NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Space Programs Summary No. 37-39, Volume IV

for the period April 1, 1966 to May 31, 1966

Supporting Research and Advanced Development

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

June 30, 1966

Preface

The *Space Programs Summary* is a six-volume, bimonthly publication that documents the current project activities and supporting research and advanced development efforts conducted or managed by JPL for the NASA space exploration programs. The titles 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,* Vol. VI consists of an unclassified digest of appropriate material from Vols. I, II, and III; an original presentation of technical supporting activities, including engineering development of environmental-test facilities, and quality assurance and reliability; and a reprint of the space science instrumentation studies of Vols. I and II.

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

Space Programs Summary No. 37-39, 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)

SYSTEMS DIVISION

I. Systems Analysis

A. Earth-Venus-Mercury Mission Opportunities in the 1970's

F. M. Sturms, Jr.

In Ref. 1, Minovitch showed that indirect missions to Mercury having a close encounter with Venus enroute, require considerably less launch energy than a direct mission. The launch energy saving is achieved by the gravitational perturbation of Venus, which removes energy from the heliocentric orbit. A detailed trajectory and guidance analysis by Sturms and Cutting (Ref. 2) of an Earth-Venus-Mercury mission in 1970 demonstrated that such missions are feasible with Earth-based radio guidance and an *Atlas–Centaur* boost vehicle. This article presents the results of a survey to find Earth-Venus-Mercury mission opportunities in the decade of the 1970's.

1. Method of Analysis

The trajectories for this survey were obtained with a new conic computer program being developed at JPL for analyzing indirect, multiple-planet trajectories and other complex advanced missions. For the present study, the option was taken to compute planetary positions and velocities from mean orbital elements taken from Ref. 3. The results agree very well with trajectories computed using JPL's planetary ephemeris tapes.

Multiple-planet trajectories for this mission are constructed by finding heliocentric Earth-Venus trajectories and heliocentric Venus-Mercury trajectories such that the arrival velocity on the Venus approach asymptote is equal to that on the departure asymptote for the Venus-Mercury leg. For any given date at Venus, there is a minimum energy for the Venus-Mercury leg. All launch date-flight time combinations on the Earth-Venus leg which result in arrival energies less than this minimum for continuing to Mercury are outside the region of possible mission opportunities. This requirement for energy matching at Venus on the two legs of the trajectory is one of three constraints which define the region of possible missions.

The second constraint requires that the point of closest approach at Venus be above the surface. After the energy matching has been obtained, the arrival and departure asymptotes define a Venus-centered hyperbola. The periapsis radius of this hyperbola must be greater than the radius of Venus, taken in this study as 6200 km.

The third constraint requires that the launch energy at Earth (C_3) be less than some selected maximum. The maximum value of C_3 chosen for this study was 21 km²/sec². This value was selected as being an approximate upper bound for which the multiple-planet mission can be performed with a sizable payload on the *Atlas–Centaur* and is therefore more attractive than a direct Earth-Mercury mission. The value is also usually the maximum C_3 found on Earth-Venus contour plots in publications such as Ref. 4.

The constraints are then: (1) energy match possible, (2) Venus altitude positive, (3) launch energy less than 21 km²/sec². These three constraints, when plotted on a grid of launch date versus Venus arrival date, result in a closed boundary defining the region of possible missions.

Because of the launch energy constraint, the Earth-Venus-Mercury opportunities will coincide with Earth-Venus opportunities, of which there are six in the 70's: 1970, 1972, 1973, 1975, 1977 and 1978. For each of these years, two closed regions exist inside which C_3 is less than 21 km²/sec². The two regions are known as Type I and Type II trajectories, having heliocentric transfer angles less than or greater than 180 deg, respectively.

For each date at Venus, there are similarly both Type I and Type II Venus-Mercury trajectories, each having a separate and distinct minimum energy. The energy match constraint must therefore be investigated separately for Type I and Type II Venus-Mercury legs.

For a given Type Venus-Mercury leg where an energy match is possible, two solutions will be found, denoted as Class I and Class II (see Ref. 4). The Class I solution has a shorter flight time than the Class II. The altitude constraint must be investigated separately for Class I and Class II trajectories. For each of the six launch years in the 70's, there are then eight possible solutions which may satisfy all three constraints, consisting of all combinations of:

(1) Type I and Type II Earth-Venus legs.

(2) Type I and Type II Venus-Mercury legs.

(3) Class I and Class II Venus-Mercury legs.

Computer runs were made in a parameter study which varies launch date and Venus arrival date at 1-day intervals. The time on all days is taken as 0^h UT. The steps in obtaining solutions are:

(1) Select a Venus arrival date.

(2) Compute minimum energies for Type I and Type II Venus-Mercury legs.

(3) Compute Earth-Venus legs at 1-day interval on launch date.

(4) Reject runs from Step 3 where $C_3 > 21$ km²/sec².

(5) Compare Venus arrival energy with minimums.

(6) For energies greater than minimum, compute Class I and Class II legs.

(7) Reject solutions with negative altitudes.

(8) Increment Venus arrival date by 1 day and repeat.

The number of rejected runs in Step 4 can be reduced by selecting the range of launch dates in Step 3 according to the 8-yr repetition of Earth-Venus opportunities (metonic cycle).

Table 1. Summary of results for opportunities in the 1970's

Launch year	Earth-Venus leg	Venus-Mercury leg			
		Type I		Type II	
		Class I	Class II	Class I	Class II
1970	Type I	++	+++	−	−
	Type II	0	0	−	−
1972	Type I	−	−	−	−
	Type II	−	+	−	−
1973	Type I	+++	++	−	−
	Type II	0	0	−	−
1975	Type I	0	0	+	−
	Type II	−	−	−	−
1977	Type I	−	−	−	−
	Type II	−	−	−	−
1978	Type I	0	0	−	−
	Type II	−	−	−	−

Key:
0 no energy match for $C_3 < 21$ km²/sec².
− all negative altitudes at Venus.
\+ small positive altitudes.
++ positive altitudes, short launch period.
+++ positive altitudes, good launch period.

2. Results

A summary of results is given in Table 1. Solutions were found in only four of the six launch years in the 70's. Energy matches are possible in 1977 and 1978, but all altitudes at Venus were negative. Of the four years where solutions were found, 1972 and 1975 have such low altitudes at Venus that these missions are less attractive. The most attractive Earth-Venus-Mercury missions in the 70's are in 1970 and 1973.

The region of possible missions for 1970, 1972, 1973 and 1975 are shown in Figs. 1 to 4, respectively. The

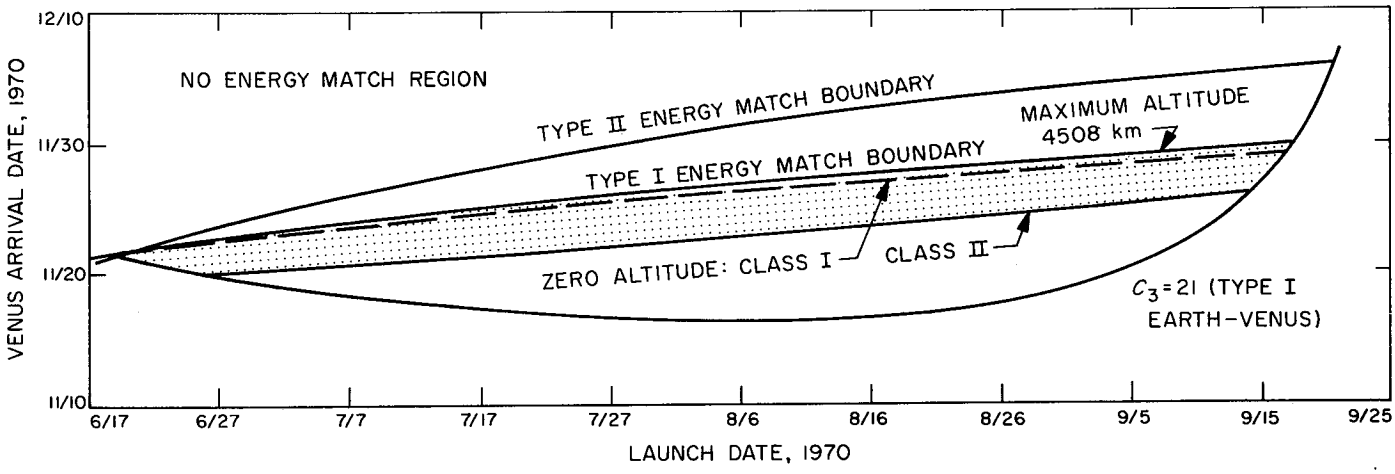


Fig. 1. Region of possible missions in 1970

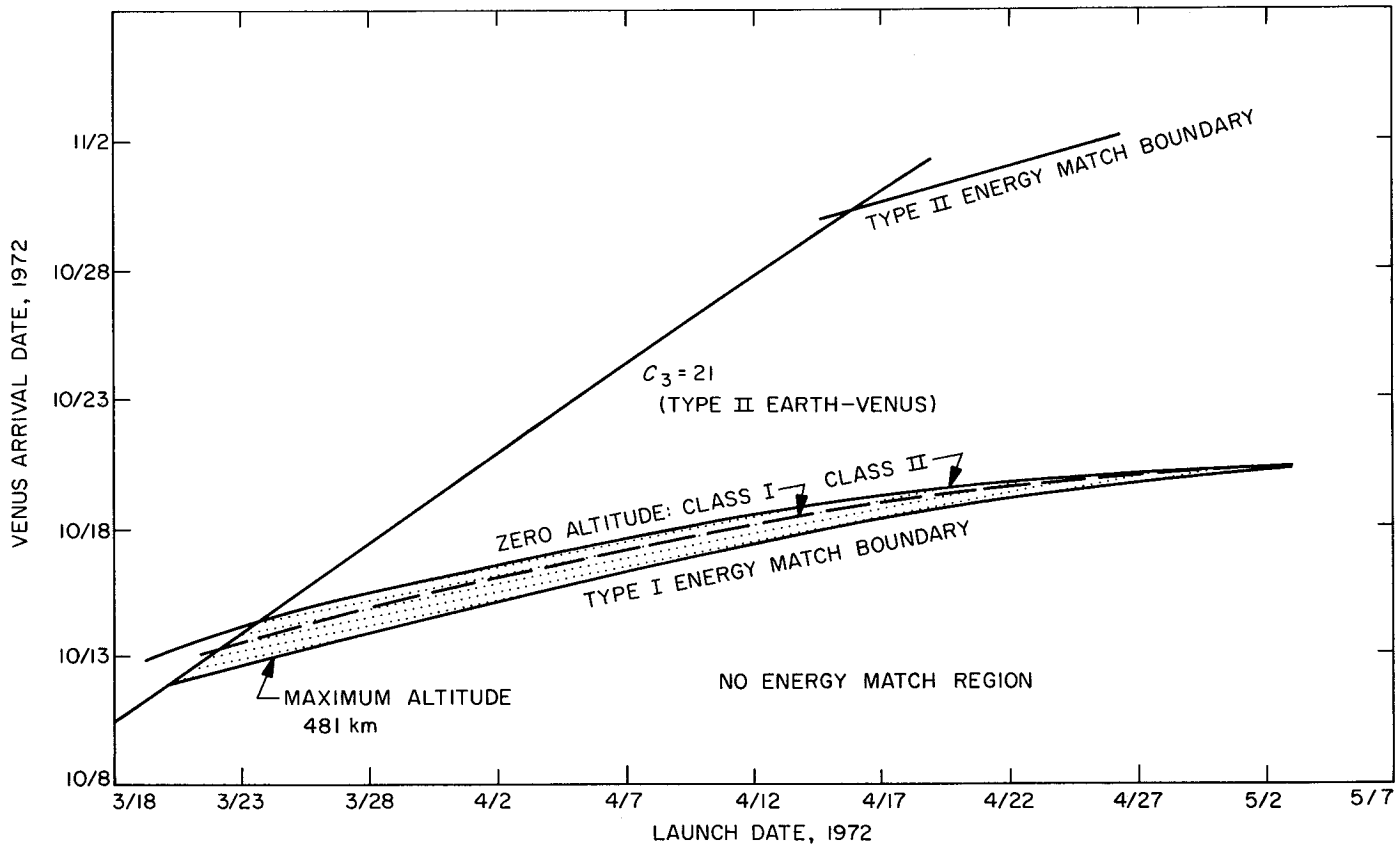


Fig. 2. Region of possible missions in 1972

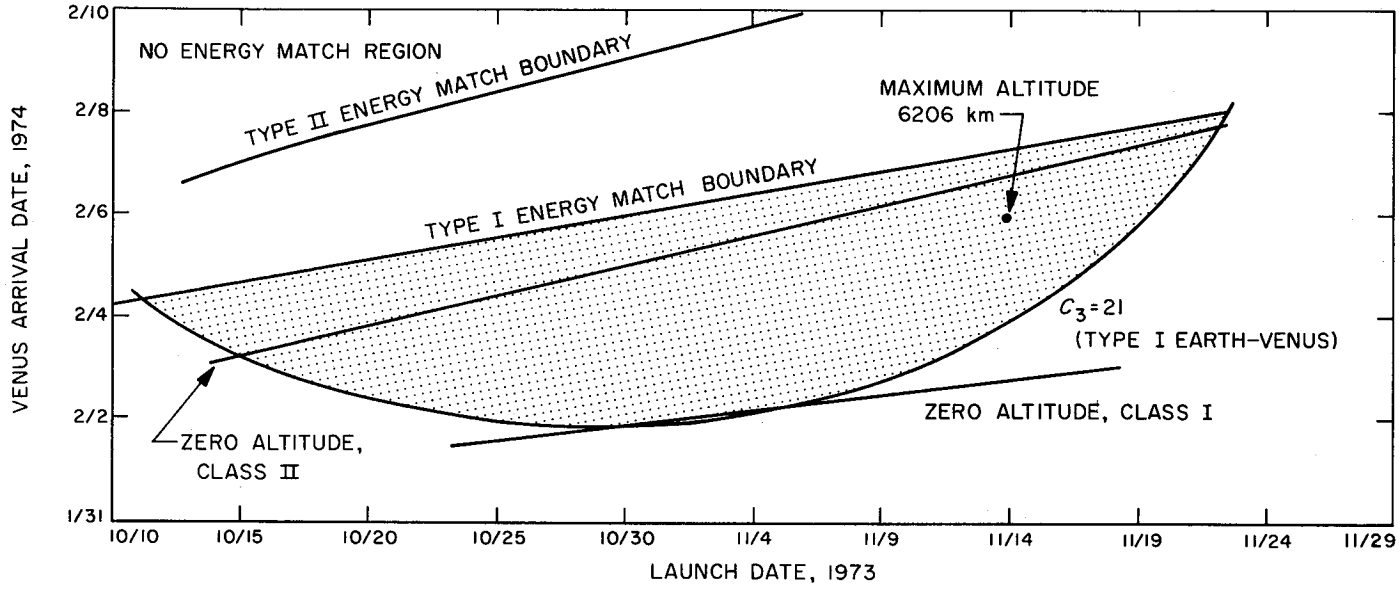


Fig. 3. Region of possible missions in 1973

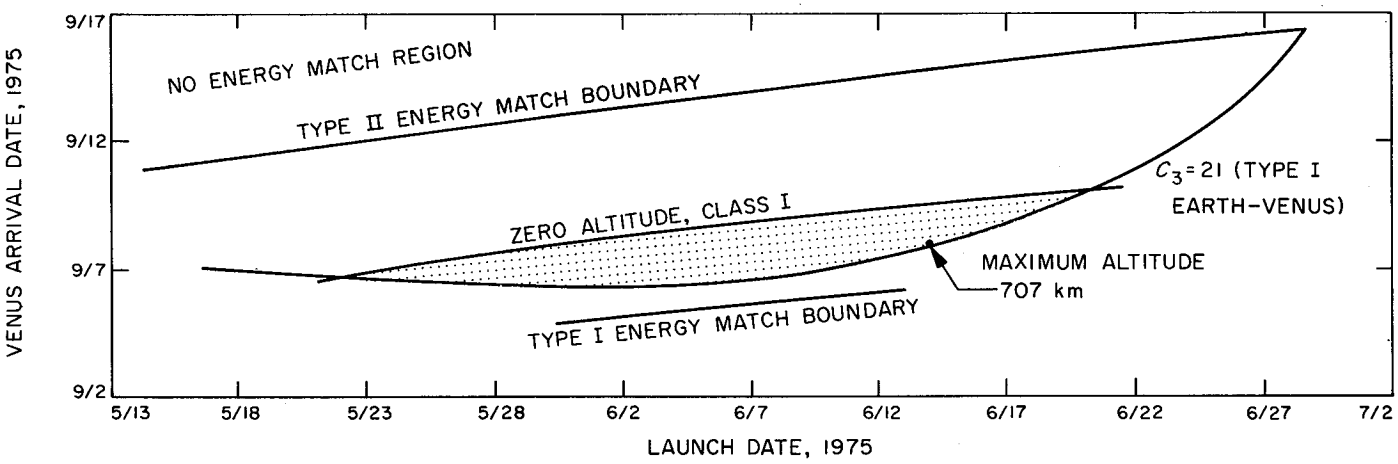


Fig. 4. Region of possible missions in 1975

figures show a grid of launch date versus Venus arrival date. On this grid are shown the boundaries corresponding to the three constraints, namely:

(1) Boundary of energy match.

(2) Locus of zero Venus altitude.

(3) Locus of $C_3 = 21$ km²/sec².

The energy match boundary is defined as the locus of points where the Venus arrival energy is just equal to the minimum energy for the Type I or Type II Venus-Mercury leg.

The shaded region is that satisfying all three constraints. For 1970 and 1972, the Class II shaded region overlaps the Class I region. For 1973, the Class I region overlaps the Class II region. In all three cases, the energy match boundary is common to both classes. On each figure, the location and value of the maximum Venus altitude for the whole region is called out.

3. Discussion

The results of this study are important, since any consideration to perform an Earth-Venus-Mercury mission must be limited to 1970 or 1973, or else postponed to the

80's. A survey of opportunities in the 80's will be conducted at some future date.

Efforts to find a repeating cycle of Earth-Venus-Mercury trajectories have not been successful. The planetary configuration approximately repeats every 8 yr. A comparison of results for 1970 and 1978, however, shows drastically different results. The high inclination, eccentricity, and mean motion of Mercury evidently make a very accurate repeat of planetary positions necessary to obtain similar results.

Both the 1970 and 1973 opportunities have Type I trajectories on both legs of the mission, resulting in the shortest flight times. Somewhat lower launch energies are available for 1970 than for 1973, but the altitudes are greater for 1973. In most regards the 1973 mission looks more favorable than 1970. Effort at JPL is being concentrated on the 1973 mission. A detailed trajectory and guidance study of the 1970 mission was reported in Ref. 2. A similar study for the 1973 mission has been completed and will be reported on in a JPL Technical Report in the near future.

B. Graphical Calculation of Ground Station/Satellite Viewing Characteristics

R. D. Bourke

This article discusses a recent graphical check of analytically-obtained ground station/satellite viewing characteristics. Previous reports (*SPS 37-34*, Vol. IV, pp. 3–9 and *37-37*, Vol. IV, pp. 11–14) have described an effort devoted to the study of satellite viewing characteristics in anticipation of a possible relay system for returning data from the surface of a distant planet (Ref. 5).

The viewing characteristics have been obtained to date by integration over a selected region of the satellite surface as discussed in *SPS 37-34*, Vol. IV, pp. 3–9. (It may be recalled that the satellite surface is, in general, a toroidal body of revolution on which the satellite always lies.) To graphically check these calculations, a model has been constructed and is shown in Fig. 5. The edges of the various ribs represent cross sections of the satellite surface. The particular orbit represented is that of the first Russian communications satellite Molniya with elements:

$$a = 4.18 \text{ planetary radii}$$

$$e = 0.74$$

$$i = 65.2 \text{ deg}$$

$$\omega = -36.5 \text{ deg}$$

An elevation angle mask consisting of a Lucite cylinder with a conical hole of appropriate half-angle is fixed to the surface of the planet and can be seen in the figure. The location of the ground station is indicated by the pin in the center of the mask. By sighting in along the conical hole, it is possible to ascertain the portion of the boundary of each of the ribs that can be seen from the ground station, and hence the latitude limits of visibility at each of the rib longitudes; i.e., ϕ_{11}, ϕ_{12} on the inner surface and ϕ_{21}, ϕ_{22} on the outer. Having obtained the latitudes, the entry and exit anomalies are available as $f(\phi_{11})$ etc., and consequently the fraction of time in view at the ith longitude may be calculated from:

$$v_i = \frac{t(f_{12}) - t(f_{11}) + t(f_{22}) - t(f_{21})}{T}$$

where T is the period. The overall approximate fraction may be obtained by summing over the n ribs spaced in longitude by Δl (equal to 30 deg in the model shown):

$$v = \frac{\Delta l}{2\pi} \sum_{i=1}^{n} v_i$$

This calculation procedure was applied to a ground station at 43.1°N latitude with minimum elevation angle 7.5 deg. The view fraction from the model was 0.49, whereas the analytically-based computer program gave 0.487. A completely independent simulation method gave 0.472 (Ref. 6).

Other graphical checks are possible. For example, by separately projecting the inner and outer satellite surfaces on to the surface of the planet and distorting this projection in the north-south direction to reflect different rates of change of satellite latitude with time, it is possible to measure directly the fraction of time in view (Ref. 7). A three-dimensional model, however, such as that shown in Fig. 5, has the advantage that it graphically displays the range characteristics as well.

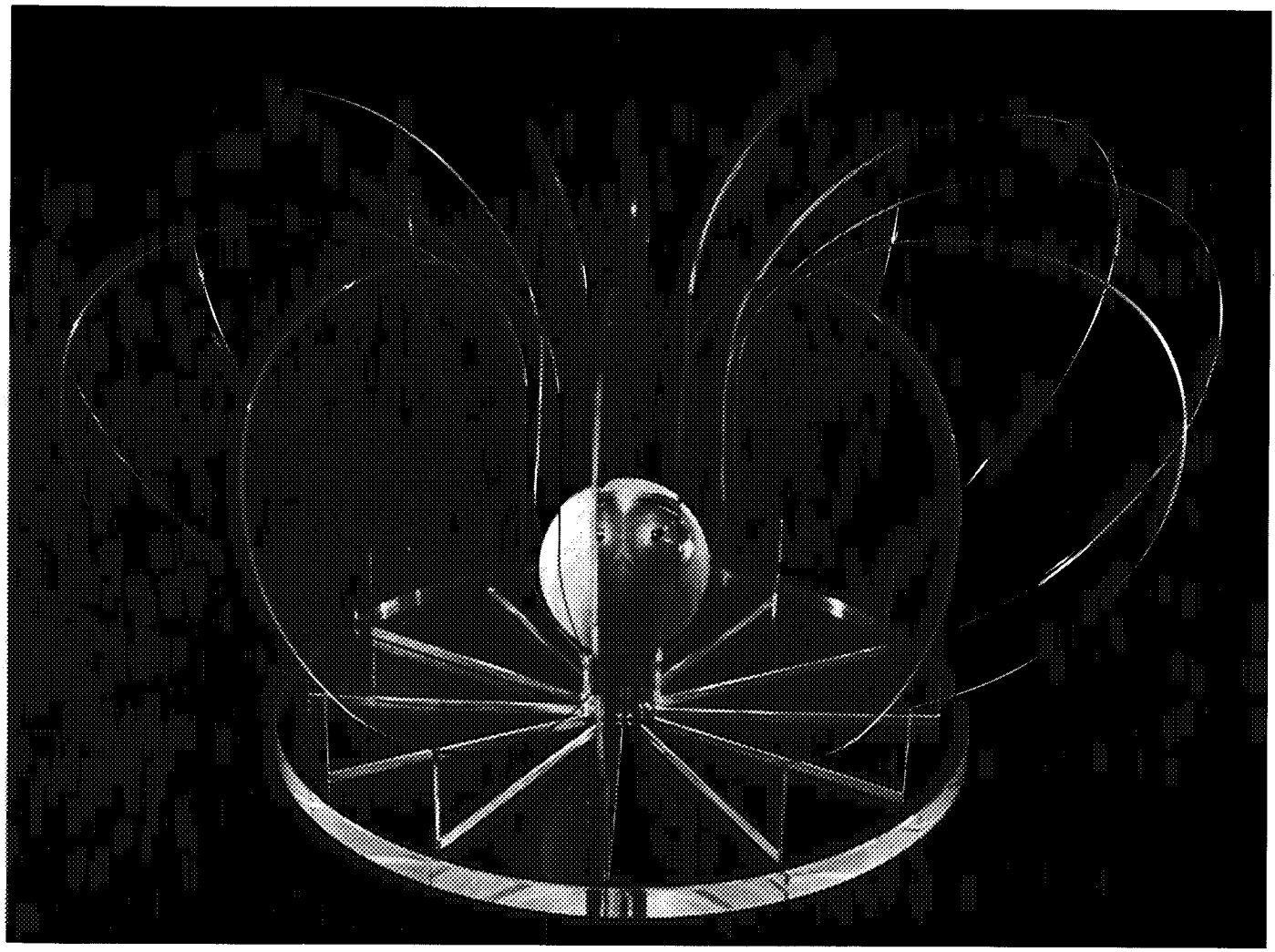


Fig. 5. Model of the surface on which the satellite always lies

It is the conclusion of the author that the validity of the satellite surface calculation is amply established by the results presented here.

C. Taylor Series Solution of the Equations of Motion in General Relativity

R. Broucke

It is well known that the solution of the equations of motion of a particle in the gravitation field of a central body can be expressed in terms of Weierstrass functions or Jacobi functions. The use of this solution for practical numerical calculations supposes that the necessary algorithms or subroutines to compute these functions are available.

Besides the existing solutions, we have developed the solution in simple recurrent Taylor series which are easy to manipulate numerically even on a small computer, and this with a high degree of accuracy. We have taken the polar angle ϕ as the independent variable and the radius r and time t as dependent variables.

Knowing that the solution is planar, we may restrict ourselves to the two-dimensional equations of motion which are derived from the Schwarzchild metric:

$$ds^2 = -\left[1 - \frac{2km}{c^2 r}\right]^{-1} dr^2 - r^2\, d\phi^2 + c^2\left[1 - \frac{2km}{c^2 r}\right] dt^2. \tag{1}$$

The equations of motion which are derived from the metric (1) have normally s as independent variable, but s may be easily eliminated and any other variable, for instance ϕ, may be used as independent variable. The final differential equations are then

$$\begin{aligned} \frac{d^2u}{d\phi^2} + u &= \frac{km}{h^2} + \frac{3km}{c^2}\,u^2, \\ \frac{dt}{d\phi} &= \frac{r^2}{h^2}\left[1 - \frac{2km}{c^2 r}\right]^{-1}. \end{aligned} \tag{2}$$

The variable u is the reciprocal of r, and h^2 is a constant of the motion, corresponding to the first integral

$$r^2\,\frac{d\phi}{dt} = h^2\left[1 - \frac{2km}{c^2 r}\right]. \tag{3}$$

In order to make the Taylor series solution more simple, we perform first a few modifications of the system (2). The last of the two equations shall be simplified and written in the form

$$\frac{dt}{d\phi} = \frac{r^2}{h^2}\left[1 + \frac{2km}{c^2 r}\right] = \frac{r^2}{h^2} + \frac{2km\,r}{c^2 h^2}. \tag{4}$$

The first equation (2) which is of second order shall be replaced by two first order differential equations, and a new unknown v is introduced.

$$\begin{aligned} \frac{du}{d\phi} &= v \\ \frac{dv}{d\phi} &= \frac{km}{h^2} + \frac{3km}{c^2}\,u^2 - u. \end{aligned} \tag{5}$$

Finally, a new first order differential equation is used to express the relationship between u and r:

$$u\,\frac{dr}{d\phi} = -\,rv. \tag{6}$$

The differential equations (4), (5) and (6) form now a system which is well prepared for a solution by Taylor series expansions in ϕ. We represent each of the four dependent variables u, v, t and r by a power series in ϕ in the following way:

$$\begin{aligned} u &= \sum_{n=1}^{\infty} U_n\,\phi^{n-1}, \\ v &= \sum_{n=1}^{\infty} V_n\,\phi^{n-1}, \\ t &= \sum_{n=1}^{\infty} T_n\,\phi^{n-1}, \\ r &= \sum_{n=1}^{\infty} R_n\,\phi^{n-1}. \end{aligned} \tag{7}$$

It is our intention to obtain for all the coefficients of the above infinite series simple expressions which are functions of the initial conditions only. It is mainly because we want these expressions to be simple and recurrent that we have taken the system in this redundant form. If we make the substitution of the series (7) in the differential equations, and if we set the coefficients of the term ϕ^{n-1} at the left side equal to the coefficient of the same power of ϕ at the right side, we then obtain a system of recurrent algebraic equations for these coefficients:

$$\begin{aligned} n\,U_{n+1} &= V_n, \\ n\,V_{n+1} &= \frac{3km}{c^2} \sum_{\substack{p=1,\,\ldots,\,n \\ q=n,\,\ldots,\,1 \\ p+q=n+1}} U_p\,U_q - U_n, \\ n\,T_{n+1} &= \frac{1}{h^2} \sum_{\substack{p=1,\,\ldots,\,n \\ q=n,\,\ldots,\,1 \\ p+q=n+1}} R_p\,R_q + \frac{2km}{c^2h^2}\,R_n, \\ n\,U_1\,R_{n+1} &= -\sum_{\substack{p=2,\,\ldots,\,n \\ q=n-1,\,\ldots,\,1 \\ p+q=n+1}} qU_p\,R_{q+1} - \sum_{\substack{p=1,\,\ldots,\,n \\ q=n,\,\ldots,\,1 \\ p+q=n+1}} R_p\,V_q \end{aligned} \tag{8}$$

The above system of equations for the coefficients is said to be recurrent because it is solved explicitly for the highest order terms $(n+1)$ on the left side. On the right side occur only coefficients up to the order n. The equations (8) allow thus to determine by recurrence all the coefficients up to the desired order, as functions of the zero-order coefficients only, which are nothing else than the initial conditions.

We have programmed the equations (8) for both the 1620 and 7094 computers, in double precision. We have restricted the series (7) to a maximum of 16 terms in most practical cases and for values of ϕ which do not exceed six degrees, the precision which is given by the truncated series (7) is of the order of twelve places, or better. Repeated use of the series over a small arc ϕ permits us to integrate the relativistic trajectory over several revolutions. We have tested the precision of the solutions by using the two well known first integrals of the motion, and also with the additional first integral which exists here because of the redundancy of our system

$$ur = 1. \tag{9}$$

The use of these relations between the variables gives us the possibility to make an estimate of the truncation error, although it does not determine what the magnitude of the error is in each variable.

D. On the Estimation of Random Spacecraft Accelerations

C. G. Pfeiffer

1. Introduction

The orbit determination program presently employed at the Jet Propulsion Laboratory is designed to obtain a minimum variance estimate of constant parameters. This form, sometimes called the "weighted least squares" estimator, is derived by assuming that the functional forms of the spacecraft accelerations are known, and that these accelerations can be completely characterized by a set of constant parameters, such as the coefficients of a polynomial in time. This is certainly a reasonable model for the gravitational accelerations, where the masses and orbital elements of the bodies of the solar system are the acceleration parameters to be estimated.

The nongravitational forces, such as arise from solar radiation pressure and gas leakage in the attitude control system, have been analyzed by Flandro (*SPS 37-30*, Vol. IV, pp. 6–14) and others, and it appears reasonable that the major portion of this type of acceleration might also be characterized by a deterministic model. Recent work by Null and Bouvier (Refs. 8, 9), however, indicates that the *Mariner IV* spacecraft was subject to random accelerations arising from unbalanced attitude control torques, and these accelerations apparently cannot be represented by any a-priori specified functional form. Although the accelerations were small, perhaps 10^{-10} g, their effect on the trajectory appears to be significant, perhaps several hundred kilometers error at Mars after a 6-month flight time. Effort is now underway to analyze these accelerations by assuming that they can be represented as polynomials in time, with undetermined constant coefficients, but it is likely that such a model will not be completely satisfactory. If so, very large estimation error might accrue, for Nishimura (Ref. 10) has analyzed a simple problem to show that the effect of employing the constant parameter weighted least squares estimator when random accelerations are indeed present can lead to an error in the estimate which goes to infinity with time. Thus it appears that random accelerations must be treated if the tracking data is to be put to the best possible use.

Consistent with the minimum variance estimation philosophy, one might treat the random accelerations as "process noise" with known statistics, and employ well-known results to derive the estimator (Ref. 11). The statistical data available on these accelerations is rather sparse, but it appears that a reasonable stochastic model can be constructed. Indeed, in Ref. 12, Friedland, Thau and Sarachik describe a simple, intuitively justified stochastic model which may be adequate to explain the data presented in Ref. 9. The statistics may be questionable, but such a model will certainly be superior to the one presently postulated, where it is assumed that the process noise has zero variance. It is the case that in the presence of measurement noise and/or process noise it is impossible to estimate the coordinates of a spacecraft exactly. The minimum variance estimator technique can be thought of as nothing more than a rational guess of the unknowns, given the best available a-priori information about the system. Considering the average error in this estimate over the ensemble of all missions described by the a-priori statistics, this estimator is the best possible. It may be that some other approach could yield an estimate which has better "worst case" behavior and is still unacceptable, but, in general, such estimators are hard to devise.

Thus, given an intuitively reasonable statistical description of the process and measurement noise, and assuming that the minimum variance estimate is sought, the theoretical derivation of the estimator is straightforward (Ref. 11). As a practical matter, however, the estimator is not easy to calculate, for one must first determine the optimal data weighting function by computing the covariance of the error in the estimate. This matrix has dimension equal to the number of parameters being estimated, and is specified as the solution of a non-linear differential equation (a matrix Ricatti equation) if the continuous form of the estimator is employed, or, for the discrete form, as a non-linear difference equation. In the present orbit determination program, where constant parameters are estimated, the difference equation for the inverse covariance matrix takes a simple form. That is, the covariance matrix at any time can be obtained by summing the so-called normal matrices, adding the inverse of the a-priori matrix, and inverting the sum. When random accelerations are present, however, this simple algorithm no longer applies, and it would appear that the non-linear Ricatti equation must be solved numerically. This is not an attractive alternative, for the dimension of the covariance matrix may be 50. This implies that the Ricatti equation may not be stable, excessive computer time and computer storage may be required, and numerical errors may be excessive. In any case, one would like to obtain an estimator

which is a simple extension of the present form. Intuitively, it seems possible to do this, because random accelerations applied in the form of impulses at prespecified times enter the estimation equations in precisely the same fashion as impulsive guidance maneuvers with random execution errors. Such maneuvers are presently treated by the simple device of adding an increment to the calculation covariance matrix at the time of application of the impulse, and apparently a similar approach could be employed for treating random accelerations. The device of incrementally increasing the covariance matrix elements at selected times in order to account for previously neglected uncertainties is well known, but the statistical assumptions implicit in this approach are often only heuristically stated. The purpose of this paper, then, is to develop an impulsive model of random accelerations; to relate this to the well-known continuous white noise model; to show how one obtains the discrete model statistics from the continuous model statistics, and to derive expressions for the covariance matrix and estimator which are simply applied extensions of the present weighted least squares form. Following Nishimura, we shall also obtain an expression for the true covariance of the error in the estimate which results if random acceleration are ignored, and the spacecraft coordinates are estimated by applying the weighted least squares form.

2. The Estimator and Its Covariance Matrix

The weighted least squares estimator of constant parameters is formulated as follows: Let $\delta\mathbf{q}$ be an n-dimensional parameter vector composed of trajectory initial condition variations and constants which affect the tracking data and spacecraft acceleration, and let the sole data type be a sequence of discrete "counted doppler" points of the form $\delta(\Delta\dot{\rho}_i)$, for $i = 1, 2, \cdots N$, where

$$\delta[\Delta\dot{\rho}_i] = A_i\,\delta\mathbf{q} + n_i \tag{1}$$

The $\rho(t)$ is the instantaneous station-spacecraft distance, $\delta(\cdots)$ refers to a variation from the standard trajectory at a fixed time t, $\Delta t_i = t_i - t_{i-1}$

$$[\Delta\dot{\rho}_i] = \left(\frac{1}{\Delta t_i}\right)\left[\int_{t_{i-1}}^{t_i} \dot{\rho}(t)\,dt\right] \tag{2}$$

$$A_i = \left(\frac{1}{\Delta t_i}\right)\int_{t_{i-1}}^{t_i}\left[\frac{\partial\dot{\rho}(t)}{\partial\mathbf{q}}\right]dt \qquad (A_i \text{ is a } 1 \times n \text{ matrix}) \tag{3}$$

and n_i is integrated white noise $(\mathrm{w}(\tau))$ divided by Δt_i, that is,

$$n_i = \left(\frac{1}{\Delta t_i}\right)\int_{t_{i-1}}^{t_i} w(t)\,dt \tag{4}$$

The variance of n_i is given by

$$E[n_i^2] = \left(\frac{1}{\Delta t_i}\right)^2\int_{t_{i-1}}^{t_i}\int w(\tau_1)\,w(\tau_2)\,d\tau_1\,d\tau_2 = \left(\frac{\sigma^2}{\Delta t_i}\right) \tag{5}$$

since $E[w(\tau_1)\,w(\tau_2)] = \sigma^2 \times$ (dirac delta function). The well-known expression for the covariance of the error in the linear, unbiased minimum variance (LUMV) estimate of $\delta\mathbf{q}$, given data up to the including time t_N is

$$\Lambda^*(t_N) = \left[\Lambda_0^{-1} + \sum_{i=1}^{N} G_i\right]^{-1} \tag{6}$$

where Λ_0 is the initial value of $\Lambda(t)$,

$$G_i = \left(\frac{\Delta t_i}{\sigma^2}\right)[A_i^T A_i] \tag{7}$$

Assuming that the a-priori estimate of $\delta\mathbf{q}$ is zero, the LUMV estimate of $\delta\mathbf{q}$ is

$$\delta\mathbf{q}^*(t_N) = [\Lambda^*(t_N)]\left\{\sum_{i=1}^{N}\left(\frac{\Delta t_i}{\sigma^2}\right)A_i^T\,\delta(\Delta\dot{\rho}_i)\right\} \tag{8}$$

Letting $\Delta t_i \to 0$ for all i, Eqs. (6) and (8) yield the continuous equations

$$\Lambda^*(t) = \left[\Lambda_0^{-1} + \left(\frac{1}{\sigma^2}\right)\int_0^t A^T(\tau)\,A(\tau)\,d\tau\right]^{-1} \tag{9}$$

$$\delta\mathbf{q}^*(t) = \Lambda^*(t)\left[\left(\frac{1}{\sigma^2}\right)\int_0^t A^T(\tau)\,\delta\dot{\rho}(\tau)\,d\tau\right]^{-1} \tag{10}$$

where $A(\tau) = \left[\dfrac{\partial\dot{\rho}(\tau)}{\partial\mathbf{q}}\right]$

Eq. (6) is the solution of the matrix Ricatti equation

$$\frac{d\Lambda^*}{dt} = -\Lambda^* G \Lambda^* \tag{11}$$

which is a special case of the more general form (Ref. 11)

$$\frac{d\Lambda^*}{dt} = F\Lambda^* + \Lambda^* F^T - \Lambda^* G A^* + H \tag{12}$$

where $F(t)$ and $H(t)$ are given time-varying matrices. The terms $(F\Lambda^* + \Lambda^* F^T)$ arise when one estimates a time-varying state vector $\delta\mathbf{q}(t)$, such as would occur

when the initial condition variations of the trajectory are continuously "epoch forwarded" to obtain the instantaneous coordinates of the spacecraft. The $H(t)$ term arises when there are random forces continuously perturbing the motion of the spacecraft, such as would be caused by leaks in the attitude control system, or by variations in the effective solar radiation pressure, or from apparent accelerations which arise from errors in calculating the spacecraft trajectory. Although the solution of (11) as given by (6) is easily extended to treat the $(F\Lambda^* + \Lambda^* F^T)$ terms if $H(t)$ is zero, the inclusion of the $H(t)$ term radically changes the character of $\Lambda^*(t)$, and no simple method for solving (12) by means of a quadrature and inverse is known. Thus, it would appear that (12) must be solved numerically if low thrust forces are to be treated, which may result in large numerical errors, and may lead to prohibitive computer requirements. The purpose of this memo is to suggest a simplified model of low thrust forces on a spacecraft, which leads to a form of the Ricatti equation which can be solved by a rather simple extension of the algorithm presently employed to obtain Eq. (6).

3. An Impulsive Model of Low Thrust Forces on the Spacecraft

Let $\delta\mathbf{x}(t)$ be the six-dimensional coordinate deviation vector of the spacecraft, that is, $\delta\mathbf{x}^T(t) = [\delta\mathbf{v}^T(t), \delta\mathbf{r}^T(t)]$, where $\delta\mathbf{v}(t)$ is the instantaneous deviation from standard of the velocity vector, and $\delta\mathbf{r}(t)$ is the instantaneous deviation from standard of the position vector. If $\delta\mathbf{z}(s)$ is the random acceleration continuously perturbating the spacecraft, we introduce the well-known state transition matrix $U(t,\tau)$ to obtain

$$\delta\mathbf{x}(t) = U(t,0)\,[\delta\mathbf{x}(0) + \delta\mathbf{p}(t)] \tag{13}$$

where

$$\delta\mathbf{p}(t) = U^{-1}(t,0)\int_0^t U(t,\tau)\,D(\tau)\,\delta\mathbf{z}(\tau)\,d\tau$$

$$= \int_0^t U^{-1}(\tau,0)\,D(\tau)\,\delta\mathbf{z}(\tau)\,d\tau \tag{14}$$

$$U(t,\tau) = \left[\frac{\partial\mathbf{x}(t)}{\partial\mathbf{x}(\tau)}\right] \tag{15}$$

$$D(t) = \left[\frac{\partial\dot{\mathbf{x}}(t)}{\partial\mathbf{z}(t)}\right] = \left[\begin{array}{c}\left[\frac{\partial\dot{\mathbf{v}}(t)}{\partial\mathbf{z}(t)}\right] \\ \hline 0\end{array}\right] \triangleq \left[\begin{array}{c}D_1(t) \\ \hline 0\end{array}\right] \tag{16}$$

In obtaining Eq. (14), we have used the property $U(t,0) = U(t,\tau)\,U(\tau,0)$. In Eq. (13), let us replace $\delta\mathbf{p}(t)$ with a sequence of delta functions which perturb the initial conditions, that is, define

$$\frac{d}{dt}\,[\delta\mathbf{x}(0)] = \sum_{i=1}^{\max i} \Delta\mathbf{p}_i \times [\text{dirac delta}\,(t_i - t)] \tag{17}$$

where $\{t_1, t_2, \cdots, t_i \cdots\}$ is some arbitrarily defined sequence of times, and

$$\Delta\mathbf{p}_i = \int_{t_{i-1}}^{t_i} U^{-1}(\tau,0)\,D(\tau)\,\delta\mathbf{z}(\tau)\,d\tau \tag{18}$$

Then in Eq. (13), we replace $\delta\mathbf{p}(t)$ with

$$\int_0^t \frac{d}{dt}\,[\delta\mathbf{x}(0)]\,d\tau = \sum_{i=1}^{N} \Delta\mathbf{p}_i \quad t_N \leq t < t_{N+1} \tag{19}$$

The $\Delta\mathbf{p}_i$ represent impulsive perturbations in the "effective" initial conditions corresponding to time t; Eqs. (17)–(19) describe the impulsive analogue of the model discussed in Ref. 13.

If $\delta\mathbf{z}(t)$ is zero-mean "white noise," with autocorrelation function

$$E[\delta\mathbf{z}(\tau_1)\,\delta\mathbf{z}(\tau_2)^T] = C(\tau_1) \times [\text{dirac delta}\,(\tau_1 - \tau_2)] \tag{20}$$

where $C(\tau)$ is a given 3×3 covariance matrix, then

$$E[\delta\mathbf{x}(t)] = 0 \tag{21}$$

$$E[\delta\mathbf{x}(t)\,\delta\mathbf{x}^T(t)] = U(t,0)\left\{E[\delta\mathbf{x}(0)\,\delta\mathbf{x}(0)^T] + \sum_{i=1}^{N} E[\Delta\mathbf{p}_i\,\Delta\mathbf{p}_i^T]\right\} U^T(t,0) \qquad t_N \leq t < t_{N+1} \tag{22}$$

where

$$E[\Delta\mathbf{p}_i\,\Delta\mathbf{p}_i^T] = \int_{t_{i-1}}^{t_i} U^{-1}(\tau,0)\,D(\tau)\,C(\tau)\,D^T(\tau)\,U^{-1}(\tau,0)^T\,d\tau \tag{23}$$

As a computational convenience, one might introduce the well-known "inversion by inspection" formula

$$U^{-1}(t,0) = \left[\begin{array}{c|c}\left[\frac{\partial\mathbf{r}(t)}{\partial\mathbf{r}(0)}\right]^T & -\left[\frac{\partial\mathbf{v}(t)}{\partial\mathbf{r}(0)}\right]^T \\ \hline -\left[\frac{\partial\mathbf{r}(t)}{\partial\mathbf{v}(0)}\right]^T & \left[\frac{\partial\mathbf{v}(t)}{\partial\mathbf{v}(0)}\right]^T\end{array}\right] \tag{24}$$

Better still, if t is small, we have (approximately)

$$U^{-1}(t,0) \cong \begin{bmatrix} I & 0 \\ -tI & I \end{bmatrix} \tag{25}$$

where I is the 3×3 identity matrix. Then if $C(t)$ and $D_1(t)$ are both constant over the interval (t_{i-1}, t_i), Eq. (23) becomes

$$E[\Delta\mathbf{p}_i\ \Delta\mathbf{p}_i^T] = \left[\begin{array}{c|c} [D_1CD_1^T]\ (t_i - t_{i-1}) & -[D_1CD_1^T]\left(\frac{t_i^2 - t_{i-1}^2}{2}\right) \\ \hline -[D_1CD_1^T]\left(\frac{t_i^2 - t_{i-1}^2}{2}\right) & [D_1CD_1^T]\left(\frac{t_i^3 - t_{i-1}^3}{3}\right) \end{array}\right] \tag{26}$$

4. The Covariance Matrix and Estimator for the Impulsive Model

Let us consider the continuous case corresponding to Eqs. (9) and (10), but where $\delta\mathbf{x}(t)$ is as given by Eq. (13). If the parameter vector to be estimated for the case of no random acceleration is

$$\delta\mathbf{q}^T = [\delta\mathbf{x}^T(0), \delta\mathbf{y}^T],$$

where $\delta\mathbf{y}$ is an $(n-6)$ dimensional vector of variations in the constants which affect the tracking data and spacecraft accelerations, then in the presence of low thrust forces we must treat

$$\delta\mathbf{q}_t^T = \left[\delta\mathbf{x}^T(0) + \sum_{i=1}^{N} \Delta\mathbf{p}_i^T, \delta\mathbf{y}^T\right] \qquad t_N \leq t < t_{N+1} \tag{27}$$

Thus, for all t, we have

$$\delta\dot{\rho}(t) = A(t)\ \delta\mathbf{q}_t + w(t) \tag{28}$$

where, as in Sect. 2,

$$A(t) = \left[\frac{\partial\dot{\rho}(t)}{\partial\mathbf{q}}\right] = \left[\frac{\partial\dot{\rho}(t)}{\partial\mathbf{q}_t}\right].$$

Note that $\delta\dot{\rho}(t)$ has jump discontinuities at times t_i, according to this model. The Ricatti equation becomes

$$\frac{d\Lambda^*}{dt} = -\Lambda^* G \Lambda^* + \sum_{i=1}^{\max i} \Delta\Lambda_i^* \times [\text{dirac delta}\ (t - t_i)] \tag{29}$$

where

$$\Delta\Lambda_i^* = \left[\begin{array}{c|c} E[\Delta\mathbf{p}_i\ \Delta\mathbf{p}_i^T] & 0 \\ \hline 0 & 0 \end{array}\right] \tag{30}$$

The solution of Eq. (29) is given by the recursive relationship

$$\Lambda^*(t) = \left\{[\Lambda^*(t_N^+)]^{-1} + \left(\frac{1}{\sigma^2}\right)\int_{t_N}^{t} A^T(\tau)\ A(\tau)\ d\tau\right\}^{-1} \tag{31}$$

$$t_N \leq t \leq t_{N+1}^-$$

$$\Lambda^*(t_{N+1}^+) = \Lambda^*(t_{N+1}^-) + \Delta\Lambda_{N+1}^* \tag{32}$$

where (t_{N+1}^-) and (t_{N+1}^+) correspond, respectively, to the instant prior to and immediately following the application of the random impulse $\mathbf{p}_{N+1}$. The solution for $\delta\mathbf{q}^*(t)$ is identical to Eq. (10), except that the $\Lambda^*(t)$ is as given above. Note that the estimate $\delta\mathbf{q}_t^*(t)$ has jump discontinuities at times t_i, according to this model, which represent the LUMV estimates of the $\Delta\mathbf{p}_i$.

The recursive relationships (31) and (32) might lead to large numerical errors and exorbitant computing time requirements if many impulses $\Delta\mathbf{p}_i$ were applied, for the inverse of $\Lambda^*(t)$ must be calculated at each such time. Suppose, however, that all the elements of $\Delta\Lambda_i^*$ are small relative to those of $\Lambda^*(t_i^-)$, so that (approximately)

$$[\Lambda^*(t_{N+1}^+)]^{-1} \cong [\Lambda^*(t_{N+1}^-)]^{-1} \times \{I - [\Delta\Lambda_{N+1}^*]\ [\Lambda^*(t_{N+1}^-)]^{-1}\} \tag{33}$$

where I is the $n \times n$ identity matrix, and $\Lambda^*(t_{N+1}^-)$ is given by Eq. (31) in terms of $\Lambda^*(t_N^+)$ and the integral of $[A^T(\tau)\ A(\tau)]$ between t_N and t_{N+1}. Thus, given $[\Lambda^*(0^+)]^{-1}$, it is possible to approximately obtain $[\Lambda^*(t)]^{-1}$ without performing any inverse operations by recursively applying Eq. (33).

5. The "Consider Option"

Suppose that the impulsive model of low thrust forces on the spacecraft applies, but we nevertheless ignore these accelerations and compute the estimate of $\delta\mathbf{q}$ from

Eqs. (9) and (10). In this part we shall derive the covariance of the error in the non-optimal estimate so obtained. This estimate we shall denote as $\delta\tilde{\mathbf{q}}_t^*$, and its incorrectly computed covariance matrix, Eq. (9), as $\tilde{\Lambda}^*$.

From Eqs. (9), (10), (27), and (28) we have for $t_N \leq t \leq t_{N+1}^-$

$$[\tilde{\Lambda}^*(t)]^{-1}\,[\delta\tilde{\mathbf{q}}_t^*(t) - \delta\mathbf{q}_t] = \left(\frac{1}{\sigma^2}\right)\int_0^t A^T(\tau)\,\delta\dot{\rho}(\tau)\,d\tau$$

$$- [\tilde{\Lambda}^*(t)]^{-1}\,\delta\mathbf{q}_t = \left(\frac{1}{\sigma^2}\right)\int_0^t A^T(\tau)\,A(\tau)\,\delta\mathbf{q}_t\,d\tau$$

$$+ \left(\frac{1}{\sigma^2}\right)\int_0^t A^T(\tau)\,w(\tau)\,d\tau - [\tilde{\Lambda}^*(t)]^{-1}\,\delta\mathbf{q}_t \tag{34}$$

Setting $\Delta\mathbf{q}_i^T = [p_i^T, 0]$, and applying Eq. (27), the sum of the first and last terms on the right-hand side of Eq. (34) is

$$\left(\frac{1}{\sigma^2}\right)\left\{\int_0^t A^T(\tau)\,A(\tau)\,\delta\mathbf{q}_0\,d\tau + \sum_{i=1}^{N}\int_{t_i}^{t} A^T(\tau)\,A(\tau)\,\Delta\mathbf{q}_i\,d\tau\right\}$$

$$- [\tilde{\Lambda}^*(t)]^{-1}\,\delta\mathbf{q}_t = \left(\frac{1}{\sigma^2}\right)\left\{\int_0^t A^T(\tau)\,A(\tau)\,\delta\mathbf{q}_t\,d\tau\right.$$

$$\left. - \sum_{i=1}^{N}\int_0^{t_i} A^T(\tau)\,A(\tau)\,\Delta\mathbf{q}_i\,d\tau\right\} - [\tilde{\Lambda}^*(t)]^{-1}\,\delta\mathbf{q}_t$$

$$= -\left\{\Lambda_0^{-1}\,\delta\mathbf{q}_0 + \sum_{i=1}^{N}\left[\Lambda_0^{-1} + \left(\frac{1}{\sigma^2}\right)\int_0^{t_i} A^T(\tau)\,\tilde{A}(\tau)\,d\tau\right]\Delta\mathbf{q}_i\right\}$$

$$= -\left\{\Lambda_0^{-1}\,\delta\mathbf{q}_0 + \sum_{i=1}^{N}[\tilde{\Lambda}^*(t_i)]^{-1}\,\Delta\mathbf{q}_i\right\} \tag{35}$$

Add the term

$$\left\{\left(\frac{1}{\sigma^2}\right)\int_0^t A^T(\tau)\,w(\tau)\,d\tau\right\}$$

to Eq. (35), pre-multiply by $\tilde{\Lambda}^*(t)$, square the result, and take the ensemble average to obtain

$$\Lambda^*(t)_{true} = E[\delta\tilde{\mathbf{q}}_t^*(t) - \delta\mathbf{q}_t]\,[\delta\tilde{\mathbf{q}}_t^*(t) - \delta\mathbf{q}_t]^T$$

$$= \tilde{\Lambda}^*(t) + [\tilde{\Lambda}^*(t)]\sum_{i=1}^{N}[\tilde{\Lambda}^*(t_i)]^{-1}$$

$$\times E\,[\Delta\mathbf{q}_i\,\Delta\mathbf{q}_i^T]\,[\tilde{\Lambda}^*(t_i)]^{-1}\,[\tilde{\Lambda}^*(t)] \tag{36}$$

The continuous form of the difference between $\Lambda^*(t)_{true}$ and $\tilde{\Lambda}^*(t)$ is obtained from Eq. (36) by letting $\Delta t_i \to 0$, where $\Delta t_i = t_i - t_{i-1}$. Thus

$$R(t) = \Lambda^*(t)_{true} - \tilde{\Lambda}^*(t) = \tilde{\Lambda}^*(t)\int_0^t [\tilde{\Lambda}^*(\tau)]^{-1}$$

$$\times \left[\begin{array}{c|c} M(\tau) & 0 \\ \hline 0 & 0 \end{array}\right][\tilde{\Lambda}^*(\tau)]^{-1}\,\tilde{\Lambda}^*(t)\,d\tau \tag{37}$$

where the 6×6 matrix $M(\tau)$ is given by

$$M(\tau) = [U^{-1}(t,0)\,D(\tau)\,C(\tau)\,D^T(\tau)\,U^{-1}(\tau,0)^T] \tag{38}$$

and $U^{-1}(\tau,0)$, $D(\tau)$, and $C(\tau)$ are as given in Sect. 3. The $R(t)$ obtained above is essentially the same as Eq. (104) of Ref. 10.

One should expect the magnitude of certain elements of $R(t)$ to go to infinity with time, for suppose we diagonalize $\tilde{\Lambda}^*(t)$ and assume that the diagonal terms are (at least approximately) given by

$$\tilde{\lambda}_{ii}^*(t) = \frac{\tilde{\lambda}_{ii}^*(0)}{1 + \alpha_i t} \tag{39}$$

where α_i is a positive constant. Then the iith element of $R(t)$, for example, is given by

$$r_{ii}(t) = \int_0^t \left(\frac{1 + \alpha_j\tau}{1 + \alpha_j t}\right)^2 m_{ii}(\tau)\,d\tau$$

$$\geq \left[\frac{m_{ii}(\tau)_{min}}{(3\alpha_i)\,(1 + \alpha_i t)^2}\right]\left[(1 + \alpha_i t)^3 - 1\right] \tag{40}$$

which goes to infinity with time if $[m_{ii}(\tau)_{min}]$ is greater than zero. This phenomenon is demonstrated by analysis of a simple example in Ref. 10.

6. Discussion

We have developed an impulsive model of low-thrust forces on a spacecraft, and have shown that the resulting orbit determination equations can be solved by what appears to be a simple extension of the presently employed "weight least squares" algorithm. It remains to be seen if the method will be practical to apply. The initial condition impulses $\Delta\mathbf{p}_i$ have no meaningful physical interpretation, but, when the estimate of the initial condition is mapped to the final time to obtain a prediction of target error, the transformed impulses represent perturbations in that prediction due to low thrust forces acting on the spacecraft during tracking. It was noted

that this approach yields a jump discontinuity in the estimate at the times that the impulses are assumed to occur, but this troublesome behavior becomes a problem only if one insists upon the continuous form of the estimator. In practice, one would deal with counted doppler $[\Delta\rho_i/\Delta t_i]$ and the discontinuity in the state vector would be absorbed into the discontinuity in the data. Indeed, it would seem reasonable to let the intervals between low-thrust impulses correspond to the intervals between data points, but this is not essential. Further work is called for to investigate the validity of the impulsive model for representing the actual low thrust forces on a spacecraft, and to determine how the parameters in the model relate to observed behavior of the spacecraft.

References

1. Minovitch, M. A., *The Determination and Characteristics of Ballistic Interplanetary Trajectories Under the Influence of Multiple Planetary Attractions,* TR 32-464, Jet Propulsion Laboratory, Pasadena, Calif., October 31, 1963.
2. Sturms, F. M., Jr., and Cutting, E., *Trajectory Analysis of a 1970 Mission to Mercury Via a Close Encounter with Venus,* 2nd AIAA Aerospace Sciences Meeting (Paper 65-90), New York, January 25-27, 1965.
3. *Explanatory Supplement to the Ephemeris,* Her Majesty's Stationery Office, London, 1961.
4. Clarke, V. C., Jr., et al., *Design Parameters for Ballistic Interplanetary Trajectories, Part 1. One-Way Transfers to Mars and Venus,* TR 32-77, Jet Propulsion Laboratory, Pasadena, Calif., January 16, 1963.
5. Barber, T. A., Billy, J. M., Bourke, R. D., *Systems Comparison of Direct and Relay Link Data Return Modes for Advanced Planetary Missions,* TM 33-228, Jet Propulsion Laboratory, Pasadena, Calif., February 15, 1966.
6. Sugai, I., *personal communication,* System Sciences Corp., Fall Church, Virginia, 1966.
7. Heppe, R., "Graphic Methods for Calculating Coverage Attainable with Communication Satellites," *Electrical Communication,* Vol. 39 (No. 1): pp. 132–143, 1964.
8. Null, G., *Current State of the Art Capability to Control and/or Measure Solar Pressure and Attitude Control Translational Forces as Demonstrated by Mariner IV,* TM 312-626, Jet Propulsion Laboratory, Pasadena, December 9, 1965.
9. Bouvier, H. K., *Mariner IV Disturbance Torques,* IOM 344-543, Jet Propulsion Laboratory, Pasadena, September 24, 1965.
10. Nishimura, T., *Evaluation of Error Caused by Application of Non-Optimal Estimators to the Orbit Determination of Low-Thrusted Spacecraft,* TM 312-683, Jet Propulsion Laboratory, Pasadena, April 21, 1966.
11. Nishimura, T., *Continuous Estimation of Time Varying Parameters of Low-Thrusted Spacecraft,* TM 312-667, Jet Propulsion Laboratory, Pasadena, March 4, 1966.

References (Cont'd)

12. Friedland, B., Thau, F. E., and Sarachik, P. E., *Stability Problems in Randomly Excited Dynamic Systems,* paper to be presented at 1966 Joint Automatic Control Conference, Seattle, Washington, August 17–19, 1966.

13. Nishimura, T., *Continuous Estimation of Injection Conditions of Low-Thrust Spacecraft,* TM 312-679, Jet Propulsion Laboratory, Pasadena, April 11, 1966.

GUIDANCE AND CONTROL DIVISION

II. Spacecraft Power

A. Applied Thermionic Research

O. S. Merrill

1. Introduction

The results presented here are a summary of work performed between May 13, 1965, and April 1, 1966, under JPL Contract Number 951225 with Electro-Optical Systems, Inc. (EOS), Pasadena, California. This effort, entitled "High-Performance Thermionic Converter," is a program to (1) investigate basic materials, processes, and operating parameters affecting the stability and optimization of cesium-vapor thermionic converters, and (2) apply the results of these investigations to the fabrication of practical, high-performance, high-efficiency, long-life cesium-vapor thermionic converters. The stated program goals are: (1) to generate fundamental data on the various converter operational parameters such as interelectrode spacing, Langmuir-Taylor type cesiated electron emission, and work function of various electrode materials, and to establish optimum electrode materials processing, all to be applicable to practical cesium-vapor thermionic converters; (2) to conduct auxiliary experiments pertinent to the engineering design and converter fabrication in a manner such that the results are applicable to practical, high-efficiency, long-life cesium-vapor converters; and (3) to design, fabricate, and test a maximum of six cesium-vapor thermionic converters utilizing the results of the auxiliary experiments, leading to a performance of 20 watts/cm^2 at 0.8 volt output and an efficiency exceeding 14% for an emitter of 1735°C.

2. Variable Parameter Test Vehicle

The first major task of the EOS effort was the design, fabrication, and utilization of a variable parameter test vehicle in performing parametric studies of different emitter and collector materials at various interelectrode spacings, emitter temperatures, collector temperatures and cesium reservoir temperatures. (See Fig. 1.) It was originally intended that the electrodes in the test vehicle would be interchangeable. Due to technical difficulties encountered, it was decided that separate test vehicles should be fabricated for only two of the four pairs of electrode materials. Consequently, instead of evaluating the four sets of electrodes as originally intended, viz.,

Emitter	Collector
Tantalum	Molybdenum
Rhenium	Molybdenum
Tantalum	Tantalum
Rhenium	Rhenium

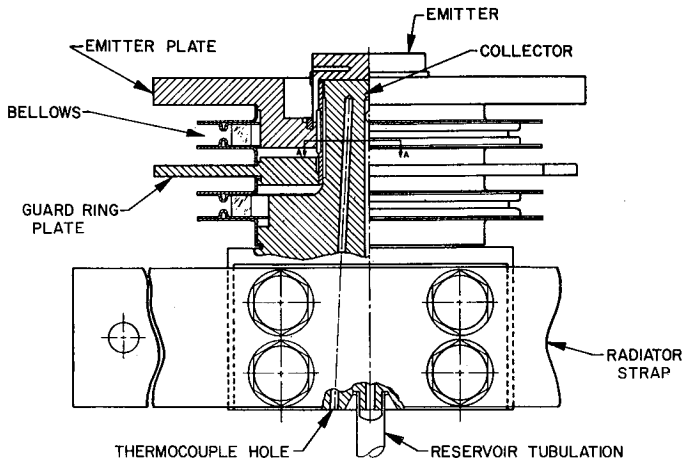


Fig. 1. Variable parameter test vehicle

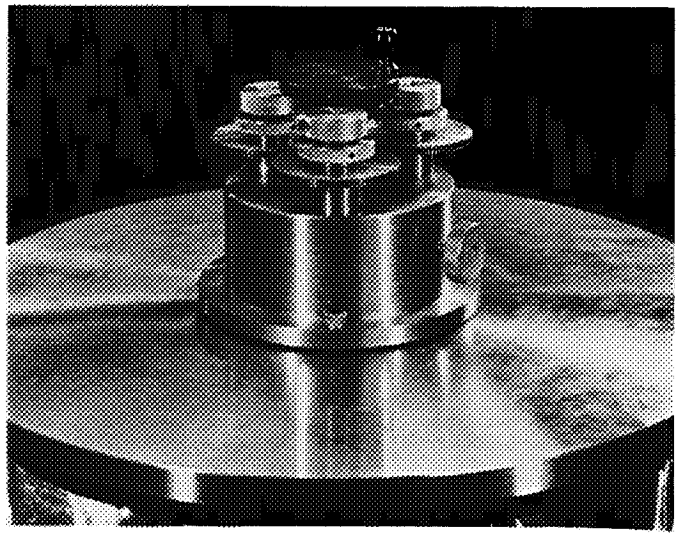


Fig. 2. Test vehicle drive mechanism

only the rhenium-rhenium electrode pair has been investigated at the time of this reporting. It is intended that the rhenium-molybdenum electrodes also be investigated before the end of the program, but the other two combinations have been dropped.

One of the most significant aspects of the design of the variable parameter test vehicle is its ability to measure interelectrode spacings to an accuracy of 0.1 mil, (0.0001 in.), an accuracy of measurement not heretofore achieved in a variable spacing test vehicle. The emitter-collector spacing is varied by applying force to three spring-loaded rods which are inserted from the belljar top plate. These rods apply pressure to the emitter plate and are controlled by a differential thread drive to allow independent movement of each rod. One complete rotation of the control nut on the individual rods results in a translational movement of 0.002 in. The three rods can be ganged for uniform motion controlled with a single drive. One complete rotation of the main control knob results in a uniform movement of 0.0008 in. by the ganged rods. The spacing measurement device is temperature-compensated, which allows it to indicate the true mechanical movement of the emitter with respect to the collector. The measuring device is a dial-gage depth indicator accurate to 0.0001 in. The test vehicle drive mechanism is shown in Fig. 2, and the test vehicle in the instrumented vacuum chamber is shown in Fig. 3.

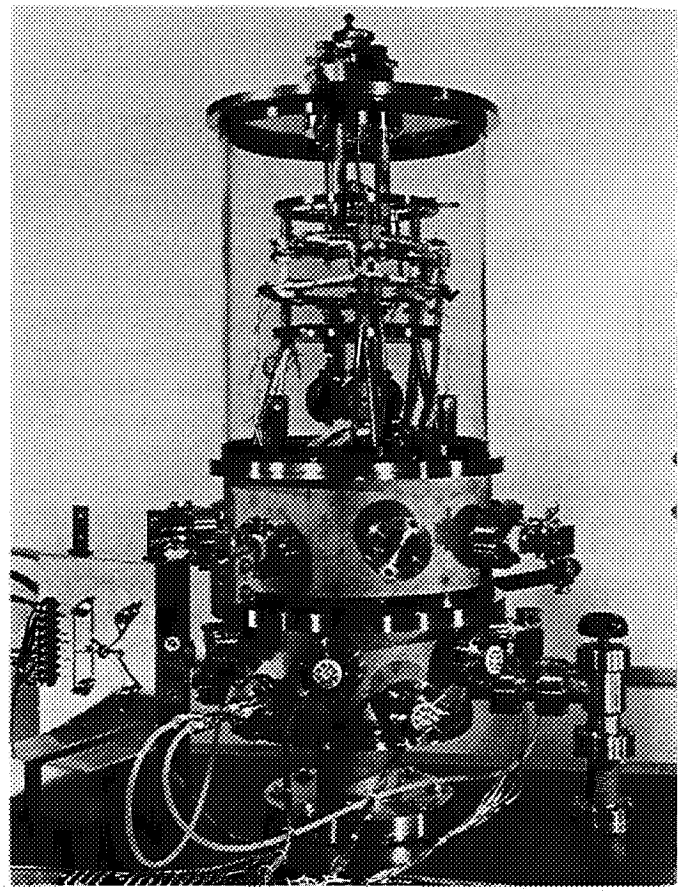


Fig. 3. Test vehicle drive mechanism, test circuitry, and instrumented vacuum chamber

3. *Results From the Rhenium-Rhenium Variable Parameter Test Vehicle*

At the time of this reporting, 1108 hr of operational test time have been logged on the rhenium-rhenium test

vehicle with no observed performance degradation during that time. The vehicle was operated at a variety of interelectrode spacings and temperatures as follows:

$T_{emitter}$ (T_e) 700°C to 2050°C

$T_{reservoir}$ (T_r) 122°C to 450°C

$T_{collector\ root}$ (T_c) 560°C to 685°C

Spacing 0.0001 in. to 0.015 in.

$I_{saturation}$ up to 180 amp/cm²

DC voltage was measured using an 0.5% accurate millivolt meter across a 0.1% accurate shunt. The vehicle ran in an ion-pumped vacuum environment at pressures of 10^{-7} torr or less. Table 1 gives results that are typical of the test data, illustrating the absence of degradation.

Table 1. Results from the rhenium-rhenium variable parameter thermionic test vehicle, illustrating the absence of degradation with time

Spacing, in.	Original voltage at 38A	875 hr, voltage at 38A	1108 hr, voltage at 38A
0.0020	0.780	0.760	0.790
0.0025	0.795	—	0.798
0.0030	0.802	0.792	0.802
0.0035	0.808	0.798	0.800
0.0040	0.805	—	0.805
0.0045	—	0.790	—
0.0050	0.788	—	0.790

T_R = 331°C, T_C = 610°C, T_E = 1735°C true, I = 38 amp.

Figs. 4 and 5 show the voltage-distance relationship at a constant current of 60 amp and 38 amp, respectively, for the cesium temperatures, collector root temperatures, and emitter temperatures shown. Similar data were taken for variable current at constant voltages of 0.6, 0.7 and 0.8 V as well as data for variable voltages for a constant current of 70 amp. The significance of this data is that there exists an optimum spacing, for a given set of electrodes and given operating conditions, which is not necessarily an extremely close spacing. In both Figs. 4 and 5, it can be observed that a spacing of 3 to 4 mils yields the optimum output in the arc-mode operation, indicated on these curves by the region to the right of the minimum. The region to the left of the minimum is the space-charge region, and the minimum itself represents the transition point between these two operational modes. Hence, for practical converter design, it is unnecessary to go to the extremely close spacing of 1 mil or less which one would have to do to achieve a comparable power output to that obtainable at the optimum 3- to 4-mil spacing.

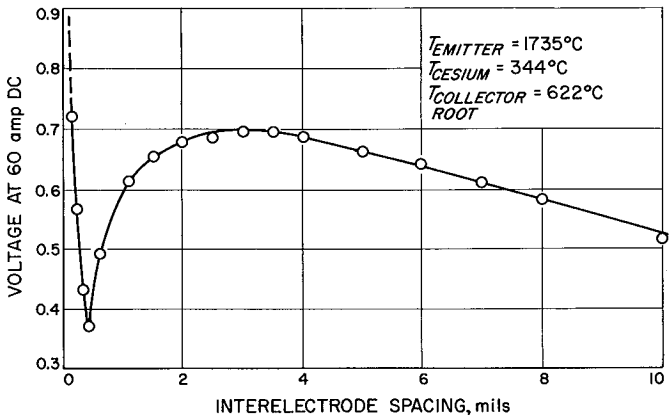


Fig. 4. Interelectrode spacing versus voltage output at a constant current of 60 amp

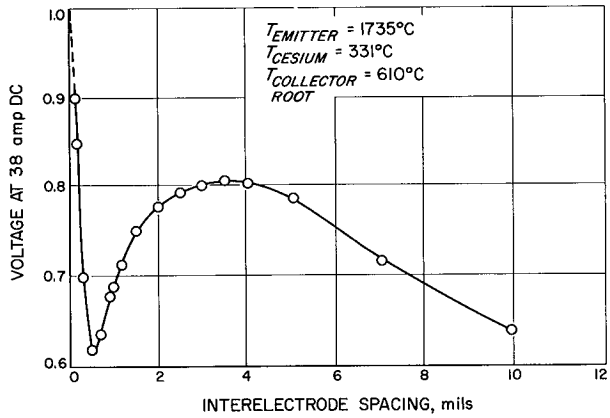


Fig. 5. Interelectrode spacing versus voltage output at a constant current of 38 amp

4. Converter Design and Testing

The second major task of the EOS effort was the design, fabrication and testing of several fixed parameter converters. The first and only such converter built to the date of this reporting, converter SN-101, was built to reproduce the performance from the rhenium electrode system studied in the variable parameter vehicle. This, in fact, is the principal goal of the program: to reproduce research data in practical hardware. To this end, converter SN-101 was highly successful since it reproduced the performance from rhenium emitter-rhenium collector test vehicle within experimental accuracy.

Concurrent with the test vehicle effort, feasibility studies were conducted to evolve advanced and reliable techniques for fabricating high-performance converters.

Three of these techniques were used to build SN-101. They were: (1) the electron-beam welding of a rhenium envelope to an all-rhenium emitter; (2) the fabrication of a long-life, high-temperature, ceramic-metal seal; and (3) the electron-beam welding of prefabricated seals and subassembly structures as the final assembly procedure before exhaust.

The rhenium emitter for SN-101 was electron-beam welded directly to a rhenium envelope. The weld was a pierce weld which requires no special geometry or consumable joint designs; instead, the emitter is slipped into the envelope, held in place by spring-loading, and directly welded.

High-temperature ceramic-metal seals have been tested on this program which have withstood several hundred thermal cycles at 100°C/min and remained leak-tight. After thermal cycling, a randomly selected seal was operated at 700°C for over 1500 hr. It has remained leak-tight, as determined by a 5×10^{-10} cc-atm/sec calibrated leak detector. These seals were prefabricated and may be selected either for testing or for converter fabrication.

The final assembly of SN-101 consisted of electron-beam welding a prefabricated seal assembly to collector and emitter subassemblies. After leak-checking and determining that the welded areas were leak-tight, the rhenium emitter was attached as previously described.

It was established from the variable parameter test vehicle that the maximum power output from a practical rhenium-rhenium system at 0.8 V output occurred at 3.5 mils spacing; SN-101 was designed for this spacing. It was experimentally determined that, with the emitter and collector in contact and stress-free, the converter structure, upon reaching operating temperature, experienced a differential expansion of 1.8 mils. The additional 1.7 mils was obtained by grinding away the rhenium surface, leaving four small 1.7-mil high pegs, permitting the emitter to be brought in contact with the collector at room temperature and welded in place to insure parallelism. At operating temperature, 99% of the area emits current at an interelectrode spacing of 3.5 mils; the remaining 1%, the surface area of the ends of the pegs, emits at a spacing of 1.8 mils.

By way of this technique, any spacing greater than 1.8 mils could be achieved in the SN-101 design. To fabricate at closer than 1.8 mils would require seal-brazing the emitter and collector subassemblies in a stressed condition. In addition to suffering the power output degradation at 2 mils spacing, the advantages of prefabrication and beam welding would be lost.

Since the test vehicle contained a guard ring to define the emitter area and a converter does not, it was necessary to design a converter to minimize side wall emission from the envelope to obtain a definitive correlation of data from the two devices. Converter SN-101 was designed with an 11-mil spacing between the collector and the envelope, thus essentially eliminating emission to the collector.

Converter SN-101 has a collecting area opposite the emitter proper of 1.88 cm^2. It has a 2.0 cm^2 emitter using the convention of the inside diameter of the envelope as the diameter of the emitter.

5. SN-101 Performance and Correlation to the Re-Re Test Vehicle Data

Fig. 6 is a performance plot of the dc output from SN-101 under the test conditions of:

(1) An emitter temperature of 1735°C (true) blackbody hole temperature.

(2) Potential measurement leads placed at the converter terminals.

(3) All data points recorded under steady-state, dc conditions.

A comparison of the performance of converter SN-101 and the rhenium-rhenium test vehicle is shown in Table 2. Reservoir temperatures have been included on this chart as an indication that the interelectrode spacing in the two devices was nearly identical. Converter SN-101 exhibited a temperature drop of 178°C from the collector surface to the collector root as compared to a calculated drop of 163°C at a current output level of 65 amp. No heat input was required to optimize performance output at 0.7 V, nor were heavy lead straps required for artificial cooling.

6. Miscellaneous Measurements from SN-101

Bare work function, or more properly, minimum coverage work function data were obtained on the emitter from SN-101. The observed values were 4.80 to 4.85 eV as compared to 4.75 eV from the guard-ringed test vehicle.

Table 2. Performance comparison of SN-101 and rhenium-rhenium test vehicle (no sidewall emission)

SN-101-emitter area of 2 cm²			SN-101 collector area of 1.88 cm²			Re-Re test vehicle-collector and guard-ring area of 2 cm²		
Voltage output, volts	Power output watts/cm²	Cesium reservoir temp, °C	Voltage output, volts	Power output, watts/cm²	Cesium reservoir temp, °C	Voltage output, volts	Power output, watts/cm²	Cesium reservoir temp, °C
1.0	7.5	319	1.0	7.9	319	—	—	—
0.8	15.2	330	0.8	16.1	330	0.8	16.1	331
0.7	21.0	343	0.7	22.2	343	0.7	21.3	344
0.6	24.6	351	0.6	26.0	351	0.6	25.2	351
0.5	25.0	361	0.5	26.6	361	—	—	—

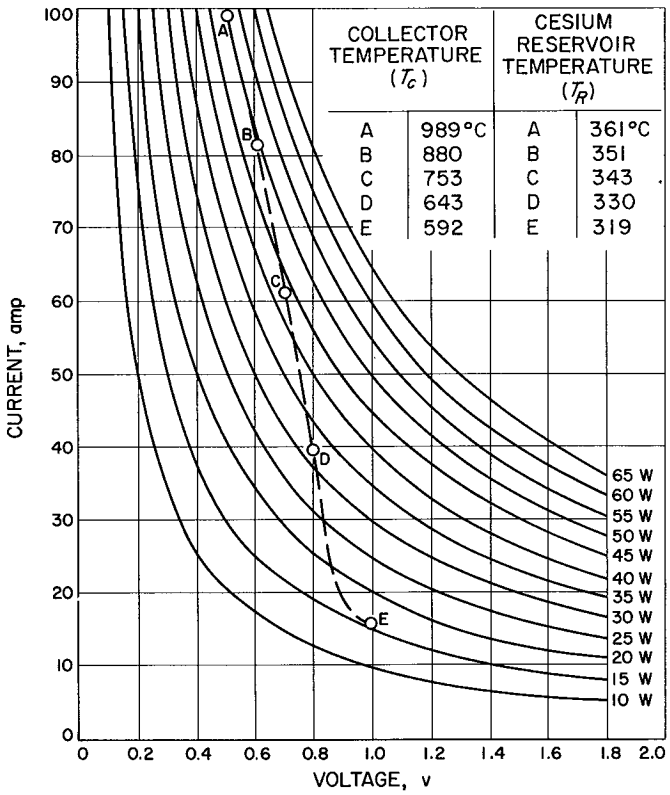


Fig. 6. Converter SN-101 performance at 1735°C emitter temperature

Cesium conduction measurements were taken and are presently being reduced in an attempt to relate the inter-electrode spacing of SN-101 to the phenomenon of gas atom conduction.

Saturated electron emission data were also taken from the emitter of SN-101; and, at an emitter temperature of 1735°C (true) hohlraum and at 330°C cesium reservoir, 27 amp/cm² of saturated electron current were recorded, compared to 30 amp/cm² from the test vehicle.

B. Electrolytic Determination of the Effective Surface Area of the Silver Electrode

G. L. Juvinall

1. Introduction

Brigham Young University is under contract to JPL to investigate the reaction geometry in oxidation and reduction of the alkaline silver electrode. Dr. Eliot Butler is the principal investigator on the present contract (No. 951554), and he also directed the work under the preceding contract (No. 951157). The results reported here are from work performed under the latter contract. Part of this effort is concerned with the electrolytic measurement of the effective area of the silver electrode.

Although the surface area of a sintered plate can be measured by various techniques, such as gas adsorption or dye adsorption, it would be far more valuable to be able to make the measurement by a method which depends upon electrolytic reaction, since the area that is effective in such a reaction is the important area in controlling current densities. Such information would be of great value in the investigation of basic battery processes as well in the area of battery design and the determination of battery failure mechanisms.

In the oxidation of silver electrodes at constant current, the potential of the silver electrode is observed to go through the changes shown in Fig. 7. The potential rises quite rapidly at first and then levels off, sometimes giving a small peak to the plateau, corresponding to the production of Ag_2O. At the end of this first plateau, the potential rises again, giving a definite peak before leveling off to the plateau corresponding to the production of AgO. At the end of this second plateau, the potential rises to that value necessary for evolution of oxygen. Of principle interest in this report is the length of the plateau which occurs between peaks A and B.

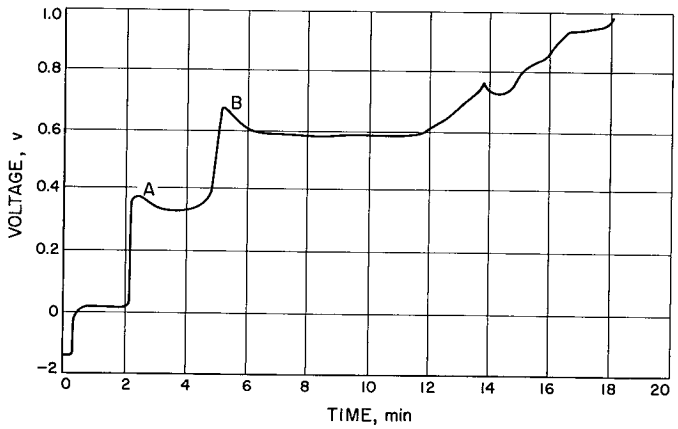


Fig. 7. Potential versus time curve obtained in the oxidation of silver electrodes at constant current

It has been suggested by several workers (Refs. 1, 2) that the first plateau of the oxidation curve corresponds to the covering of available surface area with a thin layer (perhaps monolayer) of Ag_2O. In a series of experiments in which plates of the same surface area were oxidized at various current densities and the lengths of the first plateau compared, it was found that if the current density is kept the same, the quantity $(t \times I)/a$ is constant. In this expression, t is the time in seconds which corresponds to the length of the plateau, I is the current in milliamps which flows, and a is the area of the silver plate. In the cases in which unpolished or polished plates were used, the surface area was taken as the calculated geometric surface area. This quantity $(t \times I)/a$ was found to take on a different constant value at each different current density, but when a plot was made of $(t \times I)/a$ against I/a, the smooth curve of Fig. 8 was obtained. When the log $(t \times I)/a$ was plotted against I/a, an almost straight line was obtained. With this information, a possible method for determining effective surface area is suggested. An estimation of the surface area which is effective in the electrolytic charge of a silver plate might be made upon the basis of the elapsed time between points A and B in the potential-time curve (Fig. 7).

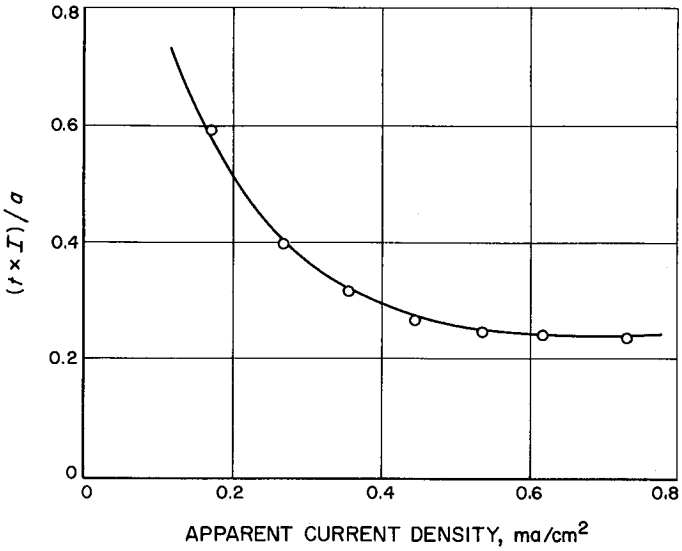


Fig. 8. Charge per unit area versus apparent current density

2. Experimental Evaluation

In all cases, the oxidation runs were made in the following way: The silver electrode to be oxidized was placed in the cell and supported at a constant depth in the cell solution (0.1N KOH). It was made the reducing electrode for a period of about 2 min before the polarity was reversed and the oxidation begun. Cycling experiments were also run in which the silver electrode was oxidized until the potential necessary for oxygen evolution was reached, and then the polarity was reversed and reduction was completed on the same electrode until three complete cycles were obtained. The potential was recorded continuously, and the current was measured periodically.

Most of the experimental work has been concentrated upon determination of reproducibility of measurement of the length of the first plateau (A to B, Fig. 7) at various current densities upon electrodes prepared in as identical a manner as possible. If thoroughly cleaned electrodes of silver foil are charged at constant apparent current densities ranging from 0.2 ma/cm^2 to 0.8 ma/cm^2, the reproducibility of measurement length of the first plateau is $\pm 10\%$. Table 3 shows the data obtained when such electrodes were oxidized at various current densities.

Table 3. Charge characteristics of silver foil

Approximate current, μa	Measured current, total spread, μa	Apparent current density, $\frac{ma}{cm^2}$	Length of first plateau, sec	Deviation from average, %
200	205.5—205.0	0.181	330	±3
	205.3—204.9		320	
	205.4—205.1		315	
300	309.5—308.9	0.273	135	
	309.5—308.7		140	
	309.4—307.8		150	
	309.4—307.8		155	±7
	309.2—307.7		140	
400	415.3—413.2	0.366	90	
	415.5—413.0		86	±7
	415.6—413.1		79	
	414.0—412.7		90	
500	510.0—508.7	0.451	64	
	511.0—508.8		60	±7
	511.1—508.5		56	
600	616.4—613.3	0.543	42	
	616.7—613.3		48	±7
	616.7—613.1		45	
700	700.0—697.5	0.619	39	
	701.0—697.3		41	
	700.8—696.9		37	±7
800	820.7—816.6	0.730	27	
	820.8—814.8		32	±10
	820.7—815.5		30	

Table 4. Electrode oxidation data

Type of electrode	Length of plateau, sec	Deviation, %	Approximate current, ma
Cleaned, unpolished electrodes	135	±7	0.30
	140		
	150		
	155		
	140		
Polished electrodes, cleaned with a detergent in the ultrasonic cleaner	108	±17	0.30
	105		
	93		
	107		
	106		
	124		
	126		
Polished electrodes, cleaned with CCl_4 and then with a detergent, in the ultrasonic cleaner	94	±27	0.30
	128		
	141		
	135		
	109		
Polished electrodes, cleaned with abrasive cleaner, anodized for 30 min in plating solution and then electroplated for 2 hr	227	±4	0.30
	223		
	219		
	210		
	220		
Polished electrode, cleaned with abrasive cleaner and electroplated for 2 hr	156	±4	0.30
	161		
	153		
	159		

Microscopic examination of the cleaned silver foil electrodes showed such gross irregularities in the surface characteristics that an attempt was made to polish these electrodes in order to obtain more uniform surfaces, and hopefully, to gain reproducibility in the establishment of a baseline for surface area determination. Table 4 shows the data obtained when various sorts of electrodes were oxidized and the lengths of the plateau for A to B in Fig. 7 were measured. Finally, an attempt was made to prepare a smooth, uniform surface by electroplating a layer of silver onto a silver electrode. Although some of these electrodes appeared to plate very smoothly and uniformly and the results from the oxidation runs were promising (see Table 4), consistently good plating was not achieved. More effort will be required in this area, and work is underway using silver electrode surfaces formed by vapor deposition. Additional details are available in the final report (Ref. 3).

3. *Results*

It appears that a realistic measurement of effective surface area by purely electrochemical means is possible. Further support for this conclusion is furnished in the case of the cycling experiments by the fact that the effective surface area increased on the second and third cycles. In the oxidation and reduction of a polished silver plate, the arrangement of the silver atoms is apparently sufficiently distorted and altered so that the effective surface area of the plate is increased and thus the plateau is lengthened. This observation correlates with the behavior of silver-zinc batteries which have been observed to increase their power output after several charge-discharge cycles.

However, more work is required to improve the reproducibility of the measurements and to extend them to other surfaces.

References

1. Cahan, B. D., Ockerman, J. B., Amlie, R. F., and Ruetschi, P., *Journal of the Electrochemical Society, 107*: 725, 1960.

2. Dirkse, T. P., *Journal of the Electrochemical Society, 106*: 920, 1959.

3. Butler, E. A., *Studies of the Reaction Geometry in Oxidation and Reduction of the Alkaline Silver Electrode,* Final Report, JPL Contract 951157, Brigham Young University, Provo, Utah, 1965.

III. Spacecraft Control

A. Wide-Angle Planet Tracker

D. G. Carpenter

1. Introduction

This is the final report on the development of a two-axis wide-angle planet tracker utilizing cadmium sulfide photoconductive detectors. The project has been divided into three phases: *Phase I*—Prototype Planet Tracker, design and evaluation; the results were reported in *SPS 37-31,* Vol. IV, pp. 82–88. *Phase II*—Planet Tracker Geometry Study. *Phase III*—Engineering Evaluation Model (EEM) Planet Tracker, design and evaluation. The results for Phases II and III are contained in this report.

This development has not been directed toward any specific flight projects but rather to develop and prove the concepts involved in this type of device. Specific requirements and constraints are required if the best possible design for a given task is to be realized.

The basic operation of the planet tracker depends on a null-sensing operation of two cadmium sulfide photocells operating in a bridge circuit configuration. These cells are shadowed by knife edge apertures which cause one cell to receive more illumination than the other as a function of angular displacement of the tracker. The change in illumination causes the cell conductance to vary, which produces an electrical offset in the bridge circuit. The magnitude of this offset is proportional to the angular displacement. There are two cells for each axis and a single acquisition cell for a total of five cells per tracker. The acquisition cell together with an electronic switch circuit provides an acquisition signal and an intensity signal which is proportional to the total illumination on the acquisition cell.

Testing has indicated the device in its present configuration will track and null on planets of from 5 deg subtended angular diameter to 60 deg. For planet sizes greater than 60 deg (100 deg was the design goal), the scale factor, or output in volts per degree off axis, falls to a value that would make integration into a closed-loop tracking system difficult. Design changes would allow the large change in scale factor to be minimized. These changes and recommendations will be discussed later.

Since this type of tracker operates on visible light, it nulls on the center of planet illumination. The pointing accuracy is related to changes in both phase angle and

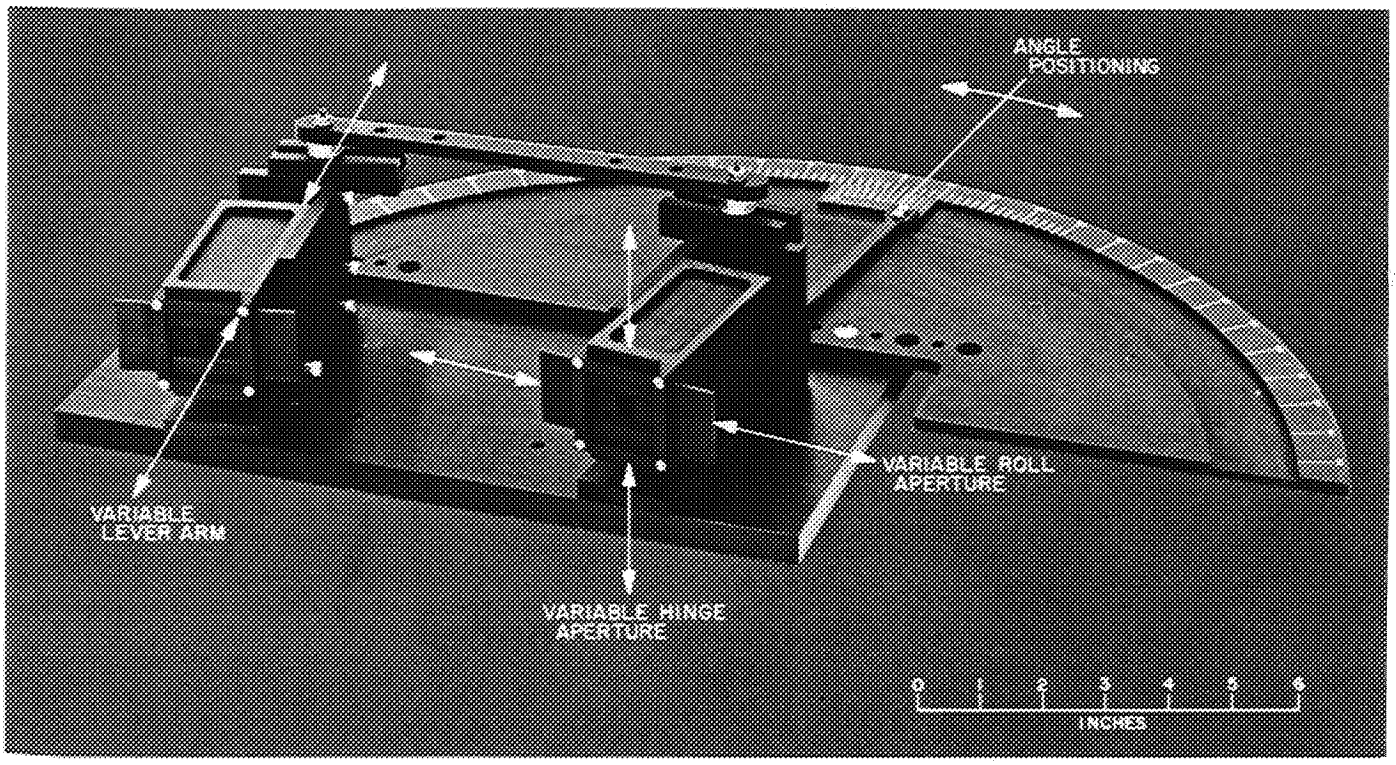


Fig. 1. Planet tracker variable geometry device

surface reflecting characteristics of the planet. In pointing to the planet's geometrical center, the unknowns of irregular surface illumination can account for null-offset errors as great as 30% of the planet's radius. A sensor such as this has only limited accuracy; however, the sensor is inherently reliable, is light in weight, has no moving parts, and has a very low power consumption. For these reasons, whenever the center-pointing accuracy is not of paramount importance, a device of this type should be considered.

2. *Planet Tracker Geometry Study*

Large changes in scale factor for varying planet sizes were obtained from an evaluation of the Phase I prototype planet tracker. It was evident that the lever arm needed to be shortened (distance from knife edge aperture to cell face) in order to improve scale factor characteristics. To obtain the optimum mechanical design parameters needed for the design of the EEM planet tracker, the device shown in Fig. 1 was designed and fabricated.

The device is called the planet tracker variable geometry device. It has the capability to vary the size of the aperture, the position of the shadow with respect to the cell, and the length of the lever arm from 0.15 to 3.0 in. The geometry errors associated with simulating large planets with a noncollimated source can also be cancelled by individual adjustment of the pointing angles.

Conclusions drawn from results of tests with the geometry device indicated that the optimum lever arm for the EEM was 0.175 in. The best shadowing was the half-masked cell at null. Using the above parameters, tests with the geometry device in a tracker configuration gave a change of 11-to-1 in the scale factor from a 5-deg planet to a 100-deg planet. This scale factor variation was still fairly large, but considering the large dynamic range of planet sizes covered, it was believed that the scale factor variation would be acceptable from a system standpoint.

3. *Design and Development of the Engineering Evaluation Model Planet Tracker*

a. Mechanical design. Fig. 2 (a) and (b) shows the mechanical configuration of the planet tracker. Instead of the conventional method of placing side by side the two cells that make up one axis of rotation, the cells are stacked one over the other. The reason for this configuration is that simulation of a large planet using a flat disk causes geometry errors when the cells are side by side.

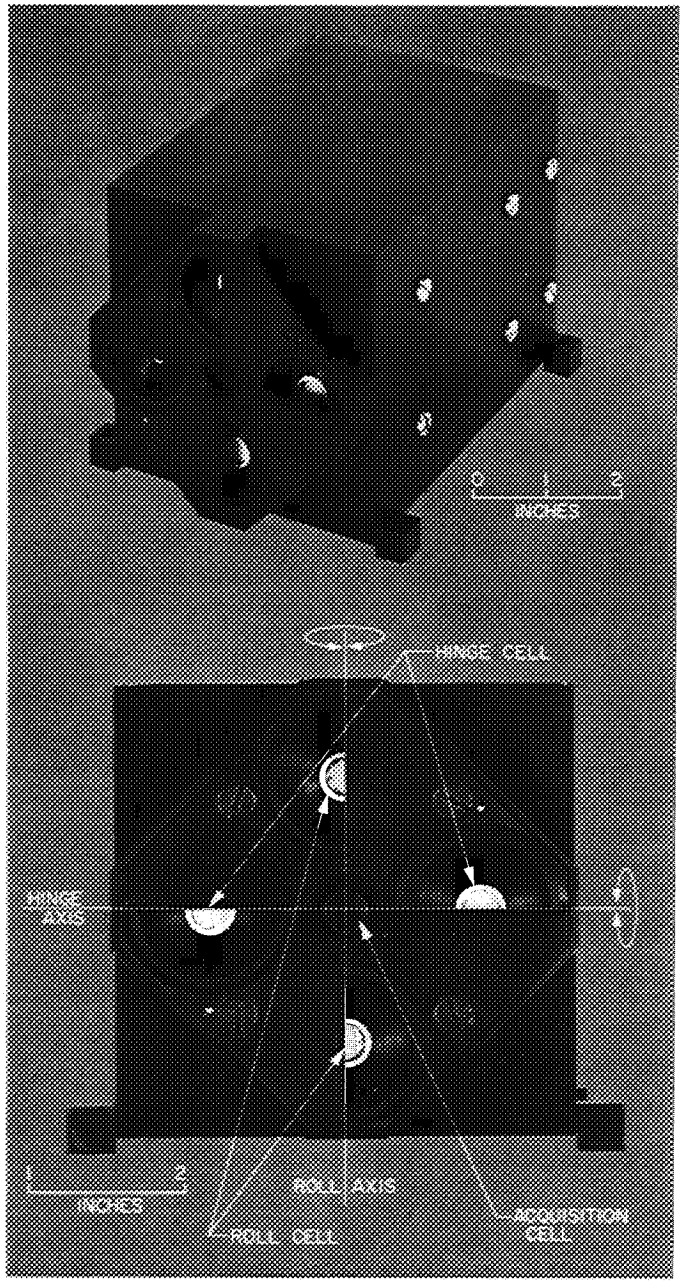


Fig. 2. Engineering evaluation model (EEM) planet tracker

(The two cells look at different parts of the planet when swung off axis.) By stacking the cells, these errors are effectively cancelled eliminating the need for a complex optically collimated planet simulator, yet not affecting flight characteristics.

The tracker uses a total of five cadmium sulfide photoconductors, the center cell being the acquisition cell, the other four comprising the two control-axis pairs. Directly in back of the cells is an area that would be used by a welded module electronic package comprising the signal processing circuitry.

Fig. 3 gives a description of the planet tracker fields of view. Angles α_1 and α_2, shown in Fig. 3(b), equal the linear fields of view of their respective cells to a point source of light. Angles β_1 and β_2 are the cells' total fields of view. However, for a planet which has a finite size, the pertinent fields of view become more complex. Referring to Fig. 3(c), angle γ equals the total tracker field of view. As the limb of the planet encroaches on angle β_1 in the direction shown on Fig. 3(c), Cell 1 becomes partially illuminated. This illumination level increases as more of the planet comes into the field of view, causing the cell conductance to vary directly. During this time, Cell 2 is still dark and remains so until the planet limb encroaches on angle β_2. At this time the conductance of

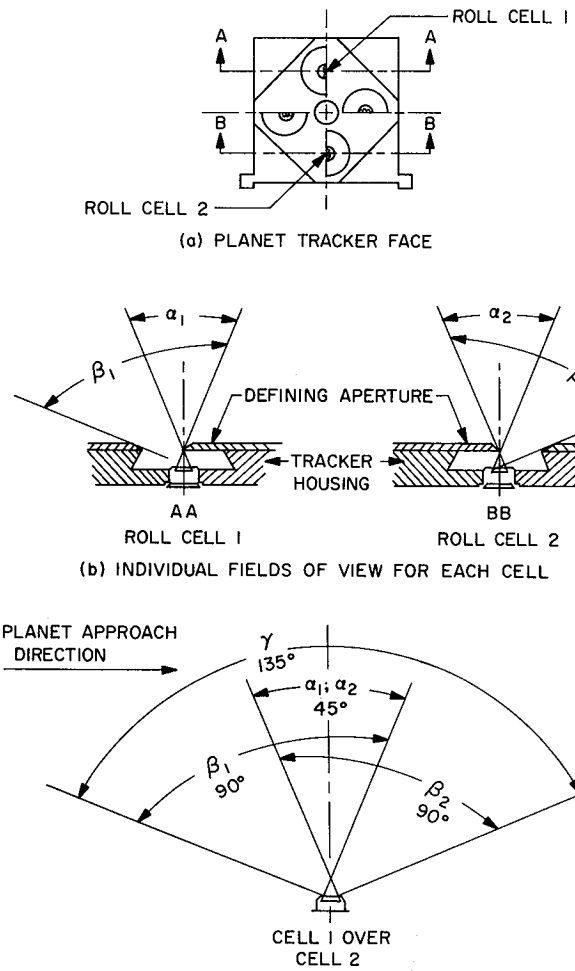


Fig. 3. Planet tracker fields of view

Cell 2 begins to increase as more of the planet comes into its field of view. Both cells now receive luminous flux from the planet in proportion to its position. When the planet's center falls on the tracker's optical axis, both cells are equally illuminated and an electrical null condition exists. The planet tracker's effective field of view is its total field of view, angle γ, plus the angular diameter of the planet.

Through the use of two cells, one increasing in resistance while the other decreases in resistance, a positive or negative voltage output is obtained depending on the relative position of the planet to the optical axis of the tracker. After processing the signal, a dc error signal is obtained which will be described later.

Fig. 4 shows the acquisition cell field of view. This cell plus its associated circuitry gives a fixed voltage output whenever the planet (or part of it, depending on its angular size) is within the ± 25 deg field of view.

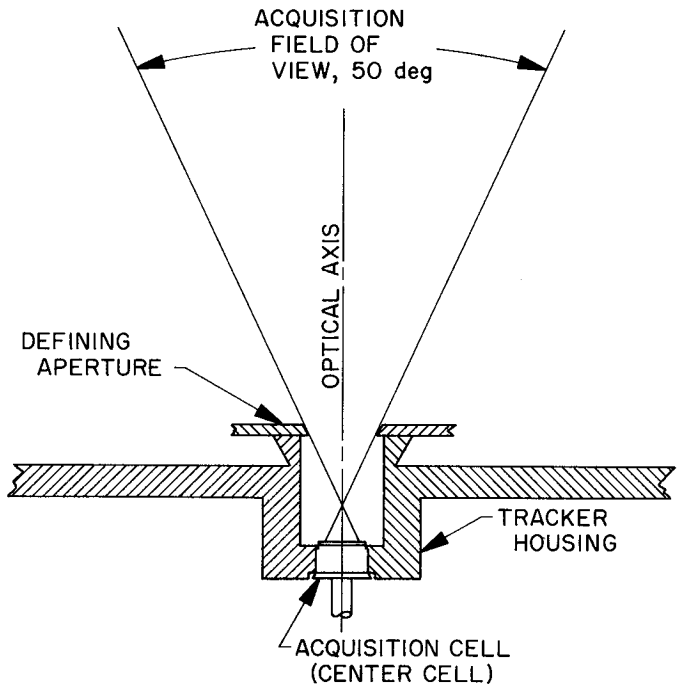


Fig. 4. Acquisition field of view

b. Electrical design. The photoconductors used in the planet tracker are cadmium sulfide, Type CL-705L, manufactured by Clairex Corporation. Conductance versus illumination curves were run on a group of 25 off-the-shelf cells. Pairs of cells with closely matched conductance curves were chosen for each control axis. No special characteristics were used for the choice of the acquisition cell.

The identical electronic breadboard, with no changes, has been used throughout the three phases of this project.

Fig. 5(a) gives the block diagram for the wide-angle planet tracker. Fig. 5(b) shows the complete schematic of the planet tracker.

An excitation voltage of 2.0 v ac is used across the photocell bridge in order to limit the power dissipation per cell below 125 mv at full solar illumination. AC voltage is used to make amplification simpler. The voltage produced due to an unbalanced bridge is amplified by the single-stage amplifier, which has a gain of approximately 20, and coupled to a demodulator and filter for a dc output signal. Typical error signals are shown in Fig. 6.

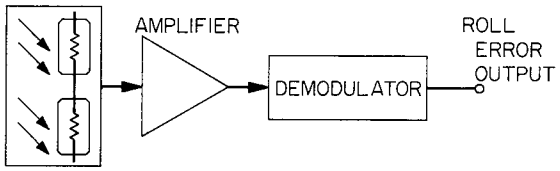
The resistance of the acquisition cell varies with the intensity of the planet causing Q_1 emitter voltage to vary. Therefore, the intensity output taken from the 24K–3K resistance voltage divider varies directly with planet intensity from approximately 1.0 v to 2.6 v, corresponding to a range of intensities from 1.0 to 720 ft-cd. Fig. 10 shows a typical intensity curve. The acquisition trigger circuit of Q_3 and Q_4 is presently set to trigger at an intensity level of approximately 0.06 ft-cd and due to hysteresis will lose acquisition at approximately 0.045 ft-cd. The acquisition output level is approximately 8.2 v when in the acquired state.

Since this project has been basically that of proving concepts, no attempt at electronic packaging has been made.

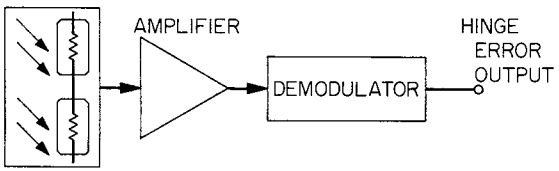
4. Test and Evaluation of EEM

a. Planet tracker field of view. The various fields of view pertaining to the tracker are shown in Figs. 3 and 4. Angles α_1 and α_2 are the linear fields of view for the respective detectors, and are both equal to 45 deg total included angle. Angles β_1 and β_2 are the total fields of view of each detector and are both equal to 90 deg. Angle γ represents the total tracker field of view and is equal to 135 deg. The acquisition field of view, Fig. 4, is 50 deg.

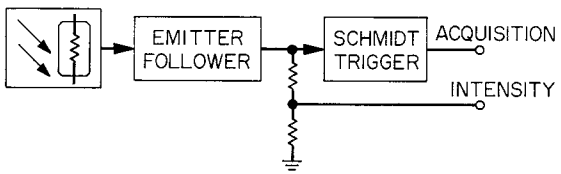
b. Characteristic error curves. The roll axis characteristic curves of the tracker are given in Fig. 6. The hinge



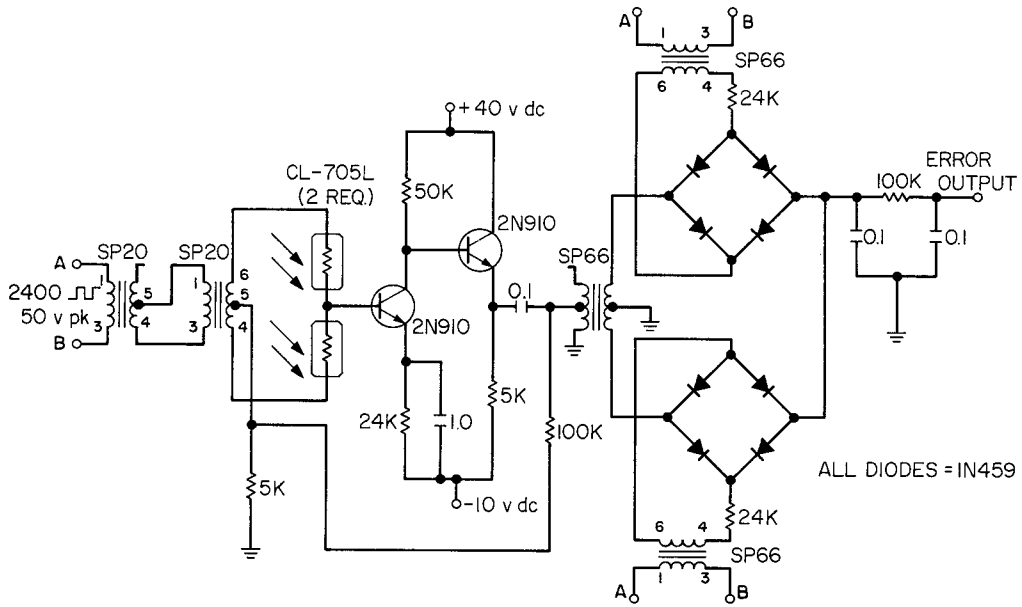
ROLL ERROR CIRCUIT, BLOCK DIAGRAM



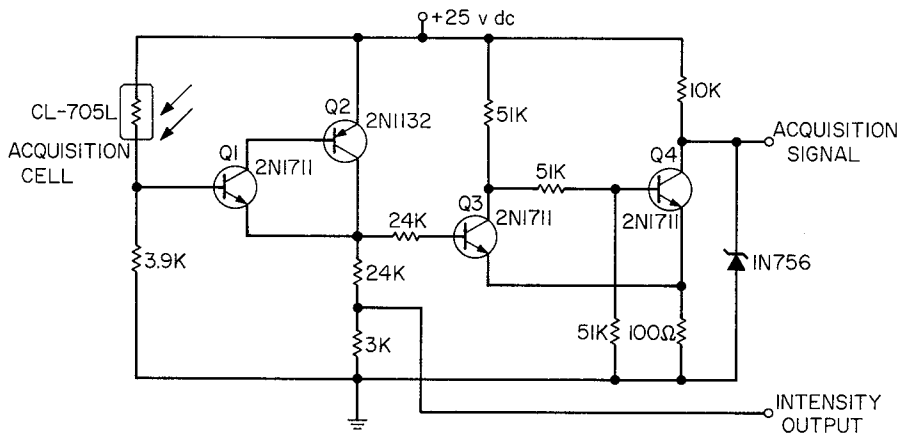
HINGE ERROR CIRCUIT, BLOCK DIAGRAM



ACQUISITION-INTENSITY CIRCUIT, BLOCK DIAGRAM



ROLL ERROR CIRCUIT; HINGE ERROR CIRCUIT;
SCHEMATIC FOR TWO IDENTICAL CIRCUITS



ACQUISITION-INTENSITY CIRCUIT, SCHEMATIC

Fig. 5. Schematic and block diagrams of wide-angle planet tracker

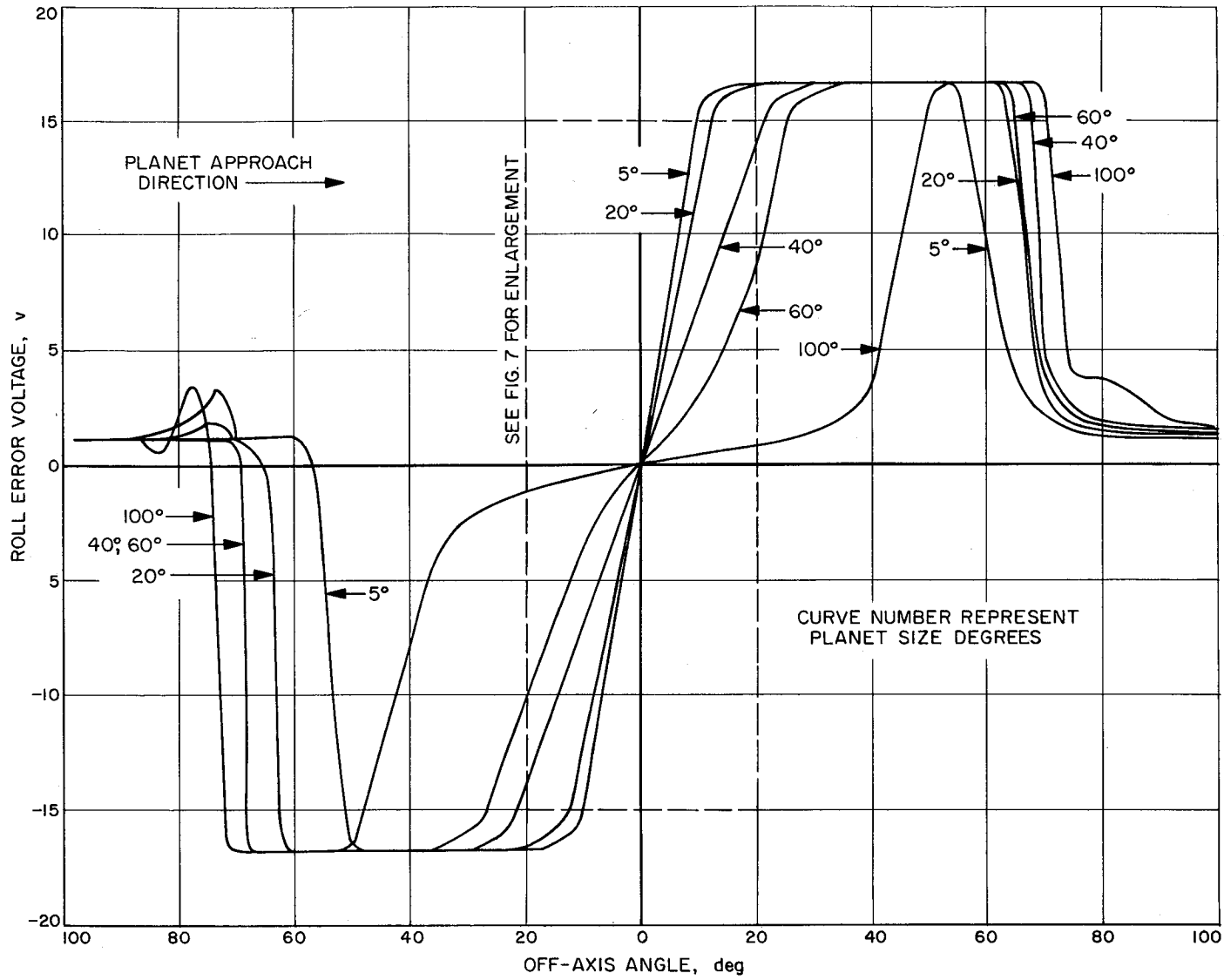


Fig. 6. Planet tracker roll error characteristic curves

axis would be identical to these curves since the two circuits are the same but isolated so that no cross-coupling between the two axes exists. The roll curves of Fig. 6 are obtained using a simulated planet with an illumination uniformity of 25% and an average luminance of 1400 foot-lamberts. The 1400 foot-lambert planet is a factor of 1.75 greater luminance than the 800 foot-lambert peak luminance of Mars. However this uniform increase in luminance has very little effect on the error curves.

Up to a planet size of 40 deg, the curves are very linear through null. After 40 deg, the curves become nonlinear as is exhibited by the great change in the 100-deg curve. The reason for this change in linearity is as follows: Planets that subtend an angle of 45 deg or less lie totally within the linear field of view of the detectors (α_1 and α_2; Fig. 3) while on axis. In this region, as the planet moves off axis, there is a linear movement of the shadow of the planet across the cell face causing a linear change in detector conductance. Hence, the linear output characteristics. For planets greater than 45 deg, there is no position in the operating range that prohibits any portion of the cell from receiving at least some luminous flux from the planet. Therefore a nonuniform condition exists caused by a variable amount of luminous flux received from the change in planet position. The larger the planet, the greater nonuniform portion of luminous flux is received and the curve assumes the shape of the 60 deg curve of Fig. 6.

To extend the linear range of operation beyond the 40-deg planet size the lever arm needs to be shortened, which will widen the linear field of view. This will have the effect of reducing the scale factor for all curves, but they can be increased again, electronically. To obtain a linear range of 100 deg total included angle, the knife edge aperture would have to be approximately 0.030 in. from the cell face. From a system standpoint, linearity all the way to 100-deg planets might not be necessary. This would then allow the knife edge position to be a compromise between scale factor linearity, accuracy considerations, and practical limitations.

c. Scale factor variations. Fig. 7 is an enlarged portion of Fig. 6, showing the scale factor variation in the region of ±20 deg. As shown, the scale factor changes a factor of 4.7 over a range of planets, from 5 deg to 60 deg angular diameter. This is very tolerable from a system standpoint. The scale factor for the 100-deg planet, however, varies from the scale factor of the 5-deg planet by a factor of 14.5. This would probably be considered unworkable. However, as mentioned before, a workable scale factor out to a 100-deg planet should be possible by shortening the lever arm. With a constant planet size, the scale factor will vary approximately 10% if the planet luminance is increased by a factor of 1.5. This is of minor importance since the planet luminance is relatively constant. The intensity at the tracker does, of course, increase with a decrease of tracker-to-planet distance, but the size of the planet is also increasing, and it is the size of the planet that is the predominant factor in scale factor change.

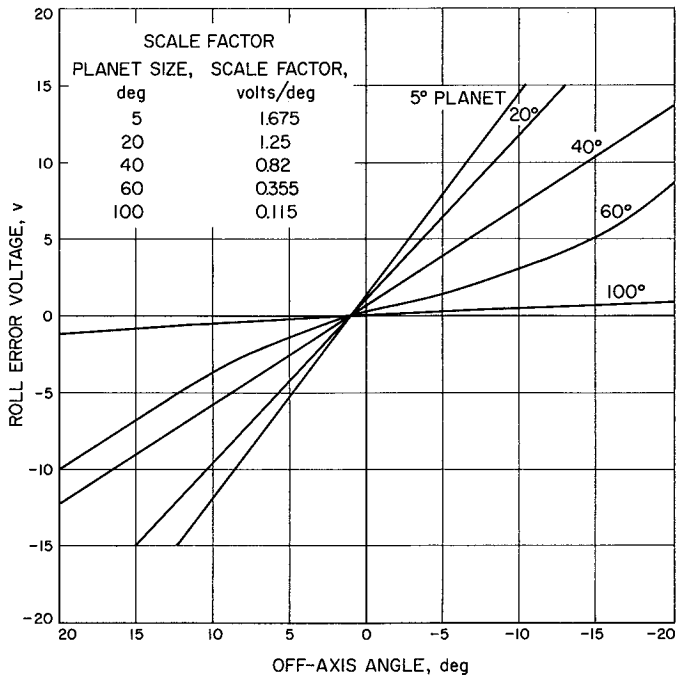


Fig. 7. Scale factor versus planet size

d. Null stability. The over-all null stability of the tracker is a function of: (1) variation of cell conductance with temperature, (2) variation of cell conductance with light intensity, and (3) electronic stability. Tests are now being conducted to investigate null offset versus temperature. These tests are being run using a separate device with the same mechanical parameters as the EEM, and utilizing the same type of detectors. Results so far are inconclusive as to numbers of degrees drift versus a given temperature. However, relative drifts (shifts between the two detectors) of 30% have been seen. Tests are continuing in order to establish reliable figures for temperature stability and null offsets produced by temperature variations. Work is also under way to develop better quality cadmium sulfide photoconductors with superior temperature characteristics.

A change in light intensity has no significant effect in shifting the null, as long as the two cells have matched conductance versus illumination curves.

Electronic stability has not been tested directly with these breadboards since it was believed that it would pose no problems to the over-all stability of the planet tracker. The circuitry is straightforward, uses standard components, and has very low noise characteristics.

e. Pointing accuracies. The over-all pointing accuracy of the planet tracker is a function of: (1) Device accuracy (null offset and electronic noise), and (2) irregularities in planet surface illumination. Item 1 was discussed in the previous section. Item 2 will be discussed here.

The planet tracker will null on the center of illumination of the planet to a fairly high degree of accuracy. However, the center of illumination is quite variable with relation to the geometric center of the planet.

For instance, assume the planet to be a perfectly lambert reflector. At 0-deg phase angle, the center of illumination will coincide with the geometric center of the planet. At 90-deg phase angle (half-illumination), the center of illumination will now be a distance of 0.542 radii from the center of the planet. Add to this the fact that, due to the many light and dark areas on the planet's surface, the center of illumination will shift still more in a rather unknown amount and direction.

Therefore the pointing accuracy of the planet tracker to any specific point on the surface is unknown, but its accuracy in pointing to the center of illumination is fairly accurate.

As an example of the magnitude of null offset that can arise with a non-uniformly illuminated planet, see Fig. 8 and Table 1. Fig. 8 represents one of the simulated non-uniform planets used to measure the amount of null offset from that of a uniformly illuminated planet. The Fig. 8 planet represents a factor of 2.5 difference in illumination levels from one edge to the other. The null offsets obtained are listed in Table 1, both in degrees and percentages of planet radii. It can be seen that the percentage change in planet radius remains quite constant through a 60-deg planet, as would be anticipated. The increased value for the 100-deg planet can be explained on the basis of the difficulty in simulating such a large planet with a 12 in. diameter flat disc.

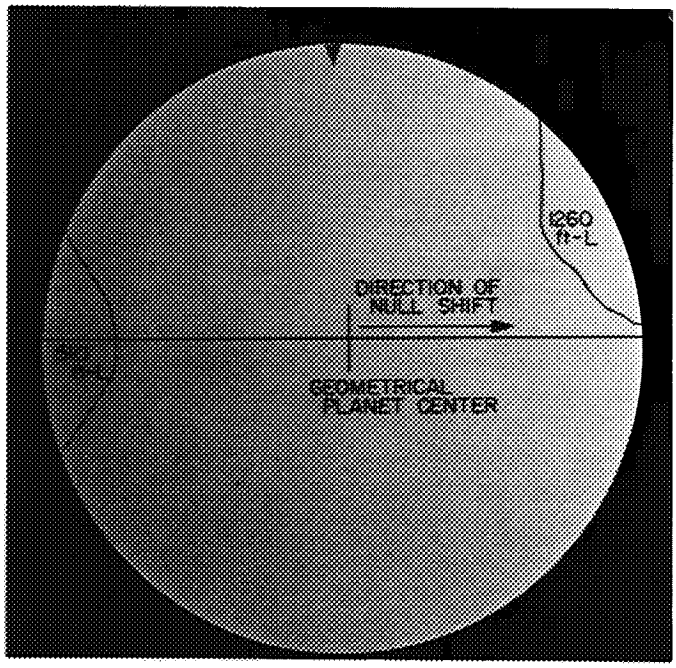


Fig. 8. Simulated planet used for null offset study

f. Acquisition signal. The planet tracker should provide a fixed output voltage of ≥ 7.0 volts whenever the planet, or some light source of great enough intensity, is within the acquisition cell field of view. In testing the tracker, it was found that only the 5-deg planet was within the field of view at acquisition. The edges of all other size planets were approximately 45 to 50 deg off

Table 1. Null shift due to non-uniform planet illumination

Planet size, deg	Shift of roll axis from null position[a]	
	Angle, deg	Planet radius, %
5	0.4	7.4
20	1.4	7.0
40	2.9	7.2
60	5.0	8.3
100[b]	17.0	17.0

[a] See Fig. 8 for simulated planet

[b] Offset values questionable for this size planet due to planet simulation difficulties.

axis when the acquisition signal was triggered. This indicates that a stray light problem exists. The cell is being triggered by reflected light off the walls of the well in which it is mounted (see Fig. 4). To eliminate this problem, special care needs to be taken in designing the acquisition cell mounting to exclude stray light reflections.

Fig. 9 shows the present characteristics of the planet tracker acquisition output signal. The acquire points are measured from the optical axis of the tracker (0 deg) to the geometrical center of the planet.

g. Intensity output signal. Fig. 10 is a plot of intensity output voltage versus planet size (using a 670 ft-lambert planet) for the present tracker. The shape of the curve is not significant since it would be modified if the stray light problem with respect to the acquisition signal were solved. The saturation point of the intensity circuit should be increased to accommodate a detectable voltage variation for the larger, higher intensity planets.

h. Stray light rejection. No stray light rejection capability tests were run since stray light rejection was not a design criteria for the EEM. Stray light rejection must be designed into a planet tracker with specific design parameters chosen from the trajectory characteristics of a mission.

5. Planet Simulator

The planet simulator used for the testing of the planet tracker consists of a 12-in. diameter opal glass disc. This disc is back-lighted by a series of six, independently

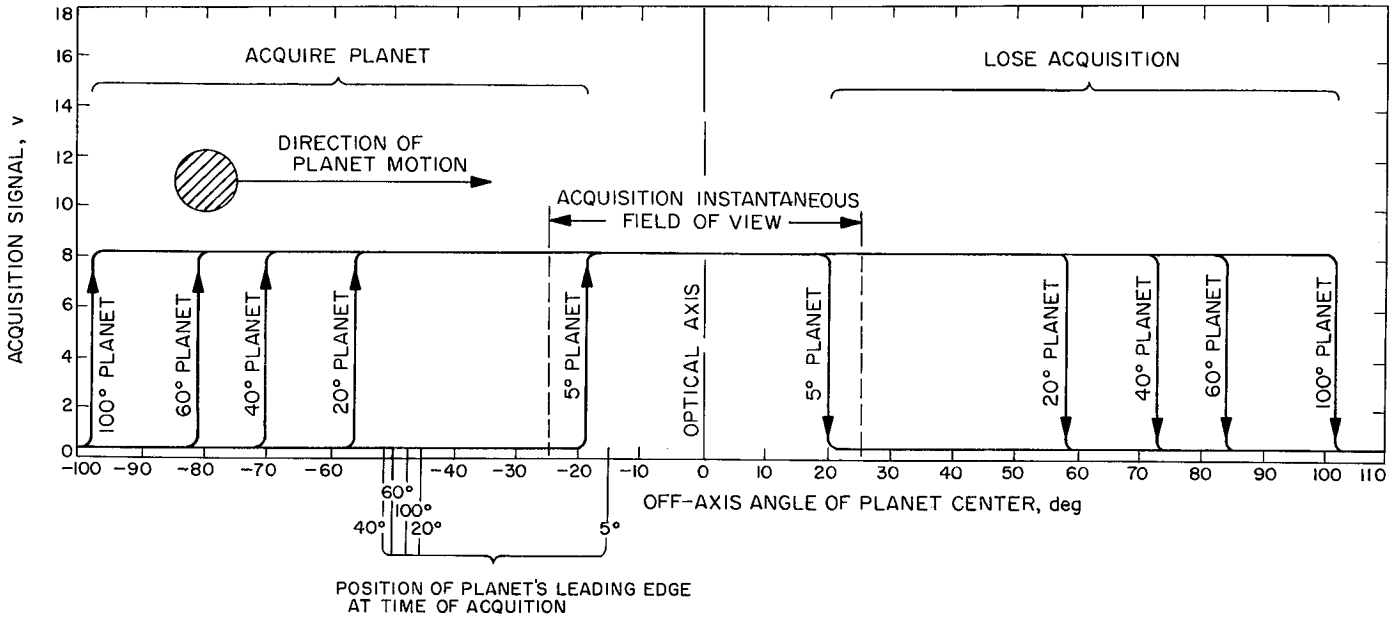


Fig. 9. Planet tracker acquisition characteristics

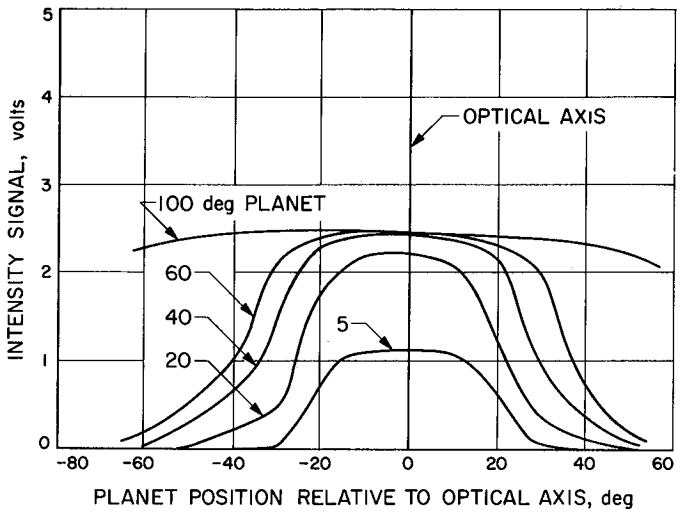


Fig. 10. Intensity output versus planet size and position

switched, 500-watt quartz iodine lamps located in back of, and outside the parameter of the disc. Most of the light energy is reflected off a white plate and then through the opal glass diffusor. The color temperature of this source is approximately 2800°K. To simulate the different sizes of planets (5 deg to 100 deg angular diameter) used during the testing of the planet tracker, the tracker was located at the proper distance from the simulator to give the desired angular diameter. The simulated planet was then moved through the field of view of the tracker, and the error signal and acquisition signal were automatically recorded on an *X-YY'* recorder. Fig. 11 shows the planet tracker test equipment set up in the celestial simulation laboratory (Celestarium) of the Jet Propulsion Laboratory. This laboratory is 40-ft diameter hemisphere, the walls of which are covered with black honeycomb. This is very effective in eliminating the stray light associated with a simulator of the type used in testing the planet tracker.

6. Summary of Planet Tracker Evaluation

Table 2 gives the operating, mechanical, and electrical characteristics of the planet tracker in its present configuration. With the present mechanical parameters, the planet tracker will track and null, on the "center of illumination," a planet subtending a total angular diameter of from 5 deg to 60 deg. By shortening the lever arm and modifying the electronics to increase amplifier gain, a planet as large as 100-deg angular diameter can be tracked with acceptable operating characteristics. The planet tracker is simple in design and operation, uses no moving parts, and uses commercial, highly reliable, electronic components. The sensor is small in size, light in weight, and has a low power consumption. Possible applications include pointing of a science platform on approach to or orbiting a planet, pointing of an antenna toward the Earth, or other similar tasks, where the limited accuracies of a center of illumination type planet tracker are acceptable.

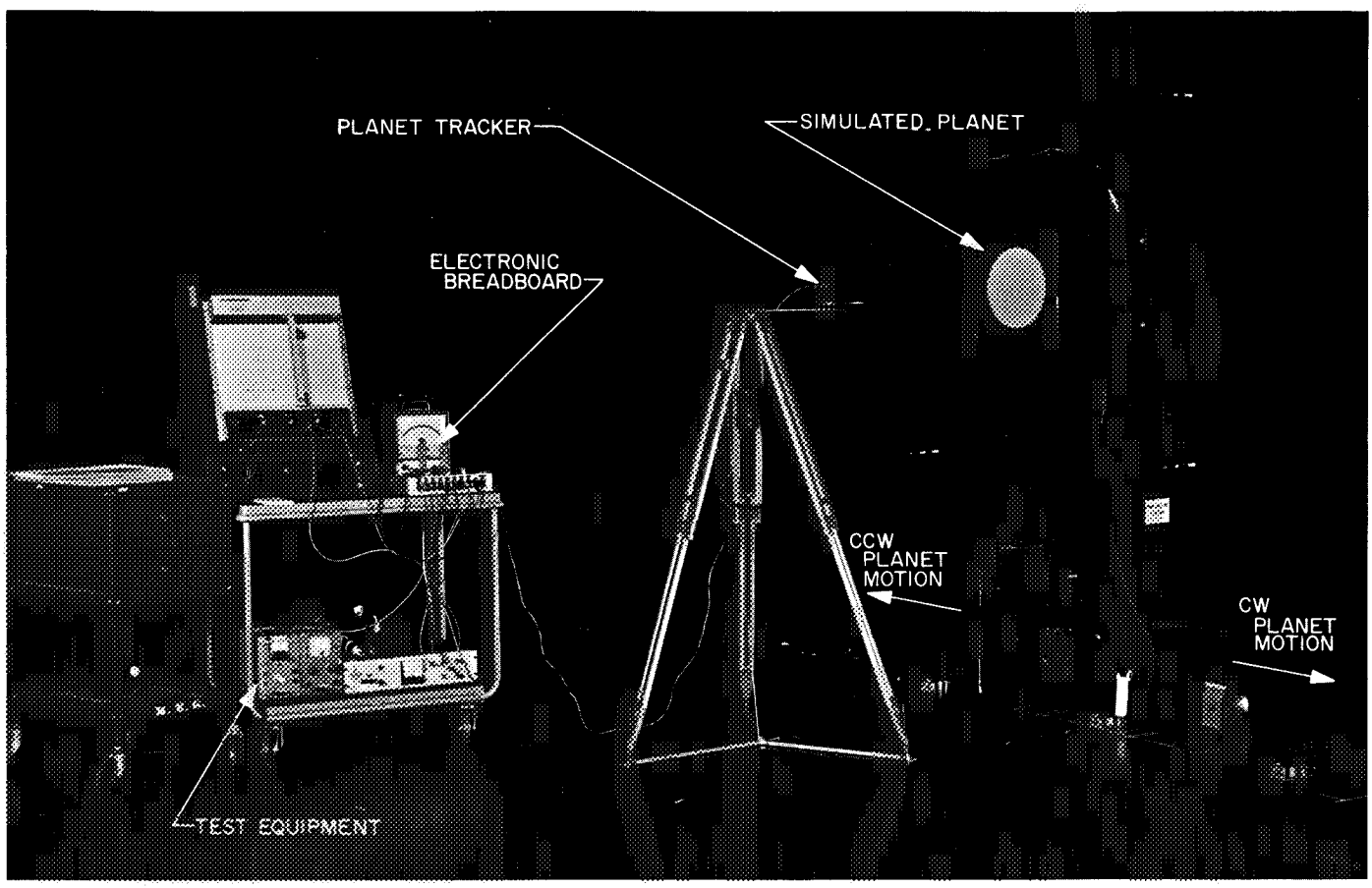


Fig. 11. Planet tracker test facilities

Table 2. Wide-angle planet tracker characteristics

Operational characteristics	
Basic operation	center of illumination tracker
Detectors	cadmium sulfide photoconductors (Clairex-705-L)
Operating range	5 deg to 100 deg (angular diameter)
Field of view	
Total tracker (γ)	135 deg
Linear (α)	45 deg
Acquisition	50 deg
Sensitivity	operate from 1.0 ft-cd to 720 ft-cd
Sun protection	able to withstand full solar illumination with no Sun shutter
Unit characteristics	
Total power dissipation	$\leq$0.5 w
Size	32 in.^{3} (4.0 x 2.8 x 2.8 inches)
Weight	$<$0.85 lb

Unit characteristics (cont'd)	
Lever arm length	0.175 in.
Shadowing	knife edge half-shadows detectors at null
Output characteristics	
Error signal	two axis (roll and hinge)
Scale factor	
5-deg planet	1.675 v/deg
20-deg planet	1.25 v/deg
40-deg planet	0.82 v/deg
60-deg planet	0.355 v/deg
100-deg planet	0.115 v/deg
Acquisition signal	not acquired at $\leq$0.5 v is acquired at $\geq$8.0 v
Intensity signal	1.0 v to 2.6 v (for inputs of 1.2 to 720 ft-cd)
Time constant	$\leq$0.5 sec at 1.0 ft-cd illumination

B. Advanced Scan Platform

R. Mankovitz

1. Introduction

This report presents an analysis of the horizon scan platform control system using an induction or synchronous motor to drive the platform. This analysis is a continuation of the study reported in *SPS 37-35*, Vol. IV, pp. 33–35. The study showed the functional and analytical diagrams for both stepper motor and induction motor systems. Preliminary investigations of stepper motor systems indicate that, due to the low damping of the structure and support dynamics, and the low platform jitter requirements (on the order of 100 μ rad/sec), this type of mechanization may prove unfeasible.

The analysis and preliminary results for a system, using a proportional drive (induction or synchronous motor), follow.

2. Analysis

a. Horizon scanner output. Fig. 12 shows the horizon scanner photocells, with the planet limbs falling on the opposing edges of symmetrical photocells. The scanner output is zero. Assume now, that the scanner output is fed into an N-pulse detector, which has the following properties: For an input pulse equal to or less than $(N-1)$ pulses wide (a unit pulse is taken as a standard width equal to 1), the N-pulse detector output is zero. For an input pulse N pulses wide or greater, the output of the detector is ± 1, where the sign is that of the incoming pulse. It now remains to determine the position errors (degrees) corresponding to the output from an N-pulse detector (± 1). Referring to Fig. 12, for the planet moving in a direction "A," the output of the N-pulse detector is

$$(N/2-1)^\circ \text{ of position error if } N \text{ is even} \quad (1)$$

$$(N/2-0.5)^\circ \text{ of position error if } N \text{ is odd} \quad (2)$$

If the planet moves in a direction "B" (Fig. 12), the output of the N-pulse detector is

$$(N/2)^\circ \text{ of position error if } N \text{ is even} \quad (3)$$

$$(N/2+0.5)^\circ \text{ of position error if } N \text{ is odd} \quad (4)$$

Eqs. (1) and (2) represent the minimum position errors which can cause a pulse detector output. Eqs. (3) and (4) indicate the maximum position errors to "just cause" a pulse detector output. For the control system analysis to follow, a pulse detector is connected to the scanner output to insure that a finite system deadband exists. Eqs. (1) and (2) indicate that a minimum of a 3-pulse detector is required to insure a finite width deadband. Using this detector, yields a minimum deadband of ± 1 deg, and a maximum deadband of ± 2 deg.

The system mechanization also requires a signal indicating that the position error is at least 2 deg greater than the deadband edge. Referring to Eqs. (2) and (4), a 9-pulse detector will yield an output for a ± 4 deg minimum error, to a ± 5 deg maximum. (For more details on the scanner operation, refer to the *SPS 37-35* report mentioned above.)

b. Control system description. Fig. 13 is the basic block diagram of the control system. Operation is as follows: Assume that when the system is energized, a large (>5 deg) position error exists between the scanner

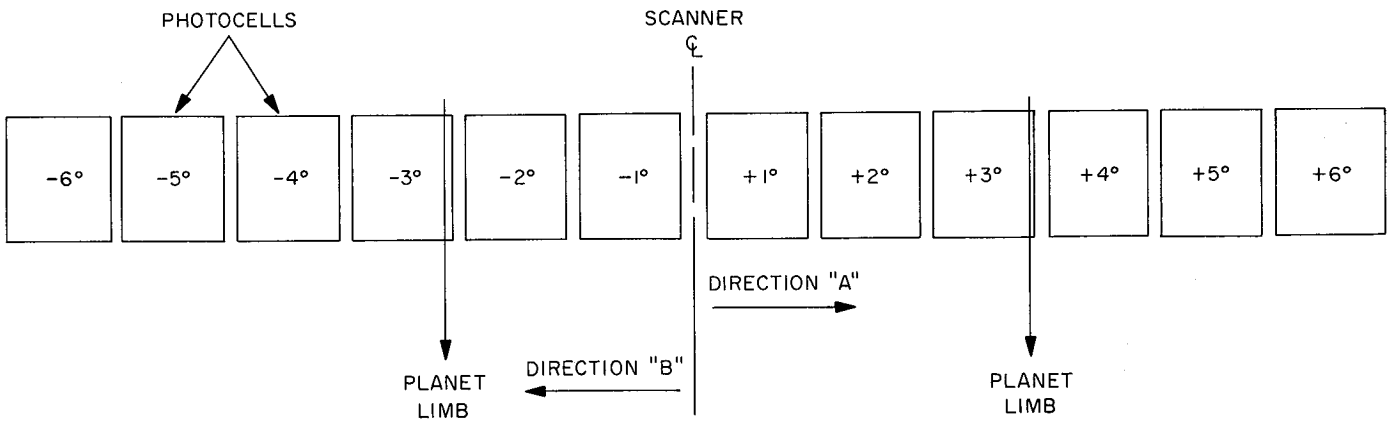


Fig. 12. Horizon scanner photocell output analysis

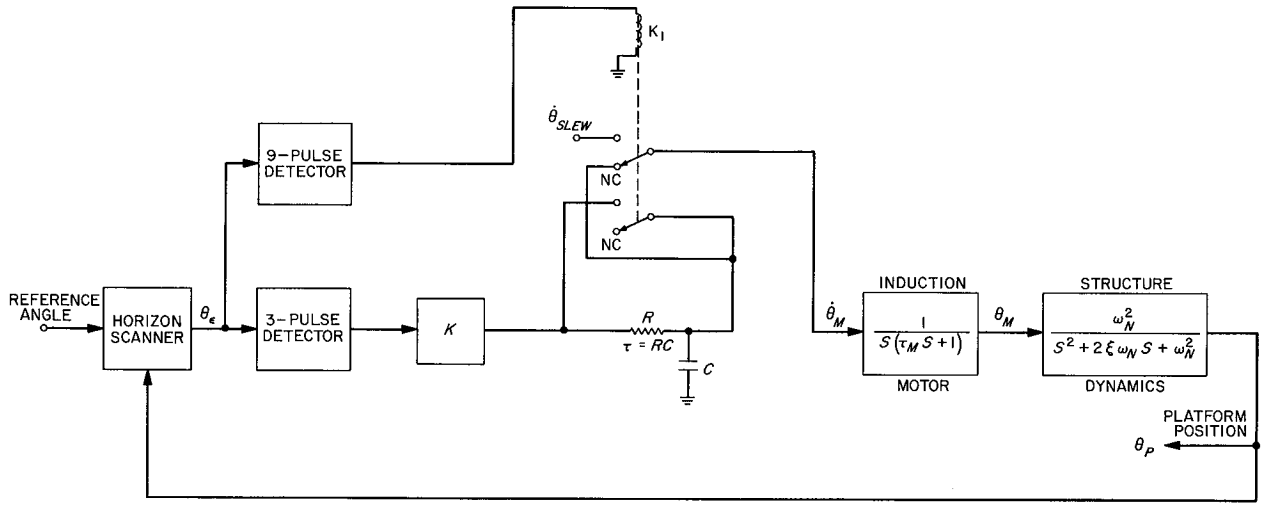


Fig. 13. Scanner control system block diagram

centerline and the planet local-vertical. For this error, both pulse detectors have a +1 output. With K_1 energized, the motor is connected to a voltage corresponding to $\dot{\theta}_{slew}$, a fairly high rate. At the same time the capacitor C is charged to a voltage K. When the angular error is reduced (due to $\dot{\theta}_{slew}$) sufficiently to de-energize the 9-pulse detector, the motor is connected to the RC network, so that $\dot{\theta}_m$ (motor rate) is now equal to K. The angular error is reduced at this rate, until it is within the deadband (between 1 deg and 2 deg), at which time the 3-pulse detector output is zero, and the R-C network begins to discharge, decreasing platform rate. It should be noted that the pulse detectors are mechanized as latched types (using flip-flops) which can only change state at clock pulse intervals, corresponding to the sensor sweep. The detectors thus provide the function of a sample—hold circuit.

c. System response. For a step input, the system will go into a limit cycle from edge to edge of the deadband, with exponentially decreasing rate, until the capacitor voltage is less than the motor threshold voltage, at which time the system will come to rest inside the deadband. The above condition represents a hypothetical case, however, since the scan platform is always subjected to a non-zero rate input.

For a ramp (or any non-zero rate) input, after the initial overshoot, the system will go into a one-sided limit cycle against the leading edge of the deadband. This limit cycle will, theoretically, decrease in amplitude (exponentially) until an infinitesimal oscillation results. Due to the platform dynamics and the sampling nature of the system, however, a minimum limit cycle amplitude will result.

The three system parameters varied to optimize transient and steady-state performance are τ, the first-order-lag time constant; $\dot{\theta}_{slew}$, the angular rate during the slew mode; K, the maximum angular rate between the deadband edge and the point where the slew mode commences (θ_{slew}).

The system performance goals are the following:

(1) Maximum time to acquire (inside deadband) from maximum angular error of 60 deg = 60 sec

(2) Accept input rates up to 0.2 deg/sec (3.5 mrad/sec)

(3) Platform jitter in the order of 100 μrad/sec.

(4) Maximum overshoot of platform from deadband should be less than θ_{slew} (4 deg).

Item (4) is considered a requirement for proper system performance, since if the overshoot is sufficient to initiate the slew mode, rate reduction will not take place, and the system will limit-cycle between the bounds $\pm\theta_{slew}$.

3. Computer Simulation

a. Description. Using the IBM 1620 computer, and a digital-analog simulator program (DIANA), the control system was simulated and the parameters optimized. The following were the system constants:

(1) Horizon scanner scan period = 100 Ms

(2) Motor time constant (τ_m) = 50 Ms

(3) Structure dynamics natural freuency (ω_n) = 15.2 rad

(4) Structure dynamics damping ratio (ζ) = 0.07

In general, it is desirable to maximize the time constant τ, and slew rate $\dot{\theta}_{slew}$, while optimizing the system performance goals.

The value of K should be held to a minimum, to decrease platform jitter; however, it must be greater than the maximum input rate (3.5 mr/s) to permit acquisition. In addition, K must be sufficiently large so that the capacitor charging rate is greater than the anticipated input acceleration.

b. Results. Since there exist two extreme values (a maximum and minimum for the deadband (θ_{DB}) and slew point (θ_{slew}), these can be combined to form four distinct cases. It is obvious, however, that the two combinations of minimum ($\pm$17.5 mr) and maximum deadband ($\pm$35 mr), with the minimum θ_{slew} ($\pm$70 mr), yield the "worst-case" system performance (maximum overshoots and setting times), and hence these cases were used for system optimization, yielding:

$$\dot{\theta}_{slew} = 21 \text{ mr/sec}$$

$$\tau = 30 \text{ sec}$$

$$K = 5.25 \text{ mr/s}$$

In an effort to determine the stability of the system using classical methods, the following two assumptions were made: First, the sampling rate of the scanner is sufficiently high (compared to the system time constants) to neglect the effects of the sampling process and treat the system as continuous. The basis for this assumption is that the sampling frequency is 10 cps, while the dominant system pole is at 0.005 cps. Second, the system nonlinearity (pulse detector) can be represented by the describing function for a deadband, which is shown in Fig. 14. The describing function represents a pure gain (K_{DB}) which is a function of the ratio of the input signal amplitude to the deadband level. The function has a maximum value of 0.63, which represents the worst case (maximum gain in the stability analysis).

Considering the deadband as a gain of 1, the root locus of the system was investigated. Using a digital computer program for computing and plotting root loci, the plots

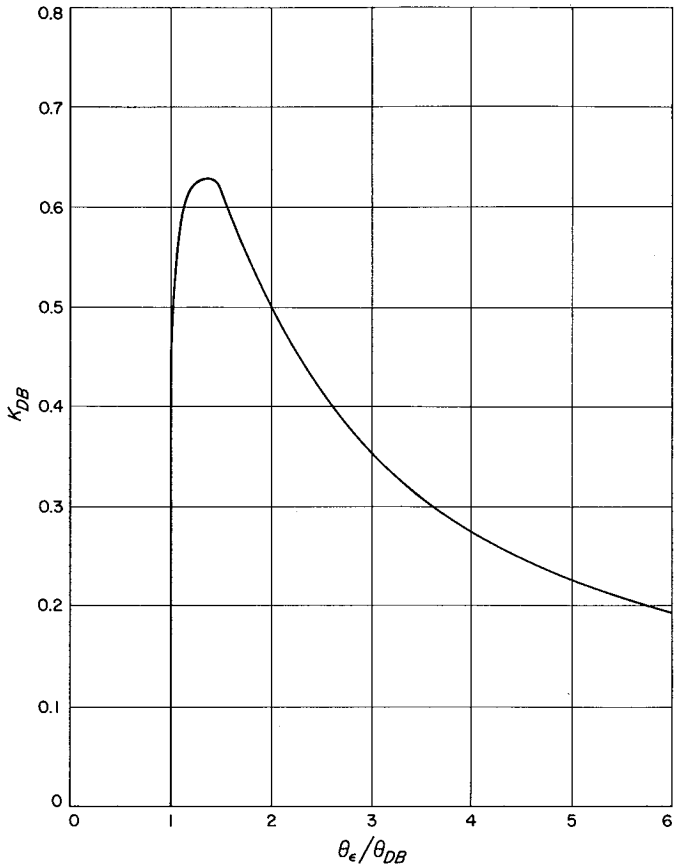


Fig. 14. Control system deadband describing function

shown in Fig. 15 were constructed, and represent the two loci for the system:

$$\frac{K \cdot K_{DB}\,(1/\tau)\,\omega_n^2}{S(S + 1/\tau)\,(S^2 + 2\zeta\,\omega_n S + \omega_n^2)} \qquad (5)$$

From Eq. (5), the system gain is calculated as 0.041, which establishes the operating points on the loci. From Fig. 15, it can be seen that the system is stable for the region of gains being considered, i.e., zero to 0.041.

Using the optimized parameters listed above, the system step responses are shown in Fig. 16, for $\pm$1 deg and $\pm$2 deg deadbands. The plot represents the system position error versus time, with the initial position equal to $+\theta_{slew}$ (70 mr) and initial platform rate equal to $\dot{\theta}_{slew}$ (21 mr/s). (These initial conditions are used for all subsequent runs.) For the two cases, the maximum overshoots are nearly equal at approximately 36 mr.

Since a step input represents the maximum overshoot, these plots indicate that performance requirement (4), discussed in Sect. 2c, is satisfied, with a 34 mr safety

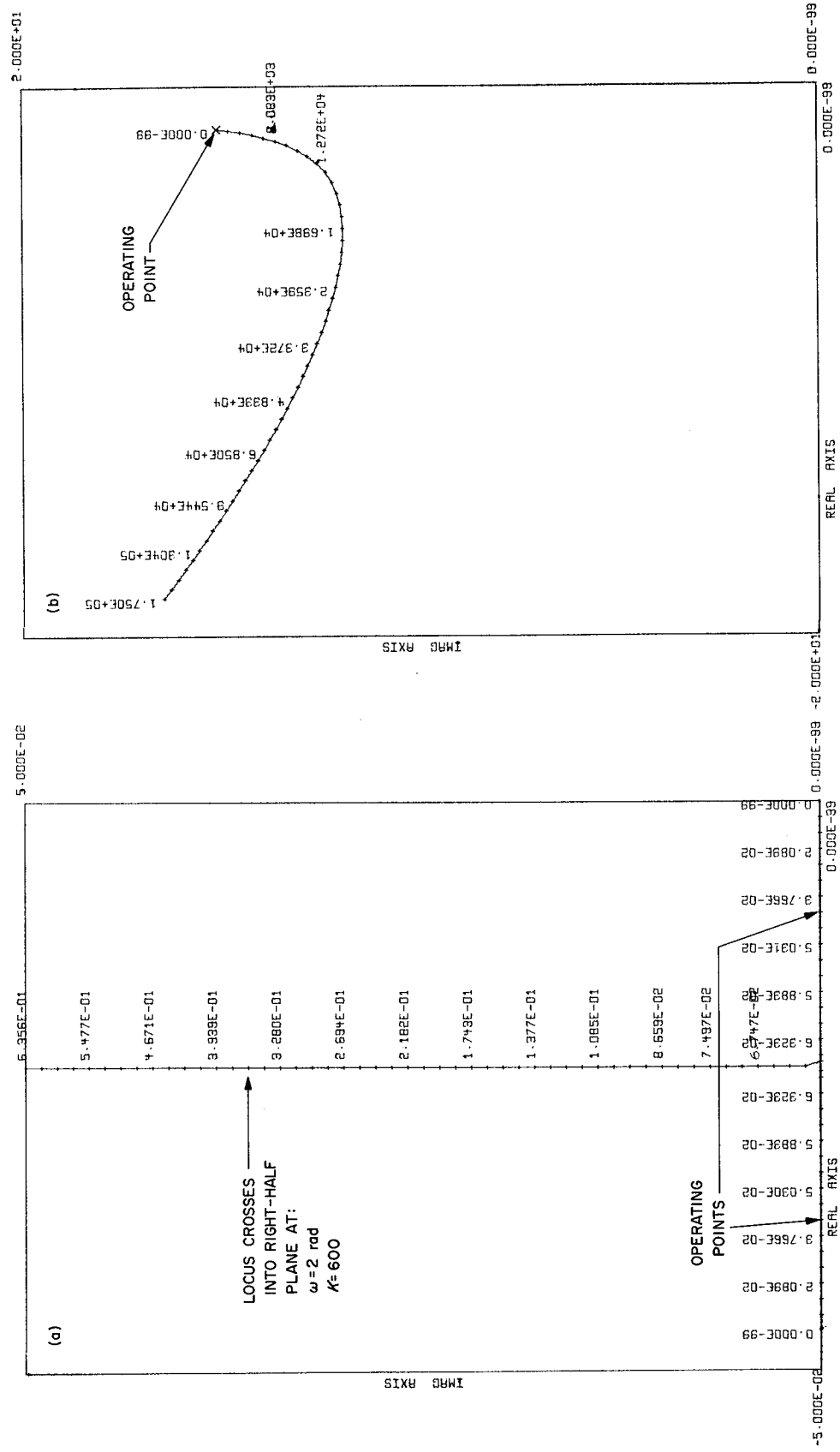


Fig. 15. Root loci and operating points

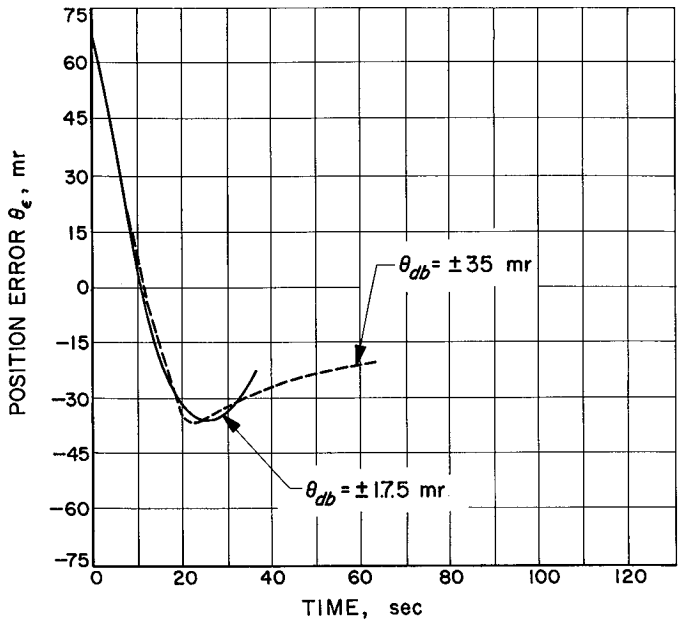


Fig. 16. Control system step response

margin. After the initial overshoot, the platform coasts into the deadband with an exponentially decreasing rate.

Fig. 17 shows the system response to the maximum input rate (0.2 deg/sec). This case was run with a ±2 deg deadband. However, since the overshoot into the deadband was less than 2 deg, the transient behavior for a ±1 deg deadband is identical, with the limit cycle occurring about the +17.5 mr point. The steady-state limit-cycle amplitudes had an RMS value of 119 μrad. The RMS rate error (defined as the difference between the platform and input rates) was 105 μrad/sec. This rate error could be considered as the platform jitter; however it should be pointed out that this limit cycling represents a smooth, slowly varying (approx 0.1 cps) platform motion, and hence only a small component of this motion may propogate to the platform instruments as jitter.

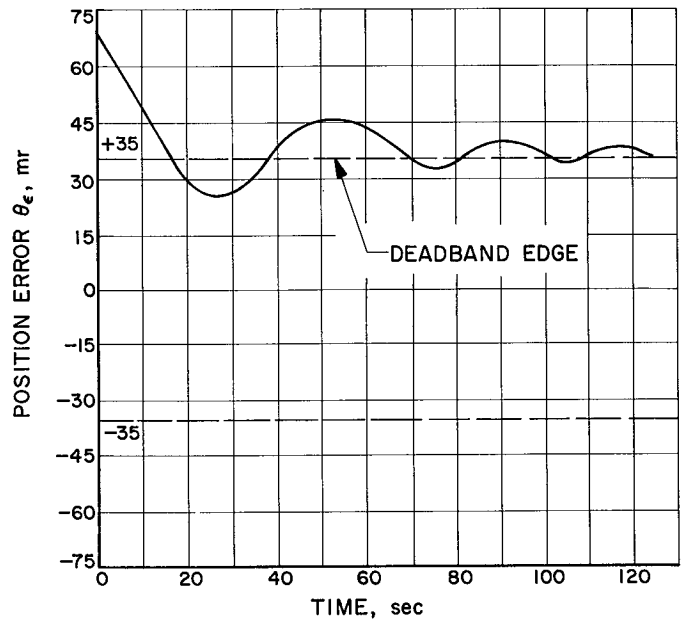


Fig. 17. Response to maximum input rate, 0.2 deg/sec

To investigate the effects of different values of input rate, as well as the effects of switching the deadband from its maximum to minimum limits (and vice versa) during scanning, the cases shown in Figs. 18 and 19 were run.

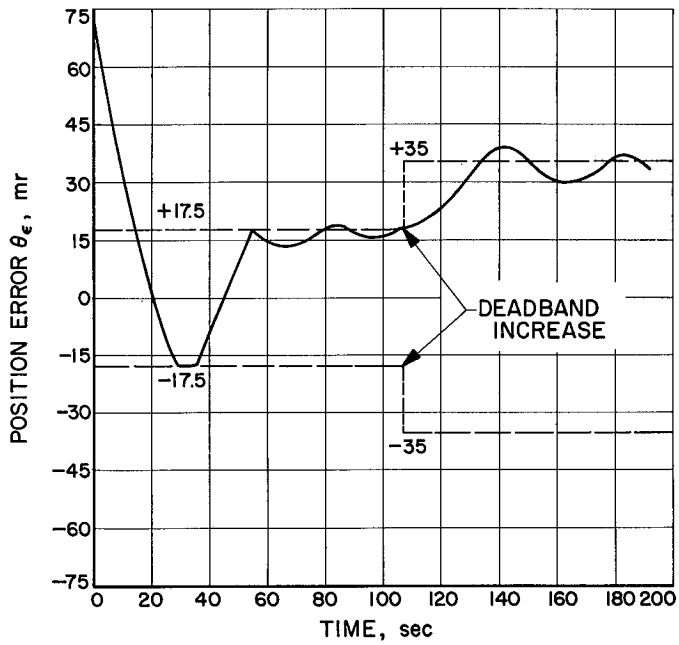


Fig. 18. Response to 0.1 deg/sec input rate and a deadband increase

The initial conditions for the case shown in Fig. 18 were: an input rate of 1.75 mr/s, and a deadband of ±1 deg. After the initial overshoot, which carries the position error to the deadband edge (where it "hangs" for a few seconds), the platform begins to establish a limit cycle at the +1 deg deadband edge. At 108 sec, the system deadband is instantaneously *increased* to ±2 deg. From Fig. 18, it is evident that the system responds correctly to this transition by moving the platform position to the new deadband edge, and establishing a limit cycle about this point.

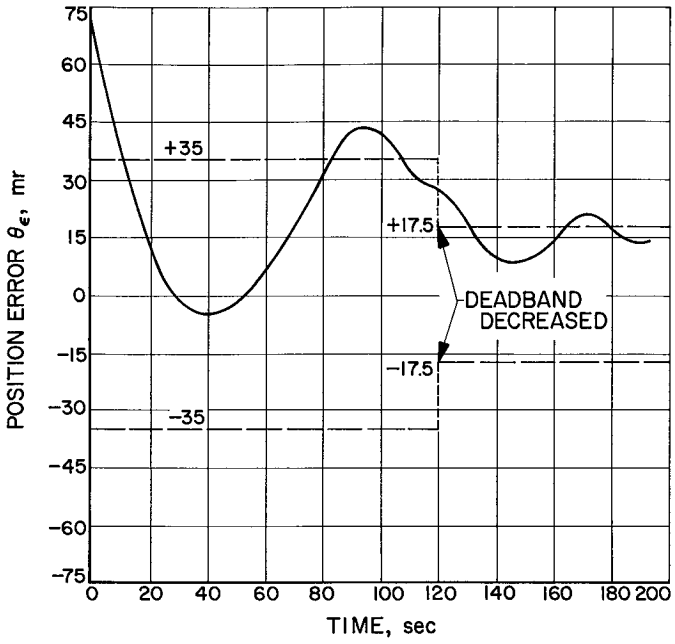


Figure 19. Response to 0.112 deg/sec input rate and a deadband decrease

For the case shown in Fig. 19, the initial conditions were: an input rate of 2 mr/s, and a deadband of ±2 deg. In a manner similar to the previous case, as the system is establishing a limit cycle about +2 deg, the deadband is instantaneously *decreased* (at $t = 117$ sec) to ±1 deg. From the position error plot, it can be seen that the system responds satisfactorily to this type of transition by decreasing the position error to the smaller deadband edge, and establishing a limit cycle about this point.

c. Conclusions. Comparing the simulation results discussed above, with the performance goals listed in Sect. 2c, it appears that this type of system can meet the scan platform requirements satisfactorily. To verify the simulation results, and determine the effects of hardware limitations, it is necessary to mechanize the system as a breadboard, and subject it to the input conditions mentioned above. The following paragraphs describe the system mechanization and future breadboard testing.

4. Control System Mechanization

Fig. 20 is a mechanization block diagram, showing induction motor control circuits. (Fig. 21 shows the motor control circuit for a synchronous motor drive.) The pulse detector and motor control mechanization details follow.

a. Pulse detector circuit. Flip-flop (1) is used to determine the error polarity, which is determined by which sensor head output pulse occurs first. The "exclusive-or" prevents the F/F from resetting on the second pulse. The first output pulse also gates on the pulse generator (which has a pulse period equal to a single scanner pulse width). If the scanner pulse width is equal to or greater than 3, the shift register will have an output, permitting the AND gates to trigger the F/F's (2) and (3) when the horizon scanner "scan reset" (clock pulse) occurs.

As an example of the circuit operation, assume that a pulse from head "*A*" has occurred first, and at some later time greater than *N*, the pulse from head "*B*" occurs. The "*A*" pulse will set F/F_1 ($Q = 1$, $\bar{Q} = 0$) since the "exclusive or" output will be true, and pulse "*A*" will also start the pulse generator. The pulse generator will continue until the "exclusive or" is false, which happens when the "*B*" pulse occurs. Since this is $\geq N$ seconds after pulse "*A*," the shift register output is true, enabling the AND gates (AND_1 is true AND_2 is false). When the scanner clock pulse occurs, the shift register is reset, and the F/F's (2) and (3) are allowed to change state, (F/F_2 is set, F/F_3 is reset) yielding an output of (+1). If the time between "*A*" and "*B*" pulses had been less than *N*, the shift register output would have been false, inhibiting both AND gates. When the clock pulse occurs, F/F's (2) and (3) are reset, and the circuit output is zero. The circuit for a 9-pulse detector is similar, with the only change being the number of shift-register stages.

The 3-pulse detector output is fed into a gain, a low-pass filter, and into the slew-mode relay switch (shown in detail in Fig. 13).

b. Induction motor control circuit. The motor control circuit shown in Fig. 20 operates as follows: The reference motor winding is powered by a 400-cps source. The motor control signal is fed through a chopper, synchronized to the reference supply frequency +90 deg. This mechanization allows a (+,0,−) DC signal to be converted to a phased 400-cps voltage to drive the motor control winding.

c. Synchronous motor control circuit. Referring to Fig. 21, the operation of this motor control circuit is as follows: Two motors are connected to a differential, with the reference motor driven from the 400-cps supply. This arrangement establishes the supply frequency as the zero reference. The DC input signal is used to vary the VCO frequency up or down around its center frequency. The VCO output is digitally divided by 10 to increase the overall accuracy and resolution of the system.

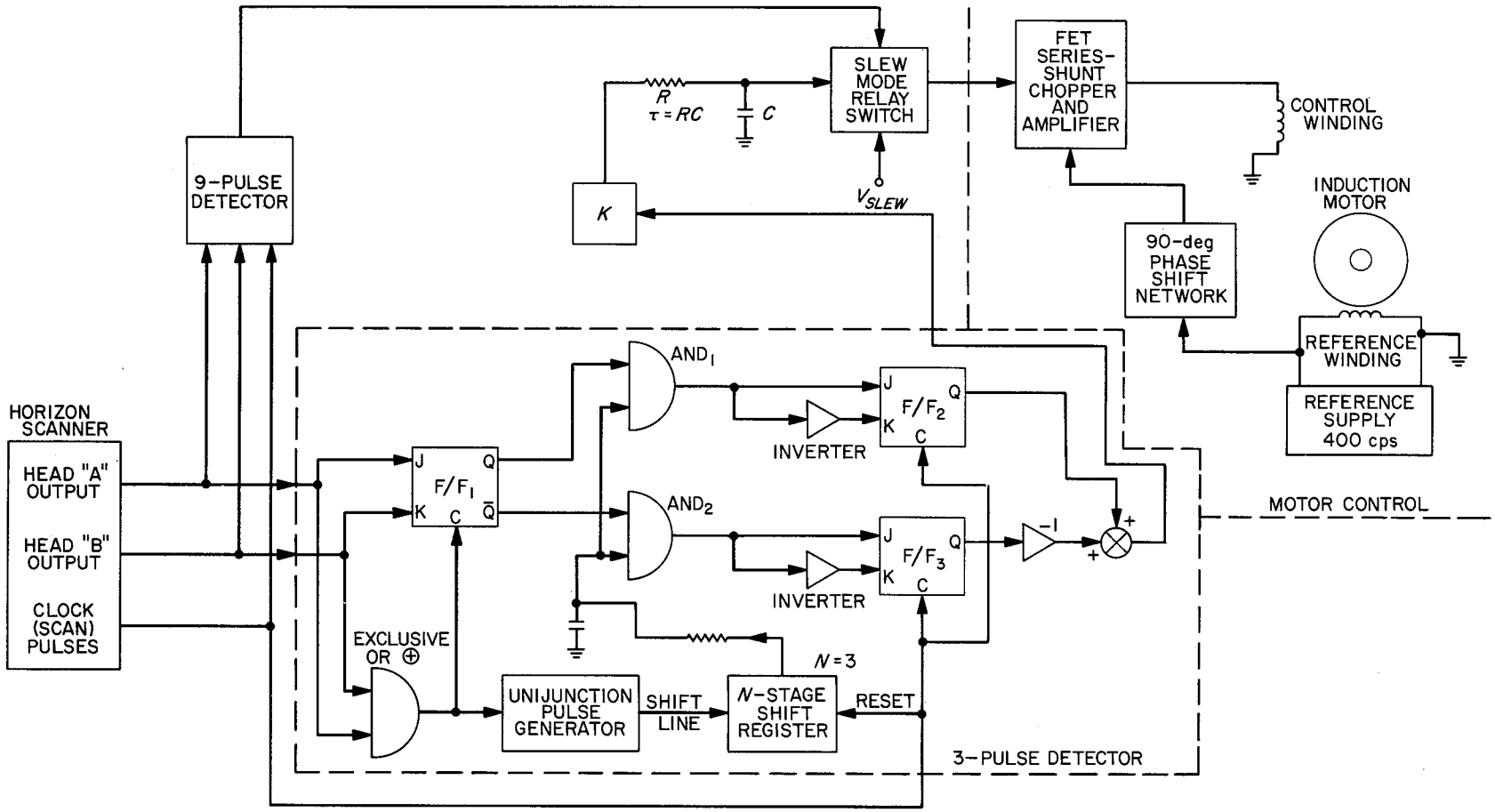


Fig. 20. Scanner control system mechanization block diagram

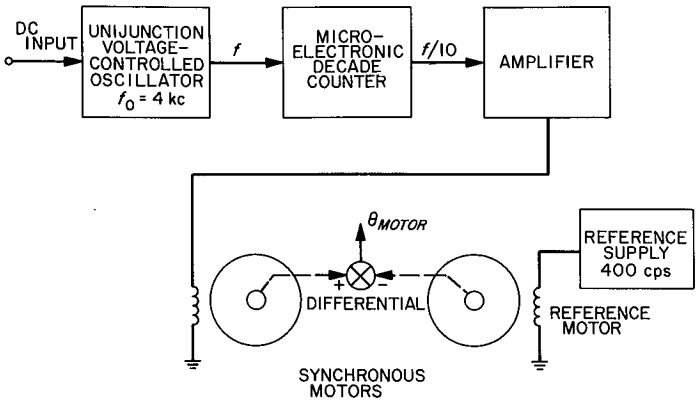


Fig. 21. Synchronous motor control circuit

5. Future Breadboard Testing

Components have been purchased to construct breadboard test units for the induction, synchronous and stepper motor systems. The initial breadboards will simulate the platform and actuator dynamics with torsion rods and inertia wheels. The horizon scanner will be simulated with precision potentiometer position sensors and sampling circuits, until the actual scanner becomes available.

Eventually, a complete dynamic simulation will be performed with a scanning platform, horizon scanner and simulated planet characteristics.

C. Attitude Control Thrust-Nozzle Measuring Techniques

J. C. Randall

As a result of the higher than anticipated limit cycle rates of the *Mariner IV* spacecraft, this task was started to develop accurate measuring techniques of attitude control thrusters in the pulse mode. As reported in the *SPS 37-35*, Vol. IV, it was originally the intent to make a dynamic measurement of thrust versus time during the standard 20-msec valve "on" time. A force rebalance system was selected utilizing a gas-bearing table, a set of rotor position sensors, a servoamplifier, and a set of torquers mounted to the gas-bearing table. The thruster to be tested was then placed on the rotor of the gas-bearing table and energized. The positional sensors detected rotor movement and fed a signal to the torquers via the servoamplifier. A measure of the thrust versus time profile was obtained by monitoring the servoamplifier output.

In attempting to utilize this setup, however, several problems were encountered. Random vibration within the building, for example, caused disturbances in the system on the same order of magnitude as the thrust level would produce. In addition, operation of the solenoid valve caused transients in the gas table when the plunger reached the end of travel.

As a backup to the gas-bearing testing, a cantilever beam and strain gage approach was started. Strain gages were attached to a 0.25-in. stainless steel tube which also served as the gas feed to the thruster. Although the beam natural frequency was about 300 cps, a severe 50 cps ringing was noticed when attempting to measure thrust using this technique. Subsequent analysis revealed that a 50-cps torsion mode in the beam was being excited. Beam configurations were looked into which had high torsional frequencies in relation to the desired beam frequency. For the configurations investigated, however, the beam frequency only became lower than the torsion frequency at very low frequencies.

With the problems that cropped up in trying to measure thrust versus time profiles, the task was re-evaluated as to what type of information was desired. For spacecraft attitude control, the two important parameters are steady-state thrust level and impulse bit per valve actuation. The thrust/time profile is an interesting problem but is not necessary for attitude control purposes; furthermore, steady-state thrust and impulse bit should be relatively easy to measure.

A beam was designed with a natural frequency of 4.35 cps and a length of 21.5 in. (limited by the height of the vacuum belljar). The cantilever beam displacement is measured with a linear position transducer. The setup is shown in Fig. 22. The low frequency was selected so that the 20 msec valve "on" time appears as a pure impulse to the beam. Monitoring the subsequent excursions of the beam allows the impulse to be determined.

Fig. 23 is a scope trace of the transducer output for a 0.0262 in. diameter nozzle. The valve was held open with scope settings of 5 volts per vertical division and 100 msec per horizontal division. It can be seen that the beam first overshoots, then damps down to the steady-state offset. The steady-state thrust level is:

$$(2.1 \text{ div})\ (5 \text{ volts/div})\ (1.23 \times 10^{-3} \text{ lb/volt})^* = 0.0129 \text{ lb}$$

* Obtained from calibration of beam and transducer under known load.

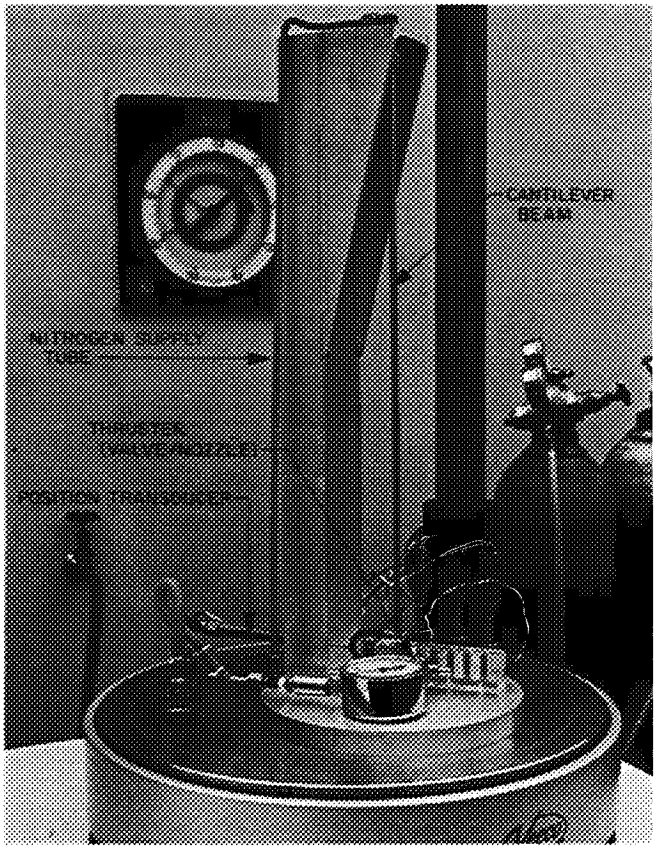


Fig. 22. Cantilever beam thrust level and impulse measuring setup

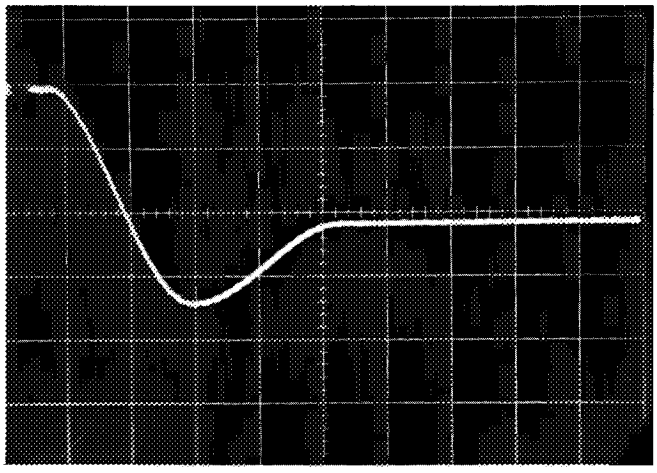


Fig. 23. Steady-state thrust

This compares with a theoretical predicted steady-state thrust level of 0.0125 lb.

Fig. 24 is a scope trace of the transducer output for a 0.0467 in. diameter nozzle. The valve was energized for 20 msec and then turned off. The scope settings are 5 volts per vertical division and 50 msec per horizontal division. The impulse can be calculated by measuring the amplitude of the first excursion. Since the beam frequency is 4.35 cps and the spring constant is 0.76 lb/in., the effective mass is:

$$M_{eff} = \frac{K}{W^2} = \frac{0.76}{[2\pi(4.35)]^2} = 1.02 \times 10^{-3} \text{ lb-sec}^2/\text{in.}$$

The amplitude of the first excursion is:

$$\frac{(3.4 \text{ div})\,(5 \text{ volts/div})\,(1.23 \times 10^{-3} \text{ lb/volt})^*}{0.76 \text{ lb/in.}} = 0.0275 \text{ in.}$$

If a sine wave shape is assumed, the maximum velocity is:

$$v = 2\pi\,(4.35)\,(0.0275) = 0.75 \text{ in./sec},$$

and the momentum transferred to the beam by the thruster is:

$$v\,M_{eff} = (0.75 \text{ in./sec})\,(1.02 \times 10^{-3} \text{ lb-sec}^2/\text{in.})$$
$$= 7.65 \times 10^{-4} \text{ lb-sec.}$$

This compares with a theoretical predicted impulse bit of 7.55×10^{-4} lb-sec.

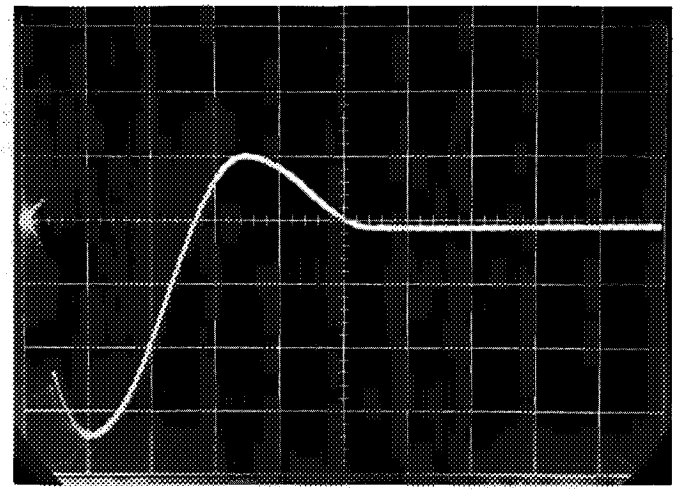


Fig. 24. Impulse bit

Thus, from these two data points it appears that very good steady-state thrust level and impulse bit measuring can be accomplished using a low-frequency beam. The steady-state thrust level measured is within 3.2% of theory, while the impulse bit size correlates within 1.3%.

* Obtained from calibration of beam and transducer under known load.

As can be seen from Figs. 23 and 24, there is considerable damping in the beam motion. This has been attributed to contact between the plunger and the housing of the linear transducer. The damping could have a small effect on the steady-state thrust level but a very large effect on the impulse bit measurement. Therefore, work in the future is being directed toward elimination of the linear transducer by the use of strain gages. Once the damping has been reduced, more data will be collected on various sizes and configurations of valve/nozzles.

IV. Guidance and Control Research

A. Sound Propagation in Liquid Helium: Comparison of Velocity and Attenuation Data with the New Theory of Khalatnikov and Chernikova

W. M. Whitney

In a recent report (*SPS 37-36,* Vol. IV, pp. 61-63), measurements of the velocity of sound in liquid helium under saturated vapor pressure (Ref. 1) were compared with the predictions of a theory developed by I. M. Khalatnikov in 1950 (Ref. 2). The curves used in the previous discussion are shown again in Fig. 1. The two unbroken lines represent smoothed values of differences in the measured velocity u_1 over the frequency intervals 1.00 to 3.91 Mc/sec and 1.00 to 11.9 Mc/sec, plotted against temperature. The corresponding theoretical values obtained from the 1950 theory, which are shown as dashed lines, give satisfactory representation of the heights and positions of the dispersion peaks, but they fall considerably below the experimental values on the low-temperature side. There is similar disagreement between measurements of the absorption coefficient and values given by the same theory at low temperatures (Refs. 3, 4).

Recently, Khalatnikov and Chernikova (Refs. 5-7) have published a new description of sound propagation in liquid helium that accounts, in a much more satisfactory way than the earlier theory, for the behavior of the absorption coefficient in the neighborhood of its peak and at temperatures just below. In the present paper we compare the predicted dispersion with our velocity measurements and examine how well the calculated absorption agrees with existing experimental data below 0.5°K, where the validity of the theory might be questioned. This work has been done in collaboration with C. E. Chase (National Magnet Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts).

Since the Khalatnikov-Chernikova theory represents a considerable advance in the description of transport behavior in liquid helium, we shall briefly describe the underlying physical model. It was shown by Landau (Ref. 8) that many of the equilibrium thermal properties

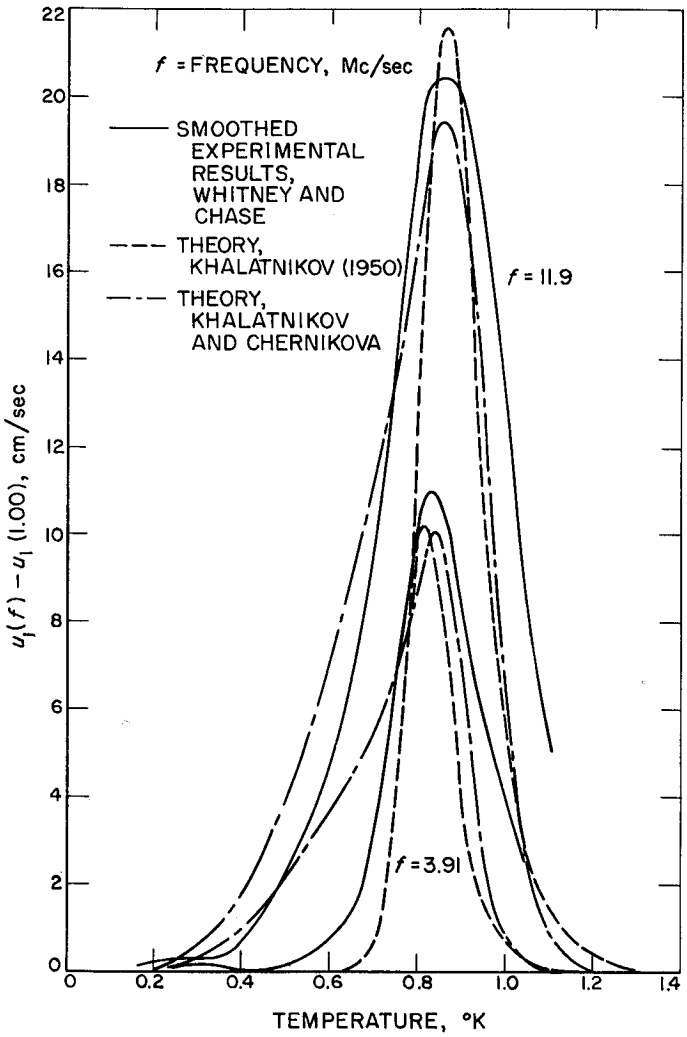


Fig. 1. Velocity difference over the frequency intervals 1.00 to 3.91 Mc/sec

of liquid helium can be understood if it is assumed that the energy of the liquid is distributed among two types of particle-like excitations or "quasi-particles;" these are called phonons and rotons. Phonons are sound quanta with energy $\epsilon = pc$ and momentum $p = h/\lambda$, where h is Planck's constant, c the velocity of the phonon, and λ its wavelength. A collection of these excitations, suitably distributed in energy, can account for all the properties of liquid helium that arise from its elasticity, just as a collection of photons (light quanta) can represent all the properties of the electromagnetic field. As for rotons, they were originally conceived by Landau to be collective motions of helium atoms associated in some way with quantized angular momentum or vorticity in the fluid. Landau postulated that they had an energy spectrum

$$\epsilon = \Delta + (p - P_0)^2/2\mu,$$

so that, to excite a roton, a certain threshold energy Δ is required. The question concerning what rotons are has yet to be given an entirely satisfactory answer, although it is believed that they are not elementary vortices, but there is nevertheless abundant experimental evidence that Landau's model is essentially correct. The most striking confirmation is provided by experiments, originally suggested by Cohen and Feynman (Ref. 9), in which neutrons are scattered inelastically from liquid helium. In their interactions with the liquid, the neutrons undergo changes in energy and momentum that are found to be fully compatible with the energy-vs-momentum relations proposed by Landau for phonons and rotons (Refs. 10-12). Furthermore, values of the roton constants Δ, P_0, and μ obtained from the neutron studies are in full agreement with those deduced indirectly from the analysis of measurements of the specific heat, entropy, and other thermal properties.

Thus the phonon-roton model successfully describes the equilibrium thermodynamic properties of liquid helium. Can it be extended to give an account of transport behavior as well? In a gas, whose energy is distributed among its atoms and molecules, the transport properties — first and second viscosity, thermal conduction, and self-diffusion — arise from atomic collisions, and the temperature dependence of the transport coefficients depends upon the detailed characteristics of the atomic interactions, as reflected in the scattering cross-sections. The first attempt to predict transport behavior in liquid helium from a study of interactions among the phonons and rotons was made by Landau and Khalatnikov in 1949 (Ref. 13), when they calculated the scattering cross-sections for certain collision processes and determined the coefficient of ordinary viscosity by solving a Boltzmann equation. This work was expanded by Khalatnikov in his 1950 paper on the absorption of first and second sound in liquid helium (Ref. 2), for which it was necessary to take additional scattering processes into account and to calculate the three independent coefficients of second viscosity.

In the recent work of Khalatnikov and Chernikova, the question of what interactions are important in restoring thermal equilibrium in liquid helium receives a detailed investigation, and some earlier conclusions are modified. Their theory rests upon the validity of a number of statements concerning the relative importance of certain interactions; these are based upon their present studies and upon the earlier ones mentioned above.

Below approximately 1.6°K, liquid helium may be considered a mixture of two weakly interacting (but nonetheless interacting) quasi-particle gases — the phonons and rotons. In a real gas, the total number of atoms of course remains constant as the temperature and pressure change. In liquid helium, however, the numbers of phonons and rotons, N_{ph} and N_r, per unit volume can change, and in thermal equilibrium have definite values that depend upon the local temperature and pressure. When these conditions are altered, for example when the pressure is increased by a passing sound wave, equilibrium is disturbed — the phonon number density and the energy density go down, while the energy of an individual phonon increases; and the roton number density and energy density go up, while the energy of a roton decreases. Energy equilibrium is achieved through "elastic" collisions among the phonons and rotons, in which the phonon and roton number densities do not change. The number densities themselves arrive at their equilibrium values through "inelastic" collisions.

Let us consider first the interactions of phonons with each other. If pressure changes were strictly proportional to density changes in liquid helium, phonons could not interact, and the establishment of a thermodynamic state of the system would be impossible. In fact, the pressure-density relation is nonlinear, and sound waves at different frequencies undergo mixing and suffer harmonic distortion; in other words, energy is exchanged among the interacting phonons and new phonons are generated. The analyses by Landau and Khalatnikov and by Khalatnikov and Chernikova show that energy equilibrium among all the phonons travelling in a given direction is rapidly brought about by the so-called four-phonon process, in which two phonons interact to form two others with the same total energy and momentum. Changes in the number density N_{ph} are brought about principally by the five-phonon process, in which two phonons coalesce and three emerge (or vice versa). This process, too, takes place most rapidly among phonons whose wave-vectors are parallel.

To calculate the cross-section for the scattering of rotons by rotons has proved difficult because the interaction potential is not known. Landau and Khalatnikov assumed that collisions among rotons could be treated like those among billiard balls. This hypothesis made subsequent analysis quite tractable, and led to the prediction that the viscosity of liquid helium above approximately 1.5°K would be independent of temperature, as is observed. The strength of the hard-core interaction can be evaluated from experimental data, and it is found that the roton collision cross-section is sufficiently large that the rotons may always be considered to be in local thermodynamic equilibrium, at least at temperatures above approximately 0.6°K and at frequencies comparable with those used in past studies of sound propagation. This fact, whose recognition is one of the new features in the Khalatnikov-Chernikova treatment, enables the roton distribution function to be characterized by an effective roton temperature T_r. Because of the smallness of the roton collision time t_{rr}, roton-roton interactions turn out to have no appreciable influence on the propagation of ordinary sound in liquid helium, although they do affect second sound.

Interactions between rotons and phonons can occur because the roton energy gap Δ varies with density. Thus, when the fluid is squeezed, the equilibrium values for the thermodynamic functions for the rotons change. According to Khalatnikov and Chernikova, it is the elastic scattering of phonons by rotons that brings the phonon and roton gases into thermal equilibrium with each other, and that is responsible — because the process is a relatively slow one — for the anomaly in the absorption and dispersion of ordinary sound near 1°K. Above 0.9°K it is apparently phonon-roton scattering that also brings about equilibrium among the phonons travelling in different directions. The phonon-roton collision time τ_{phr} increases rapidly as the temperature falls, however, and below 0.9°K the relaxation of the non-equilibrium phonon distribution toward an isotropic one proceeds through large-angle phonon-phonon scattering. Below 0.6°K, the characteristics of sound propagation should depend entirely upon phonon-phonon interactions.

One of the strengths of the Khalatnikov-Chernikova treatment is the explicitness of the above physical model, which serves as a basis for their subsequent calculation of the temperature and frequency dependence of the velocity and the coefficient of absorption of both first and second sound.

In the usual approach, the characteristics of sound propagation in a fluid are derived from a set of differential equations that impose certain mechanical and thermal constraints on the wave motion. Euler's equation relates the acceleration of a given fluid element to the forces acting on it — which may include viscous forces as well as those arising from gravity or from a pressure gradient. A continuity equation requires that the fluid mass be conserved, and an additional equation takes account of energy transfer by the fluid and of the conversion of mechanical energy to heat (entropy production). The

complete set of equations is linearized and solved for small periodic deviations of the variables about their equilibrium values. The condition that a solution exist is summed up in a so-called compatibility equation, which relates the frequency to a complex wave vector, to phenomenological transport coefficients (the viscosity and thermal conductivity, for example), and to thermodynamic functions of state. The wave velocity and the coefficient of absorption are derived from the real and imaginary parts of the wave vector.

The above procedure, which is also applicable to liquid helium above 1°K, yields useful expressions for the velocity and absorption coefficient provided that the period $2\pi/\omega$ of the sound wave is shorter than the average time τ required for collisions to bring about thermal equilibrium — otherwise it is meaningless to speak of equilibrium states of the system. In ordinary fluids, very high frequencies are required to enter the domain $\omega\tau >> 1$, but, in liquid helium, $\omega\tau_{phr} \simeq 1$ just below 1°K at frequencies of a few Mc/sec. To obtain a description of sound propagation that is applicable here, a different treatment is required. Following lines indicated in an earlier paper by Andreev and Khalatnikov (Ref. 14), Khalatnikov and Chernikova replace Euler's equation and the heat-transfer equation by two separate Boltzmann equations, one each for the roton and phonon distribution functions and with collision terms included. The solutions of these equations are required to satisfy the equation of continuity and one additional constraint, applicable only to liquid helium, that we need not discuss here. After lengthy analysis, Khalatnikov and Chernikova obtain expressions for the sound velocity $u_1(\omega,T)$ and the absorption coefficient $\alpha(\omega,T)$. We shall give their results for the temperature region below 1.2°K:

$$u_1(\omega,T) = u_1(0,T) - \frac{1}{2}c\left(\frac{\rho_{nph}}{\rho}\right)\mathcal{R}e\ \phi(\tau_{phr},\tau_{phph}),$$

$$\alpha(\omega,T) = \frac{1}{2}\frac{\omega}{c}\left(\frac{\rho_{nph}}{\rho}\right)\mathcal{I}m\ \phi(\tau_{phr},\tau_{phph}),$$

where ρ is the density, and

$$\rho_{nph}/\rho = \frac{16\pi^5}{45}\frac{k_B^4}{h^3}\frac{T^4}{\rho c^5}$$

(k_B = Boltzmann's constant). For our purposes it may be assumed that the phonon velocity $c = u_1(0,0) = 2.383 \times 10^4$ cm/sec (Ref. 1). The time constants τ_{phr} and τ_{phph} characterizing phonon-roton scattering and small-angle phonon-phonon scattering are given by the expressions

$$\tau_{phr}^{-1} = (2\pi)^{17/2}\,\Gamma\frac{k_B^{9/2}}{h^7}\left(\frac{P_0^4\mu^{1/2}}{\rho^2c^5}\right)T^{9/2}\,e^{-\Delta/kT},$$

$$\tau_{phph}^{-1} = \frac{9(13!)}{2^{13}}\frac{k_B^9}{h^7}\frac{(u+1)^4}{(\rho c^5)^2}\,T^9,$$

where $u = (\rho/c)\partial c/\partial\rho$, and

$$\Gamma = \frac{2}{9} + \frac{1}{25}\left(\frac{P_0}{\mu c}\right)^2 + \frac{2}{9}\left(\frac{P_0}{\mu c}\right)A + A^2,$$

with

$$A = \left(\frac{\rho^2}{P_0 c}\right)\frac{\partial^2\Delta}{\partial^2\rho} + \left(\frac{P_0}{\mu c}\right)\left(\frac{\rho}{P_0}\frac{\partial P_0}{\partial\rho}\right)^2.$$

To simplify later expressions we introduce the notation $\theta_{phr} = \omega\tau_{phr}, \theta_{phph} = \omega\tau_{phph}, \theta = \omega\tau$, where $\tau^{-1} = \tau_{phr}^{-1} + \tau_{phph}^{-1}$.

The general expression given by Khalatnikov and Chernikova for the function $\phi\ (\tau_{phr}, \tau_{phph})$ is very complicated. We shall give its real and imaginary parts here in the forms we have used for our computations:

$$\mathcal{R}e\ \phi\ (\tau_{phr}, \tau_{phph}) = 1 - 3\,(N_rD_r + N_iD_i)\,(D_r^2 + D_i^2)^{-1},$$

$$\mathcal{I}m\ \phi\ (\tau_{phr}, \tau_{phph}) = \theta_{phr}^{-1} - 3\,(N_iD_r - N_rD_i)\,(D_r^2 + D_i^2)^{-1},$$

$$N_r = \frac{1}{2}u^2\ln\,[1 + 4\,\theta^2] + B_rC_r - B_iC_i,$$

$$N_i = -u^2\tan^{-1}2\theta + B_iC_r + B_rC_i,$$

$$B_r = 2u + 1 - (2\beta + 1)\,\theta_{phr}^{-2},$$

$$B_i = [2\,(u + 1) + \beta\,(1 - \theta_{phr}^{-2})]\theta_{phr}^{-1} - 3\,u^2\,\theta_{phph}^{-1},$$

$$C_r = \frac{1}{2}\ln\,[1 + 4\,\theta^2] + \theta^{-1}\tan^{-1}2\,\theta - 2,$$

$$C_i = (2\,\theta)^{-1}\ln\,[1 + 4\,\theta^2] - \tan^{-1}2\,\theta,$$

$$D_r = 2 - \{\theta^{-1} - \beta\,\theta_{phr}^{-1}\}\tan^{-1}2\,\theta + 3\,\theta_{phph}^{-1}\{C_i + \beta\,C_r\theta_{phr}^{-1}\},$$

$$D_i = -\frac{1}{2}\{\theta^{-1} - \beta\,\theta_{phr}^{-1}\}\ln\,[1 + 4\,\theta^2] - 3\,\theta_{phph}^{-1}\{C_r - \beta\,C_i\theta_{phr}^{-1}\},$$

$$\beta = 3\,kT/\mu c^2.$$

Note that B_r and B_i increase rapidly at high temperatures where θ_r becomes much smaller than unity. At these temperatures, C_r and C_i are roughly proportional to θ_r^2 and to θ_r^3, so to obtain accurate values for N_r and N_i it is necessary to evaluate the term $\ln\,(1 + 4\,\theta^2)$ and $\tan^{-1}(2\,\theta)$ very carefully.

We have calculated $u_1(\omega,T)$ from these expressions, with $u_1(0,0) = c = 2.383 \times 10^4$ cm/sec (Ref. 1), $\rho = 0.1451$ g/cm^3 (Ref. 15), $u = (\rho/c)\partial c/\partial\rho = 2.64$ (Ref. 16), $\Delta/k_B = 8.65\,K°$, $\mu = 1.06 \times 10^{-24}$ g, $P_0 = 2.02 \times 10^{-19}$ g-cm/sec (Ref. 11), and $\Gamma = 2.6$. The last number is obtained if $A = -0.1$, as suggested by Khalatnikov and Chernikova (Ref. 5) after an analysis of neutron-scattering data. The predicted velocity differences for the frequency intervals 1.00 to 3.91 and 1.00 to 11.9 Mc/sec are shown as broken lines in Fig. 1.

On the whole, the agreement between the experimental curves and those derived from the new theory is quite good. The heights of the dispersion maxima and their locations are accounted for, as is the disappearance of the dispersion in the neighborhood of 1.2°K. It is particularly noteworthy that all of the parameters needed for the evaluation of the theoretical equations can be obtained from independent experiments. In contrast, Khalatnikov's 1950 equations contained two adjustable constants. These were related to the transition probabilities for the inelastic five-phonon process, and for the process in which an energetic phonon collides with a roton and is converted to a roton. According to the earlier theory, these two scattering mechanisms were responsible for the anomalous absorption below 1°K. On the basis of their more recent analysis, Khalatnikov and Chernikova state that the five-phonon process contributes to the absorption only at temperatures higher than 1.2°K, well above the absorption and dispersion peaks, and that the influence of inelastic phonon-roton scattering can be ignored altogether. In view of these assertions, it is remarkable that the older theory gave such good account of the absorption and dispersion in the neighborhood of their peaks.

Despite the good general agreement, there are nevertheless discrepancies between the experimental results and the new theoretical predictions. Differences above 0.9°K are compatible with estimated errors in both temperature and velocity and cannot be considered significant. Because of the methods used in making measurements and in converting the experimental data to velocity values, however, the results plotted in Fig. 1 are most reliable on the low-pressure side of the velocity maximum. Here the predicted dispersion exceeds what was measured by amounts that lie well outside the experimental uncertainty, particularly for the results corresponding to the interval 1.00 to 3.91 Mc/sec.

More accurate measurements will be required to establish whether the disagreement is the fault of the theory or the experiment. It should be pointed out, however, that predictions of the new theory concerning the behavior of the attenuation coefficient $\alpha(\omega,T)$ are also not in perfect agreement with experimental results. The equations given above for α accurately reproduce the height of the absorption peak (see Fig. 2) over the frequency interval 1 to 30 Mc/sec, and yield good agreement with experimental data on the high-temperature side (although above 1.2°K an adjustable constant relating to the five-phonon process comes into play). Below 0.8°K, however, the theory predicts a behavior qualitatively different from what has been observed. We illustrate the nature of the disagreement with data drawn from previous experiments. In Fig. 3, the attenuation is plotted against temperature for three experiments, all carried out at frequencies in the neighborhood of 12 Mc/sec. Although the three sets of points exhibit some small systematic discrepancies, they are consistent with the dashed line, for which $\alpha \propto T^3$. The theoretical values, calculated for the frequency 11.8 Mc/sec, lie considerably below the experimental ones and vary as T^4. In Fig. 4, the absorption coefficient at 0.4°K is plotted against frequency, the points representing seven different experiments. On the basis of measurements of the absorption coefficient at frequencies below

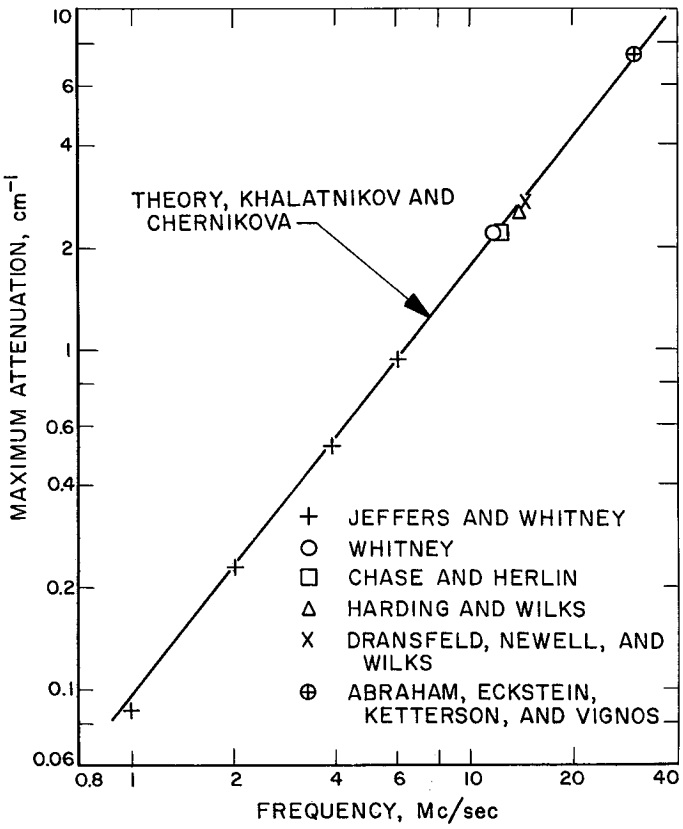


Fig. 2. Height of attenuation peak versus frequency

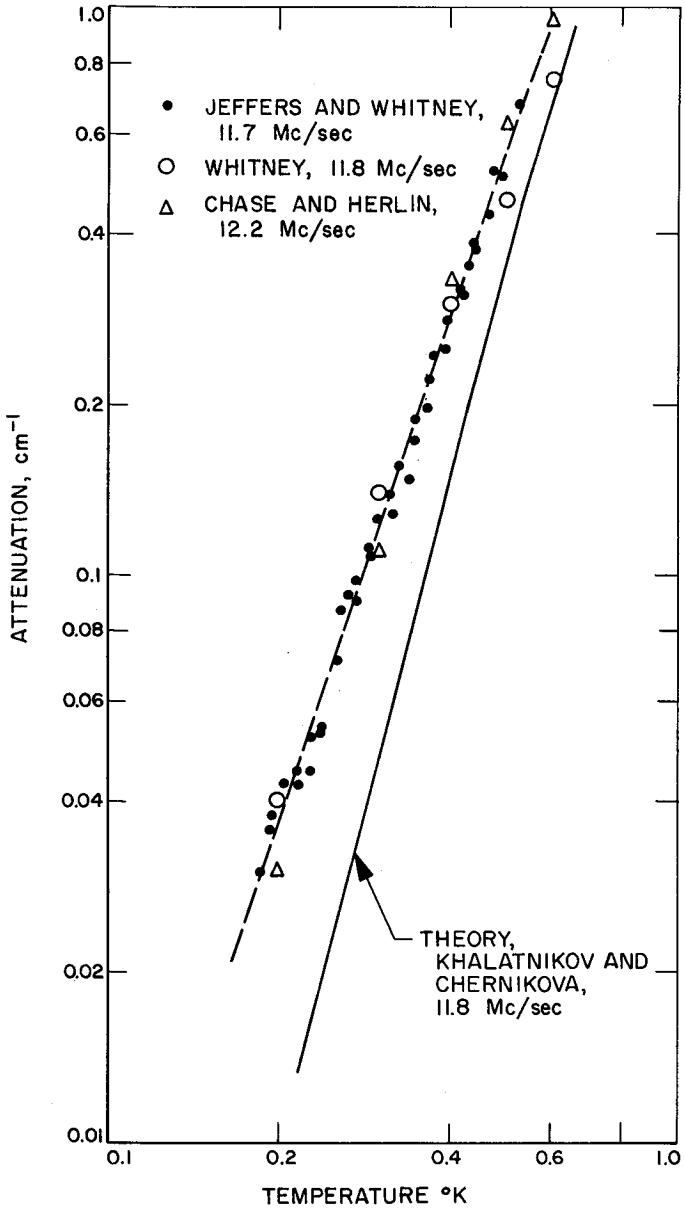


Fig. 3. Attenuation versus temperature

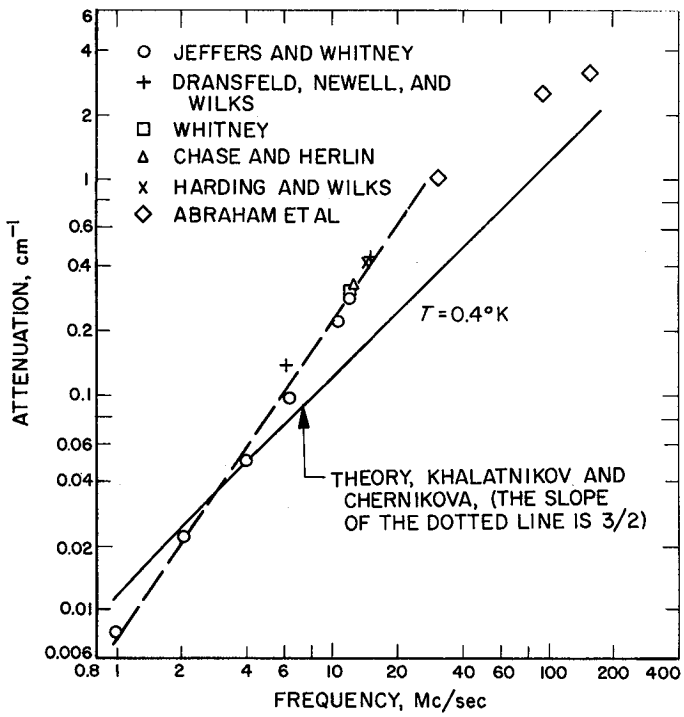


Fig. 4. Attenuation versus frequency

12 Mc/sec and temperatures below 0.8°K, it was concluded in an earlier paper (Ref. 17) that, within these regions, $\alpha \propto \omega^{3/2} T^3$. Recent measurements by Abraham, Eckstein, Ketterson, and Vignos (Ref. 18), whose results are included in Fig. 4, show that the $\omega^{3/2}$ frequency dependence does not persist at higher frequencies, but it is apparent that the (approximately) linear variation of the absorption with frequency that is predicted by the Khalatnikov-Chernikova theory (solid line in Fig. 4) is not followed. Abraham et al. find that, at 30, 90, and 150 Mc/sec, $\alpha \propto T^4$, as predicted by Khalatnikov and Chernikova and by a number of theories based upon the three-phonon process (Ref. 19), but the constant of proportionality must be considerably different from what would be expected.

Thus we conclude that, while the new theory of Khalatnikov and Chernikova gives an excellent account of the observed behavior of the velocity and absorption of sound in liquid helium near 0.9°K and above, it does not yet provide the final answer at lower temperatures. It is, in fact, explicitly stated by Khalatnikov and Chernikova that their results are applicable only at temperatures and frequencies for which ωt_{rr}, $\omega t_{phph} << 1$, where t_{rr} is the collision time for roton-roton scattering and t_{phph} is a time constant for small angle phonon-phonon scattering, related to τ_{phph} defined above. Below 0.5°K, these inequalities are violated, even at 1 Mc/sec, and it may be incorrect to expect the theory to be valid here, even though Khalatnikov and Chernikova extended their own numerical calculations to temperatures as low as 0.4°K in comparing the results of their treatment with the attenuation measurements of Jeffers and Whitney (Ref. 17). In the neighborhood of the absorption and dispersion peaks, certainly, the new theory is manifestly better than the earlier one, and represents a significant advance in the description of transport processes in liquid helium.

B. Space-Charge-Limited Electron Current in Germanium

A. Shumka

1. Introduction

Germanium $n^+-\pi-n^+$ planar solid-state diodes were fabricated (*SPS 37-32*, Vol. IV, pp. 64–66) and evaluated *SPS 37-35*, Vol. IV, pp. 49–52) at JPL. These structures were found to be suitable for the purpose of investigating space-charge-limited (SCL) electron current in near-intrinsic high-purity germanium. The effect of space-charge on limiting the flow of current in the solid-state diodes was determined from the *I-V* characteristics measured at various ambient temperatures.

2. I-V Characteristic Measurements

Saw-tooth voltage signals were applied across a solid-state diode and the resulting *I-V* characteristic was displayed on the screen of an oscilloscope. The repetition rate, the width, and the amplitude of the saw-tooth signal were varied to observe if there were any transient or thermal effects in the solid-state diode. These effects were not observed over wide ranges of pulse widths and repetition rates (duty cycles as high as 50% and pulse widths ranging from 10μs to 10ms). A duty cycle of 10% and a pulse width of 1 ms were arbitrarily chosen for the *I-V* measurements. The numerical values of the current and voltage were read directly from the oscilloscope trace. A block diagram of the measuring apparatus is shown in Fig. 5.

Different temperature baths were used for obtaining various ambient temperatures. The solid-state diodes were immersed in a liquid nitrogen bath for an ambient temperature of 78°K. Liquid pentane baths maintained at dry ice and ice temperatures were also used for ambient temperatures of 195°K and 273°K, respectively. The measured *I-V* characteristics for an $n^+-\pi-n^+$ germanium solid-state diode for ambient temperatures of 78°K, 195°K, and 273°K are shown in Figs. 6(a), (b), (c), respectively. This particular solid-state diode had an

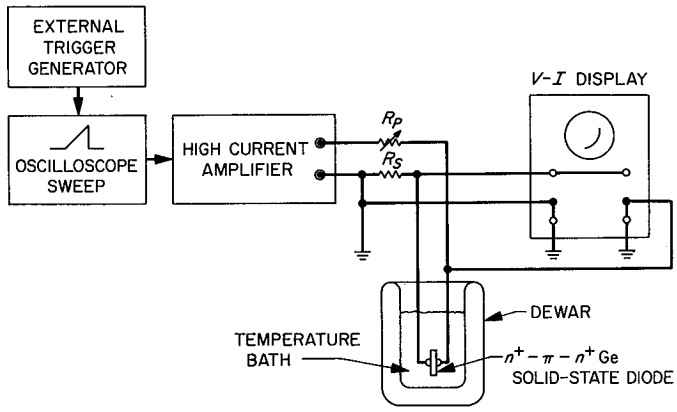


Fig. 5. Block diagram of the electrical measuring apparatus

area, $A \cong 3.8 \times 10^{-3}\text{cm}^2$, a base-width, $W = 110\ \mu$, and a punch-through voltage, $V_{pt} = 7.6$ V.

3. SCL Electron Current

Theoretical *I-V* characteristics for SCL electron current are fitted to the experimental points in Fig. 6. These theoretical curves were obtained by Dacey (Ref. 24) and are based on the field-dependent mobility observed by E. J. Ryder (Ref. 25) for electrons in germanium. The agreement between theory and experiment is very good, particularly in Figs. 6(a), (b). In Fig. 6(c), the deviation that is observed in the low-current range is due to the masking effect of the larger saturation current expected at the higher ambient temperature (Ref. 26). The theory does not take into account the saturation of drift velocity. Consequently a deviation between theory and experiment would be expected where drift velocity saturation occurs, that is, for electric fields greater than 5×10^3V/cm. The deviation observed in the high-current range is where drift velocity saturation is expected to begin.

There are no previously published results on SCL electron current in germanium. Consequently experimental results on $n^+-\pi-n^+$ solid-state diodes of various base-width will be analyzed in greater detail to determine more conclusively that observed *I-V* characteristics are that of SCL electron flow. These results will be published in a future *Space Programs Summary*.

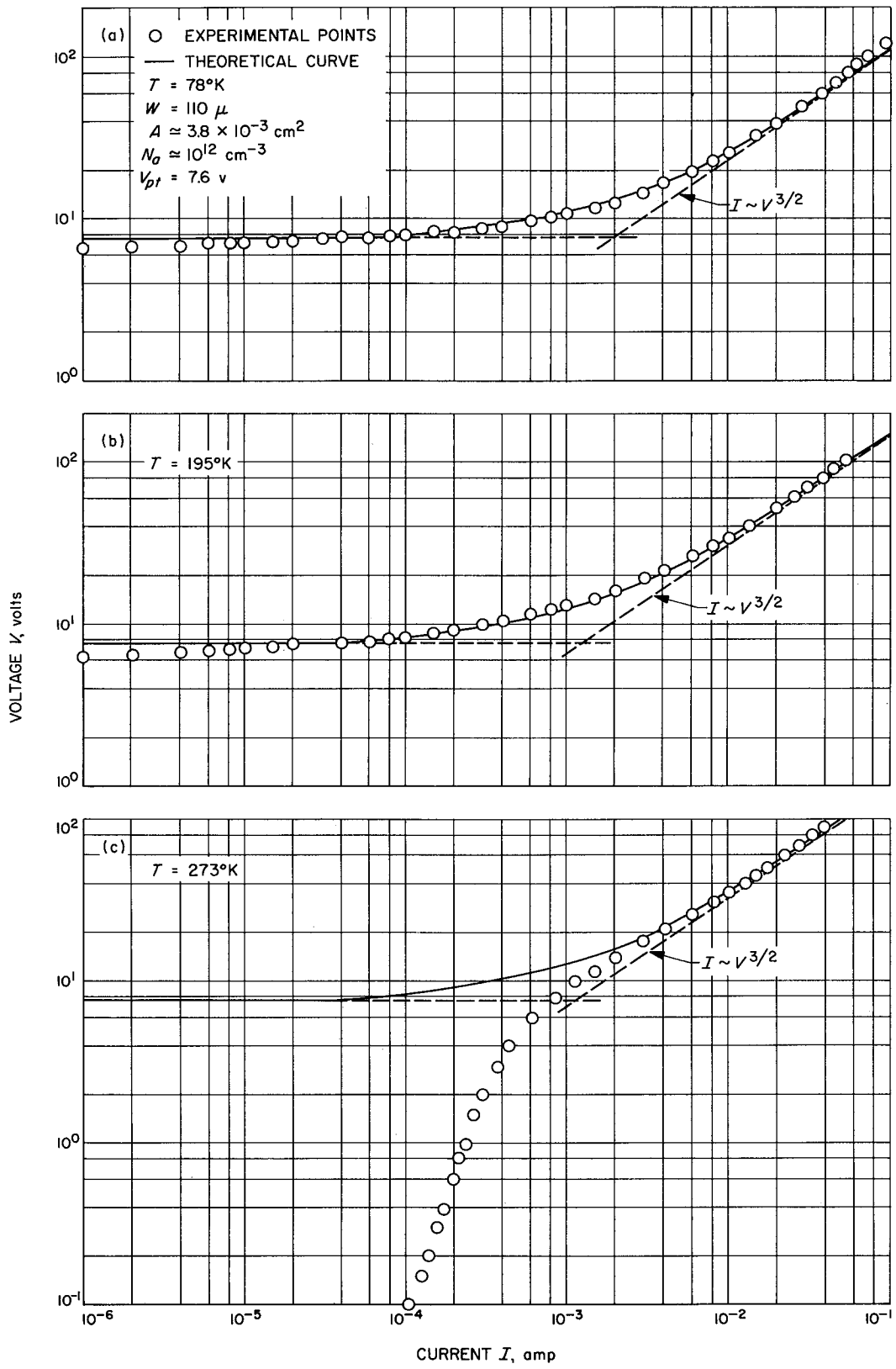


Fig. 6. A comparison between theory and experiment for SCL current in an $n^+ - \pi - n^+$ germanium solid-state diode at various temperatures

References

1. Whitney, W. M., and Chase, C. E., *Ultrasonic Velocity and Dispersion in Liquid Helium II,* to be published.
2. Khalatnikov, I. M., *JETP, 20:* 243, 1950.
3. Arkhipov, R. G., *Dokl. Akad, Nauk SSSR, 98:* 747, 1954.
4. Chase, C. E., *American Journal of Physics, 24:* 136, 1956.
5. Khalatnikov, I. M., and Chernikova, D. M., *JETP, 49:* 1957, 1965.
6. Khalatnikov, I. M., and Chernikova, D. M., *JETP, 50:* 411, 1966.
7. Khalatnikov, I. M., and Chernikova, D. M., *JETP Pis'ma, 2:* 566, 1965; English transl.: *Soviet Physics—JETP Letters, 2:* 351, 1966.
8. Landau, L. D., *J. Phys. (Moscow), 5:* 71, 1941; 11: 91, 1947.
9. Cohen, M., and Feynman, R. P., *Physical Review, 107:* 13, 1957.
10. Palevsky, H., Otnes, K., Larrson, K. E., Pauli, R., and Stedman, R., *Physical Review, 108:* 1346, 1957.
11. Yarnell, J. L., Arnold, G. P., Bendt, P. J., and Kerr, E. C., *Physical Review, 113:* 1379, 1959.
12. Henshaw, D. G., and Woods, A. D. B., *Physical Review, 121:* 1266, 1961.
13. Landau, L. D., and Khalatnikov, I. M., *JETP, 19:* 637, 709, 1949.
14. Andreev, A., and Khalatnikov, I., *JETP, 44:* 2058, 1963; English transl.: *Soviet Physics—JETP, 17:* 1384, 1963.
15. Kerr, E. C., and Taylor, R. D., *Annals of Physics (N. Y.), 26:* 292, 1964.
16. Atkins, K. R., and Stasior, R. A., *Canadian Journal of Physics, 31:* 1165, 1953.
17. Jeffers, W. A., Jr., and Whitney, W. M., *Physical Review, 139:* 1082, 1965.
18. Abraham, B. M., Eckstein, Y., Ketterson, J. B., and Vignos, J. H., *Physical Review Letters, 16:* 1039, 1966.
19. Papers dealing with the three-phonon theories are cited in Ref. 17.
20. Whitney, W. M., *Physical Review, 105:* 38, 1957.
21. Chase, C. E., and Herlin, M. A., *Physical Review, 97:* 1447, 1955.
22. Harding, G. O., and Wilks, J., *Proceedings of the Royal Society (London) A268:* 424, 1962.
23. Dransfeld, K., Newell, J. A., and Wilks, J., *Proceedings of the Royal Society (London), A243:* 500, 1958.
24. Dacey, G. C., *Physical Review, 90:* 759-764, 1953.
25. Ryder, E. J., *Physical Review, 90:* 766-769, 1953.
26. Shumka, A., *Space-Charge-Limited Current in Germanium,* pp. 50-54, Division of Engineering and Applied Science Report, California Institute of Technology, Pasadena, Calif. May 1964.

ENGINEERING MECHANICS DIVISION

V. Applied Mechanics

A. Thermal Joint Conductance

J. A. Hultberg

The effect of a variable thermal interface conductance has been discussed earlier by T. J. Lardner (*SPS 37-39*, Vol. IV, p. 83; *SPS 37-21*, Vol. IV, p. 53) and by the author (*SPS 37-38*, Vol. IV, p. 61). The geometry used in both of these investigations was a circular region with a radially variable interface conductance of the form

$$h = h_0\left[1 - C\left(\frac{r}{a}\right)\right] \tag{1}$$

where h is the conductance, C is the slope of the linear variation, and r and a are respectively a coordinate and a dimension defined as shown in Fig. 1. In the former investigation, a constant flux boundary was used along the lower surface; in the latter, an isothermal boundary was used along both the upper and the lower surfaces. In both investigations, the conducting area encompassed the entire interface area.

In the present investigation, the linearly variable interface conductance has the form given by Eq. 1, with the parameter C having a value greater than unity. The resulting conducting area will be less than the total interface area. A related problem has been investigated by A. M. Clausing for two cases: in the first (Ref. 1), the interface region is composed of both ideal insulating and zero interface resistance areas; in the second (Ref. 2), it is composed of ideal insulating and uniform finite resistance areas. For the investigation reported here, the interface region consists of both ideal insulating and nonuniform finite resistance areas.

The governing equations for the circular contact geometry in Fig. 1, with isothermal boundary conditions along

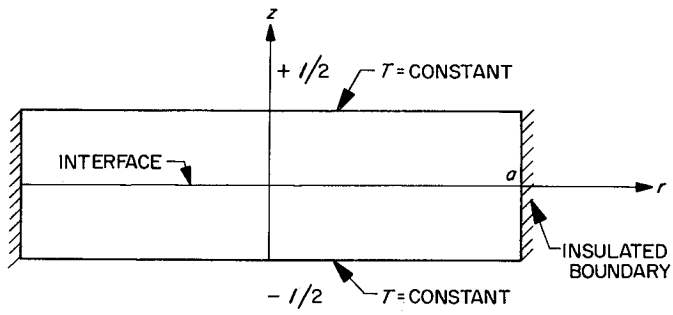


Fig. 1. Geometry of a circular region with an interface

the upper and lower surfaces, insulating boundary conditions along the periphery, and a variable interface conductance of the form stated in Eq. 1, are given in *SPS 37-38*, Vol. IV, p. 61. The numerical results may be presented in terms of a dimensionless resistance

$$R^* = k\pi a\left[\frac{T(r, +l/2) - T(r, -l/2)}{q_{total}} - \left(\frac{1}{k\pi a^2} + \frac{1}{\bar{h}\pi a^2}\right)\right] = \frac{\Delta l}{a} \tag{2}$$

where k is the thermal conductivity, Δl is the additional effective length caused by the combined effect of the geometric constriction and the variable interface conductance, and $\bar{h}$, the average value of the interface conductance, is defined by

$$\bar{h} = \frac{2}{a^2}\int_0^a h(r)\, r\, dr \tag{3}$$

The governing equations and the appropriate boundary conditions were solved numerically on an IBM 7094 computer, with $\bar{h}l/k = 1.0$ for various values of l/a, and the value of C varied parametrically for values of 4.0, 2.0, 1.3333, and 1.0. The Biot modulus ($\bar{h}l/k$) is a measure of the importance of the interface conductance relative to the conductance of the material.

The interface conductance distribution for the four values of C is given in Fig. 2, where the limit of the radius of the circular conducting region is 0.25, 0.50, 0.75, and 1.0 r/a. From the resulting dimensionless resistances (Fig. 3), it is apparent that the resistance caused by the constriction and the variable interface conductance becomes independent of l/a for large values of l/a. The temperature distributions for the upper and lower interfaces for the four values of C previously listed, for a particular value of l/a and $\bar{h}l/k$, are shown in Fig. 4. Extreme bending of the heat flow lines for large values of C may be inferred from the indicated temperature distributions. The large values of dimensionless resistance are seen to have the correct trend. A typical distribution of the temperatures throughout the circular region is shown in Fig. 5. The actual conductance of a circular region with an interface would be given by

$$K_{eff} = \frac{k\pi a^2}{l\left[1 + R^*\frac{a}{l} + \frac{k}{\bar{h}l}\right]} \tag{4}$$

Future work will include investigations of a radiation boundary condition along one of the flat surfaces and

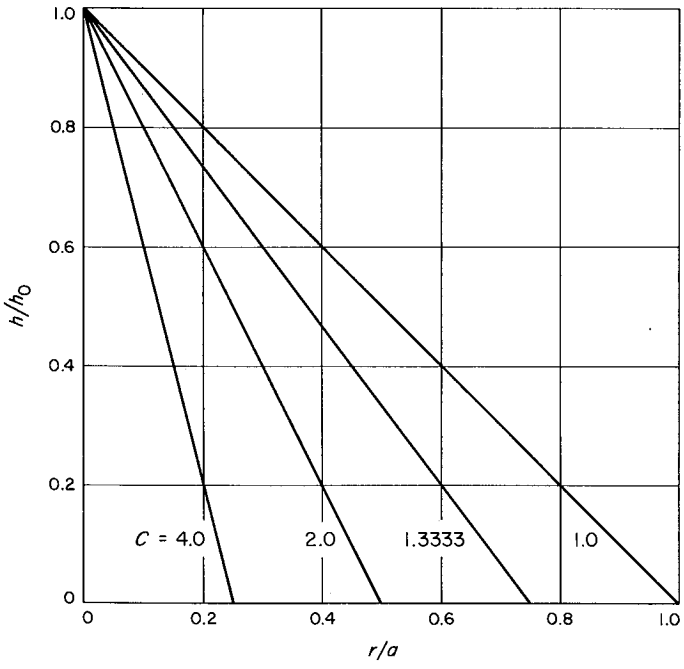


Fig. 2. Interface conductance vs r/a

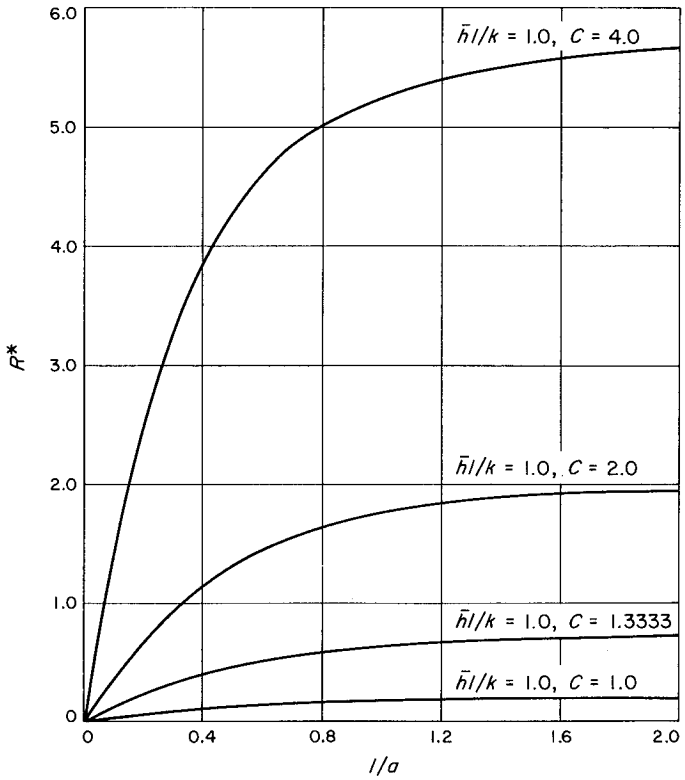


Fig. 3. Dimensionless resistances vs l/a

interface conductances generated by combining pressure distributions around circular contacts with conductance versus pressure curves.

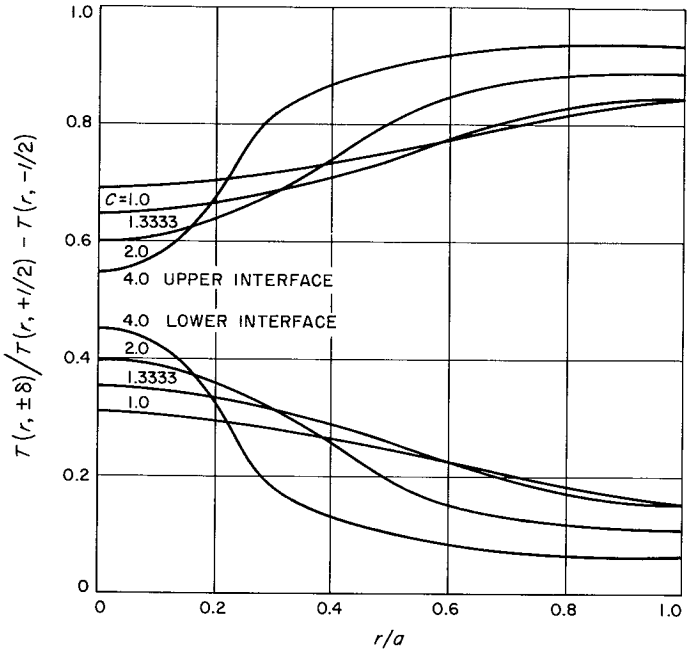


Fig. 4. Interface temperatures for $\bar{h}l/k = 1.0$ and $l/a = 2.0$

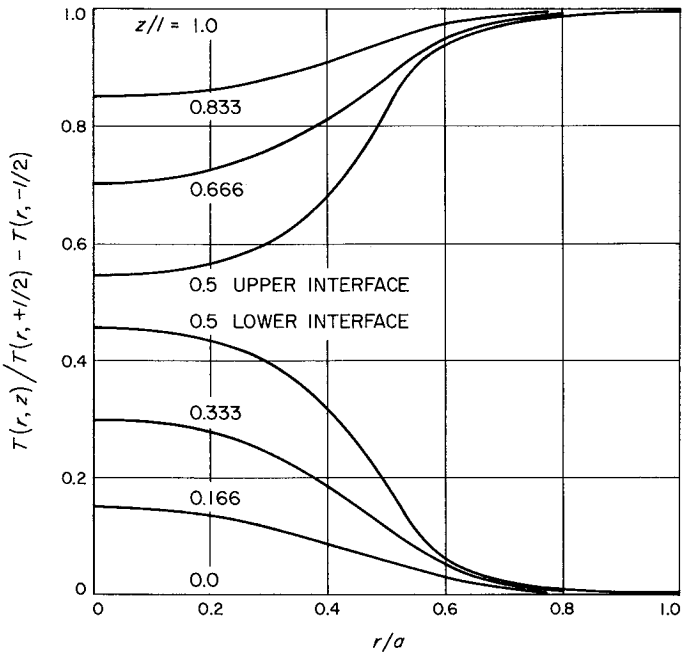


Fig. 5. Temperature distribution for $\bar{h}l/k = 1.0$, $l/a = 0.4$, and $C = 0.5$

B. Additional Tests on the Half-Scale Thermal Model of the *Mariner IV* Spacecraft

C. A. Rhodes

Tests were performed on the half-scale thermal model (TSM) of the *Mariner IV* spacecraft in the JPL 10-ft space simulator. Temperatures obtained were compared with actual *Mariner IV* flight data and with data previously obtained from the TSM when tested in the NASA Lewis Research Center space simulator. *Mariner IV* flight temperatures used in the comparison are referenced in *SPS 37-35*, Vol. II, pp. 12–19; those from the previous TSM tests are listed in Ref. 3.

Tests were performed at conditions simulating the 2nd, 98th, 180th and 234th days of the *Mariner IV* flight. The latter three tests had been carried out earlier in the NASA Lewis Research Center space simulator. Table 1 indicates the TSM power dissipation and solar intensities used in each test. In the first, identified as Earth cruise 1, the cavity amplifier and battery charger were on and the TWT was not operating, as was the case for the *Mariner IV* flight. For Earth cruise 2 and the following two tests, the cavity amplifier and battery charger were switched off and the TWT was operating. Mars cruise was identical to Earth cruise 2, except for the lower solar intensity. In addition to a reduction in dissipation within the bus during Mars playback, zero power was dissipated in the magnetometer ion chamber, cosmic dust and trapped radiation detectors, cosmic ray telescope, and television assembly.

Equilibrium temperatures were measured at 75 locations within the model. Twenty of the temperature measurements were compared with measurements made at homologous locations on the TSM and *Mariner IV*. Comparisons were also made between the *Mariner IV* solar panel temperature and the average of all model solar panel temperature measurements. The remarkable accuracy with which the TSM was able to predict the 20 *Mariner IV* flight temperatures is shown in Table 2. Approximately 65% of the temperatures from the JPL test corresponded within 10°F, and 89% within 25°F, a slight improvement over the previously obtained results. Average TSM solar panel temperatures were within 8°F of the *Mariner IV* flight temperatures.

The maximum temperature difference between the TSM and *Mariner IV* occurred in the Canopus tracker, which was approximately 50°F lower in the TSM. The

Table 1. TSM test conditions

Mode	Flight day	Heater power, w	Solar intensity, w/ft^2
Earth cruise 1	2	28.3	134
Earth cruise 2	98	28.0	84.5
Mars cruise	180	28.0	58.5
Mars playback	234	25.4	54.8

Table 2. *Mariner IV* and TSM temperature comparison for 20 locations

Temperature, °F	Number of temperature readings within indicated error			
	Earth cruise 1	Earth cruise 2	Mars cruise	Mars playback
5	7	10 (14)[a]	8 (3)	9 (1)
10	11	14 (15)	13 (11)	14 (3)
15	16	18 (16)	15 (16)	15 (13)
25	18	19 (19)	18 (17)	17 (15)

[a]Values in parentheses are for earlier TSM tests in NASA Lewis simulator.

reason for this difference is being investigated. The trapped radiation detector temperature (which was approximately 32°F below *Mariner IV* in the previous TSM tests in the NASA simulator because of the solar beam shadowing) was within 15°F of the *Mariner IV* data in the JPL tests.

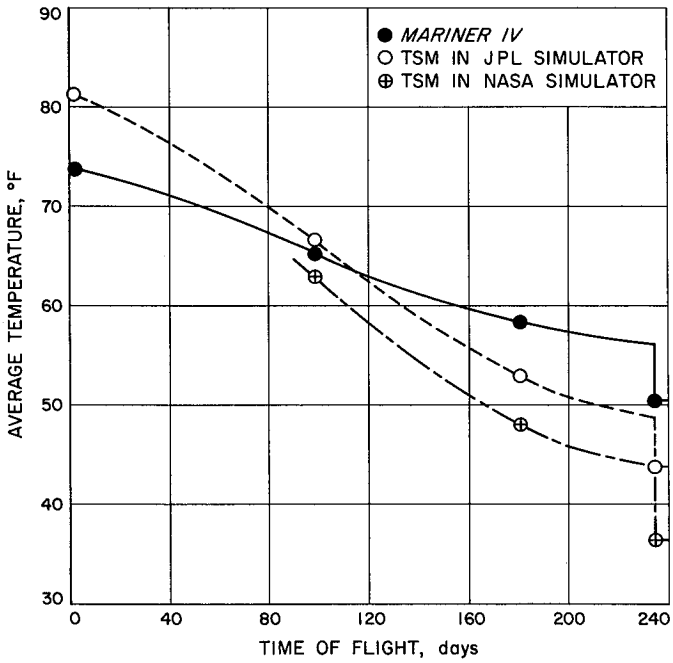


Fig. 6. Average bus temperature vs time of flight

Comparison of the TSM temperatures with flight data indicates that the TSM bus is more sensitive than the prototype to the variation in solar intensity. This phenomenon is illustrated in Fig. 6, which is a plot of the average of the bus temperatures of the TSM and *Mariner IV* bus as a function of time of flight. The condition is attributed to modeling errors in either the upper and lower thermal shields or the louvered areas. In any case, as indicated by Fig. 6, the error is not large.

It is also shown (Fig. 6) that the TSM bus temperature is approximately 5°F higher in the JPL chamber than in the NASA Lewis chamber. This difference is undoubtedly due to a higher simulated solar intensity. A 5% increase in solar intensity, which is probably within the accuracy of the radiometers, would be sufficient to cause this difference.

C. Comparison of Heat Transfer Modes for Planetary Entry

J. M. Spiegel

Some estimate of the relative importance of nonequilibrium and equilibrium radiation heat transfer, as well as radiation modes compared with convective heat transfer,

is required as a basis for planning future planetary entry research. A revised version of the trajectory and reference heating program described in Vol. I of Ref. 4 was used to make such an estimate for a spherically blunted 60-deg half-angle cone.

The radiation model used by AVCO to approximate the stagnation shock-layer region of a blunt body is shown in Fig. 7. The radiance profile[1] (solid line in Fig. 7a) is replaced by a triangularly peaked nonequilibrium portion and a constant equilibrium region (dashed lines). In the computer program, this geometric model also accounts for two truncated conditions (Figs. 7b and 7c) that can be encountered during entry, in addition to the basic profile of Fig. 7a.

[1]Such as can be obtained in shock tubes.

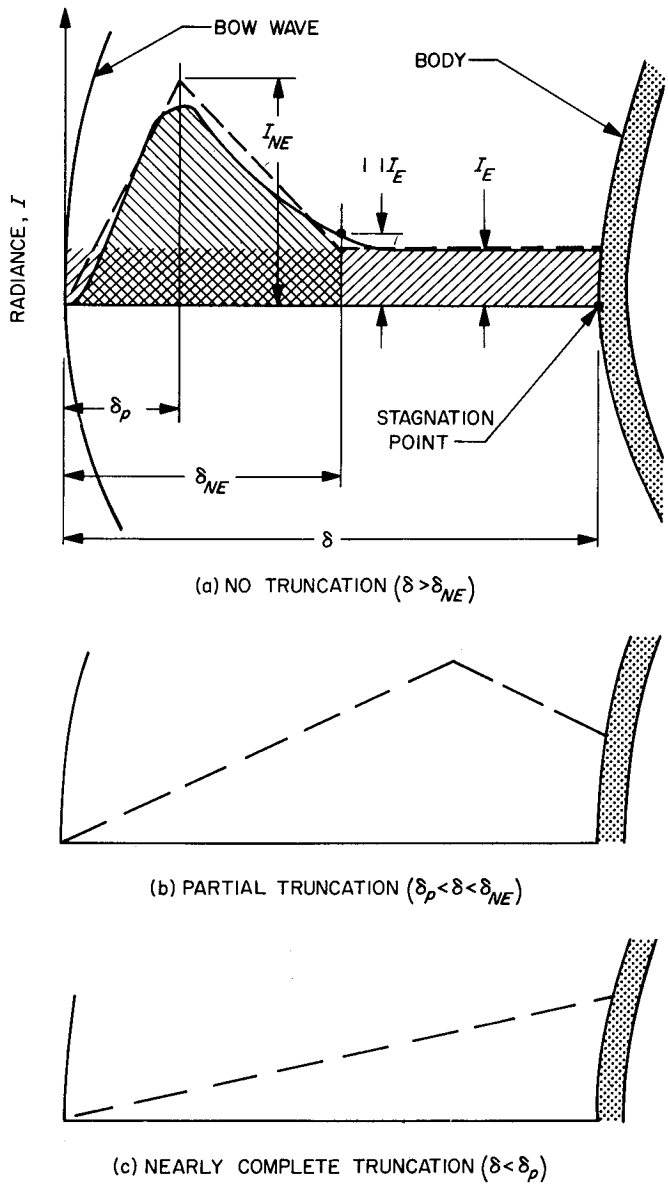


Fig. 7. AVCO stagnation point radiation model

To compare heat transfer modes, several cases of hyperbolic entry to Mars were examined for a 12-ft-diameter body of the type described above, having a nose radius one-tenth of the diameter. The results are shown in Table 3 for two possible model atmospheres of different reciprocal scale height β, both containing 20% CO_2 and 80% N_2 (by volume). Cases 3 through 10 are for a $\beta = 2.15 \times 10^{-5}$ ft^{-1}, and case 12 is for a $\beta = 4.30 \times 10^{-5}$ ft^{-1}. For each case, three trajectory points are given: the first for 0.5 of $q_{R_{max}}$ prior to peak heating, the second for $q_{R_{max}}$, and the third for 0.5 of $q_{R_{max}}$ after peak heating.

Of the seven cases shown, 3 and 9 are for the stagnation region and the remainder for the cone edge (see insert on Fig. 8). The latter is the more important region from a heat-shield weight standpoint, because of

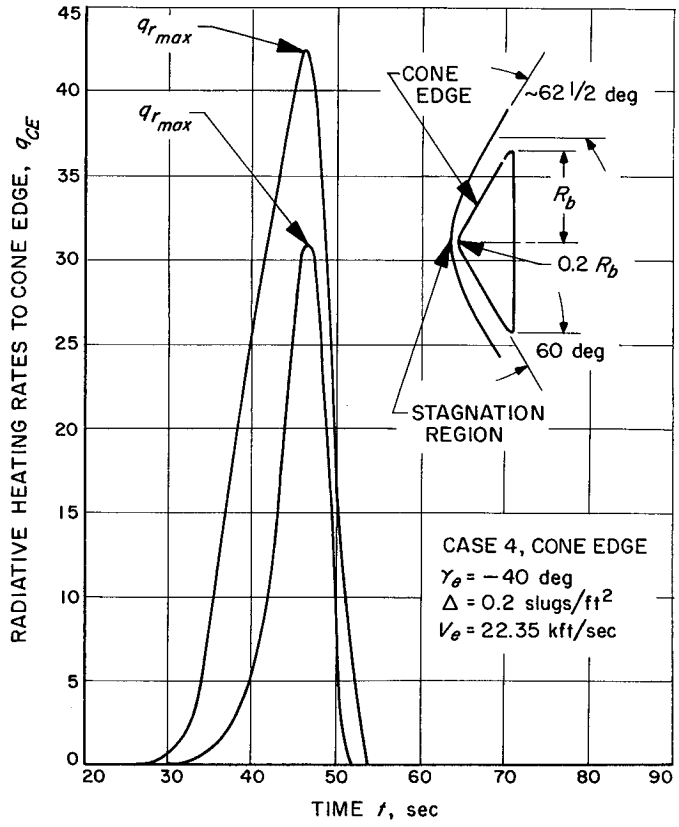


Fig. 8. Equilibrium and nonequilibrium radiative heating pulses

Table 3. Comparison of heat transfer modes for entry to Mars

Case	Body location	Velocity[a] kft/sec	$-\gamma_e$ deg	Δ slugs/ft^2	ρ_∞ slugs/ft^3	$q_{R_{max}}$ Btu/ft^2 sec	$q_{E_{max}}$ Btu/ft^2 sec	$q_{C_{max}}$[b] Btu/ft^2 sec	Q_R Btu/ft^2	Q_E Btu/ft^2	Q_C[b] Btu/ft^2	$\frac{\delta_{NE}}{\delta}$	$\frac{\delta_p}{\delta}$	$\frac{I_{NE}}{I_E}$	$\beta \times 10^{-5}$
3	Stag-nation	21.650 19.590 14.970	40	0.20	2.14×10^{-7} 7.02×10^{-7} 2.00×10^{-6}	16.2	5.8	123	154	42	2180	3.31 1.44 1.44	0.72 0.28 0.22	7.02 3.86 —	2.15
4	Cone edge	22.190 19.590 16.300	40	0.20	9.21×10^{-8} 7.02×10^{-7} 1.59×10^{-6}	42.4	30.4	30.8	469	220	544	1.44 0.27 0.24	0.32 0.05 0.04	9.93 3.92 7.13	
6	Cone edge	21.690 19.360 16.530	90	0.20	3.46×10^{-7} 1.26×10^{-6} 2.55×10^{-6}	75.0	64.7	40.2	445	277	426	0.39 0.16 0.14	0.08 0.03 0.02	5.57 3.12 5.14	
7	Cone edge	21.450 19.010 16.740	90	0.40	8.90×10^{-7} 2.85×10^{-6} 4.93×10^{-6}	185.0	177.1	57.7	943	753	594	0.15 0.08 0.07	0.03 0.01 0.01	3.54 2.28 3.78	
8	Cone edge	21.600 19.270 16.600	40	0.40	4.59×10^{-7} 1.57×10^{-6} 3.02×10^{-6}	94.3	84.6	44.2	892	605	774	0.29 0.13 0.11	0.06 0.02 0.02	4.75 2.85 4.84	
9	Stag-nation	21.990 19.870 16.080	40	0.40	2.83×10^{-7} 1.27×10^{-6} 3.33×10^{-6}	26.2	15.5	174	273	115	3094	2.34 0.73 0.66	0.52 0.15 0.11	5.93 2.98 6.38	
10	Cone edge	25.180 21.000 17.460	40	0.20	1.97×10^{-7} 1.07×10^{-6} 1.95×10^{-6}	103.6	90.9	47.2	1010	558	738	0.39 0.14 0.15	0.10 0.03 0.03	12.08 3.20 3.73	2.15
12	Cone edge	21.64 19.12 16.71	40	0.20	4.59×10^{-7} 1.66×10^{-6} 2.96×10^{-6}	95.0	85.5	44.4	446	301	396	0.30 0.13 0.13	0.06 0.02 0.02	3.32 7.95 7.91	4.30 4.30

[a] Three points for each case are $0.5q_{R_{max}}$, $q_{R_{max}}$, $0.5q_{R_{max}}$.

[b] For edge points, stagnation data reduced to $0.25q_{stag\ conv}$.

Definition of symbols

I = radiation per unit volume

$Q = \int q dt$ over entry trajectory

q = heat transfer rate

β = reciprocal scale height

γ_e = entry flight path angle

Δ = entry mass/1.4 × cross-sectional area of body

Definition of symbols

$\delta = \delta_{NE}, \delta_p$ defined in Fig. 7

ρ_∞ = free stream density

Subscripts

C = convective

E = equilibrium

NE = nonequilibrium

R = radiation including nonequilibrium

the much larger exposed area. The entry velocity was 22.35 kft sec for all cases except number 10 which was 26.0 kft sec. Note that q_R represents the total radiation under the dashed curves of Fig. 7, and q_E is related to the rectangular area defined by I_E and δ.

The δ_{NE}/δ column indicates that the nonequilibrium region is dominant for both stagnation point cases, but the cone edge is primarily in equilibrium. This is also shown by a comparison of the q_R and q_E columns. However, the integration of these heating rates (Q_R and Q_E) is most representative of heat-shield weight requirements, and the results shown in Table 3 indicate the possibly important role of nonequilibrium radiation for hyperbolic entry for the conditions considered. A time history for case 4 is shown in Fig. 8 to illustrate the heat-load increase caused by nonequilibrium radiation. These strong nonequilibrium contributions were indicated in Ref. 4 and were also predicted independently.[2]

A comparison of Q_R and Q_C columns for the cone edge shows that the radiation is about the same as for the convective mode for $\gamma_e = -40$ deg and a Δ of 0.20, and becomes dominant for higher values of either of these parameters. As expected, increasing the entry velocity (case 10 compared with case 4) accentuates the importance of radiation heat transfer over the convective mode. A comparison of cases 4 and 12 reveals that the relative importance of the radiative to convective modes is increased further when the reciprocal scale height is doubled (steeper atmospheric density gradient).

Self-absorption is essentially not included in the present estimates, but a partially compensating omission is that of the CO (4+) radiator.

Although none of the combined heat loads ($Q_R + Q_C$) shown in Table 3 are large compared with many Earth reentry missions, heat-shield requirements can still represent 10 to 20% of the entry weight for a Mars hyperbolic entry (largely insulation), and possibly more for Venus entry. Consequently, uncertainties in predicting radiation heat transfer must be kept to known and acceptable limits for entry modes and conditions similar to or more severe than that presented here.

A more detailed discussion of this heat transfer comparison is presented in Ref. 5.

[2]Tauber, Personal Communication, NASA Ames Research Center, February 8, 1966.

D. Equilibrium Radiance of Model Planetary Atmospheres

F. Wolf

In the continuing effort to obtain experimental information on equilibrium radiation of shock layers for entry into planetary atmospheres, additional data obtained from other sources[3,4] on a mixture containing 30% by volume

[3]Thomas, G. M., Personal Communication, JPL.

[4]Gruszczynski, J. S., *Hypervelocity Heat Transfer Studies in Simulated Planetary Atmospheres,* JPL Contract No. 950297 to the General Electric Company, Space Sciences Laboratory.

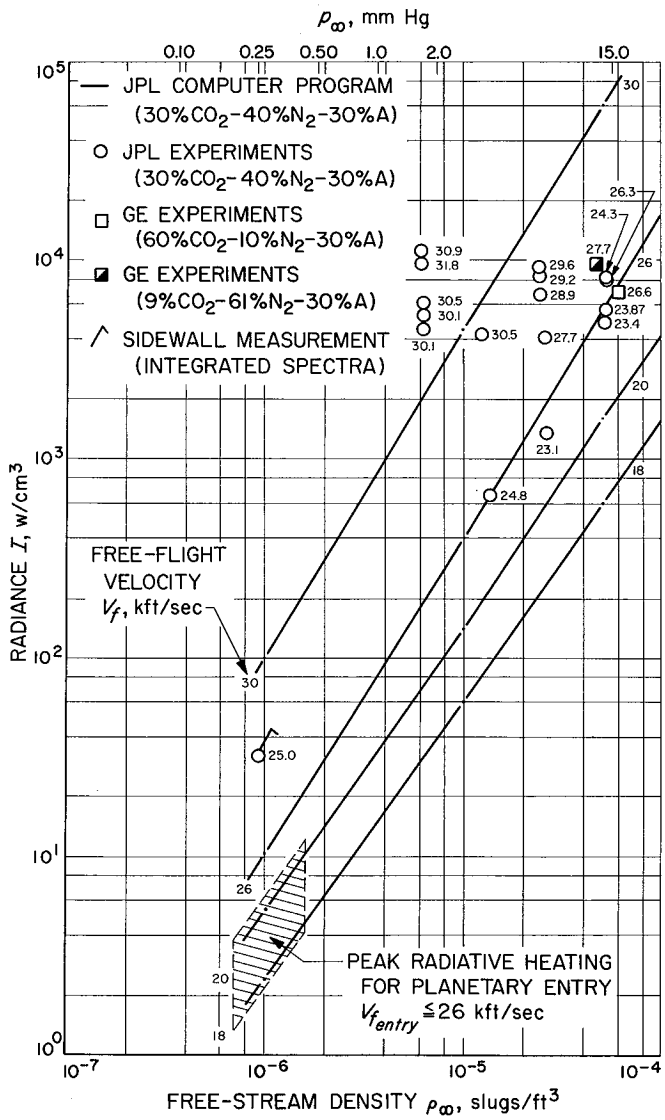


Fig. 9. Radiance of CO_2-N_2-A mixtures versus free-stream density

of argon are shown in Fig. 9. As mentioned in *SPS 37-36*, Vol. IV (p. 80), the scarcity of data at low densities applies also to this mixture. For many entry trajectories of present interest, the peak radiative heating would occur at flight conditions represented by the cross-hatched area at the lower left of the figure. The only experiment performed in this region is a spectrally integrated sidewall measurement in the JPL shock tube (flagged point).

The correlation of the experimental data by means of a fixed density dependence (see Fig. 6, *SPS 37-33*, Vol. IV, p. 102) has been reviewed and an exponent $n = 1.4$ selected to plot a normalized total equilibrium radiance as shown in Figs. 10a and 10b. The computed radiance obtained from both Ames (Ref. 6) and JPL programs is also shown. The large differences in calculated intensity values above 25 kft/sec are due to the inclusion of continuum radiation in the JPL computer program.

The scatter of experimental data in this "density-collapsed" presentation does not exceed the bandwidth of unnormalized data at any one density level (as shown in Figs. 3 and 4, *SPS 37-36*, Vol. IV, p. 83). Therefore, it seems justifiable to derive an analytic expression for the radiance from this form of the data plot; this would be useful for practical cases of entry heat transfer for Martian hyperbolic entry or entry to Venus from orbit for the mixtures cited. Two representative curves were drawn: one, an approximate mean fitted visually through the highest concentration of data points, and the other, near the upper limit of the band of scatter, to give an expression for a practical maximum of the observed values which

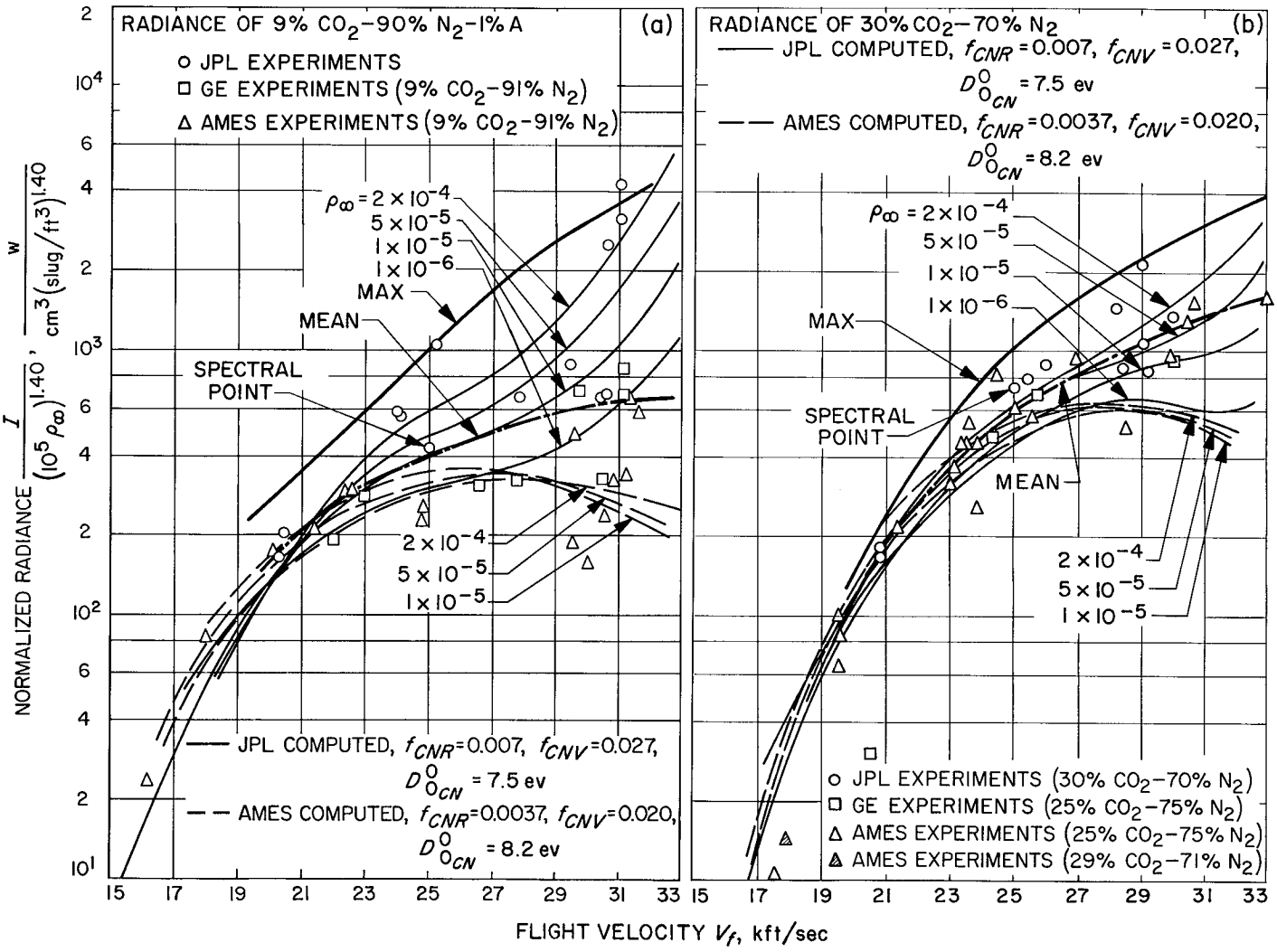


Fig. 10. (a) Radiance of 9% CO_2, 90% N_2, 1% A; (b) Radiance of 30% CO_2, 70% N_2

could serve as a conservative engineering estimate of radiance. This will include the energy transferred above $\lambda = 0.2\,\mu$, and is based on the radiance of an optically thin gas. To approximate the total radiative transfer, an estimate of the vacuum ultraviolet part of the spectrum should be added, especially in types of model planetary atmospheres which might contain considerable amounts of CO. Also, some estimate of self-absorption effects might improve the accuracy of the total radiative heat transfer predictions.

Within a limited range of flight velocities, these curves can be fitted reasonably well by the quadratic form

$$\frac{I}{(10^5\,\rho_\infty)^{1.40}} = A\left(\frac{V_f}{10^4}\right)^2 + B\left(\frac{V_f}{10^4}\right) + C$$

The coefficients A, B, and C are given in the following table:

Mixture	9% CO_2–91% N_2		30% CO_2–70% N_2	
Curve	Mean	Max	Mean	Max
A	−80	+2300	+240	+1700
B	+840	−8860	−120	−6040
C	−1200	+8800	−600	+5440

These curve fits are shown on Figs. 10a and 10b. A simple power of velocity, as often used for approximate Earth reentry radiation calculations, appears too crude for the entry velocity range shown since such a simple function would appear almost as a straight line on Fig. 10.

References

1. Clausing, A. M., *Some Influences of Macroscopic Constrictions on the Thermal Contact Resistance*, ME-TN-242-2, University of Illinois, Engineering Experiment Station, Urbana, Ill., April 1965.
2. Clausing, A. M., *Theoretical and Experimental Study of Thermal Contact Resistance in a Vacuum Environment*, Semiannual Status Report 8, University of Illinois, Dept. of Mechanical Engineering, Urbana, Ill., January 1966.
3. Gabron, F., Johnson, R. W., Vickers, J. M. F., Lucas, J. W., *Thermal Scale Modeling of the Mariner IV Spacecraft*, Paper 66-23, presented at AIAA Third Aerospace Sciences Meeting, New York, January 24–26, 1966.
4. *Mars-Venus Capsule Parameter Study*, Vols. I, II, and III, RAD TR 64-1, AVCO Corporation, Research and Advanced Development Division, Wilmington, Mass., January–March 1964.
5. Wolf, F., Spiegel, J. M., "*Status of Basic Shock Layer Radiation Information for Inner Planet Atmospheric Entry*," Paper 66-421, presented at the AIAA Fourth Aerospace Sciences Meeting, Los Angeles, Calif., June 27–29, 1966.
6. Arnold, J. O., Reis, V. H., and Woodward, H. T., *Theoretical and Experimental Studies of Equilibrium and Nonequilibrium Radiation to Bodies Entering Postulated Martian and Venusian Atmospheres at High Speeds*, Paper 65-116, presented at the AIAA Second Aerospace Sciences Meeting, New York, January 25–27, 1965.

VI. Materials

A. Pure Oxide Ceramic Research

M. H. Leipold

In the pure oxide program, a low-temperature (800–1000°C), high-pressure (greater than 10,000 psi), hot-pressing technique has been selected for the fabrication of mechanical-testing specimen blanks. By pressing in ceramic or refractory metal dies under these conditions of lower reactivity, dense specimens without significant contamination have been produced. However, considerable difficulty from die failure and die sticking is continually met in the use of this process. In addition, to fabricate the required length of the specimen, a complex welding process was developed.

A new method—hot isostatic pressing—which offers a possible solution to these problems has recently been made available (through Battelle Memorial Institute, Columbus, Ohio) on a best-efforts service basis. Powders typical of those used at JPL for more conventional fabrication were provided for processing. Particulars of the powder supply are given in Table 1. First, the sample powders were hydrostatically pressed, at 90,000 psi at ambient temperatures, into over-sized slugs of densities shown in Table 1. Next, these slugs were loaded into metal cans and baked out under a cold-trapped mechanical vacuum for 8 hr at 250–300°C. The cans were then sealed off and hot isostatically pressed at 800°C and 15,000 psi for approximately ½ to 1 hr. It was anticipated, from previous experience in uniaxial hot pressing of MgO, that these conditions would provide close to theoretical density. The compacted material was removed from the metal can by means of chemical etching.

In all cases, the hot isostatic pressings obtained showed an array of fine cracks, making them unsuitable for mechanical testing. In many cases, the compacts broke into several pieces during initial handling. However, considerable pertinent information concerning the material was gained, and is summarized in Table 2. The structure was not revealed and no porosity was visible through the optical microscope, and the grain size could not be determined. Consequently, typical electron micrographs of the polished surface were obtained from Sloan Research Industries, and are shown in Fig. 1.

Table 1. MgO powder for hot isostatic pressing

Sample number	Type	Particle size A	Density after cold compaction, g/cc	Remarks
1	JPL Special	126	2.00	
2	JPL Special	146	2.14	
3	JPL Special	155	2.10	
4	JPL Special	248	1.97	
5	Fisher M-300	225	2.04	
6	Fisher M-300	225	1.98	
7	Baker "Heavy"	NA[a]	NA	Supplied by vendor
8	Baker "Heavy"	NA	2.47	Supplied by vendor
9	Fisher M-300	225	2.03	
10	Fisher M-300	225	2.01	
11	Baker "Light"	NA	2.00	Supplied by vendor
12	JPL Special	166	NA	Leaked during pressing—subsequently discarded
13	JPL Special	177	NA	Leaked during pressing—subsequently discarded

[a]NA = not available

Table 2. MgO compacts after hot isostatic pressing

Sample number	Density, g/cc	Grain size A	Unit cell A	Remarks
1	2.99	100	ND[a]	White color, fine longitudinal cracks, severely etched, thin reaction layer with container, glassy appearance, translucent.
2	3.11	150	ND	White color, severe longitudinal cracks, severely etched, thin reaction layer with container, glassy appearance, specimen intact, translucent.
3	3.15	ND	4.2133	White color, fine longitudinal cracks, severely etched, thin reaction layer with container, glassy appearance, translucent.
4	3.33	300	ND	White color with light-blue tint, severely etched, thin reaction layer with container, glassy appearance, severe longitudinal cracking, translucent.
5	2.1	500	ND	White color, severely cracked, porous.
6	NA[b]	NA	NA	White color, severely cracked, porous.
7	NA	NA	NA	White color, fine hairline cracks, glassy appearance, appears to be translucent.
8	3.30	ND	ND	Cream color, bluish discoloration, slightly etched, glassy appearance, no apparent cracking.
9	NA	NA	NA	Container leaked specimen not densified.
10	NA	NA	NA	Container leaked specimen not densified.
11	3.27	ND	ND	Light cream color, several longitudinal cracks intact, glassy appearance, slightly etched.

[a]ND = not determined
[b]NA = not available

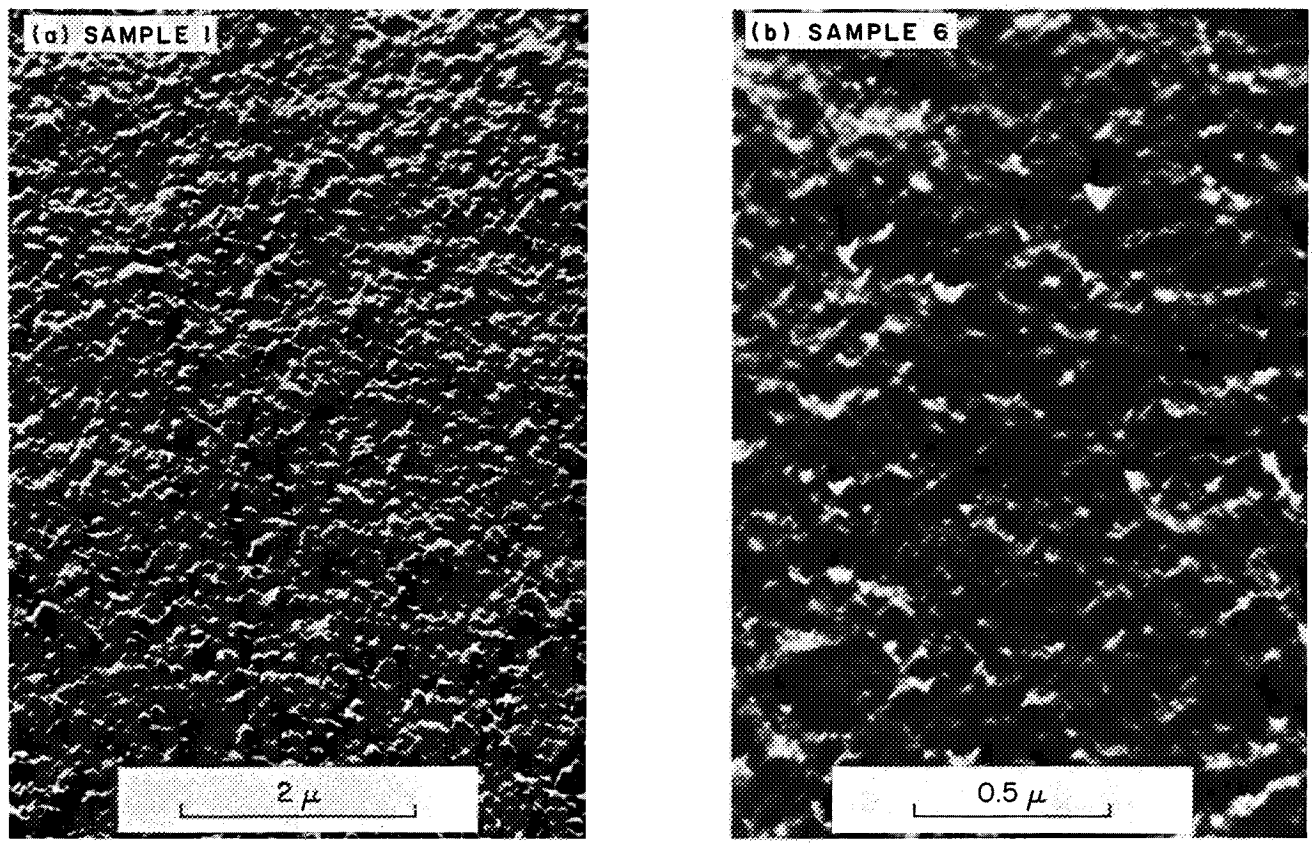


Fig. 1. Electron micrographs of hot isostatically pressed MgO (specimens steam etched and dual replicated)

Some difficulty has been encountered in correlating the observed density with the structure shown in the micrographs. (The nature of the structure is not apparent in Fig. 1. If it were porosity, the quantity would be excessive.) It might be expected that the fabricated MgO would contain considerable hydroxyl not removed in the baking process; however, analytical techniques capable of this determination are unavailable at present. To determine the effect of hydroxyl on the density of MgO, Table 3 was constructed. Here it is assumed that a magnesium defect is introduced with each pair of hydroxyls to maintain charge neutrality and that the unit cell size remains essentially constant. This latter assumption is consistent with measured unit cell size (see Table 2) in the material and with the similarity in size of hydroxyl and oxygen. Note that the observed density decrement would require excessive hydroxyl for the decrease to be ascribed entirely to the presence of OH. It must therefore be concluded that a significant portion of the reduced density is the result of porosity, although this is not apparent in the micrographs of the as-pressed material.

Table 3. Theoretical effect of OH^- impurity on density of MgO

OH^- concentration, ppm atomic	Density, g/cc
0	3.58094
10	3.58093
100	3.58084
1000	3.57994
10,000	3.57102
50,000	3.53138

To ascertain the effect of reheat treatment on the isostatically pressed MgO, some samples were reheated to 1010°C and others to 1750°C. The results (summarized in Table 4) show a weight loss with the lower temperature reheat but no additional weight loss with higher temperature reheat, suggesting that the hydroxyl loss has reached some equilibrium value and was not reduced by further heating. Previous chemical analyses indicate that this equilibrium value is not zero. Optical microscopy again failed to reveal the structure of the 1010°C reheat specimens, and electron microscopy was not attempted. Optical micrographs of the material after the 1750°C reheat are

Table 4. Effect of heat treatment on hot isostatically pressed MgO in air

Specimen	Treatment (1 hr), °C	Density, g/cc	Grain size, μ	Wt loss, % (from as-pressed condition)
2	None	3.11	0.10	—
3	None	3.15	0.15	—
4	None	3.33	0.30	—
2	1010	ND[a]	< 1[b]	3.2
3	1010	3.32	< 1[b]	3.0
4	1010	ND	< 1[b]	1.8
2	1750	3.32	40/500[c]	3.1
3	1750	3.38	50/500[c]	3.1
4	1750	3.42	40/1000[c]	1.7

[a]ND = not determined
[b] = not resolved optically
[c] = duplex structure

shown in Fig. 2. Here the agglomeration of porosity is evident. The ring of larger grains around the edges indicates a region in which grain growth has not been inhibited by the presence of structural anomalies. However, the appearance of such a structure will not serve to differentiate porosity from hydroxyl, since either normal porosity or a hydroxyl magnesium defect structure where the vacancies coalesced into porosity could create this condition.

It is evident that much more work, especially in terms of analysis, will be necessary to fully describe this material. As the technique would require considerable development to serve as a means of specimen fabrication, it will not be pursued at this time.

B. Metallurgical Examination of Thermionic Converters

E. C. Bernett

To develop better electrical power sources for spacecraft applications, the Spacecraft Power Section at JPL is

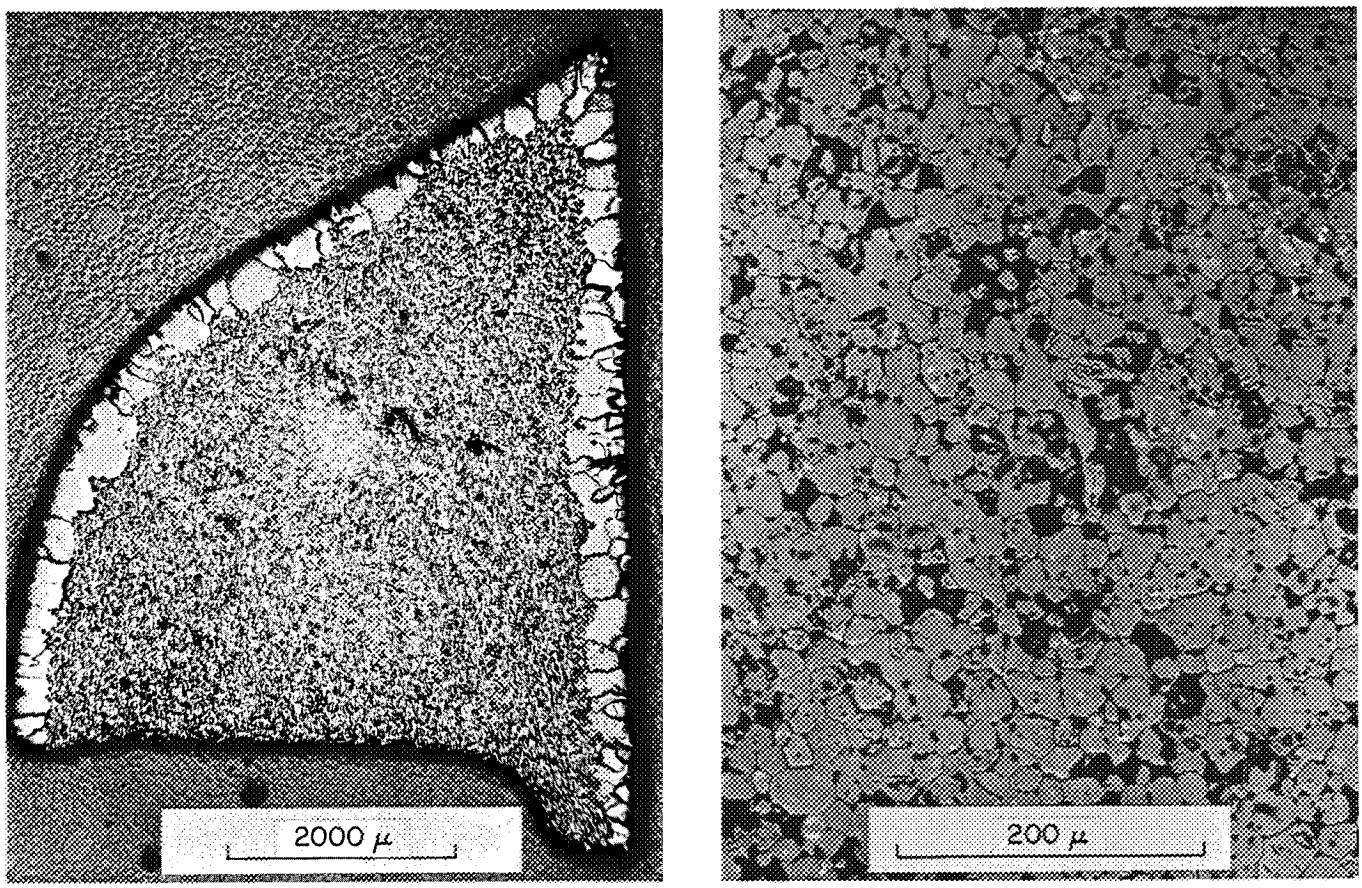


Fig. 2. Hot isostatically pressed MgO (sample 2) after reheating to 1750°C for 1 hr in air

investigating several new approaches based on the direct conversion of thermal to electrical energy. One of these, the thermionic converter, has been under development since 1961 (Ref. 1 and *SPS 37-31*, Vol. IV, p. 53). During that time, a number of contracts were let to industry as part of a program to develop a practical thermionic converter having high efficiency and a reasonably long life.

In support of this work, the Materials Section performs metallurgical evaluations by sectioning and examining those converters which have completed test or have failed for unknown reasons. These evaluations are undertaken to learn more about the behavior of these devices and to establish possible failure mechanisms which could lead to design improvements that increase efficiency, life, and reliability. A considerable amount of work has already been accomplished toward this end (*SPS 37-22*, Vol. IV, p. 58; *SPS 37-30*, Vol. IV, p. 67; and Ref. 2).

The following discussion deals with the evaluation of three Series VIII converters developed by the Thermo Electron Engineering Corporation. All three failed in test after operating for relatively short times. A summary of the test history is given in Table 5.

In outward appearance, these units differ only slightly from former Thermo Electron converters. The principal change in the Series VIII units is the design of the emitter. Instead of the one piece solid tantalum used in most previous designs, the Series VIII emitters are a composite

Table 5. Thermo Electron Series VIII thermionic converter test history

Serial No.	Total run time, hr	Observations
VIII-P2a	1500	Shorted at 300 hr; reactivated by thermal shock; shorted again at 1500 hr; cesium leak during operation.
VIII-P3	2000	Shorted at 1400 hr, at 1700 hr, and again at 2000 hr; leak tight after run.
VIII-16	1300	Good power out at first; began to degrade at 375 hr; work function remained constant; no cesium leaks apparent.

Additional performance data on Series VIII converters is given in *SPS 37-31*, Vol. IV, p. 53.

of rhenium (0.020 in. thick) bonded to tantalum, as shown in Fig. 3. A composite was used to conserve the amount of rhenium required. In the three units under examination, the rhenium facings were isostatically gas pressure bonded (by Battelle Memorial Institute) to the tantalum slug. An initial attempt to vapor-deposit the Re on the Ta was abandoned because of the unsatisfactory yield and adhesion of the Re layer; therefore, the diffusion bonding technique was adopted for production. The bonding was carried out in argon at 1565°C for 3 hr at 10,000 psi. A photomicrograph of the Ta-to-Re interface in the "as-bonded condition" is shown in Fig. 4, and a section through the interface of the emitter from converter VIII-P2a after 1500 hr of continuous operation is shown in Fig. 5. The operating temperature of this region of the

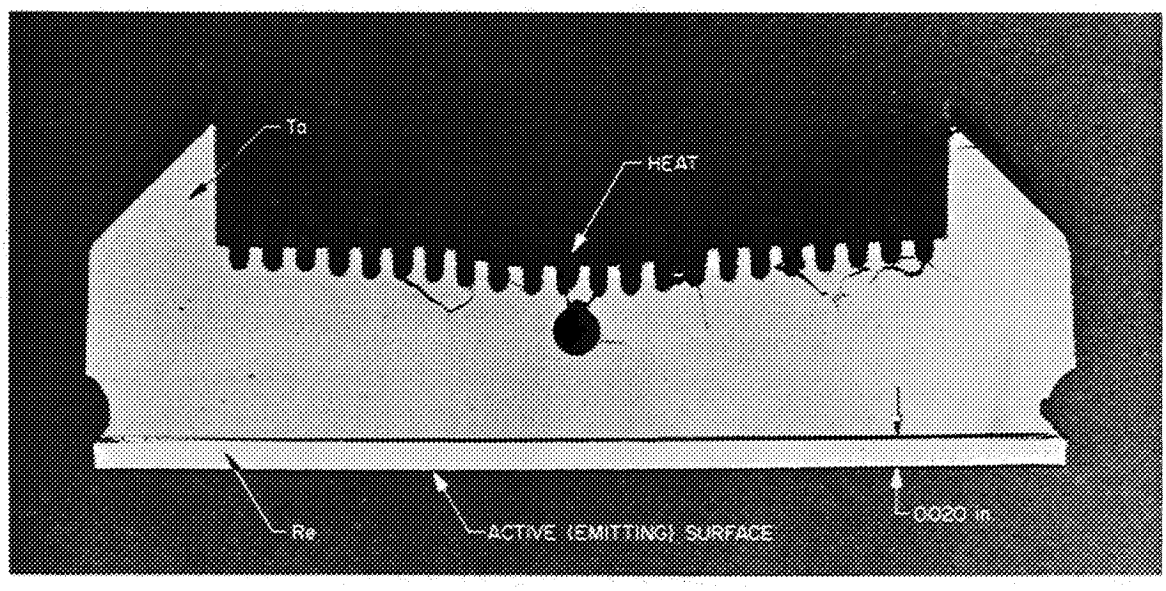


Fig. 3. Cross section of composite tantalum rhenium emitter used in Thermo Electron VIII-P3 thermionic converter

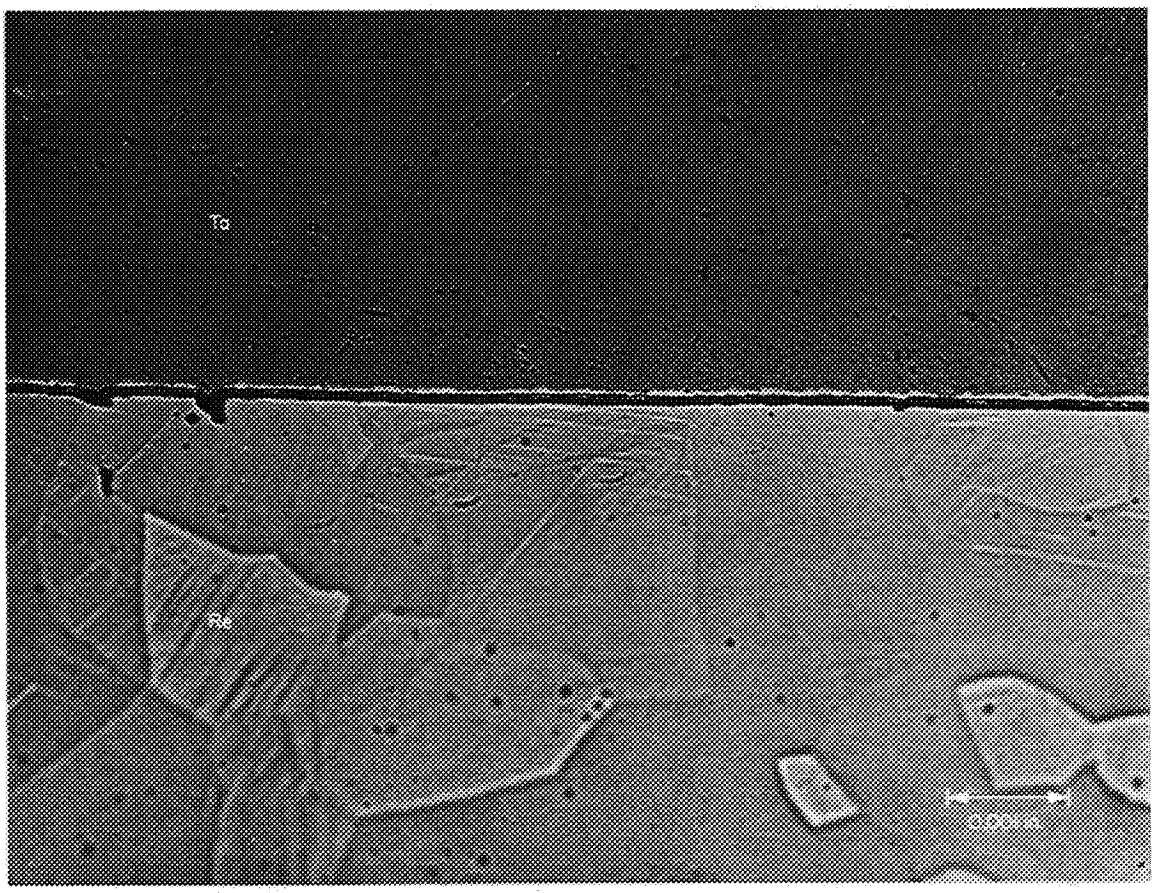


Fig. 4. Interface of Ta diffusion bonded to Re (as-bonded condition)

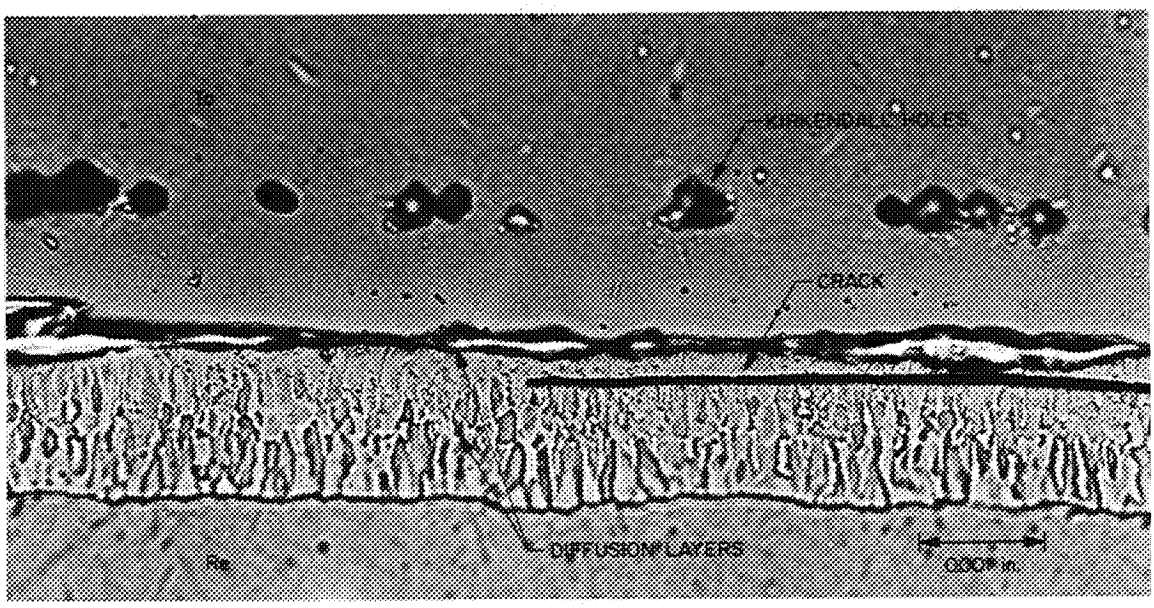


Fig. 5. Ta-to-Re interface on emitter from converter VIII-P2a after operation for 15 hr

converter is about 1650°C, and it is obvious that considerable diffusion has occurred during the run since the width of the diffused zone has increased from about 2 to 40 μ. The diffusion zone on the emitter from converter VIII-P2a (Fig. 5) is composed of two distinct layers: a thin layer on the Ta side, and a much thicker layer on the Re side. Microhardness checks on the thick layer gave readings above 2000 Knoop (15-g load), indicating that this zone is primarily an intermetallic compound(s) of rhenium and tantalum; therefore, it is not surprising that cracks were present in this layer (Fig. 5).

During converter operation, the diffusion of Ta atoms into the Re occurred at a higher rate than the Re into the Ta, causing a layer of Kirkendall holes to form in the tantalum just above the interface (Fig. 5). The vacancy count on the Ta side increases beyond the equilibrium level, and the coalescence of vacancies ultimately results in formation of holes (discussed in detail in *SPS 37-22*, Vol. IV, p. 58). From the standpoint of converter operation, the growth of Kirkendall holes is undesirable as heat transfer rates are adversely affected.

Another important aspect of the differential diffusion is the uneven mass transfer and the resultant volume shrinkage on the Ta side of the interface and expansion on the Re side, which cause these composite-type emitters to warp in service. Flatness checks were made on the active faces of the emitters under examination; in all cases, the central region was found to bulge in relation to the periphery. In converters VIII-P2a, -P3 and -16, the heights of the bulge were 0.0020, 0.0023, and 0.0015 in., respectively. These warpage values are directly proportional to the total run time of the converters; as the warpage of the emitter increases, the interelectrode spacing becomes more and more nonuniform during operation, markedly affecting converter performance.

On the basis of the problems described, it appears that unless some steps can be taken to eliminate (or drastically reduce) the diffusion rates across the interface, converters with composite-type emitters are unlikely to operate reliably for long periods of time. Because of the high operating temperatures involved, a solution to this problem will be difficult.

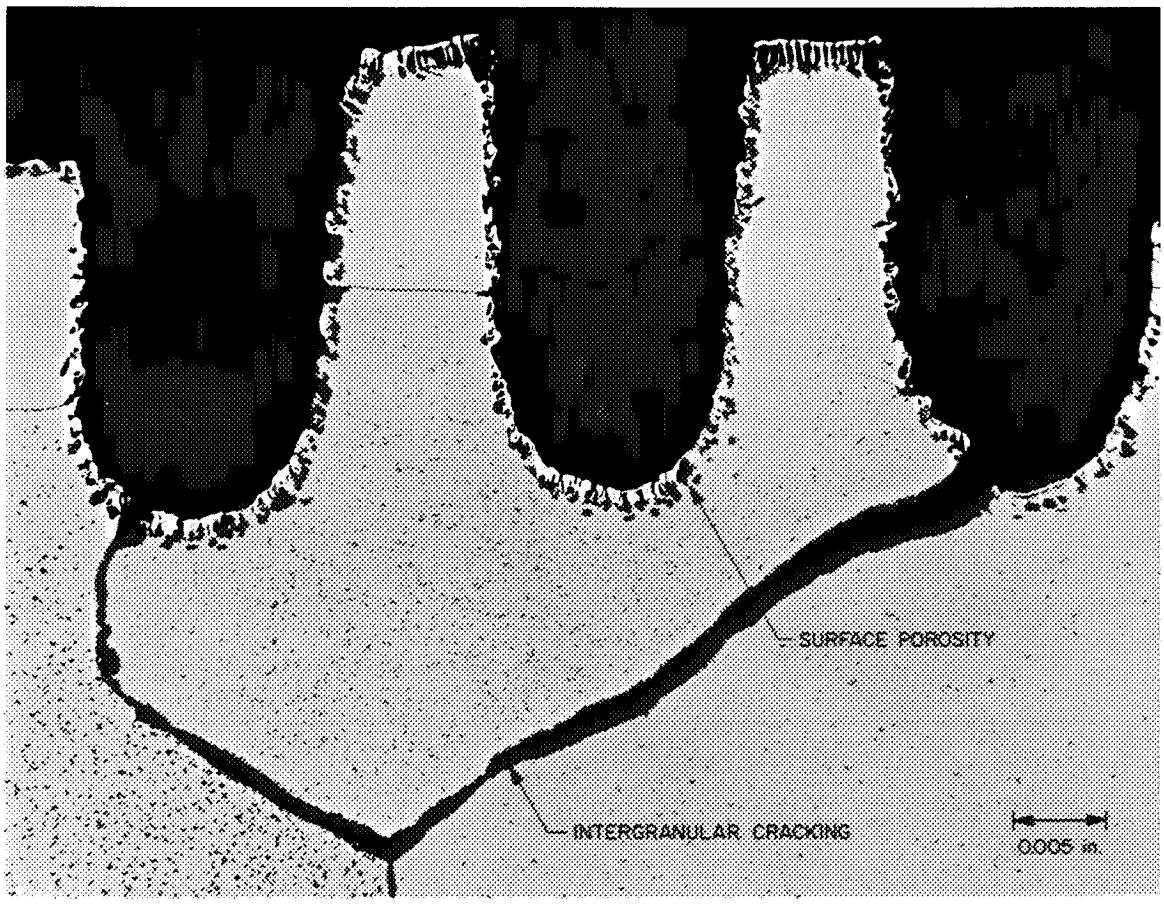


Fig. 6. Porosity and intergranular cracking on heated surface of the emitter from converter VIII-P3

It was noted that severe porosity had developed on (and just below) the surface of the machined serrations on the heated surfaces of the emitters (Fig. 6); the intergranular cracking which normally occurs in this region is also evident. Surface porosity was heaviest in the central region of the emitter. On the peripheral serrations the porosity was clearly concentrated on surfaces that were in direct line of sight of the electron bombardment heat source. The exact nature of the porosity has not yet been established; it may be an evaporation effect or possibly some type of radiation damage. In any event, both surface porosity and intergranular cracking would reduce the heat transfer rates of the emitter and therefore degrade the performance of the converter.

During the internal examination of the converters, particular attention was paid to the condition of the emitter and collector active surfaces. All emitter surfaces were clean, as was the collector surface on unit VIII-16. The collector surfaces on VIII-P2a and VIII-P3 were covered with a matte grey foreign deposit about 0.001 in. thick. X-ray diffraction checks of the deposit (in place) on the VIII-P2a collector indicated the material to be rhenium. A semiquantitative spectrochemical analysis of a micro sample scraped from this surface showed 97% Re, 2.8% Cs, and trace amounts of Au, B, Si, Mg, Fe, Mo, Cu, and Ca. It is surprising that such a heavy layer of rhenium should have been deposited on the collectors in the P2a and P3 converters, while none deposited in the P16 converter.

In previous converter examinations, some problems regarding various braze joints had been noted; therefore, some spot checks were made. The metal-ceramic braze joints were found to be strong and uniform. Checks of the molybdenum collector-copper radiator joints indicated variations in quality. Only about 80% of the bond unit VIII-P2a was sound; the remainder was not brazed at all. On VIII-P3, the copper near the braze interface was embrittled and showed gross intergranular cracking; the braze filler, however, had wet 100% of the mating copper and molybdenum surfaces. The collector-to-radiator braze appeared satisfactory on converter VIII-16. From the observations described, it is evident that additional efforts will be required to achieve uniformly high quality brazes in these devices.

References

1. Rouklove, P., "Thermionic Converters and Generators for Space Applications," *Transactions of the International Conference on Thermionic Electrical Power Generation, London, England, September 20–25, 1965*, European Nuclear Energy Agency and Institution of Electrical Engineers, 1965.
2. Bernett, E. C., *Metallurgical Examination of Development-Type Thermionic Power Converters*, Technical Report 32-548, Jet Propulsion Laboratory, Pasadena, Calif., December 15, 1963.
3. Reed-Hill, R. E., *Physical Metallurgy Principles*, D. Van Nostrand Co., Inc., New York, 1964, p. 263.

ENVIRONMENTAL SIMULATION DIVISION

VII. Instrumentation

A. Langmuir Probe Instrumentation for Electron Bombardment Ion Engines

R. Adams

The use of the Langmuir probe (Ref. 1) for studying the energy distribution of electrons in a plasma has been well documented. The Langmuir probe is of particular value for analyzing the operating parameters of the mercury and cesium plasmas in the Kaufman-type (Ref. 2) electron-bombardment ion engine in a static (no beam) condition and while a beam is being extracted. Measurements of this type are needed to evaluate the effects of various arc chamber and magnet configurations on ion density distribution in the plasma and in the beam. A thorough knowledge of these effects is necessary for electron bombardment thrustor design if acceptable operating efficiency and lifetime are to be realized.

This summary deals with the instrumentation of the Langmuir probe as applied to the operating electron-bombardment thrustor. The word "operating" is used here to indicate that accelerating voltages are applied and a beam is being extracted.

Langmuir and Mott-Smith (Ref. 3) developed the theory of probes and showed how the physical quantities of interest in a plasma could be measured.

The basic probe is an insulated conductor inserted into the plasma of a discharge. Positive and negative voltages (with respect to the cathode or anode of the discharge chamber) are applied to the probe and the resulting currents measured.

In its practical application to electron bombardment ion engine measurements, the probe is fabricated with a pure tungsten wire encased in a fused silica sheath. A typical probe is shown in Fig. 1. The tungsten wire is 0.020 in. in diameter and protrudes 0.15 in. beyond the end of the sheath. The whole assembly is tightly packed with boron nitride to prevent a discharge from forming within the sheath. The probe tip is offset from the axis by an amount determined by engine dimensions. The offset allows probing various positions in the plasma. Fig. 2

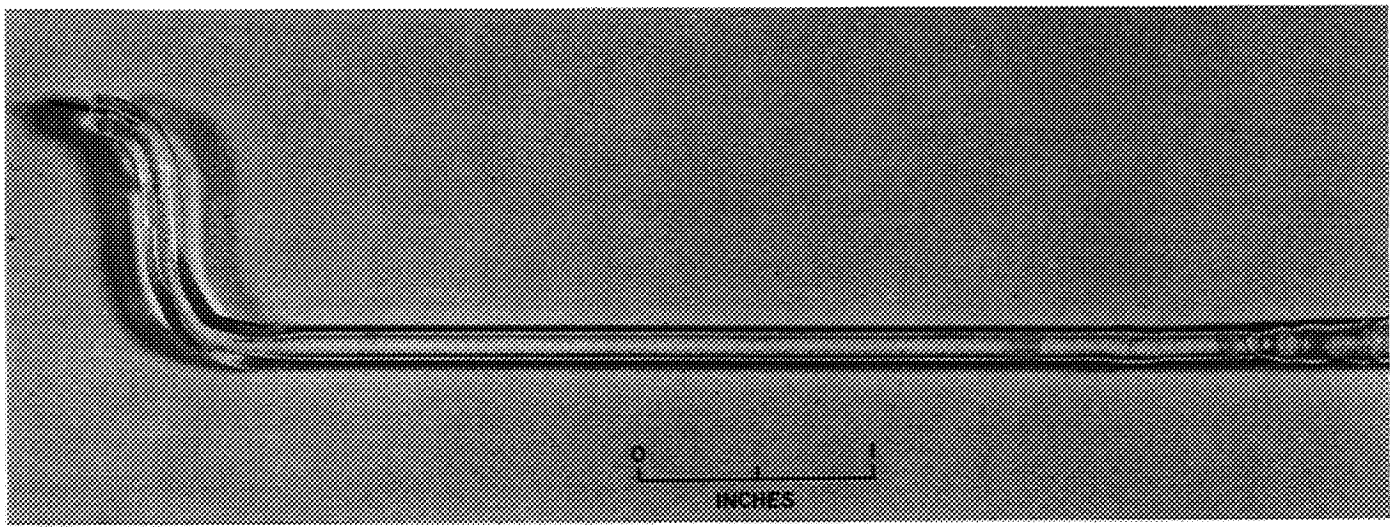


Fig. 1. Typical Langmuir probe

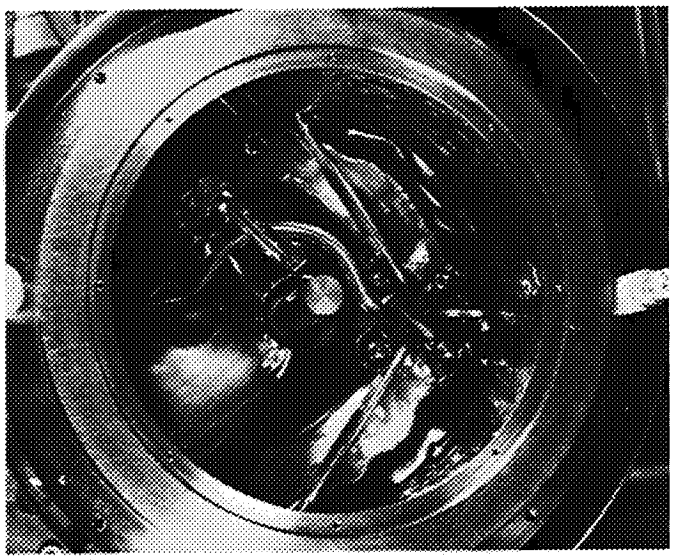


Fig. 2. Langmuir probe mounted in an experimental engine

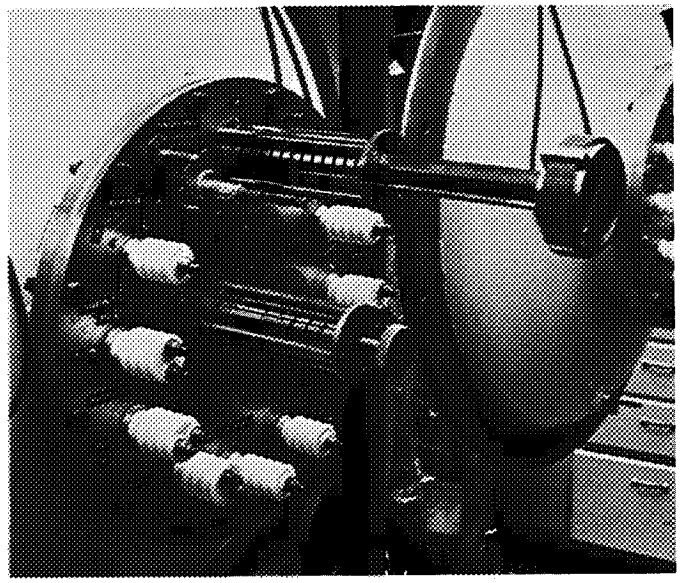


Fig. 3. Probe indexing head

shows a probe mounted in an experimental engine. In use, the probe is inserted into the plasma through the rear of the engine housing. A tightly fitting asbestos seal is used at this point to prevent propellant leakage while still allowing the probe to be moved axially or rotated for positioning.

Accurate positioning is accomplished from outside the vacuum chamber by means of an indexing head, as shown in Fig. 3 (center of header plate). The indexer shown can accommodate five axial and seven radial positions. The upper positioner shown in the photograph is for a Faraday probe. Connection to the probe is made through the high voltage bushing shown at the bottom of the header plate.

The Langmuir probe trace is made by varying the potential of the probe electrode from a negative value through zero to some positive potential where saturation occurs. The probe voltage versus current when plotted to a linear scale gives the typical Langmuir probe trace shown in Fig. 4.

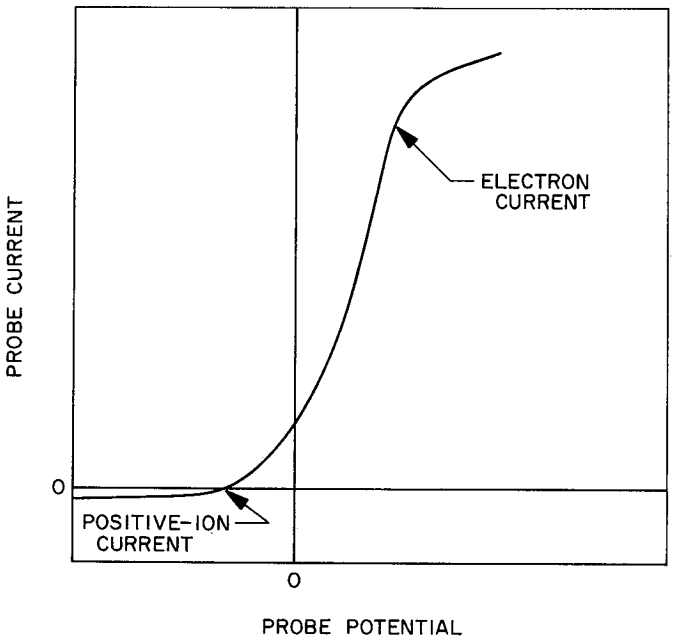


Fig. 4. Volt-ampere characteristic of a probe in the plasma of a discharge (not drawn to scale)

Instrumentation for probing the plasma of an operating electron bombardment ion engine must fulfill the following requirements:

(1) It must allow traces to be made rapidly since engine instability may be such as not to allow a data point to be maintained for more than a short period of time.

(2) It must not disturb the plasma of the engine while operating at potentials which may be as high as 5 kv relative to ground.

(3) It must provide complete safety for operating personnel.

The *X-Y* or *X-Y-Y* plotter has proved to be most suitable for recording Langmuir traces. Generally, multiple traces must be recorded simultaneously at various range settings; these are handled best, of course, on an *X-Y-Y* plotter. Fig. 5 shows a basic circuit diagram of an engine with the probe in place. Connections to the plotter are such as to make the *X*-axis proportional to the probe voltage and the *Y*-axis proportional to the probe current. The probe is

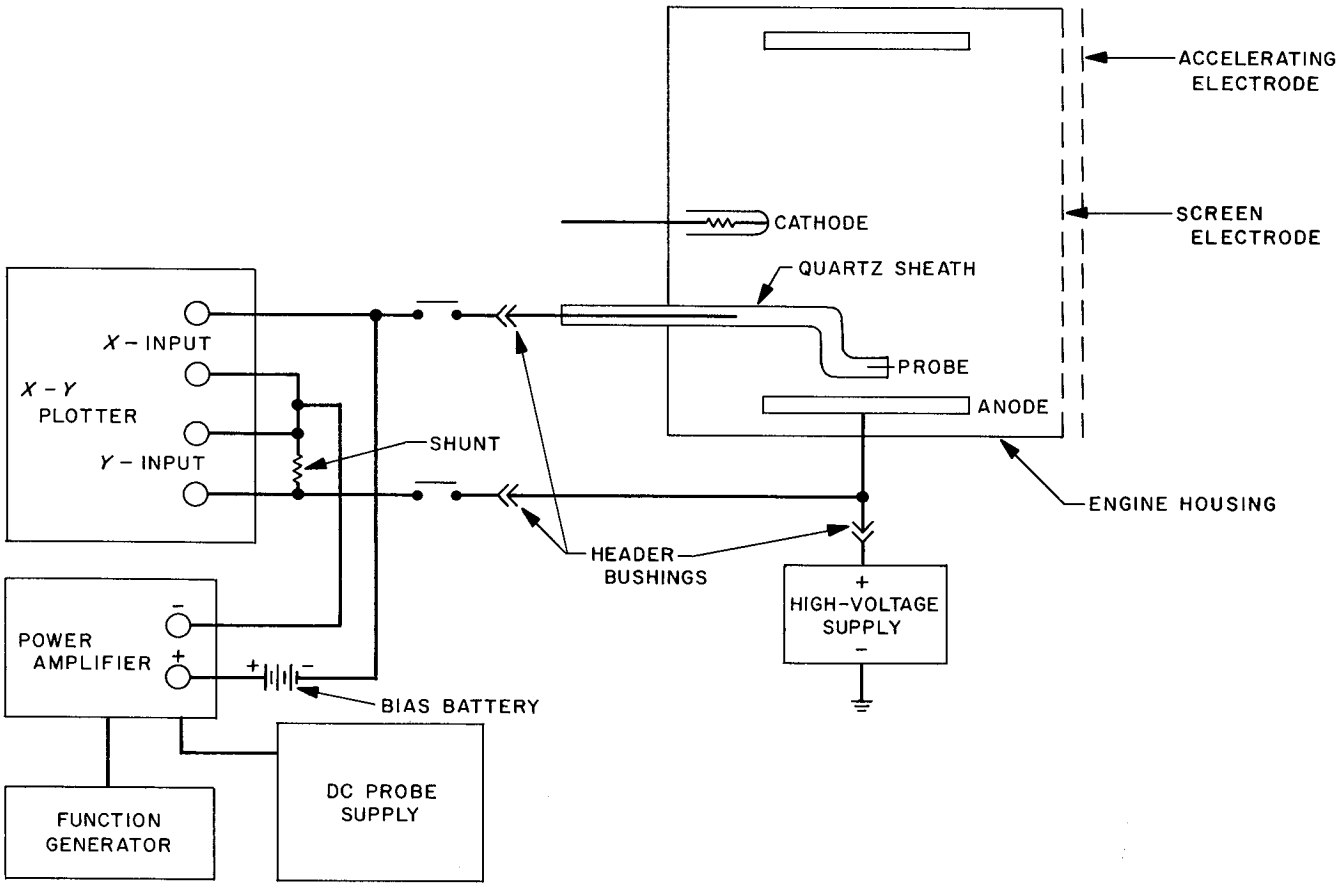


Fig. 5. Basic circuit diagram of engine and probe

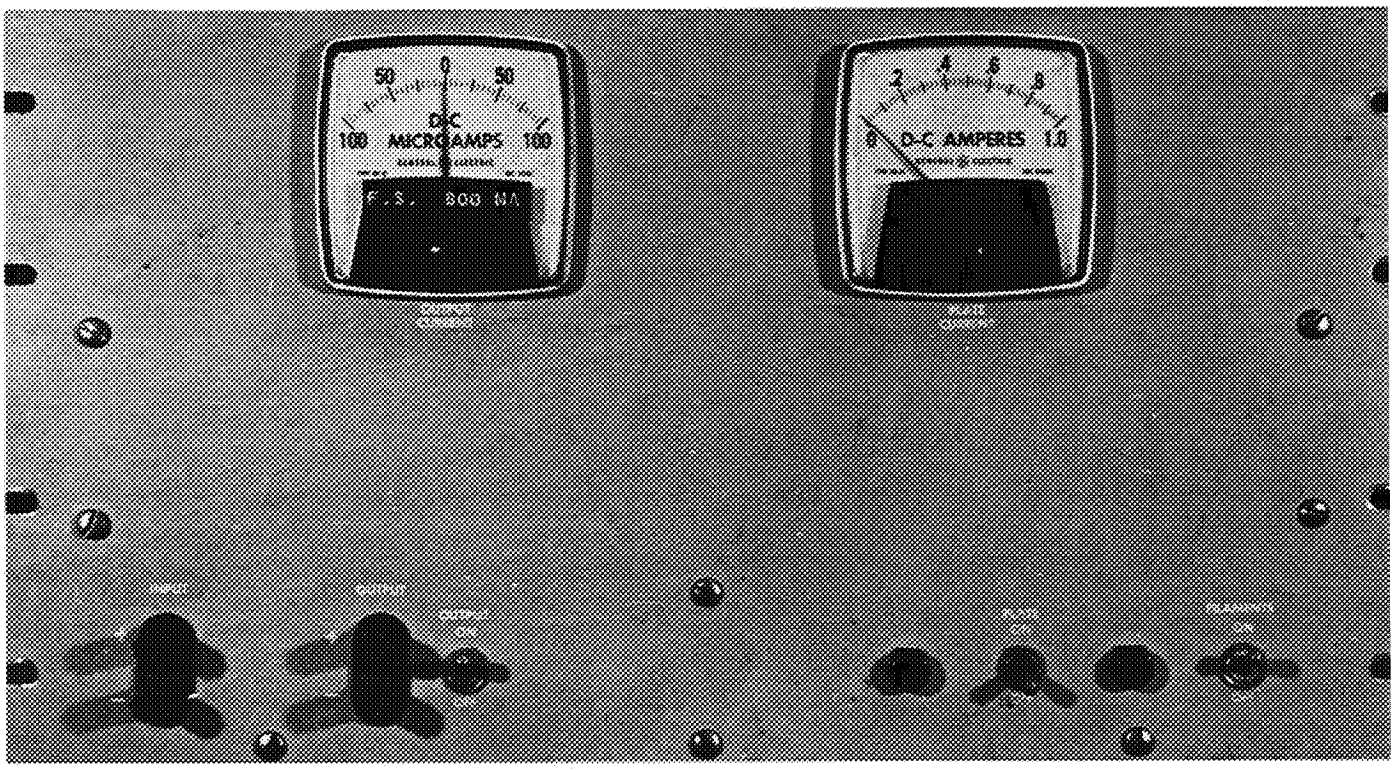


Fig. 6. Cathode follower amplifier

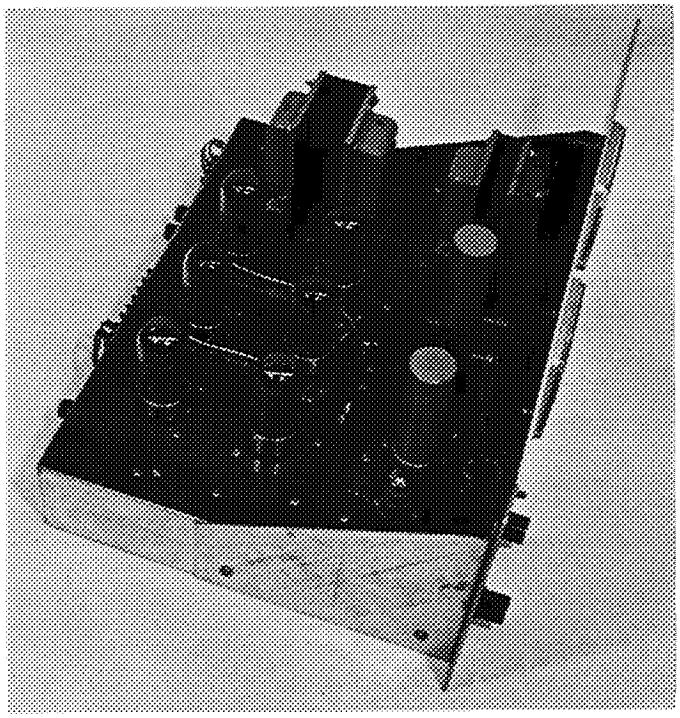


Fig. 7. Cathode follower amplifier

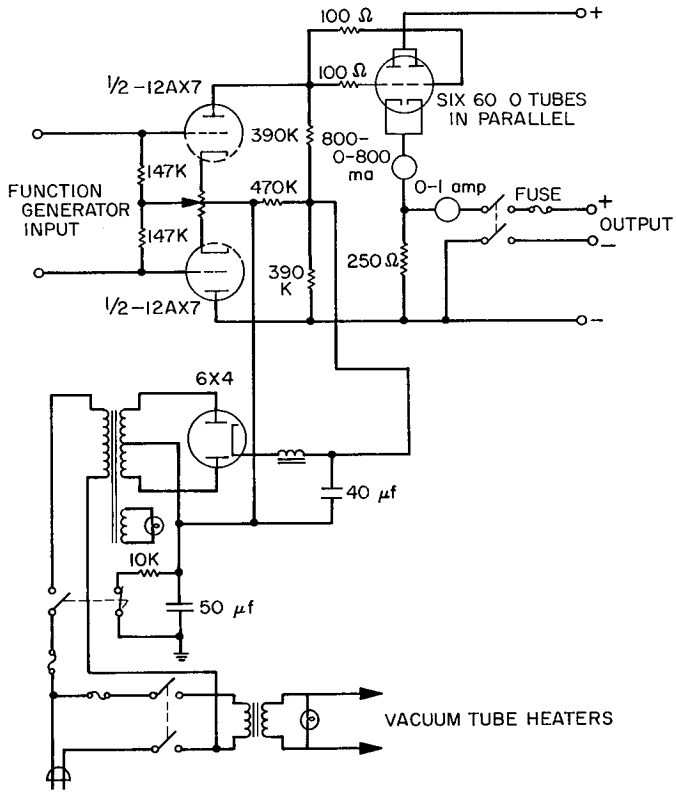


Fig. 8. Circuitry for cathode follower amplifier

driven by a special power amplifier described below. A low-frequency function generator is used to drive the power amplifier. Drive frequency is not critical; however, experience has shown that 0.05 Hz is a good compromise since it gives adequate tracing speed while remaining well within the plotter's ability to follow.

The probe input impedance varies over a wide range as it is cycled from a negative potential to a positive potential. For large negative potentials with respect to the plasma potential, only positive ions will be collected. As the probe potential becomes less negative with respect to the plasma potential, increasing numbers of electrons will be collected until saturation occurs. The ratio of the maximum electron current to ion current may be as high as 1000 to 1. Electron currents may reach 300 or 400 ma.

A special amplifier is required to drive the probe because of the nonsymmetrical probe currents and low drive frequency. Figs. 6 and 7 show a vacuum tube cathode follower amplifier that has proved very reliable. It uses six 6080 dual triodes in parallel direct-coupled to a 12AX7 driver stage. Amplifier circuitry is shown in Fig. 8.

Probe power is furnished by a separate dc supply. Power requirements are about 250 v at 600 ma.

During operation, the engine is maintained at high potential relative to ground. Since the probe will always be at engine potential, means for isolating the probe and electronics from ground must be provided. In addition, provision must be made for automatically disconnecting the probe and grounding all electronic equipment without

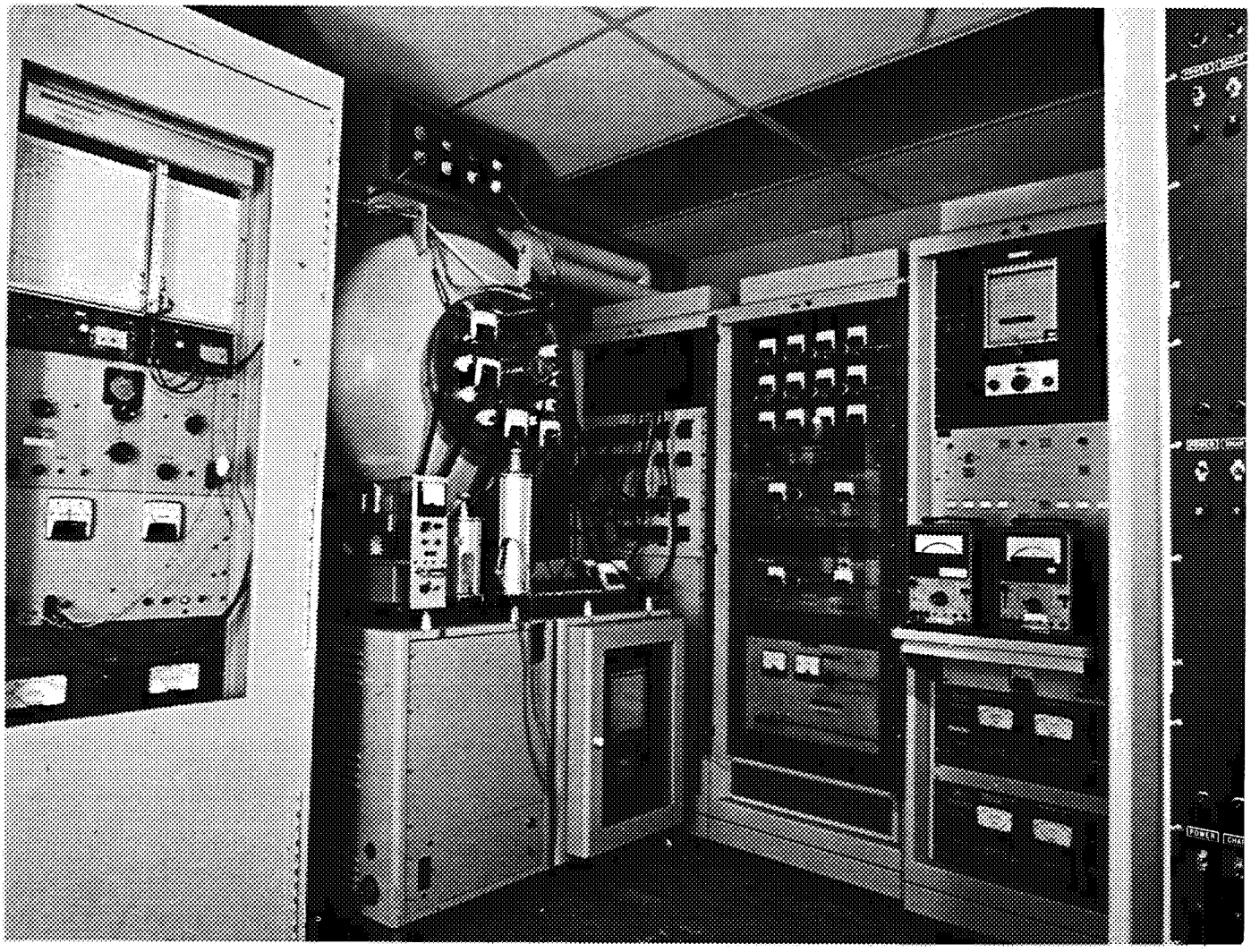


Fig. 9. Vacuum test chamber and support equipment (cabinet at left contains Langmuir probe instrumentation)

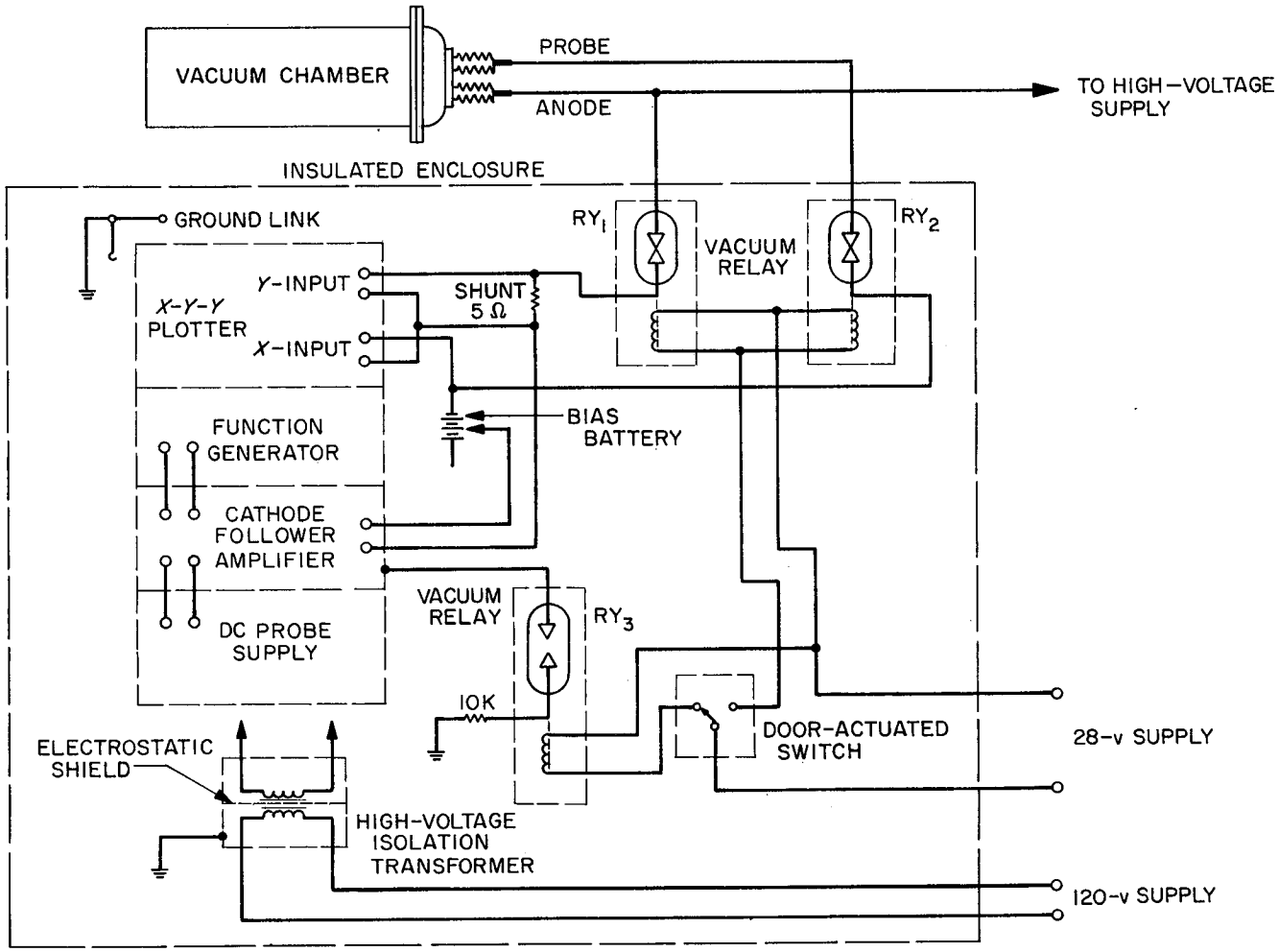


Fig. 10. Probe circuit diagram

in any way affecting engine operation. Automatic switching is required, since access to the plotter, amplifier, etc., is necessary after every trace.

Fig. 9 shows the complete instrumentation of an electron bombardment ion engine in a vacuum test chamber ready for a run. Just above the vacuum chamber, the high voltage supply terminals may be seen. The cabinets to the right and below contain low voltage power supplies and vacuum system control equipment. To the left is the Langmuir probe instrumentation. All the electronic equipment is mounted in a standard rack. High voltage isolation from the power line is provided by an oil-filled one-to-one ratio transformer. The transformer design voltage is 50 kv from secondary to primary and ground. The windings are electrostatically shielded. The complete circuitry is shown in Fig. 10.

An insulated cabinet encloses the equipment rack and affords personnel protection against high voltage. The cabinet is provided with a door for access to the equipment. Three vacuum relays (Fig. 10) are used for high voltage switching; these relays are actuated by a switch coupled to the cabinet door. When the door is open, relays RY_1 and RY_2 are actuated (open) and the system is grounded through relay RY_3 and a 10,000-Ω surge resistor. The vacuum relays are powered by an independent 28-v dc supply. One of the solenoid-operated vacuum relays is shown in Fig. 11. A redundant ground is provided by a metal link mechanically actuated by the cabinet door (Fig. 10).

Fig. 12 shows all equipment set up with a protective screen in place and ready for measurements. The two openings in the screen allow access to the Faraday and Langmuir probe positioners. Protection against high voltage is afforded by lucite cylinders supported by the two ferrules visible in the photograph. For additional protection, the operator is required to wear lineman gloves.

Recently, the cathode follower amplifier and probe power supply were replaced with a single transistorized unit. The circuitry is shown in Fig. 13. Two independent amplifiers, each with an adjustable power supply, permit independent adjustment of positive and negative probe potentials. This arrangement eliminates the need for the bias battery used with the tube amplifier. Chassis and panel layouts are shown in Figs. 14 and 15. Whether the solid state amplifier will prove to be as immune as the tube amplifier to high voltage transients must yet be determined. In Fig. 16, the amplifier is shown mounted just below the function generator.

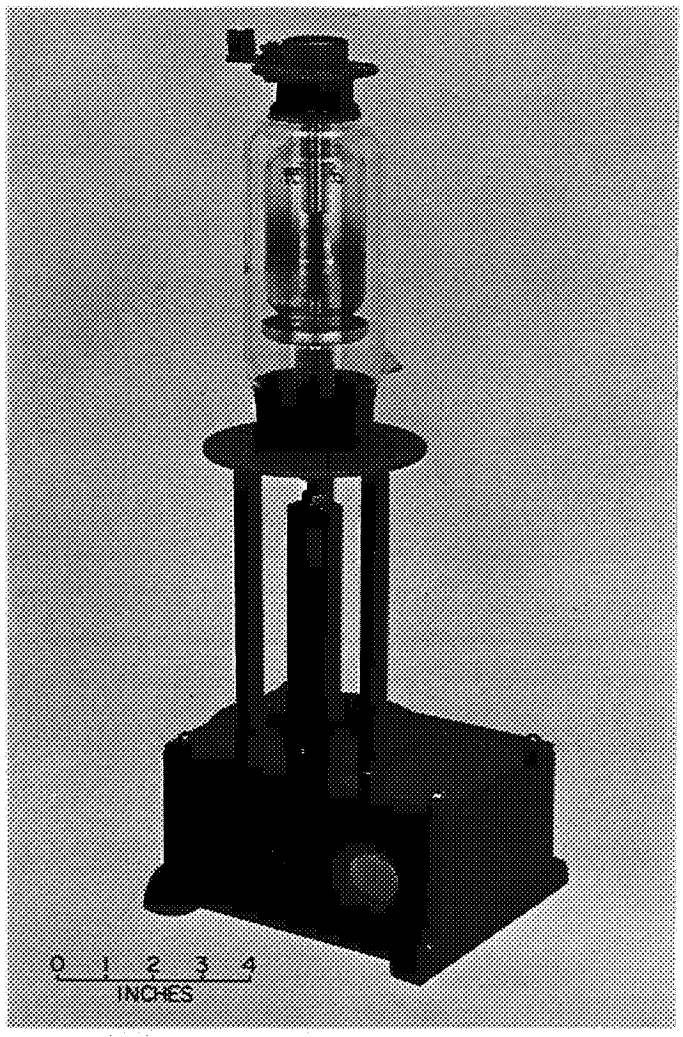


Fig. 11. Solenoid-operated vacuum relay

The usual Langmuir probe trace is recorded on a linear scale by the *X-Y* plotter. These volt-ampere data are then transferred to a semilogarithmic plot for further computation. Considerable time could be saved if the semilogarithmic plot were made directly by the plotter. Mercury could not be directly plotted on a semilogarithmic scale since a graphic correction for high energy primary electron distribution must first be made (Ref. 4). However, Langmuir data from a cesium plasma do not require such correction and the desired semilogarithmic plot could be obtained automatically.

During the next run with cesium as the propellant, it is planned to feed one channel of the *X-Y-Y* plotter from a logarithmic converter. The data thus obtained can then be compared with the semilogarithmic plot obtained from the linear scale.

Fig. 12. Vacuum test chamber with support equipment (high voltage protective screen in place)

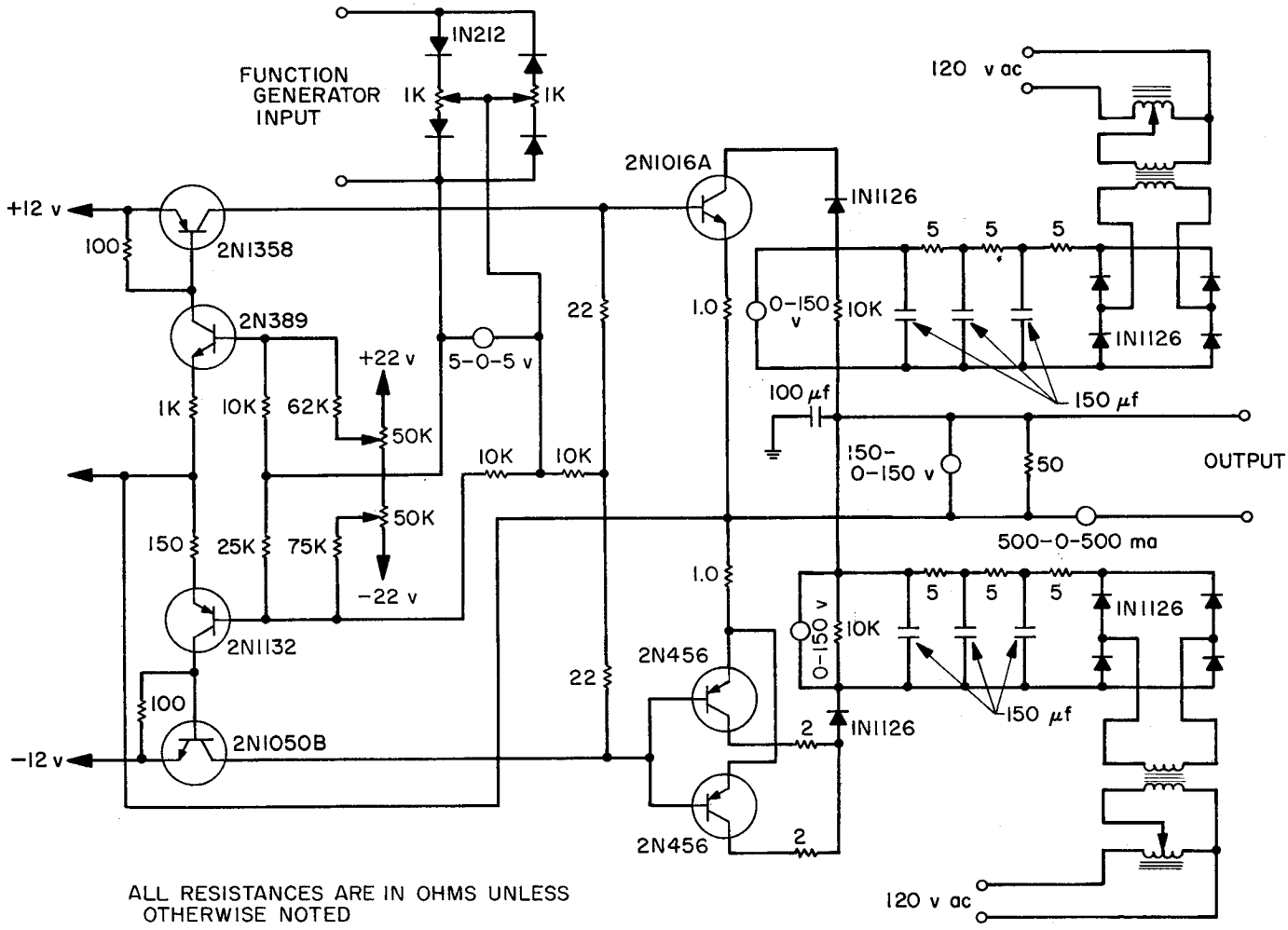


Fig. 13. Solid state probe power supply and amplifier

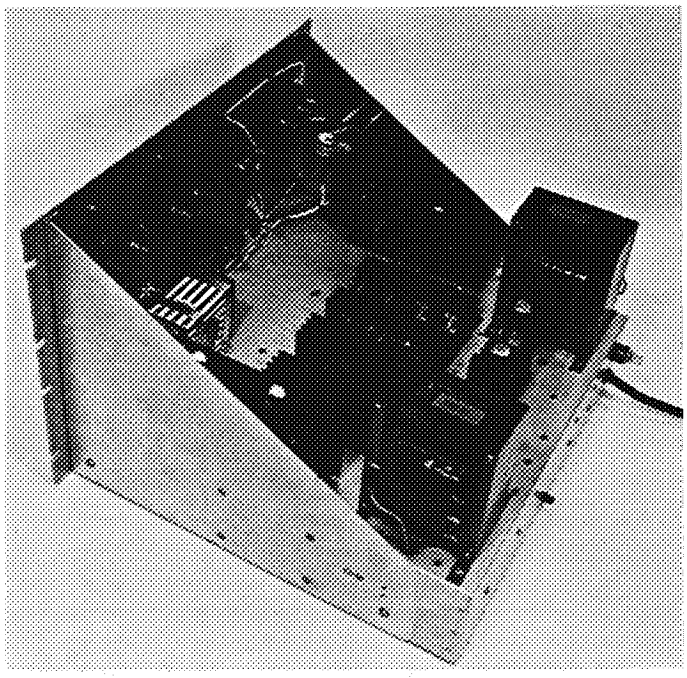


Fig. 14. Chassis of transistorized probe amplifier

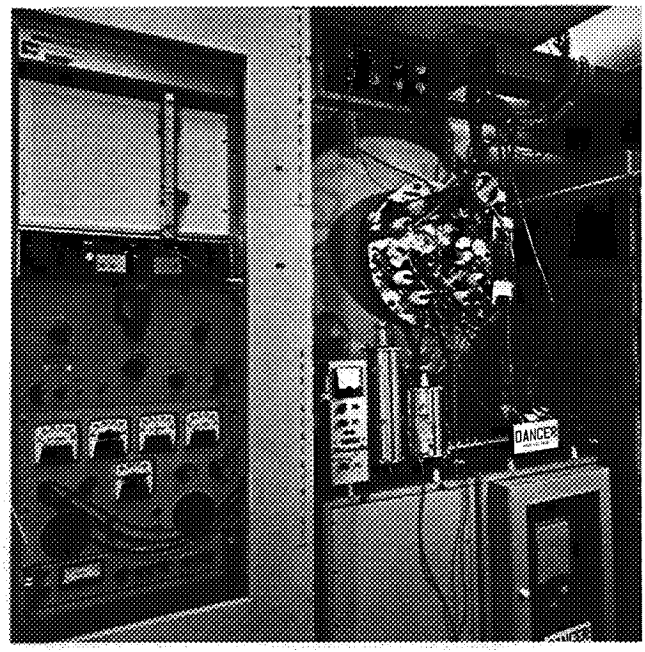


Fig. 16. Transistorized amplifier installed in Langmuir probe rack

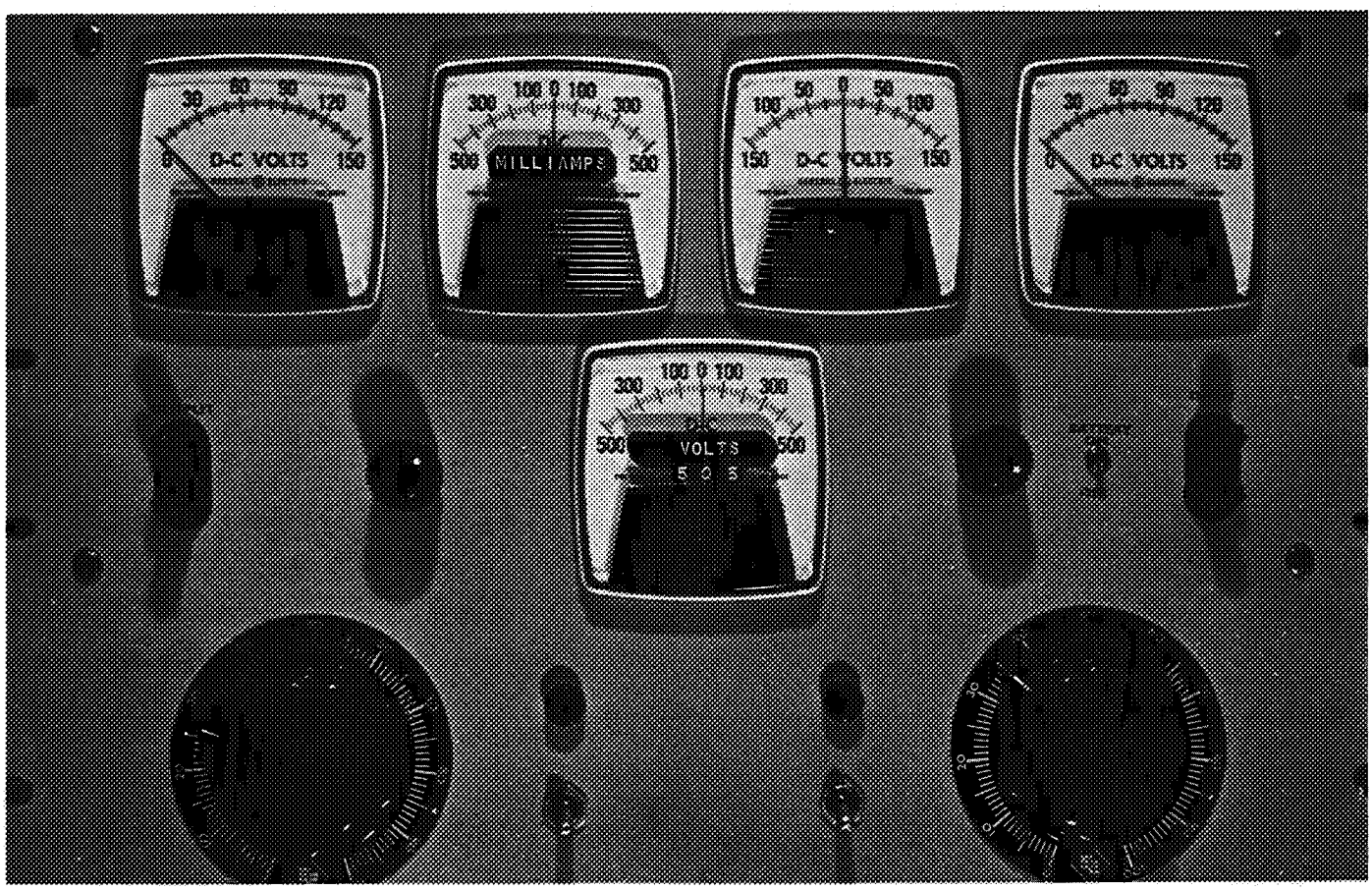


Fig. 15. Front panel of transistorized probe amplifier

In the past three years, several hundred Langmuir probe traces have been made in support of the many investigations concerning the properties of the mercury and cesium plasma inside the operating electron bombardment ion engine (Refs. 5 and 6; *SPS 37-32*, Vol. IV, p. 141; and *SPS 37-36*, Vol. IV, p. 107). During this time, the system has functioned reliably with a minimum of down time. Engine operation has not been affected, except to the extent normally expected, by the presence of the probe in the arc chamber. Repeatability of Langmuir probe traces made under identical operating conditions has been well within normal variations.

References

1. Langmuir, I., "Electron Discharges in Gases at Low Pressures," *Journal of the Franklin Institute*, Vol. 214, 1932, p. 275
2. Kaufman, H. R., *An Ion Rocket With an Electron Bombardment Ion Source*, NASA TN D-585, National Aeronautics and Space Administration, Washington D. C., January 1961.
3. Langmuir, I., and Mott-Smith, H., Jr., "Studies of Electric Discharges in Gases at Low Pressures," *General Electric Review*, Vol. 27, 1924, pp. 762, 810.
4. Strickfaden, W. B., and Geiler, K. L., *Probe Measurements of the Discharge in an Operating Electron Bombardment Engine*, Technical Report 32-417, Jet Propulsion Laboratory, Pasadena, Calif., April 19, 1963. (Also available in *AIAA Journal*, Vol. 1, 1963, pp. 1815–1823).
5. Kerrisk, D. J., and Masek, T. D., "Effects of Plasma Non-Uniformity on Grid Erosion in an Electron Bombardment Ion Engine," *AIAA Journal*, Vol. 3, 1965, p. 1060.
6. Masek, T. D., *Plasma Investigation in a Reversed Current Electron Bombardment Ion Engine*, Paper 66-246, presented at AIAA Fifth Electric Propulsion Conference, San Diego, Calif., March 7–9, 1966.

PROPULSION DIVISION

VIII. Solid Propellant Engineering

A. Applications Technology Satellite Motor Development

R. G. Anderson and D. R. Frank

1. Introduction

In January 1963 the Jet Propulsion Laboratory initiated a development program to provide a solid propellant apogee rocket motor for a second-generation *Syncom* satellite. This program, under the management of the Goddard Space Flight Center, was designated *Advanced Syncom*. It was to result in a spin-stabilized, active repeater communications satellite weighing about 750 lb, operating at synchronous altitude (23,300 mi), which would handle voice communications, teletype, and monochrome and color television signals.

In January 1964 the *Advanced Syncom* communication program was redirected to include a number of experimental instruments in addition to the original communication instruments. This expanded program is the Applications Technology Satellite (ATS) program and will result in a general-purpose satellite capable of operation at synchronous altitude with experimental instruments in the areas of meteorology, communications, radiation, navigation, gravity gradient stabilization, and various engineering experiments. For those satellites to be placed in synchronous orbit JPL will provide a solid propellant rocket motor to provide the final required velocity increment at the apogee of the elliptical transfer orbit. This rocket motor is designated the JPL SR-28-1 (steel chamber) or JPL SR-28-3 (titanium chamber) rocket motor. It is presently intended that only the JPL SR-28-3 unit will be delivered for flight use.

Previous reports of progress on the development of this motor have been published in *SPS 37-20* to *37-33*, Vol. V and *SPS 37-34* to *37-38*, Vol. IV.

2. Program Status Summary

The motor development program calls for static firing of four heavywall motors and 25 flightweight motors including two with flight design titanium chambers prior to conducting a nine-motor qualification program. To date, the four heavywall motors plus 21 flightweight

motors have been static-fired, four of which were under simulated high-altitude conditions at Arnold Engineering Development Center (AEDC) Tullahoma, Tennessee. All of the flightweight motors tested to date have been with Type 410 chromium steel chambers, with the exception of Dev. G-8T and G-9T, which used titanium chambers.

a. ATS storage units. The three storage rounds, cast during September 1965, are presently in storage. These units were removed after approximately 6 mo of storage (April 1966) and were given critical visual and physical inspections. No detrimental effects due to the 6 mo of storage were observed.

One unit will be stored for 1 yr before static testing. The remaining two units will be stored for two years before testing. Periodic visual and physical inspections will be made at designated times.

b. ATS qualification motors. On March 3, 1966 the last qualification motor for tests at AEDC from July to September of this year was cast. Of the nine motors cast for this purpose, eight of these motors will be fired while spinning at 100 rpm. The eight motors will be divided into two groups, with the first group being tested following a 40°F-temperature-conditioning period and the second group following a 110°F-temperature-conditioning period. The ninth motor will remain in reserve to be tested only if during any of the first tests the objectives are not completed.

Presently the nine qualification motors are being put through various environmental tests and dimensional checks. A schedule is presented in Table 1 which gives the tests and checks that have been completed to this date.

c. Static test of Dev. E-3T. During the first week of June 1966, Dev. E-3T will be statically fired at AEDC in Test Cell J-5 for a combination apogee-motor–spacecraft test. The apogee-motor–spacecraft assembly will be mounted on a soft test stand to evaluate the vibrational inputs to the spacecraft. Of primary importance during the test will be the evaluation of thermal inputs to the spacecraft as a result of the apogee motor firing. Approximately 80 thermocouples will be attached to the motor and spacecraft to measure the temperatures at these locations. The complete safe and arm system will also be evaluated during this test. The mechanical portion of the safe and arm device was successfully tested on May 3, 1966 at Edwards Test Station (Dev. I-1).

d. Static test of Dev. I-1. On May 3, 1966 the test of Dev. I-1 was conducted. The test was primarily used to evaluate the mechanical compatibility of the closure portion of the safe and arm device (built by the Harry Diamond Laboratory) with the apogee motor and

Table 1. Pretest schedule for the ATS qualification motors

Motor development code	Q-1T	Q-2T	Q-3T	Q-4T	Q-5T	Q-6T	Q-7T	Q-8T	Q-9T
Motor case	✓	✓	✓	✓	✓	✓	✓	✓	✓
Postcast X-ray inspection	✓	✓	✓	✓	✓	✓	✓	✓	✓
Environmental tests									
Temperature cycle	✓	✓	✓	✓	✓	✓	✓	✓	✓
Booster acceleration	✓	✓	✓	✓	✓	✓	✓	✓	✓
Booster vibration	✓	✓	✓	✓	✓	✓	✓	✓	✓
Operations following all environments									
X-ray inspection	✓	✓	✓	✓	✓	✓	✓	✓	
Special propellant grain alignment	✓	✓	✓	NP	NP	NP	NP	NP	NP
Motor assembly alignment	✓	✓	✓	✓	✓				
CG and moment-of-inertia determination	✓	✓	✓	✓	✓			NP	NP
Preshipment operations									
Visual inspection									
Alignment inspection									
Motor pressure test									

A ✓ means that this test or check has been completed.
NP means check or test is not to be performed.

to determine the ability of the device to withstand the heat loads encountered during a full-duration motor firing. Thermocouples were installed on the safe and arm hardware to assist in evaluating the temperature loads to which the hardware is subjected during and immediately following the test. The safe and arm hardware provided two special pressure taps which allowed the igniter basket pressure and the motor chamber pressure to be measured throughout the test.

Preliminary test results show that the mechanical portion of the safe and arm device can withstand the pressures and temperatures of a full-duration apogee motor firing. Detailed test results will be presented in the next issue of *SPS*.

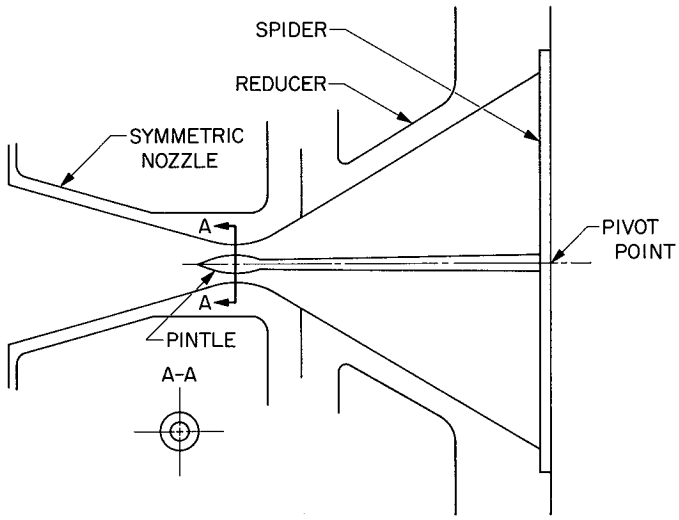


Fig. 1. Gas-flow nozzle

B. Pintle Nozzle Thrust Vector Control

L. Strand

1. Introduction

Pintle nozzles have become a part of the current development work in solid rocket technology. By traversing the pintle either forward or aft along its axis, the nozzle throat area is changed. Pintle nozzles have therefore been used to obtain both thrust magnitude control and thrust termination (L^* extinguishment). If thrust vector control (TVC) capabilities could also be incorporated into the pintle nozzle, it could result in a very competitive propulsion system. A test program was therefore initiated to attempt to investigate one approach to pintle nozzle TVC. The objectives of this program were to (1) determine if a pivoted pintle-nozzle system is capable of producing side thrust of sufficient magnitude for nozzle TVC purposes and (2) determine the criticalness of nozzle pintle position on thrust alignment for thrust magnitude control–thrust termination pintle nozzle systems.

2. Pintle Systems

A pintle system was designed to be incorporated into the gas-flow nozzle test assembly (Fig. 1) of the JPL flow channel, described in *SPS 37-35*, Vol. IV. The pintle was to be capable of being pivoted either in the horizontal or the vertical plane about the pivot point shown at

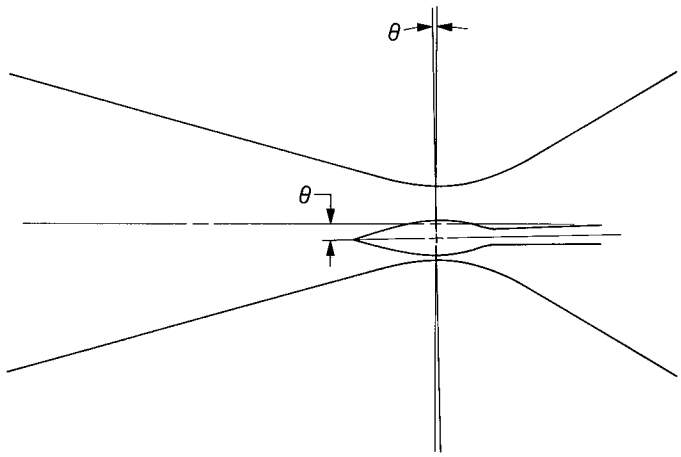


Fig. 2. Canted pintle

angular displacements (θ in Fig. 2) from the nozzle centerline. Theta values of 0°, 15′, 30′, 1°, and 1°30′ were chosen.

The new equipment, consisting of the pintle and pivot-position system and its supporting spider, is shown disassembled in Fig. 3. The pintle is 0.50 in. in diameter. Its contour is the same as that of the axisymmetric nozzle used, a 2.82-in. radius of curvature terminating in a 15° cone. The pintle is held in its canted position by one of the five threaded hubs shown. The conical cup mounts forward of the hubs to make a cleaner obstruction to the flow (because of the low velocity of the flow upstream of the nozzle reducer, no flow perturbation in the nozzle was anticipated or found). Fig. 4 shows the pintle system assembled with the 1°30′ positioning hub,

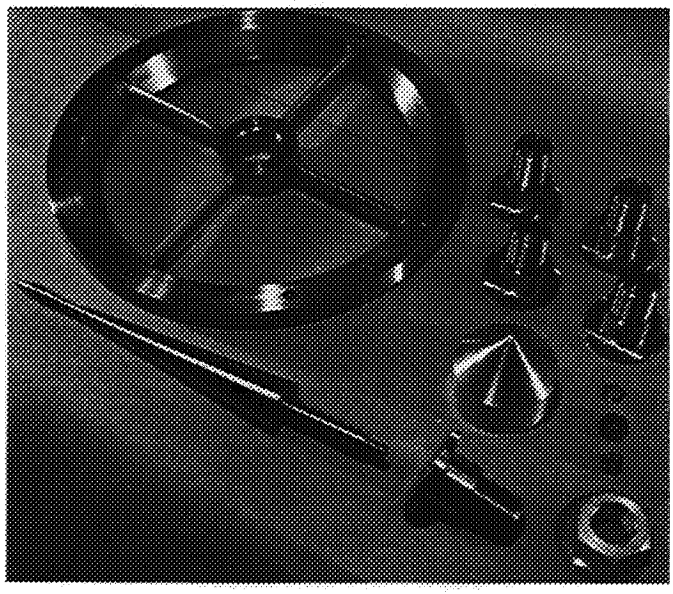


Fig. 3. Disassembled pintle system

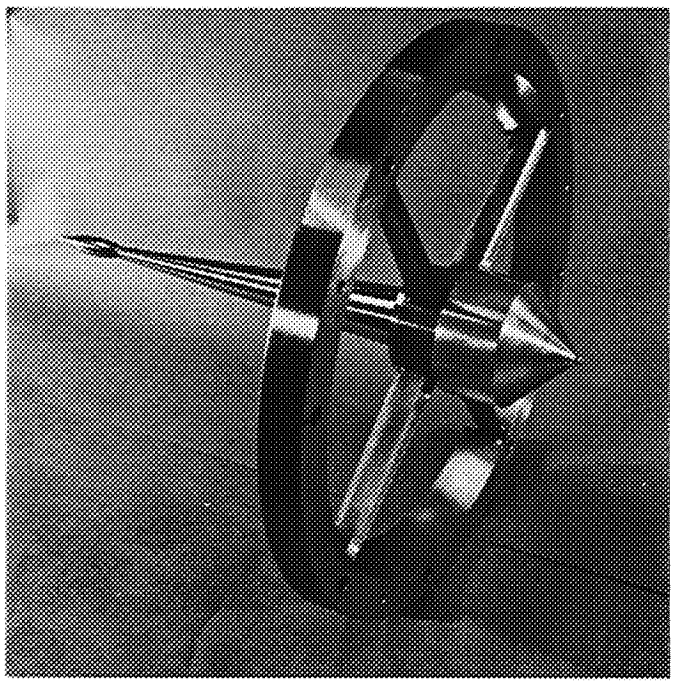


Fig. 4. Assembled pintle system

and Fig. 5 shows the pintle system mounted in the flow channel reducer, modified to accept the supporting spider. The pintle shaft, mounted with its 0° positioning hub, was measured to be out-of-round with the reducer exit by less than 0.004 in. For all pivoted-pintle tests the pintle was canted to the right in the horizontal plane, viewing the pintle from an upstream position (Fig. 5).

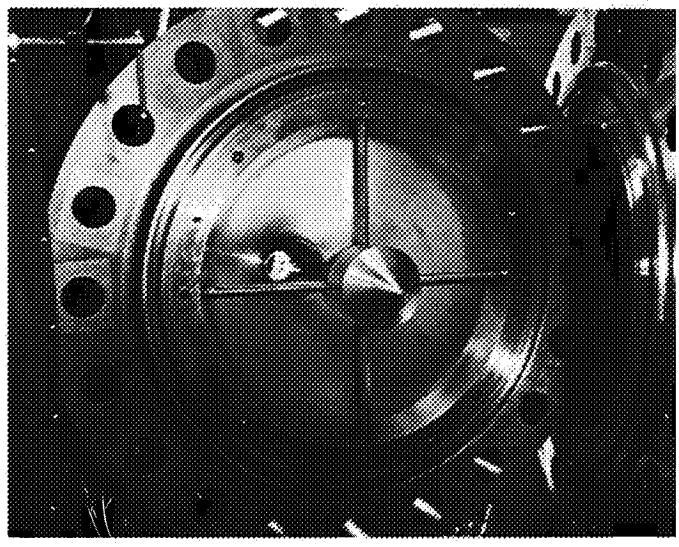


Fig. 5. Mounted pintle system

3. Nozzle

The existing 1.142-in. throat diameter axisymmetric nozzle (described in *SPS 37-35*, Vol. IV) was instrumented with approximately 50 additional static pressure taps, making a total of 140 taps in the nozzle. For this program 98 of the taps were used, giving a pressure tap net over the inner surface of the nozzle entrance section, throat section, and aft expansion cone. From the measured nozzle wall pressure distribution and an estimate of the pressure distribution over the pintle surface, an approximate value of the force unbalance normal to the nozzle axis can be obtained. Fig. 6 is a cross-sectional

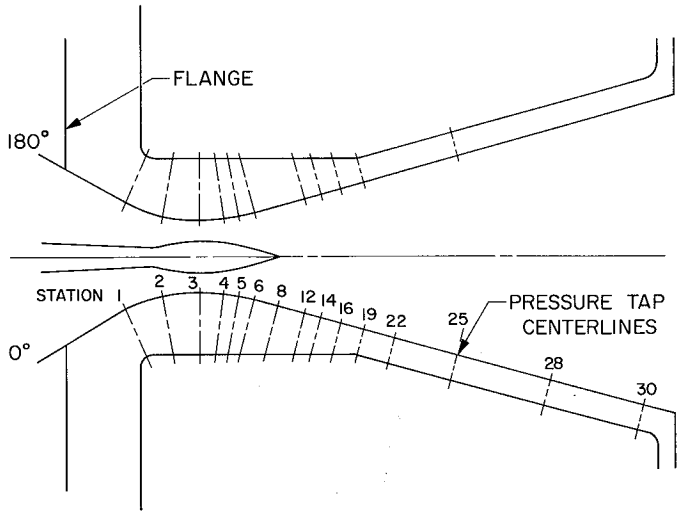


Fig. 6. Pintle nozzle cross section

sketch of the nozzle along the pressure tap centerline, showing the axial spacing of the tap positions. Fig. 7 is an in-the-flat representation of the pressure tap spacing about the same centerline. The throat area with the pintle located in the nozzle was 0.832 in. squared. The 15′, 30′, 1°, and 1°30′ angular displacements of the pintle

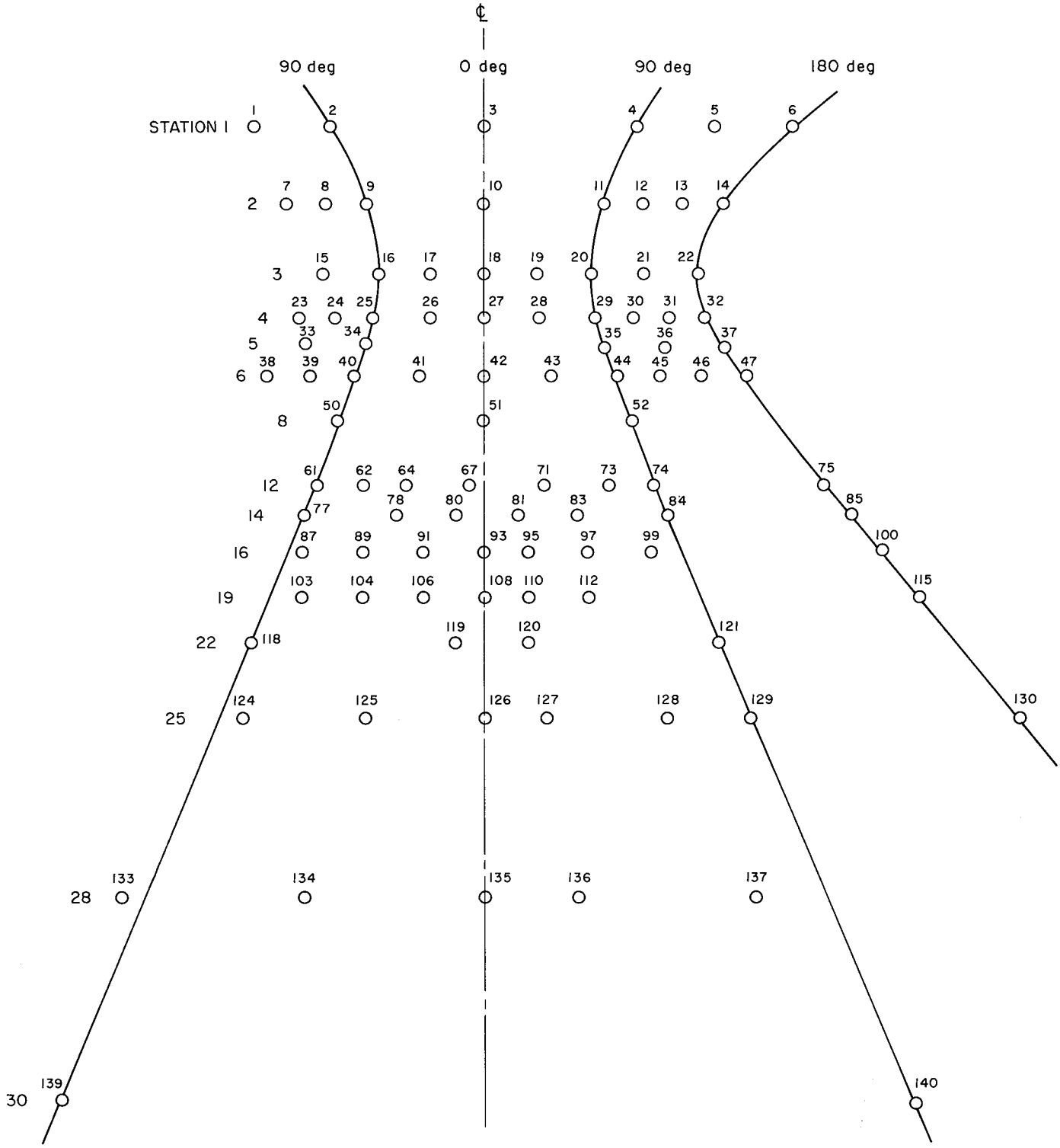


Fig. 7. Pintle nozzle in-the-flat pressure tap spacing

resulted in lateral pintle displacements at the pintle throat of 0.04, 0.09, 0.17, and 0.26 in., respectively. At the 1°30′ angular displacement the minimum separation distance between the pintle and the nozzle wall was 0.06 in.

4. Instrumentation

Static pressure measurements were made with the modified multiple pressure measuring system (MPMS). The instrumentation setup was essentially identical to that described in *SPS 37-35*, Vol. IV and will not be described further here.

5. Test Procedure

Five nozzle test configurations were used, corresponding to the five theta values. The functions of the 0° tests were both to check the flow uniformity of the pintle-nozzle system and to provide a base for interpreting the test results for the other configurations. For each configuration tests were run at three supply pressures, 100, 400, and 600 psia. The supply air temperature was only raised to a high enough level (approximately 150°F) to ensure against air liquefication in the nozzle expansion region. The air temperature was controlled by the Hypersonic Tunnel heater. At each pressure the MPMS made four scans of the nozzle static pressure taps, giving four sets of pressure data.

The nozzle test assembly was disassembled for each test configuration change. Figs. 8 and 9 show the test assembly in its assembled and disassembled conditions, respectively.

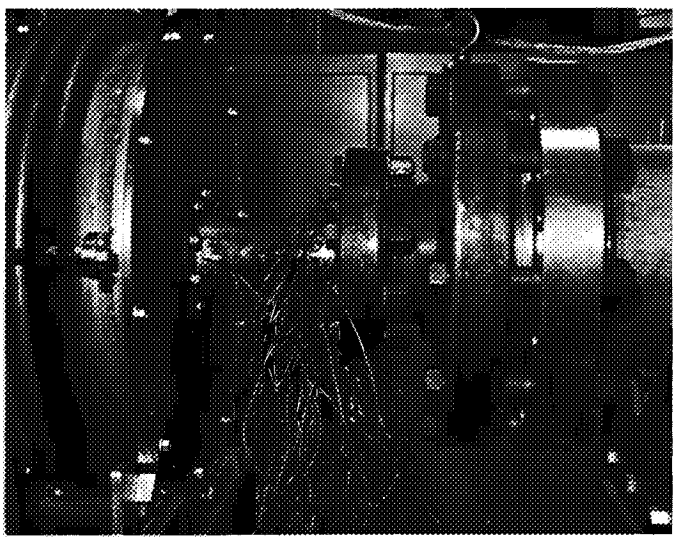


Fig. 8. Assembled nozzle test assembly

Fig. 9. Disassembled nozzle test assembly

6. Test Results

Other than one bad pressure tap that measured atmospheric pressure throughout the tests, no problems were encountered. Instrumentation system accuracy and reproducibility appear comparable to that obtained in the earlier auxiliary flow channel test (*SPS 37-35*, Vol. IV).

In Figs. 10 to 14 the ratioed static pressure data along the two 90° profiles designated in Fig. 7 are shown for five test run scans at 0°, 15′, 30′, 1°, and 1° 30′ theta values, respectively. The supply pressure is 600 psia, although the same pressure ratio profiles were found for the other two supply pressure values. The pintle was pivoted toward the 90° profile on the left and away from the 90° profile on the right in Fig. 7. A pressure perturbation, apparently from a recompression wave at the pintle tip, appears on one pressure profile in Fig. 10 at Station 19. Upstream of this point agreement between the two profiles is very good. As theta increases, the perturbation moves upstream in the nozzle and increases in length and abruptness of the pressure rise. The predicted differences in rate of expansion along the nozzle axis for the two profiles and characteristic crossover (at an axial distance of approximately 2.9 in.) appear in Fig. 11 and become more distinct as theta increases.

The experimental data are currently being reduced and analyzed, and a complete description of these results and the results of the side force summations will be reported in the next *SPS*.

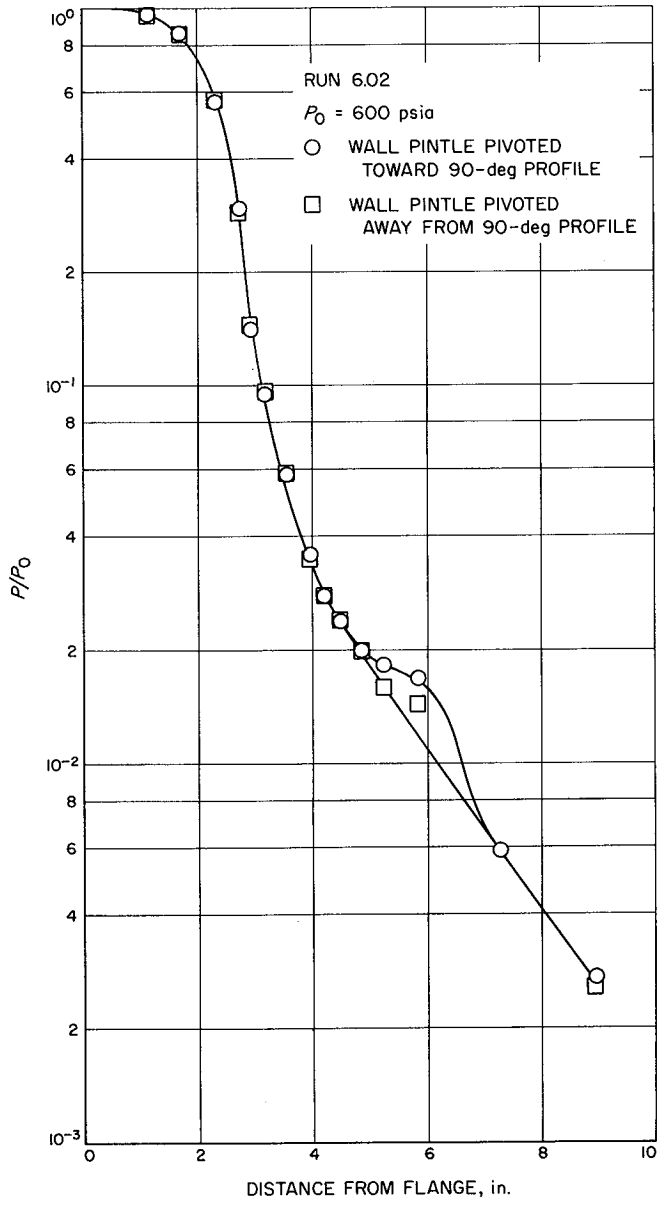


Fig. 10. Static pressure ratio versus nozzle axial distance, θ = 0°

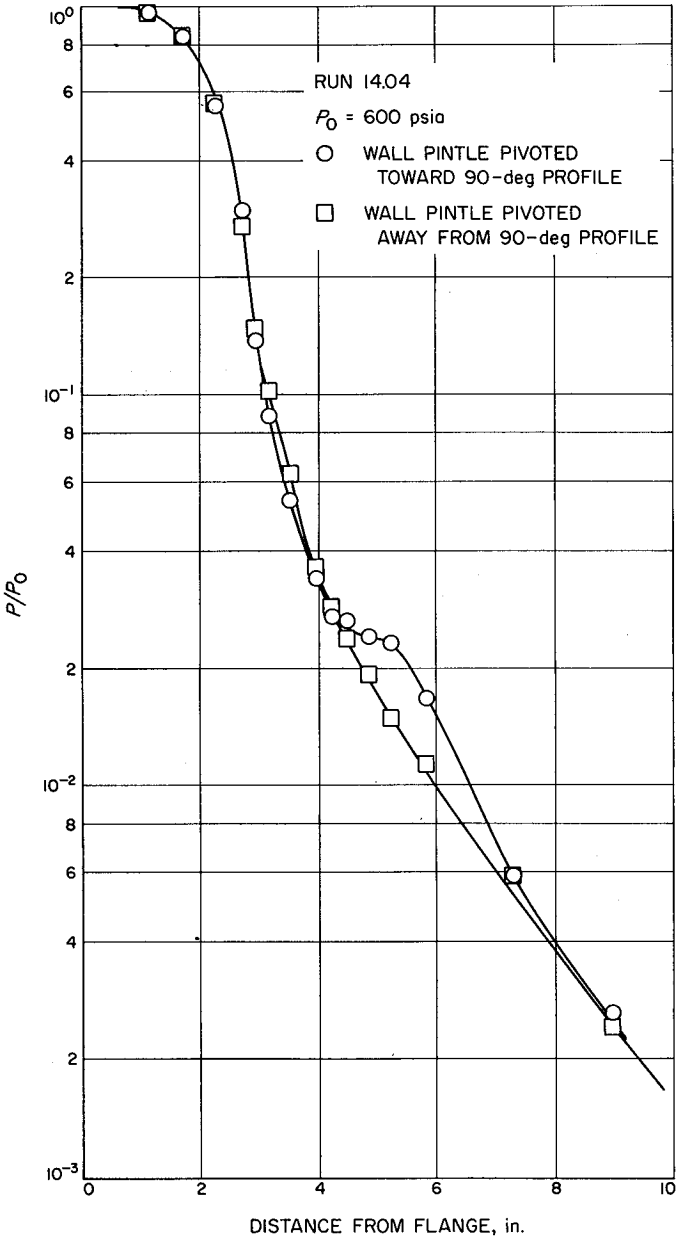


Fig. 11. Static pressure ratio versus nozzle axial distance, θ = 0° 15′

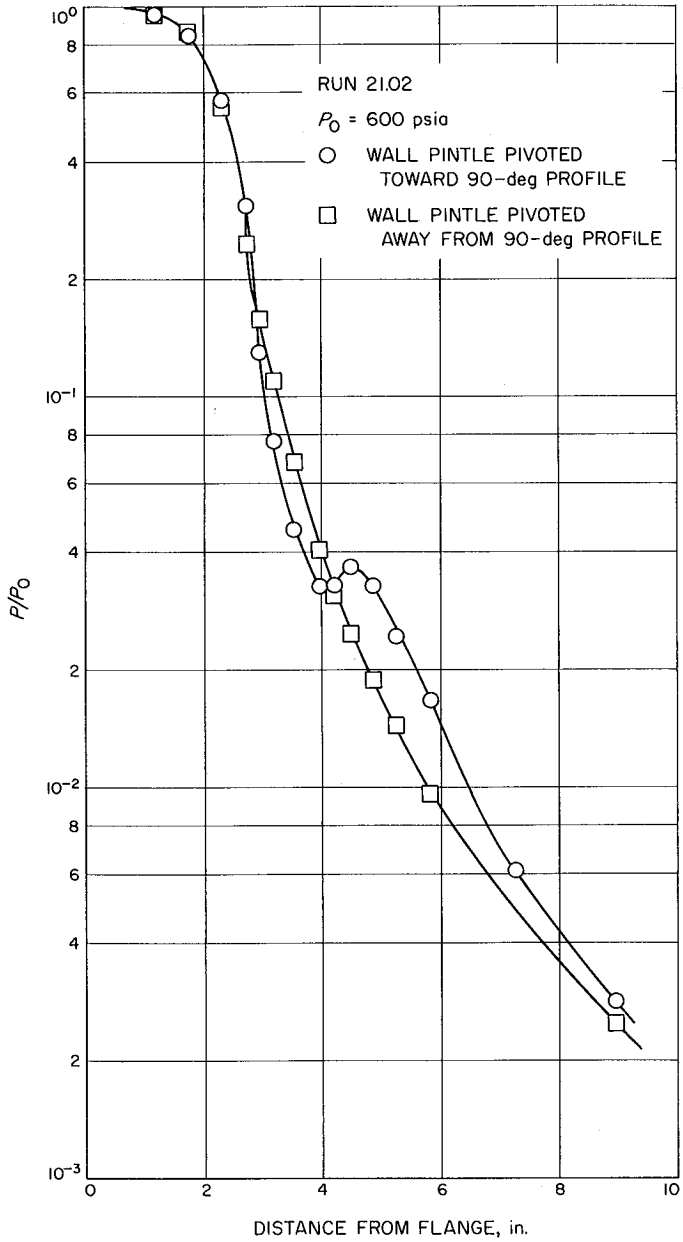


Fig. 12. Static pressure ratio versus nozzle axial distance, θ = 0° 30′

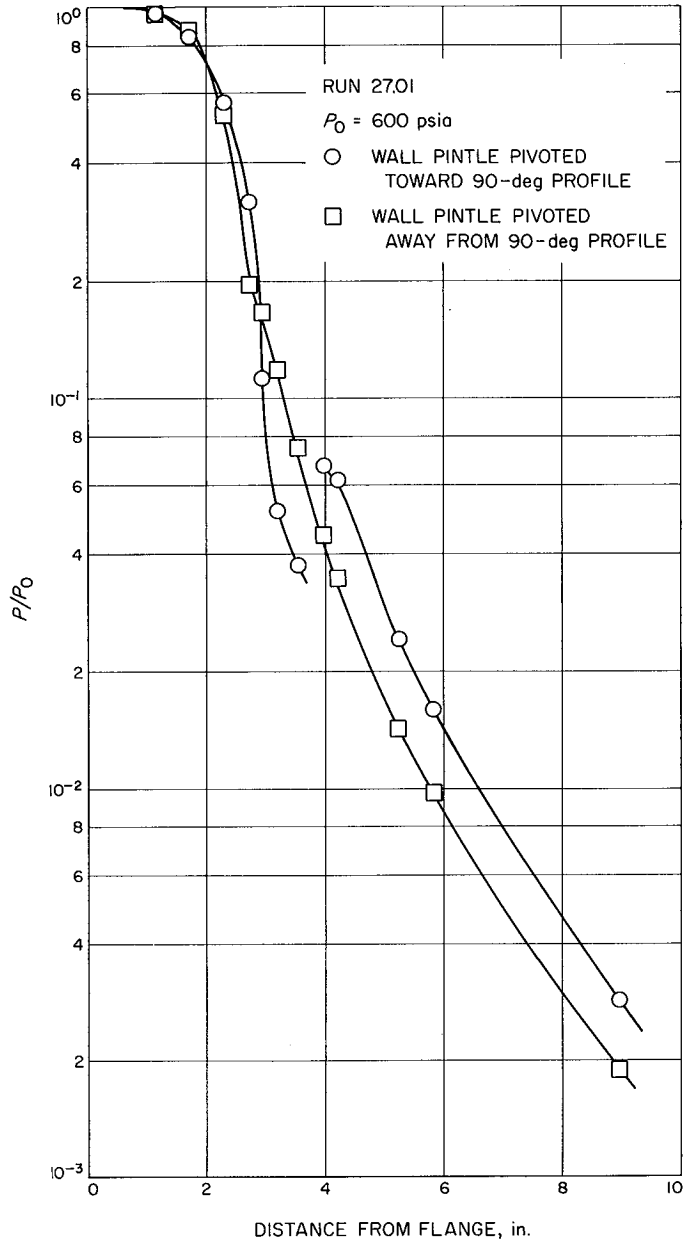


Fig. 13. Static pressure ratio versus nozzle axial distance, θ = 1°

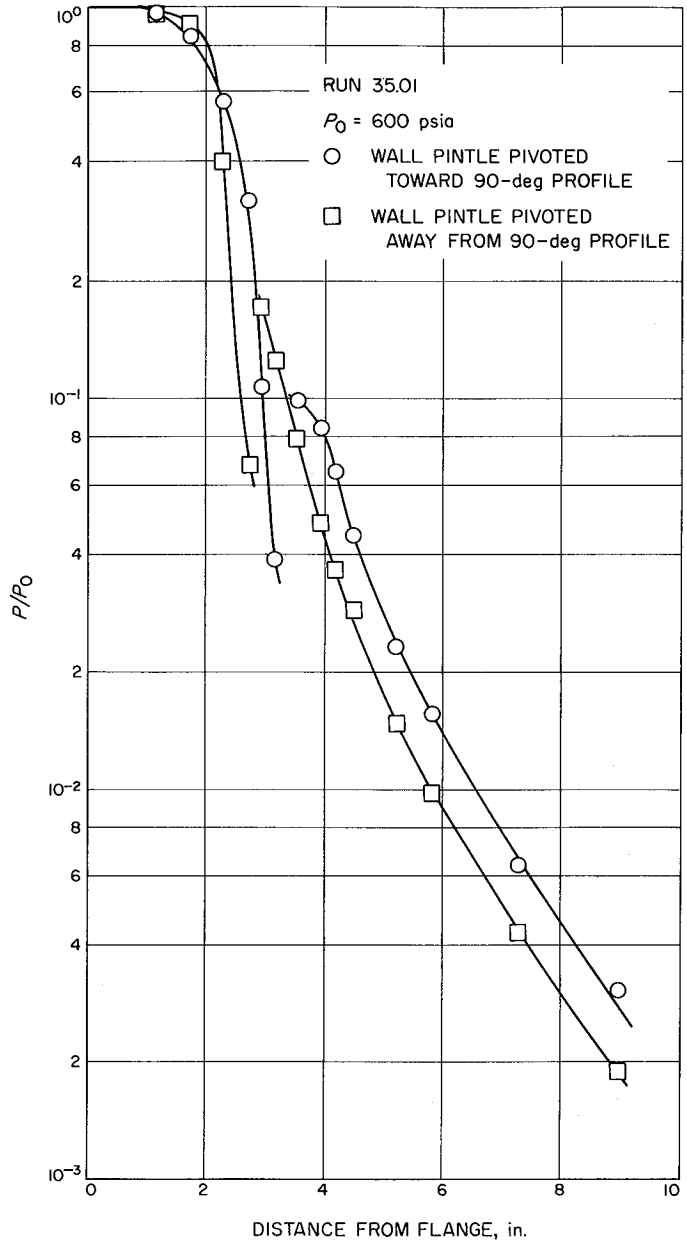


Fig. 14. Static pressure ratio versus nozzle axial distance, $\theta = 1^\circ 30'$

IX. Polymer Research

A. Calculation of a Refractive Index-Molecular Weight Relationship for Poly(Ethylene Oxide)

D. D. Lawson and J. D. Ingham

The experimental ease with which the refractive index n_D^t of a polymer may be obtained has prompted us to investigate the applicability of n_D^t-molecular weight dependency for determinations of number-average molecular weights. A previous publication demonstrated the practicality of the method if data are available for a typical polymer, poly(ethylene oxide), for the establishment of linear plots of n_D^t versus $(M_n)^{-1}$ (Ref. 1). The results indicate that accuracy approaching that for colligative properties relationships could be obtained if refined measurements (e.g., n_D^t by differential refractometry and precise density measurements) were carried out. However, the accumulation of relatively large amounts of highly precise initial data might tend to detract from the simplicity of the method. Therefore, it was believed that semiempirical calculations of the initial relationships would make either estimate methods or refined methods more amenable for routine use, by decreasing the required measurements. Although we have carried out such calculations for several different polymers, experimental data are presently available only for poly(ethylene oxide) as a check on the agreement of calculated n_D^t versus M_n with experimental results.

1. Calculation of Molecular Refraction

This method of calculation takes advantage of the molecular refraction as an additive and constitutive property for which the specific polymer of different number-average molecular weights represents a homologous series. Although the additivity of refractivity constants is not absolutely precise for all possible polymer structures, molecular refraction has been the subject of numerous empirical investigations since the work of Herschel in 1830, and it is expected that some form of molecular refraction scheme would be accurate for most polymers (Ref. 2). For poly(ethylene oxide) atomic, group and structural refractivities given by Vogel were summed to

C-2

obtain the molecular refractions (Ref. 3). Since the empirical formula is

$$HO(CH_2CH_2O)_NH$$

then

$$R_D = R_D^{(CH_2)}(2N) + R_D^{(O)}(N-1) + R_D^{(OH)}(2) \quad (I)$$

$$= 4.647(2N) + 1.764(N-1) + 2.546(2)$$

Although systems of additive bond refractivities have been established and these have a more sound theoretical basis than structural refractivities (Ref. 4), the above treatment has been found to be adequate for poly(ethylene oxide).

2. Calculation of Refractive Indices

The Lorenz-Lorentz equation (Refs. 2, 5, 6) was used to calculate refractive indices:

$$R_D = \frac{(n_D^t)^2 - 1}{(n_D^t)^2 + 2}\left(\frac{M}{d}\right) \text{ or } n_D^t = \left[\frac{M + 2R_D(d)}{M - R_D(d)}\right]^{1/2} \quad (II)$$

in which M is the number-average molecular weight, and d is the polymer density. At low degree of polymerization, and for very accurate calculations, the polymer density should be determined. However, the density dependence on molecular weight is very small for poly(ethylene oxide) so that a nominal density of 1.08 at 75° can be assumed (Ref. 1).

Calculated refractive indices at 75°C versus the reciprocal of molecular weight are shown in Fig. 1 as the solid line. The measured values (Ref. 1) are in very good agreement with the calculated curve except for a slight negative shift. Calculations on the basis of bond refractivities and the substitution of a form of the Eykman equation for the Lorenz-Lorentz equation will be carried out to see if either of these modifications results in closer correspondence of the calculated and experimental results.

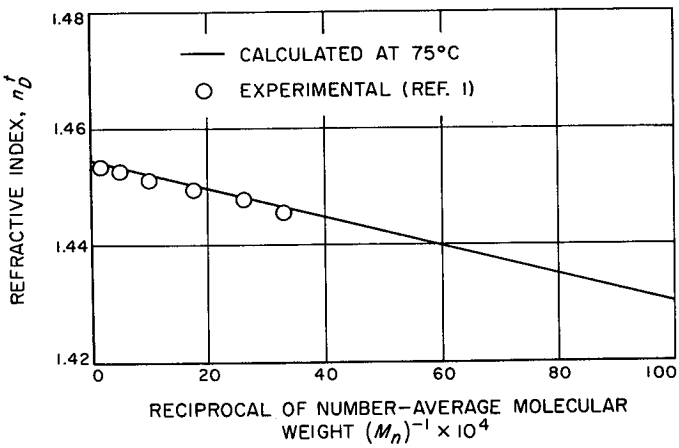


Fig. 1. Calculated and experimental relationship of n_D^t and M_n for poly(ethylene oxide)

B. Solid-State Batteries Based on Charge Transfer Complexes

F. Gutmann,[1] A. M. Hermann, and A. Rembaum

1. Introduction

This paper describes the results of studies exploring the possibilities of a truly solid-state electrochemical battery, i.e., without the use of liquid or fused electrolytes.

While devices have been reported (Ref. 7) using silver cathodes and silver halides as the electrolytes, the voltages attained were of the order of 0.2 to 1.0 v and the current densities a few 100 μa/cm^2, unless operated at elevated temperatures, e.g., 150°C and higher. The operation of such a device is governed mainly by the nature of the (solid) electrolyte: the electrolyte must give rise to an electrode reaction associated with a high free-energy change; it must not exhibit excessive electronic conductivity so as to cause effectively an internal short circuit; and it must have a sufficiently high ionic conductivity so that the reaction does not become limited by exhaustion of the supply of reacting ions at the interface. These requirements have been met by the application of charge transfer complexes, the electrical conductivity of which is being investigated concurrently (Ref. 8). Using calcium or magnesium electrodes in conjunction with an inert electrode and poly-N-vinylcarbazole-iodine or poly-2-vinyl-quinoline-iodine as the charge transfer complex, solid-state cells were constructed with voltages of up to 2.5 v per cell and initial current densities of up to 2.3 ma/cm^2. Other polymeric charge transfer complexes under investigation at present, yielded short-circuit current of the order of 25 ma/cm^2 at room temperature.

[1]Department of Physical Chemistry, University of New South Wales, Sydney, Australia.

2. Experimental

The electrolyte was compacted in a hydraulic press under a pressure of 25000 lb/in.^{2}, yielding a pellet of ½-in. D. The metals were formed into disks of the same diameter, and the resulting sandwich was assembled in a Nylon jig under positive spring pressure (Fig. 2).

Voltages and currents were measured with a Hewlett Packard 412A VTVM and a Keithley 610A electrometer.

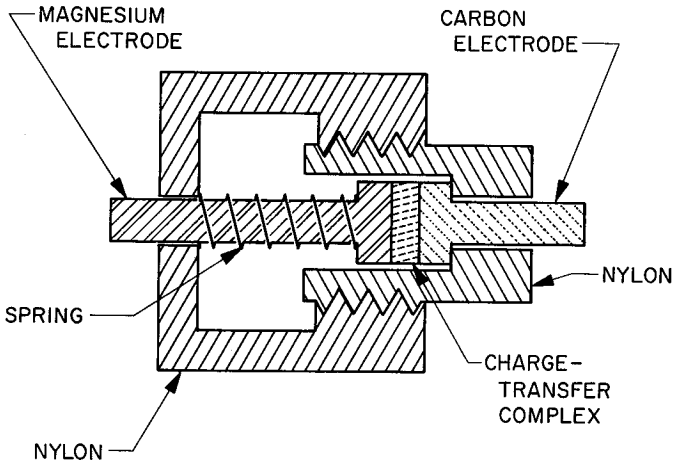


Fig. 2. Test jig

3. Results

a. Metal/pure iodine systems. In order to explore the suitability of anodes other than silver, a number of metals were tested, under standardized conditions, using a compacted pellet of pure iodine as the electrolyte and platinum, graphite, or gold as the counterelectrode, thus forming a simple and well-defined system. Preliminary tests have shown that platinum as well as graphite and gold form good ohmic contacts to iodine. The results are summarized in Table 1.

It is seen that the voltages previously reported for silver (Ref. 7) have been reproduced, but that considerably higher voltages are obtainable for barium, calcium, and magnesium in agreement with the high heat of formation of their iodides (Ref. 9). This quantity is listed in Table 1 rather than the free energy of formation, because of lack of data on the latter. Still higher voltages were observed using calcium or barium with counterelectrodes other than platinum, carbon or gold (Table 2).

The low values of the short-circuit currents obtained for calcium and barium are almost certainly due to the formation of oxide layers. Thus, while these metals offer great promise, it was decided to concentrate further work, at least at this stage, on magnesium anodes. These produce a voltage which, although substantially below that of calcium or barium, is still twice that of silver. Magnesium can be readily obtained, is machinable, and the oxide film forms sufficiently slowly to allow studies of such anodes under open atmospheric conditions.

b. The system magnesium/iodine charge transfer complex/platinum or carbon. A number of such systems were prepared and studied under identical conditions; the results are summarized in Table 3.

Table 1. Data for cells of the type: Metal/I_2/Pt, C, or Au

Metal 1	Open-circuit voltage, v	Metal 2	Heat of formation of the metal-iodide (Ref. 9), kcal-mole^{-1}
Ba	2.25	Au	−144
Ca	2.20	Pt	−127
Mg	1.85	Pt	−86
Al	0.82	Pt	−75
Ag	0.66	Pt,C	−15
Sn	0.23	Pt	
Fe	0.006	Pt	−30
Pt	0	Pt	
Pt	0	C	−21

The metal formed the internal anode in all cases; it connects to the negative pole of a voltmeter.

Table 2. Data for cells of the type: Ba or Ca/I_2/Metal 2

Metal 1 anode	Metal 2	Open-circuit voltage, v	Short-circuit current, μa
Ca	Ni	2.4	
Ca	Cu	2.5	
Ca	Al	1.2	
Ca	Pt	2.2	
Ba	W	2.43	1.62
Ba	Ni	2.4	0.24
Ba	Nb	2.35	2.76
Ba	Cu	2.31	0.72
Ba	Ti	2.32	0.6
Ba	Ta	2.37	0.46
Ba	In	1.85	2.1
Ba	Au	2.24	0.17

Table 3. Data for cells of the type: Mg/I charge transfer complex/Pt or C (electrode area 1.6 cm²)

Electrolyte, I_2 complexed with:[a]	Open-circuit voltage, v	Short-circuit current, μa	Remarks
Poly-N-vinylcarbazole + graphite	1.30	150	No separator
Poly-N-vinylcarbazole + graphite	1.72	400	Polymeric separator
Poly-N-vinylcarbazole + graphite	1.55	1000	Double-thickness polymeric separator
Poly-N-vinylcarbazole	1.45	105	No separator
Perylene	1.5	300	Freshly prepared, no separator
Perylene	1.45	1100	After 3 days' aging
Graphite	1.4	1000	No separator, freshly prepared
Graphite	1.65	2000	After 3 days' aging
Poly-2-vinylquinoline	1.45	500	No separator
Poly-2-vinylquinoline	1.65	500	Polymeric separator
Polypropylene-graphite	1.6	2300	No separator
Poly-N-vinylcarbazole-graphite	1.85	15	Polymeric separator

[a] Approximately 50% I_2 by weight.

The heat of formation of magnesium iodide (MgI_2) predicts a reversible, open-circuit potential of 1.85 v against an inert, ohmic, counterelectrode. It is seen from Table 3 that only the poly-N-vinylcarbazole-iodine-graphite complex, with a crosslinked polypropylene membrane as a separator between the electrolyte and the active electrode, yielded this voltage. In all other cases the measured value was below the calculated value.

According to the Gibbs-Helmholtz equation:

$$E = \frac{\Delta H}{zF} + T\frac{dE}{dT} \qquad (1)$$

where ΔH is the enthalpy change, viz., −86 kcal/mole, z the number of electrons involved in the reaction, F the Faraday, E the open-circuit voltage, and T the absolute temperature.

For the MgI_2 reaction, described in Eq. (3), $z = 2$. If one retains the first term only in Eq. (1) and neglects the entropy contribution, one obtains the Helmholtz-Thomson equation which yields $E = 1.85$ v. Direct measurement of dE/dT showed that it is negligible in the vicinity of room temperature. At temperatures above 45°C a value of about -5×10^{-4} v/deg was obtained. Since ΔH is negative, the voltage tends to drop with increasing temperature. The value of 1.85 v also is very close to the difference in the oxidation-reduction potentials (Ref. 3) for $Mg = Mg^{++} + 2e^-$ and $2I^- = I_2 + 2e^-$, viz., 1.840 v corresponding to the formation of MgI_2.

A value of 1.81 v is obtained from the reverse (charging) current-voltage characteristics shown in Fig. 3. These curves refer to two different electrolytes, both using magnesium anodes and inert platinum counterelectrodes. In both cases it is seen that the relation is linear at higher values of applied voltage of reverse polarity and that extrapolation yields the potentiometric zero current voltage of 1.81 v. The bromine complex behaves in a similar manner.

The effect of the separator in raising the voltage is clearly evident from the table: e.g., removal of the separator causes the voltage to drop from 1.72 to 1.3 v in the case of the same complex as the electrolyte. Incorporation of the separator causes also a substantial rise in the available current. This is due not only to the increased voltage but also to the removal of internal short circuiting. Aging has a similar effect, though it is less controllable and less

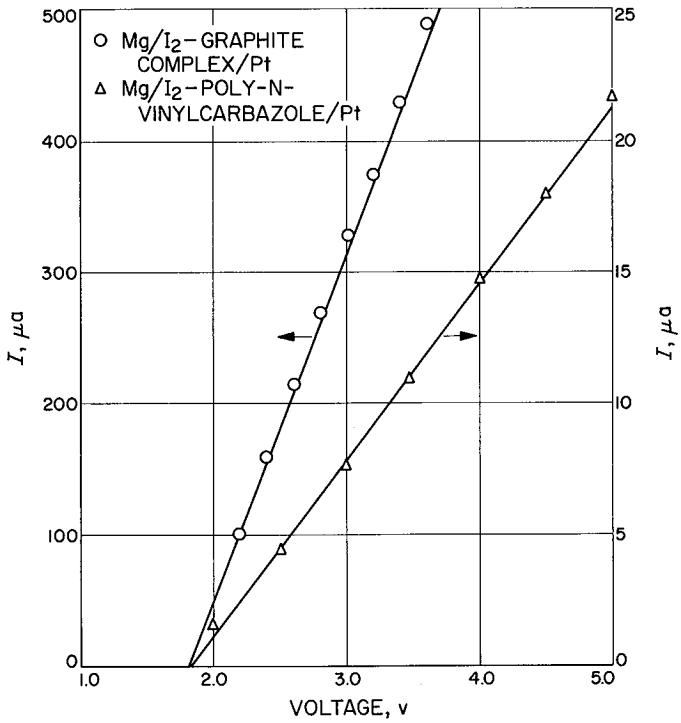


Fig. 3. Reverse (charging) characteristics

reproducible. Thus, three days' aging raises the available short-circuit current in the case of the graphite-iodine complex from 300 to 1100 μa. However, the table shows that even without a separator, a voltage of 1.6 v and a short-circuit current of 2.3 ma were obtained with an iodine-polypropylene-graphite electrolyte. The graphite-free complex polyvinyl-quinoline-iodine is also of interest. Its performance, with a separator, is not very much below that of the best complexes containing graphite.

In a series of performance tests on the magnesium/iodine-graphite-platinum system the following data were obtained:

Mean watts: 4×10^{-4} w/cm^2 electrode area

Voltage drop rate: 250 μv/min

Useful lifetime: 23 hr

Useful available charge: 7 coulombs/cm^3

Useful total energy supplied: 9×10^{-3} w/hr

Useful energy density: 8 w-hr/lb

Total charge available if discharge is continued until zero volts across load: 84 coulombs/cm^3

Open-circuit volts: 1.65 v

Short-circuit current: 2 ma/cm^2

Instantaneous efficiency of energy conversion: 78%

These values may be regarded as fairly typical. They are the result of a series of tests in which decay curves were obtained for the current I as a function of time, the cell operating under matched load conditions. The cell was loaded until its terminal voltage had dropped to one-half of its initial open-circuit value, and the voltage across the known load resistance was then applied to a recorder. It may be recalled that the cell delivers maximum useful power to an external load if both the cell internal resistance and the load resistance are matched to equal value. In taking the decay curves, the discharge was assumed to be terminated once the current had dropped to one-half of its initial value. This defines the useful lifetime. The system recovers within minutes, even after prolonged short-circuiting.

4. Reaction Mechanism

Since the open-circuit voltages of a given metal are virtually the same with pure iodine and with iodine complexes as the electrolyte, it appears that the energy-producing reaction is the formation of the metal-iodide from the elements. This is further supported by the excellent agreement between the values of the open-circuit voltage and the energy of formation (Table 1). Two electrons are thus involved in the overall reaction, and this requires that the rate determining reaction must be carried through twice for the overall reaction to advance by one unit, i.e., for the formation of 1 mole of MgI_2. Several reaction mechanisms may be devised to agree with these requirements: thus, e.g., the following mechanism is compatible with $z = 2$

$$\left.\begin{array}{l} 2\,[Mg + I^- \rightarrow MgI + e^-] \\ \underline{2\,MgI \rightleftharpoons MgI_2 + Mg} \\ Mg + 2I^- \rightarrow MgI_2 + 2e^- \end{array}\right\} \qquad (2)$$

The first step is the rate-determining one, while the second step involves a reaction at or near equilibrium. There must be, linked with the above reaction scheme, a further reaction as a source of I^- ions; this could well be

$$Mg + I_2 \rightleftharpoons Mg^{++} + 2\,I^- \qquad (3)$$

5. Conclusions

Examination of Tables 3 and 4 shows that by the use of monomeric or polymeric charge transfer complexes, the short-circuit current could be increased by three orders of magnitude. The solid-state batteries described above may be encapsulated and made as small as desired. They

Table 4. Data for cells of the type: Metal 1/solid electrolyte/metal 2 (not involving Mg), electrode area 1.6 cm^2

Metal 1	Metal 2	Electrolyte[a]	Open-circuit voltage, v	Short-circuit current, μa	Remarks
In	Ni	I_2-perylene complex	0.05	820	Liquefaction
Ba	In	I_2-perylene complex	2.4	10	BaO present
Cd	Pt	I_2-perylene complex	0.008	2	
Ba	Ni	I_2-perylene complex	2.4	0.24	BaO present
Ca	Ni	I_2-perylene complex	2.5	5000	Not well reproducible, current drops rapidly

[a] Approximately 50% I_2 by weight.

constitute promising power sources for energizing microelectronic circuits. They also lend themselves for the powering of firing mechanisms for pyrotechnic devices, e.g., squibs as well as for electromedical applications. Although the energy density of the prototypes thus far studied is below the requirements for space vehicle power, there is little doubt that proper engineering and further investigations will lead to a substantial improvement of present power capability and performance.

C. The Effect of Thermodynamic Interactions on the Viscosimetric Behavior of Low Molecular Weight Poly(Propylene Oxides)

J. Moacanin

1. Introduction

The principles underlying the proportionality between the intrinsic viscosity $[\eta]$ and a power of molecular weight M^a have been explored extensively (Ref. 10). The value of the exponent a for free-draining polymer coils is unity (Staudinger's rule), and approaches 0.5 with increasing shielding to solvent flow within the coil. The theories which lead to these results assume, however, that the molar value of the polymer is much larger than that of the solvent, so that it may be treated as a colloidal particle suspended in a continuous incompressible fluid. But when the size of the polymer becomes comparable to that of the solvent molecule, this assumption must break down. This may be the reason for $[\eta]$ values smaller that the Einstein value of 0.025 v_{sp}, which were reported for some low molecular weight polymers (Ref. 11). At the extreme of nearly equal sizes, equations which have been derived for the viscosities of liquid mixtures should be appropriate.

In order to examine the transition from the behavior of liquid mixtures to the hydrodynamic behavior of polymer solutions, a study was made of the viscosimetric behavior of poly(propylene oxides), PPO, ranging in molecular weight from about 150 up to several thousand. This system turns out to be convenient for such a study because of the relative ease of covering a wide molecular weight range. In addition, the widely different polarities of the backbone units and the hydroxyl chain ends emphasize end-group effect. Dilute solution viscosities in both poor and good solvents and melt viscosities have been determined previously (*Research Summary 36-10*). The data for intermediate concentrations are results of recent work.

The purpose of this report is to propose a theoretical interpretation for the viscosimetric behavior of this system in terms of polymer-solvent thermodynamic interactions. This approach may provide a novel method for the interpretation and prediction of the viscosimetric behavior of low molecular weight polymers.

2. Results

a. Dilute solution viscosities. Plots on a linear scale of $[\eta]$ versus M for PPO's ranging in mol wt between 120 (dipropylene glycol DPG) and 2000 for five different solvents are given in Fig. 4. The parameters for the straight lines representative of the experimental points for each of the solvents, are given in Table 5. Figs. 5 and 6 show double-log plots of $[\eta]$ versus M for methanol and benzene, respectively, and include the high molecular weight range. Table 6 gives the Mark-Houwink parameters.

For DPG in methanol and benzene $[\eta]$ values are 0.0212 and 0.0125, respectively. For the DPG monomethyl ether the respective values are 0.0128 and 0.0030. Partial specific volume data are summarized in Table 7.

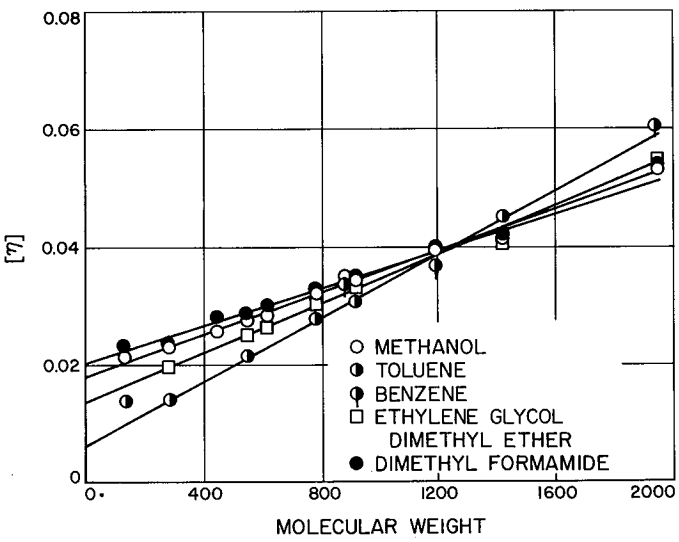


Fig. 4. Intrinsic viscosity–molecular weight relationship in five solvents at 250°C

Table 5. Intrinsic viscosity parameters for low molecular weight poly-glycols[a] at 25°C

Solvent	A	$B \times 10^3$
Dimethyl formamide	0.0201	0.0158
Methanol	0.0180	0.0175
Ethylene glycol dimethyl ether	0.0133	0.0198
Benzene	0.0062	0.0272
Toluene		

[a] Parameters in the modified Staudinger equation: $[\eta] = A + BM$

Table 6. Intrinsic viscosity parameters for high molecular weight poly(propylene oxides) at 25°C[a]

Solvent	$K' \times 10^4$	a
Methanol	7.69	0.55
Benzene	4.13	0.64

[a] Parameters in the Mark-Houwink equation: $[\eta] = K'M^a$

Table 7. Apparent specific volumes of poly(propylene oxides)

Polymer	Solvent	Concentration, g/100 g	V_2
PPG2025	Benzene	7.99	0.995
		4.01	0.998
		2.02	0.995
	Methanol	8.00	0.968
		4.02	0.959
		2.01	0.983
P400	Benzene	8.00	0.999
		4.00	1.007
		2.02	1.011
	Methanol	8.02	0.956
		4.01	0.959
		2.01	0.998
DPG	Benzene	8.02	0.994
		4.02	1.003
		2.01	0.999
	Methanol	8.02	0.949
		4.02	0.944
		2.04	0.943

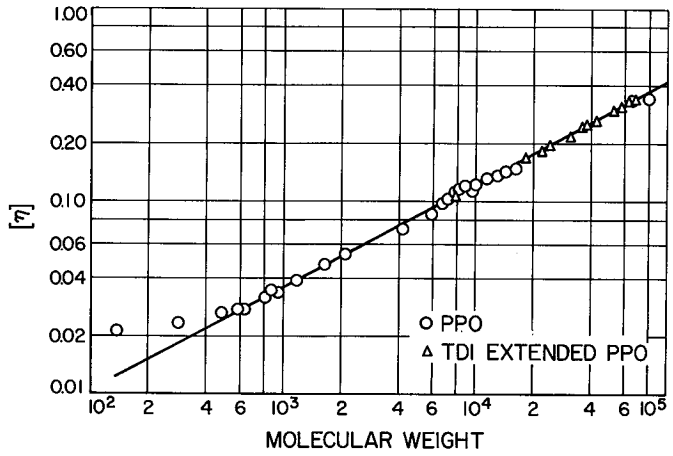


Fig. 5. Intrinsic viscosity–molecular weight relationship in methanol at 25.0°C

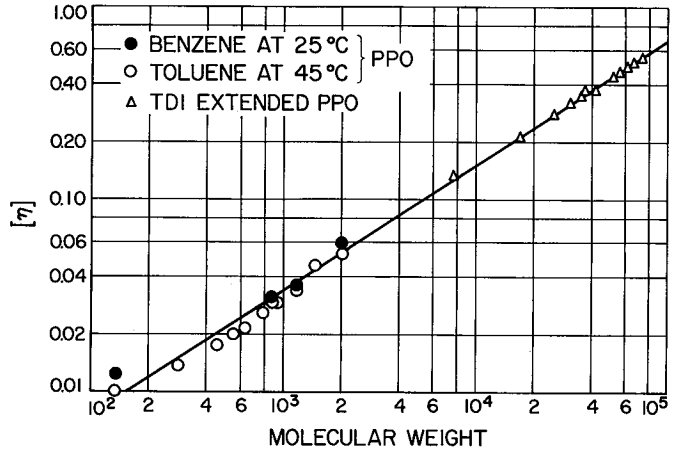


Fig. 6. Intrinsic viscosity–molecular weight relationship in benzene and toluene

b. Concentration dependence of viscosities. Following the procedure used by Utracki and Simha (Ref. 12) we show plots on a double logarithmic scale of $\eta_{sp}/c[\eta]$ versus c for PPO solutions at 25°C in methanol (Fig. 7) and benzene (Fig. 8), a poor and good solvent, respectively. For each solvent the curves were then superposed by shifting along the abcissa (Figs. 9 and 10), the shift factor γ being arbitrarily taken to be unity for the PPO with mol wt 10,000.

Inspection of the reduced curves shows that the superposition principle is obeyed up to a concentration range of $c/\gamma = 10$. Above this concentration the curves for both solvents fork into two groups, one for the high (10,000

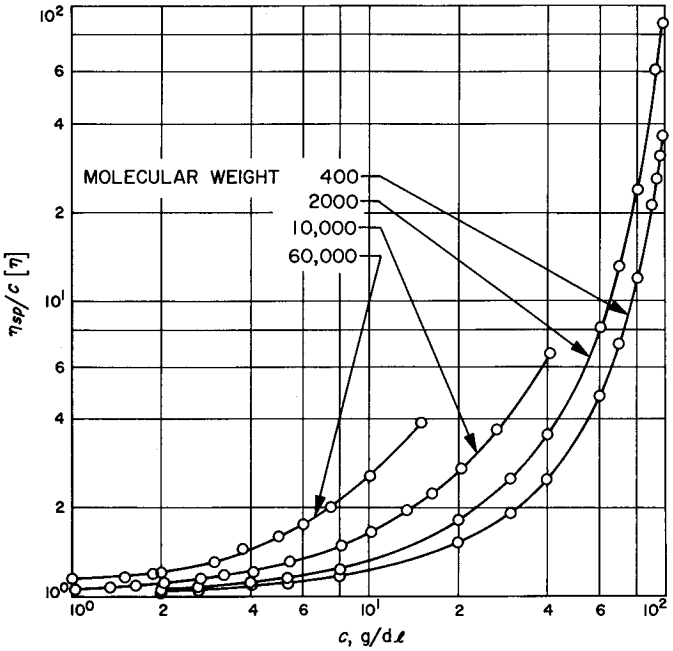


Fig. 7. $\eta_{sp}/c[\eta]$ versus c in methanol at 25°C

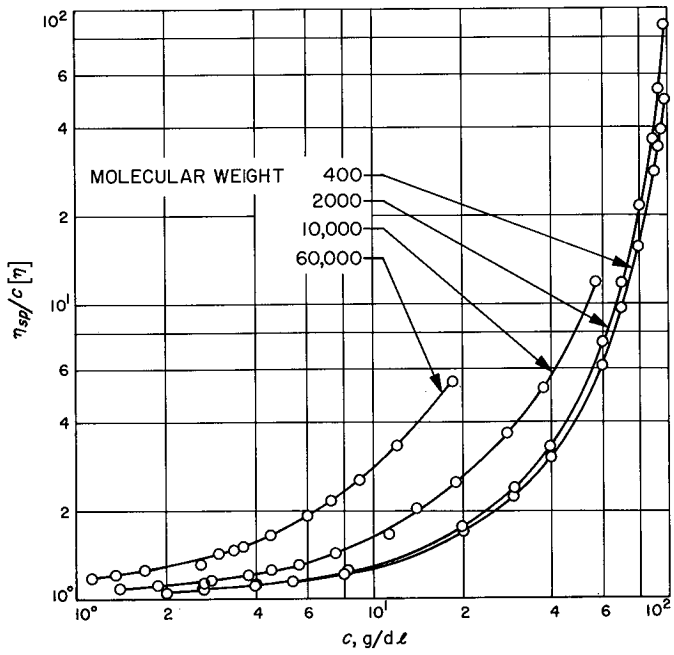


Fig. 8. $\eta_{sp}/c[\eta]$ versus c in benzene at 25°C

and 60,000) and the other for low mol wt (400 and 2000) PPO's. For both solvents the molecular weight dependence of γ is given approximately by

$$\gamma \propto M^{-0.46}$$

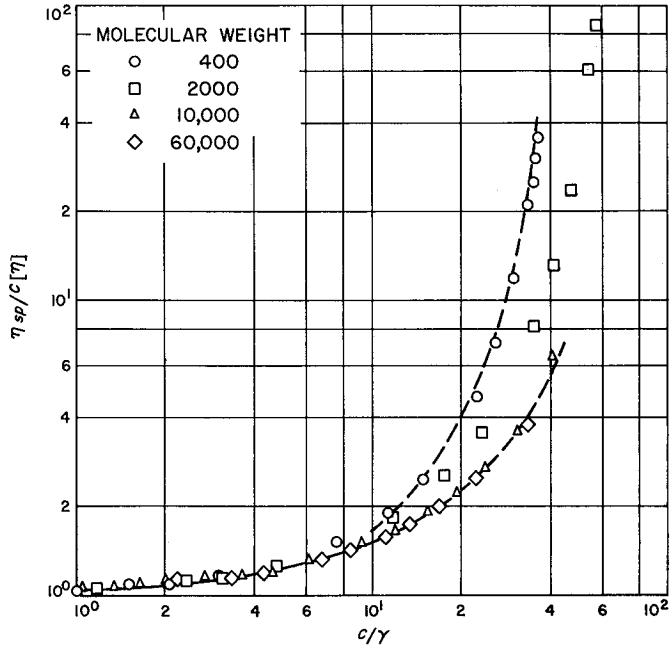


Fig. 9. Master plot of viscosity data in methanol at 25.0°C

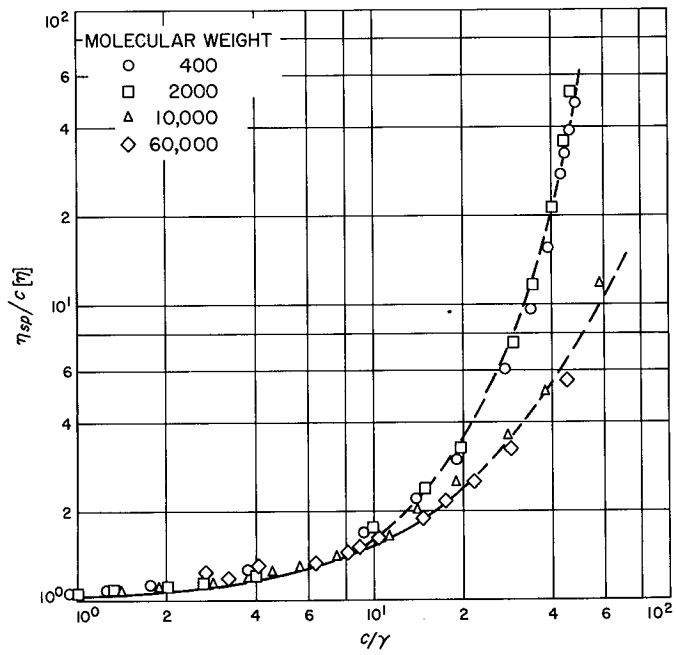


Fig. 10. Master plot of viscosity data in benzene at 25.0°C

The available results are insufficient to define more closely the relationships for each solvent. However, there appears to be a crossover in the relative magnitude of γ between molecular weights of 400 and 2000 (Fig. 11).

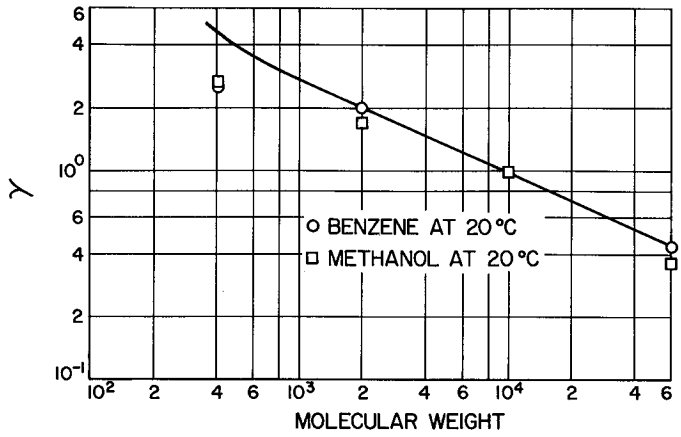


Fig. 11. Molecular weight dependence of the shift factor γ

c. Melt viscosity. Melt viscosities ν were determined on PPO with molecular weights up to 20,000. A log-log plot of ν vs M is shown in Fig. 12. A straight line of unit slope fits the data well for molecular weights between about 700 and 8000. The increase in slope above 8000 suggests this to be the critical molecular weight M_c, above which molecular entanglement becomes effective.

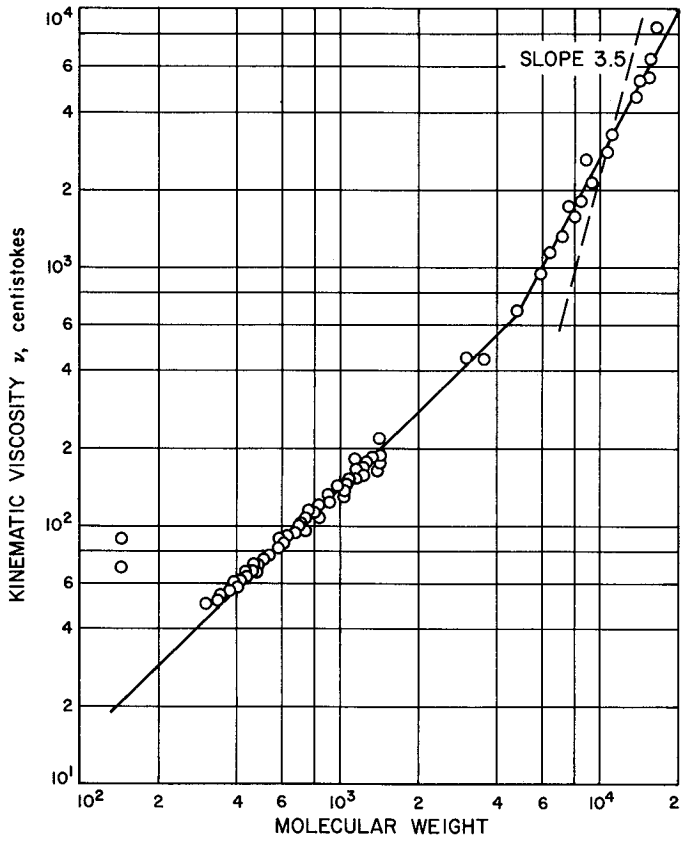


Fig. 12. Bulk viscosity–molecular weight relationship at 25.0°C

d. Discussion. Inspection of Fig. 4 suggest that $[\eta]$ data for M 2000 can be represented by the modified Staudinger's relation, viz.:

$$[\eta] = A + BM \tag{1}$$

It does not appear promising, however, to attempt to interpret these results in terms of the usual hydrodynamic theories, since for the lowest molecular weights the polymer size approaches the size of the solvent. Also, for several values $[\eta] < 0.025$ (see Introduction). Moreover, as the molecular weight decreases, the effect of end-groups must be accounted for. Motivated by the above considerations we take recourse to an equation proposed by Simha (Ref. 13), who reasoned that in view of the fact that the second virial coefficient A_2 is determined by the covolume of the solute molecule, and since $[\eta]$ also depends on this volume, one may combine certain hydrodynamic relations and theoretical formulas for A_2 to yield:

$$r\,[\eta] = 2A_2M = (1 - 2\chi)\,(v_{sp}/100)\,(V_2/V_1) \tag{2}$$

Here χ is the solvent-polymer interaction constant, r is a friction factor of the order of unity, and V_2 and V_1 are molar volumes of polymer and solvent, respectively. Similar equations can be derived from a modified approach to the viscosity of mixtures (Ref. 14).

The above equation leads to Standinger's relation if both χ and the polymer specific volume are independent of molecular weight. We find, however, that for PPO in methanol (Fig. 13) χ varies with molecular weight (Ref. 15). The low value for χ at the low molecular weight side indicates strong interactions of the solvent with hydroxyl end groups. There is a rapid rise of χ with molecular weight, and an asymptotic value of about 0.46 (for the lowest concentration) is approached at about 3,000. A fit of the molecular weight dependence can be obtained by means of the following equation:

$$\chi_M = \chi_\infty \quad (1 - Q/M) \tag{3}$$

where for the PPO-methanol system, $Q = 104$ and $\chi_\infty = 0.46$.

If we neglect variations in v_{sp} which are of the order of 2% (Table 7), we can write $V_2 = v_{sp}\,M$, which along

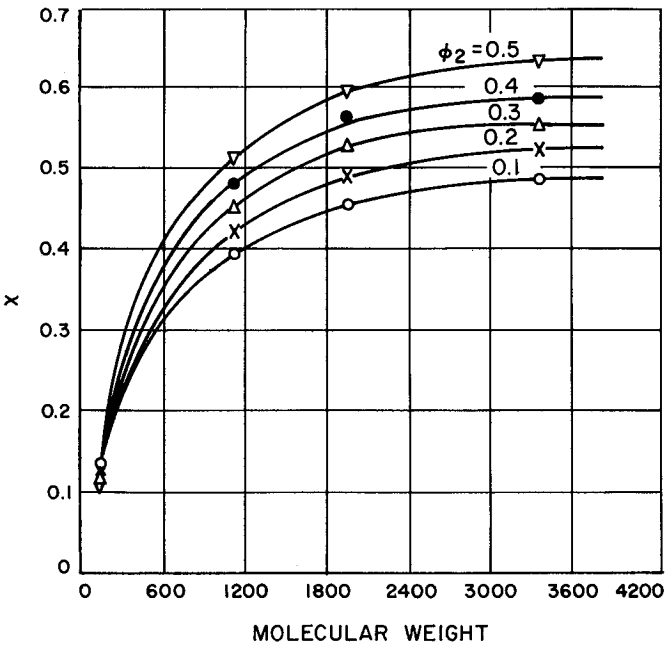


Fig. 13. Values of χ for poly(propylene oxides) at five solution concentrations

with the substitution of the expression for χ_M in Eq. (2), leads to:

$$r\,[\eta] = 2\,\chi_\infty\,v_{sp}^2\,(Q/100\,V_1) + (1 - 2\,\chi_\infty)\,v_{sp}^2/(100\,V_1)\,M \tag{4}$$

This expression gives explicity the constants A and B in Eq. (1). For the PPO-methanol system Eq. (4) yields:

$$r\,[\eta] = 0.023 + (1.82)\,(10^{-5})\,M$$

With the reasonable value $r = 1$, this result is brought into excellent agreement with the empirical straight line shown in Fig. 4:

$$[\eta] = 0.018 + (1.75)\,(10^{-5})\,M$$

It would be desirable to carry out the same calculations for the other solvents, but unfortunately the necessary thermodynamic data are not available. For benzene, however, a partial comparison can be made by using a previously published value for the second virial coefficient: $A_2 \sim 1.5 \times 10^{-3}$ (Ref. 16). Although, determined for a PPO-TDI polymer, this value should be sufficiently close to that for the pure polyether in view of the virtual identity of the $[\eta]$ versus M relationships for the pure and the TDI-extended polyether, respectively (Fig. 6). Calculations yield $\chi_\infty = 0.37$, which by substitution in Eq. (4) gives $B = 0.028$, as compared to the observed value of 0.027; $Q = 72$ would reproduce the observed straight line.

Sadron and Rempp used a semilog plot to represent their results on poly(ethylene oxides), PEO. We find that a good fit is also obtained for the low molecular weight PEO using a linear plot, as shown in Fig. 14. In contrast to PPO, the absence of crossover for the lines for the various solvents is apparent. This can be understood qualitatively by noting that owing to the polarity of the PEO ether oxygens, the relative magnitude of the hydroxyl end groups is minimized. In the PPO, however, the ether polarity is reduced by the presence of the methyl side chains.

Some measure of the relative magnitude of the end-group effect can be obtained by considering $[\eta]$ data of the PPO dimer and its monomethyl ether in methanol and benzene. For the purpose of this comparison we use the following equation for the viscosity of the liquid mixtures (Ref. 17):

$$\ln \eta = X_1^2 \ln \eta_1 + X_2^2 \ln \eta_2 + 2\,X_1\,X_2\,\eta_{12} \tag{5}$$

where η_{12} is the "mutual viscosity."

For very dilute solutions $X_1^2 = 1 - 2\,X_2$, and the mole fraction is given by $X_2 = (c/100\,\rho)\,(M_1/M_2)$; c is the concentration in g/dl, and ρ the solvent density. Substituting in Eq. (5) and rearranging, one arrives at:

$$\frac{\ln(\eta/\eta)}{c} = \frac{2}{100\,\rho}\,\frac{M_1}{M_2}\ln\frac{\eta_{12}}{\eta_1} + \left(\frac{1}{100\,\rho}\,\frac{M_1}{M_2}\right)^2 \ln\frac{\eta_1\,\eta_2}{\eta_{12}^2}\,c \tag{6}$$

clearly,

$$[\eta] = \frac{2}{100\,\rho}\,\frac{M_1}{M_2}\ln\frac{\eta_{12}}{\eta_1} \tag{7}$$

The results of the calculations are given in Table 8. Inasmuch as η_{12} is related to the solvent-solute thermodynamic interactions, the drastic decrease in their values in going from the di- to the mono-hydroxyl compound provides a measure of the end-group contribution. Also, for both compounds the methanol values are about ten times larger than those for benzene, indicating that methanol is a better solvent than benzene. This, of course, is the opposite of the situation for the high molecular weight polyethers which are characterized by Mark-Houwink exponents of 0.55 and 0.64 for methanol and benzene, respectively.

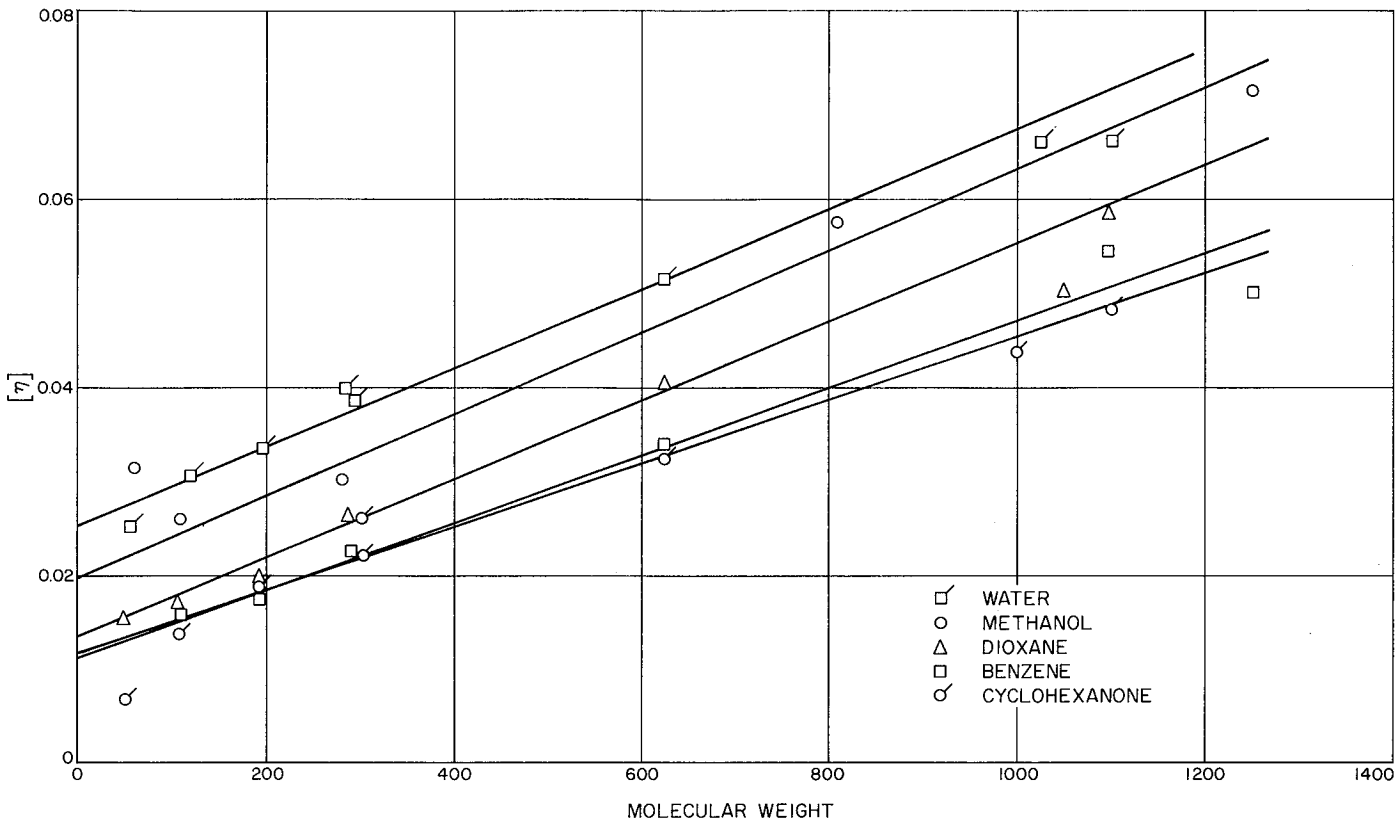


Fig. 14. Intrinsic viscosity–molecular weight relationship for poly(ethylene oxides) in five solvents at 25.0°C

Table 8. Mutual viscosities at 25°C

Polymer	M	η_2, centistokes	ρ_2, g/ml	η_{12}/η_1[a]	
				Methanol	Benzene
DPG	134	70	1.02	32.3	2.57
DPG ether	148			10.3	1.28

[a] Calculated from Eq. (7):

For methanol $\rho = 0.7868$, $\eta = 0.5445$, mol wt = 32

For benzene $\rho = 0.8738$, $\eta = 0.6028$, mol wt = 78

The results on the viscosities in dilute solutions point to the end-group effect, and are consistent with the results of thermodynamic measurements (Ref. 15). Evidence for hydrogen bonding between methanol and PPO was given recently by Heil and Prausnitz (Ref. 18). Using infrared spectroscopy these authors were able to show that the equilibrium constants for hydrogen bonding between methanol and the hydroxyl-terminated polyethers are markedly larger than those for ether-terminated species of comparable molecular weight. The disparity, however, decreases with increasing molecular weight, i.e., the end-group contribution gets diluted.

We can now proceed to show that the behavior observed in more concentrated solutions is consistent with the ideas we have developed for the dilute solutions. First, the crossover in the relationship between the shift factor and molecular weight for the two solvents parallels the reversal with increase in molecular weight in the solvent goodness for benzene and methanol. While for dilute solutions hydrogen bonding appears to be involved mainly in polymer-solvent interaction, for higher concentrations one ought to expect increasing polymer-polymer interactions. This may explain the failure of the superposition principle for the concentration range $c/\gamma > 10$, a range where for other systems the principle is still obeyed (Ref. 12). Again, for low molecular weight polymers, i.e., relatively high hydroxyl concentration, the viscosity increase on the reduced scale is faster than that for high molecular weights.

Finally, inspection of the melt viscosity dependence on chainlength shows for higher molecular weights the usual two regions: direct proportionality to M for molecular weights below a critical value M_c; for PPO $M_c = 8{,}000$. For higher molecular weights, where molecular entanglements presumably become important, the exponent tends to approach about 3.5. But, again, going below the molecular weight of 1000, the viscosity dependence is not linear on the log scale; in fact, the viscosity of DPG is of the order of magnitude of that for the polyether of mol wt 600.

D. The Reaction of Carboxyl- and Amino-Terminated Prepolymers with Polyaziridines

S. H. Kalfayan and B. A. Campbell

Previously (Ref. 19), studies on the reaction of carboxyl-terminated poly(alkylene oxides) with two triepoxides and a triaziridine were presented. The study of the reaction between polyaziridines and carboxylic prepolymers are now extended to include carboxyl-ending poly(butadienes). For comparison, two amino-ending prepolymers were also reacted with the same aziridines. Experiments on the catalysis of the carboxyl-aziridine reaction were carried out, using one of the carboxyl-terminated poly(butadienes), Telagen CT, and one of the diaziridines, HX 740, with various acidic and basic catalysts. On heating, the polyaziridines polymerized to hard resins. Results of these experiments, as well as the results obtained after the thermal sterilization of some of the products obtained in this study, are also reported.

1. Experimental

a. Materials. The following prepolymers were used:

HC 434. Carboxyl-ending poly(butadiene) eq.wt. 1968, Thiokol Chemical Corp.

Telagen CT. Carboxyl-ending poly(butadiene) eq.wt. 3030, General Tire and Rubber Corp.

Butarez CTL. Carboxyl-ending poly(butadiene) eq.wt. 2616, Phillips Petroleum Co.

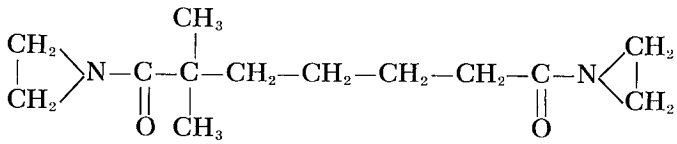
PPO-COOH. Carboxyl-ending poly(propylene oxide) eq.wt. 585 (Ref. 19)

L-3. Amine-ending, difunctional poly(propylene oxide) eq.wt. 632, Union Carbide Co.

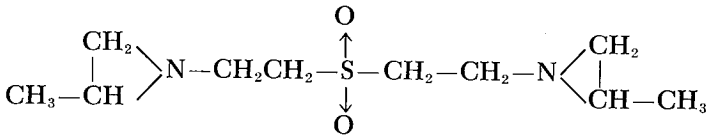
L-11. Amine-ending trifunctional poly(propylene oxide) eq.wt. 1250, Union Carbide Co.

The aziridines used were:

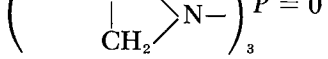
HX 740. 1,1′-(2,2-dimethylheptanedioyl)bis-aziridine, Thiokol Chemical Corp.



ZC 466. Bis(2-methylaziridinylethyl)sulfone, (American Cyanamid Co.)



MAPO. Tris-(2-methylaziridinyl)phosphine oxide, Interchemical Co.



All materials were used as received.

b. Determination of gel times. Gelation was determined by mixing stoichiometric amounts of reactants in glass test tubes and placing them in constant temperature baths after stoppering. Gel time was measured from the instant the tube was placed in the bath, until that time when no flowing of the reaction mixture was observable in the inverted test tube.

For hardness measurements, samples were cured in aluminum dishes, sprayed with Teflon mold-release agent and placed in constant-temperature air ovens.

2. Results and Discussion

a. The carboxyl-aziridine reaction. Table 9 gives gel times and Shore A hardnesses for cured carboxyl- and amino-ending prepolymers. If gel time is considered as a measure of the rates of these reactions, then there is no significant difference between the reactivities of the two diaziridines, ZC 466 and HX 740, with the carboxyl-ending poly(butadienes). MAPO, however, reacts much more slowly with these prepolymers (Experiments 3, 6, 9). The order of reactivity of the polyaziridines is found to be different with the carboxyl-ending poly(propylene oxide). Here HX 740 and MAPO showed the same reactivity (Experiments 10, 12), and ZC 466 reacted more slowly (Experiment 11).

The carboxyl-ending poly(butadienes) showed the following order of reactivity with any one of the polyaziridines tested:

Butarez CTL > Telagen CT > HC 434.

Compared with the carboxyl-ending poly(butadienes) the PPO-COOH reacted more slowly with the diaziridines (Experiments 10, 11) but somewhat faster with MAPO (Experiment 12). PPO-COOH products, in general, were too soft to be measured on the Shore A scale.

Table 9 shows that ZC 466 gave harder products than either HX 740 or MAPO with the carboxyl-ending poly(butadienes). Butarez CTL cured with MAPO is an exception (Experiment 9).

b. The amino-aziridine reaction. In one case only did the polyetheramines show gelling. This was the product obtained from the trifunctional polyetheramine L-11 and HX 740 (Experiment 13). In all other cases gelling failed to occur after 48 hr at 100°C, but the mixtures thickened, indicating an interaction between the amino-ending prepolymer and the aziridines.

c. Polymerization of the aziridines. It is a well-known fact that aziridines polymerize when heated, especially in contact with acidic reagents (Refs. 20 to 22).

The possibility of the polymerization of the aziridine as a side reaction during the carboxyl-aziridine interaction should not be overlooked. Under a set of conditions the relative rates of the two reactions will determine

Table 9. Reaction (curing) of carboxyl- and amine-terminated prepolymers with polyaziridines

Experiment No.	Prepolymer	Curing agent	Gel time at 100°C, min	Shore A hardness	
				After 18 hr at 100°C	After 48 hr at 100°C
1	HC 434	HX 740	29	5	5
2		ZC 466	29	41	43
3		MAPO	230	10	8
4	Telagen CT	HX 740	22	4	4
5		ZC 466	21	20	21
6		MAPO	180	9	12
7	Butarez CTL	HX 740	12	21	23
8		ZC 466	14	31	37
9		MAPO	170	24	37
10	PPO-COOH	HX 740	144	soft[a]	35
11		ZC 466	375	soft	soft
12		MAPO	150	soft	soft
13	Polyether-triamine (L-11)	HX 740	47	12	20
14		ZC 466	—	No gelling	No gelling
15		MAPO	—	No gelling	No gelling
16	Polyether-diamine (L-3)	HX 740	—	No gelling	No gelling
17		ZC 466	—	No gelling	No gelling
18		MAPO	—	No gelling	No gelling

[a] Not measurable on the Shore A scale.

Table 10. Polymerization of aziridines

Experiment No.	Aziridine	Catalyst	Temperature, °C	Gel time, min	Shore D hardness	
					After 18 hr	After 48 hr
1	HX 740	None	100	>480	81	85
2		None	150	15	—	—
3		HCl	100	180	83	86
4		CH_3COOH	100	300	81	86
5		HCOOH	100	>480	80	84
6	ZC 466	None	100	—	No gelling	No gelling
7		None	150	30	—	—
8		HCl	100	>480	60	85
9		CH_3COOH	100	—	No gelling	No gelling
10		HCOOH	100	—	No gelling	No gelling
11	MAPO	None	100	—	No gelling	No gelling
12		None	150	—	No gelling	No gelling
13		HCl	100	—	No gelling	No gelling
14		CH_3COOH	100	—	No gelling	No gelling
15		HCOOH	100	—	No gelling	No gelling
16		HCl	150	105		

the overall course. The data in Table 10 show that in the absence of a catalyst, HX gelled in more than 6 hr at 100°C, and ZC 466 and MAPO failed to gel even after 48 hr at 100°C. On the other hand, gelling of carboxyl-ending prepolymers with any of the aziridines took place in less than 4 hr at 100°C (Table 9). This is an indication that the rates of the uncatalyzed polymerization of the aziridines are lower than that of the carboxyl-aziridine reaction.

In the presence of catalytic amounts of the strong acid, HCl, the polymerization of the two diaziridines at 100°C took place easily (Experiments 3, 8, Table 10). Even when catalyzed, MAPO did not polymerize readily (Experiments 11, 12 and 16). The data for the aziridines indicate the following order of readiness to polymerize:

$$\text{HX 740} > \text{ZC 466} \gg \text{MAPO}$$

d. Catalysis of the carboxyl-aziridine reaction. Table 11 summarizes the test results for the catalysis of the carboxyl-aziridine reaction. Telagen CT and HX 740 were the representative reactants. Both acidic and basic catalysts were used.

In the absence of any catalyst (Experiment 9) gelation took place in 65 min at 75°C. With the exception of the strong nucleophile, the thiosulfate ion $S_2O_3^=$ (Experiment 5) and of stannous octoate (Experiment 3), none of the other reagents showed catalytic activity. The instantaneous gelation with tetrabutyl titanate (TBT) should not be misleading. This organometallic gave the same consistency of gel with Telagen CT alone. The gelling, therefore, is due to some interaction of TBT with the

Table 11. Catalyst tests for the carboxyl-aziridine reaction (Telagen CT with HX 740)

Experiment No.	Catalyst	Gel time at 75°C, min	Comments
1	Hydrochloric acid (38%)	>480	Fluid after 36 hr at 75°C and 3 days at RT
2	Acetic acid, glacial	>480	Fluid after 36 hr at 75°C and 3 days at RT
3	Stannous octoate	60	Soft gel after 36 hr at 75°C and 3 days at RT
4	Benzyl dimethyl amine	>480	Soft gel after 36 hr at 75°C and 3 days at RT
5	Sodium thiosulfate	35	Soft gel after 36 hr at 75°C and 3 days at RT
6	Tetramethyl ammonium hydroxide	>480	Fluid after 36 hr at 75°C and 3 days at RT
7	Tributyl tin oxide	120	Soft gel after 36 hr at 75°C and 3 days at RT
8	Tetrabutyl titanate	0.1	Instantaneous gelling. Very soft after 48 hr
9	None	65	Soft gel after 36 hr at 75°C and 3 days at RT

RT = room temperature.

unsaturated carboxyl-ending prepolymer, rather than to the catalysis of the carboxyl-aziridine reaction. TBT showed no gelling action on HX 740.

Results in Table 11 also show that the stronger acids and the stronger bases (Experiments 1, 2, 4, 6) retard the carboxyl-aziridine reaction.

e. Thermal sterilization of some of the products. The carboxyl-ending poly(butadienes) cured with all three aziridines of this study, were exposed to thermal sterilization conditions consisting of 3 cycles of 36 hr each at 145°C in dry nitrogen. Table 12 summarizes the results obtained. All samples had been cured 48 hr at 100°C prior to thermal exposure at 145°C. This thermal pretreatment may have been the reason for the relatively low percent weight losses encountered.

Table 12. Effect of thermal sterilization on carboxyl-terminated butadienes cured with aziridines

Experiment No.	Prepolymer	Curing agent	Weight loss after thermal exposure, %	Shore A hardness after thermal exposure[a]	
				Surface	Interior
1	HC 434	HX 740	0.045	15	15
2		ZC 466	0.365	20	13
3		MAPO	0.356	11	8
4	Telagen Ct	HX 740	0.083	12	5
5		ZC 466	0.334	11	5
6		MAPO	0.246	21	14
7	Butarez CTL	HX 740	0.129	34	21
8		ZC 466	0.724	22	11
9		MAPO	0.268	42	37

[a] After 3 cycles of 36 hr each at 145°C in dry nitrogen.

Considering the Shore hardness values obtained after thermal sterilization (Table 12) and comparing them with the hardness values presented in Table 9, the following conclusions can be made:

(1) In all thermally cycled products, except one (Experiment 1) the exposed surface of the polymer was harder than the unexposed interior.

(2) All carboxyl-ending poly(butadienes) cured with ZC 466 became softer after thermal sterilization. This was true for both the exposed surface and the unexposed interior of the products.

(3) Cures with HX 740 and MAPO showed substantially no change in hardness when the readings after the 48-hr cure at 100°C were compared with the readings on the interior of the thermally sterilized elastomers. Experiment 1 was again an exception. Thermal cycling had hardened the unexposed sections of the polymer.

The substantial loss in hardness of the ZC 466 cured prepolymers after thermal cycling, is an indication that it may not be an adequate curing agent to obtain sterilizable products.

3. Future Work

Studies on the reactions of carboxyl-terminated prepolymers with aziridines and other curing agents will be continued.

E. Outgassing Rates from Plastic-Coated Foam

E. F. Cuddihy and J. Moacanin

The diffusion properties of rigid closed-cell foams have been studied and equations were derived for predicting outgassing rates (*SPS 37-34, 37-35, 37-36,* and *37-37,* all Vol. IV). These equations are useful in establishing criteria for selecting foam systems having low outgassing rates which will then ensure the foam retaining the highest possible dielectric strength (*SPS 37-39,* Vol. IV) during exposure to vacuum.

A significant reduction in outgassing from a foam can be achieved by means of a thin plastic coating. This concept was tested by comparing the weight loss of CO_2 from a coated and an uncoated 4.5-lb/ft^3 Eccofoam FPH polyurethane foam. A polyester coating (Selectron YC 5119, Pittsburgh Plate Glass Co.) 0.055 in. thick, was formed by dipping the foam specimen into the liquid polymer. The results, shown in Fig. 15, illustrate the dramatic reduction which was achieved in the outgassing rate. After about 420 hr, the uncoated sample has lost 150 out of a possible 194 mg, whereas the coated sample has only lost about 50 mg. The coated sample required 1750 hr to lose about 100 mg of gas compared to about 150 hr for the uncoated sample. This amounts to better

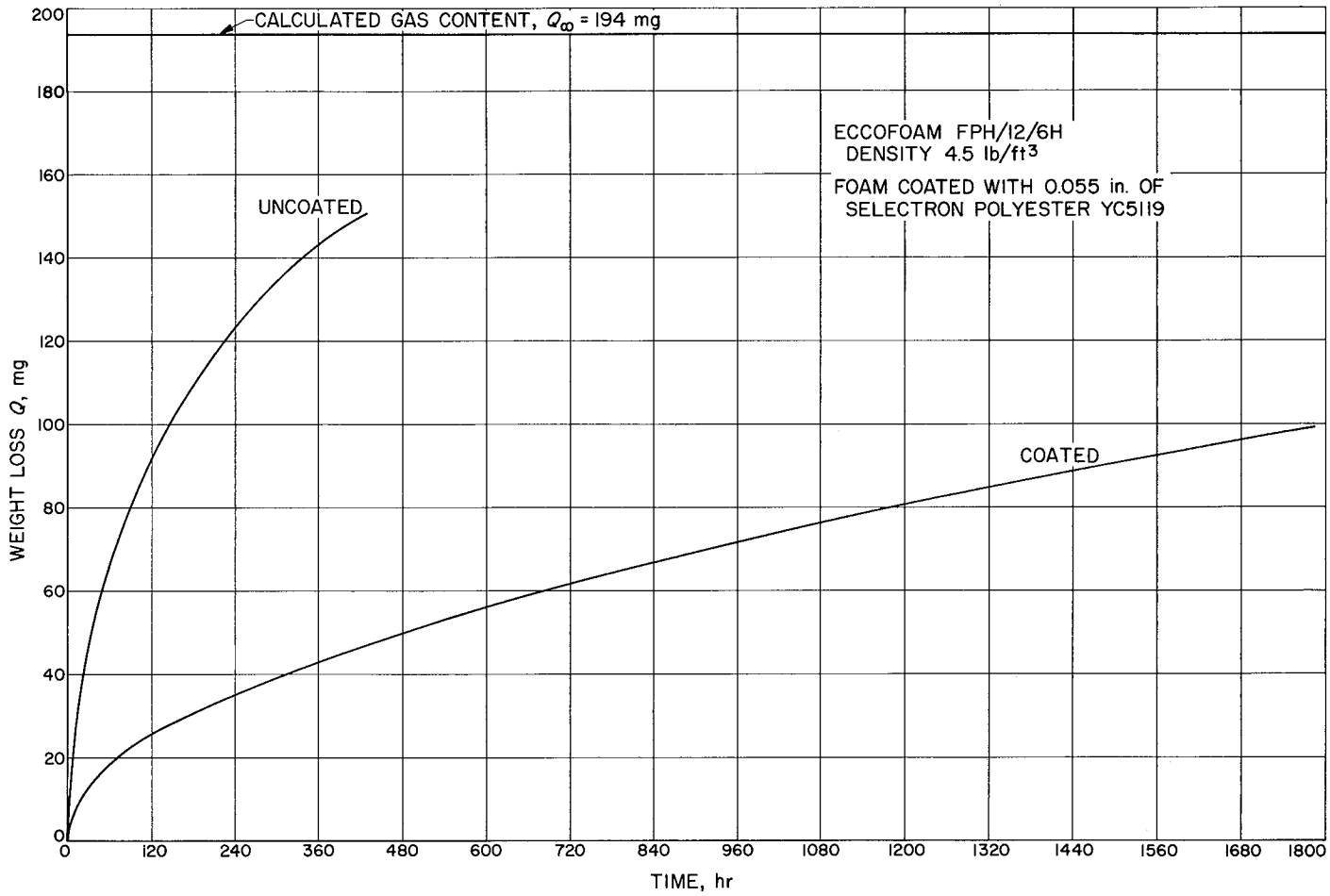


Fig. 15. Influence of plastic coating on outgassing rate

than a tenfold reduction in the outgassing rate for the coated sample. Greater reductions in outgassing rates could be achieved with coatings less permeable to CO_2.

As shown previously, the diffusion coefficient D for gases in foams can be evaluated from the slope of a plot of log $(Q_\infty - Q)$ versus time, where Q is the total gas content. Such plots are shown in Fig. 16 from which diffusion coefficients of 1.76×10^{-6} and 2.36×10^{-7} cm²/sec were obtained for the uncoated and coated samples, respectively. This method of data handling is not strictly correct for the coated foam which, in fact, is really a composite solid. However, for comparative purposes it is quite useful, and does illustrate that the coating has achieved almost a tenfold reduction in the diffusion rate which was qualitatively predicted from the data in Fig. 15.

In the past, foams were sealed in metal containers to prevent loss of gas. The use of plastic coatings should result in considerable savings in weight as well as in simpler packaging procedures.

F. Studies on Voltage Breakdown of Closed-Cell Foams

J. Farrar and J. Moacanin

1. Introduction

It is well known that electric discharge (corona, or voltage breakdown) can occur in voids or defects of plastics. Such a discharge is a disturbing source of radio-frequency interference and may also contribute to degradation of the plastic with ultimate dielectric breakdown. Therefore, if one is to consider a plastic foam as an encapsulant for high-voltage applications, it is necessary to know the conditions under which a foam could perform its function for the lifetime of a mission. Previous studies on this subject dealt with the outgassing behavior of foams in vacuum (*SPS 37-34*, Vol. IV). Now we turn our attention to the electric properties of foams under various environmental conditions.

Because of the experimental difficulties in studying corona inside a foam specimen and the fact that, in

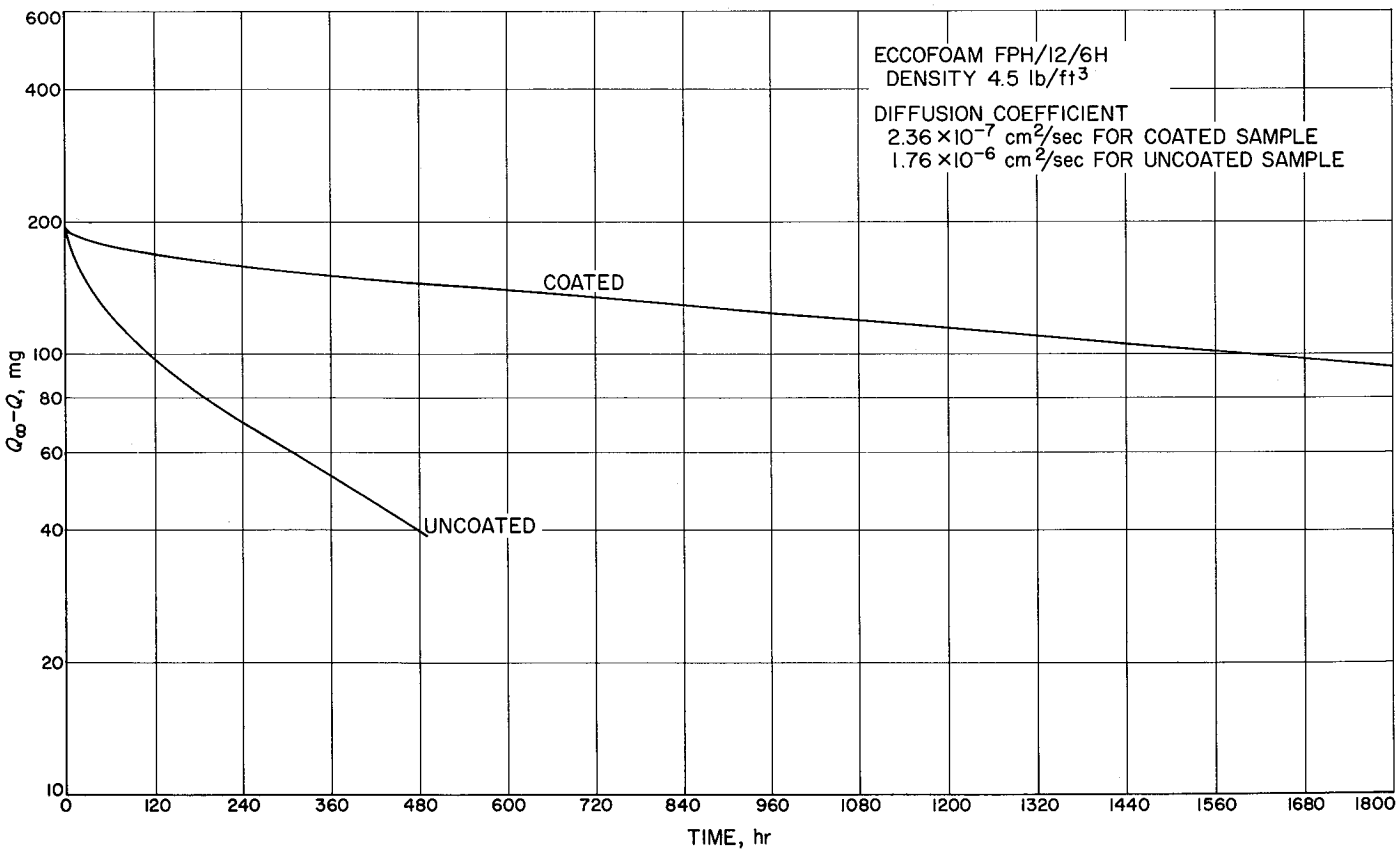


Fig. 16. Log $(Q_\infty - Q)$ versus time for a plastic-coated Eccofoam FPH/12/6H of density 4.5 lb/ft³

general, very little data exist on their high-voltage behavior it was deemed necessary first to carry out a systematic study of the effect of bulk density on the breakdown voltage of foams. Results of the effect of reduced pressure in the cells on the short-time breakdown voltage of several bulk densities are included.

2. Experiments and Results

Closed-cell polyurethane foams (Eccofoam SH, Emerson and Cumings Co.) with nominal bulk densities of 2, 4, 6, 8, 10 lb/ft^3 were purchased in the form of sheets 12 × 24 × 1 in. thick. These were then cut into test specimens 4 × 4 in. The short-time dielectric testing was carried out with a dielectric strength tester, Model PDA-1, from Industrial Instruments Co. ASTM 149 testing procedures were followed using ¼-in. electrodes (Fig. 17). First, tests were carried out on test specimens as received. For each density, five breakdown tests were made, and the length of the vertical line indicates the spread of the data points (Fig. 17). For the 2 lb/ft^3 specimen, however, 25 tests were made with no apparent increase in the spread of data points, indicating that 5 tests yield a reasonable estimate of the statistical deviation for this test procedure. Additional tests were carried out on specimens which were outgassed at 80°C in a vacuum of 1 mmHg. The specimens were allowed to cool to ambient temperature under vacuum before removal for testing. This procedure minimized back diffusion of the gas into the foam (*SPS 37-34*, Vol. IV).

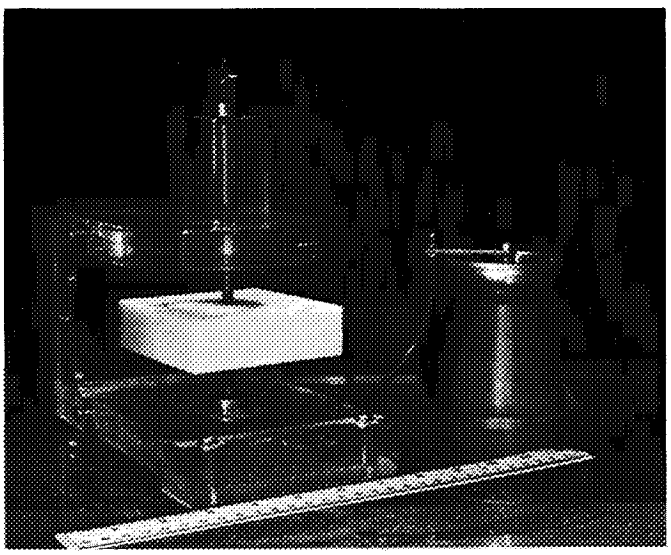


Fig. 17. Foam sample in breakdown apparatus

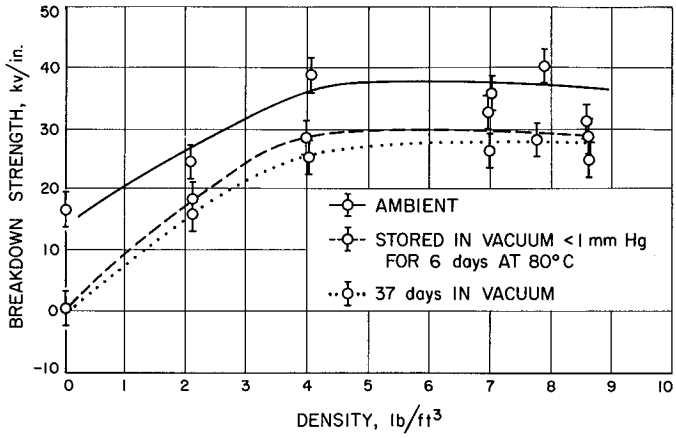


Fig. 18. Breakdown strength versus bulk density

The actual bulk density of the test specimens was determined by weighing and measuring the volume.

For the attempts to determine corona inside a foam, use was made of electrodes mounted inside a bell jar (3/32-in. brass rods, 1-in. gap) (Fig. 19). Curves of corona onset voltage versus pressure for N_2 and CO_2 were determined with this electrode configuration. These curves are shown in Fig. 20. Electrodes with the same geometry were inserted and bonded with Eccobond 55 along the axis of a cylindrical foam specimen with a radius of 2 cm and length of 4 in. A typical foam specimen with electrodes is shown in Fig. 21. The sample was put into a vacuum Bell jar and pumped down to about 5×10^{-5} mmHg within 15 min. With a 10-kv rms 60-cycle voltage, there was indication of corona. This voltage was maintained for about 20 hr with no apparent increase in the magnitude of the indicated corona. When the voltage was then

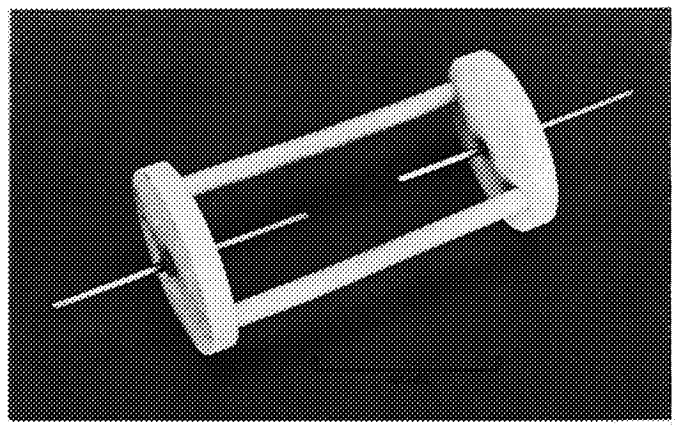


Fig. 19. Electrode configuration used to determine corona within foam sample

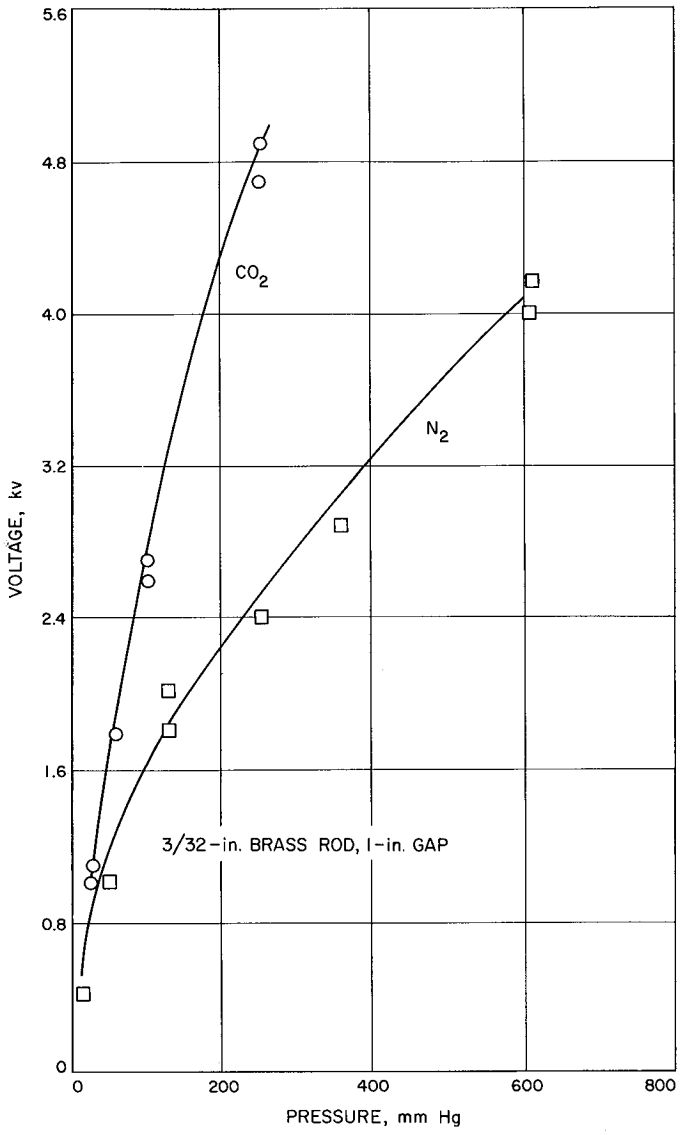


Fig. 20. Voltage at onset of corona versus pressure for CO_2 and N_2

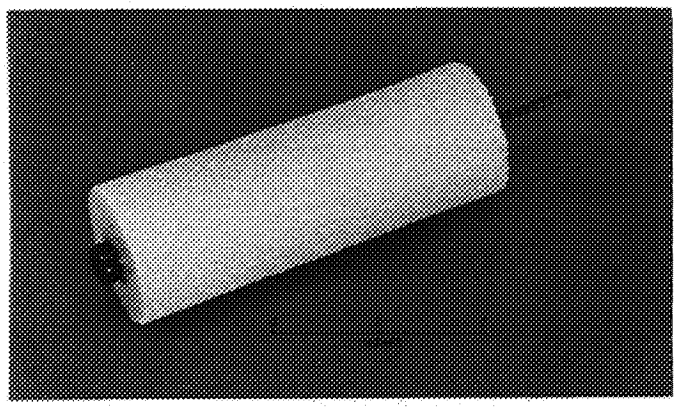


Fig. 21. Cylindrical foam specimen used for onset of corona

Fig. 22. Arcing within foam sample

increased from 10 to 12 kv, arcing took place inside the foam specimen (Fig. 22). However, the effect of reduced pressure could not be determined from this experiment, since it was subsequently found that arcing inside the foam could also be produced at atmospheric pressure at approximately the same voltage.

3. Discussion

Inspection of Fig. 17 shows that the breakdown voltage for the 2 lb/ft^3 foam is 25 kv as compared to 17 for oil at 1 atm. This increases further to about 38 kv for the 8 lb/ft^3 specimen. For specimens outgassed for 6 days a reduction of about 8 kv is observed, and an additional loss of 5 kv for those outgassed for 37 days.

The actual pressure of the blowing gas in the outgassed specimens was not determined, but reasonable estimates may be made on the basis of previous studies on the outgassing behavior of foams (*SPS 37-34*, Vol. IV). After 6 days at 80°C one should expect a loss of approximately 90% of the blowing gas leaving the remaining gas at an

average pressure of about 60 mmHg. After 37 days about 99% of the gas should be removed, and the expected pressure should be about 5 mmHg. We can now compare the reduction in the breakdown voltage of the foam with that for CO_2 for a corresponding change in pressure (Fig. 20). This comparison suggests strongly that the loss in strength for the foam can be accounted for by the loss of gas. Inasmuch as only about 1% of the gas is left after 37 days, it appears that during this period the asymptotic value for the strength of the foam has been approached. The important conclusion is that the foam will provide a minimum dielectric strength even when the blowing gas pressure is in the Paschen minimum region; e.g., for the 8 lb/ft^3 foam the minimum strength should be about 20 kv/in.

It is to be noted, however, that the above data were obtained with short-time tests (about 30 to 60 sec to failure). Limited longer time tests (2 hr to failure) show some decrease in the breakdown voltage. Further studies on rate effects are required before long-time behavior for foams can be predicted. Work is being continued and will extend to higher densities, different materials, and different cell sizes.

References

1. Ingham, J. D., and Lawson, D. D., *Journal of Polymer Science*, Vol. A-3, p. 2707, 1965.
2. LeFevre, R. J. W., in *Advances in Physical Organic Chemistry*, V. Gold, Ed., pp. 3–20, Academic Press, London, 1965.
3. Vogel, A. I., *Journal of the Chemical Society*, p. 1842, 1948.
4. Cresswell, W. T., Jeffery, G. H., Leicester, J., and Vogel, A. I., *Journal of the Chemical Society*, p. 514, 1952.
5. Lorenz, L., *Wied. Ann. Physik*, Vol. 11, p. 70, 1880.
6. Lorentz, H. A., *Wied. Ann. Physik*, Vol. 9, p. 641, 1880.
7. Lehovec, K., and Broder, J., *Journal of the Electrochemical Society*, Vol. 101, p. 208, 1954; Broder, J. D., U.S. Pat. 2,690,465, September 1954; Lehovec, K., U.S. Pat. 2,689,876. (See also Mrgudich, N., in *Encyclopedia of Electrochemistry*, Reinhold Publishing Corp., N.Y., 1964, p. 84.
8. Hermann, A. M., and Rembaum, A., *Journal of Polymer Science*, Part C: Polymer Symposia (in press).
9. *Handbook of Physics and Chemistry*, 38th ed., Chemical Rubber Publishing Co., Cleveland, Ohio, 1956/1957, p. 1691.
10. Frisch, H. L., and Simha, R., in *Rheology*, Eirich, F. R., Ed., Academic Press, New York, N.Y., 1956, Vol. I, pp. 525–614.
11. Sadron, C., and Rempp, P., *Journal of Polymer Science*, Vol. 29, pp. 127–140, 1958.
12. Utracki, L., and Simha, R., *Journal of Polymer Science*, Vol. A1, pp. 1089–1098, 1963.
13. Simha, R., *High Polymer Physics*, Robinson, H. A., Ed., Chemical Publishing Co., New York, N.Y., 1948, pp. 398–416.

References (Cont'd)

14. Powell, R. F., Roseveare, W. F., and Eyring, H., *Industrial and Engineering Chemistry,* Vol. 33, pp. 430–435, 1941.

15. Conway, B. E., and Lakhanpal, M. L., *Journal of Polymer Science,* Vol. 46, pp. 93–111, 1960.

16. Moacanin, J., *Journal of Applied Polymer Science,* Vol. 1, pp. 272–282, 1959.

17. Bondi, A., *Rheology,* Eyring, F. R., Ed., Academic Press, New York, N.Y., 1956, Vol. I, pp. 321–350.

18. Heil, J. F., and Prausnitz, J. M., *Journal of Chemical Physics,* Vol. 44, pp. 737–740, 1966.

19. Kalfayan, S. H., "Carboxyl-Ending Poly(Alkylene Oxides) Purification and Curing," *SPS 37-35,* Vol. IV, October 31, 1965, pp. 123 to 125.

20. Tarbell, D. S., and Noble, P., Jr., "Acid-Catalyzed Ring-Opening Reactions of Some Unsymmetrical Ethyleneimine Derivatives." *Journal of the American Chemical Society,* Vol. 72, p. 2657, 1950.

21. Bastian, H., "Polyalkylenepolyamines," German Patent 881,659, through *Chemical Abstracts,* Vol. 52, p. 11892, 1958.

22. Heine, H. W., et al., "The Synthesis of Some N-Arylethylenimines," *Journal of the American Chemical Society,* Vol. 76, p. 2503, 1954.

X. Research and Advanced Concepts

A. Velocity Profile Measurement in Plasma Flows Using Tracers Produced by a Laser Beam

Che Jen Chen

1. Introduction

The feasibility of using ion tracers produced by corona discharges, glow discharges, and energetic particle ionization for velocity measurement has been explored by several authors (Ref. 1). Because of the characteristics of the methods of ion production and detection, the results of the measurement can give only the average of the flow speed over an appreciable length and area rather than local ones.

In this article a method of velocity profile measurement by tracers produced by a laser beam is described. A small, highly luminous and almost fully ionized plasma drop is formed by focusing a laser beam at a point in the flow. The drop velocity is determined either by electrostatic probe or drum camera techniques. Utilizing the probe technique, the drop velocity is determined by measuring the time required for the drop to travel a known distance from the laser discharge point to the position of a fixed probe in the flow. The drum camera technique eliminates the use of probes. The velocity of the drop is obtained from the angle between the direction of the film motion (which is parallel to the streak formed by a fixed luminous point in the vicinity of the flow) and the direction of the streak formed by the plasma drop. The new features of this method are the following: (1) It can give the local velocity of flow averaged over an area (perpendicular to the flow) about 2×10^{-3} cm^2 and length about 0.5 cm; (2) the disturbances in the flow caused by the instrumentation is minimized; and (3) it is applicable to both unionized and ionized flows.

2. Experimental Instrumentation

A Q-switched giant pulse ruby laser (6943Å) capable of delivering up to 100 mw with a pulse width of 15 ns is focused in a flow by a lens at a focal point of volume less than 10^{-5} cm^3. The gas breakdown at the focal point constitutes a plasma drop of highly luminous and almost fully ionized gas. The radius of the plasma drop (assumed to be spherical in shape) during the period of measurement is growing from about 0.2 to 1.0 mm (Ref. 2). When

such a plasma drop is produced in a flow, the motion of the drop will follow the flow speed accurately. A measurement of the drop velocity will, therefore, give the velocity of the flow. The following two methods are used to detect the motion of the plasma drop produced by the laser beam:

a. Electrostatic probe method. An electrostatic probe consisting of two tungsten wires 0.5 mm in diameter, 5 mm in length and 0.4 mm apart is located at a known distance (~5 mm) downstream of the focal point of the laser beam. The signal indicating the arrival of the plasma drop at the probe is detected by the circuit as shown in Fig. 1. The time required for the drop to travel from the focal point to the probe is measured by the time elapsed between the initiation of the laser pulse (trigger of the oscilloscope) and the point of the maximum signal from the probe on the oscilloscope trace. A typical oscilloscope signal is shown in Fig. 2. For ionized flows the background charge density produces a constant and comparatively noisy probe signal which is, in most cases, more than one order of magnitude smaller than the signal, due to the arrival of the plasma drop. Thus, there is no difficulty in measuring the time of arrival of the plasma drop at the probe. The velocity of the flows in which the drop is suspended, in both unionized and ionized flow, is thus obtained.

b. Drum camera method. A drum camera having film traveling at speeds up to 100 m/sec is focused in the vertical plane containing the path of the plasma drop. The direction of the motion of the film is adjusted to be perpendicular to the axis of the flow. The velocity of the flow is obtained from the angle between the direction of film motion, which is parallel to the streak formed by a fixed luminous point in the vicinity of the flow, and the direction of the streak formed by the plasma drop. The experimental setup is also shown in Fig. 1. Typical drum camera streaks of both un-ionized and ionized flows are shown in Fig. 3. Note that, in the ionized flow, the plasma background which is much less luminous than the plasma drop, does not appear on the film because the exposure time is too small. The velocities of the flow are evaluated by knowing the angle θ as shown in Fig. 3.

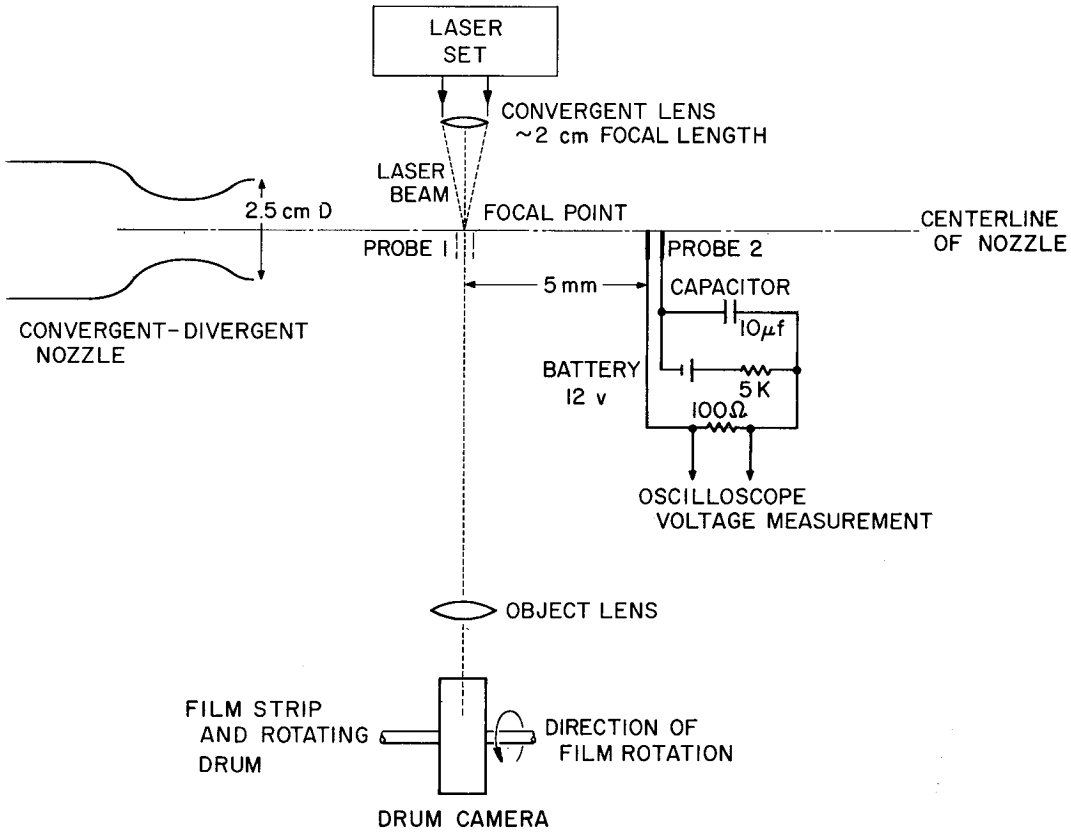


Fig. 1. Schematic for plasma velocity measurements

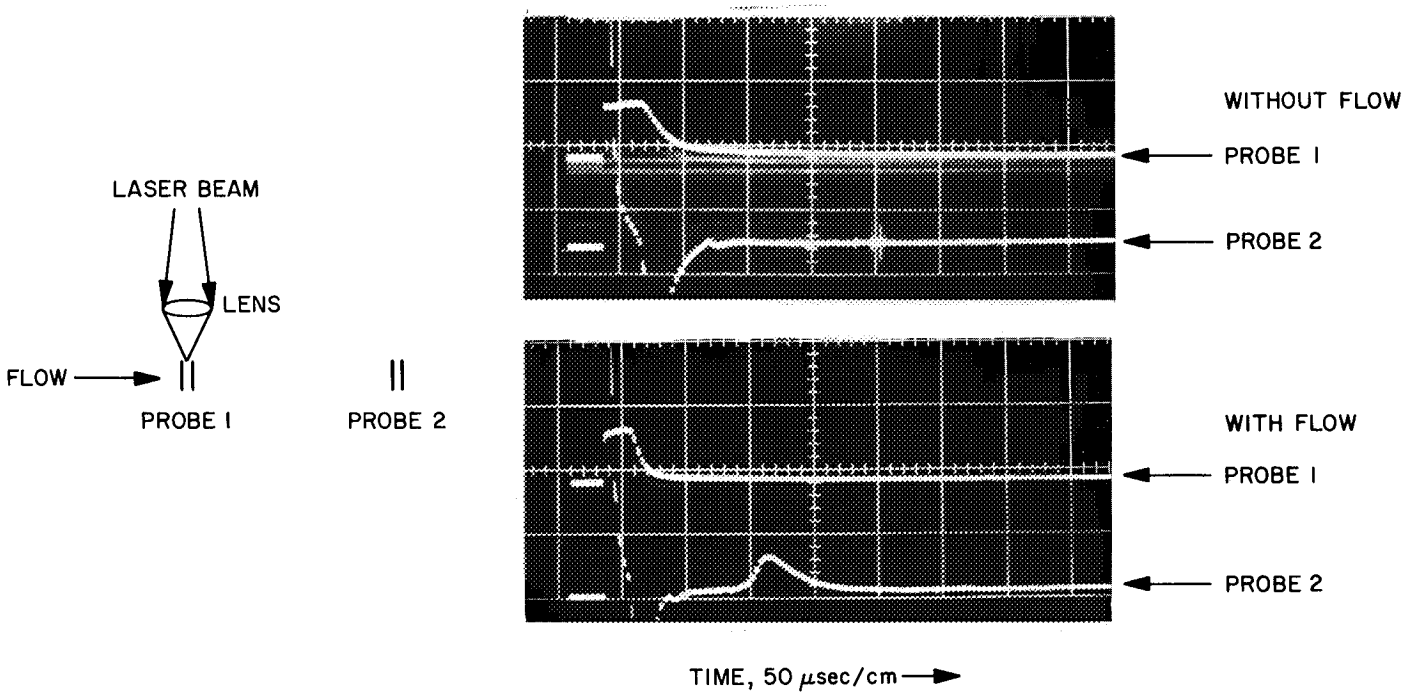


Fig. 2. Electric probe signals of moving plasma drop

3. *Error Analysis*

Errors caused by the localized heating of the plasma drop and the distortion of the emitted radiation from the plasma drop through the adjacent plasma, are discussed as follows:

a. Shock wave formation. It has been reported that the sudden heating of the gas by the concentrated photon energy causes the formation of a very fast-decaying blast wave (Ref. 3). Since the blast wave is spherical and attenuated rapidly within 1 mm, the motion of the drop center will not be altered by the energy addition from the laser beam.

b. Buoyancy effect. The size of the drop is increased due to the heating and diffusion processes in the drop, and the density of the drop becomes smaller than that of the ambient gas. Consequently, a buoyancy as well as the Stoke's viscous retardation forces are acting on the drop. The equation of motion of the drop, assuming the drop behaves like a solid body, can be written as (Ref. 4):

$$d^2y/dt^2 = du/dt = \left(\frac{\rho_g}{\rho_d} - 1\right) g - 18\eta(u/d^2)\rho_d \tag{1}$$

where u and y are the upward velocity and vertical coordinate of the drop, respectively; ρ_g and ρ_d, the densities of the ambient gas and the drop, respectively; g, the gravitational acceleration; η, the viscosity of the plasma; and d, the diameter of the drop. The solutions of Eq. (1) with the initial conditions $y = u = 0$, at $t = 0$ are

$$u = \left(\frac{\rho_g}{\rho_d} - 1\right)(g\, d^2\, \rho_d/18\eta) \times \left[1 - \exp\left(\frac{-18\,\eta t}{d^2\,\rho_d}\right)\right] \tag{2}$$

$$y = \left(\frac{\rho_g}{\rho_d} - 1\right)(g\, d^2\, \rho_d/18\eta) \times \left[t - \left(\frac{d^2\,\rho_d}{18\,\eta}\right)\left[1 - \exp\left(\frac{-18\,\eta t}{d^2\,\rho_d}\right)\right]\right] \tag{3}$$

For a time scale of the order of about 10^{-5} (see Sect. 4) with η equal to 8×10^{-4} poise, ρ_g/ρ_d, about 10; d, 0.1 cm; ρ_g, about 10^{-7} to 10^{-6} g/cm^3; y calculated from Eq. (3) is about 1.8×10^{-9} cm. This is negligible in comparison with the diameter of the drop, even at the beginning of the breakdown of the gas ($\sim 10^{-2}$ cm in diameter). Therefore, the buoyancy force is negligible. The effect of the buoyancy force on the drum camera measurement can also be shown to be negligible.

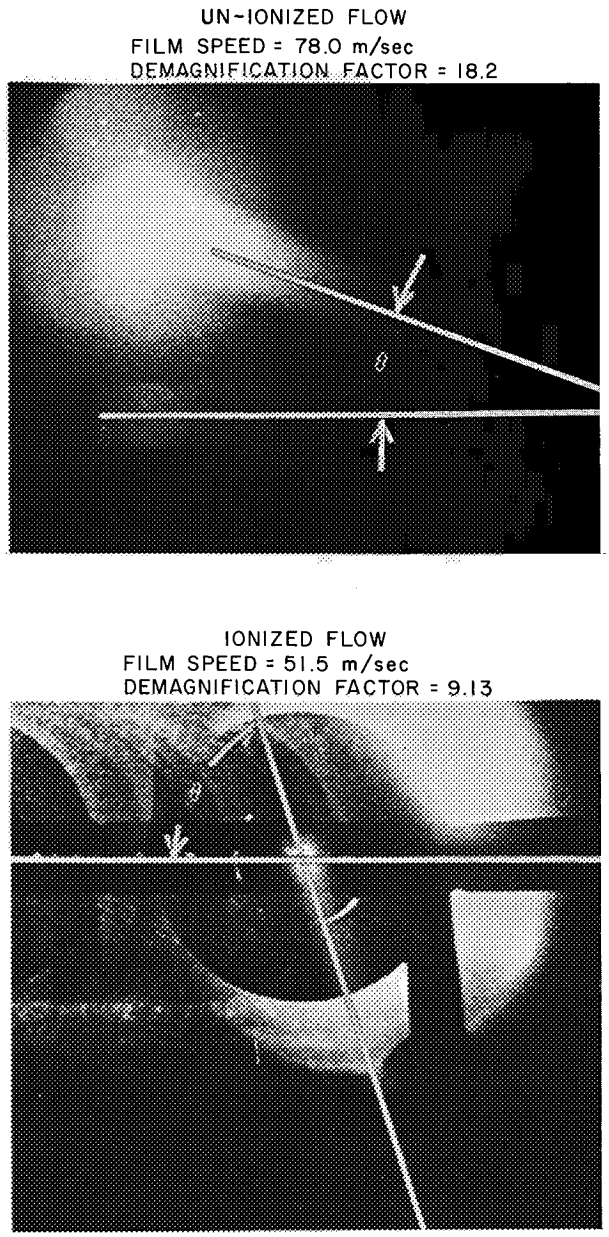


Fig. 3. Drum camera streaks of plasma drop in a flow

c. Other possible sources of error.

(1) Refraction of the optical path. This effect is negligible if the radial density gradient within the distance covered in the measurement is about constant. This is thought to be the case.

(2) Diffusion in the drop. The diffusion characteristic time in present case is calculated to be about 3×10^{-4} sec, time of drop formation is about 10^{-7} sec and time of measurement is 10^{-5} sec. Therefore, the diffusion effect will be insignificant.

(3) Motion of the drop during the formation period. Since the initial time of the measurement is at the instant of the laser pulse, the time of formation of the drop has no effect on the velocity measurement. The error in time measurement can only be the result of the inconsistency of the trigger time of the oscilloscope with respect to the laser pulse and is estimated to be less than 10^{-8} sec, which is negligible in comparison with 10^{-5} sec, the total time measured.

At very low pressures it takes about 500 ns (Ref. 2) to completely ionize the plasma drop. This long relaxation time reduces the accuracy of the measurement, because the radiation from the drop during the time of measurement is less intense than at higher pressures. To overcome this difficulty at low pressures, both the probe and drum camera signals are greatly intensified by using sublimed metallic vapor as a tracer. A metallic wire, such as tungsten, having a diameter less than the atom-atom mean free path to minimize disturbances, is introduced in the flow. The laser beam is focused on the wire to sublime a vapor drop from the surface of the wire. The linearity of drum camera streaks indicate that the sublimed vapor drop velocity is the same as the flow velocity. By using such an arrangement the utilization of the tracers produced by the laser beam for the measurement of the flow speed can be extended to even lower pressures.

4. Velocity Profile Measurements in a Supersonic Jet

To demonstrate the feasibility of this technique velocity measurements were made in a supersonic jet. A supersonic flow, approximately Mach 3, is produced by the expansion of the gas through a convergent-divergent nozzle. The flow parameters in the un-ionized jet, such as the pressure and temperature of the gas in the chamber before expanding, the nozzle exit and tank ambient gas pressure, and the pitot-tube pressure in the free jet are measured. The ionized flow is produced by a conventional vortex-stabilized plasma generator. The power input, cooling water losses and mass flow are used to calculate the total enthalpy of the ionized gas flow. The flow speed at the centerline of the jet for each case can be calculated by knowing the flow parameters mentioned above and assuming one-dimensional isentropic flow in the nozzle. For the ionized flow, calculations assuming equilibrium and frozen flow (degree of ionization constant) predict about the same velocity. The laser and the accessory optical system are actuated axially and radially with respect to the flow, making it possible to locate the focal point of the laser beam at various locations in the flow.

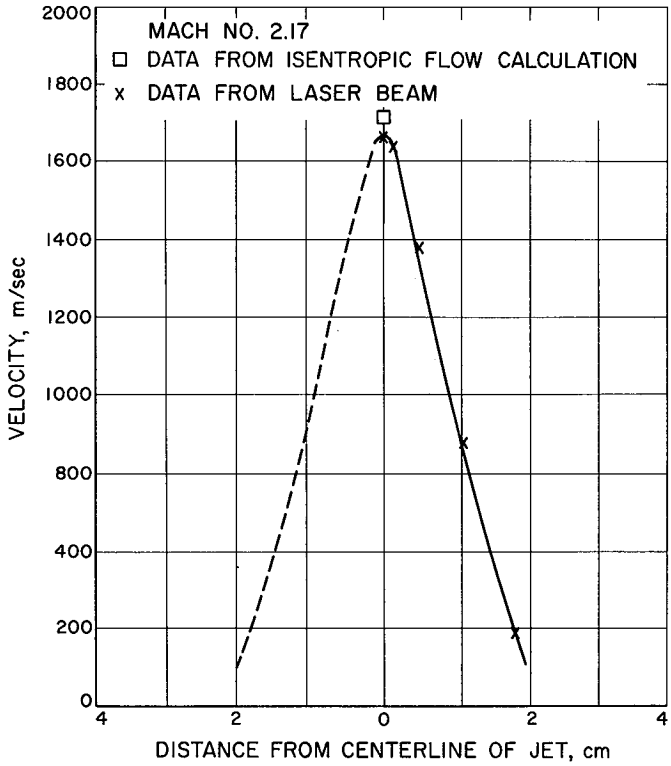


Fig. 4. Velocity profile of ionized argon jet

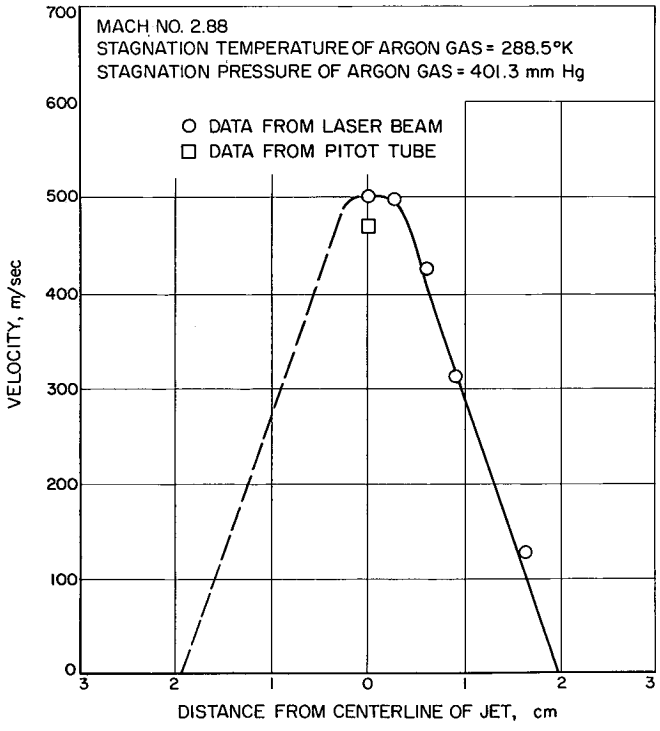


Fig. 5. Velocity profile of un-ionized argon jet

The velocity profiles measured by laser beam tracers in ionized and un-ionized argon with the drum camera technique are shown in Figs. 4 and 5, respectively. The percentage difference of the velocities measured by the electrostatic probe and drum camera techniques are within the estimated experimental error of about 3%.

B. Liquid MHD Power Conversion

D. Elliott and D. Cerini

The long lifetimes required of electric-propulsion powerplants make cycles without rotating components attractive. Such a cycle under investigation at JPL is the liquid-metal magnetohydrodynamic system shown schematically in Fig. 6. 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 (or radiator-loop condenser) as condensate, flows through an EM pump and regenerative heat exchanger to the nozzle, 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 through the regenerative heat exchanger. The lithium leaves the separator at high velocity (typically 500 ft/sec), decelerates through the production of electric power on the MHD generator, and leaves the generator with sufficient velocity (typically 300 ft/sec) to return through a diffuser to the reactor (or reactor-loop heat exchanger) where the lithium is reheated.

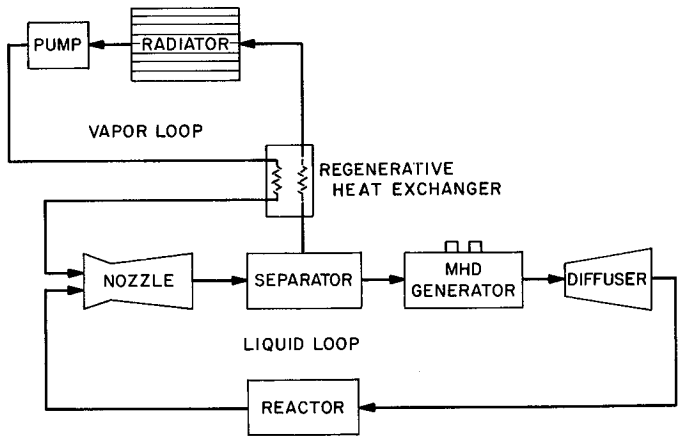


Fig. 6. Liquid MHD power conversion system

Analysis of AC induction generators and of liquid MHD cycles has been completed. Experiments on AC power generation, lithium erosion, and insulator compatibility are in progress, and a 50-kw NaK-nitrogen conversion system is under construction.

1. Analysis

a. AC power generation. Portions of the theory of compensated induction generators were presented in *SPS 37-33*, *37-36*, and *37-37*, Vol. IV. The theory has been completed and published in Ref. 5. The main result is an expression for the efficiency of the traveling-wave region of a compensated induction generator with optimum parabolic slots. The losses accounted for are fluid ohmic heating, boundary-layer shunting, wall friction, and winding power. The efficiency expression derived in Ref. 5 is

$$\eta = \frac{1}{1+s}\left[1 - \frac{1}{1+63s/(1+s)}\right] \times \left[\frac{1 - \dfrac{6\pi^4\,\alpha g^3 N^4 (1+s)^2 B_m\left[1+(sR_m b/g)^2\right]}{\mu_0^2\,\sigma\sigma_w b L^4 U^2 s B_{sat}(1-B_m/B_{sat})^3}}{1 + \dfrac{2\rho U C_f(c+b)(1+s)}{\sigma b c B_m^2\, s\left[1+(1+s)/63s\right]}}\right] \quad (1)$$

where s is the slip (difference between fluid and wave velocity, divided by wave velocity; ρ is the fluid density; U is the fluid bulk velocity (volume flow rate divided by flow area); C_f is the friction coefficient (wall shear divided by $\rho U^2/2$); c is the channel width perpendicular to the velocity and the field; b is the channel height parallel to the field; σ is the fluid electrical conductivity; B_m is the amplitude of the magnetic field; α is the ratio of total winding power to the power dissipated in the slots only, with 100% conductor fill and unity AC/DC resistance ratio; g is the iron gap; N is the electrical length of the generator in wavelengths; σ_w is the electrical conductivity of the windings; B_{sat} is the saturation flux density of the iron; R_m is the magnetic Reynolds number based on wave velocity and wavelength; and L is the generator length given by

$$L = (P_0/U^2)\left\{\sigma b c B_m^2\, s\left[1+(1+s)/63s\right]/2(1+s) + \rho U C_f(c+b)\right\}^{-1} \quad (2)$$

where P_0 is the fluid input power.

The efficiency η is the electric power output divided by P_0. Eqs. (1) and (2) were derived for constant-velocity operation where P_0 is the product of volume flow rate and pressure drop; but the same relations can be applied with little error to constant-pressure operation, where P_0 is the fluid kinetic power change, if the mean velocity and slip are employed in the equations.

The first factor in Eq. (1), $1/(1+s)$, is the efficiency with fluid ohmic heating loss only. The second factor gives the reduction in efficiency due to nonuniform slip across the channel with a 1/7-power velocity profile; the boundary layer acts as a shunt for current generated by the high-velocity core flow. The denominator of the third factor gives the reduction of efficiency due to wall friction, and the numerator gives the reduction due to winding power.

b. 300-kw lithium generator. Typical flow conditions for the generator employed in a nominal 300-kw cesium-lithium powerplant would be as follows: inlet velocity $U_1 = 150$ m/sec, exit velocity $U_2 = 90$ m/sec, flow rate $\dot{m} = 70$ kg/sec, input power $P_0 = (\dot{m}/2) \cdot (U_1^2 - U_2^2) = 500$ kw, and channel width $c = 0.25$ m. Mean velocity is $U = 120$ m/sec, lithium density is 440 kg/m^3, and the corresponding channel height is $b = 0.53$ cm. Heat-transfer estimates (*SPS 37-37*, Vol. IV) indicate that 0.2 cm of ceramic insulation, with vacuum at the interface with the iron, would provide sufficient thermal insulation, making the iron gap 0.93 cm; the fringing coefficient, due to the slots, would be about 1.15, making $g = (1.15)(0.93) = 1.07$ cm.

The Reynolds number is 2.4×10^6 and the corresponding friction coefficient is $C_f = 0.0026$. The electrical conductivity of lithium is $\sigma = 2.0 \times 10^6$ mho/m, and the magnetic Reynolds number is

$$R_m = \mu\sigma UL/2\pi N(1+s) = 48L/N(1+s) \quad (3)$$

For a slot fill of 0.8, AC/DC resistance ratio in the slot of 1.4, and external conductor resistance equal to the slot DC resistance (about the best values that could be achieved), the ratio of total winding power to slot DC power with solid copper fill would be $\alpha = 1.4/0.8 + 1.0/0.8 = 3.0$. For 200°C winding temperature, copper conductivity is $\sigma_w = 3.0 \times 10^7$ mho/m. For iron-cobalt alloy at a stacking factor of 0.95 the saturation flux density is $B_{sat} = 2.2$ webers/m^2.

These values were substituted into Eq. (1) [with length L calculated from Eq. (2)], and the efficiency for an electrical length of one wavelength is presented in Fig. 7 as a function of slip, for various values of field

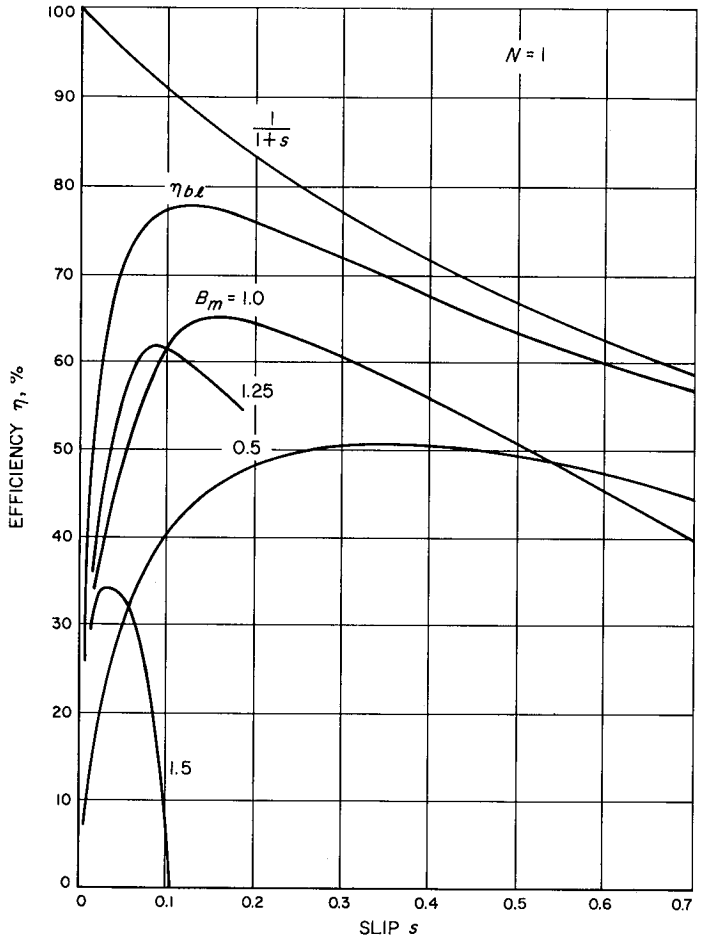


Fig. 7. Theoretical efficiency of a one-wavelength 300-kw lithium generator

amplitude B_m. Efficiencies with fluid ohmic loss only, $1/(1+s)$, and with fluid ohmic loss and boundary-layer electrical loss only, $\eta_{b\ell}$ [the product of the first two factors in Eq. (1)] are shown for comparison. The highest efficiency for the Eq. (2) curves is $\eta = 65\%$ at $B_m = 1.0$ webers/m^2 and $s = 0.15$. This efficiency is seen to be about the highest attainable for the given conditions; the peak efficiency is less for both lower and higher fields. The generator length for the $B_m = 1.0$, $s = 0.15$ condition is $L = 0.15$ m. The reactive power required from the excitation capacitors is calculated to be 1040 kvar.

Efficiencies were also calculated from Eq. (2) for electrical lengths of $N = 2$ and 3 wavelengths, and the maximum efficiency attainable with any given field was found to decrease with N and to occur at a higher slip. The maximum efficiencies for each field are presented as a function of field and number of wavelengths in Fig. 8. The maximum efficiency and corresponding optimum field for each N is: $\eta = 65\%$ at $B_m = 1.0$ for $N = 1$,

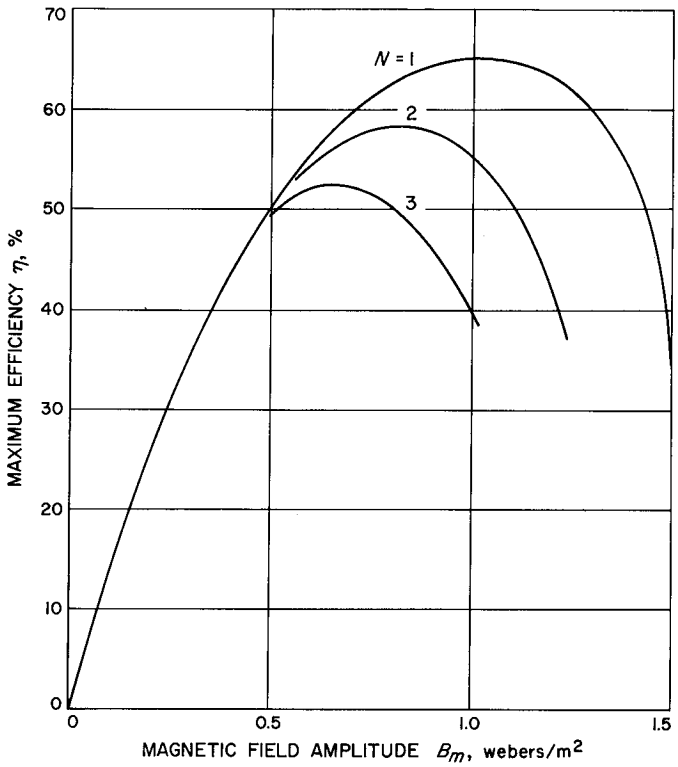


Fig. 8. Effect of electrical length on efficiency of 300-kw lithium generator

$\eta = 58\%$ at $B_m = 0.8$ for $N = 2$, and $\eta = 53\%$ at $B_m = 0.7$ for $N = 3$. Thus, a one wavelength generator should be capable of efficiencies approaching 70%, if some reduction in wall friction and flattening of the velocity profile can be accomplished as indicated possible by the water nitrogen experiments reported in *SPS 37-32*, Vol. IV. Compensating-pole losses must also be held to no more than 10% of the fluid power, and this is believed to be possible with gas injection for conductivity reduction in the compensating-pole region, a technique to be investigated in the NaK-nitrogen conversion system.

c. Cycle analysis. The efficiencies of seven different liquid-metal MHD cycles were compared in *SPS 37-37*, Vol. IV, for the ideal case of operation without separator or condenser friction loss. The efficiencies of the same cycles *with* friction, representing the efficiencies attainable with present components, are compared in Fig. 9. With the exception of the added friction, the assumptions are the same as those employed in *SPS 37-37*, Vol. IV. The major ones are: constant pressure in the separator or condenser and in the generator; regenerative heat exchanger employed in the two-component separator cycle; 80% recovery of stagnation pressure in the radiator circuit of the condenser cycles for injection of the coolant

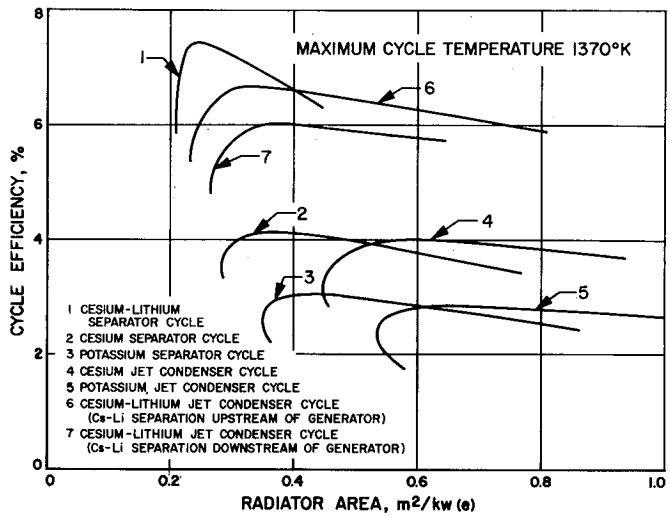


Fig. 9. Cycle efficiency and radiator area for liquid-metal MHD cycles with friction

at high velocity; 70% generator efficiency; and 85% diffuser efficiency.

The variation of cycle efficiency with radiator area at a nozzle inlet temperature of 1370°K (2000°F) is presented in Fig. 9, where the radiator area is the isothermal value (surface temperature equal to fluid temperature) at an emissivity of 0.9. Each point on the curves is at a different nozzle exit pressure, decreasing from about 30% of nozzle inlet pressure at the lower end of the curves to about 2% at the other end. Radiator inlet temperatures also decrease, being equal to the saturation temperature of the vapor. Nozzle inlet quality and separator or condenser impingement angle are optimized at each point. Initially, the cycles follow the Rankine-cycle behavior of increasing cycle efficiency and decreasing radiator area until a radiator temperature equal to about 75% of the nozzle inlet temperature is reached, beyond which the radiator area again increases. At the same time, the separator and condenser wall friction losses begin to increase rapidly, because of increased nozzle exit area, and the cycle efficiencies reach a maximum and then decrease with further pressure reduction.

The cesium-lithium separator cycle has the highest efficiency, 7.5% at a radiator area of 0.25 m^2/kw (2.7 ft^2/kw). The best cesium-lithium condenser cycle has lower efficiency because of larger nozzle exit area. The other cesium-lithium condenser cycle is further reduced in efficiency by lower lithium inlet velocity to the condenser. The other four cycles have only about half the efficiency of the cesium-lithium cycles, because of the large coolant flow rate (1 to 3 times the nozzle flow rate) in the single-component condenser cycles. Efficiencies with potassium are lower than with cesium, due to lower nozzle performance and larger nozzle exit area.

Although the cesium-lithium condenser cycle has nearly the efficiency of the cesium-lithium separator cycle, and has the advantage of generator operation at the radiator temperature instead of at the reactor temperature, its feasibility is questionable because of the low electrical conductivity of cesium and the need for separation of the cesium and lithium after condensation. It is concluded that, within the present knowledge of separators and condensers, the cesium-lithium separator cycle is the only one likely to achieve efficiencies greater than 5%.

Efficiencies greater than 12% would be attainable with all of the cycles if separator and condenser friction could be eliminated, as shown in *SPS 37-37*, Vol. IV, and this is a promising avenue to future improvement. But efficiencies of 5 to 7% are acceptable in a space powerplant, because reactor and shield weights change slowly with thermal power, and radiator weight is only 15 to 30% of the total; the weight penalty relative to a cycle with 10 to 14% efficiency, the highest attainable with a Rankine-cycle turbine system for space, would be no more than 50%, and possible only 0 to 25%, considering the weight savings that might be possible with the greater simplicity of the MHD system.

2. Experiments

a. AC generator. The second AC generator was tested with NaK in a preliminary series of runs at low NaK velocity, with the generator acting as a pump. The NaK feed pressure was set at 100 psig and the generator exit pressure at 10 psig, giving a NaK flow rate of 5.2 lb/sec and velocity of 90 ft/sec. The exciter was operated at 900 cps, giving a traveling magnetic field velocity of 135 ft/sec.

With zero magnetic field the pressure drop across the traveling-wave region of the generator was 17 psi, and with an indicated 4400-gauss field amplitude the pressure drop was reduced to zero, a pumping effect which was 60% of the theoretical value for this field.

The field measurement was uncertain, however, because of discrepancies between the search-coil field

measurements and the winding currents. Also, a short circuit had developed between the A- and C-phase windings, and the pressure drops were becoming erratic, due to some debris in the NaK system. The generator has been removed for winding repairs, search-coil calibration, and installation of filters.

b. NaK-nitrogen conversion system. Fabrication is proceeding on a schedule leading to water-nitrogen testing in September, empty-channel generator tests in December 1966, and system tests in 1967.

c. Lithium erosion loop. Repairs and modifications needed after the 45-hr initial run were completed, and the loop is being reassembled for a scheduled 500-hr run in June.

d. Insulator compatibility tests. A portion of the samples of AlN, AlB_{12}, Y_2O_3, 95% Y_2O_3-5% ThO_2 and 90% Y_2O_3-10% ThO_2 were removed from the vacuum furnace after 1034 hr exposure to 2000°F lithium in Cb-1%Zr capsules. The samples of AlN crumbled when handled. The AlB_{12} samples appeared to have reacted violently with the lithium; the capsules were distorted, and only a small amount of black residue remained of the AlB_{12}. The only samples surviving the exposure were Y_2O_3, 95%Y_2O_3-5% ThO_2 and 90% Y_2O_3-10% ThO_2. These were darkened, and some solution had taken place where three phases (Li, ceramic, and Cb-1%Zr) were in contact. In addition, black metallic globules, the composition of which are presently being analyzed, were deposited on the colder portion of the ceramics. The remaining unopened Y_2O_3, 95% Y_2O_3-5% ThO_2 and 90% Y_2O_3-10% ThO_2 samples will be exposed for an additional 4000 hr.

Tantalum-sheathed chromel-alumel thermocouples were compared during the test with Pt-Pt10%Rh and W5%Re-W26%Re thermocouples, all spot-welded to the outside of the capsules. There was no shift in output of the less expensive chromel-alumel thermocouples relative to the other two types, within 5°F, during the 660-hr portion of the test with the chromel-alumel thermocouples in place.

3. Publications

A summary of the results prior to September 1964 has been published in Ref. 6; the DC generator investigation has been published in Ref. 7; the supersonic two-phase tunnel results have been reported in Ref. 8; and the cycle studies are to be published in Ref. 9.

C. Analog Computer Study of Thermionic Reactor Space Powerplant Transients

H. Gronroos

1. Introduction

An earlier *SPS* article reported on a planned analog computer simulation of a space thermionic reactor powerplant (Ref. 10). The simulation setup is now fully operational. Only minor modifications to the original scheme were necessary, mainly because of equipment limitations.

Both of JPL's PACE analog computers and one PDP-4 digital computer are slaved together in a hybrid setup. The digital machine is used for simulation of delay times. For longer delay times (in excess of 5 sec) the usual approximations necessary in analog simulation become unreliable. In this case it was feasible to use the PDP-4 computer. In general, one would perhaps use some other simpler equipment for time-delay simulation purposes.

Concurrently with the simulation studies, analytical investigations are pursued (see Sect. E, "Linear Stability Analysis of a Thermionic Powerplant"). The aim is to obtain a complete understanding of the stability and dynamic behavior of a thermionic space reactor powerplant. To this end, during the next month's phase and gain versus frequency characteristics will be measured, using the analog computers. The remaining unused analog computer in JPL's facility will be added to the present setup to include more detail and the control system. It is planned to completely mechanize the phase and gain versus frequency readout by use of an additional PDP-4 computer.

Below is given a brief description of the simulation setup. In addition, some results from transient behavior studies are shown for illustrative purposes. Complete documentation of this phase of the study is in preparation.

2. Simulation Setup

Fig. 10 schematically shows the equipment used for the simulation. Development of the computer patching diagram was relatively straightforward. As is common with analog computers, the major problem has been with the proper balancing, noise, and drift. Because of the nature of the thermionic process, an accuracy better than ±10°K is desired for the emitter temperature in the

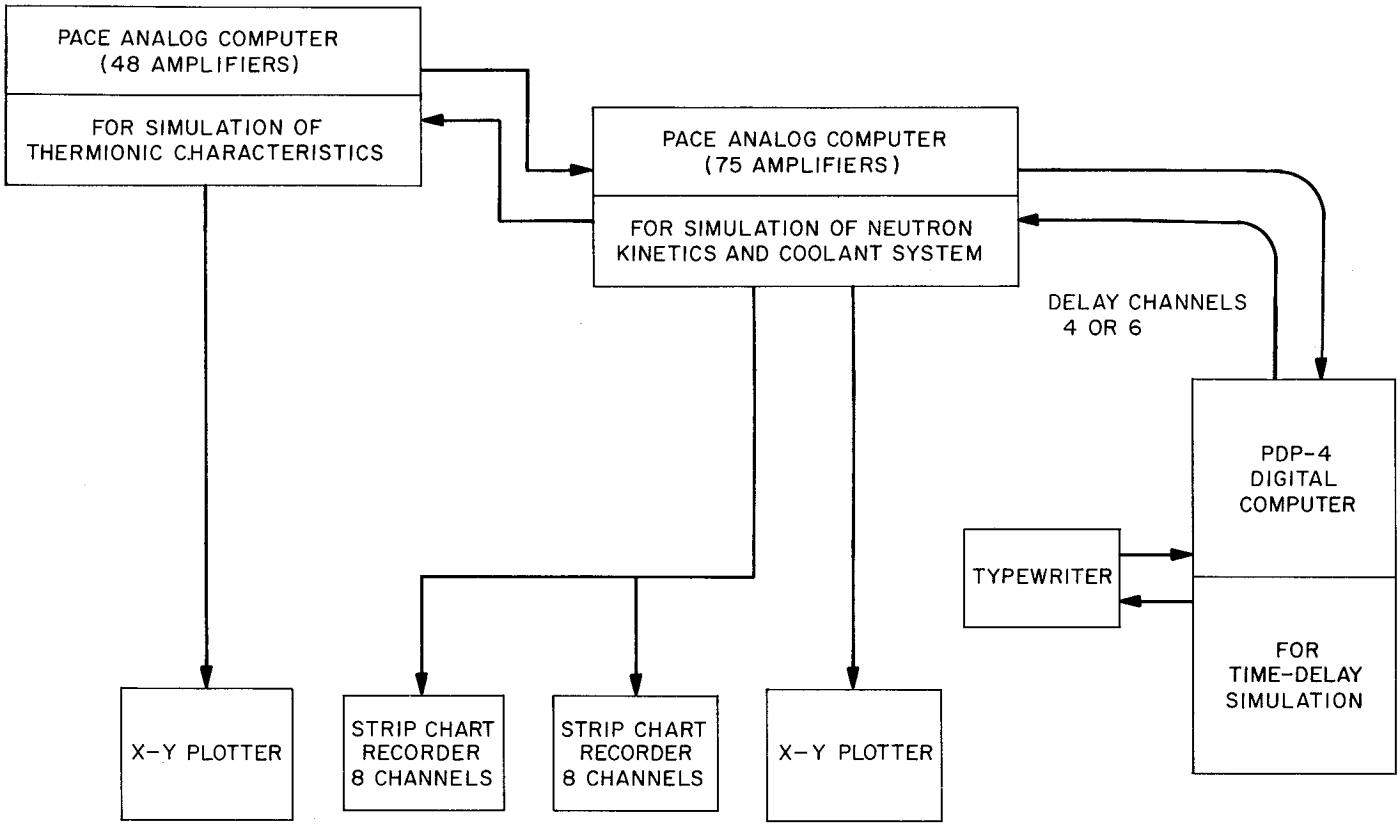


Fig. 10. Equipment setup for analog simulation of thermionic reactor space powerplant

range of 1800 to 2200°K if the errors are to be acceptable. By careful adjustments this accuracy has been achieved.

Each system component can be studied individually, and by proper switching a single-loop or two-loop coolant system is simulated. The time delays are typed in on the typewriter. This starts a run on the digital computer for the chosen delay times. A circular buffer keeps this machine running until a new command is given.

Running in real-time is feasible, but because of the relatively long time for the system to settle out at a new equilibrium — up to 5 min — production calculations have been run ten times faster than real-time.

3. Some Results

The first phase of calculations have been made to study the transient behavior following a step change or very fast ramp in some parameter. Trajectories are taken on two X-Y plotters and on strip chart recorders.

In general, it can be said that a thermionic reactor is a stable device for electric load perturbations from an initial stable condition. Even if the nuclear fuel has a small positive temperature coefficient of reactivity, the expected higher negative contributions from the rest of the system under normal operation conditions will give a highly damped response to a perturbation. The details depend on the specific design and delay times. The magnitude of the response is sensitive to the values of the temperature coefficients. The new equilibrium point is not an optimum point in the sense that the thermionic diodes are not operating at the highest efficiency for the new power input. For very small perturbation this would be tolerable, but for larger ones, in excess of 20% of the initial power output, reactivity adjustment should be done. This will also be necessary if a constant voltage or current output is to be maintained. This is different from the more conventional reactor applications where the reactor remains at a constant mean temperature without reactivity adjustment for large power perturbations taken out via the coolant.

Figs. 11 and 12 show the trajectories for an electric load perturbation in a typical case. The delay times were zero.

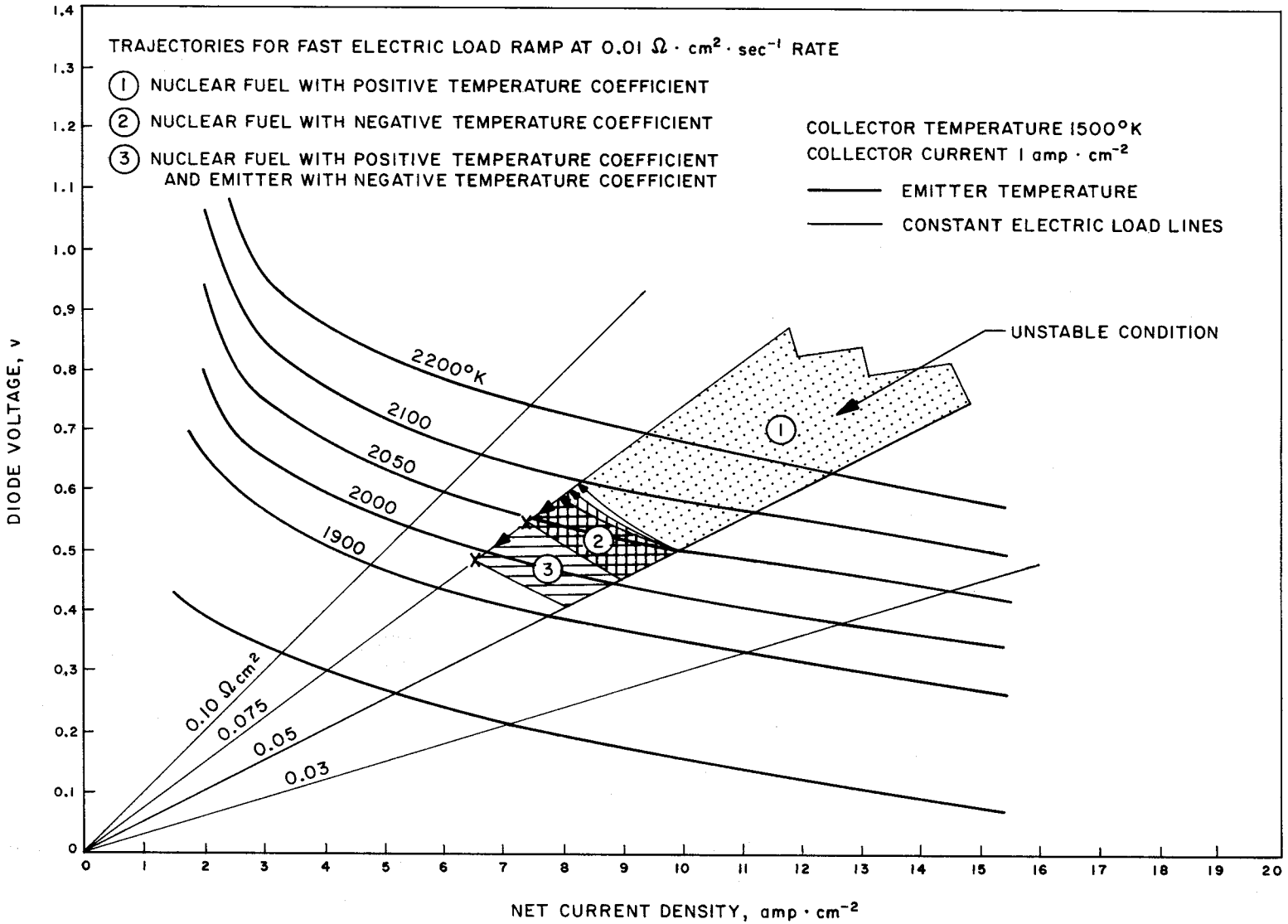


Fig. 11. Analog computer solution of thermionic diode characteristics

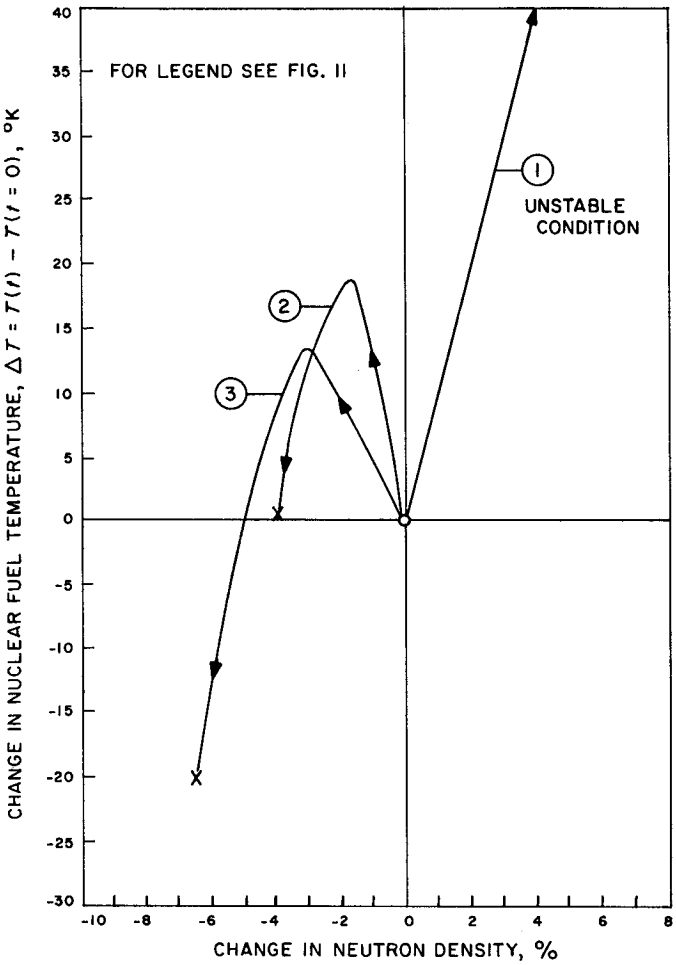


Fig. 12. Analog computer solution of change in nuclear temperature versus neutron density

D. Linear Stability Analysis of a Thermionic Powerplant

J. L. Shapiro

1. Introduction

A simplified, linearized formulation of the kinetic equations describing the behavior of a thermionic reactor has been developed. The purpose is to aid in the understanding of the results of the analog simulation (Ref. 10) as well as to estimate the possibility of spatial power instability.

2. Model

a. Neutron dynamics. The one-energy-group, one-delayed-neutron group, one-dimensional equations for a homogenized reactor are,

$$\frac{1}{v}\frac{\partial \phi\,(x,t)}{\partial t} = D\frac{\partial^2 \phi}{\partial x^2} + \left[(1-\beta)\nu\Sigma_f(T_f) - \Sigma_a(T_f) - \Sigma_a(T_c)\right]\phi + \lambda C\,(x,t) \tag{1}$$

$$\frac{\partial C\,(x,t)}{\partial t} = \beta\,\nu\,\Sigma_f\,\phi - \lambda\,C \tag{2}$$

using standard notation. The cross sections are assumed to be dependent on the fuel temperature T_f and the coolant temperature T_c.

b. Heat transfer. A two-region model of the reactor core is selected. The combination fuel-emitter region is separated by a void from the combined collector-coolant region. The coolant circulates through a heat exchanger in which the inlet temperature of the secondary coolant is fixed. The equations for this model may be in the form

$$V_f\rho_f c_{pf}\frac{\partial T_f\,(x,t)}{\partial t} = V_f q_f \Sigma_f\,(T_f)\,\phi - G_c(T_f) - H_f(T_c) \tag{3}$$

$$V_c\rho_c c_{pc}\frac{\partial T_c\,(x,t)}{\partial t} = G_c(T_f) + H_c(T_c) - wc_{pc}\Delta_r\,(x,t) \tag{4}$$

$$V_E\rho_E c_{pE}\frac{\partial T_E\,(t)}{\partial t} = -\,wc_{pc}\Delta_E - UA\,(T_E - T_0) \tag{5}$$

$$T_E(t) - \tfrac{1}{2}\Delta_E(t) = \frac{1}{H}\int_0^H T_c\,(x, t-\tau) + \tfrac{1}{2}\Delta_c\,(x, t-\tau)\,dx \tag{6}$$

$$T_c\,(x,t) - \tfrac{1}{2}\Delta_r\,(x,t) = T_E\,(t-\tau) + \Delta_E\,(t-\tau) \tag{7}$$

where the subscripts $_{f,\ c,\ E}$ refer to the fuel, coolant, and heat exchanger, respectively; q is the energy release per fission; Δ_r and Δ_E are the total temperature drops across reactor and heat exchanger, respectively; and τ is the transport delay time. The terms G and H account for the heat-transfer dependence on the fuel temperature and the collector temperature, respectively, and include the thermionic energy removal contribution.

If we expand the cross-sectional dependence in a Taylor series, keeping only the terms that are linear in temperature, and permit small perturbations of the variables, we can arrive at a set of linear equations for the perturbed quantities. Then we expand each perturbed

variable in the orthogonal modes which are eigenfunctions of the steady-state form of Eqs. (1) and (2). Thus the modal functions are solutions to

$$\alpha_m \psi_m(x) = D \frac{\partial^2 \psi_m(x)}{\partial x^2} + (\nu \Sigma_{f0} - \Sigma_{a0}) \psi_m(x) \qquad (8)$$

The new set of equations has each of the variables expressed in terms of an infinite sum. This set of infinite order may be reduced to a set of seventh order by multiplying each term of ψ_k and integrating over the entire reactor, thus utilizing the orthogonality property. This also requires the assumption of uniform flux and uniform fission cross section. Finally, Laplace transformation of the set yields a set of linear, algebraic equations whose determinant must vanish for each of the roots. The determinant appears as follows:

$$\begin{vmatrix}
\frac{p}{v} + \beta\nu\Sigma_f - \alpha_k & -\lambda & (1-\beta)\nu \frac{\partial \Sigma_f}{\partial T_f} - \frac{\partial \Sigma_a}{\partial T_f}\phi & -\frac{\partial \Sigma_a}{\partial T_c}\phi & 0 & 0 & 0 \\
-\beta\nu\Sigma_f & p+\lambda & \beta\nu \frac{\partial \Sigma_f}{\partial T_f}\phi & 0 & 0 & 0 & 0 \\
V_f q_f \Sigma_f & 0 & -V_f \rho_f c_{pf} p + V_f q_f \phi_0 \frac{\partial \Sigma_f}{\partial T_f} - \frac{\partial G_f}{\partial T_f} & -\frac{\partial H_f}{\partial T_c} & 0 & 0 & 0 \\
0 & 0 & \frac{-\partial G_c}{\partial T_f} & V_c \rho_c c_{pc} p - \frac{\partial H_c}{\partial T_c} & 0 & 0 & wc_{pc} \\
0 & 0 & 0 & 0 & V_E \rho_E c_{pE} p + UA & wc_{pc} & 0 \\
0 & 0 & 0 & 2e^{-p\tau} & -2 & 1 & e^{-p\tau} \\
0 & 0 & 0 & -2 & 2e^{-p\tau} & e^{-p\tau} & 1
\end{vmatrix}$$

The roots will be examined with the aid of a digital computer, using power series approximations for the exponential terms. Unstable conditions are associated with roots having a positive real part.

c. Further simplification. It is possible to perform hand calculations as a further check, under the assumptions that $\tau = 0$ or $\tau = \infty$. The determinant is then reduced to fifth order. The resulting algebraic equation will be analyzed, using the Routh-Hurwitz criterion for stability.

E. Approximate Thermal Radiation Properties of Spherical Cavities With and Without a Window in the Aperture

E. J. Roschke

1. Introduction

The use of a particular cavity designed for the measurement of total thermal radiation from an ionized gas was described in *SPS 37-36*, Vol. IV. This cavity was isolated from convective effects by means of a quartz window placed over the aperture. It was shown in a subsequent note (*SPS 37-37*, Vol. IV) that proper evaluation of the data could not be achieved if thermal-emission effects produced by the window were neglected. Window emission was shown to become increasingly important as the window temperature approached the cavity temperature. Thus, this effect is likely to be important in all similar applications where convective heat transfer to a window covering a cavity exists and is not negligible compared to the incident radiative flux. The problem is re-examined here in a more general way. Several simple and several limiting cases are presented to aid in establishing a more complete understanding of the window effect with and without convective heat input. The spherical cavity has been selected for study because of its simplicity, and because it possesses several convenient and unique characteristics. Only total radiation is dealt with; spectral response of cavities is not considered.

A definition of terms used appears at the end of this article.

2. Isothermal Cavity Without Window

It has been shown in Refs. 11 and 12 that the apparent absorptivity and the apparent emissivity of a cavity of

arbitrary shape are not equal unless the cavity is gray. In addition to the requirement of grayness, the incident radiation must be diffusely distributed, except in the case of the spherical cavity, which is independent of the directional distribution of the incident radiation. Exact expressions for α_a and ϵ_a are given for spherical cavities having diffusely emitting and absorbing walls (Refs. 11 and 12); these may be conveniently expressed as

$$\alpha_a = \alpha/[\alpha + f(1 - \alpha)] \quad (1)$$

$$\epsilon_a = \epsilon/[\alpha + f(1 - \alpha)] \quad (2)$$

where α and ϵ refer to the wall material and the flatness factor $f = A_1/S$. Note that A_1 is the area of the spherical cap subtended by the aperture, and S is the total internal area of the sphere. Thus, $\alpha_a/\epsilon_a = \alpha/\epsilon$, which becomes unity for gray cavities. Eq. (1) is plotted in Fig. 13 for various values of f. Values of ϵ_a may be found for a nongray cavity by multiplying the ordinate in Fig. 13 by the appropriate value of ϵ/α.

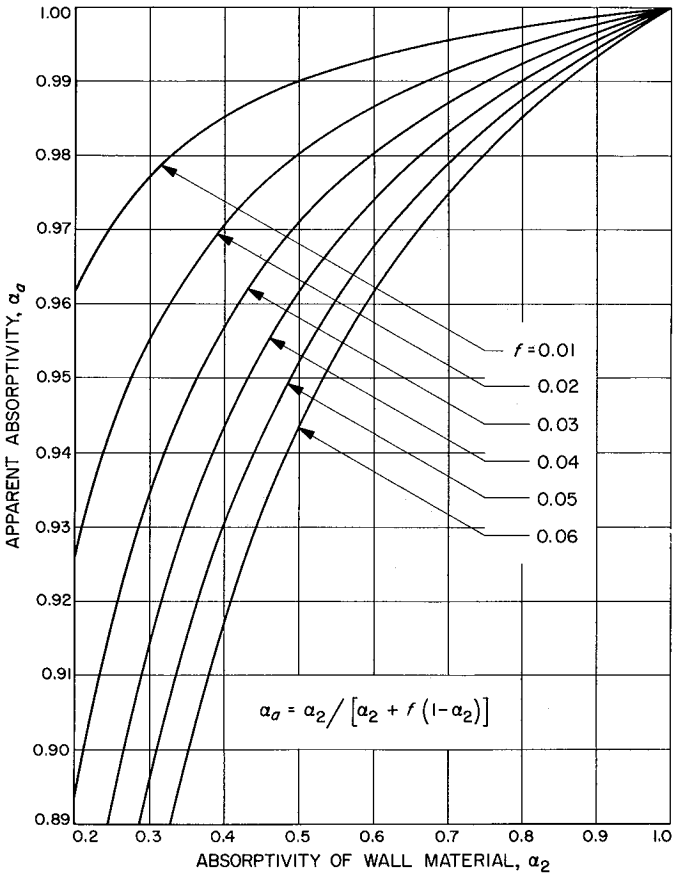


Fig. 13. Apparent absorptivity of isothermal cavity having no window in aperture

In Fig. 14(a), the cavity assumes an equilibrium temperature T_2 when subject to the incident radiative flux q_r and, if the cavity is not perfectly insulated, a loss flux denoted by q_ℓ. Emission from the cavity aperture is denoted by $\epsilon_a q_{21}$. The energy balance may be written as

$$A_1\alpha_a q_r = A_1\epsilon_a q_{21} + A_2 q_\ell$$

or

$$q_r = (\epsilon_2/\alpha_2)\sigma T_2^4 + (A_2/A_1)\,(q_\ell/\alpha_a) \quad (3)$$

The radiative flux required to produce T_2 for various values of α_2/ϵ_2 is shown in Fig. 15[1] for the case of the

[1]An error has been found in *SPS 37-36*, Vol. IV, Fig. 4, p. 91. Labels on the curves should read ϵ_2/α_2 rather than α_2/ϵ_2.

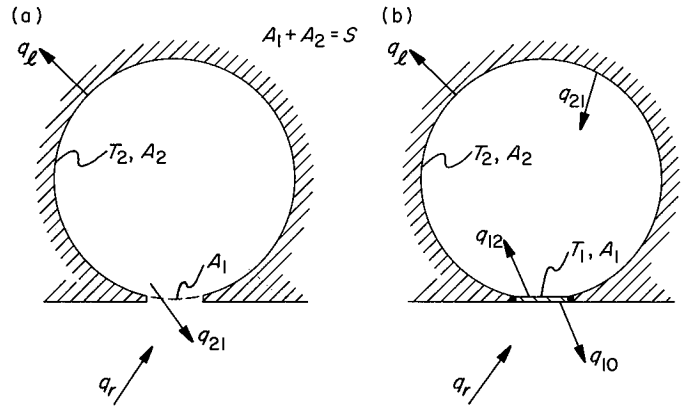


Fig. 14. Isothermal spherical cavities (a) no window (b) insulated window

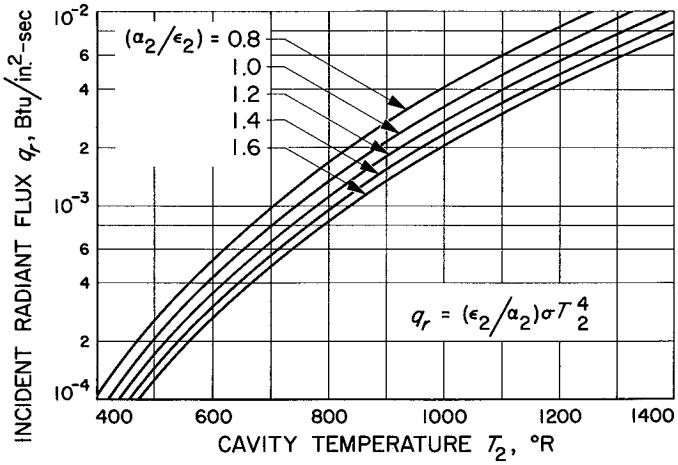


Fig. 15. Thermal radiation absorption by isothermal perfectly insulated cavity with no window in aperture

perfectly insulated cavity, $q_\ell = 0$. If q_ℓ is known as a function of q_r, or if the loss term $(A_2/A_1)q_\ell$ is expressed as a fraction K_0 of q_r, then

$$q_r[1 - (K_0/\alpha_a)] = (\epsilon_2/\alpha_2)\sigma T_2^4, \; 0 \leq (K_0/\alpha_a) < 1 \tag{4}$$

Fig. 15 may be used to obtain q_r if values read from the ordinate are divided by the factor $1 - (K_0/\alpha_a)$. Similarly, if $(A_2/A_1)q_\ell$ is expressed in terms of emission from the cavity wall $\epsilon_2\sigma T_2^4$ by a factor K_2, i.e., by $K_2\epsilon_2\sigma T_2^4$, then

$$q_r/[1 + K_2(\alpha_2/\alpha_a)] = (\epsilon_2/\alpha_2)\sigma T_2^4, \; 0 \leq K_2 < \infty \tag{5}$$

In this case, Fig. 15 again may be used to obtain q_r if values read from the ordinate are multiplied by $1 + K_2(\alpha_2/\alpha_a)$.

3. Isothermal Cavity With Insulated Isothermal Window

In Fig. 14(b), a thin, flat window of area A_1 is assumed to transmit uniformly a portion of the incident radiative flux q_r. At equilibrium, the cavity wall of area A_2 assumes a temperature T_2 and the window assumes a temperature T_1. All surfaces are assumed to absorb and emit diffusely to avoid refraction problems caused by the window; the window is assumed to emit radiation from both surfaces as it were a thin, opaque plate. Heat transfer to and from the window is by radiation only, since its edges are considered to be perfectly insulated. The medium in the cavity, e.g., a gas at low pressure, is assumed to be nonabsorbent and perfectly transmitting. Since the window is totally enclosed, the view factor from its rear surface to the cavity is unity. From view factor relations, it is easily shown that the view factor from cavity wall to window is A_1/A_2, so that the view factor from the cavity wall to itself is $1 - (A_1/A_2)$. The external view factor from window to environment is assumed unity. Flux terms q_{10}, q_{12}, and q_{21} in Fig. 14(b) include all radiative components of flux leaving the surfaces indicated. For example, q_{12} consists of emission, transmission of a portion of q_r, and reflection of a portion of q_{21}. Energy balances on the cavity and window, respectively, are

$$A_1(q_{12} - q_{21}) = A_2 q_\ell \tag{6}$$

and

$$A_1(q_r - q_{10}) = A_1(q_{12} - q_{21}) \tag{7}$$

The leaving-flux expressions are written as

$$A_1 q_{12} = A_1 e_{12} + A_1 t_{12} + A_2(A_1/A_2)(1 - \alpha_1 - \tau_1)q_{21} \tag{8}$$

$$A_2 q_{21} = A_2 e_{21} + A_1(1 - \alpha_2)q_{12} + A_2(1 - A_1/A_2)(1 - \alpha_2)q_{21} \tag{9}$$

$$A_1 q_{10} = A_1 e_{10} + A_1 t_{10} + A_1(1 - \alpha_1 - \tau_1)q_r \tag{10}$$

where $e_{10} = e_{12} = \epsilon_1\sigma T_1^4$, $\quad e_{21} = \epsilon_2\sigma T_2^4$

$$t_{12} = \tau_1 q_r \text{ and } t_{10} = \tau_1 q_{21}$$

Solutions of Eqs. (6), (8), and (9) yields,

$$\tau_1 q_r = \frac{(\alpha_1 + \tau_1)}{\alpha_2} e_{21} - e_{12} + \left(\frac{A_2}{A_1}\right) q_\ell \times \left[1 + \frac{(\alpha_1 + \tau_1)(1 - \alpha_2)}{\alpha_2}\left(\frac{A_1}{A_2}\right)\right] \tag{11}$$

which may then be used to obtain q_{10} from Eq. (10). Solution of Eq. (7) then gives

$$(\alpha_1 + \tau_1)q_r = \frac{\tau_1}{\alpha_2} e_{21} + e_{10} + \left(\frac{A_2}{A_1}\right)q_\ell \times \left[1 + \frac{\tau_1(1 - \alpha_2)}{\alpha_2}\left(\frac{A_1}{A_2}\right)\right] \tag{12}$$

Eliminating $e_{10} = e_{12}$ from Eqs. (11) and (12) yields, finally,

$$q_r = \frac{e_{21}}{\alpha_2} + \left(\frac{A_2}{A_1}\right) q_\ell \left[\frac{2}{(\alpha_1 + 2\tau_1)} + \frac{(1 - \alpha_2)}{\alpha_2}\left(\frac{A_1}{A_2}\right)\right] \tag{13}$$

Comparison of Eqs. (13) and (3) shows that the apparent absorptivity of the cavity with window in the aperture is

$$\alpha_{aw} = \alpha_2/\{2\alpha_2/[(\alpha_1 + 2\tau_1)] + (A_1/A_2)(1 - \alpha_2)\} \tag{14}$$

and the apparent emissivity may be found from the relation $\epsilon_{aw}/\alpha_{aw} = \epsilon_2/\alpha_2$.

Thus, the relationship between q_r, T_2 and q_ℓ with window present is identical in form to the relation obtained without window [Eq. (3)], that is

$$q_r = (\epsilon_2/\alpha_2)\sigma T_2^4 + (A_2/A_1)(q_\ell/\alpha_{aw}) \tag{15}$$

where α_{aw} is given by Eq. (14). If $\alpha_1 \ll \tau_1$ and $A_1 \ll A_2$ so that $f = A_1/S \simeq A_1/A_2$, Eq. (14) becomes

$$\alpha_{aw} = \alpha_2/[(\alpha_2/\tau_1) + f(1 - \alpha_2)] \tag{16}$$

Comparing the case of window with negligible absorption to the case without window, it is found that

$$\alpha_{aw}/\alpha_a = [\alpha_2 + f(1 - \alpha_2)]/[(\alpha_2/\tau_1) + f(1 - \alpha_2)] \tag{17}$$

Hence, when $\alpha_2 \to 1$,

$$\alpha_{aw}/\alpha_a \sim \tau_1 \tag{18}$$

Eq. (17) is plotted in Fig. 16 for one value of f. Values of α_{aw} may be obtained by multiplying the ordinate read from Fig. 13 by that read from Fig. 16.

A comparison of the equilibrium values of window temperature with cavity temperature is of interest. Solution of either of Eqs. (11) or (12) with Eq. (13) yields

$$e_{10} = e_{12} = \epsilon_1 \sigma T_1^4 = \alpha_1 [q_r - (A_2/A_1) q_l/(\alpha_1 + 2\tau_1)] \tag{19}$$

from which the window temperature may be determined if q_r and q_l are known. Eq. (19) shows that when $\alpha_1 = 0$, i.e., zero absorption $T_1 = 0$. If there are no cavity losses, $q_l = 0$, Eqs. (13) and (19) show that

$$\epsilon_1 \sigma T_1^4 = \alpha_1 q_r = (\alpha_1/\alpha_2)\epsilon_2 \sigma T_2^4$$

hence,

$$T_1 = (\alpha_1 \epsilon_2/\epsilon_1 \alpha_2)^{1/4} T_2 \tag{20}$$

For a gray cavity $\alpha_2 = \epsilon_2$, so that

$$T_1 = (\alpha_1/\epsilon_1)^{1/4} T_2 \tag{21}$$

This relation is plotted in Fig. 17 for various values of α_1/ϵ_1.

Just as in the case without window, it may be convenient or appropriate to express q_l in terms of either q_r as in Eq. (4), or in terms of $\epsilon_2 \sigma T_2^4$ as in Eq. (5). Parallel expressions then become

$$q_r [1 - (K_0/\alpha_{aw})] = (\epsilon_2/\alpha_2)\sigma T_2^4, \; 0 \leq (K_0/\alpha_{aw}) < 1 \tag{22}$$

and,

$$q_r/[1 + K_2(\alpha_2/\alpha_{aw})] = (\epsilon_2/\alpha_2)\sigma T_2^4, \; 0 \leq K_2 < \infty \tag{23}$$

Again, Fig. 15 may be utilized to simplify calculations by dividing or multiplying the ordinate by the appropriate factor appearing in Eqs. (22) and (23).

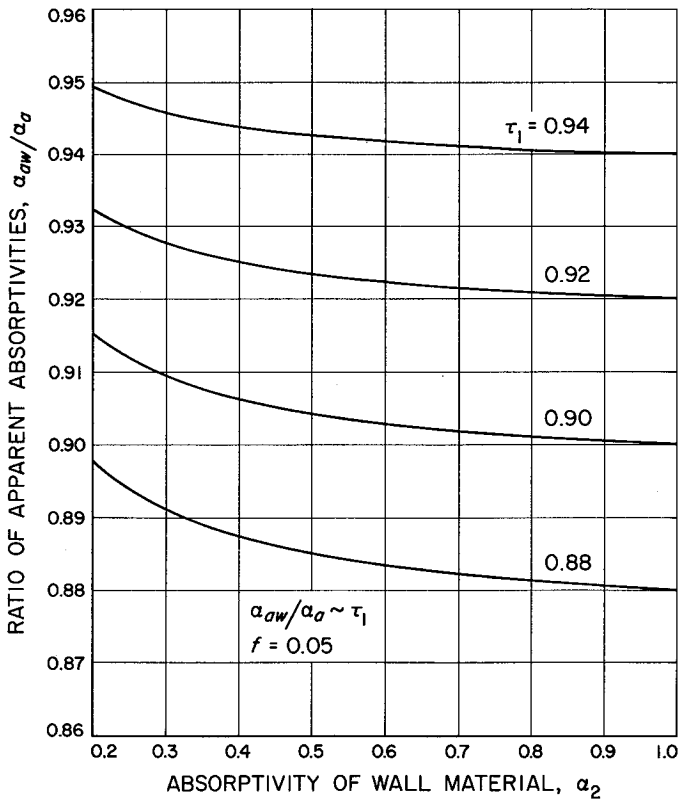


Fig. 16. Effect of window transmissivity on the apparent absorptivity of isothermal cavity when window absorption is zero

4. Further Comparisons

Two additional comparisons of cavities with and without a window are of interest: (a) a comparison of cavity

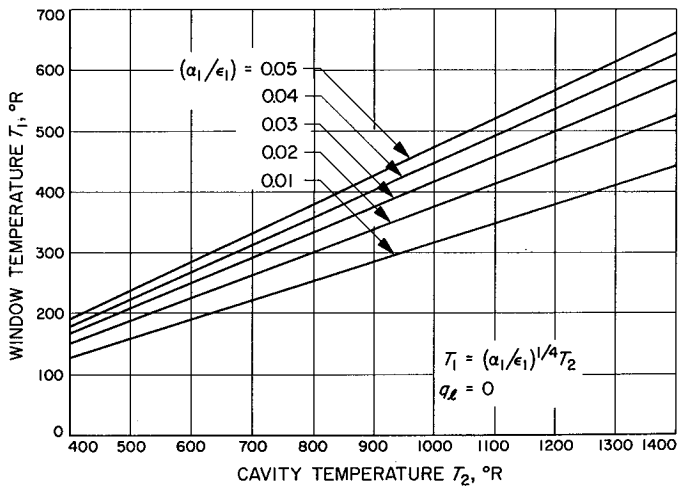


Fig. 17. Equilibrium temperature of insulated window for a gray perfectly insulated cavity

temperatures T_2 for identical incident radiant flux q_r and (b) a comparison of q_r values required to produce identical cavity temperatures. In either case the results are trivial, but identical, if $q_\ell = 0$; Fig. 15 applies without modification. If $q_\ell \neq 0$ but $(A_2/A_1)q_\ell = K_0 q_r$ and $(A_2/A_1)q_{\ell w} = K_{0w} q_{rw}$ respectively, then case (a) for $q_r = q_{rw}$ is simply expressed. Use of Eqs. (4) and (22) yields

$$\left(\frac{T_2}{T_{2w}}\right)^4 = \left(\frac{\alpha_{aw}}{\alpha_a}\right)\left[\frac{1 - (K_0/\alpha_a)}{(\alpha_{aw}/\alpha_a) - (K_{0w}/\alpha_a)}\right] \quad (24)$$

Under the approximations of Eqs. (17) and (18), Eq. (24) may be approximated by

$$\left(\frac{T_2}{T_{2w}}\right)^4 \sim \tau_1\left[\frac{1 - (K_0/\alpha_a)}{\tau_1 - (K_{0w}/\alpha_a)}\right], q_r = q_{rw} \quad (25)$$

If $K_0 = K_{0w}$, $T_2 > T_{2w}$ for $q_r = q_{rw}$. In Ref. 3 it was shown that the true temperature of a cavity having a window is less than the apparent temperature as viewed by an optical pyrometer. That effect was caused by window transmittance less than unity.

In case (b), it is likely that $q_\ell = q_{\ell w}$ when $T_2 = T_{2w}$, i.e., $K_2 = K_{2w}$. Hence,

$$\left(\frac{q_r}{q_{rw}}\right) = \left(\frac{\alpha_{aw}}{\alpha_a}\right)\left[\frac{1 + (K_2\alpha_2/\alpha_a)}{(\alpha_{aw}/\alpha_a) + (K_2\alpha_2/\alpha_a)}\right] \quad (26)$$

or, approximately

$$\left(\frac{q_r}{q_{rw}}\right) \sim \tau_1\left[\frac{1 + (K_2\alpha_2/\alpha_a)}{\tau_1 + (K_2\alpha_2/\alpha_a)}\right], T_2 = T_{2w} \quad (27)$$

Thus, $q_r < q_{rw}$ for $T_2 = T_{2w}$.

5. Isothermal Cavity With Convective Heat Input to Window

The results presented up to now have shown that cavity temperature is not significantly affected by the presence of a highly transmitting window in the aperture of an isothermal cavity when the window temperature is relatively low in comparison. This condition will prevail as long as energy transfer to the window is by the radiant mode only, and the window material has low overall absorption of the incident radiation. When, however, the window is heated externally by another source, such as gas convection, the results can be markedly different, as was shown in *SPS 37-37*, Vol. IV. If convective heat addition is sufficient to raise the equilibrium window temperature to a value equal to or greater than the cavity temperature, then the window emission becomes significant compared to other terms in the energy balances. Window emission will become important for quartz and most glass windows at least in the range from room temperature to nearly 1500°R, in which these materials have a relatively high, total emissivity.

The effects of convective heat input to the window will be further developed in this section so that direct comparisons with the previous results can be made. For illustrative purposes, a uniform heat flux is assumed to be applied convectively to the window of Fig. 14(b). The window is perfectly insulated so that no conductive, edge losses are incurred. Hence, in addition to q_r another input q_c is applied.

Eqs. (6) and (8) to (10) remain unaltered. Eq. (7) is modified as follows:

$$A_1(q_r + q_c - q_{10}) = A_1(q_{12} - q_{21}) \quad (7a)$$

The equations are solved as before to yield

$$q_r = \frac{e_{21}}{\alpha_2} + \left(\frac{A_2}{A_1}\right) q_\ell \left[\frac{2}{(\alpha_1 + 2\tau_1)} + \frac{(1 - \alpha_2)}{\alpha_2}\left(\frac{A_1}{A_2}\right)\right] - \frac{q_c}{(\alpha_1 + 2\tau_1)} \quad (28)$$

which is similar to Eq. (13) except for the convective term. It is not appropriate to define an apparent absorptivity for this case, because α_{ac} would depend on factors other than the optical and geometric properties of the cavity and window. However, in analogy to Eq. (15), the following could be written,

$$q_r = \left(\frac{\epsilon_2}{\alpha_2}\right)\sigma T_2^4 + \left(\frac{A_2}{A_1}\right)\left(\frac{q_\ell}{\alpha_{aw}}\right) - \frac{q_c}{(\alpha_1 + 2\tau_1)} \quad (29)$$

where α_{aw} is given by Eq. (14). In comparison to Eq. (19), e_{10} is now given by

$$e_{10} = e_{12} = \epsilon_1\sigma T_1^4 = \alpha_1\left[q_r - \left(\frac{A_2}{A_1}\right)q_\ell/(\alpha_1 + 2\tau_1)\right] + \frac{(\alpha_1 + \tau_1)}{(\alpha_1 + 2\tau_1)} q_c \quad (30)$$

It is evident, of course, that $T_1 \neq 0$ even when window absorption is zero with convective heat transfer to the window. The window temperature T_1 is conveniently examined by placing $q_\ell = 0$. Eqs. (29) and (30) can be combined to eliminate either q_r or q_c. In the former case the following relation is obtained,

$$\epsilon_1\sigma T_1^4 = \alpha_1 q_r + (\alpha_1 + \tau_1)\, q_c/(\alpha_1 + 2\tau_1)$$
$$= (\alpha_1\epsilon_2/\alpha_2)\,\sigma T_2^4 + \tau_1\, q_c/(\alpha_1 + 2\tau_1)$$

from which

$$T_1 = \left[(\alpha_1\epsilon_2/\epsilon_1\alpha_2)\, T_2^4 + (\tau_1 q_c/\sigma\epsilon_1)/(\alpha_1 + 2\tau_1)\right]^{1/4} \quad (31)$$

Eq. (31) may be compared to Eq. (20). For a gray cavity with negligible window absorption, Eq. (31) becomes

$$T_1 = \left[q_c/(2\sigma\epsilon_1)\right]^{1/4} \quad (32)$$

In the latter case,

$$q_r = \sigma\left[T_2^4 - (\epsilon_1/\tau_1)\, T_1^4\right] \quad (33)$$

and the condition for $q_r = 0$ occurs when

$$T_1 = (\tau_1/\epsilon_1)^{1/4}\, T_2 \quad (34)$$

If the convective heat flux is described by $q_c = K_1 q_r$ and a gray, perfectly insulated cavity with zero window absorption is considered, then Eq. (28) reduced to

$$q_r = \sigma T_2^4/(1 + K_1/2\tau_1),\ K_1 \geq 0 \quad (35)$$

Since the expression $q_r = \sigma T_2^4$ is valid for a gray, perfectly insulated cavity, with or without a window when $q_c = 0$, the following ratio

$$\left(\frac{q_r}{q_{rc}}\right) = 1 + K_1/2\tau_1,\ T_2 = T_{2c} \quad (36)$$

yields a measure of the convective effect. It is easy to see that K_1 need not be large to have a marked influence on the results. If $K_1 \gg 1$, convection dominates, and the incident radiant flux on the window has little influence on either the window temperature or the cavity temperature.

Considering similar circumstances to those above, except that equality in incident radiative flux is assumed rather than equivalent cavity temperatures, it is easily shown that

$$(T_{2c}/T_2)^4 = 1 + K_1/2\tau_1,\ q_r = q_{rc} \quad (37)$$

6. Summary

The thermal radiation properties of spherical cavities with and without windows in the aperture have been compared. If the window has low absorptivity but high transmissivity for the incident radiant flux, the window has little effect on the cavity temperature. It is found that the apparent absorptivity of the cavity is lowered by a multiplying factor approximately equal to the transmissivity of the window when a window is placed in the aperture. If the incident radiative flux remains unchanged in this process, it is found that the cavity temperature decreases slightly.

The purpose of placing a window in the aperture of a cavity may be to isolate the cavity from external convective heat-transfer effects. It is shown that convection need not be large compared to the incident radiation to cause significant changes in the cavity temperature. The last result is due to thermal emission from the window.

Definitions of terms

A	Surface area
e	Heat flux, by thermal emission
f	Flatness factor of cavity $= (A_1/S)$
K_0	Factor defined by the ratio $(A_2/A_1)\, q_l/q_r$
K_1	Factor defined by the ratio q_c/q_r
K_2	Factor defined by the ratio $(A_2/A_1)\, q_l/\epsilon_2\sigma T_2^4$
q	Heat flux, thermal energy per unit time per unit area
S	Total surface area of spherical cavity $= (A_1 + A_2)$
t	Heat flux, by transmission
T	Absolute temperature
α	Absorptivity
ϵ	Emissivity
σ	Stefan-Boltzmann constant
τ	Transmissivity

Subscripts

a	Apparent
c	Convection component, or refers to case with convection
l	Loss term
r	Radiative component
w	Refers to case with window present
0	Environment viewed by cavity
1	Window surface and/or aperture opening surface
2	Cavity surface

References

1. Mettler, R. F., "The Anemometric Application of an Electrical Glow Discharge in Transverse Air Streams," Thesis, California Institute of Technology, 1949; Werner, I. D., *Review of Scientific Instruments,* Vol. 21, p. 61, 1950; Lawrence, A. B., Final Report, ONR-478 (Contract ONR N-7), Office of Naval Research, Washington, D.C., 1951.

2. Chen, C. J., *Physical Review Letters,* Vol. 16, p. 833, 1966.

3. Ashargan, G. A., *JETP Letters,* Vol. 1, p. 162, 1965.

4. Ladenburg, R. W., Lewis, B., Pease, R. N., and Tayler, H. S., "Physical Measurements in Gas Dynamics and Combustion," p. 142, Princeton University Press, Princeton, N.J., 1954.

5. Elliott, D. G. Cerini, D. J., Hays, L. G., and Weinberg, E., "Theoretical and Experimental Investigation of Liquid-Metal MHD Power Generation," Paper No. SM-74/177, International Symposium on Magnetohydrodynamic Electrical Power Generation, Salzburg, Austria, July 4–8, 1966.

6. Elliott, D. G., Cerini, D. J., and Weinberg, E., "Liquid-Metal MHD Power Conversion," *Progress in Astronautics and Aeronautics: Space Power Systems Engineering,* Vol. 16, Academic Press, New York, N. Y., February 1966, pp. 1275–1298.

7. Elliott, D. G., "Direct-Current Liquid-Metal Magnetohydrodynamic Power Generation," *AIAA Journal,* Vol. 4, p. 4, April 1966, pp. 627–634.

8. Eddington, R. B., "Investigation of Shock Phenomena in a Supersonic Two-Phase Tunnel," Paper No. 66–87, AIAA 3rd Aerospace Sciences Meeting, New York, N. Y., January 24–26, 1966.

9. Weinberg, E., and Hays, L. G., "Comparison of Liquid-Metal Magnetohydrodynamic Power Conversion Cycles," Technical Report No. 32–946, Jet Propulsion Laboratory, Pasadena, Calif., July 1, 1966.

10. Gronroos, H., Kikin, G., "Analog Study of Thermionic Reactor Stability," *SPS 37-35,* Vol. IV, Jet Propulsion Laboratory, Pasadena, California, October 31, 1965, pp. 180 to 190.

11. Sparrow, E. M., "Radiant Emission, Absorption and Transmission Characteristics of Cavities and Passages," Symposium on Thermal Radiation of Solids, San Francisco, March 1964. Proceedings published as NASA SP-55, S. Katzoff, Ed., National Aeronautics and Space Administration, Washington, D.C., pp. 103–115, 1965.

12. Sparrow, E. M., and Jonsson, V. K., "Absorption and Emission Characteristics of Diffuse Spherical Enclosures," *Transactions of the ASME,* Series C: *Journal of Heat Transfer,* Vol. 84, No. 2, pp. 188–189, May 1962.

13. Flemming, J. C., "An Evaluation of a High-Temperature Blackbody as a Working Standard of Spectral Radiance," *Applied Optics,* Vol. 5, No. 2, pp. 195–200, February 1966.

XI. Liquid Propulsion

A. Injector Development: Investigation of Propellant Sheets

R. W. Riebling

1. Introduction

When a jet of liquid from a circular orifice is directed tangentially against a solid, concave deflector surface, it spreads to form a thin (0.005- to 0.010-in.-thick) liquid sheet of width w (Fig. 1). Upon leaving the deflectors, these free sheets spread through an angle β before finally breaking up into droplets. However, the sheets do not generally exit tangentially. Rather, the axes of the sheets are deflected through an angle δ away from the tangents to the deflector surfaces (Fig. 1).

Injector elements incorporating this effect have been under development at JPL for some time (*SPS 37-31*, Vol. IV, p. 203, and *SPS 37-35*, Vol. IV, p. 152), because they offer certain advantages over the more conventional impinging-jet varieties. An applied research program is currently under way to determine the manner in which the dimensions and spatial orientation of such propellant sheets, as well as the velocity and mass distribution within them, vary with deflector geometry, injection velocity, and propellant physical properties. Such information is necessary to the intelligent design of all injector elements based on the thin-sheet concept.

Experimental results concerning sheet dimensions and orientation were presented in *SPS 37-38*, Vol. IV, p. 116;

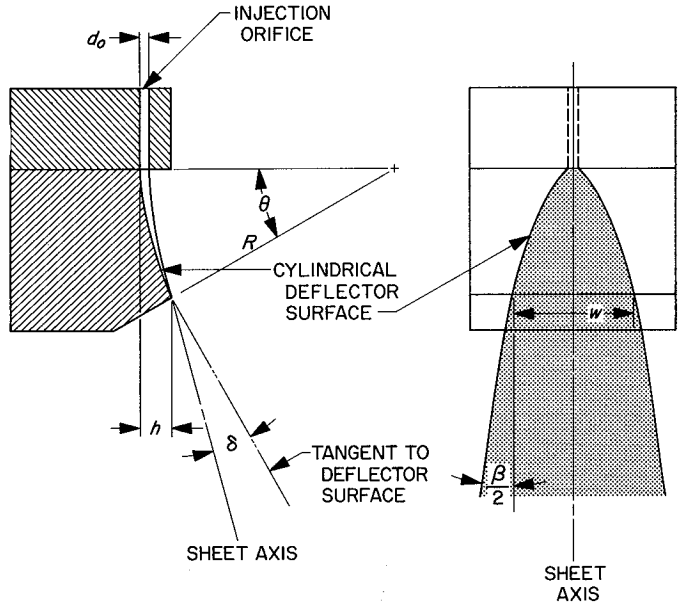


Fig. 1. Experimental apparatus for study of propellant sheets

apparatus and procedures were described in *SPS 37-37*, Vol. IV, p. 162. Presented here are a further analysis of the sheet dimension data, additional information regarding mass distribution within the sheets, and some new results of an investigation into the time variation of sheet properties.

2. Sheet Dimensions and Orientation

Empirical correlations for the width w, spreading angle β, and deflection angle δ of a free-flowing liquid sheet formed by tangential impingement of a jet from a circular orifice on a concave cylindrical deflector surface were developed in *SPS 37-38*. These correlations are repeated here for convenience.

The width of the sheets as they leave the deflectors (Fig. 1) is given by

$$w/d_o = 5.8\,(h/d_o)^{1/2}, \tag{1}$$

where

$$h = R\,(1 - \cos\theta), \tag{2}$$

R is the deflector radius, and d_o is the orifice diameter. The spreading angle β (Fig. 1) is correlated by

$$\beta = \gamma\left[1 + (1.52 \times 10^{-3})\,V\right] - 30.7 \ln\left(\frac{R}{d_o}\right), \tag{3}$$

where V is the injection velocity and γ is a weak function of liquid surface tension. The deflection angle δ was shown to depend only upon h/d_o, but the relationship was not simple enough to be expressed in the form of a power-law type of equation.

All three sheet properties were found to be essentially independent of the liquids employed and their injection velocities over ranges commonly encountered in liquid-rocket-engine technology. Thus, deflector geometry (R, θ, and d_o) exerts the primary influence on sheet dimensions and orientation. This may initially seem surprising, since, from dimensional considerations, it would be expected that these sheet properties should be functions of the velocity and liquid properties, as well as of geometric factors. Thus, if gravity and air-drag effects are dismissed as negligible for the regions of interest, we have

$$w = f_1\,[R, d_o, \theta, V, \rho, \mu, \sigma] \tag{4}$$

or, by dimensional analysis of these pertinent parameters,

$$\frac{w}{d_o} = f_2\left[\frac{R}{d_o}, \theta, Re, We\right]. \tag{5}$$

Similar expressions would be expected to hold for all geometrical properties of the jet, e.g.,

$$\delta = f_3\left[\frac{R}{d_o}, \theta, Re, We\right]. \tag{6}$$

However, the Reynolds number (Re) and Weber number (We) effects are slight over the regions of interest. This is illustrated in Figs. 2 and 3, where the geometrical parameters of the sheet are plotted against Re and We, respectively, for several typical fixed values of hardware geometry (θ and R/d_o).

Unfortunately, it is not possible to hold We constant for some fixed deflector geometry while systematically varying Re, or vice versa. A change in We can only be made by varying velocity, physical properties, or geometry, any of which variations will bring about a corresponding change in Re. However, it may be inferred from the plots of Figs. 2 and 3 that the near invariance of the sheet dimensions with Re and We accounts for the observed second-order influence of fluid velocity and physical properties. Thus, in the region of interest:

$$\left\{\begin{array}{l} 2 \leqslant R/d_o \leqslant 40 \\ 15 \leqslant \theta, \text{deg} \leqslant 45 \\ 10^4 \leqslant Re \leqslant 10^5 \\ 50 \leqslant We \leqslant 10^3 \end{array}\right\},$$

we have nearly complete geometrical similarity, allowing functional relationships such as Eq. (5) above to be very greatly simplified from a four-parameter to a two-parameter dependence; i.e.,

$$\frac{w}{d_o} \cong f_1\left[\frac{R}{d_o}, \theta\right], \tag{7}$$

$$\beta \cong f_2\left[\frac{R}{d_o}, \theta\right], \tag{8}$$

$$\delta \cong f_3\left[\frac{R}{d_o}, \theta\right]. \tag{9}$$

The empirically determined Eqs. (1) and (2) are consistent with the functional dependence of Eq. (7). Furthermore, the results presented in *SPS 37-38*, which showed the deflection angle δ as a unique function of h/d_o, are consistent with Eq. (9). Finally, the very weak dependence of β on surface tension and velocity is not consistent with Eq. (8), but the deviation is small. It is interesting to note that the dependence of β on θ that is permissible by Eq. (8) is negligibly weak [θ is absent from the empirically

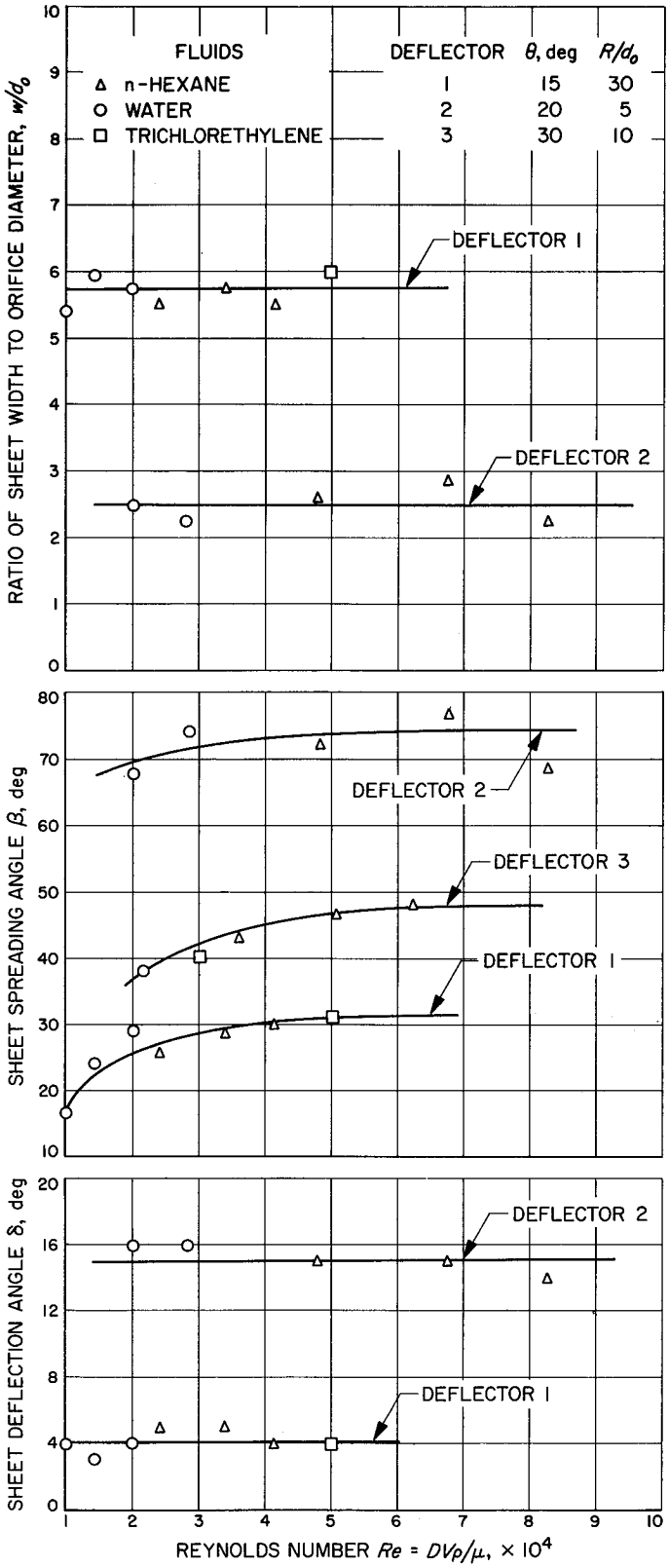


Fig. 2. Effect of Reynolds number on sheet dimensions and orientation for typical deflector geometries

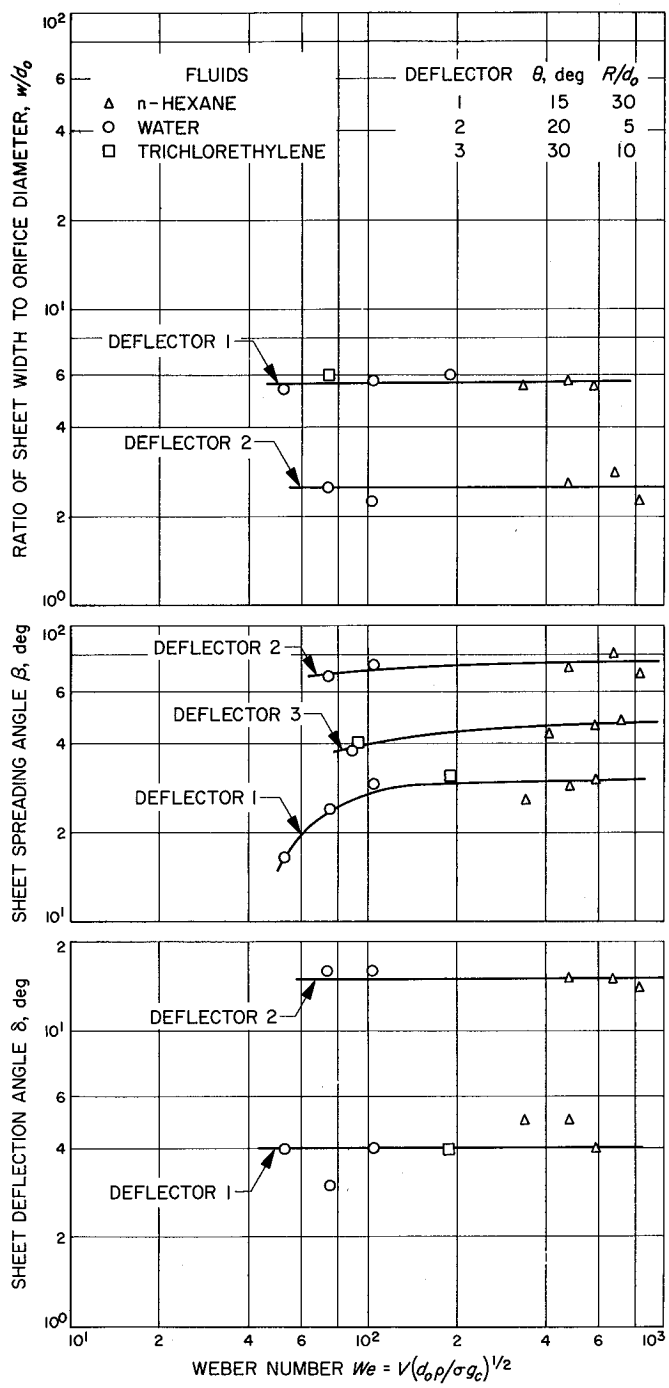


Fig. 3. Effect of Weber number on sheet dimensions and orientation for typical deflector geometries

determined Eq. (3)]. This is reasonable if one considers a crude model of the flowing sheet: assume the curved sheet to be flattened into a plane surface and to emanate from a point source (infinitely small orifice), as shown in Fig. 4. By inspection,

$$\tan\left(\frac{\beta}{2}\right) = \frac{w}{2L}, \tag{10}$$

where the deflector arc length

$$L = \frac{\theta \pi R}{180}. \tag{11}$$

Substituting Eqs. (1), (2), and (11) in Eq. (10) gives

$$\tan\left(\frac{\beta}{2}\right) = \frac{k(1 - \cos\theta)^{1/2}}{\theta (R/d_o)^{1/2}}. \tag{12}$$

Eq. (12) contains R/d_o and θ, and is therefore of the general form of Eq. (8). However, the term $(1 - \cos\theta)^{1/2}/\theta$ has the unique property of being essentially constant for $0 < \theta$, deg < 45. Accordingly, Eq. (12) simplifies to

$$\tan\left(\frac{\beta}{2}\right) \cong \frac{k'}{(R/d_o)^{1/2}}. \tag{13}$$

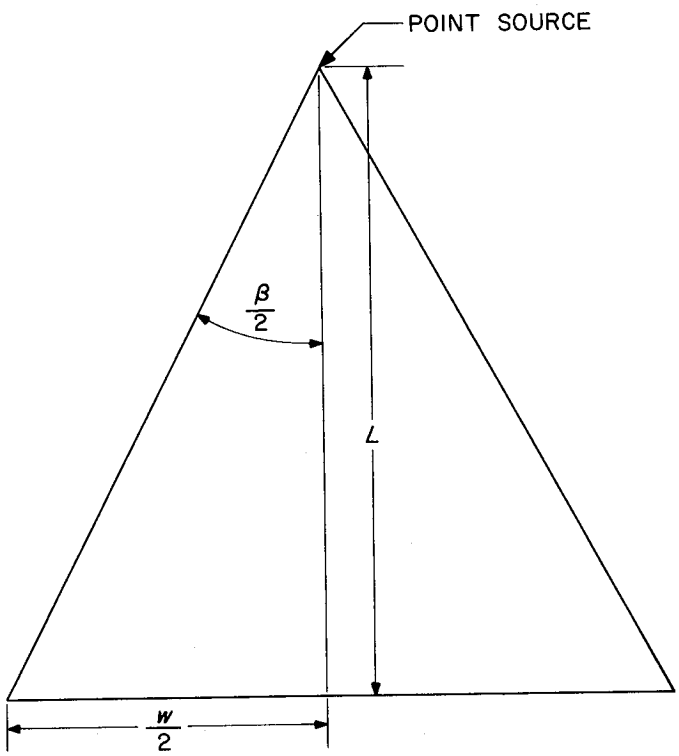


Fig. 4. Sheet geometry on deflector (simplified)

This is similar to the empirically found relationship of Eq. (3) in that $\beta \cong f(R/d_o, \theta)$ only. Plotting tan $(\beta/2)$ versus R/d_o shows the correct data trend, but, because of the simplifying assumptions involved, Eq. (13) does not correlate the experimental data as well as the empirically determined Eq. (3) or its simplified version:

$$\beta \cong 120 - 30.7 \ln\left(\frac{R}{d_o}\right). \tag{14}$$

3. Effects of Orifice Configuration on Sheet Dimensions

Thus far in the discussion of sheet dimensions, it has been presumed that these properties are time-independent. Such, however, may not be the case. The sheet width w and spreading angle β can vary randomly with time if care has not been taken to assure controlled, reproducible hydraulics in the manifold and orifice entrance. High-speed motion pictures reveal a random "writhing" motion of the sheet on the deflector when high manifold crossflow velocities and sudden orifice entrance transition conditions are combined with short orifice lengths. Through the use of long ($L_o/d_o = 50$ or greater) orifice tubes which assure fully developed turbulent flow, the time-varying flow perturbations resulting from poor manifold design can be smoothed enough to effectively eliminate the time-variation of sheet properties. The magnitude of sheet boundary variations is not very great in the former case, so that, on the average, w and β are the same regardless of orifice hydraulics.

Some of the sheet dimension experiments reported in *SPS 37-38* (originally conducted with drilled-hole orifices with length-to-diameter ratios of 6 to 12) were repeated at L_o/d_o values of 110 to assure that fully developed turbulent flow was attained. The values of w and β found with these longer orifices were, in every case, correlated by Eqs. (1) and (3), which were developed for the shorter orifices. There was considerably more scatter in the short-orifice data; however, this is not surprising since, in the short-orifice case, each dimensional data point (an instantaneous value taken from a single sheet photograph) can represent a maximum, minimum, or intermediate sheet dimension, depending on the moment the picture was taken. Eqs. (1) and (3), which are "best-fit" correlating equations, would then be expected to give time-averaged values for the shorter orifice lengths.

These results indicate that entrance conditions and orifice length exert little, if any, influence on the time-

averaged sheet dimensions. On the other hand, the existence of randomly varying propellant sheet properties can make complete description and control of the injection (and, thus, the combustion) process difficult, if not impossible. Predictable and controlled propellant injection therefore presumes controlled hydraulics, such as that obtained with fully developed turbulent flow, with its characteristic reproducibility and concomitant time-invariant sheet properties.

It should be pointed out that large orifice length-to-diameter ratios (which may involve fabrication problems) are not the only way to produce controlled injector hydraulics. Turbulence-inducing sections can produce equivalent degrees of developed turbulence in much shorter orifice passages (e.g., Ref. 1). Proper manifold design and gradual transitions between widely varying flow areas are also desirable.

4. Effects of Deflector Scale on Sheet Dimensions and Orientation

To determine the effects of deflector scale on sheet dimensions and orientation, three additional deflectors, about 20 times larger than those used in the work reported in *SPS 37-38*, were fabricated. Each deflector had a radius of 6.45 in. and an orifice diameter of 0.405 in. The three deflector angles were 15, 30, and 45 deg. All orifice length-to-diameter ratios were 100. Overhang ratios (h/d_o) varied from 0.54 to 4.66. The rates of water flow were 65 and 97 ft/sec. In every case, w was correlated by Eq. (1) and β was correlated by Eq. (14). There was no measurable velocity effect. These results indicate that the correlations in terms of dimensionless geometry ratios developed previously for the smaller deflectors are directly scalable to deflectors at least 20 times as large in linear dimensions. Sheet thickness (readily measured for the scaled-up deflectors by inserting a depth gage directly into the sheet) was found to vary from a maximum of about 0.05 in. at the center of the deflector to about 0.01 in. near the edges. This gives an indication of how thin the sheets from deflectors 20 times as small must be. Two basic differences in sheet behavior were noted with the large deflectors: First, the sheets left the large deflectors tangentially ($\delta = 0$), whereas appreciable deflection angles were encountered at the smaller scale. This implies that, at *Re* and *We* values which are orders of magnitude larger than those within the ranges called out previously, the deflection angle δ must begin to show *Re* and *We* effects. Second, no edge ribs were noted with the large deflectors, whereas prominent edge ribs characterized the sheets from the smaller deflectors.

5. Sheet Mass Flow-Rate Distribution

For the present investigation, mass flow-rate distribution within individual flowing sheets is being measured by traversing the width of the sheets, at a distance of 0.15 in. from the edge of the deflectors, with a small, flattened-end tube which collects liquid over a time interval at each station. This device is shown in Fig. 5.

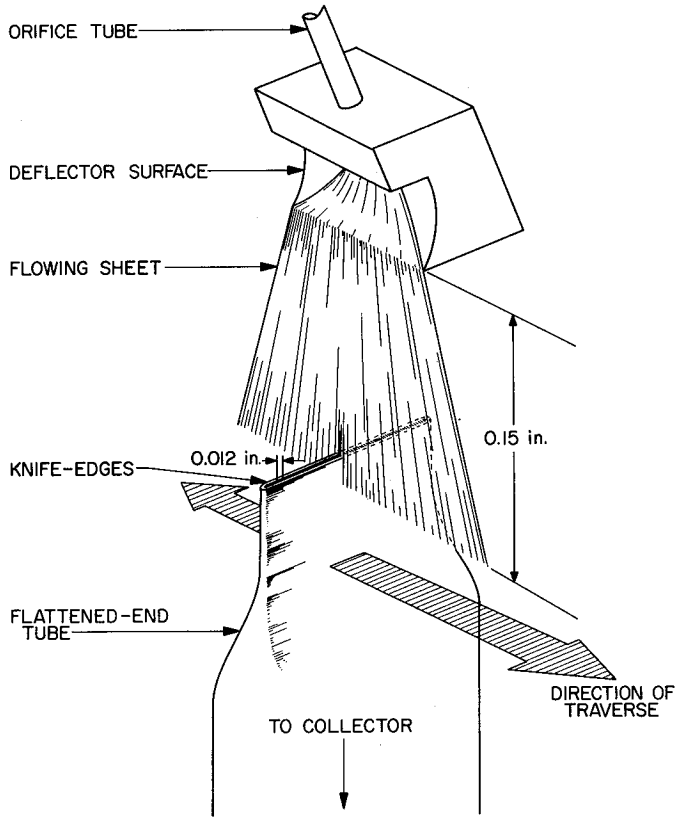


Fig. 5. Apparatus for measuring mass flow-rate distribution

Mass distribution curves obtained in this manner for water flowing at a 78-ft/sec constant velocity through an 0.018-in.-diameter orifice tube ($L_o/d_o = 110$) onto a number of different deflectors are presented in Fig. 6. To show all 16 curves clearly, Fig. 6 was divided into four parts (a–d) according to the deflector angle θ. It is seen that, in all cases, the mass flow rate is distributed symmetrically about the sheet centerline and has a sharp central peak with a well-defined maximum. The bell shape is perturbed (to a greater or lesser degree, depending on geometry) only by two smaller side peaks coinciding with the edge ribs previously noted (*SPS 37-37*). When these ribs are taken into account, integration of any distribution curve gives

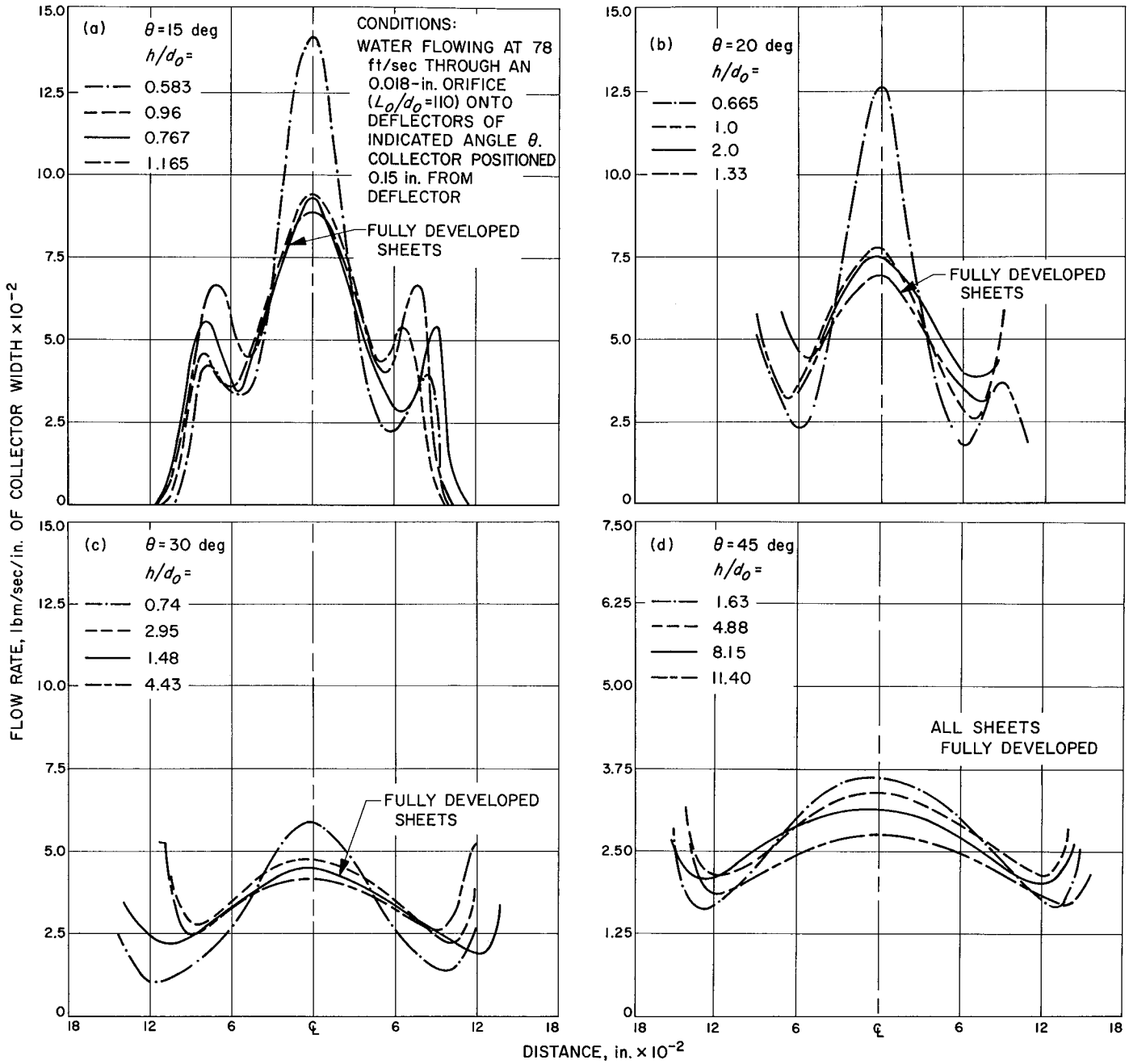


Fig. 6. Mass flow-rate distributions in free sheets of water at constant injection velocity as functions of deflector angle and overhang ratio

approximately the total mass flow rate issuing from the orifice. This basic shape does not change, regardless of how long or short the "overhang" is made relative to the orifice diameter, although the central peak tends to become broader and flatter as the deflector angle θ is increased. The curves of Fig. 6 are quite similar in form to the velocity head distribution curves reported in *SPS 37-38*.

Further examination of Fig. 6(a) reveals that, for $\theta = 15$ deg, the central portions of three of the mass flow distribution curves are essentially coincident, while one stands off by itself. Similarly, for $\theta = 20$ and 30 deg, three central peaks are nearly identical, while one appears markedly different. In each of the cases cited, the coincident mass rate distribution curves are for deflectors with

an overhang ratio $h/d_o > 0.75$, while the "different" curve is for a deflector with $h/d_o < 0.75$. [The "overhang ratio" h/d_o was introduced in *SPS 37-37*; h, defined by Eq. (2), represents the transverse distance to which the deflector protrudes into the otherwise unperturbed stream.]

Based on a detailed study of the large number of sheet photographs reported in *SPS 37-37*, it has been observed (Ref. 2) that true sheets do not form when $h/d_o \lesssim 0.75$. This critical value was apparently independent of injection velocity and fluid physical properties. Below an overhang ratio of ~ 0.75, it was found that the deflectors served only to change drastically the directions of what remained essentially solid streams. Most of the mass appeared to be concentrated in the central portion, which still resembled the original round jets in many respects.

The present results verify quantitatively what was observed qualitatively in Ref. 2. The mass flow-rate distribution in each case where $h/d_o \lesssim 0.75$ resembles that in a round jet (neglecting the end effects) much more than that of fully developed flat sheets ($h/d_o \gtrsim 0.75$).

Considering only the mass flow-rate distribution curves for true sheets in Fig. 6(a–c), it is seen that, for a constant deflector angle θ, h/d_o appears to have little or no effect on the distribution. Neglecting edge effects, the central peaks are nearly identical, regardless of the value of h/d_o (providing $h/d_o \gtrsim 0.75$). It is necessary to vary h/d_o through nearly one order of magnitude, as shown in Fig. 6(d), before any appreciable effect on the distribution curves is realized. [All curves of Fig. 6(d) represent true sheets, since no 45-deg deflectors were made with $h/d_o \lesssim 0.75$.] Deflectors with $h/d_o \gtrsim 2$ are not usually practical because of increasing fluid frictional losses on such long deflectors (Ref. 2); therefore, for a typical injection device, it may be said that the mass distribution in the central portion is virtually independent of overhang at a constant deflector angle θ.

On the other hand, θ exerts a major influence on mass flow distribution. Fig. 7 shows how θ affects the distribution curves for deflectors with a constant overhang $h/d_o = 1$ (obtained by interpolation from Fig. 6). The edge effects have been omitted for clarity, and attention is directed to the central peaks. The mass distribution in a fully developed flowing sheet is seen to become flatter (more nearly uniform) as θ increases. It is planned to study the influence of fluid physical properties and injection velocity on mass flow-rate distribution during future work.

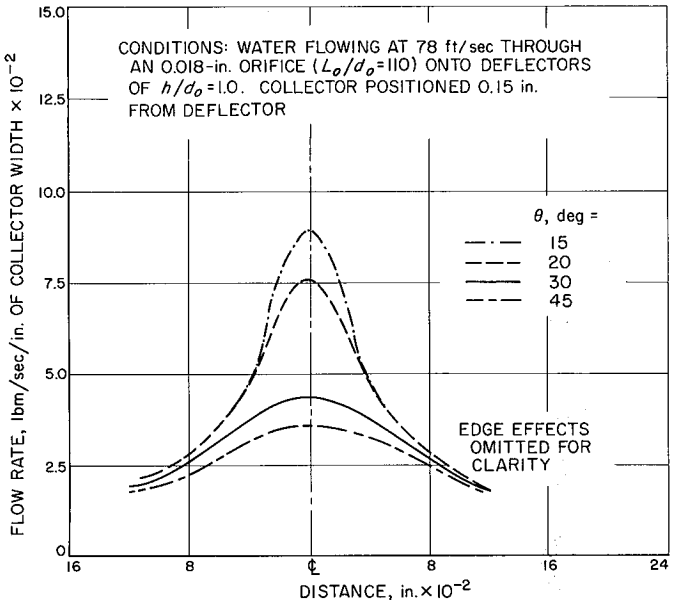
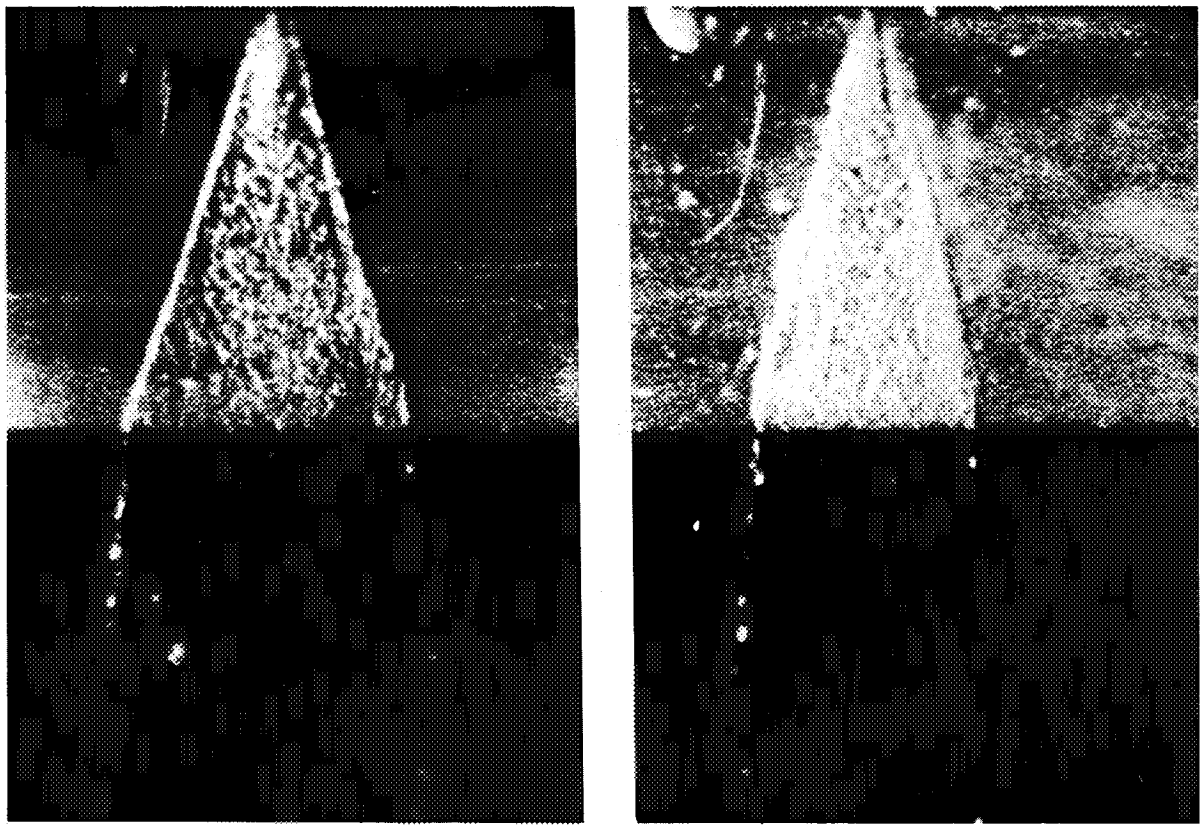


Fig. 7. Effects of deflector angle θ on mass distribution across flowing sheet at constant h/d_o

As was pointed out in *SPS 37-38*, the distribution of *velocity head* (and, therefore, of velocity itself) across a sheet of water at a constant injection velocity depends on h/d_o as well as on θ. This results from increasing viscous losses as the deflectors are made progressively longer. Therefore, it is not possible to make the distribution of fluid *momentum* across the central peak of a sheet independent of h/d_o at constant θ, even though the distribution of mass flux per unit width is.

6. Effect of Deflector Surface Wettability

A single experiment was conducted to illustrate the effect of deflector surface wettability on sheet dimensions. Water was first introduced through an 0.018-in.-diameter ($L_o/d_o = 110$) orifice tube at 100 ft/sec onto a clean (degreased) aluminum deflector surface of 0.7-in. radius and 45-deg included angle. The resulting liquid sheet is shown in Fig. 8(a). Next, a thin, uniform coating of silicone grease was applied to the deflector surface, and water was introduced at the same velocity. The appearance of the corresponding sheet is shown in Fig. 8(b). The sheet width is clearly reduced when the liquid does not wet the surface. The width reduction was 25% in this instance. The spreading angle of the free sheet is, however, unaffected. This experiment gives an indication of the sensitivity of sheet width to deflector surface conditions. It also implies that, in this extreme limit, w/d_o can depend on We, in contrast to the earlier observations. In common practice,



(a) SURFACE WET BY LIQUID SHEET

(b) SURFACE NOT WET BY LIQUID SHEET

Fig. 8. Effect of deflector surface wettability on sheet width

of course, greasy deflectors would not be used. But, if it were desired to alter the width of a propellant sheet from a given deflector surface without changing the other sheet dimensions, a coating of Teflon or other appropriate material might be applied to reduce the surface wettability. Thus, the injector designer is given an additional variable to use in the prediction and control of liquid sheets.

B. Combustion of Lithium in Air

R. A. Rhein

In some earlier experiments (Ref. 3), it was shown that lithium burns readily in air if heated to 375 ± 20°C and forms both its oxide and nitride. For these experiments, some powdered lithium (particle size ≤ 100 μ, from the Foote Mineral Company) was placed in a small stainless-steel boat, and the boat was heated with a torch until ignition occurred. The combustion products were those given below:

Substance	Quantity before combustion, millimoles		Quantity after combustion, millimoles	
	Expt. 1	Expt. 2	Expt. 1	Expt. 2
Elemental lithium	7.58	14.56	0.79	0.23
Lithium oxide	0.75	0.55	2.75	3.86
Lithium nitride	—	—	0.93	2.57

These results, together with the reaction stoichiometry and published thermodynamic properties (Ref. 4), have been used to calculate the heat of combustion of lithium metal burning in air. For the two experiments conducted, the calculated values were, respectively, 6.96 and 5.98 kcal/g of lithium consumed.

References

1. Rupe, J. H., *On the Dynamic Characteristics of Free-Liquid Jets and a Partial Correlation with Orifice Geometry,* Technical Report No. 32-207, Jet Propulsion Laboratory, Pasadena, California, January 15, 1962.

2. Riebling, R. W., and Powell, W. B., *The Properties of Flowing Sheets Formed by Impingement of Liquid Jets on Curved Surfaces,* Paper No. 66-610, presented at the American Institute of Aeronautics and Astronautics Second Propulsion Joint Specialists Conference, Colorado Springs, Colorado, June 15, 1966.

3. Rhein, R. A., "The Ignition and Combustion of Powdered Metals in the Atmospheres of Venus, Earth, and Mars," *Astronautica Acta,* Vol. 11, No. 5, pp. 322–327, 1965.

4. *Selected Values of Chemical Thermodynamic Properties,* Circular 500, U. S. Department of Commerce, National Bureau of Standards, February 1, 1952.

SPACE SCIENCES DIVISION

XII. Lunar and Planetary Sciences

A. A Sample Furnace and Conical Blackbody for Use in Infrared Thermal Emission Spectroscopy Studies

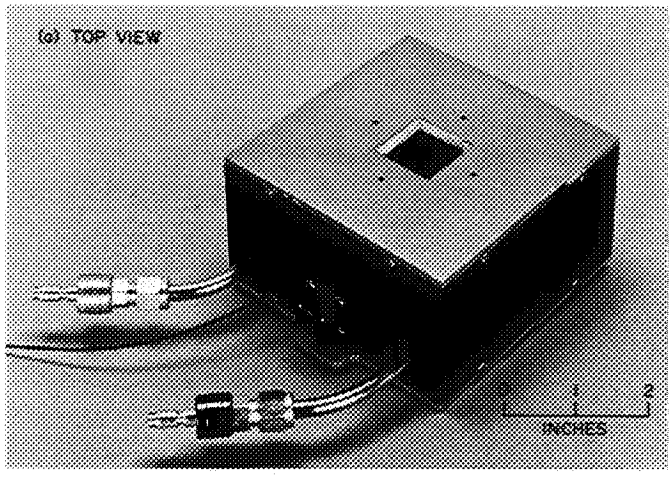
N. F. Stahlberg and J. E. Conel

An investigation of infrared spectroscopy as a method for remote compositional analysis of planetary surfaces involves determining the effective spectral emissivities of various mineral substances. Such a determination can be accomplished by comparing the measured spectral radiant emittance of a mineral substance with that of a blackbody radiator, with both sample and radiator at the same temperature.

The sample furnace and blackbody reference described here are used with a Beckman IR-7 infrared spectrophotometer in the wavelength region of 8 to 40 μ. This instrument has been modified to include a special sample compartment which houses the sample furnace or blackbody reference and the necessary transfer optics. Power for the furnace and radiator is furnished by an Infrared Industries Model 101 temperature controller, and operation is within the temperature range of ambient to 500°C.

Externally, the blackbody radiator and the sample furnace are nearly identical, as shown in Fig. 1(a) and 2(a). Each is enclosed in a thermally insulated aluminum housing to which is attached a copper plate. A square aperture on the plate defines the radiative area seen by the spectrophotometer. The plate is polished and gold-plated to reduce its emissivity and to prevent oxidation of the surface. Chilled water is circulated through tubing soldered to the back of the plate to reduce and stabilize the plate's temperature. To facilitate sample insertion and removal, the aperture plate of the sample furnace is fabricated in two parts, as shown in Fig. 2(a).

Internally, the blackbody radiator consists of an aluminum cone of 24-deg semiangle. The outside of the cone is double-threaded: one thread containing a platinum sensing wire, and the other containing a nichrome heating wire. This scheme provides uniform distribution of heat over the surface of the cone and minimizes thermal lag between heater and sensor. The surface of the cone is black-anodized to increase its emissivity and to provide electrical insulation for the heating and sensing elements. The cone is attached to the cooled aperture plate by supports made of materials having low thermal conductivity. To provide temperature monitoring, a thermocouple is attached to the apex of the cone. Fig. 1(b) shows the internal construction of the blackbody radiator.



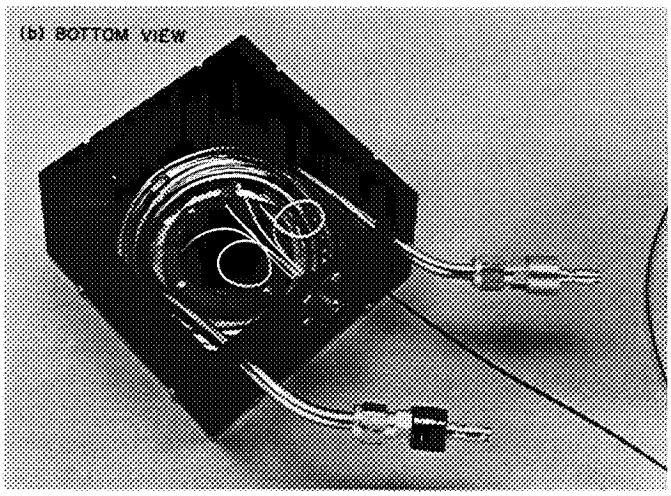
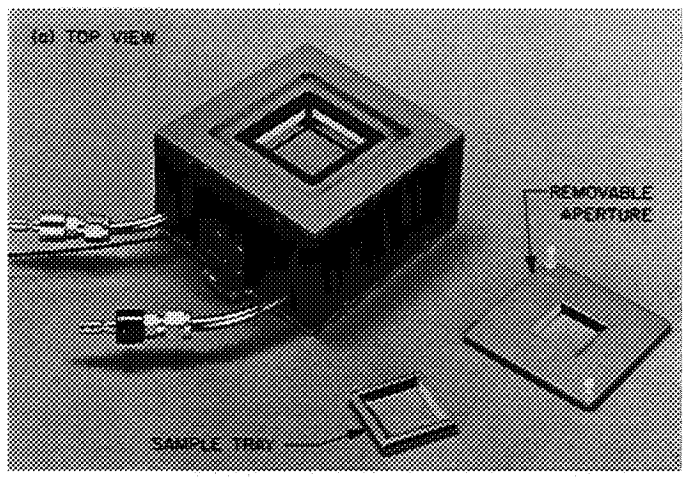


Fig. 1. Blackbody radiator



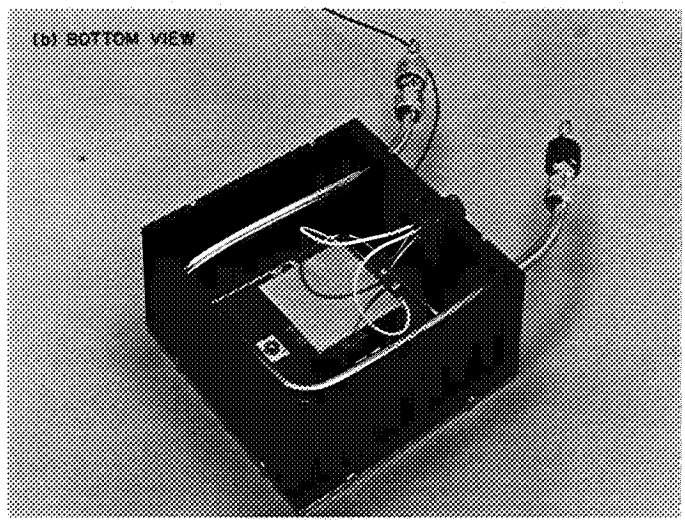


Fig. 2. Sample furnace

The sample furnace consists of a gold-plated copper block containing four heating elements of the type used in soldering irons. The sample to be examined is placed in a shallow tray (also gold-plated copper) which fits into a recess in the top of the block. The thermal conductivity and inertia of the block are such that reasonably uniform heating of the sample may be expected. The block also contains a platinum resistance thermometer for temperature control and a thermocouple for temperature monitoring. The attachment of the block to the cooled aperture plate is similar to that of the blackbody cone. The internal construction of the sample furnace is shown in Fig. 2(b).

XIII. Fluid Physics

A. Resonant Absorption by 2^3S Metastable Helium

R. E. Center and D. A. Russell

The measurement of the molecular velocity-distribution function is important for a complete understanding of high-gradient gas flows, such as free jets and shock waves. In principle, such measurements can be made by analyzing the doppler-broadened radiation from excited gas atoms; however, many of the important problems do not involve radiating gases. Muntz (Ref. 1) recently reported distribution-function measurements made using the doppler broadening from the 5016-A radiation produced when an electron beam is used to excite a thin column of relatively cold helium.

The method being studied here uses an electron beam to create a population of metastable states and to then scatter resonant radiation off these states. If the line profile of the incoming radiation is known, doppler information about the gas can be obtained from analysis of the scattered radiation. It is hoped that this approach will provide a sensitive and flexible means for measuring gas distribution functions.

The first step was to set up static experiments to study the production and diffusion of metastable gas atoms and the cross-sections for resonant scattering from them. For simplicity, the experiments were conducted using the 2^3S metastable state of helium. As shown in Fig. 1, the essential elements were the helium discharge tube, the absorption chamber with the electron beam, and the photomultiplier and interference filter. The first measurements were made with a 150-v electron beam grid-modulated at 100 cps. The resulting fluctuation in the metastable population caused a small ac absorption of the radiation from the discharge tube. This ac signal generally exceeded the direct electron-beam radiation at the photomultiplier; however, it was necessary to use phase-detection techniques to extract the signal from the statistical noise generated by the dc radiation from the discharge tube. The ratio of the ac component to the dc component of the discharge-tube radiation at the photomultiplier is the fraction of radiation that was absorbed. This fraction, $\Delta I/I$, is plotted versus the chamber pressure, P, in Fig. 2. Measurements were made on both the 10,830-A and the 3889-A triplet lines by changing the interference filter and photomultiplier. A 10,000-v television-gun electron beam was used to extend the 10,830-A measurements to conditions representative of the projected experiments.

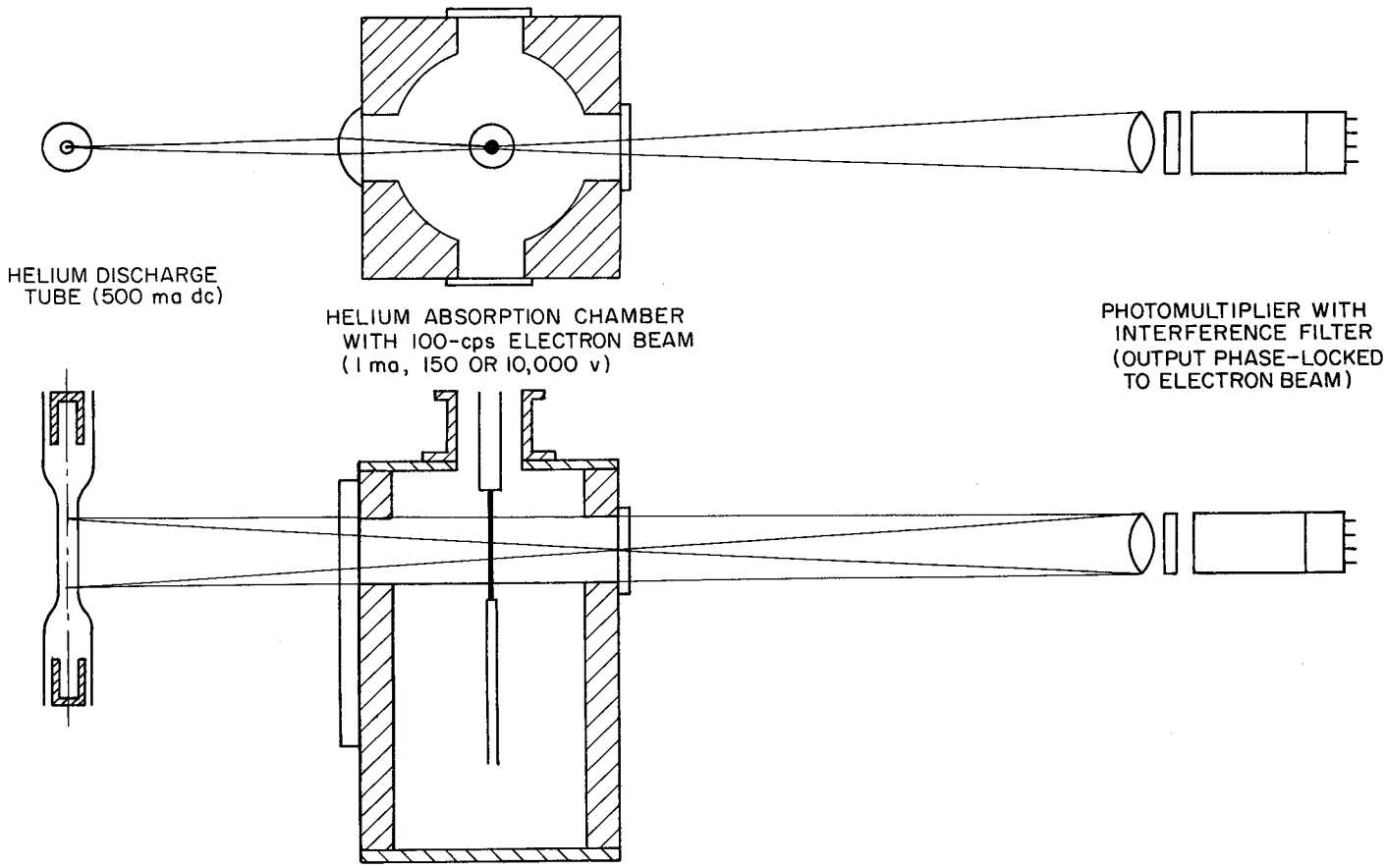


Fig. 1. Experimental setup for absorption measurements

In order to obtain cross-section or absorption coefficients from the data of Fig. 2, the average number of metastable atoms along the radiation path must be known. The metastables are produced by collisions with the beam electrons, both by direct excitation from a ground-state atom and by decay from upper excited states. Thus, absolute measurements of the number of photons given off by an atomic transition ending on the 2^3S level will provide a direct measure of the metastable production rate. The results of such a measurement are shown in Fig. 3, where the 3889-A photon flux, $\dot{N}$, is plotted versus P for both 150-v and 10,000-v beams. The departure from linearity is probably due to the radiation from secondary-electron collisions for the 150-v beam, and due to a systematic underestimate of the beam current for the 10,000-v case. It can be seen, however, that the relative photon flux varies approximately inversely with beam voltage, as expected from elementary scattering theory.

The results of Fig. 3 can be combined with measurements of the radiative photon flux on other lines ending in the 2^3S state to provide an estimate of the total production rate of metastable atoms. To obtain the actual number of metastable atoms along the radiation path, a diffusion equation is set up wherein the metastables are assumed to be created in the electron beam and destroyed at the chamber walls. The number of metastables can then be divided into the results of Fig. 2 to obtain effective cross-sections for the experiments. When this was done, the cross-sections were found to be independent of voltage and pressure as expected, and the 10,830-A radiation cross-section agreed within a factor of 2 to 3 with a calculated value of 1.8×10^{-12} cm^2. This experimental 10,830-A cross-section was approximately 10 times that for the 3889-A radiation, whereas the calculations predicted a factor of 22.

The details of both the data reduction and the theoretical calculations will be presented at a later time. The results presented here, however, are sufficient to allow the specification of a laser to replace the discharge lamp for preliminary distribution-function measurements on the 10,830-A line.

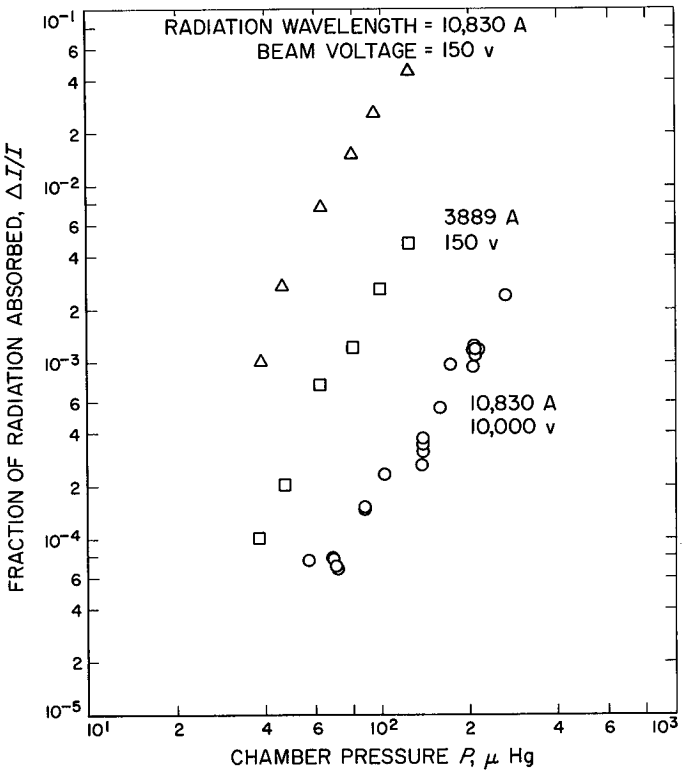


Fig. 2. Fraction of radiation absorbed by electron-beam-excited helium per ma of beam current, $\Delta I/I$, vs chamber pressure P

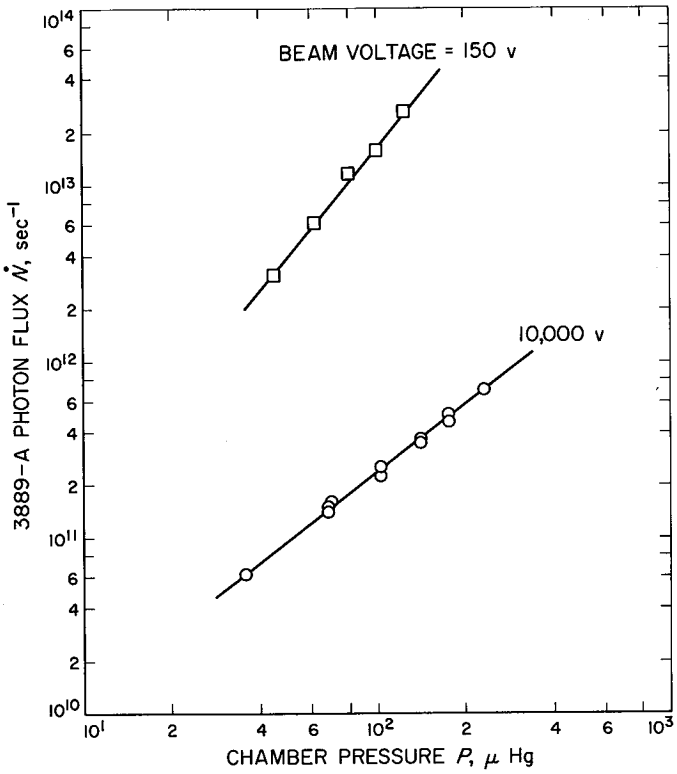


Fig. 3. 3889-A photon flux from electron-beam-excited helium (per ma of beam current per cm of beam length), $\dot{N}$, vs chamber pressure P

B. Effect of Free-Stream Temperature on the Inviscid Stability of the Compressible Laminar Boundary Layer

L. M. Mack

Most of the previous computations in this series have been made with free-stream temperatures such as are encountered in supersonic and hypersonic wind tunnels. In the computation of the boundary-layer profiles with increasing free-stream Mach number M_1, the stagnation temperature was held fixed at 311°K (100°F) until a Mach number was reached where the free-stream temperature $T_1^* = 50$°K. For higher Mach numbers, T_1^* was kept fixed at 50°K. Since the stability characteristics depend upon the boundary-layer profile, and since the profile at a fixed Mach number varies with free-stream temperature, the stability must also be a function of the free-stream temperature. The relative importance of this dependence is discussed here. All results presented are for an insulated wall.

For supersonic Mach numbers, the stability of the first mode depends strongly on the relative magnitudes of $1 - (1/M_1)$ and the mean velocity u in the boundary layer at the generalized inflection point η_s. The latter is the point at which $(u'/T)' = 0$ (the primes denoting differentiation with respect to the Blasius variable η). The minimum phase velocity c_r for a disturbance that is subsonic with respect to the free stream is $1-(1/M_1)$. The maximum phase velocity of a first-mode-amplified or neutral disturbance is the mean velocity at η_s. At a fixed Mach number, any change that brings these two phase velocities together tends to decrease the maximum amplification rate of the first mode; any change that moves them apart increases the amplification rate.

Cooling or heating the wall is one way to effect such changes. Cooling can completely stabilize the first mode by moving η_s below η_0, the point at which $u = 1 - (1/M_1)$. A pressure gradient also moves η_s with respect to η_0. A third method of changing η_s is to change the free-stream

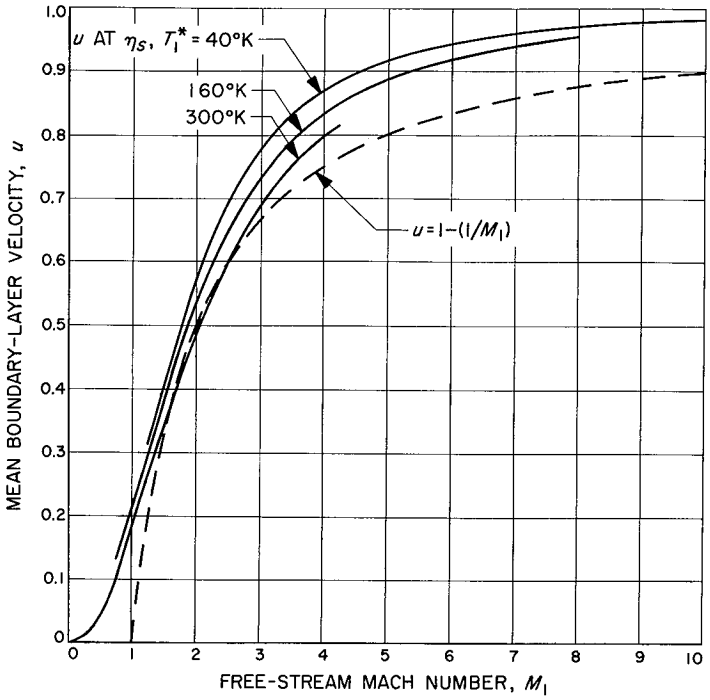


Fig. 4. Mean boundary-layer velocity u at the generalized inflection point η_s, and also $1 - (1/M_1)$, vs free-stream Mach number M_1

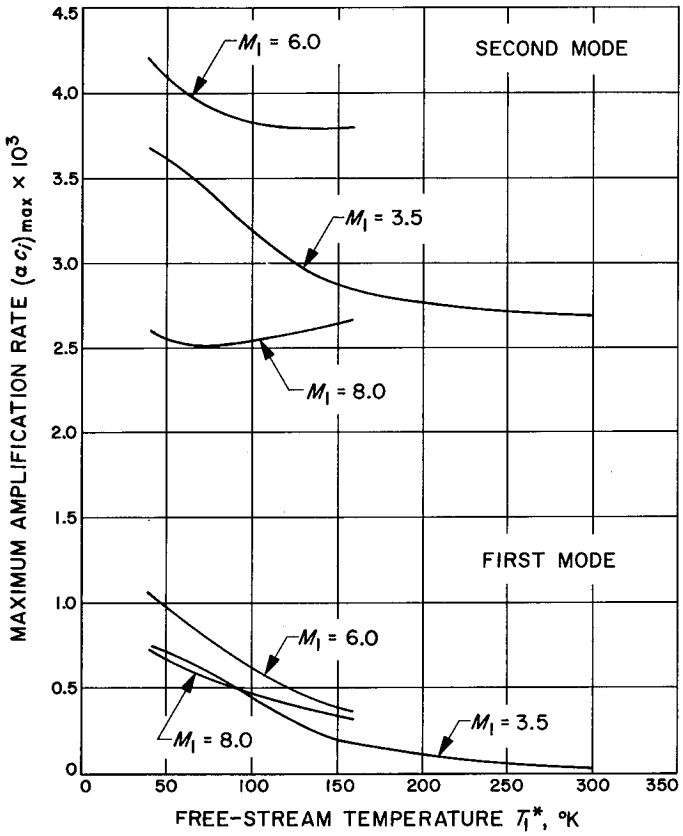


Fig. 5. Effect of free-stream temperature T_1^* on the maximum amplification rates of the first and second modes for $M_1 = 3.5$, 6.0, and 8.0

temperature. In Fig. 4, the mean boundary-layer velocity u at η_s and also $1 - (1/M_1)$ are plotted against M_1 for several free-stream temperatures. In order to obtain high-Mach-number boundary-layer profiles for the higher temperatures, the 1140°K temperature limitation in the property tables of the boundary-layer program was ignored. The program extrapolates beyond this limit, and the profiles based on extrapolated property values have been used. Fig. 4 clearly shows that increasing the free-stream temperature moves η_s toward η_0 and, consequently, can be expected to stabilize the first mode. For $T_1^* = 300$°K, the boundary layer is stable to inviscid disturbances between $M_1 = 1.6$ and 2.5. Higher free-stream temperatures will increase the Mach number range for which there is complete stability.

First- and second-mode maximum amplification rates have been computed at several free-stream temperatures for $M_1 = 3.5$, 6.0, and 8.0. These results are presented in Fig. 5. The first mode, as was predicted from Fig. 4, is strongly stabilized at $M_1 = 3.5$ by increasing T_1^*. At the two higher Mach numbers, the stabilizing effect is less pronounced. The stability of the second mode does not depend upon the relative positions of η_s and η_0; indeed, the second mode is unstable even when η_s does not exist. At $M_1 = 3.5$, the maximum amplification rate is reduced about 25% as T_1^* is increased to 300°K. At the two higher Mach numbers, the stabilizing effect not only decreases, but even reverses.

The question of whether the stabilizing effect of increasing T_1^* also applies to finite Reynolds numbers can only be answered by actual calculations. The viscous instability mechanism, which ceases to be important at about $M_1 = 3.0$ for wind-tunnel temperature conditions, may still determine the maximum amplification rates in this Mach number range at finite Reynolds numbers for the higher-temperature boundary layers. However, at a sufficiently high Mach number, regardless of the free-stream temperature, the finite-Reynolds-number amplification rates can be expected to follow the inviscid amplification rates. In any case, Fig. 5 suggests that, if transition of the laminar boundary layer is related to the instability of the first mode, the free-stream temperature itself, and not just the ratio of free-stream temperature to wall temperature, should be considered as a parameter in attempting to correlate transition results.

C. Boundary-Layer Stability Experiments

J. M. Kendall, Jr.

An experimental determination of the stability characteristics of a laminar supersonic boundary layer is under way. The results obtained shall be compared to the theoretical results of Mack presented in a series of *SPS* articles. Under present consideration are:

(1) In a certain Mach number range, there are two ranges of αc_r, the dimensionless frequency, in which disturbance waves are amplified. These are referred to as the first and second instability modes in order of increasing αc_r (*SPS 37-24*, Vol. IV, p. 271).

(2) Waves in the first-mode range generally have higher amplification rates when the waves are oblique, rather than parallel, to the crossflow dimension z or the streamwise dimension x. Waves in the second-mode range are most unstable when parallel to z (*SPS 37-36*, Vol. IV, p. 221).

The measurements are being made in the JPL supersonic wind tunnel. A Mach number of 4.5 was selected for the measurements because the theory predicts that the second mode, not previously observed experimentally, coexists with, and exhibits higher amplification rates than, the first mode. The tunnel unit Reynolds number of the tests, 0.50×10^6/ft, was slightly below the value at which the tunnel-wall boundary layers became turbulent. The boundary layer under study develops on a 72-in.-long flat plate which spans the 18-in. width of the tunnel. The plate leading edge was bevelled 10 deg on the bottom and was sharpened to an 0.0005-in. radius. The boundary layer was about 0.6 in. thick at the aft end of the plate. Disturbance waves at selected frequencies in the range of 0.9 to 30 kHz were introduced, with any one of three different angles, into the otherwise steady boundary layer by a variation of the glow discharge technique described in *SPS 37-32*, Vol. IV, p. 228.

The configuration is shown in Fig. 6. The waves were measured by means of a hot-wire anemometer supported on a traverse mechanism that could be driven in each of the three principal directions. In all cases, it was found that the fluctuation amplitude through the boundary-layer thickness δ was self-similar. Therefore, the traverse mechanism was servo-controlled to seek the value of y/δ at which the fluctuation was largest, leaving the experimenter free to vary the x and z coordinates independently. The mean flow was measured by a pitot tube.

The boundary-layer mean-flow profile shape was reasonably close to that expected from the results presented in Ref. 2. However, for some reason not yet discovered, the thickness exceeded the predicted value by approximately 10%. The hot-wire revealed that, in the first-mode range ($\alpha c_r < 0.15$), waves introduced at $\sigma = 0$ deg spontaneously became warped, and hence locally oblique, and did not amplify equally at all z stations. Conversely, waves introduced at $\sigma = 55$ deg remained straight and parallel and amplified nearly equally at various z stations. In the second-mode range, oblique waves were of such small amplitude that precise measurements were difficult. Accordingly, the amplification-rate measurements were

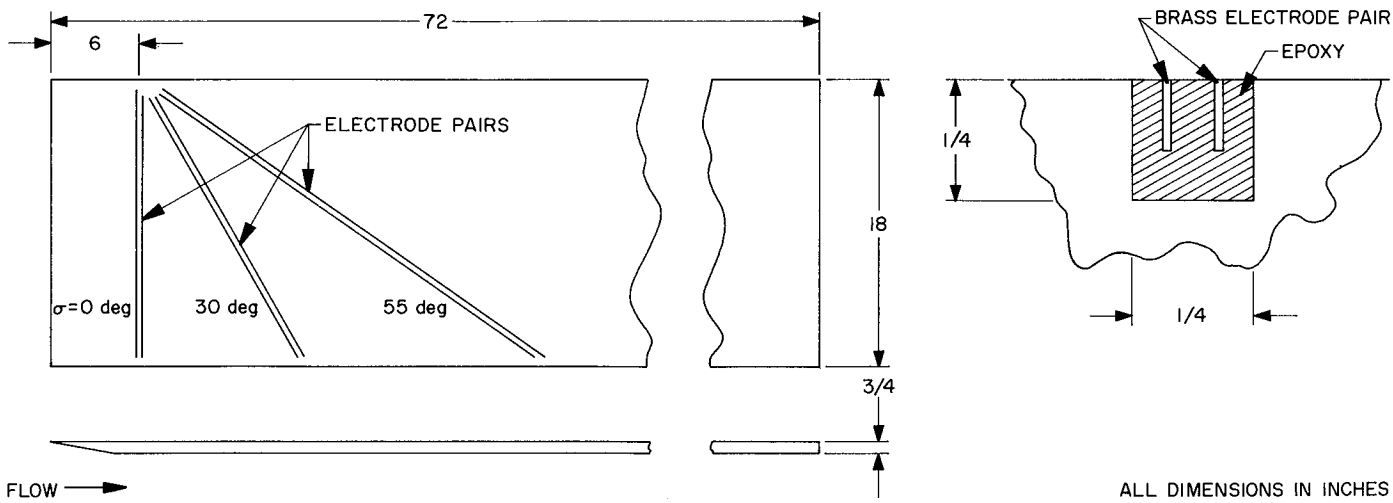


Fig. 6. Boundary-layer plate

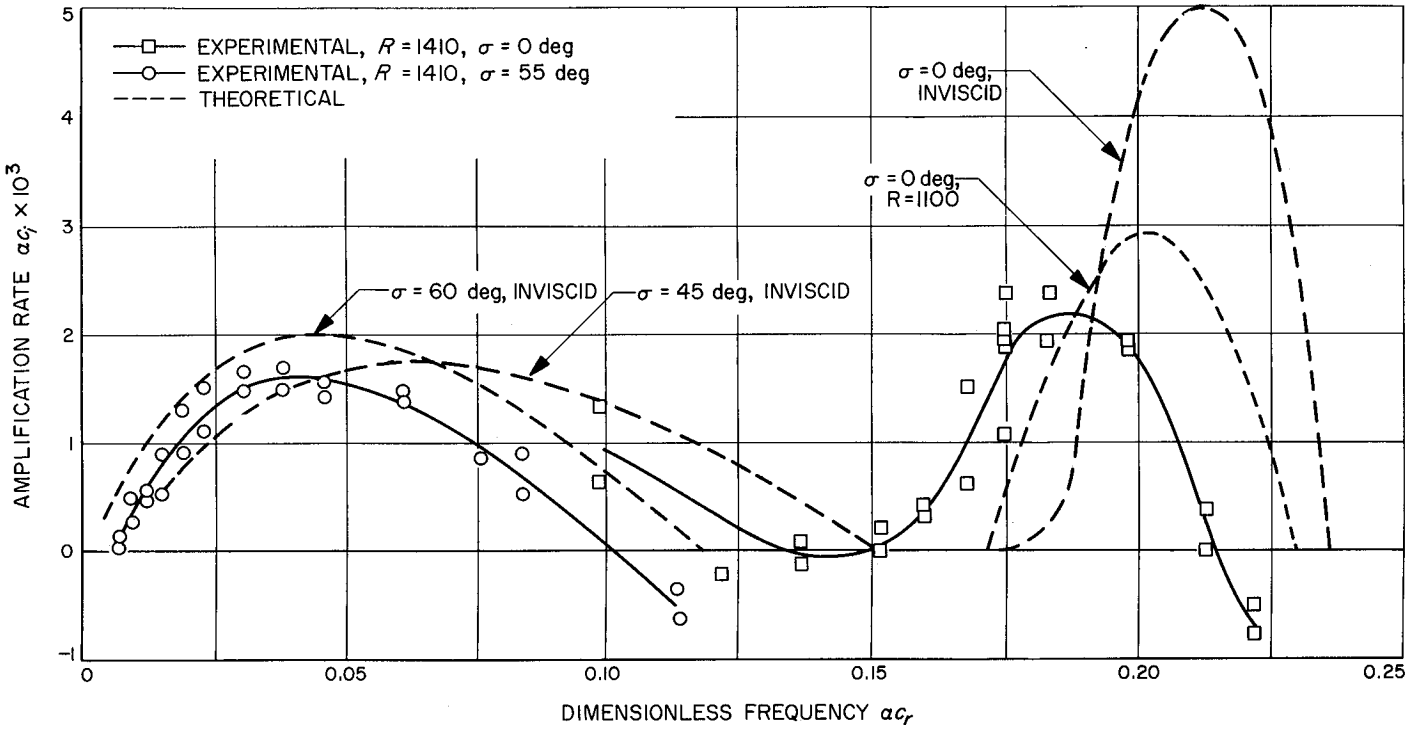


Fig. 7. Amplification rate vs dimensionless frequency at $M_1 = 4.5$

limited to 55-deg waves for the first mode and 0-deg waves for the second mode.

Fig. 7 presents the time rate of amplification αc_i versus αc_r for one value of R, defined as the square root of the x Reynolds number. Data at $R = 1525$ and 1630, not shown, differ very little. Also shown in Fig. 7 are theoretical results for inviscid oblique first-mode waves and viscous and inviscid second-mode waves for a Mach number of 4.5. Theoretical results for viscous oblique first-mode waves are not available. Although the agreement between experiment and the most appropriate theoretical curve is quite good, it would be improved if the data were corrected for the excess boundary-layer thickness of the experiment. Fig. 8 compares the measured phase velocity c_r with theoretical values. The demonstration of anomalous dispersion in the second-mode region is regarded as an especially favorable point of confirmation.

Future experiments will be conducted to obtain an improved boundary layer, assess the effect of Mach number, and investigate plate-surface cooling effects.

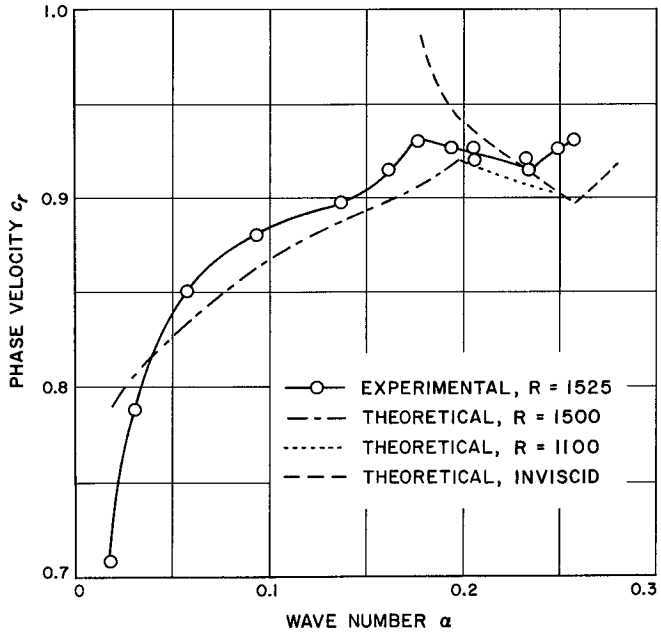


Fig. 8. Phase velocity vs wave number at $M_1 = 4.5$

References

1. Muntz, E. P., *Rarefied Gas Dynamics,* Vol. II, J. H. de Leeuw, Ed., Academic Press, New York, 1965.

2. Mack, L. M., *Calculation of the Laminar Boundary Layer on an Insulated Flat Plate by the Klunker-McLean Method*, Progress Report 20-352, Jet Propulsion Laboratory, Pasadena, California, July 1958.

XIV. Physics

A. Classification of Alcohols From the ^{19}F Spectra of Trifluoroacetates

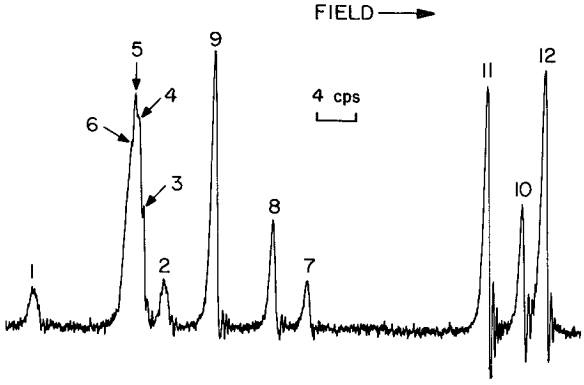
S. L. Manatt

In previous work[1], it was demonstrated that ^{19}F nuclear-magnetic-resonance (NMR) spectra of the trifluoroacetate group could be used to discriminate between primary and secondary polymer hydroxyl groups in hydroxyl-terminated polymers. Reported here is an NMR scheme for classifying hydroxyl compounds which represents a generalization of the technique previously applied at JPL to the study of polymers. This technique has generally proved more reliable and informative than several other recently described NMR techniques for classification of hydroxyl groups (Refs. 1–3).

Because a number of functional groups can be acetylated, the possibility that the acetate group might be a useful NMR probe for classification purposes was considered initially. Unfortunately, the differences in chemical shift between acetate groups of isomeric alcohols are only a few cps and provide no dependable scheme for classification. However, ^{19}F chemical shifts are, in general, about an order of magnitude larger than ^{1}H chemical shifts for a given structural perturbation. This suggested the possibility that ^{19}F NMR chemical shifts of the trifluoroacetyl derivatives of alcohols and perhaps amino, mercapto, and phenolic functional groups might provide reliable classification schemes.

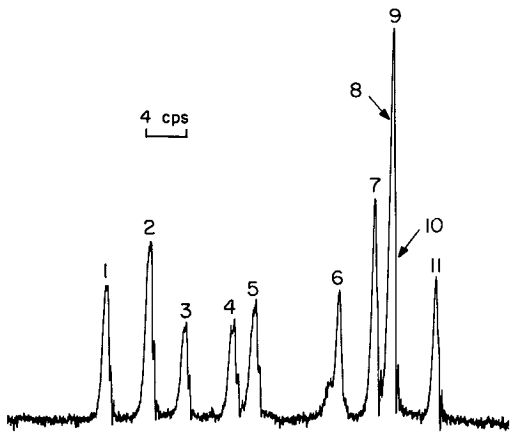
Study of the ^{19}F NMR spectra of the trifluoroacetate (TFA) esters of a large number of alcohols revealed that the TFA groups give sharp ^{19}F signals and that the order of shielding is always primary $<$ secondary $<$ tertiary. Fig. 1(a) illustrates this for a mixture of these three classes of alcohols. All ^{19}F spectra shown were taken at 56.4 Mc with a Varian Type HR instrument equipped with a field-frequency-lock system (Ref. 4). The lock signal was derived from the 5 to 10% of 1,1,2,2-tetrafluoro-1,2-dibromoethane which was added. Other spectra were recorded with a Varian A-56/60 instrument. Fig. 1(b–d) shows results for mixtures of alcohols illustrating the general nature of substituent effects on the chemical shift of the TFA group. A single electronegative substituent β to the hydroxyl group causes a downfield shift, as might be expected on purely inductive grounds (Fig. 1c); the fact that the TFA of 1,1-difluoroethanol is at a lower

[1] *SPS 37-24*, Vol. IV, p. 94; *SPS 37-26*, Vol. IV, p. 108; *SPS 37-31*, Vol. IV, p. 143; *SPS 37-33*, Vol. IV, p. 224; and *SPS 37-37*, Vol. IV, p. 145.



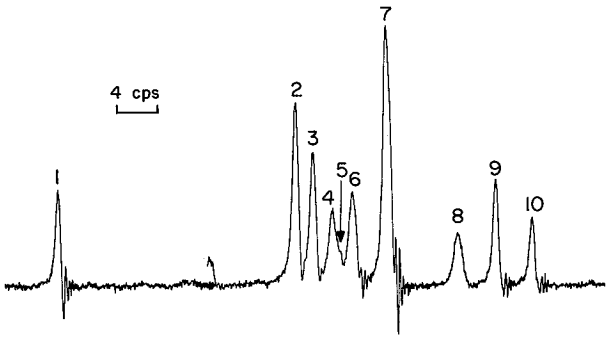
1 CH_3OH (74.21)
2 CH_3CH_2OH (74.43)
3 $CH_3CH_2CH_2OH$ (74.40)
4 $CH_3CH_2CH_2CH_2OH$ (74.39)
5 $(CH_3)_2CHCH_2OH$ (74.38)
6 $(CH_3)_3CCH_2OH$ (74.37)
7 $CH_3CHOHCH_3$ (74.68)
8 $CH_3CHOHCH_2CH_3$ (74.62)
9 $CH_3CH_2CHOHCH_2CH_3$ (74.52)
10 $(CH_3)_3COH$ (75.05)
11 $(CH_3)_2(CH_3CH_2)COH$ (74.99)
12 CF_3CO_2H (75.08)

(a) 20% ESTER MIXTURE IN DIMETHYL SULFOXIDE



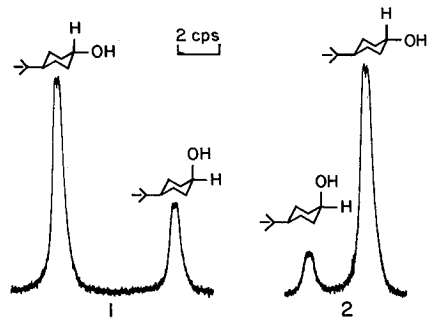
1 $CH{\equiv}CCH_2OH$ (74.95)
2 $C_6H_5CH_2OH$ (75.02)
3 $CH_2{=}CHCH_2OH$ (75.08)
4 $(CH_3)_3CCH_2OH$ (75.17)
5 $CH_3CH_2CH_2OH$ (75.20)
6 LANOSTEROL (75.35)
7 CHOLESTEROL (75.41)
8 CYCLOPENTANOL (75.43)
9 CYCLOHEXANOL (75.44)
10 CYCLOHEPTANOL (75.45)
11 $CH_3CHOHCH_3$

(b) 15% ESTER MIXTURE IN METHYLENE CHLORIDE



1 C_6H_5OH (73.85)
2 CH_2OHCH_2OH (74.27)
3 $CH_2OHCH_2CH_2OH$ (74.30)
4 $CH_3CHOHCH_2\underline{OH}$ (74.33)
5 CHF_2CH_2OH (74.34)
6 CF_3CH_2OH (74.37)
7 CH_3CH_2OH (74.43)
8 $CH_3CH\underline{OH}CH_2OH$ (74.55)
9 $CH_3CHOHCH_2CH_3$ (74.62)
10 $CH_3CHOHCH_3$

(c) 20% ESTER MIXTURE IN DIMETHYL SULFOXIDE



(d) TWO DIFFERENT MIXTURES OF 4-*t*-BUTYLCYCLOHEXANOLS

1 *TRANS*-ISOMER (74.99), *CIS*-ISOMER (75.09) (5% IN DIMETHYL SULFOXIDE)
2 *CIS*-ISOMER (76.28), *TRANS*-ISOMER (76.32) (5% IN PENTANE)

Fig. 1. ^{19}F NMR spectra of the trifluoroacetates of some hydroxyl compounds [chemical shifts in ppm ($\pm$0.01) from 10% internal $CFCl_3$ given in parentheses; chemical shifts of overlapping signals determined from more highly resolved spectra of ester mixtures with fewer components]

field than that of 1,1,1-trifluoroethanol indicates that simple inductive arguments are invalid here. Fig. 1(d) illustrates a rather interesting solvent shift for two mixtures of the TFA esters of *cis*- and *trans*-4-*t*-butylcyclohexanol.

The TFAs were prepared by the addition of trifluoroacetic anhydride to the alcohols either directly or in an inert solvent. In the latter case, the excess anhydride and the acid formed were removed either by vacuum or by extraction with dilute aqueous bicarbonate followed by drying. The reaction of primary and secondary alcohols is complete in several minutes. Tertiary and some polyhydroxyl compounds require a longer time, and perhaps more than one treatment with anhydride, to achieve complete reaction.

Although chemical-shift concentration and solvent effects have been observed, so far, for mixtures of TFAs, chemical shifts are always primary < secondary < tertiary in the order of shielding. Comparison of the data in Fig. 1 for the chemical shifts (in ppm) for the TFAs of *cis*-4-*t*-butylcyclohexanol (A) (75.09) and *trans*-4-*t*-butylcyclohexanol (B) (74.99) with those of 2-methyl-2-butanol (C) (74.99) and *t*-butyl alcohol (D) (75.05) would seem to refute this statement. However, this is an example of the concentration effects. The solution was in one case (Fig. 1d) 5% in ester and in the other case (Fig. 1a) 20% in ester. In a 20% ester mixture of DMSO (7% CCl_3F), A, B, C, and D show chemical shifts of 74.57, 74.69, 75.00, and 75.06 ppm, respectively.

For a single unknown alcohol or a mixture of alcohols, the operational procedure consists of TFA preparation, dissolution in a solvent if necessary, addition of a small amount of ethyl trifluoroacetate as an internal reference, and recording of the spectrum. Ethyl trifluoroacetate is a reliable internal standard because, at 56.4 Mc, all other primary TFAs are downfield (2–17 cps), all phenol TFAs are downfield (30 cps or more), all secondary TFAs are upfield (3–15 cps), and all tertiary TFAs are upfield (30–36 cps).

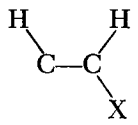
It is not claimed that the trifluoroacetylation-^{19}F NMR technique is, in every case, the simplest and best means of characterizing hydroxyl groups. There will certainly be instances where other techniques are adequate. However, the method described here does have certain obvious advantages: First, one is looking at the signal from three ^{19}F nuclei versus one ^{1}H nucleus (as in the case of several other recently described NMR techniques: Ref. 1–3) which can be a multiplet. As can be seen in the spectra shown in Fig. 1, and more readily in spectra recorded at slightly slower sweep rates, there exist small splittings (0.15–0.25 cps) of the CF_3-groups with the proton(s) five bonds removed. Thus, at worst, a signal-to-noise increase of around 2.5 should be possible. In addition, because one observes the ^{19}F spectrum, signals from the other protons in the molecules of interest are far removed, and the restrictions on the choice of solvent are essentially eliminated.

Extension of the technique described here for use with other types of organic functional groups is being investigated.

B. Effect of Conjugation on the Proton Coupling Constants in Vinyl Groups

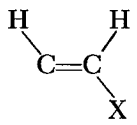
S. L. Manatt

In general, for aliphatic and olefinic nuclear-magnetic-resonance (NMR) proton-proton spin-spin couplings, as the electronegativity of a substituent X in the fragments



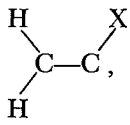
(1)

and

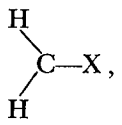


(2)

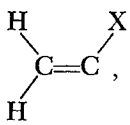
is increased, the vicinal coupling becomes smaller. The geminal couplings in the fragments



(3)

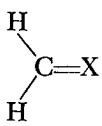


(4)



(5)

and



(6)

are likewise dependent on the electronegativity of X. For olefinic fragments, it would seem most certain that the changes in geminal and vicinal couplings result both from inductive and conjugative perturbations and possibly

through hyperconjugative effects. As yet, no clear experimental demonstration and/or separation of the various contributions to the observed coupling have been reported. Obvious extensions of the work reported here would seem to offer the possibility of effecting such an experimental separation of inductive and conjugative contributions to changes in NMR proton-proton spin-spin coupling constants in at least one system: the vinyl group.

The NMR spectra of the vinyl protons in ethylene (*7*) and styrene (*8*) are the starting points. The coupling constants for these and several other vinyl groups are given in Table 1. In (*7*), the three proton coupling constants are 2.4, 11.6, and 19.1 cps, corresponding to the geminal, *cis*-vicinal, and *trans*-vicinal couplings, respectively. In (*8*), these same couplings have values of 1.34, 10.76, and 17.64, respectively. *A priori*, the changes from (*7*) to (*8*) can be attributed to both inductive and resonance origins. In Table 1, the corresponding coupling constants of 4-vinylbiphenyl (*9*) and β-vinylnaphthalene (*10*) are seen to be very similar to those of (*8*). In (*8*)–(*10*), the vinyl group should be able to achieve maximum coplanarity and conjugation with the aromatic ring to which it is bonded. If the conjugation of the vinyl groups in (*8*)–(*10*) with their associated aromatic rings could be destroyed by confining the former to a plane perpendicular to the plane of the aromatic rings, then perhaps any change of their coupling constants from those in ethylene (*7*) could be attributed to pure inductive perturbations. This interpretation inherently suggests that there is only very small-scale redistribution in the σ-electron framework of a vinyl group when making the above-mentioned hypothetical change.

Table 1. Coupling constants in certain vinyl groups

Group	Coupling constant, cps		
	Geminal	*cis*-Vicinal	*trans*-Vicinal
Ethylene (*7*)[a]	2.4	11.6	19.1
Styrene (*8*)	1.34 ±0.02	10.76 ±0.03	17.64 ±0.03
4-Vinylbiphenyl (*9*)	1.33 ±0.03	10.73 ±0.03	17.64 ±0.03
β-Vinylnaphthalene (*10*)	1.23 ±0.02	10.75 ±0.03	17.67 ±0.03
9-Vinylanthracene (*11*)	2.38 ±0.03	10.58 ±0.03	17.83 ±0.06
2,6-Dimethylstyrene (*12*)	2.40 ±0.02	10.58 ±0.02	17.73 ±0.02
2,4,6-Trimethylstyrene (*13*)	2.45 ±0.02	10.63 ±0.02	17.71 ±0.04
2,3,5,6-Tetramethylstyrene (*14*)	2.59 ±0.03	10.40 ±0.05	17.74 ±0.04
2-Methylstyrene (*15*)	1.74 ±0.02	10.87 ±0.03	17.56 ±0.03
α-Vinylnaphthalene (*16*)[b]	1.7 ±0.1	10.8 ±0.1	17.6 ±0.1
2,5-Dimethylstyrene (*17*)	1.69 ±0.02	10.86 ±0.03	17.47 ±0.03
1,3-Butadiene (*18*)[c]	1.74 ±0.05	10.17 ±0.05	17.05 ±0.05
2-*t*-Butyl-1,3-butadiene (*19*)[c]	2.30 ±0.05	10.80 ±0.05	17.00 ±0.05
	1.70 ±0.05[d]	—	—
3,3-Dimethylbutene (20)	1.78 ±0.02	10.31 ±0.04	17.63 ±0.05

[a]Refs. 5 and 6.
[b]Analysis not complete.
[c]Ref. 7.
[d]Geminal coupling between protons on carbon atom 1.

The situation closest to the hypothetical test suggested above is found for molecules in which bulky groups, proximate to the vinyl group, constrain the latter from achieving coplanarity with its associated aromatic group. From models, it appears that the hypothetical test situation sought is closely approximated in the molecules 9-vinylanthracene (*11*); 2,6-dimethylstyrene (*12*); 2,4,6-trimethylstyrene (*13*); and 2,3,5,6-tetramethylstyrene (*14*). From the parameters given in Table 1, it is indeed significant that the geminal coupling constants of all three of these are, at most, 0.2 cps from that of ethylene (*7*); this suggests that it is *primarily a conjugative effect that accounts for the difference between the geminal coupling of ethylene and the geminal coupling of vinyl groups in complete conjugation with an aromatic ring, as in systems* (*8*)–(*10*).

Models show that the vinyl groups in 2-methylstyrene (*15*); α-vinylnaphthalene (*16*); and 2,5-dimethylstyrene (*17*) are also hindered from achieving coplanarity with their associated aromatic ring. However, the steric severity for the vinyl groups is significantly less than that for (*11*)–(*14*). The vinyl groups of (*15*), (*16*), and (*17*) can attain conformations such that a significant π-orbital resonance integral should still exist between the aromatic ring carbon at the point of attachment of the vinyl group and the adjacent π-orbital of that group. That this is the case is indicated by the fact that the geminal coupling constants for (*15*), (*16*), and (*17*) are intermediate between those for the sets (*8*)–(*10*) and (*11*)–(*14*).

A separation of inductive and conjugative contributions for a vinyl group attached to another vinyl group is possible, based on the recently reported results (Ref. 7) of the analyses of the spectra of 1,3-butadiene (*18*) and 2-*t*-butyl-1,3-butadiene (*19*). In the case of (*19*), models indicate that the 2-*t*-butyl groups severely hinder the coplanarity of the double bonds. In (*18*), where the two vinyl groups would be expected to achieve maximum conjugation, the geminal and vicinal couplings are in the range of conjugation observed for the partially conjugated styrenes (*14*)–(*16*). One of the vinyl couplings in (*19*) (2.30 cps) is close to that of ethylene (2.4 cps), while the corresponding vicinal couplings are very close to those in (*18*). The other geminal coupling in (*19*) is essentially the same as that in (*18*).

This situation may be interpreted as follows: The difference between the 1.74-cps geminal coupling in (*18*) and the 2.30-cps geminal coupling in (*19*) represents mostly the conjugative contribution of a vinyl group to the geminal coupling in another vinyl group attached to the former. The 1.70-cps geminal coupling in (*19*) is a result of the loss of conjugation with the other vinyl group and the presence of the 2-*t*-butyl substituent; this substituent modifies the coupling from the ethylene value by inductive contributions and possibly also by hyperconjugative mechanisms, which are not characteristic of the aromatic ring substituents discussed above. From the couplings in 3,3-dimethylbutene (*20*), it appears that the *t*-butyl group modifies the geminal coupling constant for a vinyl group from the ethylene value of 2.40 cps to 1.78 cps. In (*19*), steric compression may perturb certain bond angles and give rise to other contributions to the magnitude of this geminal coupling, and possibly the two vicinal couplings. However, it would appear that the difference between the 2.30-cps geminal coupling in (*19*) and the 1.70-cps geminal coupling in (*18*) demonstrates the magnitude of conjugative contribution of the vinyl group to this coupling and also suggests that the conjugative ability of a vinyl group is less than that of an aromatic ring.

Based on the above qualitative arguments, a method for isolating the conjugative contributions to geminal proton spin-spin couplings in vinyl groups has been suggested. From this discussion, it follows that an aromatic system has a greater ability to conjugate with a vinyl group than does another vinyl group. Perhaps the reverse of the arguments presented above would serve to establish the ability of a substituent to conjugate with a vinyl group.

C. Vibrational–Rotational Energies of Physically Adsorbed Molecules

J. King, Jr., and D. Merrifield

1. Introduction

In the electrostatic theory of physical adsorption developed by King and Benson (Ref. 8; and *SPS 37-33*, Vol. IV, p. 208), the assumption was made that the vibrational and rotational motions of the molecules on the surface were separable, i.e., uncoupled. This assumption facilitated the quantum mechanical calculation of the vibrational and rotational energies. In any real system, these motions are not entirely uncoupled, since the rotation of the molecule is dependent upon its distance from the surface, which, at any time, is determined by its vibration normal to the surface.

A more refined calculation must therefore include the interaction of the molecular vibration and rotation. This calculation can be performed using the variational method, which will be applied here to the adsorption of hydrogen on a gamma alumina (γ-Al_2O_3) surface to enable a comparison with the results of King and Benson.

The model is that of a diatomic molecule undergoing vibrations and rotations on a solid surface. The vibration is that of the center of mass and is normal to the surface; the rotation is that of a dumbbell with its axis, in the equilibrium position, perpendicular to the surface. The internal vibration of the molecule can be neglected, since the forces between the molecule and surface are small compared to the restoring forces involved in internal vibrations.

2. Method of Calculation

To obtain the energies, it is necessary to solve the Schroedinger equation whose Hamiltonian (Ref. 9) is

$$H = \frac{P_x^2}{2M} + \frac{P_y^2}{2M} + \frac{P_z^2}{2M} + \frac{P_\theta^2}{2I} + \frac{P_\phi^2}{2I \sin^2\theta} + \Phi(z, \theta), \quad (1)$$

where the P's are linear and angular momenta, M is the total mass, I is the moment of inertia, and $\Phi(z, \theta)$ is the potential energy of interaction of the molecule with the surface. The first two terms represent free translation parallel to the surface plane. It is assumed that there is free rotation around the azimuthal angle ϕ and hindered rotation around the angle θ.

The nature of the electrostatic interaction dictates the form that $\Phi(z, \theta)$ must take. As the molecule approaches the surface, it is polarized by the strong electric field causing it to be attracted to the surface. The energy of this attraction is given by

$$\Phi_{att} = -\frac{\alpha}{2} E_z^2, \quad (2)$$

where α is the polarizability of the adsorbed molecule and E_z is the electric-field intensity normal to the surface.

Using a reciprocal-z dependence of the field E_z and a reciprocal-z dependence for the repulsive force, King and Benson found that the total potential Φ_{tot} could be represented as follows:

$$\Phi_{tot} = -\frac{1}{2}\alpha(D/z^m)\left[1-\frac{m}{n}\left(\frac{z_0}{z}\right)^{n-m}\right], \qquad (3)$$

where m and n are the exponents for the attractive and repulsive terms, respectively; z_0 is the equilibrium distance; and D is an empirical constant. This potential has the same form as the potential energy term in Eq. (1), since α is a function only of the angle θ. In fact, the molecular polarizability is a tensor whose diagonalized component, after rotation through the angle θ, is given by (Ref. 10)

$$\alpha(\theta) = \alpha_{\parallel}\cos^2\theta + \alpha_{\perp}\sin^2\theta, \qquad (4)$$

where $\alpha_{\parallel}$ and $\alpha_{\perp}$ are the parallel and perpendicular components with respect to the internuclear axis. The potential energy can thus be written as

$$\Phi(z,\theta) = -\frac{D}{2z^m}(\alpha_{\parallel}\cos^2\theta + \alpha_{\perp}\sin^2\theta)\left[1-\frac{m}{n}\left(\frac{z_0}{z}\right)^{n-m}\right]$$

$$= A(1+B\cos^2\theta)\left[n\left(\frac{z_0}{z}\right)^m - m\left(\frac{z_0}{z}\right)^n\right]. \qquad (5)$$

At the equilibrium distance, $z = z_0$; and, at $\theta = 0$, the potential well depth becomes

$$\Phi_0 = -\frac{D\alpha_{\parallel}}{2z_0^m}\left[\frac{n-m}{n}\right]. \qquad (6)$$

In Eq. (5), the barrier to rotation is given by the product AB. Thus, the barrier to rotation is caused by the anisotropy of the molecular polarizability, as was observed by King and Benson in their original treatment (Ref. 8).

The task now is to solve the Schroedinger equation using Eqs. (1) and (5). The total eigenfunction is

$$\Psi_T = F(x,y)\,G(\phi)\,R(z,\theta), \qquad (7)$$

where the first term represents the free translation on the surface; the second term, the free rotation around the angle ϕ; and $R(z,\theta)$, the coupled z and θ motion. The energy levels for a particle in a two-dimensional box are obtained from $F(x,y)$, while the G equation yields

$$G(\phi) = 2\pi^{-\frac{1}{2}}e^{iq\phi}, \qquad q = 0, \pm1, \pm2, \qquad (8)$$

The resulting R equation is

$$ER = -\frac{\hbar^2}{2M}\frac{\partial^2 R}{\partial z^2} - \frac{\hbar^2}{2I}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial R}{\partial\theta}\right) - \frac{q^2}{\sin^2\theta}\right]R + \Phi(z,\theta)\,R. \qquad (9)$$

Because of the coupling between the vibrational and rotational motions, it is impossible to solve Eq. (9) exactly. However, the variational method allows the rotational and vibrational energies to be approximated iteratively. According to this method, separability of the motions is forced by choosing

$$R(z,\theta) = S(z)\,T(\theta). \qquad (10)$$

The next step is to minimize the expectation value of the energy to obtain the best solution under this separation.

If Eq. (9) is multiplied by R from Eq. (10) and integrated over z and θ, the expectation energy becomes

$$\int_0^\infty S^2\,dz\int_0^\pi T^2\sin\theta\,d\theta\,\langle E\rangle = \frac{-\hbar^2}{2M}\int_0^\infty SS''\,dz\int_0^\pi T^2\sin\theta\,d\theta - \frac{\hbar^2}{2I}\left\{\int_0^\pi T\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial T}{\partial\theta}\right)d\theta - q^2\int_0^\pi\frac{T^2}{\sin\theta}\,d\theta\int_0^\infty S^2\,dz\right.$$

$$\left. + A\int_0^\pi T^2(1+B\cos^2\theta)\sin\theta\,d\theta\int_0^\infty S^2\left[n\left(\frac{z_0}{z}\right)^m - m\left(\frac{z_0}{z}\right)^n\right]dz\right\}. \qquad (11)$$

In Eq. (11), Eq. (5) has been used for $\Phi(z, \theta)$, and the coefficient of $\langle E \rangle$ on the left-hand side represents the normalization factors of S and T. The requirement is that S and T be varied such that $\delta \langle E \rangle = 0$, under the restriction that S and T remain normalized, i.e., that

$$\delta \int_0^\infty S^2\, dz = 0, \qquad \delta \int_0^\pi T^2 \sin\theta d\theta = 0.$$

These conditions can be met by demanding that

$$\delta \langle E \rangle - \lambda \delta \int_0^\infty S^2\, dz - \mu \delta \int_0^\pi T^2 \sin\theta d\theta = 0, \qquad (12)$$

where λ and μ are undetermined Lagrange multipliers. Eq. (12), when combined with Eq. (11), yields

$$\int_0^\infty 2\delta S \left\{ -\frac{\hbar^2}{2M} S'' + a\left[n\left(\frac{z_0}{z}\right)^m - m\left(\frac{z_0}{z}\right)^n \right] S - \lambda S \right\} dz$$

$$+ \int_0^\pi 2\delta T \left[-\frac{\hbar^2}{2I} \frac{1}{\sin\theta} \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial T}{\partial\theta}\right) + \frac{\hbar^2 q^2}{2I} \frac{T}{\sin^2\theta} \right.$$

$$\left. + bA(1 + B\cos^2\theta)\, T - \mu T \right] \sin\theta d\theta = 0, \qquad (13)$$

where

$$a = A \int_0^\pi T^2 (1 + B\cos^2\theta) \sin\theta d\theta, \qquad (14a)$$

$$b = \int_0^\infty S^2 \left[n\left(\frac{z_0}{z}\right)^m - m\left(\frac{z_0}{z}\right)^n \right] dz. \qquad (14b)$$

Since δS and δT can be chosen independently, the separate equations which result are:

$$-\frac{\hbar^2}{2M} S'' + a\left[n\left(\frac{z_0}{z}\right)^m - m\left(\frac{z_0}{z}\right)^n \right] S = E_z S, \qquad (15)$$

$$-\frac{\hbar^2}{2I}\left[\frac{1}{\sin\theta} \frac{\partial}{\partial\theta}\left(\sin\theta \frac{\partial T}{\partial\theta}\right) - \frac{q^2 T}{\sin^2\theta} \right]$$

$$+ bA(1 + B\cos^2\theta)\, T = E_\theta T, \qquad (16)$$

where

$$E_z \equiv \lambda,$$

$$E_\theta \equiv \mu.$$

Eqs. (15) and (16) show the nature of the coupling between the vibrational and rotational motions. The

Table 2. Calculated energy states of adsorbed molecules

Φ_0, kcal/mole	z_0, A	m	n	E_{vib}, kcal/mole (uncoupled)[a]	E_{vib}, kcal/mole (coupled)	j	g	E_{rot}, kcal/mole (uncoupled)[a]	E_{rot}, kcal/mole (coupled)
1.480	2.40	5	9	0.375	0.337	0	0	0.248	0.214
					0.352	1	0	0.355	0.474
1.940	2.40	5	12	0.497	0.446	0	0	0.248	0.275
					0.466	1	0	0.355	0.513
2.220	2.40	5	15	0.595	0.533	0	0	0.248	0.310
					0.557	1	0	0.355	0.535
0.822	2.47	7	9	0.297	0.282	0	0	0.248	0.111
					0.301	1	0	0.355	0.411
1.540	2.47	7	12	0.470	0.466	0	0	0.248	0.212
					0.477	1	0	0.355	0.473
1.970	2.47	7	15	0.595	0.578	0	0	0.248	0.268
					0.603	1	0	0.355	0.508

[a]Values taken from Ref. 8.

averaged effect of hindered rotation shifts the potential for vibration, while the rotating molecule is influenced by an averaged vibrational motion. An iterative method can be used to simultaneously solve Eqs. (15) and (16), the only requirement being eventual self-consistency.

A reasonable starting point is to assume that the molecule is freely rotating and to consider the term $B\cos^2\theta$ in Eq. (15) to be a perturbation. Then,

$$T(\theta) = P_j^q(\theta),$$

where the $P_j^q(\theta)$ are Legendre polynomials, and the solution to Eq. (16) will be a mixture of these states. The vibrational states can be taken, in the first approximation, as harmonic oscillation states with the depth of the potential well dependent upon the rotational state through a. When harmonic oscillator wave functions are used for S in Eq. (15), the explicit form for b can be calculated from Eq. (14b). With this value of b, the calculation is repeated until the iteration gives self-consistent results. The final solutions must be reasonable in terms of the harmonic oscillator approximation and the assumption of hindered rotation in Eq. (16).

3. Results

Using an IBM 7094 digital computer, the calculations were performed for different values of m, n, z_0, and Φ_0. The energy values chosen were those used by King and Benson so that a direct comparison between the two methods could be made. The results are given in Table 2. Each state of the hindered rotation correlates in the limit of zero barrier height with some state of the free rotator; hence, the states have been identified with the rotational quantum numbers, j and q, of the correlating state of the free rotator whose energy is $j(j+1)\hbar^2/2I$.

The energies have been calculated for the first two rotational states (0,0) and (1,0) and for the zeroth vibrational state. In the uncoupled case, the zero point energy does not change with the rotational state. However, in the coupled case, there is an increase in zero point vibrational energy as the molecule occupies the higher rotational state. A barrier height of 0.7 kcal/mole (Ref. 8) was used to calculate the rotational states. As expected, the uncoupled rotational energy is only dependent on j.

The relationship between the coupled and uncoupled rotational energies is highly dependent upon the adjustable parameters m, n, and z_0. The $j = 1$ state always produces a higher rotational energy for the coupled case. It is to be expected that the higher rotational state is more affected by the coupling, since its axis is more strongly aligned with the surface field and is thus more tightly bound.

These results will not change qualitatively the conclusions of King and Benson. However, they will affect the quantitative results since they will lead to new values of the equilibrium distance of the adsorbed molecule from the surface. The distance from the surface will, in turn, determine the strength of the electric field in which the molecule finds itself. As a consequence of these results, all quantitative aspects of the original electrostatic theory of physical absorption are being reinvestigated.

D. Reaction of O(^{1}D) With Hydrogen, Part I: The Scavenged Case

W. B. DeMore

1. Introduction

The reaction of O(^{1}D) with H_2 has been investigated by photolyzing H_2-O_3 mixtures dissolved in liquid Ar at 87°K. In principle, this reaction is quite simple since only three initial steps are possible:

Deactivation:

$$O(^1D) + H_2 \rightarrow O(^3P) + H_2$$
$$\Delta H = -45.4\,\text{kcal/mole} \quad (1)$$

Insertion:

$$O(^1D) + H_2 \rightarrow H_2O$$
$$\Delta H = -163\,\text{kcal/mole} \quad (2)$$

Abstraction:

$$O(^1D) + H_2 \rightarrow H + OH$$
$$\Delta H = -44\,\text{kcal/mole} \quad (3)$$

Reaction (3) is particularly interesting as a possible source of the species H and OH for use in studying the reactions of these radicals with O_2 and O_3.

2. Experimental Methods

The methods used here were similar to those of a previous study of the O(^{1}D)-CH_4 reaction (*SPS 37-35*,

Vol. IV, p. 222). One exception is that the formerly all-quartz photolysis cell was replaced by a stainless-steel cell of similar design. The advantages of the new cell are a larger coolant reservoir and the capability to withstand higher operating pressures. The optical windows are sealed on the cell by means of indium gaskets.

The H_2 used in this work was Matheson Research Grade, and the D_2 was Matheson C.P. Grade. High-purity tank-grade Ar either was used without further purification or, in some cases, was passed over Ca metal at 400°C to remove traces of O_2.

The light source was a low-pressure mercury lamp with a methanol filter, which limited the radiation absorbed by O_3 almost entirely to 2537 A. Light intensities were monitored by an Eppley thermopile and were measured using a solution of O_3 in liquid N_2 as an actinometer. The quantum yield for O_3 photolysis in liquid N_2 was previously determined to be 0.016 (Ref. 11).

Solutions of H_2 in liquid Ar were prepared by saturating the liquid at known partial pressures of H_2. During the saturation process, the liquid was stirred magnetically to facilitate equilibrium. The H_2 reservoir was fitted with a piston to maintain constant pressure. The Henry's Law constant for the solubility of H_2 in liquid Ar was determined by two methods: In the first, the saturated solution was vaporized into a large bulb and analyzed for H_2 content by gas chromatography. A small correction was made for gas dead space in the liquid cell. In the second method, the amount of H_2 dissolving in the liquid was determined by measuring the piston movement in the H_2 reservoir. The results of the two methods agreed very closely and can be expressed as

$$[H_2] = k P_{H_2}, \tag{4}$$

with k = Henry's Law constant = 0.04 ±0.002 mole/liter-atm, $[H_2]$ = the H_2 concentration in moles/liter, and P_{H_2} = the partial pressure of H_2 in atm. Somewhat less accurate measurements made for the solubility of D_2 in Ar gave k = 0.045 ±0.005 mole/liter-atm.

The products expected from the O(^{1}D)-H_2 reaction are either H_2O or peroxidic compounds of hydrogen which can be decomposed to water by heating. Therefore, the problem of product analysis is essentially that of collecting and measuring the H_2O yield. Unfortunately, due to surface adsorption and also evolution of H_2O as an impurity from the walls of the vacuum system, the quantitative determination of μmole amounts of H_2O is extremely difficult. The method in this work was to obtain a product of D_2O by using pure D_2 and then to dilute the D_2O with a larger, known amount of H_2O.

The detailed procedure was as follows: After photolysis was complete, the solvent was pumped out, with the cell still cooled to 87°K. A known quantity of H_2O was then added to the cell by vacuum transfer from a 1-liter bulb containing H_2O vapor at about 2-cm pressure. To minimize surface adsorption of the H_2O vapor, the bulb was coated on the inside with silicone grease. The cell was then warmed to about 50°C to vaporize and decompose peroxides, and the D_2O-H_2O mixture was transferred by vacuum distillation to an ampul containing zinc metal. The sealed ampul was then heated with an air torch for about 15 min, thereby converting H_2O to H_2 by the reaction

$$H_2O\,(\text{or } D_2O) + Zn \rightarrow ZnO + H_2\,(\text{or } D_2) \tag{5}$$

Because of the preponderance of H_2 in the system, the D_2 appeared in the resulting mixture almost entirely as HD. The mole fraction of HD, X_{HD}, was determined by mass spectrometry, and the original yield of D_2O in moles was calculated by the expression

$$n_{D_2O} = n_{HD}/2 = \frac{n_{H_2O} X_{HD}}{2\,[1 - (X_{HD}/2)]} \cong \frac{n_{H_2O} X_{HD}}{2}, \tag{6}$$

where n_{H_2O} is the number of moles of added H_2O. The method was tested on known samples of D_2O prepared in the following way: A measured quantity of H_2O from the silicone-coated bulb was added to a dummy cell fitted with a small sidearm in which the H_2O could be frozen. The D_2 gas was then added at a known pressure and subsequently equilibrated with the H_2O by sparking with a Tesla coil. Attainment of equilibrium was demonstrated by showing that the D_2 was replaced by nearly pure H_2. The known samples were then analyzed by the above procedure. Results from several test samples are shown below. Corrections for natural D abundance and for slight HD impurity in the D_2 were small and were ignored.

n_{H_2O} Added	n_{D_2} Added	n_{D_2O} Found	% Error
108	1.5	1.4	7
103	1.9	2.1	10
108	3.6	3.5	3
113	3.5	3.6	3

3. Results

a. Quantum yields of O_3 photolysis with added O_2. Addition of O_2 to a given H_2-O_2-Ar mixture was found to lower the quantum yield of O_3 photolysis, Φ_{O_3}. At O_2/O_3 ratios greater than about 10 to 25, no further decrease in Φ_{O_3} could be detected. These results suggest scavenging of H atoms by O_2, and the attainment of a constant Φ_{O_3} shows that the limit of complete scavenging was reached. The scavenging reaction is

$$H + O_2 + Ar \rightarrow HO_2 + Ar, \qquad (7)$$

which is in competition with

$$H + O_3 \rightarrow OH + O_2. \qquad (8)$$

It is important to point out that, while the added O_2 concentration was large compared to that of O_3, it was small compared to that of H_2. Thus, one expects a negligible effect due to the removal of $O(^1D)$ by O_2. This expectation was confirmed by the fact that a lower limiting Φ_{O_3} was reached.

The rate of the initial reaction of $O(^1D)$ with H_2 is related in a simple way to Φ_{O_3} in a given H_2-O_3-Ar solution, provided that O_3 is not destroyed by secondary radical reactions. Loss of O_3 by H atom reactions is completely suppressed by O_2 scavenging, as discussed above. As for OH radicals, previous results (*SPS 37-35*) have shown that OH formed in the following reaction (analogous to Reaction 3) does not destroy O_3:

$$O(^1D) + CH_4 \rightarrow CH_3 + OH. \qquad (9)$$

This result is confirmed in the present work for OH produced in Reaction (3), as shown in Subsection *c* below.

Therefore, under conditions of complete scavenging, the rate of O_3 photolysis depends only on the rate at which $O(^1D)$ reacts with H_2 to form products as compared to the rate at which it is deactivated by Ar or H_2:

$$O_3 + h\nu\,(2537\,A) \rightarrow O_2 + O(^1D) \qquad (10)$$

$$O(^1D) + H_2 \rightarrow O(^3P) + H_2 \qquad (1)$$

$$O(^1D) + Ar \rightarrow O(^3P) + Ar \qquad (11)$$

$$O(^1D) + H_2 \rightarrow H_2O \text{ or } H + OH \qquad (2,3)$$

The O_3 quantum yield is given by

$$\Phi_{O_3} = \frac{(k_2 + k_3)[H_2]}{k_{11} + (k_1 + k_2 + k_3)[H_2]}. \qquad (12)$$

A convenient representation of Eq. (12) is

$$1/\Phi_{O_3} = \frac{(k_1 + k_2 + k_3)}{(k_2 + k_3)} + \frac{k_{11}}{(k_2 + k_3)}\frac{1}{[H_2]}. \qquad (13)$$

From Eq. (13), it can be seen that a plot of $1/\Phi_{O_3}$ versus $1/[H_2]$ should be linear. The intercept yields the ratio of deactivation by H_2 to reaction with H_2, and the slope is a measure of the reaction rate with H_2 relative to deactivation by Ar.

A plot of data according to Eq. (13) is shown in Fig. 2. The low value found for the intercept indicates that deactivation of $O(^1D)$ by H_2 is not a major contributing process relative to reaction to form products. The slope is 6.9, as compared with the value of 5.8 found in previous work for CH_4 (*SPS 37-35*). Since k_{11} is identical in each experiment, the near equality of slopes shows that $O(^1D)$ reacts with H_2 and CH_4 at approximately equal rates. Therefore, for reasons previously discussed (*SPS 37-35*), it follows that the $O(^1D)$-H_2 reaction must have an activation energy equal to or less than 0.5 kcal/mole.

b. Quantum yields of O_3 photolysis without added O_2. Striking increases in the quantum yields of O_3 decomposition are found in the photolysis of H_2-O_3-Ar solutions which have been carefully freed of O_2. As an example, Φ_{O_3} at $[H_2] = 0.076$ mole/liter is 0.013 when excess O_2 is present. However, when O_2 is rigorously excluded, values of Φ_{O_3} as high as 0.37 have been obtained at the same H_2 concentration. The upper limit seems to depend on how low the O_2 concentration can be made. Since O_2

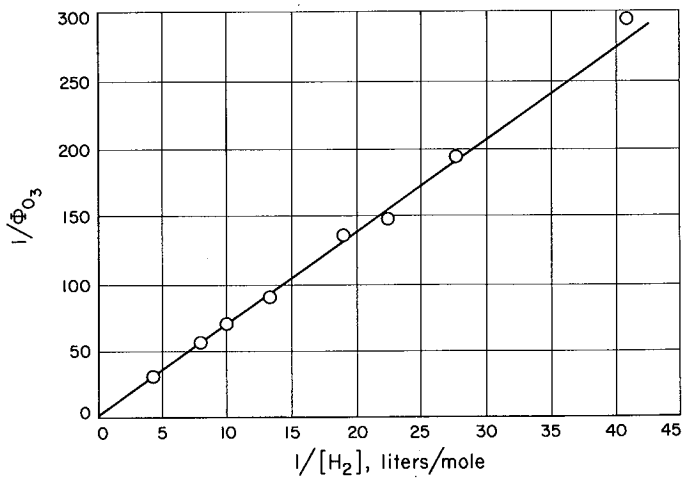


Fig. 2. Effect of $[H_2]$ on Φ_{O_3} in the limit of complete scavenging

is a by-product of the photolysis, Φ_{O_3} decreases with increasing extent of photolysis and approaches the lower, fully scavenged value.

c. Product yields with added O_2. The simplest over-all reaction stoichiometry in the photolysis of H_2-O_3 mixtures would be

$$O_3 + H_2 \rightarrow H_2O + O_2. \quad (14)$$

However, it is possible for the ratio of H_2O formed to O_3 decomposed to be either less than or greater than unity. The first case would arise if O_3 were destroyed by OH radicals, which are not scavengeable by O_2. A possible reaction is:

$$OH + O_3 \rightarrow HO_2 + O_2. \quad (15)$$

The resulting reaction stoichiometry would be

$$2O_3 + H_2 \rightarrow 2HO_2 + O_2,$$

i.e.,

$$2O_3 + H_2 \rightarrow H_2O + \tfrac{5}{2}O_2. \quad (16)$$

The second case, a yield greater than unity, would result if OH radicals reacted with H_2,

$$OH + H_2 \rightarrow H_2O + H, \quad (17)$$

followed by scavenging of H,

$$H + O_2 + Ar \rightarrow HO_2 + Ar. \quad (7)$$

Under the latter circumstances, the reaction stoichiometry would be

$$O_3 + 2H_2 \rightarrow 2H_2O + \tfrac{1}{2}O_2 \quad (18)$$

Product analysis data are listed below for three experiments with D_2:

O_3 Decomposed, μmoles	D_2O Produced, μmoles
2.6	2.6
2.4	2.4
2.2	2.1

In every case, the ratio of D_2O produced to O_3 decomposed is precisely unity. Assuming no different behavior for OH and OD, the results strongly suggest that the OH produced in Reaction (3) does not react with either O_3 or H_2. Failure of OH to react with O_3 is in agreement with conclusions from previous work (*SPS 37-35*).

4. Conclusions and Discussion

With complete scavenging by O_2, the quantum yields of O_3 decomposition in the photolysis of H_2-O_3 mixtures are very similar to those previously found for CH_4-O_3 mixtures. These results show that the $O(^1D)$-H_2 reaction has a rate very nearly equal to that of the $O(^1D)$-CH_4 reaction and an activation energy near zero.

Deactivation of $O(^1D)$ by H_2 is not a major reaction, but the data are not sufficient to exclude the possibility of a small contribution from this process.

The scavenging effect of O_2 indicates that Reaction (3), giving H and OH, is an important path. Reaction (2), which gives H_2O directly, cannot be distinguished from Reaction (3) on the basis of the product analysis data, since each eventually leads to H_2O. Neither can they be distinguished by the magnitude of the scavenging effect, since a chain decomposition of O_3 of unknown length is apparently occurring. The question of separating Reactions (2) and (3) may, in fact, be academic, since both may proceed through a common intermediate:

$$O(^1D) + H_2 \rightarrow (H_2O)^* \xrightarrow{+Ar} H_2O + Ar \quad (2a)$$

$$\longrightarrow H + OH \quad (3a)$$

If so, the relative contribution of Reactions (2a) and (3a) depends only on the rate at which $(H_2O)^*$ is deactivated as compared to the rate of unimolecular decomposition.

The most surprising result of this work is the more than 25-fold increase in the O_3 quantum yield upon exclusion of O_2 from the reaction system. This is to be compared with the 1.8-fold increase found in the CH_4 experiments (*SPS 37-35*). These results strongly suggest that a chain decomposition of O_3 is initiated by H atoms from Reaction (3). Experiments now under way to identify the chain and the nature of the chain carriers will be discussed in Part II of this report in a future issue of the *SPS*, Vol. IV.

The observation that one H_2O is formed for each O_3 decomposed in the scavenged case provides useful information on the reactivity of OH and HO_2 radicals toward O_3. Since no O_3 wastage is occurring, it may be concluded that OH is not able to initiate a Norrish-type

catalyzed chain decomposition (Refs. 12–14) of O_3. The propagating steps in the Norrish chain are

$$OH + O_3 \rightarrow HO_2 + O_2, \quad (15)$$

$$HO_2 + O_3 \rightarrow OH + 2O_2. \quad (19)$$

The results also show that OH from Reaction (3) does not react with H_2,

$$OH + H_2 \rightarrow H_2O + H, \quad (17)$$

which is not unexpected since Reaction (17) has an activation energy of about 6 kcal/mole (Ref. 15). However, it is important to note that Reaction (3) is exothermic by 44 kcal/mole and is capable of exciting OH to the 4th vibrational level. If vibrationally excited OH were efficiently formed in Reaction (3), and if the vibrational excitation persisted until a collision with H_2 occurred, then the observed reaction rate might be faster than that for vibrationally cold OH. The experimental results, of course, indicate that, even if vibrationally excited OH is formed in Reaction (3), it does not react rapidly with either H_2 or O_3.

In terms of possible vibrational excitation, there are three species of OH present: (1) vibrationally cold OH; (2) OH with up to 4 vibrational quanta from Reaction (3); and (3) OH with up to 9 vibrational quanta from Reaction (8), which is exothermic by 77 kcal/mole. Evidence for possible differences in reactivity of these three types of OH will be discussed more fully in Part II of this report.

E. Energy Momentum of the Electron–Photon Interaction and Gage Invariance

M. M. Saffren

In *SPS 37-38*, Vol. IV, p. 161, we derived a theorem which related "out" operators of the interacting photon–electron fields to the "in" operators of those fields. Using the theorem, we proved that the total energy of the interacting fields was the same whether expressed in terms of "in" operators or in terms of the "out" operators; i.e., the energy was shown to be "in"-"out" invariant. This seems strange at first, since the total energy is merely the sum of the energies of the noninteracting free fields. One may wonder what happened to the interaction energy; the answer is that this energy got cancelled. The "out" electron energy when expressed in "in" operators becomes the "in" electron energy plus an interaction energy, while the "out" photon energy when expressed in "in" operators becomes the "in" photon energy minus the same interaction energy. When the two "out" energies are added, the interaction energy disappears. This cancellation is another expression of energy conservation: any kinetic energy lost or gained by the free electron field must appear in or disappear from the free photon field and conversely.

Here we derive an expression for this interaction energy and the corresponding interaction momentum by utilizing the theorem discussed above. In deriving the interaction energy, we incidentally prove that the 4-divergence of the electromagnetic 4-potential is a quantity invariant under the interaction of the electron and electromagnetic fields by explicitly demonstrating the gage invariance of the S matrix. The form taken by Maxwell's equations in quantum electrodynamics is also discussed.

We write the interaction energy in terms of the "in" and "out" photon energy operators, which are simpler and more general than the corresponding electron energy operators. We have

$$H_{int} = S^{-1}(H_0)_{ph} S - (H_0)_{ph} = -[S^{-1}(H_0)_{el} S - (H_0)_{el}]. \quad (1)$$

Using the theorem of *SPS 37-38*, we have (taking $\hbar = c = 1$)

$$S^{-1}(H_0)_{ph} S = (H_0)_{ph} + i\int [d\rho]\, d\tau\, e^{-\alpha|\tau|} S(\tau) [\mathscr{H}(\rho,\tau), (H_0)_{ph}] S(\tau). \quad (2)$$

Thus,

$$H_{int} = \int [d\rho]\, d\tau\, e^{-\alpha|\tau|} S^{-1}(\tau) \frac{\partial a^{\mu}(\rho,\tau)}{\partial \tau} j_{\mu}(\rho,\tau) S(\tau). \quad (3)$$

We can easily write this in terms of Heisenberg operators:

$$H_{int} = \int [d\rho]\, d\tau\, e^{-\alpha|\tau|} \frac{\partial A^{\mu}}{\partial \tau}(\rho,\tau) J_{\mu}(\rho,\tau). \quad (4)$$

Note that this is not the same as the "interaction energy" $\mathscr{H}(\tau)$. Unlike H_{int}, $\mathscr{H}(\tau)$ is time-dependent and is not expressed in terms of Heisenberg operators. Even the "interaction energy" $H(\tau) = S^{-1}(\tau)\,\mathscr{H}(\tau)\,S(\tau)$, which is indeed expressed in terms of Heisenberg operators, is still quite different from the interaction energy expressed above.

In this form, the gage-invariant nature of the interaction energy is not apparent. To make it apparent, we now transform the integral that appears in Eq. (4). The essential part of the integrand can be written as

$$-\frac{\partial \mathbf{A}}{\partial \tau}\cdot\mathbf{J}+\frac{\partial A_0}{\partial \tau}J_0. \tag{5}$$

If we use current conservation $\partial_\mu J^\mu = 0$, this becomes

$$-\frac{\partial \mathbf{A}}{\partial \tau}\cdot\mathbf{J}+\frac{\partial}{\partial \tau}(A_0 J_0)+A_0\,\nabla\cdot\mathbf{J}, \tag{6}$$

which in turn becomes

$$\left(-\frac{\partial \mathbf{A}}{\partial \tau}-\nabla A_0\right)\cdot\mathbf{J}+\frac{\partial}{\partial \tau}(A_0 J_0)+\nabla\cdot(A_0\,\mathbf{J})$$

$$=\mathbf{E}\cdot\mathbf{J}+\frac{\partial}{\partial \tau}(A_0 J_0)+\nabla\cdot(A_0 J). \tag{7}$$

Thus, we obtain for the interaction energy

$$\int[d\boldsymbol{\rho}]\,d\tau\,e^{-\alpha|\tau|}\,\mathbf{E}\cdot\mathbf{J} \tag{8}$$

and, similarly, for the interaction momentum

$$\int[d\boldsymbol{\rho}]\,d\tau\,e^{-\alpha|\tau|}\,\{-\mathbf{E}\,J_0+\mathbf{B}\times\mathbf{J}\}. \tag{9}$$

Here **E** and **B** denote the Heisenberg operators of the electromagnetic fields—expressions which are quite classical in appearance. The first is the time-and-space integral of the power density due to the interaction of an electron and photon; the second is the time-and-space integral of the force density.

However, if we write the equations satisfied by **E** and **B**, we do not obtain Maxwell's equations exactly. We find instead that

$$\nabla\cdot\mathbf{E}=-\frac{\partial\nabla\cdot\mathbf{A}}{\partial\tau}-\nabla^2 A_0$$

$$=-S^{-1}(t)\left(\frac{\partial\nabla\cdot\mathbf{a}}{\partial t}+\nabla^2 a_0\right)S(t)$$

$$=-S^{-1}(t)\left(\frac{\partial\nabla\cdot\mathbf{a}}{\partial t}+\frac{\partial^2}{\partial t^2}a_0\right)S(t)$$

$$\equiv -S^{-1}(t)\frac{\partial\chi}{\partial t}S(t) \tag{10}$$

and

$$\frac{\partial\mathbf{E}}{\partial t}=-S^{-1}(t)\,\nabla\chi\,S(t)-\mathbf{J}+\nabla\times\mathbf{B}. \tag{11}$$

The other two Maxwell equations are the familiar ones. The quantity χ is the 4-divergence of the electromagnetic potential of the "in" field. We shall evaluate the term $S^{-1}(t)\,\partial^\mu\chi\,S(t)$ which appears in these equations.

To do this, it is convenient to introduce an operator which is a functional of χ. This operator, the generator of infinitesimal gage transformations, is (Ref. 16)

$$U(\Lambda)=\int[d^3\mathbf{r}]\left[\frac{\partial\Lambda}{\partial t}\chi-\frac{\partial\chi}{\partial t}\Lambda\right], \tag{12}$$

where $\partial_\mu\partial^\mu\Lambda = 0$. It is easy to see that $U(\Lambda)$ is a time-independent operator. We can verify this directly either by Fourier expansion of the integrand or by noting that the divergence of $(\partial_\mu\Lambda)\,\chi-(\partial_\mu\chi)\,\Lambda$ vanishes. Thus, by a well-known theorem, the space integral of the time component (Eq. 12) is a time-independent scalar. It is easily verified that

$$[U(\Lambda), a_\mu]=\partial_\mu\Lambda, \tag{13}$$

so that indeed $U(\Lambda)$ does generate gage transformations.

To evaluate $S^{-1}(t)\,\partial_\mu\chi\,S(t)$, we shall evaluate $S^{-1}(t)\,U(\Lambda)\,S(t)$, which can be written as (*SPS 37-38*)

$$S^{-1}(t)\,U(\Lambda)\,S(t) = U(\Lambda) + i\int_{-\infty}^{t} d\tau\, e^{-\alpha|\tau-t|}\,S^{-1}(\tau) \times [\mathscr{H}(\tau), U(\Lambda)]\,S(\tau). \quad (14)$$

But, using Eq. (13), we can write

$$[\mathscr{H}(\tau), U(\Lambda)] \text{ as } \int [d^3\rho]\, e^{-\alpha|\tau-t|}\,\partial_\mu(J^\mu\Lambda).$$

The integrand then becomes

$$\int [d^3\rho]\, e^{-\alpha|\tau|}\,J^\mu\,\partial_\mu\Lambda = \int [d^3\rho]\, e^{-\alpha|\tau-t|}\,\partial_\mu(J^\mu\Lambda). \quad (15)$$

If the surface integral which results from applying the Gauss theorem is disregarded, the above expression becomes

$$\int [d^3\rho]\, e^{-\alpha|\tau|}\,\frac{\partial}{\partial\tau}(J_0\Lambda). \quad (16)$$

Finally, we find that

$$S^{-1}(t)\,U(\Lambda)\,S(t) = U(\Lambda) - \int [d^3\rho]\,J_0(\rho,t)\,\Lambda(\rho t). \quad (17)$$

(Incidentally, by proceeding in an analogous way from the equation

$$S^{-1}\,U(\Lambda)\,S = U(\Lambda) + i\int_{-\infty}^{\infty} d\tau\, e^{-\alpha\tau}\,S^{-1}(\tau) \times [\mathscr{H}(\tau), U(\Lambda)]\,S(\tau), \quad (18)$$

we find that

$$S^{-1}\,U(\Lambda)\,S = U(\Lambda). \quad (19)$$

This shows that S is gage-invariant, even though $S(t)$ is not.)

Returning to Eq. (17), we expand

$$\Lambda(x) = \int \frac{[d^3\mathbf{k}]}{|\mathbf{k}|}\,[\Lambda_+(\mathbf{k})\,e^{-ikx} + \Lambda_-(\mathbf{k})\,e^{ikx}] \quad (20)$$

so that

$$U(\Lambda) = \int \frac{[d^3\mathbf{k}]}{|k|}\int [d^3\mathbf{x}]\left\{-\Lambda_+(\mathbf{k})\,e^{-ikx}\,\chi(x) + \Lambda_-(\mathbf{k})\,e^{ikx}\,\chi(x) - \frac{\Lambda_+(\mathbf{k})}{|k|}\,e^{-ikx}\,\frac{\partial\chi}{\partial\chi_0} - \frac{\Lambda_-(\mathbf{k})}{|k|}\,e^{ikx}\,\frac{\partial\chi}{\partial\chi_0}\right\}. \quad (21)$$

Thus, if we compare the coefficients of $\Lambda_+(\mathbf{k})$ on both sides of Eq. (17), we obtain

$$\int [d^3\mathbf{x}]\,e^{-i\mathbf{k}\mathbf{x}}\,S^{-1}(x_0)\left[\chi(x) + \frac{\partial\chi}{|\mathbf{k}|\,\partial x_0}\right]S(x_0) = \int [d^3\mathbf{x}]\,e^{-i\mathbf{k}\mathbf{x}}\left(\chi(x) + \frac{1}{|\mathbf{k}|}\,\frac{\partial\chi}{\partial x_0}\right) - \int [d^3\mathbf{x}]\,J_0(x)\,\frac{e^{-i\mathbf{k}\mathbf{x}}}{|k|}. \quad (22)$$

Using the corresponding equation obtained by comparing coefficients of $\Lambda_-(\mathbf{k})$, we find that, after adding and subtracting that equation from the one above,

$$S^{-1}(x_0)\,\chi(x)\,S(x_0) = \chi(x), \quad (23)$$

$$S^{-1}(x_0)\,\frac{\partial\chi(x)}{\partial x_0}\,S(x_0) = -J_0(x) + \frac{\partial\chi}{\partial x_0}. \quad (24)$$

The second equation, however, is not really independent of the first, since it is merely a time derivative of the first. The modified Maxwell equations now read

$$\nabla\cdot\mathbf{E} = J_0 - \frac{\partial\chi}{\partial t},$$

$$\frac{\partial\mathbf{E}}{\partial t} = -\mathbf{J} - \nabla\chi + \nabla\times\mathbf{B}. \quad (25)$$

If the following restrictions on state vectors $|\mathcal{S}\rangle$ are made,

$$\chi_P|\mathcal{S}\rangle = 0, \qquad \langle\mathcal{S}|\chi_N = 0 \tag{26}$$

(where the subscripts P and N refer to positive and negative frequency parts, respectively), then, in this subspace of Hilbert space, Maxwell's equations are obeyed. We need not apply the restrictions of (26) separately to the "out" states, since we have already explicitly proved that the S matrix commutes with χ.

That explicit proof, the reduced form of the Maxwell equations (Eq. 25), and the explicit "classical" form of the interaction energy and momentum (Eqs. 8 and 9) are the principal results presented in this article. Attention must be called to the fact that, both here and in *SPS 37-38*, theorems have been proved without recourse to a perturbation expansion of the S matrix. These proofs have used only the defining properties of the S matrix.

F. Null Electromagnetic Fields

H. D. Wahlquist and F. B. Estabrook

We have written the Maxwell–Lorentz equations for a null field in general space–time in dyadic form:

$$\nabla\cdot\mathbf{E} = 2\Omega\cdot\mathbf{k}\times\mathbf{E} + \rho_c, \tag{1}$$

$$-\mathbf{E}\times(\nabla\times\mathbf{k}) - \mathbf{k}\times(\nabla\times\mathbf{E}) + (\nabla\cdot\mathbf{E})\mathbf{k} - (\nabla\cdot\mathbf{k})\mathbf{E} - 2\mathbf{E}' + \mathbf{a}\times(\mathbf{k}\times\mathbf{E}) = \dot{\mathbf{E}} + \mathbf{j} - \mathbf{E}\cdot\mathbf{S}^* + (\mathrm{Tr}\,\mathbf{S})\mathbf{E}, \tag{2}$$

$$(\nabla\times\mathbf{k})\cdot\mathbf{E} - (\nabla\times\mathbf{E})\cdot\mathbf{k} = -2\Omega\cdot\mathbf{E}, \tag{3}$$

$$\nabla\times\mathbf{E} + \mathbf{a}\times\mathbf{E} = -\dot{\mathbf{k}}\times\mathbf{E} - \mathbf{k}\times\dot{\mathbf{E}} + \mathbf{k}\times\mathbf{E}\cdot\mathbf{S}^* - (\mathrm{Tr}\,\mathbf{S})\mathbf{k}\times\mathbf{E}, \tag{4}$$

where $\mathbf{E}$ is the electric field and $\mathbf{k}$ is a unit propagation vector normal to $\mathbf{E}$; thus, the magnetic field is $\mathbf{H} = \mathbf{k}\times\mathbf{E}$. The prime operator is $\mathbf{k}\cdot\nabla$, ρ_c is the charge density, $\mathbf{j}$ is the current density, and the other quantities refer to the dyadic reference frame.

We have analyzed these in the source-free case, $\rho = 0$, $\mathbf{j} = 0$. We write the magnitude of $\mathbf{E}$ as ε and introduce a unit polarization vector $\hat{\mathbf{E}}$ by

$$\mathbf{E} = \varepsilon\hat{\mathbf{E}}. \tag{5}$$

Following Sachs in defining a parallax distance r_p,

$$r_p = \frac{2}{\nabla\cdot\mathbf{k} + (\mathbf{I} - \mathbf{kk}):\mathbf{S}}, \tag{6}$$

we find first three propagation equations,

$$\frac{\varepsilon' + \dot{\varepsilon}}{\varepsilon} + \frac{1}{r_p} + \mathbf{a}\cdot\mathbf{k} + \mathbf{k}\cdot\mathbf{S}\cdot\mathbf{k} = 0, \tag{7}$$

$$\hat{\mathbf{E}}' + \dot{\hat{\mathbf{E}}} + \omega\times\hat{\mathbf{E}} = (\tfrac{1}{2}\mathbf{k}\cdot\nabla\times\mathbf{k} + \Omega\cdot\mathbf{k})\,\mathbf{k}\times\hat{\mathbf{E}} + (\mathbf{k}\cdot\mathbf{S} + \mathbf{a} + \Omega\times\mathbf{k})\cdot\hat{\mathbf{E}}\mathbf{k}, \tag{8}$$

$$\mathbf{k}' + \dot{\mathbf{k}} + \omega\times\mathbf{k} = -\mathbf{k}\cdot\mathbf{S} - \mathbf{a} - \Omega\times\mathbf{k} + (\mathbf{k}\cdot\mathbf{S}\cdot\mathbf{k} + \mathbf{a}\cdot\mathbf{k})\,\mathbf{k}; \tag{9}$$

secondly, equations for the divergences of $\hat{\mathbf{E}}$ and $\hat{\mathbf{H}} = \mathbf{k}\times\hat{\mathbf{E}}$,

$$\nabla\cdot\hat{\mathbf{E}} = 2\Omega\cdot\mathbf{k}\times\hat{\mathbf{E}} - \varepsilon^{-1}\hat{\mathbf{E}}\cdot\nabla\varepsilon, \tag{10}$$

$$\nabla\cdot(\mathbf{k}\times\hat{\mathbf{E}}) = -2\Omega\cdot\hat{\mathbf{E}} - \varepsilon^{-1}\mathbf{k}\times\hat{\mathbf{E}}\cdot\nabla\varepsilon; \tag{11}$$

and, finally, the Robinson equation for the projected strain dyadic $\mathbf{D}$ of the unit propagation vector $\mathbf{k}$. That is, if we set

$$\nabla\mathbf{k} = \tfrac{1}{2}\nabla\cdot\mathbf{k}\,(\mathbf{I} - \mathbf{kk}) - \tfrac{1}{2}(\mathbf{k}\cdot\nabla\times\mathbf{k})\mathbf{k}\times\mathbf{I} + \mathbf{kk}' + \mathbf{D}, \tag{12}$$

then it follows from the Maxwell–Lorentz equations that

$$\mathbf{D} = -(\mathbf{I} - \mathbf{kk}) \cdot \mathbf{S} \cdot (\mathbf{I} - \mathbf{kk}) + \tfrac{1}{2}[\mathbf{S} : (\mathbf{I} - \mathbf{kk})](\mathbf{I} - \mathbf{kk}). \tag{13}$$

In a rigid frame, in particular, we have $\mathbf{D} = 0$.

We note how the propagation equations simplify if a dyadic frame is used that is adapted to the k congruence; setting, as we may always do (*SPS 37-38*, Vol. IV, p. 159),

$$\mathbf{k} \cdot \mathbf{S} - \mathbf{k} \times \Omega + \mathbf{a} = 0, \tag{14}$$

Eqs. (7), (8), and (9) become

$$\frac{\varepsilon' + \dot{\varepsilon}}{\varepsilon} + \frac{1}{r_p} = 0, \tag{15}$$

$$\hat{\mathbf{E}}' + \dot{\hat{\mathbf{E}}} + \omega \times \hat{\mathbf{E}} = (\tfrac{1}{2}\mathbf{k} \cdot \nabla \times \mathbf{k} + \Omega \cdot \mathbf{k})\, \mathbf{k} \times \hat{\mathbf{E}}, \tag{16}$$

$$\mathbf{k}' + \dot{\mathbf{k}} + \omega \times \mathbf{k} = 0. \tag{17}$$

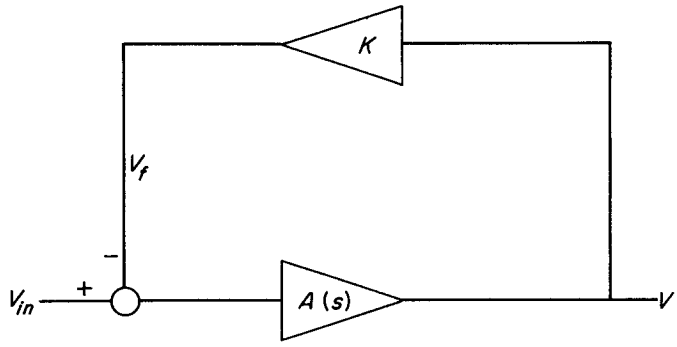
G. Relationship Between Stability, Passband, and Loop Gain for Pulse Amplifiers Used in Differential Nuclear Spectroscopy

L. L. Lewyn

The determination of the basic components of a differential spectrum is usually influenced by the stability and linearity of the measurement system. Lack of stability contributes to peak shift, and poor linearity contributes to peak spreading and improper amplitude ratios. If a monoenergetic spectral line of amplitude V is caused to shift or spread an amount ΔV by either poor stability or linearity, then a measure of the tendency to spread may be defined in terms of a spreading factor:

$$S = \Delta V / V. \tag{1}$$

Let V represent the peak output voltage of the closed-loop system shown below:



Note that, since

$$V_{in} - V_f = V_e, \tag{2}$$

$$V = V_f / K, \tag{3}$$

assuming K is constant, then

$$V = (V_{in} - V_e)/K. \tag{4}$$

Therefore,

$$S = \Delta (V_{in} - V_e)/(V_{in} - V_e). \tag{5}$$

For a monoenergetic input, V_{in} is constant. Since the error voltage is usually a small fraction of the input voltage,

$$S \cong -\Delta V_e / V_{in}. \tag{6}$$

Therefore, the system stability is intimately related to the stability of the error voltage at the time the pulse peak is measured.

The Laplace transform of the error voltage may be expressed in terms of the input voltage and the forward open-loop transfer function $A(s)$:

$$V_f(s) = V_e(s)\, A(s)\, K. \tag{7}$$

Combining Eq. (7) with the transform of Eq. (2),

$$V_e(s) = \frac{V_{in}(s)}{[1 + A(s)\, K]}. \tag{8}$$

If the input is a step function of amplitude B and the transfer function $A(s)$ represents an amplifier with low frequency gain A and single pole at $s = -1/t_0$, then:

$$V_{in}(s) = B/s, \tag{9}$$

$$A(s) = A/(t_0 s + 1). \tag{10}$$

Combining Eqs. (8), (9), and (10), we obtain:

$$V_e(s) = \frac{B}{s\{1 + [AK/(t_0 s + 1)]\}}. \tag{11}$$

Rearranging terms,

$$V_e(s) = \frac{B}{(1 + AK)} \frac{(t_0 s + 1)}{s\left(\frac{t_0}{1 + AK} s + 1\right)}. \tag{12}$$

Taking the inverse transform,

$$V_e(t) = \frac{B}{(1 + AK)}\left\{1 + AK \exp\left[\frac{-(1 + AK)\, t}{t_0}\right]\right\}. \tag{13}$$

The error voltage is therefore an exponentially decaying waveform with initial value B decaying to a final value of $B/(1 + AK)$ with time constant $t_0/(1 + AK)$. The error voltage is settled out to less than 1% of its final value by the time $t = 5\,t_0/(1 + AK)$. One might then conclude that it is appropriate to sample the system any time following $t = 5t_0/(1 + AK)$. Solving for S, we have, from Eq. (6),

$$S = -\,\Delta\,[1/(1 + AK)] \tag{14}$$

or

$$S = \frac{\Delta A}{A}\left[\frac{1}{(1 + AK)\left(1 + \frac{\Delta A}{A}\right) + \left(1 + \frac{1}{AK}\right)}\right]. \tag{15}$$

For $AK >> 1$,

$$S = \frac{\Delta A}{A}\left(\frac{1}{AK}\right)\left(\frac{1}{1 + \frac{\Delta A}{A}}\right). \tag{16}$$

Therefore, for small variation in A, the stability variation is the same as that of A divided by the loop gain. Note that, if the system is sampled at $t = 5\,t_0/(1 + AK)$ and if $t_0/(1 + AK)$ varies 20%, an additional 1% variation in stability will be introduced. This sensitivity to sampling time for $t \leq 5t_0/(1 + AK)$ is present since the initial value of the error signal is nearly equal to the amplitude of the input step.

If the input signal has an exponential rise and fall, the time at which the output signal peak is sampled is determined by the time at which the peak occurs. Under conditions of adequate loop gain and bandwidth, the time at which the output peak occurs is approximately equal to the time at which the input signal peak occurs. A typical signal from a solid-state detector preamplifier has equal rise and fall times ($t_r = t_f = t_1$). Such a signal may be expressed as:

$$V_{in} = Bt \exp(-t/t_1). \tag{17}$$

Setting

$$dV/dt = B\,[1 - (t/t_1)] \exp(-t/t_1) = 0, \tag{18}$$

we find

$$t = t_1, \tag{19}$$

$$(V_{in})_{\max} = Bt_1/e. \tag{19a}$$

Applying the equal rise- and fall-time signal to the amplifier used in the first example, we have, from Eq. (8),

$$V_e(s) = V_{in}(s)/[1 + KA(s)], \tag{20}$$

where

$$V_{in}(s) = B/[s + (1/t_1)]^2, \tag{21}$$

$$A(s) = A/(t_0 s + 1). \tag{22}$$

Substituting in Eq. (20), we obtain:

$$V_e(s) = B\frac{[s + (1/t_0)]}{[s + (1/t_1)]^2}\left(\frac{1}{s + \frac{1 + AK}{t_0}}\right). \tag{23}$$

Taking the inverse transform,

$$V_e(t) = \frac{-\frac{AK}{t_0}}{\left(\frac{1}{t_1} - \frac{1+AK}{t_0}\right)} B \exp\left[-\frac{(1+AK)}{t_0} t\right] + \left[\frac{\frac{1}{t_0} - \frac{1}{t_1}}{\frac{1+AK}{t_0} - \frac{1}{t_1}} t + \frac{\frac{AK}{t_0}}{\left(\frac{1+AK}{t_0} - \frac{1}{t_1}\right)^2}\right] B \exp\left(-\frac{t_1}{t}\right). \quad (24)$$

Substituting into Eq. (6) and letting $t = t_1$,

$$S = -\Delta\left(\frac{1}{\left(\frac{1+AK}{t_0} - \frac{1}{t_1}\right)^2}\left(\frac{AK}{t_1 t_0}\right)\left\{1 - \exp\left[-\left(\frac{1+AK}{t_0} - \frac{1}{t_1}\right)t_1\right]\right\} + \frac{\frac{1}{t_0} - \frac{1}{t_1}}{\left(\frac{1+AK}{t_0} - \frac{1}{t_1}\right)}\right). \quad (25)$$

Assuming that the loop gain $AK > 10$,

$$S \cong -\Delta\left(\frac{1}{(1+AK)\frac{t_1}{t_0}}\left\{1 - \exp\left[1 - (1+AK)\frac{t_1}{t_0}\right]\right\} + \frac{1}{1 + \frac{AK}{1 - \frac{t_0}{t_1}}}\right). \quad (26)$$

For amplifier time constants which are small compared to the input signal time constant ($t_0/t_1 << 1$), the last term of Eq. (26) dominates and we have:

$$S = \frac{\Delta A}{A}\left(\frac{1}{AK}\right)\left(\frac{1}{1 + \frac{\Delta A}{A}}\right). \quad (27)$$

For amplifier time constants which are large compared to the input signal time constant ($t_0/t_1 >> 1$), the first term of Eq. (26) dominates. When the loop gain in the frequency region near the maximum input power spectral density (AKt_1/t_0) is >10, the exponential is small and we have:

$$S \cong \frac{\Delta A}{A}\left(\frac{1}{AKt_1/t_0}\right)\left(\frac{1}{1 + \frac{\Delta A}{A}}\right). \quad (28)$$

Therefore, for amplifier time constants which are large or small compared to the input signal time constant, the spreading tendency is reduced by the loop gain at the frequency corresponding to the maximum input power spectral density. For amplifier time constants comparable to the input signal time constant, Eq. (26) is solved to yield:

$$S \cong \frac{\Delta A}{A}\left\{\frac{t_0/t_1}{(1+AK)\left(1 + \frac{\Delta A}{A}\right)} + \frac{1}{\left[1 + \frac{AK}{1 - (t_0/t_1)}\right]\left(1 + \frac{\Delta A}{A}\right) + \left[1 + \frac{1 - (t_0/t_1)}{AK}\right]}\right\} \quad (29)$$

References

1. Chapman, O. L., and King, R. W., *Journal of the American Chemical Society*, Vol. 86, p. 1256, 1964.

2. Mathias, A., *Analytica Chimica Acta*, Vol. 31, p. 598, 1964.

References (Cont'd)

3. Goodlett, V. W., *Analytical Chemistry,* Vol. 37, p. 431, 1965.
4. Elleman, D. D., Manatt, S. L., and Pearce, C. D., *Journal of Chemical Physics,* Vol. 42, p. 650, 1965.
5. Lynden-Bell, R. M., and Sheppard, N., *Proceedings of the Royal Society,* London, Vol. A269, p. 385, 1962.
6. Graham, D.M., and Holloway, C. E., *Canadian Journal of Chemistry,* Vol. 41, p. 2114, 1963.
7. Hobgood, R. T., Jr., and Goldstein, J. H., *Journal of Molecular Spectroscopy,* Vol. 12, p. 76, 1964.
8. King, J., Jr., and Benson, S. W., *Journal of Chemical Physics,* Vol. 44, p. 1007, 1966.
9. Hill, T. L., *Journal of Chemical Physics,* Vol. 16, p. 181, 1948.
10. Hirschfelder, J. O., Curtis, C. F., and Bird, R. B., *Molecular Theory of Gases and Liquids,* p. 948, John Wiley and Sons, Inc., New York, 1954.
11. DeMore, W., and Raper, O. F., *The Journal of Chemical Physics,* Vol. 37, No. 9, p. 2048, 1962.
12. Norrish, R. G. W., and Wayne, R. P., *Proceedings of the Royal Society,* London, Vol. 288A, p. 361, 1965.
13. McGrath, W. D., and Norrish, R. G. W., *Nature,* Vol. 183, p. 235, 1958.
14. McGrath, W. D., and Norrish, R. G. W., *Proceedings of the Royal Society,* London, Vol. 254A, p. 317, 1960.
15. Kaufman, F., *Annales de Geophysique,* Vol. 20, p. 106, 1964.
16. Schweber, S. S., *Introduction to Relativistic Theory of Quantized Fields,* Row Peterson, Evanston, Illinois, 1961.

TELECOMMUNICATIONS DIVISION

XV. Spacecraft Telemetry and Command

A. Advanced Data Processing Systems

R. F. Trost

1. Introduction

In *SPS 37-36*, Vol. IV, pp. 237–241, the general philosophy of an advanced engineering data handling system was presented. This report continues that discussion by presenting a more detailed explanation of one of the data compression subsystems. Specifically, the system operation of the F3HS transfer function will be explained from the functional block diagram level.

2. Engineering Data Handling System (EDHS)

a. History. In the past, the JPL Spacecraft Telemetry and Command Section has been studying two general classes of spacecraft data: engineering telemetry data and scientific video data. Although some *a priori* studies were performed in each class, most of them were done in an *a posteriori* manner.

An important result from the *a posteriori* studies of engineering telemetry data was the development of a spacecraft engineering data handling system. Direct consultations with the data users dictated the organization and characteristics of the EDHS.

b. Organization. Basically, the EDHS is divided into four subsystems, as shown in Fig. 1. Each has a transfer function which is considered best for the subset of measurements that it processes. These subsets are termed operational, operational–performance, and performance. They are processed respectively by the F1 absolute rate

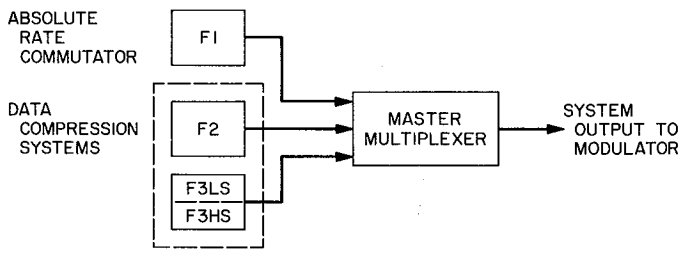


Fig. 1. Engineering data handling system

commutator subsystem, the F2 data compression (with minimum time delay feature) subsystem, and the F3 data compression subsystem.

Characteristics of the F1 subsystem are simple. Measurements are sampled at one of two constant rates, regardless of changes in the communication link bit rate. The constant rates are stored in a small memory and selected individually via a "user control line," which is sampled simultaneously with the user's measurement. Operational measurements processed by F1 are usually those which are highly critical to the success of the mission.

The characteristics of the F2 and F3 subsystems are similar, with one exception. Although both contain data compression and the attendant buffer memories, the F2 subsystem also has a minimum time delay feature. This permits some of the more important data to bypass the buffer and to be received in real time. The delay feature is necessary because many of the operational–performance measurements processed by F2 are normally very static; however, sudden abnormal changes must be relayed to the receiving station as soon as possible if corrective measures are to be initiated.

In addition to F1, F2, and F3, there is a fourth subsystem, called the master multiplexer. This is simply a time multiplexer with a programmed priority scheme for accepting the outputs from the other three subsystems and producing a continuous data bit stream to the modulator.

A command reprogramming capability exists for updating processing parameters for each channel. Also, a provision has been made for periodic confidence sampling which produces a complete readout of all channels at predetermined times.

The F3 subsystem is divided into a high speed (F3HS) and a low speed (F3LS) category because of the large number of channels and the wide dynamic range of sampling rates required. Therefore, the separation was primarily a result of the need for efficiency in the hardware implementation. With this one exception they are identical, so only an explanation of the F3HS subsystem will be given.

3. F3HS Subsystem

The F3HS subsystem processes the "performance" class of measurements using only zero-order data compression techniques. There is no direct, real-time path for data information except an apparent one when the buffer is nearly empty or when periodic confidence sampling occurs.

a. Organization. Fig. 2 is a functional block diagram of the F3HS subsystem, which consists of: a multiplexer and an analog-to-digital converter (ADC); a processor; a processing memory; and a buffer memory.

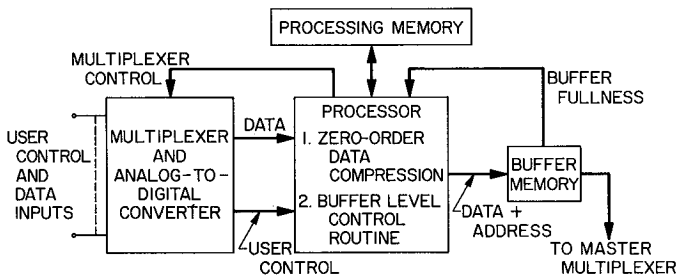


Fig. 2. F3HS subsystem

The input to the processor is several bits of digital information concerning the actual sensor measurement, and one extra bit for control information to the processor.

The processor accepts this information and performs two functions. First, a zero-order data compression routine processes the incoming sample. If it exceeds the aperture it is called a "significant sample." Second, certain parameters in the processing memory are updated and used by the processor to control the buffer level via the buffer level control routine. The significant sample (together with its address) is stored in the buffer memory.

Various constants and data (including the aperture and past sample value) used by the processor are stored in the processing memory. Each input channel has assigned to it a set of constants which permits a selective control of the buffer level. The processing memory and buffer level control routine are probably the two most important features in this subsystem.

The buffer memory is the conventional "first-in, first-out" type. Its output is time multiplexed by the master multiplexer. The buffer is needed because its input is at an irregular rate, whereas its output is a constant rate determined by the communications link bit rate.

b. Operation. All analog and digital measurements are sampled at the same rate via the multiplexer, which is under processor control. Naturally, this rate must be sufficiently high because it has to be divided among all

the channels. In the F3HS subsystem, the multiplexer is stepped once every 1/8192 sec. Since there is provision for 32 data channels, the maximum channel sampling rate is 8192/32 = 256 samples/sec. Although all measurements are sampled at this maximum rate, they are not necessarily processed this fast. It is the function of the processor to decide when and how these samples will be processed.

The processor accepts the user control signal to decide when the samples will be processed. This signal is 1 bit of digital information which is sampled simultaneously with the data and is used to either inhibit or permit the processing of the data to continue. It is a useful feature of the EDHS by which an individual channel may equivalently disconnect itself from the multiplexer during those intervals when the data is unmeaningful or undesired. Thus the bandwidth which is unused by such channels can be reallocated to other channels (other subsystems) by the master multiplexer. Hence the main purpose of user control is for efficient use of the available bandwidth.

In the following discussion, the user control will be assumed to always indicate that processing should continue, in which case the processor accepts the data output from the multiplexer and calls it the "present sample." Next, the past sample (which is also the last sample transmitted to the buffer memory for this particular channel) stored in the processing memory is retrieved and compared to the present sample. If the difference between the samples exceeds the aperture, which is also stored in the processing memory, then the present sample becomes a significant sample. This, in essence, explains the zero-order data compression routine. It merely accepts present samples and compares them to the past sample to generate significant samples.

A significant sample may or may not be stored in the buffer memory, depending on the buffer fullness factor which is fed back to the buffer level control routine in the processor. A future *SPS* article will discuss the operational details of the buffer level control routine and the results from a computer simulation of it. For the purposes of this article, it is sufficient to state that it is the buffer level control routine that decides whether or not the significant sample will be sent to the buffer memory to become a transmitted sample. In either case, the significant sample will be sent to the processing memory to become the updated past sample.

XVI. Spacecraft Radio

A. Signal-to-Noise Ratio Monitoring: Error Analysis of the Signal-to-Noise Ratio Estimator

D. W. Boyd

1. Introduction

The problem of monitoring the signal-to-noise ratio (SNR) of certain space communication channels such as carrier, telemetry, and synchronization, has been considered previously. In evaluating the practical implementation of the system, built as outlined in Vol. IV of *SPS 37-27*, *37-37*, and *37-38*, some type of error analysis must be performed. Such an analysis serves several purposes:

(1) It provides insight into the practical system and into how it compares with the ideal.

(2) It provides a means of setting performance specifications for the various system components.

(3) It tells how accurate the system measurements are.

This report describes how such an analysis was performed for the signal-to-noise ratio estimator (SNORE).

2. Model of Nonideal System

In *SPS 37-37*, it was shown that under ideal conditions the SNORE makes a consistent estimate of the actual SNR, μ^2/σ^2. The problem posed here is to describe how a practical system affects the estimate. For example, we would like answers to questions such as: what is the effect of imperfect components in the SNORE, and what is the effect of mismatched elements? We describe the ideal system, as discussed in *SPS 37-27*, by assuming that we have available n independent samples x_i, $i = 1, 2, \cdots n$ which are gaussian random variables of mean, μ, and variance, σ^2. We use the estimator

$$\hat{R}^2 = \hat{\mu}^2/\hat{\sigma}^2 \tag{1}$$

where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{3}$$

to estimate μ^2/σ^2. The practical system will be characterized by how it deviates from the ideal.

We can describe the effects of a practical system by saying the x_i are transformed in some way to give new random variables y_i with which we form estimates. Thus we describe the y_i as a function, $F[\ ]$, of the x_i and perhaps some other random variable n_i,

$$y_i = F[x_i, n_i] \tag{4}$$

The functional relation postulated in Eq. (4) is very general and may be linear or nonlinear, depending on the particular nonideal aspect of the system being modeled. As indicated, y_i may also depend on n_i, which, as might be expected, will often correspond to some kind of noise. We will illustrate some specific y_i's with several examples in Section 5.

Once we have the y_i, we form estimates just as before

$$\hat{S}^2 = \hat{m}^2/\hat{v}^2 \tag{5}$$

where

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6}$$

and

$$\hat{v}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{m})^2 \tag{7}$$

The problem is to determine how $\hat{S}^2$ is related to the actual SNR, μ^2/σ^2.

3. Measurement of Error in $\hat{S}^2$

The best way to study $\hat{S}^2$ would be to find its probability density function (pdf) just as we found the pdf for $\hat{R}^2$ in *SPS 37-37*. Then we could find $E[\hat{S}^2]$ and Var $[\hat{S}^2]$ and have an idea of how good an estimate $\hat{S}^2$ is for μ^2/σ^2. Unfortunately, it is impossible to find $\hat{p}_{S^2}(\hat{S}^2)$ in most cases and impractical in most other cases, so that we have to find some other way to measure how good an estimator $\hat{S}^2$ is.

Although we can't usually calculate $p_{\hat{S}^2}(\hat{S}^2)$, it is generally easy to calculate the moments of y_i. This suggests the following approach:

(1) Try to show that $\hat{S}^2$ is a consistent estimator for m^2/v^2 for all possible pdf's of y_i where m is the mean and v^2 is the variance of y_i.

(2) Calculate the moments of y_i and assume n is large enough that $E[\hat{S}^2] \approx m^2/v^2$.

(3) To measure the error caused by the nonideal system define

$$FM = 10 \log_{10}\left[\frac{m^2/v^2}{\mu^2/\sigma^2}\right] \tag{8}$$

so that FM gives the difference, in decibels, between the expected value of the estimate with the practical system, and the actual value of the SNR, μ^2/σ^2.

If n is large enough so that Var $[\hat{S}^2]$ is small, this should be a useful way to measure the error in the nonideal system. The obvious disadvantage with this approach is that we lose the dependence on n. With the ideal case we know that the mean and variance are given by

$$E[\hat{R}^2] = \frac{n-1}{n-3}\left(\frac{1}{n} + \frac{\mu^2}{\sigma^2}\right) \qquad n > 3$$

$$\text{Var}[\hat{R}^2] = \frac{2(n-1)^2}{(n-3)^2(n-5)}\left[\left(\frac{\mu^2}{\sigma^2}\right)^2 + 2\frac{\mu^2}{\sigma^2}\left(1 - \frac{1}{n}\right) + \frac{2}{n}\left(1 - \frac{1}{n}\right)\right] \qquad n > 5$$

so that we can gauge exactly the effect of changing n. If the approach we propose is valid, we will have expressions like

$$\lim_{n\to\infty} E[\hat{S}^2] = \frac{m^2}{v^2}$$

$$\lim_{n\to\infty} \text{Var}[\hat{S}^2] = 0$$

so that we cannot determine the effect of increasing n quite so easily. However, if we observe that: (1) in the ideal case the bias caused by the finite number of samples is quite small for values of n used in practice ($n \approx 1000$), (2) the variables we study will often be only slightly perturbed from gaussian, and (3) that a more complicated approach is impractical, we will be justified in assuming that the error caused by finite n is negligible for large n. With this justification we will establish the validity of the proposed approach.

4. Statistical Characteristics of $\hat{S}^2$

We would like to show that $\hat{S}^2$ is a consistent estimator for m^2/v^2. If it is and if n is large, we can use Eq. (8)

to measure the error caused by the nonideal system. We must show that:

$$\lim_{n\to\infty} E[\hat{S}^2] = m^2/v^2$$

$$\lim_{n\to\infty} \mathrm{Var}[\hat{S}^2] = 0$$

The only assumption that we will make about $p_{yi}(y_i)$ is that it has finite moments, $E[y_i^\nu]$, for $\nu = 1, 2, 3, 4$. This should be true of all $p_{yi}(y_i)$ realized in practice. First we note some interesting properties of the estimators we are using:

1. $E[\hat{m}] = m, \quad \mathrm{Var}[\hat{m}] = \dfrac{v^2}{n}$

2. $E[\hat{v}^2] = v^2, \quad \mathrm{Var}[\hat{v}^2] = \dfrac{v^2}{n}\left[\dfrac{\mu_4}{v^4} - \dfrac{(n-3)}{(n-1)}\right]$

where

$$\mu_4 = E[(y_i - m)^4]$$

These properties show that $\hat{m}$ and $\hat{v}^2$ are consistent estimators, that is we can use Tchebycheff's inequality to show that

$$p\lim_{n\to\infty} \hat{m} = m \quad \text{or} \quad \lim_{n\to\infty} Pr[|\hat{m} - m| \geqq \varepsilon] = 0$$

$$p\lim_{n\to\infty} \hat{v}^2 = v^2 \quad \text{or} \quad \lim_{n\to\infty} Pr[|\hat{v}^2 - v^2| \geqq \varepsilon] = 0$$

where ε is an arbitrary positive number. Expressed in words, we say: for any fixed ε the probability that $\hat{m}$ differs from m (or $\hat{v}^2$ from v^2) by more than ε becomes arbitrarily close to 0 as n tends to infinity. Further, we note that

3. $\hat{S}^2$ is continuous and has finite first and second derivatives with respect to $\hat{m}$ and $\hat{v}^2$ in the neighborhood of $\hat{m} = m$ and $\hat{v}^2 = v^2$.

This observation, coupled with the fact that $\hat{m}$ and $\hat{v}^2$ are consistent estimators, is sufficient to show (Ref. 1) that $\hat{S}^2$ converges in probability to m^2/v^2, that is,

$$p\lim_{n\to\infty} \frac{\hat{m}^2}{\hat{v}^2} = m^2/v^2 \quad \text{or} \quad \lim_{n\to\infty} Pr\left[\left|\frac{\hat{m}^2}{\hat{v}^2} - \frac{m^2}{v^2}\right| \geqq \varepsilon\right] = 0$$

In fact we can show that $\hat{S}^2$ is asymptotically gaussianly distributed with mean, m^2/v^2, and variance, constant/n. From this discussion, we conclude that $\hat{S}^2$ is indeed a consistent estimate of m^2/v^2 and that we can use Eq. (8) to measure the error.

5. Examples

There are several aspects of a practical SNORE which can cause inaccuracies in measurement. We will give examples of some of these.

a. Droop. The outputs of the integrators in the SNORE are sent to the digital instrumentation system (DIS) computer which calculates the mean and variance. Unfortunately the contents of the integrators cannot be dumped at the end of the integration period, but must be held until the DIS is free to accept the data. This delay introduces the possibility of leakage in the holding circuit before the data is dumped. We can model this effect by writing y_i as

$$y_i = e^{-n_i/\tau} x_i \tag{9}$$

The exponential factor represents the droop caused by leakage in the holding circuit. The constant τ is the time constant of the holding circuit and n_i is an independent variable representing the waiting time until the data is accepted. On the basis of the physical operation of the DIS computer, we will assume that the pdf of n_i is uniform and given by

$$p_{ni}(n_i) = \begin{cases} \dfrac{n_i}{T} & 0 < n_i < T \\ 0 & \text{elsewhere} \end{cases}$$

This behavior corresponds closely with the manner in which the SNORE interrupt commands are accepted.

We would now like to use our results to measure the error caused by the leakage or droop. As it happens, this is one of the cases in which we can calculate $p_{yi}(y_i)$. However, the result is in such a form as to be very impractical for actual use, so we will follow the procedure outlined in Section 3. Since the random variables are independent, the calculation of the various moments is straightforward and we obtain

$$FM_d = 10\log_{10}\left\{\frac{1}{\dfrac{\left(\dfrac{T}{\tau}\right)(1 - e^{-2T/\tau})}{2(1 - e^{-T/\tau})^2} + \dfrac{\mu^2}{\sigma^2}\left[\dfrac{\left(\dfrac{T}{\tau}\right)(1 - e^{-2T/\tau})}{2(1 - e^{-T/\tau})^2} - 1\right]}\right\} \tag{10}$$

where the "d" indicates the error from droop. As we might expect, the error is an increasing (more negative decibel) function of the SNR μ^2/σ^2. As long as n is large, we can use Eq. (10) to measure the error. Fig. 1 shows a plot of FM_d versus μ^2/σ^2 for various values of the ratio T/τ. For instance, if SNR $= \mu^2/\sigma^2 = 30$ db and $T/\tau = 0.1$ (10% droop in T sec), we see that the error will be -2.6 db; that is, the SNORE will estimate a SNR of 27.4 db if there are no other errors. We can use the figure for two purposes:

(1) If we have a good idea of μ^2/σ^2 and the system parameter, T/τ, we can predict the error in the SNORE's measurement.

(2) If we know that we want to measure a certain μ^2/σ^2 within a given specified error, we can specify the necessary system performance, T/τ.

b. Mismatch. Since a realizable integrator cannot be dumped in zero time and since there is a waiting time for the DIS, it is necessary to employ two integrators in the SNORE. These integrators operate alternately; while one is processing the incoming signal, the other is being dumped. For ideal operation, the gains of the two branches of the system must be identical. If they are not, the output measurement will be in error. We can model this effect on the random variable x_i as follows

$$y_i = \begin{cases} x_i & \text{if Channel 1} \\ G x_i & \text{if Channel 2} \end{cases} \tag{11}$$

where $0 < G < \infty$. There is no loss of generality in this choice, since one gain can always be picked as a reference and both gains can be normalized to this value. Physically G is a measure of the mismatch between the two integrators. For instance, $G = 1.1$ indicates that the larger gain is 10% greater than the smaller one.

With mismatch, the advantages of our technique are even more apparent, since we can find only approximate expressions for $p_{yi}(y_i)$. Proceeding as outlined before, we find

$$FM_m = 10 \log_{10} \left\{ \frac{1}{\left(\frac{1+G^2}{2}\right) \Big/ \left(\frac{1+G}{2}\right)^2 + \frac{\mu^2}{\sigma^2}\left[\left(\frac{1+G^2}{2}\right) \Big/ \left(\frac{1+G}{2}\right)^2 - 1\right]} \right\} \tag{12}$$

where the m indicates mismatch. We see that the expression has the same form as we encountered before. Plots of FM_m for various values of G are shown in Fig. 2. These can be used for the same purposes as indicated in Section 5a.

c. Line noise. Line noise picked up on the cables connecting the telemetry demodulators with the DIS is another cause of error in the SNORE measurement. Such noise could be caused by many factors, ranging from ground loops to power line and switching transients. In view of this, it is difficult to postulate and justify a statistical model for the noise. Rather than get too far from the problem, we will simply assume that the noise is independent of the signal, is additive, and has zero mean and variance σ_n^2. The assumption of zero mean is no limitation, since a non-zero mean can always be considered as an offset voltage (Section 5d). We thus write y_i as

$$y_i = x_i + n_i \tag{13}$$

where n_i is the noise described above.

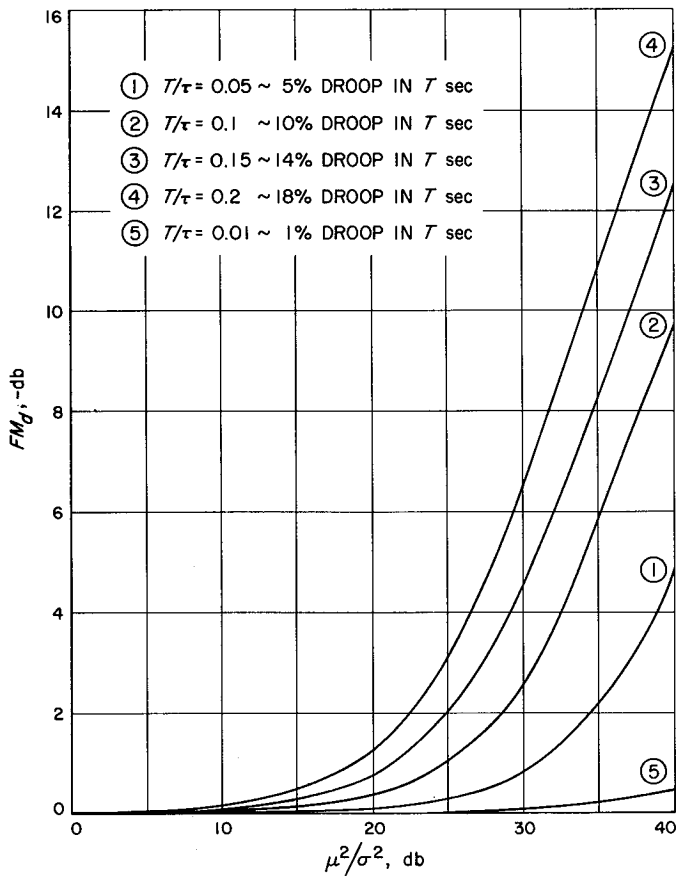


Fig. 1. FM_d versus μ^2/σ^2

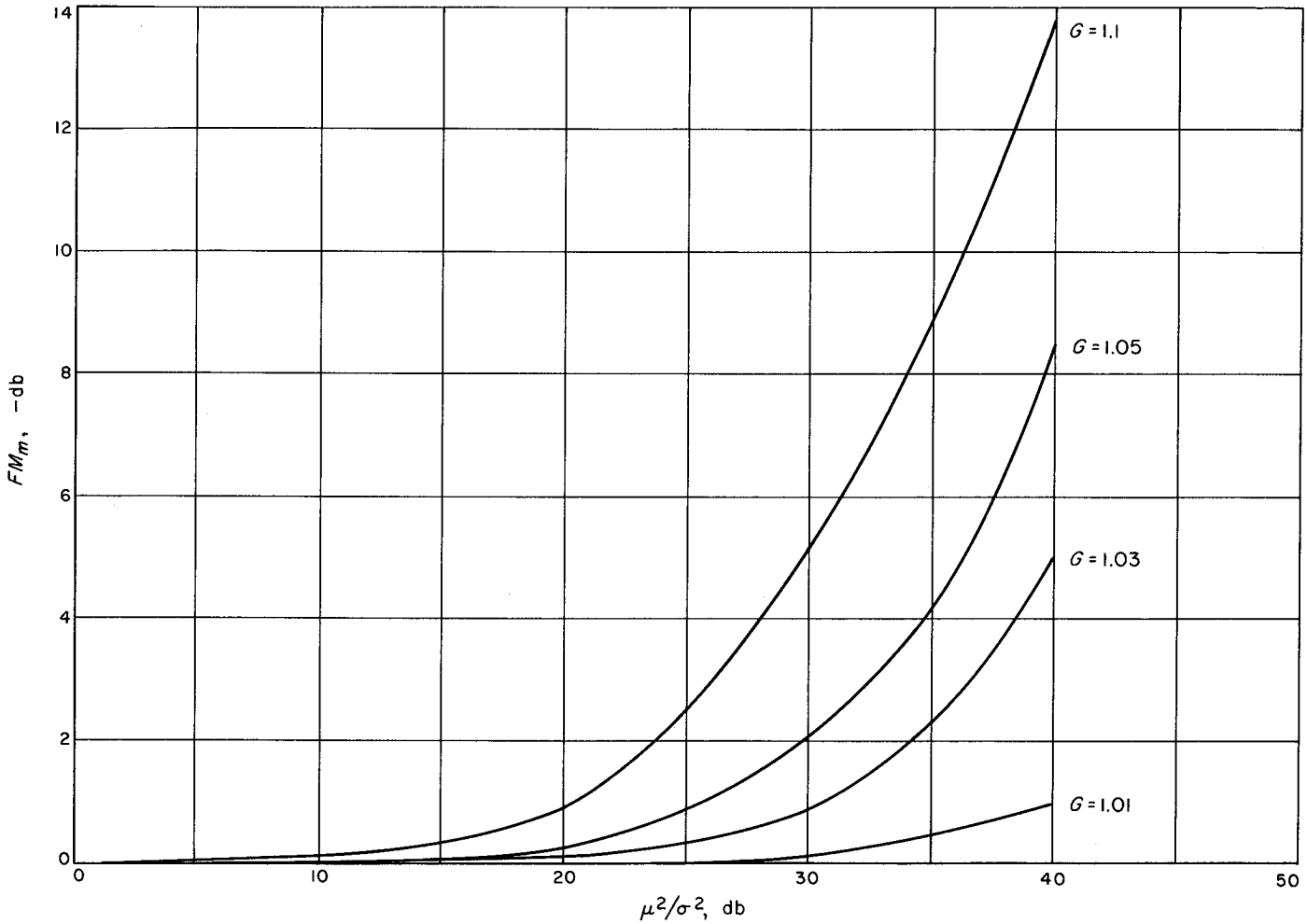


Fig. 2. FM_m versus μ^2/σ^2

In this case, the advantages of using Eq. (8) are also clear. We don't even have a pdf for n_i, but we can easily calculate the necessary moments to yield

$$FM_n = 10 \log_{10} \left(\frac{1}{1 + \frac{\sigma_n^2}{\sigma^2}} \right) \tag{14}$$

where the "n" indicates line noise. The quantity σ is determined by the physical constraints of the DIS analog-to-digital converter, the output of which is limited to the range $[0, V]$ volts. To specify σ we will make two assumptions:

(1) We will consider only SNR $(\mu^2/\sigma^2) > 10$ db. For $SNR < 10$ db σ_n will be small relative to σ and the line noise will have a negligible effect on the measurement.

(2) We will assume that the gain of the SNORE is always adjusted so that $V - \mu = \alpha\sigma$ where α is a positive number chosen to minimize the line noise error subject to the constraint $\alpha \leq 3$ (to avoid severe truncation). Although the approximately optimum value of α will be 3, a far simpler procedure practically is to let $\alpha = (SNR)^{1/2}$ which gives $\mu =$ constant $= V/2$ independent of SNR. Of course, both of these statements must be taken in an approximate sense in practice, since we don't know the SNR and are trying to estimate it. For example, the second statement would correspond to centering the output in the dynamic range of the SNORE. We will give FM's for both values of α for comparison purposes. (The optimization of the SNORE gain setting is a problem in itself and is beyond the scope of this article. The above comments indicate the important results.)

With these assumptions, we can solve for σ and write

$$FM_{no} \approx 10 \log_{10} \left[\frac{1}{1 + \frac{\mu^2}{\sigma^2} \left(\frac{\sigma_n}{V} \right)^2} \right] \qquad (14a)$$

for the optimum error, and

$$FM_{np} = 10 \log_{10} \left[\frac{1}{1 + \frac{\mu^2}{\sigma^2} \left(\frac{\sigma_n}{V/2} \right)^2} \right] \qquad (14b)$$

for the error in the practical case. Fig. 3 shows plots of FM_{np} for various values of σ_n. As can be seen from the equations, we can also use the same plots for FM_{no}. To find FM_{no} *at* $\sigma_n = X$ we simply find FM_{np} at $\sigma_n = X/2$. In addition to serving the purposes outlined in Section 5a, these figures also illustrate the trade off between a coarse and easy adjustment of gain and a more precise and complicated adjustment.

d. Offset voltages. Another cause of error in the SNORE is offset voltages appearing in the system. These undesirable biases are sometimes added to the outputs because of practical limitations in constructing the equipment. There are two distinct ways in which offset voltage problems could arise:

Case 1: an offset voltage could be added to one channel but not to the other.

Case 2: an equal offset voltage could be added to both channels.

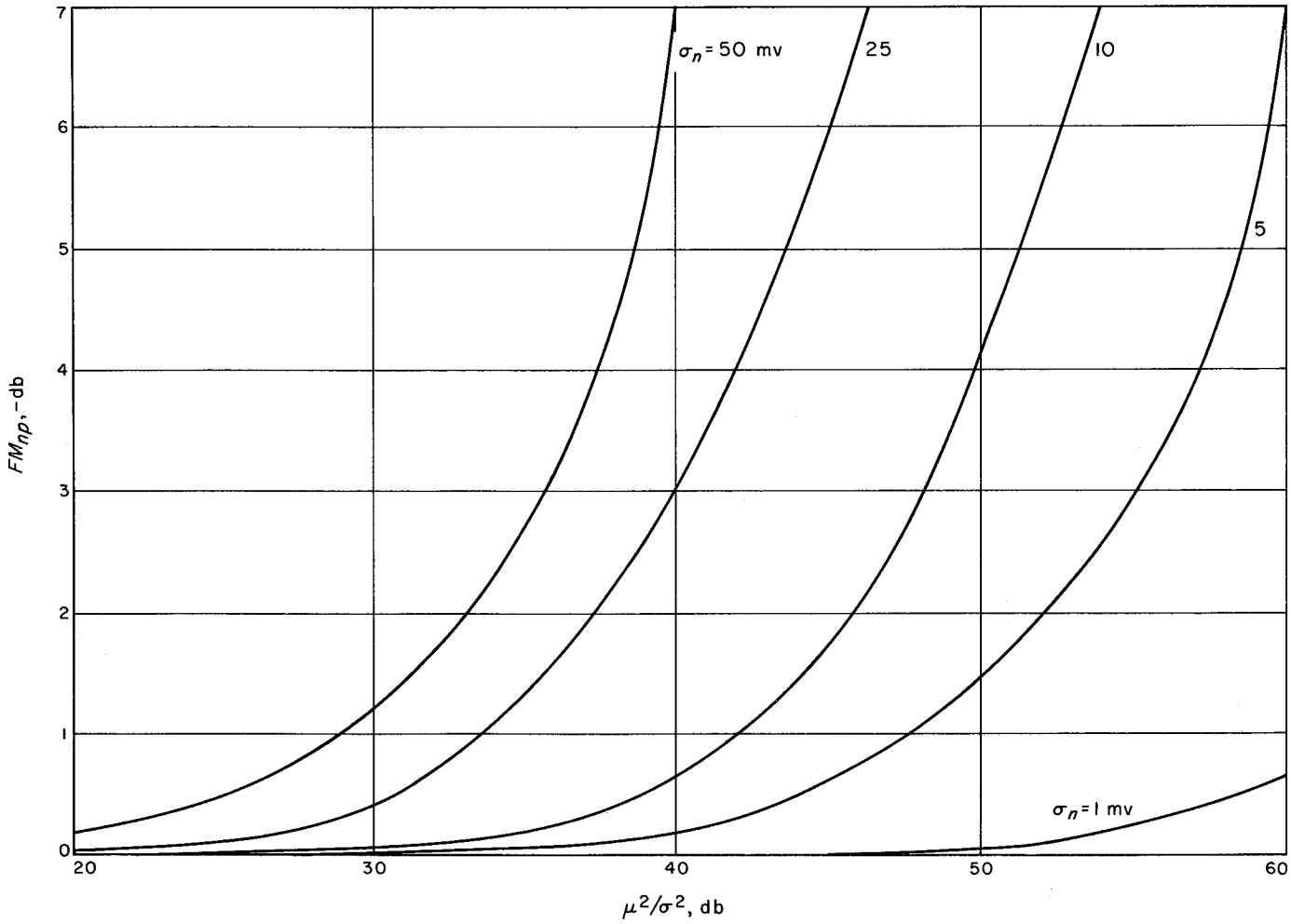


Fig. 3. FM_{np} versus μ^2/σ^2

In Case 1 we would define the y_i as

$$y_i = \begin{cases} x_i + \Delta, & \text{if Channel 1} \\ x_i, & \text{if Channel 2} \end{cases} \tag{15}$$

where Δ is a constant equal to the offset voltage. The calculation of the error is similar to that in Section 5b and yields

$$FM_o^1 = \frac{\left(1 + \frac{\Delta}{2\mu}\right)^2}{1 + \left(\frac{\Delta}{2\sigma}\right)^2} \tag{16}$$

where the 1 indicates Case 1 and the o indicates offset. Once again μ and σ will be determined by the physical constraints of the SNORE. If we proceed as outlined in Section 5c, we can find a practical FM,

$$FM_{op}^1 = \frac{(1+\delta)^2}{1 + \frac{\mu^2}{\sigma^2}\delta^2} \tag{16a}$$

and an optimum FM,

$$FM_{oo}^1 \approx \frac{\left(1 + \frac{\delta}{2}\right)^2}{1 + \frac{\mu^2}{\sigma^2}\left(\frac{\delta}{2}\right)^2} \tag{16b}$$

where $\delta = \Delta/V$. We also note that the curves for Eq. (16a) in Fig. 4 can be used to obtain values for FM_{oo}^1 as explained in Section 5c.

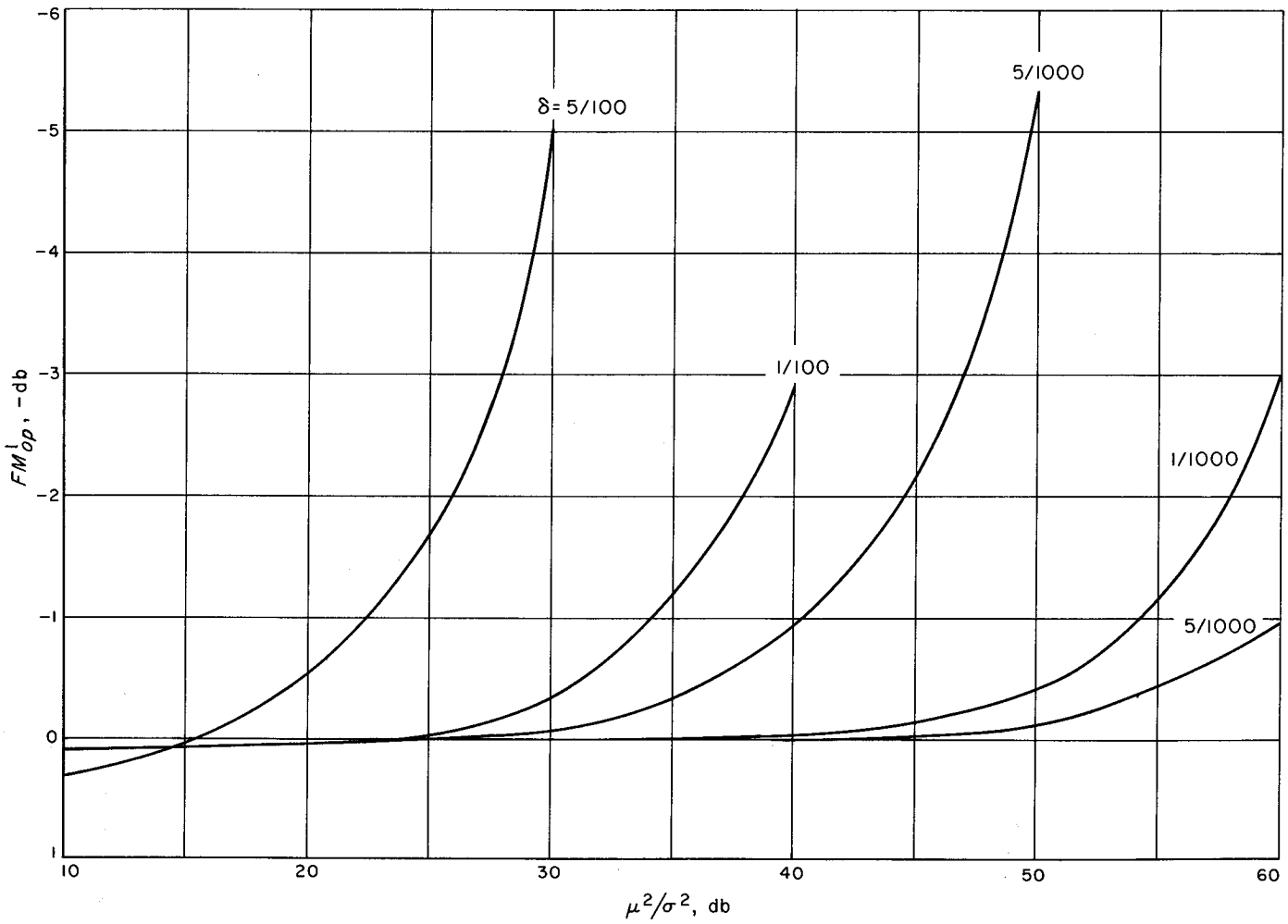


Fig. 4. FM_{op}^1 versus μ^2/σ^2

In Case 2, the results are almost trivial, since we have the y_i given by

$$y_i = x_i + \Delta \tag{17}$$

so that they are just gaussian with a different mean. The calculation yields a practical FM,

$$FM_{0p}^2 = (1 + 2\delta)^2 \tag{18a}$$

and an optimum FM,

$$FM_{0o}^2 = (1 + \delta)^2 \tag{18b}$$

e. Other errors. The same technique can be used to calculate errors from almost any source. We have discussed the most important ones, but others might include quantization, and invalid assumption of a gaussian density for the x_i.

Finally, we should emphasize that in each case the errors have been calculated as if they were the only ones existing. It is not true, in general, that errors from different sources are additive.

B. System Reliability Figure Versus Degree of Analytical Detail

M. K. Tam[1]

Prior to the launch of *Mariner IV*, analyses were made predicting the probability of success of critical functions of its telecommunication system. These numbers were typically low and not too encouraging. The subsequent success of the flight of *Mariner IV*, particularly from the reliability point of view, has prompted a review of the reliability prediction method utilized to see if there is any significant pessimistic bias in it. Continual effort has been expended in updating the method, primarily in the areas of increasing modeling accuracy and of obtaining up-to-date failure rate data. This report concerns itself with the former. Specifically, the investigation has been to determine how the system reliability figure is related to the analytical detail (or modeling accuracy) utilized. Analytical detail may be equated to cost (in manpower and delay) and should be avoided if other methods are satisfactory or can be projected to be useful.

[1]Contract affiliation.

The basic idea behind this investigation can be described briefly by the symbolic diagrams in Fig. 5.

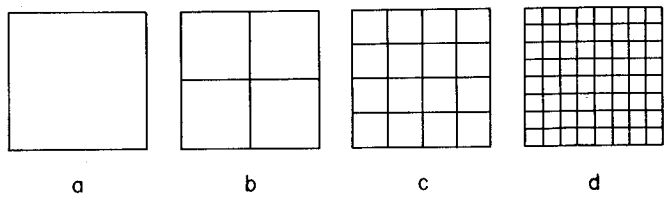


Fig. 5. Progressive diagrams for various degrees of analytical detail

Here, Fig. 5(a) represents a single body which encompasses every piece-part within the system. The reliability figure for this configuration is obtained strictly by the total parts count. Fig. 5(b) delineates the fact that the entire system is divided into a number of major functional blocks. While the reliability figure for each individual block is obtained on the parts count basis, the over-all system reliability figure is evaluated with functional dependency and redundancy of these blocks being fully considered. Fig. 5(c) is an extension of the preceding configuration which carries these same considerations to the subfunctional block level. The system reliability figure thus obtained reflects all the partial successes up to this level. Finally, Fig. 5(d) contemplates further subdivisions into the circuit and the piece-part levels so that failure modes for each component can be accounted for prior to the evaluation of the system reliability figure. It is conceivable that the more detail applied to the analysis, the higher the resultant system reliability figure will be, with the ultimate reliability figure approaching the true system probability of success when all failure mode and partial success considerations of every minute piece-part are entered into the over-all reliability figure calculation.

For systems of moderate complexity, there is a practical limit as to how far the detailed consideration can be carried. In general, this limit is on the subfunctional block level. Excessive time and effort will be required if analysis is attempted beyond this level.

In the current investigation, three different systems are selected as illustrative examples. These systems are the *Mariner C* doppler tracking equipment, data encoding equipment, and a 400-cycle static inverter. The first two systems are relatively complex and each of them consists of several thousand piece-parts. The third system has approximately 400 components and is a simple nonredundant system. Analysis has been performed for each of the systems according to the investigation outlines

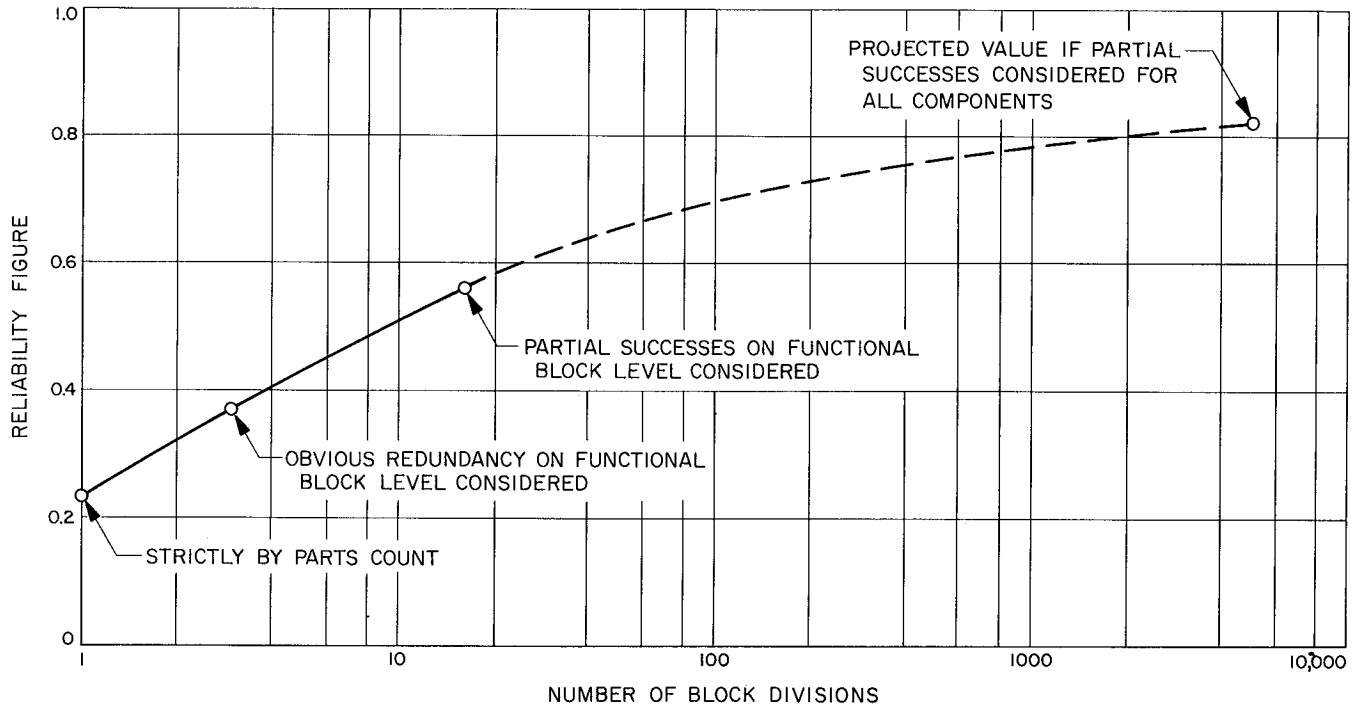


Fig. 6. *Mariner* C doppler tracking equipment reliability figure versus number of block divisions

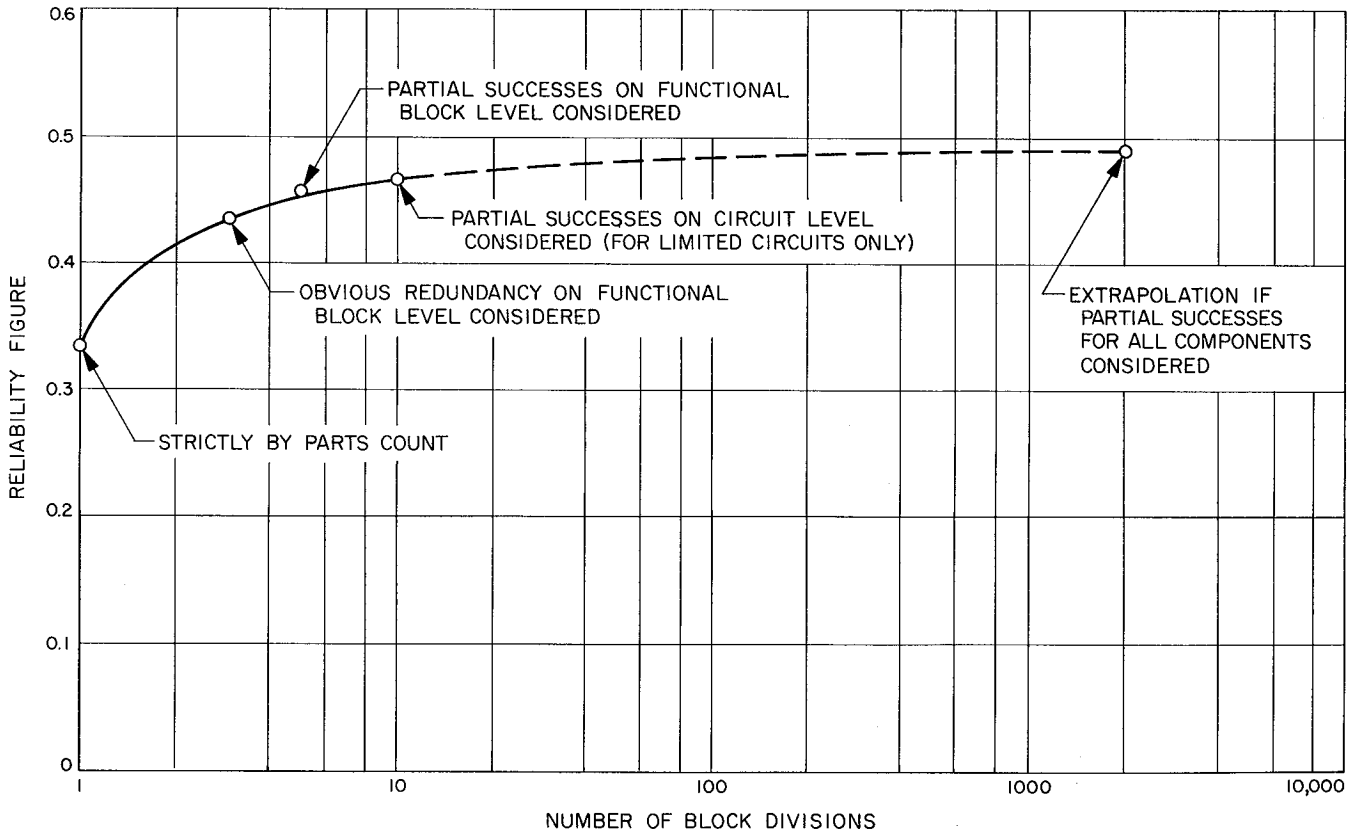


Fig. 7. *Mariner* C data encoding equipment reliability figure versus number of block divisions

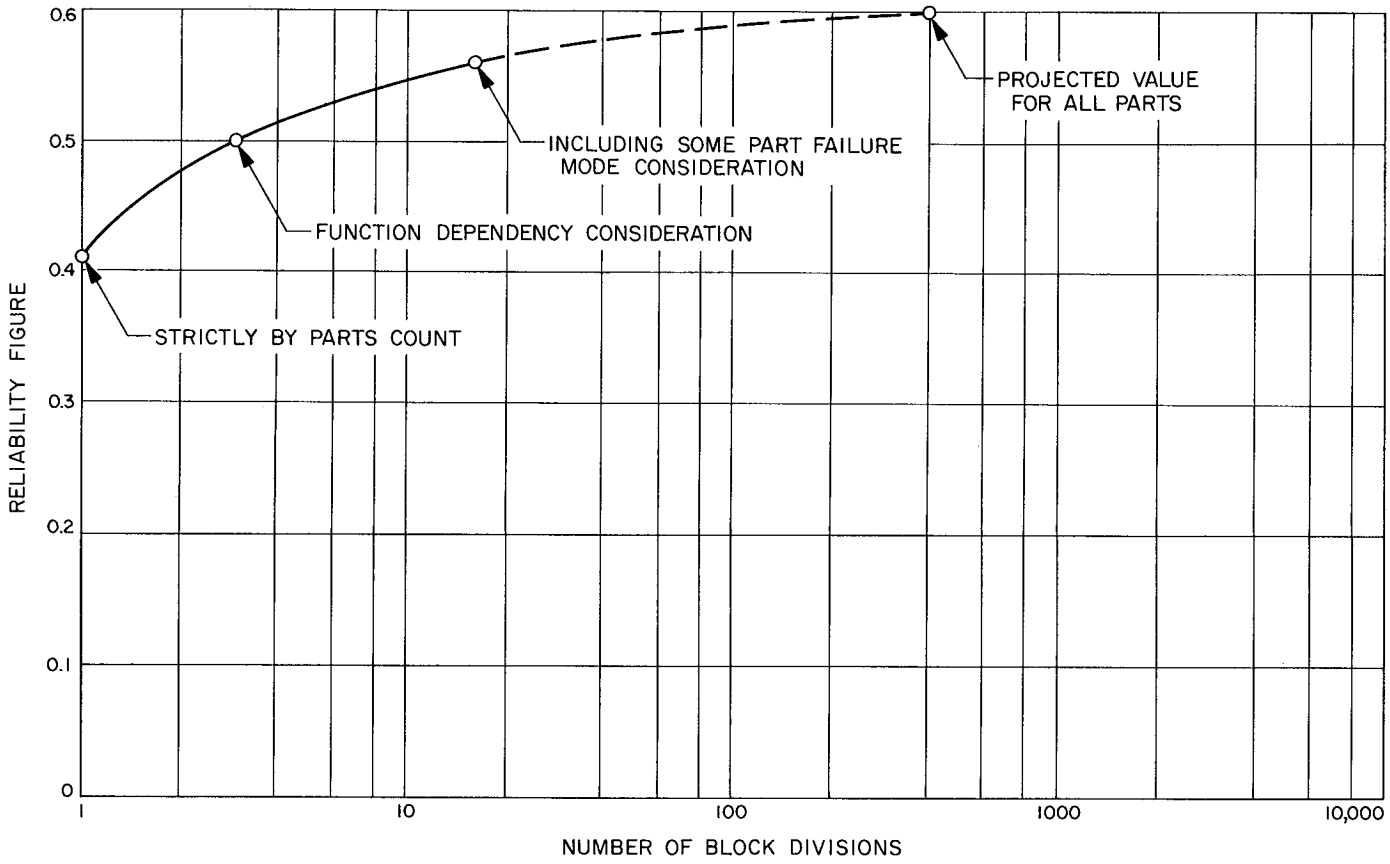


Fig. 8. Reliability figure of 400-cycle static inverter versus number of block divisions

previously described. Results of the analyses are plotted with the reliability figure versus the number of block divisions and are shown in Figs. 6, 7, and 8, respectively. These plots are further extrapolated in an attempt to show the potential system probability of success, should the analysis be carried out to the component level.

From results obtained in the foregoing analyses, the following conclusions have been made.

(1) Reliability figure of a complex system tends to approach its true probability of success as partial success and failure mode considerations are applied on the component level in the reliability analysis.

(2) The reliability figure versus number of block divisions curve displays a general characteristic, as shown in Fig. 9, provided that the curve is obtained in accordance with the following sequence.

a. Reliability figure is obtained strictly by parts count of the entire system (an uneducated and unacceptable way of system reliability estimate).

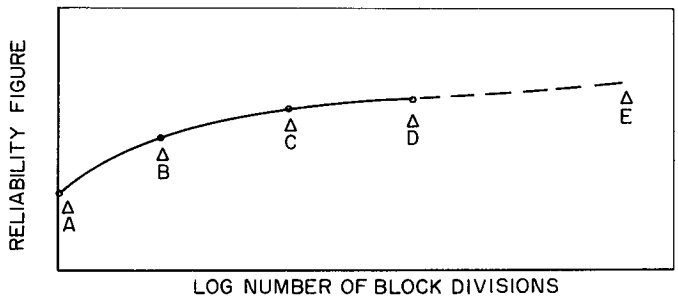


Fig. 9. General characteristics for a reliability figure versus number of block divisions

b. All obvious redundancy and function dependency on the functional block level are considered. Reliability figures for these blocks are obtained by the conventional parts count method (the commonly acceptable approach for system reliability estimate).

c. In addition to Step b, partial successes on the functional block level are included in the over-all system reliability analysis.

d. In addition to Step c, partial successes on the subfunctional block level are included in the over-all system reliability analysis.

e. Reliability figure for n block divisions is extrapolated to account for the partial successes on the component level.

(3) The specific characteristics of this curve depend on the construction of the system. Specifically, the difference between the two extreme figures is largely dependent upon the amount of redundancy incorporated and the number of nonessential or spurious parts within the system.

(4) The accuracy of the projected reliability figure as obtained in Step e is questionable. It is strongly felt that when failure mode and partial success considerations are applied to the circuit and component levels, there will be a sizable upward step apparent in the curve.

C. Analysis of Frequency-Multiplexed, PM Communication Systems

M. A. Koerner

1. Introduction

With the exception of certain FM communication systems used in the *Ranger* Block III, *Lunar Orbiter*, *Surveyor*, and *Apollo* telecommunication systems, the *Ranger* Block I, II, and III, *Mariner R* and *C*, *Pioneer*, *Lunar Orbiter*, *Surveyor*, and *Apollo* telecommunication systems are frequency-multiplexed, PM communication systems. This report presents the method used to evaluate the performance of these telecommunication systems.

A simplified block diagram of a frequency-multiplexed, PM communication system is shown in Fig. 10. Primary elements of the system are a transmitter and a receiver. The transmitter phase modulates an RF carrier with a number of sinusoidal and binary-valued signals. The receiver input is the sum of the transmitted signal and white gaussian noise. For the purpose of this analysis, the characteristics of the RF telecommunication system may be summarized by specifying the received signal power and the power spectral density of the gaussian noise at the receiver input. The receiver may be represented as a band-pass filter, which represents the filtering effect of the receiver IF stages, a partially coherent product demodulator, and a low-pass filter. A phase-locked loop tracks the carrier component of the received signal and generates the reference signal for the product demodulator. Subcarrier extractors are used to obtain the desired information from the signal at the receiver output.

This analysis provides a means of simplifying the problem of evaluating subcarrier channels in frequency-multiplexed, PM communication systems to that of evaluating the performance of the subcarrier extractors in obtaining the desired information from a sinusoidal or binary-valued subcarrier signal, distortion, and white gaussian noise. The analysis also simplifies the problem of evaluating the carrier channel of a frequency-multiplexed, PM communication system to that of evaluating the performance of a phase-locked loop receiver in generating a coherent reference for demodulating the subcarrier channels from the carrier component of the received signal, distortion, and white gaussian noise. In many cases the effect of the distortion can be neglected, simplifying the problem to that of extracting information from a carrier or subcarrier signal observed in white gaussian noise. The analysis (1) derives equations for the amplitude and power in both the carrier and distortion components of the receiver input, (2) derives equations for the amplitude and power in the subcarrier signals and the distortion components of the receiver output, and (3) demonstrates that, over the frequency band of interest, the power spectral densities of the gaussian noise at the receiver input and output are the same.

2. Analytical Model

a. Transmitter. We shall assume that each transmitter is an ideal phase modulator whose input is the sum of N_1 sinusoidal subcarriers, each of which may be phase- or frequency-modulated by information-bearing signals, and M_1 binary-valued (± 1) subcarriers of mean zero. Under these conditions the transmitter output, normalized by the RMS transmitter output level, is

$$e_{11}(t) = (2)^{1/2} \sin [\omega_1 t + \phi_1 + \theta_1(t)], \tag{1}$$

where

$$\theta_1(t) = \sum_{k=1}^{N_1} (2)^{1/2} \alpha_{1k} \cos [\omega_{1k} t + \phi_{1k} + \theta_{1k}(t)] + \sum_{k=1}^{M_1} \beta_{1k} x_{1k}(t), \tag{2}$$

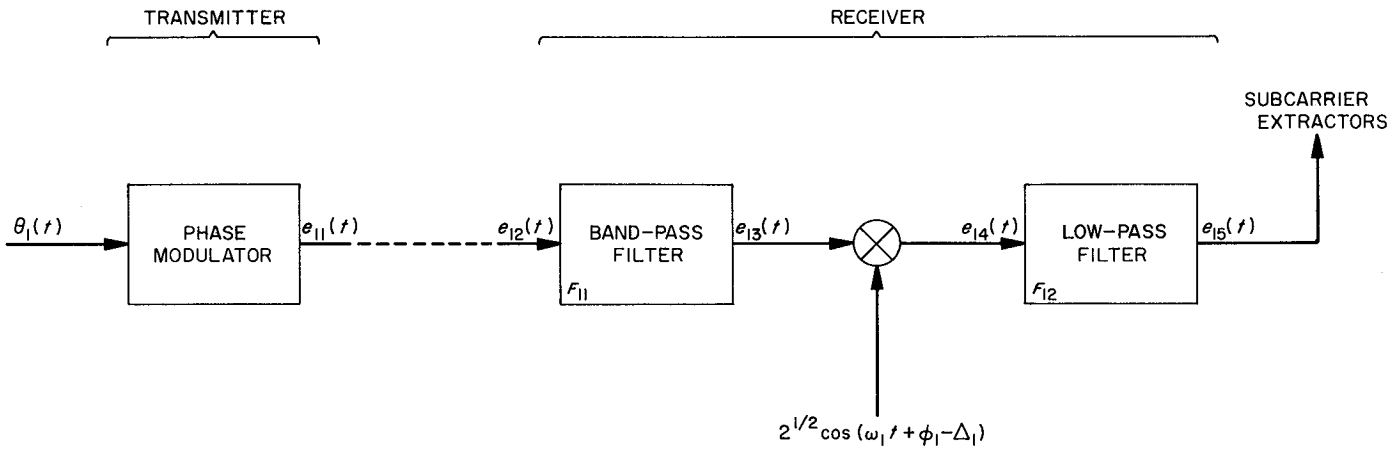


Fig. 10. Simplified functional block diagram

and ω_1 and ϕ_1 are the angular frequency and phase of the unmodulated carrier.

α_{1k} is the RMS phase deviation of the carrier produced by a sinusoidal subcarrier of angular frequency ω_{1k} and phase ϕ_{1k}. $\theta_{1k}(t)$ is the phase modulation of this subcarrier resulting from information-bearing signals phase-or frequency-modulating the subcarrier. β_{1k} is the RMS phase deviation of the carrier produced by $x_{1k}(t)$ where

$$x_{1k}(t) = \pm 1 \tag{3}$$

and

$$< x_{1k}(t) > = 0 \cdot \tag{4}$$

b. Received signal. Referring the noise introduced by the receiving system to the receiver input, the signal at the receiver input will be

$$\begin{aligned} e_{12}(t) &= s_{12}(t) + n_{12}(t) \\ &= (2P_1)^{1/2}\sin[\omega_1 t + \phi_1 + \theta_1(t)] + n_{12}(t) \end{aligned} \tag{5}$$

where P_1 is the power in the signal received from the transmitter and $n_{12}(t)$ is a white gaussian process with one-sided power spectral density Φ_1. We shall assume that provision for a time-invariant phase error in the receiver will account for the effect of a time-varying range between the transmitter and receiver of this system.

c. Receiver. For the purpose of this analysis, we shall base the mathematical model for the receiver on the simplified block diagram shown in Fig. 10. We shall assume that the time-varying component of the phase of the coherent reference that is generated by the receiver can be neglected so that the carrier component of $s_{12}(t)$ is reproduced with a phase shift $\pi/2 - \Delta_1$ to serve as a coherent reference for demodulating the received signal.

Assuming the distortion of $s_{12}(t)$ by the band-pass filter F_{11} is negligible, the band-pass filter output will be

$$e_{13}(t) = (2P_1)^{1/2}\sin[\omega_1 t_1 + \phi_1 + \theta_1(t)] + n_{13}(t) \tag{6}$$

where $n_{13}(t)$ is a narrow-band gaussian process whose spectral characteristics will be determined by the transfer function of the band-pass filter F_{11}. In Section 4 we shall find that, under the restrictions of Eqs. (61), (64), and (65), the instantaneous phase of the carrier component of the receiver input is $\omega_1 t_1 + \phi_1$. Hence, the output of the product demodulator is

$$\begin{aligned} e_{14}(t) &= \{(2P_1)^{1/2}\sin[\omega_1 t + \phi_1 + \theta_1(t)] + n_{13}(t)\} \\ &\quad \times (2)^{1/2}\cos[\omega_1 t + \phi_1 - \Delta_1] \\ &= P_1^{1/2}\sin[\theta_1(t) + \Delta_1] \\ &\quad + P_1^{1/2}\sin[2(\omega_1 t + \phi_1) + \theta_1(t) - \Delta_1] \\ &\quad + n_{14}(t) \end{aligned} \tag{7}$$

where

$$n_{14}(t) = n_{13}(t) \cdot (2)^{1/2}\cos[\omega_1 t + \phi_1 - \Delta_1] \cdot \tag{8}$$

The low-pass filter F_{12} will eliminate the components of $e_{14}(t)$ at twice the carrier frequency. Assuming the low-frequency components of $e_{14}(t)$ are passed without distortion, the receiver output is

$$e_{15}(t) = P_1^{1/2}\sin[\theta_1(t) + \Delta_1] + n_{15}(t) \tag{9}$$

where $n_{15}(t)$ represents the low-frequency component of $n_{14}(t)$.

3. Power Spectral Density of the Gaussian Noise at the Receiver Output

As stated in the introduction, one objective of this analysis is to demonstrate that, over the frequency band of interest, the power spectral densities of the gaussian noise at the receiver input and output are the same.

In the preceding section we have assumed that the one-sided power spectral density of $n_{12}(t)$, the gaussian noise at the receiver input, is Φ_1. Thus

$$G_{n_{12}}(f) = \frac{\Phi_1}{2}, \qquad -\infty < f < \infty. \tag{10}$$

Assuming the frequency response of the band-pass filter F_{11} is bounded and symmetrical about the carrier frequency and the filter has noise bandwidth $B_{F_{11}}$, $g_1(f)$ exists with the properties

$$\int_{-\infty}^{\infty} g_1(f)\,df = 1, \tag{11}$$

$$g_1(-f) = g_1(f) \tag{12}$$

and

$$0 \leqq g_1(f) \leqq \Lambda_1, \qquad -\infty < f < \infty \tag{13}$$

for some nonnegative Λ_1, such that

$$G_{n_{13}}(f) = \frac{\Phi_1}{2} B_{F_{11}} [g_1(f + f_1) + g_1(f - f_1)] \tag{14}$$

Then, if

$$\rho_1(\tau) = \int_{-\infty}^{\infty} g_1(f) \exp(i\omega\tau)\,df, \tag{15}$$

so that

$$\rho_1(0) = 1, \tag{16}$$

the autocorrelation function of $n_{13}(t)$ is

$$R_{n_{13}}(\tau) = \Phi_1 B_{F_{11}} \rho_1(\tau) \cos(\omega_1\tau). \tag{17}$$

From Eq. (8)

$$n_{14}(t) = n_{13}(t)\,(2)^{1/2} \cos(\omega_1 t_1 + \phi_1 - \Delta_1). \tag{18}$$

Assuming that

$$\begin{aligned} p(\phi_1) &= (2\pi)^{-1}, \qquad |\phi_1| \leqq \pi, \\ &= 0, \qquad\quad |\phi_1| > \pi, \end{aligned} \tag{19}$$

and that $n_{13}(t)$ and ϕ_1 are statistically independent, the autocorrelation function of $n_{14}(t)$ is

$$\begin{aligned} R_{n_{14}}(\tau) &= E[n_{14}(t)\, n_{14}(t+\tau)] \\ &= E[n_{13}(t)\, n_{13}(t+\tau)] \cdot E[2\cos(\omega_1 t_1 + \phi_1 - \Delta_1) \\ &\quad \times \cos(\omega_1 t + \omega_1\tau + \phi_1 - \Delta_1)] \\ &= R_{n_{13}}(\tau)\,\{E[\cos(\omega_1\tau)] \\ &\quad + E[\cos(2\omega_1 t + \omega_1\tau + 2\phi_1 - 2\Delta_1)]\} \\ &= R_{n_{13}}(\tau)\cos(\omega_1\tau). \end{aligned} \tag{20}$$

Using Eq. (17)

$$\begin{aligned} R_{n_{14}}(\tau) &= \Phi_1 B_{F_{11}} \rho_1(\tau) \cos^2(\omega_1\tau) \\ &= \frac{1}{2}\Phi_1 B_{F_{11}} \rho_1(\tau)[1 + \cos(2\omega_1\tau)]. \end{aligned} \tag{21}$$

Since the low-pass filter F_{12} will remove the components of $n_{14}(t)$ with power spectral density near the second harmonic of the carrier frequency, $n_{15}(t)$ is a gaussian random variable with autocorrelation function

$$R_{n_{15}}(\tau) = \frac{1}{2}\Phi_1 B_{F_{11}} \rho_1(\tau) \tag{22}$$

and power spectral density

$$G_{n_{15}}(f) = \frac{1}{2}\Phi_1 B_{F_{11}} g_1(f). \tag{23}$$

In obtaining Eq. (6), we assumed that the gain of the band-pass filter was unity over at least the frequency band occupied by $s_{12}(t)$. If $B_{s_{12}}$ is, for practical purposes, the bandwidth of $s_{12}(t)$,

$$G_{n_{13}}(f) = \frac{\Phi_1}{2}, \qquad \begin{aligned} &-f_1 - \frac{1}{2}B_{s_{12}} < f < -f_1 + \frac{1}{2}B_{s_{12}} \\ &f_1 - \frac{1}{2}B_{s_{12}} < f < f_1 + \frac{1}{2}B_{s_{12}}. \end{aligned} \tag{24}$$

Using Eq. (14)

$$\frac{1}{2}\Phi_1 B_{F_{11}} g_1(f + f_1) = \frac{\Phi_1}{2}, \qquad -f_1 - \frac{1}{2}B_{s_{12}} < f < -f_1 + \frac{1}{2}B_{s_{12}}, \tag{25}$$

and

$$\frac{1}{2}\Phi_1 B_{F_{11}} g_1(f-f_1)=\frac{\Phi_1}{2},$$

$$f_1-\frac{1}{2}B_{s_{12}}<f<f_1+\frac{1}{2}B_{s_{12}}, \tag{26}$$

or

$$g_1(f)=B_{F_{11}}^{-1}, \qquad -\frac{1}{2}B_{s_{12}}<f<\frac{1}{2}B_{s_{12}}. \tag{27}$$

Thus, using Eqs. (23) and (27),

$$G_{n_{15}}(f)=\frac{\Phi_1}{2}, \qquad -\frac{1}{2}B_{s_{12}}<f<\frac{1}{2}B_{s_{12}}. \tag{28}$$

Thus, on the frequency interval $(0, \frac{1}{2}B_{s_{12}})$, the one-sided power spectral density of the gaussian noise at the receiver output is Φ_1.

4. Signal Component of the Receiver Input

This section derives equations for the fraction of the power in both the usable and the distortion components of $s_{12}(t)$, the signal component of the receiver input. From Eqs. (5) and (2), the signal received from the transmitter is

$$s_{12}(t)=(2P_1)^{1/2}\sin[\omega_1 t+\phi_1+\theta_1(t)] \tag{29}$$

where

$$\theta_1(t)=\sum_{k=1}^{N_1}(2)^{1/2}\alpha_{1k}\cos[\omega_{1k}t+\phi_{1k}+\theta_{1k}(t)]+\sum_{k=1}^{M_1}\beta_{1k}x_{1k}(t). \tag{30}$$

Using $Re(z)$ and $Im(z)$ to denote the real and imaginary parts of a complex variable z,

$$s_{12}(t)=(2P_1)^{1/2}\,Im\{\exp[i(\omega_1 t+\phi_1)]\exp[i\theta_1(t)]\} \tag{31}$$

where

$$\exp[i\theta_1(t)]=\prod_{k=1}^{N_1}\exp\{i(2)^{1/2}\alpha_{1k}\cos[\omega_{1k}t+\phi_{1k}+\theta_{1k}(t)]\}\times\prod_{k=1}^{M_1}\exp[i\beta_{1k}x_{1k}(t)] \tag{32}$$

Since

$$\exp(iz\cos\theta)=\sum_{l=-\infty}^{\infty}i^l J_l(z)\cos(l\theta)=\sum_{l=-\infty}^{\infty}i^l J_l(z)\exp il\theta \tag{33}$$

and

$$i^{-l}J_{-l}(z)=i^l J_l(z), \tag{34}$$

$$\exp\{i(2)^{1/2}\alpha_{1k}\cos[\omega_{1k}t+\phi_{1k}+\theta_{1k}(t)]\}=\sum_{n_{1k}=-\infty}^{\infty}i^{|n_{1k}|}J_{|n_{1k}|}[(2)^{1/2}\alpha_{1k}]\times\exp\{in_{1k}[\omega_{1k}t+\phi_{1k}+\theta_{1k}(t)]\} \tag{35}$$

Moreover

$$\exp[i\beta_{1k}x_{1k}(t)]=\cos[\beta_{1k}x_{1k}(t)]+i\sin[\beta_{1k}x_{1k}(t)] \tag{36}$$

or, using Eq. (3)

$$\exp[i\beta_{1k}x_{1k}(t)]=\cos(\beta_{1k})+ix_{1k}(t)\sin(\beta_{1k})=\sum_{m_{1k}=0}^{1}i^{m_{1k}}\cos\left(\beta_{1k}-m_{1k}\frac{\pi}{2}\right)[x_{1k}(t)]^{m_{1k}} \tag{37}$$

Using Eqs. (35) and (37) in Eq. (32)

$$\exp[i\theta_1(t)]=\sum_{n_{11}=-\infty}^{\infty}\cdots\sum_{n_{1N_1}=-\infty}^{\infty}\sum_{m_{11}=0}^{1}\cdots\sum_{m_{1M_1}=0}^{1}i^{\sum_{k=1}^{N_1}|n_{1k}|+\sum_{k=1}^{M_1}m_{1k}}\times\prod_{k=1}^{N_1}J_{|n_{1k}|}[(2)^{1/2}\alpha_{1k}]\cdot\prod_{k=1}^{M_1}\cos\left(\beta_{1k}-m_{1k}\frac{\pi}{2}\right)\times\prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}}\exp\left\{i\sum_{k=1}^{N_1}n_{1k}[\omega_{1k}t+\phi_{1k}+\theta_{1k}(t)]\right\} \tag{38}$$

At this point we shall introduce vector notation to provide an abbreviated notation for the analysis that follows. Defining vectors $\mathbf{n}_1$ and $\mathbf{m}_1$ such that

$$\mathbf{n}_1=(n_{11},\ \cdots,\ n_{1N_1})' \tag{39}$$

where $\mathbf{n}'$ is the transpose of $\mathbf{n}$, and

$$\mathbf{m}_1=(m_{11},\ \cdots,\ m_{1M_1})', \tag{40}$$

where n_{1k} may be any positive or negative integer while m_{1k} must be zero or one, the symbol, $\sum_{\mathbf{n}_1, \mathbf{m}_1}$, meaning a summation over all possible values of the vectors $\mathbf{n}_1$ and $\mathbf{m}_1$, may be used to replace

$$\sum_{n_{11}=-\infty}^{\infty} \cdots \sum_{n_{1N_1}=-\infty}^{\infty} \sum_{m_{11}=0}^{1} \cdots \sum_{m_{1M_1}=0}^{1}$$

Moreover, if

$$\boldsymbol{\omega}_1 = (\omega_{11}, \cdots, \omega_{N_1})', \tag{41}$$

$$\boldsymbol{\phi}_1 = (\phi_{11}, \cdots, \phi_{1N_1})', \tag{42}$$

and

$$\boldsymbol{\theta}_1(t) = [\theta_{11}(t), \cdots, \theta_{1N_1}(t)]', \tag{43}$$

$$\sum_{k=1}^{N_1} n_{1k}[\omega_{1k}t_1 + \phi_{1k} + \theta_{1k}(t)] = \mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)]. \tag{44}$$

We shall also find it convenient to introduce the vectors

$$\boldsymbol{\alpha}_1 = (\alpha_{11}, \cdots, \alpha_{1N_1})' \tag{45}$$

and

$$\boldsymbol{\beta}_1 = (\beta_{11}, \cdots, \beta_{1N_1})' \tag{46}$$

and define the function

$$a_1(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) = \prod_{k=1}^{N_1} J_{|n_{1k}|}[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos\left(\beta_{1k} - m_{1k}\frac{\pi}{2}\right). \tag{47}$$

Then

$$\exp[i\theta_1(t)] = \sum_{\mathbf{n}_1, \mathbf{m}_1} i^{\sum_{k=1}^{N_1}|n_{1k}| + \sum_{k=1}^{M_1} m_{1k}} a_1(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) \times \prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}} \exp\{i\mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)]\}. \tag{48}$$

Since

$$i = \exp\left(i\frac{\pi}{2}\right), \tag{49}$$

if we define a L-dimensional vector

$$\mathbf{u}_L = (1,1, \cdots, 1)' \tag{50}$$

and the vector-valued function

$$\mathbf{a}(\mathbf{n}_1) = (|n_{11}|, \cdots, |n_{1N_1}|)', \tag{51}$$

$$i^{\sum_{k=1}^{N_1}|n_{1k}| + \sum_{k=1}^{M_1} m_{1k}} = \exp\left\{i[\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\mathbf{u}_{M_1}]\frac{\pi}{2}\right\} \tag{52}$$

and

$$\exp[i\theta_1(t)] = \sum_{\mathbf{n}_1, \mathbf{m}_1} a_1(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) \prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}} \exp\left[i\left\{\mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)] + [\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\mathbf{u}_{M_1}]\frac{\pi}{2}\right\}\right]. \tag{53}$$

Using Eq. (53) in Eq. (31)

$$s_{12}(t) = (2P_1)^{1/2} \sum_{\mathbf{n}_1, \mathbf{m}_1} a_1(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) \prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}} \times Im\left[\exp\left[i\left\{\omega_1 t_1 + \phi_1 + \mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)] + [\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\mathbf{u}_{M_1}]\frac{\pi}{2}\right\}\right]\right]$$

$$= (2P_1)^{1/2} \sum_{\mathbf{n}_1, \mathbf{m}_1} a_1(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) \prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}} \times \sin\left\{\omega_1 t + \phi_1 + \mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)] + [\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\mathbf{u}_{M_1}]\frac{\pi}{2}\right\}. \tag{54}$$

Thus

$$s_{12}(t) = P_1^{1/2} \sum_{\mathbf{n}_1, \mathbf{m}_1} A_{12}(\mathbf{n}_1; \mathbf{m}_1) \prod_{k=1}^{M_1} [x_{1k}(t)]^{m_{1k}}$$

$$\times \sin\left\{\omega_1 t + \phi_1 + \mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)] + [\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\mathbf{u}_{M_1}]\frac{\pi}{2}\right\} \tag{55}$$

where

$$A_{12}(\mathbf{n}_1, \mathbf{m}_1) = (2)^{1/2} a_1(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) = (2)^{1/2} \prod_{k=1}^{N_1} J_{|n_{1k}|}[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos\left(\beta_{1k} - m_{1k}\frac{\pi}{2}\right). \tag{56}$$

In computing signal-to-noise ratios, the fraction, $\eta_{12}(\mathbf{n}_1; \mathbf{m}_1)$, of the received signal power P_1 in the $(\mathbf{n}_1; \mathbf{m}_1)$ component of the receiver input is more useful than the amplitude of this component. Clearly

$$\eta_{12}(\mathbf{n}_1; \mathbf{m}_1) = \frac{1}{2} A_{12}^2(\mathbf{n}_1; \mathbf{m}_1)$$

$$= \prod_{k=1}^{N_1} J_{n_{1k}}^2[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos^2\left(\beta_{1k} - m_{1k}\frac{\pi}{2}\right). \tag{57}$$

In decibels,

$$\eta_{12}^*(\mathbf{n}_1; \mathbf{m}_1) = 10\log_{10}\eta_{12}(\mathbf{n}_1; \mathbf{m}_1)$$

$$= \sum_{k=1}^{N_1} 10\log_{10} J_{n_{1k}}^2[(2)^{1/2}\alpha_{1k}]$$

$$+ \sum_{k=1}^{M_1} 10\log_{10}\cos^2\left(\beta_{1k} - m_{1k}\frac{\pi}{2}\right). \tag{58}$$

a. Carrier component of the receiver input.

Since

$$\cos\left(a \mp \frac{\pi}{2}\right) = \cos(a)\cos\left(\pm\frac{\pi}{2}\right) + \sin(a)\sin\left(\pm\frac{\pi}{2}\right)$$

$$= \pm\sin(a), \tag{59}$$

we may modify Eq. (55) to obtain

$$s_{12}(t) =$$

$$P_1^{1/2} \sum_{\mathbf{n}_1, \mathbf{m}_1} A_{12}(\mathbf{n}_1; \mathbf{m}_1)\cos\left\{\omega_1 t + \phi_1 + \mathbf{n}_1[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)]\right.$$

$$\left. - \frac{\pi}{2}\prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}} + [\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\mathbf{u}_{M_1}]\frac{\pi}{2}\right\}. \tag{60}$$

Then, if

$$\mathbf{n}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)] - \frac{\pi}{2}\prod_{k=1}^{M_1}[x_{1k}(t)]^{m_{1k}} \neq \text{const} \tag{61}$$

unless $\mathbf{n}_1 = \mathbf{0}_{N_1}$ and $\mathbf{m}_1 = \mathbf{0}_{M_1}$, where

$$\mathbf{0}_L = (0, - - - , 0)' \tag{62}$$

is a L-dimensional vector with zero components, the carrier component of $s_{12}(t)$ is

$$s_{12}(t)_{\substack{\text{carrier}\\ \text{component}}} = P_1^{1/2} A_{12}(\mathbf{0}_{N_1}; \mathbf{0}_{M_1})\sin(\omega_1 t + \phi_1). \tag{63}$$

Therefore, provided that

$$\alpha_{1k} < \frac{2.4048}{(2)^{1/2}}, \qquad k = 1, 2, - - - , N_1 \tag{64}$$

and

$$\beta_{1k} < \frac{\pi}{2}, \qquad k = 1, 2, - - - , M_1 \tag{65}$$

so that

$$A_{12}(\mathbf{0}_{N_1}; \mathbf{0}_{M_1}) > 0, \tag{66}$$

the instantaneous phase of the carrier component of the received signal is $\omega_1 t + \phi_1$, and the fraction of the power P_1 in the carrier component of the received signal is

$$\eta_{12}(\mathbf{0}_{N_1}; \mathbf{0}_{M_1}) = \prod_{k=1}^{N_1} J_0^2[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos^2\left(\beta_{1k}\right) \tag{67}$$

or, in decibels,

$$\eta_{12}^{*}\,(\mathbf{0}_{N_1};\,\mathbf{0}_{M_1}) = \sum_{k=1}^{N_1} 10\log_{10} J_0^2\,[(2)^{1/2}\,\alpha_{1k}] + \sum_{k=1}^{M_1} 10\log_{10}\cos^2\left(\beta_{1k}\right). \tag{68}$$

5. Signal Component of Receiver Output

This section derives equations for the fraction of the power in both the usable and distortion components of $s_{15}(t)$, the signal component of the receiver output. From Eq. (9), the signal component of the receiver output is

$$s_{15}(t) = P_1^{1/2} \sin\left[\theta_1(t) + \Delta_1\right]. \tag{69}$$

Comparing Eqs. (69) and (29), we may use Eq. (54) to obtain

$$s_{15}(t) = P_1^{1/2} \sum_{\mathbf{n}_1,\,\mathbf{m}_1} a_1(\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1) \prod_{k=1}^{M_1} [x_{1k}(t)]^{m_{1k}} \sin\left\{\mathbf{n}_1'\left[\omega_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)\right] + \left[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}\right]\frac{\pi}{2} + \Delta_1\right\}. \tag{70}$$

Examining Eq. (47) we note that

$$a_1(-\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1) = a_1(\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1). \tag{71}$$

In summing over $\mathbf{n}_1$, $-\mathbf{n}_1$ will assume the same set of values as $\mathbf{n}_1$. Thus

$$\begin{aligned} s_{15}(t) &= P_1^{1/2} \sum_{\mathbf{n}_1,\,\mathbf{m}_1} a_1(\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1) \prod_{k=1}^{M_1} [x_{1k}(t_1)]^{m_{1k}} \frac{1}{2}\left[\sin\left\{\mathbf{n}_1'\left[\omega_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)\right] + \left[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}\right]\frac{\pi}{2} + \Delta_1\right\}\right. \\ &\quad \left. + \sin\left\{-\mathbf{n}_1'\left[\omega_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)\right] + \left[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}\right]\frac{\pi}{2} + \Delta_1\right\}\right] \\ &= P_1^{1/2} \sum_{\mathbf{n}_1,\,\mathbf{m}_1} a_1(\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1) \sin\left\{\left[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}\right]\frac{\pi}{2} + \Delta_1\right\} \\ &\quad \times \prod_{k=1}^{M_1} [x_{1k}(t)]^{m_{1k}} \cos\left\{\mathbf{n}_1'\left[\omega_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)\right]\right\}. \end{aligned} \tag{72}$$

Combining identical terms

$$\begin{aligned} s_{15}(t_1) &= P_1^{1/2} \sum_{\substack{\mathbf{n}_1,\,\mathbf{m}_1 \\ \mathbf{n}_1'\,\underline{\omega}_1 \geqq 0}} \varepsilon_{\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1}}\, a_1(\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1) \sin\left\{\left[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}\right]\frac{\pi}{2} + \Delta_1\right\} \\ &\quad \times \prod_{k=1}^{M_1} [x_{1k}(t)]^{m_{1k}} \cos\left\{\mathbf{n}_1'\left[\omega_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)\right]\right\}, \end{aligned} \tag{73}$$

where ε_l is one for $l = 0$ and two for $l > 0$. Thus

$$s_{15}(t_1) = P_1^{1/2} \sum_{\substack{\mathbf{n}_1,\,\mathbf{m}_1 \\ \mathbf{n}_1'\,\underline{\omega}_1 \geqq 0}} A_{15}(\mathbf{n}_1;\,\mathbf{m}_1) \prod_{k=1}^{M_1} [x_{1k}(t)]^{m_{1k}} \cos\left\{\mathbf{n}_1'\left[\omega_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)\right]\right\} \tag{74}$$

where

$$A_{15}(\mathbf{n}_1;\,\mathbf{m}_1) = \varepsilon_{\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1}}\, a_1(\mathbf{n}_1;\,\mathbf{m}_1;\,\boldsymbol{\alpha}_1;\,\boldsymbol{\beta}_1) \sin\left\{\left[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}\right]\frac{\pi}{2} + \Delta_1\right\}. \tag{75}$$

In computing signal-to-noise ratios, the fraction, $\eta_{15}(\mathbf{n}_1; \mathbf{m}_1)$, of the power P_1 in the $(\mathbf{n}_1; \mathbf{m}_1)$ component of the receiver output may be more useful than the amplitude of this component. Clearly

$$\begin{aligned} \eta_{15}(\mathbf{n}_1; \mathbf{m}_1) &= \frac{1}{2} A_{15}^2(\mathbf{n}_1; \mathbf{m}_1), \qquad \mathbf{n} \neq \mathbf{0}_{N_1} \\ &= A_{15}^2(\mathbf{0}_{N_1}; \mathbf{m}_1), \qquad \mathbf{n}_1 = \mathbf{0}_{N_1} \end{aligned} \tag{76}$$

Using Eqs. (75) and (47)

$$\begin{aligned} \eta_{15}(\mathbf{n}_1; \mathbf{m}_1) &= 2a_1^2(\mathbf{n}_1; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) \sin^2\left\{[\mathbf{a}'(\mathbf{n}_1)\,\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}]\,\frac{\pi}{2} + \Delta_1\right\} \qquad \mathbf{n}_1 \neq \mathbf{0}_{N_1} \\ &= 2\prod_{k=1}^{N_1} J_{n_{1k}}^2[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos^2\left(\beta_{1k} - m_{1k}\frac{\pi}{2}\right) \sin^2\left\{[\mathbf{a}'(\mathbf{n}_1)\mathbf{u}_{N_1} + \mathbf{m}_1'\,\mathbf{u}_{M_1}]\,\frac{\pi}{2} + \Delta_1\right\} \end{aligned} \tag{77}$$

and

$$\begin{aligned} \eta_{15}(\mathbf{0}_{N_1}; \mathbf{m}_1) &= a_1^2(\mathbf{0}_{N_1}; \mathbf{m}_1; \boldsymbol{\alpha}_1; \boldsymbol{\beta}_1) \sin^2\left(\mathbf{m}_1'\,\mathbf{u}_{M_1}\frac{\pi}{2} + \Delta_1\right) \\ &= \prod_{k=1}^{N_1} J_0^2[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos^2\left(\beta_{1k} - m_{1k}\frac{\pi}{2}\right) \sin^2\left(\mathbf{m}_1'\,\mathbf{u}_{M_1}\frac{\pi}{2} + \Delta_1\right). \end{aligned} \tag{78}$$

If, for any vector

$$\mathbf{p}_1 = (p_{11}, - - - , p_{1N_1})', \tag{79}$$

having components which may be positive or negative integers and any vector

$$\mathbf{r}_1 = (r_{11}, - - - , r_{1N_1})', \tag{80}$$

having components which are zero or one,

$$\prod_{k=1}^{M_1} [x_{1k}(t)]^{r_{1k}} \cos\{\mathbf{p}_1'[\boldsymbol{\omega}_1 t + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t)]\} \neq \text{const} \cdot \prod_{k=1}^{M_1} [x_{1k}(t+\tau)]^{n_{1k}} \cos\{\mathbf{n}_1'[\boldsymbol{\omega}_1(t+\tau) + \boldsymbol{\phi}_1 + \boldsymbol{\theta}_1(t+\tau)]\} \tag{81}$$

for any τ, unless $\mathbf{r}_1 = \mathbf{m}_1$ and $\mathbf{p}_1 = \pm\mathbf{n}_1$, each of the $(\mathbf{n}_1; \mathbf{m}_1)$ components of the receiver output is unique and the terms of Eq. (74), which reproduce the components of the modulation introduced at the transmitter, are those for which $\mathbf{n}_1 = \boldsymbol{\delta}_{lN_1}$ and $\mathbf{m}_1 = \mathbf{0}_{M_1}$ and those for which $\mathbf{n}_1 = \mathbf{0}_{N_1}$ and $\mathbf{m}_1 = \boldsymbol{\delta}_{lM_1}$, where $\boldsymbol{\delta}_{lL}$ is a L-dimensional vector with components

$$\begin{aligned} \delta_{lk} &= 1 \qquad k = l \\ &= 0 \qquad k \neq l \end{aligned} \qquad k = 1, 2, - - -, L; \qquad 1 \leq l \leq L. \tag{82}$$

The fraction of the power P_1 in each of these components of the receiver output is

$$\eta_{15}(\boldsymbol{\delta}_{lN_1}; \mathbf{0}_{M_1}) = 2J_1^2[(2)^{1/2}\alpha_{1l}] \prod_{\substack{k=1 \\ k\neq l}}^{N_1} J_0^2[(2)^{1/2}\alpha_{1k}] \prod_{k=1}^{M_1} \cos^2(\beta_{1k}) \quad \cos^2\Delta_1 \tag{83}$$

or

$$\eta_{15}(\mathbf{0}_{N_1}; \boldsymbol{\delta}_{lM_1}) = \sin^2(\beta_{1l}) \prod_{k=1}^{N_1} J_0^2[(2)^{1/2}\alpha_{1k}] \prod_{\substack{k=1 \\ k\neq l}}^{M_1} \cos^2(\beta_{1k}) \quad \cos^2\Delta_1 \tag{84}$$

C-3.

In decibels,

$$\eta_{15}^{*}(\delta_{lN_1};\mathbf{0}_{M_1}) = 10\log_{10}\eta_{15}(\delta_{lN_1};\mathbf{0}_{M_1})$$
$$= 10\log_{10}2J_1^2[(2)^{1/2}\alpha_{1l}] + \sum_{\substack{k=1\\k\neq l}}^{N_1} 10\log_{10}J_0^2[(2)^{1/2}\alpha_{1k}]$$
$$+ \sum_{k=1}^{M_1} 10\log_{10}\cos^2(\beta_{1k}) + 10\log_{10}\cos^2(\Delta_1) \quad (85)$$

and

$$\eta_{15}^{*}(\mathbf{0}_{N_1};\delta_{lN_1}) = 10\log_{10}\eta_{15}(\mathbf{0}_{N_1};\delta_{lM_1})$$
$$= 10\log_{10}\sin^2(\beta_{1l}) + \sum_{k=1}^{N_1} 10\log_{10}J_0^2[(2)^{1/2}\alpha_{1k}]$$
$$+ \sum_{\substack{k=1\\k\neq l}}^{M_1} 10\log_{10}\cos^2(\beta_{1k}) + 10\log_{10}\cos^2(\Delta_1). \quad (86)$$

If the terms of Eq. (74) are not unique, further manipulations will be required to combine those terms which differ only in amplitude or phase.

In most frequency-multiplexed, PM communication systems only those components of the receiver output, which reproduce the components of the modulation introduced at the transmitter, are used. The remaining terms are distortion.

6. Examples

To illustrate the application of these results, let us apply them to the DSN-*Mariner C* and the *Mariner C*-DSN telecommunication systems.

a. DSN-Mariner C telecommunication system. Commands, in the form of a sequence of binary digits, are transmitted to the *Mariner C* spacecraft through the DSN-*Mariner C* telecommunication system. The transmitter of this PM communication system is phase modulated by two command subcarriers, of which one is a binary sequence used for synchronization, and the other is a sine wave phase modulated by the command information.

Returning to Section 2a, we note that in this case $N_1 = M_1 = 1$ and

$$\theta_1(t) = (2)^{1/2}\alpha_{11}\cos[\omega_{11}t + \phi_{11} + \theta_1(t)] + \beta_{11}x_{11}(t) \quad (87)$$

where ω_{11} and ϕ_{11} are the angular frequency and phase of the unmodulated command data subcarrier and $x_{11}(t)$ is the binary-valued command sync subcarrier. α_{11} (=0.507 rad rms) and β_{11} (=0.655 rad rms) are the rms phase deviations of the carrier produced by these signals.

From Eqs. (63) and (56), the carrier component of the spacecraft receiver input is

$$s_{12}(t)_{\substack{\text{carrier}\\\text{component}}} = (2P_1)^{1/2}J_0[(2)^{1/2}\alpha_{11}]\cos(\beta_{11})\sin(\omega_1 t + \phi_1) \quad (88)$$

and from Eq. (68), the fraction of the power P_1 in this signal is

$$\eta_{12}^{*}(0;0) = 10\log_{10}J_0^2[(2)^{1/2}\alpha_{11}] + 10\log_{10}\cos^2(\beta_{11})$$
$$= -3.17\text{ db}. \quad (89)$$

Assuming the DSN transmitter frequency is adjusted to null, the spacecraft receiver static phase error, $\Delta_1 = 0$. Then, using Eqs. (74), (75), and (47), the signal component of the spacecraft receiver output is,

$$s_{15}(t) = P_1^{1/2}\sum_{n_{11}=0}^{\infty}\sum_{m_{11}=0}^{1}\varepsilon_{n_{11}}J_{n_{11}}[(2)^{1/2}\alpha_{11}]\cos\left(\beta_{11} - m_{11}\frac{\pi}{2}\right)$$
$$\times\sin\left[(n_{11}+m_{11})\frac{\pi}{2}\right][x_{11}(t)]^{m_{11}}$$
$$\times\cos\{n_{11}[\omega_{11}t + \phi_{11} + \theta_{11}(t)]\}. \quad (90)$$

Clearly, the terms $n_{11} = 1$, $m_{11} = 0$ and $n_{11} = 0$, $m_{11} = 1$ reproduce the two command subcarrier signals. The remaining terms of Eq. (90) are distortion.

Using Eqs. (83) and (84), at the spacecraft receiver output, the fraction of the power P_1 in the command data subcarrier is

$$\eta^*_{15}(1;0) = 10\log_{10} 2J_1^2[(2)^{1/2}\alpha_{11}] + 10\log_{10}\cos^2(\beta_{11}) = -8.48\,\text{db} \quad (91)$$

and that in the command sync subcarrier is

$$\eta^*_{15}(0;1) = 10\log_{10} J_0^2[(2)^{1/2}\alpha_{11}] + 10\log_{10}\sin^2(\beta_{11}) = -5.46\,\text{db}. \quad (92)$$

b. Mariner C-DSN telecommunication system. Telemetry in the form of a sequence of binary digits is transmitted from the *Mariner C* spacecraft to the DSN through the *Mariner C*-DSN telecommunication system. The transmitter of this PM communication system is phase modulated by two telemetry subcarriers, of which one is a binary sequence used for synchronization and the second is a binary sequence which is the product of a square wave and the sequence of telemetry bits.

Returning to Section 2a, we note that in this case $N_1 = 0$, $M_1 = 2$, and

$$\theta_1(t) = \beta_{11}x_{11}(t) + \beta_{12}x_{12}(t) \quad (93)$$

where $x_{11}(t)$ is the sync subcarrier, $x_{12}(t)$ is the data subcarrier, and β_{11} (=0.451 rad rms) and β_{12} (=0.809 rad rms) are the carrier rms phase deviations produced by these signals.

From Eqs. (63) and (56), the carrier component of the DSN receiver input is

$$s_{12}(t)_{\substack{\text{carrier}\\ \text{component}}} = (2P_1)^{1/2}\cos(\beta_{11})\cos(\beta_{12})\sin(\omega_1 t + \phi_1) \quad (94)$$

and from Eq. (68) the fraction of the power in this signal is

$$\eta^*_{12}(-;0,0) = 10\log_{10}\cos^2(\beta_{11}) + 10\log_{10}\cos^2(\beta_{12}) = -4.14\,\text{db}. \quad (95)$$

Assuming the static phase error in the DSN receiver is negligible, $\Delta_2 = 0$. Then, using Eqs. (74), (75), and (47), the signal component of the receiver output is

$$s_{15}(t) = P_1^{1/2}\sin(\beta_{11})\cos(\beta_{12})x_{11}(t) + P_1^{1/2}\cos(\beta_{11})\sin(\beta_{12})x_{12}(t). \quad (96)$$

Hence, in this special case, only the two telemetry subcarrier signals appear at the DSN receiver output.

Using Eqs. (83) and (84), at the DSN receiver output, the fraction of the power P_1 in the telemetry sync subcarrier is

$$\eta^*_{15}(-;1,0) = 10\log_{10}\sin^2(\beta_{11}) + 10\log_{10}\cos^2(\beta_{12}) = -10.43\,\text{db}, \quad (97)$$

and that in the telemetry data subcarrier is

$$\eta^*_{15}(-;0,1) = 10\log_{10}\cos^2(\beta_{11}) + 10\log_{10}\sin^2(\beta_{12}) = -3.73\,\text{db}. \quad (98)$$

The *Mariner C* telemetry demodulator recovers only the sideband power in the first harmonic of the square wave data subcarrier. This loss, however, is peculiar to the *Mariner C* telemetry demodulator and should be included in the performance analysis of the demodulator.

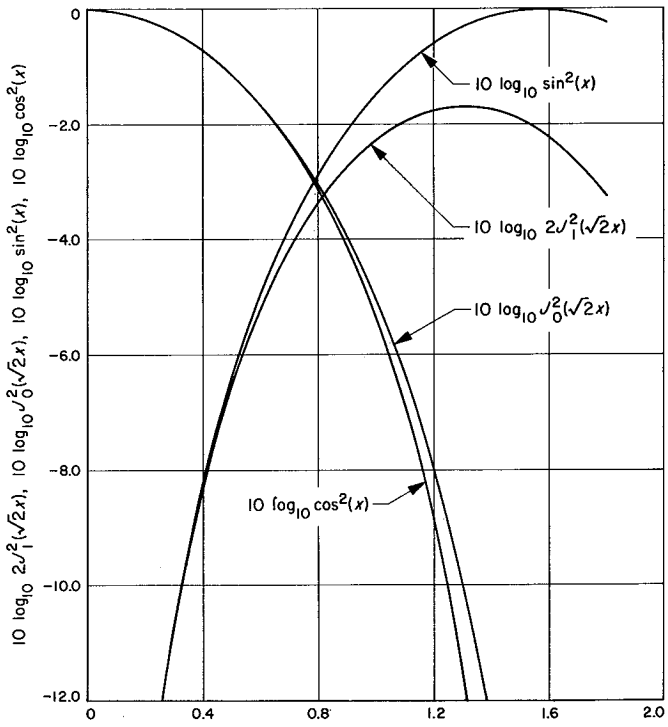


Fig. 11. The functions $10\log_{10} 2J_1^2[(2)^{1/2}x]$, $10\log_{10} J_0^2[(2)^{1/2}x]$, $10\log_{10}\sin^2(x)$, and $10\log_{10}\cos^2(x)$

7. Conclusion

This analysis provides a method of simplifying the evaluation of frequency-multiplexed, PM communication channels to that of evaluating the performance of the subcarrier extractor in obtaining the desired information from a sinusoidal or binary-valued signal, distortion, and white gaussian noise. The analysis demonstrates that, over the frequency band of interest, the power spectral densities of the gaussian noise at the receiver input and output are the same. The analysis also derives equations for the fraction of the received signal power in both the usable and distortion components of the receiver input and output. The functions $10 \log_{10} 2J_1^2 [(2)^{1/2} x]$, $10 \log_{10} J_0^2 [(2)^{1/2} x]$, $10 \log_{10} \sin^2 (x)$, and $10 \log_{10} \cos^2 (x)$, plotted in Fig. 11, will prove useful in evaluating $\eta_{12}^* (\mathbf{0}_{N_1}; \mathbf{0}_{M_1})$, $\eta_{15}^* (\boldsymbol{\delta}_{lN_1}; \mathbf{0}_{M_1})$, and $\eta_{15}^* (\mathbf{0}_{N_1}; \boldsymbol{\delta}_{lM_1})$.

Reference

1. Cramer, H., "Mathematical Methods of Statistics," Princeton University Press, New Jersey, 1961, or Wilks, S. S., "Mathematical Statistics," John Wiley and Sons, Inc., New York, 1962.

XVII. Communications Elements Research

A. Multipactor Effects

H. Erpenbach

1. Secondary Electron Emission

Investigation of the possible reduction of secondary electron emission from copper surfaces when impacted with electrons is continuing. Of three techniques mentioned in *SPS 37-38*, Vol. IV, for generating surfaces having $\delta < 1$, the cathodic etching technique seems to be the most promising. It was noted that when a nearly atomically clean surface was vented to high purity dry nitrogen, the resulting number of monolayers that was condensed and adsorbed in the surface would permit sufficient protection from oxidation for 24 hr. However, further tests now indicate that the thin layers show steady deterioration after 48 hr. Since then there has been a different approach to the problem. Two gases have been used: one for cleaning and etching of the surface (argon), and the other a hydrocarbon which dissociates to free carbon in the 5 kv discharge to form a thin carbon film of high density on the target surface. Carbon in the graphite bulk form has a $\delta < 1$. However, when pure graphite is vaporized in vacuum, by either electron beam bombardment or point-to-point resistance heating, the resulting thin carbon films are amorphous, and have a $\delta > 1$. Also, films deposited from pure graphite adhere poorly in contrast to the films from hydrocarbon dissociation, which adhere tightly to the substrate.

2. Experiment

A highly polished copper target, 1¼ in. in diameter by ¼ in. thick, is placed on an insulated water-cooled pedestal in a vacuum chamber. The air is removed and high purity (99.9999%) argon is admitted at a pressure of approximately 3 mtorr. Next, a negative potential of 5 kv is applied to the copper target, with the positive terminal connected to the base plate. The observed cathode glow is adjusted to approximately 2 ma/cm^2, a current density at which the glow is quite evenly distributed across the target. After about 1 min, a thin film of copper is deposited on the side of the glass wall of the bell jar nearest the target. This copper has been removed by the impact of 5-kv argon ions. As the material is continually removed atomically in the electrical discharge of ultra-pure gas, the surface is extremely clean at any instant. At this time, the argon is stopped and a pure hydrocarbon gas is admitted simultaneously (in this case, methane, CH_4, was used). The color of the glow discharge changes from blue,

which is characteristic of argon, to the blue-purple of hydrogen. After approximately 30 sec the target takes on a dark appearance, after 3 min the surface is completely covered with a thin carbon film.

Fig. 1 shows the secondary electron emission δ with voltage coordinate V_p of carbon in the pure graphite form; it is one of the very few materials that has $\delta < 1$. This carbon target is used from time to time as a check on the secondary electron measuring apparatus, to insure proper functioning.

The secondary electron emission δ of a polished copper target etched with argon ions and treated with methane ions is shown in Fig. 2. The 4-day exposure indicates that

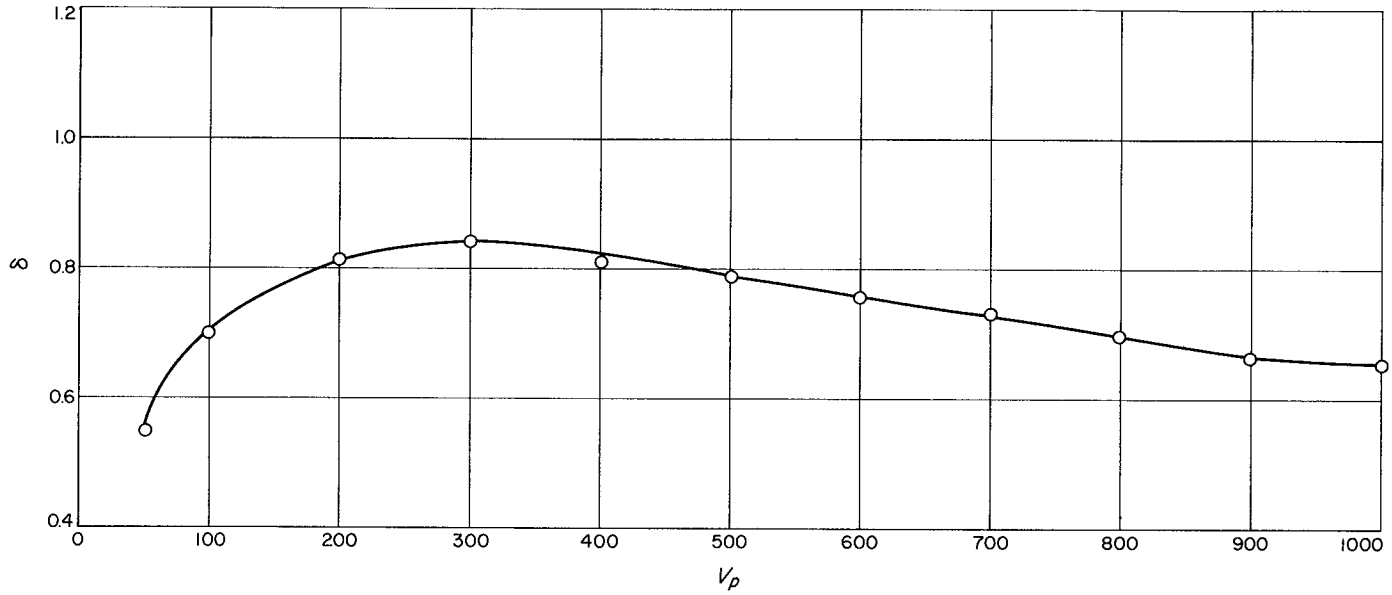


Fig. 1. Ratio δ of carbon

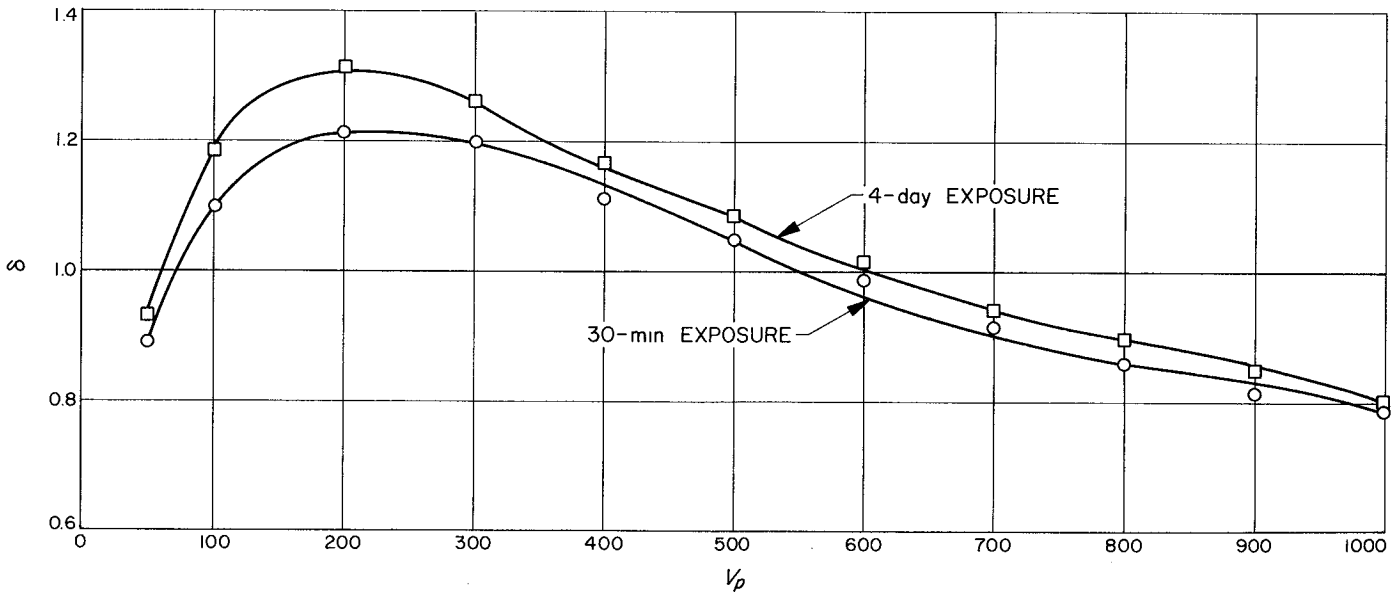


Fig. 2. Ratio δ of polished copper

an improvement has been made with the use of a very thin carbon deposit. The value of δ_{max} is 1.3, a low yield for a copper surface that has been exposed to the atmosphere for such a long time.

Fig. 3 shows δ for a copper target sanded with #400 paper, etched with argon ions and treated with methane. It has δ_{max} of 1.25. Fig. 4 shows δ for a highly polished copper target etched with argon ions for a longer period of time.

Fig. 5. shows δ for a polished copper target etched with argon ions and treated with cyclopropane. This last surface has the lowest secondary electron emission to date, apparently owing to the substitution of cyclopropane for methane. Table 1 is the data compiled.

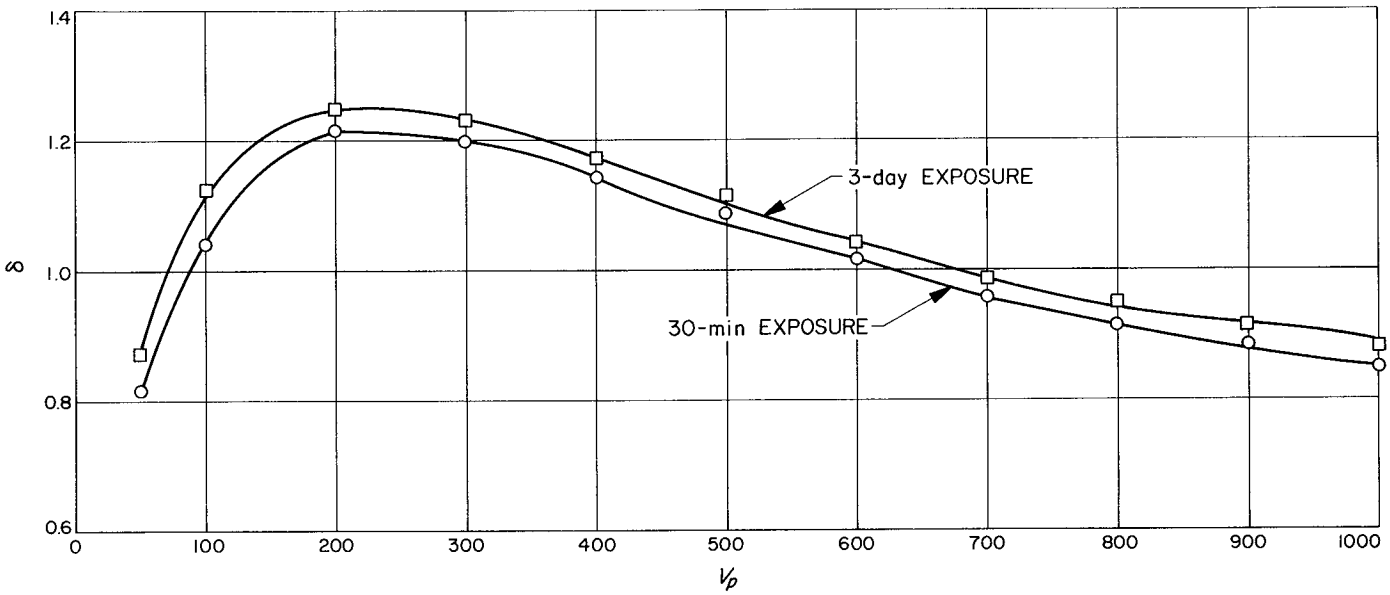


Fig. 3. Ratio δ of sanded copper

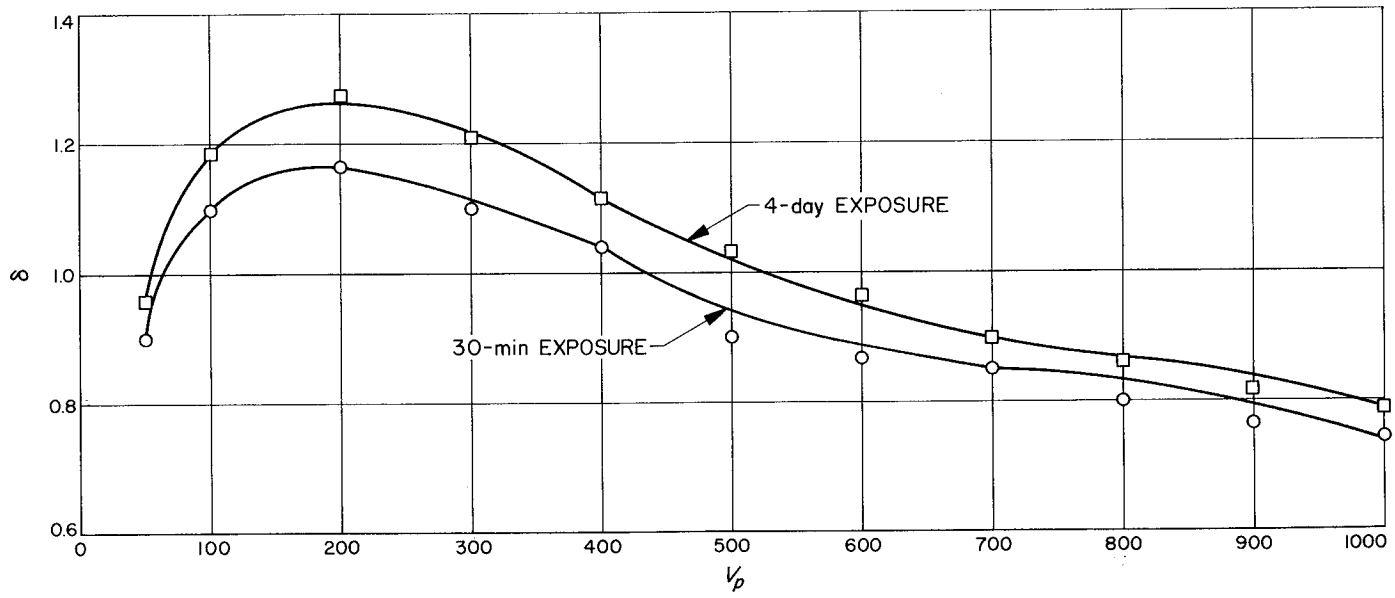


Fig. 4. Ratio δ of deeply etched copper

Table 1. Compiled data

Sample No.	δ_{max} after exposure to atmosphere		Voltage at maximum δ	Discharge time, min, at 5 kv		Pressure, mtorr		Current density, ma/cm^2	
	30 min	Days ( )		Argon	Methane or cyclopropane	Argon	Methane or cyclopropane	Argon	Methane or cyclopropane
2	1.22	1.32 (4)	200	15	M5	4	5	3	3
3	1.22	1.25 (3)	225	5	M10	4	5	3	3
4	1.17	1.27 (4)	200	120	M3	6	6	2	2
5	0.98	1.02 (5)	475	15	C3	6	4	2	4

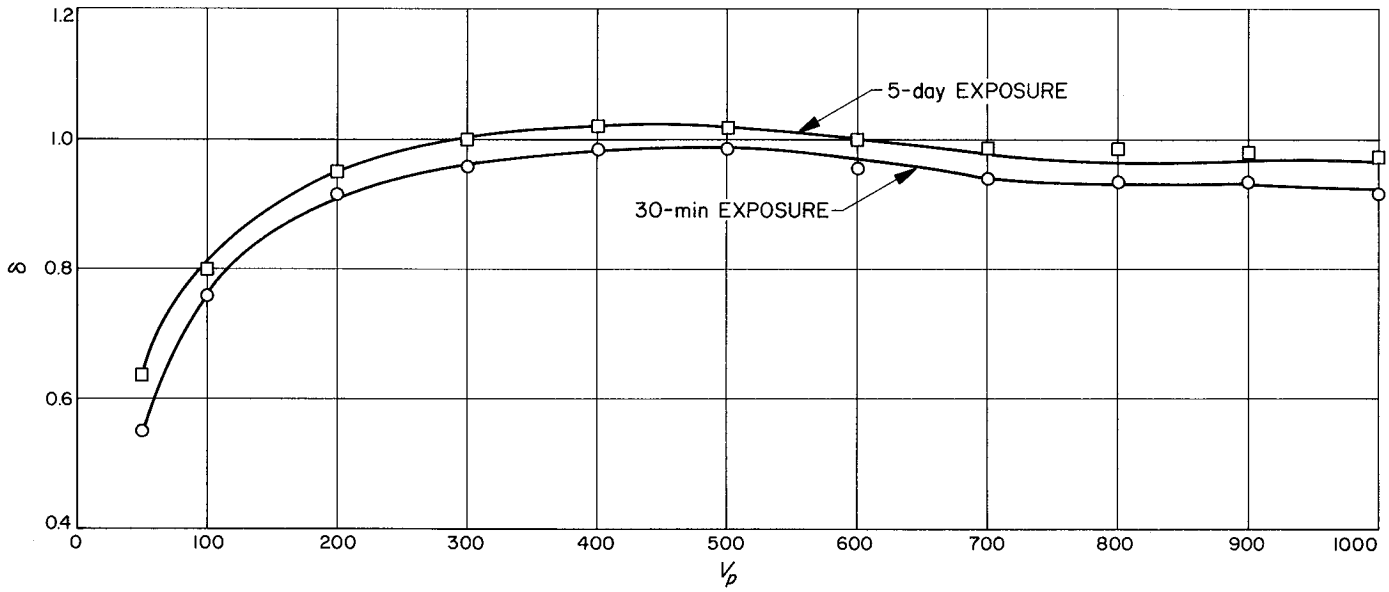


Fig. 5. Ratio δ of carbon deposit from cyclopropane

At present we cannot explain why a thin film from one hydrocarbon yields fewer secondaries than a film from any other; other compounds in the hydrocarbon series will be tested.

B. Optical Communications Components

W. H. Wells

As reported in *SPS 37-38*, Vol. IV, p. 175, the CO_2 laser at 10.6 μ is JPL's present choice for optical communications experiments. A problem in the design of stable molecular lasers for communications is the tendency for more than one line in the rotational spectrum to oscillate simultaneously. Also, small perturbations or vibrations of the resonator may cause the oscillation to jump between two or more of these lines. A brief but successful theoretical study of optimum resonator length for suppression of all but one of these spectral lines has been made.

The suppression of oscillation in multiple spectral lines is not inconsistent with the high efficiencies reported for CO_2 lasers. The mean free time between collisions in a typical CO_2 laser is of the order of 10^{-7} sec. This compares to a relaxation time of 10^{-3} sec for the vibrational laser transition. That is, the number of molecular collisions during the life of the excited state is on the order of $10{,}000 = 10^{-3}/10^{-7}$. Nearly all of these collisions change the molecule's rate of rotation among the 30 or more probable levels in the rotational fine structure. Hence there are many opportunities to catch each molecule in any of the more probable rotational levels before the molecule has time to lose its energy.

Fig. 6(a) shows a typical tuning curve for a molecular laser when only one line has enough gain to exceed the

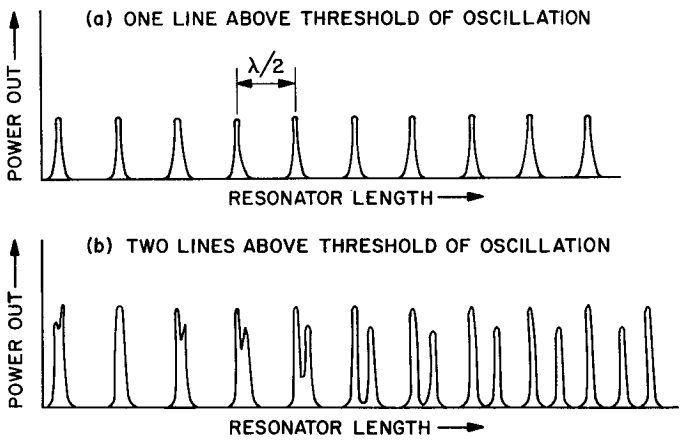


Fig. 6. Tuning curves for a laser resonator

oscillation threshold. As the resonator length is increased, the oscillation is tuned in at intervals of one-half wavelength, representing successively higher order longitudinal modes. For this case in a CO_2 laser, a fine adjustment with only 6 μ travel suffices for tuning, regardless of length. Fig. 6(b) shows a situation one step closer to normal. Here two spectral lines have sufficient gain to oscillate, but since their wavelengths differ very little, there are wide regions of interference represented by the left side of this figure. We could solve this case readily by using the concept of beats; however, typically we wish to suppress not just one, but approximately ten interfering wavelengths. In the latter case, the lengths that cause the desired line to stand clear of the others are rather rare. The widths of the lines in Fig. 6 are greatly exaggerated for illustration; nevertheless, it is undesirable to have lines too close together, even when they are resolved. For example, suppose temperature change, repairs, or something else has disturbed the resonator length. To recover the proper setting, one would have to set up a high-resolution well-calibrated spectrograph unless the coarse length of the laser was properly chosen so that the unwanted lines cluster, while the desired one stands alone on a tuning curve, such as in Fig. 6(b). To summarize, our problem is not to list precise resonator lengths that tune in the desired line, but rather to find approximate design lengths that define regions of reduced interference, as on the right side of Fig. 6(b). Such a region may be 400 μ wide, containing 75 tuning positions for each rotational line; the point is that the set of positions that tune in the desired line should be well separated from all other pertinent sets.

The theory is simple, but it required use of the IBM 1620 computer to search the practical range of resonator lengths for optimum values. It was apparent that published rotational spectra of CO_2 are not sufficiently accurate. These numbers were varied within their error limits and an entirely different set of optimum lengths was computed. Laser spectroscopy will probably improve the measured accuracy of these frequencies; if not, it is possible to invert the technique described here to accomplish almost the same thing. A laser resonator can be scanned over a wide range of lengths to determine experimentally the way in which the tuning curves of the various rotational lines mesh together.

Proceeding as though the spectrum used were sufficiently accurate serves as a demonstration and verifies that laser length can be optimized for a particular rotational line. Let F_i be the frequency intervals between the line chosen for oscillation and the ones we wish to suppress. They may be listed in arbitrary order, the index running from one to the total number of suppressed lines. The longitudinal mode frequencies of the resonator are multiples of the fundamental frequency $F_0 = c/2L$ where L is the mirror separation. The condition we wish to prevent is given by

$$F_i = NF_0 \tag{1}$$

for all values of i, and all integers N. Otherwise, some frequency interval F_i will be a multiple of the fundamental, so that the undesired line will be tuned in at the same points as the desired one, with its longitudinal mode number differing by N. Next, let us drop the factor c in order to use wave-number units and define

$$\Delta_i = 1/(2F_i)$$

Then Eq. (1) may be solved for the resonator lengths to be avoided:

$$L = N\Delta_i \tag{2}$$

Eq. (2), the basis of our problem, could hardly be simpler, but the number of these lengths between 15 and 100 cm is typically about 8,000.

Fig. 7 shows the result of a computer search that examined the sets of lengths between 15 and 100 cm for $i = 1$ to 10, then compared lengths in ascending order irrespective of i, and finally printed out every case in which the gap between bad lengths exceeded 0.4 mm. The chosen (P(22) is a strong line in the center of the usual CO_2 laser spectrum. Ten other lines are suppressed: P(12)–(32) (odd numbered lines are missing owing to symmetry of the molecule). Lengths and frequencies are given in cm and

	BAD LINES NEARBY SHORTER	LONGER	GOOD LENGTHS	SPACE BETWEEN BAD LENGTHS
1	P(24)	P(14)	14.936	.05437
2	P(32)	P(12)	17.541	.04332
3	P(30)	P(12)	19.132	.04605
4	P(14)	P(12)	21.732	.04513
5	P(16)	P(14)	26.030	.04310
6	P(14)	P(32)	26.877	.04804
7	P(14)	P(32)	29.480	.04258
8	P(18)	P(30)	31.170	.04453
9	P(26)	P(30)	33.771	.04236
10	P(16)	P(12)	33.832	.04146
11	P(12)	P(26)	36.367	.04178
12	P(32)	P(12)	36.643	.04288
13	P(14)	P(12)	43.378	.05082
14	P(16)	P(12)	47.734	.04211
15	P(16)	P(12)	50.334	.04121
16	P(12)	P(20)	52.869	.04089
17	P(12)	P(24)	55.471	.04518
18	P(32)	P(12)	55.746	.04243
19	P(14)	P(26)	62.478	.04163
20	P(14)	P(12)	63.547	.04040
21	P(16)	P(12)	64.236	.04186
22	P(14)	P(12)	65.084	.04277
23	P(32)	P(14)	65.241	.04410
24	P(16)	P(12)	66.836	.04095
25	P(12)	P(28)	69.371	.04145
26	P(32)	P(12)	69.648	.04147
27	P(18)	P(30)	70.443	.04268
28	P(30)	P(12)	71.187	.04067
29	P(12)	P(30)	81.471	.04588
30	P(24)	P(14)	81.742	.04059
31	P(12)	P(30)	84.071	.04529
32	P(32)	P(12)	84.344	.04521
33	P(12)	P(30)	89.270	.04411
34	P(32)	P(12)	89.544	.04572
35	P(20)	P(12)	91.350	.04084
36	P(14)	P(28)	93.678	.04929
37	P(32)	P(12)	93.950	.04153

DELTA L MIN = .0400
F(I) = 9.4230 8.8470 7.4850 7.1170 5.5750
SUPPRESSED P LINES (I) = 12 32 14 30 16 28 18 26 20
5.3680 3.6920 3.6000 1.8340 1.8110
24

Fig. 7. Laser lengths for oscillating the P(22) line

cm^{-1}. The "good lengths" listed in the figure are midway between the more widely-spaced bad ones.

Results show that this exercise is a worthwhile part of the design process. For example, if our input data were accurate, we would not design a CO_2 laser resonator 59 cm long. Instead, we should choose the length as close as possible to 55.471, 55.746, or 62.478 cm, then find the exact length by trial and error adjustment of one mirror. The choice of 0.4 mm minimum gap as the criterion for a good length was somewhat arbitrary. There is no danger of line widths filling up the gap, because the Q factors of both the resonator and molecular line exceed $L/\Delta L$ by about 2 orders of magnitude. If the minimum gap size was decreased while the number of suppressed lines is held constant, then the program would yield more good lengths than necessary for normal design problems. Decreasing the minimum gap will normally be accompanied by suppression of a larger number of lines. Although the ten lines already chosen exceed the number that normally oscillate, there may be a noise advantage to suppression of nonoscillating lines to prevent extra superradiant spontaneous emission (subthreshold laser behavior) into the longitudinal modes. Such a noise consideration would be important only in a laser that is already well stabilized against normal sources of noise.

C. Antennas for Space Communications: Antenna Pattern Synthesis

A. Ludwig

1. Introduction

When a fixed (nontracking) antenna is used on a spacecraft, it is usually desirable to maintain significant gain along a particular cut through the spacecraft antenna pattern which is defined by a line from the Earth to the spacecraft over the mission trajectory. In order to maximize gain over this cut, the gain at points significantly removed from the cut should be minimized. The *Mariner IV* high-gain spacecraft antenna reflector has an elliptical outline to provide such a pattern. In some cases the drop in gain, as one moves away from the peak of the beam in the broad plane of the elliptical pattern, is still too severe and it becomes desirable to raise sidelobes in this plane, preferably without disturbing the pattern in the narrow plane. An existing computer program, developed for ground antenna applications, has been adapted to synthesize a possible antenna configuration for achieving this spacecraft antenna objective.

2. Split Paraboloid Beam-Shaping Antenna

The class of antennas investigated in this study consists of two confocal paraboloids symmetrically joined, as shown in Fig. 8. The assumed apertures outline is a 46.0- × 21.20-in. ellipse, identical to the *Mariner IV* antenna aperture outline. This reflector configuration was input to a computer program which calculates the far-field pattern of a given reflector with a given illumination (*SPS 37-21*, Vol. III, pp. 28–33). Experimental amplitude patterns of the *Mariner IV* antenna feed were used for the reflector illumination. The resulting radiation patterns of this configuration are shown in Figs. 9 and 10. The assumed frequency of operation was 2295 Mc. The 0-deg axis tilt case shown in Fig. 9 identically represents the *Mariner IV* antenna, and experimental results are shown as circled points for comparison with the computed data. The agreement is satisfactory.

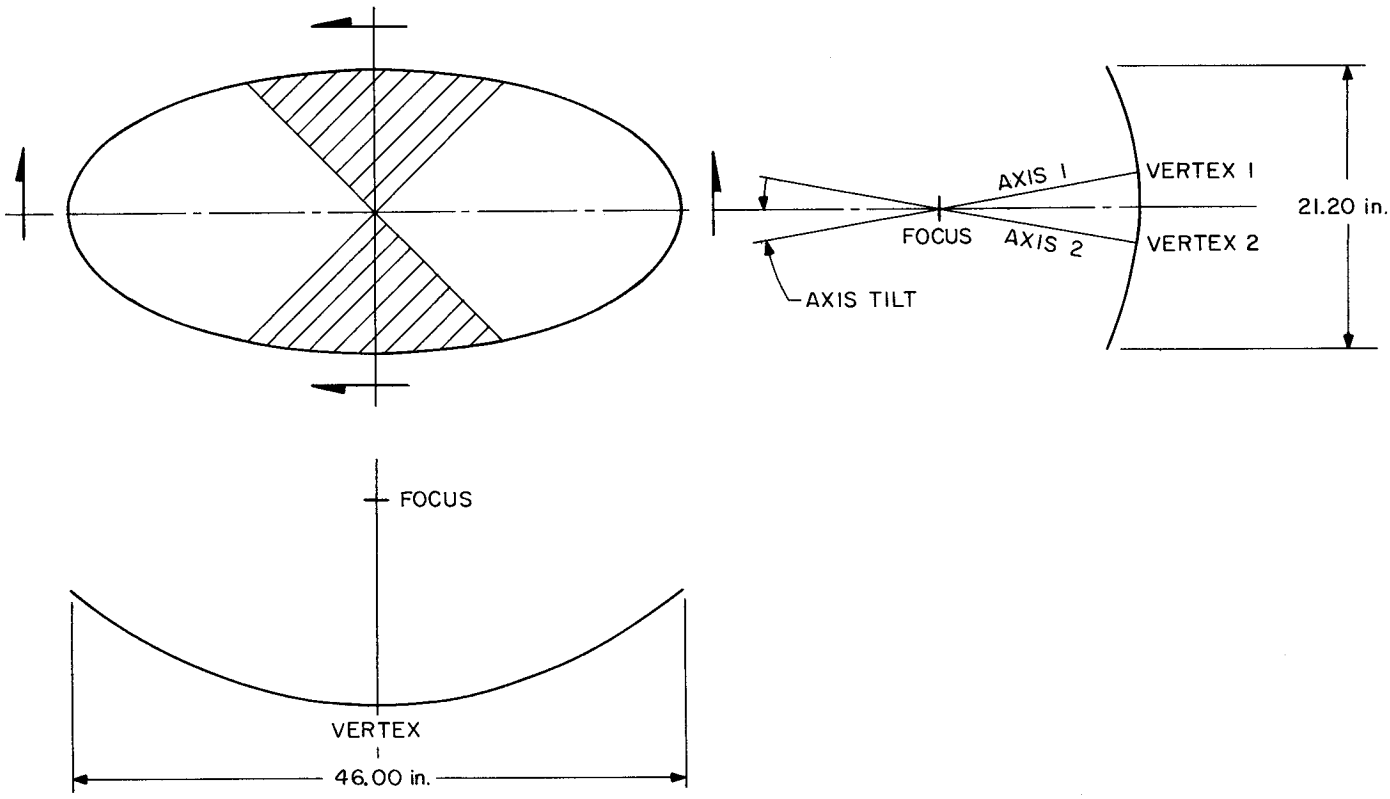


Fig. 8. Split paraboloid antenna

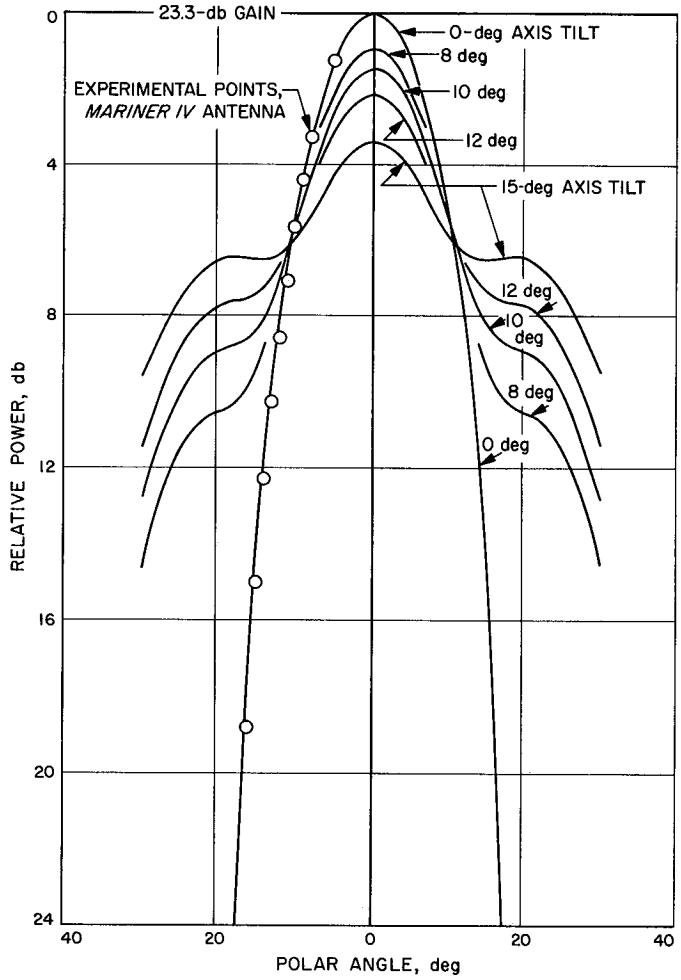


Fig. 9. Split paraboloid gain versus polar angle for various tilts, 90-deg clock angle

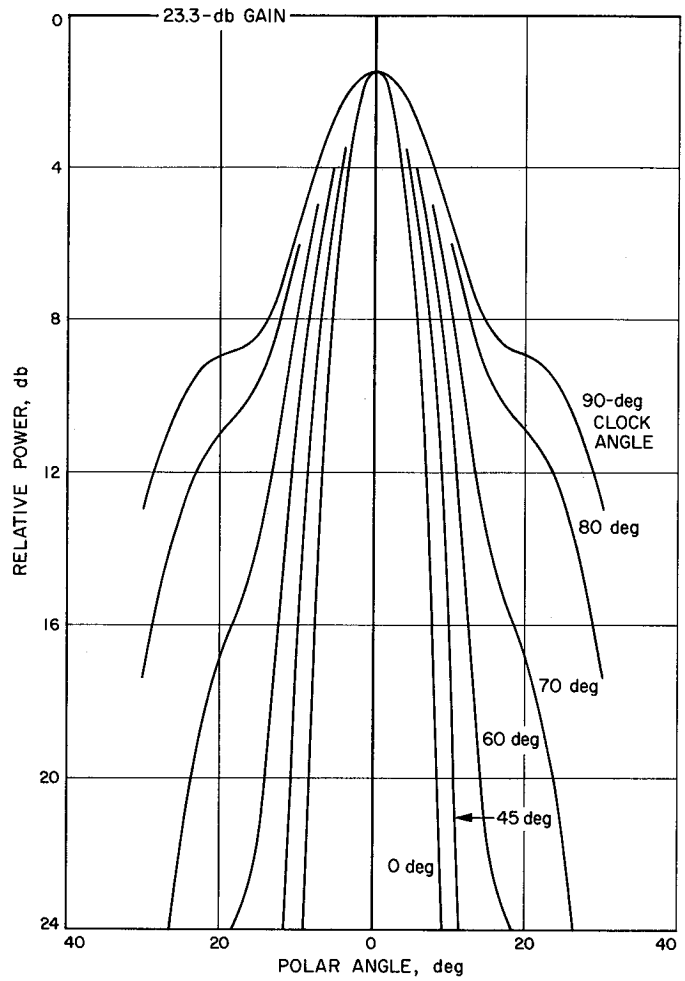


Fig. 10. Split paraboloid, each axis tilted 10 deg; gain versus polar angle for several clock angle cuts

One case was run where only a 90-deg sector on each side of the antenna (shaded area in Fig. 8) was tilted 15 deg. This raised very high sidelobes in the narrow plane of the antenna beam, and resulted in 0.5 db less on-axis gain than the case where the entire half of the paraboloid was tilted 15 deg. The conclusion drawn from this case is that, to prevent undue gain loss, the contours of the antenna which are perpendicular to the plane being synthesized should remain parabolic. For this condition, the narrow plane pattern (0 deg clock/angle) is virtually unchanged, as shown in Fig. 10.

It may be seen from Fig. 9 that the level of the sidelobes is well controlled by the tilt angle of the two-paraboloid axis, enabling a wide range of off-axis gains to be realized. Naturally, high sidelobes reduce the gain of the main beam—an unavoidable consequence of conservation of energy—but since the sidelobes do not appear in the narrow plane, the reduction of gain is minimized.

D. Deep Space Propagation Studies: Depolarization Effects Due to Solar Wind

G. Levy

1. Summary

It has been predicted that the relativistic electrons in the solar wind could produce a depolarization of the

S-band telemetry passing through. During the *Mariner IV* approach to Mars, and later as *Mariner IV* was occulted by the Sun, measurements were made to determine the magnitude of these depolarization effects. They were found to be below our threshold of unambiguous detectability.

2. Introduction

B. B. Lusignan (Ref. 1) predicted theoretically that relativistic electrons in the solar wind could alter the polarization ellipse of an electromagnetic wave passing through this medium. It was noted[1] that this phenomenon could cause significant loss of signal if circular polarization was employed in the *Mariner IV* S-band telemetry link. Although the phenomenon was predicted to have a negligible effect on the *Mariner IV* S-band communication system (Ref. 2), the suggestion of using the precision instrumentation of the NASA JPL Deep Space Network to attempt to identify the predicted phenomenon was followed.

3. Experiments

From January 19 through July 28, 1966, measurements were made to determine the ratio of right-hand circularly polarized (RCP) to left-hand circularly polarized (LCP) signals received at the Pioneer Station. This was accomplished by measuring the calibrated automatic gain control (AGC) voltage in the normal matched mode (RCP), then switching the receiver input to the mismatched (LCP) port of the antenna and again measuring the AGC voltage. On all occasions the signal level measured in the LCP mode was at least 17 db below the RCP mode, or else it was below the receiver threshold. The LCP signal measured could be completely accounted for by the combination of ellipticity of the spacecraft antenna and the 0.5-db ellipticity of the DSIF receiving antenna.

On April 1, 1966, the *Mariner IV* spacecraft was approximately 327.7 Gm from Earth, and the angle formed between the center of the Sun and the Earth and the probe was only 1.05 deg. The spacecraft was tracked, using the AAS antenna and the programmed local oscillator receiver (Ref. 3). The LCP signal was below the threshold of detectability which was estimated to be 10 db below the matched RCP mode. Fig. 11(a) shows the spectrum of the RCP signal integrated for a total of 30 min. A bandwidth of 25 cps was employed and the receiver was tuned with a programmed local oscillator.

[1]Letter from V. R. Eshleman to J. Clark, November 15, 1963.

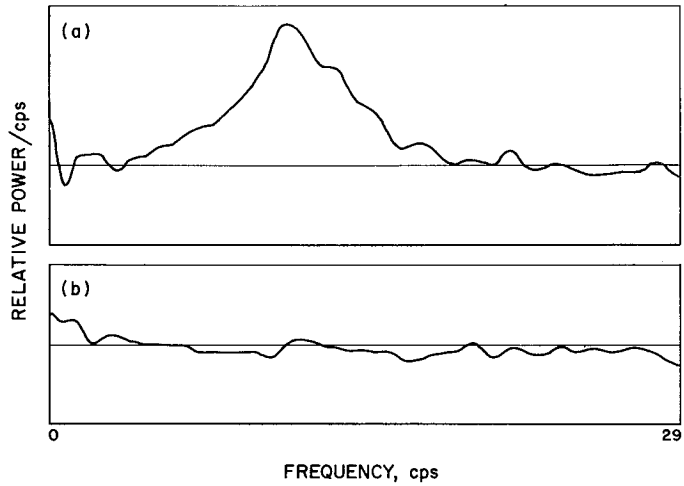


Fig. 11. *Mariner IV* depolarization measurements

Fig. 11(b) shows the spectrum of the LCP signal integrated for 77 min. There is no apparent signal present in the LCP case and it appears that one would have been detectable had it been no more than 10 db less than the RCP signal.

The cross-polarized feedthrough of the ortho-mode transducer and the ellipticity of the AAS feed system should have given a spurious cross-polarized signal of no more than 25.6 db down from the matched circular mode. Adding this to the cross-polarization signal from the spacecraft omnidirectional antenna, which had an ellipticity of 2.3 db, the total cross-polarized signal at the observed orientation angle ignoring the medium, should be between 14.6 and 21.8 db down from the RCP mode, which would be below the detection threshold.

In the April 1 experiment, the RCP signal was located and observed in a 400-cps bandwidth. The bandpass was then narrowed to 25 cps and the signal was sought on RCP and LCP. The spacecraft was transmitting on its auxiliary oscillator. There is a possibility that the oscillator might have drifted in frequency but, based on data taken during the rest of the solar occultation measurements, this seems unlikely. As a check, the last 18 min of the view period were used in measuring LCP in 100-cps filter. No signal was detectable in either the 25- or 100-cps filter for LCP.

The anisotropic nature of the solar wind can be explained by considering the effect of two linearly polarized waves propagating in a direction normal to the velocity vector of a relativistic electron (Ref. 1). One wave is polarized perpendicular to the velocity vector, and the

other parallel. The accelerations produced by the perpendicularly polarized wave on the electron make the electron's mass appear to be simply its rest mass. The mass seen by the wave polarized parallel to the velocity vector, however, is modified by a relativistic correction. The electron will therefore have an elliptical rather than circular oscillation superimposed on its linear velocity. This will result in circularly polarized signal giving rise to the opposite sense of circular polarization.

Lusignan[2] derived the following expression for the relative phase shift, in radians, between the perpendicular linear modes (into which circular polarization can be divided)

$$\phi_p = -D\left(\frac{\omega}{c}\right)\left(\frac{\omega_0^2}{\omega^2}\right)\left(\frac{1}{2}\beta^2\cos^2\theta\right)$$

where

D is the distance

ω is the wave angular frequency

ω_0 is the plasma angular frequency and is proportional to the square root of electron density

c is the velocity of light

β is the electron velocity radial from the Sun, divided by the velocity of light. It is assumed that $\beta << c$

θ is the angle between the electron velocity vector and the wave front.

If we assume that all the electrons are streaming radially from the Sun at the same velocity as the protons, then we can use the 500 km/sec value obtained by *Mariner II* (Ref. 4). We can use a value of 8 electrons/cm^3 at a distance of 1 AU and assume an inverse square law for the electron spatial distribution. Since both the angle θ and electron density change, it will be necessary to integrate the phase change along the ray path

$$\phi_p = \frac{\beta^2}{2\omega_c}\int \omega_0^2\cos^2\theta dD$$

and

$$\omega_0^2 = 3.19\times 10^9 N \text{ where } N \text{ is electrons/cm}^3$$

$$\phi_p = \frac{3.19\times 10^9 (1.67\times 10^{-3})^2}{4\pi\times 2.295\times 10^9\times 3\times 10^{10}}\int N\cos^2\theta dD$$

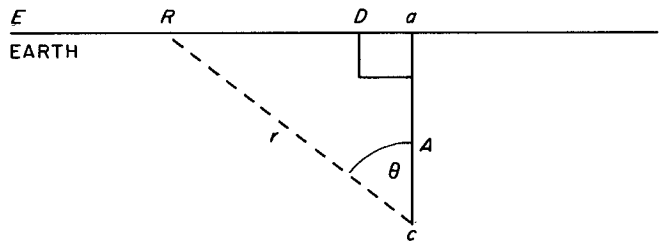
$$\phi_p = 1.028\times 10^{-17}\int N\cos^2\theta dD$$

Converting the unit of length to astronomical units and the angles to degrees:

$$\phi_\beta = 8.65\times 10^{-3}\int N\cos^2\theta dD$$

$$N = \frac{8}{r^2}$$

where r is the distance for the center of the Sun in astronomical units.



The range and angles subtended by *Mariner IV* were computed by D. Tito[3].

Let

a = point of closest approach to the Sun

A = line between center of Sun c and a

p = point along ray path

D = distance between a and p

r = distance between c and p

$\theta = \angle pca$

then

$$r = \frac{d}{\cos\theta}$$

[2]Lusignan, B. B. private communication.

[3]Tito, D., unpublished memo, November 3, 1965, "*Mariner IV* Solar Corona Equipment — Trajectory Related Information," 312.4-323.

and

$$D = A \tan\theta$$

$$\phi_\beta = 8.74 \times 10^{-3} \int \frac{8}{r^2} \cos^2\theta \, dD$$

$$\phi_\beta = 6.92 \times 10^{2} \int \frac{\cos^4\theta}{A^2} \, dD$$

$$D = A \tan\theta$$

$$dD = \frac{A \, d\theta}{\cos^2\theta}$$

$$\phi_\beta = 6.92 \times 10^{-2} \int \frac{\cos^2\theta}{A} \, d\theta$$

and for the case of April 1, $A = 0.02$ AU

Let us integrate from the Earth to the spacecraft

$$\angle ECA = 88.95 \text{ deg} = 1.552 \text{ rad}$$

$$\phi_\beta = 3.46 \int_{1.552}^{-1.552} \cos^2\theta \, d\theta$$

$$= 3.46 \left[\frac{1}{2}\theta + \frac{1}{4}\sin^2\theta \right]_{1.552}^{-1.552}$$

$$= 5.4 \text{ deg}$$

To correct the phase shift to the ratio of RCP to LCP

$$\left(\frac{1 + \tan\left(\frac{\phi}{2} + 45\right)}{1 - \tan\left(\frac{\phi}{2} + 45\right)} \right)^2 = \frac{RCP}{LCP} = 26.5 \text{ db}$$

which would be well below detectability. It is pointed out by Lusignan that we may assume either that the velocity and number of protons are equal to the velocity and number of electrons as above or we may assume, as an upper limit, that electron and proton energies are equal and that the proton and electron currents are equal so that the stream effects are proportional to N_β^2, and N is proportional to $1/\beta$ so that the stream effects are proportional to β. If we set proton and electron energies equal, then $\beta = 0.085$ and the previous value of $\beta = 0.00167$ gives a ratio of 50 for a total phase shift of approximately 275 deg, which would be a case where the energy has been completely converted to LCP and has gone slightly more than halfway back to RCP. In this case, RCP and LCP would be approximately the same magnitude.

The solar wind which has been assumed to be the sole cause of the depolarization, and the solar corona has been taken to be negligible because of the low β values in the corona at 4 solar radii.

References

1. Lusignan, B. B., "Detection of Solar Particle Streams by High Frequency Radio Waves," *Journal of Geophysical Research*, Vol. 68, No. 20, pp. 5617–5632, October 15, 1963.
2. Easterling, M., and Goldstein, R., *The Effect of the Interplanetary Medium on S-Band Telecommunications*, Technical Report No. 32-825, Jet Propulsion Laboratory, Pasadena, California, September 1, 1965.
3. Receiver and Spectrum Analyzer Sections of *The Superior Conjunction of Mariner IV*, JPL Technical Report (to be published).
4. Neugebauer, M., and Snyder, C. W., "*Mariner II* Observations of the Solar Wind, 1. Average Properties " (to be published).

XVIII. Communications Systems Research: Information Processing

A. Orthogonal Tree Codes

A. J. Viterbi

1. Introduction

Convolutional or tree codes first proposed by Elias (Ref. 1) have received considerable attention in the last several years, particularly in connection with sequential decoding algorithms (Refs. 2 and 3). We shall describe a particular convolutional encoder which generates an orthogonal tree code and a new *nonsequential* decoding algorithm which, for a given decoding complexity, provides performance superior to that of orthogonal block codes for the additive white gaussian channel.

2. Encoder

The encoder consists of a K-stage binary shift register (Fig. 1), all of whose stages are connected to a signal selector which stores or generates 2^K signal waveforms of duration $1/R$ sec. The input message is taken to be a binary data sequence fed into the shift register at a rate of R bits (binary symbols) per second. Every $1/R$ sec all the bits in the register are shifted one stage to the right,

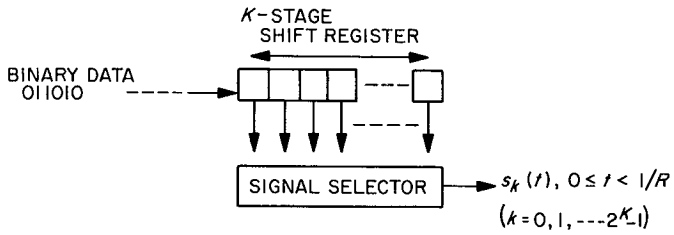


Fig. 1. Convolutional encoder for orthogonal tree code

the contents of the right-most stage is discarded and a new bit is shifted into the left-most stage. Immediately after each shift, the signal selector produces the signal waveform $s_k(t)$ of duration $1/R$ sec, where k is the integer between 0 and $2^K - 1$ whose binary expansion corresponds to the contents of the shift register, read from right to left, at the given time. We shall take the 2^K waveforms $s_k(t)$ to be orthogonal and of equal energy S/R. Thus

$$\int_0^{1/R} s_k(t)\, s_j(t)\, dt = (S/R)\, \delta_{kj}\,.$$

K is known as the *constraint length* of the code.

Assuming that the shift register initially contains all zeros, the possible sequences of signals generated may be described by following the paths on a binary tree (Fig. 2 for the case $K=3$). The convention followed is that when *zero* enters the shift register, we proceed along the upper branch while a *one* corresponds to the lower branch. It is clear that for the first K branches, two totally distinct paths (i.e., having no branches in common) produce two signals that are orthogonal over K/R sec. Thus for the example of Fig. 2, the data sequence 000 and 100 corresponding to totally distinct paths produces, respectively, the signal sequences $s_0(t)$, $s_0(t)$, $s_0(t)$ and $s_1(t)$, $s_2(t)$, $s_4(t)$, which are orthogonal. Of course, beyond the constraint length K the paths repeat themselves and orthogonality is lost. However, it should be noted that any two paths (or subpaths) whose generating data sequences are never identical over more than $K-1$ successive symbols are orthogonal over their entire lengths. This is the basis of the decoding algorithm described in the next section.

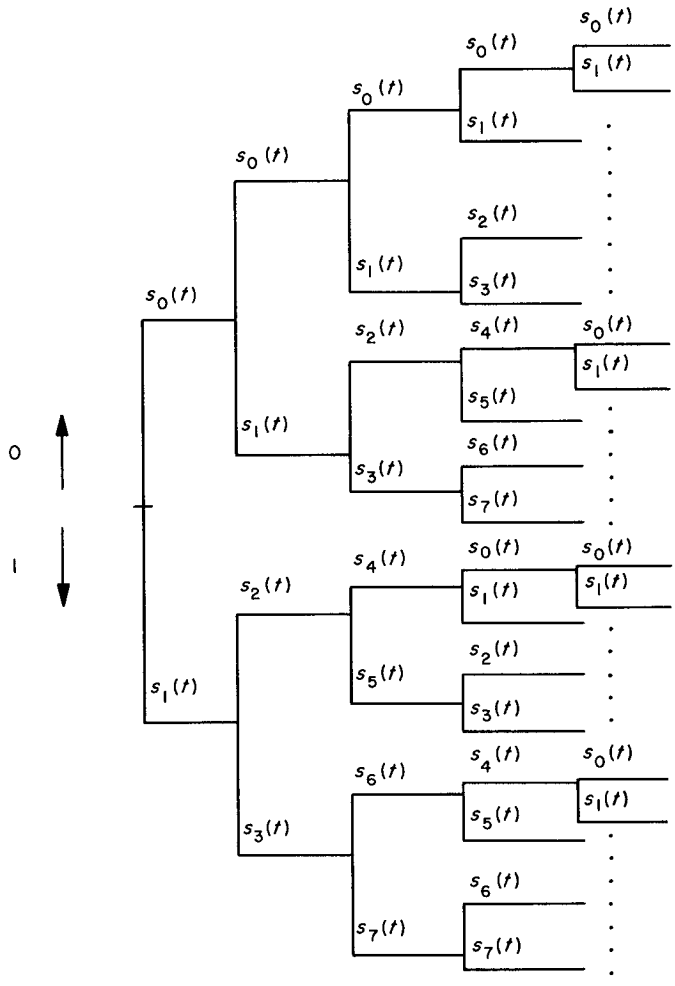


Fig. 2. Orthogonal tree code ($K = 3$)

After L bits of the data sequence have been fed to the shift register, a resynchronizing sequence of $K-1$ *zeros* is inserted, during which time the tree stops growing and after which the shift register has been reset to its initial state so that a new tree may be initiated. The sequence of L bits followed by $K-1$ *zeros* is known as a cycle and we shall be interested in the decoding error probability for a given cycle. Of course, the resynchronizing period reduces the effective transmission rate to

$$\tilde{R} = \frac{L}{L+K-1} R \text{ bits/sec.} \qquad (1)$$

3. A New Nonsequential Decoding Algorithm

The decoder must be capable of generating the 2^K integral inner products for each period of $1/R$ sec:

$$z_k^{(j)} = \int_{j/R}^{(j+1)/R} y(t)\, s_k(t)\, dt, \qquad k = 0, 1, 2, \cdots 2^K - 1$$

where

$$y(t) = s_i(t) + n(t), j/R \leq t < (j+1)/R$$

is the received waveform and n (t) is white gaussian noise of one-sided density N_0 w/cps. Also, the decoder must be capable of summing the $z_k^{(j)}$ to obtain the integral inner product of $y(t)$ with the signals of any path of the tree.

The decoding algorithm begins by computing the inner products of all 2^K paths for the first K branches, where K is the constraint (shift register) length. Denoting the paths by their generating data sequences, the decoder then performs pair-wise comparisons among path inner products for the 2^{K-1} pairs of K-branch paths as listed below:

000---00	and	100---00
000---01		100---01
000---10		100---10
⋮		⋮
011---11		111---11

We shall denote the path corresponding to the greater inner product in each comparison as the *survivor*. The algorithm then discards from further consideration all but the 2^{K-1} survivors of as many comparisons. It should be noted that there is a one-to-one correspondence between the 2^{K-1} survivors and the K-1-symbol binary sequences for the second through the Kth branches.

The algorithm then proceeds to compute the inner product for the next pair of branches for each of the 2^{K-1} surviving paths and to add to these the inner products for all preceding branches of the given path. It then compares the 2^K inner products pair-wise for the 2^{K-1} pairs of $(K+1)$-branch paths listed below:

$$\begin{array}{llll} x_{11}^{(1)} & 000\cdots00 & \text{and} \quad y_{11}^{(1)} & 100\cdots00 \\ x_{21}^{(1)} & 000\cdots01 & y_{21}^{(1)} & 100\cdots01 \\ \vdots & & \vdots & \\ x_{2^{K-1}}^{(1)} & 011\cdots11 & y_{2^{K-1},1}^{(1)} & 111\cdots11 \end{array}$$

where $x_{j1}^{(1)}$ and $y_{j1}^{(1)}$ depend on the outcome of the previous set of comparisons. Clearly there is a unique path in the above set for each combination of the last K data symbols. Consequently, all the paths considered have orthogonal signals in corresponding branches. Again only the 2^{K-1} survivors of the comparisons are preserved and the rest of the paths are discarded. The algorithm proceeds in this way, at each step eliminating half the surviving paths and then doubling again by considering the pair of branches emanating from each surviving path, so that the total population under consideration never exceeds 2^K. Also it should be noted that since the last K data symbols for each comparison are distinct, and these establish the signals $s_k(t)$ for the given branch under consideration, all paths compared have orthogonal signals in the given branch.

The procedure continues in this way until after the $(L-K+1)$th branch, after which the comparisons are among the pairs of paths

$$\begin{cases} x_{11}^{(L-K+1)}, x_{12}^{(L-K+1)} \cdots x_{1,L-K+1}^{(L-K+1)} \; 00\cdots000 \\ \text{and} \quad y_{11}^{(L-K+1)}, y_{12}^{(L-K+1)} \cdots y_{1,L-K+1}^{(L-K+1)} \quad 10\cdots000 \end{cases}$$

$$\begin{cases} x_{21}^{(L-K+1)}, x_{22}^{(L-K+1)} \cdots x_{2,L-K+1}^{(L-K+1)} \; 00\cdots010 \\ \vdots \\ \text{and} \quad y_{21}^{(L-K+1)}, y_{22}^{(L-K+1)} \cdots y_{2,L-K+1}^{(L-K+1)} \quad 10\cdots010 \\ \vdots \end{cases}$$

$$\begin{cases} x_{2^{K-2},1}^{(L-K+1)}, x_{2^{K-2},2}^{(L-K+1)} \cdots x_{2^{K-2},L-K+1}^{(L-K+1)} 01\cdots110 \\ \text{and} \quad y_{2^{K-2},1}^{(L-K+1)}, y_{2^{K-2},2}^{(L-K+1)}, \ldots, y_{2^{K-2},L-K+1}^{(L-K+1)} 11\cdots110 \end{cases}$$

where $x_{ij}^{(L-K+1)}$ and $y_{ij}^{(L-K+1)}$ depend on the outcome of the $(L-K+1)$th set of comparisons. Note, however, that there are now only 2^{K-1} paths under consideration because the last symbol must be a *zero*, corresponding to the first symbol of the $K-1$ symbol resynchronizing sequence. Thus there will be only 2^{K-2} survivors and since the next symbol must again be a *zero*, the tree fails to grow after each comparison. Therefore, just after the $(L-1)$th set of comparisons there will be only two survivors:

$$\begin{cases} x_{11}^{(L-1)}, x_{12}^{(L-1)} \cdots x_{1,L-1}^{(L-1)} \; 00\cdots00 \\ \text{and} \quad y_{11}^{(L-1)}, y_{12}^{(L-1)}, \ldots y_{1,L-1}^{(L-1)} \; 100\cdots00 \end{cases}$$

and the survivor of this last match will be accepted as the correct path.

The decoder just described for an orthogonal tree code of constraint length K is only slightly more complicated than the decoder for an orthogonal block code of block length K. Both systems require the capability to compute inner products between the received wareform and 2^K possible transmitted waveforms. However, the block decoder must make a comparison among 2^K waveforms every K bits, while the tree decoder must make comparisons after each bit; thus the decoding computation rate is K times as large as for block codes. Also, the bandwidth occupancy of the tree-coded waveforms is K times that of the block-coded waveforms. This follows from the fact that block-orthogonal signals need be orthogonal over the period K/R seconds (equal to the duration of K bits), while tree signals need be orthogonal over each bit period, $1/R$ seconds; the required time compression corresponds to a proportional frequency expansion. In the next two sections we shall show that improved performance may be worth this price.

4. An Upper Bound on the Error Probability

We now proceed to obtain an upper bound on the error probability, P_E, for orthogonal tree codes using the decoding algorithm described in the preceding section. We recall that any two divergent paths which are compared by the algorithm are orthogonal over all branches. An error will occur if the correct path is not a survivor in any one of the L comparisons in which it is involved.

Without loss of generality, we may take the transmitted data sequence to be the all-*zeros* sequence. Thus an error will be made in the first comparison if the inner product for the K-branch path $100\ldots0$ exceeds that for the

K-branch path 000 . . . 0. The total energy in the K-branch path waveforms is SK/R. Thus the probability of error in this comparison is

$$P_1\left(\frac{SK}{N_0R}\right) = \text{prob}\,(z_0 < z_1)$$

where z_0 and z_1 are unit variance independent gaussian variables with means $(SK/N_0R)^{1/2}$ and 0, respectively. In the second set of comparisons an error may occur if the $(K+1)$-branch path $y_{11}^{(1)}$ 100 . . . 0 yields an inner product greater than does the $(K+1)$-branch path 0000 . . . 0. However, $y_{11}^{(1)}$ may be either *zero* or *one*. If it is *zero*, the correct and incorrect path will be different over only K branches, while if it is *one* they will differ over $K+1$ branches. Thus using a union bound, we have that the probability of error in either the first or second comparisons is no greater than

$$2P_1\left(\frac{SK}{N_0R}\right) + P_1\left[\frac{S\,(K+1)}{N_0R}\right].$$

At the third step, an error may occur if the $(K+2)$-branch path $y_{11}^{(1)}\,y_{12}^{(1)}$ 10 . . . 0 exceeds the correct path 0000 . . . 0. Now if $y_{11}^{(1)}\,y_{12}^{(1)} = 00$, the incorrect path differs from the correct path in K branches; if $y_{11}^{(1)}\,y_{12}^{(1)} = 01$ it differs in $K+1$ branches; while if $y_{11}^{(1)}\,y_{12}^{(1)} = 10$ or 11 it differs in $K+2$ branches. Thus there is one potential adversary which differs in K branches, one which differs in $K+1$ branches, but *two* which differ in $K+2$ branches. Combinning these possibilities of making an error with those for the first and second step, we have, using a union bound, that the probability of error in any one of the first three steps

$$3P_1\left(\frac{SK}{N_0R}\right) + 2P_1\left[\frac{S\,(K+1)}{N_0R}\right] + P_2\left[\frac{S\,(K+2)}{N_0R}\right]$$

where

$$P_2\left[\frac{SJ}{N_0R}\right] = \text{prob}\left[z_0 < \max\,(z_1, z_2)\right]$$

and z_0, z_1, and z_2 are unit variance independent gaussian variables with means $(SJ/N_0R)^{1/2}$, 0, and 0 respectively. Continuing in this way we find that the over-all error probability in decoding a cycle of L bits followed by $K-1$ *zeros* is bounded by

$$P_E \leq LP_1\left(\frac{SK}{N_0R}\right) + \sum_{k=1}^{L-1}(L-k)\,P_{2^{k-1}}\left[\frac{S}{N_0R}\,(K+k)\right] \quad (2)$$

where

$$P_j\left(\frac{SJ}{N_0R}\right) = \text{prob}\left[z_0 < \max\,(z_1, z_2, \cdots z_j)\right] \quad (3)$$

and $z_0, z_1, \ldots, z_k$ are independent unit variance gaussian variables, z_0 has mean $(SJ/N_0R)^{1/2}$ and all the rest have *zero* means. The probability of Eq. (3) is just the probability of error for an orthogonal block code of j waveforms, for an energy-to-noise ratio equal to $(SJ/N_0R)^{1/2}$. It is shown in Ref. (2) that[1]

$$P_j\left(\frac{SJ}{N_0R}\right) < \exp\left[-\frac{SJ}{N_0R}\,E\,(R')\right] \quad (4)$$

where

$$R' = \frac{\log_2 j}{J}\,R,$$

$$E\,(R') = \begin{cases} \dfrac{1}{2} - \dfrac{R'}{C}, & 0 \leq \dfrac{R'}{C} \leq \dfrac{1}{4} \\ \left(1 - \sqrt{\dfrac{R'}{C}}\right)^2 & \dfrac{1}{4} \leq \dfrac{R'}{C} < 1 \end{cases}$$

and

$$C = S/(N_0 \ln 2)$$

As was noted in Section 2, since $K-1$ resynchronizing *zeros* are inserted between cycles of L bits, the effective rate $\tilde{R}$ is reduced by the factor $L/(L+K-1)$. Thus

$$\frac{S}{N_0\tilde{R}} = \frac{S\,(L+K-1)}{N_0R\,L} \quad (5)$$

is the effective energy-to-noise ratio per bit.

The upper bound on orthogonal tree codes, as a function of $S/(N_0\tilde{R})$, as determined from Eqs. (2), (4), and (5), is plotted (as a solid line) in Fig. 3 for constraint lengths $K = 20$, 20, and 30. ($L = 10K$ for $K = 10$ and $L = 20K$ for the other two.) For the sake of comparison, the bound on P_E for orthogonal block codes as computed from Eq. (4) is also shown in the figure (as a dashed line) for the block lengths $K = 10$, 20, 30. While the exact values of P_E for orthogonal block codes could be obtained from Ref. 4 for $K = 10$ and 20, the comparisons shown are more meaningful since the upper bound of Eq. (4) is used in the same way in both cases.

[1]The bound in Ref. 2 is double this, but a slightly more sensitive bounding argument yields Eq. (4).

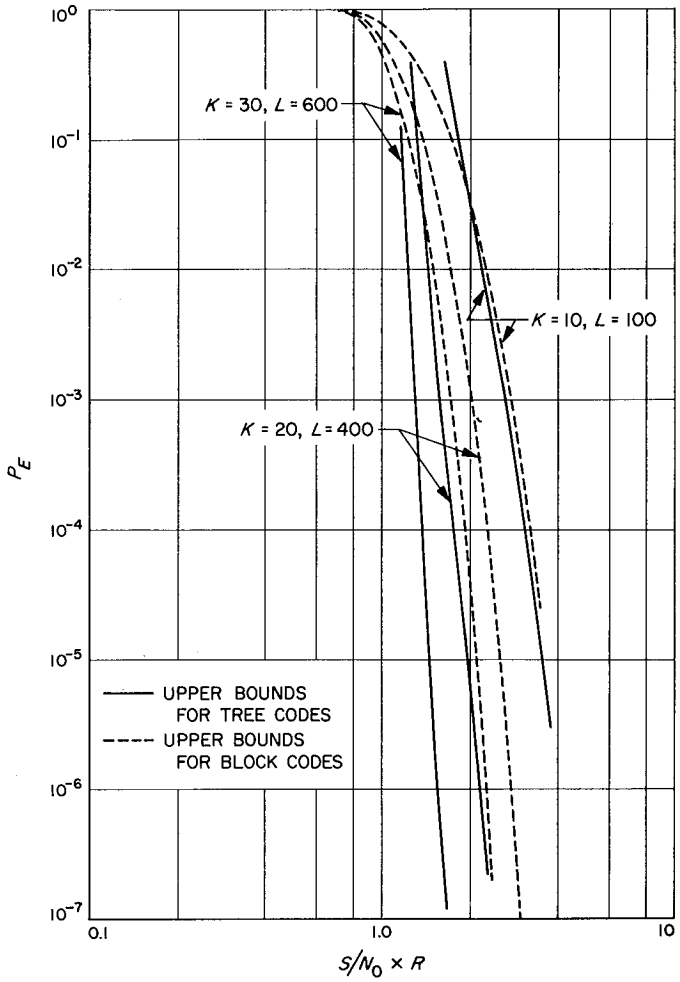


Fig. 3. Error bounds for orthogonal tree and block codes

5. An Asymptotic Result

A slightly weaker bound on P_E may be obtained directly from Eq. (2) by replacing $L-k$ by L in each term of the sum. Thus

$$P_E < L\left[P_1\left(\frac{SK}{N_0R}\right) + \sum_{k=1}^{L-1} P_{2^{k-1}}\frac{S}{N_0R}(K+k)\right]. \qquad (6)$$

It can also be shown that the probability expression, Eq. (3), can be bounded by

$$P_j\left(\frac{SJ}{N_0R}\right) < \exp\left[-\frac{SJ}{N_0R}\left(\frac{\rho}{1+\rho} - \frac{\rho R'}{C}\right)\right] \quad (0 \le \rho \le 1) \qquad (7)$$

where

$$R' = \frac{\log_2 j}{J} R \text{ and } C = S/(N_0 \ln 2).$$

Rather than giving a lengthy derivation of Eq. (7), we note that if we minimize Eq. (7) with respect to ρ (or maximize the exponent $[(\rho)/(1+\rho) - \rho R']$) we obtain Eq. (4), and consequently Eq. (7) being everywhere greater than or equal to Eq. (4) must also be a valid bound. Thus combining Eqs. (6) and (7), we have

$$\begin{aligned} P_E &< L\left\{\exp\left[-\frac{SK}{N_0R}\left(\frac{\rho}{1+\rho}\right)\right]\right. \\ &\left. + \sum_{k=1}^{L-1} \exp\left[-\frac{S}{N_0R}(K+k)\left(\frac{\rho}{1+\rho}\right) + \rho(k-1)\ln 2\right]\right\} \\ &< L\left[\exp -\frac{SK}{N_0R}\left(\frac{\rho}{1+\rho}\right)\right] \\ &+ \sum_{k=0}^{L-1} \exp\left\{-k\rho\left[\frac{S/(N_0R)}{1+\rho} - \ln 2\right]\right\} \\ &< \frac{L \exp\left[-\frac{SK}{N_0R}\left(\frac{\rho}{1+\rho}\right)\right]}{1 - \exp(-\varepsilon \ln 2/R)} \qquad (\sigma \le \rho \le 1) \end{aligned}$$

where

$$\frac{\varepsilon}{R} = \left(\frac{\rho}{1+\rho}\right)\frac{S/\ln 2}{N_0R}.$$

Since $C = S/(N_0 \ln 2)$ this may also be expressed as

$$P_E < \frac{L\,2^{-\frac{KC}{R}\left(\frac{\rho}{1+\rho}\right)}}{1 - 2^{-\varepsilon/R}} \qquad (0 \le \rho \le 1)$$

where

$$R = \frac{C}{1+\rho} - \varepsilon$$

Now, since for asymptotically large K, ε can be taken to be nearly zero, and still

$$E(R) \triangleq \frac{\log_2 \frac{1}{P_E}}{KC/R} = \frac{\rho}{1+\rho} \qquad (0 \le \rho \le 1)$$

Then, clearly, for $0 \le R \le (C/2) + \varepsilon$, we should choose $\rho = 1$ to maximize $E(R)$, while for higher rates we should take $\rho = C/(R+\varepsilon) - 1$, which yields $E(R) = (C - R - \varepsilon)/C$.

Thus we obtain

$$P_E < \frac{L\,2^{-KE(R)}}{1-2^{-\varepsilon/R}}$$

where

$$\frac{\log_2 \frac{1}{P_E}}{KC/R} \triangleq E(R) = \begin{cases} \frac{1}{2}, & 0 \le R \le \frac{C}{2} - \varepsilon \\ 1 - \frac{R+\varepsilon}{C}, & \frac{C}{2} - \varepsilon \le R < C \end{cases} \tag{8}$$

In comparison, for orthogonal block codes, it follows from Eq. (4) that for block length $K (J = K, j = 2^K)$,

$$P_E < \exp - \frac{KS}{N_0 R} E(R) \qquad = 2 \exp - \frac{KC}{R} E(R)$$

or

$$\frac{\log_2 \frac{1}{P_E}}{KC/R} \triangleq E(R) = \begin{cases} \frac{1}{2} - \frac{R}{C}, & 0 \le R \le C/4 \\ \left(1 - \sqrt{\frac{R}{C}}\right)^2 & C/4 \le R < \varepsilon \end{cases} \tag{9}$$

The results of Eq. (8) for $\varepsilon = 0$, and Eq. (9) are compared in Fig. 4.

These results can also be shown for the class of "very noisy" memoryless channels using more general convolutional codes (Ref. 5).

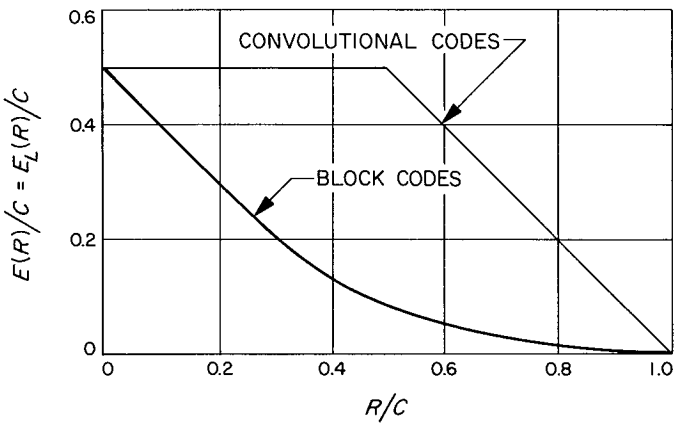


Fig. 4. Exponents of asymptotic error bounds for orthogonal tree and block codes

B. Properties of Groups of Collineations on a Certain Class of Codes

R. E. Block[2]

1. Introduction

This is a report of results on a type of code equivalence, namely collineations (*SPS 37-25*, Vol. IV, pp. 158–160, and Ref. 6), on certain, not necessarily linear, codes. The class of codes considered contains the important class of constant-distance codes (Ref. 6), and in particular includes the codes obtained by taking as code words the rows of the incidence matrix of a balanced incomplete block design (BIBD). Groups of collineations on the codes are considered—these are relevant in the search for codes and in the development of decoding schemes. The results are about the numbers of elements in the orbits of a group of collineations, and in particular about a comparison of these numbers for the orbits on the words and on the places for a code (or on the row indices and column indices for the corresponding matrix). For symmetric BIBD's (including projective planes) of order n, p-symmetry will be proved for any prime p not dividing n, i.e., the number of point (row or code-word) orbits of length divisible by any power of p is the same as for the block (column or place) orbits, and this will be generalized to nonsymmetric designs. There also will be given number-theoretic conditions on the orbit lengths, derived from a rational matrix congruence.

2. Tactical Decompositions and Groups of Collineations

Let M be a $v \times b$ matrix over a field. In particular, M may have as its rows the words of a binary code, written with ± 1, or with 1 and 0. One says that M has a *right tactical decomposition* if there is a partition of the set of row indices into disjoint classes R_i (the *row classes*) and of the set of column indices into disjoint classes C_i (the *column classes*) such that for every i and j, the submatrix of M formed by the entries with row index in R_i and column index in C_j has constant column sums s_{ij}.

The number of row and column classes will be denoted by t and t', respectively. The $t \times t'$ matrix $S = (s_{ij})$ is called the *associated matrix of column sums*. Similarly, a

[2]Consultant, Mathematics Department, University of Illinois, Urbana, Illinois.

left tactical decomposition is defined by replacing column sums by row sums, and a *tactical decomposition* is defined by requiring both conditions to hold.

An important type of tactical decomposition is obtained by taking as the row and column classes, the orbits of a *group of collineations* on M. Here a collineation means a pair π, σ of permutations, π acting on the row indices and σ on the column indices of M, such that $m_{ij} = m_{\pi(i),\sigma(j)}$ for all i, j. Thus, if the rows of M are the words of a code, a collineation is the same thing as a code equivalence.

We need a lemma which we state here.

Lemma 1. If M, of rank ρ, has a right (resp. left) tactical decomposition, then

$$t - (v - \rho) \leqq \rho_{cs} \leqq t' \text{ (resp. } t' - (b - \rho) \leqq \rho_{rs} \leqq t),$$

where ρ_{cs} (resp. ρ_{rs}) denotes the rank of the associated matrix S. In particular if the rows (resp. columns) of M are linearly independent, so are those of S.

This is a slight extension of Theorem 2 of *SPS 37-28*, Vol. IV, pp. 232–234, and follows from the proof given there.

3. An Application of an Integral Matrix Congruence

For any positive integer m, let I_m and J_m denote, respectively, the $m \times m$ identity matrix and the $m \times m$ matrix with all entries 1. For a right or left tactical decomposition of a matrix (or of an incidence structure) write $v_i\,(i = 1, \cdots, t)$ and $b_j\,(j = 1, \cdots, t')$ for the cardinalities of the ith row class and jth column class, respectively. Also write V for the $t \times t$ diagonal matrix diag $(v_1, \cdots, v_t)$ and B for the $t' \times t'$ diagonal matrix diag $(b_1, \cdots, b_{t'})$. Also S denotes the associated matrix of column or row sums.

A lemma is needed before the theorem.

Lemma 2. Suppose there is a right tactical decomposition of a $v \times b$ matrix $M = (m_{ij})$ over a field F, and suppose there are elements $\alpha \neq 0$ and β in F such that $MM' = \alpha I_v + \beta J_v$. Then

$$SBS' = \alpha VI_t + \beta VJ_tV. \tag{1}$$

The determinant of the right side W of Eq. (1) is

$$\prod_{i=1}^{t} v_i \alpha^{t-1} (\alpha + \beta v).$$

Moreover, S has rank at least $t - 1$, and rank exactly t provided $\alpha + \beta v \neq 0$. If the entries of M are integers and if $t = t'$, then

$$\alpha^{t-1} (\alpha + \beta v) \prod_{i=1}^{t} (v_i / b_i) = \det S^2 \tag{2}$$

and so is the square of an integer.

Proof. Counting

$$\sum_{l=1}^{b} \left(\sum_{k \in R_i} m_{kl} \right) \left(\sum_{k \in R_j} m_{kl} \right)$$

in two ways, one gets

$$\sum_{q=1}^{t'} b_q s_{iq} s_{jq} = v_i v_j \beta + \delta_{ij} v_i \alpha,$$

so that Eq. (1) holds. The determinant of W can be computed by subtracting v_j/v_i times the first column from the jth column, $j = 2, \cdots, t$, and then adding to the first row the rows after the first, thus making a triangular matrix with diagonal entries $\alpha v_1 + \beta v v_1, \alpha v_2, \cdots, \alpha v_t$. Since $\det MM' = \alpha^{v-1} (\alpha + \beta v)$, then if $\alpha + v\beta \neq 0$, the rank ρ of M is v, while if $\alpha + v\beta = 0$, then $\rho = v - 1$. The statements about the rank of S then follow from Lemma 1. The final statement of the lemma follows immediately from the first two conclusions.

Suppose that the hypotheses of Lemma 2 hold and that $\alpha + v\beta \neq 0$. Then S has rank t, and $t \leqq t'$. Take any t linearly independent columns of S; by reordering they may be assumed to be the first t columns. Let S_1 and S_2 denote the submatrices of S, consisting respectively of these first t columns and of the remaining $t' - t$ columns of S. Define $t' \times t'$ matrices S_0 and W_0 by

$$S_0 = \begin{bmatrix} S_1 & S_2 \\ 0 & I_{t'-t} \end{bmatrix}, \qquad W_0 = \begin{bmatrix} W & S_2 B_4 \\ B_4' S_2' & B_4 \end{bmatrix},$$

where B_4 is the $(t' - t)$, $(t' - t)$ diagonal matrix diag $(b_{t+1}, \cdots, b_{t'})$. We can now state Theorem 1.

Theorem 1. Under the hypotheses of Lemma 2, if $\alpha + v\beta \neq 0$ and S_0 and W_0 are constructed as above, then

$$S_0 B S_0' = W_0 , \tag{3}$$

and S_0 is nonsingular.

Proof. The nonsingularity of S_0 follows from its construction, and Eq. (3) follows from Eq. (1) by inspection. This completes the proof.

For a BIBD with the discrete tactical decomposition, i.e., with all v_i and b_j equal to 1, Eq. (3) becomes the congruence studied by Connor (Ref. 7).

For any prime p and any positive integer a, define $\phi_p(a)$ by writing

$$a = p^{\phi_p(a)} a^* ,$$

where a^* is an integer prime to p.

Lemma 3. Suppose that the hypotheses of Lemma 2 hold for an integral matrix M, and that $t = t'$. If p is a prime not dividing $\alpha(\alpha + \beta v)$ then p does not divide det S, and if an entry s_{ij} of S is not divisible by p then $\phi_p(b_j) \geqq \phi_p(v_i)$.

Proof. Let M_p and S_p be the matrices obtained from M and S by taking residues of the entries modulo p. Since $\alpha(\alpha + \beta v) \neq 0 \bmod p$, by Lemma 2 S_p has rank t and hence $p \nmid \det S$. Therefore $(S')^{-1}$ exists and is a rational matrix with denominators prime to p. By Eq. (1),

$$SB = (\alpha VI + \beta VJV)(S')^{-1} .$$

If $p^a | v_i$ then the ith row of the right side, and so of SB, is divisible by p^a. In particular, $p^a | s_{ij} b_j$, so that $p^a | b_j$. This gives the conclusion of the lemma.

For any set P of primes, two sequences $v_1, \cdots, v_t$ and $b_1, \cdots, b_t$ of positive integers will be called *P-symmetric* (p-symmetric if $P = \{p\}$) if the sequences can be reordered so that $\phi_p(v_i) = \phi_p(b_i)$ for $i = 1, \cdots, t$ and for every p in P. A right or left tactical decomposition with $t = t'$ will be called P-symmetric if P-symmetry holds for the corresponding sequences $v_1, \cdots, v_t$ and $b_1, \cdots, b_t$. The decomposition is called symmetric if it is P-symmetric for all P, that is, if $t = t'$ and the b_i's and v_i's can be reordered so that $b_i = v_i$, $i = 1, \cdots, t$.

Theorem 2. A tactical decomposition of a symmetric BIBD is $\{p\} \cup Q$-symmetric for any prime p not dividing $n = k - \lambda$ and any set Q of primes each greater than k.

Proof. The incidence matrix satisfies the hypotheses of Lemma 2 with $\alpha = n$ and $\beta = \lambda$, so that $\alpha + \beta v = k^2$. First suppose $p \nmid \det S$. Hence, there is a transversal of S of entries not divisible by p, and by reordering the classes we may suppose that this is the diagonal, that is, $p \nmid s_{ii}$, $i = 1, \cdots, t$. Then $q \nmid s_{ii}$ for any q in Q since $1 \leqq s_{ii} \leqq k$. Write $P = \{p\} \cup Q$. For each a in P and for $i = 1, \cdots, t$, $\phi_a(b_i) \geqq \phi_a(v_i)$ by Lemma 3, and hence $\phi_a(b_i) = \phi_a(v_i)$ by Eq. (2). This gives P-symmetry in this case.

Next suppose that $p | k$ but that $p \nmid n$. The complementary design has parameters $v, k' = v - k$, $\lambda' = v - 2k + \lambda$, and $n' = k' - \lambda' = k - \lambda = n$, and its incidence matrix has a tactical decomposition with the same row and column classes and with $s'_{ij} = v_i - s_{ij}$. Also $p \nmid n'k'$ since otherwise $p | v - k$, $p | v, p | \lambda$ since $p | k(k-1) = \lambda(v-1)$, and $p | k - \lambda = n$, a contradiction (this already proves p-symmetry). Therefore one can, as before, assume that $p \nmid s'_{ii}$, $i = 1, \cdots, t$. If j is such that $p | v_j$, then $p \nmid s_{jj} = v_j - s'_{jj}$ and $s_{jj} \neq 0$, so that no q in Q divides s_{jj}. Thus, by Lemma 3, if a is any power of an element of P, then

$$\{i : a | v_i \text{ and } p | v_i\} \subseteq \{i : a | b_i \text{ and } p | v_i\} .$$

Consideration of the left, as well as the right, version of the tactical decomposition then shows that these two sets have the same number of elements and so are equal. Thus, P-symmetry and hence also Q-symmetry hold for the subsequences of those v_i and those b_i which are divisible by p. By the first case of the proof the decomposition is Q-symmetric, and so Q-symmetry holds for the complements of the above subsequences. This gives P-symmetry for the decomposition, and the proof is complete.

This theorem generalizes results of Dembowski (Ref. 8) who proved that if p is a prime not dividing nk then the sets $\{i : p | v_i\}$ and $\{i : p | b_i\}$ have the same cardinality, and that for a p-group of collineations if $p \nmid n$ and $\lambda = 1$, then the group fixes the same number of points and lines. The theorem is also related to a work of Roth (Ref. 9) which shows that on certain planes of order n solvable collineation groups of order prime to n fix the same number of points and lines.

The result of Theorem 2 does not hold without the restriction on p dividing n — in fact the four-group acts in a nonsymmetric manner on the projective plane of order 2.

Corollary 1. Let G be a group of a collineations on a symmetric BIBD, p a prime not dividing n, Q a set of primes each greater than k, and H a normal subgroup of G, all of whose orbits have the same number m of elements. Then m divides each v_i and b_i and the sequences of quotients $v_1/m, \cdots, v_t/m$ and $b_1/m, \cdots, b_t/m$ are $\{p\} \cup Q$-symmetric. In particular, the orbits of G give a tactical decomposition which is symmetric if every prime dividing the order of G/H is in $\{p\} \cup Q$.

The proof is omitted.

Corollary 2. Let S be the associated matrix of column sums of a tactical decomposition on the incidence matrix M of a symmetric BIBD. Then $(\det S)/k$ is an integer of which every prime factor divides n, and

$$n^{t-1} \prod_{i=1}^{t} (v_i/b_i) = (\det S)^2/k^2,$$

the square of an integer. Furthermore, if $t \leq (v+1)/2$ then $\det S \mid \det M$.

The proof is omitted.

The first part of the proof of Theorem 2 actually establishes the following.

Corollary 3. Let there be given a right tactical decomposition with $t = t'$ on an integral $v \times b$ matrix M such that $MM' = \alpha I + \beta J$. If p is a prime not dividing $\alpha(\alpha + \beta v)$ then the decomposition is p-symmetric. If P is a set of primes, none of which divide $\alpha(\alpha + \beta v)$ and if there is a transversal $s_{ij_1}, \ldots, s_{tj_t}$ of the associated matrix S such that each s_{ij_i} is prime to every element of P, then the decomposition is P-symmetric.

This applies in particular to BIBD. For a BIBD with parameters v, b, r, k, λ, the incidence matrix M satisfies $MM' = \alpha I + \beta J$ with $\alpha = n = r - \lambda$, $\beta = \lambda$ and $\alpha + \beta v = r - \lambda + v\lambda = rk$. Replacing 1 and 0 in M by γ and δ respectively, one obtains a matrix A such that $AA' = (r' - \lambda')I + \lambda' J$ where $r' = r\gamma^2 + (b - r)\delta^2$ and $\lambda' = \lambda\gamma^2 + 2(r - \lambda)\gamma\delta + (b - 2r + \lambda)\delta^2$. A straightforward computation shows that $r' - \lambda' = n(\gamma - \delta)^2$ and $r' - \lambda' + v\lambda' = [k\gamma + (v - k)\delta][r\gamma + (b - r)\delta]$.

Corollary 4. Suppose there is given a right tactical decomposition with $t = t'$ on a BIBD, and a prime p not dividing $r - \lambda$. If $p \nmid (r, b)(k, v)$ or, in case $p = 2$, if $p \nmid (rk, (b - r)(v - k))$, then the decomposition is p-symmetric. If $p \nmid rk$ and if Q is a set of primes such that for every q in Q, $q > k$ and $q \nmid (r - \lambda)r$ then the decomposition is $\{p\} \cup Q$-symmetric.

Proof. When (γ, δ) has the value $(1, 0)$, $(0, 1)$ or $(-1, 1)$, then $r' - \lambda' + v\lambda'$ has the value rk, $(b - r)(v - k)$, or $(b - 2r)(v - 2k)$, respectively. The conditions on the greatest common denominator's guarantee that one of these three integers is not divisible by p [the stronger condition when $p = 2$ is needed because p must not divide $(r - \lambda)(\lambda - \delta)^2$, so that then (γ, δ) must not be $(-1, 1)$]. The first conclusion then follows from Corollary 3. As in the proof of Theorem 2, the hypotheses of the final statement of the corollary imply those of the last statement of Corollary 3, which then gives the present result.

4. An Application of the Theory of Quadratic Forms

Consider a symmetric BIBD with a tactical decomposition. Eq. (1) says that B and $V(\lambda J)V + nV$ are rationally congruent. Using this fact, Hughes (Ref. 10) and Dembowski (Ref. 8) applied the Hasse-Minkowski theory of rational congruence of quadratic forms to obtain number-theoretic conditions on the v_i's (or b_i's) for certain special symmetric decompositions. Lenz (Ref. 11) gave a simple proof that the above congruence implies the rational congruence of the $(t+1) \times (t+1)$ diagonal matrices $(b_1, \cdots, b_t, n\lambda)$ and $(nv_1, \cdots, nv_t, \lambda)$ (actually Lenz only stated this for projective planes) and used this in the case of symmetric decompositions to obtain a generalization of the results of Hughes and Dembowski. In the following, an application of the Hasse-Minkowski theory to the congruence of Lenz and of the symmetry results of the preceding section gives an extension to number-theoretic conditions for nonsymmetric decompositions. The symbols $(a, c)_p$ and (n/p) denote the Hilbert norm residue and the Legendre (quadratic residue) symbols.

Theorem 3. Suppose there is given a tactical decomposition on a symmetric BIBD. For every prime p, if t is odd then

$$\left((-1)^{(t-1)/2}\lambda \prod_{i=1}^{t} b_i, n\right)_p \prod_{1 \leq j < l \leq t} (v_j, v_l)_p (b_j, b_l)_p = 1; \tag{4}$$

if t is even, then

$$\left((-1)^{(t/2)+1}\lambda, n\right)_p \prod_{1 \leq j < l \leq t} (v_j, v_l)_p (b_j, b_l)_p = 1. \tag{5}$$

If $p \nmid n$, these both reduce to

$$\left(\frac{n}{p}\right)^{a} \prod_{\{j|\phi_p(v_j)\ \text{odd}\}} (p, v_j)_p \prod_{\{j|\phi_p(b_j)\ \text{odd}\}} (p, b_j)_p = 1, \qquad (6)$$

where a is 1 or 0 according as $\{j|\phi_p(v_j)\ \text{odd}\}$ (or $\{j|\phi_p(b_j)$ odd$\}$) has an odd or even number of elements.

The proof is obtained by computing the Hasse invariants of both sides of the congruence of Lenz; the details are omitted.

Corollary 5. Let G be a group of collineations on a symmetric BIBD and p a prime such that $p \nmid n$ and $(n/p) = -1$. If $p \leq k$ (respectively, $p > k$) suppose that $(q/p) = 1$ for every prime divisor $q \neq p$ of $|G|$ such that $q \leq k$ (respectively, such that $q|n$). Then $\phi_p(v_i)$ is odd for an even number of point orbits.

The proof is short and omitted. The hypothesis that $(q/p) = 1$ is only needed for those primes q such that the decomposition is not $\{p, q\}$-symmetric.

5. A Generalization of p-Symmetry to Nonsymmetric Designs

Suppose there is given a right tactical decomposition of a matrix with, as usual, row classes with $v_1, \cdots, v_t$ elements and column classes with $b_1, \ldots, b_{t'}$, elements. In this section, for any prime p and nonnegative integer j, we write p_j for the number of i with $\phi_p(v_i) = j$, and p'_j for the number of i with $\phi_p(b_i) = j$. Thus, p-symmetry is equivalent to $p_j = p'_j$ for all j. We now consider the following theorem for right tactical decompositions.

Theorem 4. Let there be given a right tactical decomposition of a $v \times b$ integral matrix M, where $MM' = \alpha I + \beta J$, and a prime p not dividing α. For any i, if $p \nmid (\alpha + \beta v)$ then

$$0 \leq (i+1)(p'_0 - p_0) + i(p'_1 - p_1) + \cdots + (p'_i - p_i), \qquad (7)$$

while if $p|(\alpha + \beta v)$, then

$$-(i+1) \leq (i+1)(p'_0 - p_0) + i(p'_1 - p_1) + \cdots + (p'_i - p_i). \qquad (8)$$

Proof. For any matrix A and any row indices $i_1, \cdots, i_m$ and column indices $j_1, \cdots, j_m$ let $A(i_1 \cdots, i_m; j_1, \cdots, j_m)$ denote the $m \times m$ submatrix of A formed from the given rows and columns. By the elementary expression for the determinant of the product of an $m \times l$ and an $l \times m$ matrix as a sum of products of $m \times m$ minors, Eq. (1) implies that

$$\sum_{1 \leq j_1 < j_2 < \cdots < j_m \leq t'} b_{j_1} \cdots b_{j_m} \times [\det S(i_1, \cdots, i_m; j_1, \cdots, j_m)]^2 = \det W(i_1, \cdots, i_m; i_1, \cdots, i_m)$$

for any $i_1, \cdots, i_m$ with $1 \leq i_1 < i_2 < \cdots < i_m \leq t$. A calculation just like that of $\det W$ shows that the right side equals

$$v_{i_1} \cdots v_{i_m} \alpha^{m-1} [\alpha + \beta(v_{i_1} + \cdots + v_{i_m})] = x, \text{ say.} \qquad (9)$$

Hence, if $s = \phi_p(x)$, then $\phi_p(b_{j_1} \cdots b_{j_m}) \leq s$ for some choice of $j_1, \cdots, j_m$. Applying this fact to the $m = p_0 + \cdots + p_i$ of the row indices l with $\phi_p(v_l) \leq i$, one sees that if γ is defined by

$$\gamma + \sum_{j=0}^{i} p'_j = \sum_{j=0} p_j$$

then

$$(i+1)\gamma + \sum_{j=0} j p'_j \leq \sum_{j=0} j p_j + \phi_p(\alpha)\left(-1 + \sum_{j=0} p_j\right) + \phi_p\left(\alpha + \beta \sum_{\phi_p(v_j) \leq i} v_j\right). \qquad (10)$$

By hypothesis $\phi_p(\alpha) = 0$. Moreover, the last term vanishes if $p \nmid (\alpha + \beta v)$ because $p|\beta \sum v_j$, where the sum is over those j such that $\phi_p(v_j) > i$. Hence, multiplying the equation defining γ by $i + 1$ and then subtracting Eq. (10), one gets Eq. (7).

Next, suppose that $p|(\alpha + \beta v)$. Then $p \nmid v\beta$ since $p \nmid \alpha$. Suppose the last term of Eq. (10) does not vanish. Pick an index l such that $\phi_p(v_l) = 0$ (such an index exists since $p \nmid v$) and apply the above argument this time to the $p_0 + \cdots + p_i - 1$ indices j such that $\phi_p(v_j) \leq i, j \neq l$. Then p_0 is replaced by $p_0 - 1$, and the term corresponding to the last term of Eq. (10) now vanishes, since $p \nmid \beta v_l$. Hence, Eq. (8) holds, and the proof is complete.

This theorem gives a second proof of the p-symmetry of a tactical decomposition of a symmetric BIBD when $p \nmid n$.

For a right tactical decomposition and any integer a, let g_a denote the number of i for which $a|b_i$, and h_a the number of i for which $a|v_i$. Thus

$$g_{p^j} = \sum_{i \geq j} p'_i, h_{p^j} = \sum_{i \geq j} p_i, g_{p^j} - h_{p^j} = t' - t - \sum_{i=0}^{j-1} (p'_i - p_i).$$

We can now state Theorem 5.

Theorem 5. Let M be a $v \times b$ matrix over the integers such that $MM' = \alpha I + \beta J$. Suppose that M has a right tactical decomposition and that p is a prime not dividing α. Suppose that $j \geqq 0$ is such that $g_{p^j} = t'$. Then

(a) if $p \nmid (\alpha + \beta v)$ then

$$|g_{p^{j+1}} - h_{p^{j+1}}| \leqq t' - t, \text{ and } h_{p^j} = t; \qquad (11)$$

(b) if $p|(\alpha + \beta v)$ then $|g_{p^{j+1}} - h_{p^{j+1}}| \leqq t' - t + 1$, $h_{p^j} = t$ (when $j = 0$) or $t - 1$ (when $j > 0$), and $p^j | \alpha + \beta v$.

(In the most significant case $j = 0$ and the conditions $g_1 = t'$ and $h_1 = t$ are automatically satisfied).

Proof. Suppose first that $p \nmid (\alpha + \beta v)$ and that $j = 0$. Reduction of Eq. (1) modulo p gives $S_p B_p S'_p = W_p$ for the matrices of residue classes. The rank of S_p is t (by Lemma 3); the rank of B_p is $t' - g_p$; and the rank of W_p is $t - h_p$, as can be seen using Eq. (9). Since rank $B_p \geqq$ rank $W_p = S_p B_p S'_p$, one has $t' - g_p \geqq t - h_p \leqq t - g_p - (t' - t)$, and (a) of Eq. (11) follows in this case ($j = 0$). Next suppose that $j > 0$ and that p^j divides all v_i (as well as all b_i). Then $p^{-j}B$ and $p^{-j}W$ are integral. The analog of the above argument for the case $j = 0$, applied to the equation $S_p (p^{-j}B)_p S'_p = (p^{-j}W)_p$, yields the inequality of (a). This inequality and induction then show that p^j divides all v_i whenever it divides all b_i, which completes the proof of (a).

The proof of (b) is obtained by modifying the proof of (a), in particular noting that $(p^{-j}W)_p$ has rank $t - h_{p^{j+1}} - 1$; we omit the details.

Half of the inequality of (a), as well as of (b) when $j = 0$, also follows from Theorem 4.

The following holds by the reasoning of Corollary 4.

Corollary 6. Suppose there is given a right tactical decomposition on a BIBD, and a prime p such that $p \nmid (r - \lambda)(r, b)(k, v)$, or, in case $p = 2$, $p \nmid (r - \lambda)(r, k(b - r)(v - k))$. Then Eq. (7) holds, and if $g_{p^j} = t'$ then Eq. (11) holds.

It is to be hoped that the results and techniques of this article will yield new classes of codes and designs with useful symmetry properties, and also show that for some parameter values such codes and designs can not exist.

References

1. Elias, P., "Coding for Noisy Channels," *IRE Convention Record*, Part 4, pp. 37–46, 1955.
2. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, John Wiley & Sons, Inc., New York, 1965.
3. Savage, J. E., "Sequential Decoding — The Computation Problem," *System Technical Journal*, Vol. 45, pp. 149–175, January 1966.
4. Viterbi, A. J., "On Coded Phase-Coherent Communications," *IRE Transactions on SET*, Vol. S7, No. 1, pp. 3–14, March 1961.
5. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm " (to be published).

References (Cont'd)

6. Block, R. E., "Transitive Groups of Collineations on Certain Designs," *Pacific Journal of Mathematics,* Vol. 15, pp. 13–18, 1965.

7. Connor, W. S., "On the Structure of Balanced Incomplete Block Designs," *Annals of Mathematical Statistics,* Vol. 23, pp. 57–71, 1952.

8. Dembowski, P., "Verallgemeinerunger von Transitivitätsklassen endlicher projektiver Ebenen," *Mathematische Zeitschrift,* Vol. 69, pp. 59–89, 1958.

9. Roth, R., "Collineation Groups of Finite Projective Planes," *Mathematische Zeitschrift,* Vol. 83, pp. 409–421, 1964.

10. Hughes, D. R., "Collineations and Generalized Incidence Matrices," *Transactions of the American Mathematical Society,* Vol. 86, pp. 284–296, 1957.

11. Lenz, H., "Quadratische Formen und Kollineationsgruppen," *Archiv der Mathematik,* Vol. 13, pp. 110–119, 1962.

XIX. Communications Systems Research: Combinatorial Communications

A. Asymptotic Algorithmic Complexity

L. R. Welch[1]

1. Introduction

In *SPS 37-34*, Vol. IV, pp. 298–305, and *SPS 37-35*, Vol. IV, pp. 292–304, the subject of algorithmic complexity was investigated, and upper and lower bounds were obtained for the maximum complexity for functions in a certain class. This article improves the lower bound by the same methods used in *SPS 37-34*, Vol. IV, pp. 298–305, and reports an upper bound from the published literature. The significance of these new bounds is that they allow a determination of maximum complexity within a factor of eight.

In addition, upper and lower bounds are obtained for maximum complexity of linear transformations over the field of two elements.

[1]Consultant from University of Southern California Electrical Engineering Department, Los Angeles, California.

2. Definitions

Let $V_2^k\, d$ be the set of k-tuples with 0's or 1's as components.

Let $f_{k,n}$ be the set of all mappings from V_2^k to V_2^n and let $L_{k,n}$ be the set of linear mappings.

Let $\mathcal{M}_f$ be the set of (2,2) automata without feedback which computes f. (For a definition of (2,2) automata, see *SPS 37-34*, Vol. IV, pp. 298–305.)

Let $N(M)$ be the number of stages in the automata M, and define

$$N_f = \min_{M \epsilon \mathcal{M}_f} N(M)$$

Finally, define $B_{k,n}$ by the equation:

$$B_{k,n} = \max_{f \epsilon F_{k,n}} N_f$$

3. Derivation of a Lower Bound

Theorem 1: For all n and k such that $n \cdot 2^k > 300$, the inequality

$$B_{k,n} \geq \frac{n \cdot 2^k}{2 \log_2(n \cdot 2^k)} \text{ is valid.}$$

The derivation of this lower bound will follow that of Section 2 in *SPS 37-34*, Vol. IV, pp. 298–305. In there, Stanley observes that if

$$\sum_{i=n+k}^{N} [8(i-1)(i-2)]^i < 2^{n \cdot 2^k} \tag{1}$$

then there is a mapping $f: V_2^k \rightarrow V_2^n$ such that $N_f > N$. He also observes that

$$\sum_{i=n+k}^{N} [8(i-1)(i-2)]^i \leq 2 \cdot 8^N (N)^{2N} = 2^{3N+1+2N \log N} \tag{2}$$

where logarithms are to the base 2.

Now let N be the largest integer such that

$$2^{3N+1+2N \log N} < 2^{n \cdot 2^k} \tag{3}$$

Since N is maximal, it follows that

$$2^{3(N+1)+1+2(N+1)\log(N+1)} \geq 2^{n \cdot 2^k} \tag{4}$$

Since $\log_2 x$ is a monotone increasing function we may take the $\log_2$ $\log_2$ of each side of inequality (3).

$$\log[3N+1+2N \log N] < \log(n \cdot 2^k) \tag{5}$$

If $n \cdot 2^k \geq 5$, then both sides of inequality (5) are positive, and we have

$$\frac{1}{\log N + \log(2 \log N + 3 + 1/N)} > \frac{1}{\log(n \cdot 2^k)} \tag{6}$$

Multiplying inequality (6) by the $\log_2$ of inequality (4) gives

$$2N \left[\frac{(1+1/N)\log(N+1) + 3/2 + 2/N}{\log N + \log(2 \log N + 3 + 1/N)}\right] > \frac{n \cdot 2^k}{\log(n \cdot 2^k)} \tag{7}$$

An analysis of the bracketed quantity shows it to be less than 1 when $N > 17$. Therefore,

$$N > \frac{n \cdot 2^k}{2 \log_2(n \cdot 2^k)} \tag{8}$$

when $N > 17$ or $n \cdot 2^k > 300$. Since N is a lower bound for $B_{k,n}$, the theorem follows.

4. Derivation of an Upper Bound

In Ref. 1, G. N. Povarov proves that any n functions of k variables may be realized by a device with

$$N(M) = 2n(2^{k-m}-1) + 2^{2^m} - \sum_{i=0}^{m-1} 2^{2^i} + m - 2 \tag{9}$$

components. Here, m is an arbitrary positive integer less than k, and the components are two-state devices with two binary inputs. This result is used in the next theorem.

Theorem 2: Given $1 > \epsilon > 0$, for sufficiently large k if $n < 2^{2^k} - k$ then

$$B_{k,n} \leq 4(1+5\epsilon) \cdot \frac{n \cdot 2^k}{\log_2(n \cdot 2^k)}$$

Proof: It follows from Eq. (9) that

$$B_{k,n} \leq N(M) < 2n \cdot 2^{k-m} + 2^{2^m} \tag{10}$$

Let m be the integral part and δ the fractional part of $\log_2 \log_2(n \cdot 2^k) - \epsilon$. Then m will be less than k, provided n is less than 2^{2^k-k} and inequality (10) becomes

$$B_{k,n} \leq 2\left[\frac{n \cdot 2^k \, 2^{\delta+\varepsilon}}{\log_2(n \cdot 2^k)} + (n \cdot 2^k)^{2^{-\delta-\varepsilon}}\right] \tag{11}$$

For sufficiently large k (depending only on ϵ) the second term is less than ϵ times the first. Since $2^\delta < 2$ and $2^\varepsilon < 1 + 2\epsilon$ when $1 > \epsilon > 0$, inequality (11) reduces to

$$B_{k,n} \leq 2\left[\frac{n \cdot 2^k}{\log(n \cdot 2^k)} 2(1+2\epsilon)(1+\epsilon)\right] \tag{12}$$

and the theorem follows.

5. Linear Mappings

We may define the maximum complexity of *linear* mappings by

$$C_{k,n} = \max_{f \epsilon L_{k,n}} N_f \tag{13}$$

By using the same idea used in Theorem 1 of *SPS 37-34*, Vol. IV, pp. 298–305, and in Theorem 1 of this article a lower bound for $C_{k,n}$ can be obtained. A modification of the idea that Povarov used for general mappings yields an upper bound for $C_{k,n}$.

Theorem 3: For sufficiently large k,

$$C_{k,n} \geq \frac{nk}{2 \log_2 n \cdot k}$$

Proof: The expression

$$P(N) = \sum_{i=n+k}^{N} [8(i-1)(i-2)]^i$$

is an upper bound to the number of machines of complexity N or less. There are $2^{n \cdot k}$ linear mappings in $L_{k,n}$. Therefore if $P(N_0) < 2^{n \cdot k}$, there is a linear mapping f in $L_{k,n}$ with

$$N_f > N_0 .$$

The rest of the proof is a duplication of inequalities (2) through (8) with 2^k replaced by k and will not be shown here.

Theorem 4: When n and $k/\log_2 n$ are sufficiently large,

$$C_{k,n} \leq \frac{2n \cdot k}{\log_2 n}(1 + 3\epsilon)$$

In order to describe the automata whose complexity give the upper bound we need the following lemma.

Lemma 1: The set of all linear functions of m variables can be realized with an automata with 2^m binary stages each with two or fewer inputs.

Proof of lemma: The proof is derived by induction.

(1) Suppose $m = 1$. A stage with no inputs produces the zero function, and a stage with one input connected to x_1 produces the other linear mapping.

(2) Suppose the hypothesis true for $m-1$. Then there is an automata with 2^{m-1} stages and all linear functions of $m-1$ variables as outputs. Adjoin 2^{m-1} stages each with x_m as one input and an output of the submachine as the other input. The combining function is mod 2 addition. Clearly, the enlarged machine has all 2^m linear functions of m variables as outputs uses 2^m stages.

The proof of the theorem can now proceed. For any given m, group the k variables into sets of size m with the last group of size $r = k - m\,[k/m]$ (brackets denoting integer part). For each set form all linear functions using

$$\frac{k-r}{m} 2^m + 2^r$$

stages. Now each of the n output functions is a linear combination of $(k-r)/m+1$ functions. This combination can be constructed with $(k-r)/m$ stages. The total number of stages is then

$$N(M) = \frac{n \cdot (k-r)}{m} + \frac{k-r}{m} 2^m + 2^r \qquad (14)$$

Next, set m equal to the integer part and δ equal to the fractional part of $\log_2 n$. Eq. (14) becomes

$$N(M) \leq \frac{nk}{\log_2 n}\left(\frac{\log_2 n}{\log_2 n - \delta}\right)\left[1 + 2^{-\delta}\right] + n \qquad (15)$$

For large $k/\log_2 n$, the last term is less than ϵ times the first term, and for large n the term in parentheses is less than $1 + \epsilon$. Therefore we have

$$N(M) \leq \frac{n \cdot k}{\log_2(n)} 2(1+\epsilon)(1+\epsilon) \qquad (16)$$

Since $N(M)$ is an upper bound for $C_{k,n}$, the theorem follows.

6. Concluding Remarks

The upper and lower bounds in Theorems 1 and 2 differ essentially by a factor of eight, and therefore the maximum complexity can be determined to within this order of magnitude. In the case of linear mappings the ratio of the two bounds we have obtained grows arbitrarily large. For small n and large k the lower bound becomes unrealistic, since it becomes a small fraction of k, while $k-1$ stages are needed for the mapping.

$$y_1 = \sum_{i=1}^{k} x_i .$$

Also, the upper bound is unrealistic for small n and large k. It can be shown that

$$K = k + \frac{2^n \cdot n}{\log n}$$

is an upper bound.

In the case where $n = k$, the two bounds differ essentially by a factor of eight

$$\frac{k^2}{4 \log_2 k} \leq C_{k,n} \leq \frac{2k^2}{\log_2 k}(1+3\epsilon).$$

B. Solutions of Algebraic Equations Over Fields of Characteristic 2

E. Berlekamp[2], *H. Rumsey, G. Solomon*

This article gives new fast methods for decoding certain error-correcting codes by solving certain algebraic equations. The locations of a binary Bose-Chaudhuri-Hocquenghem code (Ref. 2) are associated with the elements of a Galois field, $GF(2^k)$. The code is designed in such a way that the power-sum symmetric functions of the error locations can be obtained directly by computing appropriately chosen parity checks on the received word. Various methods (Refs. 2 and 3) are known by which the elementary symmetric functions of these error locations can be computed from the power-sum symmetric functions. These elementary symmetric functions, $\sigma_1, \sigma_2, \cdots, \sigma_t$ are the coefficients of an algebraic equation whose roots are the error locations:

$$x^t + \sigma_1 x^{t-1} + \sigma_2 x^{t-2} + \cdots + \sigma_t = 0 .$$

Previous methods for finding the roots of this equation have required an exhaustive search of all the elements in $GF(2^k)$. We present here a greatly improved procedure for extracting the roots of algebraic equations of degrees 2, 3, and 4, along with mechanization procedures for these and higher degree equations. We present the quadratic equation in Section 1 and the cubic, in Section 2. In Sections 3 and 4 (which may be read independently of Section 2) we give mechanization procedures for the quadratic, cubic and quartic. In Section 5 we present a new mechanization of higher order equations. For large k, the procedures presented here for $t=2, 3$, and 4 are superior to previously known methods in all respects. For $t \geq 5$, our procedure requires more storage than Chien's procedure (Ref. 4), but it is considerably faster.

1. Solution of the Quadratic in Fields of Characteristic 2

Let

$$x^2 + \sigma_1 x + \sigma_2 = 0 \quad , \quad \sigma_1, \sigma_2 \in GF(2^k) ; \sigma_1 \neq 0 .$$

Theorem 1: A necessary and sufficient condition for Eq. (1) to have solutions in $GF(2^k)$ is that

$$Tr\, \sigma_2 \sigma_1^{-2} = 0 , \tag{1}$$

where $TR\, \alpha = \alpha + \alpha^2 + \alpha^4 + \cdots + \alpha^{2^{k-1}}$.

Proof: One first notes that $TR\alpha = Tr(\alpha^2)$ and $TR(\alpha + \beta) = Tr\, \alpha + Tr\, \beta$. Letting $x = \sigma_1 y$ gives us

$$y^2 + y + \sigma_2 \sigma_1^{-2} = 0 . \tag{2}$$

Taking the trace of Eq. (2), we obtain

$$Tr(y^2+y) = Tr\, y^2 + Tr\, y = 0 = Tr\, \sigma_2 \sigma_1^{-2} , \tag{3}$$

giving us the necessary condition. Assuming $Tr\, \sigma_2 \sigma_1^{-2} = 0$, we multiply Eq. (2) by γ^2 and take the Trace leading to

$$Tr\, y(\gamma+\gamma^2) = Tr\, \gamma^2\, \frac{\sigma_2}{\sigma_1^2} . \tag{4}$$

Choosing (γ_i) a basis for $GF(2^k)$ over $GF(2)$; we obtain k linear equations

$$Tr\, [y(\gamma_i + \gamma_i^2)] = Tr\, (\gamma_i^2 \sigma_2 \sigma_1^{-2}) \qquad i = 1, 2, \cdots k$$

of which $(k-1)$ are linearly independent. Thus, we obtain a set of solutions which is a one-dimensional coset of $GF(2^k)$ viz., y_1, y_2. Placing $x_i = y_i \sigma_1$ gives us the quadratic solutions. Mechanization of this technique is discussed in Section 3.

If a normal basis (γ_i) is chosen with $Tr(\gamma_i) = 1$, then letting $\beta_i = \gamma_i + \gamma_i^2$ along with an element δ of Trace 1 gives us equations $Tr\, y\, \beta_i = Tr\, \gamma_i^2 \sigma_2 \sigma_1^{-2}$. Knowing a fixed formula for y in terms of $Tr(y\beta_i)$ allows us to immediately write down the solutions.

2. The Cubic

Let $x^3 + \sigma_1 x^2 + \sigma_2 x + \sigma_3 = 0$, $\sigma_i \in GF(2^k)$, $i = 1, 2, 3$. The substitution

$$x = (\sigma_2 + \sigma_1^2)^{1/2}\, y + \sigma_1$$

leads to

$$y^3 + y + q = 0 \tag{5}$$

where

$$q = \frac{(\sigma_3 + \sigma_2 \sigma_1)}{(\sigma_2 + \sigma_1^2)^{3/2}} .$$

Theorem 2: A necessary condition for all roots of Eq. (5) to be in $GF(2^k)$ is that $Tr\, q^{-2} = Tr\, 1$.

[2] Consultant from Electrical Engineering Department, University of California at Berkeley, California.

Proof: Let α_1, α_2, α_3 be the roots of Eq. (5). Then from Section 1

$$Tr\left[\frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2}+\frac{\alpha_1\alpha_3}{(\alpha_1+\alpha_3)^2}+\frac{\alpha_2\alpha_3}{(\alpha_2+\alpha_3)^2}\right]=0\,.$$

But

$$\sum_{i\neq j}\frac{\alpha_i\alpha_j}{(\alpha_i+\alpha_j)^2}=\frac{1}{q^2}+1$$

by a simple calculation.

$$Tr\,q^{-2}=Tr\,1. \qquad \text{QED}$$

Theorem 3: If $Tr\ q^{-2}\neq Tr\ 1$, then Eq. (5) has exactly one root in $GF(2^k)$.

Proof: Clearly, if Eq. (5) has 2 roots in $GF(2^k)$, it has the third, since $\alpha_1+\alpha_2+\alpha_3=0$.

Let $y=u+v$, so we obtain

$$u^3+v^3+(uv+1)(uv)+q=0\,.$$

Setting $uv=1$ gives us

$$u^3+v^3=q$$
$$(uv)^3=1\,.$$

u^3 and v^3 satisfy the quadratic

$$z^2+qz+1=0\,. \tag{6}$$

For k odd, $Tr\ 1/q^2\neq Tr\ 1=1$ implies $Tr\ q^{-2}=0$. This gives two solutions $z_1, z_2\in GF(2^k)$. Since k is odd, 3 does not divide 2^k-1. We may solve $u^3=z_1$ and $v^3=z_2$, obtaining unique values of u and v, giving us one solution $y=u+v$ and $x=(\sigma_2+\sigma_1^2)^{1/2}(u+v)+\sigma_1$. If k is even, $Tr\ q^{-2}=1$, and no solution of Eq. (6) is in $GF(2^k)$. However, $Tr\ q^{-2}=0$ in $GF(2^{2k})$ and so $z_1, z_2\in GF(2^{2k})$. Clearly $z_2=z_1^{-1}$ and $z_1^{2^k}=z_2$, giving us

$$(z_1)^{2^k+1}=1\,.$$

For k even, 3 does not divide 2^k+1, and so there exists a unique $u\in GF(2^{2k})$, such that $u^3=z_1$. We have, as above,

$$y=(\sigma_2+\sigma_1^2)^{1/2}(u+v)+\sigma_1$$

where

$$u, v\in GF(2_{2k})\ u^3=z_1 \qquad z_1^2+qz_1+1=0\,.$$

QED

We next prove the following.

Theorem 4: A necessary and sufficient condition that all three roots of the cubic polynomial x^3+x+q lie in $GF(2^k)$ is that $P_k(q)=0$, where the polynomials $P_k(x)$ may be defined recursively by the equations

$$P_1(x)=x$$
$$P_2(x)=x$$
$$P_k(x)=P_{k-1}(x)+x^{2^{k-3}}P_{k-2}(x)\,. \tag{7}$$

The proof of this theorem is in three parts. The first part consists of counting the number of q such that x^3+x+q has its roots in $GF(2^k)$; we shall call such q *admissible* over $GF(2^k)$. We shall find that the number of admissible q is $M=[(2^k+4)/6]$, where $[\alpha]$ denotes the greatest integer not exceeding α. In the second part of the proof we construct a polynomial of degree M which has each admissible q as a root. It follows that the roots of $P_k(x)$ are precisely those q which are admissible over $GF(2^k)$. We complete the proof by showing that the polynomials $P_k(x)$ satisfy the recursion relations (7).

The following lemma enables us to enumerate the set of admissible q.

Lemma: If $q\neq 0$ the roots of x^3+x+q lie in $GF(2^k)$ if and only if

$$q=\frac{v+v^{-1}}{(1+v+v^{-1})^3} \quad \text{for some } v\in GF(2^k)-GF(4)\,. \tag{8}$$

Furthermore, if Eq. (8) is satisfied, the roots of x^3+x+q are

$$a_1=\frac{v+v^{-1}}{1+v+v^{-1}},\ a_2=\frac{v}{1+v+v^{-1}},\ a_3=\frac{v^{-1}}{1+v+v^{-1}}\,. \tag{9}$$

Proof: If Eq. (8) is satisfied, a simple computation shows that $x^3+x+q=(x-a_1)(x-a_2)(x-a_3)$ where a_1, a_2, a_3 are given by Eq. (9). Hence, if Eq. (8) is satisfied, q is admissible. Conversely, let $q\neq 0$ be admissible and write

$$x^3+x+q=(x-a_1)(x-a_2)(x-a_3)$$

for some $a_1, a_2, a_3\in GF(2^k)$. Defining $v=(a_2/a_3)^{1/2}$, one may easily verify that Eqs. (8) and (9) are satisfied. We omit the details.

Eq. (8) defines a many-to-one correspondence between some elements $v\in GF(2^k)$ and the set of nonzero admissible q. To each admissible q there corresponds, at most, six v since Eq. (8) can be written as a sixth degree polynomial in v. But if $q\neq 0$ is admissible and a_1, a_2, a_3 are the roots of x^3+x+q we obtain six *distinct* values of v

by taking $v = (a_i/a_j)^{1/2}$ $i, j = 1, 2, 3$, $i \neq j$. Thus, with each admissible $q \neq 0$ there are associated exactly six $v \in GF(2^k) - GF(4)$. It follows that the number of admissible q is either $(2^k - 4)/6$ or $(2^k - 2)/6$, depending on whether k is even or odd. In either case, we have shown: the number of admissible $q \neq 0$ over $GF(2^k)$ is $M = [(2^k - 2)/6]$.

Now we shall construct the polynomials $P_k(x)$. Let $q \neq 0$ be any number algebraic over $GF(2)$, and let a_1, a_2, a_3 be the roots of the polynomial $x^3 + x + q$. A necessary and sufficient condition that a_1, a_2, a_3 be in $GF(2^k)$ is that

$$a_1^{2^{k-1}} = a_2^{2^{k-1}} = a_3^{2^{k-1}} = 1 . \tag{10}$$

Eqs. (10) imply

$$a_1^n + a_2^n + a_3^n = 1$$

$$(a_1a_2)^n + (a_2a_3)^n + (a_3a_1)^n = 1 \tag{11}$$

where we have set $n = 2^k - 1$. The left hand sides of Eqs. (11) are symmetric polynomials in a_1, a_2, a_3; hence, they may be expressed as polynomials in the elementary symmetric functions on a_1, a_2, a_3. For example, by the formula on page 82 of Ref. 5, we have

$$a_1^n + a_2^n + a_3^n = \sum n \frac{(\lambda_1 + \lambda_2 + \lambda_3 - 1)!}{\lambda_1!\,\lambda_2!\,\lambda_3!} \sigma_1^{\lambda_1}\sigma_2^{\lambda_2}\sigma_3^{\lambda_3}, \tag{12}$$

where the sum is over all non-negative λ_1, λ_2, λ_3, such that $\lambda_1 + 2\lambda_2 + 3\lambda_3 = n$. In our case $\sigma_1 = 0$, $\sigma_2 = 1$, $\sigma_3 = q$, and $n = 2^k - 1$, so that Eq. (12) may be simplified to

$$a_1^n + a_2^n + a_3^n = \sum_{2\lambda_2 + 3\lambda_3 = 2^k - 1} \frac{(\lambda_2 + \lambda_3 - 1)!}{\lambda_2!\,\lambda_3!} q^{\lambda_3}$$

$$= \sum_j \binom{2^{k-1} - 2 - j}{2j} q^{2j+1},$$

where we have used the fact that the coefficients in this polynomial are to be computed mod 2. Similarly,

$$(a_1a_2)^n + (a_2a_3)^n + (a_3a_1)^n = \sum \binom{2^k - 1 - 2j}{j} q^{2j} . \tag{13}$$

We have shown that a necessary condition that the roots of $x^3 + x + q$ lie in $GF(2^k)$ is that

$$A(q) \equiv \sum \binom{2^{k-1} - 2 - j}{2j} q^{2j+1} = 1 , \tag{14}$$

and

$$B(q) \equiv \sum \binom{2^k - 1 - 2j}{j} q^j = 1 \tag{15}$$

(Eq. (15) follows from Eqs. (11) and (13) by extracting a square root.)

In order to simplify Eqs. (14) and (15) further we need a technique for computing binomial coefficients mod 2. Such a technique is contained in the following lemma, (Ref. 6).

Lemma: Let $n = n_1n_2 \cdots n_s$, $j = j_1j_2 \cdots j_s$ be the binary expansions of n and j. Then

$$\binom{n}{j} \equiv 1 \pmod 2 ,$$

if and only if $j_i = 1$ implies $n_i = 1$ for all $i = 1, \cdots, s$.

This lemma can be proved by induction on s, using the relations

$$\binom{2n}{2j} = \binom{2n+1}{2j} = \binom{2n+1}{2j+1} = \binom{n}{j} \pmod 2$$

$$\binom{2n}{2j+1} = 0 \pmod 2 .$$

The details are omitted.

By applying this lemma to Eq. (15) we find

$$B(q) = \sum q^j ,$$

where the sum is over those non-negative $j \leq 2^k - 1/3$ whose binary expansion does *not* contain two consecutive ones not separated by a zero. Similarly, we find

$$A(q) = \sum q^j ,$$

where we sum over the *odd* j which occur in $B(q)$. For example, if $k = 4$, we have

$B(q) = 1 + q + q^2 + q^4 + q^5$ (binary exponents 0,1,10,100,101).

$A(q) = q + q^5$ (binary exponents 1,101).

It follows that the polynomial $A(q) + B(q)$ contains only even powers of q and that its degree is at most $(2^k - 1)/3 - 1/3 = (2^k - 2)/3$. Thus, by adding Eqs. (14) and (15) and extracting the square root of both sides we find that a necessary condition for q to be admissible is

$$C(q) = \sum q^j = 0 , \tag{16}$$

where $0 \leq j \leq [2^k - 2/6]$ and the binary expansion of j does not contain two consecutive ones. The number of admissible $q \neq 0$ is the same as the degree of $C(q)$; hence, the roots are precisely the set of admissible $q \neq 0$. We define $P_k(x)$ by the equation

$$P_k(x) = xC(x) = x \sum x^j , \tag{17}$$

where $j \leq [(2^k - 2)/6]$, and the binary expansion of j does not contain two consecutive ones. Since $q = 0$ is admissible for any k, we have shown that the roots of $P_k(x)$ are precisely the set of those q such that $x^3 + x + q$ splits into linear factors over $GF(2^k)$.

Now we show that the $P_k(x)$ satisfy the recursion Eqs. (6). The equations

$$P_1(x) = x$$

$$P_2(x) = x$$

follow from Eq. (17). It remains to verify that for $k \geq 3$,

$$P_k(x) = P_{k-1}(x) + x^{2^{k-3}} P_{k-2}(x). \tag{18}$$

To this end we examine

$$M = [(2^k - 2)/6] = [(2^{k-1} - 1)/3].$$

The binary expansion of $2^{k-1} - 1$ consists of a string of $k - 1$ ones. The binary expansion of 3 is 11. It follows that $M = [(2^{k-1} - 1)/3]$ is the largest integer of binary length $k - 2$ bits whose binary expansion does not contain two consecutive ones. It follows from Eq. (12) that

$$P_k = x \sum x^j ,$$

where j ranges over all binary integers of length at most $k - 2$ whose binary expansion does not contain consecutive ones. Eq. (18) is an immediate consequence of this representation. For example, if $k = 5$ and if we express the exponents in binary notation,

$$\begin{aligned} P_5(x) &= x(x^0 + x^1 + x^{10} + x^{100} + x^{101}) \\ &= x(x^0 + x^1 + x^{10}) + x^{100} [x(x^0 + x^1)] \\ &= P_4(x) + x^4 P_3(x) . \end{aligned}$$

In general, $P_k(x) + P_{k-1}(x) = x \sum x^j$, where j has binary length exactly $k-2$ and contains no consecutive ones in its binary expansion. The set of such j may be obtained by adding 2^{k-3} to the set of permissible exponents of length $k - 4$. In other words

$$P_k(x) + P_{k-1}(x) = x^{2^{k-3}} P_{k-2}(x) ,$$

which completes the proof.

3. Mechanization of the Solution of Quadratic Equations in Fields of Characteristic 2

Case 1 (With Repeated Roots).

In order to solve a quadratic equation of the type $x^2 + c = 0$, where c and $x \in GF(2^k)$, we must extract the square root of c. Since in any field of characteristic two we have the identity $(x + y)^2 = x^2 + y^2$ and similarly, $(x + y)^{1/2} = (x)^{1/2} + (y)^{1/2}$, the square root is a linear operation. In terms of a fixed basis of $GF(2^k)$, namely $u_1, u_2, \cdots, u_m$, we may write $c = \sum_{i=1}^{k} c_i u_i$, where $c_i \in GF(2)$. Because of the linearity of the square root, we then have $(c)^{1/2} = \sum_{i=1}^{k} c_i (u_i)^{1/2}$. Of course, $(u_i)^{1/2}$ can also be represented in terms of this same basis, with $(u_i)^{1/2} = \sum_j R_{i,j} u_j$, with $R_{i,j} \in GF(2)$. We then have

$$\begin{aligned} (c)^{1/2} &= \sum_{i=1}^{m} \sum_{j=1}^{m} c_i R_{i,j} u_j \\ &= \sum_{j=1}^{m} \left(\sum_{i=1}^{m} c_i R_{i,j} \right) u_j . \end{aligned}$$

For example, in $GF(2^5)$, let us take the basis consisting of $u_i = \alpha^{5-i}$ for $i = 1, 2, 3, 4, 5$, where α satisfies the equation $\alpha^5 + \alpha^2 + 1 = 0$. Then

$$(\mu_1)^{1/2} = (\alpha^4)^{1/2} = \alpha^2$$

$$(\mu_2)^{1/2} = (\alpha^3)^{1/2} = \alpha^{17} = \alpha^4 + \alpha + 1$$

$$(\mu_3)^{1/2} = (\alpha^2)^{1/2} = \alpha$$

$$(\mu_4)^{1/2} = (\alpha)^{1/2} = \alpha^{16} = \alpha^4 + \alpha^3 + \alpha + 1$$

$$(\mu_5)^{1/2} = (1)^{1/2} = 1 .$$

Hence, the matrix R is given by

$$R = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

For example, if we wish to take the square root of $c = \alpha^4 + \alpha^2 + \alpha + 1$ we write

$$(c)^{1/2} = cR = [1\ 0\ 1\ 1\ 1] \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = [1\ 1\ 1\ 0\ 0] .$$

We can verify this by checking that $c = \alpha^{26}$; $(c)^{1/2} = \alpha^{13}$.

Case 2: (Without Repeated Roots.)

In general, the quadratic equation may be written as $x^2 + bx + c = 0$. We have just seen that if $b = 0$, this equation has a unique solution in $GF(2^k)$ and that this solution may be found by multiplying the vector representing c by the matrix R which extracts square roots.

If $b \neq 0$, a more devious procedure is required. We first transform the equation by introducing the new variable $y = x/b$. This new variable satisfies the equation $b^2y^2 + b^2y + c = 0$, or $y^2 + y = d$, where $d = c/b^2$. We now notice that if $y_i^2 + y_i = v_i$ and $y_j^2 + y_j = v_j$, then $(y_i + y_j)^2 + (y_i + y_j) = v_i + v_j$. Hence, a solution of the equation $y^2 + y = d = \Sigma\, d_i v_i$; $d_i \in GF(2)$, is given by $y = \Sigma\, d_i y_i$, where y_i is a solution of the equation $y_i^2 + y_i = v_i$. This shows that the set of v for which the equation $y^2 + y = v$ has a solution in $GF(2^k)$ forms a subspace of the vector space $GF(2^k)$. Since each v in the subspace corresponds to two distinct roots in $GF(2^k)$, and since each value of y corresponds to one value of v, the dimensionality of the subspace is evidently 2^{k-1}. Consequently, the solutions of the equation $y^2 + y + d = 0$ may be represented in terms of solutions to the equations $y_i^2 + y_i = v_i = 0$, for $i = 1, 2, \cdots, k-1$, where the v_i span the space of v's for which $y^2 + y + v = 0$ has solutions in $GF(2^k)$. If d is not expressable as a sum of v's, the equation $y^2 + y + d$ has no solution in $GF(2^k)$. This corresponds to the condition $Tr\ d = 0$ in Section 1. If $d = \Sigma\, d_i v_i$, then $y = \Sigma\, d_i y_i$ is a solution of $y^2 + y + d$. The other solution is found by adding to the first solution a solution of $y^2 + y = 0$, namely $y = 1$.

For example in $GF(2^5)$, with $\alpha^5 + \alpha^2 + 1 = 0$, the equations $y_i^2 + y_i = v_i$ have the solutions

$$v_1 = \alpha \qquad y_1 = \alpha^3$$

$$v_2 = \alpha^2 \qquad y_2 = \alpha^6 = \alpha^3 + \alpha$$

$$v_3 = \alpha^4 \qquad y_3 = \alpha^{12} = \alpha^3 + \alpha^2 + \alpha$$

$$v_4 = \alpha^8 = \alpha^3 + \alpha^2 + 1 \qquad y_4 = \alpha^{24} = \alpha^4 + \alpha^3 + \alpha^2 + \alpha$$

or preferably

$$v^4 = \alpha^3 + 1 \qquad y_4 = \alpha^4 + \alpha^2\,.$$

There are no solutions to the equations $y^2 + y + 1 = 0$, or $y^2 + y + \alpha^3 = 0$. In terms of our previous basis $u_i = \alpha^{5-i}$, with $y = \Sigma\, y_i u_i$; $d = \Sigma\, d_i u_i$, the solution of the equation $y^2 + y + d = 0$ is given by

$$[y_1, y_2, y_3, y_4] = [d_1, d_2, d_3, d_4] \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

if $d_5 = d_2$. If $d_5 \neq d_2$, no solutions exist. If solutions exist, y_5 is arbitrary.

4. Mechanization of the Solution of Cubic and Quartic Equations in Fields of Characteristic 2

In Section 2 we considered the cubic equation in $GF(2^k)$. We gave necessary and sufficient conditions that all three roots be in $GF(2^k)$, and we succeeded in obtaining expressions for these roots in closed form. Not unexpectedly, those expressions involved cube roots. While conceptually simple, cube roots turn out to be rather difficult to implement. For computational purposes, elements in $GF(2^k)$ are usually represented by k binary digits, which are the coefficients of the field elements in terms of some standard basis. This representation makes addition of two elements extremely simple. To facilitate the operations of multiplication and division, this standard basis is usually chosen to consist of consecutive powers of a primitive element of the field. The operation of extracting the square root poses no major difficulty, because it is linear over the standard basis. Thus, square roots can be extracted by multiplying the $(k-1)$-dimensional binary vector by a $(k-1)$ by $(k-1)$ constant, wired in matrix, as we saw in Section 3.

If the nonzero elements of the field were expressed as powers of a fixed primitive element, then cube roots could readily be extracted by dividing the exponent by three. Although extremely helpful for hand computations, this method proves difficult to mechanize, because it is not easy to convert from the basis representation of elements to the exponential representation without the aid of a stored conversion table.

Our solution is to convert the cubic into a quartic, which, like the quadratic, can be "linearized."

Let the general quartic be represented by

$$ax^4 + bx^3 + cx^2 + dx + e = 0\,,$$

where x is an unknown element of $GF(2^k)$, and a, b, c, d, e are known elements of $GF(2^k)$. Although we allow the

possibility of $a = 0$ (the cubic case), we exclude the previously considered quadratic case with $a = b = 0$.

Our first step is to reduce the quartic to the standard form:

$$z^4 + Az^2 + fz = g\,,$$

where z is an unknown element in $GF(2^k)$ and f and g are known elements of $GF(2^k)$ and A is a known element in $GF(2)$. Here $g = 0$ if the original equation was cubic.

If $b = 0$ and $c = 0$, the transformation is accomplished immediately by dividing the original equation through by a. In this case $z = x$.

If $b = 0$, $c \neq 0$, then the transformation is accomplished by the substitution $x = hz$; $z = x/h$ where $h = (c/a)^{1/2}$.

If $b \neq 0$, we substitute $x = y + \Delta$. The choice $\Delta = (d/b)^{1/2}$ eliminates the linear term, giving

$$ay^4 + by^3 + c'y^2 + e' = 0\,,$$

where

$$\Delta = (d/b)^{1/2}$$

$$c' = c + b\Delta$$

$$e' = a\Delta^4 + b\Delta^3 + c\Delta^2 + d\Delta + e\,.$$

If $e' = 0$, y^2 may be factored out, leaving a quadratic equation.

If $e' \neq 0$, $c' = 0$, the equation is reduced to standard form by the transformation $y = 1/z$. In that case $A = 0$; $f = b/e'$; $g = a/e'$.

Finally, if $e' \neq 0$, $c' \neq 0$ then the equation is reduced to standard form by the transformation $y = h/z$; $z = h/y$, where $h = (e'/c')^{1/2}$. In that case $A = 1$; $f = bh^3/e'$; $g = ah^4/e'$.

Once the equation is reduced to standard form, the left side of the equation is linear in z. For if

$$z_1^4 + Az_1^2 + fz_1 = g_1$$

and

$$z_2^4 + Az_2^2 + fz_2 = g_2$$

then

$$(z_1 + z_2)^4 + A(z_1 + z_2)^2 + f(z_1 + z_2)^2 = g_1 + g_2\,.$$

For this reason, the quartic operator on the left side of the standard form equation may be represented as a $k \times k$ binary matrix

$$Q = S^2 + AS + f\,,$$

where S represents the squaring matrix and f represents the matrix which multiplies by the constant f.

Thus, in order to solve quartic equations in $GF(2^k)$, we precalculate and permanently store the three matrices S^2, $S^2 + S$, and $R = S^{-1}$. Then by appropriate use of the arithmetic operations and the R matrix, we reduce the quartic to standard form. We then construct the $k \times k$ binary matrix $Q = S^2 + AS + f$, by adding the f matrix to the stored S^2 matrix (if $A = 0$) or to the stored $S^2 + S$ matrix (if $A = 1$). We then solve the binary equations $zQ = g$. The rank of Q may be k, $k - 1$, or $k - 2$, in which cases we will have 1 solution, 0 or 2 solutions, or 0 or 4 solutions respectively. These solutions give the roots of the standard quartic in $GF(2^k)$. Notice that the quartic may have 0, 1, 2, or 4 roots in $GF(2^k)$. If the original equation is cubic, then it reduces to a standard quartic which has zero as an extraneous root. In that case the standard quartic may have 1, 2, or 4 solutions, of which 0, 1, or 3 will be nonzero.

For example, let us again consider $GF(2^5)$, represented with respect to the basis α^4, α^3, α^2, α^1, 1. Here α satisfies $\alpha^5 + \alpha^2 + 1 = 0$. For this representation of this field we have

$$S = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$S^2 = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$S^2 + S = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R = S^{-1} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Let us now solve the particular quartic

$$ax^4 + bx^3 + cx^2 + dx + e = 0,$$

where

$$\begin{aligned} a &= \alpha^0 = 00001 \\ b &= \alpha^{14} = 11101 \\ c &= \alpha^8 = 01101 \\ d &= \alpha^1 = 00010 \\ e &= \alpha^{20} = 01100 \end{aligned}$$

We give here both the representation with respect to our standard basis and the exponents of the representation as a power of α. This latter representation is included here for convenience of the reader in following the multiplications and divisions; in an actual decoder, this representation would probably never appear.

To solve this quartic, we first reduce it to standard form. To accomplish this, we set $x = y + \Delta$; $\Delta = (d/b)^{1/2} = (\alpha^{18})^{1/2} = \alpha^9 = 11010$. If we did not have the exponents available, we would determine Δ by

$$\Delta = (d/b)R = [00011]\begin{bmatrix} 0&0&1&0&0 \\ 1&0&0&1&1 \\ 0&0&0&1&0 \\ 1&1&0&1&1 \\ 0&0&0&0&1 \end{bmatrix} = [11010].$$

We then have

$$ay^4 + by^3 + c'y^2 + e' = 0,$$

where

$$c' = c + b\Delta = 00010 = \alpha$$

$$e' = \Delta^4 + b\Delta^3 + c\Delta^2 + d\Delta + e = \begin{matrix} 00101 \\ 10001 \\ 10111 \\ 10001 \\ 01100 \\ \hline 11110 = \alpha^{24} \end{matrix}$$

$$e' =$$

We next set $h = (e'/c')^{1/2} = (\alpha^{23})^{1/2} = \alpha^{27} = \alpha^{-4} = 11011$. (Again this could be determined by the R matrix instead of the exponents.) Setting $z = h/y$ gives

$$z^4 + Az^2 + fz = g$$

where $A = 1$, $f = bh^3/e' = \alpha^9 = 11010$; $g = ah^4/e' = \alpha^{-9} = 10101$.

The jth row of the matrix form of the constant f gives the representation of $\alpha^{m-j}f$. For this reason, this matrix is readily constructed from the bottom up, taking higher rows as the successive outputs of a primitive shift register starting from the state $f = 11010$. The matrix is

$$f = \begin{bmatrix} 1&1&1&0&0 \\ 0&1&1&1&0 \\ 0&0&1&1&1 \\ 1&0&0&0&1 \\ 1&1&0&1&0 \end{bmatrix}$$

Since $A = 1$, we next compute Q by taking

$$Q = (S^2 + S) + f$$

$$Q = \begin{bmatrix} 0&1&0&1&0 \\ 0&1&0&1&0 \\ 1&1&0&1&0 \\ 0&0&1&0&1 \\ 1&1&0&1&0 \end{bmatrix}$$

To solve the system $zQ = g$, we triangularize the augmented matrix by column operations, insofar as possible.

$$[z,1] \cdot \begin{bmatrix} 0&1&0&1&0 \\ 0&1&0&1&0 \\ 1&1&0&1&0 \\ 0&0&1&0&1 \\ 1&1&0&1&0 \\ 1&0&1&0&1 \end{bmatrix} = \underline{0}$$

$$[z,1] \cdot \begin{bmatrix} 0&1&0&0&0 \\ 0&1&0&0&0 \\ 1&1&0&0&0 \\ 0&0&1&0&1 \\ 1&1&0&0&0 \\ 1&0&1&0&1 \end{bmatrix} = \underline{0}$$

$$[z,1] \cdot \begin{bmatrix} 0&1&0&0&0 \\ 0&1&0&0&0 \\ 1&0&0&0&0 \\ 0&0&1&0&1 \\ 1&0&0&0&0 \\ 1&1&1&0&1 \end{bmatrix} = \underline{0}$$

$$[z,1] \cdot \begin{bmatrix} 0&1&0 \\ 0&1&0 \\ 1&0&0 \\ 0&0&1 \\ 1&0&0 \\ 1&1&1 \end{bmatrix} = \underline{0}$$

The four solutions are

$$z = \begin{cases} 10110 = \alpha^{28} \\ 01110 = \alpha^{12} \\ 01011 = \alpha^{27} \\ 10011 = \alpha^{17} \end{cases}$$

$$y = k/z = \begin{cases} \alpha^{30} = 10010 \\ \alpha^{15} = 11111 \\ \alpha^{0} = 00001 \\ \alpha^{10} = 10001 \end{cases}$$

$$x = y + \Delta = \begin{cases} 01000 \\ 00101 \\ 11011 \\ 01011 \end{cases}$$

5. Mechanization of the Quintic and Higher Degree Equations

The technique is best clarified by an example. Let

$$x^5 + \sigma_1 x^4 + \sigma_2 x^3 + \sigma_3 x^2 + \sigma_4 x + \sigma_5 = 0, \sigma_2 \epsilon GF(2k). \tag{19}$$

We can remove the cubic term by letting $x = \sigma_3/\sigma_2 + y$ and then setting $z = 1/y$. This gives

$$z^5 + Az^4 + Bz^2 + Cz + D = 0.$$

Multiplying by γ^5 and taking the Trace gives us

$$Tr(\gamma z)^5 + Tr\, z\,[E(\gamma)] = Tr\, D\gamma^5$$

where

$$E(\gamma) = [(A\gamma^5)^{1/4} + (B\gamma^5)^{1/2} + C\gamma^5].$$

Let us form the vectors $(T(z))$, $(T(z^5))$ and store them. $T(z)$ is the vector whose ith component is $T(\beta^i)$, β a generator of $GF(2^k)$. For $\gamma = \beta^i$, $i = 0, 1, \cdots, k-1$ form the sum of the vectors

$$(T(\gamma z)^5) + (T(zE(\gamma))) + (Tr\,\gamma^5)^C = f_i,$$

where C denotes the mod 2 complement. Note that $(T(\gamma z)^5)$ is a cyclic shift of length j of the vector $T(z^5)$ if $\gamma = \beta^j$ and $(TzE(\gamma))$ is a cyclic shift of length l along the vector $T(z)$, for $E(\gamma) = \beta^l$. The coordinatewise "anding" of the vectors f_i, $i = 0, 1, 2, \cdots, k-1$ yields the error vector, i.e., a vector whose one positions are the roots of the quintic.

Example:

$x^5 + x + 1 = 0 \quad k = 3$

$Tr\, x = (1001011)$

$Tr\, x^5 = (1110100)$

$Tr(\gamma x)^5 + Tr\,\gamma^5 x = Tr\,\gamma^5 \quad \gamma = \beta^i \quad i = 0,1,2$

$i = 0 \quad f_0 = (0111111)$

$i = 1 \quad f_1 = \begin{array}{r} 1101001 \\ 1110010 \\ \hline 0011011 \end{array} \quad f_1 \cdot f_2 \cdot f_3 = 0001011$

$i = 2 \quad f_2 = 0001111$ i.e., $x = \beta^3, \beta^6, \beta^5$.

For higher order equations, the technique is the same. We need store (Tx^i) for those i odd remaining after reduction. For each i odd, we need store only the first k bits along with the recursion rule corresponding to that power of x, i.e.,

$$k = 3 \quad \{Trx\} = \{100\cdots\}\ a_{n+3} = a_{n+1} + a_n$$
$$\{Trx^5\} = \{111\cdots\}\ a_{n+3} = a_{n+2} + a_n.$$

C. Analysis of Channels With Unidirectional Drift

E. R. Berlekamp[3] *and L. Kleinrock*[4]

This article calculates the error bounds for an unusual class of channels, for which these bounds take an unexpectedly simple form. Because of this simplification, these channels provide an interesting example of the theorems of R. Gallager.

1. The Channel Description

In this article we analyze the properties of an unusual class of discrete memoryless channels, which we refer to as *ladder* and *star* channels. The main feature of these channels is that the noise causes confusion only by shifting the input symbols in a single direction.

[3] Consultant, Electrical Engineering Department, University of California—Berkeley.

[4] Consultant, Engineering Department, University of California—Los Angeles.

The first member of the class of ladder channels is the well-known *Z-channel* shown in Fig. 1. We see that the "drift" for this channel is in the "upward" direction only. The next member of this class is shown in Fig. 2. Again we note the upward drift; symbol 3 may be received as either symbols 1, 2, or 3; symbol 2 may be received only as symbols 1 or 2; and symbol 1 is noiseless. Observe that the Z-channel is imbedded within this channel. To get the general ladder channel, we continue to add input (and output) symbols, giving each new symbol a nonzero transition probability to all previous symbols, but allowing no previous symbols to have transitions to the new symbol, as shown in Fig. 3. We note that all the ladder channels for $k = 2, 3, \cdots, K-1$ are imbedded in this K-symbol channel. The transition probabilities p_{jk} (where $p_{jk} = p_r\,[\text{output} = j \mid \text{input} = k]$) are chosen such that

$$P_{jk} = \begin{cases} 0 & j > k \\ \beta\alpha^{k-j} & j = 2, 3, \cdots, k \\ \alpha^{k-1} & j = 1 \end{cases} \qquad (1)$$

where $\alpha + \beta = 1$.

It is interesting to observe the way in which these transition probabilities might come about. Consider a K-input, K-output channel, as shown in the *flow* diagram of Fig. 4. In this figure, we allow an impulse to be fed in at any one of the input terminals. The directed branches show the possible directions this impulse may travel, and the conditional probability of traversing a particular branch is given by the branch label. The node where the impulse finally emerges identifies the output symbol. It is clear that the probability of emerging at node j, given that the impulse was fed in at node k, is merely p_{jk}, as given in

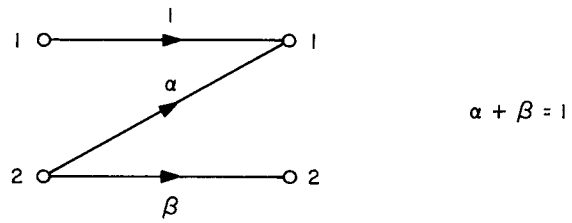


Fig. 1. The Z-channel (ladder channel with 2 symbols)

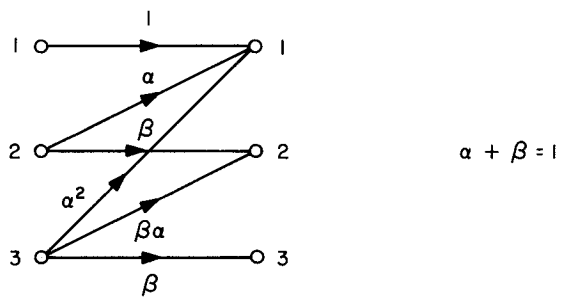


Fig. 2. The ladder channel with 3 symbols

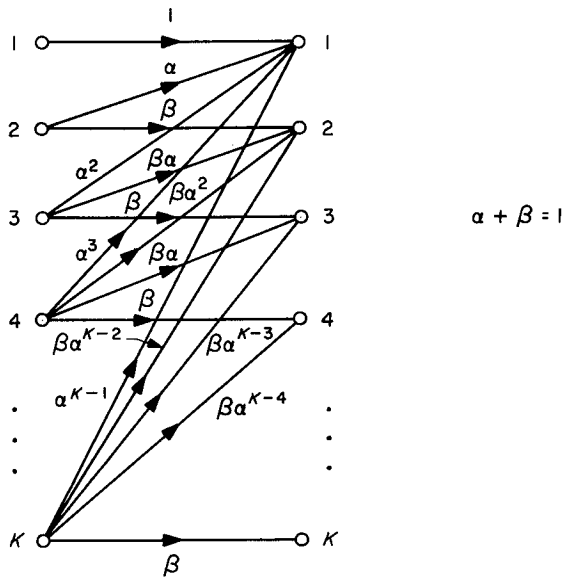


Fig. 3. The ladder channel with *K* symbols

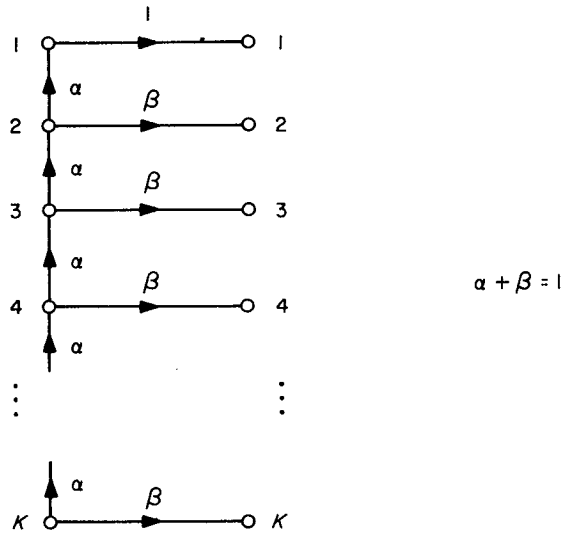


Fig. 4. The ladder channel as a flow diagram

Eq. (1). From the topological structure of this flow diagram, we obtain the name "ladder" channel. After presenting our results for this class of channels below, we then further generalize the ladder channel to a class of central-drift channels, which we refer to as "star" channels.

2. Definitions

Following the approach taken by Gallager (Ref. 7) we observe that for any discrete memoryless channel, there exist codes for which the average error probability, P_e, satisfies

$$P_e \leq e^{-NE(R)}$$

where

N = block code length

R = transmission rate $= \frac{1}{N} \ln M$

M = number of code words of length N

and

$$E(R) = E_0(\rho) - \rho R \tag{2}$$

where

$$E_0(\rho) = \max_{\underline{q}} \left\{ - \ln \sum_j \left[\sum_k q_k(\rho)\, p_{jk}^{\frac{1}{1+\rho}} \right]^{1+\rho} \right\} \tag{3}$$

$$R = \frac{\partial E_0(\rho)}{\partial \rho} \tag{4}$$

$$0 \leq \rho \leq 1$$

and

$\underline{q} = [q_1, q_2, \cdots, q_K]$ = input probability vector.

In general $\underline{q} = \underline{q}(\rho)$. The tilted output probability distribution is defined by

$$\underline{f}(\rho) = [f_1(\rho), \cdots, f_J(\rho)]$$

where

$$f_j(\rho) = \frac{\left(\sum_k q_k(\rho)\, p_{jk}^{\frac{1}{1+\rho}} \right)^{1+\rho}}{\exp\left[- E_0(\rho) \right]}$$

and where q is the optimum input probability distribution. $\underline{f}(\rho)$ satisfies the conditions

$$\left(\sum_j f_j(\rho)^{\frac{\rho}{1+\rho}}\, p_{jk}^{\frac{1}{1+\rho}} \right)^{1+\rho} \geq \exp - E_0(\rho)$$

for all k, with equality unless $q_k = 0$. For any given ρ, $f(\rho)$ is unique, although $q(\rho)$ may not be unique.

Note that E_0 and R are related through the parameter ρ for $0 \leq \rho \leq 1$. This gives $E(R)$ as a function of R for the range

$$\frac{\partial E_0(\rho)}{\partial \rho}\Big|_{\rho=1} \leq R \leq \frac{\partial E_0(\rho)}{\partial \rho}\Big|_{\rho=0} = C = \text{capacity.}$$

For $0 \leq R \leq \frac{\partial E_0(\rho)}{\partial \rho}\Big|_{\rho=1}$ we have

$$E(R) = E_0(1) - R \tag{5}$$

Furthermore, for low rates, we can improve this bound by "expurgation" methods. In this case, in Eq. (1) we use, in place of $E_0(\rho)$, the quantity

$$- \rho \frac{\ln 4}{N} + E_x(\rho)$$

where

$$E_x(\rho) = \max_{\underline{q}} \left\{ - \rho \ln \sum_k \sum_i q_k q_i \left[\sum_j (p_{jk} p_{ji})^{1/2} \right]^{1/\rho} \right\} \tag{6}$$

This improved bound is applicable in the region

$$\lim_{\rho \to \infty} \frac{\partial E_x(\rho)}{\partial \rho} \leq R + \frac{\ln 4}{N} \leq \frac{\partial E_x(\rho)}{\partial \rho}\rho = 1$$

For further details, refer to Refs. 7 and 8.

3. Results

We have calculated the error bounds, defined in Section 2, for the general ladder channel (Fig. 3, Section 1), and we present the results of these calculations here.

The surprising result is that the optimum input distribution uses all the middle inputs ($k = 2, 3, \cdots, K-1$) with *equal* probability, although the optimum distribution at both endpoints ($k = 1, K$) depends upon the rate. Furthermore, the output distribution resulting from this input distribution uses all but the first ($k=1$) letter with equal probability.

The input distribution $q(\rho)$ for the sphere packing bound (or equivalently, for random coding with a large list) is, for $0 \leq \rho \leq \infty$

$$\left.\begin{aligned} q_K &= \left[\xi + (K-1)\left(1-\alpha^{\frac{1}{1+\rho}}\right)\right]^{-1} \\ q_k &= q_K\left(1-\alpha^{\frac{1}{1+\rho}}\right) \qquad k = 2, 3, \cdots, K-1 \\ q_1 &= q_K\left(\xi-\alpha^{\frac{1}{1+\rho}}\right) \end{aligned}\right\} \tag{7}$$

where

$$\xi = \left[\frac{\beta}{1-\alpha^{1/1+\rho}}\right]^{\frac{1}{\rho}} \tag{8}$$

Also

$$E_0(\rho) = \ln\left[1 + \frac{(K-1)\left(1-\alpha^{\frac{1}{1+\rho}}\right)}{\xi}\right] \tag{9}$$

We have that

$$\lim_{\rho\to 0} \xi = \alpha^{-\frac{\alpha}{1-\alpha}}$$

$$\lim_{\rho\to\infty} \xi = 1$$

The resulting tilted output distribution is

$$\left.\begin{aligned} f_1 &= \left[1 + \frac{(K-1)(1-\alpha^{1/1+\rho})}{\xi}\right]^{-1} \\ f_k &= \left[K-1 + \frac{\xi}{1-\alpha^{1/1+\rho}}\right]^{-1} \qquad k = 2, 3, \cdots, K \end{aligned}\right\} \tag{10}$$

The channel capacity, C, is

$$C = \ln\left[1 + (K-1)\beta\,\alpha^{\frac{\alpha}{1-\alpha}}\right] \tag{11}$$

Thus,

$$\lim_{\beta\to 0} \frac{C}{\beta} = \frac{(K-1)}{e}$$

$$\lim_{\beta\to 1} C = \ln K$$

Also

$$\lim_{\rho\to\infty} E_0(\rho) = (K-1)\ln\frac{1}{\alpha} \tag{12}$$

For the expurgated codes, we get

$$\left.\begin{aligned} q_1 = q_K &= \left[2 + (K-2)\left(1-\alpha^{\frac{1}{2\rho}}\right)\right]^{-1} \\ q_k &= q_K\left(1-\alpha^{\frac{1}{2\rho}}\right) \qquad k = 2, 3, \cdots, K-1 \end{aligned}\right\} \tag{13}$$

and

$$E_x(\rho) = -\rho \ln\left[\frac{1 + \alpha^{\frac{1}{2\rho}}}{2 + (K-2)(1-\alpha^{1/2\rho})}\right] \tag{14}$$

As $\rho\to\infty$, $q_1 = q_K = \frac{1}{2}$, $q_k = 0$ $(k = 2, 3, \cdots, K-1)$ and

$$E_x(\infty) = \frac{K-1}{4}\ln\frac{1}{\alpha} \tag{15}$$

Note that $E_0(\infty)/E_x(\infty) = 4$ for these channels.

Equations (7)–(15) completely describe the (upper) error bounds for these ladder channels. As shown in Ref. 8, Eqs. (9) and (14) also give exponentially correct lower bounds.

4. Extension to Star Channels

We now consider a collection of L ladder channels, the lth having K_l input (and output) symbols ($l = 1, 2, \cdots, L$) and each with the same parameter α describing their transition probabilities. [See Eq. (1).] We consider that the first symbol of each channel is a *common* symbol to all channels. We have, then, a total of $K - L + 1$ input and output symbols for this one larger channel (which we refer to as a *star* channel), where

$$K = \sum_{l=1}^{L} K_l \tag{16}$$

The observation here is that all inputs tend to drift toward the "central" symbol 1. An example is given in Fig. 5 for $L = 2$, $K_1 = 3$, $K_2 = 2$; this is a 4-input 4-output symbol channel. For the star channels, we can get results

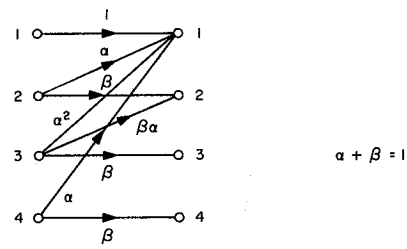


Fig. 5. A star channel with $K_1 = 3, K_2 = 2$

very much like those for the ladder channels ($L = 1$); however, the expurgated bounds do not come out simply and so we omit them. We have shown the following but omit proofs. We use the shorthand notation:

q_1 = optimum probability of using input symbol 1.

q_{K_l} = optimum probability of using the K_lth input symbol of the lth ladder channel.

q_{int} = optimum probability of using any of the other "internal" input symbols of the star channel.

We have then,

$$\left.\begin{aligned} q_{K_1} = q_{K_2} = \cdots = q_{K_L} &= \left[\xi + (K-L)(1-\alpha^{1/1+\rho})\right]^{-1} \\ q_1 &= q_{K_L}(\xi - L\alpha^{1/1+\rho}) \\ q_{\text{int}} &= q_{K_L}(1-\alpha^{1/1+\rho}) \end{aligned}\right\} \tag{17}$$

and

$$\left.\begin{aligned} f_1 &= \left[1 + \frac{(K-L)(1-\alpha^{1/1+\rho})}{\xi}\right]^{-1} \\ f_k &= \left[K - L + \frac{\xi}{1-\alpha^{1/1+\rho}}\right]^{-1} \qquad k \neq 1 \end{aligned}\right\} \tag{18}$$

and

$$E_0(\rho) = \rho \ln\left[1 + \frac{(K-L)(1-\alpha^{1/1+\rho})}{\xi}\right] \tag{19}$$

$$C = \ln\left[1 + (K-L)\beta\,\alpha^{\alpha/1-\alpha}\right] \tag{20}$$

Also, we see that

$$\lim_{\beta \to 0} \frac{C}{\beta} = \frac{K-L}{e}$$

$$\lim_{\beta \to 1} C = \ln(K - L + 1)$$

D. A Combinatorial Identity in Order Statistics

R. J. McEliece

From time to time in the mathematical literature, the following elementary combinatorial identity occurs.

$$\sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} \frac{1}{j} = 1 + \frac{1}{2} + \cdots + \frac{1}{n}. \tag{1}$$

The most recent rediscovery of this is in Ref. 9. Identities of this form turn out to be very useful in many statistical calculations, especially in investigations of order statistics. In particular, the identity (1) can be used to calculate the mean value of the largest order statistic from a uniform distribution on [0, 1].

We present here a generalization of Eq. (1), which can be used, for example, to give formulas for all moments of *any* of the order statistics from such a distribution.

1. Two Functions

We define for $s \geq 0$ and $r \leq n$ two functions

$$F_s(r,n) = \sum_{j=r}^{n} (-1)^{j+r} \binom{n}{j} \binom{j-1}{r-1} \frac{1}{j^s}$$

and

$$G_s(r,n) = \sum \frac{1}{k_1 k_2 \cdots k_s},$$

where the summation is extended over all s-tuples $(k_1, k_2, \cdots, k_s)$ for which $r \leq k_1 \leq k_2 \leq \cdots \leq k_s \leq n$; $G_0(r,n) = 1$. The identity (1) may then be written as $F_1(1,n) = G_1(1,n)$.

2. Main Result

Theorem. $F_s(r, n) = G_s(r, n)$ for all admissible values of the parameters.

Remark. The presence of the factor $\binom{j-1}{r-1}$ in $F_s(r,n)$ is unfortunate. However, some such factor seems to be necessary for an identity of this sort. It is apparently the price one must pay for the luxury of allowing the lower limit of the summation to vary.

Proof. Since the proof of this theorem, as the proof of nearly all such theorems, is mostly computational, we will sketch the proof before giving details. The reader may then spare himself the trouble of reading all the proof if he merely wishes to get an idea of why the theorem is true.

We shall show that both F and G satisfy the same recurrence relation, and that they have the same boundary values. Thus, let $H_s(r,n)$ represent either $F_s(r,n)$ or $G_s(r,n)$. Then we shall show that

$$H_s(r,n) = H_s(r,n-1) + \frac{1}{n} H_{s-1}(r,n), \tag{2}$$

$$H_0(r,n) = 1 \tag{3}$$

$$H_s(n,n) = \frac{1}{n^s} \tag{4}$$

Eq. (4) is trivial for both F and G. Eq. (3) is the definition of $G_0(r,n)$, and this definition is forced if Eq. (2) is to be true for G with $s = 1$. Eq. (2) is almost obvious for G with $s>1$, since the summands for $G_s(r,n)$ may be divided into two classes: those for which $k_s = n$ and those for which $k_s < n$. The first sum is clearly $(1/n)\, G_{s-1}(r,n)$, and the second is $G_s(r,n-1)$.

Thus it remains to prove Eqs. (2) and (3) for F. Here we cannot escape some calculation. First we prove Eq. (2).

$$F_s(r,n) = \sum_{j=r}^{n} (-1)^{j+r} \binom{n}{j}\binom{j-1}{r-1}\frac{1}{j^s}$$

$$= \sum_{j=r}^{n-1} (-1)^{j+r} \binom{n-1}{j}\binom{j-1}{r-1}\frac{1}{j^s} + \sum_{j=r}^{n} (-1)^{j+r} \binom{n-1}{j-1}\binom{j-1}{r-1}\frac{1}{j^s}$$

$$= F_s(r,n-1) + \frac{1}{n}\sum_{j=r}^{n} (-1)^{j+r} \binom{n}{j}\binom{j-1}{r-1}\frac{1}{j^{s-1}}$$

$$= F_s(r,n-1) + \frac{1}{n} F_{s-1}(r,n)\,.$$

This proves Eq. (2) for F. We turn now to Eq. (3). We need to prove that

$$\phi(r,n) = \sum_{j=r}^{n} (-1)^j \binom{n}{j}\binom{j-1}{r-1} = (-1)^r\,.$$

To this end, let $\psi(r,n) = \sum_{j=r}^{n} (-1)^j \binom{n}{j}\binom{j}{r}$.

Then using

$$\binom{j}{r} = \binom{j-1}{r} + \binom{j-1}{r-1}$$

we quickly obtain that $\psi(r,n) = \phi(r,n) + \phi(r+1,n)$. Iterating this once, $\phi(r,n) = \phi(r+2,n)$.

We need only calculate $\phi(1,n)$ and $\phi(2,n)$.

$$\phi(1,n) = \sum_{j=1}^{n} (-1)^j \binom{n}{j}$$

$$= \sum_{j=0}^{n} (-1)^j \binom{n}{j} - 1 = (1-1)^n - 1 = -1\,.$$

$$\phi(2,n) = \sum_{j=2}^{n} (-1)^j \binom{n}{j}(j-1)$$

$$= 1 + \sum_{j=0}^{n} (-1)^j \binom{n}{j}(j-1)$$

$$= 1 + \sum_{j=0}^{n} (-1)^j \binom{n}{j} j - \sum_{j=0}^{n} (-1)^j \binom{n}{j}$$

$$= 1 + \sum_{j=0}^{n} (-1)^j \binom{n}{j} j = 1\,.$$

(This last sum may be computed by induction on n in the obvious manner. It occurs as a special case of an exercise in Feller [Ref. 10, p. 63, exercise 16].) This proves Eq. (3) for F and so completes the proof of the theorem.

3. Simpler Form

It is apparent that for large s, the calculation of G, though much simpler than the calculation of F, becomes forbiddingly complex. It is, in general, possible to put G into more manageable form. To this end, let us define

$$\zeta_s(r,n) = \sum_{j=r}^{n} \frac{1}{j^s}$$

Where no confusion will arise, we write ζ_s for $\zeta_s(r,n)$.

Now by a *partition* of the integer $n > 0$ we mean a collection of integers $n_i > 0$ such that $\Sigma n_i = n$. If $\lambda = (n_1,n_2,\cdots)$ represents such a partition, we write $\zeta_\lambda = \zeta_{n_1}\zeta_{n_2}\cdots$. [For example, if $n = 7$, $\lambda = (1,1,2,3)$, $\zeta_\lambda = \zeta_1^2\zeta_2\zeta_3$.] Also, if the integer j occurs in λ r_j times, we put

$$\mu(\lambda) = \frac{n!}{1^{r_1}2^{r_2}\cdots r_1!\, r_2!\cdots}\,.$$

Then it can be shown that $F_n = 1/n!\ \Sigma\mu(\lambda)\zeta_\lambda$, where the summation is extended over all partitions of the integer n. Thus, for example

$$F_1 = \zeta_1$$

$$F_2 = \frac{1}{2}(\zeta_1^2 + \zeta_2)$$

$$F_3 = \frac{1}{6}(\zeta_1^3 + 3\zeta_1\zeta_2 + 2\zeta_3)$$

$$F_4 = \frac{1}{24}(\zeta_1^4 + 6\zeta_1^2\zeta_2 + 3\zeta_2^2 + 8\zeta_1\zeta_3 + 6\zeta_4).$$

These and related formulas are proved using some techniques for Möbius Inversion which were developed by G. Solomon and me earlier (Ref. 11), for use in error-correcting coding work. (Also, see *SPS 37-35*, Vol. IV, pp. 340–348.) The proofs will not be given here.

References

1. G. N. Povarov, "Synthesis of Contact Multipoles," paper reviewed in *IRE Transactions on Circuit Theory,* March 1956.

2. Peterson, *Error Correcting Codes,* Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1961.

3. Berlekamp, E., "Practical BCH Decoders," to appear in *IEEE Transactions on Information Theory,* 1967.

4. Chien, "Cyclic Decoding Procedures for BCH Codes," *IEEE Transactions On Information Theory,* Vol. IT-10, No. 4, pp. 357–362, 1964.

5. Van der Waerden, *Modern Algebra,* Vol. I, F. Ungar Publishing Company, New York, New York, 1949.

6. Lucas, M. E., "Sur les congruences des nombres euleriennes et des coefficients differentiels des fonctions trigonometriques, suivant un-module premier," *Bulletin Societé Mathematiques,* France 6, pp. 49–54, 1878.

7. Gallager, R. G., "A Simple Proof of the Coding Theorem, With Applications," *IEEE Transactions on Information Theory,* 1965.

8. Shannon, C. E., Gallager, R. G., and Berlekamp, E. R., "Lower Bounds to the Probability of Error," *Information and Control,* (forthcoming).

9. Hietala, H. J., and Winter, B. B., "Note on a Combinatorial Identity," *Math. Mag.,* 38, pp. 149–151, 1965.

10. Feller, W., *An Introduction to Probability Theory and Its Applications,* 2nd edition, Vol. I, John Wiley & Sons, New York, 1961.

11. Solomon, G., and McEliece, R. J., "Weights of Cyclic Codes," *Journal of Combinatorial Theory* (to appear).

XX. Communications Systems Research: Efficient Data Systems

A. Carrier Suppression in Coherent Two-Way Communication Systems

W. C. Lindsey

1. Introduction

In this article we make use of the second moment theory of random processes to determine the effects which the random modulation existing on the vehicle's VCO output produces on the down-link carrier. This random modulation, say $\hat{\theta}_1(t)$, reduces the power (which remains in the down-link carrier component) below that value which would be obtained if the system were to transmit back to the reference system a clean carrier. In practice, this suppression of the down-link carrier may be measured near the end of a mission. We make use of the notation given in *SPS's 37-34* through *37-37*, Vol. IV, pp. 242, 339, 298 and 287, respectively.

2. Carrier Suppression (Linear PLL Theory)

The waveform transmitted back to the reference system is

$$\eta(t) = (2P_{c2})^{1/2} \sin\left[\omega_{10}t + \hat{\theta}_1(t)\right] \tag{1}$$

where $\omega_{10} = G\omega_1$, ω_1 is received carrier frequency from the up-link, G is the static phase gain in the vehicle system, and $\hat{\theta}_1(t)$ is the random phase modulation due to the additive noise on the up link. If we invoke the linear PLL theory, then the stationary process $\theta_1(t)$ is well approximated by a normal or Gaussian process with zero mean and a variance $\sigma^2_{\hat{\theta}_1} = N_{01}W_{L1}G^2/2P_{c1}$. If we denote the normalized covariance function of the phase modulation $\hat{\theta}_1(t)$ by $K_{\hat{\theta}_1}(\tau)$ with $K_{\hat{\theta}_1}(0) = 1$, it can be shown that the covariance function $K_\eta(\tau)$ of $\eta(t)$ is given by

$$K_\eta(\tau) = P_{c2} \exp\left[-\sigma^2_{\hat{\theta}_1}\left\{K_{\hat{\theta}_1}(0) - K_{\hat{\theta}_1}(\tau)\right\}\right] \cos\omega_{10}\tau\,. \tag{2}$$

At $\tau = 0$, we have $K_\eta(0) = P_{c2}$; i.e., as expected, the total down-link power is P_{c2}, inasmuch as the modulation is entirely in the phase term. The intensity of the carrier component is

$$I_c = K_\eta(\infty) = P_{c2} \exp\left(-N_{01}W_{L1}G^2/2P_{c1}\right) \tag{3}$$

and that of the remaining random component is

$$I = K_\eta(0) - K_\eta(\infty) = P_{c2}\left[1 - \exp\left(-N_{01}W_{L1}G^2/2P_{c1}\right)\right]. \tag{4}$$

This I represents the total power in the "self-noise" that is created by the random down-link phase modulation $\theta_1(t)$. It should be pointed out that in the ground receiver of a two-way system, this component of noise must be considered in addition to receiver noise when computing the threshold characteristic of the receiver. That is, the suppression of the down-link carrier alone does not reflect the complete picture, insofar as predicting what the receiver phase error will be. However, the effect we derive does predict what sort of signal strength (i.e., antenna gain control) variations may be expected in a two-way mode. For Gaussian phase modulation processes then, it is clear that $I_c > 0$, so that there is always a discrete carrier component, although it may represent a trivial fraction of the total power P_{c2} if $\sigma_{\hat{\theta}_1}$ is large.

Denote the carrier suppression factor S as the ratio of the power remaining in the carrier component when the up-link additive noise effects the down-link transmission to the power remaining in the carrier component if the down-link carrier is derived from a free-running oscillator in the vehicle. Thus, from Eq. (3) we have

$$S = \frac{I_c}{P_{c2}} = \exp\left(-N_{01}W_{L1}G^2/2P_{c1}\right) = \exp\left[-\sigma^2_{\hat{\theta}_1}\right]. \tag{5}$$

Plotted in Fig. 1 is the carrier suppression factor S versus $\sigma^2_{\hat{\theta}_1}$.

3. Carrier Suppression (Nonlinear PLL Theory)

If the phase modulation process is not Gaussian, then the covariance function $K_\eta(\tau)$ is not mathematically tractable; however, the carrier suppression factor can still be computed. The distribution of the phase estimate $\hat{\theta}_1$ is well approximated (in the range where PLL's are generally expected to operate as carrier tracking filters) by

$$p(\hat{\theta}_1) = \frac{\exp\left[\sigma^2_{\hat{\theta}_1}\cos\hat{\theta}_1\right]}{2\pi I_0\left(\sigma^2_{\hat{\theta}_1}\right)} \quad ; \quad |\hat{\theta}_1| \leq \pi \tag{6}$$

where $I_k(x)$ is the imaginary Bessel function of order k and argument x, and where $\sigma^2_{\hat{\theta}_1}$ is the phase variance as determined from the linear PLL theory, viz,

$$\sigma^2_{\hat{\theta}_1} = \frac{N_{01}W_{L1}G^2}{2P_{c1}}. \tag{7}$$

The waveform $\eta(t)$ may be written as

$$\eta(t) = (2P_{c2})^{1/2}\cos\hat{\theta}_1\cos\omega_{01}t + (2P_{c2})^{1/2}\sin\hat{\theta}_1\sin\omega_{01}t \tag{8}$$

from which it follows that the intensity of the carrier component is

$$I_c = \left[(P_{c2})^{1/2}\int_{-\pi}^{\pi}(\cos\hat{\theta}_1 + \sin\hat{\theta}_1)\,p(\hat{\theta}_1)\,d\hat{\theta}_1\right]^2 \tag{9}$$

which, through Eq. (6) reduces to

$$I_c = P_{c2}\left(\left[\overline{\cos\hat{\theta}_1}\right]\right)^2 = P_{c2}\left\{\frac{I_1\left(\sigma^2_{\hat{\theta}_1}\right)}{I_0\left(\sigma^2_{\hat{\theta}_1}\right)}\right\}^2. \tag{10}$$

Thus, the corresponding carrier suppressions are given by

$$S = \frac{I_c}{P_{c2}} = \left\{\frac{I_1\left(\sigma^2_{\hat{\theta}_1}\right)}{I_0\left(\sigma^2_{\hat{\theta}_1}\right)}\right\}^2. \tag{11}$$

The factor given in Eq. (11) is plotted in Fig. 1 for comparison with that obtained using the linear PLL theory. Note again that as $\sigma^2_{\hat{\theta}_1}$ approaches zero, the down-link carrier component is completely suppressed.

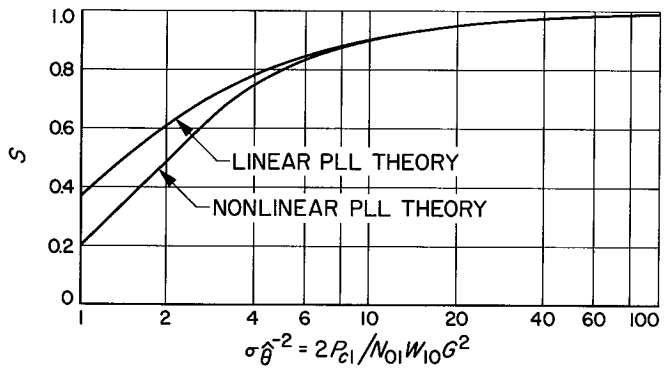


Fig. 1. Carrier suppression S vs the parameter x_1/G^2. Linear and nonlinear PLL theory is assumed, and no limiters are present. $x_1 = 2P_{c1}/N_{01}W_{L1}$ is the SNR existing in the vehicle carrier tracking loop

4. Carrier Suppression (Linear PLL Theory With Bandpass Limiters Preceding the Loop)

Next, let us assume that the bandwidth of the vehicle carrier tracking loop is small enough to have the process $\hat{\theta}_1(t)$ Gaussian. Then the variance of the process $\hat{\theta}_1(t)$ adequately characterizes the carrier suppression. Thus,

the variance of $\hat{\theta}_1$, assuming the presence of a limiter in the loop, is equal to (Ref. 1).

$$\sigma^2_{\hat{\theta}_1} = \frac{N_{01}W_{10}}{2P_{c1}} \cdot G^2\Gamma_1 \cdot \frac{1 + r_{10}/\mu_1}{1 + r_{10}} = 1/\alpha \tag{12}$$

in which Γ_1 is the limiter performance factor; $\mu_1 = \alpha_{11}/\alpha_{01}$ is the limiter suppression factor; B_{10} is the loop bandwidth at "threshold"; and r_{10} is determined by the damping in the vehicle system (Ref. 1, pp. 30–31). The carrier suppression is again given by

$$S = \exp\left(-\sigma^2_{\hat{\theta}_1}\right) \tag{13}$$

but $\sigma^2_{\hat{\theta}_1}$ is now defined by Eq. (12). This suppression is plotted in Fig. 2 versus $2P_{c1}/N_{01}B_{10}G^2$ for various system mechanizations. The parameter $2P_{c1}/N_{01}B_{10}$ is the signal-to-noise ratio (SNR) existing in the threshold loop bandwidth B_{10}.

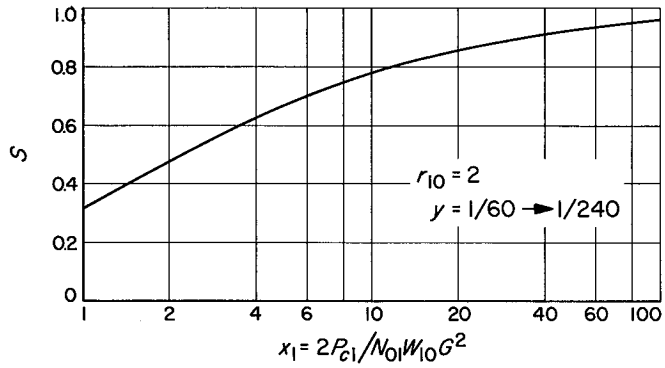


Fig. 2. Carrier suppression S vs the parameter x_1/G^2, as determined from the linear PLL theory with bandpass limiters present. $x_1 = 2P_{c1}/N_{01}W_{10}$ is the SNR existing in the vehicle carrier tracking loop

5. Carrier Suppression (Nonlinear PLL Theory With Bandpass Limiter Preceding the Loop)

In this case the covariance $K_\eta(\tau)$ is impossible to determine mathematically; however, on the basis of the results presented in *SPS 37-35*, Vol. IV, p. 339 we have

$$S = \left[\frac{I_1(\alpha)}{I_0(\alpha)}\right]^2 \tag{14}$$

And, as before, we note that the down-link carrier is completely suppressed when the up-link SNR approaches zero. Plotted in Fig. 3 is Eq. (14) versus $2P_{c1}/N_{01}W_{10}G^2$

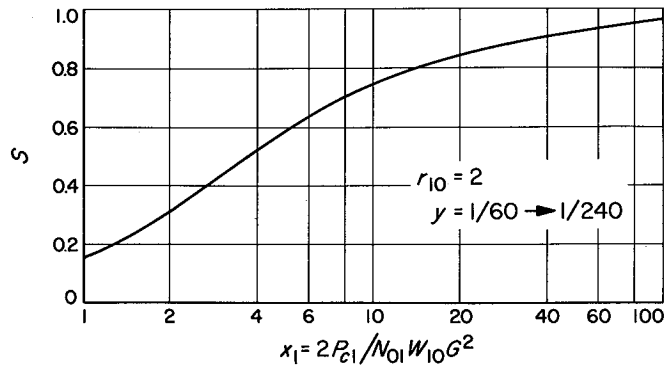


Fig. 3. Carrier suppression S vs the parameter x_1/G^2, as determined from the nonlinear PLL theory with bandpass limiters present. $x_1 = 2P_{c1}/N_{01}W_{10}$ is the SNR existing in the vehicle's carrier tracking loop

for various system mechanizations. As before, the parameter $2P/_{c1}/N_{01}W_{10}$ is the SNR existing in the threshold loop bandwidth W_{10}.

B. A Recursion Formula For Prefix Codes Over an *r*-ary Alphabet

J. J. Stiffler

1. Introduction

The prefix codes are one class of block length codes designed for use over the noiseless channel. Like comma-free codes, no nontrivial overlap of any two prefix code words is itself a code word. And prefix codes are further constrained by the requirement that all code words have the same m-symbol prefix and that this prefix occurs nowhere else within a code word or in the overlap of any two code words. This additional constraint enables the received sequence of symbols to be separated into code words as soon as a prefix is identified. Since the prefix can be considerably shorter than the code word, the synchronization of the decoder can presumably be simplified with these codes vis à vis the comma-free codes.

A recursion formula for the number of words in a prefix code dictionary was derived in Ref. 2 for the binary case. In this article results of this derivation are extended, by a different approach, to an arbitrary r-ary alphabet.

2. The Recursion Formula

If the prefix consists of the m symbols $\gamma_1\gamma_2\cdots\gamma_m$ and a code word is represented by $\gamma_1\gamma_2\cdots\gamma_m\alpha_1\alpha_2\cdots\alpha_n$, the prefix constraint is that none of the m-tuples

$$\gamma_2\gamma_3\cdots\gamma_m\alpha_1$$

$$\gamma_3\gamma_4\cdots\gamma_m\alpha_1\alpha_2$$

$$\vdots$$

$$\gamma_m\alpha_1\cdots\alpha_{m-1}$$

$$\alpha_1\alpha_2\cdots\alpha_m$$

$$\alpha_2\alpha_3\cdots\alpha_{m+1}$$

$$\vdots$$

$$\alpha_{n-m+1}\cdots\alpha_n$$

$$\vdots$$

$$\alpha_n\gamma_1\cdots\gamma_{m-1}$$

be equal to the prefix $\gamma_1\gamma_2\cdots\gamma_m$. Let $N_m(n)$ denote the maximum number of words in a prefix dictionary having the m-symbol prefix $\gamma_1\gamma_2\cdots\gamma_m$, and with words of total length $n+m$. We seek to demonstrate a recursive relationship on the terms $N_m(n)$.

A prefix is said to be repetitive with period ν ($\nu \leq m$) if $\gamma_1\gamma_2\cdots\gamma_\nu\gamma_1\gamma_2\cdots\gamma_{m-\nu} = \gamma_1\gamma_2\cdots\gamma_m$. Note that all prefixes of length m are repetitive with period m. The distinction between repetitive sequences and periodic sequences should be emphasized. All periodic sequences are repetitive, but not conversely. The nonperiodic sequence 101, for example, is repetitive with period two.

Lemma 1. If a prefix of length m is repetitive with periods ν_1 and ν_2, either $\nu_1+\nu_2 > m$ or $\nu_1+\nu_2$ is also a repetitive period of the prefix.

Proof. By definition

$$\begin{aligned}\gamma_1\gamma_2\cdots\gamma_m &= \gamma_1\gamma_2\cdots\gamma_{\nu_1}\gamma_1\cdots\gamma_{m-\nu_1}\\ &= \gamma_1\gamma_2\cdots\gamma_{\nu_1}\gamma_1\cdots\gamma_{\nu_2}\gamma_1\cdots\gamma_{m-(\nu_1-\nu_2)}\\ &= \gamma_1\cdots\gamma_{\nu_1+\nu_2}\gamma_1\cdots\gamma_{m-(\nu_1+\nu_2)}\,.\end{aligned}$$

Thus, if $\nu_1+\nu_2 \leq m$, $\nu_1+\nu_2$ is a repetitive period.

Lemma 2. If ν_1 and ν_2 are repetitive periods of an m-symbol prefix $\gamma_1\gamma_2\cdots\gamma_m$, and if $\gamma_1\gamma_2\cdots\gamma_{\nu_2} = \gamma_1\gamma_2\cdots\gamma_{\nu_2-\nu_1}\gamma_1\gamma_2\cdots\gamma_{\nu_1}$, $\nu_2-\nu_1$ is also a repetitive period.

Proof. The following three m-tuples are identical

$$\gamma_1\gamma_2\cdots\cdots\cdots\cdots\gamma_m$$

$$\gamma_1\gamma_2\cdots\cdots\cdots\gamma_{\nu_2}\gamma_1\cdots\gamma_{m-\nu_2}$$

$$\gamma_1\gamma_2\cdots\gamma_{\nu_2-\nu_1}\gamma_1\cdots\gamma_{\nu_1}\gamma_1\cdots\gamma_{m-\nu_2}\,.$$

But $\gamma_1\cdots\gamma_{\nu_1}\gamma_1\cdots\gamma_{m-\nu_2} = \gamma_1\cdots\gamma_{m-(\nu_2-\nu_1)}$ and the lemma is proved.

Now we are ready to prove a recursion on the number of works in the maximal prefix dictionaries.

Theorem 1. Let $\gamma_1\gamma_2\cdots\gamma_m$, the prefix of the $(n+m)$-symbol r-ary words of a prefix dictionary, have repetitive periods $\nu_1, \nu_2, \cdots, \nu_l, \nu_{l+1} = m$. Define Δ_i such that

$$\Delta_i = \begin{cases}1 & i = 0 \text{ or } i = \nu_j \text{ for some } j,\ 1 \leq j \leq l+1\\ 0 & \text{otherwise}\end{cases}$$

Then the number of code words $N_m(n)$ in the maximal dictionary is given, for $n > m$, by the recursion

$$N_m(n) = \sum_{i=1}^{m} (r\Delta_{i-1} - \Delta_i)\, N_m(n-1)\,. \qquad (1)$$

Proof. Let $D_m(n)$ represent the maximal prefix dictionary with an m-symbol prefix and $(n+m)$-symbol words; let $w_m(n)$ denote any word in $D_m(n)$; and let the set $\{w_m(n)\,\alpha_1\alpha_2\cdots\alpha_k\}$ be the set of all words in $w_m(n)$ followed by the k-tuple $\alpha_1\alpha_2\cdots\alpha_k$. Further, let $N_m(n)$ be the number of words in $D_m(n) = \{w_m(n)\}$. Then

$$N_m(n) = rN_m(n-1) - N(S_1) + N(S_2), \qquad (2)$$

where S_1 is the set of $(m+n)$-tuples in some set $\{w_m(n-1)a\}$ for some symbol a which are not in the set $\{w_m(n)\}$; S_2 is the set of $(m+n)$-tuples in $\{w_m(n)\}$ but not in $\{w_m(n-1)a\}$ for any a; and $N(S_i)$ is the number of elements in the set S_i, $i = 1,2$. This statement (2) follows because the union of the sets $\{w_m(n-1)a\}$ (where a assumes all r possible values) contains $rN_m(n-1)$ elements. We first determine $N(S_1)$:

(1) All words in $\{w_m(n-1)a\}$ not in $\{w_m(n)\}$ end with some suffix $\gamma_1\gamma_2\cdots\gamma_{\nu_i}$ for some repetitive period $\nu_i \leq m$. The only other possible factor preventing some word $w_m(n-1)a$ from being in $\{w_m(n)\}$ would be the occurrence of the entire m-tuple $\gamma_1\gamma_2\cdots\gamma_m$ at some point in $w_m(n-1)$ other than as a prefix. But since $\{w_m(n-1)\} = D_m(n-1)$ this cannot happen, if $w_m(n-1)a$ ends with $\gamma_1\gamma_2\cdots\gamma_{\nu_i-1}$. Thus, words ending with $\gamma_1\gamma_2\cdots\gamma_{\nu_i}$ will be in $w_m(n-1)a$ only if ν_i-1 is not also a period of repetition of the prefix.

(2) All words in $\{w_m(n-1)\}$ ending with $\gamma_1\gamma_2\cdots\gamma_{\nu_{i-1}}$ with ν_i-1 not a period of repetition and with no longer sequence $\gamma_1\gamma_2\cdots\gamma_{\nu_{j-1}}$, for any $\nu_j > \nu_i$, are in the set $\{w_m(n-\nu_i)\gamma_1\cdots\gamma_{\nu_{i-1}}\}$ and conversely. For any word in $\{w_m(n-\nu_i)\gamma_1\cdots\gamma_{\nu_{i-1}}\}$ and not in $\{w_m(n-1)\}$ must contain the prefix beginning at some internal point (other than as its actual prefix). The prefix cannot be contained entirely in the first $m+n-\nu_i$ symbols since $w_m(n-\nu_i)$ is a word in $D_m(n-\nu_i)$. Thus, such a word $w_m(n-\nu_i)$ must end with the sequence $\gamma_1\gamma_2\cdots\gamma_{\nu_j}$ for some period of repetition ν_j. But this too is impossible, since $w_m(n-\nu_i)$ is in $D_m(n-\nu_i)$. Thus, all words in $\{w_m(n-\nu_i)\nu_1\cdots\gamma_{\nu_{i-1}}\}$ are in $\{w_m(n-1)\}$.

If there is a word in $\{w_m(n-1)\}$ ending with the suffix $\gamma_1\cdots\gamma_{\nu_{i-1}}$ but with no longer suffix $\gamma_1\gamma_2\cdots\gamma_{\nu_{j-1}}$ which is not of the form $w_m(n-\nu_i)\gamma_1\cdots\gamma_{\nu_{i-1}}$, the word $w_m(n-\nu_i)$ must end with the suffix $\gamma_1\gamma_2\cdots\gamma_{\nu_l}$ for some period of repetition ν_l. If this were not true, deleting the last ν_i-1 symbols from some words in $\{w_m(n-1)\}$ ending with $\gamma_1\cdots\gamma_{\nu_{i-1}}$ would have to produce a word $w_m(n-\nu_i)$ in $D_m(n-\nu_1)$, since the latter could not contain the prefix internally. But if $w_m(n-\nu_i)$ did end with the suffix $\gamma_1\gamma_2\cdots\gamma_{\nu_l}$ (and as a consequence was not in $w_m(n-\nu_i)$), then ν_l and ν_i would both be periods of repetition of the prefix, and by lemma 1, so would their sum (unless $\nu_l+\nu_i > m$). The word $w_m(n-1)$ would have as a suffix the $(\nu_l+\nu_i-1)$-tuple $\gamma_1\gamma_2\cdots\gamma_{\nu_l}\gamma_1\gamma_2\cdots\gamma_{\nu_{i-1}} = \gamma_1\gamma_2\cdots\gamma_{\nu_l+\nu_{i-1}}$, violating the stipulation that it ended with no suffix of the form $\gamma_1\gamma_2\cdots\gamma_{\nu_{j-1}}$ for any $\nu_j > \nu_i$. (Or, if $\nu_l+\nu_i > m$, the suffix of $w_m(n-1)$ would contain the forbidden prefix.) Accordingly, all words in $\{w_m(n-1)\}$ of the specified form are also in $w_m(n-\nu_i)\gamma_1\cdots\gamma_{\nu_{i-1}}$.

(3) All sets $\{w_m(n-\nu_i)\gamma_1\gamma_2\cdots\gamma_{\nu_{i-1}}\}$ are disjoint. Suppose, to the contrary, that $w_m(n-\nu_i)\gamma_1\gamma_2\cdots\gamma_{\nu_{i-1}} = w_m(n-\nu_j)\gamma_1\cdots\gamma_{\nu_{j-1}}$ with $\nu_i < \nu_j$. Then $\gamma_1\gamma_2\cdots\gamma_{\nu_{j-1}} = \alpha_1\alpha_2\cdots\alpha_{\nu_j-\nu_i}\gamma_1\gamma_2\cdots\gamma_{\nu_i}$ where $\alpha_1\alpha_2\cdots\alpha_{\nu_j-\nu_i} = \gamma_1\cdots\gamma_{\nu_j-\nu_i}$ is the suffix of $w_m(n-\nu_i)$. But, by lemma 2, if ν_j and ν_i are periods of repetition so also is $\nu_j-\nu_i$, and $w_m(n-\nu_i)$ could not be in the set $\{w_m(n-\nu_i)\} = D_m(n-\nu_i)$.

(4) Defining $\delta_i = \begin{cases} 1 & \text{if } i=\nu_j \text{ is a repetitive period and } \nu_j-1 \text{ is not} \\ 0 & \text{otherwise} \end{cases}$

we have

$$N(S_1) = \sum_{i=1}^{m} \delta_i\, N_m(n-i)\,.$$

For, beginning with the longest repetitive period ($\nu_j=m$), we count $N_m(n-\nu_j)$ terms in S_1 ending with $\gamma_1\gamma_2\cdots\gamma_{\nu_j}$, if ν_j-1 is not a period of repetition, and none otherwise. Continuing through the shorter periods, paragraphs (1) and (2) assert that we count all elements in S_1 once, and from paragraph (3) only once.

Essentially, the same four steps may be repeated to count the number of elements in S_2:

(5) Any word in $\{w_m(n)\}$ and not in $\{w_m(n-1)a\}$ must end with the suffix $\gamma_1\gamma_2\cdots\gamma_{\nu_i}a$ for some period of repetition ν_i, since any other such word would be in $\{w_m(n-1)\}$ upon deleting the last symbol.

(6) All words $\{w_m(n)\}$ ending with $\gamma_1\gamma_2\cdots\gamma_{\nu_i}a$, and with no longer sequence $\gamma_1\gamma_2\cdots\gamma_{\nu_j}a$, $\nu_j > \nu_i$, are in some set $\{w_m(n-\nu_i-1)\,\gamma_1\gamma_2\cdots\gamma_{\nu_i}a\}$ and conversely, where a can be any symbol if ν_i+1 is not a period of repetition and can be any symbol but $\gamma_{\nu_{i+1}}$ if it is. The proof is identical to that of paragraph (2).

(7) All sets $\{w_m(n-\nu_i-1)\,\gamma_1\gamma_2\cdots\gamma_{\nu_i}a\}$ are disjoint. See paragraph (3).

(8) $N(S_2) = \sum_{i=1}^{m-1}[r-(1-\delta_{i+1})]\;N_m(n-i-1)$. This follows from paragraphs (5), (6), and (7) (see paragraph (4)). If $i+1=\nu_j+1$ is a repetitive period ($\delta_{i+1}=0$), there are only $r-1$ possible values for a, if the word $w_m(n-\nu_i-1)\,\gamma_1\gamma_2\cdots\gamma_{\nu_i}a$ is to be in $D_m(n)$. If $i+1=\nu_j+1$ is not a repetitive period ($\delta_{i+1}=1$), there are r possible values for a.

Thus, we have

$$N_m(n) = rN_m(n-1) - \sum_{i=1}^{m}\delta_i\,N_m(n-i) + \sum_{i=1}^{m-1}[r-(1-\delta_{i+1})]\,N_m(n-i-1)\,. \tag{3}$$

But if $\delta_i = 0$, $\nu_j - 1 = \nu_{j-1}$, and since δ_1 must be 1 by definition,

$$\sum_{i=1}^{l}(1-\delta_{i+1})\,N_m(n-\nu_i-1) = \sum_{j=2}^{l+1}(1-\delta_j)\,N_m(n-\nu_{j-1}-1) = \sum_{j=1}^{l+1}(1-\delta_j)\,N_m(n-\nu_j)\,.$$

Consequently,

$$N_m(n) = rN_m(n-1) - \sum_{i=1}^{l+1} (\delta_i + 1 - \delta_i)\, N_m(n - \nu_i) + r \sum_{i=1}^{l} N_m(n - \nu_i - 1)\,. \tag{4}$$

Defining $\Delta_i = \begin{cases} 1 & i = 0 \text{ or } i = \nu_j \text{ for some } j,\ 1 \le j \le m \\ 0 & \text{otherwise} \end{cases}$, it follows that

$$N_m(n) = rN_m(n-1) - \sum_{i=1}^{m} \Delta_i\, N_m(n-i) + r \sum_{i=2}^{m} \Delta_{i-1}\, N_m(n-i)$$

$$= \sum_{i=1}^{m} (r\,\Delta_{i-1} - \Delta_i)\, N_m(n-i)\,, \tag{5}$$

which was to be proved.

C. Optimum Word Synchronization

J. W. Layland

1. Introduction

The optimization of binary communication links with respect to carrier/subcarrier/and bit synchronization pilot tone(s) has been carried out in several contexts (Refs. 3–5). This article is concerned with the optimization of a system in which word synchronization is obtained by use of a properly designed pilot tone and with the comparison of such a system to a self-synchronizing one-channel system.

The analysis given here applies whenever timing for the data channel must be obtained independently of any ranging to be done with the same signal. It does not apply when the data channel is synchronous with and of secondary importance to a ranging channel.

The optimum allocation of power is determined between a message-bearing signal and a pilot tone, from which tone word synchronization for that message signal is derived. The analysis assumes that bit-timing is known, and considers that power allocation optimum which minimizes the overall message error probability under conditions of broad-band Gaussian noise and correlation detection. Results are obtained for either coded or uncoded message signals and for either delay-lock tracking or maximum-likelihood detection of the sync-signal. Under the criterion used, asymptotic large-signal results are valid for all signal-to-noise ratio (SNR) of interest in most cases considered. Numerical results are given for the one case in which this does not hold.

The main conclusion of this article is that for a *Voyager*-class telemetry system, the best way of obtaining synchronization is by the use of self-synchronizing block codes.

2. Optimum Two-Channel System

It is assumed in the following that synchronization of carrier, subcarrier, and symbol-pulses is known; that word synchronization is the only unknown level of timing. (This condition is somewhat relaxed later.) The received signal is constrained to a total power P, αP of which is message-bearing signal and $(1 - \alpha)P$ of which is the word-synchronizing pilot tone. The message consists of words of W pulses each, such that each pulse lasts T_P sec, and the modulating signal takes value either $+1$ or -1 during any given pulse. The pilot signal is similarly constructed such that it has the two-level autocorrelation function of a pseudo-noise sequence. Both signals are subject to statistically identical noises.

Four cases are of interest, the combinations of the following two conditions on the message signal and modes of timing detection.

Message signal.

A. Each pulse represents one bit of information. The message is detected by a bit-by-bit maximum-likelihood detector with $T_b = T_P$.

B. All $2^k - 1$ pulses represent one k-bit transorthogonal word. The message is detected by a word-by-word maximum-likelihood detector with $T_b = (2^k - 1)/k\ T_p$.

Timing detection.

C. Timing is obtained by a maximum-likelihood detector for the phase shift of the pilot tone, using as long an integration time as is practicable with the channel involved.

D. Timing, having once been obtained, is maintained by the delay-lock loop used for the tracking of a ranging signal. The tracking-loop bandwidth is made as small as practicable with the channel involved.

In each of the four cases, AC, AD, BC, and BD, the best value of α is that value which minimizes the expected error probability in message detection, using $\overline{P}_\varepsilon = E_{\text{timing}}\{P_\varepsilon|\text{timing}\}$.

The timing error is a random process with mean zero and autocorrelation function similar to that of the channel's incremental timing variations. As such, it can be characterized by a sequence of independent samples spaced T_c apart. Each sample dominates the actual (continuous) timing error for a duration which averages T_c. Since the bit-timing is assumed known, and the word-timing instant is selected as that bit-instant nearest the continuous word-timing estimate, word-timing will either be correct or incorrect during the time each timing sample dominates the timing error. Hence,

$$\overline{P}_\varepsilon = 1 - \Pi_i E_{\tau_i}\{P_r\{\text{no error during } T_c|\tau_i\}\} \quad (1)$$

where it has been assumed that an approximately integral number of message-decisions occur during T_c. When the τ_i are treated as if independent and identically distributed, minimizing $\overline{P}_\varepsilon$ is equivalent to minimizing

$$\overline{P}_\varepsilon = 1 - E_\tau\{P_r\{\text{no error during } T_c|\overline{\tau}\}\} \quad (2)$$

If $P_\varepsilon^m|\tau$ is the error probability of each message decision conditioned by the word-timing and if message decisions are made independently of each other, once each T_m sec, Eq. (2) becomes

$$\overline{P}_e = 1 - E_\tau\{(1 - P_\varepsilon^m|\tau)^{T_c/T_m}\} \quad (3)$$

Case A D. The average error probability is quite simply derived in this case. If word-timing is correct, the conditional bit-error probability is just the bit-error probability with timing assumed, or

$$P_\varepsilon^b|_{\text{on time}} = \int_{\left(\frac{\alpha P T_b}{N_2}\right)^{1/2}}^{\infty} N_{01}(\eta)\, d\eta \quad (4)$$

Where $N_{\mu,\sigma^2}(x)$ designates the normal density function, T_b is the bit-pulse duration, and N_2 is the (two-sided) spectral density of the noise, assumed Gaussian and spectrally flat over the region of interest. If word-timing is not correct, the detector's output is independent of the bit supposedly being received, and

$$P_\varepsilon^b|_{\text{off-time}} = \frac{1}{2} \quad (5)$$

Because bit-timing is known, the detected bit is that one which is closest to the (continuous) timing estimate produced by the tracking filter. Fokker-Planck techniques (Ref. 6) can be applied to determine the probability density $P(t)$ of the error t of this estimate. For a first-order loop, this density is

$$P(t) = \begin{cases} k \exp\left\{-\frac{S}{2}\left(2\frac{t}{T_P}\right)^2\right\} & |t| \le \frac{T_P}{2} \\ k \exp\left\{-S\left(\frac{-5}{4} + 2\left(2\frac{t}{T_P}\right) - \frac{1}{4}\left(2\frac{t}{T_P}\right)^2\right)\right\} & \frac{T_P}{2} \le |t| \le 3\frac{T_P}{2} \\ k \exp\left\{\frac{-5}{2}S\right\} & 3\frac{T_P}{2} \le |t| \le \left(W - \frac{3}{2}\right)T_P \end{cases} \quad (6)$$

where k is a normalizing constant, W is the word-length, and

$$S = \frac{(1-\alpha)P}{2\,B_L\,N_2}\,\frac{W+1}{2\,W}$$

B_L = tracking loop bandwidth.

For reasonably strong signals, Eq. (4) can be approximated closely by a Gaussian density

$$P(t) \approx N_{0,\sigma^2}(t)$$

$$\sigma^2 = \frac{1}{S}\left(\frac{T_P}{2}\right)^2 \tag{7}$$

The probability that the detector is "off-time" is the probability that the timing estimator is in error by more than $T_P/2$, or

$$Pr\,\{\text{off-time}\} = 2\int_{\frac{T_P}{2}}^{\infty} N_{0,\sigma^2}(\tau)d\tau$$

$$= 2\int_{S^{1/2}}^{\infty} N_{01}(\tau)d\tau \tag{8}$$

The expected message error probability is

$$\bar{P}_e = 1 - 2^{-\frac{T_c/}{T_b}} - 2\int_0^{S^{1/2}} N_{01}(\tau)d\tau \left\{\left(\int_{-\infty}^{\left(\frac{\alpha PT}{N_2}\right)^{1/2}} N_{01}(\eta)d\eta\right)^{\frac{T_c}{T_b}-\frac{T_c}{T_b}} - 2\right\} \tag{9}$$

Now for a convenient notation, let

$$\rho \triangleq \left(\frac{PT_b}{N_2}\right)^{1/2}$$

$$\beta \triangleq \left(\frac{1}{2B_LT_b}\right)^{1/2} = \left(\frac{T_c}{T_b}\right)^{1/2}$$

$$\gamma \triangleq \left(\frac{W+1}{W}\right)^{1/2} \tag{10}$$

Then

$$P_e = 1 - 2^{-\beta^2} - 2\int_0^{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}} N_{01}(\tau)d\tau \left[\left(\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{\beta^2} - 2^{-\beta^2}\right] \tag{11}$$

This expression for $\bar{P}_e$ is unimodal, taking a maximum value of $1-2^{-\beta^2}$ at $\alpha=0$ and at $\alpha=1$, and possessing a single minimum on the open interval $0 < \alpha < 1$. This minimum can be determined by solving $\bar{P}'_e = (d\,\bar{P}_e/d\,\alpha) = 0$.

$$\bar{P}'_e = \rho^2\beta^2\int_0^{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}} N_{01}(\tau)dt\left[\left(\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{\beta^2} - 2^{-\beta^2}\right]$$

$$\times\left\{\gamma^2\frac{1}{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}}\,\frac{N_{01}\left(\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}\right)}{2\int_0^{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}} N_{01}(\tau)d\tau} - \frac{1}{\rho\alpha^{1/2}}\,\frac{N_{01}(\rho\alpha^{1/2})}{\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta}\,\frac{1}{1-\left(2\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{-\beta^2}}\right\} \tag{12}$$

If bit error probability is to be even moderately small

$$2\int_0^{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}} N_{01}(\tau)d\tau \approx \int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta \approx 1$$

and if the word-length is more than a few symbols, $\gamma^2 \approx 1$. Under these conditions, $\overline{P}'_e$ can be rewritten in the form

$$\overline{P}_e = C\left\{\frac{(2)^{1/2}}{\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}\left(\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}\right) - \frac{1}{\rho\alpha^{1/2}} N_{01}(\rho\alpha^{1/2})\right\} \tag{13}$$

where "C" is the positive coefficient external to the brackets of Eq. (12). The value α^* of α which minimizes $\overline{P}_e$ is, therefore,

$$\alpha^* = \frac{1}{1+\frac{2}{\gamma^2\beta^2}} \tag{14}$$

and the solution is valid over all SNR of interest.

Case B D. The timing probabilities are exactly as derived in the previous case. Message error probabilities for the transorthogonal code set are given in Ref. 7 as

$$P_\varepsilon^m\Big|_{\text{on time}} = 1 - \int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\left(\frac{k\alpha PT_b}{\frac{N}{2B}}\times\frac{2^k}{2^{k-1}}\right)^{1/2}} N_{01}(\xi)\,d\xi\right]^{2^k-1} d\eta \tag{15}$$

The actual error probability "off-time" is a complicated function of the time-shift and of the adjacent word. Rather than obtain a rigorous result which is specific to a given dictionary, we assume that

$$P_\varepsilon^m\Big|_{\text{off time}} = 1 - 2^{-k} \tag{16}$$

The effects of this assumption will be examined later. In the notation of Eq. (10), the expected error probability is

$$\overline{P}_e = 1 - 2^{-\beta^2} - 2\int_0^{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}} N_{01}(\tau)d\tau \left\{\left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{0_1}(\xi)d\xi\right]^{2^k-1}\right)^{\frac{\beta^2}{k}} - 2^{-\beta^2}\right\} \tag{17}$$

$\overline{P}_e$ assumes its maximum value of $1 - 2^{-\beta^2}$ at $\alpha = 0$ and at $\alpha = 1$, is strictly less than this for $0 < \alpha < 1$, and is unimodal in this interval. The minimizing α is the solution to

$$\begin{aligned} 0 = \frac{d\overline{P}_e}{d\alpha} &= \frac{1}{2}(2)^{1/2}\rho\beta\gamma(1-\alpha)^{1/2} N_{01}\left(\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}\right) \\ &\times\left\{\left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-1} d\eta\right)^{\frac{\beta^2}{k}} - 2^{-\beta^2}\right\} \\ &- \int_0^{\rho\beta\gamma\left(\frac{1-\alpha}{2}\right)^{1/2}} N_{01}(\tau)d\tau \left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-1} d\eta\right)^{\frac{\beta^2}{k}-1} \\ &\times \frac{\beta^2}{k}(2^k-1)\rho\gamma\left(\frac{k}{\alpha}\right)^{1/2} N_{02}(\rho\gamma(k\alpha)^{1/2}) \int_{-\infty}^{\infty} N_{0\frac{1}{2}}(\theta)\left[\int_{-\infty}^{\theta+\frac{1}{2}\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-2} d\theta \end{aligned} \tag{18}$$

As ρ becomes very large, the solution α^* for α in Eq. (18) approaches

$$\alpha^* = \frac{1}{1 + \frac{k}{\beta^2}} \tag{19}$$

while for moderately large ρ, the solution is approximately

$$\alpha \approx 1 - \frac{1}{1 + \frac{\beta^2}{k}} \left(1 - \frac{4\,ln\,2}{\gamma^2 \rho^2}\right) \tag{20}$$

which is only slightly greater than α^* for most ρ of interest.

For $k = 1 = 2^k - 1 = W$, Eqs. (19) and (14) are identical, while for larger k, considerably more importance is given to the timing of the coded system than to the one without coding. The assumption made earlier, that $P_\varepsilon^m|_{\text{off time}} = 1 - 2^{-k}$, has no effect on the results of Eq. (19) or (20), since for any transorthogonal dictionary, $P_\varepsilon^m|_{\text{off time}} \geq 1 - 2^{-k}$, and the 2^{-k} term appears only as $2^{-\beta^2}$ in Eq. (18), which is negligible for all SNR interest.

Case A C. The conditional error probabilities for this case are the same for case A D, but the distribution of timing must still be determined. The W cyclic phase-shifts of the pilot tone are exactly W of the $W + 1$ words of a transorthogonal code. The probability of being "on time" is exactly the probability of correct decision using a transorthogonal code and knowing that one word is never sent, or

$$Pr\{\text{on time}\} = \int_{-\infty}^{\infty} N_{0_1}(\tau) \left[\int_{-\infty}^{\tau+\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\theta)d\theta\right]^{W-1} d\tau \tag{21}$$

where again, the maximum-likelihood detector has averaged over as long a period as the channel stability will allow.

The expected error probability is then

$$P_e = 1 - 2^{-\beta^2} - \int_{-\infty}^{\infty} N_{0_1}(\tau) \left[\int_{-\infty}^{\tau+\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\theta)d\theta\right]^{W-1} d\tau \times \left\{\left(\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{\beta^2} - 2^{-\beta^2}\right\} \tag{22}$$

As before, P_e takes a maximum value of $1 - 2^{-\beta^2}$ at $\alpha = 0$ but now is slightly less than this at $\alpha = 1$; P_e possesses a single minimum in the open interval $0 < \alpha < 1$. Solving $dP_e/d\alpha = 0$ gives this minimum

$$P'_e = (W-1)\frac{1}{2}\rho\beta\gamma(1-\alpha)^{-1/2} N_{02}(\rho\beta\gamma(1-\alpha)^{1/2}) \int_{-\infty}^{-\infty} N^{0\frac{1}{2}}(\tau) \left[\int_{-\infty}^{\tau+\frac{1}{2}\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{W-2} d\tau \tag{23}$$

$$\times \left\{\left(\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{\beta^2} - 2^{-\beta^2}\right\}$$

$$- \int_{-\infty}^{\infty} N_{01}(\tau) \left[\int_{-\infty}^{\tau+\frac{1}{2}\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{W-1} d\tau \quad \beta^2 \left(\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{\beta^2-1} \frac{1}{2}\,\rho\alpha^{-1/2} N_{01}(\rho\alpha^{1/2})$$

For large ρ, P'_e is approximately given by

$$P'_e|_{\text{lge }\rho} \approx C\left\{(W-1)\frac{1}{2}(2)^{1/2}\frac{\gamma}{\beta}\left(\frac{\alpha}{1-\alpha}\right)^{1/2} e^{\frac{\rho^2}{2}\left[\alpha - \frac{\beta^2\gamma^2}{2}(1-\alpha)\right]} - 1\right\} \tag{24}$$

where C is a positive coefficient. As ρ becomes extremely large, the sign of (24) is the sign of $[\alpha - \frac{1}{2}\beta^2\gamma^2(1-\alpha)]$, and the zero of (24) approaches

$$\alpha^* = \frac{1}{1 + \frac{2}{\beta^2\gamma^2}} \tag{25}$$

or exactly the same result as obtained with the delay-lock timing. However, for moderately large ρ, α is approximately given by

$$\alpha \approx 1 - \frac{1}{1 + \frac{\beta^2\gamma^2}{2}}\left(1 + \frac{2}{\rho^2}\, ln\left(\frac{2}{\gamma^2}\,[W-1]\right)\right) \tag{26}$$

which should be considerably smaller than α^* for most ρ and W of interest. The solution, in fact, seems to approach $\alpha = 1/(1+\log_2 W)/(\beta^2)^{-1}$ for small values of ρ.

By means of the substitution

$$\alpha = \frac{1}{1 + \theta\,\frac{\log_2 W}{\beta^2}} \tag{27}$$

the dependence of α upon β^2 can be made implicit and the numerical solution simplified by the removal of one free parameter. If $\rho\alpha^{1/2} > 0$ and β^2 is large enough that

$$\left(2\int_{-\infty}^{\rho\alpha^{1/2}} N_{01}(\eta)d\eta\right)^{\beta^2} >> 1\,,$$

and if W is large enough that $\gamma^2 \approx 1$, substitution of Eq. (27) into Eq. (23) and some manipulation leads to

$$P_e' = C\,\frac{d}{dx}\bigg|_{x=\rho\alpha^{1/2}} ln\left\{\frac{\int_{-\infty}^{\infty} N_{01}(\tau)\left[\int_{-\infty}^{\tau+x(\theta\log_2 W)^{1/2}} N_{01}(\xi)d\xi\right]^{W-1} d\tau}{\left(\int_{-\infty}^{x} N_{01}(\eta)d\eta\right)^{\theta\log_2 W}}\right\} \tag{28}$$

where C is a positive constant. As both integrals in Eq. (28) have been tabulated, the solution for the optimal θ as a function of $\rho\alpha^{1/2}$, and W is fairly straightforward. Fig. 4 shows the optimal θ, θ^* as a function of $\rho^2\alpha/2$ for several values of $\log_2 W$. The optimal α^* can be readily determined from this curve using Eq. (27).

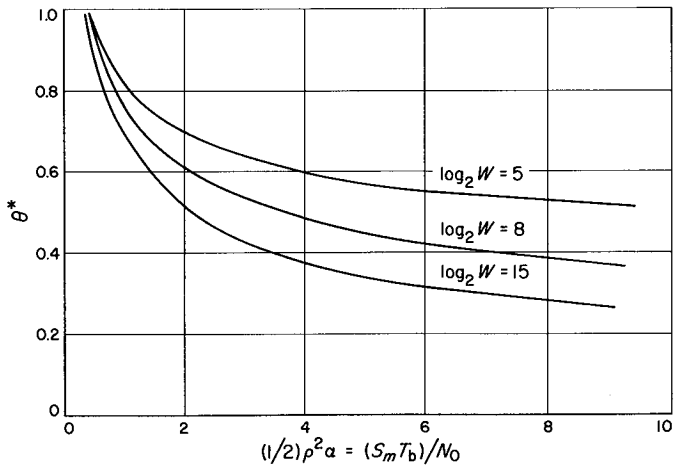


Fig. 4. Optimal θ vs $\frac{1}{2}\rho^2\alpha$ for $\log_2 W = 5$, 8, and 15

Case BC. Each of the terms which make up $\overline{P}_e$ has been computed with one of the previous cases. Thus, $\overline{P}_e$ may be immediately written as

$$\overline{P}_e = 1 - 2^{-\beta^2} - \int_{-\infty}^{\infty} N_{01}(\tau)\left[\int_{-\infty}^{\tau+\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-2} d\tau \tag{29}$$

$$\times\left\{\left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^{k-1}} d\eta\right)^{\frac{\beta^2}{k}} - 2^{-\beta^2}\right\}$$

This, as before, is unimodal in α over (0, 1); and, hence, the solution to $d\overline{P}_e/d\alpha = 0$ gives the minimizing α

$$\overline{P}'_e = (2^k-2)\tfrac{1}{2}\rho\beta\gamma(1-\alpha)^{-1/2} N_{02}(\rho\beta\gamma(1-\alpha)^{1/2}) \int_{-\infty}^{\infty} N_{0\frac{1}{2}}(\tau)\left[\int_{-\infty}^{\tau+\frac{1}{2}\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-3} d\tau$$

$$\times\left\{\left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^{k-1}} d\eta\right)^{\frac{\beta^2}{k}} - 2^{-\beta^2}\right\}$$

$$-\frac{\beta^2}{k}(2^k-1)\frac{\rho\gamma}{2}\left(\frac{k}{\alpha}\right)^{1/2} N_{02}(\rho\gamma(k\alpha)^{1/2}) \int_{-\infty}^{\infty} N_{0\frac{1}{2}}(\eta)\left[\int_{-\infty}^{\tau+\frac{1}{2}\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-2} d\eta$$

$$\times\left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^{k-1}} d\eta\right)^{\frac{\beta^2}{k}-1}$$

$$\times\int_{-\infty}^{\infty} N_{01}(\tau)\left[\int_{-\infty}^{\tau+\rho\beta\gamma(1-\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^k-2} d\tau \tag{30}$$

For any ρ of interest, word error probability will be considerably less than $1 - 2^{-k}$ and

$$\left(\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma(k\alpha)^{1/2}} N_{01}(\xi)d\xi\right]^{2^{k-1}} d\eta\right)^{\frac{\beta^2}{k}} >> 2^{-\beta^2}$$

Furthermore, for most useful codes, k is large enough that $2^k - 1 \approx 2^k - 2 \approx 2^k - 3$. Under these two conditions, Eq. (30) may be rewritten in the form

$$\overline{P}'_e = C\left\{F(k\alpha, \beta^2(1-\alpha)) - F(\beta^2(1-\alpha), k\alpha)\right\} \tag{31}$$

where "C" is a positive coefficient and

$$F(x,y) = y^{-1/2} N_{02}(\rho\gamma y^{1/2}) \int_{-\infty}^{\infty} N_{0\frac{1}{2}}(\tau)\left[\int_{-\infty}^{\tau+\frac{1}{2}\rho\gamma y^{1/2}} N_{01}(\theta)\,d\theta\right]^{2^k} d\tau$$

$$\times\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\gamma x^{1/2}} N_{01}(\phi)\,d\phi\right]^{2^k} d\eta$$

$F(x,y)$ is monotone decreasing in y, increasing in x. Hence, Eq. (31) is monotone increasing in α and has exactly one zero.

$$\alpha^* = \frac{1}{1 + k(\beta^2)^{-1}} \tag{32}$$

This value of α is valid for all ρ of interest and most code lengths of interest.

In each of the four cases, a weak-signal asymptotic solution produces the result

$$\alpha_{ws} \to 1 \tag{33}$$

(Note: For the tracking filter, the solution was obtained using Eq. (6), as the Gaussian approximation is invalid for weak signals.) The result, however, is of little interest, since no communication can take place in this region.

Figs. 5–7 give the word error probability which results in case BC when the power is divided optimally. P_e is plotted as a function of $(ST)/N_0$ with k = 2, 6, and 20,

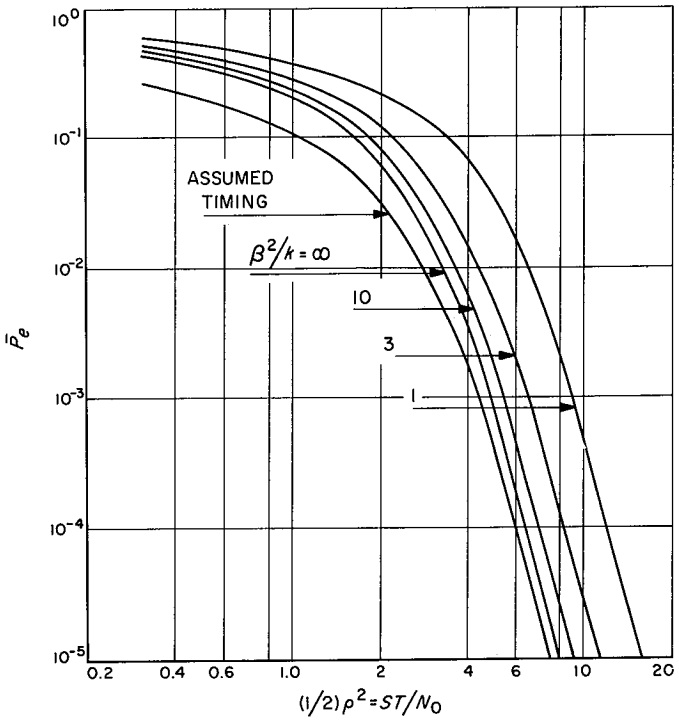


Fig. 5. Error probability of case BC using optimal power allocation, for $k = 2$

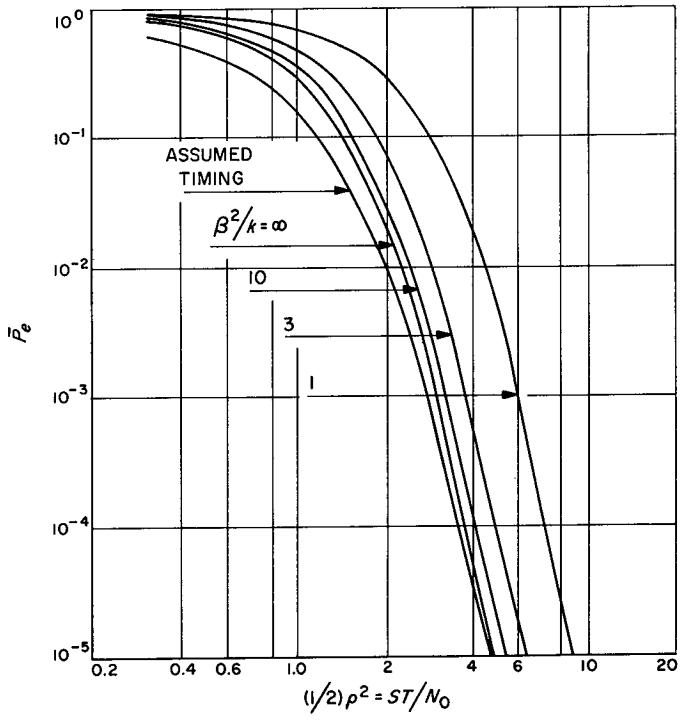


Fig. 6. Error probability of case BC using optimal power allocation, for $k = 6$

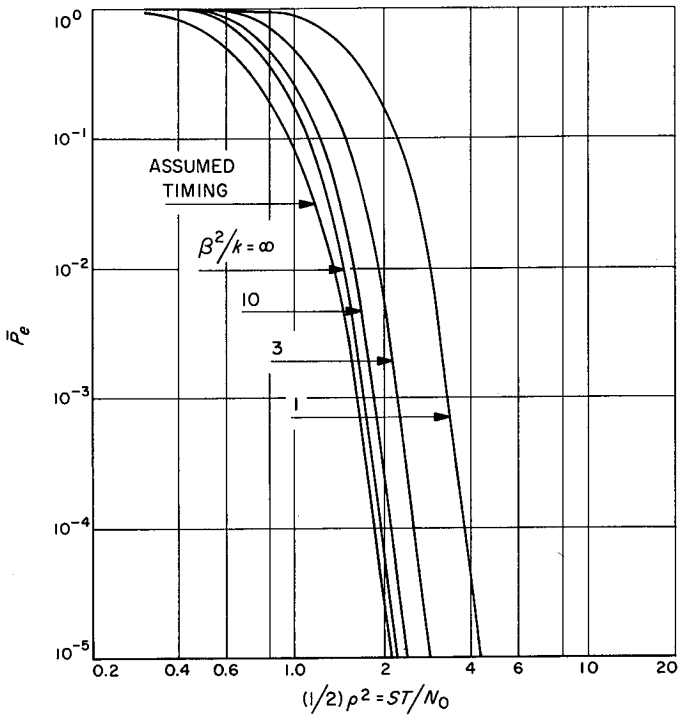


Fig. 7. Error probability of case BC using optimal power allocation, for $k = 20$

and $\beta^2/k = 1$, 3, 10, and ∞, (and for comparison for assumed perfect timing). The curves were graphically generated from Fig. 26 of Ref. (7).

3. A Case Against Coding

It is interesting at this point to examine the most powerful configuration, case BC, as the number of bits per code word becomes large. If α^* is substituted for α in Eq. (29), $\overline{P}_\varepsilon$ becomes

$$\overline{P}_\varepsilon = 1 - 2^{-k} - \int_{-\infty}^{\infty} N_{01}(\tau)\left[\int_{-\infty}^{\tau+\rho\left(\frac{2k\beta^2}{k+\beta^2}\right)^{1/2}} N_{01}(\theta)\,d\theta\right]^{2^{k-2}} d\tau$$

$$\times\left\{\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\left(\frac{2k\beta^2}{k+\beta^2}\right)^{1/2}} N_{01}(\xi)\,d\xi\right]^{2^{k-1}} d\eta - 2^{-k}\right\} \qquad (34)$$

For $k << \beta^2$, each of the product-integral terms of Eq. (34) behaves like Eq. (15) and $\overline{P}_\varepsilon$ decreases for increasing k. However, for large $k >> \beta^2$, the signal term in each of these integrals approaches the fixed value $\rho\beta(2)^{1/2}$, causing $\overline{P}_\varepsilon$ to approach 1 for very large k. Thus the infinitely long block code, which can (in theory) be used to communicate at rates approaching channel capacity

if no timing problem exists, cannot be so used in a time-varying channel, because too much information is needed to maintain adequate timing.

The situation may, in reality, be far worse than so far indicated, for β^2 has been treated as if it were independent of k. Such is probably *not* the case, for as k increases, the pulse duration T_P decreases. The time required for a given channel timing process to shift one pulse-duration decreases with decreasing T_P at a rate determinable from the spectrum of the timing process. Since β^2 is directly proportional to this one-pulse-shift time, β^2 is *at best* a nonincreasing function of k. It may be rapidly decreasing.

In one important situation, in fact, β^2 is proportional to $T_P{}^2$. If, for example, the communication channel's basic synchronization is by means of a phase-locked loop on the data subcarrier, the channel variations which remain are the result of skipping in the phase-locked loop. This skipping can be viewed as a Markov process, with a fixed step size $((2\pi)/\omega_{sc})$ and an average skipping rate. For a Markov process, the expected number of time-steps necessary to exceed n space-steps from a given starting point is approximately $(n + 1)^2$. (See Ref. (8) for the appropriate probability density.) Thus, the expected lifetime of channel stability at the word-sync level is proportional to the square of the number of subcarrier lock-steps per data pulse, and, hence, to the square of the data-pulse width. Therefore, $\beta_k{}^2 = \beta_1{}^2\,(k/(2^k-1))^2$, where $\beta_k{}^2$ refers to the value of β^2 existing with the coded system of kth order, and

$$P_\varepsilon(k) = 1 - 2^{-k} - \int_{-\infty}^{\infty} N_{01}(\tau)\left[\int_{-\infty}^{\tau+\rho\beta_1\left(\frac{2k^2}{k+(2^k-1)^2}\right)^{1/2}} N_{01}(\theta)\,d\theta\right]^{2^k-2} d\tau$$

$$\times\left\{\int_{-\infty}^{\infty} N_{01}(\eta)\left[\int_{-\infty}^{\eta+\rho\beta_1\left(\frac{2k^2}{k+(2^k-1)^2}\right)^{1/2}} N_{01}(\phi)\,d\phi\right]^{2^k-1} d\eta - 2^{-k}\right\} \tag{35}$$

which is monotone increasing in k. In this channel, all gain expected from coding appears to be destroyed by the increased difficulty of synchronization.

This result does not hold if the coded message is self-synchronizing as, e.g., the comma-free orthogonal block code. Nor does it hold intact if the β^2 is equipment limited instead of channel limited.

4. A Heuristic Comparison to Prefix Coding

Prefix coding is one of the alternatives to the use of a separate synchronizing channel. With it, certain pulses of each word are constrained into a fixed pattern, and the remaining pulses are constrained such that this pattern does not exist in them or in any possible word-overlap. Synchronization is obtained by identifying the location of this pattern. The transmitter's resources are allocated between synchronization and message in terms of number of symbols instead of in terms of power, as with the pilot-tone system.

It is known that for a W-symbol word to be prefix-synchronizable, at least $\log_2 W$ of the symbols must be used in the prefix (Ref. 9). If the symbols each carry equal energy, the prefix uses $1/W \log_2 W$ of the available signaling energy. Section 2, however, showed that the optimum fraction of signaling energy, which should be applied toward synchronization, is *at most* $\beta^{-2}(1+\beta^{-2}\log_2 W)^{-1}\log_2 W$. Since no synchronizable word can last longer than the channel is stably timed, $W \leq \beta^2$, and the optimal pilot tone uses less energy than does any prefix constraint.

Given the amount of energy in the prefix, the W measurements for the sync position will indicate the true sync with no smaller error probability than if these measurements were performed upon an optimal signal set—i.e. if the measurements were upon a pilot tone of the simplex form.

The prefix form is still further degraded from the optimal by direct use of symbols rather than energy for the synchronizing signal. Since the channel under discussion is binary, the information handling capability of the channel is not linear with signal energy, and for signal levels of interest, where the symbol error probability is low, removal of symbols from the message is far more costly than direct removal of the same amount of energy for use as a synchronizing pilot tone.

The conclusion reached is that the optimal pilot-tone synchronization out-performs any prefix-coded system which is synchronizable over the same channel.

5. Comparison to Comma-free Block Codes

The comma-free block code developed by Stiffler (Ref. 10) is another self-synchronizing one-channel system. These codes are produced by complementing or interchanging symbol-positions of all words of a standard orthogonal/biorthogonal code dictionary so that the

code's error correcting properties are unchanged but off-time correlations are minimized. This code-form supplies at least a part of the synchronization power which is supplied by the pilot-tone of the two-channel scheme. In almost all situations of interest, it supplies all that is necessary.

One may generalize from the results of Section 2 to say that the power is optimally divided when the probability of a word error given correct synchronization is equal to the probability of a synchronization error. Consequently, a one-channel system will be called "optimally self-synchronizable" if its synchronization error probability after reception of β^2/k words is at most equal to the synchronized word error probability. That this condition is satisfied for most situations of interest may be inferred from Table 1, extracted from Tables 6.2 and 6.3 of Ref. (10). The table is for orthogonal codes with word error probability $= 10^{-3}$. $(\beta^2/k)_{\min}$ is the minimum value of that parameter for which the synchronization error probability $\leq 10^{-3}$. $E\{(\beta^2/k)_{\min}\}$ is the expected value assuming all code vectors are equally likely, whereas $\max\{(\beta^2/k)_{\min}\}$ is an absolute upper-bound.

Table 1. Minimum values of (β^2/k) for which a comma-free code is optimally self-synchronizable

$W = 2^k$	$E\left\{\left(\frac{\beta^2}{k}\right)_{\min}\right\}$	$\max\left\{\left(\frac{\beta^2}{k}\right)_{\min}\right\}$
8	10	—
16	8	16
32	5	15
64	3	13
128	~1	4

The timing variations which exist at the word-synchronization level are almost always the low-frequency error, i.e. the "skipping", of phase-locked loop operating at the RF carrier or subcarrier level. The rate of this skipping is usually quite low (Ref. 6), and hence β^2/k is apt to be several orders of magnitude larger than the constraint values of Table 1. Knowing this, it is difficult to envision a design situation in which the synchronizability of the comma-free code would not be adequate.

6. Discussion

In most channels, the constraint upon both the comma-free code and the pilot-tone system is the number of words which the receiving equipment is able to use to determine the sync position. Since the comma-free code provides its own sync after receiving only a small number of words, while a pilot tone synchronizable in the same time would require a fairly large fraction of the available power for sync, the comma-free code would seem to be preferable in all cases.

However, the self-synchronization property of the comma-free block codes can only be utilized through sophisticated receiver processing. If transmitter power is cheap, and complex receiving equipment expensive, as in a ground-to-vehicle space telemetry application, a two-channel system is preferable; but if transmitter power is severely limited, and the sophisticated receiver processing no problem, as in a vehicle-to-ground telemetry application, the one-channel self-synchronizing system is far more preferable to one using pilot-tone synchronization.

The conclusion is that for a *Voyager*-class telemetry system, self-synchronizing codes provide the best synchronization method.

The analysis has assumed that bit timing is known. Imperfect bit timing causes an identical decrease in the effective signal strength at both the message detector and the maximum-likelihood word-timing detector. If "g" denotes the ratio of the effective signal power to the true signal power (given the degree of bit-timing uncertainty), a first-order correction for bit timing can be obtained by substituting "$g\alpha P$" for the message signal power and "$g(1-\alpha)P$" for the maximum-likelihood-detected pilot-signal power in the foregoing analyses. In most cases, this would require only slight modification of the results.

D. A Serial Orthogonal Decoder

R. R. Green

In this article a new approach to the decoding problem for certain block coded communication systems is presented. A simple and efficient decoder is presented.

Assume that a code word is selected from one of the 2^n code words in the dictionary H_n, where H_n is defined by

$$H_n = H_{n-1} \otimes H_1$$

and

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

($\otimes$ is the symbol for the Kroneck Product). This word is then transmitted over a channel which adds to it white Gaussian noise and is then available as a received signal $x(t)$. If we let τ be the time required to transmit each symbol of the code word, then the time required to transmit the whole word is $T = 2^n \tau$. It has been shown (Ref. 11) that to do optimal decoding we want to find k such that

$$1 \leq k \leq 2^n \text{ and } c_k = \max_{j=1, \cdots, 2^n} \{c_j\}$$

where

$$c_i = \int_0^T x(t) h_i(t) dt$$

and where $h_i(t)$ is one of the 2^n possible code words (or one of the 2^n rows of H_n).

Since $h_i(t) = \pm 1$ for all t we have

$$c_i = \sum_{j=1}^{2^n} x_j h_{ij}$$

where

$x_j = \int_{(j-1)\tau}^{j\tau} x(t) dt$ and h_{ij} is the jth bit of $h_i(t)$. So:

$$\begin{aligned} c_k &= \max_{j=1, \cdots, 2^n} \{c_j\} \\ &= \max_{j=1, \cdots, 2^n} \{(H_n x)_j\} \\ &= \max_{j=1, \cdots, 2^n} \{y_j\} \end{aligned}$$

where

$$y = H_n x.$$

If we assume that the components of the vector x are available sequentially as 2^n q bit serial binary words, we would like to find a machine which would perform the operation $H_n x$. However, as this operation requires 2^{n+1} additions or subtractions, it is rather inefficient and difficult to mechanize.

Instead, a more efficient and more easily mechanized procedure is as follows.

Define:

1. $P_1 = I_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$$P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and inductively;}$$

$$P_{n+1} = (I_1 \otimes P_n)(P_2 \otimes I_{n-1}) \qquad n \geq 1$$

2. $R_1 = H_1$ and inductively;

$$R_{n+1} = (P_2 \otimes I_{n-1})(I_1 \otimes R_n) \qquad n \geq 1$$

Note that P_n is a $2^n \times 2^n$ permutation matrix; therefore, $P_n^{-1} = P_n^T$.

Similarly P_n^k for any $k \geq 1$ must be a $2^n \times 2^n$ permutation matrix, therefore

$$\begin{aligned} (P_n^k)^{-1} &= (P_n^k)^T \\ &= (P_n^T)^k \\ &= (P_n^{-1})^k \end{aligned}$$

Also, the matrix R_n has been described by Koerner in *SPS 37-17*, Vol. IV, page 72.

Lemma 1:

$$P_n = (I_k \otimes P_{n-k})(P_{k+1} \otimes I_{n-k-1}) \qquad \text{for } n - 1 \geq k \geq 0$$

Proof by induction: Trivial for n arbitrary $k = 0$. True by definition for $k = 1$.

Assume true for all $n \geq k' + 1$ where $k \geq k' \geq 0$, prove for $k + 1$ for all $n \geq (k + 1) + 1 = k + 2$

$$\begin{aligned} P_n &= (I_k \otimes P_{n-k})(P_{k+1} \otimes I_{n-k-1}) \\ &= [I_k \otimes (I_1 \otimes P_{n-k-1})(P_2 \otimes I_{n-k-2})](P_{k+1} \otimes I_{n-k-1}) \\ &= (I_{k+1} \otimes P_{n-k-1})(I_k \otimes P_2 \otimes I_{n-k-2})(P_{k+1} \otimes I_{n-k-1}) \\ &= (I_{k+1} \otimes P_{n-k-1})[(I_k \otimes P_2)(P_{k+1} \otimes I_1) \otimes I_{n-k-2}] \\ &= (I_{k+1} \otimes P_{n-k-1})(P_{k+2} \otimes I_{n-k+2}) \end{aligned}$$

Lemma 2:

$$P_{n+1}(I_1 \otimes P_n^T) = (P_n^T \otimes I_1)P_{n+1} \qquad n \geq 1$$

Proof by induction: Trivial for $n = 1$. For $n = 2$ see Fig. 8. Assume true for $n = k$, prove for $n = k + 1$

$$P_3(I_1 \otimes P_2) = \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix}$$

$$= \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix}$$

$$(P_2 \otimes I_1)P_3 = \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix}$$

$$= \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix} = P_3(I_1 \otimes P_2)$$

Fig. 8. Lemma 2 — $n = 1$

$$\begin{aligned} P_{k+2}(I_1 \otimes P_{k+1}^T) &= (I_k \otimes P_2)(P_{k+1} \otimes I_1)(I_1 \otimes P_k^T \otimes I_1)(I_k \otimes P_2) \\ &= (I_k \otimes P_2)[P_{k+1}(I_1 \otimes P_k^T) \otimes I_1](I_k \otimes P_2) \\ &= (I_k \otimes P_2)[(P_k^T \otimes I_1)P_{k+1} \otimes I_1](I_k \otimes P_2) \\ &= (I_k \otimes P_2)(P_k^T \otimes I_2)(P_{k+1} \otimes I_1)(I_k \otimes P_2) \\ &= (P_k^T \otimes I_2)(I_k \otimes P_2)(P_{k+1} \otimes I_1)(I_k \otimes P_2) \\ &= (P_k^T \otimes I_2)(I_{k-1} \otimes P_3)(P_k \otimes I_2)(I_k \otimes P_2) \\ &= (P_k^T \otimes I_2)(I_{k-1} \otimes P_3)(I_k \otimes P_2)(P_k \otimes I_2) \\ &= (P_k^T \otimes I_2)[I_{k-1} \otimes P_3(I_1 \otimes P_2)](P_k \otimes I_2) \\ &= (P_k^T \otimes I_2)(I_{k-1} \otimes P_2 \otimes I_1)(I_{k-1} \otimes P_3)(P_k \otimes I_2) \\ &= (P_{k+1}^T \otimes I_1)P_{k+2} \end{aligned}$$

Theorem 1: $P_n^k = (I_1 \otimes P_{n-1}^k)(P_{k+1}^T \otimes I_{n-k-1})$ for $n - 1 \geq k \geq 1$

Proof by induction: True by definition for $k = 1$. Assume true for k prove for $k + 1$ for $n \geq k + 2$:

$$\begin{aligned}
P_n^{k+1} &= P_n^k P_n = (I_1 \otimes P_{n-1}^k)(P_{k+1}^T \otimes I_{n-k-1})(I_{k+1} \otimes P_{n-k-1})(P_{k+2} \otimes I_{n-k-2}) \\
&= (I_1 \otimes P_{n-1}^k)(I_{k+1} \otimes P_{n-k-1})(P_{k+1}^T \otimes I_{n-k-1})(P_{k+2} \otimes I_{n-k-2}) \\
&= (I_1 \otimes P_{n-1}^k)(I_{k+1} \otimes P_{n-k-1})[(P_{k+1}^T \otimes I_1) P_{k+2} \otimes I_{n-k-2}] \\
&= (I_1 \otimes P_{n-1}^k)(I_{k+1} \otimes P_{n-k-1})(P_{k+2} \otimes I_{n-k-2})(I_1 \otimes P_{k+1}^T \otimes I_{n-k-2}) \\
&= (I_1 \otimes P_{n-1}^k)(I_1 \otimes P_{n-1})(P_2 \otimes I_{n-2})(I_1 \otimes P_{k+1}^T \otimes I_{n-k-2}) \\
&= (I_1 \otimes P_{n-1}^{k+1})(P_{k+2}^T \otimes I_{n-k-2})
\end{aligned}$$

Corollary: $P_n^n = I_n$

Proof by induction: $P_1^1 = I_1$. Assume true for n prove for $n + 1$:

$$\begin{aligned}
P_{n+1}^{n+1} = P_{n+1}^n P_{n+1} = (I_1 \otimes P_n^n)(P_{n+1}^T \otimes I_0) P_{n+1} &= I_{n+1} P_{n+1}^T P_{n+1} \\
&= I_{n+1}
\end{aligned}$$

Note that $P_u^n = I_n$ implies $P_n^{n-1} = P_n^{-1} = P_n^T$.

Lemma 3: $R_n = (P_{k+1}^T \otimes I_{n-k-1})(I_k \otimes R_{n-k})$ for $n - 1 \geq k \geq 0$

Proof by induction: Trivial for $k = 0$, true by definition for $k = 1$. Assume true for all $n \geq k' + 1$ where $k \geq k' \geq 0$, prove for $k + 1$ for all $n \geq k + 2$:

$$\begin{aligned}
R_n &= (P_{k+1}^T \otimes I_{n-k-1})(I_k \otimes R_{n-k}) \\
&= (P_{k+1}^T \otimes I_{n-k-1})(I_k \otimes P_2 \otimes I_{n-k-2})(I_{k+1} \otimes R_{n-k-1}) \\
&= (P_{k+2}^T \otimes I_{n-k-2})(I_{k+1} \otimes R_{n-k-1})
\end{aligned}$$

Lemma 4: $P_{n+1}(R_1 \otimes I_n) = (I_n \otimes R_1) P_{n+1}$ for $n \geq 1$

Proof by induction:

$$\text{for } n = 1,\ P_2 (R_1 \otimes I_1) = \begin{bmatrix} 1&0&0&0 \\ 0&0&1&0 \\ 0&1&0&0 \\ 0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&1&0 \\ 0&1&0&1 \\ 1&0&-1&0 \\ 0&1&0&-1 \end{bmatrix} = \begin{bmatrix} 1&0&1&0 \\ 1&0&-1&0 \\ 0&1&0&1 \\ 0&1&0&-1 \end{bmatrix} = \begin{bmatrix} 1&1&0&0 \\ 1&-1&0&0 \\ 0&0&1&1 \\ 0&0&1&-1 \end{bmatrix} \begin{bmatrix} 1&0&0&0 \\ 0&0&1&0 \\ 0&1&0&0 \\ 0&0&0&1 \end{bmatrix} = (I_1 \otimes R_1) P_2$$

Assume true for n, prove for $n + 1$:

$$\begin{aligned}
P_{n+2}(R_1 \otimes I_{n+1}) &= (I_n \otimes P_2)(P_{n+1} \otimes I_1)(R_1 \otimes I_{n+1}) \\
&= (I_n \otimes P_2)(I_n \otimes R_1 \otimes I_1)(P_{n+1} \otimes I_1) \\
&= (I_{n+1} \otimes R_1)(I_n \otimes P_2)(P_{n+1} \otimes I_1) \\
&= (I_{n+1} \otimes R_1) P_{n+2}
\end{aligned}$$

Theorem 2: $R_n^k = (P_{k+1} \otimes I_{n-k-1})(I_1 \otimes R_{n-1}^k)$ for $n-1 \geq k \geq 1$

Proof by induction: True by definition for $k = 1$. Assume true for k, prove for $k+1$ for $n \geq k+2$:

$$\begin{aligned}
R_n^{k+1} &= R_n R_n^k = (P_{k+2}^T \otimes I_{n-k-2})(I_{k+1} \otimes R_{n-k-1})(P_{k+1} \otimes I_{n-k-1})(I_1 \otimes R_{n-1}^k) \\
&= (P_{k+2} \otimes I_{n-k-2})(P_{k+2}^k \otimes I_{n-k-2})(P_{k+1} \otimes I_{n-k-1})(I_{k+1} \otimes R_{n-k-1})(I_1 \otimes R_{n-1}^k) \\
&= (P_{k+2} \otimes I_{n-k-2})(I_1 \otimes P_{k+1}^T \otimes I_{n-k-2})(P_{k+1}^T \otimes I_{n-k-1})(P_{k+1} \otimes I_{n-k-1})(I_{k+1} \otimes R_{n-k-1})(I_1 \otimes R_{n-1}^k) \\
&= (P_{k+2} \otimes I_{n-k-2})[I_1 \otimes (P_{k+1}^T \otimes I_{n-k-2})(I_k \otimes R_{n-k-1})](I_1 \otimes R_{n-1}^k) \\
&= (P_{k+2} \otimes I_{n-k-2})(I_1 \otimes R_{n-1})(I_1 \otimes R_{n-1}^k) \\
&= (P_{k+2} \otimes I_{n-k-2})(I_1 \otimes R_{n-1}^{k+1})
\end{aligned}$$

Corollary: $R_n^n = H_n$

Proof by induction: $R_1^1 = H_1$. Assume true for n, prove for $n+1$:

$$\begin{aligned}
R_{n+1}^{n+1} &= R_{n+1} R_{n+1}^n = R_{n+1} P_{n+1}(I_1 \otimes R_n^n) \\
&= P_{n+1}^T (I_n \otimes R_1) P_{n+1}(I_1 \otimes H_n) \\
&= P_{n+1}^T P_{n+1}(H_1 \otimes I_n)(I_1 \otimes H_n) \\
&= H_{n+1}
\end{aligned}$$

Theorem 3: $P_n^k R_n^k = I_{n-k} \otimes H_k$ for $n \geq k \geq 1$

Proof by induction: $P_1^1 R_1^1 = I_1 H_1 = H_1$. Assume true for $P_n^k R_n^k$, prove for $P_{n+1}^k R_{n+1}^k$:

$$\begin{aligned}
P_{n+1}^k R_{n+1}^k &= (I_1 \otimes P_n^k)(P_{k+1}^T \otimes I_{n-k})(P_{k+1} \otimes I_{n-k})(I_1 \otimes R_n^k) \\
&= I_1 \otimes P_n^k R_n^k = I_1 \otimes I_{n-k} \otimes H_k = I_{n+1-k} \otimes H_k
\end{aligned}$$

and prove for $P_{n+1}^{k+1} R_{n+1}^{k+1}$:

$$\begin{aligned}
P_{n+1}^{k+1} R_{n+1}^{k+1} &= P_{n+1}(I_{n+1-k} \otimes H_k) R_{n+1} = P_{n+1}(I_{n+1-k} \otimes H_k) P_{n+1}^T (I_n \otimes R_1) \\
&= (I_{n-k} \otimes P_{k+1})(P_{n-k+1} \otimes I_k)(I_{n+1-k} \otimes H_k)(P_{n+1-k}^T \otimes I_k)(I_{n-k} \otimes P_{k+1}^T)(I_n \otimes R_1) \\
&= [I_{n-k} \otimes P_{k+1}(I_1 \otimes H_k) P_{k+1}^T](I_n \otimes R_1) \\
&= I_{n-k} \otimes P_{k+1}(I_1 \otimes H_k) P_{k+1}^T (I_k \otimes R_1) \\
&= I_{n-k} \otimes P_{k+1}(P_{k+1}^k R_{k+1}^k) R_{k+1} \\
&= I_{n-k} \otimes P_{k+1}^{k+1} R_{k+1}^{k+1} = I_{n-k} \otimes H_{k+1}
\end{aligned}$$

Thus we see that if we can build a machine which consists of n stages, the ith one of which, $1 \leq i \leq n$, performs the operation $P_n^i R_n (P_n^{i-1})^T x$, cascading these n stages would give the desired operation of

$$\begin{aligned}
y &= P_n^n R_n (P_n^{n-1})^T P_n^{n-1} R_n (P^{n-2})^T \cdots P_n^2 R_n P_n P_n^T R_n P_n^0 x \\
&= P_n^n R_n^n x = R_n^n x = H_n x
\end{aligned}$$

For an example of the second stage operation for a length 8 code, see Fig. 9. One possible mechanization of a machine to accomplish the job of the ith stage is shown in Fig. 10, where w_{i-1} is the output of the ith stage of a binary counter which is pulsed every word time (τ). Thus w_{i-1} changes states every 2^{i-1} word times, and w_{i-1} is time to go true as the first component of $P_n^{i-1}\ R_n^{i-1}\ x$ appears at the ith stage.

Since we add 2^n binary numbers of q bits each, the digital word length must be $m \geq q + n$. Since symbols are received at the rate of $1/\tau$ per second, the decoder must operate at $s = m/\tau$ bits per second. If we take both m and s to be fixed, the data rate which the decoder can handle for an orthogonal code of length 2^n is $r = (ns)/(m\,2^n)$. For example, letting $n = q = 7$, $s = 10^7$ we have $r > 35{,}000$ data bits per second.

This decoder has several advantages. Since every component of y involves all of the 2^n components of x, any decoder must have at least $2^n - 1$ words of memory. This decoder involves

$$\sum_{i=1}^{n} 2^{i-1} = \sum_{i=0}^{n-1} 2^i = 2^n - 1$$

words of memory. As is shown in Theorem 3, a decoder of N stages will decode any code for which $N \geq n \geq 1$, and, further, if it is desired to expand the decoder to the case of $N + 1$, no redesign is needed. To accomplish the expansion, simply add one more decoding stage and one more flip-flop to the w counter. The final advantage is the previously mentioned one of being able to accommodate quite high data rates.

$$P_3^2 R_3 P_3^T = \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&1&0&0&0&0&0&0\\ 0&0&1&1&0&0&0&0\\ 0&0&0&0&1&1&0&0\\ 0&0&0&0&0&0&1&1\\ 1&-1&0&0&0&0&0&0\\ 0&0&1&-1&0&0&0&0\\ 0&0&0&0&1&-1&0&0\\ 0&0&0&0&0&0&1&-1 \end{bmatrix} \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix}$$

$$= \begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0\\ 0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 1&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&1&0\\ 0&1&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&1\\ 1&0&-1&0&0&0&0&0\\ 0&0&0&0&1&0&-1&0\\ 0&1&0&-1&0&0&0&0\\ 0&0&0&0&0&1&0&-1 \end{bmatrix} = \begin{bmatrix} 1&0&1&0&0&0&0&0\\ 0&1&0&1&0&0&0&0\\ 1&0&-1&0&0&0&0&0\\ 0&1&0&-1&0&0&0&0\\ 0&0&0&0&1&0&1&0\\ 0&0&0&0&0&1&0&1\\ 0&0&0&0&1&0&-1&0\\ 0&0&0&0&0&1&0&-1 \end{bmatrix}$$

Fig. 9. Second stage for 8/3 Code

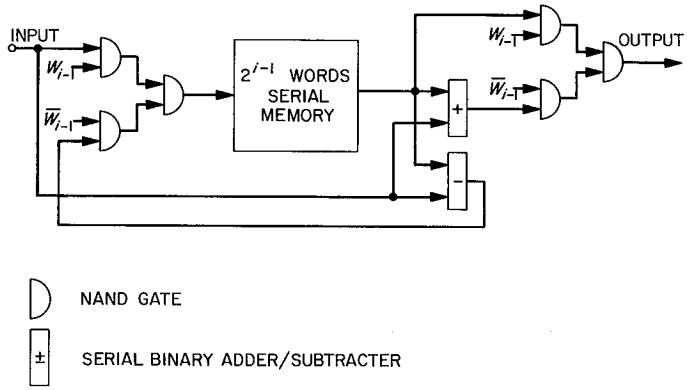


Fig. 10. ith stage of decoder

References

1. Tausworthe, R. C., *Theory and Practical Design of Phase-Locked Receivers,* Technical Report No. 32-819, Jet Propulsion Laboratory, Pasadena, California, February 1966.
2. Gilbert, E. N., "Synchronization of Binary Messages," *IRE Transactions on Information Theory,* Vol. IT-6, No. 4, September 1960, pp. 470–477.
3. Van Trees, H. L., "Optimum Power Division in Coherent Communications Systems," *IEEE Trans., SET-10,* No. 1, March, 1964.
4. Stiffler, J. J., "On the Allocation of Power in a Synchronous Binary PSK Communcations System," *National Telemetry Conference* Paper No. 5–1, June, 1964.

References (Cont'd)

5. Kaneko, H., "A Statistical Analysis of the Synchronization of a Binary Receiver," *IEEE Trans.*, CS-11, No. 4, December 1963.

6. Viterbi, A. J., "Phase-locked Loop Dynamics in the Presence of Noise by Fokker-Planck Techniques," *Proc. IEEE*, Vol. 51, December 1963.

7. Viterbi, A. J., et. al., *Coding Theory and its Application to Communications Systems*, Technical Report No. 32-67, Jet Propulsion Laboratory, Pasadena, California, March 1961.

8. Kac, M., "Brownian Motion—First Passage Time Problem," *Ann. Math. Statist.*, Vol. 16, p. 62, 1945.

9. Gilbert, E. N., "Synchronization of Binary Messages," *I.R.E. Trans. IT-6*, pp. 470–477, September 1960.

10. Stiffler, J. J., *Self-Synchronizing Binary Telemetry Codes*, PhD Thesis, California Institute of Technology, Pasadena, California, 1962.

11. Viterbi, A. J., *On Phase Coherent Communications*, Technical Report No. 32-25, Jet Propulsion Laboratory, Pasadena, August 15, 1960.

XXI. Communications Systems Research: Astrometrics

A. Optimal Combination of Estimates

P. Reichley

1. Summary

A problem arising in radar astronomy as well as in other fields of the physical sciences is the combination of estimates of a set of parameters in such a way that an optimal estimate is obtained. If we have two independent estimate vectors x and y of a set of parameters and their associated covariance matrices $\mathbf{V}_x$ and $\mathbf{V}_y$, then the usual optimal estimate of the set of parameters used by most authors is $(\mathbf{V}_x^{-1} + \mathbf{V}_y^{-1})^{-1}(\mathbf{V}_x^{-1}\mathbf{x} + \mathbf{V}_y^{-1}\mathbf{y})$. We have yet to see a satisfactory explanation why this particular estimate is used and in what sense it is an optimal estimate. We shall examine this same problem for the case of m independent estimate vectors and show in what sense the estimate obtained is optimal.

2. Introduction

When we speak of a positive definite Hermitian matrix **A** being minimal with respect to a set of restrictions, we mean that for any other positive definite Hermitian matrix **B** found subject to the same restrictions, the difference $\mathbf{B} - \mathbf{A}$ is positive definite.

We must also introduce the concept of a stationary point of a matrix. Suppose we have a matrix **A** each of whose elements a_{ij} is a function of a set of variables t_i, $i = 1, \cdots, m$. We define the variation of **A** as

$$\delta\mathbf{A} = \{\delta a_{ij}\} = \left\{\sum_{k=1}^{m} \frac{\partial a_{ij}}{\partial t_k}\,\delta t_k\right\}$$

and the stationary point of **A** as the set of variables t_i, $i = 1, \cdots, m$, which satisfy $\delta\mathbf{A} = \mathbf{0}$ for arbitrary variation.

The covariance matrix of an estimate vector **x** is defined by $E[(\mathbf{x} - \mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t)^{\mathrm{T}}]$ (E stands for expected value and $\mathbf{A}^{\mathrm{T}}$ for **A** transpose), where $\mathbf{x}_t$ is the true value of **x**. When we say two estimate vectors **x** and **y** are independent, we mean that

$$E[(\mathbf{x} - \mathbf{x}_t)(\mathbf{y} - \mathbf{y}_t)^{\mathrm{T}}] = E[(\mathbf{y} - \mathbf{y}_t)(\mathbf{x} - \mathbf{x}_t)^{\mathrm{T}}] = \mathbf{0}.$$

When we say an estimate vector is unbiased, we mean that $E(\mathbf{x} - \mathbf{x}_t) = \mathbf{0}$.

3. Statement of the Problem

Suppose we have m independent unbiased estimate vectors $\mathbf{x}_i$, $i = 1, \cdots, m$, of n parameters. We shall assume that we know the covariance matrix $\mathbf{V}_i$ corresponding to each estimate vector $\mathbf{x}_i$. We would like to find a set of $n \times n$ matrices $\mathbf{W}_i$, $i = 1, \cdots, m$, such that

$$\mathbf{x} = \sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i \tag{1}$$

is an unbiased minimum variance estimate of $\mathbf{x}_t$, the true value vector of the n parameters. By minimum variance we mean that the covariance matrix of $\mathbf{x}$ is minimum.

Since we wish $\mathbf{x}$ unbiased, we have

$$\mathbf{0} = E(\mathbf{x} - \mathbf{x}_t) = E\left(\sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i - \mathbf{x}_t\right)$$
$$= E\left[\sum_{i=1}^{m} \mathbf{W}_i (\mathbf{x}_i - \mathbf{x}_t) + \left(\sum_{i=1}^{m} \mathbf{W}_i - \mathbf{I}\right)\mathbf{x}_t\right].$$

But, since the $\mathbf{x}_i$ are unbiased,

$$E(\mathbf{x}_i - \mathbf{x}_t) = \mathbf{0}, \qquad i = 1, \cdots, m,$$

and the condition on $\mathbf{W}_i$, $i = 1, \cdots, m$, for $\mathbf{x}$ to be an unbiased estimate of $\mathbf{x}_t$, is

$$\sum_{i=1}^{m} \mathbf{W}_i = \mathbf{I}. \tag{2}$$

Let us now consider the covariance matrix of $\mathbf{x}$:

$$\mathbf{V} = E\left[\left(\sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i - \mathbf{x}_t\right)\left(\sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i - \mathbf{x}_t\right)^{\mathrm{T}}\right].$$

Since

$$\sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i - \mathbf{x}_t = \sum_{i=1}^{m} \mathbf{W}_i (\mathbf{x}_i - \mathbf{x}_t),$$

then, from Eq. (2), we have

$$\mathbf{V} = E\left\{\left[\sum_{i=1}^{m} \mathbf{W}_i (\mathbf{x}_i - \mathbf{x}_t)\right]\left[\sum_{i=1}^{m} (\mathbf{x}_i - \mathbf{x}_t)^{\mathrm{T}} \mathbf{W}_i^{\mathrm{T}}\right]\right\}.$$

Since the $\mathbf{x}_i$, $i = 1, \cdots, m$ are independent, then $E[(\mathbf{x}_i - \mathbf{x}_t)(\mathbf{x}_j - \mathbf{x}_t)^{\mathrm{T}}] = \mathbf{0}$, $i \neq j$, $i, j = 1, \cdots, m$. Hence,

$$\mathbf{V} = \sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} \tag{3}$$

since $\mathbf{V}_i = \mathrm{E}[(\mathbf{x}_i - \mathbf{x}_t)(\mathbf{x}_i - \mathbf{x}_t)^{\mathrm{T}}]$.

The problem we therefore wish to consider is the minimization of $\mathbf{V}$ (as given by Eq. 3), subject to the restriction on $\mathbf{W}_i$, $i = 1, \cdots, m$ (as given by Eq. 2), where $\mathbf{V}_i$, $i = 1, \cdots, m$ is a given set of covariance matrices; i.e., positive definite Hermitian.

4. Determination of the Stationary Point of V

Since $\mathbf{W}_i$, $i = 1, \cdots, m$ must satisfy Eq. (2), we introduce an $n \times n$ matrix $\mathbf{\Lambda}$ of Lagrange multipliers and consider the variation of the following function subject to Eq. (2):

$$\mathbf{F} = \sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} + \left(\sum_{i=1}^{m} \mathbf{W}_i - \mathbf{I}\right)\mathbf{\Lambda}.$$

The variation of $\mathbf{F}$ with respect to $\mathbf{W}_i$ is given by

$$\delta\mathbf{F}_i = \delta\mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} + \mathbf{W}_i \mathbf{V}_i \delta\mathbf{W}_i^{\mathrm{T}} + \delta\mathbf{W}_i \mathbf{\Lambda} = \mathbf{0}.$$

Since $\delta\mathbf{W}_i^{\mathrm{T}} = (\delta\mathbf{W}_i)^{\mathrm{T}}$, it follows that $\delta\mathbf{W}_i\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{T}}\delta\mathbf{W}_i^{\mathrm{T}}$, and we may rewrite the above equation as

$$\delta\mathbf{W}_i (2\mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} + \mathbf{\Lambda}) + (2\mathbf{W}_i \mathbf{V}_i + \mathbf{\Lambda}^{\mathrm{T}})\, \delta\mathbf{W}_i^{\mathrm{T}} = \mathbf{0}.$$

Since $\delta\mathbf{W}_i$ and $\delta\mathbf{W}_i^{\mathrm{T}}$ are arbitrary, we have

$$2\mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} + \mathbf{\Lambda} = \mathbf{0}.$$

Premultiplying the above equation by $\mathbf{W}_i$ and summing, it follows that

$$\mathbf{\Lambda} = -2 \sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}}.$$

Hence,

$$\mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} = \sum_{j=1}^{m} \mathbf{W}_j \mathbf{V}_j \mathbf{W}_j^{\mathrm{T}}, \qquad i = 1, \cdots, m.$$

Premultiplying the above equation by $\mathbf{V}_i^{-1}$ and summing, it follows that

$$\sum_{j=1}^{m} \mathbf{W}_j \mathbf{V}_j \mathbf{W}_j^{\mathrm{T}} = \left(\sum_{i=1}^{m} \mathbf{V}_i^{-1}\right)^{-1}.$$

Thus,

$$\mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} = \left(\sum_{j=1}^{m} \mathbf{V}_j^{-1}\right)^{-1}, \qquad i = 1, \cdots, m.$$

Hence,

$$\mathbf{W}_i = \left(\sum_{j=1}^{m} \mathbf{V}_j^{-1}\right)^{-1} \mathbf{V}_i^{-1}, \qquad i = 1, \cdots, m, \tag{4}$$

is the solution of the variation of $\mathbf{F}$ subject to Eq. (2).

We may summarize the above result in the following theorem:

Theorem 1. Let $\mathbf{V}_i$, $i = 1, \cdots, m$ be a given set of positive definite Hermetian $n \times n$ matrices. If $\mathbf{W}_i$, $i = 1, \cdots, m$ is a set of $n \times n$ matrices which are allowed to vary, then the stationary point of

$$\mathbf{V} = \sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} \tag{3}$$

subject to

$$\sum_{i=1}^{m} \mathbf{W}_i = \mathbf{I} \tag{2}$$

is given by

$$\mathbf{W}_i = \left(\sum_{j=1}^{m} \mathbf{V}_j^{-1} \right)^{-1} \mathbf{V}_i^{-1}, \qquad i = 1, \cdots, m. \tag{4}$$

5. An Optimal Estimate for x

We wish to find a set of $n \times n$ matrices $\mathbf{W}_i, i = 1, \cdots, m$, such that the covariance matrix (Eq. 3) of x as given by Eq. (1), subject to Eq. (2), is minimal. Again we emphasize that the word *minimal* is used in the sense that, given any other matrix $\mathbf{V}^*$ found subject to the above restrictions, $\mathbf{V}^* - \mathbf{V}$ is positive definite. The set $\mathbf{W}_i, i = 1, \cdots, m$ thus found will then constitute what we call an optimal solution. To this end, let us show that the stationary point of $\mathbf{V}$ given by Eq. (4), subject to Eq. (2), yields an optimal solution; i.e., an unbiased minimum variance solution.

Let $\mathbf{U}_i$, $i = 1, \cdots, m$ be any other set of $n \times n$ matrices satisfying Eq. (2). We may assume without loss of generality that $\mathbf{U}_i = \boldsymbol{W}_i + \boldsymbol{D}_i$, $i = 1, \cdots, m$, where $\mathbf{D}_i$, $i = 1, \cdots, m$ is a set of $n \times n$ matrices. Let us consider the matrix

$$\mathbf{V}^* = \sum_{i=1}^{m} \mathbf{U}_i \mathbf{V}_i \mathbf{U}_i^{\mathrm{T}} = \mathbf{V} + \sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}} + \sum_{i=1}^{m} \mathbf{D}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} + \sum_{i=1}^{m} \mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}.$$

From Eq. (4), we have

$$\sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}} = \left(\sum_{j=1}^{m} \mathbf{V}_j^{-1} \right)^{-1} \sum_{i=1}^{m} \mathbf{D}_i^{\mathrm{T}}.$$

But, since $\mathbf{U}_i$, $i = 1, \cdots, m$ satisfies Eq. (2),

$$\mathbf{I} = \sum_{i=1}^{m} \mathbf{U}_i = \sum_{i=1}^{m} \mathbf{W}_i + \sum_{i=1}^{m} \mathbf{D}_i = \mathbf{I} + \sum_{i=1}^{m} \mathbf{D}_i$$

and

$$\sum_{i=1}^{m} \mathbf{D}_i = \mathbf{0}.$$

Hence,

$$\sum_{i=1}^{m} \mathbf{D}_i^{\mathrm{T}} = \mathbf{0}$$

and

$$\sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}} = \sum_{i=1}^{m} \mathbf{D}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} = \mathbf{0}$$

since

$$\sum_{i=1}^{m} \mathbf{D}_i \mathbf{V}_i \mathbf{W}_i^{\mathrm{T}} = \left(\sum_{i=1}^{m} \mathbf{W}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}} \right)^{\mathrm{T}}.$$

Therefore,

$$\mathbf{V}^* = \mathbf{V} + \sum_{i=1}^{m} \mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}.$$

Since $(\mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}})^{\mathrm{T}} = \mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}$, then $\mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}$ is symmetric. Consider any eigenvalue λ of $\mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}$ and its corresponding eigenvector e. We have $\mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}} \mathbf{e} = \lambda \mathbf{e}$. Premultiplying by $\mathbf{e}^{\mathrm{T}}$, we have $\mathbf{e}^{\mathrm{T}} \mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}} \mathbf{e} = \lambda \mathbf{e}^{\mathrm{T}} \mathbf{e}$ and

$$\lambda = \frac{(\mathbf{D}_i^{\mathrm{T}} \mathbf{e})^{\mathrm{T}} \mathbf{V}_i (\mathbf{D}_i^{\mathrm{T}} \mathbf{e})}{\mathbf{e}^{\mathrm{T}} \mathbf{e}} > 0,$$

which follows from $\mathbf{V}_i$ being positive definite. It follows that $\mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}$ is positive definite, and hence

$$\sum_{i=1}^{m} \mathbf{D}_i \mathbf{V}_i \mathbf{D}_i^{\mathrm{T}}$$

is positive definite. Hence, $\mathbf{V}^* - \mathbf{V}$ is positive definite, and $\mathbf{W}_i$, $i = 1, \cdots, m$ is the set which minimizes $\mathbf{V}$. We summarize these results by the following theorem:

Theorem 2. Suppose we have m independent unbiased estimate vectors $\mathbf{x}_i$, $i = 1, \cdots, m$, of the same n parameters, and their corresponding covariance matrices $\mathbf{V}_i$, $i = 1, \cdots, m$. Then an optimal estimate (unbiased minimum variance) of the n parameters is given by

$$\mathbf{x} = \sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i \tag{1}$$

where the $\mathbf{W}_i$, $i = 1, \cdots, m$ are $n \times n$ matrices given by

$$\mathbf{W}_i = \left(\sum_{j=1}^{m} \mathbf{V}_j^{-1} \right)^{-1} \mathbf{V}_i^{-1} . \tag{4}$$

Moreover, the covariance matrix of this estimate is

$$\mathbf{V} = \left(\sum_{j=1}^{m} \mathbf{V}_j^{-1} \right)^{-1} . \tag{5}$$

It follows easily from Eqs. (3) and (4) that the covariance matrix of $\mathbf{x}$ is given by Eq. (5).

6. Geometrical Significance

Let $\mathbf{D}$ be a positive definite Hermetian $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$. Then the k^{th} elementary symmetric function of the eigenvalues is defined as

$$f_k(\mathbf{D}) = \sum \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_k}$$

where the summation is taken over all permutations of $i_1, i_2, \cdots, i_k$ over $1, 2, \cdots, n$.

Theorem 3. If $\mathbf{A}$ and $\mathbf{B}$ are positive definite Hermitian $n \times n$ matrices and $\mathbf{A} - \mathbf{B}$ is positive definite, then $f_k(\mathbf{B}) < f_k(\mathbf{A})$, $k = 1, \cdots, n$.

Proof. Let $\mathbf{A} - \mathbf{B} = \mathbf{C}$. Then $\mathbf{A} = \mathbf{B} + \mathbf{C}$. But a result of Marcus and Lopes (Ref. 1) states that $f_k(\mathbf{B} + \mathbf{C}) \geqq f_k(\mathbf{B}) + f_k(\mathbf{C})$ if $\mathbf{B}$ and $\mathbf{C}$ are positive definite Hermitian matrices. Since $f_k(\mathbf{C}) > 0$, the theorem follows.

From the above theorem, we see that the $\mathbf{W}_i$, $i = 1, \cdots, m$ obtained in Theorem 2 minimize all the elementary symmetric functions of the covariance matrix $\mathbf{V}$ (Eq. 5) of our optimal estimate.

Let us consider the ellipsoid of concentration corresponding to $\mathbf{V}$ and some given distribution. The semi-axes of the ellipsoid are given by the eigenvalues of $\mathbf{V}$. Hence, we have minimized all the elementary symmetric functions of the semi-axis of the ellipsoid. In particular, since $f_n(\mathbf{V})$ is simply the determinant of $\mathbf{V}$, we have minimized the volume of the ellipsoid, since the volume is a direct function of the determinant of $\mathbf{V}$ (Ref. 2). We may summarize with the following theorem:

Theorem 4. Consider the ellipsoid of concentration, corresponding to some distribution, of the covariance matrix of

$$\mathbf{x} = \sum_{i=1}^{m} \mathbf{W}_i \mathbf{x}_i .$$

Then the elementary symmetric functions of the semi-axes of the ellipsoid are minimized by the choice of the optimal estimate set of matrices $\mathbf{W}_i$, $i = 1, \cdots, m$, given by Eq. (4).

Corollary. The volume of the ellipsoid of concentration is minimized by this choice.

7. Conclusions

We hope that some insight into the combination of estimates has been provided. Although we have chosen a situation which arises frequently, it is by no means inclusive. Situations also arise in which the estimates are either biased or dependent or both. The most common estimation scheme – the Gauss-Markoff estimation scheme (Ref. 3) – fits well into the framework of our results.

B. The Ray Equation in the Solar Corona

P. Reichley

1. Summary

If a spacecraft is being tracked from Earth, and it happens that the signal passes through the solar corona, the returning wave has a doppler shift and delay time due to some effects of the corona. These effects are important in high-precision tracking; conversely, they also provide information on the corona.

The *ray equation* is an equation arising in the application of geometrical optics to approximate wave propagation. As a wave propagates through a nonhomogeneous medium such as the corona, the nonhomogeneity causes a change in the phase index and group index of the wave. These changes then cause corresponding changes in the doppler shift and delay time.

In the case of a corona with spherical symmetry, the doppler and range equations are functions of the distance of closest approach of the ray to the center of the corona. As this information is contained in the ray equation, we are therefore concerned with the solution of the ray equation for the distance of closest approach. A numerical

solution for the ray equation in the solor corona was presented in *SPS 37-32*, Vol. IV, pp. 273–276. Since this solution takes large amounts of computer time, an analytical solution was sought. As a result, the ray equation was finally solved for a model of the solar corona.

2. Introduction

If the index of refraction of a medium has spherical symmetry, then the path of a ray through the medium lies in a plane. As we will consider such an index, it is convenient to work in polar coordinates. Let r_1 and r_2 be the radius vectors to the endpoints of the ray path and let θ be their angular separation. Then, for a medium with index of refraction $n(r)$, the ray equation is

$$\theta = \int_{r_m}^{r_1} \frac{bdr}{r(r^2n^2(r) - b^2)^{1/2}} + \int_{r_m}^{r_2} \frac{bdr}{r(r^2n^2(r) - b^2)^{1/2}} \tag{1}$$

where r_m is the radius vector to the point of closest approach, and $b = r_m n(r_m)$ is the impact parameter of the ray.

For the solar corona, the index of refraction we wish to consider is given by Allen's revision (Ref. 4) of Baumbach's formula for electron distribution, and is of the form (r in solar radii)

$$n^2(r) = 1 - \alpha\left(\frac{\beta}{r^6} + \frac{\gamma}{r^{16}}\right) \tag{2}$$

where $\beta = 1.55$, $\gamma = 2.99$ and $\alpha = 80.45 \times 10^{14}/f^2$, where f is the frequency of the wave. (This formula is good only if we neglect the magnetic field of the Sun.) This index is, of course, the phase index of refraction as that is the index which applies to the ray equation.

As the frequencies we wish to consider are such that $f \sim 10^9$ cps or greater, we see that $\alpha << 1$. As a result, we will seek solutions that are of second order in α and we will neglect higher order terms. *For convenience we shall not even indicate the presence of higher order terms.*

To solve the ray equation (Eq. 1) for r_m involves solving an integral equation. We are given θ, r_1, r_2, α, β, and γ, and we wish to determine r_m. As this is not an integral equation in the classical sense, a new method of attack must necessarily be found.

3. An Approximation for θ

Let us look at the function in the square root of the integrand of the ray equation. We have

$$\begin{aligned} r^2n^2(r) - b^2 &= r^2 - \alpha\left(\frac{\beta}{r^4} + \frac{\gamma}{r^{14}}\right) - r_m^2 + \alpha\left(\frac{\beta}{r_m^4} + \frac{\gamma}{r_m^{14}}\right) \\ &= r^2 - r_m^2 + \alpha\left[\beta\left(\frac{1}{r_m^4} - \frac{1}{r^4}\right) + \gamma\left(\frac{1}{r_m^{14}} - \frac{1}{r^{14}}\right)\right] \\ &= (r^2 - r_m^2)\{1 + \alpha[\beta p(r) + \gamma q(r)]\} \end{aligned}$$

where

$$\begin{aligned} p(r) &= \frac{1}{r_m^4 r^2} + \frac{1}{r_m^2 r^4} \\ q(r) &= \frac{1}{r_m^{14} r^2} + \frac{1}{r_m^{12} r^4} + \cdots + \frac{1}{r_m^4 r^{12}} + \frac{1}{r_m^2 r^{14}}. \end{aligned} \tag{3}$$

Since r is expressed in solar radii and, therefore, has minimum value 1, and since $p(r)$ and $q(r)$ are monitonically decreasing functions, we have $\max \alpha[\beta p(r) + \gamma q(r)] = \alpha(2\beta + 7\gamma) = 3.39 \times 10^{-2}$.

Let us then consider the expansion of the square root. We have

$$(r^2n^2 - b^2)^{-1/2} = (r^2 - r_m^2)^{-1/2}\left\{1 - \frac{1}{2}\alpha[\beta p(r) + \gamma q(r)] + \frac{3}{8}\alpha^2[\beta p(r) + \gamma q(r)]^2\right\}$$

We can now express

$$\begin{aligned} \theta = b\Bigg\{&\int_{r_m}^{r_1} \frac{dr}{r(r^2 - r_m^2)^{1/2}} + \int_{r_m}^{r_2} \frac{dr}{r(r^2 - r_m^2)^{1/2}} \\ &- \frac{1}{2}\alpha\left(\int_{r_m}^{r_1} \frac{[\beta p(r) + \gamma q(r)]\,dr}{r(r^2 - r_m^2)^{1/2}} + \int_{r_m}^{r_2} \frac{[\beta p(r) + \gamma q(r)]\,dr}{r(r^2 - r_m^2)^{1/2}}\right) \\ &+ \frac{3}{8}\alpha^2\left(\int_{r_m}^{r_1} \frac{[\beta p(r) + \gamma q(r)]^2\,dr}{r(r^2 - r_m^2)^{1/2}} + \int_{r_m}^{r_2} \frac{[\beta p(r) + \gamma q(r)]^2\,dr}{r(r^2 - r_m^2)^{1/2}}\right)\Bigg\}. \end{aligned} \tag{4}$$

Upon examining Eqs. (3) and (4), we see that all the integrals involved in terms with coefficients α and α^2 are of the form

$$I_n \equiv \int_{r_m}^{r_a} \frac{dr}{r^n(r^2 - r_m^2)^{1/2}} \qquad n \text{ odd}, \; r_a = r_1, r_2.$$

At this point we shall assume that $r_a \geqq 10^2$. This includes a wide number of possibilities of orbits as the orbit of Venus is approximately 151 solar radii. We may then approximate I_n. We have

$$I_n = \int_{r_m}^{\infty} \frac{dr}{r^n (r^2 - r_m^2)^{1/2}} - \int_{r_a}^{\infty} \frac{dr}{r^n (r^2 - r_m^2)^{1/2}}. \tag{5}$$

But

$$\int_{r_a}^{\infty} \frac{dr}{r^n (r^2 - r_m^2)^{1/2}} \leqq \frac{1}{r^{n-1}} \frac{\pi}{2} \leqq \frac{1}{10^{2n-2}} \frac{\pi}{2}. \tag{6}$$

The lowest degree of n appearing is $n = 3$ with an α coefficient, and the next lowest is $n = 5$ with both α and α^2 as coefficients. For $n = 3$

$$I_3 = \frac{1}{2}\left\{(r_a^2 - r_m^2)^{1/2} \frac{1}{r_m^2 r_a^2} + \frac{1}{r_m^3}\left(\frac{\pi}{2} - \sin^{-1}\frac{r_m}{r_a}\right)\right\}$$
$$= \frac{1}{2}\left(\frac{1}{r_m^3}\frac{\pi}{2} - \frac{2}{3}\frac{1}{r_a^3} - \frac{1}{5}\frac{r_m^2}{r_a^5} - \cdots\right)$$

and since $r_a \geqq 10^2$,

$$I_3 = \frac{1}{2}\frac{1}{r_m^3}\frac{\pi}{2} = \int_{r_m}^{\infty} \frac{dr}{r^3 (r^2 - r_m^2)^{1/2}}$$

for our order of approximation. For $n \geqq 5$, we see from Eqs. (5) and (6) that we can use

$$I_n = \int_{r_m}^{\infty} \frac{dr}{r^n (r^2 - r_m^2)^{1/2}}$$

for our approximation. Hence, for

$$n = 2k + 1; \qquad k = 1, 2, \cdots,$$

we have

$$I_n = \binom{2k}{k} \frac{1}{2^{2k} r_m^{2k+1}} \frac{\pi}{2} \equiv A_n \frac{1}{r_m^{2k+1}} \frac{\pi}{2}. \tag{7}$$

We may now express Eq. (4) as

$$\theta = b\left\{\int_{r_m}^{r_1} \frac{dr}{r(r^2 - r_m^2)^{1/2}} + \int_{r_m}^{r_2} \frac{dr}{r(r^2 - r_m^2)^{1/2}} - \alpha \int_{r_m}^{\infty} \frac{[\beta p(r) + \gamma q(r)]\, dr}{r(r^2 - r_m^2)^{1/2}} + \frac{3\alpha^2}{4} \int_{r_m}^{\infty} \frac{[\beta p(r) + \gamma q(r)]^2 dr}{r(r^2 - r_m^2)^{1/2}}\right\}$$

which yields, from using Eqs. (3) and (7),

$$\theta = b\left[\frac{1}{r_m}\left(\frac{\pi}{2} - \sin^{-1}\frac{r_m}{r_1} - \sin^{-1}\frac{r_m}{r_2}\right) - \alpha\left(\frac{d_1\beta}{r_m^7} + \frac{d_2\gamma}{r_m^{17}}\right)\frac{\pi}{2} + \frac{3\alpha^2}{4}\left(\frac{d_3\beta^2}{r_m^{13}} + \frac{d_4\beta\gamma}{r_m^{23}} + \frac{d_5\gamma^2}{r_m^{33}}\right)\frac{\pi}{2}\right] \tag{8}$$

where

$$d_1 = A_3 + A_5$$

$$d_2 = A_3 + A_5 + \cdots + A_{15}$$

$$d_3 = A_5 + A_7 + A_9$$

$$d_4 = A_5 + 2A_7 + \cdots + 2A_{17} + A_{19}$$

$$d_5 = A_5 + 2A_7 + 3A_9 + \cdots + 6A_{15} + 7A_{17} + 6A_{19} + \cdots + 3A_{25} + 2A_{27} + A_{29}.$$

Since

$$b = r_m\left[1 - \alpha\left(\frac{\beta}{r_m^6} + \frac{\gamma}{r_m^{16}}\right)\right]^{1/2}$$
$$= r_m\left[1 - \frac{\alpha}{2}\left(\frac{\beta}{r_m^6} + \frac{\gamma}{r_m^{16}}\right) - \frac{3\alpha^2}{8}\left(\frac{\beta}{r_m^6} + \frac{\gamma}{r_m^{16}}\right)^2\right]$$

we have, from Eq. (8),

$$\theta = \pi - \sin^{-1}\frac{r_m}{r_1} - \sin^{-1}\frac{r_m}{r_2} - \frac{\alpha}{2}\left[\left(\frac{a_1\beta}{r_m^6} + \frac{a_2\gamma}{r_m^{16}}\right)\pi - \left(\frac{\beta}{r_m^6} + \frac{\gamma}{r_m^{16}}\right)\left(\frac{r_m}{r_1} + \frac{r_m}{r_2}\right)\right] + \frac{\alpha^2}{8}\left[\left(\frac{a_3\beta^2}{r_m^{12}} + \frac{a_4\beta\gamma}{r_m^{22}} + \frac{a_5\gamma^2}{r_m^{32}}\right)\pi\right] \tag{9}$$

where

$$a_1 = d_1 + 1$$
$$a_2 = d_2 + 1$$
$$a_3 = 3d_3 + 2d_1 - 1$$
$$a_4 = 3d_4 + 2(d_1 + d_2) - 2$$
$$a_5 = 3d_5 + 2d_2 - 1.$$

We have used the approximation $\sin^{-1} r_m/r_a = 1/r_a$ in Eq. (9) for terms involving α and neglected $\sin^{-1} r_m/r_a$

entirely in terms containing α^2, since this is within our order of approximation. Using the expressions for d_i, $i = 1, \cdots, 5$, and A_n as given by Eq. (7), we have

$$\begin{aligned} a_1 &= 15/8 \\ a_2 &= 6435/2048 \\ a_3 &= 585/128 \\ a_4 &= 817845/32768 \\ a_5 &= 1134308445/33554432 . \end{aligned} \tag{10}$$

Hence, Eqs. (9) and (10) yield the desired solution for θ, with the restrictions that (1) the frequency of the wave must be on the order of 10^9 cps or greater, and (2) the radius vectors to the endpoints of the ray must be 10^2 solar radii or greater.

In Fig. 1, we have plotted θ versus r_m for different frequencies, using Eq. (9). We have set $r_1 = 215$ solar radii, which is the approximate Earth radius vector, and $r_2 = 100$ solar radii. We see that as the frequency increases the ray geometry approaches that of the straight line.

Since θ first reaches a maximum and then is monotonically decreasing with increasing r_m, we see that r_m is a double valued function of θ. The effect of this double-valued property on range and doppler was discussed in *SPS 37-32*, Vol. IV, pp. 273–276. In Table 1, we have analyzed this property further. We note that the region of double valuedness and the value of r_m for which θ is maximum both increase with increasing r_2 for a fixed frequency f. For r_2 fixed, the region of double valuedness and the r_m for which θ is maximum both decrease with increasing frequency.

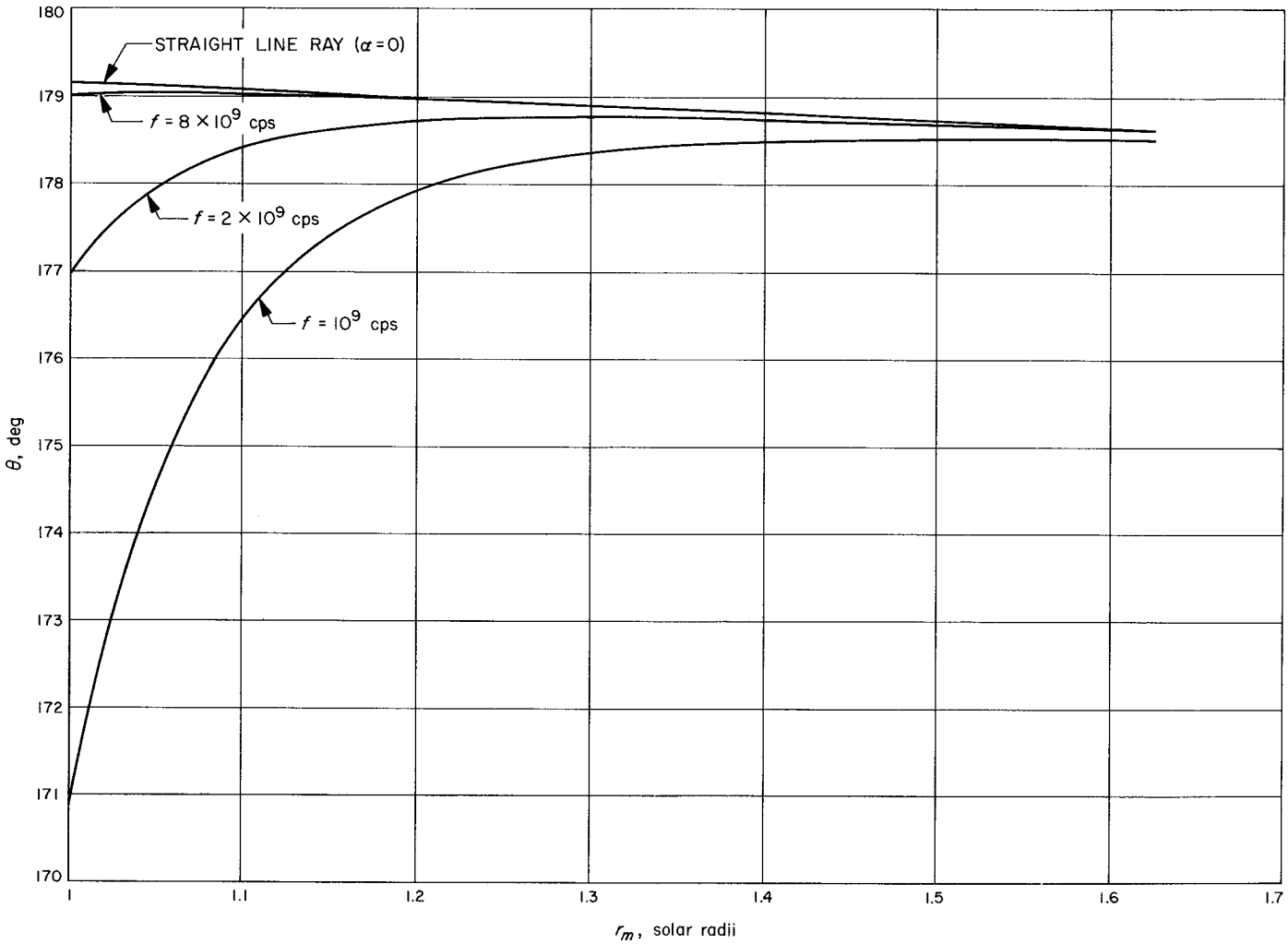


Fig. 1. θ versus r_m for different f

Table 1. Values of θ for $r_1 = 215$ solar radii

r_2, solar radii	θ at $r_m = 1$, deg	θ max, deg	r_m at θ max, solar radii	(θ max) − (θ at $r_m = 1$), deg	Frequency f, cps
100	170.894	178.547	1.5013	7.653	10^9
150	171.081	178.838	1.5500	7.756	
200	171.175	178.987	1.5818	7.812	
250	171.231	179.078	1.6044	7.847	
300	171.269	179.140	1.6212	7.871	
350	171.296	179.184	1.6343	7.889	
100	176.977	178.775	1.2965	1.798	2×10^9
150	177.167	179.025	1.3285	1.858	
200	177.262	179.153	1.3495	1.891	
250	177.319	179.231	1.3644	1.911	
300	177.357	179.283	1.3758	1.926	
350	177.384	179.231	1.3847	1.936	
100	178.180	178.874	1.2103	0.694	3×10^9
150	178.371	179.108	1.2358	0.737	
200	178.466	179.226	1.2525	0.760	
250	178.523	179.298	1.2643	0.775	
300	178.561	179.347	1.2727	0.785	
350	178.588	179.382	1.2804	0.793	
100	178.607	178.933	1.1578	0.326	4×10^9
150	178.798	179.156	1.1810	0.358	
200	178.893	179.270	1.1953	0.376	
250	178.951	179.338	1.2057	0.388	
300	178.989	179.384	1.2134	0.396	
350	179.016	179.418	1.2195	0.402	
100	179.022	179.046	1.0534	0.024	8×10^9
150	179.213	179.248	1.0715	0.036	
200	179.308	179.351	1.0829	0.043	
250	179.365	179.413	1.0909	0.048	
300	179.404	179.455	1.0976	0.052	
350	179.431	179.485	1.1014	0.054	

θ and r_m at θ max obtained from $d\theta/dr_m = 0$.

4. The Solution for r_m

The double valuedness of r_m with respect to θ reveals that for a given geometry—i.e., θ, r_1, and r_2—there are actually *two* paths in the double-valued region. We shall consider only the path corresponding to the greatest r_m for a given θ; i.e., the path farthest from the Sun. We choose this path because the path passing closest to the Sun presents technical problems in the reception of the signal, such as high background noise temperature due to the Sun lying in the beam. Hence, when we now refer to θ and r_m, we shall mean the larger r_m if θ is in the region of double valuedness. The function $\theta = \theta(r_m)$ is then monotonically decreasing in this region.

We see from Fig. 1 that the function $\theta = \theta(r_m)$ for a straight-line path is a close approximation to the curved path. Since the straight line is obtained by setting $\alpha = 0$ in Eq. (9), this suggests using a perturbation technique to solve for r_m. Hence, let

$$r_m = R_0 + \alpha R_1 + \alpha^2 R_2 \,. \tag{11}$$

Substituting Eq. (11) in Eq. (9) yields

$$\theta = \pi - \sin^{-1}\frac{R_0}{r_1} - \sin^{-1}\frac{R_0}{r_2} - \alpha\left(\frac{R_1}{(r_1^2 - R_0^2)^{1/2}} + \frac{R_1}{(r_2^2 - R_0^2)^{1/2}}\right)$$
$$- \alpha^2\left(\frac{R_2}{(r_1^2 - R_0^2)^{1/2}} + \frac{R_2}{(r_2^2 - R_0^2)^{1/2}} + \frac{1}{2}\frac{R_0 R_1^2}{(r_1^2 - R_0^2)^{3/2}} + \frac{1}{2}\frac{R_0 R_1^2}{(r_2^2 - R_0^2)^{3/2}}\right)$$
$$- \frac{\alpha}{2}\left[\left(\frac{a_1\beta}{R_0^6} + \frac{a_2\gamma}{R_0^{16}}\right)\pi - \left(\frac{\beta}{R_0^5} + \frac{\gamma}{R_0^{15}}\right)\left(\frac{1}{r_1} + \frac{1}{r_2}\right)\right]$$
$$+ \alpha^2\left[\frac{1}{8}\left(\frac{a_3\beta^2}{R_0^{12}} + \frac{a_4\beta\gamma}{R_0^{22}} + \frac{a_5\gamma^2}{R_0^{32}}\right)\pi + \frac{1}{2}\left(\frac{6a_1\beta R_1}{R_0^7} + \frac{16a_2\gamma R_1}{R_0^{17}}\right)\pi\right].$$

Equating coefficients of powers of α, we have the following set of equations:

$$\theta = \pi - \sin^{-1}\frac{R_0}{r_1} - \sin^{-1}\frac{R_0}{r_2}$$

$$\frac{R_1}{(r_1^2 - R_0^2)^{1/2}} + \frac{R_1}{(r_2^2 - R_0^2)^{1/2}} = -\frac{1}{2}\left[\left(\frac{a_1\beta}{R_0^6} + \frac{a_2\gamma}{R_0^{16}}\right)\pi - \left(\frac{\beta}{R_0^5} + \frac{\gamma}{R_0^{15}}\right)\left(\frac{1}{r_1} + \frac{1}{r_2}\right)\right]$$

$$\frac{R_2}{(r_1^2 - R_0^2)^{1/2}} + \frac{R_2}{(r_2^2 - R_0^2)^{1/2}} + \frac{1}{2}\frac{R_0R_1^2}{(r_1^2 - R_0^2)^{3/2}} + \frac{1}{2}\frac{R_0R_1^2}{(r_2^2 - R_0^2)^{3/2}} = \left[\frac{1}{8}\left(\frac{a_3\beta^2}{R_0^{12}} + \frac{a_4\beta\gamma}{R_0^{22}} + \frac{a_5\gamma^2}{R_0^{32}}\right) + \frac{1}{2}\left(\frac{6a_1\beta R_1}{R_0^7} + \frac{16a_2\gamma R_1}{R_0^{17}}\right)\right]\pi .$$

Solving for R_0, R_1, and R_2, we have

$$R_0 = \frac{r_1r_2\sin\theta}{(r_1^2 + r_2^2 - 2r_1r_2\cos\theta)^{1/2}} \equiv r_s \tag{12a}$$

$$R_1 = -\frac{1}{2}\frac{(r_1^2 - r_s^2)^{1/2}(r_2^2 - r_s^2)^{1/2}}{(r_1^2 - r_s^2)^{1/2} + (r_2^2 - r_s^2)^{1/2}}\left[\left(\frac{a_1\beta}{r_s^6} + \frac{a_2\gamma}{r_s^{16}}\right)\pi - \left(\frac{\beta}{r_s^5} + \frac{\gamma}{r_s^{15}}\right)\left(\frac{1}{r_1} + \frac{1}{r_2}\right)\right] \tag{12b}$$

$$\begin{aligned} R_2 = {} & \frac{1}{8}\frac{(r_1^2 - r_s^2)^{1/2}(r_2^2 - r_s^2)^{1/2}}{(r_1^2 - r_s^2)^{1/2} + (r_2^2 - r_s^2)^{1/2}}\left(\frac{a_3\beta^2}{r_s^{12}} + \frac{a_4\beta\gamma}{r_s^{22}} + \frac{a_5\gamma^2}{r_s^{32}}\right)\pi \\ & - \frac{1}{4}\frac{(r_1^2 - r_s^2)(r_2^2 - r_s^2)}{[(r_1^2 - r_s^2)^{1/2} + (r_2^2 - r_s^2)^{1/2}]^2}\left(\frac{a_1\beta}{r_s^6} + \frac{a_2\gamma}{r_s^{16}}\right)\left(\frac{6a_1\beta}{r_s^7} + \frac{16a_2\gamma}{r_s^{17}}\right)\pi^2 \\ & - \frac{1}{8}r_s\frac{(r_1^2 - r_s^2)^{3/2} + (r_2^2 - r_s^2)^{3/2}}{[(r_1^2 - r_s^2)^{1/2} + (r_2^2 - r_s^2)^{1/2}]^3}\left(\frac{a_1\beta}{r_s^6} + \frac{a_2\gamma}{r_s^{16}}\right)^2\pi^2 . \end{aligned} \tag{12c}$$

Eqs. (10), (11), and (12) form the desired solution. Note that in Eq. (12a) we have observed that $R_0 = r_s$, where r_s is the distance of closest approach of the straight line connecting the two ends of the ray path. Hence, our solution for r_m is a function of r_s; this property is very convenient for purposes of comparing range and doppler in the medium against range and doppler provided by the ephemeris.

To sum up, the restrictions on our solution for r_m are:

(1) The frequency of the wave must be approximately 10^9 cps or greater.

(2) The radius vectors to the endpoints of the ray path must be 100 solar radii or greater.

(3) When r_m is double valued with respect to θ, we select the larger value of r_m; i.e., the path passing farthest from the Sun.

5. Conclusions

The method used in this paper to solve the ray equation for r_m can be generalized to any medium with spherical symmetry and an index of refraction $n(r)$ of the form

$$n^2(r) = 1 \pm \varepsilon\sum_{k=0}^{m}\frac{a_k}{r^{2k}}, \qquad a_k \geqq 0, k = 1, 2, \cdots, m \tag{13}$$

where $\varepsilon << 1$ and

$$\sum_{k=0}^{m} a_k \leqq 1 .$$

(We have assumed that r has been normalized so that its smallest value is one.) One first obtains

$$\begin{aligned} r^2n^2(r) - b^2 &= r^2 \pm \varepsilon\sum_{k=0}^{m}\frac{a_k}{r^{2k-2}} - r_m^2 \mp \varepsilon\sum_{k=0}^{m}\frac{a_k}{r_m^{2k-2}} \\ &= r^2 - r_m^2 \mp \varepsilon\sum_{k=0}^{m} a_k\left(\frac{1}{r_m^{2k-2}} - \frac{1}{r^{2k-2}}\right) \\ &= (r^2 - r_m^2)\left(1 \mp \varepsilon\sum_{k=0}^{m} a_k\sum_{i=1}^{k}\frac{1}{r^{2i}r_m^{2k-2i+2}}\right) \\ &= (r^2 - r_m^2)\left[1 \mp \varepsilon\sum_{i=1}^{m} r_m^{2i}\left(\sum_{k=i}^{m}\frac{a_k}{r_m^{2k+2}}\right)\frac{1}{r^{2i}}\right]. \end{aligned}$$

We see that this equation is precisely in the form in which we obtained $r^2n^2(r) - b^2$ in this paper. Hence, the index given by Eq. (13) can be used together with this technique.

References

1. Marcus, M., and Lopes, L., "Inequalities For Symmetric Functions and Hermitian Matrices," *Canadian Journal of Mathematics*, Vol. 9, 1957, pp. 305–312.

2. Cramer, H., *Mathematical Methods of Statistics*, Princeton Mathematical Series 9, Princeton University Press, Princeton, N. J., 1957, pp. 300–301.

3. Zelen, M., *The Role of Constraints in the Theory of Least Squares*, MRC Technical Summary Report 314, University of Wisconsin, Mathematics Research Center, Madison, Wis., 1962.

4. Allen, C. W., "Interpretation of Electron Densities from Corona Brightness," *Monthly Notices of the Royal Astronomical Society*, Vol. 107, 1947, pp. 426–430.