the first harmonic. We will also slide the outer limits of the range to eliminate higher orders.

The author is grateful to Wiley R. Bunton for programming this problem for the IBM 7090 computer.

C. RF Techniques

T. Otoshi

1. Ac Ratio Transformer Technique, Capacitive Phase Shifter

a. Summary. The development of the ac ratio transformer technique for precision insertion loss measurements has been described in previous reports (*SPS 37-22*, Vol. IV, pp. 189–196, *SPS 37-25*, Vol. IV, pp. 128–132, and *SPS 37-27*, Vol. IV, pp. 165–167). This report presents a description of the operating principle of the ac ratio transformer technique and the results of an analysis made of the error introduced by a necessary adjustment of the phase shift capacitance of the insertion loss set during a measurement.

Graphs of insertion loss error and phase shift as a function of phase shift capacitance and capacitance change are presented. The error values may be used as a correction term to the measured insertion loss for improved accuracy or used to determine if a correction should be made. For the phase shift capacitance and capacitance change normally used for an insertion loss measurement with this technique, it has been found from making many insertion loss measurements that the error (due to the capacitance change) is typically much less than 0.001 db.

b. Recent work.

Principle of operation. The ac ratio transformer technique uses an audio-frequency substitution principle for insertion loss measurements. An accurate ac ratio transformer is used as the audio-frequency attenuation standard. Fig. 27 is a block diagram of the test setup. The RF

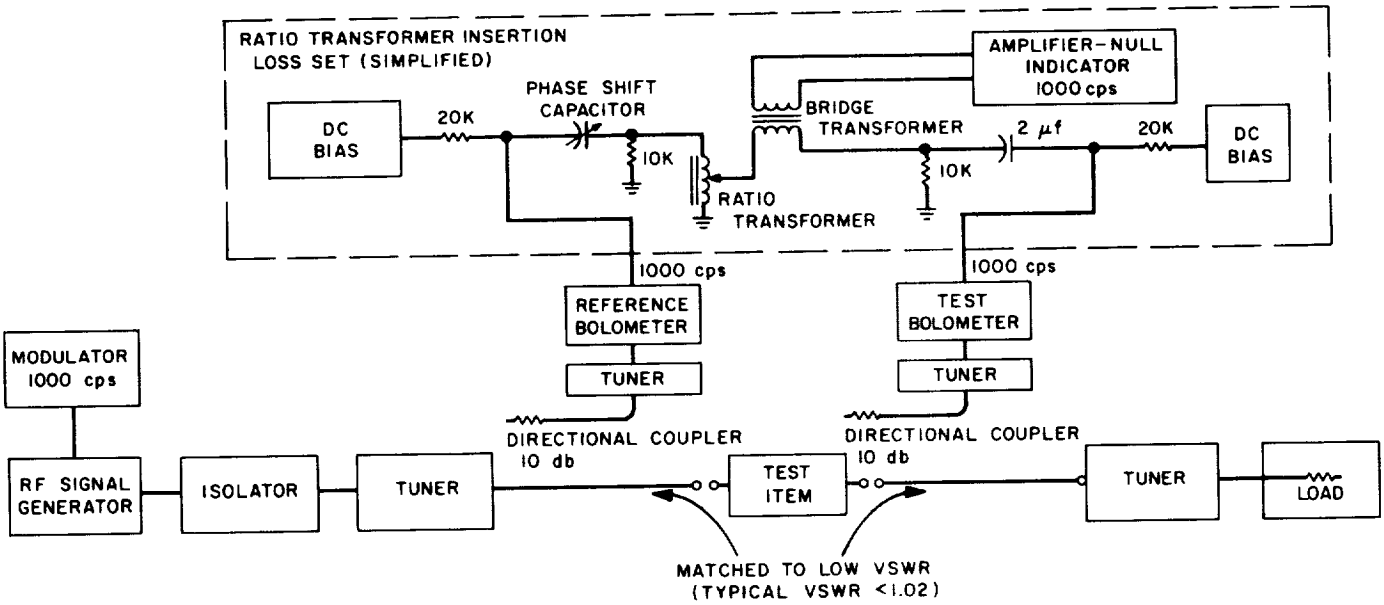


Fig. 27. Test setup block diagram

signal generator is 100% square-wave modulated at 1000 cps. Detection is accomplished by means of a dual channel bolometer system. The RF power delivered to the test bolometer is simultaneously compared to that delivered to the reference bolometer. Any amplitude changes in the output of the RF signal generator changes the power in the respective channels in the same proportion so that the power ratio remains constant.

If the bolometers for the test set are truly square-law detectors, comparisons may be made of their detected ac voltages to give the ratio of the RF power levels at the reference and test bolometers. Comparison of the detected ac reference and test signals is accomplished by a null detection technique. The divided reference signal voltage which is developed across the output terminals of the ratio transformer is fed into one of the bridge arms and is used to buck the test signal voltage being fed into the opposite arm of the bridge transformer nulling circuit (Fig. 27).

The operating procedure is to adjust the ratio transformer setting and phase shift capacitors until a null occurs in the bridge transformer output. At this null condition, the ratio of the test to the reference signal is equal to the dial setting of the ratio transformer. This ratio of test to reference signals is determined before and after insertion of the test item in the RF system. The RF power ratios indicated on the ratio transformer dial before and after insertion are each converted to decibels with the use of conversion tables. The difference of the decibel values gives the insertion loss of the test item in decibels.

Generally, there exists a small amount of audio-frequency phase shift in the system. In order to correct this to obtain the best null, a set of decade capacitors is used. The value of phase shift difference will vary when bolometers are changed. During an actual measurement, when large values of insertion loss are inserted and removed, the capacitance change required corresponds to a phase shift of a small fraction of 1 deg. This capacitance change will produce a small change in the amplitude of the reference signal output and hence, produces an error.

Error analysis. For a dual channel insertion loss measurement system, it can be shown that the insertion loss of a test item is given by the expression,

$$IL = -10\log_{10}\left[\frac{{}^{i}P_{RF_1}}{{}^{i}P_{RF_2}}\right] + 10\log_{10}\left[\frac{{}^{f}P_{RF_1}}{{}^{f}P_{RF_2}}\right] \qquad (1)$$

where P_{RF_1} is the RF power into the reference channel bolometer, and P_{RF_2} the RF power into the test channel bolometer. The front superscripts i and f refer to the initial value (before insertion) and final value (after insertion), respectively. The ratios in the parentheses of Eq. (1) will remain constant for a dual channel measurement system.

If the bolometers are truly square-law detectors, then the amplitudes of their ac output voltages will be directly proportional to the respective RF input powers. Therefore,

$$IL = -10\log_{10}\left[\frac{|{}^{i}e_{b_1}|}{|{}^{i}e_{b_2}|}\right] + 10\log_{10}\left[\frac{|{}^{f}e_{b_1}|}{|{}^{f}e_{b_2}|}\right] \qquad (2)$$

For the ac ratio transformer system, the equivalent 1-kc circuit is shown in Fig. 28. The 20-kΩ minimum source impedances of the dc bias supplies have been omitted from the equivalent representation since their major effect in this analysis is to change the effective bolometer resistances of 200 Ω by only 1%.

When the phase shifter and ratio transformer are adjusted for a null at the output of the bridge transformer, it may be seen from Fig. 28 that

$$se_1 = e_2$$

or

$$s = \left(\frac{e_2}{e_1}\right) \qquad (3)$$

where s is the ratio transformer setting. The transfer functions of the reference and test channel circuits for initial and final nulling conditions are given by

$${}^{i}K_1 = \frac{{}^{i}e_1}{{}^{i}e_{b1}}, \quad {}^{i}K_2 = \frac{{}^{i}e_2}{{}^{i}e_{b2}} \qquad (4)$$

and

$${}^{f}K_1 = \frac{{}^{f}e_1}{{}^{f}e_{b1}}, \quad {}^{f}K_2 = \frac{{}^{f}e_2}{{}^{f}e_{b2}} \qquad (5)$$

Substitution of Eq. (4) and (5) into Eq. (2) gives

$$IL = -10\log_{10}\left[\frac{|{}^{i}e_1|}{|{}^{i}K_1|}\frac{|{}^{i}K_2|}{|{}^{i}e_2|}\right] + 10\log_{10}\left[\frac{|{}^{f}e_1|}{|{}^{f}K_1|}\frac{|{}^{f}K_2|}{|{}^{f}e_2|}\right] \qquad (6)$$

$$= -10\log_{10}\left[\frac{|{}^{i}e_1|}{|{}^{i}e_2|}\right] + 10\log_{10}\left[\frac{|{}^{f}e_1|}{|{}^{f}e_2|}\right]$$

$$-10\log_{10}\left[\frac{|{}^{i}K_2|}{|{}^{i}K_1|}\frac{|{}^{f}K_1|}{|{}^{f}K_2|}\right] \qquad (7)$$

Substitution of Eq. (3) into (7) gives

$$IL = -10\log_{10}\left[\frac{1}{{}^{i}s}\right] + 10\log_{10}\left[\frac{1}{{}^{f}s}\right]$$

$$-10\log_{10}\left[\frac{|{}^{i}K_2|}{|{}^{i}K_1|}\frac{|{}^{f}K_1|}{|{}^{f}K_2|}\right] \qquad (8)$$

The sum of the first two terms of Eq. (8) is actually the measured decibel insertion loss while the last term is the error term which will be denoted ϵ_{db}.

$$\text{True } IL = \text{measured } IL - \epsilon_{db} \qquad (9)$$

Since no passive element (capacitance or resistance) changes are made for the test channel circuit during the measurement, then

$$|{}^{f}K_2| = |{}^{i}K_2|$$

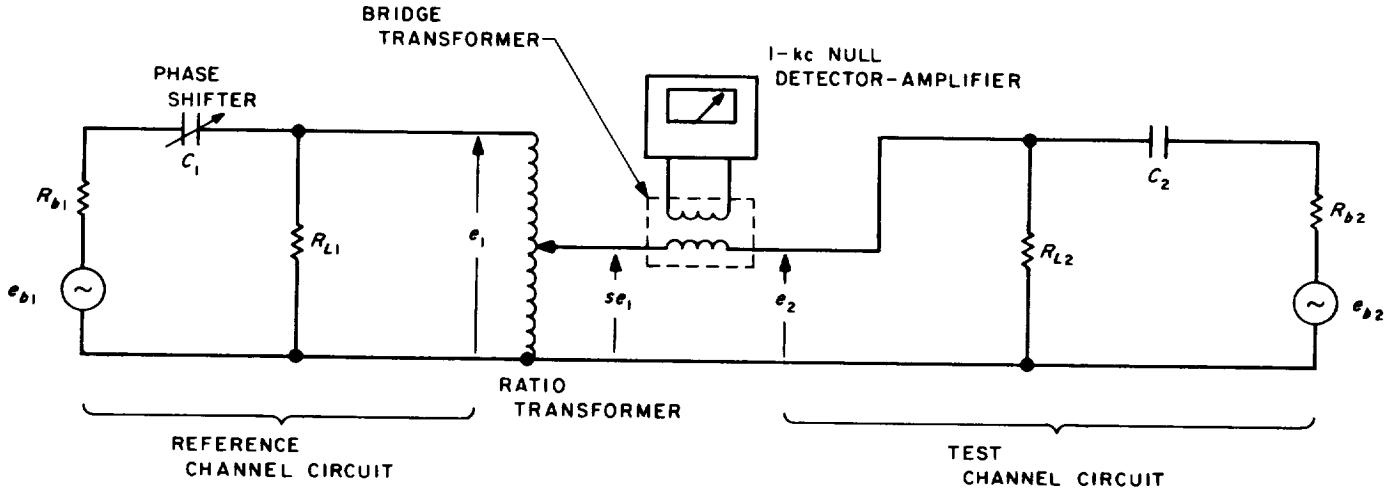


Fig. 28. Equivalent 1-kc circuit for the ac ratio transformer insertion loss set

and the error term becomes

$$\epsilon_{db} = 10 \log_{10} \left[\frac{|{}^{f}K_1|}{|{}^{i}K_1|} \right] \tag{10}$$

For determining the error term, we are concerned only with the transfer voltage of the reference channel. Fig. 29 shows a simplified equivalent circuit of the reference channel when the ratio transformer and capacitive phase shifter have been adjusted to produce a null at the output of the bridge transformer. The following assumption has been made in order to simplify analysis:

It has been assumed that the ratio transformer input impedance may be neglected and varying the ratio transformer dial setting will have negligible effect on the transfer function of the reference circuit. The input impedance of the ratio transformer used for the insertion loss set is typically greater than 100 kΩ at 1 kc. A more detailed preliminary analysis, based on typical operating conditions and the manufacturer's formulation of the equivalent circuit of the ratio transformer input impedance as a function of dial setting, shows that the errors due to the above assumption are negligible.

From Fig. 29, the transfer function is derived as

$$K_1 = \left(\frac{R_{L1}}{R_{b1} + R_{L1}} \right) \frac{1}{\left[1 - j \dfrac{1}{\omega(R_{b1} + R_{L1})C} \right]} \tag{11}$$

Substitution of Eq. (11) into Eq. (10) gives

$$\epsilon_{db} = 10 \log_{10} \left\{ \frac{\sqrt{1 + \left[\dfrac{1}{\omega(R_{b1} + R_{L1})^{i}C_1} \right]^2}}{\sqrt{1 + \left[\dfrac{1}{\omega(R_{b1} + R_{L1})^{f}C_1} \right]^2}} \right\} \tag{12}$$

The corresponding phase in degrees of the transfer function is given by

$$\phi_0 = \frac{180°}{\pi} \tan^{-1} \left[\frac{Im\,(K_1)}{Re\,(K_1)} \right] \tag{13}$$

and the phase shift in degrees produced by a capacitance change is given by

$$\Delta\phi_0 = {}^{i}\phi_0 - {}^{f}\phi_0$$

$$= \frac{180°}{\pi} \left\{ \tan^{-1} \left[\frac{1}{\omega(R_{b1} + R_{L1})\,{}^{i}C_1} \right] - \tan^{-1} \left[\frac{1}{\omega(R_{b1} + R_{L1})\,{}^{f}C_1} \right] \right\} \tag{14}$$

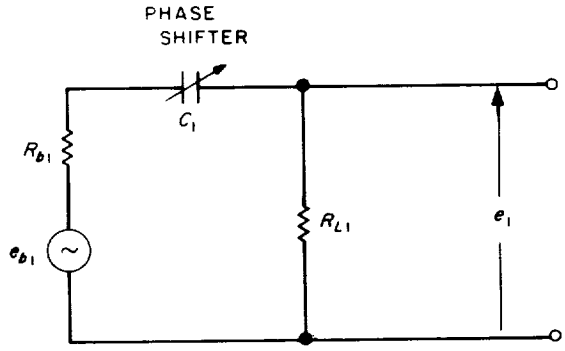


Fig. 29. Simplified equivalent 1-kc circuit for the reference channel at a null setting

For the graphs plotted on Figs. 30 and 31, we have let

$${}^{f}C_1 = {}^{i}C_1 + \Delta C_1$$

in Eqs. (12) and (14) where ΔC_1 is the capacitance change, and

$$R_{b1} = 200\ \Omega$$

$$R_{L1} = 10\ \text{k}\Omega$$

and $\omega = 2\pi \times 10^3$ rad/sec

The values plotted on Figs. 30 and 31 were obtained from the output of the IBM 7090 computer. It may be seen from Fig. 30 that if the initial operating capacitance is 0.2 μf, then in order to limit the insertion loss measurement error (contributed by the phase shifter) to 0.001-db maximum, the capacitance change, required for a null at the output of the bridge transformer after the test device is inserted into the RF system, should not be greater than 0.008 μf. The corresponding audio phase shift produced by this initial operating capacitance and capacitance change is found, from Fig. 31, to be 0.17 deg.

For most commercial bolometers tested in this system, it was possible to obtain the initial null with a total phase shift capacitance of 0.4 μf or greater and capacitance changes of less than 0.02 μf to obtain the null after insertion of the test item. The corresponding insertion loss error is about 0.0003 db. The measured insertion losses could be improved in accuracy by using the error term as a correction term in Eq. (9). However, based on many measurements made with this insertion loss system, it has been found that a correction has not been necessary.

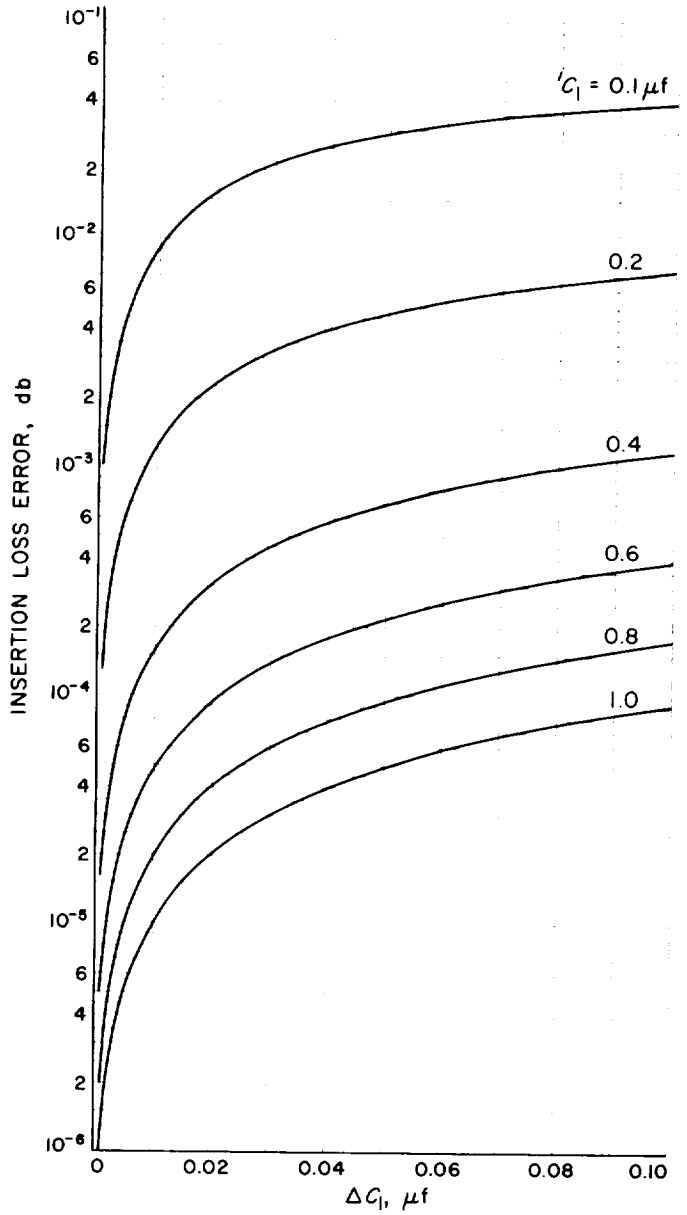


Fig. 30. Insertion loss error due to capacitance change

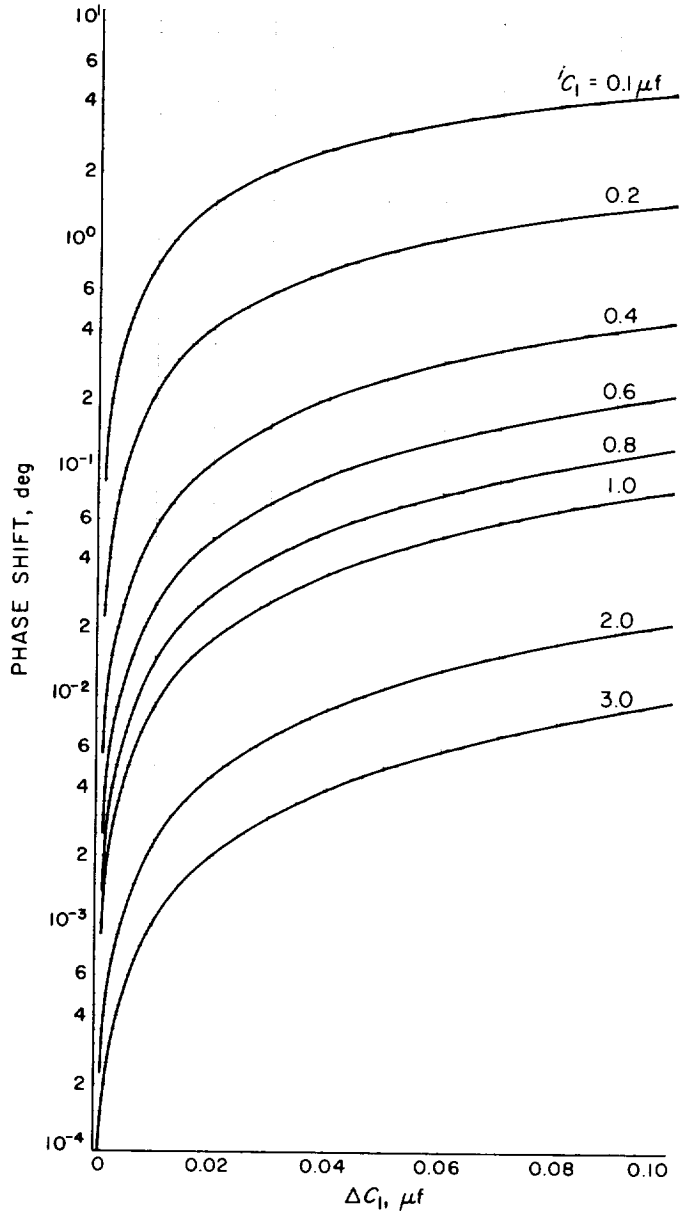


Fig. 31. Phase shift due to capacitance change

References

1. Friis, H. T., "Noise Figure of Radio Receivers," *Proceedings of IRE*, p. 420, July 1944.

2. Colvin, R. S., *A Study of Radio-Astronomy Receivers*, Scientific Report No. 18, Stanford Radio Astronomy Institute, Publication 18A, p. 112, October 31, 1961.

References (Cont'd)

3. Stelzried, C. T., Petty, S., "Microwave Insertion Loss Test Set," *IEEE Transactions on Microwave Theory and Techniques*, p. 475, July 1964.

4. Boyd, G., and Gordon, J., "Confocal Multimode Resonators for Millimeter through Optical Wavelength," *Bell System Technical Journal*, Vol. 40, pp. 489–507, March 1961.

5. Fox, A. G., and Li, T., "Resonant Modes in a Maser Interferometer," *Bell System Technical Journal*, Vol. 40, pp. 453–488, March 1961.

6. Boyd, G., and Kogelnik, H., "Generalized Confocal Resonator Theory," *Bell System Technical Journal*, Vol. 41, p. 1347, July 1962.

7. Fox, A. G., and Li, T., "Modes in a Maser Interferometer with Curved and Tilted Mirrors," *Proceedings of the IEEE*, Vol. 51, pp. 80–89, January 1963.

8. Feynman, R. P., Vernon, F. L., and Hellwarth, R. W., "Geometrical Representation of the Schrodinger Equation for Solving Maser Problems," *Journal of Applied Physics*, Vol. 28, pp. 49–52, January 1957.

XXI. Communications Systems Research: Mathematical Research

A. Higher-Order Distributions of Pseudo-Random Numbers

R. C. Titsworth

1. Introduction

In *SPS 37-27*, Vol. IV, pp. 185–189, the author presented a construction procedure for a sequence of random numbers which were uniformly distributed on the interval (0, 1) and in which all numbers were uncorrelated, regardless of their separation in the sequence. It has since been found that *pairs* of these numbers are not *independent* but, to the contrary, the two-dimensional distribution is sometimes extremely pathological. This article shows that a slight alteration of the generating procedure yields a sequence of numbers equally as uncorrelated as the first, with the additional property that any M-tuple of adjacent numbers lie equally distributed over the unit M-cube, for any prespecified integer M. With this improvement, the method can be used to generate a sequence with any preselected degree of randomness desired.

2. The Result

Let $a = \{a_k\}$ be the sequence of *zeros* and *ones* generated by the linear recursion relation

$$a_k = c_1 a_{k-1} + c_2 a_{k-2} \cdots + c_n a_{k-n} \pmod 2$$

for any given set of integers c_i $(i = 1, 2, \cdots, n)$, each having the value 0 or 1. We require $c_n = 1$, and say that the sequence has degree n.

The period p of a linear recurring sequence cannot be greater than $2^n - 1$, and the necessary and sufficient condition that $p = 2^n - 1$ is that the polynomial

$$f(x) = 1 + c_1 x + c_2 x^2 + \cdots + x^n$$

be primitive over $GF(2)$ (Refs. 1 and 2).

We shall assume in the remainder of this article that $f(x)$ is a primitive n^{th} degree polynomial over $GF(2)$; the sequence a is then a *maximal-length linearly recurring sequence modulo 2.*

Based on this binary stream of *zeros* and *ones*, a sequence of random numbers can be generated with the following properties:

Theorem. Let $\{a_k\}$ be a (0, 1) binary sequence generated by an n^{th} degree maximal-length linear recursion relation modulo 2. Let $(q, 2^n - 1) = 1$, L be a positive integer, and $q \geqq L$; write $y_k = \cdot a_{kq-1} a_{kq-2} \cdots a_{kq-L}$, the binary expansion of a real positive number in the interval (0,1). Let w_k be a real number in the interval $(-1, +1)$ related to y_k by $w_k = 1 - 2y_k - 2^{-L}$. Then, averaged over all possible (assumed equally likely) initial values y_0 (or w_0):

(1) The mean value μ of the sequence w_k

$$\mu = -2^{-n}\left(\frac{1 - 2^{-L}}{1 - 2^{-n}}\right) \approx 0$$

and variance σ^2

$$\sigma^2 = \frac{1}{3} + 2^{-n}\left[\frac{1}{3}\left(\frac{1 - 2^{-2L}}{1 - 2^{-n}}\right) - \frac{(1 - 2^{-L})^2}{1 - 2^{-n}} - 2^{-n}\left(\frac{1 - 2^{-L}}{1 - 2^{-n}}\right)^2\right] \approx \frac{1}{3}$$

(2) The sample autocorrelation function, defined by

$$\hat{R}(m) = \frac{1}{N}\sum_{k=1}^{N} w_k w_{k+m}$$

has its mean value $R(m)$, given by

$$R(m) = -2^{-n}\left(\frac{1 - 2^{-L}}{1 - 2^{-n}}\right) \approx 0$$

for non-zero integral values of $|m|$ less than $(p - L)/q$. The variance of $\hat{R}(m)$ about $R(m)$ is bounded by

$$var\,[\hat{R}(m)] < \frac{1}{N}\left[1 + \frac{1}{(2^n - 1)}\right] \approx \frac{1}{N}$$

(3) The relative number of times $\hat{T}$ that y_k falls in the interval for which the first d positions of the binary expansion are fixed [i.e., a neighborhood of length 2^{-d} in the interval (0,1)], has mean

$$T = E\,[\hat{T}] = 2^{-d}\left[1 + \frac{1}{(2^n - 1)}\right] + \frac{1}{2}\,[g(0) - 1]\left(\frac{1}{2^n - 1}\right) \approx 2^{-d}, \qquad |g(0)| = 1$$

for any number N of points y_k. The variance of $\hat{T}$ is bounded by

$$var\,[\hat{T}] < \frac{1}{4}\left[1 + \frac{1}{(2^n - 1)}\right]\left[\frac{1}{N} + \frac{2}{(2^n - 1)}\right] \approx \frac{1}{4N}$$

(4) The relative number of times $\hat{T}$ that $(y_k, y_{k-l_2}, \cdots, y_{k-l_M})$ falls in the interval of the unit M-cube for which the first d_i positions of the binary expansion of y_{k-l_i} is fixed (i.e., in a $2^{-d_1} \times 2^{-d_2} \times \cdots \times 2^{-d_M}$ interval in the unit M-cube) has mean value

$$T = E\,[\hat{T}] = 2^{-(d_1 + \cdots + d_M)}\left(1 + \frac{1}{2^n - 1}\right) + 2^{-n-1}\left(\frac{g(0) - 1}{1 - 2^{-n}}\right) \approx 2^{-(d_1 + d_2 + \cdots + d_M)}$$

for any number N of points $(y_k, y_{k-l_2}, \cdots, y_{k-l_M})$, provided $0 < l_2 < \cdots < l_M < (n/q) - 1$. The variance of $\hat{T}$ is then bounded by

$$var\,[\hat{T}] < \frac{1}{4}\left[\frac{1}{N} + \frac{2}{2^n - 1}\right]\left[1 + \frac{1}{2^n + 1}\right] \approx \frac{1}{4N}.$$

The only difference between these numbers and those of the author's previous article is that here the number of binary digits in each random number is L, and random numbers are spaced q bits apart in $\{a_k\}$. The condition $q \geqq L$ is put in to allow no overlap of digits in the random numbers, and the condition $l_M < (n/q) - 1$ ensures that the maximum distance between any two binary digits of $\{a_k\}$ in the M-tuple $(y_k, y_{k+l_2}, \cdots, y_{k+l_M})$ is less than the degree n of the recursion formula. The former condition gives an uncorrelated, or white, sequence, and the latter produces uniform distribution over the M-cube. This property was called "equidistribution by M's" in Ref. 3.

Proof. We shall use the same notation and type of analysis as the previous work (*SPS 37-27*, Vol. IV, pp. 185–189). The analysis presented originally is valid for the modified sequences as well, and the properties (1), (2), and (3) announced by the theorem follow directly. Consequently, all that need be investigated more fully here is the higher-order distribution property of the sequence. We shall indicate the method which will allow the proof of (3) above to be altered to fit (4).

Consider the distribution of $(y_k, y_{k-l_2}, \cdots, y_{k-l_M})$ where $0 = l_1 < l_2 < \cdots < l_M$. It can be shown that this distribution can be far from uniform if $q(l_M + 1) > n$. For $q(l_M + 1) \leq n$, however, the distribution is uniform over the unit M-cube. To show this is the case, we shall count the relative number of times $(y_k, y_{k-l_2}, \cdots, y_{k-l_M})$ lies in an arbitrary given $2^{-d_1} \times \cdots \times 2^{-d_M}$ interval. Let the initial positions in the binary expansion of y_{k+l_i} be zero. $e_1^i\, e_2^i, \cdots, e_{d_i}^i$ for $i = 1, 2, \cdots, M$.

Since we are considering binary expansions of numbers, intervals of width 2^{-d} are most conveniently considered, and these will surely be sufficient to our needs. This is done efficiently by considering the first d positions of the vectors representing y_k for $k = 1, 2, \cdots, N$, and counting the number of these having a specified pattern. This is equivalent to forming a Boolean function on the first d positions of y_k, whose value is, say -1, if y_k has this initial pattern and $+1$ otherwise. Now define $g(\mathbf{x})$ as follows:

$$g(\mathbf{x}) = \begin{cases} -1, & \text{if } x_{l_i q + j} = e_j^i \\ & \text{for } i = 1, 2, \cdots, M \\ & \text{and } j = 1, 2, \cdots, d_i, \\ +1, & \text{otherwise.} \end{cases}$$

Since $q(l_M + 1) \leq n$, if we let the Boolean function variables be

$$x_t = a_{qk+r-t}$$

then the sequence $\gamma_k = g(a_{qk+r-1}, \cdots, a_{qk+r-n})$ can be expressed as

$$\gamma_k = G(0) + \sum_{s \neq 0} G(s)\, \alpha_{kq+r+v(s)}$$

where $G(s)$ is the Boolean transform of $g(\mathbf{x})$ as in *SPS 37-27*, Vol. IV, pp. 185–189.

The relative number of times $\hat{T}$ that the M-tuple $(y_k, y_{k+l_i}, \cdots, y_{k+l_M})$ lies in that $2^{d_1} \times \cdots \times 2^{d_M}$ interval in M-space where y_{k+l_i} is prefixed by $e_1^i\, e_2^i \cdots e_{d_i}^i$ for $i = 1, 2, \cdots, M$ is

$$\hat{T} = \frac{1}{2N}\left[N - \sum_{k=1}^{N} \gamma_k\right].$$

The method in the original work used to prove property (3) above is now valid, with $d = d_1 + d_2 + \cdots + d_M$, and, therefore, the relative number of times $\hat{T}$ that $(y_k, y_{k-l}, \cdots, y_{k-l_M})$ lies in the specified interval has mean value

$$T = E(\hat{T}) = \left(1 + \frac{1}{p}\right)2^{-(d_1 + \cdots + d_M)} + \frac{1}{2p}[g(0) - 1],$$

and the variance about this mean is bounded by

$$\text{var}(\hat{T}) < \frac{1}{4}\left(1 + \frac{1}{p}\right)\left(\frac{1}{N} + \frac{2}{p}\right).$$

B. A New Construction for Hadamard Matrices

L. Baumert

1. Summary

In this paper a new construction for certain Hadamard matrices is given. This new construction yields, among others, an Hadamard matrix of order 156, of which no previous example was known.

2. Hadamard Matrices and Optimal Codes

The primary problem of space communications is the transmission of information through a channel perturbed by stationary white Gaussian noise. For this channel the optimum receiver uses a correlation detector, and the optimal codes are related to those whose waveforms are as mutually uncorrelated as possible. There are three related classes of codes which most nearly approximate this situation: the regular simplex codes (trans-orthogonal), the orthogonal codes, and the bi-orthogonal codes (JPL *TR 32-67*, "Coding Theory and Its Applications to Communications Systems"), all of which can be constructed from Hadamard matrices. That is, given an Hadamard

matrix there is an associated regular simplex code, an associated orthogonal code, and an associated bi-orthogonal code. Conversely, given a code (simplex, orthogonal, bi-orthogonal) one can easily derive the associated Hadamard matrix. A more detailed discussion of this matter together with applications and an evaluation of the efficiency of these codes may be found in JPL *Technical Report No. 32-67*.

3. Hadamard Matrices

An *Hadamard matrix* H is a square matrix of *ones* and *minus ones* whose row (and hence column) vectors are orthogonal. The order n of an Hadamard matrix is necessarily 1, 2, or $4t$ with $t = 1, 2, 3, \cdots$. It has been conjectured that this condition ($n = 1$, 2, or $4t$) also ensures the existence of an Hadamard matrix. Constructions have been given for particular values of n ($n = 92$ was done in a previous article in Ref. 4 and *RS 36-11*, p. 31), and even for various infinite classes of values. While other constructions exist, those given by *RS 36-11* and Refs. 4 through 10 exhaust the previously known values of n. This paper gives a new construction which yields, among others, the previously unknown value $n = 156$, leaving only two undecided values of $n = 4t \leq 200$ (these are 116 and 188).

4. Williamson Type

An Hadamard matrix is said to be of the *Williamson type* if it has the structure imposed by Williamson (Ref. 9):

$$H = \begin{vmatrix} A & B & C & D \\ -B & A & -D & C \\ -C & D & A & -B \\ -D & -C & B & A \end{vmatrix}$$

where each of A, B, C, D is a symmetric circulatory $t \times t$ matrix. Marshall Hall, Jr. noticed (in connection with the work of *RS 36-11* and Ref. 4) that if a Williamson type matrix exists for $n = 4t$, then an Hadamard matrix (not obviously Williamson) of order $n = 12t$ would exist provided one could find a 12×12 matrix with the following properties: Each row and column must contain precisely three $\pm A$'s, three $\pm B$'s, three $\pm C$'s, three $\pm D$'s and the rows must be formally orthogonal (i.e., A, B, C, D are to be considered as independent quantities).

5. The 12 × 12 Matrix

Fig. 1 displays a 12×12 matrix with the special properties necessary for the multiplication by 3.

$H =$

A	A	A	B	−B	C	−C	−D	B	C	−D	−D
A	−A	B	−A	−B	−D	D	−C	−B	−D	−C	−C
A	−B	−A	A	−D	D	−B	B	−C	−D	C	−C
B	A	−A	−A	D	D	D	C	C	−B	−B	−C
B	−D	D	D	A	A	A	C	−C	B	−C	B
B	C	−D	D	A	−A	C	−A	−D	C	B	−B
D	−C	B	−B	A	−C	−A	A	B	C	D	−D
−C	−D	−C	−D	C	A	−A	−A	−D	B	−B	−B
D	−C	−B	−B	−B	C	C	−D	A	A	A	D
−D	−B	C	C	C	B	B	−D	A	−A	D	−A
C	−B	−C	C	D	−B	−D	−B	A	−D	−A	A
−C	−D	−D	C	−C	−B	B	B	D	A	−A	−A

Fig. 1. The 12 × 12 Williamson extension matrix

Among the known orders of Williamson type matrices (Refs. 4, 9), only the value 52 yields a new value of n by this construction. This gives an Hadamard matrix of order 156. For definiteness the first rows of A, B, C, D for one of the Williamson H_{52}'s are given in Fig. 2.

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	+	+	−	−	+	−	+	+	−	+	−	−	+
B	+	−	−	−	+	+	+	+	+	+	−	−	−
C	+	+	+	−	+	+	−	−	+	+	−	+	+
D	+	+	−	+	−	+	+	+	+	−	+	−	+

Fig. 2. The Williamson H_{52} matrix generators

C. Some Comments Concerning the Integral Equations for Maximum Entropy

A. M. Garsia

1. Introduction

In a recent work (*SPS* 37-25, Vol. IV, pp. 186–194; also to appear in *Annals of Mathematical Statistics* in February 1965), Posner and Rumsey introduced the following

problem. Let $p(x)$ and $q(y)$ be two non-negative functions integrable in $(-\infty, +\infty)$ such that

$$\int_{-\infty}^{+\infty} p(x)\,dx = \int_{-\infty}^{+\infty} q(y)\,dy = 1.$$

A non-negative integrable function $f(x, y)$ is sought which satisfies

$$\int_{-\infty}^{+\infty} f(x, y)\,dy = p(x), \tag{1}$$

$$\int_{-\infty}^{+\infty} f(x, y)\,dx = q(y), \tag{2}$$

and in addition, $f(x, y)$ is required to have the special form

$$f(x, y) = a(x)\,b(y)\,e^{\alpha r(x, y)}, \tag{3}$$

where $r(x, y)$ is a given non-negative measurable function which is almost everywhere finite and $a(x)$, $b(y)$ are unknown, also non-negative functions; α is a constant.

Posner and Rumsey prove that, under certain conditions, this problem admits one and only one solution.

In the present note we shall be concerned with the uniqueness part of the proof. We shall obtain some simplifications in the treatment and a sharpening of their results.

2. Fixing α

The problem can be formulated in a different manner. Find two non-negative measurable functions $a(x)$ and $b(y)$ and a constant α such that

$$a(x)\int_{-\infty}^{+\infty} b(y)\,e^{\alpha r(x, y)}\,dy = p(x) \tag{4}$$

$$b(y)\int_{-\infty}^{+\infty} a(x)\,e^{\alpha r(x, y)}\,dx = q(y) \tag{5}$$

when $r(x, y)$, $p(x)$, $q(y) \geqq 0$ are given functions such that

$$\int_{-\infty}^{+\infty} p(x)\,dx = \int_{-\infty}^{+\infty} q(y)\,dy = 1.$$

The question of uniqueness has two different aspects. First of all, we do not expect that Eqs. (4) and (5) are sufficient to determine $a(x)$ and $b(y)$ without some further condition. The cited work suggests that there are two sets of natural side conditions:

(1) To fix the value of α.

(2) To fix the value of the integral

$$\mathcal{J} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} a(x)\,b(y)\,e^{\alpha r(x, y)}\,r(x, y)\,dx\,dy.$$

The first of these conditions is the simplest one, and in this case we shall be able to show uniqueness without further assumptions. The second is clearly more complex, for in order to formulate it we have to require the existence of the integral in question. To prove uniqueness in this case we shall need further assumptions. It would be interesting to see if our assumptions can be further reduced.

We shall proceed to work under Condition (1). To this end we shall need the following:

Lemma. Two distribution densities $f(x, y)$ and $g(x, y)$ which have the same marginals and a degenerate ratio (function of x times function of y) are identical.

Proof. We are assuming that for almost all x and y,

$$\int_{-\infty}^{+\infty}[f(x, y) - g(x, y)]\,dy = \int_{-\infty}^{+\infty}[f(x, y) - g(x, y)]\,dx = 0, \tag{6}$$

and in addition, that for some almost everywhere positive functions $\alpha(x)$, $\beta(y)$ we have

$$\beta(y)\,f(x, y) = \alpha(x)\,g(x, y) \qquad \text{a.e.} \tag{7}$$

Let then $\theta(u)$ denote a function which increases strictly from 0 to 1 as u varies from $-\infty$ to $+\infty$. Note for almost all x and y that $f > g$ if and only if $\alpha(x) > \beta(y)$. Consequently, we shall have

$$[f(x, y) - g(x, y)]\,[\theta(\alpha(x)) - \theta(\beta(y))] \geqq 0.$$

On the other hand, by Eq. (6), we obtain

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[f(x, y) - g(x, y)]\,\theta(\alpha(x))\,dx\,dy =$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[f(x, y) - g(x, y)]\,\theta(\beta(y))\,dx\,dy = 0.$$

Thus, the assertion necessarily follows.

We therefore obtain:

Theorem b.1. The integral equations (4) and (5) for fixed α have a unique solution.

Proof. Suppose that $f(x,y) = a(x)\,b(y)\,e^{\alpha r(x,y)}$ and $g(x,y) = \tilde{a}(x)\,\tilde{b}(y)\,e^{\alpha r(x,y)}$ are both solutions of Eqs. (4) and (5); we can suppose without loss that $r(x,y)$ is everywhere finite. This implies that the set on which $f > 0$ is a product set $P = E \times F$, where

$$E = \{x: a(x) > 0\} \quad \text{and} \quad F = \{y: b(y) > 0\}.$$

For almost all $x \in E$, we have

$$\int_{-\infty}^{+\infty} f(x,y)\,dy = a(x) \int_{-\infty}^{+\infty} b(y)\,e^{\alpha r(x,y)}\,dy > 0.$$

In fact, if it were zero, we would obtain $b(y) = 0$ almost everywhere, and we could not have

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1.$$

Thus, for almost all $x \in E$ we have $\tilde{a}(x) > 0$ as well. We can thus suppose without loss that

$$E = \{x: a(x) > 0\} = \{x: \tilde{a}(x) > 0\},$$

$$F = \{y: b(y) > 0\} = \{y: \tilde{b}(y) > 0\}.$$

Now define

$$\alpha(x) = \begin{cases} a(x)/\tilde{a}(x) & \text{for } x \in E, \\ \text{otherwise } 1; \end{cases}$$

$$\beta(y) = \begin{cases} \tilde{b}(y)/b(y) & \text{for } y \in F, \\ \text{otherwise } 1. \end{cases}$$

We then clearly have Eq. (7); thus, the hypotheses of the lemma are fulfilled and we must have

$$f(x,y) = g(x,y) \quad \text{a.e.}$$

However, this implies that for almost all $x \in E$ and almost all $y \in F$

$$a(x) = \tilde{a}(x),$$
$$b(y) = \tilde{b}(y).$$

3. Fixing the r(x, y)-Integral

To formulate the second uniqueness theorem, we shall assume that $e^{\alpha r(x,y)}$ itself is not degenerate (i.e., the product of a function of x by a function of y), and in addition require that the integrals

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,r(x,y)\,dx\,dy, \tag{8}$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,|\log a(x)|\,dx\,dy, \tag{9}$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,|\log b(y)|\,dx\,dy, \tag{10}$$

when $f(x,y) = a(x)\,b(y)\,e^{\alpha r(x,y)}$ are all finite. The result can be stated as follows.

Theorem 2. If the integral $\mathcal{J}$ is to take a specified value, there is at most one couple $a(x)$, $b(y)$ and only one α for which the integrals in Eqs. (8), (9), and (10) are finite and Eqs. (4) and (5) are satisfied.

Proof. Let then (if possible) $a(x)$, $b(y)$; $\tilde{a}(x)$, $\tilde{b}(y)$ be two such couples and let the functions

$$f(x,y) = a(x)\,b(y)\,e^{\alpha r(x,y)},$$

$$g(x,y) = \tilde{a}(x)\,\tilde{b}(y)\,e^{\tilde{\alpha} r(x,y)},$$

be solutions of Eqs. (1) and (2). Note then that the mixed integrals

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,|\log \tilde{a}(x)|\,dx\,dy,$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\,|\log a(x)|\,dx\,dy,$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,|\log \tilde{b}(y)|\,dx\,dy,$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\,|\log b(y)|\,dx\,dy,$$

are also finite. It will be sufficient to show this fact only for the first of these integrals.

By our assumptions we have

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,|\log \tilde{a}(x)|\,dx\,dy$$

$$= \int_{-\infty}^{+\infty} |\log \tilde{a}(x)| \int_{-\infty}^{+\infty} f(x,y)\,dy\,dx$$

$$= \int_{-\infty}^{+\infty} |\log \tilde{a}(x)| \int_{-\infty}^{+\infty} g(x,y)\,dy\,dx.$$

Thus, the assertion follows from Fubini's theorem. The consequences are then that the integrals

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f|\log g|\,dx\,dy,$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g|\log f|\,dx\,dy$$

are both finite. Now note that

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log f\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log a(x)\,dx\,dy$$

$$+ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log b(y)\,dx\,dy$$

$$+ \alpha\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} fr(x,y)\,dx\,dy.$$

However, by our assumptions we get

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log a(x)\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g\log a(x)\,dx\,dy,$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log b(y)\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g\log b(y)\,dx\,dy,$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} fr(x,y)\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} gr(x,y)\,dx\,dy.$$

Consequently, we shall have

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log f\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g\log f\,dx\,dy$$

and similarly

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f\log g\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g\log g\,dx\,dy.$$

Subtracting we then get

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (f-g)\log(f/g)\,dx\,dy = 0. \qquad (11)$$

Since the integrand in the equation is non-negative and vanishes only when $f = g$, we must then have $f = g$ almost everywhere. By an argument similar to that carried out in the proof of Theorem 1 (and using the assumption that $e^{\alpha r(x,y)}$ is not degenerate), we obtain that

$$a(x) = \tilde{a}(x) \qquad \text{a.e.}$$

$$b(y) = \tilde{b}(y) \qquad \text{a.e.,}$$

and $\alpha = \tilde{\alpha}$.

Remark. Since the equality in Eq. (11) implies $f = g$, one would expect that when the integral in Eq. (11) is small, in some sense f must be close to g. In fact, we can estimate the L_1 norm of $f - g$ in terms of the integral equation (11). We have indeed

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |f-g|\,dx\,dy$$

$$\leq 2^{1/2}\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (f-g)\log(f/g)\,dx\,dy\right]^{1/2}.$$

Proof. We start by observing that for $|x| \leq 1$

$$\left|\log\frac{1+x}{1-x}\right| \geq |x|.$$

Thus, setting $w = f - g$, $v = f + g$, we get

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (f-g)\log(f/g)\,dx\,dy =$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |w|\left|\log\frac{1+\frac{w}{v}}{1-\frac{w}{v}}\right| dx\,dy$$

$$\geq \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{|w|^2}{v}\,dx\,dy \geq \frac{1}{2}\left(\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|w|\,dx\,dy\right)^2$$

the last step following from Schwartz's inequality. This establishes Eq. (12).

D. Asymptotic Behavior of Stirling Numbers

L. Harper

1. Summary

The Stirling Numbers $\{\sigma_n^j\}$ of the Second Kind are asymptotically normal. This result is similar to results achieved by Feller (Ref. 11) and Goncărov (Ref. 12) for other combinatorial distributions. Here the technique of proof is different; one of the most general forms of the Central Limit Theorem is used.

Interesting qualitative information about the Stirling Numbers is also obtained from this result. Asymptotic estimates on the value of $\max_j \{\sigma_n^j\}$ are given.

2. Introduction

Mathematicians have been aware for quite awhile that probability theory has combinatorial applications. The classical De Moivre–La Place theorem, for example, can be interpreted as a theorem about binomial coefficients, the binomial coefficients being the solution to the difference equation

$$A_{nj} = A_{n-1,j} + A_{n-1,j-1}$$

with the boundary conditions

$$A_{0j} = \begin{cases} 1, & j = 0 \\ 0, & j \neq 0 \end{cases}.$$

Feller (Ref. 11, p. 241) uses more general versions of the Central Limit Theorem to show that the distributions B_{nj}, the number of permutations of n elements with j inversions, and C_{nj}, the number of permutations of n elements with j cycles, are asymptotically normal. There he defines random variables on the set of all permutations to count either inversions or cycles. He shows that these random variables are independent and satisfy a condition known as the "Lindeberg Condition" (Ref. 13, p. 239), and thus have asymptotically normal distributions.

It can be verified that $B_{n,j}$ and $C_{n,j}$ are the solutions of the difference equations

$$B_{n,j} = \sum_{k=M}^{j} B_{n-1,k}, \qquad M = \max(0, j - n + 1),$$

and

$$C_{n,j} = (n-1)\,C_{n-1,j} + C_{n-1,j-1}$$

respectively, with boundary conditions

$$B_{0j} = \begin{cases} 1, & j = 0 \\ 0, & j \neq 0 \end{cases}$$

and

$$C_{0,j} = \begin{cases} 1, & j = 0 \\ 0, & j \neq 0 \end{cases}.$$

In view of the similarity of these three results, it seems appropriate to call any "generalized Pascal's triangle" on the lattice points of the positive quadrant, defined by a difference equation with the common boundary condition above, a *combinatorial distribution*.

V. Goncărov (Ref. 12) has shown asymptotic normality for B_{nj}, C_{nj} and other combinatorial distributions by a different and less elegant method. By brute force he tortuously manipulates the characteristic functions of the distributions until they approach $\exp(-x^2/c)$, c a positive constant. His attack is certainly the most general conceptually, but hardly the most efficient, and not even feasible in cases where there is lack of knowledge about the characteristic functions involved.

This paper presents a program for showing asymptotic normality of combinatorial distributions, much like Feller's in that it gives the problem a probabilistic interpretation, and uses the Central Limit Theorem in an essential way.

3. The Stirling Numbers of the Second Kind

The Stirling Numbers of the Second Kind are combinatorially distributed by the following difference equation:

$$\sigma_n^j = j\sigma_{n-1}^j + \sigma_{n-1}^{j-1}.$$

Several preliminary lemmas are needed.

Lemma 1. If

$$P_n(x) = \sum_{j=0}^{n} \sigma_n^j x^j,$$

then the roots of P_n are real, distinct and non-positive for all $n = 1, 2, \cdots$.

Proof. By induction: $P_0(x) = 1$, so the statement is vacuously true for $n = 0$; for other values of n,

$$P_n(x) = \sum_{j=0}^{n} \sigma_n^j x^j = \sum_{j=0}^{n} j\sigma_{n-1}^j\, x^j + \sum_{j=0}^{n} \sigma_{n-1}^{j-1}\, x^j$$

$$= x\left[\sum_{j=0}^{n-1} j\sigma_{n-1}^j\, x^{j-1} + \sum_{j=0}^{n-1} \sigma_{n-1}^j\, x^j\right]$$

$$= x\left[\frac{dP_{n-1}(x)}{dx} + P_{n-1}(x)\right].$$

Therefore, $P_1(x) = x$, and $P_2(x) = x(1 + x) = x + x^2$, so the statement still holds for $n = 1, 2$. Now suppose $n \geqq 2$. By hypothesis, P_n has n distinct real roots. Let x_1, x_2 be two of them which are consecutive: that is, $P_n(x_1) = 0$, $P_n(x_2) = 0$, and $P_n(x) \neq 0$ if $x_1 < x < x_2$. Since the roots are distinct, $dP_n(x_1)/dx$, $dP_n(x_2)/dx$ must be of opposite signs. Assume, without loss in generality, that

$$\frac{dP_n(x_1)}{dx} > 0, \qquad \frac{dP_n(x_2)}{dx} < 0.$$

Now since $f(x) = P_n(x) + dP_n(x)/dx$ is continuous on the interval $[x_1, x_2]$, and $f(x_1) > 0$, $f(x_2) < 0$, then by Bolzano's theorem there then exists $x_0, x_1 < x_0 < x_2$ such that $f(x_0) = 0$. Therefore, $P_{n+1}(x_0) = x_0 f(x_0) = 0$.

Thus, between any consecutive pair of roots of P_n, we have found a root of P_{n+1}. This gives us $n - 1$ distinct, negative roots for P_{n+1}. The number zero is obviously another root.

Now note that if $H_n = P_n e^x$, then

$$H_{n+1}(x) = x\frac{dH_n(x)}{dx},$$

and furthermore, the zeros of H_n are exactly those of P_n. But $H_n(-\infty) = 0$, so there must be a turning point x_t to the left of the most negative zero x_0 of H_n. Thus,

$$H_n(x_0) = 0,$$

$$H_n(x) \neq 0 \text{ if } x < x_0;$$

$$\frac{dH_n(x_t)}{dx} = 0, x_t < x_0.$$

Therefore,

$$H_{n+1}(x_t) = 0 = P_{n+1}(x_t).$$

Consequently, we have found $(n + 1)$ real, distinct, nonpositive roots for P_{n+1}, $n > 2$. Since the degree of P_{n+1} is exactly $n + 1$, we have found all the roots and completed the proof of Lemma 1.

4. Bell Numbers

The sum

$$B_n = \sum_{j=0}^{n} \sigma_n^j$$

is called the Bell Number of order n. We now show

Lemma 2.

$$\frac{B_{n+2}}{B_n} - \left(\frac{B_{n+1}}{B_n}\right)^2 \to \infty \text{ as } n \to \infty.$$

Proof. By the classical formula of Dobinski (Ref. 14),

$$B_n = \sum_{k=0}^{\infty} \frac{k^n}{k!},$$

so that the sequences

$$\left\{\frac{k^n}{B_n k!}\right\}_{k=0}^{\infty}, \qquad n = 0, 1, 2, \cdots,$$

can be considered as density functions of the random variables T_n; viz.,

$$\Pr[T_n = k] = \frac{k^n}{B_n k!}.$$

Then

$$E(T_n) = \sum_{k=0}^{\infty} k\frac{k^n}{B_n k!} = \frac{B_{n+1}}{B_n},$$

$$\mathrm{Var}^2(T_n) = E(T_n^2) - E^2(T_n)$$

$$= \sum_{k=0}^{\infty} k^2 \frac{k^n}{B_n k!} - \left(\frac{B_{n+1}}{B_n}\right)^2$$

$$= \frac{B_{n+2}}{B_n} - \left(\frac{B_{n+1}}{B_n}\right)^2.$$

One way to assure that the variance of a sequence of random variables approaches infinity is to show that the density functions of the random variables uniformly approach zero. To prove this, we will use the relation (Ref. 11, p. 52)

$$m! > (2\pi)^{1/2} m^{m+1/2} \exp\left(m + \frac{1}{12m+1}\right).$$

Also, from Ref. 15, we have

$$B_n \sim (R+1)^{-\frac{1}{2}} \exp\left[n(R+R^{-1}-1)-1\right] \times\left(1-\frac{R^2(2R^2+7R+10)}{24n(R+1)^3}\right),$$

where R is the unique real solution of the transcendental equation $Re^R = n$.

Using the inequality above, we have

$$\frac{j^n}{j!} < \frac{j^n}{(2\pi)^{\frac{1}{2}} j^{j+\frac{1}{2}} \exp -\left(j+\frac{1}{12j+1}\right)} = f(j), \text{ say.}$$

Then if we take logarithms and replace j by the continuous variable x,

$$\ln f(x) = n\ln(x) - \tfrac{1}{2}\ln(2\pi) - (x+\tfrac{1}{2})\ln(x) + x - \frac{1}{12x+1}, \quad \text{for } x>0.$$

Therefore,

$$\frac{1}{f(x)}f'(x) = \frac{n}{x} - \frac{x+\frac{1}{2}}{x} - \ln x + 1 + \frac{12}{(12x+1)^2} = \frac{n}{x} - \frac{1}{2x} - \ln x + \frac{12}{(12x+1)^2}.$$

It is clear then that as $n\to\infty$ the solutions x_n' of the equation $f'(x)/f(x) = 0$ go to infinity. Thus, $x_n' - x_n \to 0$ where $x_n \ln x_n = n$. Consequently,

$$\frac{\sup_j \frac{j^n}{j!}}{eB_n} < \frac{\sup_j \frac{j^n}{(2\pi)^{\frac{1}{2}} j^{j+\frac{1}{2}} \exp -(j+[1/(12j+1)])}}{eB_n} \sim \frac{x_n^n}{(2\pi)^{\frac{1}{2}} x_n^{x_n+\frac{1}{2}} e^{-x_n+1} B_n} \sim \frac{x_n^n (R+1)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}} x_n^{x_n+\frac{1}{2}} e^{-x_n+1} \exp\left[n(R+R^{-1}-1)-1\right]} = \left(\frac{R+1}{2\pi x_n}\right)^{\frac{1}{2}} \to 0.$$

Lemma 2 is proved.

5. Main Theorem

We are now in a position to prove the main theorem.

Theorem. The Stirling Numbers of the Second Kind are asymptotically normal in the sense that

$$\sum_{j=1}^{x_n} \sigma_n^j \sim \frac{B_n}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{t} e^{-x^2/2}\, dx,$$

where

$$x_n = \left\{\frac{B_{n+2}}{B_n} - \left(\frac{B_{n-1}}{B_n}\right)^2 - 1\right\}^{\frac{1}{2}} t + \left\{\frac{B_{n-1}}{B_n} - 1\right\}.$$

Proof. The result is an application of the "Bounded Variance Normal Convergence Criterion" in Ref. 14, p. 295. It is stated: "Let the independent summands $\{X_{nk}\}_{k=1}^{n_n}$, centered at expectations, be such that $\Sigma \operatorname{Var}^2(X_{nk}) = 1$ for all n. Let F_{nk} be the distribution function of X_{nk}. Then

$$S_n = \sum_k X_{nk}$$

converges normally with mean zero, unit variance and

$$\left(\max_k \operatorname{Var}(X_{nk})\right) \to 0,$$

if and only if: for all $\varepsilon > 0$,

$$\lim_{n\to\infty} g_n(\varepsilon) = \lim_{n\to\infty} \sum_k \int_{|x|\geq\varepsilon} x^2\, dF_{nk} = 0.$$

In order to use this theorem, we need a "hat" from which to pull a double sequence of independent (for fixed n) random variables. The hat is Lemma 1. That the roots of the polynomial P_n are real and non-positive is equivalent to the fact that P_n can be factored into linear terms with real non-negative coefficients. If we normalize each of these terms suitably, we see that the distribution whose density function is $\{\sigma_n^j/B_n\}_{j=0}^n$ is the distribution of a sum of independent random variables taking on only the values zero and one. If $-x_{nj}$ is a root of P_n, then define the random variable X_{nk}' by

$$\Pr[X_{nk}' = y] = \begin{cases} \dfrac{x_{nk}}{1+x_{nk}}, & \text{if } y = 0; \\ \dfrac{1}{1+x_{nk}}, & \text{if } y = 1. \end{cases}$$

Letting

$$S_n' = \sum_k X_{nk}',$$

we have

$$E(S_n') = \sum_{j=0}^{n} j\frac{\sigma_n^j}{B_n} = \frac{B_{n+1}}{B_n} - 1$$

and

$$\mathrm{Var}\,(S'_n) = \sum_{j=0}^{n} j^2 \frac{\sigma_n^j}{B_n} - \left(\frac{B_{n+1}}{B_n} - 1\right)^2$$

$$= \frac{B_{n+2}}{B_n} - \left(\frac{B_{n+1}}{B_n}\right)^2 - 1.$$

Thus, by Lemma 2, $\mathrm{Var}\,(S'_n) \to \infty$ as $n \to \infty$.

Now we normalize and let

$$S_n = \frac{S'_n - E\,(S'_n)}{\mathrm{Var}\,(S'_n)} = \sum_k \frac{X'_{nk} - E\,(X'_{nk})}{\mathrm{Var}\,S_n} = \sum_k X_{nk}.$$

Since $0 \leq X'_{nk} \leq 1$ and $-1 \leq X'_{nk} - E\,(X'_{nk}) \leq 1$, and since given $\varepsilon > 0$ there exists N such that $|X_{nk}| < \varepsilon$ for all $n \geq N$, we conclude:

$$\lim_{n\to\infty} g_n(\varepsilon) = \lim_{n\to\infty} \sum_k \int_{|x|\geq\varepsilon} x^2\, dF_{nk} = 0, \quad \text{for all } n \geq N.$$

Thus, the hypotheses of the Normal Convergence Criterion are fulfilled. This finally proves the main theorem.

6. Corollaries

The theorem gives us interesting qualitative information about the Stirling Numbers of the Second Kind which we state here as corollaries, but will not prove.

Corollary 1.

$$S_n^{x_n} \sim \frac{B_n}{(2\pi)^{1/2}}\, e^{-x^2/2},$$

where

$$B_n = \sum_j \sigma_n^j, \; x_n = x\left(\frac{B_{n+2}}{B_n} - \left[\frac{B_{n+1}}{B_n}\right]^2 - 1\right)^{1/2} + \left(\frac{B_{n+1}}{B_n} - 1\right).$$

Corollary 2. Given $\varepsilon = 0$, there exists an N such that $n \geq N$ implies

$$\left| J_n - \left(\frac{B_{n+1}}{B_n} - 1\right)\right| < \varepsilon\left(\frac{B_{n+2}}{B_n} - \left[\frac{B_{n+1}}{B_n}\right]^2 - 1\right)^{1/2},$$

where J_n is defined as that integer j such that $\sigma_n^j = \max_j \sigma_n^j$. Thus,

$$J_n \sim \left(\frac{B_{n+1}}{B_n} - 1\right) \text{ and } \max_j \sigma_n^j \sim \sigma_n^{[(B_{n+1}/B_n)-1]}.$$

E. Optimal Communication Nets

R. J. McEliece

1. Summary

Recently Kleinrock (Ref. 16) has investigated the problem of finding an optimal (with respect to certain reasonable criteria) configuration of communication levels between a given set of terminals. In this paper, a graph-theoretic quantity related to Kleinrock's problem is introduced and studied. In the first part of the discussion, this quantity is minimized under certain conditions. In the second part, certain estimates on the quantity are derived under much more general circumstances. Throughout the paper, it is seen that the "star" configuration (Fig. 3) is either optimal or very close to optimal.

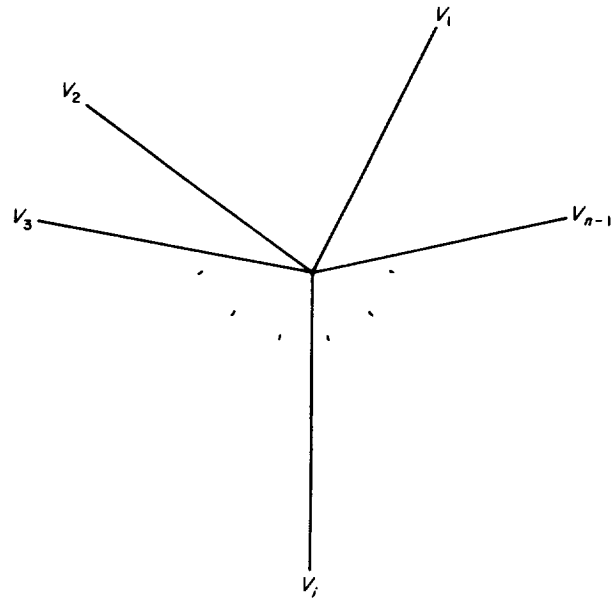


Fig. 3. The star S_{n-1}

2. Introduction

In a recent article (Ref. 16), Kleinrock has considered the following problem: given a set of n terminals, and a pattern of the communication traffic between them (and certain other constraints), what is the best possible configuration of communication links between the terminals? Here "best possible" has been interpreted to mean the configuration which minimizes the mean time a message is in the communication net. Kleinrock's discussion has suggested the following graph-theoretic investigations.

If we define, as is customary, the *distance* $\mu(a,b)$ between two vertices a and b of a graph as the minimum number of edges in a path which joins the two vertices, then we may speak of the *average separation* in a finite undirected connected graph; this is naturally defined by

$$\bar{\lambda} = \frac{\sum_{(a,b)} \mu(a,b)}{\binom{n}{2}}. \tag{1}$$

The summation is taken over all unordered pairs of vertices (a,b) where $a \neq b$, and $n = |G|$ is the number of vertices of G (frequently called the *order* of G). In this discussion, we shall always assume that G is connected and undirected. The problem of finding the minimum of this quantity with respect to all graphs on n vertices is a trivial one; $\bar{\lambda} = 1$ when and only when G is the complete graph U_n on n vertices. (The problem of maximizing $\bar{\lambda}$ is less easy, but the answer is that $\bar{\lambda}_{max} = (n+1)/2$, attained by the chain L_n to be defined below.) But we get the feeling that U_n "uses too many edges" in attaining the minimum, and is therefore in some sense inefficient. This leads us to consider the quantity $m\bar{\lambda}$, where m is the number of edges in G. Here we will consider the problem of minimizing the quantity $m\bar{\lambda}$ (for a fixed n) under various circumstances.

First note that in Ref. 16, $\bar{\lambda}$ was computed with the assumption that all paths were to be given equal weight (in this case that weight was $2/n(n-1)$. In the first path of this discussion we will retain that assumption; we will relax it later.

It now seems appropriate to compute the quantity $m\bar{\lambda}$ for several of the simplest graphs:

(1) The *complete graph* U_n on n vertices:

$$m\bar{\lambda} = \binom{n}{2}\frac{\binom{n}{2}}{\binom{n}{2}} = \frac{n(n-1)}{2} \sim \frac{n^2}{2}.$$

(2) The *chain* L_n (the graph formed by joining the vertices V_i to V_j if and only if $|i-j| = 1$):

$$m\bar{\lambda} = (n-1)\frac{\sum_{n \geq i > j \geq 1}(i-j)}{\binom{n}{2}} = \frac{2}{n} \times \frac{n}{6}(n^2-1) = \frac{n^2-1}{3} \sim \frac{n^2}{3}.$$

(3) The *ring* R_n (here the edges are $V_1V_2, V_2V_3, \cdots, V_{n-1}V_n, V_nV_1$): Here it can be shown that

$$m\bar{\lambda} = \begin{cases} \dfrac{n^3}{4(n-1)}, & n \text{ even} \\ \dfrac{n(n+1)}{4}, & n \text{ odd} \end{cases} \sim \frac{n^2}{4}.$$

(4) The *star* S_{n-1} (Fig. 3):

$$m\bar{\lambda} = (n-1)\frac{(n-1) + 2(n(n-1)/2 - (n-1))}{n(n-1)/2} = \frac{2(n-1)^2}{n} \sim 2n.$$

In the above examples, we see that only for the star S_{n-1} does the quantity $m\bar{\lambda}$ behave linearly in n; for the rest it grows as n^2. (In fact, we see that while the complete graph U_n minimizes $\bar{\lambda}$, it exhibits the *maximum* $m\bar{\lambda}$ of all the examples we have given.) We now proceed to prove that S_{n-1} does, in fact, minimize the quantity $m\bar{\lambda}$, and that no other graph on n vertices does as well.

3. Results in the Case of Equal Weights

Theorem 1. If G has n vertices and m edges,

$$m\bar{\lambda} \geq 2m\left(1 - \frac{m}{n(n-1)}\right).$$

Proof. For a graph with m edges, exactly m of the distances between vertices will be 1. Hence, the remaining

$$\binom{n}{2} - m$$

paths must each be of length 2 or greater. Thus,

$$m\bar{\lambda} \geq m\,\frac{m + 2\left[\binom{n}{2} - m\right]}{\binom{n}{2}} = 2m\left(1 - \frac{m}{n(n-1)}\right).$$

This completes the proof.

In order to now obtain a *lower* bound on $m\bar{\lambda}$ which is independent of m (depends only on n), we need to examine the expression $2m[1 - m/n(n-1)]$ as a function of m. The graph of this can be seen to be a parabola with maximum at $m = n(n-1)/2$, (the maximum is also $n(n-1)/2$) and since $m \leq n(n-1)/2$, the quantity $2m(1 - m/n(n-1))$ will be minimized if m is as small

as possible. But it can be shown (e.g., Ref. 17, p. 152) that $m \geqq n-1$ if the graph G is connected. Consequently, we have proved:

Theorem 2. If G has n vertices, then

$$m\bar{\lambda} \geqq \frac{2(n-1)^2}{n}.$$

This theorem, coupled with Example (4), shows that the star S_{n-1} minimizes the quantity $m\bar{\lambda}$. Now let us suppose that for some graph G, $m\bar{\lambda} = 2(n-1)^2/n$. Then it is clear from what has been said above that G has $n-1$ edges, and furthermore that all distances are either 1 or 2. In Ref. 17, p. 153, it is shown that a graph on n vertices with $n-1$ edges has at least two "pendant" vertices. (A pendant vertex is one which has only one edge incident to it.) Let one of these vertices be denoted by A, and let the other endpoint of the edge incident to A be denoted by K. Now let B be any other vertex of the graph. Then B must be connected to K with an edge, since otherwise the distance from A to B would be greater than 2. But the choice of B was arbitrary. Therefore, all vertices of G are joined to K, and since G has only $n-1$ edges, all edges are of the form V_iK. We have therefore proved:

Theorem 3. The bound of the Theorem 2 is attained if and only if $G = S_{n-1}$.

In view of this result, it might be expected that a similar result would be true for the bound of Theorem 1: if $m = n - 1 + \nu$, $\nu > 0$, then the bound of Theroem 1 is attained only for graphs G which have been obtained by adding ν arbitrary edges to S_{n-1}. But although the bound of theorem is attained for all such graphs, the bound can also be attained by others, as illustrated in Figs. 4 and 5 for the case $n = 5$, $\nu = 1$. Clearly the graph in Fig. 5 cannot be transformed into S_4 by the removal of an edge. But there is a "partial uniqueness theorem" analogous to Theorem 3.

Theorem 4. If for a graph G we write $\nu = m - n + 1$, *and if* $n > 2\nu + 3$, *then the bound of Theorem 1 is attained only for graphs G which have been obtained from* S_{n-1} *by the addition of* ν *(arbitrary) edges. (The number* $\nu = \nu(G)$ *is sometimes called the cyclomatic number of G.)*

Proof. From the proof of Theorem 1, it is easy to see that the bound is attained only if all distances in the graph are either 1 or 2.

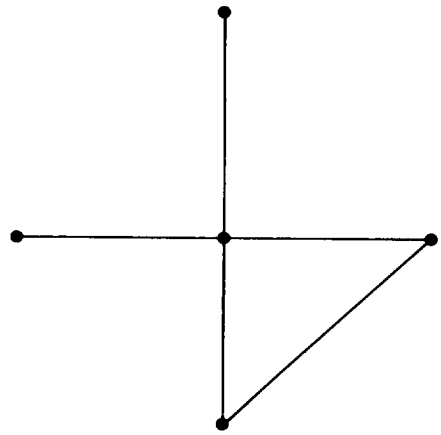


Fig. 4. A "star-like" graph attaining the bound of Theorem 1

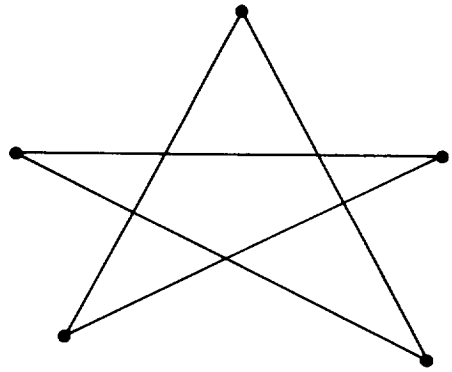


Fig. 5. A graph attaining the bound of Theorem 1 which is not "star-like"

Following the notation of Berge (Ref. 17), let $|\Gamma A|$ represent the number of edges incident to the vertex A. Then

$$\sum_{A \in G} |\Gamma A| = 2m = 2(n - 1 + \nu),$$

since an edge incident to A_1 and A_2 is counted in $|\Gamma A_1|$ and $|\Gamma A_2|$. Consequently, the *average* number of edges incident to a vertex is $2 + (2(\nu - 1)/n)$. When $2(\nu - 1)/n < 1$, i.e., when $n > 2\nu - 2$, the average is less than 3, and so there is at least one vertex A such that $|\Gamma A| \leqq 2$. If $|\Gamma A| = 1$, A is a pendant vertex, and we may then conclude that G is of the required type by a modification of the proof given for Theorem 3.

If, now $|\Gamma A| = 2$, let B and C be the (distinct) vertices of G which are joined to A by an edge. If K is any other vertex of the graph (which is of course not joined

to A by an edge) then K must be joined to either B or C by an edge (or both), since otherwise the distance from A to K would be greater than 2. This shows $|\Gamma A| + |\Gamma B| + |\Gamma C| \geq n + 1$, and so the average number of vertices incident to the remaining (we assume $n > 3$) vertices is $(1 + 2\nu)/(n - 3)$. When this is less than 2, i.e., when $n > 2\nu + 3$, there must be at least one vertex K such that $|\Gamma K| = 1$. But we have seen that this implies that G is of the required type. Theorem 4 is proved. Note here that if $\nu = 0$, Theorem 4 tells us that for $n > 3$, S_{n-1} is the only graph which attains the minimum value of Theorem 2. The cases $n = 1, 2, 3$ are easily disposed of, so that Theorem 4 gives an alternate proof of Theorem 3. With $\nu = 1$, Theorem 4 also shows that no graph "larger" than that of Fig. 5 can attain the bound of Theorem 2 unless it has a star subgraph.

4. Arbitrary Weights

We now proceed to the more general case, where in Ref. 16 we assign a weight $w_{ij} > 0$ to each path:

$$m\bar{\lambda} = \sum w_{ij}\mu_{ij}.$$

Here we normalize $\sum w_{ij} = 1$ (summations are taken over all unordered pairs (i, j) where $i \neq j$), and μ_{ij} denotes the distance between the i^{th} and j^{th} vertices. Here the problem becomes a more realistic one; the w_{ij} may be considered as measures of the traffic between terminals, and in general the traffic is not the same between each pair of terminals. In our previous discussion, of course, we set $w_{ij} = 2/n(n-1)$ for all i, j. To facilitate the discussion which is to follow, let us now renumber the w_{ij}'s (and the μ_{ij}'s correspondingly) with a single subscript so that $w_1 \geq w_2 \geq \cdots \geq w_{N-1} \geq w_N$ (here $N = n(n-1)/2$). This renumbering may often be accomplished in several ways, in the general case.

Theorem 5. If the connected graph G has n vertices and m edges, then

$$m\bar{\lambda} \geq m\left(1 + \sum_{k=m+1} w_k\right).$$

This bound is a minimum for $m = n - 1$. Consequently, for a given set of weights $\{w_i\}$,

$$m\bar{\lambda} \geq (n-1)\left(1 + \sum_{k=n}^{N} w_k\right).$$

Proof. As in the proof of Theorem 1, notice that in G there are exactly $m\mu$'s equal to 1, and so the $N - \bar{\lambda}$ remaining μ's must be 2 or greater. To minimize $\bar{\lambda}$, we can do no better than to have the m greatest w's correspond to the μ's which are 1, and the $N - m$ remaining w's correspond to the μ's which are 2. Thus,

$$m\bar{\lambda} \geq m\left(\sum_{k=1}^{N} w_k + 2\sum_{k=m+1}^{N} w_k\right) = m\left(1 + \sum_{k=m+1}^{N} w_k\right).$$

Let us now attempt to find the minimum of the expression

$$F(m) = m\left(1 + \sum_{k=1}^{N} w_k\right)$$

with respect to m. We have

$$F(m-1) = (m-1)\left(1 + \sum_{k=m}^{N} w_k\right)$$

$$= m + m\sum_{k=m}^{N} w_k - 1 - \sum_{k=m}^{N} w_k,$$

$$F(m-1) = m\left(1 + \sum_{k=m+1}^{N} w_k\right) + \left(mw_m - 1 - \sum_{k=m}^{N} w_k\right)$$

$$= F(m) + \left(mw_m - 1 - \sum_{k=m}^{N} w_k\right).$$

But since $w_1 + w_2 + \cdots + w_m + \cdots + w_N = 1$, $w_1 \geq w_2 \geq \cdots \geq w_N \geq 0$, we have $mw_m \leq 1$ and so

$$mw_m - 1 - \sum_{k=m}^{N} w_k \leq 0.$$

Actually, this is a strict inequality, since if $mw_m = 1$, $w_m > 0$ and so

$$\sum_{k=m}^{N} w_k > 0.$$

Consequently, $F(m)$ is a *decreasing* function of m, thus minimized when m is as small as possible. But we have seen that if G is connected, $m \geq n - 1$. Theorem 5 is proved.

Let us now define $e_s(G)$ to be the minimum of $m\lambda$ taken over *all* connected graphs G with respect to a given set $S = \{w_{ij}\}$ of weights. Although for $w_{ij} = 2/n(n-1)$ we have seen that the bound of Theorem 5 is always attained, this is not the case generally. In fact, it is easy to see that the second (m-independent) bound of Theorem 5 is only attained *by a star* if the $n-1$ pairs $(i_1, j_1), \cdots, (i_{n-1}, j_{n-1})$ corresponding to the $n-1$ largest weights all share a common coordinate. Here, of course, if we make this common point the center of a star S_{n-1}, the bound is attained.

For example, in Fig. 6 if we assign the weight $AB = CD = 0.4$, $AC = BD = 0.1$, $AD = BC = 0$, the bound of Theorem 3 gives $e_s(G) \geqq 3.3$; but it is relatively easy to see that the best possible configuration gives $e_s(G) = 4.5$ (choose the star formed by AB, AC, AD).

Next, we might be tempted to conjecture that although the bound of Theorem 5 is not always attained, the best possible graph is always a star; but this is not true: for let $AB = AC = BD = CD = 0.1$, $AD = BC = 0.03$ in Fig. 6. Here the bound of Theorem 5 is 3.9, all stars have the same value of $m\bar{\lambda} = 4.5$, but the graph formed by the edges AD, AB, and BC has $m\bar{\lambda} = 4.2$.

Finally, we might hope that the best configuration always is attained by a graph with $n-1$ edges, but even this is not the case. Let $AB = BC = CD = AD = 0.23$, $AC = BD = 0.04$ in Fig. 6. Here the graph formed by AB, BC, DC, AD has $m\bar{\lambda} = 4.32$, but the best graph with three edges has $m\bar{\lambda} = 4.5$ (any star has $m\bar{\lambda} = 4.5$).

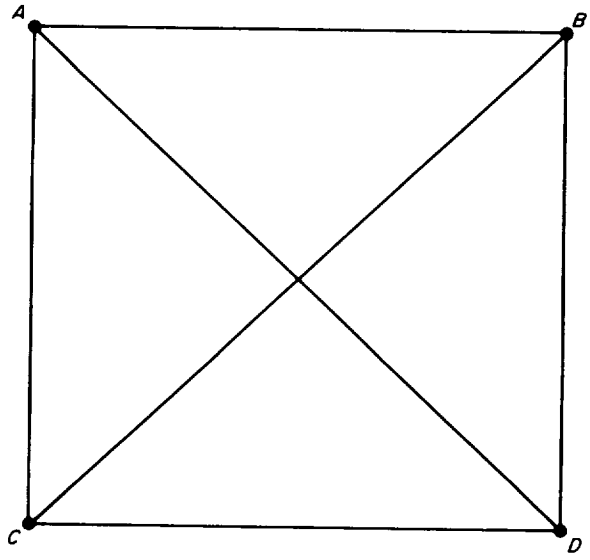


Fig. 6. Graph of order 4

In the previous examples ($n = 4$), we have seen that the minimum of $m\bar{\lambda}$ for the *best star* was always 4.5, even though the weights assigned to the various pairs of vertices were different in the three cases considered. Further, 4.5 is the bound of *Theorem 2* for $n = 4$. This behavior can be explained by the following result:

Theorem 6. For a given set of weights $S = \{w_{ij}\}$ on a set of n vertices, let $e_s^*(G)$ represent the minimum value of $m\bar{\lambda}$ attained by any star S_{n-1}. Then it is always the case that $e^*(G) \leqq 2(n-1)^2/n$.

Remark. Compare this result to that of Theorem 2; it states that any deviation from a flat distribution of weights can only result in a *decrease* in the minimum value of $m\bar{\lambda}$.

Proof. Denote the n vertices by $V_1, V_2, \cdots, V_n$, and with each vertex V_i associate a positive number u_i as follows:

$$u_i = \sum_j w_{ij}.$$

If we form the star with V_i as center, it has

$$m\bar{\lambda} = (n-1)(u_i + 2(1-u_i)) = (n-1)(2-u_i),$$

and so the best star corresponds to the vertex V_i for which u_i is a minimum.

But

$$\sum_{i=1}^{n} u_i = 2,$$

since in this sum each w_{ij} occurs exactly twice, i.e., once in u_i and once in u_j, and $\sum w_{ij} = 1$. Hence,

$$u_{max} = \max_i (u_i) \geqq \frac{2}{n},$$

and

$$e_s^*(G) \leqq (n-1)\left(2 - \frac{2}{n}\right) = 2(n-1)/n.$$

Comparing the results of Theorems 5 and 6, we see that the best S_{n-1} has $m\bar{\lambda} < 2(n-1)$, while the best possible $m\bar{\lambda}$ is $\geqq n-1$. So although a star may not always be optimal, we see that the best star is never worse than a factor of 2 from optimal. An extreme case of the relationship of $e_s^*(G)$ and $e_s(G)$ is given by the following example:

Let the vertices of G be denoted by $V_1, V_2, \cdots, V_n$, and let $w_{ij} = 1/(n-1)$ if $j = i+1$, and 0 otherwise.

Here the bound of Theorem 3 is attained by the chain L_n from V_1 to V_n, while the best star has

$$m\bar{\lambda} = (n-1)\left(2 - \frac{2}{n-1}\right);$$

here the ratio

$$e^*(G)/e(G) = 2 - \frac{2}{n-1} \to 2$$

as $n \to \infty$, and we have seen that 2 is the largest ratio possible.

But in a statistical sense, this is a pessimistic example, as the following section will help to show.

5. Probabilistic Weights

Lemma. *Let F be a distribution function with $F(x) = 0$ for $x \leq 0$. Suppose F has (finite) mean μ and (finite) variance σ^2. Let $X_1, X_2, \cdots, X_n$ be n independent variables with identical distribution functions F, and let $X_{(n)}$ represent the largest of $X_1, X_2, \cdots, X_n$. Let Y_n be the random variable defined by $Y_n = X_{(n)}/(X_1 + X_2 + \cdots + X_n)$. Then for every $\varepsilon > 0$, $Pr\{N^{1/2}Y_n > \varepsilon\} \to 0$ as $n \to \infty$.*

Proof. To find the distribution of $X_{(n)}$, note that

$$\Pr\{X_{(n)} < x\} = \prod_{i=1}^{n} \Pr\{X_i < x\} = F^n(x),$$

since the random variables X_i are independent.

Write

$$S_n = X_1 + X_2 + \cdots + X_n,$$

and

$$G_n(y) = \Pr\{n^{1/2}X_{(n)} \leq yS_n\}.$$

Then

$$1 \geq G_n(y) \geq \Pr\left\{X_{(n)} \leq \frac{y}{n^{1/2}} S_n \,\middle|\, \left|\frac{S_n}{n} - \mu\right| < \varepsilon\right\} \Pr\left\{\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right\}.$$

By the Chebyschev inequality

$$1 \geq G_n(y) \geq \Pr\left\{X_{(n)} \leq \frac{y}{n^{1/2}}(n\mu - n\varepsilon)\right\}\left(1 - \frac{\sigma^2}{n\varepsilon^2}\right),$$

and by substitution,

$$1 \geq G_n(y) \geq F^n(yn^{1/2}(\mu - \varepsilon))\left(1 - \frac{\sigma^2}{n\varepsilon^2}\right), \quad \text{for all } \varepsilon > 0. \tag{3}$$

We now compute

$$\lim_{n \to \infty} F^n(yn^{1/2}(\mu - \varepsilon)).$$

F has finite variance, so

$$\int_0^\infty x^2 f(x)\,dx = \mu_2 < \infty,$$

where $F' = f$. Consequently,

$$\lim_{y \to \infty} \int_y^\infty x^2 f(x)\,dx = 0,$$

which means that

$$\lim_{y \to \infty} y^2 \int_y^\infty f(x)\,dx = 0.$$

But now

$$F^n(yn^{1/2}(\mu - \varepsilon)) = \left(\int_0^{yn^{1/2}(\mu-\varepsilon)} f(x)\,dx\right)^n = \left(1 - \int_{yn^{1/2}(\mu-\varepsilon)}^\infty f(x)\,dx\right)^n = \left(1 - \frac{A(n)}{n}\right)^n,$$

where

$$A(n) = n \int_{yn^{1/2}(\mu-\varepsilon)}^\infty f(x)\,dx.$$

But from the fact that

$$\lim_{y \to \infty} y^2 \int_y^\infty f(x)\,dx = 0$$

we see that

$$\lim_{n \to \infty} y^2 n(\mu - \varepsilon)^2 \int_{yn^{1/2}(\mu-\varepsilon)}^\infty f(x)\,dx = 0,$$

and so

$$\lim_{n \to \infty} A(n) = 0$$

as well. Hence,

$$\lim_{n\to\infty} F^n(yn^{1/2}(\mu-\varepsilon)) = \left(1-\frac{A(n)}{n}\right)^n = 1$$

from elementary limit considerations. So from Eq. (3) we see that

$$\lim_{n\to\infty} G_n(y) = 1.$$

But $\Pr\{n^{1/2}Y_n > \varepsilon\} = 1 - G_n(\varepsilon)$, and so

$$\lim_{n\to\infty} \Pr\{n^{1/2}Y_n > \varepsilon\} = 0.$$

This proves the lemma.

Now in Theorem 5, the bound can be rewritten as follows:

$$e_s(G) \geqq (n-1)\left(2 - \sum_{k=1}^{n-1} w_k\right),$$

Hence,

$$e_s(G) \geqq (n-1)(2-(n-1)w_1),$$

and so

$$\frac{e_s^*(G)}{e_s(G)} \leqq 2(n-1)/n(2-(n-1)w_1) \leqq 2/(2-(n-1)w_1).$$

If now, for example, the weights are considered to be $n(n-1)/2$ random samples from a distribution function with finite mean and variance (normalized so that their sum is 1), then an easy conclusion of the lemma shows that $\Pr\{(n-1)w_1 > \varepsilon\} \to 0$. Consequently,

$$\Pr\{e^*(G)/e(G) > 1+\varepsilon\} \to 0 \text{ as } n \to \infty$$

for every $\varepsilon > 0$. In a meaningful sense, therefore, the star "asymptotically minimizes" $m\bar{\lambda}$.

Remark. The lemma can be modified to show the stated result when the traffic between two terminals is assumed to be proportional to the sum or product of the "sizes" of the vertices, where now the "sizes" are assumed to be distributed according to some distribution with finite mean and variance.

References

1. Golomb, S. W., *Sequences with Randomness Properties*, The Martin Co., Baltimore, 1955.
2. Zierler, N., "Linear Recurring Sequences," *Journal of the Society of Applied Mathematics*, Vol. 7, No. 1, March 1959.
3. Franklin, J., "Deterministic Simulation of Random Processes," *Mathematics of Computation*, Vol. 27, pp. 28–59, 1963.
4. Baumert, L. D., Golomb, S. W., and Hall, M., Jr., "Discovery of an Hadamard Matrix of Order 92," *Bulletin of the American Mathematical Society*, Vol. 68, pp. 237–238, 1962.
5. Brauer, A., "On a New Class of Hadamard Determinants," *Mathematische Zeitschrift*, Vol. 58, pp. 219–225, 1953.
6. Goldberg, K., "Hadamard Matrices of Order Cube Plus One," *Notices of the American Mathematical Society*, No. 7, p. 348, 1960.
7. Paley, R. E. A. C., "On Orthogonal Matrices," *Journal of Mathematics and Physics*, Vol. 12, pp. 311–320, 1933.

References (Cont'd)

8. Stanton, R. G., and Sprott, D. A., "A Family of Difference Sets," *Canadian Journal of Mathematics,* Vol. 10, pp. 73–77, 1958.

9. Williamson, J., "Hadamard's Determinant Theorem and the Sum of Four Squares," *Duke Mathematical Journal,* Vol. 11, pp. 65–81, 1944.

10. Williamson, J., "Note on Hadamard's Determinant Theorem," *Bulletin of the American Mathematical Society,* Vol. 53, pp. 608–613, 1947.

11. Feller, W., *An Introduction to Probability Theory and Its Applications,* Second Edition, Wiley & Sons, New York, 1957.

12. Goncarov, V., "Du domaine d'analyse combinatoire," Bulletin de l'Academie Sciences USSR, Ser. Math, Vol. 8, pp. 3–48, 1944 (in Russian with French summary. Also available in English translation as "On the Field of Combinatory Analysis," American Mathematical Society Translation, Providence, 1955).

13. Loeve, M., *Probability Theory,* Second Edition, Van Nostrand, Princeton, 1960.

14. Rota, G. C., "The Number of Partitions of a Set," *American Mathematical Monthly,* Vol. 71, pp. 498–503, 1964.

15. Moser, L., and Wyman, M., "An Asymptotic Formula for the Bell Numbers," *Transactions of the Royal Society of Canada,* Vol. 49. Series 3, pp. 49–53, 1955.

16. Kleinrock, L., "Communication Nets," MIT Electrical Engineering Department, Doctoral Dissertation, Cambridge, Mass., 1963.

17. Berge, C., *The Theory of Graphs,* Methuen & Co. Ltd., London, 1962.

XXII. Communications Systems Research: Radar Astronomy

A. Radar Occultations

P. Reichley and D. O. Muhleman

1. Summary

Radar observations of Venus can be successfully obtained over the entire orbit of the planet, including the time of superior conjunction. During the time of superior conjunction the radar beam must twice traverse the solar corona; this traverse affects the ray path, signal loss, and polarization of the signal through the magneto-ionic effects of the electron plasma in the corona. These effects become more pronounced at the longer wavelengths but are present, of course, in the centimeter range utilized in the JPL Radar Astronomy Program. The effects of the plasma on the ray path are observable in terms of a spurious doppler shift and a delay in range measurements. This article discusses these effects, from the point of view of measuring them experimentally.

2. Introduction

We wish to investigate the magnitude of the corona effects on radar beams at a wavelength of 12.5 cm. It is necessary to observe Venus when it is very close to the Sun in angle in order to obtain measurable effects at this wavelength. In carrying out this analysis, we have the use of a 210-ft antenna in mind which, according to preliminary design studies, will have a half-power beam width of about 0.1 deg with its side lobes suppressed by from 50 to 60 db. With this system it should be possible to observe Venus to within a fraction of a degree from the Sun without a serious degradation in the system noise temperature. Although the calculations presented below refer to occultations of Venus by the Sun, we are considering the equally interesting problem of the radar occultations of the moons of Jupiter by the atmosphere of the planet. It should be noted that the radar detectability of Jupiter's two large satellites is comparable to the detectability of Jupiter itself due to differences in their rotations.

The calculations of interest here can best be carried out using the index of refraction of the corona which, from magneto-ionic theory, is approximately a linear function of the electron density N_e, a fairly well-known function of the distance from the center of the Sun. This electron density is given by

$$N_e(r) = 10^8\left(\frac{1.55}{r^6} + \frac{2.99}{r^{16}}\right) \text{cm}^{-3},$$

where r is expressed in solar radii. This relationship is discussed in detail below.

The electron density in terms of the index of refraction has been programmed for a digital computer for exact calculations. However, in this preliminary paper we have analyzed the problem by using an r^{-2} approximation to the solar density, since the entire problem can then be analyzed in closed form. In the absence of a magnetic field, the index of refraction of an electron plasma (neglecting collisions) is given by

$$n^2 \simeq 1 - \frac{N_e}{N_{e_c}(\nu)},$$

where $N_{e_c}(\nu)$ is the critical electron density for a frequency ν and is given by

$$N_{e_c}(\nu) = 1.24 \times 10^{-8}\nu^2.$$

For $\nu = 2.388 \times 10^9$ cps, the above equation becomes

$$N_{e_c} = 7.1 \times 10^{10} \text{ electrons/cm}^3.$$

It should be noticed that the actual solar electron density does not become critical (at $\lambda = 12.5$ cm) even at $r = 1$ solar radius.

We wish to approximate the electron density in the form

$$n^2 = 1 - \frac{\alpha^2}{r^2},$$

where we need the value of α^2 in the region of interest. If we aim a ray from the Earth at an angle of 0.4 deg from the center of the Sun, it will approach no nearer than 1.5 solar radii. Therefore, we wish to approximate the index of refraction in the region from $r = 1.5$ out to the Earth, $r = 215$. This yields $\alpha^2 = 2 \times 10^{-4}$. As a preliminary check on this figure, we can compare the integrated electron density over the path between the Earth and $r = 1.5$ from our approximation to the same integral for the true density. Most of the refraction effects are essentially dependent of such an integration (not, however, the absorption). This yields 6.6×10^{17} electrons/cm^2 for $\alpha = 2 \times 10^{-4}$ and 2.9×10^{17}cm^{-2} for the true density, which is more than satisfactory agreement. Further justification for our choice of α^2 will be found below.

We assume that a measurement of Earth–Venus range is accomplished with a *group-velocity* ranging device. Furthermore, measurements of the doppler shift in the radar echo are measured with a *phase-velocity* device. Define a group index of refraction n_g, such that

$$v_g = \frac{c}{n_g},$$

analogous to the phase velocity

$$v = \frac{c}{n}.$$

c is the vacuum speed of light and v_g is the group velocity. Then the group index is related to the natural index of refraction by

$$n_g = n + \nu \frac{dn}{d\nu}.$$

The later formula can be shown to be correct except in the region of anomolous dispersion which occurs near the critical electron density. The group index must then be used in range equations and the natural index in the doppler equations.

3. Ray Paths In An Inhomogeneous Medium

a. Fermat's principle and the ray equation. In considering the path a radar beam takes as it passes through an inhomogeneous medium such as a planetary atmosphere or the Sun's corona, we shall use the principles of geometrical optics. From Fermat's principle, the path of least time of a ray through a medium from Point P_1 to Point P_2 is given by the minimum of the line integral

$$\frac{1}{c}\int_{P_1}^{P_2} n\, ds, \tag{1}$$

where n is the index of refraction of the medium and c is the speed of light.

Since we will be considering media bounding planetary bodies, we shall assume that n is a function of altitude alone. We shall also work in only two dimensions, since the coordinate system may always be taken in the plane of the ray. Hence, writing Eq. (1) in polar coordinates we have

$$\frac{1}{c}\int_{P_1}^{P_2} n(r)\,(dr^2 + r^2\, d\theta^2)^{1/2}.$$

The variational problem is then given by

$$\delta \int_{P_1}^{P_2} n(r)\left[1 + r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2} dr = 0,$$

where we choose the integration to be with respect to r for convenience. From Euler's equation, we then have

$$-\frac{d}{dr}\left\{\frac{r^2 n(r)\left(\frac{d\theta}{dr}\right)}{\left[1+r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2}}\right\}=0,$$

or

$$\frac{r^2 n(r)\left(\frac{d\theta}{dr}\right)}{\left[1+r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2}}=b,$$

where b is a constant of integration. Solving for $d\theta/dr$, we have

$$\frac{d\theta}{dr}=\frac{b}{r\left[r^2n^2(r)-b^2\right]^{1/2}}. \tag{2}$$

We shall refer to this equation as the *ray equation.*

In working with the ray equation, we shall use a coordinate system in which $\theta = 0$ when r is minimal. We see from Eq. (2) that this requires that

$$r^2n^2(r)-b^2=0. \tag{3}$$

Let us call the r which satisfies this equation r_m. A point on the ray path $(\hat{\theta},\hat{r})$ is then given by

$$\theta(\hat{r})=\int_{r_m}^{\hat{r}}\frac{b\,dr}{r\left[r^2n^2(r)-b^2\right]^{1/2}}. \tag{4}$$

Before we are able to evaluate Eq. (4), we must determine b. Let i be defined as the angle between the ray tangent vector and the radius vector from the planet at any point on the ray path. Then from vector calculus

$$\sin i=\frac{r}{\left[r^2+\left(\frac{dr}{d\theta}\right)^2\right]^{1/2}}.$$

But from Eq. (2), this yields

$$rn(r)\sin i=b, \tag{5}$$

a result known as *Bouger's Theorem.* We shall call b the impact parameter of the ray. If we construct a line parallel to the asymptote of the ray and passing through the origin of our coordinate system, then b is simply the distance between the asymptote and the line (Fig. 1).

b. Ray bending. If we observe that

$$\frac{d}{dr}\left[\sin^{-1}\frac{b}{rn(r)}\right]=\frac{-bn'(r)}{n(r)\left[r^2n^2(r)-b^2\right]^{1/2}}-\frac{b}{r\left[r^2n^2(r)-b^2\right]^{1/2}},$$

then from Eq. (4), we have

$$\theta(\hat{r})=-\sin^{-1}\frac{b}{rn(r)}\Bigg|_{r_m}^{\hat{r}}-\int_{r_m}^{\hat{r}}\frac{bn'(r)\,dr}{n(r)\left[r^2n^2(r)-b^2\right]^{1/2}}.$$

Hence,

$$\theta(\hat{r})=\frac{\pi}{2}-\left\{\sin^{-1}\frac{b}{\hat{r}n(\hat{r})}+\int_{r_m}^{\hat{r}}\frac{bn'(r)\,dr}{n(r)\left[r^2n^2(r)-b^2\right]^{1/2}}\right\} \tag{6}$$

The angle between the tangent vector to the ray at r_m and the tangent vector to the ray at $\hat{r}$, which we shall call the "one-way bending" $B(\hat{r})$, is given by

$$B(\hat{r})=-\int_{r_m}^{\hat{r}}\frac{bn'(r)\,dr}{n(r)\left[r^2n^2(r)-b^2\right]^{1/2}}, \tag{7}$$

since $\sin^{-1}\left[b/\hat{r}n(\hat{r})\right]$ is simply $i(\hat{r})$ from Bouger's Theorem.

If a ray travels from a point at $\hat{r}$ to a point in a different direction at $\bar{r}$, then the "two-way bending" is simply $B(\hat{r})+B(\bar{r})$. When the ray is considered as approaching and leaving from points at infinite distances, we call this the *total bending* B, which is given by

$$B=-2\int_{r_m}^{\infty}\frac{bn'(r)\,dr}{n(r)\left[r^2n^2(r)-b^2\right]^{1/2}}. \tag{8}$$

We consider two examples of the total bending both of which were handled numerically:

Example 1: The Earth. An acceptable model of the index of refraction of the Earth's atmosphere is given by

$$n^2(r)=1+\alpha e^{-\beta(r-r_0)}, \tag{9}$$

where

$$\begin{aligned}&\alpha = n_0^2-1,\ n_0 \text{ the index of refraction at sea level,}\\ &\beta = 1/h,\ h \text{ the scale height of the atmosphere,}\\ &r_0 = 1 \text{ Earth radius.}\end{aligned} \tag{9a}$$

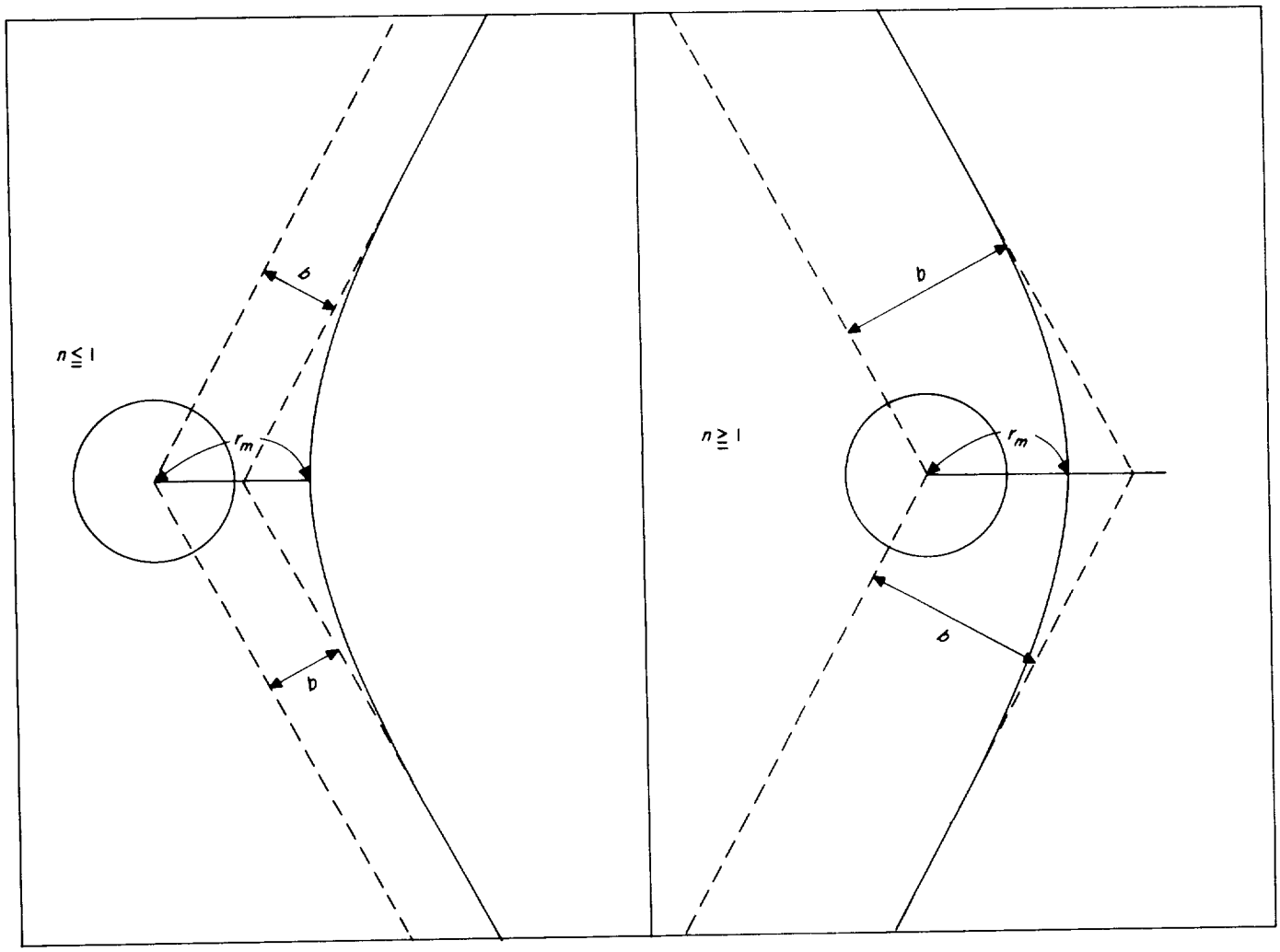


Fig. 1. Relation of impact parameter b to ray path

We chose to use a scale height of 10 km, and to use Earth radii as units of measurement. The constants (9a) are then

$$\alpha = 6.0009 \times 10^{-4},$$

$$\beta = 600,$$

$$r_0 = 1.$$

To handle Eq. (8) numerically, we first made the change of variable $s = 1/n(r)$, which yielded

$$B = 2b \int_{s_m}^{1} \frac{ds}{[f^2(s) - b^2 s^2]^{1/2}},$$

where

$$s_m = \frac{1}{n(r_m)}$$

and from Eq. (9)

$$f(s) = \frac{1}{\beta} \ln\left[\frac{\alpha s^2}{1 - s^2}\right] + 1.$$

To remove the singularity at $s = s_m$, we use the method of subtraction of the singularity, which leads to

$$B = 2b \int_{s_m}^{1} \left\{ \frac{1}{[f^2(s) - b^2 s^2]^{1/2}} - \frac{k}{(s - s_m)^{1/2}} \right\} ds + 4bk(1 - s_m)^{1/2}, \quad (10)$$

where

$$k = \frac{1}{[2f(s_m) f'(s_m) - 2b^2 s_m]^{1/2}}.$$

Using the constants (9a), we programmed Eq. (10) with the results as shown in Fig. 2. With the impact parameter $b = 1.0003$, i.e., a grazing ray, we see that the ray is

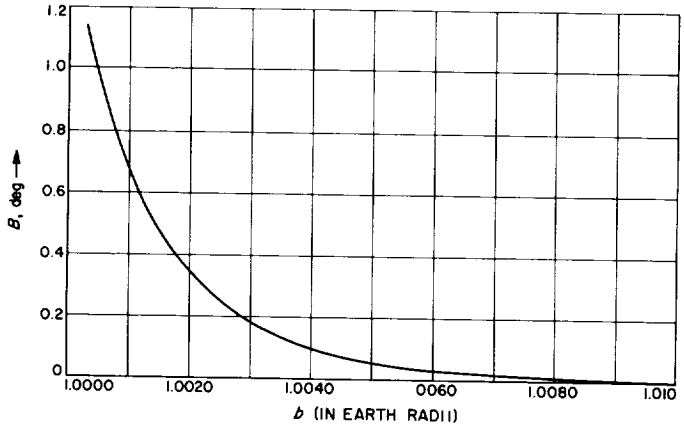


Fig. 2. Total bending versus impact parameter for the Earth

bent 1.14 deg and as b increases, B falls off exponentially as should be expected.

Example 2: The Sun. An acceptable model of the refractive index of the Sun's corona is given by

$$n^2(r) = 1 - \frac{e^2 T_e N(r)}{\varepsilon_0 m(\omega^2 + \nu^2)},$$

where

e = charge of electron = 1.6×10^{-19} coulombs

m = mass of electron = 9.11×10^{-31} kgs

T_e = kinetic temperature of electrons $\cong 10^{6}$°K

ε_0 = permittivity of free space = 8.85×10^{-12} F/m

$N(r)$ = electron density/m^3 depending on r

ω = angular frequency of ray, cps

ν = collisions/sec of N electrons

If we assume $\nu << \omega$, then we have

$$n^2(r) = 1 - 80.45 \times 10^6 \frac{N(r)}{f^2}.$$

For $N(r)$ we shall use Allen's (Ref. 1) revision of Baumbach's formula, which is given by

$$N(r) = 10^8\left(\frac{1.55}{r^6} + \frac{2.99}{r^{16}}\right), \quad \text{cm}^{-3}$$

where r is measured in solar radii. The frequency of interest is the frequency of the Goldstone radar, i.e., 2388 Mc. Hence, we have

$$n^2(r) = 1 - 1.4108 \times 10^{-8}\left(\frac{1.55}{r^6} + \frac{2.99}{r^{16}}\right), \qquad (11)$$

and for brevity we shall use

$$n^2(r) = 1 - \alpha\left(\frac{\beta}{r^6} + \frac{\gamma}{r^{16}}\right).$$

To handle Eq. (8) numerically, we first make the (change of) variable $s = r^{-1}$, which yields

$$B = -2b\int_0^{s_m} \frac{n'(s^{-1})\,ds}{sn(s^{-1})\,[n^2(s^{-1}) - b^2s^2]^{1/2}},$$

where

$$s_m = r_m^{-1},$$

and

$$n(s^{-1}) = [1 - \alpha(\beta s^6 + \gamma s^{16})]^{1/2},$$

$$n'(s^{-1}) = \frac{\alpha(3\beta s^7 + 8\gamma s^{17})}{n(s^{-1})}.$$

We then remove the singularity at $s = s_m$ by the method of subtraction of the singularity, which yields

$$B = -2b\int_0^{s_m} \left\{\frac{n'(s^{-1})}{sn(s^{-1})\,[n^2(s^{-1}) - b^2s^2]^{1/2}} - \frac{k}{(s_m - s)^{1/2}}\right\} ds - 4bk(s_m)^{1/2}, \qquad (12)$$

where

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

We then programmed Eq. (12) using (11) with the results as shown in Fig. 3. With the impact parameter

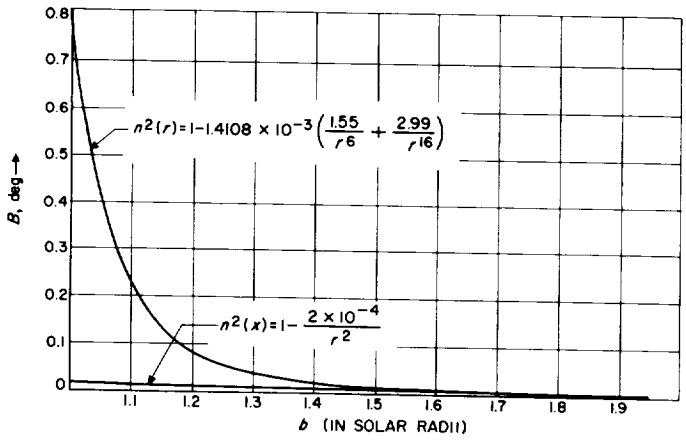


Fig. 3. Total bending versus impact parameter for the Sun

$b = 0.9986$, i.e., a grazing ray, we get a maximum bending of 0.784 deg, with B falling off as the index (11) with increasing b.

4. Range Differences Due to Refraction

If Venus is occulted by the Sun, then by tracking Venus by radar prior to occultation, we may compare the optical range to the true range. This difference, as a function of the angular separation of Venus and Earth with respect to the Sun, would then give us a measurement of the size and nature of the Sun's corona. In our analysis, we shall assume that both Venus and Earth have circular orbits, and that their orbits are coplanar.

Since the optical path length of the ray is given by cT_0, where T_0 is the time the ray takes on its minimal path with respect to group velocity, then from Eq. (1), the optical path length for a round-trip path is given by

$$\bar{\rho}_0 = 2\int_{P_1}^{P_2} n_g\, ds = 2\int_{r_m}^{r_\oplus} n_g(r)\left[1 + r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2} dr + 2\int_{r_m}^{r_\varphi} n_g(r)\left[1 + r^2\left(\frac{d\theta}{dr}\right)\right]^{1/2} dr,$$

where $d\theta/dr$ is simply the ray equation (2) and $n_g(r)$ is the group index of refraction as opposed to the phase index $n(r)$ in the ray equation. Hence,

$$\bar{\rho}_0 = 2\int_{r_m}^{r_\oplus} \frac{r n_g(r)\, n(r)\, dr}{[r^2 n^2(r) - b^2]^{1/2}} + 2\int_{r_m}^{r_\varphi} \frac{r n_g(r)\, n(r)\, dr}{[r^2 n^2(r) - b^2]^{1/2}}. \tag{13}$$

Since we have neglected terms of order higher than the second in the model of the phase index we have chosen as acceptable, i.e.,

$$n^2(r) = 1 - 80.45 \times 10^6 \frac{N(r)}{f^2},$$

and since the group index is given by

$$n_g^2(r) \simeq 1 + 80.45 \times 10^6 \frac{N(r)}{f^2},$$

then we may take $n_g(r)\, n(r) = 1$, neglecting terms of order higher than the second, and Eq. (13) becomes

$$\bar{\rho}_0 = 2\int_{r_m}^{r_\oplus} \frac{r\, dr}{[r^2 n^2(r) - b^2]^{1/2}} + 2\int_{r_m}^{r_\varphi} \frac{r\, dr}{[r^2 n^2(r) - b^2]^{1/2}}. \tag{14}$$

Since the only known quantities are $r_\oplus$, r_φ, and θ_d, the angular separation, we have to determine b and r_m. Once we have b, we simply solve Eq. (3) for r_m. Hence, our problem is then to find a way to determine b.

From Eq. (4), we have

$$\theta_d = \theta(r_\oplus) + \theta(r_\varphi) = \int_{r_m}^{r_\oplus} \frac{b\, dr}{r[r^2 n^2(r) - b^2]^{1/2}} + \int_{r_m}^{r_\varphi} \frac{b\, dr}{[r^2 n^2(r) - b^2]^{1/2}}. \tag{15}$$

In general, using accepted models of the index of refraction of the corona, Eq. (15) will be unintegrable in closed form. We therefore will have to invert Eq. (15) numerically, so that we may express b as a function of θ_d. Let us first find a set of points, (θ_{di}, C_i), $i = 1, n$, by numerically integrating Eq. (15) for $b = b_i$.

We then make a least squares fit to (θ_{di}, b_i), using θ_d as the abscissa and b as the ordinate. We then obtain a polynomial, $P(\theta_d)$, which expresses b as a function of θ_d, i.e.,

$$b = P(\theta_d). \tag{16}$$

If Eq. (15) is integrable in closed form, as it would be for certain simplified $n(r)$, we then obtain

$$\theta_d = F(b).$$

We may then solve this equation either analytically or numerically to obtain

$$b = G(\theta_d) = F^{-1}(\theta_d). \tag{17}$$

Once we have obtained either Eq. (16) or (17), we can then find r_m, and hence solve (14) to obtain range $\bar{\rho}_0$ as a function of θ_d. Since the straight line or true round-trip range is given by

$$\bar{\rho}_t = 2[r_\oplus^2 + r_\varphi^2 - 2r_\oplus r_\varphi \cos\theta_d]^{1/2}, \tag{18}$$

then the round-trip range difference due to refraction is

$$\delta\bar{\rho} = \bar{\rho}_0 - \bar{\rho}_t = 2\left\{\int_{r_m}^{r_\oplus} \frac{r\, dr}{[r^2 n^2(r) - b^2]^{1/2}} + \int_{r_m}^{r_\varphi} \frac{r\, dr}{[r^2 n^2(r) - b^2]^{1/2}} - [r_\oplus^2 + r_\varphi^2 - 2r_\oplus r_\varphi \cos\theta_d]^{1/2}\right\}, \tag{19}$$

where b is obtained either from Eq. (16) or (17).

We are now writing a program to find $\delta\bar{\rho}$ as a function of θ_d using the index (11), and this report will follow in a subsequent *SPS* article. Presently, however, we will consider the simplified model of the index, such that Eq. (19) is easily found analytically.

Let us take

$$n^2(r) = 1 - \frac{\alpha^2}{r^2} \qquad (\alpha^2 = 2 \times 10^{-4}) \tag{20}$$

as our simplified model of the index. As a comparison of Eq. (20) with Eq. (11), let us find the total bending B yielded by Eq. (20). We have from Eq. (8)

$$B = -2\int_{r_m}^{\infty} \frac{b\alpha^2\, dr}{r(r^2 - \alpha^2)[r^2 - (\alpha^2 + b^2)]^{1/2}},$$

which yields

$$B = \left[\frac{b}{(\alpha^2 + b^2)^{1/2}} - 1\right]\pi. \tag{21}$$

We have plotted Eq. (21) in Fig. 3. We notice that the index (11) and the index (20) yield results that disagree badly for $b \leq 1.3$. But for $B \geq 1.3$, they yield results that are very close, with the index (20) yielding a result that is slightly larger than the index (11) for $b \geq 1.7$ and slightly smaller for $b \leq 1.7$. Since small values of b are of no interest, due to physical limitations of the Goldstone antennas, we may conclude that we have chosen a suitable model.

We shall now proceed with the solution of Eq. (19) using Eq. (20). Solving Eq. (15) using Eq. (20), we have

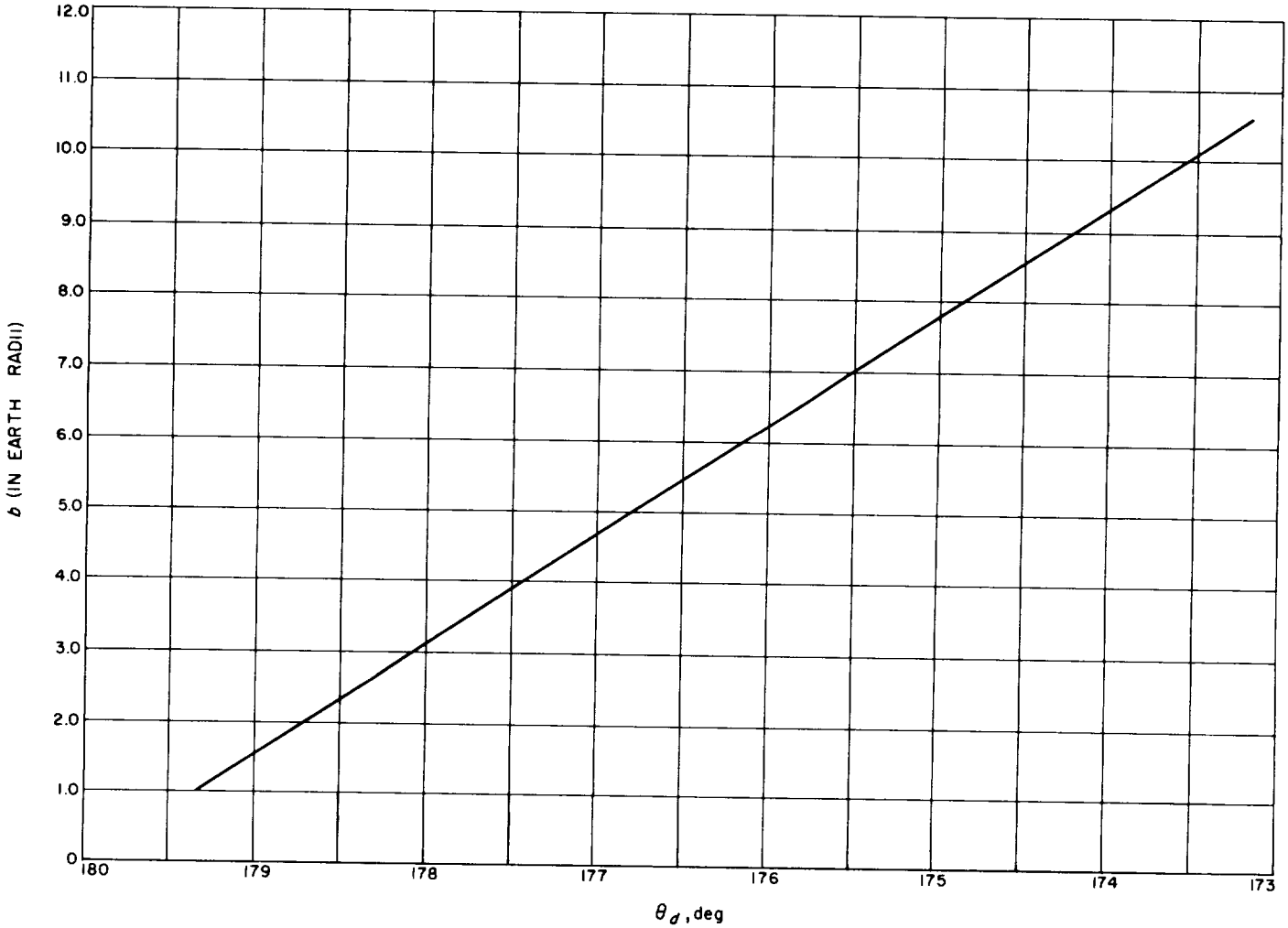


Fig. 4. Impact parameter versus Earth–Venus heliocentric angle

$$\theta_d = \int_{r_m}^{r_\oplus} \frac{b\,dr}{r[r^2-(\alpha^2+b^2)]^{1/2}} + \int_{r_m}^{r_\varphi} \frac{b\,dr}{r[r^2-(\alpha^2+b^2)]^{1/2}},$$

$$r_m = [\alpha^2+b^2]^{1/2},$$

which yields

$$\theta_d = \frac{b}{(\alpha^2+b^2)^{1/2}}\left[\cos^{-1}\frac{(\alpha^2+b^2)^{1/2}}{r_\oplus} + \cos^{-1}\frac{(\alpha^2+b^2)^{1/2}}{r_\varphi}\right] \quad (22)$$

We may now solve Eq. (22) to get a numerical solution of the form (17), i.e., $b = G(\theta_d)$.

We now wish to solve Eq. (14) for $\bar{\rho}_0$ using Eq. (20). We have

$$\bar{\rho}_0 = 2\int_{r_m}^{r_\oplus} \frac{r\,dr}{[r^2-(\alpha^2+b^2)]^{1/2}} + 2\int_{r_m}^{r_\varphi} \frac{r\,dr}{[r^2-(\alpha^2+b^2)]^{1/2}},$$

$$r_m = (\alpha^2+b^2)^{1/2},$$

which yields

$$\bar{\rho}_0 = 2\{[r_\oplus^2-(\alpha^2+b^2)]^{1/2} + [r_\varphi^2-(\alpha^2+b^2)]^{1/2}\}. \quad (23)$$

Hence, from Eq. (19), we have

$$\delta\bar{\rho} = 2\{[r_\oplus^2-(\alpha^2+b^2)]^{1/2} + [r_\varphi^2-(\alpha^2+b^2)]^{1/2} - [r_\oplus^2+r_\varphi^2-2r_\oplus r_\varphi\cos\theta_d]^{1/2}\}, \quad (24)$$

where $b = G(\theta_d)$ from Eq. (22).

Using $r_\oplus = 215r_\odot$, $r_\varphi = 151r_\odot$, and $\alpha^2 = 2\times10^{-4}$, $b = G(\theta_d)$ is shown in Fig. 4. Using the same constants, the result (24) for $\delta\bar{\rho}$ is shown in Fig. 5. We see that we get a significant range difference with an angular separation as small as 174 deg. These results are encouraging, as the 210-ft radar antenna ceases to have sidelobe interference of significance at approximately 179 deg of angular separation (b = 1.5 from Fig. 4).

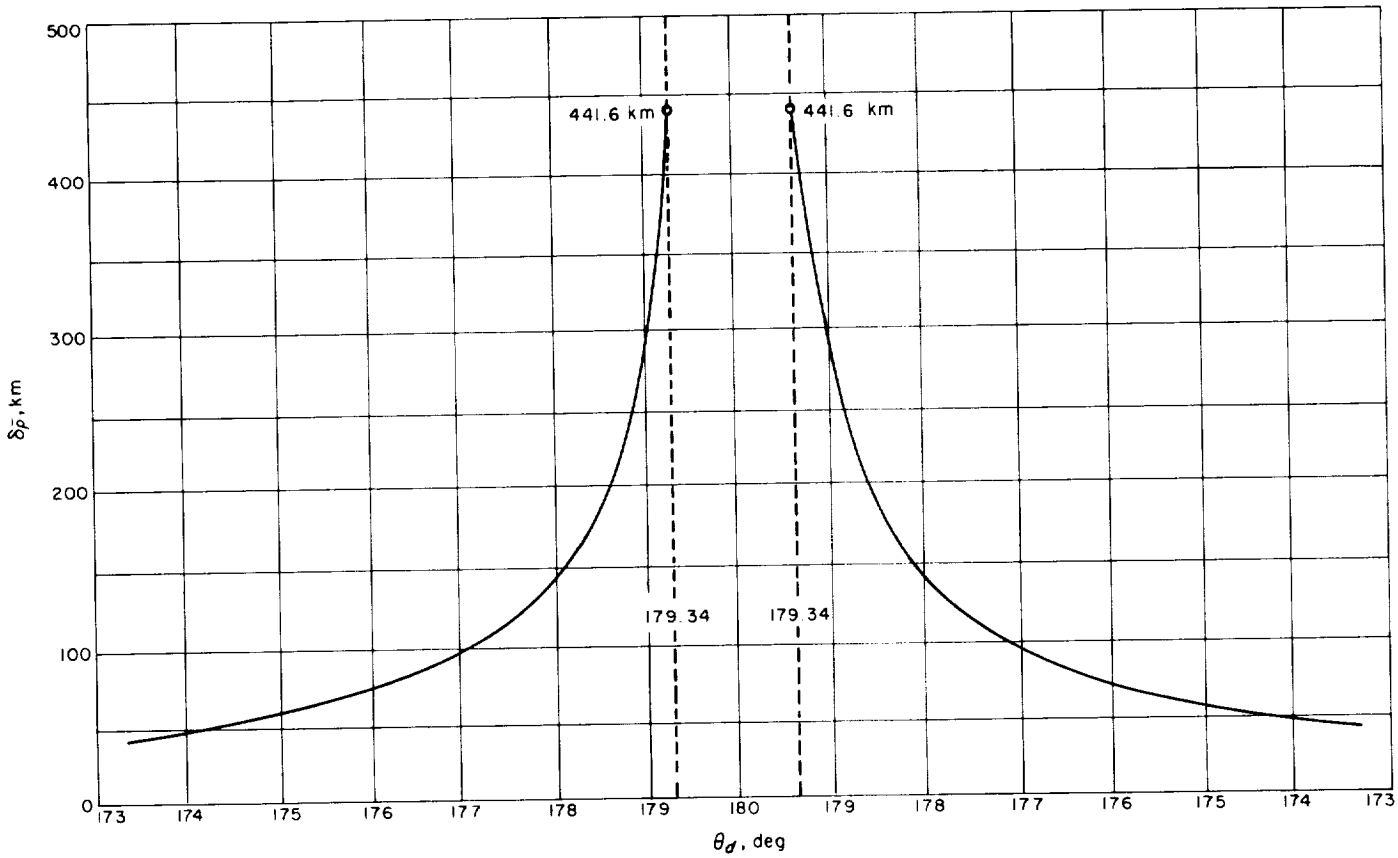


Fig. 5. Range difference versus Earth–Venus heliocentric angle

5. Doppler Shift Due to Refraction

If we are measuring the doppler shift as we track Venus prior to occultation by the Sun, then by removing the doppler shift contributed by the motion of Venus, we obtain the doppler shift due to refraction by the corona. This information, like the range difference $\delta\bar{\rho}$, would then be useful in determining the size and nature of the corona. As in our analysis of $\delta\bar{\rho}$, we shall assume that Venus and Earth have circular orbits and that their orbits are coplanar.

We shall consider the doppler shift in terms of the range rate. As a result, the range we shall consider shall be in terms of the phase index of refraction rather than the group index. Our range equation is then

$$\rho = \int_{S_1}^{S_2} n\, ds. \tag{25}$$

The doppler shift of Venus as it moves away from the Earth is given by

$$f_d = -\frac{2\dot{\rho}_0 f_0}{c},$$

where $\dot{\rho}_0$ is the range rate of the optical path length, f_0 is the transmitter frequency, and c is the speed of light. The doppler shift of Venus due to its motion alone as it moves away from the Earth is given by

$$f_m = -\frac{2\dot{\rho}_t f_0}{c},$$

where $\dot{\rho}_t$ is the range rate of the true path length. Hence, the doppler shift due to refraction is given by

$$f_r = \frac{-2f_0}{c}(\dot{\rho}_0 - \dot{\rho}_t). \tag{26}$$

To find $\dot{\rho}$ we must solve Eq. (25) considering b to be a function of time. This follows since if we consider θ_d to be a function of t, as given by the ephemeris, then upon solving Eq. (15), we obtain b as a function of θ_d, i.e., t. We have from Eq. (25)

$$\rho(t) = \int_{r_m}^{r_\oplus} n(r)\left[1 + r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2} dr + \int_{r_m}^{r_\varphi} n(r)\left[1 + r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2} dr,$$

where $d\theta/dr$ is simply the ray equation (2). Hence,

$$\rho(t) = \int_{r_{m(t)}}^{r_\oplus} \frac{rn^2(r)\,dr}{[r^2n^2(r) - b^2(t)]^{1/2}} + \int_{r_{m(t)}}^{r_\varphi} \frac{rn^2(r)\,dr}{[r^2n^2(r) - b^2(t)]^{1/2}}. \tag{27}$$

We shall now modify Eq. (27) so that differentiation will be easier. We have

$$\begin{aligned}\rho(t) = {} & \int_{r_{m(t)}}^{r_\oplus} \frac{rn^2(r)\,dr}{[r^2n^2(r) - b^2(t)]^{1/2}} \\ & - b(t)\int_{r_{m(t)}}^{r_\oplus} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}} \\ & + b(t)\int_{r_{m(t)}}^{r_\oplus} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}} \\ & + \int_{r_{m(t)}}^{r_\varphi} \frac{rn^2(r)\,dr}{[r^2n^2(r) - b^2(t)]^{1/2}} \\ & - b(t)\int_{r_{m(t)}}^{r_\varphi} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}} \\ & + b(t)\int_{r_{m(t)}}^{r_\varphi} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}},\end{aligned}$$

and

$$\begin{aligned}\rho(t) = {} & \int_{r_{m(t)}}^{r_\oplus} \frac{[r^2n^2(r) - b^2(t)]^{1/2}}{r}\,dr \\ & + b(t)\int_{r_{m(t)}}^{r_\oplus} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}} \\ & + \int_{r_m}^{r_\varphi} \frac{[r^2n^2(r) - b^2(t)]^{1/2}}{r}\,dr \\ & + b(t)\int_{r_{m(t)}}^{r_\varphi} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}}\end{aligned}$$

which yields upon differentiation,

$$\dot{\rho}(t) = b(t)\frac{d}{dt}\left[\int_{r_{m(t)}}^{r_\oplus} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}} + \int_{r_{m(t)}}^{r_\varphi} \frac{b(t)\,dr}{r[r^2n^2(r) - b^2(t)]^{1/2}}\right]$$

But from Eq. (15), the quantity in brackets is simply $\theta_d(t)$. Hence,

$$\dot{\rho}(t) = b(t)\,\dot{\theta}_d(t),$$

and since we are considering circular orbits, then $\theta_d(t) = \omega_s t$ where ω_s is the synodic angular rate of Venus, which leads to

$$\dot{\rho}(t) = \omega_s b(t). \tag{28}$$

Since the impact parameter $b_t(t)$ of the true path is given by elementary trigonometry as

$$b_t(t) = \frac{r_\oplus r_\varphi \sin(\omega_s t)}{[r_\oplus^2 + r_\varphi^2 - 2r_\oplus r_\varphi \cos(\omega_s t)]^{1/2}}$$

then from Eq. (26), we have

$$f_r(t) = \frac{-2f_0\omega_s}{c}\left[b(t) - \frac{r_\oplus r_\varphi \sin(\omega_s t)}{[r_\oplus^2 + r_\varphi^2 - 2r_\oplus r_\varphi \cos(\omega_s t)]^{1/2}}\right],$$

where $b(t)$ is obtained by inverting Eq. (15).

We are now writing a program to evaluate $f_r(t)$ for variable parameter t using the index (11). This report will follow in a subsequent *SPS* article. We shall again consider the simplified index (20) in (29) as we did for the range difference $\delta\bar{\rho}$.

We recall that to solve for $b(t)$ using the simplified index, we had to solve Eq. (22) numerically to find $b(t)$, given $\theta_d(t)$. Using the following constants:

$$\alpha = 2 \times 10^{-4},$$
$$f_0 = 2.388 \times 10^9 \text{ cps},$$
$$c = 0.43092 r_\odot/\text{sec},$$
$$\omega_s = 1.24311 \times 10^{-7} \text{ rad/sec},$$
$$r_\oplus = 215 r_\odot,$$
$$r_\varphi = 151 r_\odot,$$

the result (29) shown in Fig. 6 is obtained. As in the case of the range difference, we see that at 174 deg of angular

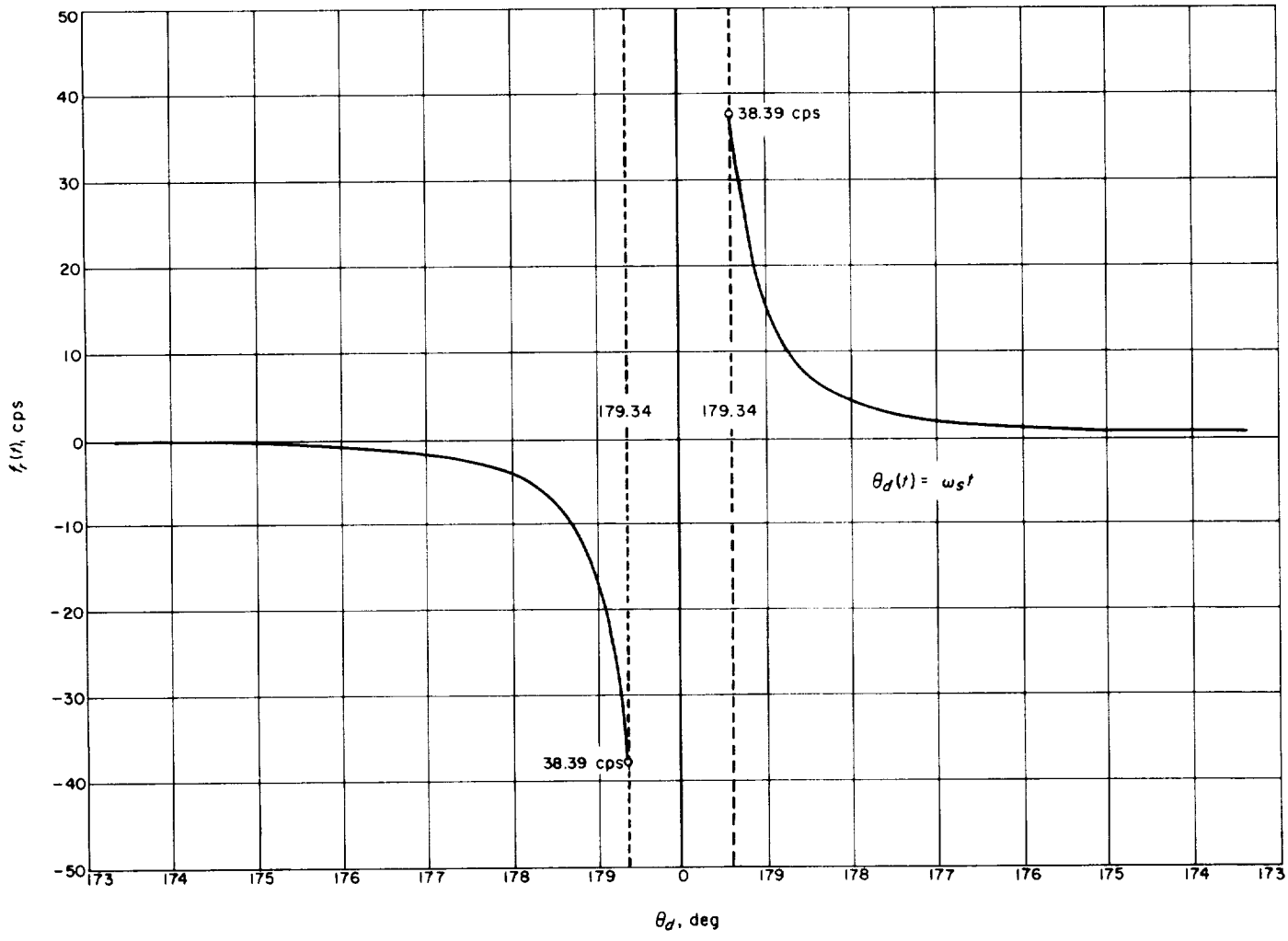


Fig. 6. Refraction doppler shift versus Earth–Venus heliocentric angle

separation we have a doppler shift of approximately 0.5 cps; at 179 deg we have approximately 17 cps. This again is an encouraging result.

The reliability of the numerical results in terms of range differences and doppler differences is difficult to assess even though we have been careful in formulating the index of refraction approximation. It appears, however, that numerical calculations utilizing the more exact model of coronal electron density will yield similar results. In such a circumstance highly meaningful experiments can be carried out at such times when Venus (or possibly even Mercury) is in the correct celestial position. Future studies will include an investigation of the Jupiter atmosphere about which very little is known. Detailed polarization and absorption effects remain to be investigated.

B. Effects of General Relativity on Planetary Radar Distance Measurements

D. O. Muhleman and P. Reichley

1. Summary

The Earth–Venus distance has been measured with the Goldstone Venus site radar over a 3-mo period centered around the 1964 inferior conjunction of Venus with an apparent accuracy of about 5 μsec in round-trip propagation time. These measurements, in conjunction with previous observations, are being utilized to redetermine the Astronomical Unit and the elliptical parameters of the orbits of the two planets. The computational procedure is to compare the measurements to theoretical observables computed from the ephemerides and its associated *a priori* constants. For the most part, the ephemerides have been formulated from the Newtonian theory of gravitation, and the observables have been computed with regard to the special theory of relativity. However, Einstein's Theory of Gravitation predicts a direct effect on the time-of-propagation measurements due to the mass of the Sun. These effects are separate from the secular motions of the planets themselves arising from Einstein's theory. In this report we estimate the magnitude of the former effect.

2. Introduction

It is well-known from classical physics that the wave front of an electromagnetic wave will propagate in free space according to the *eikonal equation*

$$\frac{1}{c^2}\left(\frac{\partial \Psi}{\partial t}\right)^2 - (\text{grad}\,\Psi)^2 = 0, \tag{1}$$

where Ψ is the phase and c the velocity of light in a vacuum. If the wave is propagating in a medium of index of refraction n, Eq. (1) holds with c replaced by the phase velocity in the medium, or

$$\left(\frac{n}{c}\right)^2\left(\frac{\partial \Psi}{\partial t}\right)^2 - (\text{grad}\,\Psi)^2 = 0. \tag{2}$$

A modern discussion with the pertinent formulas for the treatment of the problem of propagation in vacuum in the presence of a gravitational field (in our case the Sun) is presented by Fock in Ref. 2. The mechanism for solving our particular problem arises from the solution of Einstein's gravitational equations for a light ray in the presence of a single, concentrated mass which leads to the well-known Schwarzschild metric. The relativistic aspects of this metric are contained in a single parameter, α, called the "gravitation radius of the mass M," and are given by the expression

$$\alpha = \frac{\gamma M}{c^2}, \tag{3}$$

where γ is the Newtonian gravitational constant. The parameter α is equal to 1.48 km for the Sun and 0.433 cm for the Earth, which justifies neglecting the masses of the Earth and Venus. From the Schwarzschild metric and the differential equation for the null geodesic (Ref. 2), it can be shown that to first order in α/r (where r is the distance from the center of the Sun to any point of interest) the eikonal equation (1) becomes

$$\frac{n^2(r)}{c^2}\left(\frac{\partial \Psi}{\partial t}\right)^2 - (\text{grad}\,\Psi)^2 = 0, \tag{2'}$$

where

$$n^2(r) = 1 + \frac{4\alpha}{r}. \tag{4}$$

Thus, to first order in α/r, the propagation of a wave front in vacuum in the presence of a concentrated mass M is mathematically equivalent to the propagation of the wave in a medium of index of refraction n given by Eq. (4). *Consequently, the ray trajectory and propagation time can be obtained in the same way as we have done in the study of radar rays in the solar corona* as reported in Sect. XXII-A.

We will make no attempt to give a "deep" physical argument as to the reality of this effect. It is sufficient here to say that it is a mathematical consequence of Einstein's equations. From the point of view of classical physics, one can say that from the spherical symmetry of the gravitational index of refraction which is everywhere greater than unity a light ray will be bent toward the Sun and, that propagation between two points in the solar system will "take longer" than one would expect. The theory that has been developed has been applied to the angular bending of light rays as they pass near the Sun. Our application and treatment, however, is apparently new.

3. Propagation Times

Specifically, our problem is to compute the propagation time between the Earth and Venus in the presence of a medium of index of refraction $n(r)$ [Eq. (4)]. The propagation time is given by the line integral *along the ray trajectory* between the two bodies,

$$\Delta t = \frac{1}{c}\int_1^2 n(r)\, ds, \tag{5}$$

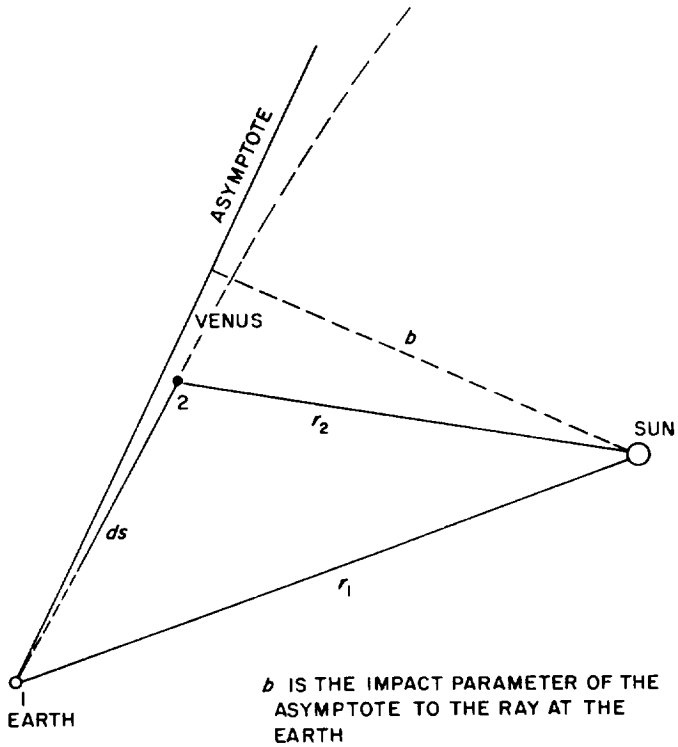


Fig. 7. Geometry of the heliocentric Earth–Venus system

where ds is the "Euclidian" line element given by the following expression in cylindrical coordinates:

$$ds^2 = dr^2 + r^2\, d\theta^2. \tag{6}$$

The angle θ is measured in the plane containing the Sun, Earth, and Venus at the Sun. The ray will remain in this plane due to the dependence on r alone in $n(r)$. The two-dimensional geometry is shown in Fig. 7, where b is the impact parameter of the asymptotic to the ray at the Earth.

The differential equation of the ray trajectory can be obtained from Fermat's principle as shown in Sect. XXII-A and in Ref. 3. In terms of a general $n(r)$ one obtains

$$\left(\frac{d\theta}{dr}\right)^2 = \frac{b^2}{r^4 n^2(r) - r^2 b^2}, \tag{7}$$

where b is a constant of integration which may be taken as the impact parameter of the ray whose definition is given in Fig. 7. From Eq. (6)

$$ds = \left[1 + r^2\left(\frac{d\theta}{dr}\right)^2\right]^{1/2} dr \tag{8}$$

which when inserted into Eq. (5) yields, after the substitution of Eqs. (4) and (7),

$$\Delta t = \frac{1}{c}\int_{r_1}^{r_2} \frac{(r + 4\alpha)\, dr}{[r^2 + 4\alpha r - b^2]^{1/2}}. \tag{9}$$

Eq. (9) is now a simple integral between r_1 and r_2 and can be carried out in closed form to yield

$$\Delta t = \frac{1}{c}[r_2^2 + 4\alpha r_2 - b^2]^{1/2} - \frac{1}{c}[r_1^2 + 4\alpha r_1 - b^2]^{1/2} + \frac{2\alpha}{c}\ln\left\{\frac{2[r_2^2 + 4\alpha r_2 - b^2]^{1/2} + 2r_2 + 4\alpha}{2[r_1^2 + 4\alpha r_1 - b^2]^{1/2} + 2r_1 + 4\alpha}\right\}. \tag{10}$$

The first two terms essentially represent the geometrical distance between the two planets and the third term a relativistic correction. Eq. (10) is exact to first order in α/r and is the solution of the problem if b is known. One can obtain b by finding the ray trajectory that passes through both the position of the Earth and that of Venus and satisfies Eq. (7). At any given instant of time, the angle θ_0 between the Earth and Venus is known from the planetary ephemerides, and we can integrate Eq. (7) after inserting the index of refraction, Eq. (4)

$$\theta_0 = -\int_{r_1}^{r_2} \frac{b\, dr}{r[r^2 + 4\alpha r - b^2]^{1/2}} \tag{11}$$

or

$$\theta_0 = \left[-\sin^{-1}\left\{\frac{2\alpha r - b^2}{r(2\alpha^2 + b^2)^{1/2}}\right\}\right]_{r=r_1}^{r=r_2}$$

Eq. (12) is a rather difficult transcendental equation for b which is best solved numerically for particular values of θ_0, r_1, and r_2.

4. Example

As a numerical example, we will take $r_1 = 1.00$ AU, $r_2 = 0.72$ AU, and $\theta_0 = 30$ deg (about 50 days on both sides of inferior conjunction). Then from Eq. (12): $b \simeq 0.69$; $2\Delta t$ (the round-trip propagation time) is, from Eq. (10), given by

$$2\Delta t = 521 + 12.6 \times 10^{-6} \text{ sec.}$$

The first term of the above equation represents the geometrical distance, and the second term represents the relativistic correction. At an ideal inferior conjunction $b = 0$ and Eq. (10) yields

$$2\Delta t = 290 + 6.0 \times 10^{-6} \text{ sec.}$$

If indeed our Venus range measurements are accurate to 5.0 μsec, the relativistic effect should be observable. However, the difficulty of separating this effect from the other errors in the planetary theory and ephemerides should not be underestimated. It is quite possible that the data on hand is sufficient for this check on Einstein's Theory of Gravitation, but an exhaustive analysis will be required to accomplish this goal.

Doppler measurements of the range change will also be affected by the gravitational index of refraction due to changes with time of the integral contained in Eq. (9). However, this effect is smaller than our *present* capability of measurement.

References

1. Allen, C. W., "Interpretation of Electron Densities from Corona Brightness," *Monthly Notices of The Royal Astronomical Society*, Vol. 107, pp. 426–430, 1947.

2. Fock, V., *The Theory of Space, Time, and Gravitation*, p. 221, The Macmillan Co., New York, 1964.

3. Born, M., and Wolf, E., *Principles of Optics*, p. 716, Pergamon Press, New York, 1959.

XXIII. Communications Systems Research: Detection and Filter Theory

A. Asymptotic Results for Optimum Equally Spaced Quantization of Gaussian Data

R. C. Titsworth

1. Introduction

In many data-transmission systems, analog input signals are first converted to discrete form at the transmitter, transmitted digitally over a channel, and reconstructed at the receiver as analog signals. The resulting output normally resembles the input process but is not precisely the same, since the quantization process at the transmitter introduces infidelity, and noise in the channel causes further corruption. Quantization for minimum distortion has been treated by the author (*SPS 37-24*, Vol. IV, pp. 196–200) and by Max (Ref. 1) assuming no noise in the channel. The former article indicated that uniform quantization of Gaussian data using a large number of quantization levels causes little extra distortion in the output. Based on the largest interval obtained for optimum quantization, the author placed a rather crude bound on performance.

The work of Max gives tables of optimum equally spaced quantization parameters for Gaussian data, up to 36 levels. This article gives an asymptotic treatment for these parameters when the number of levels is very large.

2. Quantization Distortion

Let $x(t)$ be a stationary, normally distributed, unit-variance random process, and for any value of $x(t)$, say x, we shall quantize as follows: For any preassigned even number L of levels, define quantization threshold q_k by

$$q_0 = -\infty$$

$$q_L = +\infty$$

$$q_k = \left(k - \frac{L}{2}\right)\Delta$$

for $k = 1, 2, \cdots, L-1$ and reconstruction values v_i by

$$v_i = \left(k - \frac{L}{2} - \frac{1}{2}\right)\Delta.$$

The spacing between levels, and between finite thresholds is Δ. The final finite threshold q_{L-1} we denote by q, given by

$$q = \left(\frac{L}{2} - 1\right)\Delta.$$

The final value v_L we denote by v,

$$v = \left(\frac{L-1}{2}\right)\Delta = q + \frac{\Delta}{2}.$$

The error between x and $v(x)$ is

$$\varepsilon = x - v(x)$$

whose mean-squared-value is a measure of the distortion or quantization noise present

$$N = \overline{\varepsilon^2} = \sum_{i=1}^{L} \int_{q_{i-1}}^{q_i} (x - v_i)^2 p(x)\, dx.$$

We have, for convenience, set $p(x) = [1/(2\pi)^{1/2}]\, e^{-x^2/2}$. When L is large, we see that approximately

$$N = \sum_{i=2}^{L-1} \int_{-\Delta/2}^{\Delta/2} x^2 p(v_i)\, dx + 2\int_q^\infty (x - v)^2 p(x)\, dx$$

$$N = \frac{1}{12}\Delta^2 \sum_{i=2}^{L-1} p(v_i)\Delta + 2\int_q^\infty (x - v)^2 p(x)\, dx$$

$$= \frac{1}{6}\Delta^2 \int_0^q p(x)\, dx + 2\int_q^\infty (x - v)^2 p(x)\, dx.$$

By introducing $M = (L/2) - 1$, the expression for N is

$$N = \frac{q^2}{6M^2}\int_0^q p(x)\, dx + 2\int_q^\infty \left(x - q\left(1 + \frac{1}{2M}\right)\right)^2 p(x)\, dx.$$

3. Minimization of Distortion

By choosing q properly, we can minimize N. This value of q is the solution to

$$\frac{\partial N}{\partial q} = \frac{q}{6M^2} - \frac{q^2 p(q)}{3M^2} - 4\left(1 + \frac{1}{2M}\right)p(q)$$

$$+ \left[4\left(1 + \frac{1}{2M}\right)^2 - \frac{1}{3M^2}\right] q\int_q^\infty p(x)\, dx = 0.$$

The integral in the expression above is approximately

$$\int_q^\infty p(x)\, dx = p(x)\left(\frac{1}{q} - \frac{1}{q^3}\right),$$

an approximation slightly underestimating the integral. The condition on q is then

$$\frac{q}{6M^2} = p(q)\left\{\frac{q^2}{3M^2} - \frac{2}{M}\left(1 + \frac{1}{3M}\right)\right.$$

$$\left. + \left(\frac{4}{q^2}\right)\left(1 + \frac{1}{M} + \frac{1}{6M^2}\right)\right\}$$

or equivalently, q is the solution to the transcendental equation

$$q^3 = 2p(q)\{q^4 - (6M + 2)q^2 + (12M^2 + 12M + 2)\}.$$

Solution of this equation for q using Newton's method on a digital computer is very easy, and yields the results in Table 1. When M is very large, the transcendental equation is approximately

$$q^3 = 24p(q)M^2$$

or, for a given value of q, L is approximately

$$L = 0.646q^{3/2}e^{q^2/4} + 1.$$

In the formula for N, the tail-quantization error is outweighed by the remainder of the terms, and can be omitted whenever L is large. Also,

$$\int_0^q p(x)\, dx \approx \tfrac{1}{2}.$$

This and the approximation above yield

$$N \approx \frac{2p(q)}{q}.$$

Max's table gives $q = 3.065$ for $L = 36$. The formula above gives $L = 37$, and the value in Table 1 is $L = 36$ for $q = 2.98$.

The quantization error in Max's table for $L = 36$ is 0.002843. The corresponding value of 0.00203 in Table 1 is only a slight underestimation of the error, and the value using the approximation above for N yields a value of 0.00315. The asymptotic values thus agree very well with the actual ones—certainly accurate enough for most engineering applications.

Table 1. Optimum uniform quantization

L	Δ	v	Error
8	6.74×10^{-1}	2.3622	1.15×10^{-2}
10	5.45×10^{-1}	2.4549	1.05×10^{-2}
12	4.61×10^{-1}	2.5373	9.04×10^{-3}
14	4.01×10^{-1}	2.6098	7.67×10^{-3}
16	3.56×10^{-1}	2.6739	6.53×10^{-3}
18	3.21×10^{-1}	2.7311	5.61×10^{-3}
20	2.92×10^{-1}	2.7825	4.87×10^{-3}
22	2.69×10^{-1}	2.8290	4.26×10^{-3}
24	2.49×10^{-1}	2.8716	3.76×10^{-3}
26	2.32×10^{-1}	2.9108	3.34×10^{-3}
28	2.18×10^{-1}	2.9470	2.99×10^{-3}
30	2.05×10^{-1}	2.9806	2.70×10^{-3}
32	1.94×10^{-1}	3.0120	2.44×10^{-3}
34	1.84×10^{-1}	3.0414	2.22×10^{-3}
36	1.75×10^{-1}	3.0691	2.03×10^{-3}
38	1.67×10^{-1}	3.0951	1.87×10^{-3}
40	1.59×10^{-1}	3.1198	1.72×10^{-3}
42	1.53×10^{-1}	3.1432	1.59×10^{-3}
44	1.47×10^{-1}	3.1655	1.48×10^{-3}
46	1.41×10^{-1}	3.1867	1.38×10^{-3}
48	1.36×10^{-1}	3.2070	1.28×10^{-3}
50	1.31×10^{-1}	3.2264	1.20×10^{-3}
52	1.27×10^{-1}	3.2450	1.13×10^{-3}
54	1.23×10^{-1}	3.2629	1.06×10^{-3}
56	1.19×10^{-1}	3.2800	1.00×10^{-3}
58	1.15×10^{-1}	3.2965	9.47×10^{-4}
60	1.12×10^{-1}	3.3125	8.95×10^{-4}
62	1.09×10^{-1}	3.3278	8.48×10^{-4}
64	1.06×10^{-1}	3.3427	8.05×10^{-4}
66	1.03×10^{-1}	3.3570	7.65×10^{-4}
68	1.00×10^{-1}	3.3709	7.28×10^{-4}
70	9.80×10^{-2}	3.3844	6.93×10^{-4}
72	9.57×10^{-2}	3.3975	6.61×10^{-4}
74	9.34×10^{-2}	3.4101	6.32×10^{-4}
76	9.12×10^{-2}	3.4224	6.04×10^{-4}
78	8.92×10^{-2}	3.4344	5.78×10^{-4}
80	8.72×10^{-2}	3.4461	5.54×10^{-4}
82	8.53×10^{-2}	3.4574	5.32×10^{-4}
84	8.35×10^{-2}	3.4684	5.10×10^{-4}
86	8.18×10^{-2}	3.4792	4.90×10^{-4}
88	8.02×10^{-2}	3.4897	4.72×10^{-4}
90	7.86×10^{-2}	3.5000	4.54×10^{-4}
92	7.71×10^{-2}	3.5101	4.38×10^{-4}
94	7.56×10^{-2}	3.5198	4.22×10^{-4}
96	7.43×10^{-2}	3.5294	4.07×10^{-4}
98	7.29×10^{-2}	3.5388	3.93×10^{-4}
100	7.16×10^{-2}	3.5479	3.80×10^{-4}

B. Optimum Quantization Through Noisy Channel

T. Nieh

1. Summary

When a band-limited analog signal from a stationary stochastic process is quantized, coded, and then transmitted over a noisy channel, it is perturbed both by the quantization process and the noise in the channel. The joint effect of these two sources of disturbance can be minimized by picking the parameters of the quantizer judiciously according to some prescribed criterion. Using either the criterion of: (1) maximizing the output signal-to-noise ratio (SNR) or (2) minimizing the mean-squared-difference between the input signal to the quantizer and the output of the optimum receiver when constant-distance codes are employed, it is found that the parameters for the respective optimum quantizers must satisfy the same basic set of recursive equations. The latter criterion, however, is a stronger one than the former. By means of a computer program, explicit numerical results have been obtained for two particular input processes: uniform and Gaussian statistics. It is noted that with increasing channel noise (or, what is the same thing, decreasing SNR in the channel), the quantization levels of the optimum quantizer tend to cluster rapidly around the mean of the input process. This fact indicates the strong dependence of the quantization process on the channel noise. A performance comparison is made between optimum quantizers and the ones obtained without taking into account the effect of channel noise. The optimum number of quantization levels is also obtained as a function of the channel SNR.

2. Introduction

The problem of optimally selecting the parameters of a quantizer under the assumption of errorless transmission of the quantized data has been analyzed by Max (Ref. 1), and asymptotic results are given by Titsworth in Sect. XXIII-A of this volume. This article treats essentially the same problem but with noise in the channel taken into account.

Fig. 1 represents the system under study. An analog stationary signal $x(t)$ with known amplitude statistics is uniformly sampled at the Nyquist rate, T seconds/sample. The sampled data at time t_j (i.e., $x(t_j)$) is fed into the L-level quantizer which sends out v_j if $x(t_j)$ lies between the threshold settings $q_{j-1} < v_j$ and $q_j > v_j$. The quantized output v_j is then coded and transmitted through

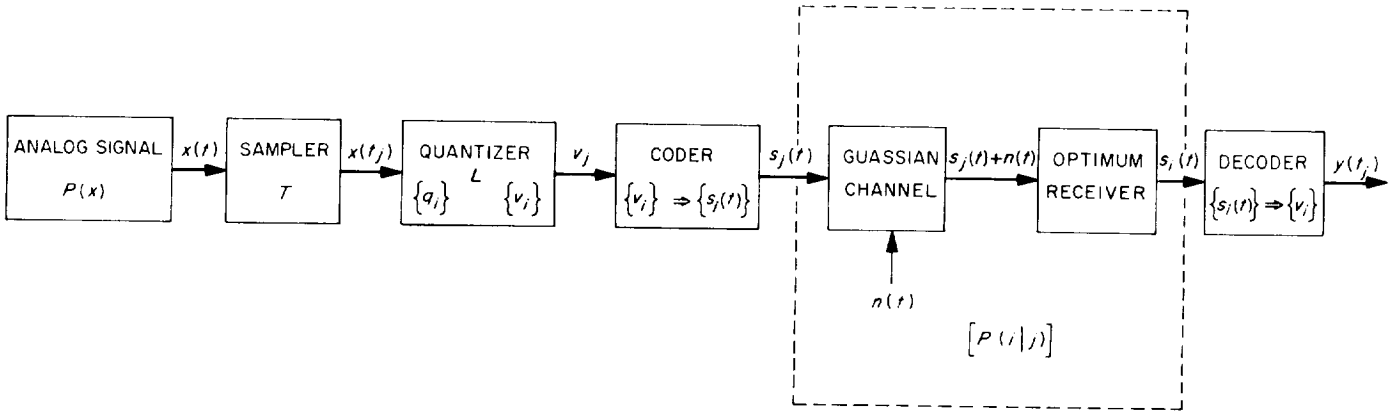


Fig. 1. Communication system under study

a Gaussian channel whose effect on the L coded signals can be described by a $L \times L$ matrix $P(i|j)$, where $P(i|j)$ denotes the probability of receiving the i^{th} signal when the j^{th} signal was actually transmitted. Both the codes and the receiver are tacitly assumed to be optimum under either coherent or noncoherent environment, whichever the case may be.

If the quantizer is constrained to have L levels, the problem is then to choose the L quantizing levels $\{v_i\}$ and $(L+1)$ threshold settings $\{q_i\}$ in such a way so as to satisfy some prescribed optimality criterion. Notice that in the error-free transmission case, the system performance (under any reasonable definition) is a strictly increasing function of the number of quantization levels L. However, with channel noise taken to be nonzero, the error rate in the channel increases with the number of signals used, and hence the over-all system performance is no longer a monotonic function of L. In fact, with a given noise level in the channel there exists an optimum number of quantization levels L_{opt} with which the system achieves its best performance.

Quantitative results of the above discussion will now follow.

3. Performance Criteria and the Derivation of Necessary Conditions For Optimality

Two performance criteria are chosen for this study:

(1) Minimization of mean-squared-difference between the quantizer input x and the receiver output y.

(2) Maximization of signal-to-noise ratio (SNR) at the output.

It will be seen shortly that, with SNR properly defined, the former criterion for optimality implies the latter.

a. Minimization of the mean-square-error. Without any loss of generality, the mean and variance of the input process are taken to be zero and one, respectively. Furthermore, for purpose of analysis, it is also assumed that the input process is symmetric [i.e., $p(x) = p(-x)$ where $p(x)$ denotes the amplitude probability density function of $x(t)$].

The mean-squared-error between x and y can thus be written as:

$$\begin{aligned}\epsilon^2 &= E[(x-y)^2] \\ &= E[x^2] - 2E[xy] + E[y^2] \\ &= 1 - 2\sum_{j=1}^{L}\int_{q_{j-1}}^{q_j} x\left(\sum_{i=1}^{L} v_i P(i|j)\right) p(x)\,dx \\ &\quad + \sum_{j=1}^{L}\int_{q_{j-1}}^{q_j}\left(\sum_{i=1}^{L} v_i^2 P(i|j)\right) p(x)\,dx. \qquad (1)\end{aligned}$$

By differentiating Eq. (1) with respect to both q_k, $k = 1, 2, \cdots, L-1$ (note that q_0 and q_L are always set equal to x_{min} and x_{max}, respectively), and v_k $(k = 1, 2, \cdots, L)$, and then by setting derivatives equal to zero, the following necessary conditions for optimality are obtained:

$$q_k = \frac{\sum_{i=1}^{L} v_i^2 (P(i|k) - P(i|k+1))}{2\left[\sum_{i=1}^{L} v_i (P(i|k) - P(i|k+1))\right]} \qquad (2)$$

for $(k = 1, 2, \cdots, L-1)$, and

$$v_k = \frac{\sum_{i=1}^{L} P(k|i)\, m_i}{\sum_{i=1}^{L} P(k|i)\, P_i} \tag{3}$$

for $(k = 1, 2, \cdots, L)$, where

$$m_i \equiv \int_{q_{i-1}}^{q_i} xp(x)\, dx$$

$$P_i \equiv \int_{q_{i-1}}^{q_i} p(x)\, dx.$$

4. Constant-Distance Coding

If a constant-distance coding scheme (such as a regular simplex code or an orthogonal code for coherent and noncoherent reception, respectively) is employed, the conditional probabilities $P(i|k)$ $(i = 1, 2, \cdots, L;$ $k = 1, 2, \cdots, L)$ characterizing the Gaussian channel are given by the following expressions:

For coherent reception using a Regular Simplex Code (Ref. 2):

$$P(i|i) \equiv P = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y+A} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}t^2}\, dt \right]^{L-1} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}y^2}\, dy \tag{4}$$

for $i = 1, 2, \cdots, L$, where

$$A = \left[\frac{2STL}{N_0 (L-1)} \right]^{1/2}$$

$\frac{S}{N_0}$ = SNR per cycle bandwidth in the channel

T = sampling rate or duration of the coded signal

$$P(i|k) \underset{i \neq k}{\equiv} q = \frac{1-P}{1-L}.$$

For noncoherent reception (using orthogonal signals) (Ref. 3):

$$P(i|i) \equiv P = 1 - \frac{e^{-ST/N_0}}{L} \sum_{r=1}^{L} (-1)^r \binom{L}{r} e^{-ST/N_0 r} \tag{5}$$

and $P(i|k) \equiv q = (1-P)/(1-L)$ for $i \neq k$.

With this set of signals, Eqs. (2) and (3) can be simplified to the following:

$$q_k = \frac{(P-q)(v_k^2 - v_{k+1}^2)}{2(P-q)(v_k - v_{k+1})} = \frac{v_k + v_{k+1}}{2} \tag{6}$$

$$v_k = \frac{m_k P + q \sum_{j \neq k} m_j}{P_k P + q \sum_{j \neq k} P_k} = \frac{m_k (P-q)}{P_k (P-q) + q} \tag{7}$$

$$= \frac{m_k}{P_k + Q}$$

where $Q \equiv q/P - q$.

Eqs. (6) and (7) constitute the necessary recursive relations between the q_k's and v_k's which must be satisfied in order to achieve optimality. With the optimum quantizer, it can easily be shown that the mean-squared-error is given by:

$$\varepsilon_{opt}^2 = 1 - (p-q) \sum_{i=1}^{L} v_i m_i = 1 - (p-q) \sum_{i=1}^{L} \frac{m_i^2}{P_i + Q}. \tag{8}$$

The method of obtaining optimum parameters from Eqs. (6) and (7) will be elucidated later.

5. Maximization of the Signal-to-Noise Ratio

Since the noise of the over-all system consists of the joint effect of the quantization process and the noisy channel, a meaningful definition of system noise N in this case is: *that part of the output y which is uncorrelated to the input x.* The output y can thus be expressed as:

$$y = \rho x + N \tag{9}$$

where $\rho = E[xy]/E[x^2]$. The signal-to-noise ratio of the system is then equal to:

$$SNR = \frac{\rho^2 E[x^2]}{E[N^2]} = \frac{\rho^2 E[x^2]}{E[y^2] - \rho^2 E[x^2]}. \tag{10}$$

The necessary conditions for the maximization of SNR will now be derived under the same set of assumptions used for the previous criterion of minimizing the mean-squared-error. Differentiating Eq. (10) with respect to

the q_k's and v_k's and setting the derivatives equal to zero, the following expressions are obtained:

$$\left.\begin{aligned}\frac{E[y^2]}{\rho} &= \frac{\partial E[y^2]/\partial q_k}{2\dfrac{\partial \rho}{\partial q_k}} \\ \frac{E[y^2]}{\rho} &= \frac{\partial E[y^2]/\partial v_k}{2\dfrac{\partial \rho}{\partial v_k}}\end{aligned}\right\} \tag{11}$$

The necessary conditions for optimality follow after straightforward (but lengthy) substitution:

$$\frac{E[y^2]}{\rho} = \frac{\sum_{i=1}^{L} v_i^2 (P_i + Q)}{\sum_{i=1}^{L} v_i m_i} = \frac{v_k + v_{k+1}}{2q_k} = \frac{v_k}{m_k}(P_k + Q)$$

$$\text{for all } k \tag{12}$$

Notice $E[y^2]/\rho$ is a *number* (i.e., dimensionless); let it be denoted by ϕ. Eq. (12) can be equivalently rewritten as the following system of two equations:

$$2q_k = \frac{m_k}{P_k + Q} + \frac{m_{k+1}}{P_k + Q} \tag{13}$$

$$v_k = \frac{m_k}{P_k + Q}\phi \tag{14}$$

It is easy to see from Eqs. (14) and (10) that the output *SNR* is independent of ϕ. Hence, the value for ϕ can be *arbitrary* if *only* the *SNR* is to be maximized. However, if the criterion of minimizing the mean-squared-error is *also* imposed, Eq. (13) is still necessary [Eqs. (6) and (7)], and *in addition*, ϕ is constrained to be exactly equal to one. This shows that: *maximizing SNR is a necessary condition for minimizing the mean-squared-error*, or equivalently, *the minimization of mean-squared-error implies the maximization of output SNR.*

The output *SNR* using this optimum quantizer can be easily verified to be

$$(SNR)_{opt} = \frac{\rho_{opt}}{1 - \rho_{opt}} = \frac{(P - q)\sum_{i=1}^{L} v_i m_i}{1 - (P - q)\sum_{i=1}^{L} v_i m_i}. \tag{15}$$

A comparison with Eq. (8) yields the following simple relationship between $(SNR)_{opt}$ and ε^2_{opt}:

$$(SNR)_{opt} = \frac{1}{\varepsilon^2_{opt}} - 1. \tag{16}$$

6. Numerical Solutions of Necessary Conditions for Optimality

Having shown that the minimization of mean-squared-error implies the maximization of *SNR*, the former criterion is thus sufficient to obtain quantizers optimum with respect to both criteria. Hence, only the mean-squared-error criterion need be considered.

Owing to the fact that Eqs. (13) and (14) are not easily accessible by any analytical means, they must be treated numerically. A computer algorithm thus becomes necessary to solve for the optimum parameters. A brief description of an algorithm developed for the purpose of this study is outlined below for the case of L even; the procedure for L odd is similar.

For L even, the three initial conditions available to the set of recursive equations are: $q_0 = x_{min}$, $q_{L/2} = 0$, and $q_L = x_{max}$, respectively. With regard to Eq. (13), any given $p(x)$ is an implicit function of q_{k-1}, q_k, and q_{k+1}. Any third variable can be found if the other two are given. It is clear that since the input distribution is assumed to be symmetrical, the parameters should likewise be symmetrically located. Hence, only half of them need to be found. One way of solving Eq. (13) successively is to first pick q_{L-1} arbitrarily, and with $q_L = x_{max}$ then given, q_{L-2} can be obtained. Each succeeding q_i is then obtained from the previous q_{i+1} and q_{i+2} until $q_{L/2}$ is reached. By comparing $q_{L/2}$ with 0, the initial choice of q_{L-1} can then be readjusted systematically since all subsequent q_i's $(i = L - 2, L - 3, \cdots, L/2)$ are monotonically increasing functions of q_{L-1}. This process is iterated until the difference between the computed $q_{L/2}$ and 0 is less than some preset tolerance. Having obtained the optimum q_i's, it is a simple matter to calculate the optimum v_i's by means of Eq. (14) with $\phi = 1$.

7. Summary of Results and Conclusions

Parameters of optimum quantizers have been obtained as a function of q with various numbers of output levels L for two particular input signals of interest, viz., the uniform and Gaussian processes. Optimum quantizers for the uniform case are tabulated in Table 2 for $L = 2(2)10$, and various q. The results have also been obtained for $L = 20(10)70$, but are not shown. Table 3

Table 2. Parameters for optimum quantizers with uniform input distribution

$$\begin{pmatrix} P(x) = 1/[2(3)^{1/2}]\ ; & -3^{1/2} \leq x \leq 3^{1/2} \\ = 0 \qquad\quad ; & \text{elsewhere} \end{pmatrix}$$

	$q = 10^{-2}$		$q = 10^{-3}$		$q = 10^{-4}$		$q = 10^{-5}$		$q = 0$	
	q_i	v_i	q_i	v_i	q_i	v_i	q_i	v_i	q_i	v_i
$L = 2; i = 1$	0.0		0.0						0.0	
2	1.732×10^{0}	8.635×10^{-1}	1.732×10^{0}	8.658×10^{-1}					1.732×10^{0}	8.660×10^{-1}
$L = 4; i = 2$	0.0		0.0						0.0	
3	8.015×10^{-1}	3.981×10^{-1}	8.592×10^{-1}	4.293×10^{-1}					8.66×10^{-1}	4.33×10^{-1}
4	1.732×10^{0}	1.260×10^{0}	1.732×10^{0}	1.295×10^{0}					1.732×10^{0}	1.299×10^{0}
$L = 6; i = 3$	0.0		0.0						0.0	
4	4.249×10^{-1}	2.098×10^{-1}	5.593×10^{-1}	2.794×10^{-1}					5.773×10^{-1}	2.887×10^{-1}
5	9.719×10^{-1}	6.917×10^{-1}	1.132×10^{0}	8.451×10^{-1}					1.155×10^{0}	8.66×10^{-1}
6	1.732×10^{0}	1.343×10^{0}	1.732×10^{0}	1.431×10^{0}					1.732×10^{0}	1.443×10^{0}
$L = 8; i = 4$	0.0		0.0						0.0	
5	1.993×10^{-1}	9.702×10^{-2}	4.001×10^{-1}	1.998×10^{-1}					4.33×10^{-1}	2.165×10^{-1}
6	5.073×10^{-1}	3.472×10^{-1}	8.138×10^{-1}	6.062×10^{-1}					8.66×10^{-1}	6.495×10^{-1}
7	9.994×10^{-1}	7.451×10^{-1}	1.254×10^{0}	1.033×10^{0}					1.299×10^{0}	1.083×10^{0}
8	1.732×10^{0}	1.356×10^{0}	1.732×10^{0}	1.491×10^{0}					1.732×10^{0}	1.516×10^{0}
$L = 10; i = 5$	0.0		0.0		0.0		0.0		0.0	
6	6.742×10^{-2}	3.115×10^{-2}	2.959×10^{-1}	1.477×10^{-1}	3.409×10^{-1}	1.704×10^{-1}	3.459×10^{-1}	1.729×10^{-1}	3.464×10^{-1}	1.732×10^{-1}
7	2.175×10^{-1}	1.860×10^{-1}	6.054×10^{-1}	4.499×10^{-1}	6.832×10^{-1}	5.120×10^{-1}	6.919×10^{-1}	5.188×10^{-1}	6.928×10^{-1}	5.196×10^{-1}
8	5.103×10^{-1}	3.571×10^{-1}	9.410×10^{-1}	7.720×10^{-1}	1.028×10^{0}	8.556×10^{-1}	1.038×10^{0}	8.650×10^{-1}	1.039×10^{0}	8.66×10^{-1}
9	9.980×10^{-1}	7.457×10^{-1}	1.313×10^{0}	1.126×10^{0}	1.377×10^{0}	1.203×10^{0}	1.385×10^{0}	1.211×10^{0}	1.386×10^{0}	1.212×10^{0}
10	1.732×10^{0}	1.355×10^{0}	1.732×10^{0}	1.521×10^{0}	1.732×10^{0}	1.554×10^{0}	1.732×10^{0}	1.558×10^{0}	1.732×10^{0}	1.559×10^{0}

Table 3. Parameters for optimum quantizers with Gaussian input process $[P(x) = 1/(2\pi)^{1/2}\, e^{-x^2/2}]$

	Optimum quantizers								Optimum uniform quantizer	
	$q = 10^{-2}$		$q = 10^{-3}$		$q = 10^{-4}$		$q = 0$		$q = 0$	
	q_i	v_i	q_i	v_i	q_i	v_i	q_i	v_i	q_i	v_i
$L = 4; i = 2$	0.0		0.0		0.0		0.0		0.0	
3	8.742×10^{-1}	3.968×10^{-1}	9.689×10^{-1}	4.465×10^{-1}	9.803×10^{-1}	4.522×10^{-1}	9.816×10^{-1}	4.528×10^{-1}	9.957×10^{-1}	4.979×10^{-1}
4	$+\infty$	1.352×10^{0}	$+\infty$	1.491×10^{0}	$+\infty$	1.508×10^{0}	$+\infty$	1.510×10^{0}	$+\infty$	1.493×10^{0}
$L = 6; i = 3$	0.0		0.0		0.0		0.0		0.0	
4	3.558×10^{-1}	6.720×10^{-2}	6.317×10^{-1}	3.042×10^{-1}	6.561×10^{-1}	3.163×10^{-1}	6.589×10^{-1}	3.177×10^{-1}	7.334×10^{-1}	3.667×10^{-1}
5	1.053×10^{0}	6.445×10^{-1}	1.392×10^{0}	9.593×10^{-1}	1.441×10^{0}	9.958×10^{-1}	1.447×10^{0}	1.00×10^{0}	1.467×10^{0}	1.100×10^{0}
6	$+\infty$	1.461×10^{0}	$+\infty$	1.825×10^{0}	$+\infty$	1.886×10^{0}	$+\infty$	1.894×10^{0}	$+\infty$	1.834×10^{0}
$L = 8; i = 4$	0.0		0.0		0.0		0.0		0.0	
5	2.414×10^{-1}	1.079×10^{-1}	4.562×10^{-1}	2.229×10^{-1}	4.955×10^{-1}	2.426×10^{-1}	5.006×10^{-1}	2.451×10^{-1}	5.860×10^{-1}	2.930×10^{-1}
6	5.821×10^{-1}	3.75×10^{-1}	9.615×10^{-1}	6.894×10^{-1}	1.040×10^{0}	7.484×10^{-1}	1.050×10^{0}	7.560×10^{-1}	1.172×10^{0}	8.79×10^{-1}
7	1.153×10^{0}	7.892×10^{-1}	1.617×10^{0}	1.233×10^{0}	1.732×10^{0}	1.331×10^{0}	1.748×10^{0}	1.344×10^{0}	1.758×10^{0}	1.465×10^{0}
8	$+\infty$	1.516×10^{0}	$+\infty$	2.00×10^{0}	$+\infty$	2.133×10^{0}	$+\infty$	2.152×10^{0}	$+\infty$	2.051×10^{0}
$L = 12; i = 6$	0.0		0.0		0.0		0.0		0.0	
7	2.790×10^{-2}	6.920×10^{-3}	2.598×10^{-1}	1.280×10^{-1}	3.293×10^{-1}	1.631×10^{-1}	3.401×10^{-1}	1.684×10^{-1}	4.238×10^{-1}	2.119×10^{-1}
8	1.059×10^{-1}	4.894×10^{-2}	5.364×10^{-1}	3.917×10^{-1}	6.724×10^{-1}	4.955×10^{-1}	6.943×10^{-1}	5.119×10^{-1}	8.476×10^{-1}	6.357×10^{-1}
9	2.762×10^{-1}	1.628×10^{-1}	8.513×10^{-1}	6.811×10^{-1}	1.048×10^{0}	8.492×10^{-1}	1.081×10^{0}	8.768×10^{-1}	1.271×10^{0}	1.069×10^{0}
10	5.883×10^{-1}	3.896×10^{-1}	1.241×10^{0}	1.021×10^{0}	1.489×10^{0}	1.246×10^{0}	1.534×10^{0}	1.286×10^{0}	1.695×10^{0}	1.493×10^{0}
11	1.148×10^{0}	7.870×10^{-1}	1.798×10^{0}	1.460×10^{0}	2.082×10^{0}	1.731×10^{0}	2.141×10^{0}	1.783×10^{0}	2.119×10^{0}	1.917×10^{0}
12	$+\infty$	1.508×10^{0}	$+\infty$	2.136×10^{0}	$+\infty$	2.433×10^{0}	$+\infty$	2.499×10^{0}	$+\infty$	2.341×10^{0}
$L = 16; i = 8$			0.0		0.0		0.0		0.0	
9			1.458×10^{-1}	7.153×10^{-2}	2.403×10^{-1}	1.195×10^{-1}	2.582×10^{-1}	1.284×10^{-1}	3.352×10^{-1}	1.676×10^{-1}
10			3.026×10^{-1}	2.201×10^{-1}	4.864×10^{-1}	3.611×10^{-1}	5.224×10^{-1}	3.881×10^{-1}	6.704×10^{-1}	5.028×10^{-1}
11			4.816×10^{-1}	3.851×10^{-1}	7.452×10^{-1}	6.116×10^{-1}	7.996×10^{-1}	6.568×10^{-1}	1.006×10^{0}	8.380×10^{-1}
12			6.954×10^{-1}	5.781×10^{-1}	1.026×10^{0}	8.788×10^{-1}	1.099×10^{0}	9.424×10^{-1}	1.341×10^{0}	1.173×10^{0}
13			9.618×10^{-1}	8.128×10^{-1}	1.345×10^{0}	1.174×10^{0}	1.437×10^{0}	1.256×10^{0}	1.676×10^{0}	1.508×10^{0}
14			1.314×10^{0}	1.111×10^{0}	1.732×10^{0}	1.516×10^{0}	1.844×10^{0}	1.618×10^{0}	2.011×10^{0}	1.944×10^{0}
15			1.842×10^{0}	1.517×10^{0}	2.269×10^{0}	1.947×10^{0}	2.401×10^{0}	2.069×10^{0}	2.346×10^{0}	2.280×10^{0}
16			$+\infty$	2.167×10^{0}	$+\infty$	2.591×10^{0}	$+\infty$	2.733×10^{0}	$+\infty$	2.615×10^{0}

Table 3. Parameters for optimum quantizers with Gaussian input process [$P(x) = 1/(2\pi)^{1/2} e^{-x^2/2}$] (cont'd)

	Optimum quantizers								Optimum uniform quantizer	
	$q = 10^{-2}$		$q = 10^{-3}$		$q = 10^{-4}$		$q = 0$		$q = 0$	
	q_i	v_i	q_i	v_i	q_i	v_i	q_i	v_i	q_i	v_i
$L = 20; i = 10$			0.0		0.0		0.0		0.0	
11			7.174×10^{-2}	3.465×10^{-2}	1.827×10^{-1}	9.102×10^{-2}	2.083×10^{-1}	1.038×10^{-1}	2.788×10^{-1}	1.394×10^{-1}
12			1.529×10^{-1}	1.088×10^{-1}	3.684×10^{-1}	2.744×10^{-1}	4.197×10^{-1}	3.128×10^{-1}	5.576×10^{-1}	4.182×10^{-1}
13			2.519×10^{-1}	1.970×10^{-1}	5.607×10^{-1}	4.625×10^{-1}	6.375×10^{-1}	5.265×10^{-1}	8.364×10^{-1}	6.970×10^{-1}
14			3.757×10^{-1}	3.067×10^{-1}	7.638×10^{-1}	6.590×10^{-1}	8.661×10^{-1}	7.486×10^{-1}	1.115×10^{0}	9.758×10^{-1}
15			5.316×10^{-1}	4.447×10^{-1}	9.834×10^{-1}	8.686×10^{-1}	1.111×10^{0}	9.837×10^{-1}	1.394×10^{0}	1.255×10^{0}
16			7.293×10^{-1}	6.186×10^{-1}	1.228×10^{0}	1.098×10^{0}	1.381×10^{0}	1.239×10^{0}	1.672×10^{0}	1.533×10^{0}
17			9.841×10^{-1}	8.399×10^{-1}	1.512×10^{0}	1.360×10^{0}	1.690×10^{0}	1.524×10^{0}	1.951×10^{0}	1.812×10^{0}
18			1.328×10^{0}	1.128×10^{0}	1.865×10^{0}	1.667×10^{0}	2.068×10^{0}	1.857×10^{0}	2.230×10^{0}	2.091×10^{0}
19			1.850×10^{0}	1.527×10^{0}	2.369×10^{0}	2.064×10^{0}	2.594×10^{0}	2.279×10^{0}	2.508×10^{0}	2.370×10^{0}
20			$+\infty$	2.172×10^{0}	$+\infty$	2.674×10^{0}	$+\infty$	2.908×10^{0}	$+\infty$	2.649×10^{0}
$L = 26; i = 13$			0.0		0.0		0.0		0.0	
14			1.177×10^{-2}	4.866×10^{-3}	1.245×10^{-1}	6.206×10^{-2}	1.616×10^{-1}	8.06×10^{-2}	2.240×10^{-1}	1.12×10^{-1}
15			3.067×10^{-2}	1.867×10^{-2}	2.505×10^{-1}	1.868×10^{-1}	3.245×10^{-1}	2.425×10^{-1}	4.48×10^{-1}	3.36×10^{-1}
16			6.173×10^{-2}	4.266×10^{-2}	3.800×10^{-1}	3.142×10^{-1}	4.905×10^{-1}	4.066×10^{-1}	6.72×10^{-1}	5.60×10^{-1}
17			1.088×10^{-1}	8.081×10^{-2}	5.148×10^{-1}	4.458×10^{-1}	6.61×10^{-1}	5.743×10^{-1}	8.96×10^{-1}	7.84×10^{-1}
18			1.755×10^{-1}	1.368×10^{-1}	6.572×10^{-1}	5.838×10^{-1}	8.383×10^{-1}	7.477×10^{-1}	1.12×10^{0}	1.008×10^{0}
19			2.658×10^{-1}	2.142×10^{-1}	8.101×10^{-1}	7.307×10^{-1}	1.025×10^{0}	9.289×10^{-1}	1.344×10^{0}	1.232×10^{0}
20			3.842×10^{-1}	3.174×10^{-1}	9.771×10^{-1}	8.895×10^{-1}	1.224×10^{0}	1.121×10^{0}	1.568×10^{0}	1.456×10^{0}
21			5.367×10^{-1}	4.511×10^{-1}	1.164×10^{0}	1.065×10^{0}	1.442×10^{0}	1.328×10^{0}	1.792×10^{0}	1.680×10^{0}
22			7.320×10^{-1}	6.223×10^{-1}	1.378×10^{0}	1.262×10^{0}	1.685×10^{0}	1.556×10^{0}	2.016×10^{0}	1.904×10^{0}
23			9.853×10^{-1}	8.417×10^{-1}	1.633×10^{0}	1.493×10^{0}	1.968×10^{0}	1.814×10^{0}	2.24×10^{0}	2.128×10^{0}
24			1.328×10^{0}	1.129×10^{0}	1.960×10^{0}	1.774×10^{0}	2.318×10^{0}	2.121×10^{0}	2.464×10^{0}	2.352×10^{0}
25			1.849×10^{0}	1.527×10^{0}	2.437×10^{0}	2.145×10^{0}	2.811×10^{0}	2.514×10^{0}	2.688×10^{0}	2.576×10^{0}
26			$+\infty$	2.171×10^{0}	$+\infty$	2.728×10^{0}	$+\infty$	3.109×10^{0}	$+\infty$	2.80×10^{0}

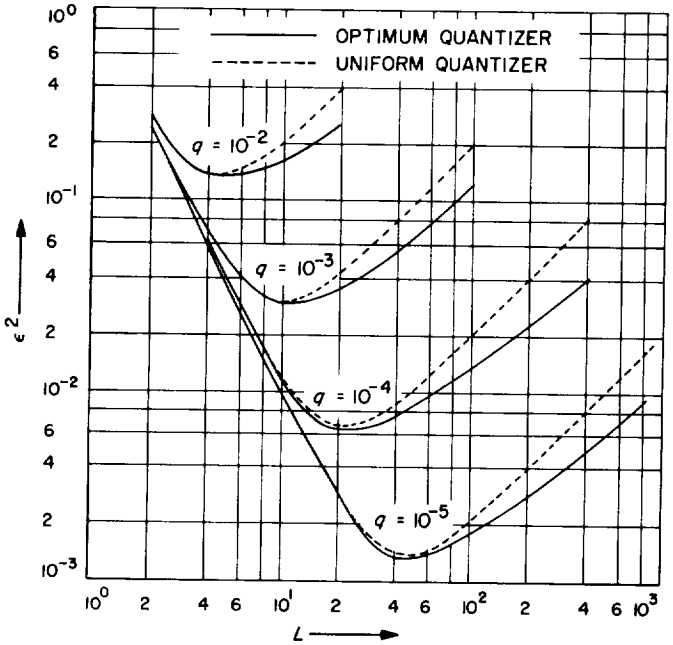


Fig. 2. Mean-square-error versus number of quantization levels with uniform input distribution

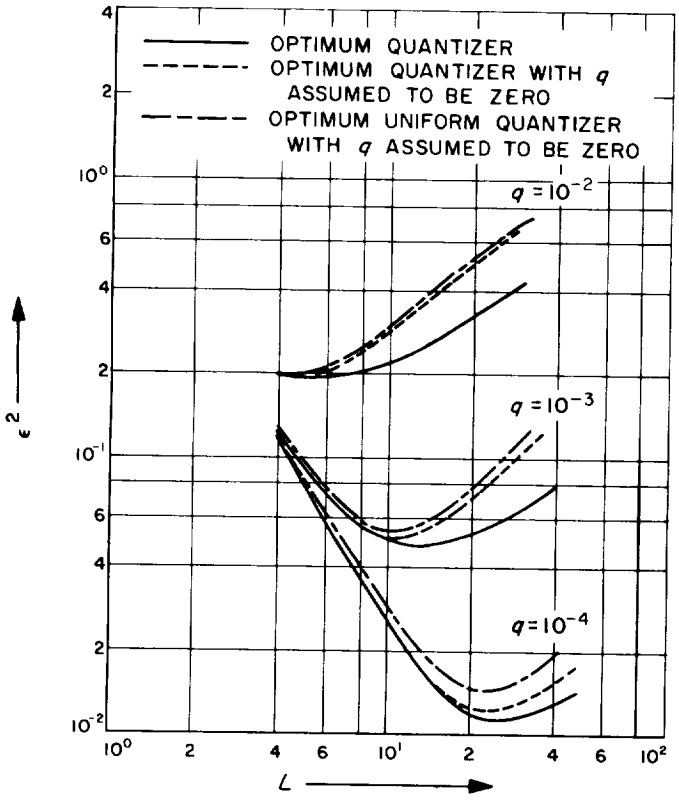


Fig. 3. Mean-square-error versus number of quantization levels with Gaussian input distribution

gives similar results for the Gaussian case $L = 4, 6, 8, 12, 16$. Calculations for $L = 20$ and 26 have also been made but are not shown here. For purpose of comparison, the results obtained by Max and Titsworth without taking the channel noise into account are also shown. Take particular note on the "clustered in" phenomenon on the parameter settings when the effect of the channel noise becomes large (i.e., for large L's and q's). Both (SNR) and (ϵ^2) have been calculated using the optimum quantizers and are plotted as functions of L with q as a parameter (Figs. 2 through 5). It is seen that the degradation in performance due to non-optimal quantization is relatively small at reasonable channel noise levels for L less than L_{opt} [the point at which (SNR) and ϵ^2 reach their respective maximum and minimum]. The degradation, however, increases rapidly with increasing L. Thus, if the system is designed to operate with the optimum

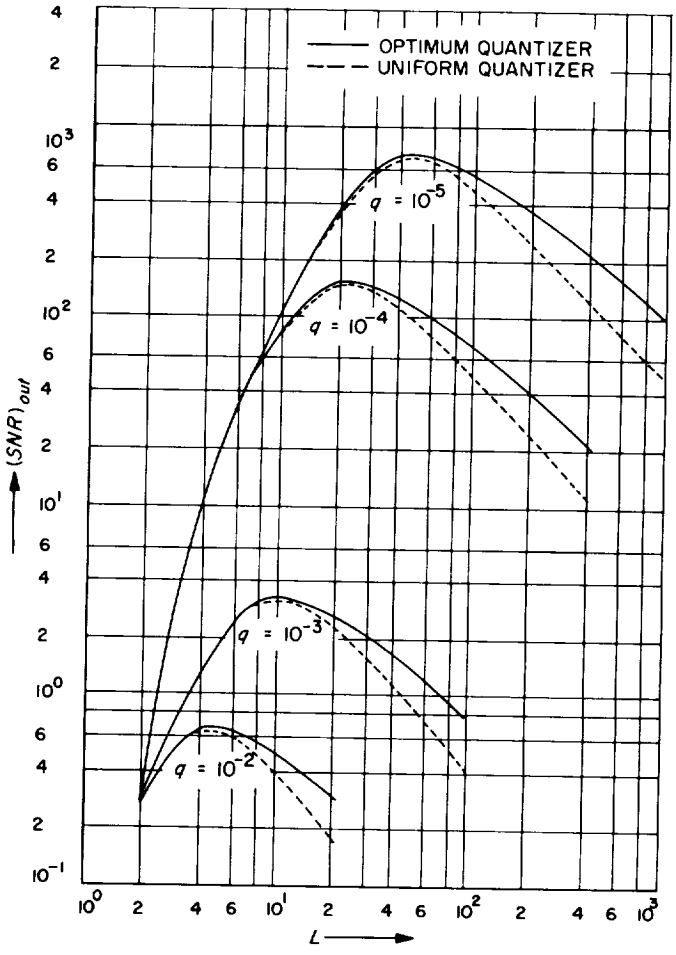


Fig. 4. Output signal-to-noise ratio versus number of quantization levels with uniform input distribution

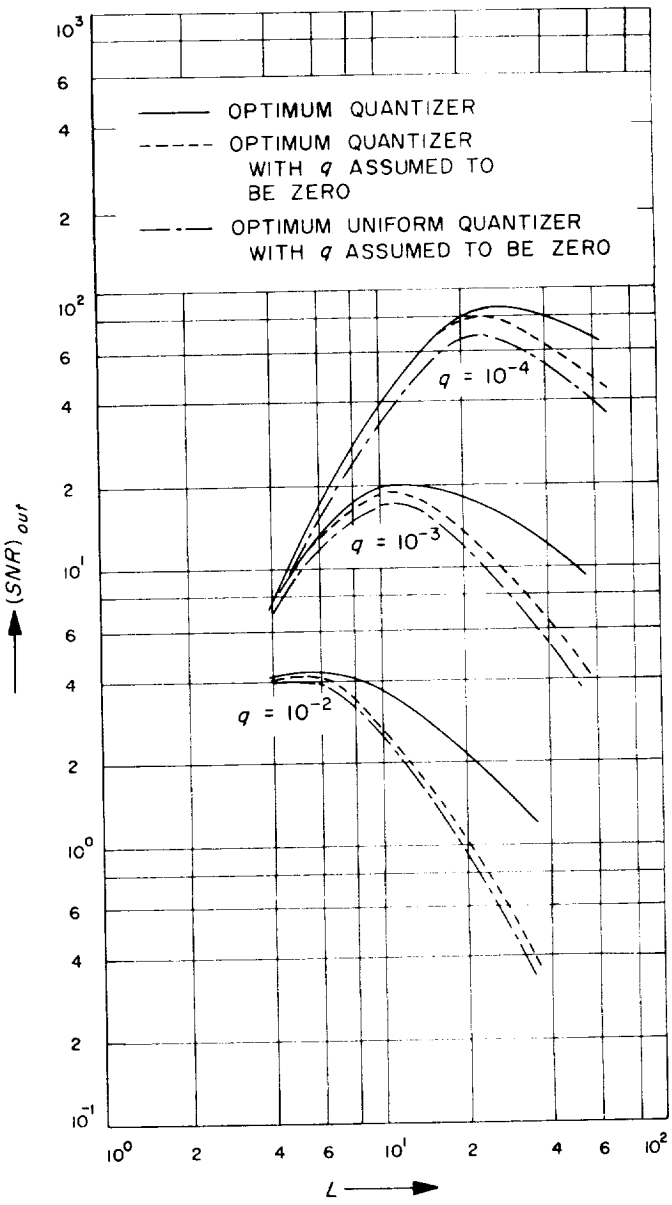


Fig. 5. Output signal-to-noise ratio versus number of quantization levels with Gaussian input distribution

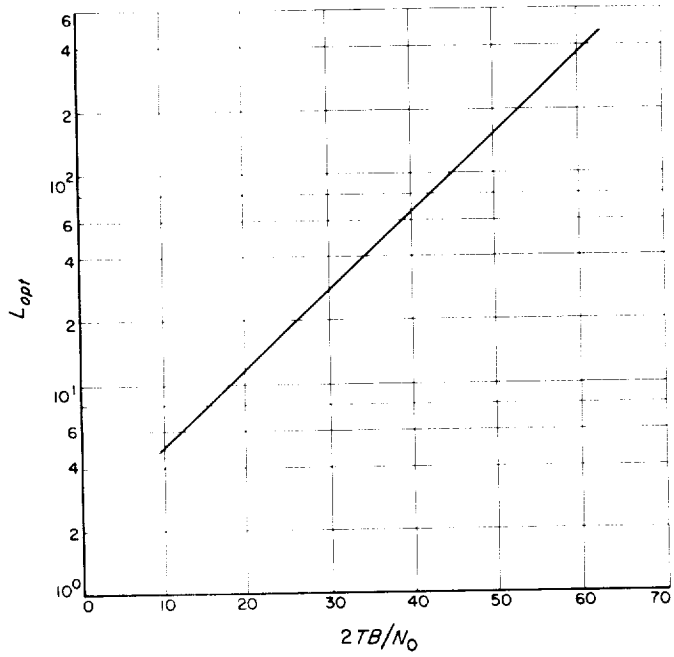


Fig. 6. Optimum number of quantization levels with uniform input distribution

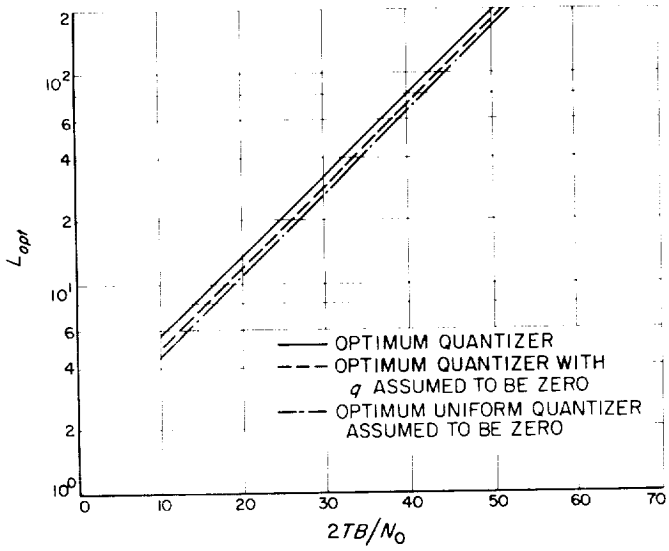


Fig. 7. Optimum number of quantization levels with Gaussian input distribution

number of quantization levels L_{opt}, it is actually of secondary importance in obtaining the exact optimum parameters. Figs. 6 and 7 show L_{opt} as a function of the physical parameter $2TB/N_0$, using the following upper bound for both Eqs. (4) and (5) (Refs. 3 and 4):

$$q < e^{-ST/2N_0}. \tag{17}$$

Notice for the uniform statistics case, where the performance gained by using an optimum quantizer is extremely small over that of a uniform quantizer (for $L \leqq L_{opt}$), L_{opt} is essentially the same using either quantizers. However, in the case of a Gaussian signal where the improvement due to optimal selection of parameters is more prominent, the optimum quantizer yields a higher value of L_{opt} than any other quantizer (Figs. 3 and 5). This observation is essentially attributed to the fact that with an optimum quantizer, more noise in the channel can be

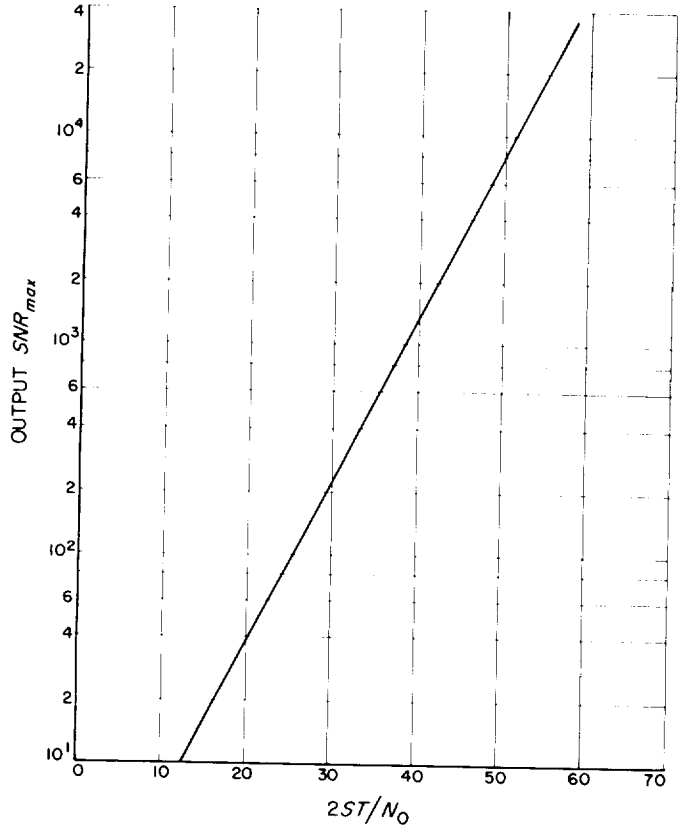


Fig. 8. Maximum output signal-to-noise ratio with uniform input distribution

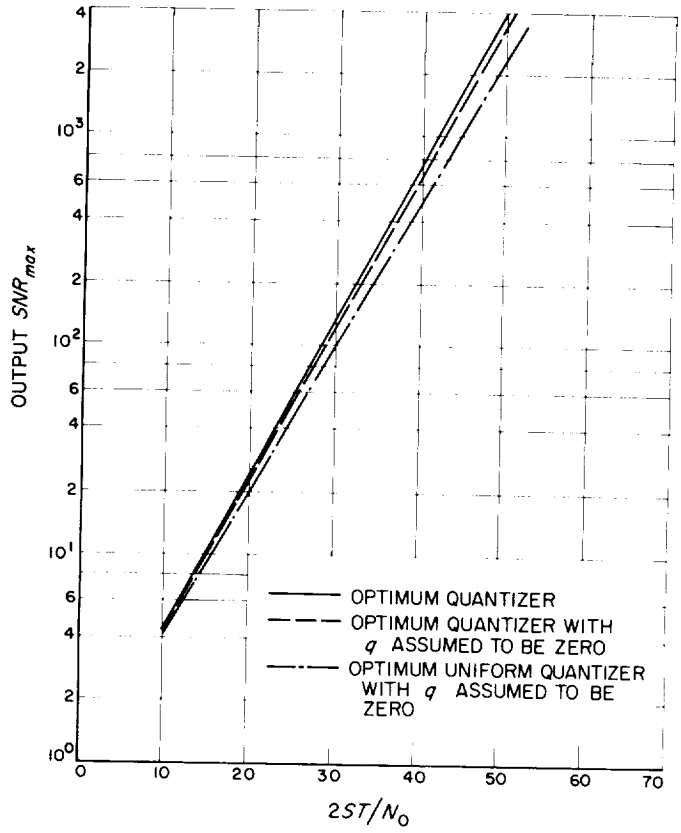


Fig. 9. Maximum output signal-to-noise ratio with Gaussian input distribution

tolerated. Corresponding plots (Figs. 8 and 9) of $(SNR)_{max}$ as a function of $2ST/N_0$ for both input processes are also included.

C. Effect of Bandwidth Constraints and Quantization on Coding for the Gaussian Channel

A. J. Viterbi

1. Introduction

It is well-known (Ref. 5) that, for a coherent white Gaussian channel without bandwidth constraints, it is possible to transmit a k bit sequence at rate R bits/sec by sending one of 2^k signals in such a way that the error probability for a k bit sequence is bounded by

$$K'2^{-k(C_\infty \alpha/R)} < P_E < K2^{-k(C_\infty \alpha/R)} \tag{1}$$

where

$$\alpha = \begin{cases} \frac{1}{2} - \frac{R}{C_\infty}, & \text{for } 0 \leq \frac{R}{C_\infty} \leq \frac{1}{4} \\ \left[1 - \left(\frac{R}{C_\infty}\right)^{1/2}\right]^2, & \text{for } \frac{1}{4} \leq \frac{R}{C_\infty} < 1 \end{cases}$$

$$C_\infty = \frac{S}{N_0 \ln 2}$$

and where S is the received signal power and N_0 is one-sided noise density. Also $K < 2$, while K' is a slowly varying function of R/C_∞. If the signals have n degrees of freedom, the bandwidth occupancy of the channel is

$$W = \frac{n}{2k} R. \tag{2}$$

This result is attained by using one of 2^k orthogonal signals to transmit a k bit sequence. This requires that the signals have $n = 2^k$ degrees of freedom which means that $W = (2^{k-1}/k)\,R$. Such signals can be generated as sequences of $n = 2^k$ binary signals so that the transmitted signal may be quantized to two levels. We note in particular that as k approaches infinity so does W in this case.

In the following parts of this article, we shall consider first the effect on the error probability of limiting the bandwidth. We shall also treat the effect of using two-level transmitted signals for a bandwidth constrained channel and the effect of quantizing the corresponding received signals to two levels, thus reducing the channel to a binary symmetric channel.

The effect of bandwidth constraints on the Gaussian channel has been studied by Shannon (Ref. 6) and the effect of quantization by Ziv (Ref. 7). However, the analysis was severely complicated by the restriction that all signals have exactly the same energy. By removing this restriction and imposing only an average power restriction, we shall obtain results much more readily.

2. Error Bounds for the Bandwidth-Constrained Gaussian Channel

Fano has shown (Ref. 5) that for very general time-discrete (sampled) channels, for a transmitted signal x and a received signal y, the error probability is bounded by

$$P_E < Ke^{-n\tilde{\alpha}}, \tag{3}$$

where $K < 2$, and

$$\tilde{\alpha} = \begin{cases} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q(x,y)\ln\left[\frac{q(x,y)}{p(x)\,p(y|x)}\right]dx\,dy \leq \tilde{\alpha}_c, \\ \qquad\qquad \text{for } \tilde{R}_c \leq \tilde{R},\ 0 \leq s \leq \frac{1}{4} \\ \tilde{\alpha}_c + \tilde{R}_c - \tilde{R} \qquad \text{for } 0 \leq \tilde{R} \leq \tilde{R}_c \end{cases}$$

$$\tilde{R} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q(x,y)\ln\left[\frac{q(x|y)}{p(x)}\right]dx\,dy, \qquad \text{for } \tilde{R}_c \leq \tilde{R},\ 0 \leq s \leq \frac{1}{2}$$

Here $p(x)$ and $p(y|x)$ are the transmitted signal probability density and the received signal conditional probability density, respectively, and

$$q(x,y) = q(y)\,q(x|y)$$

$$q(x|y) = \frac{p(x)\,p(y|x)^{1-s}}{\displaystyle\int_{-\infty}^{\infty} p(x)\,p(y|x)^{1-s}\,dx}$$

$$q(y) = \frac{\left[\displaystyle\int_{-\infty}^{\infty} p(x)\,p(y|x)^{1-s}\,dx\right]^{1/1-s}}{\displaystyle\int_{-\infty}^{\infty} dy\left[\int_{-\infty}^{\infty} p(x)\,p(y|x)^{1-s}\,dx\right]^{1/1-s}}$$

Also, $\tilde{R}_c = [\tilde{R}]_{s=\frac{1}{2}}$, $\tilde{\alpha}_c = [\tilde{\alpha}]_{s=\frac{1}{2}}$, and R is the rate in natural units (nats/symbol). Thus, to convert to the parameters of (1) for the continuous bandwidth-constrained channel, we must use

$$R\left(\frac{\text{bits}}{\text{sec}}\right) = \tilde{R}\left(\frac{\text{nats}}{\text{symbol}}\right)2W\left(\frac{\text{symbols}}{\text{sec}}\right)\frac{1}{\ln 2}\left(\frac{\text{bits}}{\text{nat}}\right)$$

$$= \frac{2W}{\ln 2}\,\tilde{R}\left(\frac{\text{bits}}{\text{sec}}\right). \tag{4}$$

Eq. (4) is based on the fact that, by transmitting a constant-amplitude sinusoidal pulse every $1/2W$ sec, it is possible to maintain the bandwidth occupancy (the minimum separation between similarly modulated channels) equal to W.

Also using Eqs. (2) and (4), we may rewrite Eq. (3) as

$$P_E < K2^{-k(C_\infty \alpha/R)}$$

where

$$\alpha = \frac{\tilde{\alpha}}{\ln 2}\left(\frac{n}{k}\right)\left(\frac{R}{C_\infty}\right) = \frac{\tilde{\alpha}}{\ln 2}\left(\frac{2W}{R}\right)\left(\frac{R}{C_\infty}\right) = \frac{2N_0W}{S}\tilde{\alpha}. \tag{5}$$

If there are no amplitude constraints on the transmitted signals other than that the ensemble average of the power be S watts, then we may generate the signals of bandwidth occupancy W by selecting from a Gaussian distribution of variance S the amplitudes of the pulses of duration of $1/2W$. Thus,

$$p(x) = \frac{e^{-x^2/2S}}{(2\pi S)^{1/2}}. \tag{6}$$

If the channel is Gaussian, the optimum receiver consists of a set of correlators. For the signals generated as described above, the equivalent of the correlation operation is to integrate over each pulse of duration $1/2W$ and then combine the successive integrations linearly. Thus, if we let y be the integrator output over a pulse duration, we have the conditional probability

$$p(y|x) = \frac{e^{-(y-x)^2/2N_0W}}{(2\pi N_0W)^{1/2}}, \tag{7}$$

and thus we have reduced the continuous channel to the time-discrete amplitude-continuous case whose performance is bounded by Eq. (3).

Substituting Eqs. (6) and (7) in the various expressions involved in Eq. (3), after considerable computation, we obtain

$$\tilde{\alpha} = \begin{cases} \dfrac{S}{2N_0W}\dfrac{s^2}{1+(1-s)S/N_0W}, & \text{for } \begin{cases} \tilde{R}_c \leq \tilde{R} \\ 0 \leq s \leq \dfrac{1}{2} \end{cases} \\ \tilde{\alpha}_c + \tilde{R}_c = \tilde{R}, & \text{for } 0 \leq \tilde{R} \leq \tilde{R}_c \end{cases}$$

$$\tilde{R} = \frac{1}{2}\ln\left[1+(1-s)(S/N_0W)\right] - \frac{S}{2N_0W}\frac{s(1-s)}{1+(1-s)(S/N_0W)} \quad \text{for } 0 \leq s \leq \frac{1}{2}$$

and

$$\tilde{\alpha}_c = \tilde{\alpha}|_{s=1/2} = \frac{S}{2N_0W}\frac{1/4}{1+\dfrac{S}{2N_0W}}$$

$$\tilde{R}_c = \tilde{R}|_{s=1/2} = \frac{1}{2}\ln\left(1+\frac{S}{2N_0W}\right) - \frac{S}{2N_0W}\frac{1/4}{1+\dfrac{S}{2N_0W}}. \tag{8}$$

Thus, from Eqs. (4), (5), and (8) we obtain

$$P_E < K2^{-k(C_\infty \alpha/R)} \tag{9}$$

where

$$\alpha = \begin{cases} \dfrac{s^2}{1+(1-s)S/N_0W}, & \text{for } R_c \leq R \\ \dfrac{N_0W}{S}\ln\left(1+\dfrac{S}{2N_0W}\right) - \dfrac{R}{C_\infty}, & \text{for } 0 \leq R \leq R_c \end{cases}$$

and

$$\frac{R}{C_\infty} = \frac{N_0W}{S}\ln\left[1+(1-s)\left(\frac{S}{N_0W}\right)\right] - \frac{s(1-s)}{1+(1-s)(S/N_0W)}, \quad \text{for } 0 \leq s \leq \frac{1}{2}$$

$$\frac{R_c}{C_\infty} = \frac{N_0W}{S}\ln\left[1+\frac{S}{2N_0W}\right] - \frac{1/4}{1+\dfrac{S}{2N_0W}}.$$

Thus, α is implicitly related to R by the parameter s. Note that the expressions of Eq. (9) approach the corresponding expressions of (1) in the limit at $W \to \infty$.

This α is shown in Fig. 10 as a function of R/C_∞ for $N_0W/S = 1, 2, 3, 5, 10$, and ∞. It appears that the performance for finite bandwidth rapidly approaches the

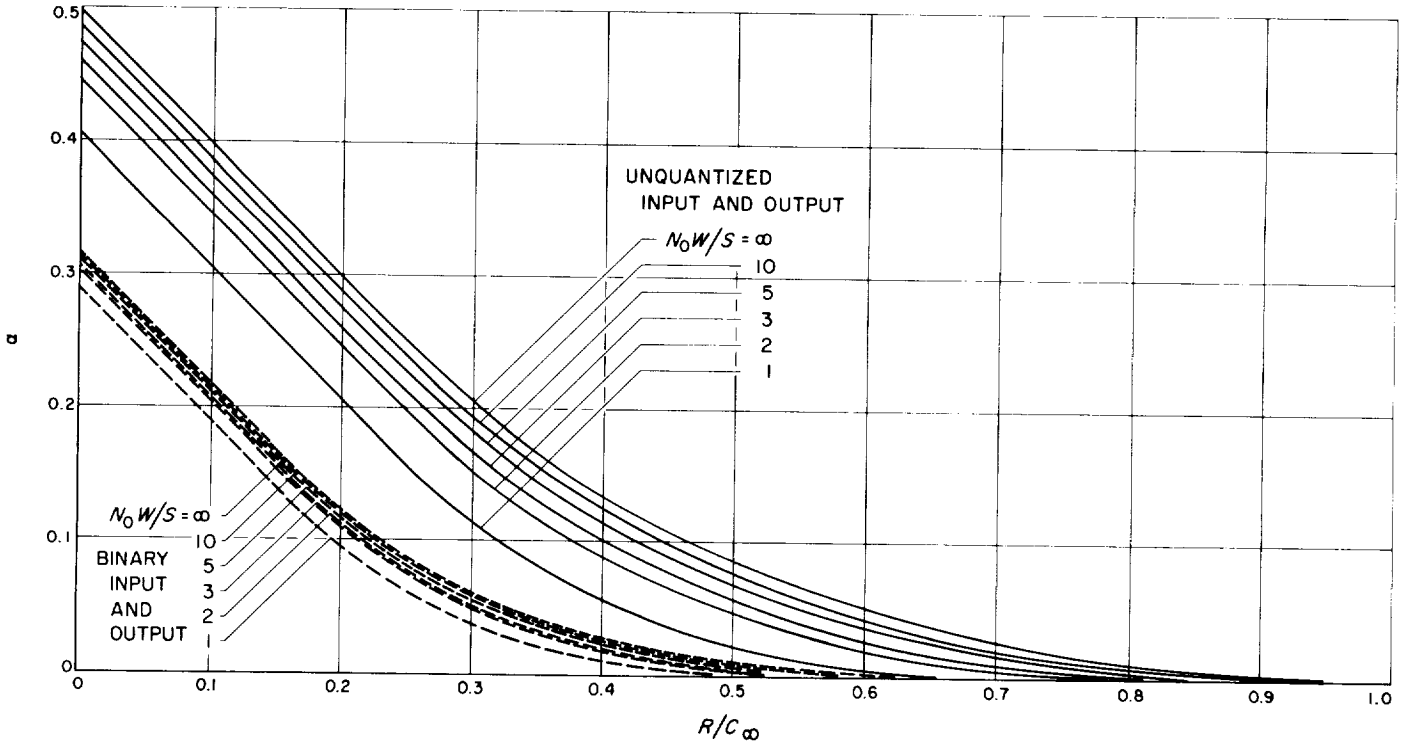


Fig. 10. Error-bound exponents

limiting case of Eq. (1). However, while for the limiting case of unconstrained bandwidth the coding procedure (for orthogonal signals) is well-known, for finite bandwidth limitations no deterministic coding procedure exists, although by random selection it is possible to generate rapidly a code which is nearly as good as the bound.

Also of interest are the intercepts of the curves with the α and R axes. These are, from Eq. (9):

$$\alpha(R=0)=\frac{N_0W}{S}\ln\left[1+\frac{S}{2N_0W}\right] \tag{10}$$

and

$$\frac{R(\alpha=0,s=0)}{C_\infty}=\frac{N_0W}{S}\ln\left[1+\frac{S}{N_0W}\right]$$

$$=\frac{W\log_2\left(1+\frac{S}{N_0W}\right)}{S/N_0\ln 2}=\frac{C}{C_\infty}, \tag{11}$$

where C is the capacity of the bandwidth constrained channel. In Figs. 11 and 12, $\alpha(R=0)$ and C/C_∞ are plotted as a function of N_0W/S. Note that as $W\to\infty$ $\alpha(R=0)\to 1/2$ and $C/C_\infty\to 1$.

3. Effect of Quantizing the Input to Two Levels

For purposes of generation it is often convenient to restrict the input to two levels, $S^{1/2}$ and $-S^{1/2}$, for example. This also serves the purpose of maintaining a constant transmitter power. It is well-known that orthogonal signals can be generated as sequences of such binary signals,

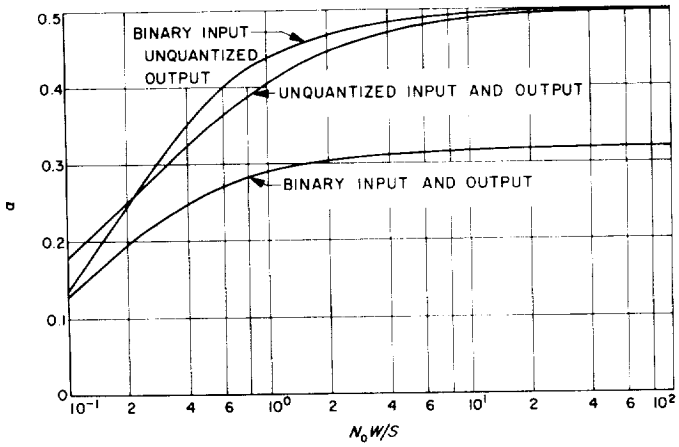


Fig. 11. Zero-rate exponents

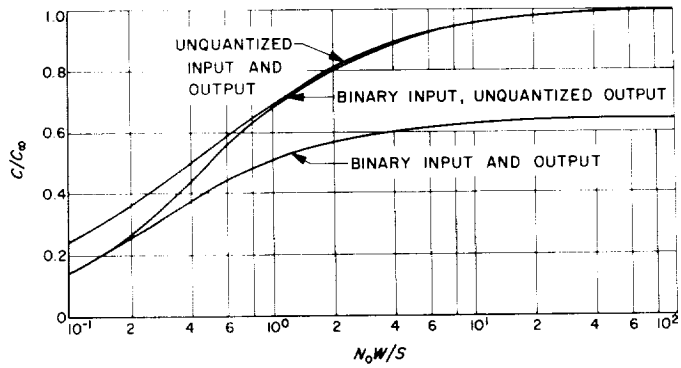


Fig. 12. Zero exponent rates (channel capacities) normalized by $C_\infty = S/N_0 \ln$

and thus for the bandwidth-unconstrained channel quantizing the input has no effect.

If the bandwidth is constrained, we may randomly generate signals consisting of sequences of binary signals by flipping an unbiased coin. Then

$$p(x)=\frac{1}{2}\left[\delta(x-S^{1/2})+\delta(x+S^{1/2})\right] \tag{12}$$

and since the channel is Gaussian, $p(y|x)$ is given by Eq. (7). While, in principle, we could compute the bound (5) as before by substituting Eqs. (7) and (12) in the expressions of (3), the result can not be obtained in closed form for this case. However, it is possible to evaluate the intercepts of the curves as was done in the previous case in Eqs. (10) and (11). We have from Eq. (3)

$$\tilde{\alpha}(\tilde{R}=0)=\tilde{\alpha}_r+\tilde{R}_r=[\tilde{\alpha}+\tilde{R}]_{s=1/2}$$

$$=-\ln\int_{-\infty}^{\infty}dy\left[\int_{-\infty}^{\infty}p(x)\,[p(y|x)]^{1/2}\,dx\right]^2. \tag{13}$$

Substituting Eqs. (7) and (12) in Eq. (13) yields

$$\tilde{\alpha}(\tilde{R}=0)=\ln 2-\ln(1+e^{-S/(2N_0W)}).$$

Thus, from Eq. (5), we have

$$\alpha(R=0)=\frac{2N_0W}{S}[\ln 2-\ln(1+e^{-S/(2N_0W)})]. \tag{14}$$

This function is also plotted in Fig. 11.

To obtain the intercept of the R-axis, we have from Eq. (3)

$$\widetilde{R}(\alpha=0, s=0)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln\frac{p(x|y)}{p(x)}\,dx\,dy \tag{15}$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln\frac{p(y|x)}{p(y)}\,dx\,dy$$

$$=-\int_{-\infty}^{\infty} p(y)\ln p(y)\,dy$$

$$+\int_{-\infty}^{\infty} p(x,y)\ln p(y|x)\,dx\,dy$$

where

$$p(y)=\frac{e^{-[y-(S)^{1/2}]^2/2N_0W}+e^{[y+(S)^{1/2}]^2/2N_0W}}{2(2\pi N_0W)^{1/2}}.$$

Then, integrating and using Eq. (4), we have

$$\frac{R(\alpha=0)}{C_\infty}=\frac{2N_0W}{S}\widetilde{R}(\alpha=0)$$

$$=-\frac{2N_0W}{S(\pi)^{1/2}}\int_{\infty}^{\infty} e^{-u^2}\ln\{1+e^{-4(S/2N_0W)^{1/2}[u+(S/2N_0W)^{1/2}]}\}\,du. \tag{16}$$

The integral in Eq. (16) was first computed by Bloom, Chang, et al. (Ref. 8), and we have used their numerical results to obtain the plot in Fig. 12.

It is rather surprising that for $N_0W/S > 1/4$, the zero rate exponent α is greater for the quantized case than for the amplitude-unconstrained case. This may be explained by the fact that these results represent merely upper bounds on P_E. In any case, it appears again that for $N_0W/S > 1/4$ there is very little loss in performance due to quantization of the input signals.

4. Effect of Quantizing the Output

Suppose now that in addition to quantizing the input signal, the output is integrated over each symbol duration $1/2W$ seconds and is then quantized into the same two levels. This reduces the Gaussian channel to a binary symmetric channel. For this channel Fano (Ref. 5) has shown that the error probability of the optimum code is bounded by $P_E < Ke^{-n\tilde{\alpha}}$ where $K < 2$, and

$$\tilde{\alpha}=\begin{cases} s\ln\left(\frac{s}{p}\right)+(1-s)\ln\left(\frac{1-s}{1-p}\right), & p<s<s_C \\ \ln 2-2\ln(p+1-p)-R, & s_C-s \end{cases} \tag{17}$$

$$R=\begin{cases} \ln 2+s\ln s+(1-s)\ln(1-s), & p\leq s<s_C \\ \ln 2+s\ln s_C+(1-s)\ln(1-s_C), & s_C\leq s\leq\frac{1}{2} \end{cases}$$

where

$$s_C=\frac{p^{1/2}}{p^{1/2}+(1-p)^{1/2}}$$

and p is the probability that $x=+S^{1/2}$ and $y=-S^{1/2}$ (or vice-versa). Thus,

$$p=\int_{-\infty}^{0} e^{-[y-(S)^{1/2}]^2/2N_0W}=\operatorname{erfc}\left(\frac{S}{N_0W}\right)^{1/2}, \tag{18}$$

erfc the error-function complement. Then using Eqs. (4) and (5), we have

$$P_E<K2^{-k(C_\infty\alpha/R)} \tag{19}$$

where

$$\alpha=\begin{cases} \dfrac{2N_0W}{S}\left[s\ln\left(\dfrac{s}{p}\right)+(1-s)\ln\left(\dfrac{1-s}{1-p}\right)\right], & \text{for } \begin{cases} R_c<R \\ p<s<s_c \end{cases} \\ \dfrac{2N_0W}{S}\left\{\ln 2-2\ln[p^{1/2}+(1-p)^{1/2}]-\dfrac{R}{C_\infty}\right\}, & \text{for } 0<R<R_c \end{cases}$$

$$\frac{R}{C_\infty}=\frac{2N_0W}{S}[\ln 2+s\ln s+(1-s)\ln(1-s)], \quad \text{for } \begin{cases} R_c<R \\ p\leq s<s_c \end{cases}$$

Thus, α and R are related by the parameter s and the transition probability p. Fig. 10 shows α as a function of R/C_∞ for $N_0W/S = 1, 2, 3, 5, 10$, and ∞.

The intercepts of the α and R axes are

$$\alpha(R=0)=\frac{2N_0W}{S}\{\ln 2-2\ln[p^{1/2}+(1-p)^{1/2}]\} \quad (20)$$

$$\frac{R(\alpha=0,s=p)}{C_\infty}=\frac{2N_0W}{S}\times[\ln 2+p\ln p+(1-p)\ln(1-p)] \quad (21)$$

$$=\frac{2W}{C_\infty}[1+p\log_2 p+(1-p)\log_2(1-p)]$$

$$=\frac{2WC_b}{C_\infty}$$

where C_b is the capacity of the binary channel in bits/symbol. Eqs. (20) and (21) are plotted in Figs. 11 and 12. Note from Eq. (18) that the asymptotic expression

$$p\sim\frac{1}{2}-\left(\frac{S}{2\pi N_0W}\right)^{1/2}$$

holds as $W\to\infty$, so that Eqs. (20) and (21) become, respectively,

$$\lim_{W\to\infty}\alpha(R=0)=\frac{1}{\pi},$$

$$\lim_{W\to\infty}\frac{R(\alpha=0)}{C_\infty}=\frac{2}{\pi}$$

which are exactly $2/\pi$ times the limiting values for the cases of unquantized output. Furthermore, it appears from Figs. 10, 11, and 12 that for all W quantization of the output causes a loss of approximately $2/\pi$ (or approximately 2 db) in the signal power to restore the performance to that without quantization.

D. Accuracy of an Angle-Measuring System

W. B. Kendall

1. Introduction

If two separate antennas are illuminated by radiation with plane wave fronts, such as RF energy from a spacecraft, then the phase difference between their outputs provides a measure of the direction from which the energy is arriving. This basic notion can be (and has been) used to construct various systems for angle-tracking satellites which radiate an RF carrier. An important problem with such interferometer systems is that when the two antennas are more than half a wavelength apart, the resulting angle information is ambiguous. The simplest way to resolve this ambiguity is to simply make another measurement with a different antenna spacing (and thus different ambiguities) and combine the two ambiguous measurements into a single unambiguous measurement. In this paper we examine the effect of receiver noise on such a system. The optimum way to process the required waveforms is determined, as well as the optimum spacings derived for the antennas. Furthermore, the requirements on signal-to-noise ratio are found; these requirements must be met to insure that with a given probability the final unambiguous measurement is not in error by more than some specified amount. Though the results and discussions are couched in the terminology of angle measurements made with an interferometer, they are applicable to any ambiguous measurements for which the number of ambiguities is inversely proportional to the (ambiguous) accuracy. For example, they can be used to study the unambiguous range accuracy of a pulsed radar in which the number of ambiguities is directly proportional to the PRF and the energy in each pulse (and thus the ambiguous range accuracy) is (in order to maintain fixed power) inversely proportional to the PRF.

2. Maximum-Likelihood Estimation of Arrival Angle

The basic interferometer we wish to consider is shown schematically in Fig. 13, which is drawn in the plane containing the signal source and the two antennas. (Two such systems are required to determine the line of sight in three-dimensional space.) The two signals are effectively received at the phase centers of their respective antennas. Though the position of these phase centers

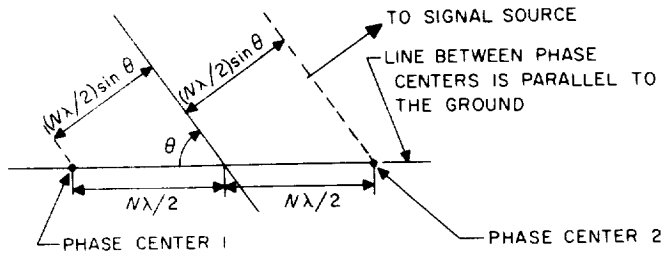


Fig. 13. Basic interferometer

relative to the antennas is, in general, a function of the angle of arrival of the wave fronts, the slope of the line between them and the distance between them is independent of this if the two antennas are identical.

If we denote the signals received at the two antennas by $s_1(t)$ and $s_2(t)$, then

$$s_1(t) = a \sin(\omega t + \phi - N\pi \sin\theta)$$

and

$$s_2(t) = a \sin(\omega t + \phi + N\pi \sin\theta)$$

where a is an unknown amplitude, ω is the RF angular frequency, ϕ is the unknown RF phase (at a reference position taken arbitrarily midway between the two antennas), N is the distance between the two antennas measured in wavelengths, and θ is the angle of arrival of the received wave fronts. After amplification, these signals become

$$x_1(t) = n_1(t) + a_1 \sin(\omega t + \phi - y/2)$$

and

$$x_2(t) = n_2(t) + a_2 \sin(\omega t + \phi + y/2)$$

where $y = 2N\pi \sin\theta$, and $n_1(t)$ and $n_2(t)$ represent wideband Gaussian noise sample functions added by the receivers. The amplitudes a_1 and a_2 are now possibly different due to different amplifier gains, but the phase ϕ is taken to be the same since we assume identical phase shifts through the two amplifiers.

The problem now is to form a maximum-likelihood estimate of the quantity y by processing the available waveforms $x_1(t)$ and $x_2(t)$. We assume that the two receiver noises are independent and Gaussian and have (one-sided) noise spectral densities N_0 (w/cps). The joint conditional probability-density function of the waveforms $x_1(t)$ and $x_2(t)$, given ϕ and y, is

$$p(x_1, x_2 | \phi, y) =$$

$$k \exp\left\{ -\frac{1}{N_0}\int_0^T [x_1(t) - a_1 \sin(\omega t + \phi - y/2)]^2 + [x_2(t) - a_2 \sin(\omega t + \phi + y/2)]^2\, dt \right\}$$

where k is a constant and the observation period is $(0, T)$.

This can be simplified to

$$p(x_1, x_2 | \phi, y) = k \exp\left\{ \frac{2}{N_0} M \cos(\phi + \Psi) - \frac{(a_1^2 + a_2^2)T}{2N_0} - \frac{1}{N_0}\int_0^T [x_1(t)^2 + x_2(t)^2]\, dt \right\}$$

where

$$M^2 = X^2 + Y^2$$

$$\Psi = \tan^{-1}(Y/X)$$

$$X = (a_1X_1 + a_2X_2)\cos(y/2) - (a_1Y_1 - a_2Y_2)\sin(y/2)$$

$$Y = (a_1Y_1 + a_2Y_2)\cos(y/2) + (a_1X_1 - a_2X_2)\sin(y/2)$$

$$\left.\begin{aligned} X_i &= \int_0^T x_i(t) \sin\omega t\, dt \\ Y_i &= \int_0^T x_i(t) \cos\omega t\, dt \end{aligned}\right\} \quad i = 1, 2$$

Now, since

$$p(x_1, x_2 | y) = \int_{-\infty}^{\infty} p(\phi)\, p(x_1, x_2 | \phi, y)\, d\phi,$$

and since ϕ is equally likely to be any value between 0 and 2π, we have

$$p(x_1, x_2 | y) = k \exp\left\{ -\frac{(a_1^2 + a_2^2)T}{2N_0} - \frac{1}{N_0}\int_0^T [x_1(t)^2 + x_2(t)^2]\, dt \right\} I_0\left(\frac{2M}{N_0}\right),$$

where I_0 is the modified Bessel function of the first kind and order zero, and we have made use of the relation

$$I_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{x \cos(\phi + \Psi)\}\, d\phi.$$

The maximum-likelihood estimate of y is obtained by choosing the value of y which maximizes $p(x_1, x_2 | y)$. This is equivalent to maximizing M (or M^2).

$$M^2 = X^2 + Y^2 = [(a_1X_1 + a_2X_2)^2 + (a_1Y_1 + a_2Y_2)^2]\cos\frac{2y}{2} - 2[(a_1X_1 + a_2X_2)(a_1Y_1 - a_2Y_2) - (a_1X_1 - a_2X_2)(a_1Y_1 + a_2Y_2)]\sin\frac{Y}{2}\cos\frac{Y}{2} + [(a_1X_1 - a_2X_2)^2 + (a_1Y_1 - a_2Y_2)^2]\sin^2\frac{y}{2}.$$

The derivative of this with respect to y can be set equal to zero to yield the estimate

$$y = \tan^{-1}\left\{\frac{X_1Y_2 - X_2Y_1}{X_1X_2 + Y_1Y_2}\right\}.$$

Note that this is independent of a_1 and a_2, so it is not necessary to know the signal strength or amplifier gains. Furthermore, the quantities X_1, X_2, Y_1, and Y_2 are simply the result of detecting the quadrature components of $x_1(t)$ and $x_2(t)$. We must, however, accurately know the RF frequency ω. It is assumed that this is either already known or is determined by locking onto the signal from one of the antennas with a phase-locked loop before the determination of the quadrature components of $x_1(t)$ and $x_2(t)$ is begun.

3. Phase-Difference Measurements

The maximum-likelihood estimate of the phase of a sine wave in Gaussian noise is obtained by first getting the quadrature components X and Y where

$$X = \int_0^T x(t) \sin \omega t \, dt$$

$$Y = \int_0^T x(t) \cos \omega t \, dt$$

and then using as the estimate

$$\Psi = \tan^{-1}(Y/X).$$

Thus, if we estimate the phases of the two received signals $x_1(t)$ and $x_2(t)$ and then take the difference, we obtain

$$\tan(\Psi_2 - \Psi_1) = \frac{\tan\Psi_2 - \tan\Psi_1}{1 + \tan\Psi_2 \tan\Psi_1}$$

$$= \frac{(Y_2/X_2) - (Y_1/X_1)}{1 + Y_1Y_2/X_1X_2}$$

$$= \frac{X_1Y_2 - X_2Y_1}{X_1X_2 + Y_1Y_2}.$$

Thus, as might be expected, the maximum-likelihood estimate of y is simply

$$y = \Psi_2 - \Psi_1,$$

where Ψ_1 and Ψ_2 are maximum-likelihood estimates of the phase of $s_1(t)$ and $s_2(t)$.

4. Ambiguous Accuracy

For moderately high signal-to-noise ratios the errors in the estimates of Ψ_1 and Ψ_2 are nearly Gaussian with zero means and variance given by $N_0/(2E_i)$, $i = 1, 2$, where

$$E_i = \int_0^T s_i(t)^2 \, dt = \frac{a_i^2}{2}$$

is the energy in the received signal s_i. Here we assume equal signal strengths in the two channels, so $E_1 = E_2 = E$. The error in the phase-difference measurement thus has a variance of N_0/E, and the standard deviation of the *angular* error in the (still ambiguous) estimate of $\theta = \sin^{-1}(y/2N\pi) = \sin^{-1}[(\Psi_2 - \Psi_1)/2N\pi]$ is approximately

$$\sigma_\theta = \frac{\sec\theta}{2N\pi(E/N_0)^{1/2}}.$$

Since $\sec\theta \to \infty$ as $\theta \to \infty$ ± 90 deg, it is evident that this system will fail when the signal source lies near the line through the two antennas. This can be prevented by using three antennas arranged in an equilateral triangle. Then a source will never be within 30 deg of the line through more than one pair of antennas, and angle information can be obtained from the other two pairs. We then have

$$\sigma_\theta \leq \frac{1}{N\pi(E/N_0)^{1/2}}.$$

5. Ambiguity Resolution

Since the phase difference $\Psi_1 - \Psi_2$ can only be measured modulo 2π, it follows that $\sin\theta = (\Psi_1 - \Psi_2)/2N\pi$ can only be measured modulo $1/N$. Thus, for $N > \frac{1}{2}$, i.e., for antennas spaced more than half a wavelength apart, the determination of $\sin\theta$, and thus of θ, is ambiguous. Let v denote the primary measured value of $\sin\theta$, i.e.,

$$v = \frac{\Psi_1 - \Psi_2}{2N\pi}$$

with

$$0 \leq v < 1/N.$$

Then all that is known about the actual value of $\sin\theta$ is that

$$\sin\theta = v + j/N$$

for some integer j satisfying

$$-1 \leq v + j/N \leq 1$$

or

$$-N(1+v) \leq j \leq N(1-v).$$

Thus, if no information about the true value of θ (except that from this measurement) is available, then $\sin\theta$ is known only to be one of about[1] $2N$ values. In many cases, however, some *a priori* information about θ is available and can be used to a number less than this.

The correct value of j can be determined by duplicating the measurement of $\sin\theta$ with a different antenna spacing. Denote by u the (ambiguous) value of $\sin\theta$ measured with antennas spaced M wavelengths apart. Then, in addition to the above, we have

$$\sin\theta = u + i/M; \qquad 0 \leq u < 1/M$$

for some integer i satisfying

$$-M(1+u) \leq i \leq M(1-u).$$

This is shown schematically in Fig. 14, where the small open arrows indicate which points along a line form -1 to $+1$ are possible values of $\sin\theta$ for a given v and for a given u. For a properly chosen M/N there is only one pair of values i, j consistent with the requirement

$$v + j/N = \sin\theta = u + i/M;$$

so in the absence of noise, ambiguity resolution is a simple matter.

When the measurements u and v are corrupted by noise there will not, in general, be any solution to the equation

$$-1 \leq v + j/N = u + i/M \leq 1.$$

This situation is shown schematically in Fig. 15. However, the ambiguities can still be resolved from the two measurements u and v by choosing the i and j which minimize the magnitude of the difference $v + j/N - u = i/M$, i.e., by choosing the two arrows in Fig. 15 which most nearly oppose each other. Then the probability of correct ambiguity resolution can be maximized by choosing the antenna spacings M and N which minimize the probability that $|v + (j/N) - u - (i/M)|$ has its minimum for any but the correct values of i and j. If we denote by J the correct value of j, this is equivalent to maximizing the ratio of the minimum over i and $j \neq J$ of the expected

[1]If $2N$ is an integer, this is exact. Otherwise, $\sin\theta$ is known only to be one of $[2N]$ or $[2N]+1$ values (depending on the value of v), where $[2N]$ denotes the integer part of $2N$.

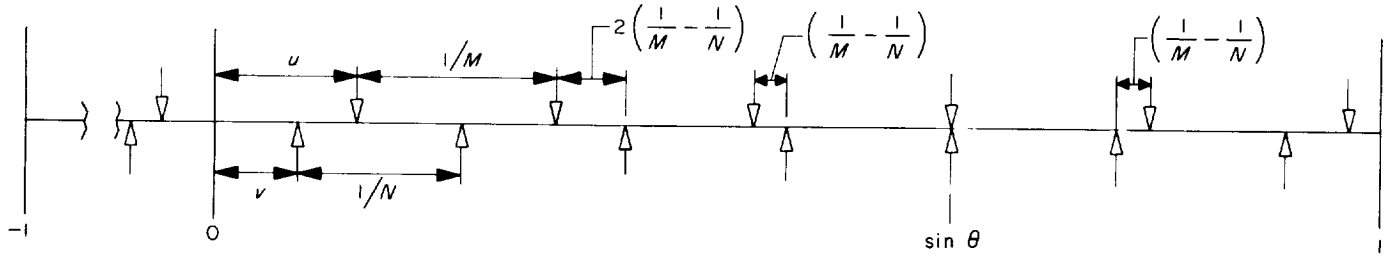


Fig. 14. Possible values of sin θ in the absence of noise

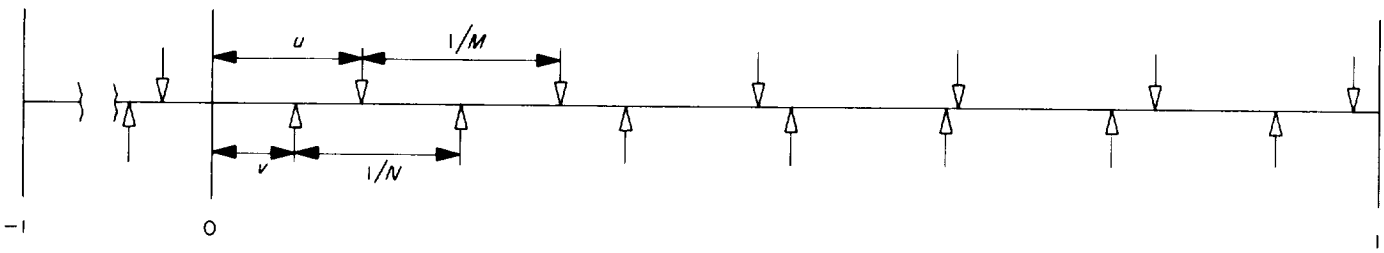


Fig. 15. Possible values of sin θ in the presence of noise

value of $|v + (j/N) - y - (i/M)|$ divided by the standard deviation of $(v - u)$.

The expected value of $|v + (j/N) - u = (i/M)|$ is just its value in the absence of noise, and its minimum over i, i.e., the distance between nearest pairs of opposite arrows in Fig. 14, is simply

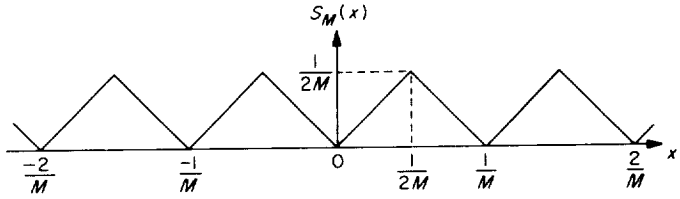
$$D_{J_j} = S_M\left[(J-j)\left(\frac{1}{M} - \frac{1}{N}\right)\right],$$

where we assume $M < N$, and $S_M(x)$ is the saw-tooth function which, along with some of its properties, is shown in Fig. 16. From the properties of $S_M(x)$ it follows that

$$D_{J_j} = \frac{1}{M} S_1\left[(J-j)\frac{M}{N}\right] = \frac{1}{M} S_1\left[(J-j)1 - \frac{M}{N}\right].$$

Next we let K denote the maximum possible value of $|J - j|$. Of course, this maximum value depends on both the *a priori* uncertainty about θ, i.e., on the range of values j may take on, and also it depends on the true value of θ, i.e., on J. In most system design problems, however, it is the probability of incorrect ambiguity resolution in the *most ambiguous case* which must be minimized. Thus, in this situation, K is given its maximum possible value. As noted above, in the absence of a *priori* information about θ, this maximum possible value of K is $2N - 1$ if $2N$ is an integer, and is $2N$ otherwise. Now, if we let x correspond to either M/N or $1 - M/N$, then the numerator of the function we wish to maximize depends on the function

$$D_{min}(K,x) = 1 \leqq \min_k \leqq K^{S_1(kx)} \qquad 0 < x < \tfrac{1}{2}$$



$S_M(-x) = S_M(x)$

$S_M(ax) = aS_{aM}(x) \qquad a > 0$

$S_M(x + k/M) = S_M(x) \qquad k = \text{INTEGER}$

$S_M(kx) = S_M[kS_M(x)] \qquad k = \text{INTEGER}$

Fig. 16. The saw-tooth function and some of its properties

In general, this is a complicated function. For example, $D_{min}(8,x)$ and $D_{min}(9,x)$ are shown in Figs. 17 and 18. Fortunately, however, we will need only the following simple properties, which are illustrated in Figs. 17 and 18.

(1) $D_{min}(K,x) = D_{min}(K, 1-x)$

(2) $D_{min}(K,x) = x, \qquad \text{for } 0 \leqq x \leqq \dfrac{1}{K+1}$

(3) $D_{min}(K,x) \leqq \dfrac{x}{l}\dfrac{l-1}{(K+1)/2}$ $\Big\}$ $l = 1, 2, \cdots, L$

$< x < \dfrac{l}{K+1}$

(4) $D_{min}(K,x) < \dfrac{1}{K+1}\dfrac{l}{K+1}$ $\Big\}$ $L = \begin{cases} \dfrac{K}{2}, & K \text{ even} \\ \dfrac{K+1}{2}, & K \text{ odd} \end{cases}$

$< x < \dfrac{l}{(K+1)/2}$

(5) $D_{min}\left(K, \dfrac{l}{K+1}\right) = \begin{cases} 0, & \text{if } l \text{ and } K+1 \text{ have any common factor} \\ \dfrac{1}{K+1}, & \text{if } l \text{ and } K+1 \text{ have no common factor} \end{cases}$

(6) $D_{min}(K,x) = 1 - 2x \qquad \dfrac{L}{2L+1} \leqq x \leqq \dfrac{1}{2}$

For our purposes these properties can be summarized as follows. The function $D_{min}(K,x)$ increases linearly from zero at $x = 0$ to a maximum value of $1/(K+1)$ at $x = 1/(K+1)$. Also, it decreases linearly to zero at $x = \frac{1}{2}$ from a peak value of $1/(K+1)$ at $x = K/2(K+1)$ if K is even, or from a peak value of $1/(K+2)$ at $x = (K+1)/2(K+2)$ if K is odd.

The other quantity we must have, in order to maximize the probability of correct ambiguity resolution, is the variance of the difference $(v - u)$. This variance equals

$$\frac{N_0[1 - 2\rho(N/M) + (N/M)^2]}{4\pi^2 N^2 E},$$

where ρ is the correlation coefficient between u and v. If u and v are obtained from separate pairs of antennas, then ρ is zero. If u and v are obtained from three colinear antennas, then ρ is $-\frac{1}{2}$ if the common antenna is between the other two, and is $+\frac{1}{2}$ otherwise.

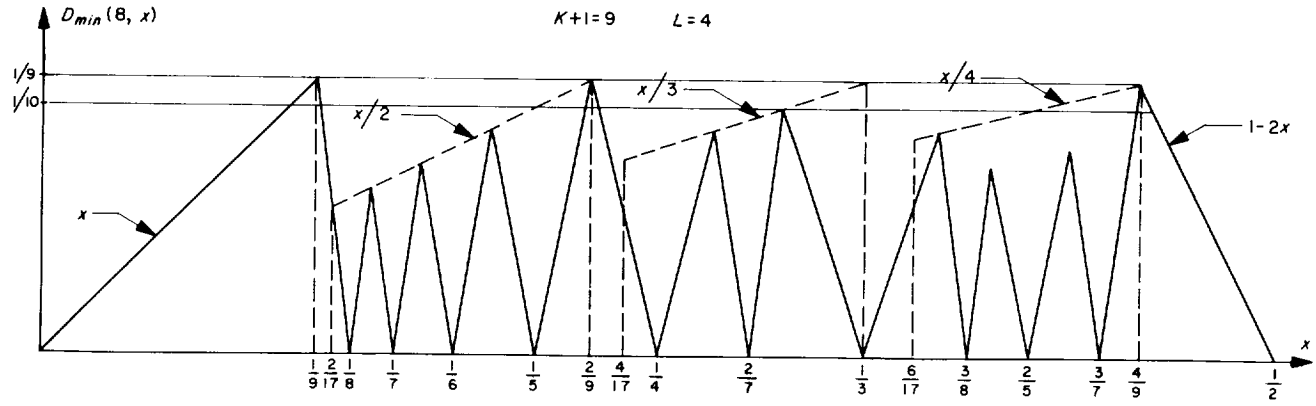


Fig. 17. The function $D_{min}(8, x)$ as a function of x

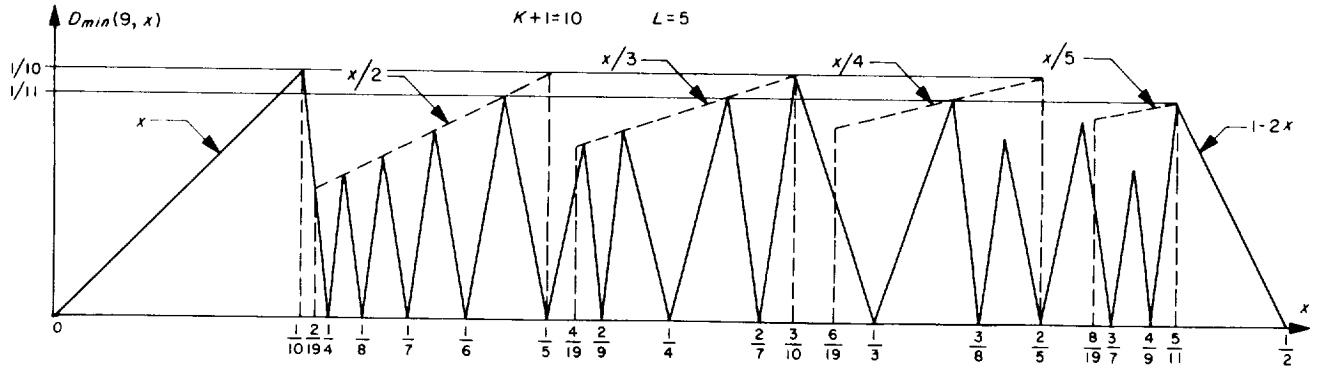


Fig. 18. The function $D_{min}(9, x)$ as a function of x

Now the ratio we wish to maximize is

$$R = \frac{\frac{1}{M} D_{min}\left(K, \frac{M}{N}\right)}{\left\{\frac{N_0\left[1 - 2\rho\frac{N}{M} + \left(\frac{N}{M}\right)^2\right]}{4\pi^2 N^2 E}\right\}^{1/2}}$$

$$= 2\pi\left(\frac{E}{N_0}\right)^{1/2} \frac{D_{min}\left(K, \frac{M}{N}\right)}{\left[\left(\frac{M}{N}\right)^2 - 2\rho\frac{M}{N} + 1\right]^{1/2}}.$$

The denominator is shown as a function of M/N in Fig. 19. From the properties of $D_{min}(K, x)$ and the fact that the denominator always increases less than linearly with M/N, it is evident that for $\rho = -1/2$ or $\rho = 0$ we should choose[2] $M/N = 1/(K+1)$, and for $\rho = 1/2$ and K even we should choose[2] $M/N = L/(2L+1) = K/[2(K+1)]$. For $\rho = 1/2$ and K odd it is not immediately evident whether the maximum of R occurs when

$$M/N = L/(2L+1) = (K+1)/[2(K+2)]$$

and $D_{min}(K, M/N) = 1/(K+1)$. However, from the properties of $D_{min}(K, x)$ it follows that for odd K the maximum value of x for which $D_{min}(K, x) = 1/(K+1)$ is

$$x = \begin{cases} \dfrac{L-1}{K+1} = \dfrac{1}{2} - \dfrac{1}{K+L}, & \text{for } L = \dfrac{K+L}{2} \text{ even} \\ \dfrac{L-2}{K+1} = \dfrac{1}{2} - \dfrac{2}{K+1}, & \text{for } L \text{ odd} \end{cases}$$

[2]This assumes that our only interest is in minimizing the probability of incorrect ambiguity resolution. When one wishes to minimize the overall probability that the angle error exceeds some threshold value, the results are slightly different. This is considered in the next part of this article.

Calculation of R at these alternate values of M/N shows that for $\rho = \frac{1}{2}$ and K odd we should choose M/N equal to this maximum value of x for which $D_{min}(K, x) = 1/(K+1)$. The resulting values of the ratio R are

$$R = \begin{cases} \dfrac{2\pi\left(\dfrac{E}{N_0}\right)^{\frac{1}{2}}}{(K^2+3K+3)^{\frac{1}{2}}} & \rho = -\dfrac{1}{2} \\ \dfrac{2\pi\left(\dfrac{E}{N_0}\right)^{\frac{1}{2}}}{(K^2+2K+2)^{\frac{1}{2}}} & \rho = 0 \\ \dfrac{2\pi\left(\dfrac{E}{N_0}\right)^{\frac{1}{2}}}{(3L^2+3L+1)^{\frac{1}{2}}}, & K \text{ even} \\ \dfrac{2\pi\left(\dfrac{E}{N_0}\right)^{\frac{1}{2}}}{(3L^2+2)^{\frac{1}{2}}}, & K \text{ odd}, \ L \text{ even} \\ \dfrac{2\pi\left(\dfrac{E}{N_0}\right)^{\frac{1}{2}}}{(3L^2+4)^{\frac{1}{2}}}, & K \text{ odd}, \ L \text{ odd} \end{cases} \quad \rho = \frac{1}{2}$$

where, as before,

$$L = \begin{cases} \dfrac{K}{2}, & K \text{ even} \\ \dfrac{K+1}{2}, & K \text{ odd} \end{cases}$$

Since the inequalities

$$K^2 + 3K + 3 > 3L^2 + 4 > 3L^2 + 2$$

are satisfied for all $K > 1$, it follows that as far as ambiguity resolution is concerned, when three colinear antennas are used, $\rho = \frac{1}{2}$ should be chosen. This is done by obtaining v from the phase difference between the two end antennas, and obtaining u from the phase difference between the center antenna and one end antenna. It can be shown that this procedure yields a maximum-likelihood resolution of the ambiguity.

Finally, the probability that the ambiguities are incorrectly resolved, which we will denote by P_A, is simply the probability that a Gaussian random variable differs from its mean by more than $R/2$ standard deviations.

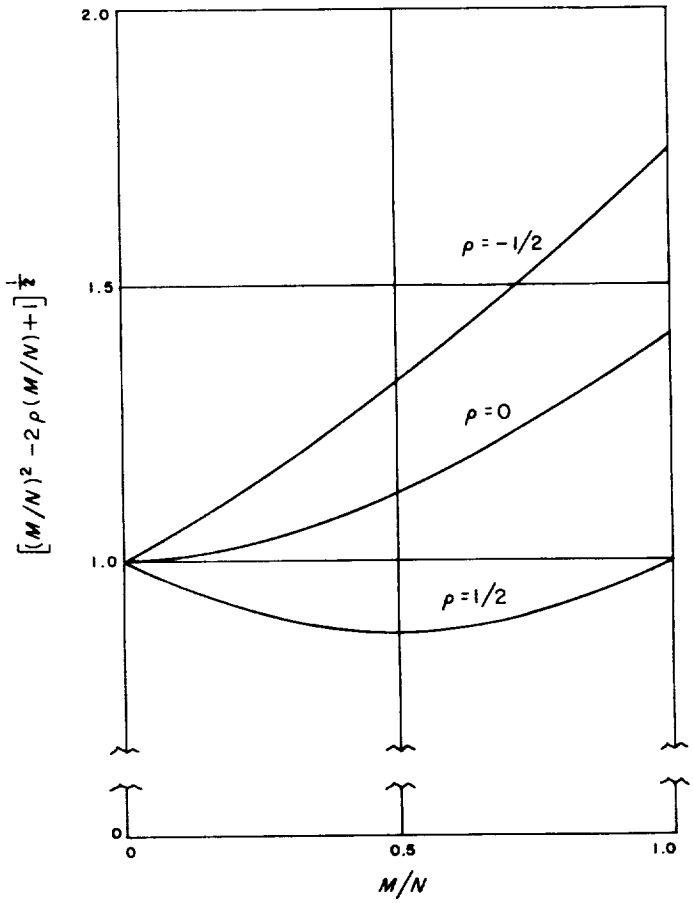


Fig. 19. The denominator of the expression for R as a function of the antenna spacing ratio M/N

Thus, we have

$$P_A = 1 - \Theta\left\{\left[\frac{R}{2(2)^{\frac{1}{2}}}\right]\right\}$$

where $\Theta(x)$ is the error integral (Ref. 9) defined by

$$\Theta(x) = \frac{2}{\pi^{\frac{1}{2}}}\int_0^x e^{-t^2}\,dt.$$

This gives the values shown in Table 4.

Table 4. The probability of incorrect ambiguity resolution as a function of the ratio R

P_A	R
0.1	3.29
0.03	4.34
0.01	5.16
0.003	5.94
0.001	6.58

6. Unambiguous Accuracy

Since the resolution of ambiguities requires that two measurements of $\sin\theta$ be made, both of these measurements should be used not only to resolve the ambiguity, but also to provide the final estimate of $\sin\theta$. After the two measurements have been made and the ambiguity has been resolved, we have the two estimates $\sin\theta = u + i/M$ and $\sin\theta = v + j/N$ which have a joint probability density function which is proportional to

$$\exp\left\{-\frac{\left[v+\frac{j}{N}-\sin\theta\right]^2-2\rho\frac{M}{N}\left[v+\frac{j}{N}-\sin\theta\right]}{2\sigma_v^2(1-\rho^2)}\right.$$
$$\left.\times\frac{\left[u+\frac{i}{M}-\sin\theta\right]\left(\frac{M}{N}\right)^2\left[u+\frac{i}{M}-\sin\theta\right]^2}{2\sigma_v^2(1-\rho^2)}\right\}$$

where

$$\sigma_v^2 = N_0/4N^2\pi^2E$$

and, as before, ρ is the correlation coefficient between u and v. (Here we have again approximated the true probability distributions of u and v by Gaussian distributions, and thus tacitly assume a moderate to high signal-to-noise ratio.) Setting the derivative of this with respect to $\sin\theta$ equal to zero yields the maximum-likelihood estimate

$$\sin\theta = \frac{(v+j/N)(1-\rho M/N)+(u+i/M)\{(M/N)^2-\rho M/N\}}{1-2\rho(M/N)+(M/N)^2}.$$

The standard deviation of this estimate is

$$\sigma_{\sin\theta} = \frac{1}{2N\pi(E/N_0)^{1/2}}\left(\frac{1-\rho^2}{1-2\rho(M/N)+(M/N)^2}\right)^{1/2}$$

In terms of the ratio R, which, as was seen above, determines the probability of incorrect ambiguity resolution, this is

$$\sigma_{\sin\theta} = \frac{R(1-\rho^2)^{1/2}}{4N\pi^2(E/N_0)D_{min}(K, M/N)}.$$

The quantity which is really of interest in most angle measuring systems is the probability P_E that the magnitude of the angle error exceeds some threshold value δ_θ. This value can be exceeded in either of two ways: either by the ambiguities being incorrectly resolved, or by the final unambiguous estimate being in error by more than δ_θ. Note that, for the system we are considering, ambiguity resolution depends only on the difference $v - u$, while the error in the estimate of $\sin\theta$, and thus of θ itself, depends only on the error in the linear combination

$$v(1-\rho M/N)+u[(M/N)^2-\rho M/N].$$

Furthermore, it can be shown that these quantities are uncorrelated, and thus, since they have Gaussian distributions, the events "correct ambiguity resolution" and "angle error magnitude less than δ_θ" are statistically independent. We have then simply

$$P_E = 1-\Theta\left[\frac{\delta_\theta}{(2)^{1/2}\sigma_\theta}\right]\Theta\left[\frac{R}{2(2)^{1/2}}\right]$$
$$= 1-\Theta\left[\frac{\delta}{2^{1/2}\sigma_{\sin\theta}}\right]\Theta\left[\frac{R}{2(2)^{1/2}}\right],$$

where $\delta = \delta_\theta\cos\theta$ is the maximum allowable error in $\sin\theta$, and $\Theta(x)$ is the error integral. This can also be written

$$P_E = 1-\Theta\left[\frac{\delta}{2^{1/2}\sigma_{\sin\theta}}\right]\Theta\left[\frac{A\sigma_{\sin\theta}}{2(2)^{1/2}}\right],$$

where

$$A = \frac{4N\pi^2(E/N_0)D_{min}(K, M/N)}{(1-\rho^2)^{1/2}}.$$

We would now like to choose A and $\sigma_{\sin\theta}$ so that P_E is minimized. It can be shown that for $x > 0$ the function $\Theta(x)\Theta(a/x)$ has a unique maximum at $x = a^{1/2}$. Thus, P_E is minimized when

$$\frac{\delta}{2^{1/2}\sigma_{\sin\theta}} = \left(\frac{\delta A}{4}\right)^{1/2}$$

or

$$\sigma_{\sin\theta} = \left(\frac{2\delta}{A}\right)^{1/2}.$$

Its value is then

$$P_E = 1-\left\{\Theta\left[\left(\frac{A\delta}{4}\right)^{1/2}\right]\right\}^2.$$

Next we must maximize, as far as is possible, the term

$$\frac{A\delta}{4} = \frac{\delta N\pi^2(E/N_0)D_{min}(K, M/N)}{(1-\rho^2)^{1/2}}.$$

This means first of all that the antenna spacing ratio M/N should be chosen to maximize $D_{min}(K, M/N)$, which we

have already seen can always be done by choosing, among other possible values, $M/N = 1/(K+1)$, so that

$$D_{min}(K, M/N) = 1/(K+1)$$

and

$$\frac{A\delta}{4} = \frac{\delta\pi^2 (E/N_0)}{(1-\rho^2)^{1/2}} \frac{N}{K+1}.$$

Now if we assume that all existing ambiguities must be resolved, i.e., that *a priori* information about θ is not used for ambiguity resolution, then, as noted in the last section, the maximum number of ambiguities to be resolved is[3]

$$K = \begin{cases} 2N-1, & 2N \text{ is an integer} \\ [2N], & \text{otherwise} \end{cases}$$

This yields $K+1 \geq 2N$ with equality only when $2N$ is an integer. Therefore, $2N$ should be an integer[4] and we have

$$\frac{A\delta}{4} = \frac{\delta\pi^2 (E/N_0)}{2(1-\rho^2)^{1/2}}$$

and

$$P_E = 1 - \left(\Theta \left\{ \left[\frac{\delta\pi^2 (E/N_0)}{2(1-\rho^2)^{1/2}} \right]^{1/2} \right\} \right)^2$$

The values of the parameter $A\delta/4$ which are required for various values of P_E are shown in Table 5. Note that they suggest the very convenient approximation $A\delta/4 \approx 2\log_{10}(1/P_E)$, which allows us to write

$$P_E \approx 10^{-\{(\pi^2/4)\delta(E/N_0)/(1-\rho^2)^{1/2}\}}.$$

Alternatively, the signal-to-noise ratio required for a given accuracy and probability of error can be expressed approximately as

$$\frac{E}{N_0} \approx \frac{4(1-\rho^2)^{1/2}\log_{10}(1/P_E)}{\pi^2\delta}.$$

7. Comparison of Two Ambiguous Measurements to One Unambiguous Measurement

An interesting result of the above analysis is that, for a given probability of error, the unambiguous accuracy does *not* depend on the antenna spacing as long as the antennas are spaced an integer number of half wavelengths apart. This suggests that the simplest system might consist of antennas spaced only one-half wavelength apart. Then $N = 1/2 = M$ and $K = 0$; i.e., then there are no ambiguities to be resolved. These parameter values yield

$$\left.\begin{aligned} &\sigma_{\sin\theta} = \frac{1}{\pi(E/N_0)^{1/2}} \left(\frac{1+\rho}{2}\right)^{1/2} \\ &\text{and} \\ &P_E = 1 - \Theta\left(\frac{\delta}{2^{1/2}\sigma_{\sin\theta}}\right) \\ &\quad = 1 - \Theta\left[\frac{\pi\delta(E/N_0)^{1/2}}{(1+\rho)^{1/2}}\right] \end{aligned}\right\} \begin{aligned} &N = M = 1/2 \\ &\text{Two measurements} \end{aligned}$$

Since in this case the second measurement is not needed for ambiguity resolution (it serves only to reduce the variance of the estimate), it can be eliminated with the result that

$$\left.\begin{aligned} &\sigma_{\sin\theta} = \frac{1}{\pi(E/N_0)^{1/2}} \\ &\text{and} \\ &P_E = 1 - \Theta\left[\frac{\pi\delta(E/N_0)^{1/2}}{2^{1/2}}\right] \end{aligned}\right\} \begin{aligned} &N = 1/2 \\ &\text{One measurement} \end{aligned}$$

In some cases this is an improvement over the performance obtained with two ambiguous measurements,

[3]Even if *a priori* information *is* used to reduce the number of ambiguities which must be resolved, the following results are substantially the same, since in any event the maximum number of ambiguities to be resolved is approximately proportional to N.

[4]Note that this gives $N = (K+1)/2$, so that the antenna spacing M is $M = (M/N)(K+1)/2 = 1/2$ if $M/N = 1/(K+1)$. Since independent antennas only one-half wavelength apart are difficult to obtain in practice, usually one would use one of the larger values of M/N which gives $D_{min}(K, M/N) = 1/(K+1)$.

Table 5. Probability of error for various values of the parameter $A\delta/4$

P_E	$A\delta/4$
0.1	1.9
0.03	3.0
0.01	3.9
0.003	5.0
0.001	6.1

and in some cases it is not. In order to make the comparison on the basis of the same *total* received signal energy, we define E_T as

$$E_T = E \times (\text{number of independent receiving antennas})$$
$$= E(4 - 2|\rho|)$$

where, as before, ρ is $\pm\frac{1}{2}$ when three colinear antennas are used, and is zero when two pairs of independent antennas are used. When only one measurement is made, ρ is effectively unity. The comparison to be made is now between the two values

$$P_E = 1 - \Theta\left[\frac{\pi\delta\,(E_T/N_0)^{1/2}}{2}\right]^{1/2} \qquad \text{One unambiguous measurement}$$

and

$$P_E = 1 - \left\{\Theta\left[\frac{\pi^2\delta\,(E_T/N_0)}{8(1-\rho^2)^{3/2}}\right]^{1/2}\right\}^2 \qquad \text{Two ambiguous measurements}$$
$$\rho = \pm\tfrac{1}{2}, 0$$

One unambiguous measurement is superior if

$$\left\{\Theta\left[\frac{\pi^2\delta\,(E_T/N_0)}{8(1-\rho^2)^{3/2}}\right]^{1/2}\right\}^2 < \Theta\left\{\left[\frac{\pi^2\delta^2\,(E_T/N_0)}{4}\right]^{1/2}\right\}.$$

Whether or not this is satisfied depends on P_E and the parameter $\Delta = \delta(1-\rho^2)^{3/2}$. This dependence is shown in Table 6, where values of Δ_{crit} are shown for several values of P_E. The significance of Δ_{crit} is that for $\Delta_{crit} < \Delta$ one unambiguous measurement is superior, and for $\Delta < \Delta_{crit}$ two ambiguous measurements are superior.

In most cases of interest Δ satisfies

$$\Delta = \delta(1-\rho^2)^{3/2}$$
$$= \delta_\theta(1-\rho^2)^{3/2}\cos\theta \leq \delta_\theta < 0.35 = 20 \text{ deg}$$

so usually two ambiguous measurements will be superior to one unambiguous measurement.

Table 6. Critical values of Δ for several values of P_E

P_E	Δ_{crit}
0.1	0.36
0.03	0.40
0.01	0.42
0.003	0.44
0.001	0.45

8. Conclusion

We have examined the use of two and three colinear antennas to make unambiguous interferometric estimates of the direction of arrival of RF energy with plane wave fronts, and have seen that the maximum-likelihood procedure for this consists of first estimating the phase of the signal received at each antenna with respect to an arbitrary reference phase, and then processing the differences between these estimates to obtain an unambiguous estimate of the direction from which the wave fronts are arriving. Furthermore, we have seen that when an accuracy of better than 20 deg is desired, it is better to make two ambiguous measurements and then resolve the ambiguity, than it is to make one unambiguous measurement. Also, we determined the best antenna spacing ratio M/N for two ambiguous measurements, and noted that the probability that the error in the direction estimate exceeds some threshold value depends on the antenna spacing ratio only, not on the absolute distance between the antennas.

E. Frequency Demodulation as an Estimation Problem

J. K. Holmes

1. Summary

This article treats the demodulation of frequency-modulated Gaussian signals in the presence of Gaussian noise as a problem in Estimation Theory. The approach is to maximize the *a posteriori* probability density of the modulation vector **m** given the data vector **y**. Two integral equations are developed that specify the best estimate of $m(t)$. The important case of white, additive noise is then considered. A system for estimating $m(t)$ is then developed along with a simpler approximate system that closely resembles the phase-locked loop.

2. Introduction

The first approach to estimation of a non-linear function of a random process was done by Lehan and Parks (Ref. 10). Youla followed with a more rigorous formulation (Ref. 11). Later, other works have generalized and extended their work (Refs. 12 and 13) in terms of fading

channels, multidimensional channels, etc. All these previous derivations, however, used eigenfunction expansions to represent the modulating process. This article uses a sampled data approach suggested by Viterbi.

3. Formulation

The transmitted signal is assumed to be of the form

$$x(t) = 2^{1/2}A \sin\left[\omega_0 t + \int_{t_0}^{t} m(\sigma)\, h(t-\sigma)\, d\sigma\right], \quad t_0 \leq \sigma \leq t_M \tag{1}$$

where $h(t)$ is the impulse response of $H(\omega)$, the preemphasis and integration filter (Fig. 20), and $m(t)$ is the modulation, assumed to be a stationary Gaussian process with zero mean and covariance function $R_m(\tau)$.

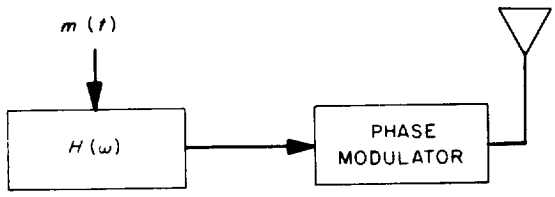


Fig. 20. Modulation system

The received waveform is assumed to be of the following form

$$y(t) = 2^{1/2}A' \sin\left[\omega_0 t + \int_{t_0}^{t} m(\sigma)\, h(t-\sigma)\, d\sigma\right] + n(t), \tag{2}$$

where the observation interval is (t_0, t_M) and $n(t)$, the noise, is a stationary Gausian process with zero mean and covariance function $R_n(\tau)$.

Consider forming an estimate of the modulation process $m(t)$ based on samples of the received waveform taken at Δt seconds apart so that

$$y(t_k) = 2^{1/2}A' \sin\left[\omega_0 t_k + \sum_{t_i \leq t_k} m(\tau_i)\, h(t_i - \tau_i)\, \Delta\tau_i\right] + n(t_k), \tag{3}$$

where

$$t_k = t_0 + k\Delta t. \tag{4}$$

We define

$$\mathbf{f}(\mathbf{m}) = (f(t_1, m_{t_1}), f(t_2, m_{t_2}), \cdots, f(t_M, m_{t_M})) \tag{5}$$

$$\mathbf{y} = (y_{t_1}, y_{t_2}, \cdots, y_{t_M}) \tag{6}$$

$$\mathbf{m} = (m_{t_1}, m_{t_2}, \cdots, m_{t_M}). \tag{7}$$

Since we desire the maximum *a posteriori* probability estimate of m, we have, from Bayes' rule,

$$p(\mathbf{m}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{m})}{P(\mathbf{y})} P(\mathbf{m}). \tag{8}$$

Now since the noise is independent of the modulation, the probability $p(\mathbf{y}|\mathbf{m})$ is just $p(\mathbf{y} = \mathbf{f}(\mathbf{m}, t))$, since $n(t) = y(t) - f(t, m_t)$ and given m_t, $n(t)$ is the only random term. Thus, $p(\mathbf{y}|\mathbf{m})$ is characterized by the multivariate normal probability density

$$p(\mathbf{y}|\mathbf{m}) = \frac{1}{(2\pi)^{M/2}|R_n|^{1/2}} \times \exp\{-\tfrac{1}{2}[\mathbf{y} - \mathbf{f}(\mathbf{m}, t)]\,[R_n]^{-1}\,[y - f(\mathbf{m}, t)]^T\} \tag{9}$$

where $[R_n]$ is the covariance matrix of the noise samples whose components are

$$R_n(t_j - t_k) = E\{n_{t_j} n_{t_k}\}, \qquad j, k = 1, 2, \cdots, M. \tag{10}$$

Since $\ln(x)$ is a continuous monotonic function of x, we can just as well maximize $\ln[p(\mathbf{m}|\mathbf{y})]$. Now

$$\ln p(\mathbf{m}|\mathbf{y}) = \ln p(\mathbf{y}|\mathbf{m}) - \ln p(\mathbf{y}) + \ln p(\mathbf{m}). \tag{11}$$

The M necessary conditions for a maximum to exist are obtained by differentiating Eq. (11) in conjunction with Eq. (9)

$$0 = \frac{\partial p(\hat{\mathbf{m}}|\mathbf{y})}{\partial \hat{m}_{t_n}} = \frac{\partial}{\partial \hat{m}_{t_n}} \ln p(\hat{\mathbf{m}}) + [\mathbf{y} - f(\hat{\mathbf{m}})]\,[R_n]^{-1}\left[\frac{\partial \mathbf{f}(\hat{\mathbf{m}})}{\partial \hat{m}_{t_n}}\right]^T \tag{12}$$

$$n = 1, 2, \cdots, M$$

so if we define the M dimensional vector $\mathbf{g}(\hat{\mathbf{m}})$ as

$$[g(\hat{\mathbf{m}})]^T \Delta t = [R_n]^{-1}\,[\mathbf{y} - \mathbf{f}(\mathbf{m})]^T, \tag{13}$$

then Eq. (12) becomes

$$\frac{\partial}{\partial \hat{m}_{t_n}} \ln p(\hat{\mathbf{m}}) + [g(\hat{\mathbf{m}})] \left[\frac{\partial \mathbf{f}(\hat{\mathbf{m}})}{\partial \hat{m}_{t_n}} \right]^T \Delta t = 0. \quad (14)$$

From Eq. (13), we find that

$$y(t_j) - f(t_j, \hat{\mathbf{m}}) = \sum_{k=1}^{M} R_n(t_j - t_k)\, g(t_k, \hat{\mathbf{m}})\, \Delta t \quad (15)$$

$$j = 1, 2, \cdots, M;$$

and from Eq. (14), we have

$$\frac{\partial}{\partial \hat{m}_{t_n}} \ln p(\hat{\mathbf{m}}) = -\sum_{k=1}^{M} g(t_k, \hat{\mathbf{m}}) \frac{\partial f(t_k, \hat{m})}{\partial \hat{m}_{t_n}} \Delta t. \quad (16)$$

Employing our assumption that $m(t)$ is a stationary Gaussian random process with covariance function $R_m(t-s)$, the probability density function of $\mathbf{m}$ is

$$p(\mathbf{m}) = \frac{1}{(2\pi)^{M/2} |R_m|^{1/2}} \times \exp\left\{ -\tfrac{1}{2} \sum_{j=1}^{M} \sum_{k=1}^{M} m_{t_j} m_{t_k} \Gamma_m(t_j - t_k) \right\} \quad (17)$$

where $|R_m|$ is the determinate of $[R_m]$ and $\Gamma_m(t_j - t_k)$ is the jk^{th} element of the inverse of $[R_m]$. Combining Eqs. (16) and (17), we have

$$\sum_{r=1}^{M} m_{t_r} \Gamma_m(t_n - t_r) = \sum_{k=1}^{M} g(t_k, \hat{\mathbf{m}}) \frac{\partial f(t_k, \hat{\mathbf{m}})}{\partial \hat{m}_{t_n}} \Delta t. \quad (18)$$

Since $[\Gamma_m] = [R_m]^{-1}$, Eq. (19) may be inverted to give

$$\hat{m}_{t_r} = \sum_{n=1}^{M} R_m(t_r - t_n) \left\{ \sum_{k=1}^{M} g(t_k, \mathbf{m}) \frac{\partial f(t_k, \hat{\mathbf{m}})}{\partial \hat{m}_{t_n}} \right\} \Delta t, \quad (19)$$

while Eq. (14) may be inverted to yield

$$[R_n]\, [g(\hat{\mathbf{m}})]^T \Delta t = [\mathbf{y} - \mathbf{f}(\hat{\mathbf{m}})]^T \quad (20)$$

or

$$\sum_{k=1}^{M} R_n(t_j - t_k)\, g(t_k, \hat{\mathbf{m}})\, \Delta t = y(t_j) - f(t_j, \hat{\mathbf{m}}). \quad (21)$$

Taking the limit as $\Delta t \to 0$ while M tends to infinity such that

$$M \Delta t = t_M - t_0, \quad (22)$$

we have from Eq. (21)

$$y(t) - f(t, \hat{m}) = \int_{t_0}^{t_1} R_n(t - \sigma)\, g[\sigma, K[\hat{m}(\sigma)]]\, d\sigma, \quad (23)$$

where $K[\hat{m}(\sigma)]$ is a functional of $\hat{m}$. Using Eq. (3) with (19) we have

$$\hat{m}_{t_r} = \sum_{n=1}^{M} R_m(t_r - t_n) \left\{ \sum_{k=n}^{M} g(t_k, \hat{\mathbf{m}})\, 2^{1/2} A' \times \cos\left[\omega_0 t_k + \sum_{t_i \le t_k} m(\tau_i)\, h(t_i - \tau_i)\, \Delta\tau \right] \times h(t_k - \tau_k) \right\} \Delta\tau \Delta t. \quad (24)$$

In the limit as $\Delta\tau$ and $\Delta t \to 0$ and $M \to \infty$ such that $M\Delta\tau$ and $M\Delta t = t_M - t_0$, we have

$$\hat{m}(t) = \int_{t_0}^{t_M} R_m(t - \tau) \int_{\tau}^{t_M} \left\{ g(t', K[\hat{m}(t')])\, 2^{1/2} A' \times \cos\left[\omega_0 t' + \int_{t_0}^{t'} m(\sigma)\, h(t' - \sigma)\, d\sigma \right] \right\} \times h(t' - \tau)\, dt'\, d\tau. \quad (25)$$

Hence, the basic integral equations to be solved are (23) and (25).

The case of greatest interest occurs when the noise is white with a covariance function given by

$$R_n(\tau - \sigma) = \frac{N_0}{2} \delta(\tau - \sigma). \quad (26)$$

Then Eq. (23) becomes

$$y(t) - f(t, \hat{m}) = \int_{t_0}^{t_M} \frac{N_0}{2} \delta(\tau - \sigma)\, g\{\sigma, K[\hat{m}(\sigma)]\}\, d\sigma \quad (27)$$

or

$$g[\tau, K[\hat{m}(\tau)]] = \frac{2}{N_0} \{ y(\tau) - f[\tau, \hat{m}(\tau)] \}, \quad (28)$$

and Eq. (25) using Eq. (28) becomes

$$\hat{m}(t) = \frac{8^{1/2}A'}{N_0}\int_{t_0}^{t_M} d\tau\, R_m(t-\tau)\left\{\int_{\tau}^{t_M}\left(y(t') - 2^{1/2}A'\right.\right.$$
$$\times \sin\left[\omega_0 t' + \int_{t_0}^{t'} \hat{m}(\sigma)\, h(t'-\sigma)\, d\sigma\right]\Bigg)$$
$$\left.\times \cos\left[\omega_0 t' + \int_{t_0}^{t'} m(\sigma)\, h(t'-\sigma)\, d\sigma\right] h(t'-\tau)\, dt'\right\} \tag{29}$$

4. Mechanizations of the Solution

Eq. (29) may be solved by the system depicted in Fig. 21. We point out that this mechanization is unrealizable due to two considerations. One is the fact that $H_m(\omega)$, the Fourier transform of $R_m(\tau)$, is, in general, non-zero on the real line, and hence is not realizable without infinite delay. The second consideration is that the conjugate filter $h^*(t)$ required to produce the following input-output relation

$$e_o(t) = \int_t^{t_M} e_i(\sigma)\, h(\sigma - t)\, d\sigma. \tag{30}$$

(Hence, the output $e_o(t)$ is determined by integrating from the present into the future to give a present output value.) One is faced with a highly unrealizable situation.

By making the assumption that the second harmonic of the signal is greatly attenuated in passing through the filter $H_m(\omega)$, we may however write

$$\hat{m}(t) = \frac{8^{1/2}A'}{N_0}\int_{t_0}^{t_M} R_m(t-\tau)\left\{\int_{\tau}^{t_M} y(t')\right.$$
$$\times \cos\left[\omega_0 t' + \int_{t_0}^{t'} \hat{m}(\sigma)\, h(t'-\sigma)\, d\sigma\right]$$
$$\left.\times\, h(\tau - t')\, dt'\right\} d\tau. \tag{31}$$

The corresponding system realization is depicted in Fig. 22.

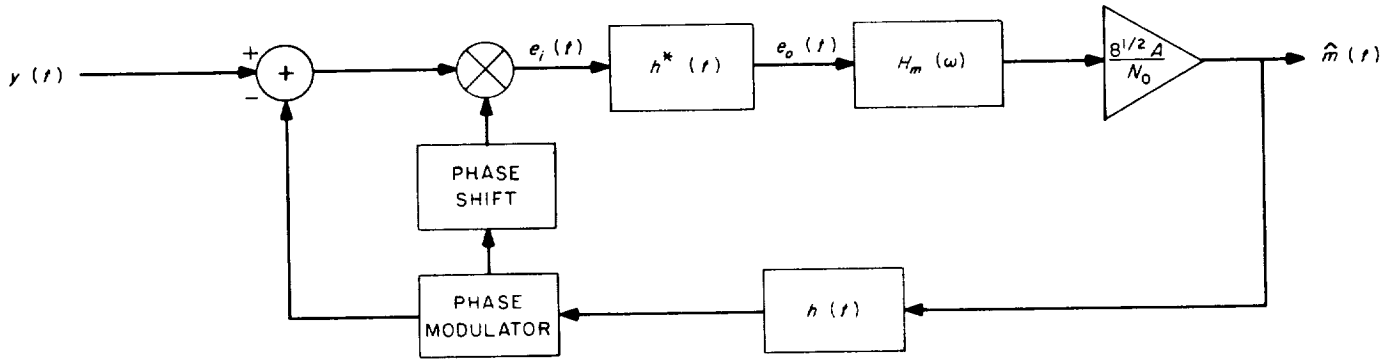


Fig. 21. Optimum FM demodulator (unrealizable)

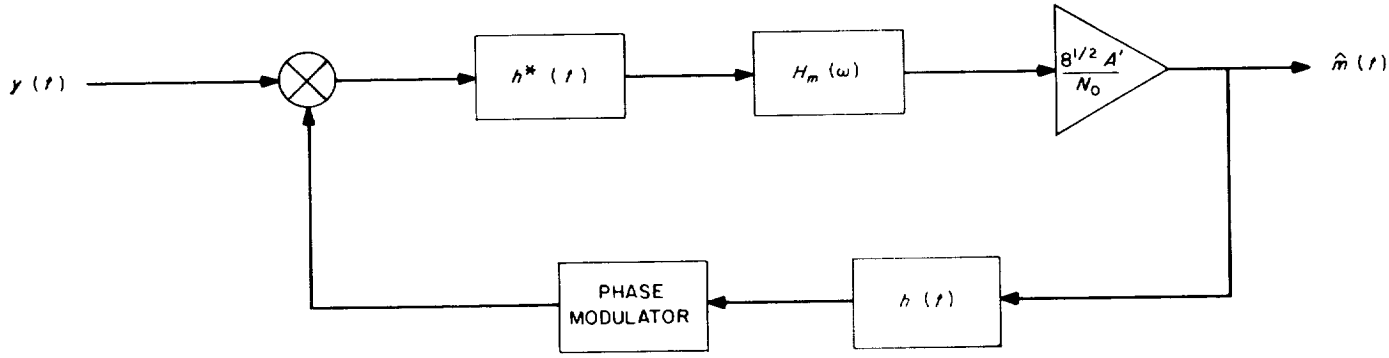


Fig. 22. Optimum FM demodulator (realizable)

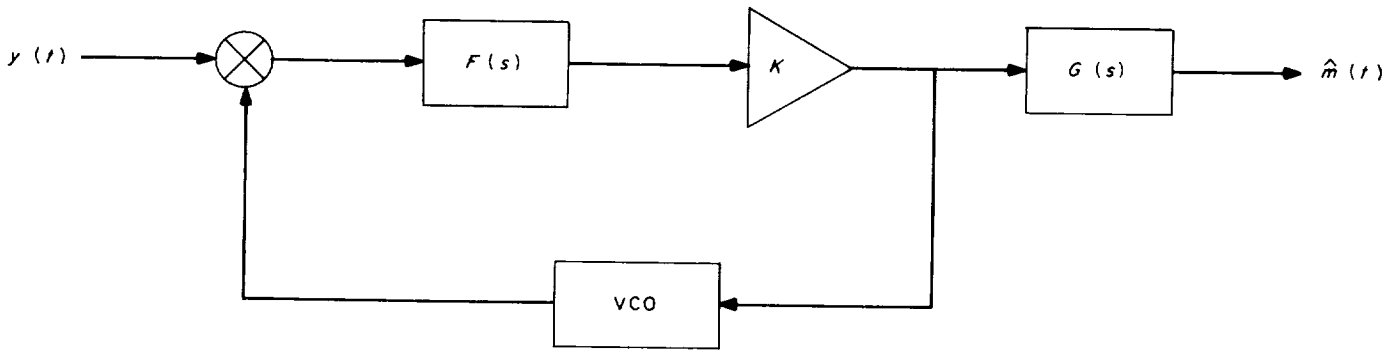


Fig. 23. Phase-locked loop as a frequency discriminator

If we now consider the phase-locked loop as a frequency demodulator (discriminator), we usually use the structure shown in Fig. 23. The integral equation for which this mechanization is given by

$$m(t) = \int_0^t \left\{ \int_0^\tau y(\sigma) \cos\left[\omega_0 \sigma + \int_0^\sigma m(t')\, dt' \right] \times F(\tau - \sigma)\, d\sigma \right\} g(t - \tau)\, d\tau. \quad (32)$$

Comparing Figs. 22 and 23, we see that the optimum demodulator and the phase-locked loop have practically the same structure. If we took the "structure preserving" approach to synthesize a realizable system from Fig. 22, we would combine $h^*(t)$ and $H_m(\omega)$ as one realizable filter $F'(s)$ and add an additional exterior filter to the loop $G'(s)$ and then adjust the filter parameters to give the least mean-squared-frequency error. It is interesting to note that in the practical design of phase-locked loops employed as frequency discriminators, the system (Fig. 23) excluding $G(s)$ is first chosen so that the mean-squared-phase error is a minimum, and then the total transfer function $H(s)$ is selected to minimize the mean-squared-frequency error from which $G(s)$ is then specified. Hence, it appears that in the structure-preserving sense, the present day phase-locked loop demodulators are near the realizable optimum.

Open problems include comparison of the variances of $m(t)$ for the realizable PLL discriminator and the optimum FM demodulator specified by Eq. (29). Another interesting study is the realization of these systems in some optimal way.

F. Conditional Frequency Uncertainty, Final Report

E. A. Yerman and E. C. Posner

1. Summary

This article completes work reported on in the last issue of this *Summary* as a preliminary report. Continuing previous work on spacecraft frequency acquisition, we considered the problem of acquiring a spacecraft in frequency, when the original dynamic coordinates were not known precisely, but were instead subject to some specified probability distribution. This report completes examination of the problem of acquisition at times later than injection.

The idea has been to produce a saving in frequency search time by conditioning the distribution of dynamic coordinates by the values of the look angles of the search antenna, since these angles are known quantities at all times. The resulting conditional distribution must have a distribution of frequencies with smaller variance. This variance reduction results in a shorter time to search the shorter frequency uncertainty interval thus obtained. Previous work reported indicated a saving in frequency search time by this method of a factor of two for a slow lunar trajectory, if the search is begun just at injection into the final orbit. In the preliminary report, we announced that if the search is begun as soon as 10 min after injection, the saving jumped to a factor of more than ten. In this report, savings are found for other post-injection times.

2. Review

Previous JPL work on the subject of speeding frequency acquisition by the use of conditional distributions appeared even before the preliminary results announced in Ref. 14 (Refs. 15, 16, 17). The model that has been considered is of the following type. Consider a Newtonian orbit, such as would be obtained on the last phase of a slow, nearly parabolic, but elliptical, lunar trajectory. At the instant of injection of the spacecraft into this Newtonian orbit, our knowledge of the six dynamical parameters determining its orbit is not precise, due, for example, to slight uncertainties in the guidance system, as well as other uncertainties of burn time, atmospheric drag, etc. However, such knowledge is sufficiently precise so that the location of the spacecraft in dynamic coordinate-space at injection can be described by a particular kind of probability distribution: a multivariate normal distribution on the six dynamical parameters, with a known covariance matrix. This assumption of joint normality is possible because the uncertainties involved are extremely small relative to the parameters themselves. Linear techniques coupled with the Central Limit Theorem then lead to a multivariate normal distribution.

One searches for the spacecraft by moving the antenna; then with the antenna then stopped (relative to the nominal trajectory), one searches for the spacecraft transponder frequency by moving the receiver VCO. (Frequency uncertainty arises from spacecraft velocity uncertainty via the doppler effect acting on the radial component of spacecraft velocity.) One searches the more likely antenna look angle first, but the frequency search then proceeds with a stopped antenna. Thus, two dynamical parameters are known during the time in which the frequency search is being carried out: the two look angles of the antenna are known. So the original distribution of six dynamic parameters is *conditioned* by the knowledge of these two parameters. The resulting four-dimensional conditional distribution results in a distribution of returning transponder frequency more tightly concentrated about its mean for the simple reason that the unconditioned six-dimensional distribution has to take into account all possible positions of the spacecraft on the celestial sphere. But the conditional four-dimensional distribution has only one unique position on the celestial sphere to take into account.

Consequently, one expects a reduction in the average time necessary to acquire the spacecraft in frequency. In Ref. 16, this saving was found to be a factor of almost two for a typical slow lunar trajectory, if one is searching right at injection.

However, the spacecraft may not have been visible at injection, or the spacecraft may not have been yet found for some other reason. It thus becomes relevant to ask what the frequency search time saving would be if the search is carried out after injection. The preliminary report (Ref. 14) initiated such a task. A saving at 10 min after injection of a factor of more than ten was found for a particular slow lunar trajectory. The saving at later times is still large, but this saving decreases after 10 or 12 min. Consideration of the geometry involved in a nearly parabolic orbit shows that this increase and then decrease is expected to occur. Far out on the nearly parabolic ellipse, conditioning the look angles has hardly any effect on toward-Earth doppler, since the spacecraft velocity vector becomes almost pointed directly at Earth at such orbital positions. The rest of this article indicates the methods and results used for finding the saving in expected search time for searches carried out after injection.

3. Coordinate Systems

Since methods of matrix theory are to be used in this work, we start by redefining the coordinate systems and transformations to be used shortly.

The geocentric coordinates (δx, δr, δv, $\delta \tau$, $\delta \zeta$, $\delta \dot{\zeta}$) as well as the station-centered spherical coordinates ($\delta \rho$, $\delta \theta$, $\delta \Phi$, $\delta \dot{\rho}$, $\delta (\rho \dot{\theta} \sin \Phi)$, $\delta (\rho \dot{\Phi})$) used in this study are defined in Refs. 14, 15, and 16. Since the multivariate normal error distribution in station-centered spherical coordinates is known at injection, the covariance matrix in station-centered spherical coordinates could, if one wished, be obtained directly by "mapping ahead" along the post-injection trajectory, staying in station-centered spherical coordinates. Approximating linear transformations, however, already exist by which multinormal error distribution in the *geocentric* coordinates are mapped ahead. Then the linear transformations derived in Ref. 16 will be used to determine the station-centered covariance matrix obtained for that time by the mapping-ahead procedure. The methods thus make use of the post-injection equations of motion for the two-body problem (Ref. 18)[5] to obtain the approximate linear mapping of spacecraft station-centered dynamical coordinates into the future.

[5]See also "Guidance for Space Missions," JPL External Publication No. 656, June 1959.

From this approximate linear mapping, the covariance matrix can then be mapped into the future by standard statistical theory as in Chapt. 9 of Ref. 19.

4. Matrices

Consider the motion of the spacecraft after injection under the assumption that only one attracting body (Earth) is present. This assumption is reasonable for our purposes since we wish to consider relatively small post-injection times, i.e., in the vicinity of 10 or 20 min. The post-injection plane of motion is called the *thrust plane*, the plane of the spacecraft velocity vector. We then map ahead along the trajectory in the familiar geocentric coordinates to obtain the covariance matrix of errors at the time in question and then utilize the approximating linear transformation mentioned above to obtain the corresponding covariance matrix in station-centered spherical coordinates (Ref. 16).

The geocentric phase of a lunar trajectory is comprised of two parts. One is the pre-injection portion which contains the parking orbit, and the second part is an ellipse with an eccentricity that is related to the time required to reach the Moon; the eccentricity is close to 1 but less than 1 for the nearly parabolic orbits used in the example. It should be mentioned, however, that the results presented below are general in the sense that they hold (with minor differences) for either elliptical, parabolic, or hyperbolic post-injection-trajectories with arbitrary eccentricities.

5. Mapping Error Distributions Along an Ellipse

We take over some results from Ref. 18 in proceeding to determine the transformation matrix $U(t, t_1) = (u_{ij})$ by which an error vector in geometric coordinate $\delta Z_1^T = (\delta x_1, \delta r_1, \delta v_1, \delta \tau_1, \delta z_1, \delta \dot{z}_1)$ at injection time t_1 can be transformed into an error vector $\delta Z^T = (\delta x, \delta r, \delta v, \delta \tau, \delta z, \delta \dot{z})$ at post-injection time t, i.e., $\delta Z = U(t, t_1)\, \delta Z_1$. Once this has been done, it is a simple matter to obtain the geocentric covariance matrix of the multinormal error distribution at time t.

We first define dynamical quantities in the equations of motion, for those unfamiliar with celestial mechanics. The equation of an ellipse (in fact, any conic, if e is allowed to vary) in polar coordinates, with center of the Earth at the right-hand focus, is given by

$r = p/1 + e\cos\theta = r(\theta_1, C_1, C_3)$ say, where:

θ = the so-called *true anomaly* (the angle from *periapsis* of the spacecraft at a particular time; periapsis of an orbit is the nearest Earth point of the spacecraft orbit);

v = the magnitude of the thrust plane velocity vector $\bar{v}$;

τ = the angle between this $\bar{v}$ and the *local horizon* (local horizon is the direction parallel to the Earth in the thrust plane at the point of intersection of the radius vector from the center of the Earth to the spacecraft and the Earth's surface);

r = the distance of the spacecraft from the center of the Earth;

r_0 = the radius of the Earth;

μ = the *gravitational constant* of the Earth (the mass of the Earth times the universal constant of gravitation);

C_1 = the *angular momentum* of the motion = $rv\cos\tau$, a constant;

ρ = the *semi-latus rectum* of the ellipse = C_1^2/μ (the half-length of the chord of the ellipse through a focus and perpendicular to the axis);

C_3 = *Vis-viva* (definition) = $v^2 - 2\mu/r$ (a negative constant for elliptical orbits);

e = $[1 + (C_1^2 C_3/\mu^2)]^{1/2}$, the *eccentricity*, which is less than 1 when C_3 is negative (as it is for elliptical orbits).

As an example, the elements of the matrix $U(t, t_1)$ appear in Eq. (2), using the matrix M of Appendix A of Ref. 18 (M is shown in Fig. 24):

$$\begin{bmatrix} 0 & \left(\frac{-\sin\theta_1}{er_1}\right) & \left(\frac{-2\sin\theta_1}{ev_1}\right) & \left(1 + \frac{e + \cos\theta_1}{e(1 + e\cos\theta_1)}\right) \\ 0 & \frac{C_1}{r_1} & \frac{C_1}{v_1} & (-C_1 \tan\tau_1) \\ 0 & \frac{2\mu}{r_1^2} & 2v_1 & 0 \end{bmatrix}$$

Fig. 24. The matrix *M*

$$
\begin{aligned}
&(u_{11}, u_{12}, u_{13}, u_{14}) = (1,0,0,0) + r_0 \\
&\qquad \times \left(\frac{\partial\theta}{\partial\theta_1} - 1, \frac{\partial\theta}{\partial C_1}, \frac{\partial\theta}{\partial C_3}\right) M; \\
&(u_{21}, u_{22}, u_{23}, u_{24}) = \left(\frac{\partial r_2}{\partial\theta}\frac{\partial\theta}{\partial\theta_1}, \frac{\partial r}{\partial C_1} + \frac{\partial r}{\partial\theta}\frac{\partial\theta}{\partial C_1},\right. \\
&\qquad \left. \frac{\partial r}{\partial C_3} + \frac{\partial r}{\partial\theta}\frac{\partial\theta}{\partial C_3}\right) M; \\
&(u_{31}, u_{32}, u_{33}, u_{34}) = \left(0, \frac{\mu}{v r_1^2}, \frac{v_1}{v}, 0\right) \\
&\qquad - \frac{\mu}{v r^2}(u_{21}, u_{22}, u_{23}, u_{24}); \\
&(u_{41}, u_{42}, u_{43}, u_{44}) = \frac{\cos^2\tau}{(1 + e\cos\theta)^2}\left[e(e + \cos\theta)\right. \\
&\qquad \times \left(\frac{\partial\theta}{\partial\theta_1}, \frac{\partial\theta}{\partial C_1}, \frac{\partial\theta}{\partial C_3}\right) \\
&\qquad \left. + \left(\frac{e^2 - 1}{e}\right)\sin\theta\left(0, \frac{1}{C_1}, \frac{1}{2C_3}\right)\right] M; \\
&(u_{55}, u_{56}) = \frac{r}{C_1}[v_1\cos(\theta - \theta_1 + \tau_1), \\
&\qquad \times r_1\sin(\theta - \theta_1)]; \\
&(u_{65}, u_{66}) = \frac{v}{C_1}[v_1\sin(\theta - \theta_1 + \tau_1 - \tau), \\
&\qquad \times r_1\sin(\theta - \theta_1)]; \\
&u_{ij} = 0 \text{ for } \left\{\begin{matrix} i = 1,2,3,4 \\ j = 5,6 \end{matrix}\right\} \text{and} \left\{\begin{matrix} i = 5,6 \\ j = 1,2,3,4 \end{matrix}\right\}.
\end{aligned}
$$

At the time t_1 of injection, the angle turned from periapsis for the particular slow lunar trajectory to be used as an example is $+3°19'$ (periapsis would occur before injection). The values of the angle θ for the post-injection times of 5, 10, 20, and 30 min are: 30°, 52°43′, 80°31′, and 96°2′, respectively, for the orbit considered.

Let the geocentric covariance matrix at post-injection time t be denoted by V. Recall that the transformation matrices $U(t, t_1)$ denote the linear transformation required to map ahead the error distributions to t the orbital positions corresponding to t. Thus, one obtains the vector and matrix equations

$$
\left.\begin{aligned}
&\delta Z = U(t, t_1)\,\delta Z_1 \\
&\text{and} \\
&V = U(t, t_1)\,V_1 U^T(t, t_1).
\end{aligned}\right\} \qquad (3)
$$

The transformation matrix $U(t, t_1)$ for $t = 10$ min is given in Fig. 25, while the corresponding covariance matrix V for the same t is given in Fig. 26.

Also, one finds that for the post-injection ellipse under consideration, $e = 0.97$, $C_1 = 71{,}813$ km²/sec, $C_3 = -1.75$ km²/sec², $\rho = 12.94 \times 10^3$ km. These four orbital parameters are determined by the initial values of the spacecraft's position and velocity at time t_1, the time of injection into the final Newtonian orbit.

1	−0.0187	1195 sec	−4144 km	0	0
0	1.444	1005 sec	499.9 km	0	0
0	0.0001 sec^{-1}	0.5013	−3.064 km/sec	0	0
0	0.0003 km^{-1}	0.3276 sec/km	0.6688	0	0
0	0	0	0	0.7639	566 sec
0	0	0	0	0.0004 sec^{-1}	0.8184

Fig. 25. The matrix $U(t, t_1)$

1626 km²	−374.1 km²	466.3 × 10^{-3} km²/sec	17.32 × 10^{-3} km	0	0
	210.8 km²	−131.7 × 10^{-3} km²/sec	19.3 × 10^{-3} km	0	0
		1.392 × 10^{-4} km²/sec²	51.13 × 10^{-8} km/sec	0	0
			4.546 × 10^{-6}	0	0
				233.7 km²	25.1 × 10^{-2} km²/sec
Symmetric					289.6 × 10^{-6} km²/sec²

Fig. 26. The geocentric covariance matrix V

The four post-injection orbital positions at $t = 5$, 10, 20, 30 min and their angular relationship to the DSIF Station zenith are depicted in Fig. 27. These angles are of interest since they play an important role in determining qualitative properties of the linear transformations used in the next section to obtain the station-centered covariance matrices corresponding to the orbital positions considered. It was of interest to note after the necessary calculations were made that the variances in the position coordinates x, r, τ, and z increased with time. From a heuristic point of view, this is plausible, since the position coordinates increase with time and it is reasonable to expect their dispersions to become larger. On the other hand, the spacecraft is slowing down after injection and it is to be expected, and was indeed verified, that the dispersions in the velocity coordinates actually decrease with time.

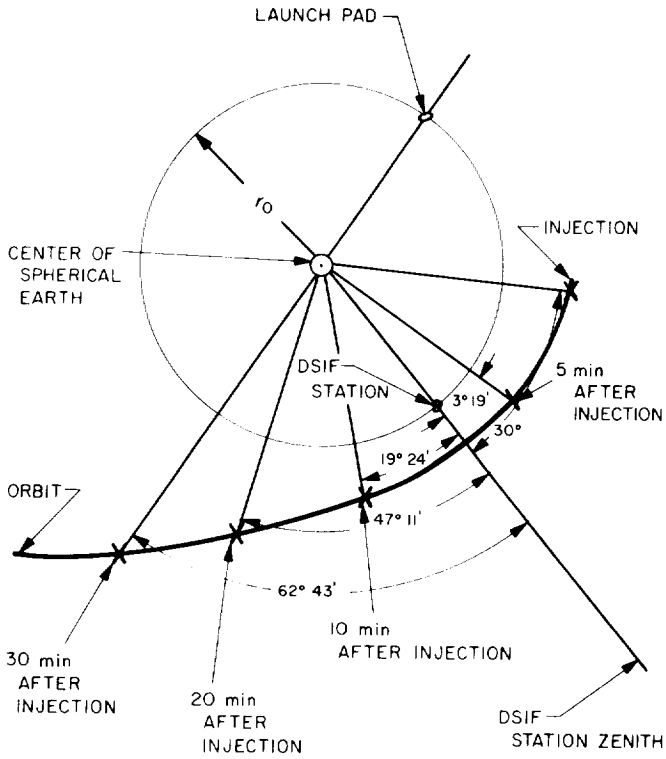


Fig. 27. Post-injection orbital positions

6. Error Distributions in Spherical Coordinates

The transformation matrices D by which a geocentric error vector $\delta Z^T = (\delta x, \delta r, \delta v, \delta \tau, \delta z, \delta \dot{z})$ at post-injection time t is transformed into an error vector in station-centered spherical coordinates $\delta S^T = (\delta \rho, \delta \theta, \delta \Phi, \delta \dot{\rho}, \delta (\rho \dot{\theta} \Phi \sin \Phi), \delta (\rho \dot{\Phi})$ was obtained by us in the previous article (Ref. 14) by using the previously mentioned linear transformations. This transformation matrix is given in Fig. 28 for $t = 10$ min. Since $\delta S = D \delta Z$, it follows that the station-centered covariance matrices $\overline{V}$ are determinable: the one for $t = 10$ min is given in Fig. 29. (The spherical covariance matrix and the transformation matrix for injection appeared in Ref. 16.)

The unconditional standard deviations of $\delta \dot{\rho}$ were then computed, and are graphed in Fig. 30. One sees that the dispersion in $\dot{\rho}$ increases and then decreases. This seems reasonable, since $\dot{\rho}$ increases, reaches a maximum directly over the DSIF Station, and then decreases with time. Thus, the uncertainty in $\dot{\rho}$ should become larger as $\dot{\rho}$ increases, then decrease when $\dot{\rho}$ decreases.

We now condition the six-dimensional error distribution by assuming the look angles θ, Φ to be known. This yields a four-dimensional error distribution in the remaining variables. Let the 4×4 covariance matrices for the (conditioned) random variables $(\delta \rho, \delta \dot{\rho}, \delta (\rho \dot{\theta} \sin \Phi), \delta (\rho \dot{\Phi}))$ be denoted by B. This matrix is given for $t = 10$ min in Fig. 31. The conditional standard deviations of $\delta \dot{\rho}$ are also plotted in Fig. 30 for this example. This means that the reduction, due to conditioning, in the standard deviation of $\delta \dot{\rho}$ is by the factors given in Fig. 32. The maximum reduction of the factor of 11.7 occurs at about 10 min post-injection. As was observed in Part 2 of this

$$\begin{bmatrix} 0.8895 & 0.6599 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0004\ \text{km}^{-1} & 0 \\ 0.0003\ \text{km}^{-1} & -0.0002\ \text{km}^{-1} & 0 & 0 & 0 & 0 \\ 0.0005\ \text{sec}^{-1} & -0.0006\ \text{sec}^{-1} & 0.9215 & 2.809\ \text{km/sec} & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.0036\ \text{sec}^{-1} & 1\ \text{sec}^{-1} \\ -0.0481\ \text{sec}^{-1} & -0.0134\ \text{sec}^{-1} & 1.615 & -1.5\ \text{km/sec} & 0 & 0 \end{bmatrix}$$

Fig. 28. The transformation matrix D

$$
\begin{bmatrix}
963\ \text{km}^2 & 0 & 40.4 \times 10^{-2}\ \text{km} & 1.11\ \text{km}^2/\text{sec}^2 & 0 & -55.3\ \text{km}^2/\text{sec}^2 \\
 & 3.74 \times 10^{-5} & 0 & 0 & -2.36 \times 10^{-4}\ \text{km/sec} & 0 \\
 & & 2 \times 10^{-4} & 5.31 \times 10^{-4}\ \text{km/sec} & 0 & -2.47 \times 10^{-2}\ \text{km/sec} \\
 & & & 1.42 \times 10^{-3}\ \text{km}^2/\text{sec}^2 & 0 & -6.71 \times 10^{-2}\ \text{km}^2/\text{sec}^2 \\
 & & & & 1.51 \times 10^{-3}\ \text{km}^2/\text{sec}^2 & 0 \\
\text{Symmetric} & & & & & 3.25\ \text{km}^2/\text{sec}^2
\end{bmatrix}
$$

Fig. 29. The station-centered covariance matrix *V*

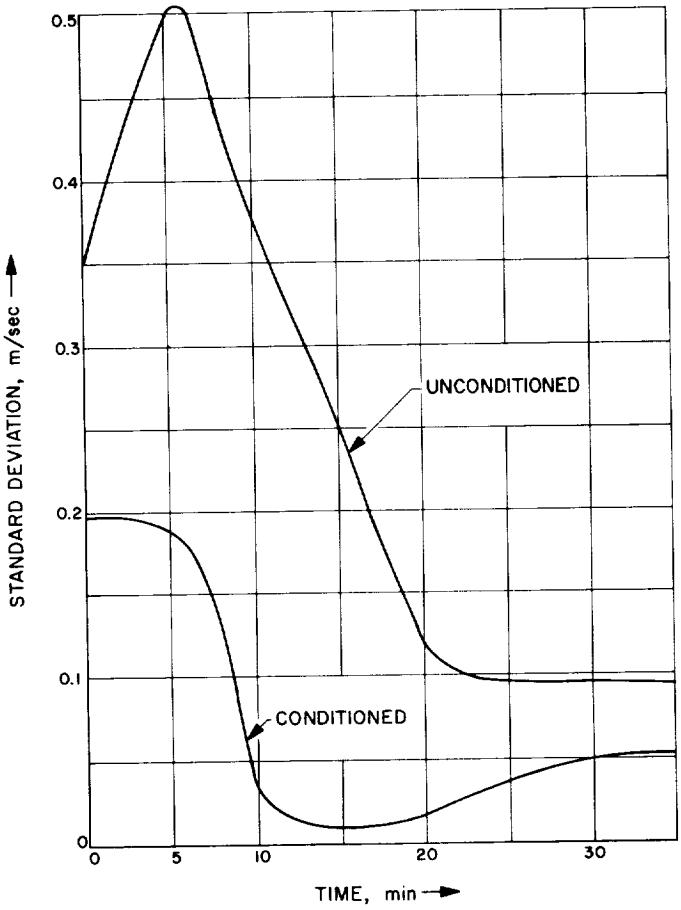


Fig. 30. Standard deviation of $\delta\dot{\rho}$ versus *t*

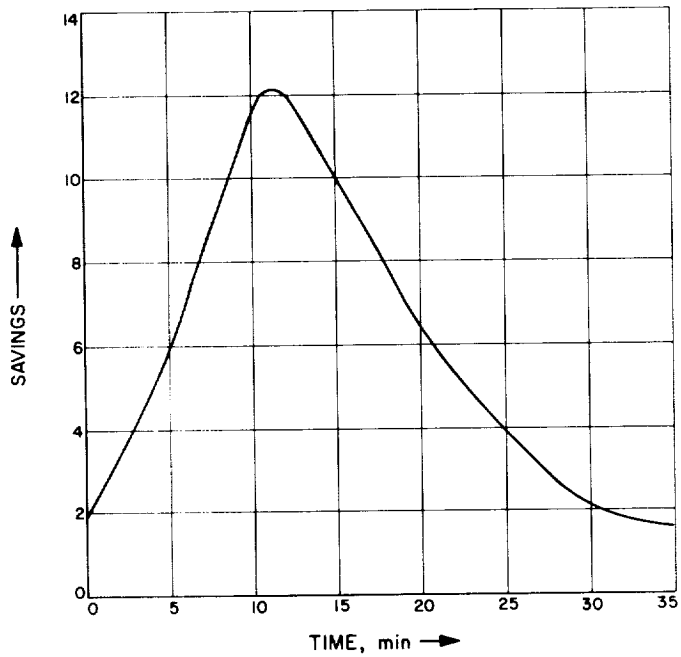


Fig. 32. Savings in expected search time

article, it is to be expected that the reductions should increase and then decrease, since the unconditional variance of $\dot{\rho}$, θ, and Φ do likewise.

7. The Conditional Expected Values of $\dot{\rho}$

Let the unconditioned expected values of $\dot{\rho}$, θ, and Φ be denoted by $\dot{\rho}_0$, θ_0, and Φ_0 for some arbitrary time t

$$
\begin{bmatrix}
136\ \text{km}^2 & 3.52 \times 10^{-2}\ \text{km}^2/\text{sec} & 0 & -4.95\ \text{km}^2/\text{sec} \\
 & 1.04 \times 10^{-5}\ \text{km}^2/\text{sec}^2 & 0 & -1.28 \times 10^{-3}\ \text{km}^2/\text{sec}^2 \\
 & & 2 \times 10^{-5}\ \text{km}^2/\text{sec}^2 & 0 \\
\text{Symmetric} & & & 0.18\ \text{km}^2/\text{sec}^2
\end{bmatrix}
$$

Fig. 31. The conditioned covariance matrix *B*

under consideration. Denote the conditional expected value of $\dot{\rho}$ by $\dot{\rho}_0 + m$, where m is the *offset* in the mean of $\dot{\rho}$ due to conditioning. Using the statistical techniques of linear regression used in Ref. 16, it follows that m is given for the example as in Fig. 33. For $t = 10$ min, then:

$$m = -(\Phi - \Phi_0)\{(3.522 \times 10^{-2}\ \text{km}^2/\text{sec})(11.362780\ \text{km}^{-1}) + (1.0402501 \times 10^{-5}\ \text{km}^2/\text{sec}^2) \times (-3.2635394 \times 10^{8}\ \text{sec/km}) + (0\ \text{km}^2/\text{sec}^2)(0\ \text{sec/km}) + (-1.2808709 \times 10^{-3}\ \text{km}^2/\text{sec}^2) \times (3.09793 \times 10^{8}\ \text{sec/km})\}.$$

Then at $t = 10$ min, $m = (\Phi - \Phi_0)(2.624\ \text{km/sec})$. The factor 2.624 km/sec is called the *offset factor.*

Thus, just as in Refs. 16 and 17, the required offset can be accomplished by a simple linear operation on only one of the look angles, a very desirable result. In particular, *the conditional expected value of $\dot{\rho}$ is independent of the conditioning value of θ.* This fact is predictable without calculation, since variations about $\theta_0 = 0$ have second-order effect on the value of $\dot{\rho}$, whereas the motion of the spacecraft nominally takes place in the thrust plane, i.e., the plane for which $\theta_0 = 0$. Also observe that for each time considered here, $-m$ always has approximately the value of one unconditioned standard deviation unit of $\dot{\rho}$, if the deviation $\Phi - \Phi_0$ is equal to one unconditional standard deviation unit of the angle $\bar{\Phi}$. This fact is also predictable in advance from the general form of the regression equation.

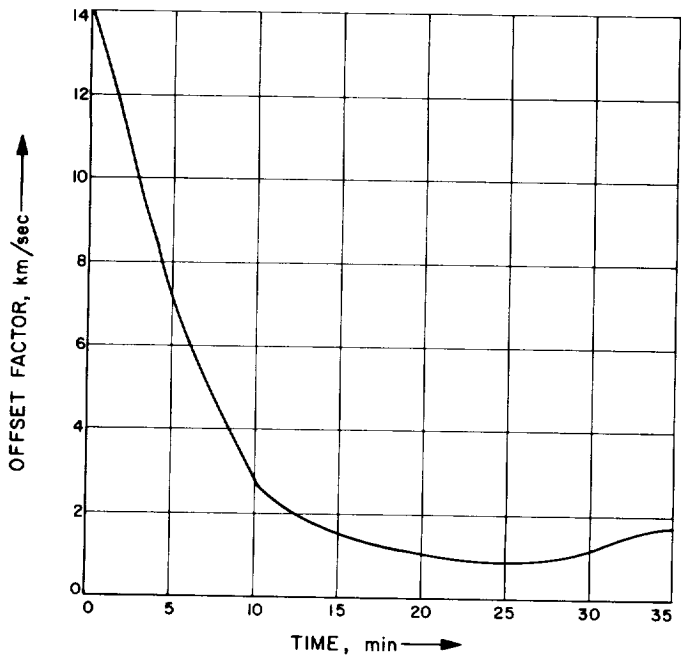


Fig. 33. Offset in mean of $\dot{\rho}$

8. Conclusions

This study has demonstrated that large savings in expected frequency search time can be obtained in spacecraft acquisition by using the theory of conditional distributions. By using the known values of the antenna look angles at a given time, the standard deviation in returning transponder frequency can be reduced by factors of more than ten for times shortly after injection, typical of times when the spacecraft is first visible over the South Africa DSIF Station after injection. Since it was shown previously (Refs. 16 and 17) that expected frequency search time is directly proportional to this standard deviation, a reduction in the average search time by the same factors of more than ten is likewise realizable. Typically a "5σ" search is used. That is, the half-width of the frequency uncertainty interval is five standard deviation units, conditioned or unconditioned, resulting in extremely high probability of acquiring the correct frequency (when the spacecraft is at the given look angles).

To realize this saving, however, it was shown in Ref. 17 that the center of the frequency uncertainty interval is changed by the conditioning even though the length of this search interval does indeed decrease by factors of more than ten. But since the necessary offset is a linear function of the offset of only one of the look angles from its nominal value, it is especially easy to accomplish this offset: merely read the difference between actual angle and nominal angle from the antenna coordinate converter, and then multiply by a number depending on the orbit and the time. Knowing both the width and center of the frequency uncertainty interval, one knows the frequency uncertainty interval itself, and the search can then proceed.

Because of the simplicity of the implementation of this system, it is therefore a good one to adopt for spacecraft frequency acquisition.

G. Optimum Coherent Amplitude Demodulation

W. C. Lindsey

1. Introduction

The process of optimum demodulation of an analog signal has recently been receiving a great deal of attention. In particular, the process of computing the set of *a posteriori* probabilities for a given set of deterministic signals in noise is well-known for a rather broad class of fixed and linear time-varying channels. On the other hand, the performance of a large subclass of these detectors has been analyzed in detail and tabulated.

The specifications of *a posteriori* probability-computing processors for detecting analog signals are not so well-known, and certainly the performance of these processors is less organized than in the case of detecting deterministic signals in noise. Particularly, the *a posteriori* probability-computing processors for frequency and phase modulation turn out to be unrealizable, and it seems that a set of integral equations particularly difficult to solve lies at the heart of the problem.

For linear amplitude modulation–demodulation systems, the realizable and unrealizable *a posteriori* probability-computing processors have been specified by Thomas and Wong and more recently by Van Trees in a more general formulation of the problem (Refs. 20 and 21). In this article we characterize the performance of a broad class of amplitude demodulation processors in which the random modulating signals may be selected from two broad classes of modulating spectra. This is done for both the realizable and unrealizable forms of the receiver structure. In the amplitude demodulation case, that signaling processor which computes the *a posteriori* probability that the modulation takes on a certain waveshape is identical with that processor which minimizes the mean-squared-error. This identity is, of course, a result of linearity of the modulation process.

2. Modulation and Spectra

A problem of current interest in the area of space communications is that of utilizing one of the neighboring planets as a parasitic antenna for reflecting an analog signal between two widely separated (or the same) points on the Earth, and then detecting the transmitted signal. The problem that the system-design engineer must face is that of selecting the modulation–demodulation technique which allows for the most unambiguous detection procedure at the receiver. In this paper we consider four types of amplitude modulation–demodulation techniques and compare each technique based on two classes of modulating spectra. These are:

(1) Amplitude-modulated, double-sideband (AM-DSB).

(2) Amplitude-modulated, double-sideband, suppressed carrier (AM-DSB/SC).

(3) Amplitude-modulated, single-sideband (AM-SSB).

(4) Amplitude-modulated, single-sideband, suppressed carrier (AM-SSB/SC).

The results obtained, however, may be applied to a wide class of problems which requires the transmission of analog information to a distant point.

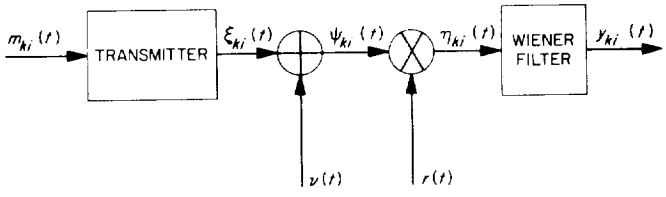


Fig. 34. Communication link—DSB

The communication links under consideration are depicted in Figs. 34 and 35. At the transmitter the k^{th} random process of the i^{th} message class $\{m_{ki}(t)\}$; $i = 1, 2$; $k = 1, 2, \infty$, is used to modulate the transmitter. The output waveform, say $\xi_{ki}(t)$, is transmitted into the channel

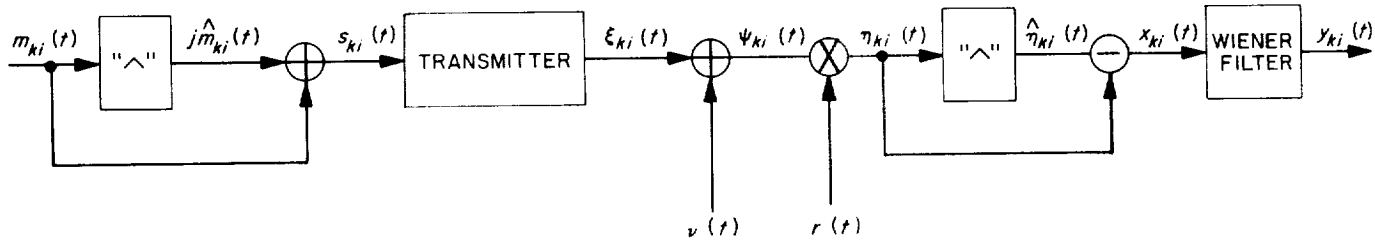


Fig. 35. Communication link—SSB

where additive white Gaussian noise of single-sided spectral density N_0 w/cps corrupts the transmitted equation display waveform resulting in the received waveform $\psi_{ki}(t) = \xi_{ki}(t) + \nu(t)$.

The detection procedure is carried out as follows: The observed data $\psi_{ki}(t)$ is multiplied by a noisy copy, say $r(t)$, of the transmitted carrier and the resulting waveform $\eta_{ki}(t)$ is filtered (after an appropriate transformation in the SSB systems) by one of the two types of Wiener filters, i.e., the appropriate linear filter which minimizes the mean-square-error (Ref. 22). A filter of Type I works as follows: The input function $\eta_{ki}(t)$ (or $x_{ki}(t)$) is recorded for a certain interval of time (theoretically for $-\infty < t < \infty$) and is then processed. For Type II filtering we assume that the filter is physically realizable and can be constructed by a circuit containing resistances, inductances, and capacitances. In certain applications, e.g., reflection of the analog signal from a neighboring planet where delay in the demodulation procedure is of no importance, Type I filtering may be practical.

The advantages of Type II filters are the simplicity with which they may be implemented and the rapidity with which the output data is delivered. The advantage of Type I filters is the more complete use they make of the input signal; consequently, the additive noise can be suppressed more effectively. A comparison of both types of filtering action will be given (for two classes of message spectra) on the basis of a "signal-to-noise ratio" related to the Wiener error versus a "signal-to-noise ratio" determined by initial design parameters.

At the transmitter we presume we have available two classes of stationary time series with spectral densities denoted by $S_1(\omega; k)$ and $S_2(\omega; k)$, $(k = 1, 2, \cdots, \infty)$. Class 1 is taken to be of the "maximally flat" form, i.e.,

$$S_1(\omega; k) = K_1(k)\,[1 + (\omega/a)^{2k}]^{-1}; \qquad k = 1, 2, \cdots, \infty \tag{1}$$

where $K_1(k)$ is a constant which is chosen such that the time series which it represents has unit variance. For the class of "maximally flat" spectra $K_1(k)$ is given by

$$K_1(k) = (\pi/a)\,\mathrm{sinc}\,(\pi/2k). \tag{2}$$

(We have adopted the notation that $\mathrm{sinc}\ x = \sin x/x$.) This process is both physically reasonable and mathematically convenient; the integer k is a measure of the rate of spectrum cutoff. For example, $k = 1$ corresponds to a dropoff of 6 db per octave, $k = 2$ corresponds to 12 db per octave, etc. Further, $a/2\pi$ may be considered to be the half-power frequency of the time series $m_{ki}(t)$. If $k = 1$, $S_1(\omega; 1)$ is the spectral density occurring at the output of an RC circuit whose input is white Gaussian noise. For $k = \infty$, we have

$$S_1(\omega; \infty) = \begin{cases} \pi/a; & |\omega| < a \\ 0; & |\omega| > a \end{cases}$$

which is the impulse power response of an ideal low-pass filter of bandwidth $a/2\pi$ cps.

Class 2 processes are taken to be the stationary "Asymptotically Gaussian" processes with a spectral density given by

$$S_2(\omega; k) = K_2(k)\,\{1 + [\omega/a\,(k)^{1/4}]^2\}^{-k}; \qquad k = 1, 2, \cdots, \infty \tag{3}$$

and $K_2(k)$ is adjusted such that the k^{th} member of the process has unit variance. For this class $K_2(k)$ turns out to be

$$K_2(k) = 4\pi\,[a\,(k)^{1/2}\,B\,(1/2, k - 1/2)]^{-1} \tag{4}$$

where $B(\mu, \nu)$ is the well-known Beta function. If $k = 1$, $S_1(\omega; 1) = S_2(\omega; 1)$ while k approaches infinity in Eq. (3), we have

$$S_2(\omega; k) = K_2(k) \exp\left[-k\left\{\frac{(\omega/a)^2}{k} - \frac{(\omega/a)^4}{2k^2} + \cdots\right\}\right]$$

or

$$\lim_{k \to \infty} S_2(\omega; k) = 2\,(\pi/a)^{1/2} \exp\,[-(\omega/a)^2] \tag{5}$$

which is the Gaussian spectrum. This unit variance process is rather interesting from the physical standpoint in that it can be generated by passing white Gaussian noise through k isolated-cascaded RC networks.

Note that the two random processes possess radically different frequency components as k becomes large. For $k = \infty$, the parameter $a/2\pi$ may be considered to be that frequency at which the spectrum has decayed to $1/e$ times the value at $\omega = 0$. These two classes of random processes are sufficiently general in that they include a broad class of signaling spectra encountered in communication engineering.

3. The Transmitted Signals

A representation of these signals which is most convenient for our purposes is to represent the transmitted waveforms as the product of a real low-pass waveform,

which depends on the modulating signal $m_{ki}(t)$, and a complex cisoidal carrier. For comparison purposes we normalize the transmitted signals such that the average transmitted power is P watts regardless of the type of modulation employed. The set of transmitted signals which possess an average power of P watts are for the double-sideband systems

$$\left.\begin{aligned} \xi_{ki}(t) &= (2P/1 + m_a^2)^{1/2}\,[1 + m_a m_{ki}(t)]\exp(j\omega t); \quad \text{AM-SSB} \\ \xi_{ki}(t) &= (2P)^{1/2}\, m_{ki}(t)\exp(j\omega t); \quad \text{AM-SSB/SC} \end{aligned}\right\} (6)$$

where we have assumed we are transmitting the k^{th} member of the i^{th} message class. In Eq. (6), ω is a suitably defined carrier frequency and $100m_a$ is a measure of the percent of amplitude modulation. It should be noted that the real parts of Eq. (6) represent the physical signal emitted by the transmitter.

The AM-SSB (Type-III) and AM-SSB/SC (Type-IV) signals are a bit more difficult to generate. For these systems we transmit the real parts of, respectively

$$\left.\begin{aligned} \xi_{ki}(t) &= (2P/1 + 2m_a^2)^{1/2}\,[1 + m_a s_{ki}(t)]\exp(j\omega t); \quad \text{AM-SSB} \\ \xi_{ki}(t) &= P^{1/2} s_{ki}(t)\exp(j\omega t); \quad \text{AM-SSB/SC} \end{aligned}\right\} (7)$$

The process $s_{ki}(t)$ is generated at the transmitter in the following manner. Assume that the k^{th} member of the i^{th} process is being transmitted. The time series $m_{ki}(t)$ is passed through a *Hilbert Transforming*, ("$\wedge$"), filter whose output has been phase-shifted by 90 deg and represented by the waveform $j\hat{m}_{ki}(t)$. This process is added to produce the signal (with one-sided frequency components)

$$s_{ki}(t) = m_{ki}(t) + j\hat{m}_{ki}(t). \tag{8}$$

(See Fig. 35 and Ref. 23.) Eqs. (6) and (7) characterize the set of "normalized" transmitted signals $\{\xi_{ki}(t); i = 1,2; k = 1, 2, \cdots, \infty\}$ using complex carriers. In physical situations the suppressed carrier signals may be generated by means of balanced modulators.

4. Characterization of the Additive Noise

We presume that the complex additive noise $\nu(t)$ is given by

$$\nu(t) = [n(t) + j\hat{n}(t)]\exp(j\omega t) \tag{9}$$

and $n(t)$ and its Hilbert Transform $\hat{n}(t)$ are white Gaussian noise processes possessing single-sided spectral densities of N_0 w/cps. The physical additive noise process is the "real part" of the complex Gaussian process $\nu(t)$, i.e.,

$$n_0(t) = Re\,[\nu(t)] = n(t)\cos\omega t - \hat{n}(t)\sin\omega t. \tag{10}$$

In carrying out the frequency-translation operation at the receiver, we use the real part of the received signal $\psi_{ki}(t)$. If we multiply the noise process $n_0(t)$ by the noisy stored carrier reference $r(t)$ and neglect the double frequency terms (the Wiener filter will not respond to them), we obtain

$$n'(t) = \frac{1}{2^{1/2}}\,[n(t)\cos\Phi + \hat{n}(t)\sin\Phi]. \tag{11}$$

We have assumed that the stored reference $r(t) = 2^{1/2}\cos(\omega t + \Phi)$ and Φ is a slowly-varying random phase variation. For example, Φ may well represent the phase error of a phased-locked loop which is tracking the sinusoid $\sin\omega t$ in the presence of additive white Gaussian noise. Several probability distributions have been derived in Refs. 23 and 24 which govern the statistics of this phase error.

Further, using Eq. (11) and the facts that n and $\hat{n}$ have zero mean and are uncorrelated, it is easily shown that the noise process $n'(t)$ at the multiplier output is white with a single-sided spectral density of N_0 w/cps.

For reasons which will become obvious later, we compute the multiplier outputs $\eta_{ki}(t)$ for all four types of modulation. To accomplish frequency-translation in the physical sense, we use the physical waveforms received, i.e., $Re\,\{\psi_{ki}(t)\}$, where *Re* denotes "real part." For the DSB systems we have, using Eqs. (6), (7) and (11) and a little labor,[6]

$$\left.\begin{aligned} \eta_{ki}(t) &= m_a\,[(P/1 + m_a^2)^{1/2}]\, m_{ki}(t) + n'(t); \quad \text{AM-DSB} \\ \eta_{ki}(t) &= P^{1/2} m_{ki}(t) + n'(t); \quad \text{AM-DSB/SC} \end{aligned}\right\} (12)$$

(We have neglected the double-frequency and dc terms.) In the SSB cases we have for the multiplier[7] outputs

[6] At this point we have assumed perfect coherence at the receiver. The noisy phase reference case will be considered later.

[7] We have assumed that $r(t) = Re\,[2^{1/2}(1 + j1)\exp(j\omega t)]$ in the SSB systems. This serves to illustrate how either $\hat{m}$ or m may be recovered at the receiver. If $\hat{m}$ is desired one must form $\eta + \hat{\eta}$ instead of the difference.

(neglecting the double-frequency and dc terms)

$$\left.\begin{aligned}\eta_{ki}(t) &= m_a(P/1+2m_a^2)^{1/2}\,[m_{ki}(t)+\hat{m}_{ki}(t)] + \frac{n''(t)}{2^{1/2}}; \qquad \text{AM-SSB}\\ \eta_{ki}(t) &= (P/2)^{1/2}\,[m_{ki}(t)+\hat{m}_{ki}(t)] + \frac{n''(t)}{2^{1/2}}; \qquad \text{AM-SSB/SC}\end{aligned}\right\} \tag{13}$$

where $n'' = n + \hat{n}$. The input to the Wiener filter for the SSB systems is $x_{ki}(t) = \hat{\eta}_{ki}(t) - \eta_{ki}(t)$ or

$$\left.\begin{aligned}x_{ki}(t) &= 2^{1/2}\,[m_a(2P/1+2m_a^2)^{1/2}\,m_{ki}(t)+n(t)]; \qquad \text{AM-SSB}\\ x_{ki}(t) &= 2^{1/2}\,[P^{1/2}m_{ki}(t)+n(t)]; \qquad \text{AM-SSB/SC}\end{aligned}\right\} \tag{14}$$

where $n(t)$ is white Gaussian noise of single-sided spectral density N_0 w/cps. Eqs. (13) and (14) represent, respectively, the inputs (signal plus noise) to the Wiener filters of Figs. 34 and 35. Note that, for the AM-DSB/SC and AM-SSB/SC systems, the signals to be filtered are essentially the same since the square root of two in Eq. (14) may be neglected because it effects both the signal and noise.

The spectral density of the input process for the four types of modulation may be written from Eqs. (1), (3), (13) and (14). Neglecting this square root of two in Eq. (14), they are

$$S_{ji}(\omega;k) = P_j S_i(\omega;k) \tag{15}$$

for $i = 1, 2$ and all k. The P_j factors are defined as

$$\left.\begin{aligned}P_1 &= m_a^2(1+m_a^2)^{-1}P = g_1P; \qquad P_2 = g_2P = P\\ P_3 &= 2m_a^2(1+2m_a^2)^{-1}P = g_3P; \quad P_4 = g_4P = P\end{aligned}\right\} \tag{16}$$

which represents the average input signal power at the receiver, i.e.,

$$P_j = \frac{1}{2\pi}\int_{\infty}^{\infty} S_{ji}(\omega;k)\,d\omega; \qquad j = 1,2,3,4.$$

5. The Wiener Error for the Two Classes of Signaling Spectra

The instantaneous value of the Wiener error may be written assuming the k^{th} member of the i^{th} stochastic class is being transmitted using the j^{th} modulation technique as (Figs. 34 and 35)

$$\varepsilon_{ki}^{j}(t) = y_{ki} - (P_j)^{1/2}m_{ki}(t); \qquad j = 1,2,3,4,$$

where the P_j's are defined in Eq. (16). Since the modulating signal and the noise vary randomly with time, it is natural to characterize the "output noise" by its mean-square intensity

$$\overline{(\varepsilon_{ki}^{j})^2} = \overline{[y_{ki}(t) - (P_j)^{1/2}m_{ki}(t)]^2} = \sigma_{ji}^2(k).$$

The Wiener filter (Types I and II) which minimizes the mean-square-error for all members of the two classes of stochastic processes is the filter which we use at the receiver for smoothing the observed data $\eta_{ki}(t)$ and $x_{ki}(t)$. The filter functions (impulse responses) are formally determined from the spectral densities of the signal and noise; however, we are not interested here in the frequency responses of the individual filters. Instead we shall be concerned primarily with determining the filtering action, i.e., computation of the Wiener error $\sigma_{ji}^2(k)$.

For Type II filters (non-realizable) it can be shown that the mean-square-error occurring when one transmits the k^{th} member of the i^{th} signal class using the j^{th} modulation technique is given by Ref. 23 as

$$\sigma_{ji}^2(k) = \int_{\infty}^{\infty}\frac{1}{2\pi}\times S_{ji}(\omega;k)\,[1+2S_{ji}(\omega;k)/N_0]^{-1}\,d\omega. \tag{17}$$

We have assumed the input noise is white and $S_{ji}(\omega;k)$ are the spectral densities given by Eq. (15).

On the other hand, for Type I filters (realizable), the mean-square-error encountered (when the k^{th} member of the i^{th} signal class is transmitted and the j^{th} modulation technique employed at the transmitter) is for white noise.

$$\sigma_{ji}^2(k) = \frac{N_0}{4\pi}\int_{-\infty}^{\infty} \ln[1+2S_{ji}(\omega;k)/N_0]\,d\omega. \tag{18}$$

This is the Wiener error obtained by Yovits and Jackson (Ref. 25). Eqs. (17) and (18) are remarkable in that the Wiener error may be evaluated without having to compute the individual filter functions.

6. Performance of Type I and II Filters Using "Maximally Flat" Spectra

The Wiener error for all four types of modulation and all members of both signaling classes may be computed from Eqs. (15) and (17). Letting $i = 1$ and substituting Eq. (1) into Eq. (17), it may be shown that

$$\sigma_{j1}^2(k) = P_j\,[1 + 2P_jK_1(k)/N_0]^{(1/2k)-1}, \tag{19}$$

where $K_1(k)$ is given by Eq. (2). Defining the signal-to-noise ratio ρ as the ratio of the mean-squared-value of the signal power P_j [Eq. (16)] to the Wiener error, we have

$$\rho_{j1}(k;\mathrm{I}) = [1 + 2\pi g_jR\,\mathrm{sinc}\,(\pi/2k)]^{1-(1/2k)}, \tag{20}$$

where $R = P/aN_0$ and the g_j's are given by Eq. (16). The notation $\rho_{ji}(k;\mathrm{I})$ signifies the signal-to-noise ratio when the k^{th} member of the i^{th} stochastic process is being transmitted using the j^{th} modulation technique and Type I filtering at the receiver. For the "Maximally Flat" case, we have for $k = \infty$

$$\rho_{j1}(\infty;\mathrm{I}) = 1 + 2\pi g_jR; \tag{21}$$

while for large values of the parameter R, Eq. (20) becomes

$$\rho_{j1}(k;\mathrm{I}) \sim [2\pi g_jR\,\mathrm{sinc}\,(\pi/2k)]^{1-(1/2k)}. \tag{22}$$

The Wiener error for Type II filters operating on signal Class 1 may be shown to be

$$\sigma_{j1}^2(k) = \frac{kN_0}{K_1(k)}\,\{[1 + 2P_jK_1(k)/N_0]^{1-(1/2k)} - 1\}.$$

The signal-to-noise ratio ρ becomes

$$\rho_{j1}(k;\mathrm{II}) = \frac{\pi g_jR\,\mathrm{sinc}\,(\pi/2k)}{k\,\{[1 + 2\pi g_jR\,\mathrm{sinc}\,(\pi/2k)]^{1/2k} - 1\}}, \tag{23}$$

which for large R is asymptotic to

$$\rho_{j1}(k;\mathrm{II}) \sim \frac{\pi g_jR\,\mathrm{sinc}\,(\pi/2k)}{k\,[(2\pi g_jR\,\mathrm{sinc}\,(\pi/2k)^{1/2k} - 1]}. \tag{24}$$

Of special interest is the case where $k = \infty$. It may be shown that

$$\rho_{j1}(\infty;\mathrm{II}) = 2\pi g_jR\,[\ln(1 + 2\pi g_jR)]^{-1}. \tag{25}$$

For large R, this is asymptotic to

$$\rho_{j1}(\infty;\mathrm{II}) \sim 2\pi g_jR\,[\ln(1 + 2\pi g_jR)]^{-1}. \tag{26}$$

Comparison of Eq. (27) with Eq. (24) shows that, for large R and small k, Type I filters have a signal-to-noise ratio of approximately $2k$ times the signal-to-noise ratio of Type II filters. As k approaches infinity, Eqs. (22) and (26) show that the performance of Type II filtering becomes inferior to Type I filtering by a factor of $\ln(2\pi g_jR)$. For $k = 1$, and large R, Type I filters outperform Type II filters by a factor of approximately 3 db.

7. Performance of Type I and II Filters Using "Asymptotically Gaussian" Spectra

Eq. (3) may be substituted into Eq. (15) yielding the spectra for the four types of modulation. This result when used in Eqs. (17) and (18) gives the required Wiener error. Due to the lengthy details and integration procedure required for general k, we evaluate the signal-to-noise ratio ρ for the special case $k = \infty$.

For Type II filtering the Wiener error for the j^{th} modulation technique is given by

$$\sigma_{j2}^2(\infty) = \frac{aN_0}{4\pi}\int_{-\infty}^{\infty} \ln[1 + 4(\pi)^{1/2}g_jR\exp(-x^2)]\,dx$$

and the signal-to-noise ratio ρ becomes

$$\rho_{j2}(\infty;\mathrm{II}) = 2\pi g_jR\left\{\int_0^{\infty}\ln[1 + 4(\pi)^{1/2}g_jR\exp(-x^2)\,dx]\right\}^{-1}. \tag{27}$$

For Type I filtering the signal-to-noise ratio for the j^{th} type of modulation is easily shown to be

$$\rho_{j2}(\infty;\mathrm{I}) = (\pi^{1/2}/2)\int_0^{\infty}[\exp(-x)^2 + 4g_j(\pi)^{1/2}R]^{-1}\,dx. \tag{28}$$

Eqs. (27) and (28) can be integrated by expanding the integrand into an infinite series and integrating term by term. Difficulty arises, however, when $4g_j(\pi)^{1/2}R > 1$. A more tractable procedure to use is to integrate Eqs. (27) and (28) numerically on a general purpose computer.

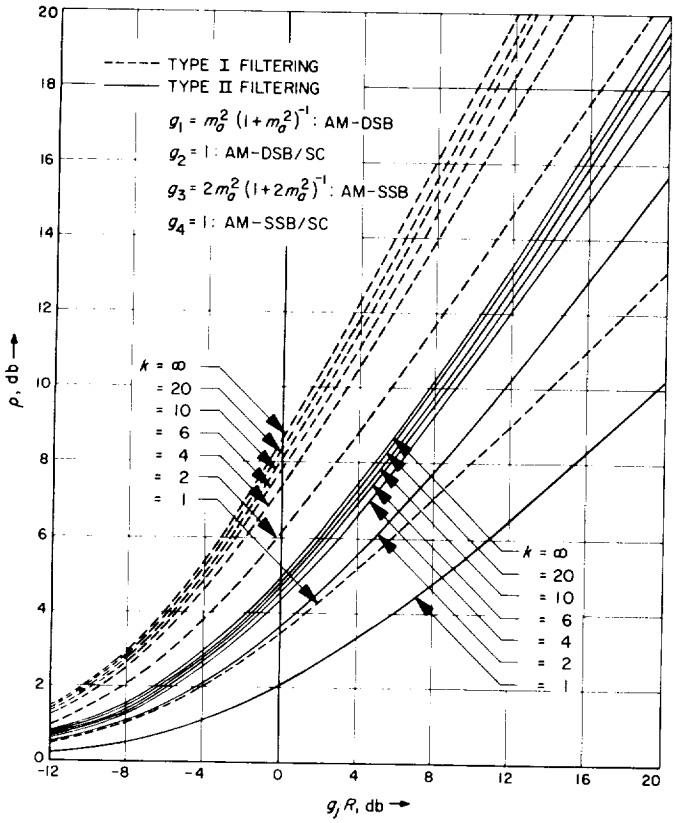


Fig. 36. System performance characteristics

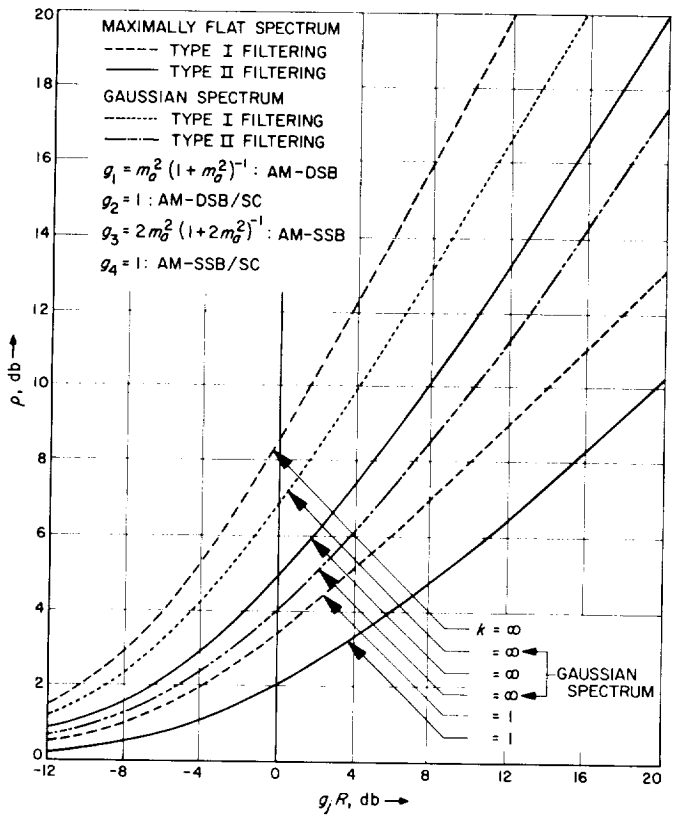


Fig. 37. Comparison of system performance characteristics

8. Calculated Performance and Comparison

Plotted in Figs. 36 and 37 is the signal-to-noise ratio ρ versus the basic parameter $g_j R$ where $R = P/aN_0$. In particular, Figs. 36 and 37 have been plotted for $g_2 = g_4 = 1$, i.e., AM-DSB/SC and AM-SSB/SC systems. Performance for the other two types of modulation can be obtained from these figures by rescaling the abscissa by g_j, e.g., if the performance of an AM-DSB system is required $j = 1$ and $g_1 = m_a^2(1 + m_a^2)^{-1}$; see Eq. (16).

The curves show that, regardless of the type of amplitude modulation employed at the transmitter, the larger k (for either class of stochastic signals) the better is the signal-to-noise ratio ρ. This is easily explained on a physical basis. For large k, energy in the signaling spectra is suppressed in the high frequency regions and accentuated in the low frequency regions. Hence, the Wiener filter, for a white noise input, accepts a smaller amount of the input noise, and the signal-to-noise ratio ρ is larger.

Note that for large k and R, Type I filters (nonrealizable) yield a value of ρ highly superior to Type II filters (realizable). In physical situations where delay in the demodulation procedure is tolerable, it is quite evident that Type I filtering should be employed with either the AM-SSB/SC or AM-DSB/SC system. If bandwidth is a premium, then the AM-SSB/SC system should be selected over AM-DSB/SC system.

All systems have the disadvantage of requiring a local copy of the carrier at the receiver. For AM-SSB and AM-DSB a carrier component is available in the observed data. Such is not the case for AM-SSB/SC and AM-DSB/SC; consequently, other means must be employed for obtaining this information at the receiver. This is obviously a disadvantage of either of these systems over the AM-DSB and AM-SSB systems. In terms of transmission bandwidth, AM-DSB and AM-DSB/SC require equal amounts while AM-SSB and AM-SSB/SC require only half as much as the AM-DSB or AM-DSB/SC system.

If we view the parameter $g_j R = g_j P/aN_0$ as a measure of the effectiveness of j^{th} modulation technique, we find that AM-DSB/SC and AM-SSB/SC perform equally well. On this basis, the AM-DSB/SC and AM-SSB/SC systems are 10 $\log_{10}$ $[m_a^{-2}(1 + m_a^2]$ db better than AM-DSB and

$10 \log_{10} [(1 + 2m_a^2)(2m_a^2)^{-1}]$ db better than AM-SSB. In terms of ρ, for a given k and R, no general conclusions may be reached; the curves in Figs. 36 and 37 must be consulted.

9. Performance Using a Noisy Phase Reference

One major difficulty with implementing any of the amplitude-demodulation systems studied here is that of providing the receiver with a copy of the transmitted carrier, i.e., synchronization of the transmitter and receiver local oscillators. One practical means of achieving carrier synchronization in the past has been to employ a phase-locked loop at the receiver. Even if one is willing to build a phase-locked loop at the receiver, there remains the question of to what component in the received signal one should try to achieve carrier lock; e.g., in the AM-DSB/SC or AM-SSD/SC systems, the received spectrum does not contain a frequency component oscillating at the carrier frequency. The best one can do (probably) is to transmit a pilot carrier for use in connection with the phase-locked loop. This, however, requires additional energy.

On the other hand, for AM-DSB and AM-SSB there exists a carrier component in the received signal spectrum. In fact, the power in this component is a function of the modulation index m_a [Eqs. (6) and (7)]. With a knowledge of this it is not at all clear how one could most effectively mechanize a phase-locked loop for synchronization purposes. Avoiding this question we assume that a phase error and the voltage controlled oscillator (VCO) in the phase-locked loop is oscillating at the carrier frequency.

Viterbi (Ref. 26) and Tikhonov (Ref. 24) have shown that the probability distribution $p(\Phi)$ for the phase error Φ is given by

$$p(\Phi) = [2\pi I_0(\alpha)]^{-1} \exp(\alpha \cos \Phi); \qquad -\pi \leq \Phi \leq \pi \tag{29}$$

where $I_0(\alpha)$ is the imaginary Bessel function evaluated at the signal-to-noise ratio existing in the tracking loop. Taking into consideration a phase error of Φ rad/sec the factor $\cos \Phi$ multiplies the signal components in Eqs. (12) and (14). As already shown the noise statistics remain unchanged. Hence, the spectral densities of Eq. (15) are multiplied by $\cos^2 \Phi$ as well as the g_j's of Eq. (16), and the signal-to-noise ratios computed for the ideal reference signal become that value of signal-to-noise ratio conditional on the fact that the phase error is Φ rad/sec, i.e., $\rho_{ji}(k; \cdot) = \rho_{ji}(k; \cdot | \Phi)$. The signal-to-noise ratio which results when all members of the phase error ensemble are taken into consideration becomes

$$\rho_{ji}(k; \text{I or II}) = \int_{-\pi}^{\pi} p(\Phi)\, \rho_{ji}(k; \text{I or II} | \Phi)\, d\Phi. \tag{30}$$

If one attempts to solve this equation using Eq. (29) for general k, j, and i, a formidable integral is immediately encountered. Special cases, e.g., $k = \infty$, can be worked out exactly. For general j, k, and i, numerical integration techniques could be applied to obtain values for $\rho_{ji}(k;$ I or II), but it appears, at this point, to be hardly worth the effort.

An alternate procedure which gives some idea as to the effect of a noisy phase reference is to average over the phase error before filtering; i.e., define the input signal component of $\eta_{ki}(t)$ (or $x_{ki}(t)$) by the following relationship

$$\eta_{ki}(t) = \int_{-\pi}^{\pi} p(\Phi)\, \eta_{ki}(t | \Phi)\, d\Phi. \tag{31}$$

Carrying out this integral using Eqs. (12) and (29) yields for the j^{th} type of modulation

$$\eta_{ki}(t) = (g_j')^{1/2} P m_{ki}(t) + n'(t), \tag{32}$$

where

$$(g_j')^{1/2} = (g_j)^{1/2} I_1(\alpha)\, [I_0(\alpha)]^{-1}$$

and $I_1(\alpha)$ is the first-order Bessel function of imaginary argument. Thus, for all types of modulation the curves of Figs. 36 and 37 still apply; however, the abscissa is now $g_j' R$ instead of $g_j R$. If $\alpha = \infty$, corresponding to perfect coherence, $g_j' = g_j$. If $\alpha = 0$, corresponding to a carrier whose phase variable is uniformly distributed over an interval of length 2π, $g_j' = 0$ for all j. For $0 < a < \infty$, we find that $g_j' < g_j$, e.g., if $\alpha = 1$ (corresponding to a signal-to-noise ratio in the tracking loop of 0 db), we find $g_j' = 0.20 g_j$. Thus, we see that a good (non-noisy) replica of the carrier is required at the receiver in order that the demodulations procedure be performed efficiently.

10. Conclusions

In this paper, four types of amplitude modulation–demodulation systems have been analyzed. The information-bearing signal used to modulate the transmitter is generated from one of two classes of stochastic processes: the "Maximally Flat" and the "Asymptotically

Gaussian" processes. We have shown that system performance depends on the type of Wiener filter (realizable or non-realizable) used to smooth the noisy data and the modulating spectrum. In particular, for $k = 1$ and large R, the nonrealizable filter performs approximately 3 db better than the realizable filter. For large k and R the nonrealizable filter performs approximately

$$10 \log_{10} [\ln 2\pi g_j R] \text{ db}$$

better than the realizable filter.

It is shown that system performance is highly dependent on the parameter k of the modulation spectrum. In fact it is advisable to shape the modulating spectrum before transmission by means of a Butterworth filter or a series of isolated-cascaded RC networks. Shaping of the modulation spectrum by a Butterworth filter proves to be more effective than that of using a series of isolated-cascaded RC networks.

The situation was considered in which the receiver utilizes in the demodulation procedure a noisy replica of the transmitted carrier. If the carrier-replica is derived at the receiver by means of a phase-locked loop, we found that the Wiener error is the least when the carrier-replica is relatively noise-free. For example, a signal-to-noise ratio of 10 db in the tracking loop reduces the effective input signal-to-noise ratio at the demodulator input by 0.4 db, while a signal-to-noise ratio of 0 db in the tracking loop reduces the effective input signal-to-noise ratio by 7 db.

Finally, we point out that these results are compared in JPL TR 32-637, "Optimum and Sub-Optimum Frequency Demodulation," with similar results obtained for frequency demodulation using phase-locked frequency discriminators.

H. On Suboptimum Binary Decisions

J. J. Stiffler

1. Summary

A signal $y(t) = \pm Af_i(t) + n(t)$ is received, where $f_i(t)$ is a known signal of time duration T seconds, and $n(t)$ is white Gaussian noise with the single-sided spectral density N_0. The constant A is such that the average signal power is A^2; i.e.,

$$\frac{1}{T}\int_0^T f_i^2(t)\,dt = 1.$$

The waveform $f_i(t)$ is repeated every T seconds, but the sign of $f_i(t)$ is equally likely to be positive or negative of its predecessor. It is desired to determine which of the signals $f_i(t)$, $(i = 1, 2, \cdots, M)$ is actually being received, and in a minimum number of observations. Two methods for making this decision are presented in this article as well as a comparison of the results using the two methods when two correlated waveforms $f_1(t)$ and $f_2(t)$ are to be distinguished.

2. Introduction

It is well-known (Ref. 27) that the maximum-likelihood detector for the signal $Af(t)$ involves the determination of the *a posteriori* probability $p(y(t)|Af(t))$ of the received signal $y(t)$ given $Af(t)$. Since the noise is white and Gaussian

$$p(y(t)|Af(t)) =$$

$$\frac{1}{(\pi N_0 T)^{1/2}} \exp\left\{-\frac{1}{N_0 T}\int_0^T [Af(t) - y(t)]^2\,dt\right\}. \quad (1)$$

If a number of signals $f_i(t)$ could have been received, the optimum decision is to select the largest with respect to i of the quantities.

$$p(Af_i(t)|y(t)) = \frac{p(y(t)|Af_i(t)\,p(Af_i(t))}{p(y(t))}. \quad (2)$$

If it is then assumed that $p(Af_i(t))$ is independent of i, one sees that since $p(y(t))$ is not a function of i, the decision procedure is simply to select the largest of the *a posteriori* probability distributions $p(y(t)|Af_i(t))$.

Suppose on the other hand that either $+Af_i(t)$ or $-Af_i(t)$ could have been received for all i, and we wish to determine which of the signals $f_i(t)$ was actually received regardless of the sign. Then, evidently, it is necessary to find the largest of the probabilities

$$p(Af_i(t) \text{ or } -Af_i(t)|y(t)) =$$

$$\frac{p(y(t)|Af_i(t))\,p(Af_i(t)) + p(y(t)|-Af_i(t))\,p(-Af_i(t))}{p(y(t))}. \quad (3)$$

If $p(Af_i(t)$ and $p(-Af_i(t))$ are equal and constant independent of i, this problem is equivalent to choosing the largest of the probabilities

$$p(y(t)|Af_i(t)) + p(y(t)|-Af_i(t)) = \frac{1}{(\pi N_0 T)^{1/2}} \exp\left\{-\frac{1}{N_0 T}\int_0^T [Af_i(t) - y(t)]^2\,dt\right\} + \frac{1}{(\pi N_0 T)^{1/2}} \exp\left\{-\frac{1}{N_0 T}\int_0^T [-Af_i(t) - y(t)]^2\,dt\right\}. \quad (4)$$

But since $y(t)$ and

$$\int_0^T f_i^2(t)\,dt$$

are assumed to be independent of i, the value of i for which Eq. (4) is a maximum is just that i which maximizes the expression

$$\exp\left\{\frac{2A}{N_0 T}\int_0^T f_i(t)\,y(t)\,dt\right\} + \exp\left\{-\frac{2A}{N_0 T}\int_0^T f_i(t)\,y(t)\,dt\right\} \quad (5)$$

Letting

$$x = \int_0^T f_i(t)\,y(t)\,dt,$$

it is then desired to determine the largest of the quantities $\cosh[(2A/N_0T)x]$.

If n observations are made (that is, the signal $\pm f_i(t)$ is repeated n times with arbitrary signs), the probability that $\pm f_i(t)$ was received, when the observed signal was $y_j(t) = y(t + jT)$, $j = 0, 1, \cdots, n-1$, is just

$$\prod_j \left\{\frac{p(y_j(t)|Af_i(t)) + p(y_j(t)|-Af_i(t))}{p(y_j(t))}\right\} \quad (6)$$

it is again assumed that $p(Af_i(t)) = p(-Af_i(t)) = \text{constant}$.

Rewriting the expression (6), we have

$$\frac{1}{(\pi N_0 T)^{1/2}} \prod_j \left\{\frac{\exp\left\{-\frac{A^2}{N_0 T}\left[\int_{jT}^{(j+1)T} f_i^2(t)\,dt + \int_{jT}^{(j+1)T} y^2(t)\,dt\right]\right\} \cosh\frac{2A}{N_0 T}x_j}{p(y_j(t))}\right\}, \quad (7)$$

where

$$x_j = \int_{jT}^{(j+1)T} f_i(t)\,y(t)\,dt.$$

The maximum likelihood decision is to choose the value of i for which the quantities (7) attain a maximum as that corresponding to the transmitted signal.

In practice, this decision is rather difficult to implement, at least in part due to the difficulty in determining the values of A and the spectral density N_0. For this reason it is useful to consider several simplifications of the expression (7). First of all it is observed that only $\cosh[(2A/N_0T)x_j]$ is dependent upon the value of i, assuming that

$$\int_0^T f_i^2(t)\,dt = 1.$$

The coefficient of $\cosh[(2A/N_0T)x_j]$ is a weighting which is small when a noisy signal has been received and large for a clean signal. If we ignore this weighting function and give all observations equal weight, then the procedure is simply to select the largest of the M functions

$$\prod_j \cosh\frac{2A}{N_0 T}x_j. \quad (8)$$

Again, since the quantities A and N_0 are not generally known precisely, some further simplification is in order. In particular, for small values of A/N_0T, we can use the partial expansion

$$\cosh\frac{2A}{N_0 T}x_j \approx 1 + \left(\frac{2A}{N_0 T}\right)^2 \frac{x_j^2}{2}$$

and, neglecting higher order terms

$$\prod_j \cosh\left(\frac{2A}{N_0 T}x_j\right) \approx 1 + \frac{1}{2}\left(\frac{2A}{N_0 T}\right)^2 \sum_j x_j^2. \quad (9)$$

If this approximation is used then it is only necessary to form the expressions

$$z(i) = \sum_j x_j^2(i) \quad (10)$$

and to determine the largest of these.

At the other extreme, when A/N_0T is large

$$\cosh\frac{2A}{N_0T}x_j \approx \frac{1}{2}e^{(2A/N_0T)|x_j|} \tag{11}$$

and the expression (8) becomes, approximately,

$$\frac{1}{2}\prod_j e^{(2A/N_0T)|x_j|} = \frac{1}{2}\exp\left\{\frac{2A}{N_0T}\sum_j |x_j|\right\}, \tag{12}$$

and it is sufficient to determine the largest of the M quantities

$$w(i) = \sum_j |x_j(i)|. \tag{13}$$

In the following two sections, the results of using the two sets of observables $z(i)$ and $w(i)$ are compared when $M = 2$, that is, when one of two correlated waveforms $f_1(t)$ or $f_2(t)$ is received.

3. The Squaring Method

Suppose $f_i(t) = f_1(t)$ is actually being received, but that the receiver has not decided whether it is observing $f_1(t)$ or another waveform $f_2(t)$. The correlation ρ between $f_1(t)$ and $f_2(t)$

$$\rho = \frac{1}{T}\int_0^T f_1(t) f_2(t)\, dt \tag{14}$$

is not assumed to be zero. The squaring method discussed in the previous part of this article involves the formation of the observables

$$x(1) = x_1 = \int_0^T f_1(t)\, y(t)\, dt$$

$$= A\int_0^T f_1^2(t)\, dt + \int_0^T f_1(t)\, n(t)\, dt \tag{15}$$

and

$$x(2) \equiv x_2 = \int_0^T f_2(t)\, y(t)\, dt$$

$$= A\int_0^T f_1(t) f_2(t)\, dt + \int_0^T f_2(t)\, n(t)\, dt. \tag{16}$$

Since the noise is white and Gaussian, x_1 and x_2 are both Gaussian random variables with

$$\left.\begin{aligned} E(x_1) &= AT \equiv \mu \\ E(x_2) &= \rho AT \equiv \rho\mu \\ \sigma_1^2 = E(x_1^2) - A^2 &= N_{0/2}T \equiv \sigma^2 \\ \sigma_2^2 = E(x_2^2) - \rho^2A^2 &= N_{0/2}T \equiv \sigma^2 \\ \frac{E(x_1x_2) - \rho A^2}{\sigma_1\sigma_2} &= \rho \end{aligned}\right\} \tag{17}$$

Thus,

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2(1-\rho^2)^{1/2}} \times \exp\left\{\frac{(x_1-\mu_2) - 2\rho(x_1-\mu)(x_2-\rho\mu) + (x_2-\rho\mu_2)}{2\sigma^2(1-\rho^2)}\right\} \tag{18}$$

The decision as to which function, $f_1(t)$ or $f_2(t)$, is actually being received is to be based upon the sum of the squares of a number of independent samples of both Gaussian variables x_1 and x_2. The larger of these two sums is assumed to correspond to the transmitted function. If a reasonably large number of samples are involved, then the two random variables

$$\left.\begin{aligned} z_1 &= \sum_{i=1}^{n} x_{1i}^2 \\ \text{and} \\ z_2 &= \sum_{i=1}^{n} x_{2i}^2 \end{aligned}\right\} \tag{19}$$

(x_{1i} and x_{2i} are the i^{th} samples of the processes x_1 and x_2, respectively) are approximately jointly Gaussianly distributed, by the Central Limit Theorem.

Clearly, therefore, the variable $\zeta = z_1 - z_2$ is asymptotically Gaussian, and a correct decision is made if the sample from the process ζ is greater than zero. That is, a correct decision is made with the probability

$$P = \int_0^\infty p(\zeta)\, d\zeta \approx \frac{1}{(2\pi)^{1/2}\sigma_\zeta}\int_0^\infty \exp\left[\frac{(\zeta-\mu_\zeta)^2}{2\sigma_\zeta^2}\right] ds$$

$$= \frac{1}{2\pi^{1/2}}\int_{-\mu_1/\sigma_1}^{\infty} e^{-\xi^2/2}\, d\zeta. \tag{20}$$

This probability is a function of the ratio μ_ζ/σ_ζ only, so we will proceed to evaluate this ratio as a function of the channel signal-to-noise ratio and of the correlation between x_1 and x_2.

The following equalities are easily proved:

$$\left.\begin{aligned} E(x_1^2) &= \sigma^2 + \mu^2 \\ E(x_2^2) &= \sigma^2 + \rho^2\mu^2 \\ E(x_1^4) &= 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4 \\ E(x_2^4) &= 3\sigma^4 + 6\rho^2\mu^2\sigma^2 + \rho^4\mu^4 \\ E(x_1x_2) &= \rho(\sigma^2 + \mu^2) \\ E(x_1^2x_2^2) &= 5\rho^2\mu^2\sigma^2 + \mu^2\sigma^2 + \rho^2\mu^4 + \sigma^4 + 2\rho^2\sigma^4 \end{aligned}\right\} \quad (21)$$

Thus,

$$\mu_\zeta = E(\zeta) = \sum_{i=1}^{n} [E(x_{1i}^2) - E(x_{2i}^2)] = n(1-\rho^2)\mu^2.$$

$$\sigma_\zeta^2 = E(\zeta^2) - E^2(\zeta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \{E(x_{1i}^2x_{1j}^2) - 2E(x_{1i}^2x_{2j}^2) + E(x_{2i}^2x_{2j}^2)\} - E^2(\zeta). \quad (22)$$

Using the equalities (21) and observing that $x_{\alpha i}$ and $x_{\beta j}$ ($\alpha = 1, 2$; $\beta = 1, 2$) are independent for $i \neq j$ (since the integrals defining the two variables are over disjoint time intervals if $i \neq j$), it follows that

$$\begin{aligned} \sigma_\zeta^2 &= nE(x_1^4) + n(n-1)E^2(x_1^2) - 2nE(x_1^2x_2^2) \\ &\quad - 2n(n-1)E(x_1^2)E(x_2^2) \\ &\quad + nE(x_2^4) + n(n-1)E^2(x_2^2) - E^2(\zeta) \\ &= 4n(1-\rho^2)\sigma^2(\sigma^2 + A^2). \end{aligned} \quad (23)$$

Consequently,

$$\frac{\mu_\zeta^2}{\sigma_\zeta^2} = \frac{n(1-\rho^2)\left(\frac{A^4}{\sigma^4}\right)}{4(1 + A^2/\sigma^2)} = \frac{n(1-\rho^2)\left(\frac{A^2T}{N_{0/2}}\right)}{4(1 + A^2T/N_{0/2})}. \quad (24)$$

4. The Absolute Value Method

The argument here parallels that of the previous section. We consider the random variable

$$\eta = w_1 - w_2 = \sum_{i=1}^{n} [|x_{1i}| - |x_{2i}|]. \quad (25)$$

Since η is asymptotically Gaussian, and since, as before, a correct decision is made when $\eta > 0$, it is again of interest to determine the ratio μ_η/σ_η.

First of all, observe that if

$$p(y) = \frac{1}{(2\pi\sigma)^{1/2}} \exp -\frac{(y-\mu)^2}{2\sigma^2},$$

then

$$\begin{aligned} E(|y|) &= \frac{1}{(2\pi)^{1/2}\sigma}\int_0^\infty y\left(\exp -\frac{(y-\mu)^2}{2\sigma^2} + \exp -\frac{(y+\mu)^2}{2\sigma^2}\right)dy \\ &= \frac{\sigma}{(2\pi)^{1/2}}\int_{\mu^2/\sigma^2}^\infty e^{-z/2}\,dz + \frac{\mu}{(2\pi)^{1/2}}\int_{\mu/\sigma}^{\mu/\sigma} e^{-z^2/2}\,dz \\ &= \frac{2\sigma}{2\pi}e^{-\mu^2/2\sigma^2} + \mu \operatorname{erf}\frac{\mu}{(2\sigma)^{1/2}} \end{aligned} \quad (26)$$

Moreover, if x and y have the joint probability density function $p(x, y)$ of Eq. (18), then

$$\begin{aligned} E(|xy|) &= \int_0^\infty\int_0^\infty xy\,p(xy)\,dx\,dy - \int_{-\infty}^0\int_0^\infty xy\,p(x,y)\,dx\,dy \\ &\quad - \int_0^\infty\int_{-\infty}^0 xy\,p(x,y)\,dx\,dy \\ &\quad + \int_{-\infty}^0\int_{-\infty}^0 xy\,p(x,y)\,dx\,dy. \end{aligned} \quad (27)$$

But

$$\begin{aligned} \int_a^b yp(x,y)\,dy &= \frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{2\pi\sigma^2(1-\rho^2)^{1/2}} \times \int_a^b y\exp\left\{-\frac{y-\rho x}{2\sigma^2(1-\rho^2)}\right\}dy \\ &= (1-\rho^2)^{1/2}\frac{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{2\pi} \times \int_{(a-\rho x)/[\sigma(1-\rho^2)^{1/2}]}^{(b-\rho x)/[\sigma(1-\rho^2)^{1/2}]} \xi e^{-\xi^2/2}\,d\xi \\ &\quad + \frac{\rho x\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{2\pi\sigma} \times \int_{(a-\rho x)/[\sigma(1-\rho^2)^{1/2}]}^{(b-\rho x)/[\sigma(1-\rho^2)^{1/2}]} e^{-\xi^2/2}\,d\xi. \end{aligned} \quad (28)$$

When $a = 0$ and $b = \infty$, we have

$$\int_0^\infty yp(x,y)\,dy = (1-\rho^2)^{1/2}\,\frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\exp\left[-\frac{\rho^2x^2}{2\sigma^2(1-\rho^2)}\right]}{2\pi} + \frac{\rho x\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{2}\left\{1+\operatorname{erf}\left[\frac{\rho x}{2\sigma(1-\rho^2)^{1/2}}\right]\right\}}{(2\pi)^{1/2}\,\sigma}; \tag{29}$$

when $a = -\infty$ and $b = 0$,

$$\int_{-\infty}^0 yp(x,y)\,dy = -(1-\rho^2)^{1/2}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\exp\left[\frac{\rho^2x^2}{2\sigma^2(1-\rho^2)}\right]\Big/2\pi \tag{30}$$

$$+\frac{\rho x\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{2}\left\{1-\operatorname{erf}\left[\frac{\rho x}{2^{1/2}\sigma(1-\rho^2)^{1/2}}\right]\right\}}{(2\pi)^{1/2}\,\sigma}.$$

Thus,

$$E(|xy|) = \frac{2(1-\rho^2)^{1/2}}{2\pi^{1/2}}\int_0^\infty x\exp\left[-\frac{x^2+\mu^2}{2\sigma^2}\right]\cosh\left(\frac{\mu x}{\sigma^2}\right) \times\left\{\left(\frac{2}{\pi}\right)^{1/2}\exp\left(-\frac{\rho^2x^2}{2\sigma^2(1-\rho^2)}\right)+\frac{\rho x}{\sigma(1-\rho^2)^{1/2}} \times\operatorname{erf}\left[\frac{\rho x}{2^{1/2}\sigma(1-\rho^2)^{1/2}}\right]\right\}dx = f(\rho,\mu,\sigma). \tag{31}$$

Using these results it can be verified that

$$E(\eta) = \sum_{i=1}^n [E|x_{1i}|) - E(|x_{2i}|)] = n\left\{\left(\frac{2}{\pi}\right)^{1/2}\sigma\left(e^{-\mu^2/2\sigma^2}-e^{-\rho^2\mu^2/2\sigma^2}\right) + \mu\operatorname{erf}\left(\frac{\mu}{2\sigma}\right)-\rho\mu\operatorname{erf}\left(\frac{\rho\mu}{2^{1/2}\sigma}\right)\equiv ng(\rho,\mu,\sigma), \tag{32}$$

and

$$E(\eta^2) = \sum_{i=1}^n [E(x_{i1}^2)+E(x_{i2}^2)-2E(|x_{i1}x_{i2}|) = n[2\sigma^2+(1+\rho^2)\mu^2-2f(\rho,\mu,\sigma)]. \tag{33}$$

($f(\rho,\mu,\sigma)$ is as defined by Eq. (31)). Consequently,

$$\frac{\mu_\eta^2}{\sigma_\eta^2} = \frac{ng^2(\rho,\mu,\sigma)}{2\sigma^2+(1+\rho^2)\mu^2-2f(\rho,\mu,\sigma)-g^2(\rho,\mu,\sigma)}, \tag{34}$$

where $g(\rho,\mu,\sigma)$ is that function defined in Eq. (32).

5. Comparison of Results

Although the form of the ratio μ_η^2/σ_η^2 is rather unwieldy, it is readily computed and is plotted in Fig. 38 along with the ratio of μ_ξ^2/σ_ξ^2 as a function of $A^2T/N_{0/2}$. In addition, it is possible to evaluate the limiting values of $f(\rho,\mu,\sigma)$ and $g(\rho,\mu,\sigma)$, and hence of the ratio in question, and to compare them with the corresponding limits of the ratio μ_ξ^2/σ_ξ^2 of Eq. (24). This is the subject of this section.

First, let $\sigma >> \mu$ but assume that $\rho \approx 1$ so that $(1-\rho^2)\sigma^2 << 1$. Then, since $\cosh(\mu x/\sigma) \approx 1$ for $\sigma >> \mu$, and

$$\operatorname{erf}\left[\frac{\rho x}{2^{1/2}\,\sigma(1-\rho^2)^{1/2}}\right] \approx 1-\left(\frac{2}{\pi}\right)^{1/2}\frac{\sigma(1-\rho^2)^{1/2}}{\rho x} \times\exp\left[-\frac{\rho^2x^2}{2\sigma^2(1-\rho^2)^{1/2}}\right] \tag{35}$$

for $\sigma^2(1-\rho^2) << 1$, it follows that

$$f(\rho,\mu,\sigma) \approx \frac{2\rho}{(2\pi)^{1/2}\,\sigma}\int_0^\infty x^2e^{-x^2/2\sigma^2}\,dx = \rho\sigma^2. \tag{36}$$

Similarly, when $(1-\rho^2)\sigma^2 << 1$ and $\sigma << \mu$, the approximation (33) is still applicable. Approximating $\cosh \mu x/\sigma^2$ by $\frac{1}{2}\exp\{\mu|x|/\sigma^2\}$, we then have

$$f(\rho,\mu,\sigma) \approx \frac{\rho}{(2\pi)^{1/2}}\int_0^\infty x^2\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma}\right\}dx \approx \rho(\sigma^2+\mu^2). \tag{37}$$

When $\rho = 0$, and $\sigma >> \mu$ so that $\cosh(\mu x/\sigma^2) \approx 1$, one has

$$f(0,\mu,\sigma)\int_0^\infty x\exp\left\{-\frac{x^2}{2\sigma^2}\right\}dx = \frac{2}{\pi}\sigma^2. \tag{38}$$

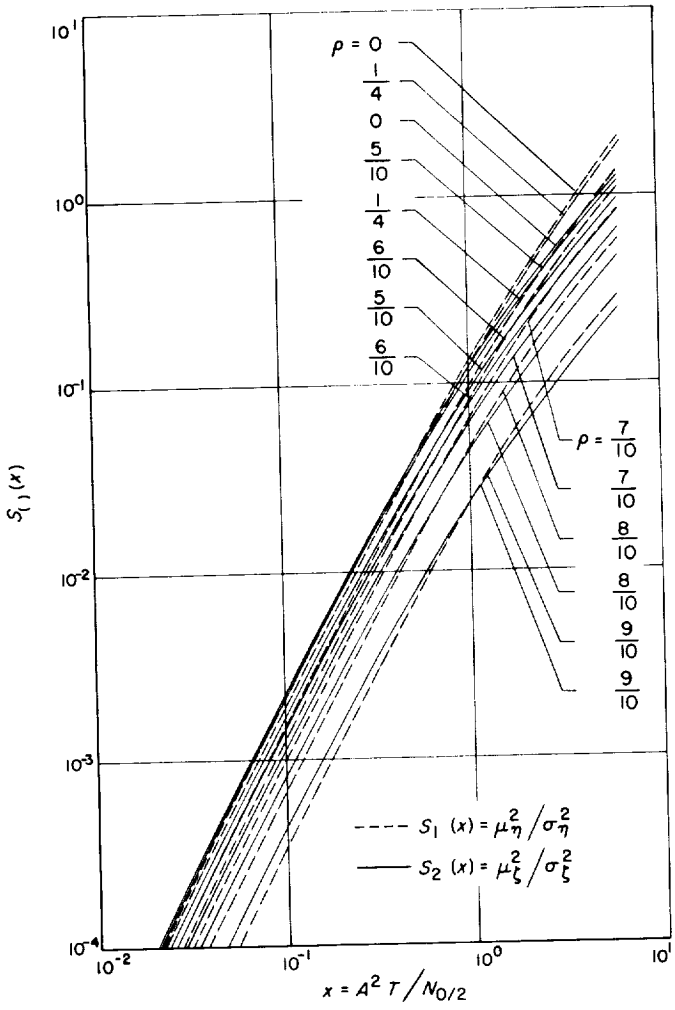


Fig. 38. Comparison of the absolute value method and the squaring method

The function $g(\rho, \mu, \sigma)$ is easily approximated for $\mu << \sigma$:

$$g(\rho, \mu, \sigma) \approx \left(\frac{2}{\pi}\right)^{1/2} \frac{(1-\rho^2)\,\sigma}{2}\,\mu^2/\sigma^2, \tag{39}$$

and for $\mu >> \sigma$

$$g(\rho, \mu, \sigma) \approx \mu(1-\rho). \tag{40}$$

Using these approximations in Eq. (32) then establishes:

$$\left.\begin{aligned} \frac{\mu_\eta^2}{\sigma_\eta^2} &\approx \frac{\eta}{4\pi}\,\frac{(1-\rho^2)}{1-\rho}\,\mu^4/\sigma^4 && \begin{cases}\mu << \sigma,\\ (1-\rho^2)\,\sigma^2 << 1\end{cases}\\ \frac{\mu_\eta^2}{\sigma_\eta^2} &\approx \eta\,\frac{(1-\rho)}{2}\,\mu^2/\sigma^2 && \begin{cases}\mu >> \sigma,\\ (1-\rho^2)\,\sigma^2 << 1\end{cases}\\ \frac{\mu_\eta^2}{\sigma_\eta^2} &\approx \left(\frac{1}{2\pi-4}\right)\mu^4/\sigma^4 && \begin{cases}\mu << \sigma,\\ \rho = 0\end{cases} \end{aligned}\right\} \tag{41}$$

Finally, using Eq. (24) it is interesting to compare the squaring method and the absolute value method at the extremes. Some of these comparisons are summarized in Table 7.

Note that at high input signal-to-noise ratios the squaring method is inferior to the absolute value method; this would be expected from the argument presented in Part 2 of this article. The two methods differ, however, by the factor $(1+\rho)/2$, which becomes arbitrarily close to 1 as ρ approaches one. At low input signal-to-noise ratios the squaring method, however, is always superior.

Table 7. Output signal-to-noise ratio in the squaring method as compared with that in the absolute value method

Conditions	Squaring method	Absolute value method
$(1-\rho^2)\,\sigma^2 << 1,\ \sigma >> \mu$	$\frac{\eta(1-\rho^2)}{4}\ (\mu^4/\sigma^4)$	$\frac{(1-\rho)}{\pi}\,\frac{\eta(1+\rho)^2}{4}(\mu^4/\sigma^4)$
$(1-\rho)\,\sigma^2 << 1,\ \sigma << \mu$	$\frac{(1+\rho)}{2}\,\eta\,\frac{(1-\rho)}{2}\ (\mu^2/\sigma^2)$	$\frac{\eta(1-\rho)}{2}\ (\mu^2/\sigma^2)$
$\rho = 0, \sigma >> \mu$	$\frac{\eta}{4}\ (\mu^4/\sigma^4)$	$\frac{\eta}{4.56}\ (\mu^4/\sigma^4)$
$(1-\rho^2)\,\sigma^2 << 1, \sigma >> \mu$	$\frac{\eta\varepsilon}{2}\ (\mu^4/\sigma^4)$	$\frac{\eta\varepsilon}{\pi}\ (\mu^4/\sigma^4)$
$\rho = 1-\varepsilon \approx 1$		

References

1. Max, J., "Quantizing for Minimum Distortion," *IRE Transactions on Information Theory*, Vol. IT-6, No. 1, pp. 7–12, March 1960.

2. Balakrishnan, A. V., "A Contribution to the Sphere-Packing Problem of Communication Theory," *Journal of Mathematical Analysis and Applications*, Vol. 3, No. 3, pp. 485–506, 1961.

3. Reiger, S., "Error Rates in Data Transmission" (correspondence), *Proceedings of the IRE*, Vol. 46, No. 5, pp. 919–920, May 1958.

4. Fano, R. M., *Transmission of Information*, Chapter 6, John Wiley–MIT Press, New York, 1961.

5. Fano, R. M., *Transmission of Information*, MIT Press, John Wiley & Sons, 1961.

6. Shannon, C. E., "Probability of Error for Optimal Codes in a Gaussian Channel," *Bell System Technical Journal*, Vol. 38, p. 611, May 1959.

7. Ziv, J., "Coding and Decoding for Time-Discrete Amplitude-Continuous Memoryless Channels," *IRE Transactions on Information Theory*, Vol. IT-8, No. 5, pp. S199–S206, September 1962.

8. Bloom, F. J., Chang, S. S. L., et al., "Improvement of Binary Transmission by Null-Zone Reception," *Proceedings of the IRE*, Vol. 45, No. 7, pp. 963–975, July 1957.

9. Jahnke, E., and Emde, F., *Tables of Functions*, p. 24, Dover Publications, Inc., New York, 1945.

10. Lehan, F. W., and Parks, R. J., "Optimum Demodulation," *IRE National Convention Record*, Part 8, pp. 101–103, 1953.

11. Youla, D. C., "The Use of Maximum Likelihood in Estimating Continuously Modulated Intelligence Which Has Been Corrupted by Noise," *IRE Transactions on Information Theory*, Vol. IT-3, pp. 90–105, March 1954.

12. Thomas, J. B., and Wong, E., "The Structure of Efficient Demodulators for Multidimensional Phase Modulated Signals," *IEEE Transactions on Communications Systems*, No. 3, pp. 261–271, September 1963.

13. Van Trees, H. L., "Analog Communication over Randomly-Time-Varying Channels," *Wescon Convention Record*, Part 13.2, August 25–28, 1964.

14. Yerman, E. A., and Posner, E. C., "Conditional Frequency Uncertainty at Post-Injection, Preliminary Report," *SPS 37-28*, Vol. IV, pp. 226–227, Jet Propulsion Laboratory, Pasadena, California, August 31, 1964.

15. Yerman, E. A., "Conditional Frequency Uncertainty in Spacecraft Acquisition," *SPS 37-24*, Vol. IV, pp. 192–195, Jet Propulsion Laboratory, Pasadena, California, December 31, 1963.

References (Cont'd)

16. Yerman, E. A., "Conditional Frequency Uncertainty in Spacecraft Acquisition," *SPS 37-27*, Vol. IV, pp. 215–219, Jet Propulsion Laboratory, Pasadena, California, June 30, 1964.

17. Yerman, E. A., "Frequency Uncertainty as a Function of Offset from Nominal," *SPS 37-27*, Vol. IV, pp. 219–222, Jet Propulsion Laboratory, Pasadena, California, June 30, 1964.

18. Pfeiffer, C. G., and Gordon, H. J., "An Approximate Method for Computing RMS Midcourse Maneuver from Injection Condition," Technical Memorandum No. 312-69, Jet Propulsion Laboratory, Pasadena, California, December 1960.

19. Mood, A. M., and Graybill, F. W., "Introduction to the Theory of Statistics," Second Edition, McGraw-Hill, New York, 1963.

20. Thomas, J. B., and Wong, E., "On the Statistical Theory of Optimum Demodulation," *IRE Transactions on Information Theory*, pp. 420–425, September 1960.

21. Van Trees, H. L., "Analog Communications over Randomly-Time-Varying Channels," Massachusetts Institute of Technology, Lexington 13, Massachusetts. Paper presented at the 1964 Wescon Show, Los Angeles, California.

22. Wainstein, L. A., and Zubakov, V. D., *Extraction of Signals from Noise*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962.

23. Deutsch, R., *Nonlinear Transformations of Random Processes*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962.

24. Tikhonov, V. I., "Phase-Lock Automatic Frequency Control Operation in the Presence of Noise," *Avtomatika i Telemekhanika*, Vol. 21, pp. 301–309, March 1961.

25. Yovits, M. C., and Jackson, J. L., "Linear Filter Optimization with Game Theory Considerations," *IRE National Convention Record*, Pt. 4, pp. 193–199, 1955.

26. Viterbi, A. J., "Phase-Locked Loop Dynamics in the Presence of Noise by Fokker-Planck Techniques," *Proceedings of the IEEE*, Vol. 51, No. 12, pp. 1737–1753, December 1963.

27. Woodward, P. M., "Probability and Information Theory, with Applications to Radar," McGraw-Hill Book Co., New York, 1955.

XXIV. Communications Systems Research: Coding Theory

A. Optimum Synchronizing Words for Fixed Word-Length Code Dictionaries

W. B. Kendall

1. Summary

If information is transmitted by sending n-symbol code words one after the other, and if the beginning of the sequence of transmitted symbols is not available at the receiver, then it is necessary for the receiver to acquire word synchronization. Here we consider the maintenance of synchronization by the periodic transmission of a special synchronizing word. So that this synchronizing word will be uniquely identifiable in an unsynchronized string of transmitted words, it is necessary that some of the possible n-symbol words not be used as code words. Restrictions on the code words are determined which, for any given synchronizing word, will ensure that when any n consecutive symbols identical to the synchronizing word are received, these n symbols are in fact the synchronizing word. Thus, any time such a sequence of symbols is received, the location of the divisions between words can be determined by simply marking off n symbols at a time starting at this synchronizing word. It is shown how the number of words which can be used for information transmission is affected by the choice of synchronizing word. The largest code dictionary results when the synchronizing word is of either the form $abb \cdots b$ or the form $aa \cdots aab$, where a and b denote any two distinct symbols. If the number of available symbols is k, then there are

$$k^n - k^{[n/2]} - \frac{k^{[(n+1)/2]} - 1}{k - 1}$$

(where [ ] denotes the interger part) words in the resulting largest dictionary.

2. Introduction

In this note we consider the synchronization problem which arises when information is transmitted by sending words, one after the other, from a code dictionary of n-symbol words. Before the individual words can be identified at the receiving point, it is necessary that the divisions between them be located. Of course, if the beginning of the sequence of transmitted symbols is available, then these divisions can be found by simply marking off k symbols at a time from the beginning. In this case it is possible to use all possible k^n words for transmission. However, it is often the case, due to such things as timing errors or loss of signal, that the beginning of the sequence of transmitted symbols is not available at the receiver. Then the receiver must acquire word timing. This is made possible by suitably restricting the code dictionary.

The simplest restriction on the code dictionary which will allow the receiver to acquire word timing is the use of a comma code, which means simply that one of the k available symbols is used exclusively as a "comma" to mark the end of each word. However, then the dictionary can contain at most $(k-1)^{n-1}$ words. A less drastic procedure which is still simple to implement, and which allows a much larger dictionary, is the use of a path-invariant comma-free code (Ref. 1). With this type of code, as with the comma codes, the divisions between words can be uniquely located by examining only n consecutive symbols. The next possibility for restricting the dictionary is the use of a comma-free code (Ref. 2), or more generally, a code with bounded synchronization delay (*SPS* 37-23, Vol. IV, pp. 146–149). Then the dictionary can be still larger,[1] and for the comma-free code, the divisions between words can be uniquely located by examining at most any $2n$ consecutive symbols.

Often times, word synchronization can be maintained a long time once it is established, and when this is true, the use of any of the above types of codes is wasteful, since they then provide too much synchronization information. In these cases a much larger dictionary can be obtained by simply using one word exclusively as a synchronizing word to be transmitted periodically. In what follows we examine the restrictions which this technique places on the code dictionary.

3. Restrictions on the Dictionary

So that the synchronizing word will be uniquely identifiable in an unsynchronized string of symbols, we will require first of all that if $a_1a_2 \cdots a_n$ and $b_1b_2 \cdots b_n$ are two words in the code dictionary (where a_i and b_i can be any of the k available symbols) then none of the overlap words:

$$\left\{\begin{array}{llllllll} a_2 & a_3 & a_4 & \cdots & a_{n-1} & a_n & b_1 & \text{(Type 1 overlap)} \\ a_3 & a_4 & a_5 & \cdots & a_n & b_1 & b_2 & \text{(Type 2 overlap)} \\ a_4 & a_5 & a_6 & \cdots & b_1 & b_2 & b_3 & \text{(Type 3 overlap)} \\ & & & & \vdots & & & \\ a_{n-1} & a_n & b_1 & \cdots & b_{n-4} & b_{n-3} & b_{n-2} & \text{(Type } n-2 \text{ overlap)} \\ a_n & b_1 & b_2 & \cdots & b_{n-3} & b_{n-2} & b_{n-1} & \text{(Type } n-1 \text{ overlap)} \end{array}\right.$$

[1]All codes which use a comma are special cases of path-invariant comma-free codes, and all path-invariant comma-free codes are special cases of comma-free codes. For a comparison of the dictionary sizes possible with these three types of codes, see Ref. 1.

is identical to the synchronizing word $s_1s_2 \cdots s_n$. Also, we will require that none of the overlap words:

$$\left\{\begin{array}{llllllll} a_2 & a_3 & a_4 & \cdots & a_{n-1} & a_n & s_1 & \text{(Type } n-1\text{,S overlap)} \\ a_3 & a_4 & a_5 & \cdots & a_n & s_1 & s_2 & \text{(Type } n-2\text{,S overlap)} \\ & & & & \vdots & & & \\ a_n & s_1 & s_2 & \cdots & s_{n-3} & s_{n-2} & s_{n-1} & \text{(Type 1,S overlap)} \\ s_2 & s_3 & s_4 & \cdots & s_{n-1} & s_n & a_1 & \text{(Type S,1 overlap)} \\ s_3 & s_4 & s_5 & \cdots & s_n & a_1 & a_2 & \text{(Type S,2 overlap)} \\ & & & & \vdots & & & \\ s_{n-1} & s_n & a_1 & \cdots & a_{n-4} & a_{n-3} & a_{n-2} & \text{(Type S,}n-2\text{ overlap)} \\ s_n & a_1 & a_2 & \cdots & a_{n-3} & a_{n-2} & a_{n-1} & \text{(Type S,}n-1\text{ overlap)} \end{array}\right.$$

resulting from the juxtaposition of a code word and the synchronizing word be identical to the synchronizing word. We will not specifically require, however, that overlap words formed by the synchronizing word and itself not be identical to the synchronizing word, since two consecutive synchronizing words will never be transmitted. Finally, we will require, of course, that the synchronizing word itself not be in the dictionary.

These requirements are necessary and sufficient to ensure that any n consecutive received symbols which are identical to the synchronizing word are in fact the synchronizing word; so once the receiver finds such a sequence, the divisions between all words can be located by simply marking off n symbols at a time starting at this n-symbol sequence. Note that these requirements are stronger than necessary to ensure that the synchronizing word is uniquely locatable, and that they give rise to "doubly instantaneous" synchronization. By this we mean that if each n-tuple of symbols is examined, starting at *either* end of a sequence, then synchronization will be accomplished the first time any n-tuple corresponding to the synchronizing word is found.

4. Dictionary Construction

As soon as the synchronizing word is chosen, one can determine which of the possible k^n words must not be included in the code dictionary. Basically, the words which must not be included are those which end with the beginning of the synchronizing word, or which begin with the end of the synchronizing word (but not all such words must necessarily be eliminated). These words, along with type designations and the number of words of

Table 1. Word types and their forms

Word type	Number of words in type	Word form	Word form	Number of words in type	Word type
$E,1$	k	$x\ s_1\ s_2\ s_3\ s_4 \cdots s_{n-3}\ s_{n-2}\ s_{n-1}$	$s_n\ x\ x\ x\ x \cdots x\ x\ x$	k^{n-1}	$B,1$
$E,2$	k^2	$x\ x\ s_1\ s_2\ s_3 \cdots s_{n-4}\ s_{n-3}\ s_{n-2}$	$s_{n-1}\ s_n\ x\ x\ x \cdots x\ x\ x$	k^{n-2}	$B,2$
$E,3$	k^3	$x\ x\ x\ s_1\ s_2 \cdots s_{n-5}\ s_{n-4}\ s_{n-3}$	$s_{n-2}\ s_{n-1}\ s_n\ x\ x \cdots x\ x\ x$	k^{n-3}	$B,3$
$E,4$	k^4	$x\ x\ x\ x\ s_1 \cdots s_{n-6}\ s_{n-5}\ s_{n-4}$	$s_{n-3}\ s_{n-2}\ s_{n-1}\ s_n\ x \cdots x\ x\ x$	k^{n-4}	$B,4$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$E,n-3$	k^{n-3}	$x\ x\ x\ x\ x \cdots s_1\ s_2\ s_3$	$s_4\ s_5\ s_6\ s_7\ s_8 \cdots x\ x\ x$	k^3	$B,n-3$
$E,n-2$	k^{n-2}	$x\ x\ x\ x\ x \cdots x\ s_1\ s_2$	$s_3\ s_4\ s_5\ s_6\ s_7 \cdots s_n\ x\ x$	k^2	$B,n-2$
$E,n-1$	k^{n-1}	$x\ x\ x\ x\ x \cdots x\ x\ s_1$	$s_2\ s_3\ s_4\ s_5\ s_6 \cdots s_{n-1}\ s_n\ x$	k	$B,n-1$

each type, are shown in Table 1. In that table an x denotes a position in a word which can be any of the k symbols. Thus, $s_n x\,x \cdots x\,x$, for example, refers to all words which begin with the last symbol of the synchronizing word.

It is clear from Table 1 that, in order to eliminate all type i overlaps, all the words of *either* type E,i *or* B,i must be excluded from the code dictionary. In order that the dictionary be as large as possible, one would ordinarily eliminate the type E,i words for $i < n/2$, and the type B,i words for $i > n/2$, and either E,i or B,i when $i = n/2$. Next, overlaps of the types i,S and S,i must be considered. Unless the symbols of the synchronizing word repeat in a particular pattern, none of these overlaps can be identical to the synchronizing word. Examination of the overlaps shows that type i,S or S,i overlaps can be identical to the synchronizing word only if the synchronizing word is periodic[2] with period i, i.e., only if $s_1 = s_{1+i} = s_{1+2i} = \cdots$ and $s_2 = s_{2+i} = \cdots$ and $\cdots$ and $s_i = s_{2i} = s_{3i} = \cdots$. If the synchronizing word is periodic with period i, then all words of types E,i *and* $B,n-i$ must be (if they have not already been) excluded from the dictionary.

5. Optimum Synchronizing Words

A word which is used as the synchronizing word will here be considered optimum if no other choice of synchronizing word can give rise to a larger code dictionary. To see how such a word is found, we first note that by choosing a non-periodic synchronizing word we can be certain that of the k^n possible words in the dictionary, we need eliminate no more than[3]

$$\sum_{i=1}^{[(n-1)/2]} k^i = \frac{k^{[(n+1)/2]} - k}{k-1}$$

words of type E,i for $i < n/2$, plus this same number of words of type B,i for $i > n/2$, plus, if n is even, $k^{n/2}$ words of either type $E,n/2$, or $B,n/2$, plus one word for the synchronizing word, for a total of

$$\frac{k^{[(n+2)/2]} + k^{[(n+1)/2]} - k - 1}{k-1}$$

words. The only way the number of eliminated words can be reduced below this is by choosing the synchronizing word so that some eliminated words are of more than one type. To this end we note that if the first $n-1$ symbols of the synchronizing word are identical, i.e., if $s_1 = s_2 = \cdots = s_{n-1}$, then all words of types $E,1$; $E,2$; $\cdots$; $E,i-1$ are also of type E,i. Then the total number of words of *all* of these types is only k^i. Similarly, if the last $n-1$ symbols of the synchronizing word are identical, then all the words of types $B,n-1$; $B,n-2$; $\cdots$; $B,i+1$ are also of type B,i. Thus, if all but the last or the first symbol of the synchronizing word are identical,

[2]Here we do not require that the word consist of a whole number of periods before it is called periodic; nor do we require that the period be minimal. For example, *aaaa* is periodic with periods 1, 2, and 3, and *ababa* is periodic with periods 2 and 4, and *abcdea* is periodic with period 5.

[3]Here $[x]$ denotes the integral part of x.

then only $k^{[n/2]}$ words of either type E,i or B,i, respectively, must be eliminated.

Note, however, that this property cannot be used to reduce the number of words of the remaining types which must be eliminated, for if both the first $n-1$ and also the last $n-1$ symbols of the synchronizing word are identical, then all symbols of the synchronizing word are identical. It is then periodic with all periods, which means that the words of all types shown in Table 1 must be eliminated. Furthermore, all the words of the remaining types which must be eliminated are of one type and one type only, so the minimum total number of words which must be eliminated when all but the first or the last symbol of the synchronizing word are identical is

$$k^{[n/2]} + \frac{k^{[(n+1)/2]} - 1}{k - 1}$$

Also, since

$$k^N > \sum_{i=0}^{N-1} k^i$$

for all $k \geq 2$, no other choice of synchronizing word could lead to fewer eliminated words.

In conclusion, we note that optimum synchronizing words are of the form $abb \cdots b$ or $aa \cdots aab$, where a and b denote any two distinct symbols. In the first case, all words $bb \cdots bxx \cdots x$ which begin with $n - [n/2] = [(n+1)/2]$ repetitions of b, and all words $xx \cdots xab \cdots b$ which end with a followed by $[n/2]$ or more repetitions of b, must be eliminated from the code dictionary. In the second case all words $xx \cdots xaa \cdots a$ end with $n - [n/2] = [(n+1)/2]$ repetitions of a, and all words $aa \cdots abx \cdots x$ which begin with $[n/2]$ or more repetitions of a followed by b, must be eliminated from the code dictionary. This is sufficient to ensure that in an unsynchronized string of words from the code dictionary, any n consecutive symbols which are identical to the synchronizing word are in fact the synchronizing word. The number of words in the resulting code dictionary is

$$k^n - k^{[n/2]} - \frac{k^{[(n+1)/2]} - 1}{k - 1}.$$

This is the result of this note.

B. Fast Decoding for a Class of Bose–Chaudhuri Codes

G. Solomon

1. Summary

This article presents a new decoding procedure for a wide class of Bose–Chaudhuri codes, which allows error-correction up to the full Bose–Chaudhuri bound. The procedure is simple enough that it can decode teletype in real-time, when the message length is more than 62 teletype words, or 310 binary bits. It then corrects up to ten errors, and will result in an output binary symbol error-probability of less than 10^{-8}, when the input binary symbol error probability is as high as 0.02 (*SPS 37-26*, Vol. IV, pp. 223–225). The procedure uses nothing more complex than evaluating 11×11 determinants over finite fields, and the decoding can be programmed for real-time operation on a fast general-purpose computer.

2. Introduction

In a recent article (*SPS 37-27*, Vol. IV, pp. 190–193), the author presented a code, together with a decoding procedure, for use in DSN interstation teletype channels. This code was a code with 62 symbols, 17 of them being information symbols. A "symbol" meant a teletype character, which is a 5-binary bit word. The symbols were then interpreted as being elements of the 32-element field $GF(2^5)$.

The (62,17) code of *SPS 37-27*, Vol. IV, pp. 190–193, was obtained from a (63,18) Bose–Chaudhuri (Ref. 3) code over $GF(2^5)$ by deleting one information symbol. The class of Bose–Chaudhuri codes to be considered here is of the form $(2^k - 1, m)$; thus, $k = 6$ for the (63,18) code. The discussion will consider a $(2^k - 1, m)$ *binary* code; the results are similar for a $(2^k - 1, m)$ code over bigger fields of characteristic 2, at least if certain number-theoretic conditions are satisfied. Or one can use a technique in *SPS 37-27*, Vol. IV, to decode the (63,18) code over $GF(2^5)$ by disassembling the code into five binary codes, decoding as binary codes, and reassembling again to get the 2^5-ary code word. Both of these procedures are equivalent in the sense of error-correcting ability.

3. Basic Formulas

The procedure to be developed shall be presented in reference to the (63,18) code. The generalization to any

$(2^k - 1, m)$ code is straightforward from the example. The (63,18) Bose–Chaudhuri code is a 10-error-correcting code generated by the recursion polynomial.

$$\begin{aligned} f(x) &= x^{18} + x^{16} + x^{11} + x^9 + x^8 + x^7 + x^5 + x^3 + x^2 + 1 \\ &= (x^6 + x + 1)(x^6 + x^5 + x^2 + x + 1) \\ &\quad \times (x^3 + 1)(x^3 + x^2 + 1). \end{aligned} \tag{1}$$

For our decoding computations, choose β a root of $x^6 + x + 1$. Then to the code word $a = (a_i)$ is associated a polynomial $g_a(x)$ as in Ref. 3:

$$\begin{aligned} g_a(x) &= c_0 + c_1 x + c_1^2 x^2 + \cdots + c_1^{32} x^{32} \\ &\quad + c_5 x^5 + c_5^2 x^{10} + \cdots + c_5^{32} x^{34} \\ &\quad + c_9 x^9 + c_9^2 x^{18} + c_9^4 x^{36} + c_{21} x^{21} + c_{21}^2 x^{42}, \end{aligned} \tag{2}$$

where $g_a(\beta^e) = a_e$. The Reed Formula (Ref. 4) gives the values of c's and d's for each code word $a = (a_i)$, with the computations performed in $GF(2^6)$. For a code word to be received correctly, the d's must all be $0\,(d_1 = d_3 = d_5 = d_{17} = d_{19} = 0)$:

$$c_j = \sum_{i=0}^{62} a_i \beta^{-ij}, \qquad d_j = \sum_{i=0}^{62} a_i \beta^{ji}.$$

Now let t errors be made in transmission at the positions $\beta_1, \beta_2, \cdots, \beta_t$. If to the transmitted word $a = (a_i)$ is associated the polynomial $g_a(x)$ with the coefficients c_0, c_1, c_5, c_9, c_{21}, then the computed coefficients of the received word are given as

$$\begin{cases} c_0' = c_0 + t \\ c_1' = c_1 + \sum_{i=1}^{t} \beta_i^{-1} \\ c_5' = c_5 + \sum_{i=1}^{t} \beta_i^{-5} \end{cases} \qquad \begin{cases} c_9' = c_9 + \sum_{i=1}^{t} \beta_i^{-9} \\ c_{21}' = c_{21} + \sum_{i=1}^{t} \beta^{-21} \end{cases}$$

We introduce the power and symmetric sums of the error positions, S_j and σ_j, with

$$S_j = \sum_{i=1}^{t} \beta_i^j$$

and σ_j = sum of the β_i taken j at a time. Then note that $d_i = S_i$ for $i = 1, 2, 3, \cdots, 19, 20$, and $c_{42}' = c_{42} + S_{21} = (c_{21}')^2$;

$$\begin{cases} c_{40}' = c_{40} + S_{23} = (c_5')^8 \\ c_{36}' = c_{36} + S_{27} = (c_9')^4 \\ c_{32}' = c_{32} + c_{32} + S_{32} = (c_1')^{32}. \end{cases}$$

If we can compute the above four values of S_i, then we may recover the correct code word very easily. We have the following sets of relations between S and σ from the Newton Formulas. If t errors are made, then $\sigma_i = 0$ for $i = t + 1, t + 2, \cdots$, and we have

$$\begin{aligned} &S_1 + \sigma_1 = 0 \\ &S_3 + S_2\sigma_1 + S_1\sigma_2 + \sigma_3 = 0 \\ &\qquad \vdots \\ &S_{21} + S_{20}\sigma_1 + S_{19}\sigma_2 + \cdots S_{21-t}\sigma_t = 0 \\ &\qquad \vdots \\ &S_{31} + S_{30}\sigma_1 + S_{29}\sigma_2 + \cdots S_{31-t}\sigma_t = 0. \end{aligned}$$

To solve for the σ's in the above equations, the $(t + 1, t + 1)$ determinant of the associated augmented matrix A must be 0. In particular, we have

$$\det \begin{vmatrix} S_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ S_3 & S_2 & S_1 & 1 & 0 & 0 & \\ S_5 & S_6 & S_3 & S_2 & & & \\ S_{2t+3} & S_{2t+2}, & \cdots, & S_{t+3} & & & \end{vmatrix} = \det A = 0.$$

Replacing the top row in A above with any of the following four rows,

$$\begin{cases} S_{21} & S_{20} & S_{19} \cdots S_{21-t} \\ S_{23} & S_{22} & S_{21} \cdots S_{23-t} \\ S_{27} & S_{26} & S_{25} \cdots S_{27-t} \\ S_{31} & S_{30} & S_{29} \cdots S_{31-t} \end{cases}$$

the determinant must still be zero.

Consider substituting the first row above, where only S_{21} is unknown. (The others are computable by the Reed

Formula.) If we are to get a zero determinant of the new matrix, we may solve for S_{21} (note that coefficient of $S_{21} \neq 0$). Now looking at the second row, after computing S_{21}, we see that only S_{23} is unknown, so placing this row in the above matrix and setting the determinant equal to zero, we may solve for S_{23}. We can repeat this for the other unknown values of S, each time using the fact that $S_{2j} = S_j^2$.

Using the fact that $c_1 = c_1' + S_1 = c_1'$, we simply add S_{31}^2 to c_1' and we obtain the correct c_1. Similarly, by adding S_{23}^8 to c_5', S_{27}^2 to c_9, S_{21}^2 to c_{21}', and t to c_0, we obtain the correct c's. Putting these correct c's in the polynomial $g_a(x)$ and plugging in the first 18 powers of β for x, we obtain the correct information bits. This method can be thought of as an "error-location number" method used in decoding Hamming codes (Ref. 3, Chapt. 5).

4. General Algorithm for Decoding

(1) From the received vector, compute the 15 coefficients c_0, c_1, c_5, c_{19}, c_{21} and d_i, $i = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$.

(2) If all the d's = 0, then word is correct; go on to the next one.

(3) For $t = 1$, form the augmented matrix for $t = 1$, and compute its determinant. If $\det A = 0$, then one error has been made; go on to the next four determinant operations and compute the S's, then the correct c's. Plug these into Formula (1) and obtain the correct information bits. If the augmented matrix A has non-zero determinant, assume $t = 2$ and proceed with same set of operations. Continue until one reaches the value $t = 10$ for which augmented determinant has determinant 0; then evaluate the appropriate S's, and correct c's to obtain information bits from the first 18 values of the corrected polynomial. See *SPS 37-29*, Vol. III, for a discussion of the implementation of this procedure in an actual teletype experiment.

5. Error-Correction Procedure for Cyclic Codes over GF (2^s)

We note that the same error-correcting procedure can be modified for use in any cyclic code over any symbol field $GF(2^l)$ where l is prime to k, and $GF(2^k)$ is the field of operations of the code. In particular, we consider here the (63,18) Bose–Chaudhuri code over $GF(2^5)$.

Using the same recursion formula as in the binary case [Eq. (1)] with initial information bits in $GF(2^5)$, we may generate a cyclic code with the same error-correcting properties as the binary, i.e., this code will correct 10 *symbol*-errors (*SPS 37-27*, Vol. IV). As in the binary case, to every vector $a = (a_i)$, $a_i \in GF(2^5)$, there corresponds a polynomial $g_a(x)$ of the same form as Eq. (2), but with exceptions:

$$g_a(x) = c_0 + c_1x + c_1^{2^5}x^{32} + c_1^{2^{10}}x^{16} + c_1^{2^{15}}x^8 + c_1^{2^{20}}x^4 + c_1^{2^{25}}x^2$$

$$+ c_5x^5 + c_5^{2^5}x^{36} + c_5^{2^{10}}x^{17} + \cdots + c_5^{2^{25}}x^{10}$$

$$+ c_9x^9 + c_9^{2^5}x^{36} + c_9^{2^{10}}x^{18} + \cdots + c_{21}x^{21} + c_1^{2^5}x^{12}.$$

Here $c_i \in GF(2^5)$, and the other c's are in $GF(2^{30})$, the smallest field containing both $GF(2^6)$ and the symbol field $GF(2^5)$ as subfields. For a code word to be correct, we still must have the d's (as before) equal to zero. Note as before, these coefficients are computed by the Reed Formula:

$$c_j = \sum_i a_i\beta^{-ij}, \qquad d_k = \sum_i a_i\beta^{ik}.$$

Defining S' as before, $S_j' = d_j$, $j \leq 19$, let us consider a new set of modified formulas. If t errors are made in the positions, $\beta_i, i = 1, \cdots, t$, with *symbol changes* α_i, define σ_j, as before, to be the sum of the t positions taken j at a time. We have the following relationship in the case of t errors:

$$\begin{cases} S_j = d_j = \sum_{i=1}^{t} \alpha_i\beta^{ij}; & j = 0, 1, 20; \\ c_0' = c_0 + S_0, & c_i' = c_i + S_i. \end{cases}$$

The modified version of the Newton Formula for arbitrary t is as follows:

$$\begin{cases} S_t + S_{t-1}\sigma_1 + S_{t-2}\sigma_2 + \cdots + S_0\sigma_t = 0 \\ S_{t+1} + S_t\sigma_1 + S_{t-1}\sigma_2 + \cdots + S_0\sigma_t = 0 \\ S_n + S_{n-1}\sigma_1 + \cdots + S_{n-1}\sigma_t = 0. \end{cases}$$

Further details on the decoding procedures are omitted.

C. Determinants for Error Correction

H. Fredericksen

1. Summary

In this article, we discuss finding the error vector to correct up to ten errors in a received block of 63 teletype symbols, 18 of which are information symbols, using the method of the previous articles (*SPS 37-26*, Vol. IV, pp. 223–225; *SPS 37-29*, Vol. III).

2. Determinants

The decoding procedure described in the foregoing article outlines a general procedure to be followed in decoding a teletype message. In summary, the critical calculation for determining the errors in the teletype message involves evaluating certain determinants. Using the notation in the previous article, for $t = 1, 2, 3, \cdots, 10$, we evaluate the $t \times t$ upper-left sub-determinants of A, where A is given in Fig. 1. The entries are elements of the 64-element field, corresponding to the splitting field of a sixth degree irreducible polynominal over $GF(2)$.

$$A = \begin{vmatrix} S_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_3 & S_2 & S_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_5 & S_4 & S_3 & S_2 & S_1 & 1 & 0 & 0 & 0 & 0 \\ S_7 & S_6 & S_5 & S_4 & S_3 & S_2 & S_1 & 1 & 0 & 0 \\ S_9 & S_8 & S_7 & S_6 & S_5 & S_4 & S_3 & S_2 & S_1 & 1 \\ S_{11} & S_{10} & S_9 & S_8 & S_7 & S_6 & S_5 & S_4 & S_3 & S_2 \\ S_{13} & S_{12} & S_{11} & S_{10} & S_9 & S_8 & S_7 & S_6 & S_5 & S_4 \\ S_{15} & S_{14} & S_{13} & S_{12} & S_{11} & S_{10} & S_9 & S_8 & S_7 & S_6 \\ S_{17} & S_{16} & S_{15} & S_{14} & S_{13} & S_{12} & S_{11} & S_{10} & S_9 & S_8 \\ S_{19} & S_{18} & S_{17} & S_{16} & S_{15} & S_{14} & S_{13} & S_{12} & S_{11} & S_{10} \end{vmatrix}$$

Fig. 1. Determinant A

3. Diagonalization

To evaluate these, we first put the determinant in lower triangular form. To evaluate the determinant, we need only multiply the elements on the diagonal. We find the first sub-determinant that is equal to zero, say $k \times k$. Then there have been $k - 1$ errors in transmission (*SPS 37-27*, Vol. IV, pp. 190–193), and we enter the algorithm for correcting them. *But since we are operating over a field, it is not even necessary to multiply the elements on the diagonal.* Instead, we need only evaluate the diagonal elements themselves and find the first one that is zero, say the k^{th} element. None of the smaller sub-determinants up to $(k-1) \times (k-1)$ can be zero, whereas all larger sub-determinants vanish. If the 10×10 determinant itself is non-zero, we handle this as a special case.

4. Decoding

Suppose k is the least integer for which the $k \times k$ sub-determinant $(1 < k < 10)$ is zero. Then there are $k - 1$ errors and there is an algorithm for correcting them! Take this $k \times k$ sub-determinant and for the k^{th} row substitute, in turn, the following four rows:

$$\begin{cases} a & S_{21} & S_{20} & S_{19} \cdots S_{21-k+1} \\ b & S_{23} & S_{22} & S_{21} \cdots S_{23-k+1} \\ c & S_{27} & S_{26} & S_{25} \cdots S_{27-k+1} \\ d & S_{31} & S_{30} & S_{29} \cdots S_{31-k+1} \end{cases}$$

and evaluate as before. Set the k^{th} element on the diagonal equal to zero and solve for S_{21}, S_{23}, S_{27}, S_{31}. For the case where the 10×10 determinant is non-zero, we assume 10 errors have been made, and we augment the 10×10 determinant by a row and a column to form an 11×11 determinant. For the last row we choose in turn the four rows

$$\begin{cases} S_{21} & S_{20} \cdots S_{11} \\ S_{23} & S_{22} \cdots S_{13} \\ S_{27} & S_{26} \cdots S_{17} \\ S_{31} & S_{30} \cdots S_{21} \end{cases}$$

and solve for S_{21}, S_{23}, S_{27}, S_{31}. When we have these values we can recover the numbers c_0, c_1, c_5, c_7, c_{21} of previous *Summaries*. With these we can recover the message that was sent.

A program is currently being written for the SDS 920 computer which will do all of the computations described above. Preliminary calculations show the decoding can be done in 0.02 word times.

D. Parallel Generation of the Check Bits of a PN Sequence

T. O. Anderson and W. A. Lushbaugh

1. Summary

This article describes a switching network whereby all the check bits in a PN sequence are generated in parallel. For k inputs, $2^k - 1 - k$ outputs (the check bits) are derived. A simple two-variable mod 2 bridge circuit with single-term inputs that can be directly cascaded is used throughout. A minimum propagation time through the network has been achieved. Also, special emphasis has been placed on an even load distribution among the original inputs as well as among the individual circuits. An example has been worked out for $k = 6$. Fifty-seven mod 2 circuits, each for two variables, are used.

2. Background

The PN sequence is normally generated in series by a k-bit shift register, whose input is a modulo 2 circuit for two or more of the k bits. The k bit register assumes all $2^k - 1$ different states other than the all zero state. Thus, sequences generated by different starting positions are merely cyclic shifts of one another (JPL TR 32-67). The $2^k - 1$ sequences each of $2^k - 1$ bits thus can be thought of as a PN code dictionary.

If the objective were to generate such a dictionary serially, a continuously recycling k bit register with a mod 2 circuit input and with the register being realigned one step per complete cycle would yield a solution. If constant bit rate is of essence an extra output control bit can be used, and once per cycle the register is shifted one extra step while the output control bit is shifted continuously at a constant rate. If instead the object were to generate the complete PN-sequence dictionary in parallel, that is, a set of $2^k - 1$ words each of $2^k - 1$ bits, a $2^k - 1$ bit shift register would be appropriate, but expensive.

Such a register would be preset to any one of the possible sequences and the complete dictionary would be generated simply by recycling the register, thus generating one complete sequence for each step. The above mentioned two methods for generating PN-sequence dictionaries are applicable in serial and parallel decoders for cyclic codes (*SPS 37-27*, Vol. III, pp. 97–103; and *SPS 37-28*, Vol. IV). In certain algebraic decoding procedures (*SPS 37-27*, Vol. IV, pp. 190–194) it becomes necessary to generate the PN-sequence check bits in parallel for a given code word.

In this case, recycling $2^k - 1$ bit shift register could be used accompanied with a k bit comparator for selection of the particular code word. The shift generation is time consuming, and would in the worst case require $2^k - 2$ clock periods. A method is sought of instantaneously generating the $2^k - 1 - n$ PN-sequence check bits from k information bits. Such a design shall now be described, using $k = 6$ as an example.

3. Design

Let the information bits of a code word be $x_0, x_1, \cdots, x_5$. The PN-sequence terms to be generated are $x_6, x_7, \cdots, x_{62}$. The recursion relation between the elements of the seequence is

$$x_n = x_{n-5} \oplus x_{n-6},$$

corresponding to the primitive polynomial $x^6 + x + 1$ (JPL TR 32-67). Each of the sequence terms $x_6 x_7, \cdots, x_{62}$ can then be expressed as some mod 2 function of the generating bits $x_0, x_1, \cdots, x_5$. An example for x_{12} will show how the function for a sequence term is derived.

$$x_{12} = x_6 \oplus x_7$$

but

$$x_6 = x_0 \oplus x_1$$

and

$$x_7 = x_1 \oplus x_2$$

so

$$x_{12} = x_0 \oplus x_1 \oplus x_1 \oplus x_2;$$

and since $x_1 \oplus x_1 = 0$, we have

$$x_{12} = x_0 \oplus x_2.$$

From this, it can be concluded that each of the terms $x_6, x_7, \cdots, x_{62}$ can be reduced to mod 2 function of from two to six of the generating bits. And all non-zero linear combinations occur.

Table 2 is a list of the variables, the mod 2 function of which generates the PN-sequence terms in order x_6, x_7, $\cdots$, x_{62}. There are

$\binom{6}{2} = 15$ terms with two variables,

$\binom{6}{3} = 20$ with three,

$\binom{6}{4} = 15$ with four,

$\binom{6}{5} = 6$ with five,

$\binom{6}{6} = 1$ with six variables.

Since mod 2 summing is an associative operation, it is evident that the outputs of the mod 2 functions for two variables can be economically used as inputs to mod 2 functions of higher orders. To derive the mod 2 function for 2, 3, 4, 5, and 6 variables a two-term mod 2 circuit may be cascaded in several different ways. The deciding factors are propagation time and load distribution. The circuit shown in Fig. 2 emphasizes maximum load of the original terms and minimum load of the individual mod 2 circuits, resulting in a five level gating structure. In

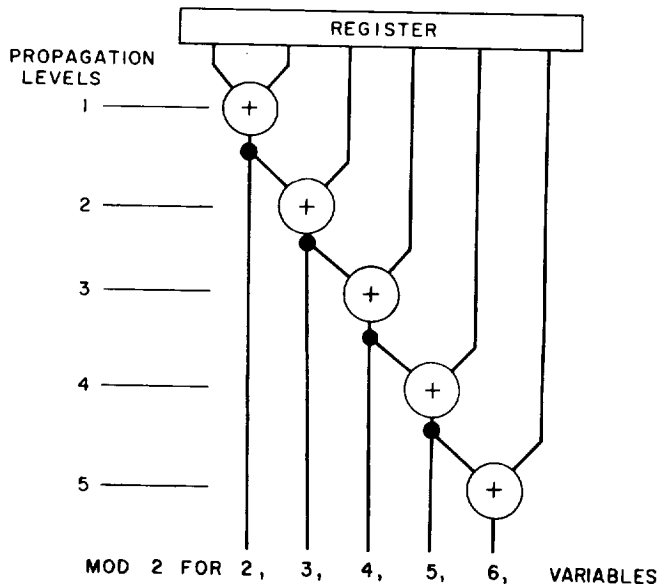


Fig. 2. Cascading mod 2 circuits. Maximum load of the original word and minimum load of individual circuits

Table 2. List of the variables, the mod 2 function of which generates the PN-sequence terms in order from x_6 to x_{62}

x_n	x_0	x_1	x_2	x_3	x_4	x_5
6	1	1				
7		1	1			
8			1	1		
9				1	1	1
10					1	1
11	1	1				1
12	1		1			
13		1		1	1	1
14			1		1	
15				1		1
16	1	1			1	
17		1	1			1
18	1	1	1	1		
19		1	1	1	1	
20			1	1	1	1
21	1	1		1	1	1
22	1		1		1	1
23	1			1		1
24	1				1	
25		1				1
26	1	1	1			
27		1	1	1		
28			1	1	1	
29				1	1	1
30	1	1			1	1
31	1		1			1
32	1			1		
33		1			1	
34			1			1
35	1	1		1		
36		1	1		1	
37			1	1		1
38	1	1		1	1	
39		1	1		1	1
40	1	1	1	1		1
41	1		1	1	1	
42		1		1	1	1
43	1	1	1		1	1
44	1		1	1		1
45	1			1	1	
46		1			1	1
47	1	1	1			1
48	1		1	1		
49		1		1	1	
50			1		1	1
51	1	1		1		1
52	1		1		1	
53		1		1		1
54	1	1	1		1	
55		1	1	1		1
56	1	1	1	1	1	
57		1	1	1	1	1
58	1	1	1	1	1	1
59	1		1	1	1	1
60	1			1	1	1
61	1				1	1
62	1					1

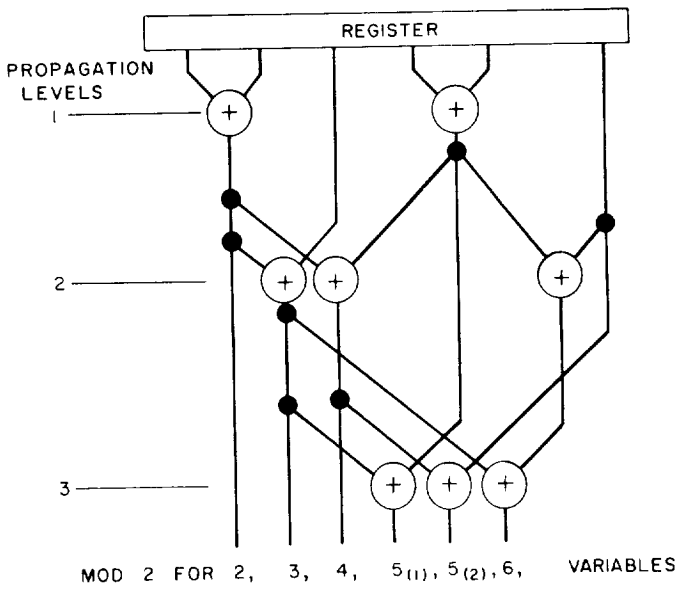


Fig. 3. Minimum propagation time; evenly distributed loads

Fig. 3, the number of gating levels is held to a minimum resulting in greater loads on the individual mod 2 circuits.

4. Optimum Arrangement

Fig. 4 shows a gate tree of 57 mod 2 circuits cascaded to minimize the number of levels and to optimally balance the loads. The numbers in parentheses give the position of that output in the PN sequence in accordance with Table 2. The 15 two-variable sequence terms are derived at the first level of the mod 2 circuits in such a way that each of the original terms drives five mod 2 inputs.

The second level of mod 2 circuits contains the 20 three-variable and the 15 four-variable terms. The three-variable terms must be formed by a mod 2 circuit which uses one two-variable input and one of the original single variables. Since there are 15 two-variable terms and 20 three-variable terms, the loading cannot be completely balanced. It is possible though, to have only the minimum of 5 two-variable outputs used twice. Fig. 4 shows one of the many solutions. It is conjectured that any five may be chosen as those to be used twice.

Generation of the 15 four-variable terms from the 15 two-variable terms is the most interesting. Since 2 two-variable terms are to be combined to get each four-variable term, it is natural to ask if each two-variable term may be used exactly twice in generating the complete set of four-variable terms. It turns out that this is possible, and the solution is unique up to relabeling the terms. Table 3 shows this solution as five sets of triples of pairs, which, when taken two at a time across any row, generate all quadruples of numbers from 1 to 6.

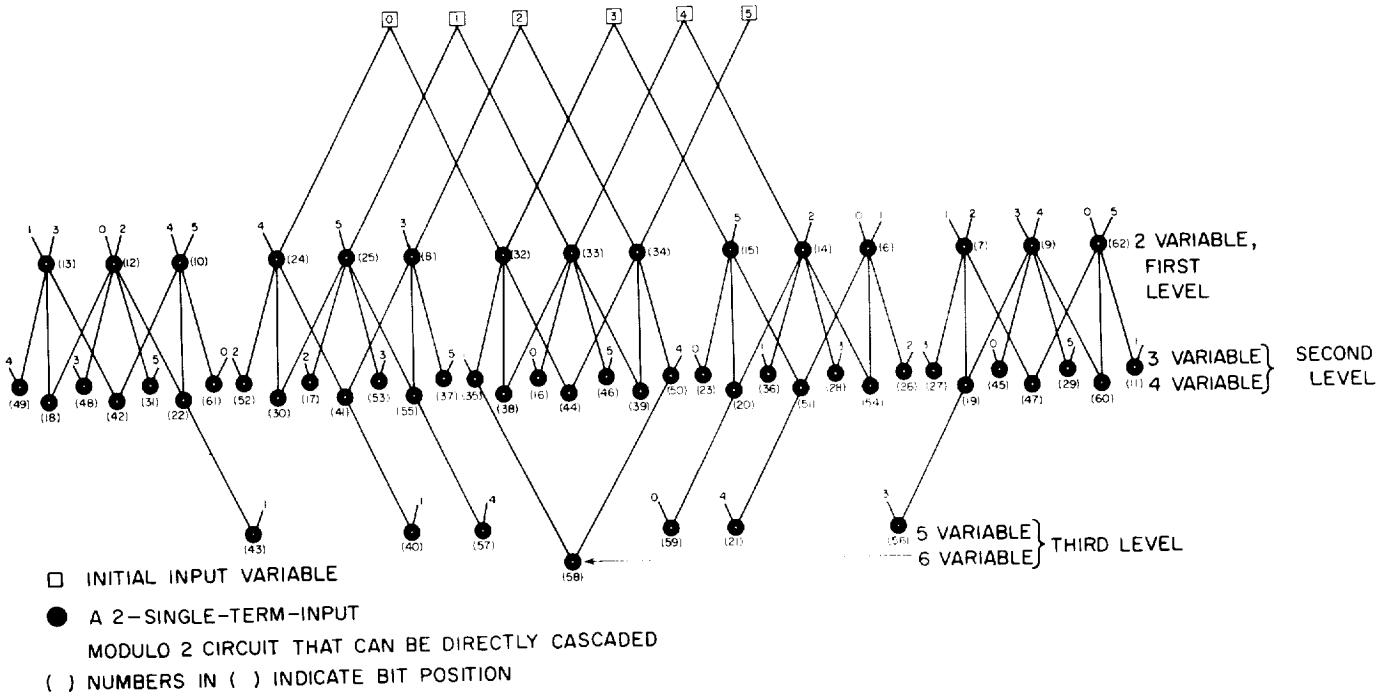


Fig. 4. Complete gate tree generating the 57 check bits

Table 3. Fifteen combinations of six things taken two at a time

Five sets of triples of pairs	1 2	3 4	5 6
	1 3	2 5	4 6
	1 4	2 6	3 5
	1 5	2 4	3 6
	1 6	2 3	4 5

5. Combinational Problem

This interesting combinatorial problem can be stated more generally as follows. Given the

$$\binom{3m}{m}$$

combinations of $3m$ numbers taken m at a time, can each combination be used once to form

$$\frac{1}{3}\binom{3m}{m}$$

triples such that each triple has each number in it. Table 3 is the solution for $m = 2$. A solution also has been discovered for $m = 3$, but no general algorithm has yet been discovered for the general case.

The 6 five-variable terms are best formed from six of the fours plus one of the original. The six-variable term can be formed from any two disjoint three-variable terms.

One extra load balancing feature to be noticed in arrangements such as that shown in Fig. 4 is that the mod 2 sum of complemented variables is equal to that of the uncomplemented variable. Thus, if the generating bits are contained in flip-flops, the two-variable sums should be derived from the complement side of the flip-flop and the other single-variable inputs should be derived from the uncomplemented side.

6. The Two-Variable Mod 2 Circuit Module

As shown in Fig. 5, the modular mod 2 bridge circuit resembles a full wave rectifier bridge. The A and B terms are connected to two opposing nodes, while the base and emitter of a transistor switch are tied to the remaining two nodes.

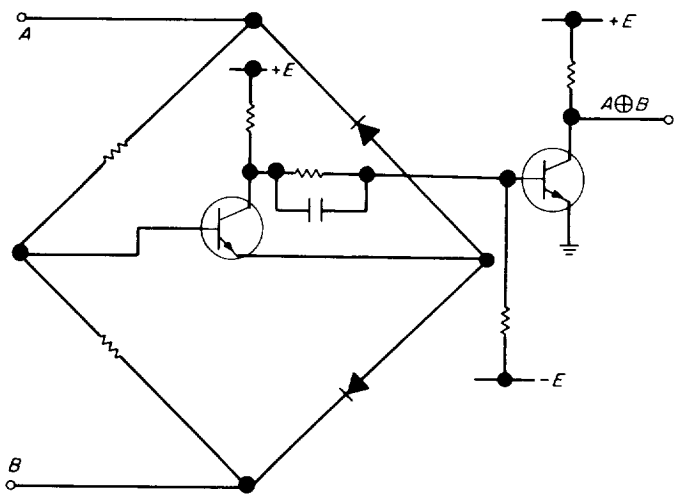


Fig. 5. Two-variable mod 2 circuit

When A and B are of the same logic value, i.e., both *ones* (high levels) or both *zeros* (ground), the base and emitter will assume the same voltage and the transistor will not conduct. Conversely, with A and B of opposing logic values, the transistor will conduct. The collector output, however, is removed from ground by two emitter-collector junctions and one diode junction, and would cause difficulty in cascading similar circuits. A second transistor switch is therefore added to restore the logic levels, to invert the logic function for *true* representation and to provide added drive capability for cascading. The input to the mod 2 circuit presents a *nor*-load that is a negative current load as opposed to a *nand*-load, which is a positive current load. The output stage is designed accordingly.

E. Weight Comparison Between Binary Code Words

T. O. Anderson and W. A. Lushbaugh

1. Summary

This article discusses methods of comparing two binary n-tuples to determine which has the greater *weight*, that is, the greater number of *ones*. The weight of a code word, or rather the comparison of the weights of two or more

code words, is of special interest in certain decoding procedures (*SPS 37-27*, Vol. III, pp. 97–103).

In fact, one difficulty that has prevented the wider adoption of error-correcting coding is the complexity of decoding procedures. Ane one of the more time-consuming tasks in many decoding procedures is the problem of deciding which of the possible error vectors (that could have been added to a code word to obtain the received word) has the least number of ones.

This article describes different methods to perform these operations in decoders built with digital modules, and also those built with threshold elements (*SPS 37-28*, Vol. IV, pp. 235–240). It is shown that the threshold element results in by far the simplest decoders.

2. Punctured-Cyclic Decoder

An error-correction encoder–decoder is considered using the theory of *SPS 37-27*, Vol. III; and *SPS 37-23*, Vol. IV, pp. 149–151. A k-bit data word is expanded through a linear feedback network to $2^k - 1$ bits. The decoding of the expanded word is executed by comparing the received word with all possible words in the dictionary. That dictionary word which causes the least number of bit-disagreements is the word most likely to have been transmitted. The number of bit-disagreements between two words is just the weight of their modulo 2 sum. Thus, it is desired to determine which of the words formed by adding the received word to each of the dictionary words has the least weight.

In a *series comparison* between the received word and a dictionary word, the two are added, and the weight of the sum can be simply obtained by tallying the number of *ones* by a binary counter. With the weight of a word presented as a binary number, the comparison is then quite simple. The result of interest in comparing the weights of these mod 2 sums is whether one weight is greater, smaller, or equal to another; one is not interested in the amount of weight by which one sum is greater or smaller than the other.

For a weight comparison to be implemented in a simple manner, the weight could well be represented in any other code but binary. Another useful such code would be one where all the *ones* in the word would be adjacent to one another starting from one end. The word with the greater weight would then be found simply by applying implication gates between corresponding bits of the two code words. All $A > B$ gates would be *or*-ed together and all $B > A$ would be *or*-ed together. Unless the weight of the two code groups is equal, at least one of the implication gates will be true, and thus will indicate which code group has the greatest weight. Fig. 6 shows the logic connection of the implication gates. A method of grouping all *ones* in a word adjacent to each other and starting from one end regardless of their original positions is shown in Fig. 7.

The method is one of shifting and accumulating all *ones* in one end. The *ones* are shifted to the right provided they are preceded by *zeros*, and the *zeros* are shifted to the left. In other words, adjacent *ones* and *zeros* are changing place, the *ones* stepping to the right and the *zeros* to the left.

In decoding schemes of the type outlined in the above example, however, these serial techniques often require decoding times incompatible with the transmission rate. Weight comparison in parallel then becomes necessary.

3. Parallel Digital Comparison

Because of its general interest, parallel digital comparison schemes will now be discussed. One such scheme is

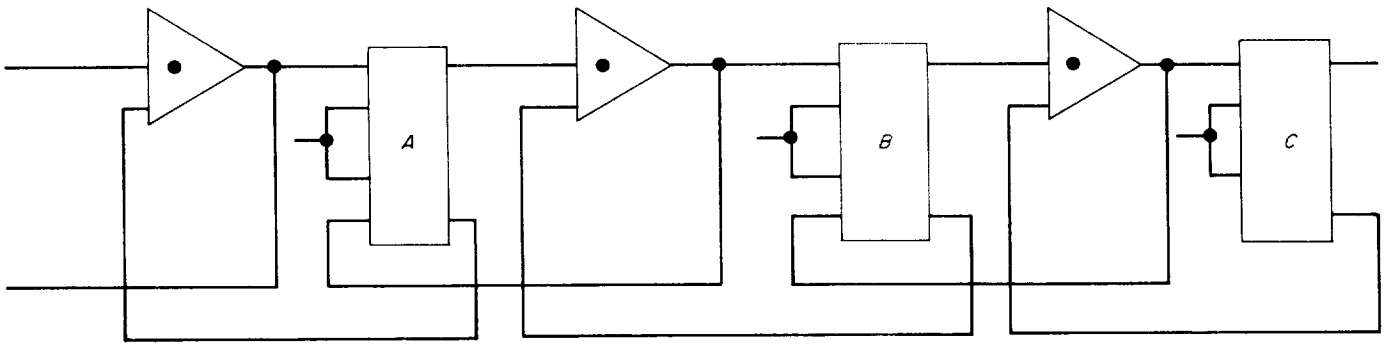


Fig. 6. Logic diagram of a register that shifts ones to the right and zeros to the left

that of instantaneously generating either the binary function for the number of *ones*, or the function where all *ones* are adjacently located. A second scheme is simply to perform a parallel binary addition on successively larger groups. And a third scheme is the successive elimination of common *ones* between the words to be compared until one group shows an all *zero* condition. The three methods outlined above will now be discussed in more detail.

It must be realized, however, that any gating matrix will be of such depth that the propagation through its numerous layers must be considered and yet looked upon as a serial operation. By intermediate buffering with registers, an isolation can be afforded which will allow several words to propagate through the network at the same time, each in different stages. The final output will then be misaligned in regard to the original word with a fixed number of clock periods, but the misalignment between any two results would only be one clock period.

a. Code conversion method. Table 4 illustrates the conversion for four variables: Truth Table (a) represents all possible combinations of four variables; (b) shows the corresponding weight expressed in binary code; and (c) shows the corresponding weight expressed in a code where all ones adjacent to one another starting from one end. The brute force method of code conversion is that of decoding all functions of $2^k - 1$ variables and group the outputs as to their weight, and then to encode the outputs of these groups into a k-bit binary number.

Table 4. Code conversion from binary code to two weight codes

A	B	C	D	A	B	C	D	A	B	C	D
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	1	0	0	0	1
0	0	1	0	0	0	1	0	0	0	0	1
0	0	1	1	0	0	1	0	0	0	1	1
0	1	0	0	0	0	0	1	0	0	0	1
0	1	0	1	0	0	1	0	0	0	1	1
0	1	1	0	0	0	1	0	0	0	1	1
0	1	1	1	0	0	1	1	0	1	1	1
1	0	0	0	0	0	0	1	0	0	0	1
1	0	0	1	0	0	1	0	0	0	1	1
1	0	1	0	0	0	1	0	0	0	1	1
1	0	1	1	0	0	1	1	0	1	1	1
1	1	0	0	0	0	1	0	0	0	1	1
1	1	0	1	0	0	1	1	0	1	1	1
1	1	1	0	0	0	1	1	0	1	1	1
1	1	1	1	0	1	0	0	1	1	1	1
(a)				(b)				(c)			

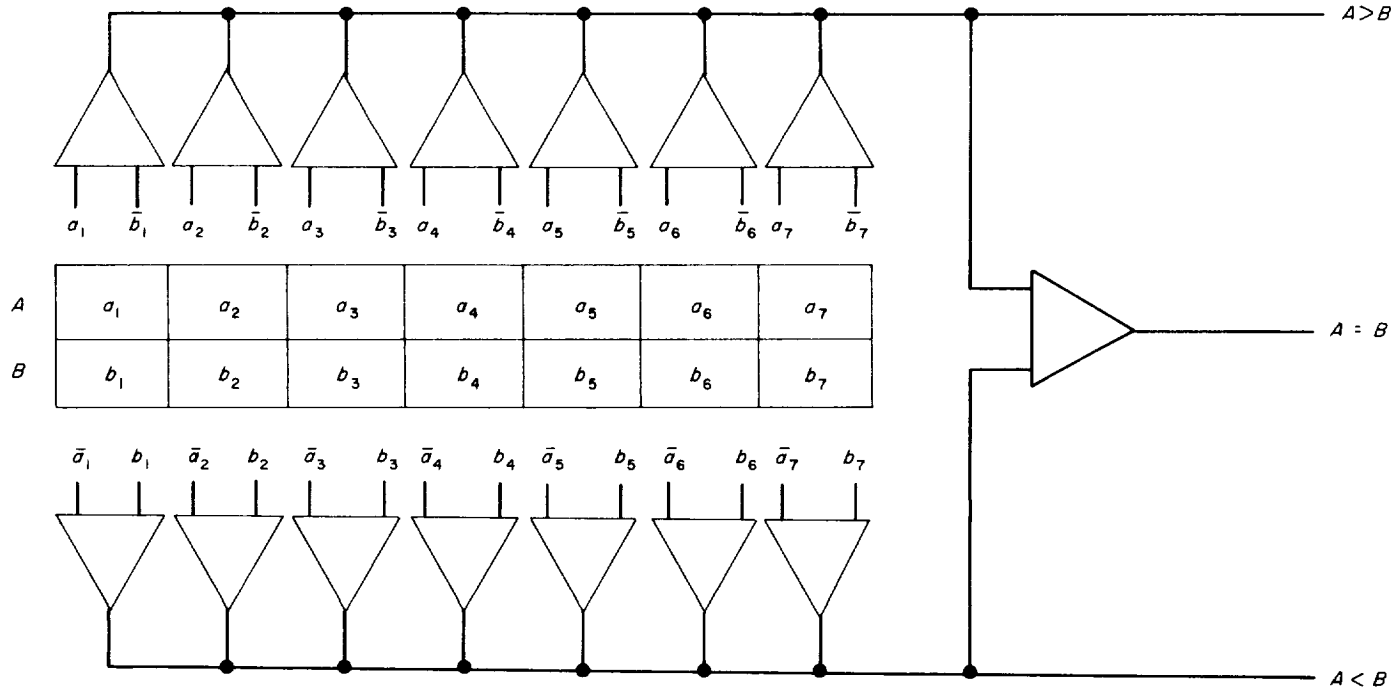


Fig. 7. Comparison of two numbers where all ones are adjacent starting from one end

b. Arithmetic method. The arithmetic method of deriving the weight function includes successive addition starting with small groups of the original word. The weight of a small group expressed in binary notation can readily be derived through combinational logic in the following manner:

For groups of two variables x_1, x_2 the 2^0 digit is $x_1 \oplus x_2$ and the 2^1 digit simply x_1x_2.

For groups of three variables $x_1x_2x_3$ the 2^0 digit is $x_1 \oplus x_2 \oplus x_3$, and the 2^1 digit the majority function of x_1, x_2, x_3, i.e., $x_1x_2x_3 + x_1x_2x_3 + x_1x_2x_3$.

The weight function for groups of seven bits consist of three binary bits which are similar in structure to those mentioned. The 2^0 digit is again the mod 2 sum of the seven variables and the 2^2 digit is the majority function for seven variables. The 2^1 digit is that function of seven variables which is a *one* for words with 2, 3, 6 or 7 *ones*. In generating this function, the 2^0 and 2^2 functions can be put to good use. It will be noticed that 2 *ones* is the only even minority except for *zero*, and 3 *ones* is the only odd minority other than 1, etc.

An example of successive parallel addition after the first grouping will now be discussed. 63 variables are partitioned into groups of three. The combinational

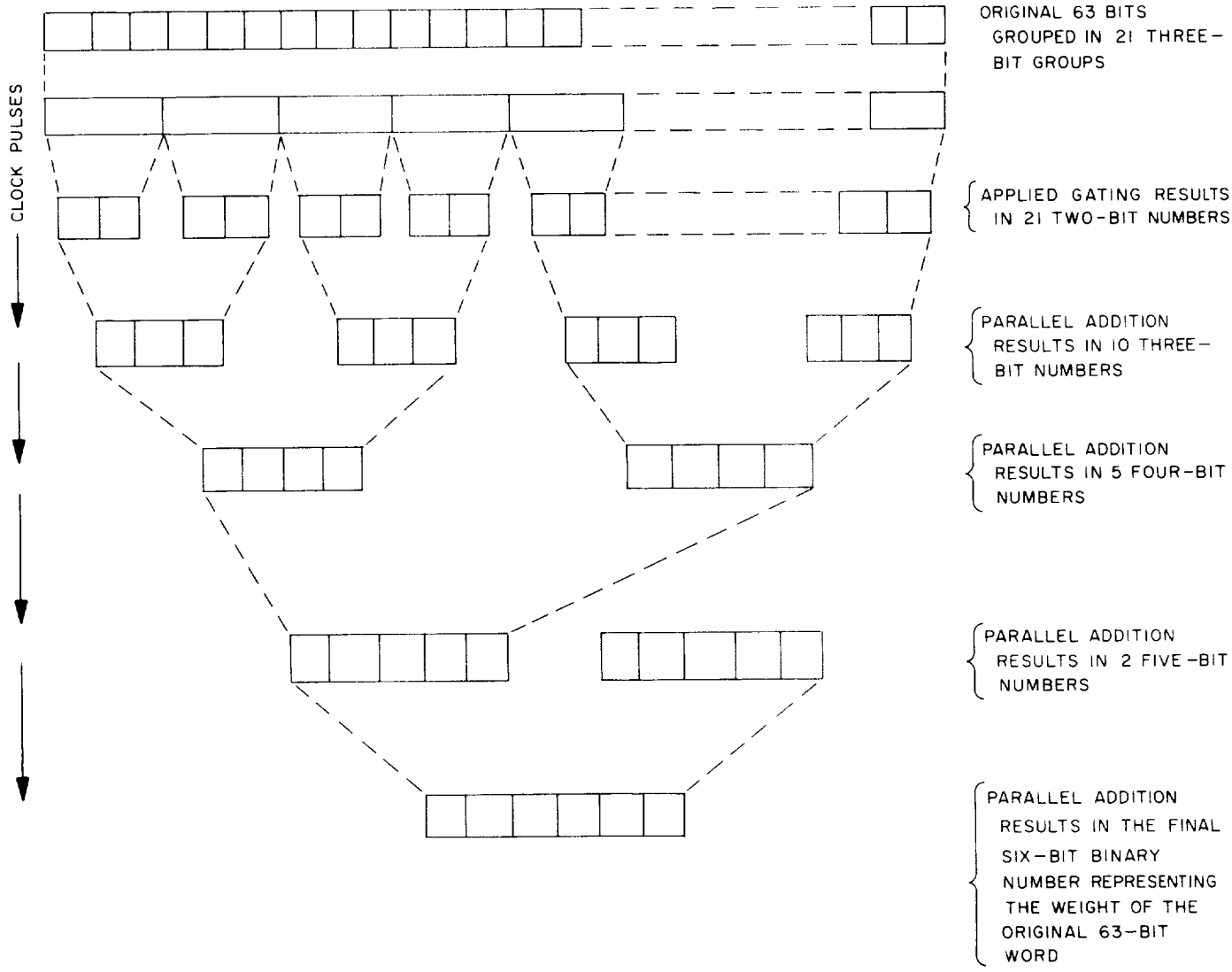


Fig. 8. Arithmetic derivation of the weight function by successive serial-parallel addition

weight function for three variables is used to form 21 two-bit binary numbers. The 21 two-bit numbers can be added in the usual fashion two at a time resulting in 10 three-bit numbers and 1 odd two-bit number. This process of adding the resulting numbers two at a time is continued until the final 6-digit result is obtained. A block diagram of the system is shown in Fig. 8.

An alternate approach is to start with the weight function for seven variables at the first stage, i.e., dividing the 63 bits into nine groups of seven bits. In the six-step process, the same number of gates as in the 21 by 2 procedure is used, but there is a saving of register flip-flops.

c. Successive elimination of common ones. This method is simple in concept and requires a moderate amount of hardware which can be arranged in an orderly fashion. A static switching network was developed on the basis of the following sequential circuit model.

The two words whose weight is to be compared are contained in two registers: *A* and *B*. Register *A* is a stationary hold register and *B* is a recycling shift register. An identity gate is formed between corresponding bits of register *A* and *B*. The output of these identity gates will reset the corresponding register bits. With register *B* recycling, each bit of *A* will be compared with each bit of *B* and common *ones* will be reset to zero. Unless both words are of equal weight, one of the two registers will, after one complete cycle, contain an all *zero* word. The conclusion is then that the other word is the word with the greater weight. If both registers contain all *zero* words, the original words are of equal weight.

This sequential synchronous-shift technique requires as many clock periods as there are bits in the words and is therefore time-consuming. And no new word can be entered until the first calculation is completed. Quite evidently, however, the same basic technique can very successfully be applied using a static switching network.

The result of each comparison would then again be compared in a second level of gating, with the inputs however, skewed one position per level as illustrated in Fig. 9. The gating structure would be the same for each level and the total time required would be the compounded propagation time. By manipulating the switching equations, two, three, four or more steps can be combined with considerable savings in both switching elements and propagation time.

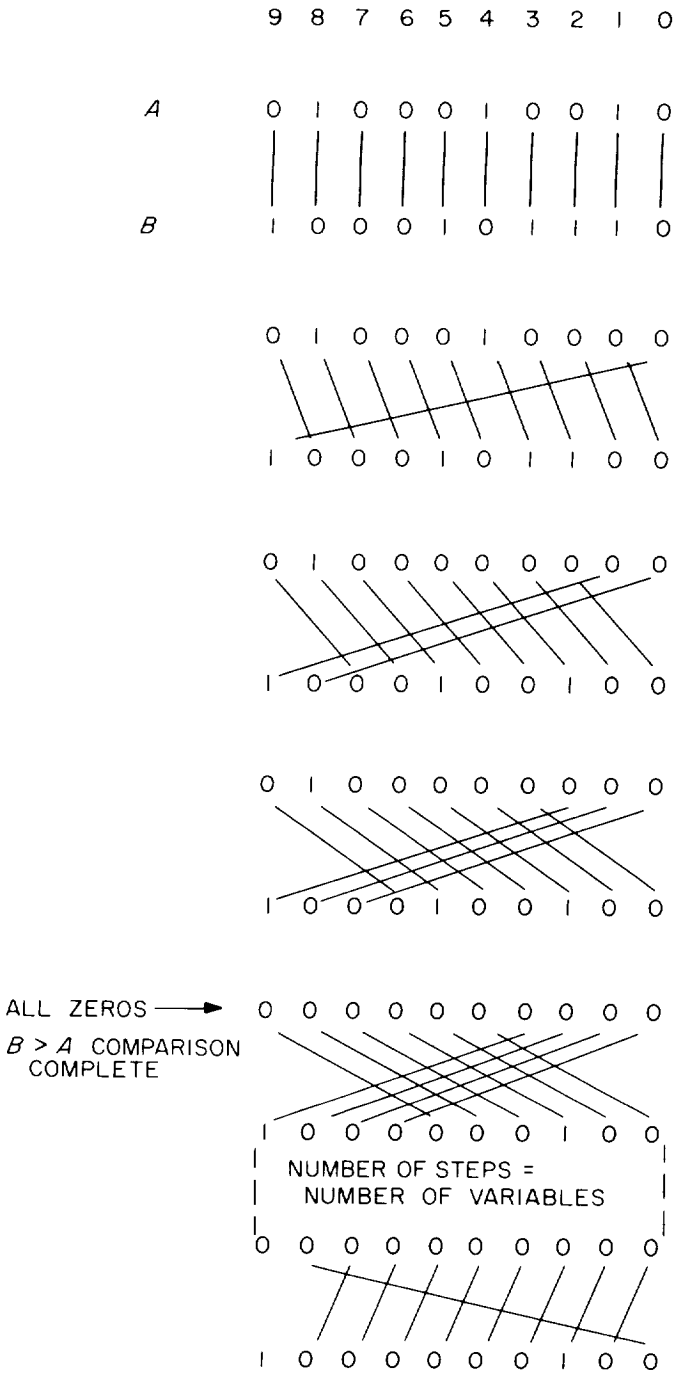


Fig. 9. Successive elimination of common ones

4. Analog Threshold Comparison

It has been seen previously that digital evaluation of the weight function is an expensive and bulky solution to the problem, and in no case could be calculated without considerable delay. An analog approach, however, was shown in *SPS* 37-28, Vol. IV, pp. 235–240, to require a

reasonably small amount of hardware and perform the calculation instantaneously. We now review that procedure.

If each bit of a word is entered into an analog summing network, the output voltage will be proportional to the weight of the word. Now the decoding technique assumes that some code word, which is necessarily unique, will cause at most e disagreements with the received word; e is the number of errors which the code is capable of correcting, i.e., $e = 2^{k-2}$. Then the correct word can be chosen by means of a threshold amplifier. An output from the amplifier would mean that the last word tested was at most e units away from the received word, and hence the unique correct transmitted word.

The accuracy of the analog device used would have to be one part in 2^k, where k is the number of information bits in the code word. Commercial summing networks and threshold amplifiers are readily available then to handle decoding for k up through 12.

For words containing a large number of digits when the accuracy of the analog devices becomes questionable, the analog networks may be divided into groups and the output from each group converted to a digital binary number. Continued comparison is then performed digitally.

One concludes that parallel calculation of the weight function, essential to many decoding schemes, is best performed by an analog device. All digital schemes fall short of the threshold scheme in either hardware or time of calculation, or both.

F. Lattice-Structure of Codes over Binary Aysmmetric Channel

R. J. McEliece

1. Summary

A structure is developed for codes over the binary asymmetric channel, roughly analogous to the group structure for codes over the binary symmetric channel. The structure is defined in terms of lattices of sets, where the order relation is a simple dominance relation.

2. Introduction

Recently, interest has been shown at JPL in the binary asymmetric channel (*SPS 37-27*, Vol. IV, pp. 193–195). This channel arises, for example, in the *Ranger* Block III command detector (*SPS 37-27*, Vol. III, pp. 222–225), where a bit threshold detector is used, set asymmetrically. In fact, the threshold is set so asymmetrically that essentially only one of the two kinds of error can occur. Thus, the "1 → 0" transition occurs with a certain probability when a *one* is sent, but the "0 → 1" transition has never been observed when a *zero* is sent. This total asymmetry was taken in *SPS 37-27* as the definition of the binary asymmetric channel.

One trouble with developing a theory of error-correcting codes for use in such asymmetric channel has been the lack of a structure theory for such codes, corresponding to Hamming distance and modulo 2 addition in the symmetric case. This paper puts a lattice structure on binary codes, where the lattice operation has special advantage to the asymmetric channel.

3. The Lattice

Let us consider binary words of *zero* and *one* of fixed length n; the transition "1 → 0" occurs with probability $p > 0$, but the "0 → 1" transition does not occur at all. In such a block code κ of length n, there are 2^n binary words possible. Let us define a partial ordering "$\geqq$" of κ' in the following way: We say $W_1' \geqq W_2'$ for two words of κ' if and only if whenever W_1' is a transmitted word; then W_2' has a positive probability of being received. (Alternatively, in the channel we are considering, W_2' has a zero in every position that W_1' does.) The relation $\geqq$ is a partial ordering; it satisfies the following three laws:

(P1) $W' \geqq W'$ for all $W' \in \kappa'$.

(P2) If $W_1' \geqq W_2'$ and $W_2' \geqq W_3'$, then $W_1' \geqq W_3'$.

(P3) If $W_1' \geqq W_2'$ and $W_2' \geqq W_1'$, then $W_1' = W_2'$.

Consider a set S of n distinct objects $a_1, a_2, \cdots, a_n$. Then there are 2^n subsets of S. Let κ be the collection of the subsets of S. If we make a correspondence between κ and κ' by associating the word W' of κ' with the subset W of S if and only if W' has a *one* in the i^{th} position if and only if $a_i \in W$, then we see that the $\geqq$ relation introduced an κ' is the ordinary set inclusion relation of κ.

Since there is really no difference, then, between κ' and κ, in the rest of the discussion we will only consider κ.

κ is called the *Boolean Algebra* on n objects. It forms a lattice, where l.u.b $(W_1,W_2) = W_1 \cup W_2$ (ordinary set union), and g.l.b. $(W_1,W_2) = W_1 \cap W_2$ (ordinary set intersection.) Also, we shall later need the notation $A \setminus B$ for $A \cap B^c$ (= B complement).

4. Dimension Function

We define the *dimension* d of an element of κ as the number of elements it contains. Thus, $0 \leq d(W) \leq n$ for all $W \in \kappa$.

Lemma. $d(W_1) + d(W_2) = d(W_1 \cup W_2) + d(W_1 \cap W_2)$.

Proof. On the left-hand side, every element in W_1 and W_2 is counted once. The only elements counted twice on the left-hand side are then those which occur in both W_1 and W_2, i.e, in $W_1 \cap W_2$. The right-hand side simply counts all the elements which are in either W_1 or W_2 ($d(W_1 \cup W_2)$), and then gets the repeats ($d(W_1 \cap W_2)$).

5. Error Correction

We say that a set $L \subseteq \kappa$ is an *e-error-correcting code* if for L_1, L_2, $\in L$, if $L_1 \geq W$, $L_2 \geq W$, then either $d(L_1 \setminus W) \geq e + 1$ or $d(L_2 \setminus W) \geq e + 1$. (In the usual context this says that if two transmitted words L_1,L_2 can give rise to the same received word W, then there have been at least $e+1$ errors made on one of the words.) We then have the following result:

Theroem. A set $L \subseteq \kappa$ is an e-error-correcting code if and only if for all pairs L_1, L_2 for which $d(L_1) \geq d(L_2)$ we have $d(L_1 \cap L_2^c) \leq e + 1$.

Proof. In κ, if $W_1 \geq W_2$, then $d(W_1) \geq d(W_2)$. Now if $L_1 \geq W$, $L_2 \geq W$, then $L_1 \cap L_2 \geq W$. (This is the fundamental property of "$\cap$" in a lattice.) Consequently, $d(L_i \setminus W) \geq d(L_i \setminus L_1 \cap L_2)$, $i = 1,2$.

The assumption that $d(L_1) \geq d(L_2)$ means that $d(L_1 \setminus W) \geq d(L_2 \setminus W)$ for all $W \in \kappa$. In this case, $\max\{d(L_1 \setminus W),\ d(L_2 \setminus W)\} \geq e + 1$ if and only if $d(L_1 \setminus W) \geq e + 1$. But as remarked above,

$$d(L_1 \setminus W) \geq d(L_1 \setminus L_1 \cap L_2).$$

And

$$L_1 \setminus (L_1 \cap L_2) = L_1 \cap (L_1 \cap L_2)^c = L_1 \cap (L_1^c \cup L_2^c)$$
$$= L_1 \cap L_1^c) \cap (L_1 \cap L_2^c) = L_1 \cap L_2^c$$

by ordinary set algebra. Hence, $d(L_1 \setminus W) \geq d(L_1 \cap L)^c$. The proof is complete if we notice that equality is attained here when $W = L_1 \cap L_2$. (Obviously $L_1 \geq L_1 \cap L_2$, $i = 1,2$.) This completes the proof of the theorem.

6. Symmetric Channel

It is interesting here to see what the requirement for e-error-correcting codes over the ordinary symmetric channel becomes in the lattice-theoretic content. Of course, the definition of e-error-correcting codes given above is not suitable, since it referred to the asymmetric channel we have been considering. What is required here is that the "distance" between any two code words be $\geq 2e + 1$; where the "distance" between two words is defined to be the number of positions in which they disagree. If we start with L_1, say, and begin changing its coordinates in order to reach L_2, we may proceed as follows: First, change all *ones* in L_1 which do not occur in L_2 into *zeros*. The resulting word is $L_1 \cap L_2$. Then change all *zeros* of $L_1 \cap L_2$ which are *ones* in L_2 to *ones*. This brings us to L_2. The number of changes required was

$$d\left(\frac{L_1}{L_1 \cap L_2}\right) + d\left(\frac{L_2}{L_1 \cap L_2}\right) = d(L_1 \cap L_2^c) + d(L_2 \cap L_1^c)$$

by arguments used in the proof of the theorem. Thus, in the lattice-theoretic context the requirements for an e-error-correcting code may be given for the symmetric channel as well as for the asymmetric:

(1) $\max\{d(L_1 \cap L_2^c),\ d(L_2 \cap L_1^c)\} \geq e + 1$ (asymmetric channel);

(2) $d(L_1 \cap L_2^c) + d(L_2 \cap L_1^c) \geq 2e + 1$ (symmetric channel).

When the requirements are stated this way, the relationship between the two concepts is clarified. We immediately see that the code over the asymmetric channel is a weaker notion since, formally, (2) implies (1). But it also points up the fact that the cascade of an asymmetric channel and its complement yields the symmetric channel; it is thus a more basic concept. And in fact any degree of asymmetry can be studied by proper cascades.

References

1. Kendall, W. B., and Reed, I. S., "Path-Invariant Comma-Free Codes," *IRE Transactions on Information Theory*, Vol. IT-8, No. 6, pp. 350–355, October 1962.

2. Golomb, S. W., Gordon, B., and Welch, L. R., "Comma-Free Codes," *Canadian Journal of Mathematics*, Vol. 10, No. 2, pp. 202–209, 1958.

3. Peterson, W. W., "Error-Correcting Codes" (see Chapt. 9), Wiley & Sons, New York, 1961.

4. Solomon, G., "A Note on a New Class of Codes," *Information and Control*, Vol. IV, 1961.

XXV. Communications Systems Research: Information Processing

A. Statistics of Pseudo-Random Pulses

E. C. Posner

This article extends the mathematical theory of random pulses developed in Ref. 1 and demonstrates the agreement between theory and experiments performed using the random-pulse generator recently built in JPL Section 331 (Refs. 2 and 3). The mean and variance of the number of pulses in n bit times are derived for the various options obtainable with the random pulse generator. Experiments performed on the output of the random pulse generator show that the generator behaves according to theory.

1. Review

We shall first review the principles of the random pulse generator of Refs. 2 and 3. This device produces an output train of pulses for which the average density of pulses is a variable, as is also the conditional probability of a pulse at an arbitrary time, given a short segment of the previous pulse history. In effect, what was desired was the ability to produce trains of pulses with a variable covariance function.

Such variability is useful in many contexts. For example, as a *noise generator*, various covariance functions of the pulse train correspond to various "colorings" of white gaussian noise. When the "colored" noise is quantized along with the signal, the errors produced (which correspond to pulses) have such varying covariance functions. In particular, a large positive autocorrelation function corresponds to bursts of errors, useful in testing error-correcting coding–decoding systems (Ref. 4).

As an artificial data source, the conditional probability of a pulse, given that a pulse has occurred, depends on the type of particle counter used; one is then thinking of "pulses" as meaning "particles counted." For example, *counters of Type I* have the property that the occurrence of a pulse at a given bit time inhibits the occurrence of the pulse for a specific number of bit times thereafter. This phenomenon is caused by a "dead time" in the counter.

The way pulses are produced in the generator is by word detection on words of length seven or less in a stream of "pseudo-random" digits. A most convenient source of such digit streams is the so-called *maximal-length shift-register generators* of Ref. 5. Such generators generate digits by a linear recurrence implemented by a shift register. The shift register that has been chosen here corresponds to the *primitive polynomial* $x^{36} + x^{11} + 1$ of Ref. 6, Chapter II. That is, the recursion is

$$a_{n+36} = a_{n+11} + a_n,$$

addition being mod 2. Since this polynomial is primitive, the period of any sequence generated by any initial condition (other than the *all-zero* state) is $2^{36} - 1$ digits (longer than 19 hr at the 1-Mc rate being used).

Now that the bit stream has been described, the methods of word detection used to define whether a pulse is to be produced can also be defined. There are three statistical options available with the pulse generator to define when a pulse occurs:

Option 1 requires the user to select a *window* of length $m = 7$ or less. A window is defined as tuple of 0's 1's, and x's (an x corresponding to logical "don't-care"). A pulse is produced whenever the window occurs in the sequence. But the user can *inhibit* the occurrence of a pulse for up to $l - 1 \leq 6$ bits after a pulse by setting the *inhibit number* l (as a binary number in three *cumulator capacity switches*). If one is interested in windows of length less than seven, he calls the extra positions "don't-cares." If the inhibit number equals the word length, the output process is called a *renewal process* (Ref. 7, Chapter IX), a special kind of output for counters of Type I.

Option 2 outputs a pulse when the cumulator has counted up to r 1's, $1 \leq r \leq 7$. The value of r is chosen by the same three cumulator capacity switches; there is no inhibition.

Option 3 allows windows to be chosen, but the search for a pulse is on consecutive non-overlapping blocks of words of length m, where m is again chosen by the cumulator capacity switches (with no inhibition). This feature is known as *strobing*.

Note that the three options agree when the length m is set equal to 1 (r set to 1 in Option 2), and the window 1 is chosen. For then every occurrence of a "1" produces a pulse, and every pulse is produced only in this way.

2. Statistics Under Option 1: Mean Number of Pulses, No "Don't-Cares"

Option 1 is by far the most interesting of the three options due to the fact that the inhibit time l can vary independently of the window length. We shall now demonstrate how to study the random variable representing the number of pulses in n bits for large n. If m is the length of the window and l the inhibit number, then the case $l = m$ is the known case of renewal processes. The case in which $l \neq m$ is, however, new (except that some results on the case $l = 1$ were obtained in Ref. 1).

We now show how to compute the mean and variance of the number of pulses in n bit times for large n, using techniques developed in Ref. 1 for the case $l = 1$ (no inhibitions). Let E_n be the random variable denoting the number of pulses in n bits. Define a set of n random variables Y_i such that $Y_i = 1$ or 0 depending on whether a pulse does or does not occur at bit i. Thus, one has the identity

$$E_n = \sum_{i=1}^{n} Y_i. \tag{1}$$

First, let us consider the case in which the window of length m has no "don't-care" positions. Define

$$u_i = \Pr(Y_i = 1) = \Pr(\text{pulse at bit } i).$$

Then, as in the case of renewal theory, one can prove that

$$u = \lim_{i \to \infty} u_i \tag{2}$$

exists. Thus, the mean number of pulses in n bits, e_n say, is given by the formula asymptotic in n:

$$e_n \sim nu. \tag{3}$$

We therefore must find the limiting probability of a pulse, u. First, assume that $l \leq m$. The probability (for $i \geq m$) that the window from $i - m + 1$ to i is the right window for a pulse is $1/2^m$. But there would not be a pulse at i if there was a pulse earlier in the window and the inhibit rule still held sway. This inhibiting pulse could have occurred at one of the $l - 1$ positions

$$i - 1, i - 2, \cdots, i - l + 1.$$

Furthermore, some windows make it impossible to have the right window for a pulse and yet have a pulse

earlier in the window. For the window can cause a disagreement between its last k places and its first k places, for every k, $1 \leq k \leq m-1$. Suppose, for example, that the window is 0010111. Then, when a pulse occurs, no pulse can occur until seven bits later, regardless of whether the inhibit number l is set to 7 or less than 7.

We therefore define $g(k)$, $1 \leq k \leq m$, to be 1 or 0 according to whether the first k bits of the window agree or disagree in at least one place with the last k bits of the window; $g(m) = 1$. Then, if the window occurs in positions $i - m + 1$ through i, there could have been a pulse at any position earlier in the window for which $g(k) = 1$, and the inhibit could be taking effect, preventing a pulse at the bit in question. Furthermore, these events are mutually exclusive. Consequently, we can write

$$\frac{1}{2^m} = \sum_{j=0}^{l-1} g(m-j)\, 2^{-j}\, u_j. \tag{4}$$

For if a pulse occurred at $m - j$ in the window,

$$0 \leq j \leq l-1,$$

then the next j positions must occur as in the window. This extra event has probability 2^{-j} and is independent of the event representing a pulse at $m - j$.

Using the fact that $u_i \to u$ as $i \to \infty$, we can write from Eq. (4)

$$\frac{1}{2^m} = \sum_{j=0}^{l-1} g(m-j) 2^{-j}\, u, \tag{5}$$

or finally

$$u = \left\{ \sum_{j=0}^{l-1} g(m-j) 2^{m-j} \right\}^{-1}, \tag{6}$$

the required expression for the limiting probability of a pulse.

For example, consider the window 1110111 with $m = 7$, $l = 5$. Then $g(7) = g(3) = g(2) = g(1) = 1$; all other g's are equal to 0. Eq. (6) becomes

$$u = \frac{1}{\sum_{j=0}^{4} g(7-j) 2^{7-j}} = \frac{1}{(2^7 + 2^{7-4})} = \frac{1}{136}.$$

In 10^6 bits, then, the average number of pulses is $10^6/136 = 7352.9$.

To treat the case in which the inhibit length l is, in fact, larger than the window length m, we proceed as follows: First, consider the renewal case $l = m$, and let $u_{(m)}$ be the limiting probability of a pulse so obtained. If now $l = m + t$, $t > 0$, write $u_{(m+t)}$ for the limiting probability of a pulse for this case. In n bits there are, for large n, close to $u_{(m+t)}n$ pulses. Each of these pulses has a "wastage" of t bits on which no pulse occurs, after which the process starts over as if it were a renewal process with no wastage. Thus, in n bits there are close to $nu_{(m+t)}t$ wasted bits. If these bits are removed from the n bits, there are $n - nu_{(m+t)}t$ bits left, and the process behaves as if it were an ordinary renewal process with the reduced number of bits. Hence, one can write

$$u_{(m)}\,(n - nu_{(m+t)}\, t) \sim nu_{(m+t)}. \tag{7}$$

Thus, one concludes

$$u_{(m)}\,(1 - u_{(m+t)}\, t) = u_{(m+t)}, \tag{8}$$

or

$$u_{(m+t)} = \frac{u_{(m)}}{(1 + u_{(m)}t)}, \tag{9}$$

the required expression. The same comment applies for the following discussion, when "don't-cares" are allowed.

3. Option 1, "Don't-Cares"

When "don't-cares" are allowed in Option 1, the procedure for finding the asymptotic probability of a pulse is more complicated. Thus, if there are m^* unconditioned positions in the window, one defines 2^{m^*} auxiliary variables $u^{\{\epsilon\}}$, where $\{\epsilon\}$ is an m^*-tuple of 0's and 1's. The meaning of $u^{\{\epsilon\}}$ is the limiting probability of a pulse which has, for its m^*-tuple of unconditioned positions, the binary vector $\{\epsilon\}$. One then has

$$u = \sum_{\{\epsilon\}} u^{\{\epsilon\}}, \tag{10}$$

and it is the 2^{m^*} $u^{\{\epsilon\}}$ that are actually found by the procedures of the preceding section. But then one obtains 2^{m^*} linear equations in the 2^{m^*} unknowns $u^{\{\epsilon\}}$.

Thus, suppose $m = 3$, $l = 2$, $m^* = 1$, and the window is $11x$, x denoting a "don't-care." There are then two variables $u^{(0)}$ and $u^{(1)}$, the probabilities of a 110 and a 111 pulse, respectively. If a 110 occurs, then a pulse of the

$u^{(1)}$ kind could have occurred at the first of the two equations:

$$\frac{1}{8} = u^{(0)} + \frac{1}{4}\,u^{(1)}. \tag{11}$$

Similarly, if 111 occurs, a pulse of the $u^{(1)}$ kind could have occurred at the first position, and one obtains

$$\frac{1}{8} = u^{(1)} + \frac{1}{4}\,u^{(1)}. \tag{12}$$

The solution of this pair of equations [Eqs. (11) and (12)] is $u^{(0)} = u^{(1)} = 1/10$, so $u = u^{(0)} + u^{(1)} = 1/5$.

One can regard this procedure as yielding a method for solving simultaneous linear equations in 2^{m^*} unknowns by a "Monte-Carlo" technique. The coefficients would have to be special to arise in this way, but a large class of equations is included (how large is not yet known). One would choose the proper m, m^*, and window, as well as l, to give the equations one wants, and would then let the sequence generator run for a large number n of bits, recording the number of pulses of both types. Since, as we shall see, the variances one obtains are reasonably small for large n, three-place accuracy is readily obtainable with, say, 10^6 bits, or 1 sec of "computation" at 1 Mc.

4. Burst Errors

In Ref. 4, the production of bursts of pulses was desired to simulate the noise in a burst-error channel for testing coding systems. Here we shall do this for the case of windows with no unconditioned positions. Furthermore, one sees that, for a burst (of length greater than 1) to occur, the inhibit length must be 1 (no inhibition). A burst is defined as a string of two or more *consecutive* pulses. (Note the difference in definition of a burst in the error-correction sense, Ref. 8, Chapter 10.) Since no "don't-cares" are allowed, the only windows that can produce bursts are the constant windows $00\cdots0$ or $11\cdots1$. The *length* of a burst is defined as the number of pulses in sequence, flanked by no pulses at both ends.

Thus, if the detector is to sense 111, a burst of length 1 corresponds to the occurrence of 01110; a burst of length 2 to 011110; and so on. One then uses the theory for these augmented windows to find the expected number of bursts of a given length in n bits.

Thus, let $w_{(t)}$ be the probability of a burst of length t, corresponding to the window $0\ 111\cdots1\ 0$ containing $m + t$ 1's. One finds, since $l = 1$, that

$$w_{(t)} = \frac{1}{2^{m+1+t}}. \tag{13}$$

Since when a burst of length t occurs, t pulses are made, we have the identity

$$u = \sum_{t=1}^{\infty} t w_{(t)}, \tag{14}$$

where u is the probability of a pulse. Now u corresponds to the window of m 1's, so $u = 1/2^m$. These considerations lead to the identity

$$\frac{1}{2^m} = \sum_{t=1}^{\infty} \frac{t}{2^{m+1+t}}, \tag{15}$$

or

$$\sum_{t=1}^{\infty} \frac{t}{2^{t+1}} = 1. \tag{16}$$

This identity can, of course, be derived by other means.

5. Variances Under Option 1

We saw in Ref. 1 that the asymptotic distribution of the number of pulses in n bits is asymptotically normal when n is large, at least for the cases considered in that *SPS*. The asymptotic normality result can be extended to cover all the cases under all three options considered in this article. Thus, it becomes of interest to find the variance of the number of pulses in n bits, since the variance, together with the mean, completely determine the asymptotic distribution. The case $l = 1$ was done in Ref. 1, and we modify the method to handle all of Option 1. The variances can be proved to be asymptotically proportional to n, as in Ref. 1. Thus, it is the coefficient of n in the asymptotic formula that is needed to obtain the asymptotic distribution of the number of pulses.

Rather than do the general case here, which is a slight extension of the discussion in Ref. 1, but more complicated (especially in the case of "don't-cares"), we shall instead do the particular case of the window 1110111

with $l = 5$. The pulse-probability u was found to be 1/136 in Section 2.

We had for the number of pulses E_n in Eq. (1):

$$E_n = \sum_{i=1}^{n} Y_i,$$

where $Y_i = 1$ or 0 according to whether a pulse does or does not occur at i. Then, if $\mathrm{Cov}(Y_i, Y_j)$ denotes the covariance between Y_i and Y_{i+k}, one has

$$\mathrm{Var}\ E_n = \sum_{i,j=1}^{n} \mathrm{Cov}(Y_i, Y_j). \tag{17}$$

Since $\mathrm{Cov}\ (Y_i, Y_j)$ just depends on $k = j - 1$, at least for $i, j \geq 7$, one defines $f(k)$ as $\mathrm{Cov}(Y_i, Y_{i+k})$ for $i \geq 7$ and finds, as in Ref. 1, that

$$\mathrm{Var}\ E_n \sim n \sum_{k=0}^{n} f(k). \tag{18}$$

One therefore wants

$$\sum_{k=0}^{n} f(k).$$

But, as in Ref. 1, Y_i and Y_{i+k} are, in fact, independent for $i, k \geq 7$, so $f(k) = 0$ for $i, k \geq 7$. Thus,

$$\mathrm{Var}\ E_n \sim n \sum_{k=0}^{6} f(k). \tag{19}$$

We therefore turn to the computation of the seven values of $f(k)$. First, $f(0)$ is just

$$\mathrm{Var}(Y_i) = E(Y_i^2) - E(Y_i)^2,$$

where E denotes expected value. Since $Y_i^2 = Y_i$,

$$f(0) = u - u^2 = \frac{135}{(136)^2}. \tag{20}$$

To compute the other six values of $f(k)$, note from the form of the word 1110111 that Y_i and Y_{i+k} cannot both be equal to 1 for $k = 1, 2, 3, 4$. Hence, $Y_i Y_{i+k} = 0$, $1 \leq k \leq 4$. So, since

$$\mathrm{Cov}(Y_i, Y_{i+k}) = E(Y_i Y_{i+k}) - E(Y_i)\, E(Y_{i+k}),$$

one concludes

$$f(k) = u^2 = \frac{1}{(136)^2},\ 1 \leq k \leq 4. \tag{21}$$

We now turn to $f(5)$ and $f(6)$. To find $E(Y_i Y_{i+k})$, use the fact that

$$\begin{aligned} E(Y_i Y_{i+k}) &= \Pr(Y_i = 1, Y_{i+k} = 1) \\ &= \Pr(Y_i = 1)\ \Pr(Y_{i+k} = 1 \mid Y_i = 1), \end{aligned} \tag{22}$$

where $\Pr(A|B)$ denotes the probability of A given B. Now if $Y_i = 1$ (that is, if there is a pulse at i), then the probability of a pulse at $i + 5$ is $1/2^5$, since the first two positions of the new window are picked up "free" from the given window. Similarly, the probability of a pulse at $i + 6$ is $1/2^6$, given that a pulse occurs at i. Thus,

$$E(Y_i Y_{i+5}) = \frac{u}{2^5};\ E(Y_i Y_{i+6}) = \frac{u}{2^6}. \tag{23}$$

That is,

$$\begin{aligned} f(5) &= \frac{u}{2^5} - u^2 \\ f(6) &= \frac{u}{2^6} - u^2 \end{aligned} \tag{24}$$

We finally have

$$\begin{aligned} \sum_{k=0}^{6} f(k) = \frac{135}{(136)^2} &- \frac{4}{(136)^2} + \frac{1}{(136)2^5} \\ &- \frac{1}{(136)^2} + \frac{1}{(136)2^6} - \frac{1}{(136)^2}, \end{aligned} \tag{25}$$

or

$$\mathrm{Var}\ E_n \sim 0.00732n. \tag{26}$$

The final conclusion is that the distribution of the number of pulses in n bits, for large n, is approximately normal with mean $e_n = n/136$, standard deviation $0.0856n^{1/2} = \sigma_n$ say. For $n = 10^6$, we have

$$e_n = 7352.9\cdots,\ \sigma_n = 85.6\cdots. \tag{27}$$

Thus, the ratio of standard deviation to mean is quite small.

Note that if we had a binomial distribution with probability of head $p = 1/136 = u$ (independent coin flips), the standard deviation would be slightly lower: $[np(1-p)]^{1/2} = 83.0\cdots$. But it is not merely a certain mean and variance that one wants, but also the actual properties of how the pulses are distributed in time. That is, the covariance function $f(k)$ of the stationary binary process is what one really wants to control in experimental applications.

6. Options 2 and 3

Options 2 and 3 are much easier to work out, and the results are quite well-known. Option 2, accumulation of r 1's can be looked at as follows: In n pulses, there are very likely to be $n/2$ 1's, hence $(n/2)/r = n/2r$ pulses. Thus, the limiting probability u of a pulse is given by $u = 1/2r$.

The variance of the number of pulses in n bits under Option 2 is also easy to obtain, since the number of pulses in n bits is $[w/r]$, where w is the number of 1's in n bits and $[\cdots]$ denotes the greatest integer function. Hence, the variance of the number of pulses is asymptotic to $(1/r^2)$ Var(w), that is, to $n/4r^2$.

Finally, Option 3 is nothing but the familiar binomial distribution, with probability of head p given as $p = 1/2^{m-m^*}$; m^* is, as before, the number of unconditioned positions in the window. However, we must remember to divide by the window length m, since the probability p of a pulse is defined above for a block of m bits. Thus, the probability of a pulse is finally $1/m2^{m^*}$ The variance of the number of pulses in n bits is

$$\left[\frac{n}{m}\right]\left(\frac{1}{2^{m^*}}\right)\left(1 - \frac{1}{2^{m^*}}\right). \tag{28}$$

7. An Experiment

An experiment was performed with the random pulse generator using Option 1 with window 1110111 and $l = 5$; n was 10^6. The quantity $s = 100$ samples of n bits each was taken, and the number of pulses for each sample of n recorded.

The mean number of pulses found was 7354.4; Section 2 gave the mean as $7352.9\cdots$. The standard deviation of the mean of a sample of 100 is $8.56 = s$, say. The variance of the 100 samples was found to be 5920; Section 5 gives the expected value of this variance as 7320.

The deviation of the sample mean from the actual mean was 1.5 or, in s-units, $0.164s$. Deviations this large or larger, according to the theory of the normal distribution, occur with probability .67. Thus, as far as the mean is concerned, the fit is excellent.

To test whether the observed sample variance could have arisen from sampling effects from the calculated population variance, we use a result from Ref. 9, Chapter 10: the variance of the sample variance computed from s samples is $2(s-1)/s^2$ times the square of the population variance. The standard deviation of the sample variance is thus 1030; the sample variance has mean 7320. The sample variance is also approximately normal. Using the same techniques as in the above paragraph, we find that deviations of the sample variance from the population variance as large or larger than the observed difference of 1400 occur with probability .47. The fit is still excellent, and we are indeed obtaining the statistics predicted by this article.

B. A Test of Independence and Estimation of the Correlation Coefficient Using Quantiles

I. Eisenberger

This article presents further results stemming from the investigation into the use of quantiles in data compression of space telemetry. Previous results of this investigation are given in Ref. 10; *SPS 37-25*, Vol. IV, pp. 194–198; and *SPS 37-27*, Vol. IV, pp. 229–234. In addition, a mechanization of a quantile system is considered in *SPS 37-27*, Vol. III, pp. 103–112.

1. Introduction

Given a set of n independent pairs of observations, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ taken from two normally distributed populations with known means and variances, one is interested in the answers to the following two questions:

(1) Can we assert that the set of observations $x = \{x_1, x_2, \cdots, x_n\}$ is independent of the set of observations $y = \{y_1, y_2, \cdots, y_n\}$?

(2) What can be said about the correlation between them, if any?

To answer the first question, the problem of testing the null hypothesis [$N(0, 1)$ is the class of random variables having the unit normal distribution]:

$$H_0\colon\ g_1(x) \in N(0,1),\quad g_2(y) \in N(0,1),\quad \rho = 0$$

against the alternative hypothesis:

$$H_1\colon\ g_1(x) \in N(0,1),\quad g_2(y) \in N(0,1),\quad \rho \neq 0,$$

will be considered. (Because of the assumption of known means and variances, we can, without loss of generality, assume standard normal distributions.)

To answer the second question, unbiased estimators of ρ are constructed. The power functions of the tests are derived, and the efficiencies of the tests are determined relative to the best test using the entire sample. The efficiencies of the estimators are also determined relative to the sample correlation coefficient for the case $\rho = 0$. One, two, and four pairs of quantiles are used for the tests and estimators; n is assumed to be large (≥ 200). An application is given to telemetering micrometeoroid measurements from spacecraft.

2. Review of Quantiles

To define a quantile, consider a sample of n independent values $x_1, x_2, \cdots, x_n$ taken from a distribution of a continuous type, with distribution function $G(x)$ and density function $g(x)$. The pth quantile, or the quantile of order p, of the distribution or population, denoted by ζ_p, is defined as the root of the equation $G(\zeta) = p$; that is,

$$p = \int_{-\infty}^{\zeta_p} dG(x) = \int_{-\infty}^{\zeta_p} g(x)\, dx.$$

The corresponding sample quantile z_p is defined as follows: If the sample values are arranged in nondecreasing order of magnitude

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

then $x_{(i)}$ is called the ith order statistic and

$$z_p = x_{([np]+1)},$$

where $[np]$ is the greatest integer $\leq np$.

If $g(x)$ is differentiable in some neighborhood of each quantile value considered, one knows (as in Ref. 10) that the joint distribution of any number of quantiles is asymptotically normal as $n \to \infty$ and that, asymptotically,

$$E(z_p) = \zeta_p,$$

$$\mathrm{Var}(z_p) = \frac{p(1-p)}{ng^2(\zeta_p)},$$

$$\rho_{12} = \left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)^{1/2},$$

where ρ_{12} is the correlation between z_{p_1} and z_{p_2}, $p_1 < p_2$.

3. The Test Using One Pair of Quantiles

Assume that the limiting distribution and moments of the quantiles hold. Denote by $F(x)$ and $f(x) = F'(x)$ the distribution function and density function, respectively, of the standard normal distribution; that is,

$$F(x) = \int_{-\infty}^{x} f(t)dt, \text{ where } f(x) = (2\pi)^{-1/2} \exp(-\tfrac{1}{2}x^2).$$

It is necessary at this point to form two new sets of values $\{u_i\}$ and $\{v_i\}$ from the sample values $\{x_i\}$ and $\{y_i\}$ by means of the linear transformations:

$$u_i = \frac{2^{1/2}}{2}(x_i + y_i),$$

$$v_i = \frac{2^{1/2}}{2}(-x_i + y_i).$$

It is easily verified that, under H_0,

$$E(u_i) = E(v_i) = 0,$$

$$\mathrm{Var}(u_i) = \mathrm{Var}(v_i) = 1,$$

$$E(u_i v_i) = 0,$$

and, under H_1,

$$E(u_i) = E(v_i) = 0,$$

$$\mathrm{Var}(u_i) = 1 + \rho,$$

$$\mathrm{Var}(v_i) = 1 - \rho,$$

$$E(u_i v_i) = 0,$$

so that the set of values $\{u_i\}$ is independent of the set of values $\{y_i\}$ *under both hypotheses*. All the tests and estimators will be based on the quantiles of the transformed sets of variables $\{u_i\}$ and $\{v_i\}$.

Now let z and z' denote the quantiles of order p of the $\{u_i\}$ and $\{v_i\}$, respectively, and let ζ denote the corresponding population quantile of the standard normal distribution. The test will be made on the statistic $w = z - z'$.

Under H_0,

$$E(z) = E(z') = \zeta,$$

$$\mathrm{Var}(z) = \mathrm{Var}(z') = a^2,$$

where

$$a^2 = \frac{F(\zeta)\,[1 - F(\zeta)]}{nf^2(\zeta)},$$

$$E(w) = 0,$$

$$\mathrm{Var}(w) = 2a^2.$$

Under H_1,

$$E(z) = \zeta(1+\rho)^{1/2},$$
$$\mathrm{Var}(z) = a^2(1+\rho)$$
$$E(z') = \zeta(1-\rho)^{1/2},$$
$$\mathrm{Var}(z') = a^2(1-\rho)$$
$$E(w) = \zeta\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right],$$
$$\mathrm{Var}(w) = a^2(1+\rho) + a^2(1-\rho) = 2a^2.$$

The best critical (or rejection) region is determined by the likelihood ratio inequality

$$\frac{L(w|H_0)}{L(w|H_1)} = \frac{\exp\left[-\frac{1}{2}\left(\frac{w^2}{2a^2}\right)\right]}{\exp\left[-\frac{1}{2}\left(\frac{\{w-\zeta[(1+\rho)^{1/2}-(1-\rho)^{1/2}]\}^2}{2a^2}\right)\right]} < c,$$

which, by taking logarithms and simplifying, reduces to

$$\begin{aligned} &w > k_1, \text{ for } \rho > 0, \\ &w < k_2, \text{ for } \rho < 0, \end{aligned} \tag{1}$$

as the regions providing the maximum power. Here, k_1 and k_2 are determined such that, given H_0, the probability of inequality (1) occurring is equal to ϵ, the significance level of the test.

The power of the test, P_0, is determined as follows: Under H_0, for $\rho > 0$,

$$\Pr(w < k_1) = F\left(\frac{k_1}{a2^{1/2}}\right) = F(b_1) = 1-\epsilon; \; k_1 = ab_1 2^{1/2}.$$

Under H_1,

$$\Pr(w < k_1) = F\left\{\frac{k_1 - \zeta\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right]}{a2^{1/2}}\right\}$$
$$= F\left\{b_1 - \frac{\zeta\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right]}{a2^{1/2}}\right\} = 1 - P_0. \tag{2}$$

P_0 will be maximized if the order of the quantiles z and z' is chosen to maximize

$$\frac{\zeta}{a} = \frac{\zeta f(\zeta)\, n^{1/2}}{[F(\zeta)]^{1/2}\,[1-F(\zeta)]^{1/2}}.$$

Setting equal to zero the derivative of this quantity with respect to ζ, one finds that the maximum occurs at $p = F(\zeta) = 0.9424$. For this value of p, $\zeta = 1.575$ and $a = 2.0193/n^{1/2}$. Inserting these values into Eq. (2), one obtains, as the optimum power function of the test using one pair of quantiles,

$$P_0 = 1 - F\left\{b_1 - 0.5515\, n^{1/2}\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right]\right\}.$$

Efficiency of the test. If one defines the efficiency of the test as the ratio P_0/P_0', where P_0' is the power of the best test using all the transformed values $\{u_i\}$ and $\{v_i\}$, it is of interest to determine P_0' in order to see how "good" the quantile tests are when compared to what can be achieved using all the values at our disposal. The likelihood ratio inequality for this case gives, as the best critical region,

$$w' = \sum_{i=1}^{n}\left[(\rho-1)\,u_i^2 + (\rho+1)\,v_i^2\right] \begin{cases} < k_3, \text{ for } \rho > 0 \\ > k_4, \text{ for } \rho < 0. \end{cases}$$

Thus, w' is asymptotically normally distributed, and under H_0

$$E(w') = 2n\rho,$$
$$\mathrm{Var}(w') = 4n(\rho^2+1),$$

while under H_1

$$E(w') = 0,$$
$$\mathrm{Var}(w') = 4n(1-\rho^2)^2.$$

Hence, assuming the limiting distribution of w' (which is a very good approximation for $n \geq 200$), one has under H_0, for $\rho > 0$,

$$\Pr(w' < k_3) = F\left\{\frac{k_3 - 2n\rho}{2\left[n(\rho^2+1)\right]^{1/2}}\right\} = F(b_2) = \epsilon,$$
$$k_3 = 2b_2\left[n(\rho^2+1)\right]^{1/2} + 2n\rho.$$

Under H_1,

$$\Pr(w' < k_3) = F\left[\frac{k_3}{2(1-\rho^2)\,n^{1/2}}\right]$$
$$= F\left\{\frac{1}{1-\rho^2}\left[b_2(\rho^2+1)^{1/2} + \rho n^{1/2}\right]\right\} = P_0'.$$

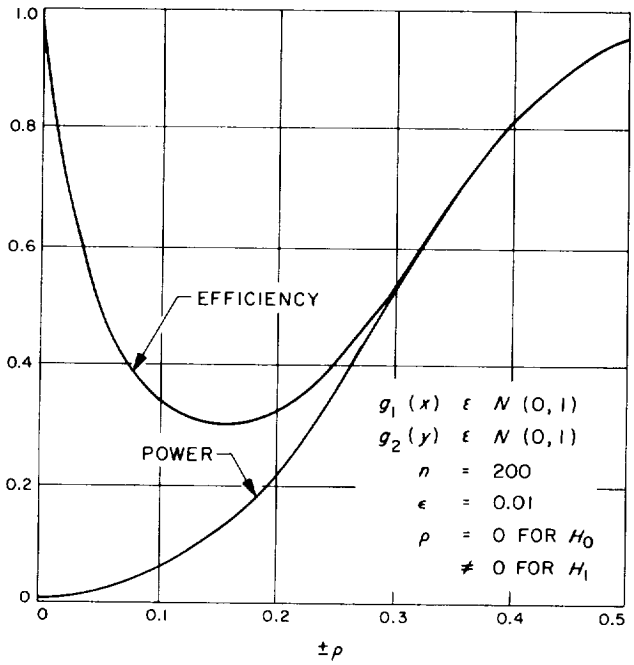


Fig. 1. Power and efficiency of test using one pair of quantiles

For $n = 200$, $\epsilon = 0.01$, Fig. 1 shows the power and efficiency of the test using one pair of optimum quantiles. The efficiency is never less than 0.30, approaches 1 as $\rho \to 0$, and approaches a number close to, but not equal to, 1 as $|\rho| \to 1$. Under these conditions, $b_1 = 2.326, b_2 = -2.326$, and $k_1 = 6.6424/n^{1/2} = 0.4697$.

Thus, for $\rho > 0$, if $w = z(0.9424) - z'(0.9424) < 0.4697$, *accept* H_0. Otherwise, reject H_0. The test will be made at a significance level of 0.01. If $\rho < 0$, $k_2 = -0.4697$, and hence *accept* H_0 if $w > -0.4697$. Otherwise, reject H_0.

4. The Test Using Two Pairs of Quantiles

Let z_1 and z_2 be the quantiles of $\{u_i\}$ of orders p_1 and $p_2 = 1 - p_1 > p_1$; z'_1 and z'_2, the corresponding quantiles of $\{v_i\}$; and ζ_1 and ζ_2, the corresponding *population* quantiles of the standard normal. Let $w_1 = z_1 - z'_1$ and $w_2 = z_2 - z'_2$. The test will be based on the values of w_1 and w_2.

Under H_0,

$$E(z_1) = E(z'_1) = -\zeta_2,$$

$$E(z_2) = E(z'_2) = \zeta_2,$$

$$\mathrm{Var}(z_1) = \mathrm{Var}(z'_1) = \mathrm{Var}(z_2) = \mathrm{Var}(z'_2) = \frac{F(\zeta_2)[1-F(\zeta_2)]}{nf^2(\zeta_2)} = a^2.$$

Hence, $E(w_1) = E(w_2) = 0$, $\mathrm{Var}(w_1) = \mathrm{Var}(w_2) = 2a^2$. Under H_1,

$$\begin{aligned}
E(z_1) &= -\zeta_2(1+\rho)^{1/2},\\
E(z'_1) &= -\zeta_2(1-\rho)^{1/2},\\
E(z_2) &= \zeta_2(1+\rho)^{1/2},\\
E(z'_2) &= \zeta_2(1-\rho)^{1/2},\\
\mathrm{Var}(z_1) &= \mathrm{Var}(z_2) = (1+\rho)a^2,\\
\mathrm{Var}(z'_1) &= \mathrm{Var}(z'_2) = (1-\rho)a^2,\\
E(w_1) &= -\zeta_2[(1+\rho)^{1/2} - (1-\rho)^{1/2}],\\
E(w_2) &= \zeta_2[(1+\rho)^{1/2} - (1-\rho)^{1/2}],\\
\mathrm{Var}(w_1) &= \mathrm{Var}(w_2) = 2a^2.
\end{aligned}$$

Simplifying the likelihood ratio inequality, for which $L(w_1,w_2\,H_0)$ and $L(w_1,w_2\,H_1)$ are the joint distributions of w_1 and w_2, given H_0 and H_1, respectively, results in the best critical region:

$$S = (w_2 - w_1) \begin{cases} > k_5, \text{ for } \rho > 0 \\ < k_6, \text{ for } \rho < 0. \end{cases}$$

Under H_0,

$$E(S) = 0,$$

$$\mathrm{Var}(S) = 4a^2(1 - \rho_{12}),$$

where ρ_{12} denotes the correlation between z_1 and z_2 (z'_1 and z'_2). For $\rho > 0$,

$$\Pr(S < k_5) = F\left[\frac{k_5}{2a(1-\rho_{12})^{1/2}}\right] = F(b_3) = 1 - \epsilon,$$

$$k_5 = 2ab_3(1-\rho_{12})^{1/2}.$$

Under H_1,

$$E(S) = 2\zeta_2[(1+\rho)^{1/2} - (1-\rho)^{1/2}],$$

$$\mathrm{Var}(S) = 4a^2(1-\rho_{12}),$$

$$\begin{aligned}
\Pr(S < k_5) &= F\left\{\frac{k_5 - 2\zeta_2[(1+\rho)^{1/2} - (1-\rho)^{1/2}]}{2a(1-\rho_{12})^{1/2}}\right\}\\
&= F\left\{b_3 - \frac{\zeta_2[(1+\rho)^{1/2} - (1-\rho)^{1/2}]}{a(1-\rho_{12})^{1/2}}\right\}\\
&= 1 - P_0.
\end{aligned}$$

P_0 is maximized by determining the value of p_2 which maximizes

$$\frac{\zeta_2}{a(1-\rho_{12})^{1/2}} = \frac{\zeta_2 f(\zeta_2)}{\{F(\zeta_2)\,[1-2F(\zeta_2)]\}^{1/2}}\,.$$

This maximum occurs at $p_2 = 0.9310$. For this value of p_2, $\zeta_1 = -1.483$, $\zeta_2 = 1.483$, $a = 1.9085/n^{1/2}$, $\rho_{12} = 0.07416$, and the optimum power function is given by

$$P_0 = 1 - F\left\{b_3 - 0.8076\, n^{1/2}\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right]\right\}.$$

For $n = 200$, $\epsilon = 0.01$, Fig. 2 shows the power and efficiency of the test using two pairs of optimum symmetric quantiles. For this case, $b_3 = 2.326$, $k_5 = 0.6041$, and hence for $\rho > 0$, if $S = z\,(0.9310) - z'\,(0.9310) - z\,(0.0690) + z'\,(0.0690) < 0.6041$, accept H_0. Otherwise, reject H_0. If $\rho < 0$, $k_6 = -0.6041$, and so if $S > -0.6041$ accept H_0. Otherwise, reject H_0. The efficiency never drops below 0.66.

5. The Test Using Four Pairs of Quantiles

For this case, let z_i and z_i', $i = 1, 2, 3, 4$, be four quantiles of $\{u_i\}$ and $\{v_i\}$, respectively, such that $p_1 < p_2 < p_3 < p_4$ and $p_1 + p_4 = p_2 + p_3 = 1$. Let ζ_i denote the corresponding population quantiles of the standard normal,

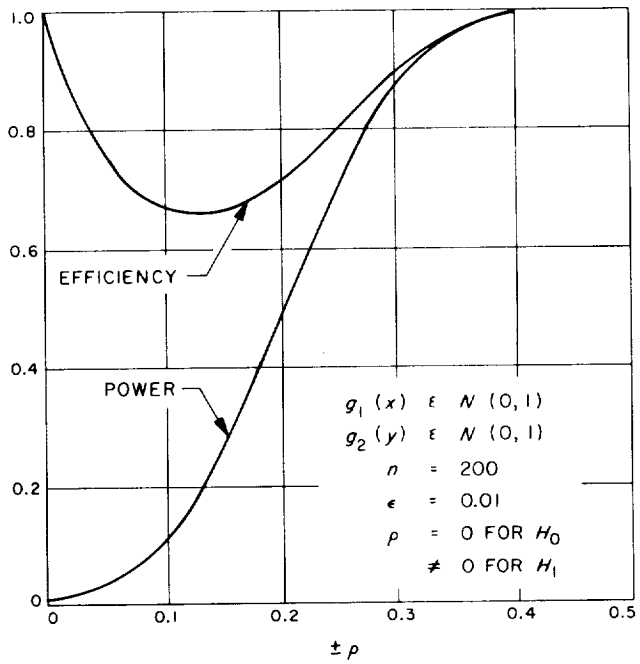


Fig. 2. Power and efficiency of test using two pairs of quantiles

$w_i = z_i - z_i'$, and let $S_1 = w_4 - w_1$ and $S_2 = w_3 - w_2$. The test will be based on the linear combination

$$t = \alpha S_1 + \beta S_2\,,$$

determining α and β as well as the orders of the quantiles so as to maximize P_0. Under H_0,

$$E(S_1) = E(S_2) = 0,$$

$$\mathrm{Var}(S_1) = 4a_4^2\,(1-\rho_{14}),$$

$$\mathrm{Var}(S_2) = 4a_3^2\,(1-\rho_{23}),$$

$$E(t) = 0,$$

$$\mathrm{Var}(t) = 4\{\alpha^2 a_4^2\,(1-\rho_{14}) + \beta^2 a_3^2\,(1-\rho_{23}) + 2\alpha\beta a_3 a_4(\rho_{12}-\rho_{13})\} = 4\gamma^2,$$

where

$$a_i^2 = \frac{F(\zeta_i)[1-F(\zeta_i)]}{nf^2(\zeta_i)},\quad i = 3, 4,$$

and ρ_{ij} denotes the correlation between z_i and z_j, as well as that between z_i' and z_j'. Under H_1,

$$E(S_1) = 2\zeta_4\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right],$$

$$\mathrm{Var}(S_1) = 4a_4^2\,(1-\rho_{14}),$$

$$E(S_2) = 2\zeta_3\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right],$$

$$\mathrm{Var}(S_2) = 4a_3^2\,(1-\rho_{23}),$$

$$E(t) = 2\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right](\alpha\zeta_4 + \beta\zeta_3),$$

$$\mathrm{Var}(t) = 4\gamma^2.$$

Omitting the details, one obtains as the critical region

$$t\begin{cases} > k_7 \text{ for } \rho > 0 \\ < k_8 \text{ for } \rho < 0 \end{cases}$$

The power function, for $\rho > 0$, is given by

$$F\left\{b_4 - \frac{\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right](\alpha\zeta_4+\beta\zeta_3)}{\gamma}\right\} = 1 - P_0,$$

where $F(b_4) = 1 - \epsilon$ and $k_7 = 2\gamma b_4$.

From previous investigations of the use of quantiles in estimating the parameters of normal distribution (Ref. 10), it is known that, for

$$\alpha = 0.116, \qquad \beta = 0.236,$$
$$p_1 = 0.0230, \qquad p_2 = 0.1269,$$
$$p_3 = 0.8731, \qquad p_4 = 0.9770,$$

the quantity $(\alpha\zeta_4 + \beta\zeta_3)/\gamma$ (and hence also P_0) will be a maximum. Thus, the optimum power function is given by

$$P_0 = 1-F\left\{b_4 - 0.9080\, n^{1/2}\left[(1+\rho)^{1/2} - (1-\rho)^{1/2}\right]\right\} .$$

For $n = 200$, $\epsilon = 0.01$, Fig. 3 shows the power and efficiency of the test using four pairs of optimum symmetric quantiles. For this case, $b_4 = 2.326$ and $k_7 = 0.1814$.

Hence, for $\rho > 0$, if $t = 0.116[z(0.9770) - z'(0.9770) - z(0.0230) + z'(0.0230)] + 0.236[z(0.8731) - z'(0.8731) - z(0.1269) + z'(0.1269)] < 0.1814$, accept H_0.

If $\rho < 0$, $k_8 = -0.1814$, so if $t > -0.1814$, accept H_0. Otherwise, reject H_0. The efficiency for this case never drops below 0.84.

6. An Unbiased Estimator of ρ Using One Pair of Quantiles

With respect to the set of pairs of sample values (x_1,y_1), $\cdots$, (x_n,y_n), the sets $\{x_i\}$ and $\{y_i\}$ are sample values of the random variables x and y with a joint distribution given by

$$g_3(x,y) = \frac{1}{2\pi(1-\rho^2)^{1/2}}\exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right].$$

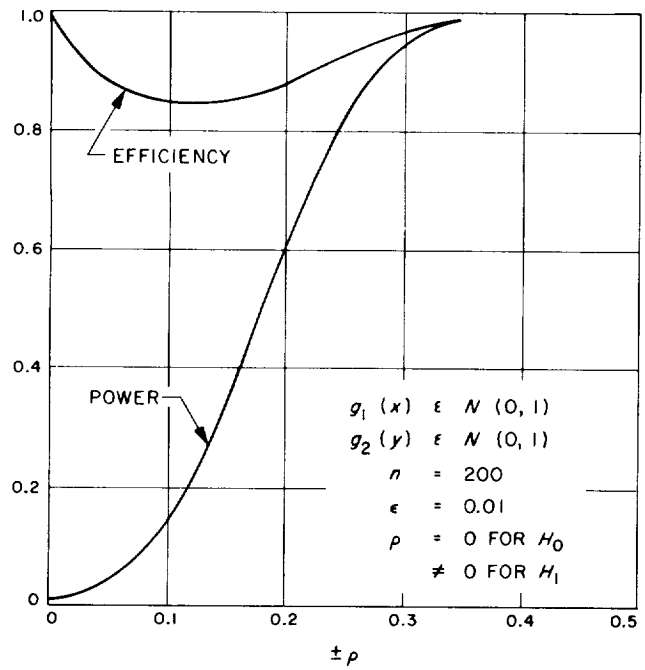


Fig. 3. Power and efficiency of test using four pairs of quantiles

The problem considered here is that of estimating ρ. By means of the linear transformation given above, a new set of pairs of sample values (u_1,v_1), $\cdots$, (u_n,v_n) is generated for which the sets $\{u_i\}$ and $\{v_i\}$ can be considered as sample values of the random variables u and v with marginal distributions

$$h_1(u) = \frac{1}{[2\pi(1+\rho)]^{1/2}}\exp\left[-\frac{u^2}{2(1+\rho)}\right] ,$$

$$h_2(v) = \frac{1}{[2\pi(1-\rho)]^{1/2}}\exp\left[-\frac{v^2}{2(1-\rho)}\right] ,$$

and joint distribution

$$g_4(u,v) = h_1(u)\, h_2(v) .$$

Unbiased estimators of ρ will be constructed using quantiles of $\{u_i\}$ and $\{v_i\}$, and the efficiencies of these estimators will be determined relative to the sample correlation r, the minimum-variance unbiased estimator of ρ, given by

$$r = \frac{\sum_{i=1}^{n}(u-\bar{u})(v-\bar{v})}{\left[\sum_{i=1}^{n}(u_i-\bar{u})^2\sum_{i=1}^{n}(v_i-\bar{v})^2\right]^{1/2}},$$

where

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n}u_i ,$$

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n}v_i ,$$

for the special case $\rho = 0$. The efficiency will be defined as in Ref. 11:

$$\mathrm{Eff}(\hat{\rho}) = \frac{\mathrm{Var}(r \mid \rho = 0)}{\mathrm{Var}(\hat{\rho} \mid \rho = 0)} = \frac{1}{(n-1)\,\mathrm{Var}(\hat{\rho} \mid \rho = 0)}.$$

Hence, as in Section 3 above, let z and z' denote the quantiles of order p of the $\{u_i\}$ and $\{v_i\}$, respectively, and let ζ denote the corresponding population quantile of the standard normal. Then, an unbiased estimator of ρ in terms of z and z' is given by

$$\hat{\rho} = \frac{z^2 - (z')^2}{2(a^2 + \zeta^2)},$$

where

$$a^2 = \frac{F(\zeta)\,[1 - F(\zeta)]}{nf^2(\zeta)}.$$

Since

$$E(z) = \zeta(1+\rho)^{1/2} ,$$

$$E(z') = \zeta(1-\rho)^{1/2} ,$$

$$\mathrm{Var}(z) = a^2(1+\rho) ,$$

$$\mathrm{Var}(z') = a^2(1-\rho) ,$$

one has

$$E(\hat{\rho}) = \frac{1}{2(a^2+\zeta^2)} \times \left[a^2(1+\rho) - a^2(1-\rho) + \zeta^2(1+\rho) - \zeta^2(1-\rho)\right] = \rho .$$

Noting the fact that if x is distributed $N(m, \sigma)$ (normal, mean m, variance σ^2), one has

$$\mathrm{Var}(x^2) = 2\sigma^4 + 4m^2\sigma^2 ,$$

and, hence,

$$\mathrm{Var}(\hat{\rho}) = \frac{1}{4(a^2+\zeta^2)^2}\left\{\mathrm{Var}(z^2) + \mathrm{Var}\left[(z')^2\right]\right\}$$

$$= \frac{1}{4(a^2+\zeta^2)^2}\left[2a^4(1+\rho)^2 + 4\zeta^2a^2(1+\rho)^2 + 2a^4(1-\rho)^2 + 4\zeta^2a^2(1-\rho)^2\right]$$

$$= \frac{a^2(a^2+2\zeta^2)(1+\rho^2)}{(a^2+\zeta^2)^2} . \qquad (3)$$

For $\rho = 0$, Eq. (3) can be written as

$$\mathrm{Var}(\hat{\rho} \mid \rho = 0) = \frac{a^4 + 2\zeta^2a^2}{a^4 + 2a^2\zeta^2 + \zeta^4} \cong \frac{1}{1 + (\zeta^2/2a^2)} , \qquad (4)$$

if one neglects the a^4 term in the numerator and denominator of Eq. (4) (this term is small compared to ζ^2a^2 and ζ^4 for large n). Now the approximate value of $\mathrm{Var}(\hat{\rho} \mid \rho = 0)$ in Eq. (4) is minimized if one chooses the order of z and z' to maximize ζ^2/a^2. In Section 1 above, it was found that $p = 0.9424$ will maximize ζ/a and will, of course, also maximize ζ^2/a^2. Moreover, since the curve defined by $\mathrm{Var}(\hat{\rho} \mid \rho = 0)$ is quite flat around its minimum, the error involved in using $p = 0.9424$ instead of the true value of p which minimizes $\mathrm{Var}(\rho \mid \rho = 0)$ is small. Thus, using $p = 0.9424$, one obtains

$$\hat{\rho} = \frac{z^2 - (z')^2}{\frac{8.1552}{n} + 4.9612} ,$$

$$\mathrm{Var}(\hat{\rho} \mid \rho = 0) = \frac{\frac{16.627}{n^2} + \frac{20.230}{n}}{\frac{16.627}{n^2} + \frac{20.230}{n} + 6.1535} \cong \frac{1}{1 + 0.3042n} .$$

For $n = 200$,

$$\hat{\rho} = 0.200\left[z^2 - (z')^2\right],$$

$$\mathrm{Var}(\hat{\rho} \mid \rho = 0) = 0.01624,$$

$$\frac{1}{1 + 0.3042n} = 0.01617,$$

$$\mathrm{Eff}(\hat{\rho}) = 0.3094.$$

7. An Unbiased Estimator of ρ Using Two Pairs of Quantiles

Let z_1, z_1', z_2, z_2', ζ_1, and ζ_2 be defined as in Section 4. Then an estimator of ρ using two pairs of symmetric transformed sample quantiles is given by

$$\hat{\rho} = \frac{z_2^2 - (z_2')^2 + z_1^2 - (z_1')^2}{4(a^2+\zeta_2^2)} ,$$

where

$$a^2 = \frac{F(\zeta_2)\left[1 - F(\zeta_2)\right]}{nf^2(\zeta_2)} .$$

Then,

$$E(\hat{\rho}) = \frac{1}{4(a^2+\zeta_2^2)}\left[2\zeta_2^2(1+\rho) + 2a^2(1+\rho) - 2\zeta_2^2 \times (1-\rho) - 2a^2(1-\rho)\right] = \rho .$$

Noting that if x and y are distributed $N(m_1, \sigma_1)$ and $N(m_2, \sigma_2)$, respectively, one has

$$E(x^2y^2) = m_1^2m_2^2 + \sigma_1^2m_2^2 + \sigma_2^2m_1^2 + 4\rho_{xy}\sigma_1\sigma_2m_1m_2 + \sigma_1^2\sigma_2^2(1 + 2\rho_{xy}^2) ,$$

and, hence,

$$\begin{aligned}\mathrm{Var}(\hat{\rho}) &= \frac{1}{16(a^2+\zeta_2^2)^2} \\ &\quad \times \left\{\mathrm{Var}(z_2^2+z_1^2)+\mathrm{Var}\left[(z_2')^2+(z_1')^2\right]\right\} \\ &= \frac{1}{16(a^2+\zeta_2^2)^2} \\ &\quad \times \left[4a^4(1+\rho_{12}^2)+8a^2\zeta_2^2(1-\rho_{12})\right] \\ &\quad \times \left[(1+\rho)^2+(1-\rho)^2\right] \\ &= \frac{a^2\left[a^2(1+\rho_{12}^2)+2\zeta_2^2(1-\rho_{12})\right](1+\rho^2)}{2(a^2+\zeta_2^2)^2}, \end{aligned} \tag{5}$$

where ρ_{12} denotes the correlation between z_1 and z_2.

For $\rho = 0$, Eq. (5) can be written as

$$\begin{aligned}\mathrm{Var}(\hat{\rho}\mid \rho=0) &= \frac{a^4(1+\rho_{12}^2)+2a^2\zeta_2^2(1-\rho_{12})}{2a^4+4a^2\zeta_2^2+2\zeta_2^4} \\ &\cong \frac{1}{2+\dfrac{\zeta_2^2}{a^2(1-\rho_{12})}}, \end{aligned} \tag{6}$$

neglecting the a^4 term in the numerator and denominator of Eq. (6) and taking $1-\rho_{12} \cong 1$. The approximate value of $\mathrm{Var}(\hat{\rho}\mid \rho=0)$ in Eq. (6) is minimized by maximizing $\zeta_2^2/a^2(1-\rho_{12})$. This maximum occurs, as found in Section 4, at $p_2 = 0.9310$. Using this value of p_2, one obtains

$$\hat{\rho} = \frac{z_2^2-(z_2')^2+z_1^2-(z_1')^2}{\dfrac{14.5694}{n}+8.7972},$$

$$\mathrm{Var}(\hat{\rho}\mid \rho=0) = \frac{\dfrac{13.3397}{n^2}+\dfrac{14.8330}{n}}{\dfrac{26.5334}{n^2}+\dfrac{32.042}{n}+9.6738} \cong \frac{1}{2+0.6522n}.$$

For $n = 200$,

$$\hat{\rho} = 0.1127\left[z_2^2-(z_2')^2+z_1^2-(z_1')^2\right],$$

$$\mathrm{Var}(\hat{\rho}\mid \rho=0) = 0.007575,$$

$$\frac{1}{2+0.6522n} = 0.007551,$$

$$\mathrm{Eff}(\hat{\rho}) = 0.6655.$$

8. An Unbiased Estimator of ρ Using Four Pairs of Quantiles

Let z_i, z_i', and ζ_i', $i = 1, 2, 3, 4$, be defined as in Section 5. Then an unbiased estimator of ρ using a linear combination of four pairs of symmetric transformed quantiles is given by

$$\hat{\rho} = \frac{\alpha\left[z_4^2-(z_4')^2+z_1^2-(z_1')^2\right]}{4\left[\alpha(a_4^2+\zeta_4^2)+\beta(a_3^2+\zeta_3^2)\right]} + \frac{\beta\left[z_3^2-(z_3')^2+z_2^2+(z_2')^2\right]}{4\left[\alpha(a_4^2+\zeta_4^2)+\beta(a_3^2+\zeta_3^2)\right]}$$

where

$$a_i^2 = \frac{F(\zeta_i)\left[1-F(\zeta_i)\right]}{nf^2(\zeta_i)}, \quad i = 3, 4.$$

Then,

$$\begin{aligned} E(\hat{\rho}) &= \frac{1}{4\left[\alpha(a_4^2+\zeta_4^2)+\beta(a_3^2+\zeta_3^2)\right]} \\ &\times \left\{2\alpha\left[\zeta_4^2(1+\rho)+a_4^2(1+\rho)-\zeta_4^2(1-\rho)-a_4^2(1-\rho)\right]\right. \\ &\left. +2\beta\left[\zeta_3^2(1+\rho)+a_3^2(1+\rho)-\zeta_3^2(1-\rho)-a_3^2(1-\rho)\right]\right\} \\ &= \rho, \end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\hat{\rho}) &= \frac{1}{16\left[\alpha(a_4^2+\zeta_4^2)+\beta(a_3^2+\zeta_3^2)\right]^2} \\ &\quad \times \left(\alpha^2\left\{\mathrm{Var}(z_4^2+z_1^2)+\mathrm{Var}\left[(z_4')^2+(z_1')^2\right]\right\}\right. \\ &\quad +\beta^2\left\{\mathrm{Var}(z_3^2+z_2^2)+\mathrm{Var}\left[(z_3')^2+(z_2')^2\right]\right\} \\ &\quad +2\alpha\beta\left\{\mathrm{Cov}(z_4^2+z_1^2, z_3^2+z_2^2)\right. \\ &\quad \left.\left.+\mathrm{Cov}\left[(z_4')^2+(z_1')^2, (z_3')^2+(z_2')^2\right]\right\}\right) \\ &= \frac{1}{2\left[\alpha(a_4^2+\zeta_4^2)+\beta(a_3^2+\zeta_3^2)\right]^2} \\ &\quad \times \left\{\alpha^2\left[a_4^4(1+\rho_{14}^2)+2a_4^2\zeta_4^2(1-\rho_{14})\right]\right. \\ &\quad +\beta^2\left[a_3^4(1+\rho_{23}^2)+2a_3^2\zeta_3^2(1-\rho_{23})\right] \\ &\quad +2\alpha\beta\left[2a_3a_4\zeta_3\zeta_4(\rho_{12}-\rho_{13})\right. \\ &\quad \left.\left.+a_3^2a_4^2(\rho_{12}^2+\rho_{13}^2)\right]\right\}(1+\rho^2), \end{aligned} \tag{7}$$

where ρ_{ij} denotes the correlation between z_i and z_j. As in the previous cases, we will use the values of the parameters as given in Section 5. These are

$$\alpha = 0.116\,, \qquad \beta = 0.236\,,$$

$$p_1 = 0.0230\,, \qquad p_2 = 0.1269\,,$$

$$p_3 = 0.8731\,, \qquad p_4 = 0.9770\,.$$

For these values

$$\hat{\rho} = \frac{0.116\left[z_4^2 - (z_4')^2 + z_1^2 - (z_1')^2\right]}{\frac{5.930}{n} + 3.0754} + \frac{0.236\left[z_3^2 - (z_3')^2 + z_2^2 - (z_2')^2\right]}{\frac{5.930}{n} + 3.0754}$$

and, for $n = 200$,

$$\hat{\rho} = 0.0374\left[z_4^2 - (z_4')^2 + z_1^2 - (z_1')^2\right] + 0.0760\left[z_3^2 - (z_3')^2 + z_2^2 - (z_2')^2\right],$$

$$\mathrm{Var}(\hat{\rho}\,|\,\rho = 0) = 0.006286\,,$$

$$\mathrm{Eff}(\hat{\rho}) = 0.7994\,.$$

This efficiency is quite high and makes this method of data compression very attractive.

It is of interest to compare the present method of estimating ρ from quantiles to a method proposed by F. Mosteller (Ref. 12), using bivariate order statistics. The technique he uses is to construct lines $y = 0$ and $x = \pm k$, which cut the xy-plane into six parts. The estimate of ρ is based on the number of pairs of observations falling in the four corners. Briefly, let

n_1 = the number of pairs of observations (x_i, y_i) such that $(x_i > k, y_i > 0)$,

n_2 = the number of pairs of observations (x_i, y_i) such that $(x_i < -k, y_i > 0)$,

n_3 = the number of pairs of observations (x_i, y_i) such that $(x_i < -k, y_i < 0)$,

n_4 = the number of pairs of observations (x_i, y_i) such that $(x_i > k, y_i < 0)$,

$$S_1 = \int_0^\infty \int_k^\infty g_3(x,y)\,dx\,dy\,, \quad \lambda = \int_k^\infty f(x)\,dx\,.$$

The maximum likelihood estimator $\hat{\rho}$ of ρ based on the n_i is then found by solving, for $\hat{\rho}$, the equation

$$\frac{n_1 + n_3}{\sum_{i=1}^{4} n_i} = \left(\frac{S_1}{\lambda}\right)_{\rho=\hat{\rho}}\,.$$

The $\mathrm{Var}(\hat{\rho}|\rho=0)$ is a minimum for $k = 0.6121$ and is given by $1.939/n$. Thus,

$$\mathrm{Eff}(\hat{\rho}) = \frac{n}{(n-1)(1.939)} = 0.5183\,, \text{ for } n = 200,$$

which is greater than the efficiency of $\hat{\rho}$ using one pair of quantiles, but less than the efficiency of the estimators using more than one pair, as we have seen.

9. Application

An application of this method of data compression will here be given to the problem of determining the correlation coefficient between micrometeoroid intensities measured at two different places on a deep-space probe. It can be assumed that, from previous flights, the mean and variance of the number of counts per second is known before the flight. What is desired is a measure of the correlation coefficient between the counts per second at two different positions and orientations of the counters on the spacecraft, for this correlation gives a measure of the direction from which the micrometeoroids are arriving.

The spacecraft would require a Quantiler (*SPS 37-27*, Vol. III, pp. 103–112) to do this data compression, but it may very well have a Quantiler onboard anyway for compression of other data. The extra equipment needed would merely serve to form the linear combinations of the x_i and y_i to get the u_i and v_i. This operation is easy to perform. As is usual with the use of quantiles (Ref. 13), data compression ratios on the order of 100:1 are obtainable with the use of four quantiles, with no loss of statistical efficiency. And, as usual, the quantile method requires less equipment than would be necessary to compute the sample correlation coefficient onboard the spacecraft.

C. Two-Sample Tests Using Quantiles

I. Eisenberger

This article continues recent JPL work on the use of quantiles for data compression of space telemetry. The problem considered here is that of discriminating between two hypotheses about a parent normal distribution. We consider testing for equality of two independent normal populations by the use of quantiles, where one of the two parameters of the normal distributions is given to be equal in both populations. Optimal choices of quantiles are found, and the efficiency of the quantile tests relative to the optimum test using all the samples is shown to be quite high.

1. Introduction

The use of quantiles to effect data compression of space telemetry was considered previously (Ref. 14; *SPS 37-25*, Vol. IV, pp. 194–198; and *SPS 37-27*, Vol. IV, pp. 229–234). A mechanization of a quantile system was discussed in *SPS 37-27*, Vol. III, pp. 103–112. In all previous cases, a set of sample values taken from a single normally distributed population was given, and methods for extracting certain kinds of statistical information using sample quantiles were derived. For example, in Ref. 14 the problem of estimating the mean and standard deviation of the population was studied, and, in addition, goodness-of-fit tests were considered. In *SPS 37-25*, Vol. IV, quantile tests were given for discriminating between the means of normal distributions with the same known variance; in *SPS 37-27*, Vol. IV, quantile tests were given for discriminating between the variances of normal distributions with the same known mean. The definitions in the theory of quantiles were summarized in the above *SPS*.

In this report, however, it is assumed that we are given sets of independent samples taken from *two* independent normally distributed populations with density functions $g_1(x;\ \mu_1,\ \sigma_1)$ and $g_2(y;\ \mu_2,\ \sigma_2)$. We consider the following two problems [$N(\mu,\ \sigma)$ is the class of random variables having the normal distribution of mean μ, variance σ^2]:

(1) If $\sigma = \sigma_1 = \sigma_2$ is known and μ_1 is *not* known, is $\mu_2 = \mu_1$ or $\mu_1 + \theta, \theta \neq 0$?

(2) If μ_1 and μ_2 are known and σ_1 is *not* known, is $\sigma_2 = \sigma_1$ or $\theta\sigma_1, \theta > 0$?

More formally, we consider the following two tests:

$$\text{(Test A)}\quad H_0:\ g_1(x) \in N(\mu,\ \sigma),\ g_2(y) \in N(\mu,\ \sigma),$$
$$H_1:\ g_1(x) \in N(\mu,\ \sigma),\ g_2(y) \in N(\mu + \theta,\ \sigma),\ \theta \neq 0,$$

where σ is known and μ is not known;

$$\text{(Test B)}\quad H_0:\ g_1(x) \in N(\mu_1,\ \sigma),\ g_2(y) \in N(\mu_2,\ \sigma),$$
$$H_1:\ g_1(x) \in N(\mu_1,\ \sigma),\ g_2(y) \in N(\mu_2,\ \theta\sigma),\ \theta > 0,$$

where μ_1 and μ_2 are known and σ is not known. Since μ_1 and μ_2 are assumed known in Test B, without loss of generality we can put $\mu_1 = \mu_2 = 0$, so that one has

$$\text{(Test B')}\quad H_0:\ g_1(x) \in N(0,\ \sigma),\ g_2(y) \in N(0,\ \sigma),$$
$$H_1:\ g_1(x) \in N(0,\ \sigma),\ g_2(y) \in N(0,\ \theta\sigma),\ \theta > 0,$$

where σ is not known. For both Tests A and B′, tests using one, two, and four pairs of quantiles, one of the pair from each of the two distributions, are given. Sample sizes of n_1 and n_2 are assumed, where n_1 and n_2 are large (≥ 200). In each case, the power function is derived and the efficiency determined.

2. Test A Using One Pair of Quantiles

Assume one has the limiting distribution and moments of the quantiles, and denote by $F(x)$ and $f(x) = F'(x)$ the distribution function and density function, respectively, of the standard normal distribution; that is,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt\,,$$

where

$$f(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\tfrac{1}{2}x^2\right).$$

Now let z be the sample quantile of order p (p the same for both distributions) of the samples taken from the first population, that with density function $g_1(x)$, and let z' be the corresponding sample quantile from the second population. Furthermore, let ζ be the corresponding population quantile of the standard normal. Then one has under H_0:

$$E(z) = E(z') = \sigma\zeta + \mu,$$

$$\text{Var}\,(z_1) = \frac{\sigma^2 a^2}{n_1},\quad \text{Var}\,(z_2) = \frac{\sigma^2 a^2}{n_2}\,,$$

where

$$a^2 = \frac{F(\zeta)\,[1-F(\zeta)]}{f^2(\zeta)},$$

and under H_1:

$$E(z) = \sigma\zeta + \mu,$$

$$E(z') = \sigma\zeta + \mu + \theta,$$

$$\text{Var}\,(z) = \frac{\sigma^2 a^2}{n_1},$$

$$\text{Var}\,(z') = \frac{\sigma^2 a^2}{n_2}.$$

Since μ is unknown, the distribution of the test statistic cannot depend on μ. Hence, the test statistic to be used should be given by

$$w = z - z'.$$

Under H_0:

$$E(w) = 0,$$

$$\text{Var}\,(w) = \sigma^2 a^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \sigma_w^2,$$

and under H_1:

$$E(w) = -\theta,$$

$$\text{Var}\,(w) = \sigma_w^2.$$

The best critical region (or rejection region) is determined by the likelihood ratio inequality

$$\frac{L(w\,|\,H_0)}{L(w\,|\,H_1)} = \frac{\exp\left[-\frac{1}{2}\left(\frac{w^2}{\sigma_w^2}\right)\right]}{\exp\left\{-\frac{1}{2}\left[\frac{(w+\theta)^2}{\sigma_w^2}\right]\right\}} < c, \tag{1}$$

where c is determined so that, under H_0, the probability of inequality (1) occurring is equal to ϵ, the *significance level* of the test. By taking logarithms and simplifying, the critical region of significance level ϵ providing the maximum power for a given order of quantile is the region in which

$$w < k_1, \theta > 0,$$

$$w > k_2, \theta < 0.$$

The value of k_1 and the power of the test, P_0, are determined as follows, assuming $\theta > 0$: Under H_0,

$$\Pr(w < k_1) = F\left(\frac{k_1}{\sigma_w}\right) = F(b) = \epsilon, k_1 = \sigma_w b;$$

and under H_1,

$$\Pr(w < k_1) = F\frac{(k_1+\theta)}{\sigma_w}$$

$$= F\left[b + \frac{\theta}{\sigma a\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2}}\right] = P_0.$$

For fixed ϵ and θ/σ, and with sample sizes n_1 and n_2, P_0 depends only upon the quantity a, which in turn depends only upon the order of the pair of quantiles chosen for the test. It is natural, then, to choose that quantile which minimizes a and hence maximizes P_0. It is easy to determine that the *median*, the quantile of order 0.5, should be used. Inserting this value, the power function becomes:

$$P_0 = F\left[b + 0.7979\,\frac{\theta}{\sigma}\left(\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}\right].$$

In order to determine the efficiency of the test, the power function of the best test using all the sample values, denoted by P_0', must be derived. The efficiency is then defined as P_0/P_0'. The test statistic based on all the sample values is given by

$$v = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i - \frac{1}{n_2}\sum_{i=1}^{n_2} y_i.$$

The best critical region is easily found to be the region for which

$$v < k_3, \theta > 0,$$

$$v > k_4, \theta < 0.$$

For $\theta > 0$, P_0' is given by

$$P_0' = F\left[b + \frac{\theta}{\sigma}\left(\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}\right].$$

If now we take $n_1 = n_2 = n = 200$, $\epsilon = 0.01$, then

$$P_0 = F\left[-2.326 + 0.7979\frac{\theta}{\sigma}\left(\frac{n}{2}\right)^{1/2}\right]$$

$$= F\left(-2.326 + 7.979\frac{\theta}{\sigma}\right),$$

$$P_0' = F\left[-2.326 + \frac{\theta}{\sigma}\left(\frac{n}{2}\right)^{1/2}\right]$$

$$= F\left(-2.326 + 10\frac{\theta}{\sigma}\right).$$

Fig. 4 shows the power and efficiency of the test using one optimal pair of quantiles. Under the given conditions, $\sigma_w = 0.1253\sigma$ and $k_1 = -0.2914\sigma$. Thus, if the difference between the median of the samples taken from the first population and the median of the samples taken from the second is less than -0.2914σ, reject H_0 if $\theta > 0$. Otherwise, accept H_0. The test is made at a significance level of 0.01. If $\theta < 0$, then

$$w > k_2 = 0.2914\sigma$$

is the critical region for the single-quantile test, and

$$P_0 = 1 - F\left(2.326 + 7.979\frac{\theta}{\sigma}\right),$$

$$P_0' = 1 - F\left(2.326 + 10\frac{\theta}{\sigma}\right).$$

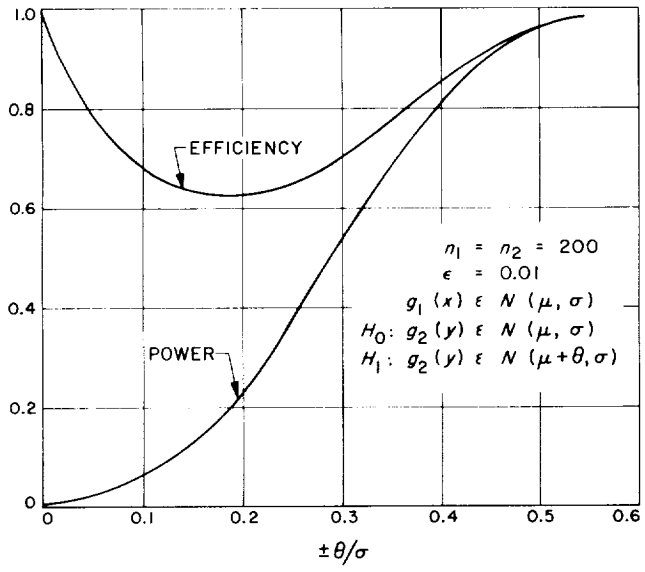


Fig. 4. Power and efficiency of Test A using one pair of quantiles

It is easily seen, therefore, that if $\theta_1 = -\theta_2 < 0$,

$$P_0(\theta_1) = P_0(\theta_2),$$

$$P_0'(\theta_1) = P_0'(\theta_2).$$

It is interesting to note that, as $n_1 \to \infty$ for fixed $n_2 = n$, the power functions P_0 and P_0' increase and approach, for $\theta > 0$,

$$P_0 = F\left(-2.326 + 0.7979\frac{\theta}{\sigma}n^{1/2}\right),$$

$$P_0' = F\left(-2.326 + \frac{\theta}{\sigma}n^{1/2}\right). \qquad (2)$$

Eqs. (2) are the power functions obtained for the one-quantile test on a single set of sample values described in *SPS 37-25*, Vol. IV, which, in the present terminology, is given by

$$H_0: g(x) = g_1(x) \in N(\mu, \sigma),$$

$$H_1: g(x) = g_2(x) \in N(\mu + \theta, \sigma),$$

where μ and σ are known.

This phenomenon can be explained by the fact that, although μ is unknown in the present test, its value can be estimated exactly with probability one as $n_1 \to \infty$, so that the test is essentially made on the second set of sample values with known μ and σ. This same phenomenon also occurs in the two- and four-quantile cases.

3. *Test A Using Two Pairs of Quantiles*

Now let z_1 and z_2 be the sample quantiles of the first population of orders p_1 and $p_2 = 1 - p_1 > p_1$; z_1' and z_2', the corresponding sample quantiles of the second population; and ζ_1 and ζ_2, the corresponding population quantiles of the standard normal. To eliminate μ, let

$$w_1 = z_1 - z_1', \quad w_2 = z_2 - z_2'.$$

Then one has the following: Under H_0,

$$E(z_1) = \sigma\zeta_1 + \mu,$$

$$\mathrm{Var}(z_1) = \frac{\sigma^2 a^2}{n_1},$$

$$E(z_2) = \sigma\zeta_2 + \mu = -\sigma\zeta_1 + \mu,$$

$$\mathrm{Var}(z_2) = \mathrm{Var}(z_1),$$

$$E(z_1') = \sigma\zeta_1 + \mu,$$

$$\mathrm{Var}(z_1') = \frac{\sigma^2 a^2}{n_2},$$

$$E(z_2') = -\sigma\zeta_1 + \mu,$$

$$\mathrm{Var}(z_2') = \mathrm{Var}(z_1'),$$

$$E(w_1) = E(w_2) = 0,$$

$$\mathrm{Var}(w_1) = \mathrm{Var}(w_2) = \sigma^2 a^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

where

$$a^2 = \frac{F(\zeta_1)\,[1 - F(\zeta_1)]}{f^2(\zeta_1)}.$$

Under H_1,

$$E(z_1) = \sigma\zeta_1 + \mu,$$

$$\mathrm{Var}(z_1) = \frac{\sigma^2 a^2}{n_1},$$

$$E(z_2) = -\sigma\zeta_1 + \mu,$$

$$\mathrm{Var}(z_2) = \mathrm{Var}(z_1),$$

$$E(z_1') = \sigma\zeta_1 + \mu + \theta,$$

$$\mathrm{Var}(z_1') = \frac{\sigma^2 a^2}{n_2},$$

$$E(z_2') = -\sigma\zeta_1 + \mu + \theta,$$

$$\mathrm{Var}(z_2') = \mathrm{Var}(z_1'),$$

$$E(w_1) = E(w_2) = -\theta,$$

$$\mathrm{Var}(w_1) = \mathrm{Var}(w_2) = \sigma^2 a^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

The two-quantile test will be made on the statistic $y = w_1 + w_2$. Under H_0,

$$E(y) = 0,$$

$$\mathrm{Var}(y) = 2\sigma^2 a^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(1 + \rho),$$

where ρ = the correlation between z_1 and z_2 and also that between z_1' and z_2'. Under H_1,

$$E(y) = -2\theta,$$

$$\mathrm{Var}(y) = 2\sigma^2 a^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(1 + \rho).$$

From the likelihood ratio, the critical regions are easily determined to be the regions in which

$$y < k_5, \theta > 0,$$

$$y > k_6, \theta < 0.$$

The power function for $\theta > 0$ is given as

$$P_0 = F\left\{b + \frac{2\theta}{\sigma a\left[\frac{2(n_1 + n_2)(1 + \rho)}{n_1 n_2}\right]^{1/2}}\right\},$$

$$F(b) = \epsilon,$$

$$k_5 = \sigma_y b.$$

The two symmetric quantiles that maximize P_0 are those of orders $p_1 = 0.2703$ and $p_2 = 0.7297$.

For $n_1 = n_2 = 200$, $\epsilon = 0.01$, using the optimum quantiles, one obtains the following:

$$P_0 = F\left(-2.326 + 9.0\frac{\theta}{\sigma}\right),$$

$$\sigma_y = 0.2222\sigma,$$

$$k_5 = -0.5168\sigma.$$

Hence, for $\theta > 0$, if $y < -0.5168\sigma$, reject H_0. Otherwise, accept H_0. For $\theta < 0$, if $y > 0.5168\sigma$, reject H_0. Otherwise, accept H_0. Fig. 5 shows the power and efficiency of the test using the two optimum pairs of quantiles. Only symmetric quantiles are considered; they have been shown to have the optimum spacing for estimating the mean of a normal population with an even number of quantiles.

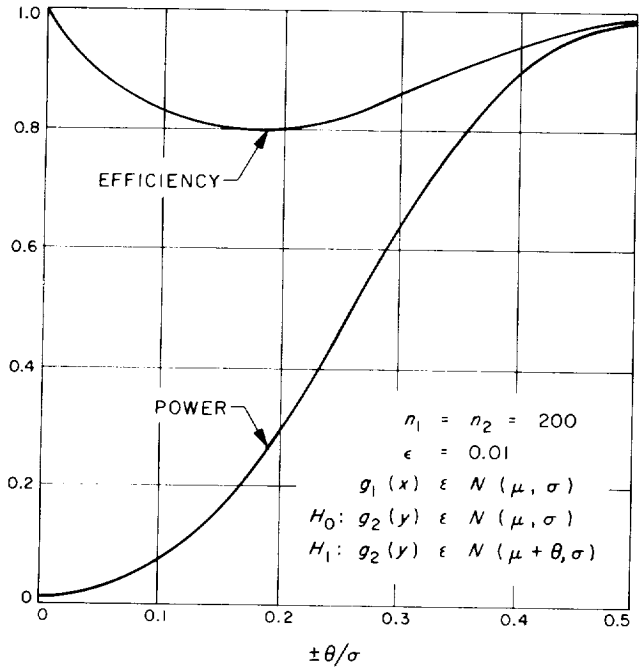


Fig. 5. Power and efficiency of Test A using two pairs of quantiles

4. Test A Using Four Pairs of Quantiles

The procedure in the four-quantile case is slightly more involved, but still straightforward. Let z_i, $i = 1, 2, 3, 4$, be the sample quantiles of the first population of orders p_i, such that $p_1 < p_2 < p_3 < p_4$ and $p_1 + p_4 = p_2 + p_3 = 1$. Let z'_i be the corresponding sample quantiles from the second population and ζ_i, the corresponding population quantiles of the standard normal. Forming

$$w_i = z_i - z'_i\,, \qquad i = 1, 2, 3, 4,$$

$$x_1 = w_1 + w_4, \qquad x_2 = w_2 + w_3\,,$$

the test will be based on the statistic given by the linear combination

$$y = \alpha\, x_1 + \beta\, x_2\,.$$

The parameters α and β, as well as the optimum orders of the quantiles, will be determined so as to maximize P_0. Omitting many of the details, one has, for $n_1 = n_2 = n$, under H_0:

$$E(y) = 0\,,$$

$$\mathrm{Var}(y) = \frac{8\alpha\beta\, a_1 a_2 \sigma^2}{n}(\rho_{12} + \rho_{13}) + \frac{4\sigma^2}{n}\left[\alpha^2 a_1^2 (1 + \rho_{14}) + \beta^2 a_2^2 (1 + \rho_{13})\right] = \sigma^2\gamma^2\,,$$

where

$$a_1^2 = \frac{F(\zeta_1)\,[1 - F(\zeta_1)]}{f^2(\zeta_1)}\,,$$

$$a_2^2 = \frac{F(\zeta_2)\,[1 - F(\zeta_2)]}{f^2(\zeta_2)}\,,$$

and ρ_{ij} denotes the correlation between z_i and z_j as well as that between z'_i and z'_j. The power function for $\theta > 0$ is given by

$$P_0 = F\left(b + \frac{2(\alpha + \beta)\theta}{\sigma\gamma}\right), \quad F(b) = \epsilon\,, k = \sigma\gamma b\,.$$

It has been shown in Ref. 14 that the four quantiles which maximize P_0 are

$$p_1 = 0.1068\,, \; p_2 = 0.3512\,,$$

$$p_3 = 0.6488\,, \; p_4 = 0.8932\,,$$

and that weights $\alpha = 0.192$ and $\beta = 0.308$ are optimum. Inserting these values and assuming $n = 200$, $\epsilon = 0.01$, one has, finally,

$$P_0 = F\left(-2.326 + \frac{9.586\theta}{\sigma}\right).$$

Fig. 6 shows the power and efficiency of the test using four optimum pairs of quantiles. For this case,

$$\sigma_y = 0.1043\sigma\,,$$

$$k = -0.2426\sigma\,,$$

so that, for $\theta > 0$, if $y = 0.192\,(z_1 - z'_1 + z_4 - z'_4) + 0.308\,(z_2 - z'_2 + z_3 - z'_3) < -0.2426\sigma$, reject H_0. Otherwise, accept H_0. For $\theta < 0$, if $y > 0.2426\sigma$, reject H_0. Otherwise, accept H_0.

In each case, since the critical region depends upon σ and *not upon* θ, the tests are seen to be uniformly most powerful (among quantile tests). Figs. 4–6 show that, as $|\,\theta/\sigma\,|$ increases from zero, the efficiency of each test decreases from one to a minimum value and then increases and approaches one asymptotically. The minimum efficiency is about 0.62 using one pair of quantiles, about 0.80 using two pairs of quantiles, and about 0.91 using four pairs of quantiles. Thus, the efficiency, as is now familiar to students of quantiles, is quite high.

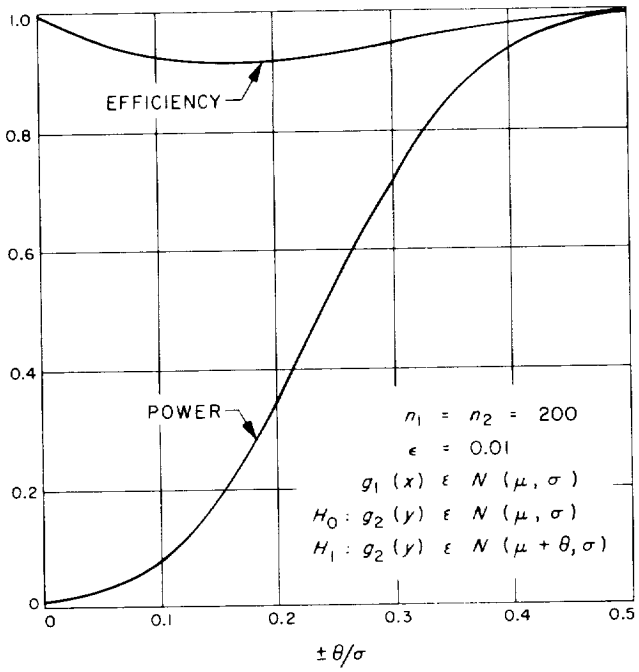


Fig. 6. Power and efficiency of Test A using four pairs of quantiles

5. Test B′, Determination of P'_0

In Test B′, we are testing the null hypothesis:

$$H_0: g_1(x) \in N(0,\sigma), g_2(y) \in N(0,\sigma)$$

against the alternative hypothesis:

$$H_1: g_1(x) \in N(0,\sigma), g_2(y) \in N(0,\theta\sigma), \theta > 0,$$

where σ is unknown; we will assume $n_1 = n_2 = n = 200$. Since σ is not known, the distribution of any test statistic must be independent of σ. Under this restraint, the best test statistic using all the sample values is given by

$$S = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} .$$

However, it is more convenient to use instead the equivalent statistic

$$r = \frac{1}{2} \ln S = \frac{1}{2} \ln \left(\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\frac{1}{n}\sum_{i=1}^{n} y_i^2} \right).$$

Now, under H_0,

$$r = \frac{1}{2} \ln S = \frac{1}{2} \ln \left(\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2/\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} y_i^2/\sigma^2} \right),$$

and S has the F-distribution with n degrees of freedom in the numerator and denominator. Since n is large, r is approximately normal (Ref. 15), with zero mean and variance equal to $1/n$. This approximation will be used to compute P'_0. Under H_1,

$$r = \frac{1}{2} \ln S = \frac{1}{2} \ln \left(\frac{\frac{1}{\theta^2 n}\sum_{i=1}^{n} x_i^2/\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} y_i^2/\theta^2\sigma^2} \right)$$

$$= \frac{1}{2} \ln \left(\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2/\sigma^2}{\frac{1}{n}\sum_{i=1}^{n} y_i^2/\theta^2\sigma^2} \right) - \frac{1}{2} \ln \theta^2 .$$

Thus, under the alternative hypothesis, r is also approximately normal, with mean equal to $-\frac{1}{2} \ln \theta^2$ and variance equal to $1/n$. The critical regions are easily determined to be the regions for which

$$r < k_7, \theta > 1 ,$$

$$r > k_8, \theta < 1 .$$

For $\epsilon = 0.01$,

$$P'_0 = F(-2.326 + 7.07 \ln \theta^2), \theta > 1, k_7 = -\frac{2.326}{n^{1/2}},$$

$$P'_0 = 1 - F\left(2.326 - 7.07 \ln \frac{1}{\theta^2}\right), \theta < 1, k_8 = \frac{2.326}{n^{1/2}} .$$

6. Test B′ Using One Pair of Quantiles

The orders of the quantiles used in the quantile tests will be those which minimize the variance of the estimate of σ from a single set of samples. Since μ_1 and μ_2 are known, it is possible to obtain a consistent estimate of σ using one optimum quantile of order $p = 0.058$ or $p = 0.942$, as determined by J. Ogawa in Ref. 16. (However, if μ_1 and μ_2 are not known, it is still possible to obtain a non-consistent test statistic for Test B′ by using the medians, but, since the power is extremely poor and is also independent of n, this test will not be considered.)

Thus, let z be the sample quantile of order 0.942 of the samples taken from the first population; z', the corresponding sample quantile from the second population; and ζ, the corresponding population quantile of the standard normal. The test statistic that will be used to eliminate dependence on σ is given by $u = z/z'$. In order to specify a critical region for a given ϵ, it is necessary to determine the distribution of u. In general, if x and y are normal random variables and are distributed $N(a_1,\sigma_1)$ and $N(a_2,\sigma_2)$, respectively, then $u = x/y$ can be shown to have a density function given by

$$h(u) = \frac{\sigma_1\sigma_2 \exp\left[-\frac{1}{2}\left(\frac{\sigma_1^2 a_2^2 + \sigma_2^2 a_1^2}{\sigma_1^2 \sigma_2^2}\right)\right]}{\pi(\sigma_2^2 u^2 + \sigma_1^2)}$$

$$+ \frac{\sigma_2^2 a_1 u + \sigma_1^2 a_2}{(2\pi)^{1/2}(\sigma_2^2 u^2 + \sigma_1^2)^{3/2}} \exp\left\{-\frac{1}{2}\left[\frac{(a_2 u - a_1)^2}{\sigma_2^2 u^2 + \sigma_1^2}\right]\right\}$$

$$\times \left\{-1 + 2F\left[\frac{\sigma_2^2 a_1 u + \sigma_1^2 a_2}{\sigma_1\sigma_2(\sigma_2^2 u^2 + \sigma_1^2)^{1/2}}\right]\right\},$$

$$-\infty < u < \infty .$$

Thus, since under H_0

$$E(z) = E(z') = \sigma\zeta ,$$

$$\mathrm{Var}(z) = \mathrm{Var}(z') = \sigma^2 a^2 ,$$

where

$$a^2 = \frac{F(\zeta)[1-F(\zeta)]}{nf^2(\zeta)} ,$$

whereas under H_1

$$E(z) = \sigma\zeta , E(z') = \sigma\theta\zeta ,$$

$$\mathrm{Var}(z) = \sigma^2 a^2 , \mathrm{Var}(z') = \sigma^2\theta^2 a^2 ,$$

one has, for $p = 0.942$,

$$h(u|H_0) = \frac{\exp(-121.679)}{\pi(u^2+1)} + \frac{4.4007(u+1)}{(u^2+1)^{3/2}}$$

$$\times \exp\left\{-\left[\frac{60.841(u-1)^2}{u^2+1}\right]\right\}$$

$$\times \left\{-1 + 2F\left[\frac{11.031(u+1)}{(u^2+1)^{1/2}}\right]\right\} ,$$

$$-\infty < u < \infty ,$$

$$h(u|H_1) = \frac{\theta\exp(-121.679)}{\pi(\theta^2u^2+1)} + \frac{4.4007\theta(\theta u+1)}{(\theta^2u^2+1)^{3/2}}$$

$$\times \exp\left\{-\left[\frac{60.841(\theta u-1)^2}{\theta^2u^2+1}\right]\right\}$$

$$\times \left\{-1 + 2F\left[\frac{11.031(\theta u+1)}{(\theta^2u^2+1)^{1/2}}\right]\right\} ,$$

$$-\infty < u < \infty .$$

Since

$$\int_{-\infty}^{\infty} \frac{\theta}{\pi(\theta^2u^2+1)}\, du = 1 ,$$

and

$$2F\left[\frac{11.031(\theta u+1)}{(\theta^2u^2+1)^{1/2}}\right] \cong 2$$

for all values of u and θ for which

$$\exp\left[-\frac{60.841(\theta u-1)^2}{(u^2+1)}\right]$$

is not nearly equal to zero, $h(u|H_0)$ and $h(u|H_1)$ can be written as follows:

$$h(u|H_0) = \frac{4.4007(u+1)}{(u^2+1)^{3/2}} \exp\left[-\frac{60.841(u-1)^2}{u^2+1}\right] ,$$

$$h(u|H_1) = \frac{4.4007\theta(\theta u+1)}{(\theta^2u^2+1)^{3/2}} \exp\left[-\frac{60.841(\theta u-1)^2}{\theta^2u^2+1}\right]$$

Moreover, the mean and mode of $h(u|H_0)$ and $h(u|H_1)$ are approximately $u = 1$ and $u = 1/\theta$, respectively, so that, for $\theta > 1$, $h(u|H_1)$ is shifted to the left of $h(u|H_0)$ and, for $\theta < 1$, $h(u|H_1)$ is shifted to the right of $h(u|H_0)$. Hence, the critical regions will be taken as the regions for which

$$u < k_9 , \theta > 1 ,$$

$$u > k_{10} , \theta < 1 ,$$

where k_9 and k_{10} are defined by

$$\left.\begin{aligned} \int_{-\infty}^{k_9} h(u|H_0)du &= \epsilon \\ \int_{-\infty}^{k_9} h(u|H_1)du &= P_0 \end{aligned}\right\}, \theta > 1 ;$$

$$\left.\begin{aligned} \int_{-\infty}^{k_{10}} h(u|H_0)du &= 1-\epsilon \\ \int_{-\infty}^{k_{10}} h(u|H_1)du &= 1-P_0 \end{aligned}\right\}, \theta < 1 .$$

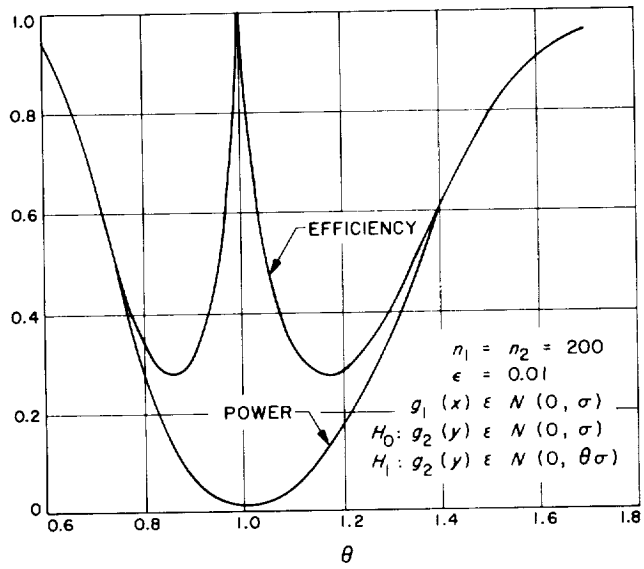


Fig. 7. Power and efficiency of Test B′ using one pair of quantiles

For $\epsilon = 0.01$, the values of k_9 and k_{10} were determined to be 0.738 and 1.354, respectively. Fig. 7 shows the efficiency and power of the test using one quantile as functions of θ.

7. Test B′ Using Two Pairs of Quantiles

The orders of the optimum two pairs of symmetric quantiles that will be used in the two-quantile test are $p_1 = 0.069$ and $p_2 = 0.931$. Hence, let z_1 and z_2 be the sample quantiles taken from the first population of orders p_1 and p_2; z_1' and z_2', the corresponding sample quantiles from the second population; and ζ_1 and ζ_2, the corresponding population quantiles from the standard normal. Now let $x_1 = z_2 - z_1$ and $x_2 = z_2' - z_1'$. Under H_0:

$$E(x_1) = E(x_2) = 2\sigma\zeta_2 \ ,$$

$$\mathrm{Var}(x_1) = \mathrm{Var}(x_2) = 2\sigma^2 a^2(1-\rho) \ ,$$

where

$$a^2 = \frac{F(\zeta_1)[1 - F(\zeta_1)]}{nf^2(\zeta_1)}$$

and ρ is the correlation between z_1 and z_2. Under H_1:

$$E(x_1) = 2\sigma\zeta_2 \ , \ \mathrm{Var}(x_1) = 2\sigma^2 a^2(1-\rho) \ ,$$

$$E(x_2) = 2\theta\sigma\zeta_2 \ , \ \mathrm{Var}(x_2) = 2\sigma^2\theta^2 a^2(1-\rho) \ .$$

Now, using $u = x_1/x_2$ as the test statistic, one has, for $p_1 = 0.069$ and $p_2 = 0.931$,

$$h(u|H_0) = \frac{\exp(-260.76)}{\pi(u^2+1)} + \frac{6.442(u+1)}{(u^2+1)^{3/2}} \times \exp\left[-\frac{130.38(u-1)^2}{u^2+1}\right] \times \left\{-1+2F\left[\frac{16.14(u+1)}{(u^2+1)^{1/2}}\right]\right\} \ ,$$

$$h(u|H_1) = \frac{\theta\exp(-260.76)}{\pi(\theta^2u^2+1)} + \frac{6.442\theta(\theta u+1)}{(\theta^2u^2+1)^{3/2}} \times \exp\left[-\frac{130.38(\theta u-1)^2}{\theta^2u^2+1}\right] \times \left\{-1+2F\left[\frac{16.14(\theta u+1)}{(\theta^2u^2+1)^{1/2}}\right]\right\} \ .$$

It is easily seen again that, in the same sense as before, one can express $h(u|H_0)$ and $h(u|H_1)$ as

$$h(u|H_0) = \frac{6.442(u+1)}{(u^2+1)^{3/2}} \exp\left[-\frac{130.38(u-1)^2}{u^2+1}\right] \ ,$$

$$h(u|H_1) = \frac{6.442\theta(\theta u+1)}{(\theta^2u^2+1)^{3/2}} \exp\left[-\frac{130.38(\theta u-1)^2}{\theta^2u^2+1}\right] \ .$$

Taking the critical regions as

$$u < k_{11} \ , \theta > 1 \ ,$$

$$u > k_{12} \ , \theta < 1 \ ,$$

k_{11} and k_{12} are found to be 0.814 and 1.228, respectively, for $\epsilon = 0.01$. Fig. 8 shows the efficiency and power of the

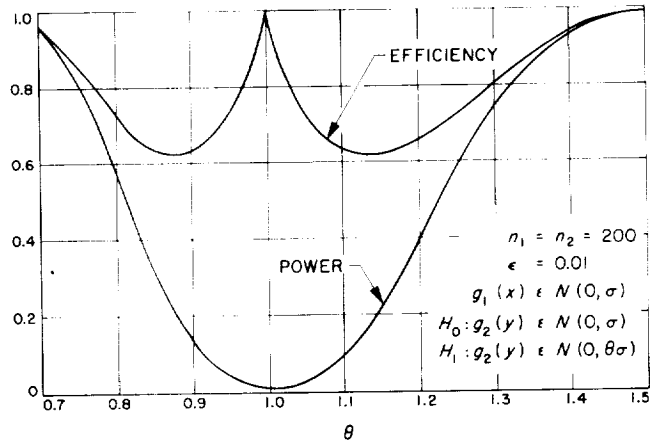


Fig. 8. Power and efficiency of Test B′ using two pairs of quantiles

test using the best pair of symmetric quantiles. We consider only symmetric quantiles; it is conjectured that this is the optimum choice.

8. Test B′ Using Four Pairs of Quantiles

The orders of the optimum four pairs of symmetric sample quantiles to be used in the four-quantile test are

$$p_1 = 0.023, p_2 = 0.127,$$
$$p_3 = 0.873, p_4 = 0.977.$$

Hence, let z_i, $i = 1, 2, 3, 4$, be four sample quantiles from the first population with orders p_i; z_i', the corresponding sample quantiles from the second population; and ζ_i, the corresponding population quantiles of the standard normal. Furthermore, define

$$x_1 = \alpha(z_4 - z_1), \quad x_1' = \alpha(z_4' - z_1'),$$
$$x_2 = \beta(z_3 - z_2), \quad x_2' = \beta(z_3' - z_2'),$$
$$y_1 = x_1 + x_2, \quad y_2 = x_1' + x_2'.$$

With weights $\alpha = 0.116$, $\beta = 0.236$, and the given orders of the quantiles, y_1 and y_2 are the best estimators of the standard deviation of the first and second populations, respectively, using two pairs of symmetric quantiles. Under H_0:

$$E(y_1) = E(y_2) = 2\sigma(\alpha\zeta_4 + \beta\zeta_3),$$
$$\mathrm{Var}(y_1) = \mathrm{Var}(y_2) = 2\sigma^2[\alpha^2 a_1^2(1-\rho_{14}) + \beta^2 a_2^2(1-\rho_{23}) + 2\alpha\beta a_1 a_2(\rho_{12} - \rho_{13})] = 2\sigma^2\gamma^2,$$

where

$$a_1^2 = \frac{F(\zeta_1)[1 - F(\zeta_1)]}{nf^2(\zeta_1)},$$
$$a_2^2 = \frac{F(\zeta_2)[1 - F(\zeta_2)]}{nf^2(\zeta_2)},$$

and ρ_{ij} is the correlation between z_i and z_j. Under H_1:

$$E(y_1) = 2\sigma(\alpha\zeta_4 + \beta\zeta_3), \ \mathrm{Var}(z_1) = 2\sigma^2\gamma^2,$$
$$E(y_2) = 2\sigma\theta(\alpha\zeta_4 + \beta\zeta_3), \ \mathrm{Var}(z_2) = 2\sigma^2\theta^2\gamma^2.$$

Using $u = y_1/y_2$ as the test statistic, one uses for the density of u:

$$h(u \mid H_0) = \frac{7.257(u+1)}{(u^2+1)^{3/2}} \exp\left[-\frac{165.445(u-1)^2}{u^2+1}\right],$$
$$h(u \mid H_1) = \frac{7.257\theta(\theta u+1)}{(\theta^2u^2+1)^{3/2}} \exp\left[-\frac{165.445(\theta u-1)^2}{\theta^2u^2+1}\right]$$

Taking the critical regions as

$$u < k_{13}, \theta > 1,$$
$$u > k_{14}, \theta < 1,$$

k_{13} and k_{14} are found to be 0.834 and 1.200, respectively, for $\epsilon = 0.01$. Fig. 9 shows the efficiency and power of the test using the best four pairs of symmetric quantiles.

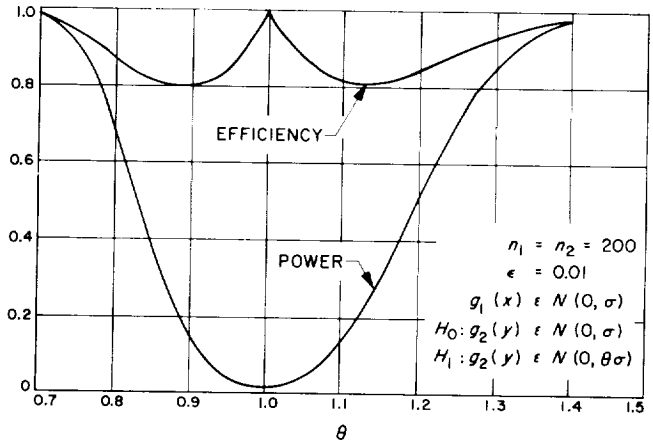


Fig. 9. Power and efficiency of Test B′ using four pairs of quantiles

It can be seen from Figs. 7–9 that, as θ increases or decreases from $\theta = 1$, the efficiency decreases from one to a minimum and then increases and approaches one asymptotically. The minimum efficiency is about 0.28 using one pair of quantiles, about 0.62 using two pairs of quantiles, and about 0.81 using four pairs of quantiles.

To convert Test B to Test B′, if z and z' are sample quantiles of order p taken from the first and second populations, respectively, under Test B, then $z - u_1$ and $z' - u_2$ should be used as the corresponding sample quantiles under Test B′.

D. Error-Probability Estimation

E. C. Posner

This article ties together some loose ends in the error-probability estimation theory being developed by the author in conjunction with J. Ashlock of JPL Section 334 (see Refs. 17 and 18 for previous work in this area).

The idea was developed to estimate the low error probabilities that occur in the *Ranger* Block III command detector. This article gives an improved method for obtaining confidence intervals for the error probability and provides another check on the validity of the theory. This theory is being considered for use on *Rangers* 8 and 9.

1. Review

Errors in many communication systems that use a binary digital channel are caused by voltage fluctuations so large that a hard limiter mistakes a voltage representing a "1" for a voltage representing a "0." This is the situation in the *Ranger* command detector, where the thresholds are set in such a way that the "0→1" error can be assumed not to occur. The specification[1] for the *Ranger* command detector calls for error probabilities out of the detector of less than 10^{-5}.

One method that is often used to check error probabilities is to count the number of errors made in many bits. But the error probabilities involved in the *Ranger* command detector are quite low. Thus, if this method were used, the number of bits that would have to be examined before good error-probability estimates could be obtained would be prohibitively large.

The way that has been adopted to avoid this problem is to use some of the physical information available concerning the cause of the error. Errors are caused by large voltage fluctuations, so, instead of just recording whether an error occurs on a given bit, one records the magnitude of the voltage fluctuation. Since one is interested in deviations above a certain amount, the distribution of maximum deviations is relevant.

In Refs. 17 and 18, the maximum deviation of n successive samples of the voltage out of the detector was recorded. The distribution of these maxima tends to follow a Gumbel distribution $\exp\{-\exp[-\alpha(x-u)]\}$ (Ref. 19) for some positive parameters α and u. One then estimates α and u from a block of N maxima and, in turn, uses α and u to estimate the probability that the maximum voltage deviation in n bits exceeds the threshold. In this way, one obtains an estimate for the probability of at least one error in n bits, from which it is an easy matter to estimate the error probability in a single bit. Thus, the desired error probability is found.

[1] JPL Function Specification RCK-31041-DSN-C.

This program was carried out in Refs. 17 and 18. It was shown that large savings in the time required to estimate low error probabilities (within a given accuracy) are obtained. Confidence intervals for the error probability were also found. In addition, a goodness-of-fit test which was applied to the data showed that theory fit experiment extremely well (see also *E. A Distribution-Free Goodness-of-Fit Test for Use in Extreme-Value Theory* for another test applied to these data).

An improved method for obtaining confidence intervals is given here. Also, a test is given which shows that the estimates of error probability do not fluctuate from each other any more than by the effect of random sampling.

2. Confidence Intervals

In Ref. 18, confidence intervals for the error probability were obtained in the following way: We assumed that the distribution of the maximum voltage deviation in n bits (n was 100) followed a Gumbel distribution $\exp\{-\exp[-\alpha(x-u)]\}$ for some positive parameters α and u. We took N (N was 30) independent samples from this Gumbel distribution and, from these N, obtained a confidence interval. If an error occurs whenever the voltage deviation x exceeds the threshold x_0, define $\nu = \alpha(x_0 - u)$ and write the distribution function $F(x)$ of x as $F(x) = \exp\{-\exp[-\alpha(x-x_0)+\nu]\}$. It is actually α and ν that were used as parameters in Ref. 18; the exceedance probability is $1 - F(x_0) = \exp[-\exp(-\nu)]$.

Large-sample theory was then used to obtain the distribution of the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\nu}$ of α and ν; $N = 30$ is thought to be large enough for this theory to hold. The values of the estimators $\hat{\alpha}$ and $\hat{\nu}$ were obtained by numerical calculation from the data.

We used the result that the large-sample distribution of $\hat{\alpha}$ and $\hat{\nu}$ is a bivariate normal distribution with means α and ν and with covariance matrix $(1/N)V$, where $V = R^{-1}$ and $R = -E[(\partial^2/\partial_i\partial_j)\log F(x)]$. Here $i, j = 1, 2$ and $\partial_1 = \partial\alpha$, $\partial_2 = \partial\nu$; E denotes expected value with respect to the distribution F. We then found that the large-sample marginal distribution of $\hat{\nu}$ was normal, with mean ν and variance given by

$$\operatorname{Var}\hat{\nu} = \frac{6}{N\pi^2}\left[(1-\gamma+\nu)^2 + \frac{\pi^2}{6}\right], \qquad (1)$$

where γ is Euler's constant $0.5772\cdots$.

In Ref. 18, we replaced ν by the $\hat{\nu}$ calculated from the data to obtain the variance of $\hat{\nu}$ to be used in obtaining confidence intervals. But, since the square root of the variance of $\hat{\nu}$ turned out to be non-negligible in comparison with $\hat{\nu}$, this procedure did not give the best answer. In this article, we obtain confidence intervals for ν without replacing ν by $\hat{\nu}$ in Eq. (1).

What is desired is a *one-sided* confidence interval on the error probability. That is, we wish to say that, unless an event with a certain low probability has occurred, the true error probability is less than a certain value. This is the criterion upon which acceptance or rejection of a given detector under test is to be based. A one-sided-on-the-left confidence interval for the probability that x exceeds x_0 corresponds to another confidence interval for ν, one-sided on the right. We now proceed to obtain such an interval.

If a confidence interval of confidence $\lambda(\lambda$ = a number slightly less than 1) is desired, we demand a ν_0 such that

$$\Pr(\nu > \nu_0) = \lambda \tag{2}$$

in the *a posteriori* sense. To do this, we seek a ν_1 such that

$$\Pr(\hat{\nu} < \nu_1) = \lambda\,, \tag{3}$$

where probability is interpreted according to the marginal distribution of $\hat{\nu}$. Since $\hat{\nu}$ has mean ν and variance σ^2 given by Eq. (1), we write Eq. (3) as

$$\Pr\left(\frac{\hat{\nu}-\nu}{\sigma} < \frac{\nu_1-\nu}{\sigma}\right) = \lambda\,. \tag{4}$$

Since $\hat{\nu}$ is normal, $(\hat{\nu}-\nu)/\sigma$ has the unit normal distribution. Now define Φ_λ as the quantile of order λ of the unit-variance normal distribution (that is, with probability λ, a unit normal variate is less than Φ_λ). We then have, from Eq. (1),

$$\frac{\nu_1-\nu}{\left\{\frac{6}{N\pi^2}\left[(1-\gamma+\nu)^2+\frac{\pi^2}{6}\right]\right\}^{1/2}} = \Phi_\lambda\,, \tag{5}$$

so that one obtains a quadratic equation for ν:

$$\nu^2\left(1-\frac{6\Phi_\lambda^2}{N\pi^2}\right) - 2\nu\left[\nu_1+\frac{6\Phi_\lambda^2}{N\pi^2}(1-\gamma)\right] + \nu_1^2 - \frac{6\Phi_\lambda^2}{N\pi^2}\left[(1-\gamma)^2+\frac{\pi^2}{6}\right] = 0\,. \tag{6}$$

Only one of the two roots of Eq. (6) is acceptable, since ν_1 must be greater than ν when $\Phi_\lambda > 0$ (the case of interest corresponds to $\lambda > \frac{1}{2}$). Thus,

$$\nu_1 = \nu + \left\{\frac{6}{N\pi^2}\left[(1-\gamma+\nu)^2+\frac{\pi^2}{6}\right]\right\}^{1/2}\Phi_\lambda\,. \tag{7}$$

$$\Pr\left(\hat{\nu} < \nu + \left\{\frac{6}{N\pi^2}\left[(1-\gamma+\nu)^2+\frac{\pi^2}{6}\right]\right\}^{1/2}\Phi_\lambda\right) = \lambda\,. \tag{8}$$

Consider the region in which

$$\hat{\nu} < \nu + c\left[(\nu+a)^2+b^2\right]^{1/2}\,, \tag{9}$$

for a,b,c positive constants and $c < 1$. One has

$$(\hat{\nu}-\nu)^2 < c^2\left[(\nu+a)^2+b^2\right]\,, \tag{10}$$

which ultimately reduces to

$$\nu > \frac{1}{1-c^2}\left\{ac^2+\hat{\nu}-c\left[(\nu+a)^2+(1-c^2)b^2\right]^{1/2}\right\}\,. \tag{11}$$

The minus sign must be taken in Eq. (11) since the plus sign will ultimately make ν_0 greater than $\hat{\nu}$. But ν_0 must be less than ν, for $\hat{\nu}$ corresponds to a confidence of 0.5, whereas we are interested in larger confidences. In the problem at hand,

$$c = \left(\frac{6}{N\pi^2}\right)^{1/2}\Phi_\lambda,\ a = 1-\gamma,\ b = \pi^2/6\,.$$

We now have, as in Eq. (2), that $\Pr(\hat{\nu} > \nu_0) = \lambda$, where ν_0 is given by

$$\nu_0 = \frac{1}{1-c^2}\left\{ac^2+\hat{\nu}-c\left[(\nu+a)^2+(1-c^2)b^2\right]^{1/2}\right\}\,. \tag{12}$$

But, since $\hat{\nu}$ is computed from the data and a,b,c are known, Eq. (12) can be turned around to be interpreted as an *a posteriori* statement about the unknown ν. That is, ν_0 is the left-hand endpoint of the one-sided confidence interval for ν and is computable from the data.

We shall now do a sample calculation using this method. In Ref. 18 N was 30, λ was taken as 0.9 (Φ_λ was therefore 1.282), c was 0.1828, and $\hat{\nu}$ was 3.52; a,b were given above. We find that $\nu_0 = 2.87$. Previously, using the method which replaces ν by $\hat{\nu}$ in Eq. (1), we had obtained $\nu_0 = 2.76$. That is, the more exact method gives

a smaller confidence interval. This happens because the interval given by the previous method gives a confidence greater than 0.9. Thus, the method presented here allows sharper inferences to be made.

Translating back to exceedance probability $1 - \exp[-\exp(-\nu_0)]$, we find that the right-hand endpoint of the confidence interval for the exceedance probability is 0.0531; the previous method gave 0.0613. Since the error probability is related to the exceedance probability by virtue of the fact that the probability of no exceedance is the probability of no error in n trials, the error probability is closely approximated by $1/n$ times the exceedance probability (for small probabilities). Since $n = 100$, we can say that, unless an event with probability $1 - \lambda = .1$ has occurred, the bit error probability is less than 5.3×10^{-4}, instead of 6.1×10^{-4} obtained by the previous method. This improvement is desirable.

3. Internal Consistency of Method

To further demonstrate the applicability of the Gumbel method to deviations in the *Ranger* command detector, the following experiment was performed. If the theory is to hold, exceedance probability estimates from a short run of data should yield estimates related to, but not as good as, the estimates obtained with further data. To test this, the 30 extremes of Ref. 18 were divided into two groups of 15 each at random. One of the groups was chosen, and a "Gumbel line" was fitted on extreme-value probability paper for these 15 values. On the same paper, the line for all 30 points was drawn for comparison. These plots are given in Fig. 10.

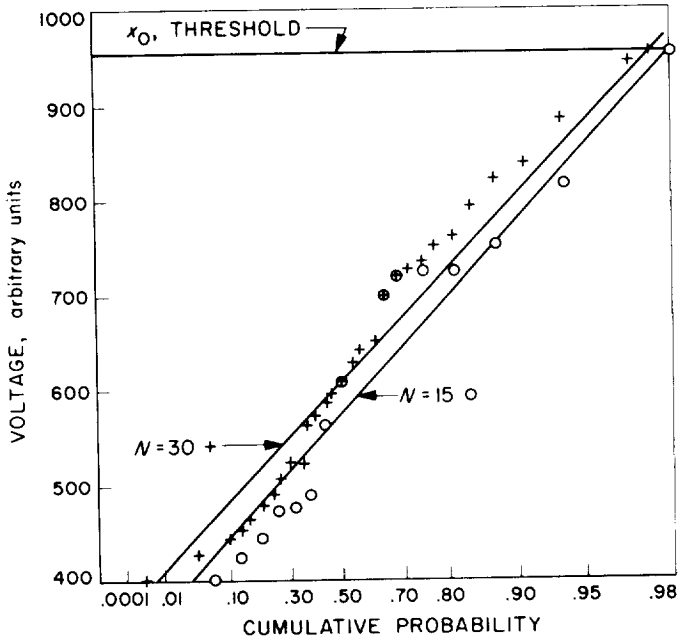


Fig. 10. Plots demonstrating consistency

Note how close the two lines are—well within an error explainable by random deviations in sampling. The maximum-likelihood exceedance probability estimated from all 30 extremes was .024; the set of 15 gives .020. Thus, the theory is internally consistent within the data. This justifies the use of the method for error-probability estimation for command detectors.

E. A Distribution-Free Goodness-of-Fit Test for Use in Extreme-Value Theory

E. C. Posner and S. Zohar

The distribution-free goodness-of-fit test for use in extreme-value theory (see Ref. 20) described in this article is of the one-sided Kolmogorov–Smirnov type and uses as the test statistic

$$P_n^+ = n^{1/2} \sup_x [F(x) - F_n(x)],$$

where F is the assumed distribution function and F_n is the sample distribution function from n samples. A table of the distribution for $n = 30$ is obtained by using the Franklin–Koksma random numbers. This table is used to test whether the Gumbel distribution holds for extreme voltages in the *Ranger* command detector, as studied in Ref. 18 (see also *D. Error-Probability Estimation*).

1. Introduction

A *goodness-of-fit test* in statistics is a test which determines whether it is reasonable to assume that a supposed distribution function actually holds after the random variable whose distribution is being questioned has been sampled. A test is called *distribution-free* if the distribution of the test statistic being used, under the null hypothesis that the assumed distribution is the true one, is independent of the true distribution. One advantage of distribution-free tests is that only one table is required, instead of a different table for each null hypothesis.

In Ref. 21, a family of distribution-free goodness-of-fit tests is described. A test is called a one-sided *Kolmogorov–Smirnov test* if the test statistic used is of one of the two forms

$$n^{1/2} \sup_x \{\pm [F(x) - F_n(x)] \Psi [F(x)]\},$$

where F is the assumed distribution; supposed continuous everywhere; F_n, the sample distribution; and $\Psi(F)$, a fixed non-negative weight function (selected according to what the alternative hypothesis to F might be). In Ref. 22, Doob used the theory of Markov processes to derive the asymptotic distribution of the test statistic for large n (this explains the use of the factor $n^{1/2}$ when Ψ is identically equal to 1). The answer is surprisingly simple—the asymptotic distribution of the test statistic is $1 - \exp(-2t^2)$.

In Refs. 18 and 20, interest was expressed in Gumbel's extreme-value distributions $\exp\{-\exp[-\alpha(x-u)]\}$, $\alpha, u > 0$. In using Gumbel's theory, one should first test whether Gumbel's distribution can be assumed to hold. Since Gumbel's theory is concerned with extreme values, the alternatives to the null hypothesis that one wishes to avoid are deviations in the right-hand end of the distribution. This is especially true when using Gumbel's theory to estimate error probabilities in the *Ranger* command detector, as in Ref. 18. One fears that the values of x given by the true distribution are even larger than the Gumbel distribution would indicate, since then detectors would be accepted as good when in fact they yield too high an error probability.

This suggests that the test statistic should weight values of $F(x)$ close to 1 more than values close to 0. However, since one is also interested in whether the theory holds at all and not just in whether the large values of x are given correct probabilities, the test statistic should also give some weight to small x. These facts suggest $\Psi(F) = F$ as the test statistic. That is, define

$$P_n^+ = n^{1/2} \sup_x \{[F(x) - F_n(x)]F(x)\}.$$

The "max" rather than the "min" is chosen because, when F_n is below F by a large amount, the true values of x tend to be larger than the null hypothesis would indicate. For if fewer x are below x_0 [that is, if $F_n(x_0) < F(x_0)$], then x exceeds x_0 more than F would indicate. The weight function $F(x)$ weights the larger x_0 more than the smaller, so that if $F_n(x_0)$ were less than $F(x_0)$ near the left-hand tail, the statistic P_n^+ might not pick up this deviation, as it need not by the heuristic motivation for the test. Ref. 20 attempted to find the distribution of P_n^+ analytically, but did not succeed in doing so. This article finds the distribution by a Monte-Carlo method.

2. Transformation of the Statistic

As in Ref. 22, let u be the random variable $F(x)$, where the function F is assumed everywhere continuous. Then the random variable u has the uniform distribution on the unit interval $[0, 1]$, and the statistic P_n^+ becomes (under the null hypothesis that $F(x)$ is the true distribution of x):

$$P_n^+ = n^{1/2} \sup_{0 \le u \le 1} \{[u - G_n(u)]u\}, \qquad (1)$$

where u has the uniform distribution on $[0, 1]$, and $G_n(u)$ is the sample distribution of n samples from a random variable having this uniform distribution. Fig. 11 illustrates this definition.

We now consider a simplification of Eq. (1). Namely, let the n samples be u_i, $1 \le i \le n$. Order the n samples as $u_{(1)} \le u_{(2)} \le \cdots \le u_{(n)}$ [$u_{(j)}$ is the jth *order statistic* in n samples]. Then G_n has a jump of $1/n$ (ignoring the existence of ties in rank for simplicity of exposition) at each $u_{(j)}$ and has no other jumps. In fact, $G_n(u_{(j)}) = (j-1)/n$, $G_n(u_{(j)} + \epsilon) = j/n$ for all sufficiently small $\epsilon > 0$.

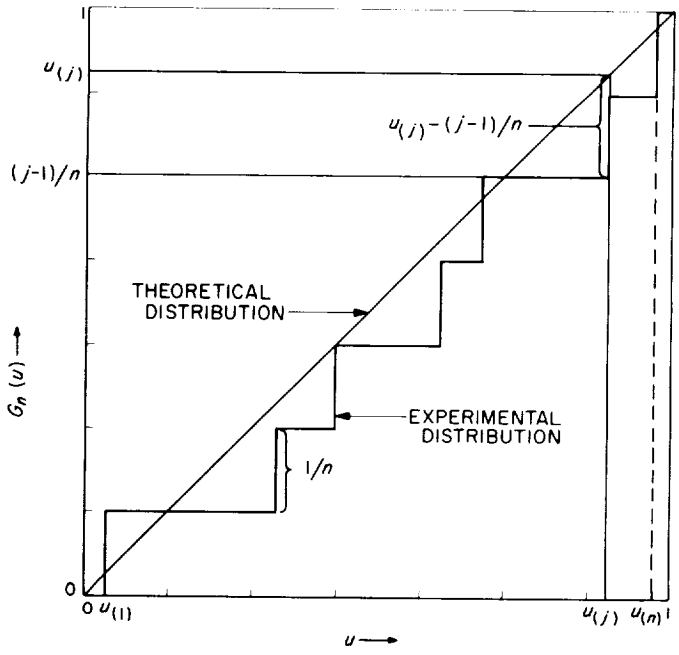


Fig. 11. Definition of P_n^+

Eq. (1) can then be written, as we shall see, as

$$P_n^+ = n^{1/2} \max_{1 \le j \le n} \left[\left(u_{(j)} - \frac{j-1}{n} \right) u_{(j)} \right], \qquad (2)$$

and it is in this form that P_n^+ is computed.

To show that the computation of P_n^+ is easier than it would appear from Eq. (1), we proceed as follows: Note that $G_n(u)$ is an increasing step function. Consider the interval (closed on the left, open on the right) in which $G_n(u)$ is the constant a. The function $(u-a)u$ is, however, increasing if $u \ge a/2$. Now $G_n(u) = 0$, $0 \le u < u_{(1)}$, so $[u - G_n(u)]u \ge 0$ for an initial interval of u, with probability 1. Then $a - G_n(a) \ge 0$ on the interval $[u_{(j-1)}, u_{(j)}]$, on which the supremum of $[u - G_n(u)]u$ is attained. That is, $u \ge a$ on the supremum interval, and certainly $u \ge a/2$ there. Then, on this interval, $(u - a)\,u$ is indeed increasing, and the supremum does indeed occur at the right-hand endpoint of the interval. We conclude that the supremum of $[u - G_n(u)]u$ occurs at $u = u_{(j)}$, where the value

$$G_n(u_{(j-1)}) = \frac{(j-1)}{n}$$

is to be substituted for $G_n(u)$. This proves Eq. (2).

3. Computation of the Table

In the *Ranger* command detector extreme-value-theory error-probability estimator (Ref. 17), the value $n = 30$ is of particular interest, rather than the asymptotic distribution of P_n^+ when n is large. In order to prepare a table of P_{30}^+, 30 values from a synthetic uniform distribution were generated using the Franklin–Koksma random number generator of Ref. 23, a copy of which program was supplied by the Booth Computing Center, California Institute of Technology. A large number of such random numbers (almost 20,000) was obtained to estimate the distribution of P_{30}^+. This yields almost 660 independent samples of the proper distribution. The program was run on the Mod II stored program controller of JPL Section 331 (Refs. 24 and 25). The number of runs r necessary to estimate the distribution of P_n^+ within a certain maximum error with given probability is given by the two-sided Kolmogorov–Smirnov test, using its asymptotic distribution given in Ref. 26. For 660 samples, the maximum error is less than 0.06 with probability .98. Thus, Fig. 12 should be interpreted as having a confidence band around the distribution function so obtained. Some quantiles (percentage points) at the right-hand end of the distribution are given below:

Order	Quantile
.1	.083
.2	.159
.3	.247
.4	.274
.5	.312
.6	.373
.7	.425
.8	.491
.9	.583
.95	.676
.975	.740
.99	.857
.995	.885

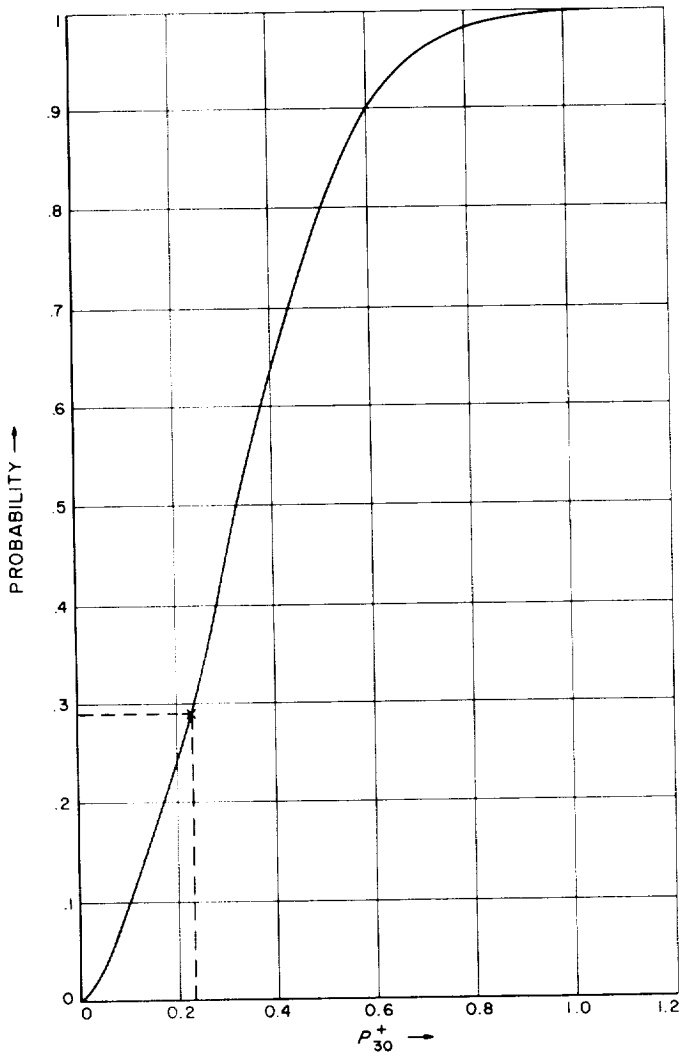


Fig. 12. Distribution of P_{30}^+

4. Checks on Computation

Let us discuss the checks that were made to ensure that the program was running correctly. First, the program was run with $n = 30$ using the known (Ref. 27) statistic

$$D_n^+ = n^{1/2} \sup_{0 \le u \le 1} [u - G_n(u)],$$

again with 660 independent samples. The answer obtained by our method was compared with that obtained using the exact method of Ref. 27. The maximum deviation was 0.05. According to the two-sided Kolmogorov–Smirnov test, deviations this large or larger occur with probability about .25. Thus, the fit is excellent.

The random numbers of Ref. 28 were tried before the Franklin–Koksma numbers were tried, but the deviations, when run for the known statistic, were found to be much too large to be explainable by random variation. Since the Franklin–Koksma numbers appear to be random enough for the known statistic, we accept the results also for the unknown statistic.

5. Application

We now apply this test to the Gumbel distribution which arises in extreme-value theory applied to the *Ranger* command detector, as studied in Ref. 20. From data in Ref. 20, one is able to find the value of the statistic P_{30}^+. Here the assumed distribution is a certain Gumbel distribution, and the empirical distribution is the distribution of minimum voltages in 30 runs of 100 independent samples out of the *Ranger* command detector. We find $P_{30}^+ = 0.22$, which deviation or larger occurs according to Fig. 12 with probability .72. Thus, the fit is quite good according to this special one-sided Kolmogorov–Smirnov test.

The advantage of using this special Kolmogorov–Smirnov test with weight function $\Psi(F) = F$, rather than the more usual one with weight function $\Psi(F) = 1$, is that the new test has higher power against alternatives of the form "the true distribution at a given x is less than the assumed distribution." That is, the test has higher power against alternatives of the form "the true random variable is stochastically greater than the assumed one." We gain this extra power where we need it by giving up our power against alternatives of the form "the true random variable is unequal to the assumed one."

References

1. Posner, E. C., "Statistics of Data Sources," *SPS 37-27*, Vol. IV, pp. 225–229, Jet Propulsion Laboratory, Pasadena, California, June 30, 1964.
2. "A Random Pulse Generator for Use as a Data and Noise Source," *SPS 37-27*, Vol. III, pp. 112–115, Jet Propulsion Laboratory, Pasadena, California, May 31, 1964.
3. "Construction of Random Pulse Generator," *SPS 37-29*, Vol. III, pp. 92–95, Jet Propulsion Laboratory, Pasadena, California, September 30, 1964.
4. "Design of Punctured-Cyclic Coder–Decoder," *SPS 37-27*, Vol. III, pp. 97–103, Jet Propulsion Laboratory, Pasadena, California, May 31, 1964.
5. Baumert, L., Easterling, M., Golomb, S. W., and Viterbi, A., *Coding Theory and its Applications to Communications Systems*, Technical Report No. 32-67, Jet Propulsion Laboratory, Pasadena, California, March 31, 1961.
6. Golomb, S. W., Welch, L. R., and Hales, A., *On the Factorization of Trinomials over GF(2)*, Memorandum No. 20-189, Jet Propulsion Laboratory, Pasadena, California, July 12, 1959.

References (Cont'd)

7. Feller, William, *An Introduction to Probability Theory and its Applications,* Second Edition, John Wiley & Sons, Inc., New York, 1957.

8. Peterson, W. W., *Error-Correcting Codes,* John Wiley & Sons, Inc., New York, 1961.

9. Mood, Alexander M., and Gaybill, Franklin W., *Introduction to the Theory of Statistics,* Second Edition, McGraw-Hill Book Company, Inc., New York, 1963.

10. Eisenberger, I., and Posner, E. C., *Systematic Statistics Used for Data Compression in Space Telemetry,* Technical Report No. 32-510, Jet Propulsion Laboratory, Pasadena, California, October 1, 1963.

11. Pearson, E. S., and Hartley, H. O., *Biometric Tables for Statisticians,* Vol. I, p. 28, Cambridge University Press, 1962.

12. Mosteller, F., "On Some Useful 'Inefficient' Statistics," *Annals of Mathematical Statistics,* Vol. 17, pp. 377–407, 1946.

13. Posner, E. C., "The Use of Quantiles for Space Telemetry Data Compression," *Proceedings of the National Telemetering Conference,* 1964, Sect. 1–3, pp. 1–6.

14. Eisenberger, I., and Posner, E., *Systematic Statistics Used for Data Compression in Space Telemetry,* Technical Report No. 32-510, Jet Propulsion Laboratory, Pasadena, California, October 1, 1963.

15. Kendall, M., and Stuart, A., *The Advanced Theory of Statistics,* pp. 378, 379, Charles Griffen & Company Limited, London, 1958.

16. Sarhan, A., and Greenberg, B., Editors, *Contributions to Order Statistics,* Table 10 E.4, John Wiley & Sons, Inc., New York, 1962.

17. Posner, E. C., "Extreme-Value Statistics for Estimation of Error Probabilities in Ranger Block III Command Detector," *SPS 37-27,* Vol. IV, pp. 222–225, Jet Propulsion Laboratory, Pasadena, California, June 30, 1964.

18. Posner, E. C., "Extension of Extreme-Value Theory for Error-Probability Estimation," *SPS 37-28,* Vol. IV, pp. 213–219, Jet Propulsion Laboratory, Pasadena, California, August 31, 1964.

19. Gumbel, Emil J., *Statistics of Extremes,* Columbia University Press, New York, 1958.

20. Harper, L., and Posner, E. C., "Random Walk for Kolmogorov–Smirnov Tests," *SPS 37-28,* Vol. IV, pp. 182–186, Jet Propulsion Laboratory, Pasadena, California, August 31, 1964.

21. Darling, D. A., "The Kolmogorov–Smirnov, Cramér-von Mises Goodness-of-Fit Test," *Annals of Mathematical Statistics,* Vol. 28, pp. 823–838, 1957.

22. Doob, J. L., "Heuristic Approach to the Kolmogorov–Smirnov Theorems," *Annals of Mathematical Statistics,* Vol. 20, pp. 393–403, 1949.

23. Franklin, Joel N., "Complete Equidistribution," *Journal of the Association for Computing Machinery,* Vol. 9, 1964.

24. "Stored Program Controller, Mod II," *SPS 37-14,* Vol. I, pp. 83–90, Jet Propulsion Laboratory, Pasadena, California, April 1, 1962.

References (Cont'd)

25. "Mod II Coders," *SPS 37-14*, Vol. I, pp. 91, 92, Jet Propulsion Laboratory, Pasadena, California, April 1, 1962.

26. Birnbaum, Z. W., "Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size," *Journal of the American Statistical Association*, Vol. 47, pp. 425–441, 1952.

27. Miller, Leslie H., "Table of Percentage Points of Kolmogorov Statistics," *Journal of the American Statistical Association*, Vol. 51, pp. 111–121, 1956.

28. Coveyou, R. R., "Serial Correlation in the Generation of Pseudo-Random Numbers," *Journal of the Association for Computing Machinery*, Vol. 5, pp. 72–74, 1960.