NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Space Programs Summary 37-53, Vol. III

Supporting Research and Advanced Development

For the Period August 1 to September 30, 1968

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

October 31, 1968

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Space Programs Summary 37-53, Vol. III

Supporting Research and Advanced Development

For the Period August 1 to September 30, 1968

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

October 31, 1968

SPACE PROGRAMS SUMMARY 37-53, VOL. III

Prepared Under Contract No. NAS 7-100
National Aeronautics and Space Administration

Preface

The Space Programs Summary is a multivolume, bimonthly publication that presents a review of technical information resulting from current engineering and scientific work performed, or managed, by the Jet Propulsion Laboratory for the National Aeronautics and Space Administration. The Space Programs Summary is currently composed of four volumes:

Vol. I. *Flight Projects* (Unclassified)

Vol. II. *The Deep Space Network* (Unclassified)

Vol. III. *Supporting Research and Advanced Development* (Unclassified)

Vol. IV. *Flight Projects and Supporting Research and Advanced Development* (Confidential)

Contents

SYSTEMS DIVISION

PROJECT ENGINEERING DIVISION

GUIDANCE AND CONTROL DIVISION

Contents (contd)

Contents (contd)

PROPULSION DIVISION

SPACE SCIENCES DIVISION

Contents (contd)

Contents (contd)

I. Systems Analysis Research

SYSTEMS DIVISION

A. Recent Development Ephemerides and the Mass of Mercury, *W. G. Melbourne and D. A. O'Handley*

The ephemeris development activity has completed the revision of the radar data set to reflect the current available information. A new determination of the mass of Mercury has been made.

Developmental ephemeris (DE) 40[1] was the last DE fit to planetary radar and optical data. A version of DE 40 (DE 43[2]) has an updated lunar ephemeris (LE) 6 (Ref. 1) instead of LE 4, which is on the DE 40 tapes.

Since the announcement of DE 40 on March 29, 1968, the radar data set has been expanded (Table 1).

Eighty-six observations from Arecibo, taken in 1967, have been removed from the current data set because of their anomalous character with respect to other observations made during the same time interval. An additional four Venus observations were removed because they had residuals of over 3σ when compared with DE 40.

Having completed this updating of the data set, a series of ephemerides were made. Initially, the data set was compared with DE 40 and a solution made for 21 parameters (SPS 37-51, Vol. III, pp. 4–13). Most of the corrections, although small, were significant to at least one figure with respect to the formal standard deviations (Table 2).

[1]O'Handley, D. A., *Announcement of JPL Developmental Ephemerides 39 and 40*, Mar. 29, 1968 (JPL internal documents).

[2]Mulholland, J. D., *Announcement of Developmental Ephemeris 43*, July 24, 1968 (JPL internal document).

As seen in Table 2, the astronomical unit (AU) and radii of Venus and Mars were not changed significantly from the DE 40 values. The correction to the radius of

Table 1. Radar-range data status

Planet	Observatory	Number	Period
Mercury[a]	Arecibo	157	1964–1968
	Haystack	63	1967
Venus[b]	Arecibo	106	1964, 1965/1966, 1968
	JPL	284	1964–1967
	Haystack	49	1967
	Millstone	101	1964–1967
Mars	Arecibo	39	1964/1965
	Haystack	10	1967
[a]*This data set includes the following additions:*			
Mercury	Arecibo	11	14 May, 1968–9 June, 1968
	Haystack	63	26 Oct, 1966–12 Sept, 1967
[b]This data set includes the following additions:			
Venus	Arecibo	4	10 May, 1968–21 May, 1968
Note: The additional Mercury and Venus Arecibo, and Mercury Haystack, observations were provided by I. I. Shapiro in private communications to the authors.			

Table 2. Planetary radius and astronomical unit values

Radius unit	DE 40 value, km	Correction to DE 40 value, km
Planetary radius		
Mercury	2437.3	$+8.7 \pm 0.5$
Venus	6055.8	$+0.7 \pm 0.4$
Mars	3375.3	$+0.3 \pm 12.4$
Astronomical unit		
—	149, 597, 895.8	$+1.3 \pm 0.4$

Mercury was significant and reflects the expansion of the data set.

The DE 45 was created by applying these corrections to the DE 40 starting conditions at epoch JD 244 0800.5. The time span is JD 243 8400.5 to 244 0800.5 (January 6, 1964 to August 2, 1970). The planetary masses used in this integration were those given in SPS 37-45, Vol. IV, pp. 17–19. A comparison of DE 45 with the radar observations and a subsequent solution indicated that no further iteration was required for this set of data.

As reported in SPS 37-51, Vol. III, a definite signature appeared in the residuals of Venus time-delay measurements taken during 1965–1966 (see p. 10, Fig. 7, of SPS 37-51, Vol. III). At that time, it was conjectured that this effect might be due to second-order effects of fixed parameters. Also, intensive studies were made in search of program errors in the solar-system data processing system (SSDPS), and the choice of planetary masses came under scrutiny. Although Venus, earth–moon, and Mars mass values are very precisely known due to spacecraft radio tracking, the mass of Mercury has been poorly determined since its value depends on perturbation analyses of neighboring and minor planets. Clemence (Ref. 2) discusses the various determinations of the mass of Mercury and, in terms of reciprocal solar masses, the variation of the individual determinations amount to 8% of the quoted value. Subsequently, a brief analysis of "periodic perturbations of the longitude and radius vector of Venus" from Newcomb's Tables verified that a variation of this size in the mass of Mercury produced an effect on Venus time-delay observables of the same order of magnitude as the observed signature. This suggests the possibility of improving the mass of Mercury with radar observations.

Since the SSDPS does not presently generate partial derivatives of observables with respect to the mass of a perturbing body, the direct method of searching for the minimum of the weighted sum of the squared residuals was followed. Toward this end, several DEs spanning the 1964–1968 radar observation period were generated by setting the mass of Mercury to different values covering the neighborhood of uncertainty. The masses of the other planets were fixed at the values given in Ref. 3. The initial conditions in each of these ephemerides are identical with those of DE 45. In each case, a weighted least-squares fit of the up-dated radar data set discussed above, using the 21-parameter model, was performed to provide parameter corrections and predicted residuals for the subsequent iteration. However, subsequent iteration was unnecessary because the corrections were small enough that nonlinear effects could be neglected. Table 3 identifies the developmental ephemeris, the reciprocal mass of Mercury, $M_{\mathrm{☿}}^{-1}$, and the weighted sum-of-squares of the residuals after the fit, Σv^2. Here, Σv^2 is given by

$$\Sigma v^2 = \sum_{i=1}^{N} \left(\frac{O_i - C_i}{\sigma_i} \right)^2$$

where O is the observation, C is the predicted observation, and σ is the assigned standard deviation of the measurement.

Figure 1 exhibits the quadratic variation of Σv^2 with $M_{\mathrm{☿}}^{-1}$. The minimum corresponds to a value of $M_{\mathrm{☿}}^{-1} = 5.988 \times 10^6 \pm 10{,}000$. The formal standard devi-

Table 3. DEs spanning the 1964–1968 radar observation period

DE	$M_{\mathrm{☿}}^{-1} \times 10^{-6}$	Σv^2
47	5.845	1818.22
51	5.890	1641.02
46	5.935	1537.42
48	6.025	1538.17
52	5.984	1504.52

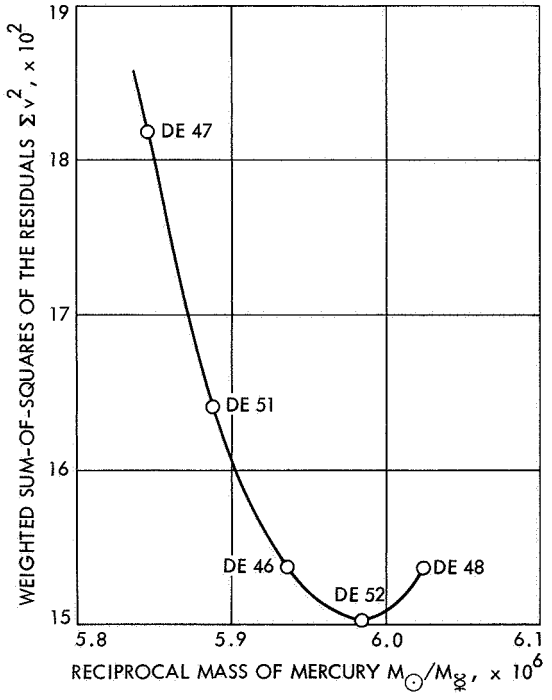


Fig. 1. The quadratic variation of Σv^2 with $M_{\mathrm{☿}}^{-1}$

ation of 10,000 quoted here is obtained from the curvature of the quadratic at the minimum point. It may be shown from estimation theory that

$$\sigma_{M_{\mathrm{Q}}^{-1}} = \left[\frac{1}{2}\frac{d^2\Sigma v^2}{d(M_{\mathrm{Q}}^{-1})^2}\right]^{-1/2}$$

evaluated at the minimum point corresponds to the standard deviation of the estimated parameter obtained from the covariance matrix for the case where M_{Q}^{-1} is included as one of the simultaneously estimated least-squares parameters. It should be stressed that this is a formal error and that this result is predicated on fixed mass values of the other planets.

It is possible to determine mass values of all of the inner planets in a simultaneous solution using the radar and optical observations of the planets. This has been

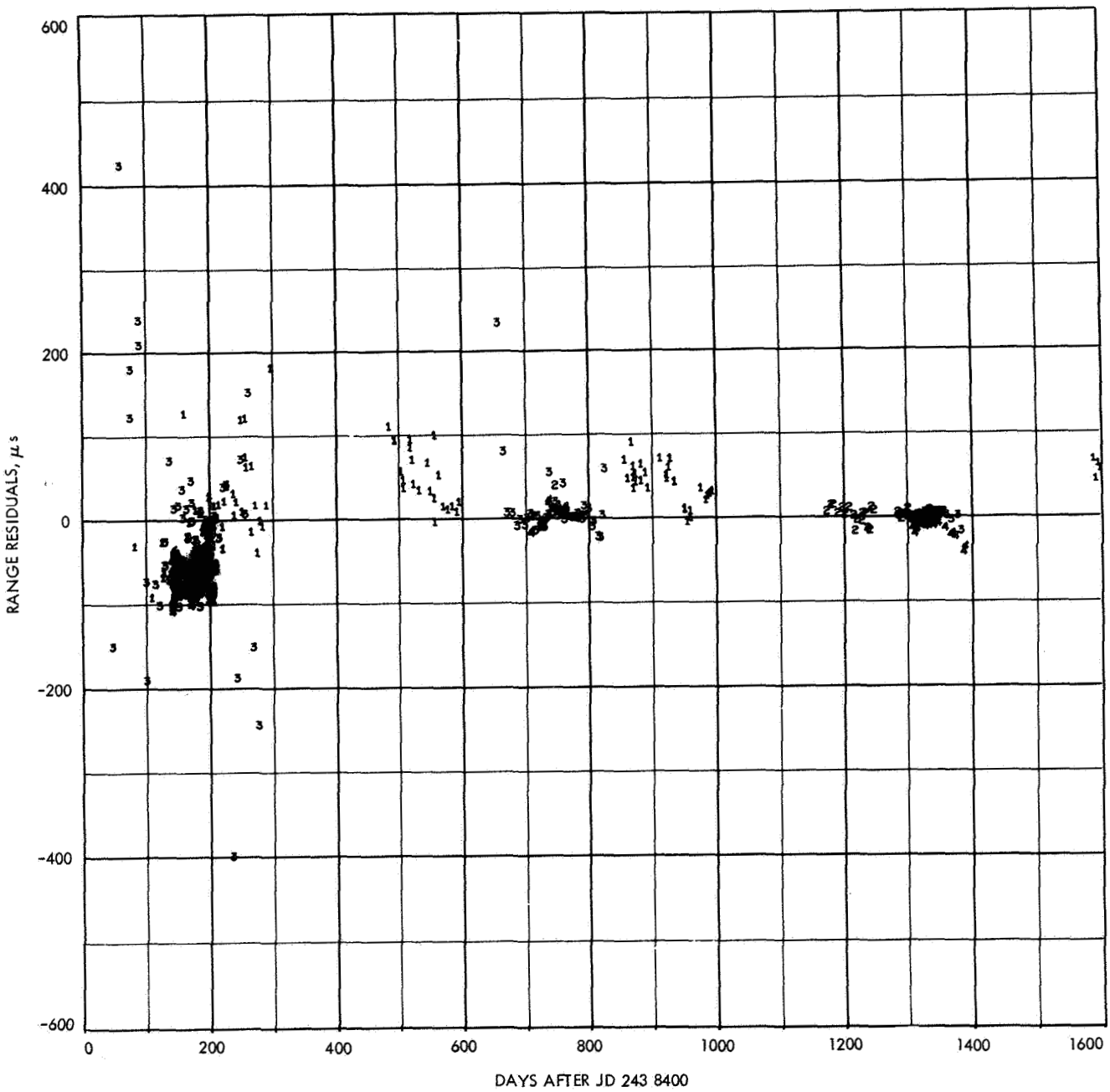


Fig. 2. Venus range residuals after fit to DE 52

done at MIT and the values of the reciprocal masses and the formal probable errors are as follows:[3]

Mercury	5,935,000 ± 45,000
Venus	408,536.5 ± 95
Earth–moon	328,896 ± 57
Mars	3,116,350 ± 6,400

The mass ratio of the earth–moon is 81.301 ±0.002. While these MIT mass values for Venus, earth–moon, and Mars are consistent with the values obtained for these quantities from the *Ranger* and *Mariner* series of spacecraft, the latter estimates are about two orders of magnitude more precise. For this reason, it has been decided to fix the mass values of these planets at their spacecraft-determined values until it becomes practicable to simultaneously process radio tracking and radar data.

Figure 2 shows the Venus range residuals after the fit to DE 52; the anomaly around the January 1966 inferior conjunction has nearly disappeared. The conclusion to be drawn from this work is that the feature in the radar-range residuals must be regarded as an anomaly in the modeling of the masses. If one considers the masses better known from spacecraft tracking, and therefore sets the values as known, the reduction of degrees of freedom will cause the "feature" to appear in the Venus residuals. By altering the mass of Mercury, the feature does disappear and the over-all sum of squared residuals is diminished.

References

1. Mulholland, J. D., *JPL Lunar Ephemeris Number 6*, Technical Memorandum 33-408, Jet Propulsion Laboratory, Pasadena, Calif., Oct. 15, 1968.
2. Clemence, G. M., "Masses of the Principal Planets." Joint Discussion on the Report of the Working Group on the IAU System of Astronomical Constants. *Trans. IAU*, Vol. XII B, p. 610, 1966.
3. Melbourne, W. G., et al., *Constants and Related Information for Astrodynamic Calculations, 1968*, Technical Report 32-1306, Jet Propulsion Laboratory, Pasadena, Calif., July 15, 1968.

B. A New Variation-of-Parameters Method With Universal Variables, *R. Broucke*

The classical variation-of-parameters methods, with the classical osculating orbit elements, have the disadvantage of not being uniformly valid for all types of conic orbits and all eccentricities or inclinations. This is essentially due to the particular representation of the conic orbit that is used. It is well known that solutions of the two-body problem exist that include, in one single formulation, the elliptic, parabolic, hyperbolic, and rectilinear cases. In these formulations, no particular attention has to be given to the small eccentricity or small inclination situations. In the recent past, several perturbation techniques have been produced that take advantage of the universal formulation of the two-body problem; e.g., S. Herrick (Ref. 1), R. H. Battin (Ref. 2), S. Pines (Ref. 3), W. H. Goodyear (Refs. 4 and 5), and others.

In the present article, a new method of variation of parameters (variation of arbitrary constants) with universal variables is given that seems to be more simple in form than most of the existing methods. It is generally accepted that a method of variation of parameters uses osculating orbits, and that the Encke method of computing perturbations uses a fixed Keplerian reference orbit. The remarkable feature of the new method described in this article is that it uses a fixed Keplerian reference orbit, but no osculating orbits or osculating orbit elements.

In previous works on variation-of-parameters methods with universal variables, six first-order differential equations for the initial positions, $\mathbf{x}_0$, and velocity, $\mathbf{v}_0$, of the osculating Keplerian orbits are always obtained. In the work discussed herein, six similar first-order differential equations are also obtained, but a fixed Keplerian reference orbit is used. This reference orbit could be osculating either at some instant or at the epoch, or it could be some well-chosen mean orbit, but this is by no means essential.

In terms of the $\mathbf{x}_0$ and $\mathbf{v}_0$ vectors, the equation for the two-body reference orbit may be written as

$$\left.\begin{aligned} \mathbf{x}_R &= f\mathbf{x}_0 + g\mathbf{v}_0 \\ \dot{\mathbf{x}}_R &= \dot{f}\mathbf{x}_0 + \dot{g}\mathbf{v}_0 \end{aligned}\right\} \tag{1}$$

For an elliptic orbit, the functions f, g and their derivatives are

$$\left.\begin{aligned} f &= \frac{a}{r_0}(\cos\theta - 1) + 1 \\ g &= t - \frac{1}{n}(\theta - \sin\theta) \\ \dot{f} &= -\frac{a^2 n}{r_R r_0}\sin\theta \\ \dot{g} &= 1 + \frac{a}{r_R}(\cos\theta - 1) \end{aligned}\right\} \tag{2}$$

[3]Private correspondence from I. I. Shapiro.

where $\theta = E - E_0$ is the difference between the eccentric anomalies at time $t(E)$ and time $t = 0(E_0)$. In general, the subscript 0 refers here to the values at the time $t = 0$.

The different elements used in Expression (2) refer to the reference orbit rather than the true perturbed orbit. The functions f and g have several interesting properties; for instance, the Wronskian determinant of f and g is

$$f\dot{g} - \dot{f}g = +1 \tag{3}$$

In fact, Eq. (3) is equivalent to Kepler's equation. On the other hand, the f and g functions are solutions of the fundamental two-body differential equations

$$\left.\begin{aligned} \ddot{f} &= -\mu \frac{f}{r_R^3} \\ \ddot{g} &= -\mu \frac{g}{r_R^3} \end{aligned}\right\} \tag{4}$$

Expression (1) is valid for all types of conics at the condition to choose the right f and g functions. A universal formulation has been produced by R. H. Battin (Ref. 2, p. 52, Eqs. 2.43 and 2.44) wherein he gives f and g functions containing two special transcendental functions S and C that generalize sin θ and cos θ in Expression (2). Also, S. Herrick has given several different forms of f and g functions in his study on the use of universal variables in the two-body problem (Ref. 1). The following derivation of a variation-of-parameters method is valid if either Battin's or anyone of the sets of f and g functions given by Herrick are used.

Let us first derive the variational equations. The equations of motion for the unperturbed reference motion are

$$\ddot{\mathbf{x}}_R = -\mu \frac{\mathbf{x}_R}{r_R} \tag{5}$$

and the corresponding equations for the perturbed motion are

$$\ddot{\mathbf{x}} = -\mu \frac{\mathbf{x}}{r^3} + \mathbf{X} \tag{6}$$

where $\mathbf{X}$ is the perturbing acceleration vector. Now defining the perturbations by

$$\delta\mathbf{x} = \mathbf{s} = \mathbf{x} - \mathbf{x}_R \tag{7}$$

we obtain, by subtracting Eq. (5) from Eq. (6), the following second-order differential equations for the components of $\mathbf{s}$:

$$\mathbf{s} + \mu \frac{\mathbf{s}}{r_R^3} = -\mu \left(\frac{1}{r^3} - \frac{1}{r_R^3}\right)\mathbf{x} + \mathbf{X} = \mathbf{F} \tag{8}$$

Thus, this is a system of non-linear differential equations called the variational equations or the "perturbation equations." Since Expression (8) is exact and is not limited to the first order, it should not be confused with the well-known first-order variational equations because no approximation and no expansion has been made in deriving it. By comparison with Expression (4), we see that the homogeneous differential equations corresponding to Expression (8) have the general solution

$$\mathbf{s} = \mathbf{K}_1 f + \mathbf{K}_2 g \tag{9}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are the six arbitrary constants. We will transform Eq. (9) in such a way that by replacing the constants $\mathbf{K}_1$ and $\mathbf{K}_2$ with the appropriate functions of time, the same expression for $\mathbf{s}$ is a solution of the non-homogeneous equations. This is accomplished using the standard method of solving non-homogeneous linear differential equations with Lagrange's method of variation of arbitrary constants. We may *a priori* give the relation between the components of $\mathbf{K}_1$ and $\mathbf{K}_2$ as

$$\dot{\mathbf{K}}_1 f + \dot{\mathbf{K}}_2 g = 0 \tag{10}$$

so that the derivative of $\mathbf{s}$ in Eq. (9) is

$$\dot{\mathbf{s}} = \mathbf{K}_1 \dot{f} + \mathbf{K}_2 \dot{g} \tag{11}$$

Taking the second derivative of $\mathbf{s}$, and using the relations given in Expression (4), we obtain the equation

$$\dot{\mathbf{K}}_1 \dot{f} + \dot{\mathbf{K}}_2 \dot{g} = \mathbf{F} \tag{12}$$

Combining Eqs. (10) and (12) now gives a system of six ordinary equations in the first derivatives of the six components of the vectors $\mathbf{K}_1$ and $\mathbf{K}_2$:

$$\left.\begin{aligned} \dot{\mathbf{K}}_1 f + \dot{\mathbf{K}}_2 g &= 0 \\ \dot{\mathbf{K}}_1 \dot{f} + \dot{\mathbf{K}}_2 \dot{g} &= \mathbf{F} \end{aligned}\right\} \tag{13}$$

Using the property given in Eq. (3), we can write the solution of Expression (13) in the form

$$\left.\begin{aligned} \dot{\mathbf{K}}_1 &= -g\mathbf{F} \\ \dot{\mathbf{K}}_2 &= +f\mathbf{F} \end{aligned}\right\} \tag{14}$$

and this is the desired final result. In Expression (14), there are six first-order differential equations that can be used for solution by numerical integration as a variation-of-parameters method, or also for the generation of a solution in the form of an iterative general perturbation theory.

The result given in Expression (14) may be written in a slightly different form:

$$\mathbf{x} = \mathbf{x}_R + \mathbf{s} = (\mathbf{x}_0 + \mathbf{K}_1) f + (\mathbf{v}_0 + \mathbf{K}_2) g \tag{15}$$

Now we introduce two new vectors

$$\left.\begin{aligned} \mathbf{X}_0 &= \mathbf{x}_0 + \mathbf{K}_1 \\ \mathbf{V}_0 &= \mathbf{v}_0 + \mathbf{K}_2 \end{aligned}\right\} \tag{16}$$

which differ from $\mathbf{K}_1$ and $\mathbf{K}_2$ only by the addition of a constant vector so that for $\mathbf{X}_0, \mathbf{V}_0$, we have the same differential equations as in Expression (14):

$$\left.\begin{aligned} \dot{\mathbf{X}}_0 &= -g\mathbf{F} \\ \dot{\mathbf{V}}_0 &= +f\mathbf{F} \end{aligned}\right\} \tag{17}$$

The coordinates in the perturbed orbit are thus given by

$$\mathbf{x} = \mathbf{X}_0 f + \mathbf{V}_0 g \tag{18}$$

As a consequence of Eq. (10), the velocity components are given by

$$\dot{\mathbf{x}} = \mathbf{X}_0 \dot{f} + \mathbf{V}_0 \dot{g} \tag{19}$$

i.e., Eq. (18) has to be differentiated as if $\mathbf{X}_0$ and $\mathbf{V}_0$ were constants, although they are functions of time, to be obtained by the integration of the first-order differential Expression (17). It is remarkable that the expressions for the coordinates and velocity in the perturbed orbit, as given by Eqs. (18) and (19), are so similar to the corresponding Expression (1) in a Keplerian orbit.

References

1. Herrick, S., "Universal Variables," *Astron. J.*, Vol. 70, No. 4, pp. 309–315, May 1965.
2. Battin, R. H., *Astronautical Guidance*, McGraw-Hill Book Co., Inc., New York, 1964.
3. Pines, S., "Variation of Parameters for Elliptic and Near Circular Orbits," *Astron. J.*, Vol. 66, No. 1, pp. 5–7, Feb. 1961.
4. Goodyear, W. H., "Completely General Closed-Form Solution for Coordinates and Partial Derivatives of the Two-Body Problem," *Astron. J.*, Vol. 70, No. 3, pp. 189–192, April 1965.
5. Goodyear, W. H., "A General Method of Variation of Parameters for Numerical Integration," *Astron. J.*, Vol. 70, No. 8, pp. 524–526, Oct. 1965.

C. Optimization of a Solar Electric Propulsion Planetary Orbiter Spacecraft, C. G. Sauer, Jr.

1. Introduction

For the past several years there has been a great deal of interest in combining large solar-array power systems with electric-propulsion thrusters for unmanned exploration of the solar system. In particular, a Jupiter flyby mission has been studied quite extensively (Refs. 1–5). In order to determine the feasibility of a solar electric propulsion (SEP) spacecraft for a particular mission, the increase in performance must be balanced against the greater cost and increased complexity, and hence possibly lower reliability, of the SEP spacecraft as compared with that of an equivalent ballistic spacecraft.

Since the SEP spacecraft being considered in the various studies have a relatively low specific power, the thrust acceleration is also quite low (10^{-5} to 5×10^{-5} g), and the use of low-thrust spiral trajectories for the escape and capture phases of the mission is precluded because of the extremely long flight times that would result. Thus, the missions being analyzed use the launch vehicle to provide some fraction of the energy required for the mission over that energy required for the initial earth parking orbit. Also, for an orbiter mission, a relatively high-thrust chemical retro-propulsion maneuver is used to place the spacecraft into the specified planetary orbit. It is, consequently, impossible to make a clear distinction between the launch vehicle and the SEP spacecraft in the mission analysis, and an optimization cannot be performed that separates the heliocentric and planet-centered phases of the mission.

2. Trajectory Optimization

The kinematic aspects of the trajectory optimization are only briefly considered here since a thorough analysis has been made by a number of previous investigators (Refs. 3 and 6). A two-body optimization of the trajectory of a thrusting SEP low-thrust spacecraft is made that simultaneously solves the equations of motion and the variational equations. The program used in this analysis has the capability of optimizing not only the path of the spacecraft, but also certain vehicle parameters such as thruster specific impulse or solar-panel output power. In addition, the program allows for coast phases of flight during the mission.

To account for the effects of the departure and arrival planets on the performance estimates to be obtained, an asymptotic velocity bias method is employed (SPS 37-36, Vol. IV, pp. 14–19, and Ref. 7). This method is based upon the observation that for the case of a thrusting spacecraft in the gravitational field of a planet, the planet-centered velocity of the spacecraft approaches an asymptotic form as the spacecraft recedes from the planet and the gravitational effects of the planet become negligible. The extrapolation of this asymptotic form to "zero" time serves to define an initial velocity bias such that a thrusting spacecraft departing from a massless planet with this initial velocity would have this asymptotic form as its velocity profile. This asymptotic velocity bias is then used to bias the initial velocity or final velocity as specified by the ephemeris of the departure or arrival planets.

The differential equations serving to define the velocity $\mathbf{V}$ and position $\mathbf{R}$ of the spacecraft are given by

$$\dot{\mathbf{V}} = -\nabla U + a\zeta \tag{1}$$

$$\dot{\mathbf{R}} = \mathbf{V} \tag{2}$$

where the gravitational potential U is given by

$$U = -\frac{GM}{r} \tag{3}$$

for an inverse-square central force field. The magnitude of the vector position $\mathbf{R}$ is given by r. The magnitude a of the thrust acceleration in Eq. (1) is given by

$$a = 2\frac{P_T}{mc} \tag{4}$$

where P_T is the thruster output power, m is the vehicle mass, and c is the effective exhaust velocity of the thrusters. The thrust is aligned in a direction given by the unit thrust vector ζ. The mass of the spacecraft is found by solving the differential equation

$$\dot{m} = -2\frac{P_T}{c^2} \tag{5}$$

The P_T is equal to the thruster electrical input power P_I decreased by the efficiency factor η of the thrusters:

$$P_T = \eta P_I \tag{6}$$

This thruster efficiency is a function of the power conditioning efficiency, thruster mass utilization, and other losses in the thrusters proportional to the thruster specific impulse.

When a requirement exists for an additional power drain from the solar panels for a spacecraft auxiliary power requirement, the P_I is not equal to the solar-panel output power P, but rather to this power decreased by the auxiliary power requirement ΔP:

$$P_I = P - \Delta P \tag{7}$$

Since P and P_I are not necessarily equal, an overall powerplant specific mass, which includes both solar-panel and thruster specific masses, must be defined and used with caution. When an auxiliary power requirement exists, the overall specific mass must be separated into a solar-panel specific mass α_w defined by

$$\alpha_w = \frac{m_w}{P_0} \tag{8}$$

and a thruster subsystem specific mass α_{th} defined by

$$\alpha_{th} = \frac{m_{th}}{P_I} = \frac{m_{th}}{P_0 - \Delta P} \tag{9}$$

The P_0 at 1 AU is used to define α_w.

The overall powerplant mass m_{pp}, defined as the sum of the solar-panel mass m_w, and the thruster mass m_{th} is given by

$$m_{pp} = m_w + m_{th} = (\alpha_w + \alpha_{th})\, P_0 - \alpha_{th}\, \Delta P \tag{10}$$

or

$$m_{pp} = \alpha P_0 - \alpha_{th}\, \Delta P \tag{11}$$

with the overall specific mass α being given by

$$\alpha = \alpha_w + \alpha_{th} \tag{12}$$

A constant power drain from the solar panels complicates the analysis since the panel output power varies as a function of the distance of the spacecraft from the sun. Denoting $\gamma_p(r)$ as the normalized variation of P with solar distance r, P is given by

$$P = P_0\, \gamma_p(r) \tag{13}$$

and the P_I is consequently

$$P_I = P_0\, \gamma_p(r) - \Delta P \tag{14}$$

The thrust acceleration [Eq. (4)] and mass [Eq. (5)] can thus be given as a function of P_0:

$$a = \frac{2\eta}{mc}\left[P_0\,\gamma_p(r) - \Delta P\right] \tag{15}$$

$$\dot{m} = -\frac{2\eta}{c^2}\left[P_0\,\gamma_p(r) - \Delta P\right] \tag{16}$$

Two additional differential equations are considered in the formulation, although they are not actually used in the trajectory program. These equations

$$\dot{P}_0 = 0 \tag{17}$$

and

$$\dot{c} = 0 \tag{18}$$

serve to define two vehicle parameters that can be optimized.

The modified Hamiltonian H for this problem is

$$H = -\left(\boldsymbol{\lambda}\cdot\mathbf{V} + \dot{\boldsymbol{\lambda}}\cdot\nabla U\right) + \frac{2\eta}{m_0 c}\left[P_0\gamma_p(r) - \Delta P\right] L \tag{19}$$

where m_0 denotes the initial mass of the spacecraft and L is the so-called thrust switching function (Ref. 6) defined by

$$L = \frac{\zeta\cdot\boldsymbol{\lambda}}{m/m_0} - \frac{m_0}{c}\lambda_m \tag{20}$$

and determines the periods of propulsion and coasting from

$$P_T = P_0\gamma_p(r) - \Delta P\,, \qquad L > 0 \tag{21}$$

$$P_T = 0\,, \qquad L < 0 \tag{22}$$

In Eq. (19), the $\boldsymbol{\lambda}$ are a set of LaGrange multipliers conjugate to the position and velocity of the spacecraft, and λ_m is the multiplier conjugate to the mass of the spacecraft. The Euler–LaGrange equations resulting from Eq. (19) are

$$\ddot{\boldsymbol{\lambda}} = -\left(\boldsymbol{\lambda}\cdot\nabla\right)\nabla U + \frac{2\eta}{m_0 c}\left[P_0\,\nabla\gamma_p(r)\right] L \tag{23}$$

for position and velocity and

$$\dot{\lambda}_m = \frac{2\eta}{m_0 c}\left[P_0\gamma_p(r) - \Delta P\right]\frac{\zeta\cdot\boldsymbol{\lambda}}{m/m_0}\frac{1}{m} \tag{24}$$

for spacecraft mass. During a coast period, the term containing $\nabla\gamma_p(r)$ in Eq. (23), and also the right-hand side of Eq. (24), are zero.

It is, perhaps, more convenient in the formulation to determine the thrust switching function from the differential equation

$$\dot{L} = \frac{\zeta\cdot\dot{\boldsymbol{\lambda}}}{m/m_0} \tag{25}$$

and then determine λ_m from

$$\lambda_m = \frac{c}{m_0}\left(\frac{\zeta\cdot\boldsymbol{\lambda}}{m/m_0} - L\right) \tag{26}$$

The calculus of variations also yields the condition that the unit thrust vector is aligned in the direction specified by the multiplier vector $\boldsymbol{\lambda}$:

$$\zeta = \frac{\boldsymbol{\lambda}}{\lambda} \tag{27}$$

where λ is the magnitude of the multiplier vector $\boldsymbol{\lambda}$. In addition, the multipliers conjugate to the power P_0 and to the thruster exhaust velocity c are given by the differential equations

$$\dot{\lambda}_P = -\frac{2\eta}{m_0 c}\gamma_p(r)\,L \tag{28}$$

$$\dot{\lambda}_c = \frac{2\eta}{m_0 c^2}\left[P_0\gamma_p(r) - \Delta P\right]\left[\left(2 - c\frac{\eta'}{\eta}\right)L - \frac{\zeta\cdot\boldsymbol{\lambda}}{m/m_0}\right] \tag{29}$$

where η' is the derivative of η with respect to c.

The transversality condition that must be satisfied for this problem is given by

$$\left[-H\,dt + \boldsymbol{\lambda}\cdot d\mathbf{V} + \dot{\boldsymbol{\lambda}}\cdot d\mathbf{R} + \lambda_m\,dm + \lambda_P\,dP_0 + \lambda_c\,dc\right]_0^T = 0 \tag{30}$$

The conditions required for optimizing P, c, and departure and arrival energy are derived from the transversality condition given in Eq. (30).

In the discussion that follows, we will consider the initial and final time specified so that $dt = 0$ at both endpoints of the trajectory. In addition we will assume an ephemeris is employed so that $d\mathbf{R} = 0$ at both endpoints

and that $d\mathbf{V}$ is given by

$$d\mathbf{V} = V_B\, d\xi + dV_B \xi \tag{31}$$

where V_B is both the magnitude of the velocity bias and a function of the departure or arrival energy, thrust acceleration, and gravitational attraction of the attracting planet. The unit vector ξ defines the direction in which the velocity bias is directed. The transversality condition given in Eq. (30) can thus be rewritten as

$$V_{BA}\boldsymbol{\lambda}\cdot d\xi + dV_{BA}\boldsymbol{\lambda}\cdot\xi - V_{BD}\boldsymbol{\lambda}^0\cdot d\xi - dV_{BD}\boldsymbol{\lambda}^0\cdot\xi + \lambda_m\, dm - \lambda_m^0\, dm_0 + (\lambda_P - \lambda_P^0)\, dP_0 + (\lambda_c - \lambda_c^0)\, dc = 0 \tag{32}$$

where subscripts A and D refer to values of V_B at arrival and departure, respectively, and superscript zeros denote the initial values of the multipliers. In order to optimize the direction in which the velocity bias is applied, the term $\boldsymbol{\lambda}\cdot d\xi$ in Eq. (32) must be zero, implying that the velocity bias is aligned in, or opposed to, the direction defined by the multiplier vector $\boldsymbol{\lambda}$. In addition, we can set

$$\lambda_c^0 = 0 \tag{33a}$$

$$\lambda_P^0 = 0 \tag{33b}$$

without any loss of generality.

In terms of the variables appearing in Eq. (32), dV_B can be expanded to

$$dV_B = \frac{\partial V_B}{\partial C_3}\, dC_3 + \frac{\partial V_B}{\partial a}\, da \tag{34}$$

where C_3 is the *vis viva* energy. Since the thrust acceleration is a function of the m, c, and P, Eq. (34) becomes

$$dV_B = \frac{\partial V_B}{\partial C_3}\, dC_3 + \frac{\partial V_B}{\partial a}\frac{\partial a}{\partial c}\, dc + \frac{\partial V_B}{\partial a}\frac{\partial a}{\partial m}\, dm + \frac{\partial V_B}{\partial a}\frac{\partial a}{\partial P_0}\, dP_0 \tag{35}$$

where dV_B is to be evaluated at each endpoint.

The transversality condition given in Eq. (32) can also be expanded into

$$\begin{aligned} &-\boldsymbol{\lambda}\frac{\partial V_{BA}}{\partial C_{3A}}\, dC_{3A} - \lambda^0\frac{\partial V_{BD}}{\partial C_{3D}}\, dC_{3D} + \left\{\lambda_m - \lambda\frac{\partial V_{BA}}{\partial a}\frac{\partial a}{\partial m}\right\} dm - \left\{\lambda_m^0 + \lambda^0\frac{\partial V_{BD}}{\partial a}\frac{\partial a}{\partial m_0}\right\} dm_0 \\ &+ \left\{\lambda_P - \lambda\frac{\partial V_{BA}}{\partial a}\frac{\partial a}{\partial P_0} - \lambda^0\frac{\partial V_{BD}}{\partial a}\frac{\partial a}{\partial P_0}\right\} dP_0 + \left\{\lambda_c - \lambda\frac{\partial V_{BA}}{\partial a}\frac{\partial a}{\partial c} - \lambda^0\frac{\partial V_{BD}}{\partial a}\frac{\partial a}{\partial c}\right\} dc = 0 \end{aligned} \tag{36}$$

where λ is the magnitude of the multiplier $\boldsymbol{\lambda}$ at arrival and λ^0 is the magnitude at departure, and subscripts A and D denote values of V_B and C_3 at arrival and departure, respectively. The direction of the velocity bias is opposite to that of the multiplier $\boldsymbol{\lambda}$ at arrival; hence, the terms containing V_{BA} have a negative sign associated with them.

Since c does not appear in any of the differentials in Eq. (36) except the last, this last term in Eq. (36) is the expression that must equal zero for c to be optimum.

3. Spacecraft and Payload Optimization

The previous development has been devoted exclusively to the SEP spacecraft. The m_0 of the spacecraft is a function only of the initial *vis viva* energy C_{3D} through the injected weight capability of the launch vehicle. Thus, the variations in the initial mass of the spacecraft can be directly related to variations in the value of the departure energy:

$$dm_0 = \frac{dm_0}{dC_{3D}}\, dC_{3D} \tag{37}$$

The m_0 can be expressed in the following form as a function of the launch-vehicle characteristics m_s, m_d, and I_{sp}:

$$m_0 = m_s \exp\left(-\frac{V_C - V_P}{g_0 I_{sp}}\right) - m_d \tag{38}$$

where V_P is the local parabolic velocity and V_C is given by

$$V_C = (V_P^2 + C_{3D})^{1/2} \tag{39}$$

In order to match the net injected payload capability of the launch vehicle with the above constants, the specific impulse I_{sp} appearing in Eq. (38) will not necessarily equal the I_{sp} of the last stage. The term m_d appearing in Eq. (38) represents the net inert mass of the last stage of the injection vehicle, including payload adapter, that is discarded. The value of $m_s - m_d$ represents the net injected weight capability at a C_3 equal to zero. From Eq. (38), the variation of m_0 can be expressed as

$$dm_0 = -\frac{m_0 + m_d}{2V_C g_0 I_{sp}} dC_{3D} \tag{40}$$

There are two additional mass components of the low-thrust spacecraft that can be considered in the optimization. These are the structural mass m_{st} given by

$$m_{st} = k_{st} m \tag{41}$$

where k_{st} is the structural factor of the spacecraft. The low-thrust propellant tankage mass m_{pt} is defined as

$$m_{pt} = k_{pt} m_p = k_{pt}(m_0 - m) \tag{42}$$

where k_{pt} is the propellant tankage factor.

For the orbiter or rendezvous mission, there is an additional system to be defined—that of the retro-propulsion system. This system will be employed for the final capture maneuver at the target planet and consists of two parts: (1) the propellant mass, and (2) the retro-system inert mass including tankage and thruster. For the purposes of this analysis, the inert retro-system mass will be defined as a fixed fraction of the propellant mass m_{fl}:

$$m_r = k_r m_{fl} \tag{43}$$

where k_r is the retro-system inert mass fraction.

At the point in the trajectory prior to the retro-maneuver, there are several options available. For example, we can jettison parts of the spacecraft for which there is no further use, such as the electric-propulsion engines or perhaps some fraction of the solar panels. We could also consider a re-entry package being discharged at this point. The particular spacecraft we will consider, however, is one in which the entire spacecraft less retro-propellant is orbited.

At the point where the retro-maneuver is made, the speed V_C of the spacecraft with respect to the target planet is given as a function of the *vis viva* energy C_{3A}:

$$V_C^2 = V_P^2 + C_{3A} \tag{44}$$

where V_P is the local parabolic velocity at this point. Denoting the desired orbital speed after the capture maneuver by V_F, the velocity increment V_R due to the deboost maneuver is

$$V_R = V_C - V_F \tag{45}$$

and the spacecraft mass m_g after the deboost maneuver is given by

$$m_g = m \exp\left(\frac{-V_R}{g_0 I_{sp}}\right) = m E \tag{46}$$

The net payload m_f that we wish to maximize will be defined as the orbiting m_g less solar panels, electric-propulsion thrusters, structural mass, low-thrust propellant tankage mass, and retro-system inert mass:

$$m_f = m_g - m_w - m_{th} - m_{st} - m_{pt} - m_r \tag{47}$$

Since the m_{fl} is given by

$$m_{fl} = m - m_g \tag{48}$$

the m_f can be rewritten as, using Eqs. (41), (42), and (43),

$$m_f = [(1 + k_r) E - k_{st} + k_{pt} - k_r] m - m_w - m_{th} - k_{pt} m_0 \tag{49}$$

The change in m_f due to changes in m, m_w, m_{th}, m_0, and C_{3A} is consequently given by

$$dm_f = [(1 + k_r) E - k_{st} + k_{pt} - k_r] dm - dm_w - dm_{th} - k_{pt} dm_0 + \left[(1 + k_r) m \frac{\partial E}{\partial C_{3A}}\right] dC_{3A} \tag{50}$$

From Eqs. (8) and (9)

$$dm_w = \alpha_w \, dP_0 \tag{51}$$

and

$$dm_{th} = \alpha_{th} \, dP_0 \tag{52}$$

so that Eq. (50) can be rewritten as

$$dm = \frac{1}{(1+k_r)E - k_{st} + k_{pt} - k_r} \left\{ dm_f + \alpha dP_0 + k_{pt} \, dm_0 - \left[(1+k_r) \, m \frac{\partial E}{\partial C_{3A}} \right] dC_{3A} \right\} \tag{53}$$

By substituting Eqs. (53) and (37) into Eq. (36)

$$\lambda_m^* \, dm_f = \left[\lambda_m^* (1+k_r) \, m \frac{\partial E}{\partial C_{3A}} + \lambda \frac{\partial V_{BA}}{\partial C_{3A}} \right] dC_{3A} + \left[\lambda^0 \frac{\partial V_{BD}}{\partial C_{3D}} + \left(\lambda_m^0 + \lambda^0 \frac{\partial V_{BD}}{\partial m_0} - \lambda_m^* \, k_{pt} \right) \frac{dm_0}{dC_{3D}} \right] dC_{3D}$$

$$+ \left(- \lambda_P + \lambda \frac{\partial V_{BA}}{\partial P_0} + \lambda^0 \frac{\partial V_{BD}}{\partial P_0} - \lambda_m^* \, \alpha \right) dP_0 + \left(- \lambda_c + \lambda \frac{\partial V_{BA}}{\partial c} + \lambda^0 \frac{\partial V_{BD}}{\partial c} \right) dc \tag{54}$$

where λ_m^* is given by

$$\lambda_m^* = \frac{\lambda_m - \lambda \dfrac{\partial V_{BA}}{\partial a} \dfrac{\partial a}{\partial m}}{(1+k_r)E - k_{st} + k_{pt} - k_r} \tag{55}$$

The functions in Eq. (54) are those to be used in maximizing m_f with respect to arrival C_3, departure C_3, P, and c.

By considering the definition of E in Eq. (46), we can set

$$\frac{\partial E}{\partial C_{3A}} = - \frac{E}{g_0 I_{sp}} \frac{1}{2V_{CA}} \tag{56}$$

in Eq. (54). The remaining terms appearing in Eqs. (54) and (55) still to be defined are those that are functions of the velocity bias at the departure or arrival planets.

Rather than using the expressions derived in SPS 37-36, Vol. IV and Ref. 7, an approximation is made that allows the partial derivatives to be more easily calculated and, furthermore, eliminates a discontinuity in the derivatives. The velocity bias is of the form

$$V_B = F(x) \, (GM \, a)^{1/4} \tag{57}$$

where

$$x = \frac{C_3}{4 \, (GM \, a)^{1/2}} \tag{58}$$

and the function $F(x)$, which in the references quoted contains elliptic integrals, is of the form

$$F(x) = 2 \frac{(x + 0.651630)(x + 4.113609)(x + 1.214342)}{(x + 4.169068)(x + 1.303312)(x+1)^{1/2}} \tag{59}$$

Note that the above approximation is invalid for values of the parameter x that are less than or equal to -1; in fact, the approximation should not be used for values of x less than around -0.6 since the asymptotic velocity bias method that is used to calculate V_B starts to deteriorate in accuracy at about this point. The actual path of the spacecraft for escape at this value of x would appear more like a one-turn skewed spiral escape. An additional observation of the above approximation is that the velocity bias asymptotically approaches the hyperbolic excess velocity as the *vis viva* energy C_3 becomes large.

References

1. *Solar-Powered Electric Propulsion Summary Report,* Report SD-60374R, Hughes Aircraft Co., Culver City, Calif., Dec. 1966.
2. Stearns, J. W., and Kerrisk, D. J., "Solar-Powered Electric Propulsion Systems—Engineering and Applications," Paper 66-576, presented at the AIAA Second Propulsion Joint Specialist Conference, Colorado Springs, Colo., June 1966.
3. Dickerson, W. D., and Smith, D. B., "Trajectory Optimization for Solar-Electric Powered Vehicles," Paper 67-583, presented at the AIAA Guidance, Control and Flight Dynamics Conference, Huntsville, Ala., Aug. 1967.

4. Flandro, G. A., and Barber, T. A., "Mission Analysis for Interplanetary Vehicles With Solar-Electric Propulsion," Paper 67-708, presented at the AIAA Electric Propulsion and Plasmadynamics Conference, Colorado Springs, Colo., Sept. 1967.
5. Sauer, C. G., Jr., "Trajectory Analysis and Optimization of a Low-Thrust Solar-Electric Jupiter Flyby Mission," Paper 67-710, presented at the AIAA Electric Propulsion and Plasmadynamics Conference, Colorado Springs, Colo., Sept. 1967.
6. Melbourne, W. G., and Sauer, C. G., "Optimum Thrust Programs for Power-Limited Propulsion Systems," *Astronaut. Acta*, Vol. VIII, Fasc. 4, 1962.
7. Fimple, W. R., and Edelbaum, T. N., *Study of Low-Acceleration Space Transportation Systems*, Report D-910262-3, United Aircraft Corp., East Hartford, Conn., July 1965.

D. Representation of Point Masses by Spherical Harmonics, *J. Lorell*

1. Introduction

The recent work by W. L. Sjogren and P. M. Muller (Ref. 1) suggests that part of the lunar mass inhomogeneity consists of a few isolated mass concentrations scattered across the lunar surface. Sjogren and Muller have used *Lunar Orbiter* tracking data to identify several of these high-density regions and named them "mascons." By incorporating these mascons in the moon gravity model, it is hoped that the *Lunar Orbiter* tracking data fits can be much improved.

The basic computer programs presently used for *Lunar Orbiter* orbit determination (OD) use a moon gravity model expressed in spherical harmonics. Inclusion of the mascons as point mass potentials would require some modification of these programs and would result in a hybrid gravity model. On the other hand, the mascons can be represented by spherical harmonics expansions and the resulting coefficients used directly in the OD program.

In this article, we will write the equations for spherical harmonic expansion of mascons and apply same to a preliminary moon model obtained from Sjogren and Muller.

2. Spherical Harmonic Expansion of Potential Due to Point Mass

The potential due to a point mass is given by the expression $-\mu/\rho$, in which μ is the gravity constant giving the size of the mass and ρ is the distance to the mass. This potential is easily represented as a spherical harmonic expansion centered at the point mass. What is needed here, however, is the expansion about the center of the moon, not the point mass that lies just below the moon's surface. Let the mascon be located at a point $Q_i\,(r_i, \theta_i, \lambda_i)$ located within the body of the moon whose mass center is at 0 (see Fig. 3). Then the potential at $P\,(r, \theta, \lambda)$ denoted by $\phi_i\,(r, \theta, \lambda)$ is given by

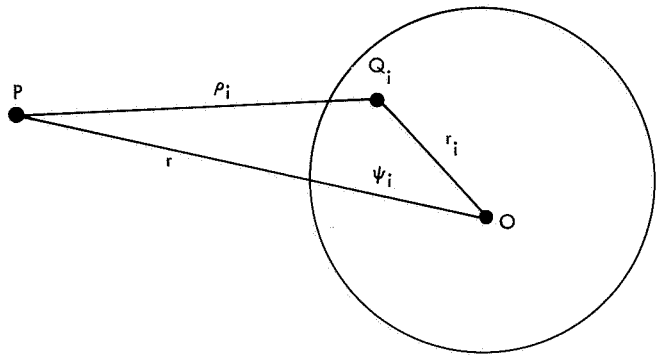


Fig. 3. Potential at P due to mascon at Q_i

$$\phi_i = -\mu_i/\rho_i \tag{1}$$

where μ_i is the gravity constant of the mascon and ρ_i is the distance from Q_i to P. Expanding in spherical harmonics about Q_i gives

$$\phi_i = -\frac{\mu_i}{(r^2 + r_i^2 - 2r\,r_i\cos\psi_i)^{1/2}}$$

$$= -\frac{\mu_i}{r}\sum_{n=0}^{\infty}\left(\frac{r_i}{r}\right)^n P_n(\cos\psi_i) \tag{2}$$

where

$$\cos\psi_i = \frac{\mathbf{r}\cdot\mathbf{r}_i}{rr_i}$$

$$= \sin\theta\sin\theta_i + \cos\theta\cos\theta_i\cos(\lambda - \lambda_i) \tag{3}$$

Then, applying the addition theorem for Legendre polynomials (Ref. 2, p. 328)

$$P_n(\cos\psi_i) = P_n(\sin\theta)\,P_n(\sin\theta_i) + 2\sum_{m=1}^{n}\frac{(n-m)!}{(n+m)!}P_n^m(\sin\theta)\,P_n^m(\sin\theta_i)\cos m(\lambda - \lambda_i) \tag{4}$$

Hence, if we write

$$C^i_{no} = \left(\frac{r_i}{R}\right)^n P_n(\sin\theta_i) \tag{5}$$

$$C^i_{nk} = 2\left(\frac{r_i}{R}\right)^n \frac{(n-k)!}{(n+k)!} P^k_n(\sin\theta_i)\cos k\lambda_i \tag{6}$$

$$S^i_{nk} = 2\left(\frac{r_i}{R}\right)^n \frac{(n-k)!}{(n+k)!} P^k_n(\sin\theta_i)\sin k\lambda_i \tag{7}$$

it follows that the potential ϕ_i may be written in standard form

$$\phi_i = -\frac{\mu_i}{r}\left[1 + \sum_{n=1}^{\infty}\sum_{m=o}^{n}\left(\frac{R}{r}\right)^n P^m_n(\sin\theta)\left(C^i_{nm}\cos m\lambda + S^i_{nm}\sin m\lambda\right)\right] \tag{8}$$

as an expansion about the center of the moon.

3. Example

An estimate of the Sjogren–Muller mascon distribution given in Table 4 has been expanded in spherical harmonics up to degree 15 according to the formulas of *Subsection 2*. A contour map based on the harmonic expansion and showing the corresponding bulges on a uniform-density moon are shown in Figs. 4 and 5. Table 5 identifies the level of each contour line by letter. It is interesting to note that this contour map, based on a fifteenth degree harmonic set, is quite adequate for identification of the mascon locations. A similar map using all the fourth degree harmonics plus zonals through degree eight completely failed to show the mascon locations.

Table 4. Mascon locations and magnitudes

Number	Latitude, deg	Longitude, deg	Magnitude (fraction of lunar mass)
1	32	−17	0.23×10^{-4}
2	25	19	0.18×10^{-4}
3	18	56	0.10×10^{-4}
4	−15	33	0.09×10^{-4}
5	−23	−38	0.06×10^{-4}
6	−20	−95	0.12×10^{-4}
7	6	−6	0.07×10^{-4}
8	−6	−7	-0.10×10^{-4}
9	45	−35	-0.06×10^{-4}
10	11	15	-0.08×10^{-4}
11	−2	19	-0.08×10^{-4}

Table 5. Elevation contour key

Arc	Contour value, m	Arc	Contour value, m
A	−3000	N	250
B	−2750	O	500
C	−2500	P	750
D	−2250	Q	1000
E	−2000	R	1250
F	−1750	S	1500
G	−1500	T	1750
H	−1250	U	2000
I	−1000	V	2250
J	−750	W	2500
K	−500	X	2750
L	−250	Y	3000
M	0		

References

1. Muller, P. M., and Sjogren, W. L., "Mascons: Lunar Mass Concentrations," *Science*, Vol. 161, No. 3842, pp. 680–684, Aug. 16, 1968.
2. Whittaker, E. T., and Watson, G. N., *A Course of Modern Analysis*, Fourth Edition. Cambridge University Press, 1927.

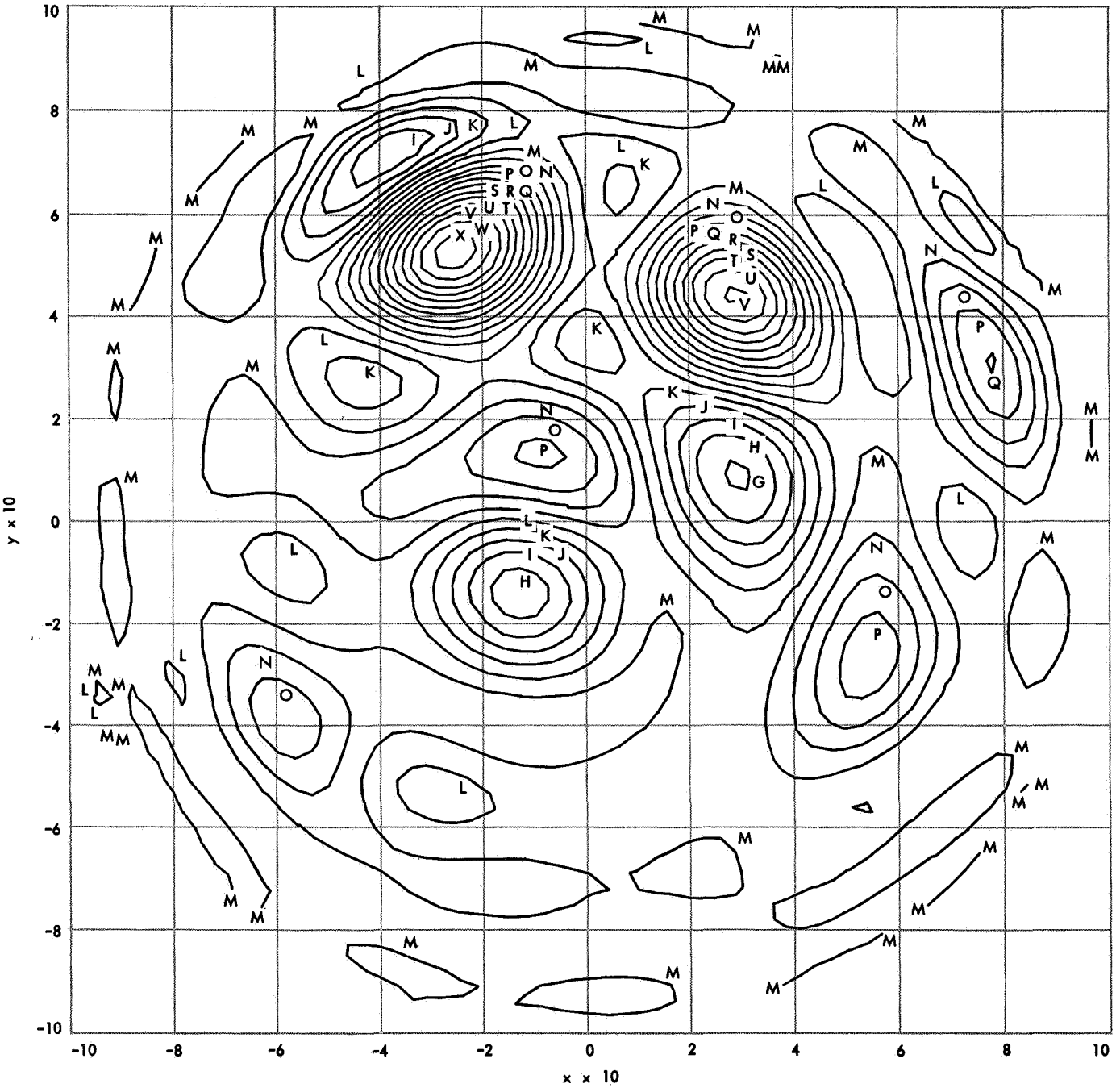


Fig. 4. Mascon distribution on moon—front

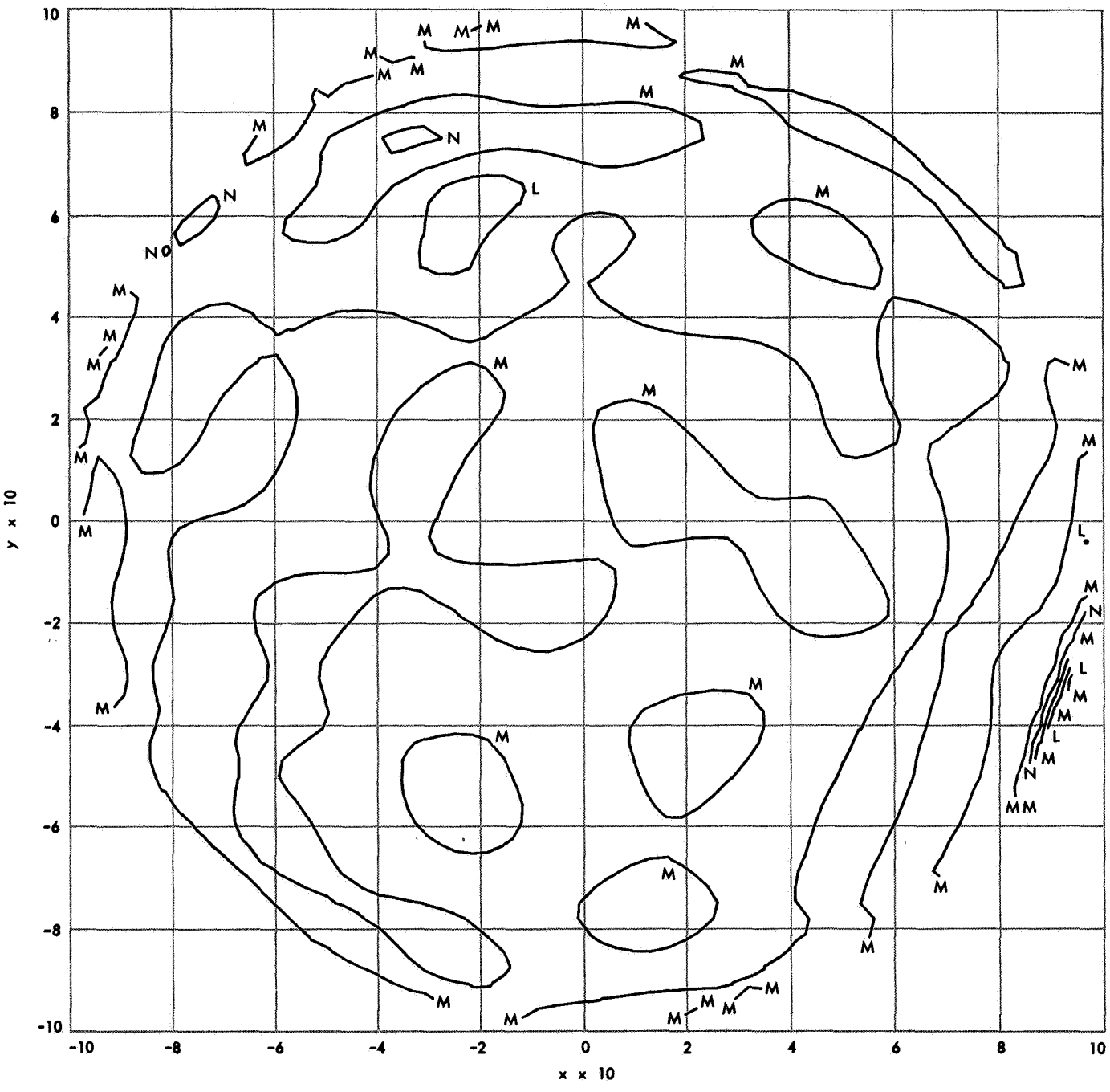


Fig. 5. Mascon distribution on moon—back

II. Computation and Analysis

SYSTEMS DIVISION

A. The Direct Summation of Series Involving Higher Transcendental Functions,[1] *E. W. Ng*

In many problems of physics there is often the need to evaluate or compute a series of the form

$$S_N(x) = \sum_{n=j}^{N} a_n(x) f_n(x) \tag{1}$$

where $f_n(x)$ is a higher transcendental function and $a_n(x)$ are given coefficients. Well-known examples of such are truncated series involving Chebychev polynomials, Bessel functions, and Legendre functions. In most applications we have $j = 0$ or 1. Many higher transcendental functions satisfy a three-term recurrence relation of the form

$$f_{n+1}(x) = B_n(x) f_n(x) + C_n(x) f_{n-1}(x), \qquad n = 0, 1, \cdots \tag{2}$$

For the orthogonal polynomials, we define $f_{-1}(x) = 0$.

It is well known that recurrence relations form a basic mathematical tool for the computation of many functions. We have, for example, Miller's algorithm for computing Bessel functions. For a recent detailed survey and analysis of such algorithms, the reader is referred to Ref. 1. Whereas these relations are simple to use, one must attend to the problem of numerical stability. For example, Gautschi shows that given the Bessel functions $J_0(1)$ and $J_1(1)$ accurate to 10 significant figures and generating the next values of $J_n(1)$ by forward recursion, one loses all significance for $n \geqq 7$. Abramowitz (Ref. 2) summarizes the caution one must take in using such recurrence relations. In particular, the direction of recurrence is important. For example, the Bessel functions J_n and I_n are stable only in backward recurrence, whereas Y_n and K_n are stable only in forward recurrences.

Clenshaw (Ref. 3) recommends an algorithm to sum a Chebychev series directly. Here we shall generalize the algorithm to other special functions satisfying Eq. (2).

Consider the recurrence formula (with the functional dependence on x understood)

$$b_k = b_{k+1} B_k + b_{k+2} C_{k+1} + a_k, \qquad k = N, N-1, \cdots, j$$
$$b_{N+1} = b_{N+2} = 0 \tag{3}$$

Multiply Eq. (3) by f_k and form a "system of equations" as follows:

$$\left.\begin{aligned}
b_N f_N &= && + a_N f_N \\
b_{N-1} f_{N-1} &= b_N f_{N-1} B_{N-1} && + a_{N-1} f_{N-1} \\
b_{N-2} f_{N-2} &= b_{N-1} f_{N-2} B_{N-2} + b_N f_{N-2} C_{N-1} && + a_{N-2} f_{N-2} \\
&\;\;\vdots \\
b_j f_j &= b_{j+1} f_j B_j + b_{j+2} f_j C_{j+1} + a_j f_j
\end{aligned}\right\} \tag{4}$$

[1] A more detailed version of this article will appear in a future issue of *J. Comp. Phys.*, Vol. III.

Adding all equations of Expression (4) and using Eq. (2), we obtain

$$S_N = \sum_{n=j}^{N} a_n f_n = b_j f_j + b_{j+1}(f_{j+1} - B_j f_j) \tag{5}$$

Notice that Eq. (3) is a backward recurrence, but not as the nonhomogeneous counterpart of Eq. (2), because the role of C_k is displaced. Obviously one can also derive a recurrence scheme expressing S_N in terms of f_N and f_{N-1}. Notice that for $j = 0$, $S_N = b_0$ for the orthogonal polynomials. Thus Eq. (3) represents a formalism for computing the series S_N. It is mainly useful for the case $j = 0$ or 1, because here f_0 and f_1 are readily obtainable. But the applicability will, of course, depend on the stability of Eq. (3), which in turn depends on the function in question. In the following, we shall describe some numerical experiments with this algorithm by applying it to the following simple series (Ref. 4):

$$J_0(\pi) + 2\sum_{n=1}^{\infty}(-1)^n J_{2n}(\pi)\, T_{2n}(x) = \cos \pi x, \qquad -1 \leq x \leq 1 \tag{6}$$

$$I_0(1) + 2\sum_{n=1}^{\infty}(-1)^n I_n(1)\, T_n(x) = e^x, \qquad -1 \leq x \leq 1 \tag{7}$$

$$\sum_{n=0}^{\infty}(2n+1)\, z^n P_n(x) = \frac{1 - z^2}{(1 - 2xz + z^2)^{3/2}}, \qquad -1 \leq x \leq 1, |z| \leq 0.6 \tag{8}$$

$$\sum_{n=0}^{\infty}\frac{1}{nz^n} P_n(x) = \log\left[\frac{2z}{z - x + (1 - 2zx + z^2)^{1/2}}\right], \qquad -1 \leq x \leq 1, \quad 2 \leq z \leq 10 \tag{9}$$

$$\sum_{h=0}^{\infty}\frac{z^n}{\Gamma(n+1)} P_n(x) = e^{xz} J_0[z(1 - x^2)^{1/2}], \qquad -1 \leq x \leq 1, |z| \leq 4 \tag{10}$$

$$\sum_{n=0}^{N} A_n P_n(x), \qquad -1 \leq x \leq 1, \quad N = 20, 30, 40, \quad 0 \leq A_n \leq 100 \tag{11}$$

$$\sum_{n=0}^{N} A_n L_n(x), \qquad 0 \leq x \leq 100, \quad N = 20, 30, 40, \quad 0 \leq A_n \leq 100 \tag{12}$$

All computations were performed on an IBM 7094 computer using double precision (16-decimal digit) arithmetic. For Eqs. (6) to (10), we terminate the series when the coefficient is less than 10^{-17}. For each series we generate 1000 uniformly distributed pseudorandom numbers for the variables x and z in the indicated range, which does not necessarily cover the whole range of theoretical convergence. The choice of range is obviously for practicality. For example, for Eq. (8), at $|z| = 0.6$ one needs about 100 terms to satisfy our criterion. In Eqs. (11) and (12) the A_n's are a set of pseudorandom numbers uniformly distributed in the indicated range. In all of the above series, we also compute the sum by generating the special functions by forward recurrence and then summing. Thus we have three different results for Eqs. (6) to (10) and two for Eqs. (11) and (12). In all cases we compute the relative differences among the two or three different methods. These differences, of course, depend on the values of x, z, A_n, S_N, and N. They range from 1×10^{-6} to 1×10^{-14}, but are in no case greater than the last number.

References

1. Gautschi, W., *SIAM Rev.*, Vol. 9, p. 24, 1967.
2. Abramowitz, M., "Handbook of Mathematical Functions," *NBS Appl. Math.*, Ser. 55, p. XIII, 1965.
3. Clenshaw, C. W., *National Physical Laboratory Mathematical Tables*, Vol. 5. Her Majesty's Stationery Office, London, 1962.
4. Margulis, V., *Handbook of Series*, Academic Press, New York, 1965.

B. Integrals of Confluent Hypergeometric Functions, Part II,[2] *E. W. Ng*

This article is a direct continuation of Part I in SPS 37-46, Vol. IV, p. 34. A closely related set of integrals appear in Ref. 1.

[2]A condensed version of Parts I and II will appear in a future issue of *J. Res. NBS, Sec. B.*

1. Reduction Formulas for $\Theta_n(a, b, \alpha, z)$ and $\Upsilon_n(a, b, \alpha, z)$

As before, four different cases of the parameters a and b will be considered.

a. Case 1

$$a \neq \text{integer}, \qquad b \neq \text{integer}$$

In this case two formulas equivalent to Eqs. (34) and (35) of Part I can be derived. However, the results will not be too useful, because the right-hand side will have a term with subscript n, due to the derivative of $(z^n e^{\alpha z})$. Instead, the equivalent of Eqs. (38) and (39) of Part I is written as

$$\Theta_n(a, b, \alpha, z) = (b-1)\,\Theta_{n-1}(a, b-1, \alpha, z) - (b-1)\,\Theta_{n-1}(a-1, b-1, \alpha, z) \tag{1}$$

$$\Upsilon_n(a, b, \alpha, z) = (b-a-1)\,\Upsilon_{n-1}(a, b-1, \alpha, z) + \Upsilon_{n-1}(a-1, b-1, \alpha, z) \tag{2}$$

b. Case 2

$$a \neq \text{integer}, \qquad b = \text{integer}$$

With the help of Eqs. (12) and (14) of Part I, $\Theta_{n-1}(a, b-1, \alpha, z)$ can be expressed in terms of $\Theta_{n-1}(a+1, b, \alpha, z)$, thereby obtaining

$$\Theta_n(a, b, \alpha, z) = 2\Theta_{n-1}(a+1, b, \alpha, z) - (2a-b)\,\Theta_{n-1}(a, b, \alpha, z) + (a-b)\,\Theta_{n-1}(a-1, b, \alpha, z) \tag{3}$$

$$\Upsilon_n(a, b, \alpha, z) = a(a+1-b)\,\Upsilon_{n-1}(a+1, b, \alpha, z) + (b-2a)\,\Upsilon_{n-1}(a, b, \alpha, z) + \Upsilon_{n-1}(a-1, b, \alpha, z) \tag{4}$$

c. Case 3

$$a = \text{integer}, \qquad b \neq \text{integer}$$

In this case, Eqs. (1) and (2) can be used again to reduce Θ_n and Υ_n to Θ_0 and Υ_0, and $\Theta_n(0, \beta, \alpha, z)$ and $\Upsilon_n(0, \beta, \alpha, z)$, where the last two are just elementary integrals. Kummer's first theorem (Ref. 2, p. 6) can also be used to transform Θ_n, in this case to that of case 1, as follows:

$$\Theta_n(a, b, \alpha, z) = \int z^n e^{(\alpha+1)z} M(b-a, b, -z)\,dz = (-1)^{n+1}\,\Theta_n(b-a, b, \alpha+1, -z) \tag{5}$$

d. Case 4

$$a = \text{integer}, \qquad b = \text{integer}$$

For $a > b$, a reduction formula equivalent to Eq. (42) can be used:

$$\Theta_n(a, b, \alpha, z) = \Theta_n(a-1, b, \alpha, z) + 1/b\,[\Theta_{n+1}(a, b+1, \alpha, z)] \tag{6}$$

Again, successive application will reduce the right-hand side of the equation to elementary integrals. For $b > a > 0$, Eqs. (1) and (2) can be conveniently applied. for $a < 0$, Eqs. (30) and (31) of Part I can again be used, but with a replaced by $(a+1)$ and "recur upward" in n. However, for all integer values of a and b, Eqs. (3) and (4) are applicable.

Therefore, it can be seen that Θ_n and Υ_n can be reduced to a finite combination of Λ, M, Ω, U, or elementary integrals. Properties of Λ and Ω will be discussed in subsequent investigations.

References

1. Ng, E. W., *J. Math. Phys.*, Vol. 46, p. 223, 1967.
2. Slater, L. J., *Confluent Hypergeometric Functions*, Cambridge University Press, London, 1960.

C. Survey of Computer Methods for Fitting Curves to Discrete Data or Approximating Continuous Functions, C. L. Lawson

1. Introduction

In preparation for this survey, a classified bibliography of recent publications[3] was compiled that includes 394 references. As is clear from this bibliography, approximation theory has wide application in the mathematics of computation; e.g., approximation of functions or data; quadrature; solution of ordinary differential equations, partial differential equations, and integral equations; and graphical displays. On the other hand, approximation algorithms often depend upon more general computational techniques, such as the solution of linear or nonlinear systems of equations and/or inequalities and general minimization methods. A selection of references on these latter topics is included in the bibliography.

[3]Lawson, C. L., "Bibliography of Recent Publications in Approximation Theory With Emphasis on Computer Applications," *Comput. Rev.*, Vol. 9 (to be published).

This survey treats primarily the problems of fitting curves to discrete data and approximating continuous functions. The point of view is that of practical scientific computation.

The choice of a mathematical model in an approximation problem can often be conveniently described as the choice of form and norm, i.e., the choice of approximating form such as polynomial, rational, or spline, and the choice of norm such as ℓ_2, ℓ_∞, or ℓ_1. There are, of course, other considerations such as constraints and transformation of variables.

Often the problem objectives are such that there is some freedom in the choice of form or norm. Then the choice should be made on the basis of properties such as numerical stability and economy of computation. These properties are discussed in *Subsections* 2 and 3. Most of *Subsections* 2 and 3 generalizes to fitting functions of two or more real variables or complex variables; however, from the practical point of view, such fits are often limited to applications requiring only moderate accuracy (e.g., 10^{-4}) because of the very large number of parameters needed for higher accuracy.

2. Choice of Form

a. Polynomial forms. Polynomial forms are, in a sense, the simplest, and a variety of parameterizations is possible. If a polynomial is expressed as $\Sigma a_i x^i$, which will be called the monomial basis parameterization, it can be evaluated in n multiplications and n additions. The matrix of basis function values is typically very poorly conditioned. This conditioning is generally significantly improved by translating the domain of the independent variable to be centered at zero. Exponent overflow is avoided by scaling to, e.g., $[-1, 1]$. Even with these precautions, polynomials of degree higher than about 7 in monomial basis form are essentially useless in 8-decimal digit arithmetic.

Other bases for parameterization, such as Chebyshev polynomials, typically provide remarkable stability. For example, polynomials of degrees 533 through 223 have been computed to represent the positions of the five outer planets, Jupiter through Pluto, over a period of 200 yr. This work used 16-decimal digits and preserved at least 5-digit accuracy. A polynomial of degree n represented as a linear combination of Chebyshev polynomials can be evaluated in n multiplications and $2n$ additions. In general, the Chebyshev basis is preferable to the monomial basis, independent of other factors such as the method for determining coefficients or the choice of norm.

Other polynomial parameterizations include the Forsythe parameterization for polynomials determined to be orthogonal over a specific point set, the product-of-roots form, and streamlined forms. The product-of-roots form is very stable if the roots are in the x-interval of interest, but the determination of parameters may be inconvenient. The streamlined forms reduce the number of multiplications needed in evaluation but are often very unstable. The Forsythe parameterization is redundant, requiring about $3n$ parameters to specify an nth-degree polynomial. It exhibits very good numerical stability, and the algorithm for determining the parameters is very efficient, since the execution time depends upon mn, rather than mn^2, where m is the number of data points.

b. Rational forms. Various special properties (such as remaining bounded at infinity, having poles, and having abrupt changes of curvature) make rational forms more useful than polynomial forms in some cases. Since the parameters occur nonlinearly, their determination requires iterative procedures which entail various practical difficulties: (1) the absence of zeros from the denominator must always be verified. (2) Best rational approximations on discrete sets do not always exist. The use of rational functions for fitting discrete data can probably be largely supplanted by the use of spline polynomials.

Rational functions have been very successfully used as approximating forms for many analytic functions such as the exponential and arctangent. The effective design of such approximations depends more upon a thorough understanding of the function being approximated (leading to the use of special identities and changes of variables) than upon the actual method of computation of the approximation.

All polynomial parameterizations can be used for rational function parameterization. There is also the possibility of using continued fraction forms; however, these are frequently unstable and must be tested for growth of rounding error in each case.

c. Spline forms. A spline function s, defined on an interval $[a, b]$ partitioned into k segments, is a polynomial of degree n on each segment with continuous derivatives through order m $(m < n)$ throughout $[a, b]$. A spline will generally have discontinuities in its $(m + 1)$st derivative at the partition points. The splines which have received

the most study are those for which $m = n - 1$ and, more particularly, cubic splines with second-order continuity. Such splines, having k segments, can be parameterized by $k - 1 + 4k$ parameters, giving the abscissas of the $k - 1$ partition points and 4 coefficients for each of the k cubic polynomial segments. The discussion given previously for the parameterization of polynomials is then applicable to each segment.

This parameterization, though convenient for evaluation, is redundant. Other parameterizations having less redundancy have been given in the literature. Some, such as

$$\sum_{i=0}^{3} c_i x^i + \sum_{i=1}^{k-1} a_i \{\max [0, (x - b_i)]\}^3$$

are of theoretical use only and are definitely not recommended for computational use.

If the partition points are fixed, $4k$ parameters remain, and these occur linearly. The second-order continuity requirement constitutes $3(k - 1)$ linear equality constraints, reducing the number of degrees of freedom to $k + 3$. One basis consisting of $k + 3$ linearly independent splines, which has been recommended as having favorable properties in practical use, can be defined as follows:

By introducing 3 auxiliary segments to the left and 3 to the right of the interval $[a, b]$, $k + 6$ segments are defined. For each set of four contiguous segments, a spline function is constructed that is nonzero on that set and is zero elsewhere. This defines $k + 3$ basis functions (each uniquely determined to within an arbitrary scalar multiple). The associated matrix in curve fitting has a block diagonal structure that can be used to conserve computer time and storage.

Although spline forms have received intensive study in recent years, the best strategies for parameterizing and manipulating splines and treating the problem with variable breakpoints have yet to be evolved. With their extreme flexibility in changing curvature, stability of low-degree polynomials, and linearity of coefficients (for fixed partition points), spline forms provide a very attractive approach to general data fitting.

The second derivative of a cubic spline with second-order continuity is a linear spline with zero-order continuity. Thus, the sign of the second derivative can be constrained throughout $[a, b]$ by constraining it only at the partition points. This fact has been used to obtain some very satisfactory data fits where oscillations were to be avoided.

3. Choice of Norm

For fitting data subject to random errors, it can be argued that the ℓ_2 norm is most appropriate if the error is normally distributed; ℓ_1 is most appropriate if the error distribution has broad tails; and ℓ_∞ is most appropriate if the error distribution has narrow or no tails, e.g., a uniform distribution over a finite interval.

In practice, the ℓ_1 approximation is probably very rarely used, since the broad tail problem is usually treated by some ad hoc wild-point exclusion logic. Discrete ℓ_1 approximations can be nonunique even with the Haar condition, and characterization of a best ℓ_1 approximation is complex. The discrete linear ℓ_1 problem is a linear programming problem; however, a linear programming code should have a full capability to treat degenerate cases if it is to be trusted for ℓ_1 fitting.

Discrete ℓ_2 (least-squares) approximation is, of course, widely used. With linear parameters it is a linear problem, i.e., no iteration is needed. Orthonormal methods such as Householder transformations or modified Gram-Schmidt orthogonalization (numerically superior to Gram-Schmidt orthogonalization) can be used to avoid the squaring of the condition number associated with the formation of normal equations. The number of multiplications and additions is approximately doubled with the orthonormal methods, and thus these methods must be compared with the use of normal equations in double-length arithmetic to determine which is more efficient and reliable in a given application.

Discrete linear ℓ_∞ approximations can be treated as the linear programming problem it is or by specially adapted equivalent algorithms such as the exchange algorithm. Two other distinct methods for the discrete ℓ_∞ problem, although probably not competitive with the exchange algorithm for the linear Haar ℓ_∞ problem, appear to generalize to the nonlinear or non-Haar cases in a more natural way. These are: (1) the Polya algorithm, which relies on the ℓ_∞ solution being the limit of ℓ_p solutions as $p \to \infty$; and (2) the Lawson algorithm, which adjusts weights in a weighted ℓ_2 approximation so that the ℓ_∞ approximation is approached via a sequence of weighted ℓ_2 approximations.

For the approximation of continuous functions by curve fitting, primary interest has been with the ℓ_∞ approximation. Exchange-type algorithms have been used very

effectively and efficiently for both polynomial and rational approximations.

Such approximations are commonly produced for use in function subprograms. The construction of an efficient function subprogram also depends strongly upon the use of special properties of the function being approximated and machine-dependent considerations. For some functions, particularly functions of more than one variable, efficient constructive representations have been derived entirely from mathematical analysis of the functions without the use of fitting.

III. Environmental Requirements

PROJECT ENGINEERING DIVISION

A. Engineering Models of the Venus Atmosphere, *R. A. Schiffer*

Additional scientific measurements and theoretical studies are required before a clear understanding of the structure of the Venus atmosphere can be evolved. In the meantime, Venus atmosphere engineering models reflecting the best current knowledge are still needed for space vehicle design and mission planning. Accordingly, an interim set of standard models[1] based on the latest scientific data has been prepared; however, they should not be considered as new scientific models of the Venus atmosphere.

Although the combination of parameters into "worst-case" models is a recognized function of the specific mission design, it is not certain that all extremes have been met by the models presented. However, these models may be regarded as a state-of-the-art approximation that envelop current uncertainties of the Venus atmospheric parameters with an estimated confidence of at least 95%.

[1]Schiffer, R. A., "Engineering Models of the Venus Atmosphere Based on an Interpretation of Recent Space Vehicle Observations of Venus," paper to be presented at the AIAA 7th Aerospace Sciences Meeting, Jan. 1969.

In preparing these models, particular attention was given to assessing the atmospheric environmental interactions that influence the integrity and performance of a planetary vehicle and its major subsystems. These atmospheric interactions are both aerodynamic and thermal and are directly related to the structure, composition, and dynamics of the atmosphere. Table 1 summarizes these interactions with the principal space vehicle subsystems, and identifies the atmospheric parameters involved in each case. The vertical distribution of mass density is regarded as the most critical parameter for design functions that involve aerodynamic interactions. However, adequate definition of other quantities as chemical composition and temperature structure is also important because they are implicit in the calculation of density and appear as parameters in thermal calculations. In addition, the viscosity, specific heat, and speed of sound influence the vehicle aerothermodynamic analyses, while atmospheric winds primarily affect terminal descent entry dynamics. Finally, the atmospheric aerosol content and opacity constrain the design of landed solar power systems and influence the performance of communications equipment.

Six atmospheric engineering models of the Venus atmosphere are proposed for space vehicle design based

Table 1. Orbiter, entry, and lander vehicle atmospheric environmental interactions

Space vehicle subsystem	Pressure profile	Temperature profile	Density profile	Specific heat	Viscosity	Speed of sound	Gas composition	Winds	Opacity and aerosols
Structural			√	√	√	√		√	
Retardation			√	√	√	√	√	√	
Propulsion	√	√	√	√	√	√	√	√	√
Heat shield			√	√	√	√	√		
Guidance	√	√	√			√		√	
Attitude control	√	√	√			√		√	
Communications	√		√				√		√
Power supply	√	√	√				√	√	√
Electronics	√	√	√				√		
Mechanical devices	√	√	√					√	√
Thermal control	√	√	√	√	√		√	√	√
Systems analyses	√	√	√	√	√	√	√	√	√

on the theoretical Venus thermal model of McElroy (Ref. 1) and data from the recent *Mariner V* and *Venera 4* space probes (Refs. 2 and 3). The models, which are based on the constraints summarized in Table 2, describe profiles for temperature, density, pressure, speed of sound, molecular mass, density scale height, number density, mean free path, and viscosity. These parameters were calculated by numerical integration of the hydrostatic equation with the aid of thermodynamic relationships (Footnote 1). Figures 1 and 2 illustrate the profiles of temperature and density for each model. Models MV-1 and -2 correspond to minimum solar activity, MV-3 and -4 to moderate solar activity, and MV-5 and -6 to maximum solar activity. Models MV-1, -3, and -5 are high-density models characterized by high pressure and low molecular weight. Models MV-2, -4, and -6 are low-density models characterized by low pressure and high molecular weight. Uncertainties in the knowledge of the dynamics of the Venus atmosphere preclude the specification of a realistic wind model. In addition, no acceptable model describing the aerosol content and opacity is currently available.

The models can be described in the form of probability density functions. However, limitations in the number of data points and in the identification of the experimental errors of currently available scientific data do not permit statistical treatment of the uncertainty ranges for such key parameters as surface pressure, temperature, and composition.

A superposition of the *Mariner V* and *Venera 4* temperature and pressure data interpretations (Figs. 3 and 4) results in profiles that agree remarkably well, provided the planetary surface at the *Venera 4* final data transmission point is located at a radius of approximately 6078 km.

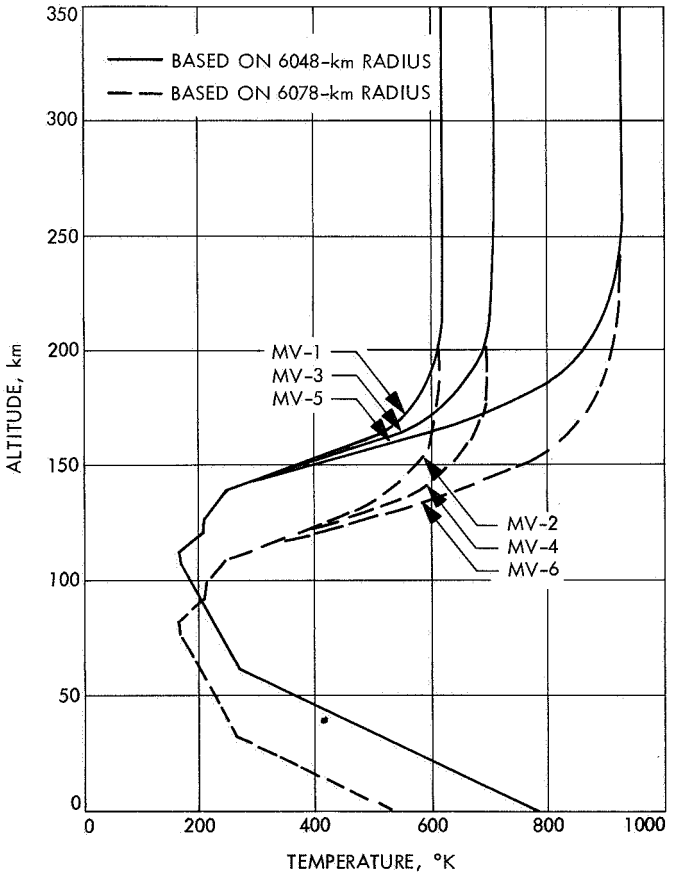


Fig. 1. Temperature vs altitude in Venus atmosphere

Table 2. Parameters for Venus atmosphere models MV-1 through MV-6

Parameter	Minimum solar activity		Mean solar activity		Maximum solar activity	
	MV-1 (high density)	MV-2 (low density)	MV-3 (high density)	MV-4 (low density)	MV-5 (high density)	MV-6 (low density)
Surface pressure, atm	167	16.4	167	16.4	167	16.4
Composition, mole fraction:						
CO_2	0.81	0.9998	0.81	0.9998	0.81	0.9998
N_2	0.0998	0	0.0998	0	0.0998	0
CO	0.045	0	0.045	0	0.045	0
O	0.045	0	0.045	0	0.045	0
He	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
H_2	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Molecular mass, gm/mole	40.42	44	40.42	44	40.42	44
Surface temperature, °K	770	534	770	534	770	534
Exosphere temperature, °K	625	625	710	710	931	931
Planetary radius, km	6048	6078	6048	6078	6048	6078
Surface gravity, cm/s^2	888.1	879.4	888.1	879.4	888.1	879.4
Density at turbopause, g/cm^3	3.6×10^{-11}	3.6×10^{-11}	3.6×10^{-11}	3.6×10^{-11}	3.6×10^{-11}	3.6×10^{-11}

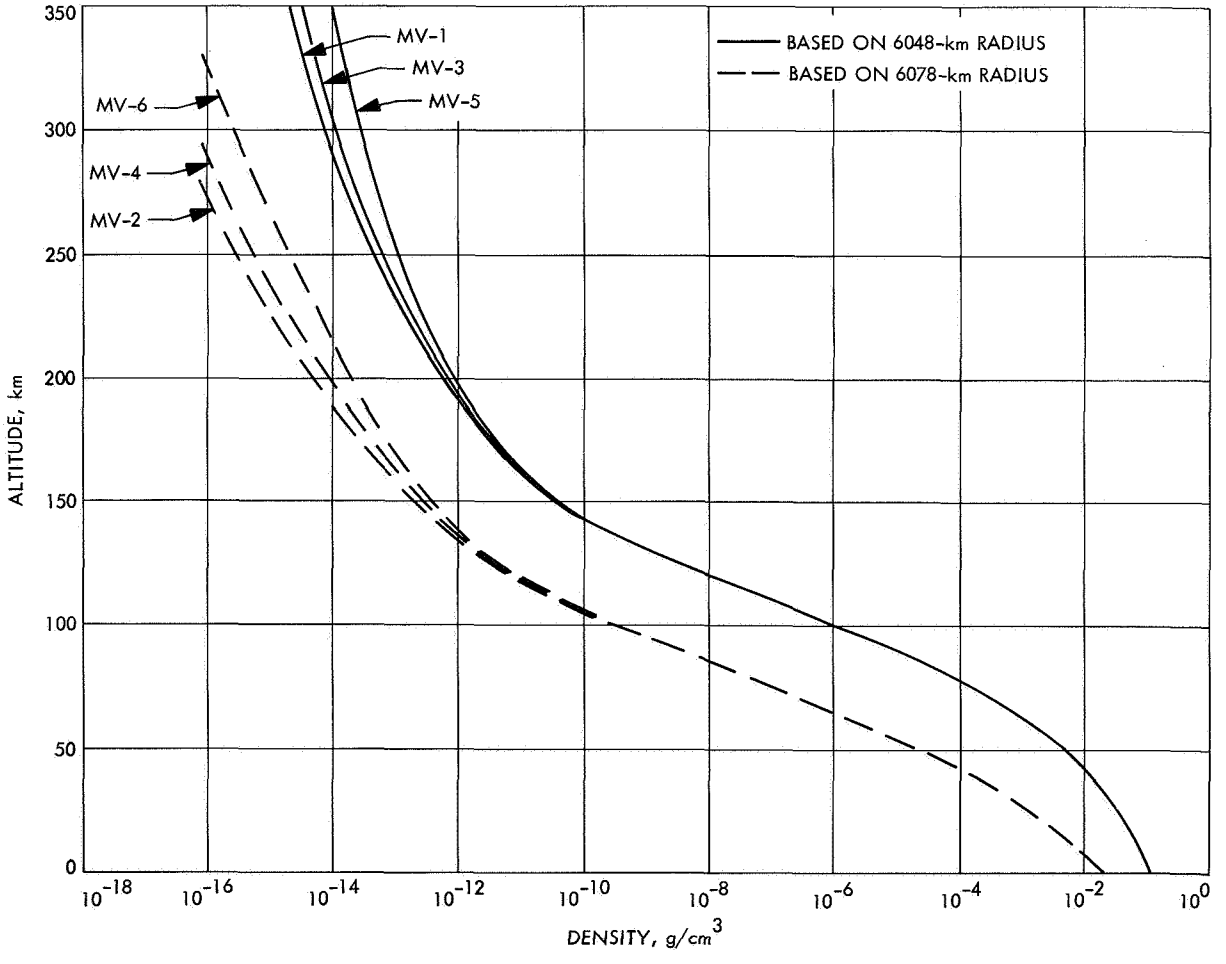


Fig. 2. Density vs altitude in Venus atmosphere

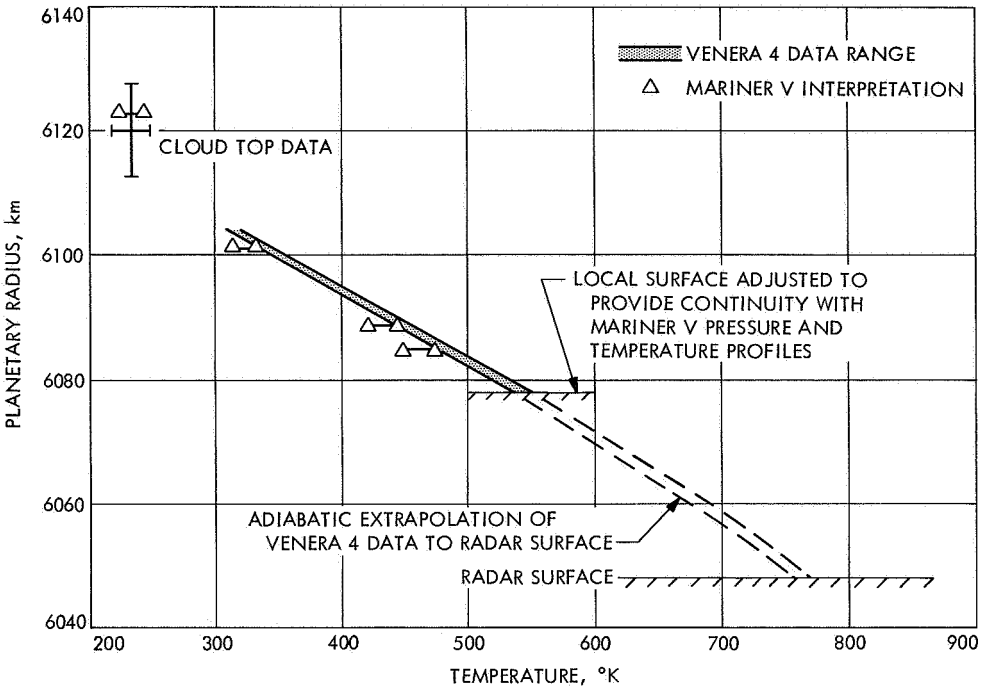


Fig. 3. Venus atmosphere temperature data

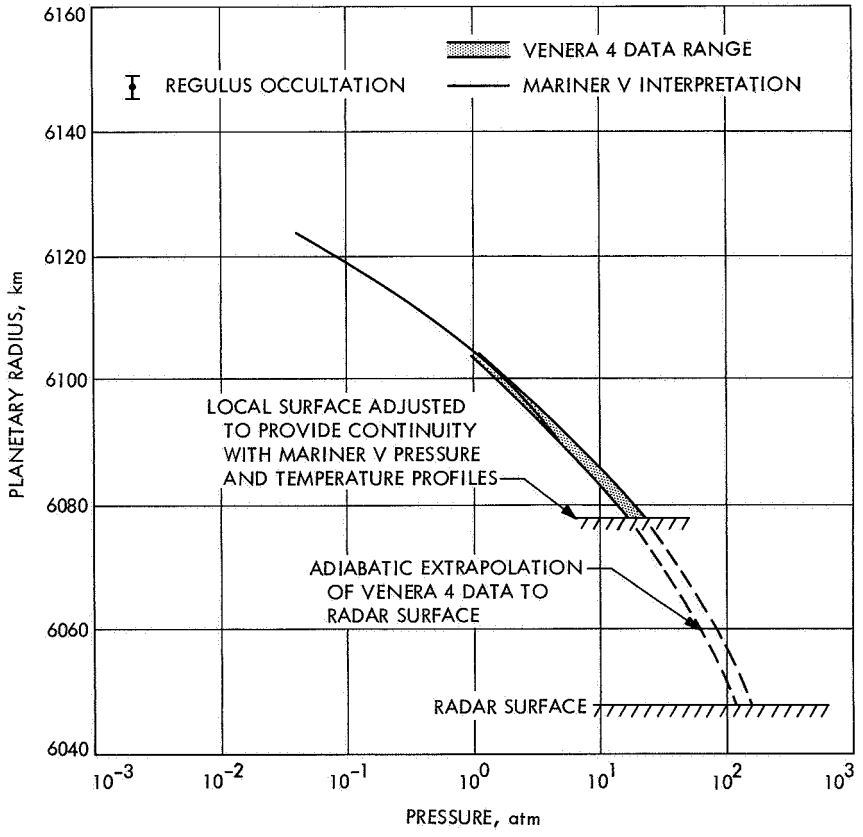


Fig. 4. Venus atmosphere pressure data

However, recent Massachusetts Institute of Technology radar studies of Venus (Ref. 4), suggest a value of 6050 ±0.5 km for the radius of Venus. In addition, *Mariner V* ranging data combined with simultaneous radar data[2] give a radius of 6056.6 ±2.1 km. The radar radius of 6048 km measured at Arecibo (Ref. 4) was selected as the gravitational potential surface defining the upper bound surface pressure for the models. It is doubtful that the entire discrepancy in radius can be attributed to topography. Asymmetry in the shape of the planet is estimated to be on the order of only 1 km.

Thus, the resolution of an appropriate uncertainty range for the mean Venus surface pressure is a direct consequence of the uncertainty in the location of the planetary mean surface radius. The surface pressure, as interpreted from the *Venera 4* data (16.4 to 20.3 atm), could be as high as 167 atm if the radar radius is correct and the *Venera 4* probe did not, in fact, impact at the instant of final data transmission. Although arguments have been made that the Soviet probe did indeed transmit data up to the time of surface impact (Ref. 5), the case in favor of the radar surface seems the most plausible. Thus, the data are bimodal in nature, the planetary surface and atmospheric parameters being related to the radius defined by superimposing the *Mariner V* and *Venera 4* data in one case, and to the radar radius in the other. The likelihood of the planetary radius being somewhere in between appears remote. Consequently, the specification of a mean model would appear unjustified.

References

1. McElroy, M. B., "The Upper Atmosphere of Venus," *J. Geophys. Res.*, Vol. 73, No. 5, Mar. 1, 1968.
2. Kliore, A., et al., "Atmosphere and Ionosphere of Venus from the *Mariner V* S-Band Radio Occultation Measurement," *Science*, Vol. 160, Dec. 29, 1967.
3. Avduevskiy, V. S., Marov, M. Ya., and Rozhdestvenskiy, M. K., "The Model of the Atmosphere of the Planet Venus on the Results of Measurements Made by the Soviet Automatic Interplanetary Station Venera 4," *J. Atmos. Sci.*, Vol. 35, No. 4, July, 1968.
4. Ash, M., et al., "The Case for the Radar Radius of Venus," *Science*, Vol. 160, May 31, 1968.
5. Reese, D., and Swan, P., "Venera 4 Probes the Atmosphere of Venus," *Science*, Vol. 159, Mar. 15, 1968.

[2]Anderson, J. D., et al., "The Radius of Venus As Determined By Planetary Radar and *Mariner V* Radio Tracking Data," paper presented at the American Astronomical Society meeting, Victoria, British Columbia, Aug. 20–23, 1968.

B. Application of Thermal Modeling to Space Vehicle Sterilization, *A. R. Hoffman and J. T. Wang*

1. Introduction

Planetary quarantine constraints may necessitate the application of a dry-heat thermal sterilization process to a planetary capsule prior to launch. To minimize the severity of the sterilization cycle and also to assure the desired level of sterility, it is necessary to account for the reductions in microbial population that occur during the transient phases of heating and cooling, as well as the reductions that occur during the steady-state phase. Geometric and analytic capsule models have been developed and applied to (1) provide insight into the relationships existing between the characteristics of the microbial populations and the thermal characteristics of the space vehicle and heating medium, and (2) perform sensitivity studies *prior* to subjecting the hardware to a sterilization environment.

2. Geometric Analytic Model

Numeric analytic techniques used to establish sterilization processes in the food and pharmaceutical industries were adapted to provide a first approximation of the calculation necessary for the development of capsule dry-heat sterilization process parameters. A simplified geometric conceptual model of a space vehicle was constructed. The space vehicle was assumed to be a series of cylindrical shells (see SPS 37-47, Vol. III, Fig. 1, p. 32) made of homogeneous material with insulated ends. Each shell was mated to the other in such a manner that the heat flow through the model was as through an infinite cylinder. The dimensions of the model were arbitrary but were chosen to approximate the dimensions of a large planetary landing vehicle. The model is not representative of any space vehicle configuration but was developed to facilitate the transition of the numeric analytic techniques from food containers to space vehicles.

Using the geometric model, some important conclusions were drawn (SPS 37-47, Vol. III, pp. 31–35, and Refs. 1 and 2):

(1) Verification that consideration of the microbial reduction that occurs during the transient phases of the sterilization cycle can result in a significant reduction in total process time.

(2) Indication that the distribution of microbial load upon the space vehicle may significantly affect the calculations of the required process parameters and therefore is necessary information for proper process calculation.

(3) Indication that, as the effective thermal conductivity increases, the required sterilization process time will decrease.

(4) Demonstration that the process calculations are sensitive to changes in certain microbial heat resistance parameters (D values[3]) and relatively insensitive to other heat resistance parameters (z values[3]).

3. Capsule Analytic Model

To apply the numeric analytic concept to hardware, the feasibility capsule of a possible Mars entry and landing vehicle (Fig. 5) was analytically modeled as illustrated in Fig. 6. The capsule analytic model was divided

[3]Term D is the decimal reduction time, or time at temperature required to destroy 90% of the microorganisms. Term z is numerically equal to the number of degrees Fahrenheit (or centrigrade) required for a thermal destruction curve to traverse one logarithm cycle.

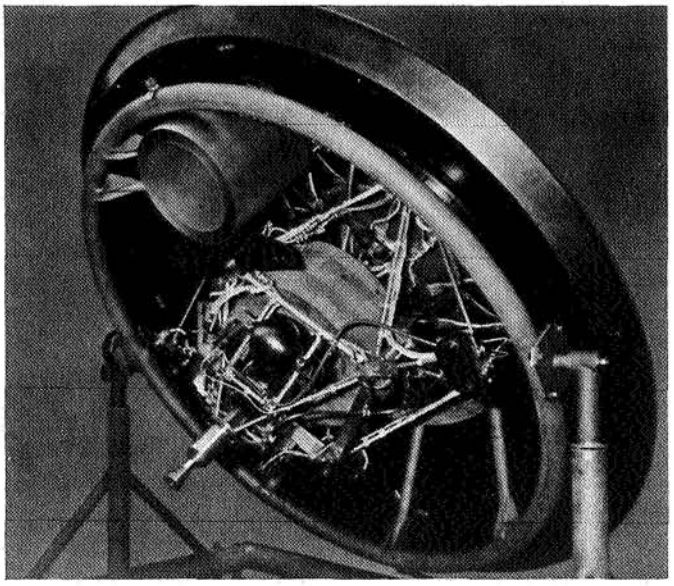


Fig. 5. Feasibility model—separation configuration

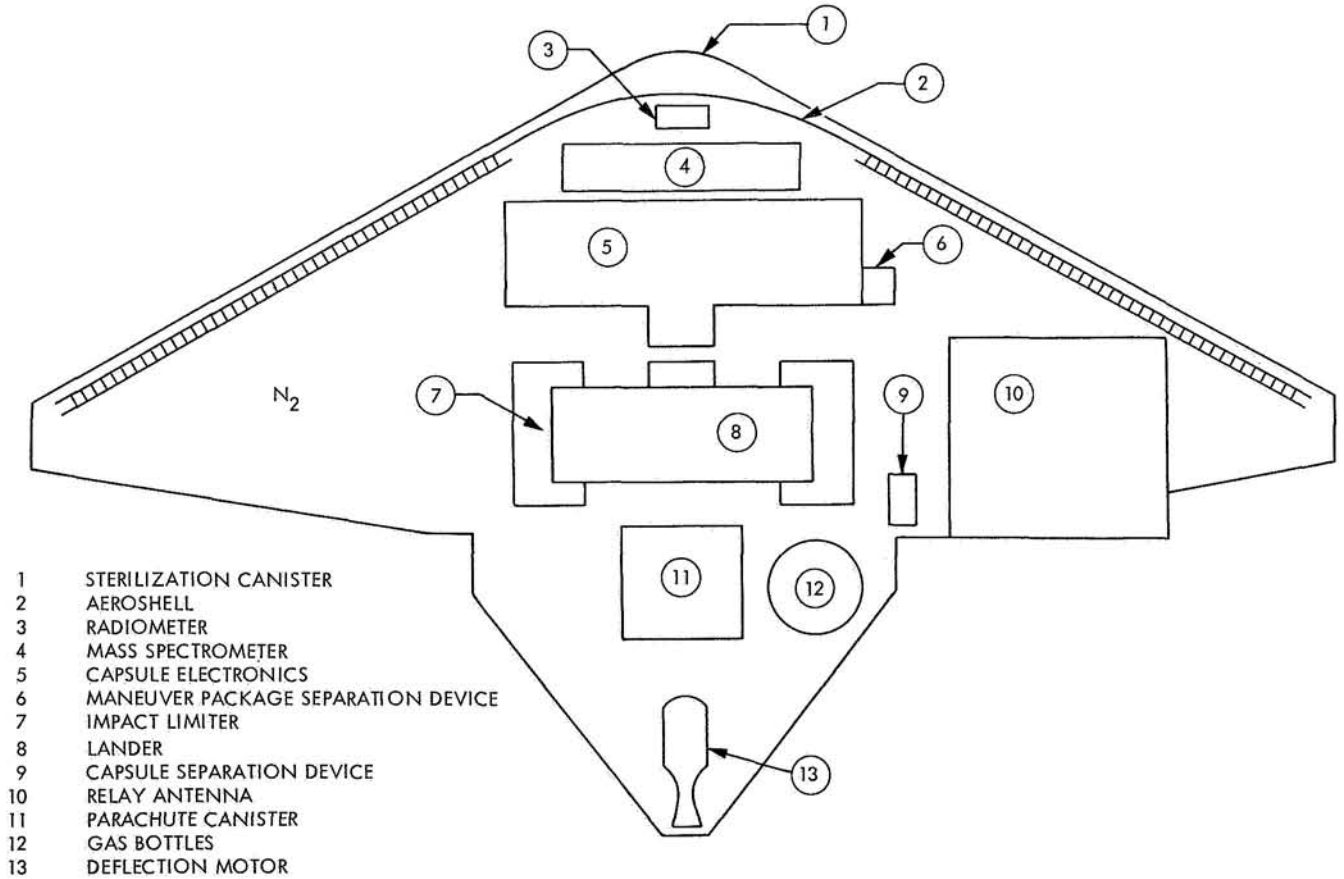


Fig. 6. CSAD thermal model

into 72 thermal nodes with 4 gaseous nitrogen nodes located within the canister; the canister itself was divided into 6 nodes to provide the capability of accounting for nonuniformity of temperatures. In order to provide length-of-cycle alternatives during the sterilization process, the thermal analysis bracketed a wide range of possible temperature responses and a wide range of possible microbial burden numbers that could exist after the cycle had begun. The predictions for a family of heating profiles used in determining the length-of-cycle alternatives are shown in Fig. 7. These curves were used for the sterilization of the capsule system advanced development (CSAD) flight model (Ref. 3).

An attempt was also made to optimize the sterilization cycle for the capsule by analyzing the effects of variations in heating and cooling rates on total process times. The boundary conditions for the heating and cooling rate evaluation included the following cases:

(1) A driving temperature rate R_c of 11°C/h was applied to the sterilization canister. The interior portions of the capsule were heated and cooled by natural convection and conduction through the nitrogen gas atmosphere inside the canister. Then, higher heating rates R_c of 19, 25, and 40°C/h were individually applied to the canister. [A rate of 40°C/h was believed to be the maximum capability of the terminal sterilization chamber (TSC).]

(2) A hot gas with a heating and cooling rate R_g of 11°C/h was forced through the capsule (while in the TSC with R_c of 11°C/h) with a flow rate of 40 ft³/min through an 8-in. port.

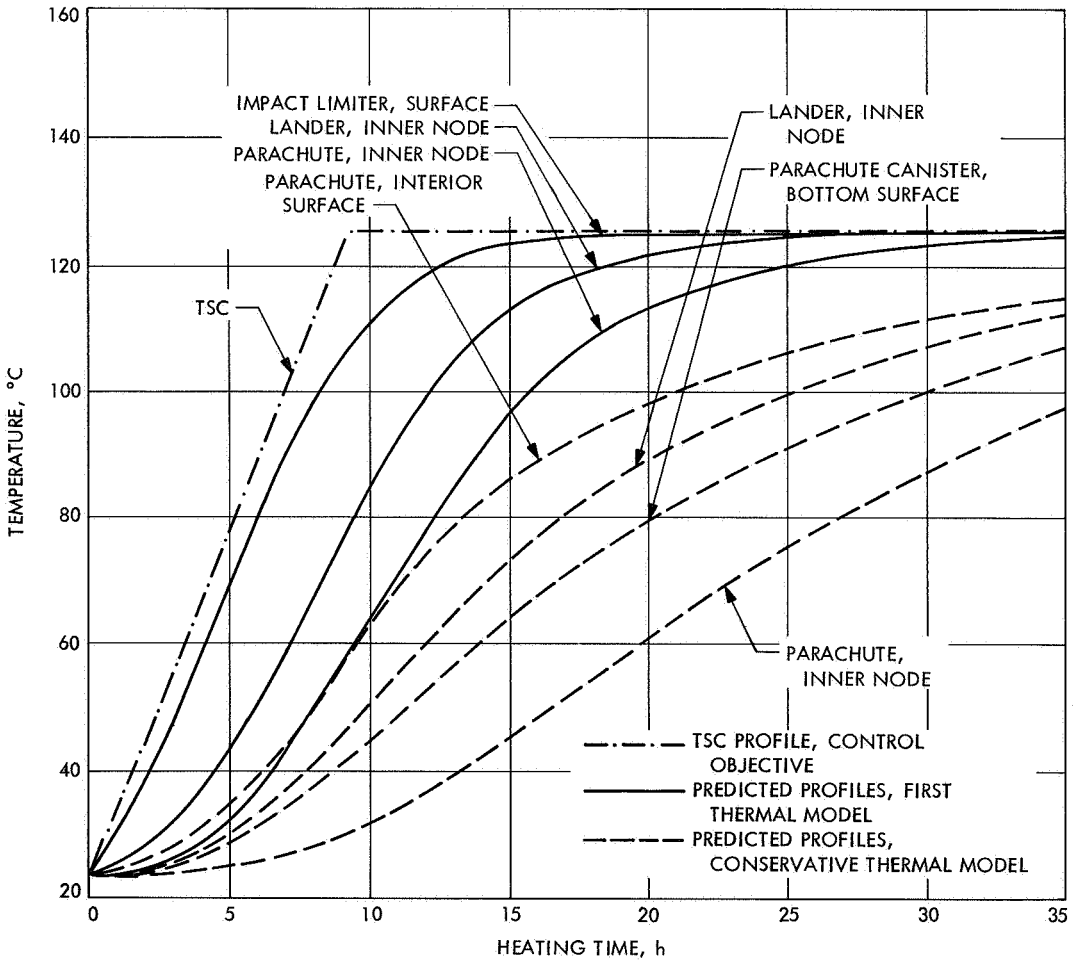


Fig. 7. CSAD heating profiles—system sterilization

4. Results

Important results of this analysis to the particular capsule configuration considered included:

(1) As R_c increased, the heat application time decreased, but the time needed to be at 125°C increased (Table 3).[4] This is attributed to the larger lethality occurring during the transient phases for the cycles with slow heating and cooling rates. If the time a subsystem is at 125°C can be used as a measure of severity, a subsystem with low-thermal mass, such as the radiometer, would experience a more severe sterilization environment at the 40°C/h rates than at the 11°C/h rate even though the same sterility level is achieved.

(2) There was no significant reduction noted in heat application time between the case where the capsule was "baked" in a gas environment and the case where the gas was forced through the canister.

Further applications using thermal models are being performed to better define and understand the parametric relationships existing in space vehicle sterilization processes.

[4]Lethality calculation assumptions: initial number of microorganisms $N_0 = 10^4$, probability of survival $P_s = 10^{-4}$, $z = 25$°C, lethality begins at 100°C.

References

1. Hoffman, A. R., and Stern, J. A., *Terminal Sterilization Process Calculation for Spacecraft*, Technical Report 32-1209. Jet Propulsion Laboratory, Pasadena, Calif., Nov. 15, 1967. (Also appears in *Developments in Industrial Microbiology*, Vol. IX, pp. 49–64, 1968.)
2. Stern, J. A., and Hoffman, A. R., *Determination of Terminal Sterilization Process Parameters*, Technical Report 32-1191. Jet Propulsion Laboratory, Pasadena, Calif., Oct. 1, 1967. (Also appears in *Proceedings of COSPAR*, London, July 1967.)
3. Hoffman, A. R., "Determination of the Terminal Sterilization Cycle for a Possible Mars Capsule," *Proceedings of the AAS-AIAA Rocky Mountain Symposium*, Denver, Colo., July 1968.

Table 3. CSAD process times for different heating and cooling rates

Parameter	Analysis case 1				Analysis case 2, gas following 11°C/h	Assumed D_{125} value
	11°C/h	19°C/h	25°C/h	40°C/h		
Aeroshell surface						
Maximum temperature, °C	118	118	119	118	120	20 min
Heat application, h	13.5	11.2	10.7	9.0	11.6	
TSC at 125°C, h	4.1	5.8	6.6	6.4	2.2[a]	
Total process,[b] h	32.0	27.2	26.9[c]	25.4	25.6	
Parachute canister surface						
Maximum temperature, °C	117	118	117	117	118	20 min
Heat application, h	14.7	12.8	12.0	11.0	13.7	
TSC at 125°C, h	5.3	7.4	7.9	8.4	4.3[a]	
Total process,[b] h	36.0	32.7	31.0[c]	29.5	32.3	
Lander inner node						
Maximum temperature, °C	122	122	122	122	122	40 min
Heat application, h	20.0	18.0	17.5	17.0	19.2	
TSC at 125°, h	10.6	12.6	13.4	14.4	9.8[a]	
Total process,[b] h	40.9	36.5	35.8[c]	35.1	37.5	

[a]Time canister atmosphere at 125°C.

[b]Sum of heat application (time from 23°C to maximum temperature) and cooling (time from maximum temperature to 25°C).

[c]Estimated, cooling profile not complete.

IV. Spacecraft Power

GUIDANCE AND CONTROL DIVISION

A. Solar Power System Definition Studies,

H. M. Wick

1. Introduction

The overall objective of the solar power system definition studies is to investigate problems associated with the development of spacecraft power systems and to develop the technology required to solve specific system design problems for JPL missions. One task which is presently being undertaken is the investigation and development of computer programs for power system design and analysis.

2. Shepherd's Equation Battery Discharge Computer Program

a. Method. A Fortran IV computer program has recently been developed for predicting battery discharge characteristics. Experimental data, at a constant discharge current, are fitted by the program to an empirical equation derived by C. M. Shepherd. This equation describes the battery potential during discharge as a function of discharge time, current density, and other factors. The method used by the computer program for determining the empirical constants of Shepherd's equation is essentially the same as described in Refs. 1 and 2.

The battery potential during discharge is given as a function of time, current density, polarization, internal resistance, and other factors:

$$Y = E_s - B\left(\frac{C}{C-X}\right)Z + D\exp\left(-\frac{EX}{C}\right) - LZ \tag{1}$$

Empirical values for E_s, L, B, C, D, and E are determined by the computer program by numerically fitting experimental discharge data to the above equation. (See Table 1 for definition of symbols.)

Battery data consisting of discharge voltage–time data, current density, number of voltage plateaus, and plateau base potentials are input to the program. A capacity versus voltage curve is computed, then the empirical constants E_s, L, B, C, D, and E for the first plateau of the discharge curve are determined by using least-square curve-fitting techniques. Similarly, the empirical constants are computed for the second plateau, provided that one exists. The method is flowcharted as shown in Fig. 1.

b. Output. The first page of the output for the computer program is shown in Fig. 2. Lines one through

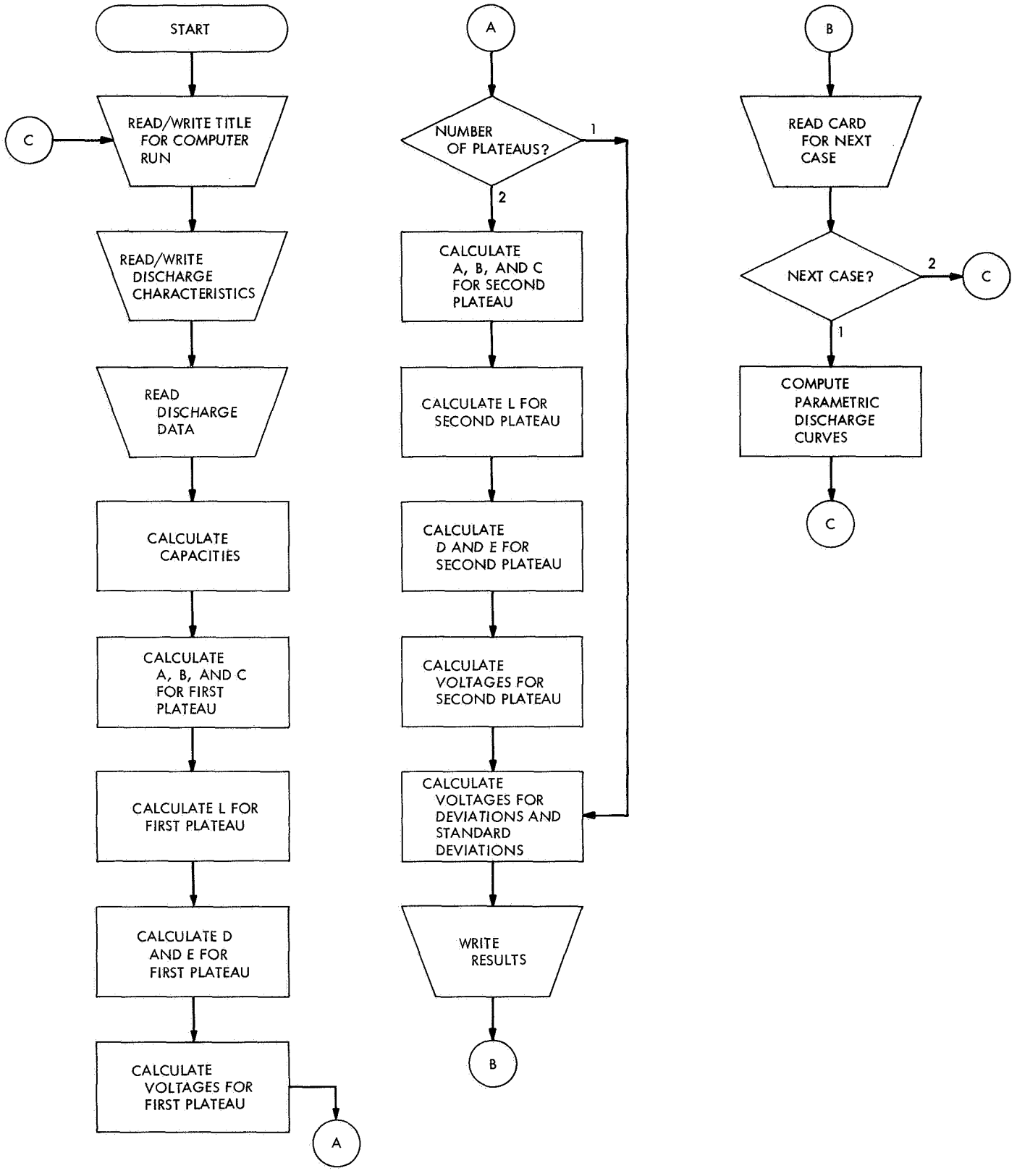


Fig. 1. Flowchart of Shepherd's equation battery discharge computer program

DELCO-REMY SILVER-ZINC 13 PLATE CELL TEMP. = 75 F HMW N3=N2

Z	ES1	ES2	NP	N1	N2	N3	N4
2.0000E 00	1.8600E 00	1.6000E 00	2	5	13	13	24

A1	B1	C1	D1	E1	L1
1.8942E 00	3.6682E-02	1.4654E 01	4.3349E-02	7.7696E 00	-1.7115E-02

A2	B2	C2	D2	E2	L2
1.5705E 00	1.9135E-03	4.5443E 01	0.0000E-39	0.0000E-39	1.4735E-02

T	X	V	Y	Y-V
0.0000E-39	0.0000E-39	1.8750E 00	1.8642E 00	-1.0785E-02
5.0000E-01	1.0000E 00	1.8400E 00	1.8410E 00	1.0028E-03
1.0000E 00	2.0000E 00	1.8200E 00	1.8243E 00	4.2825E-03
1.5000E 00	3.0000E 00	1.8100E 00	1.8108E 00	8.1447E-04
2.0000E 00	4.0000E 00	1.8000E 00	1.7985E 00	-1.4799E-03
2.5000E 00	5.0000E 00	1.7850E 00	1.7859E 00	9.2790E-04
3.0000E 00	6.0000E 00	1.7700E 00	1.7718E 00	1.8004E-03
3.6000E 00	7.2000E 00	1.7500E 00	1.7510E 00	9.5287E-04
4.0000E 00	8.0000E 00	1.7300E 00	1.7333E 00	3.2824E-03
4.6000E 00	9.2000E 00	1.7000E 00	1.6974E 00	-2.5614E-03
5.0000E 00	1.0000E 01	1.6600E 00	1.6634E 00	3.4391E-03
5.3000E 00	1.0600E 01	1.6300E 00	1.6292E 00	-8.1067E-04
5.5000E 00	1.1000E 01	1.6000E 00	1.6001E 00	1.2706E-04
5.7000E 00	1.1400E 01	1.5800E 00	1.5654E 00	-1.4578E-02
6.0000E 00	1.2000E 01	1.5700E 00	1.5653E 00	-4.6692E-03
7.0000E 00	1.4000E 01	1.5650E 00	1.5650E 00	-1.4901E-08
8.0000E 00	1.6000E 01	1.5650E 00	1.5646E 00	-3.7572E-04
1.0000E 01	2.0000E 01	1.5600E 00	1.5637E 00	3.6957E-03
1.2000E 01	2.4000E 01	1.5550E 00	1.5624E 00	7.4206E-03
1.4000E 01	2.8000E 01	1.5500E 00	1.5606E 00	1.0561E-02
1.6000E 01	3.2000E 01	1.5450E 00	1.5576E 00	1.2594E-02
2.1000E 01	4.2000E 01	1.5350E 00	1.5200E 00	-1.4982E-02
2.2000E 01	4.4000E 01	1.4500E 00	1.4500E 00	-1.4901E-08
2.2400E 01	4.4800E 01	1.3000E 00	1.3000E 00	0.0000E-39

AVERAGE DEVIATION = 6.3611E-03

Fig. 2. Typical output from Shepherd's equation battery discharge computer program, showing empirical constants and accuracy of curve fit

Table 1. Nomenclature

Y	battery potential during discharge, V
A	$E_s - LZ$
B	polarization coefficient, Ω-cm²
C	available active material, C/unit area
X	energy removed from battery during time t
Z	current density, A/cm²
E_s	constant base potential, V
L	internal resistance/unit area, Ω/cm²
D	empirical constant
E	empirical constant
$ES1$	first plateau base potential
$ES2$	second plateau base potential

three of the printout identify the battery, current density, number of plateaus, number of data points inputted for each plateau, location of the end of the first plateau and beginning of the second plateau, and the plateau base potentials. The empirical constants required in Shepherd's equation, and computed by the program, are printed on lines four through seven. The remaining information contained in the printout consists of discharge voltage, time, capacity removed, and computed terminal voltage Y. The last column of this printout lists the difference between the measured voltage and computed voltage. This difference is useful in determining how well Shepherd's equation describes the inputted battery discharge curve.

Using Shepherd's equation computer program, a parametric discharge curve was predicted for a discharge current of 2 A. Figure 3 shows this predicted curve plotted with the actual discharge curve at 2 A obtained from experimental data. Figure 4 compares the predicted discharge curve for a discharge current density of 5 A with the actual discharge curve at 5 A. In this case, Shepherd's equation constants were evaluated from the 2 A case.

3. Conclusion

This computer program provides an excellent means for modeling a wide variety of battery or cell discharge characteristics through the use of Shepherd's equation. It provides a complete description of battery discharge characteristics, using a minimum of experimental data, and facilitates the pinpointing of experimental error in the discharge data. Capacity and discharge voltage can

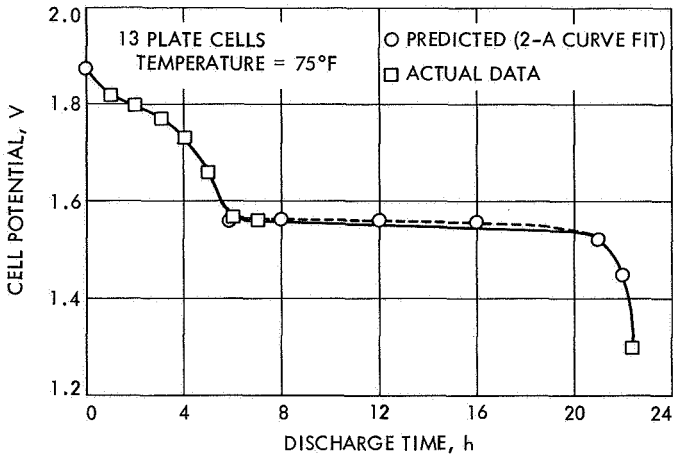


Fig. 3. Comparison of actual discharge curve and predicted curve (I = 2 A)

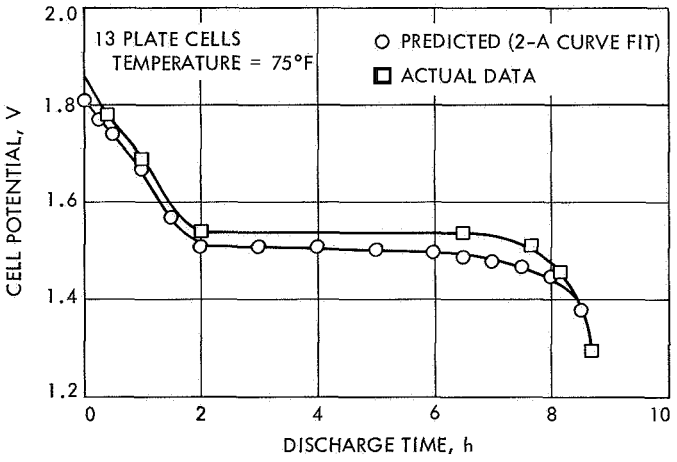


Fig. 4. Comparison of actual discharge curve and predicted curve (I = 5 A)

be predicted for a wide range of current densities by a single equation. This greatly facilitates battery analysis for the power systems engineer.

References

1. Shepherd, C. M., *Theoretical Design of Primary and Secondary Cells, Part III–Battery Discharge Equation*, Report 5908. U. S. Naval Research Laboratory, May 2, 1963.
2. Shepherd, C. M., "Design of Primary and Secondary Cells, II. An Equation Describing Battery Discharge," *J. Electrochem. Soc.*, Vol. 112, No. 7, July 1965.

B. Mars Spacecraft Power System Development,

H. M. Wick

1. Introduction

A two-phase study was initiated to design an improved *Mariner* spacecraft power system for possible future Mars

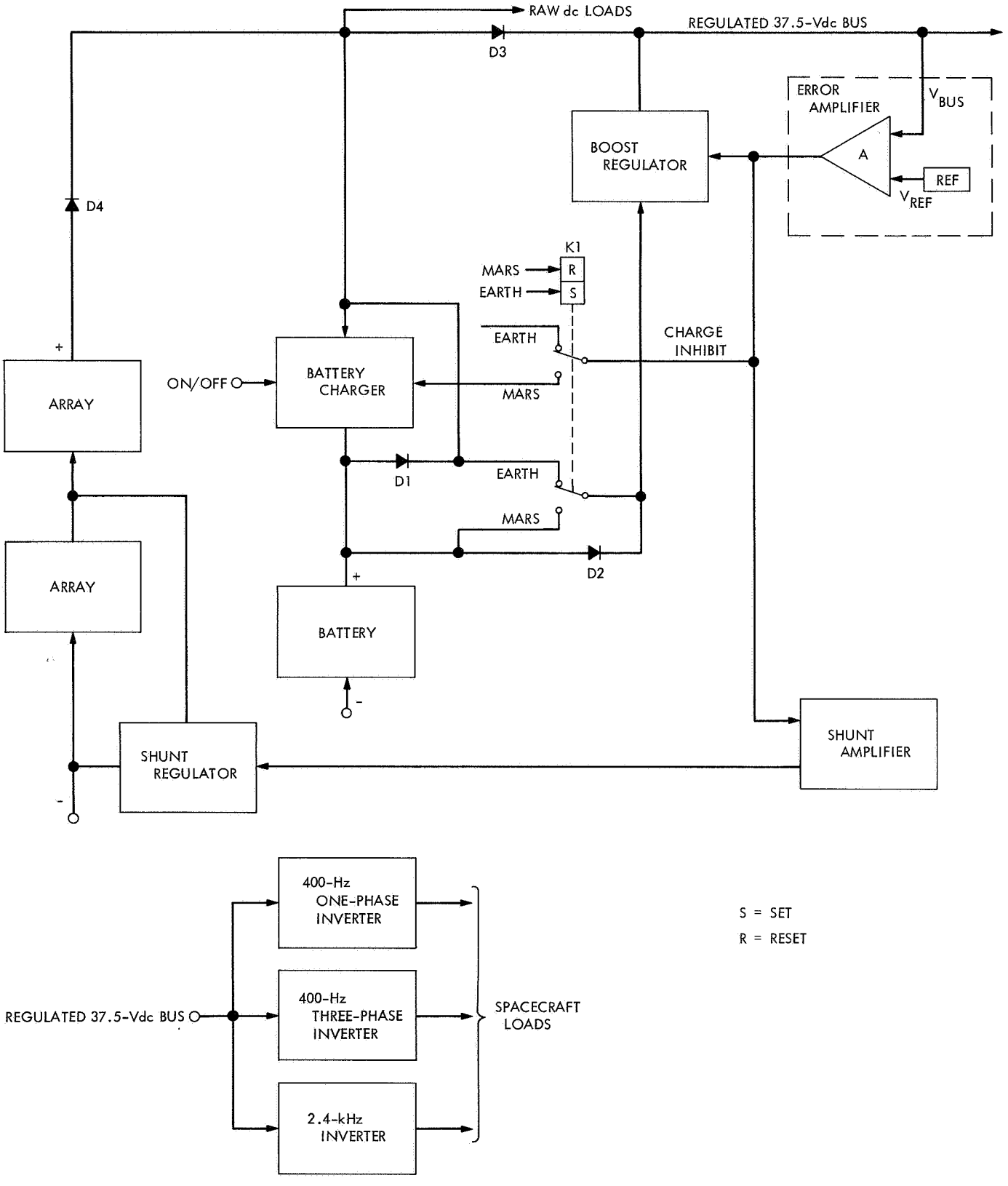


Fig. 5. GE shunt power system

missions. The latest system design techniques and component technology are being employed to develop optimum power systems for both Mars flyby and orbiter spacecraft.

In Phase I, General Electric Co. and TRW Systems investigated and analyzed various candidate power system designs. Each contractor then selected one power system for recommendation to JPL.

2. General Electric Co. Study

Using load power requirements supplied by JPL for a typical early 1970 Mars orbiting mission, the General Electric study resulted in the design of a shunt power system.

A functional block diagram of the shunt regulation system is shown in Fig. 5. Inverter power (both 2.4 kHz and 400 Hz) is derived from a regulated 37.5-V dc bus. Direct-current regulation of this bus is maintained by sequentially controlling the array shunt regulators, battery charger, and the boost regulator in response to separate regions (Fig. 6) of an error amplifier input voltage (see upper right of Fig. 5). The shunt regulator operates in the highest error region with maximum solar array shunting occurring at V_4 down to little or no shunting at V_3. At this voltage, the available array power at the regulated bus just satisfies the spacecraft load demand along with any battery charging power that may be required. Increased load demands or a decrease in available solar array power cause the battery charging power to be diverted to the load to maintain regulation. This occurs in the voltage range V_3 to V_2. With further load demands or further decrease in array power, the boost regulator comes on to maintain the regulated bus voltage between V_2 and V_1.

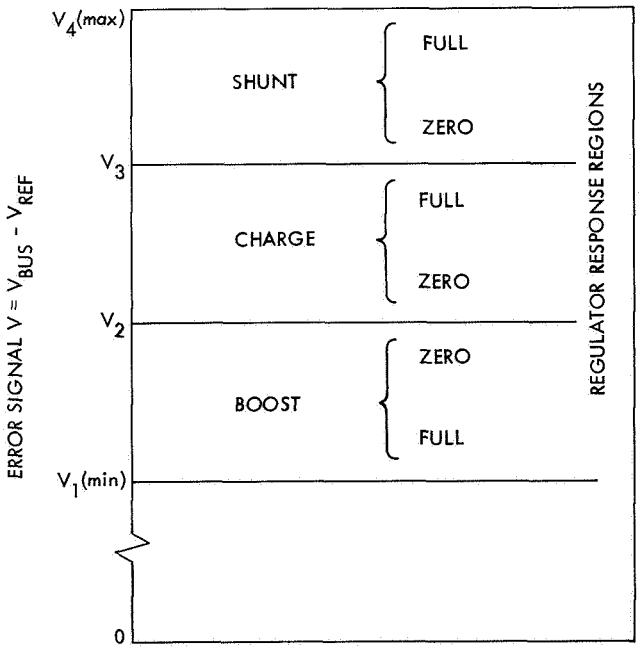


Fig. 6. Response regions of shunt, boost, and charge regulators

To reduce solar array matching problems which arise from a change in the array current–voltage characteristics in the earth-to-Mars transit, diode D3 and relay K1 (Fig. 5) have been incorporated in the design of the shunt system. The solar array power–voltage characteristics shown in Fig. 7 help to illustrate the nature of this problem. For the 1.5-AU curve (Fig. 7), which is typical for a Mars orbiting mission 90 days after encounter, the normalized voltage at the maximum power point is 1.3 V. If the system-regulated voltage is selected to be 1.3 V or higher, no power would be available at this voltage from the solar array near earth (1.0 AU). By selecting a system voltage of 1.2 V, equivalent power can be obtained for the 1.5- and 1.0-AU conditions but with a sacrifice in the power at 1.5 AU of about 6%. The in-line diode D3 and the earth/Mars mode relay K1 solve this problem.

During launch and the early cruise phase, K1 is set in the "earth" position. If the solar array voltage is low, the boost regulator operates to maintain voltage regulation. When array voltage is high, the regulated dc bus voltage is maintained by the shunt regulator and the array current passes directly through diode D3. In both cases, battery discharge diodes D1 and D2 are back-biased.

Approximately 2 mo after launch, the change in array characteristics permits transferring the earth/Mars mode relay K1 to the "Mars" position M. This prevents array/battery load sharing during later phases when the solar array power capability becomes limited. Two system modes of operation will predominate.

(1) *Near earth,* the boost regulator will be operating from solar array power. Battery charging is not inhibited since array power is much greater than required to support the spacecraft power demand. The shunt regulator is "standing by" on line. During emergence from a solar occultation near earth, the shunt regulator will be operating with the boost regulator off.

(2) *Near Mars,* the shunt regulator will be operating from solar array power with the boost regulator off.

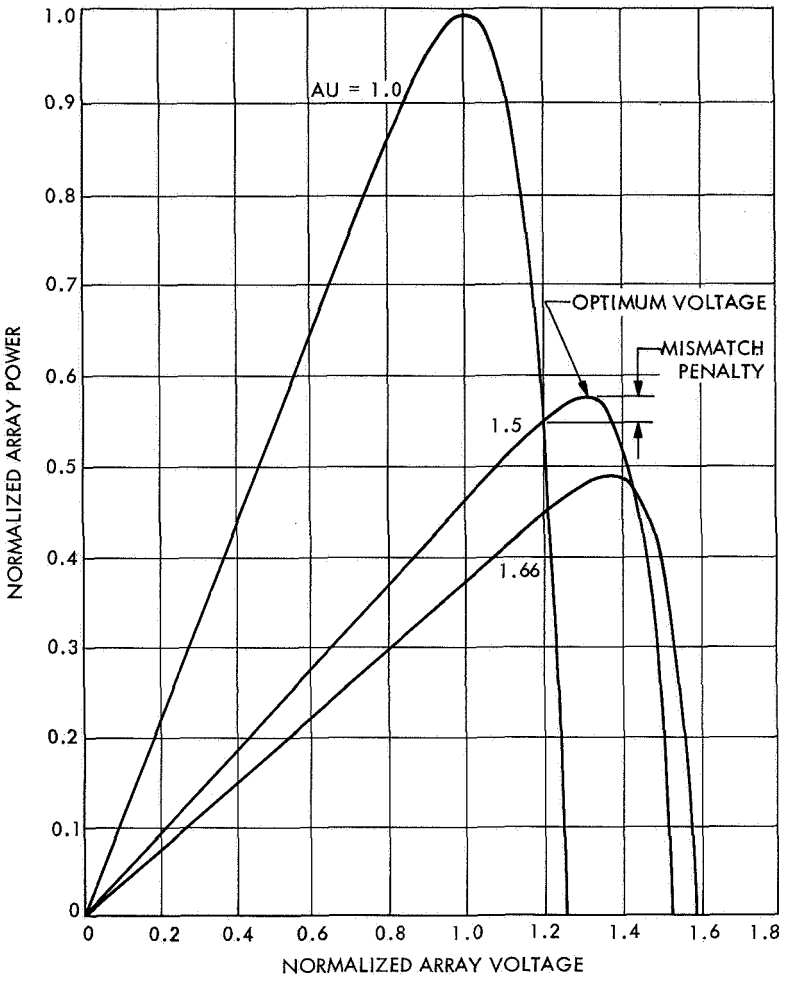


Fig. 7. Predicted *Mariner* Mars solar array power–voltage curves

Battery charging may be inhibited, if necessary, to maintain the system-regulated voltage.

The power system design recommended by General Electric eliminates the need for the array zener diodes, share mode detector, and share boost converter; but requires the addition of the earth/Mars mode relay and the array-mounted shunt regulators. Unregulated bus voltage range has been reduced from 25–50 V dc (*Mariner* Mars 1969) to 25–38.2 V dc. The power system recommended by General Electric weighs 0.3 lb more than the present *Mariner* Mars 1969 power system.

3. TRW Systems Study

TRW recommended a buck–boost power system. A simplified block diagram of this system is shown in Fig. 8. A regulated 50-V dc bus supplies power to the 400-Hz and 2.4-kHz inverters and to the various spacecraft dc loads (traveling-wave tube, heaters, etc.). Regulation is provided by a buck–boost line regulator. Redundant silver–zinc 25 A-h batteries have been included in the system design. Battery charging power is obtained from the regulated dc bus. A current-limiting resistor and a relay are used to control charging. Ground command backup capability is provided to override the automatic charge control circuitry that controls the relay.

Power system reliability has been enhanced through incorporation of current limiting and operational redundancy in the buck–boost regulator. The need for zener diode voltage limiters on the solar array has been eliminated. Packaging requirements reduce the number of power conditioning modules from 10 (*Mariner* Mars 1969) to 6. Power system weight has been reduced by about 4 lb.

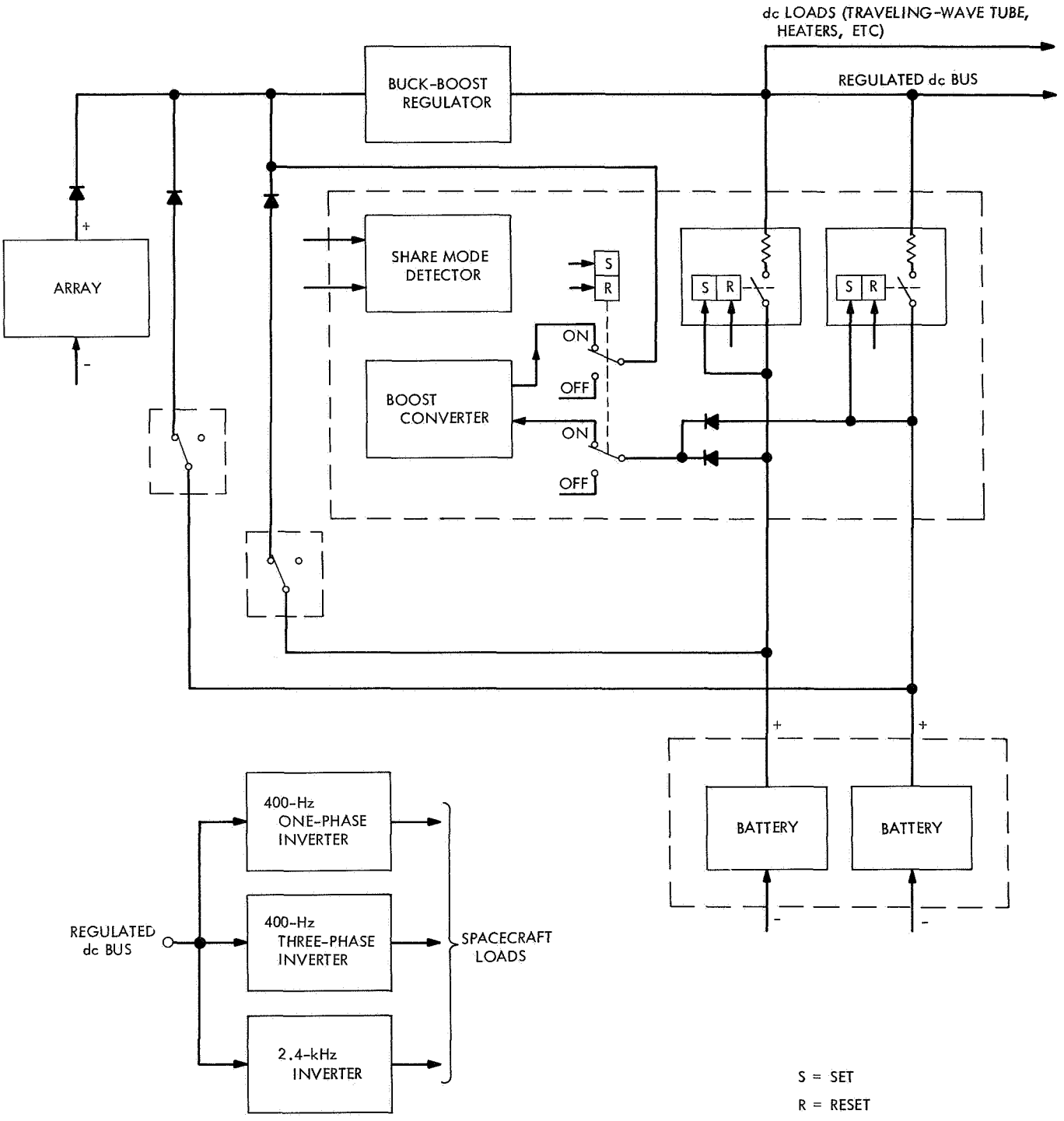


Fig. 8. TRW buck-boost power system

C. Solar Cell Contact Studies, *P. A. Berman and G. P. Rolik*

1. Introduction

The purpose of these studies is to determine the characteristics of solar cell contacts, especially as a function of exposure to various environmental conditions. Many problems exist with present-day state-of-the-art titanium–silver contacts, with and without solder coating. The manufacturers have not analyzed solar cell contacts sufficiently to determine what characteristics the cells will have as a result of exposure to environmental extremes.

2. Testing

a. Ferranti Electric, Inc. cells. These cells have contacts fabricated by plating nickel onto the silicon and layers of copper and gold over the nickel. Forty-eight cells, 24 fabricated from 1 Ω-cm material and 24 from 10 Ω-cm, have been exposed to an environment consisting of 95% relative humidity at a temperature of 60°C for a period of 48 h.

Before and after environmental exposure, the cells were electrically tested at a cell temperature of 28 ± 1°C in a tungsten simulator having a color temperature of 2800°K and an equivalent solar intensity of approximately 100 mW/cm². The tungsten simulator will be used extensively during these studies because of the excellent stability and reproducibility of the spectral and intensity characteristics of this source. Since the nature of the measurements to be made is, for the most part, comparative (i.e., pre- and post-environmental exposure), the fact that the tungsten spectrum differs from the sunlight spectrum should not affect the analysis.

It was determined that the environmental exposure did not adversely affect the electrical characteristics of the cells. Table 2 lists the average parameter values before and after exposure for both groups of cells.

b. Ion Physics cells. In these cells, developed by Ion Physics under the improved solar cell contact development program, the aluminum contacts were deposited by high-vacuum sputtering. The first lot was subjected to a temperature–humidity environment of 95% relative humidity at a temperature of 80°C for 30 days.

Electrical tests before and after exposure of the cells to this rather severe environment were performed under tungsten illumination, the intensity being adjusted to correspond to a solar intensity of 100 mW/cm². Five control cells were held back from the total lot of 15 cells. The test temperature was 28 ± 1°C. Table 3 lists the average parameter values before and after exposure for the 10 cells tested. The test results show an extremely small degradation in cell characteristics.

Visual inspection after the environmental test showed the aluminum contacts to be oxidized. It is thought that this film of Al_2O_3 is the cause of the slightly higher series resistance experienced in the post-humidity test.

A tape test utilizing Scotch 810 tape was also conducted after exposure to the humidity environment. The

Table 2. Ni-Cu-Au contact cell characteristics before and after 48-h exposure to 95% relative humidity at 60°C

Test group/ time	Open-circuit voltage, V	Short-circuit current, mA	Current at max power, mA	Voltage at max power, V	Max power, mW	Curve power factor,[a] %
1 Ω-cm						
Before exposure	0.583	99.6	88.9	0.480	43.0	0.73
After exposure	0.584	98.7	88.7	0.477	42.4	0.73
10 Ω-cm						
Before exposure	0.542	100.3	93.0	0.436	40.6	0.74
After exposure	0.545	99.01	91.8	0.439	40.4	0.74

[a]The curve power factor is a measure of the "squareness" of the curve:

$$\text{curve power factor} = \frac{\text{max power}}{\text{short-circuit current} \times \text{open-circuit voltage}}$$

Table 3. High vacuum sputtered Al contact cell characteristics before and after 30-day exposure to 95% relative humidity at 80°C

Test time	Cell series resistance, Ω	Short-circuit current, mA	Open-circuit voltage, V	Current at max power, mA	Voltage at max power, V	Max power, mW	Curve power factor, %
Before exposure	0.54	95.7	0.541	87.7	0.430	37.7	0.728
After exposure	0.63	96.7	0.541	87.8	0.429	37.7	0.714

"N" contact strip (not grids) and back sides were tested by the peel method, with the peel starting at the edges of the cells. Results showed excellent contact adherence. There was no evidence of contact material residue on the tape.

3. Conclusions

The nickel–copper–gold contact cells procured from Ferranti Electric, Inc. appear to successfully survive a 48-h exposure to a 95% relative humidity, 60°C environment. The series resistance before and after exposure was not measured, but the fact that the pre- and post-exposure curve power factors were identical strongly indicates that there was no large change in series resistance.

The high-vacuum sputtered aluminum contact cells developed by Ion Physics under JPL contract exhibited no significant electrical degradation as a result of a 30-day exposure to the severe environment of 95% relative humidity at 80°C. A slight increase in series resistance was observed, but this is believed to be primarily due to a layer of aluminum oxide which formed as a result of the test, and which acted as a thin insulating layer between the test probes and the aluminum. In many cases, it was found that a slight scraping of the contact significantly lowered the series resistance. It is anticipated that in actual flight-use the cells would be interconnected prior to such exposures so that oxidation, at least of the magnitude observed here, would not result in an increase of series resistance.

D. Solar Cell Standardization, *R. F. Greenwood*

1. Introduction

Standard solar cells calibrated above 97% of the earth's atmosphere using high-altitude balloons have been effectively used over the past several years to aid in the prediction of solar array output. The standard cells are mounted in modular form, permitting temperature control and providing a means to electrically load the cell and monitor its output. Two 1- × 2-cm cells or one 2- × 2-cm cell can be mounted on the module. The modular form also provides protection for the cell during normal handling and during payload impact upon balloon flight termination.

Standard solar cells are also used to aid in establishing the light intensity and evaluating the spectral content of solar simulators. Balloon-calibrated standard solar cells in conjunction with a simulator are presently being used to classify solar cells according to power output during cell procurements, thus eliminating the use of a pyrheliometer.

Cooperative efforts between JPL and other NASA and government agencies have provided standard solar cells at minimum expense for space flight programs and advanced solar cell development.

2. Description of Calibrated Solar Cells

Solar cells submitted for calibration on the 1968 balloon flight series were from several sources. The NASA Goddard Space Flight Center supplied eight modules containing 1- × 2-cm N/P Heliotek solar cells which are intended for use with the *Orbiting Astronomical Observatory* program. The German Research Satellite Corp., in cooperation with NASA Goddard Space Flight Center, submitted seven 1-cell modules, six of which were balloon-calibrated and all of which were correlated in the JPL X-25L solar simulator. Four of the cells were manufactured by Siemens AG and three by AEG-Telefunken. Both suppliers are located in West Germany.

The Air Force Aero Propulsion Laboratory supplied eight modules containing advanced development solar cells. Three modules contained ion-implanted silicon cells. Three were assembled with cadmium sulfide solar cells having an H-film (Kapton) covering, and two modules contained cadmium telluride solar cells.

Two modules fabricated by the NASA Langley Research Center to aid in solar cell calibration were included on the flights. Heliotek 1- × 2-cm N/P cells having a base resistivity of 10 Ω-cm were used in the module assembly.

The Applied Physics Laboratory submitted a solar cell and bandpass filter experiment. The experiment consisted of five solar cells and four interference-type bandpass optical filters designed to divide the solar cell response into four equal energy bands. The fifth cell was unfiltered. Calibration data from the cell-filter combinations are intended for use with laboratory solar simulators.

The Jet Propulsion Laboratory included several experimental modules, as well as modules for use on a possible *Mariner* Mars 1971 flight. A unique experiment employing a filter wheel in conjunction with four solar cells and two radiometers was flown. The radiometers, more accurately described as enclosed standard cavity active radiometers, were designed and built at JPL.

3. Results of the 1968 Balloon Flights

A series of three 80,000-ft balloon flights was conducted in the vicinity of Minneapolis, Minnesota during the months of July and August, 1968. The solar trackers which had been modified to increase the payload capacity functioned perfectly throughout the flights. Figure 9 shows the modified tracker with payload mounted for flight 1. The tracker provides space for as many as 26 solar cell modules in the event that all modules should contain a single 2- × 2-cm cell. Two additional spaces are available to accommodate temperature-monitoring modules.

Good data were returned from each flight. Also, excellent correlation with solar simulator measurements was obtained prior to and following the flights. All cells were recovered, although two modules suffered minor damage upon impact of the second flight payload. The solar tracker was extensively damaged and a spare tracker was used for the third flight.

Flight data on all solar cells have been reduced through a computer program and all modules, along with calibration data, have been returned to the respective agencies. A formal report on the 1968 balloon flights is now in progress.

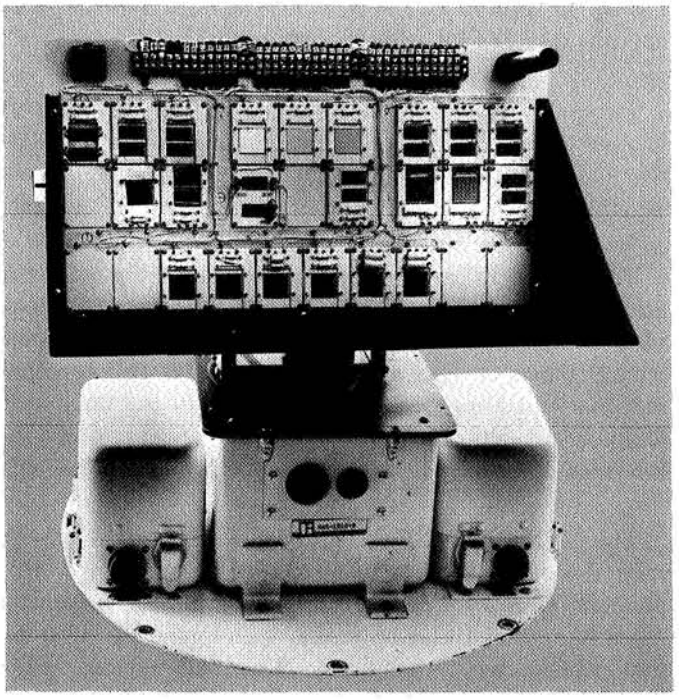


Fig. 9. Modified solar tracker

E. Advanced Roll-Up Solar Array Concept,

W. A. Hasbach

1. Program Objective

A program was initiated by JPL in June 1967 to investigate the feasibility of developing a 10-kW solar array system which would have a specific power capability of 30 W/lb and could be deployed after launch through a roll-out technique similar to that employed in a window shade.

The ever-increasing power requirements of the spacecraft, coupled with limited launch vehicle storage capacities, dictated the need to evaluate new deployment and structural designs of solar power panels. Improvement in the solar cell conversion efficiency is not anticipated in the immediate future. However, improvements in the power-to-weight ratio appear feasible using lightweight structural concepts and packaging techniques. Hence, emphasis is placed upon the mechanical–structural aspects of the solar array.

Studies performed by General Electric Co., Fairchild-Hiller Corp., and Ryan Aeronautical Co. indicate that the program objective of a 30-W/lb power-to-weight ratio system can be achieved. The total array will consist of four roll-out panels, each containing 250 ft^2 of deployed surface area. The four roll-out panels will be mounted symmetrically about the base of the spacecraft and will deploy uniformly without disturbing the center of gravity of the vehicle.

Previous accomplishments in this program were reported in SPS 37-48, Vol. III, pp. 51–57 and SPS 37–49, Vol. III, pp. 93–99.

2. Configuration Studies

For analysis, the roll-out array can be divided into three major components:

(1) The deployment mechanism or extendible boom.

(2) The storage drum upon which the solar cell substrate is wrapped during storage and launch.

(3) The substrate upon which the solar cells are attached.

Although the feasibility of 30 W/lb solar cell power systems appears to be within the current state-of-the-art, significant questions still exist:

(1) Will the substrate track uniformly during the retracting cycle?

(2) Will thermal cycling degrade the array output?

(3) Will the solar cell/coverglass combination be adequately protected during launch?

In this article, the greatest emphasis will be placed on the deployment mechanisms proposed by each of the three contractors. Both the substrates and storage drums are sufficiently common to all designs to allow for a general discussion of each.

During the study of deployment methods, it was noted by the three contractors that no one extendible boom system (Fig. 10) was clearly the preferred choice to achieve the desired objectives of this program. Designs selected differ broadly in concept yet still achieve the

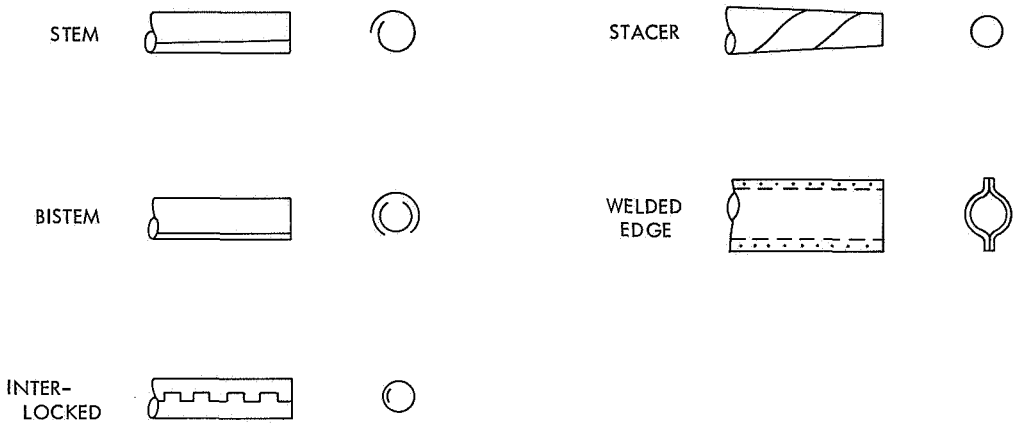


Fig. 10. Types of deployment booms

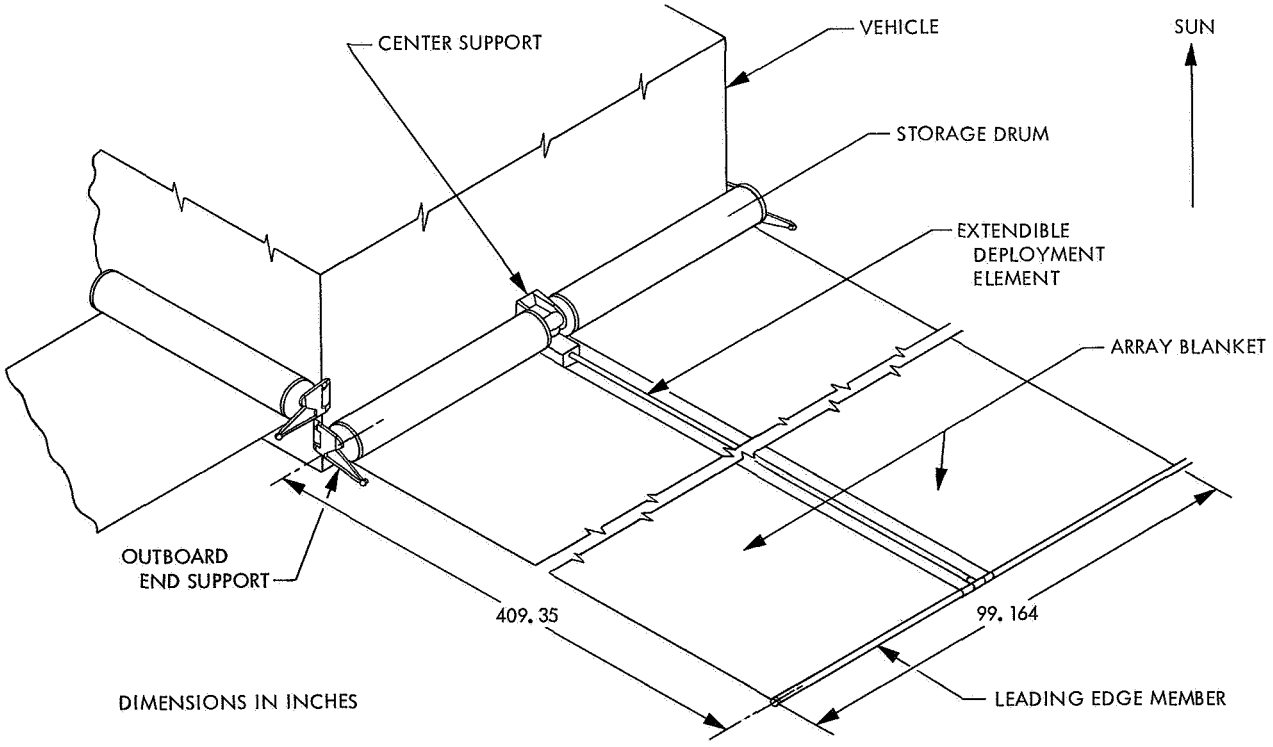


Fig. 11. Deployable 30-W/lb solar array (General Electric Co.)

30-W/lb weight requirement. Also, analysis of the various approaches has shown that they will survive the environments defined by applicable specifications.

3. Deployment Methods

a. General Electric Co. Figure 11 shows a single-rod deployment scheme, using a de Havilland bi-stem extendible element. This element consists of two stainless steel tapes which are pre-stressed and wrapped on a storage drum. As the two tapes are unrolled from the storage drum, they form tubes, one within the other. Fastened at the leading edge of the bi-stem is a cross member to which the substrate material is attached. This cross member is free to rotate through a ball bearing joint. The free floating end member is so designed as to allow the extendible rod torsional freedom during deployment. As the bi-stem rod is extended, it unrolls the solar cell substrate from the storage drum. Positive tension is maintained on the substrate at all times by means of a "negator" spring in the storage drum. Substrate tension is required to keep the solar cell surface plane flat within 10 deg and to provide a minimum natural frequency of 0.04 Hz. To achieve these conditions, it has been calculated that a substrate tension of 4.0 lb is required.

The bi-stem is fabricated of 0.007-in.-thick No. 301 stainless steel (silver-plated). Its extended length is 33.5 ft and its diameter is 1.34 in. The extension and retraction rates are 1.5 in./s.

b. Fairchild–Hiller Corp. Figure 12 also shows a single-boom deployment system, but the extension mechanism design is considerably different from other designs investigated during this study. Deployment is accomplished

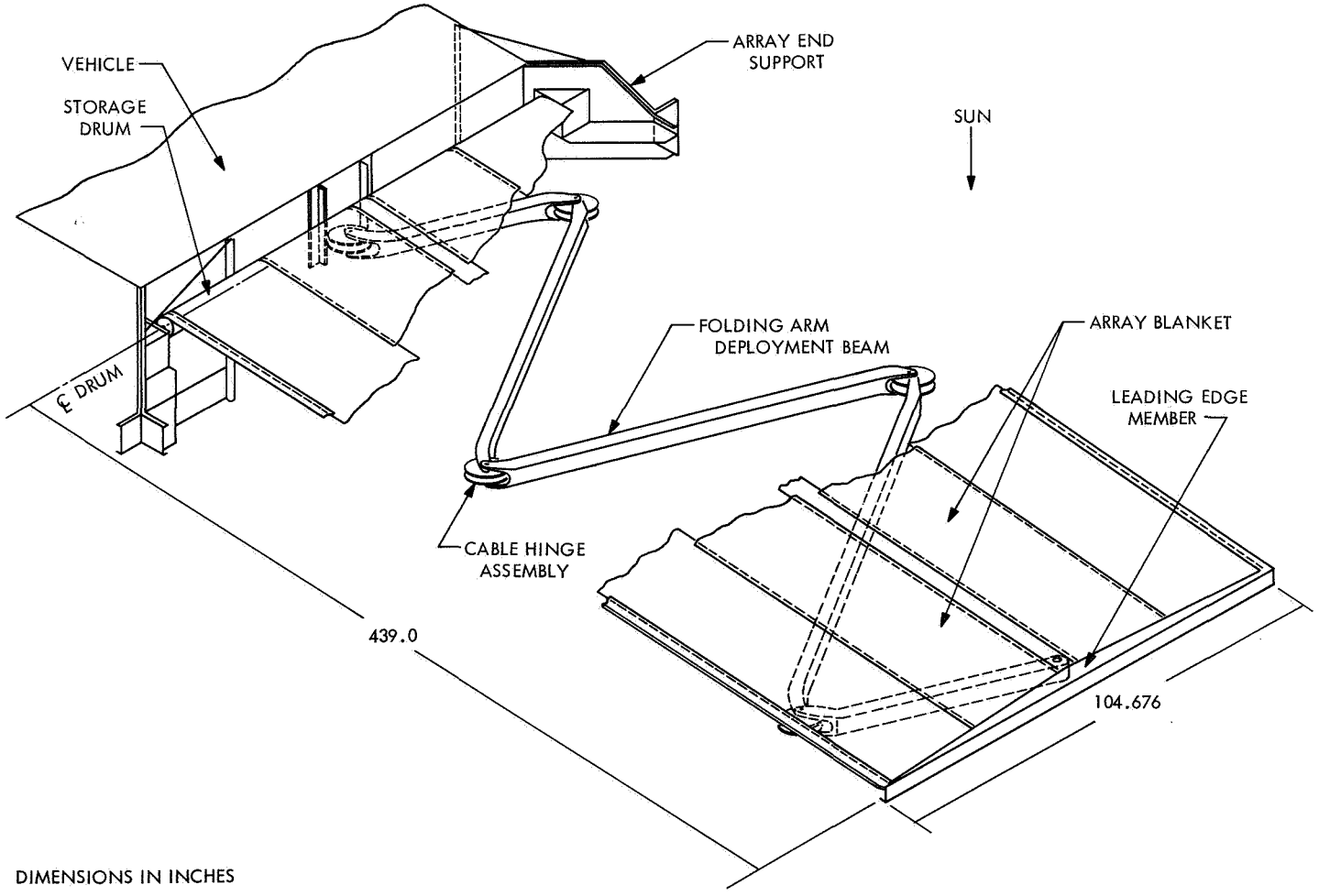


Fig. 12. Deployable 30-W/lb solar array (Fairchild–Hiller Corp.)

by means of programmed folding arms which, during the launch and stowed period, are stacked parallel to each other adjacent to the substrate storage drum. The folding arms consist of three full-length arms and two half-length arms. The half-length arms are end members of the folding arm linkage with the full arms in between. The outboard half link is pinned to the cross member which pulls the solar cell substrate from the storage drum.

Interconnecting each arm section is a fitting operated by a pulley cable arrangement. This technique is best described by comparing it to a drafting machine. In the programming of the joint action, it is essential that the arm linkages extend the end member in a straight line to avoid distorting the solar cell substrate during deployment and retraction. During deployment, tension is maintained on the substrate by means of a negator spring, as in the General Electric design, to assure a minimum natural frequency of 0.04 Hz in the substrate and to keep the substrate surface plane flat within 10 deg. To achieve these conditions, Fairchild–Hiller has determined that a force of 10.0 lb should be applied.

In the folding arm design, stainless steel, titanium, aluminum, and beryllium are used for the control cables, pulleys, and associated hardware. Boron/epoxy composites have been selected for the cross member to which the solar cell substrate is attached and for the arm segments. The arm segments are corrugated 3.0-in.^{2} tubes with a wall thickness of 0.911 in. When extended, the total length of the arms is 39.1 ft. The deployment and retraction rates are not yet established.

c. Ryan Aeronautical Co. Figure 13 shows a system using two deployment booms. Unlike the General Electric and Fairchild–Hiller systems, the solar cell substrate is fastened between the two booms, along the edges, by means of tabs at intervals of approximately 4.0 in. The two booms, which collapse as they are wound about a drum, store on the same drum as the solar cell substrate.

The booms are pre-stressed titanium, constructed of identical halves, welded along their entire length. In their natural state, the two halves are expanded, forming a hollow tube. When compressed, they can be wrapped tightly about a drum or cylinder.

As the booms are deployed, the solar cell substrate is carried off of the same storage drum. No substrate tension is considered necessary because of the edge support by the two-beam system. Thermal deflection is maintained within the allowed 10-deg bending, by sizing of the two beams and thermal control coating.

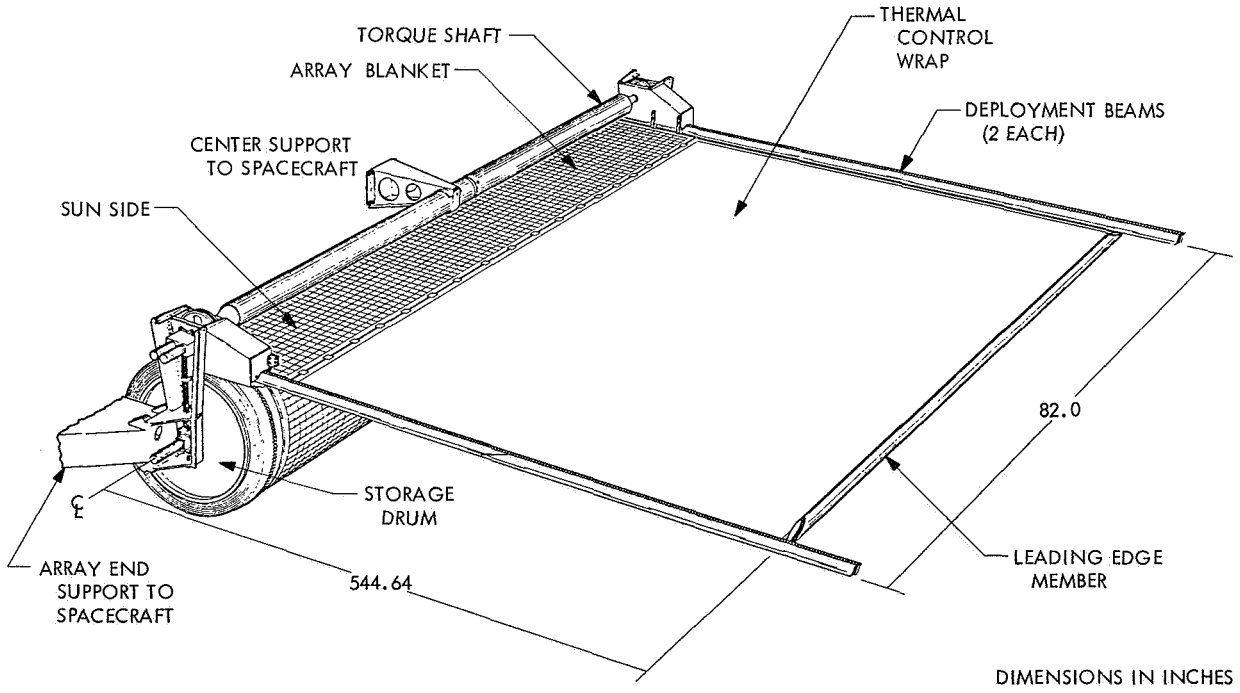


Fig. 13. Deployable 30-W/lb solar array (Ryan Aeronautical Co.)

The two booms, constructed of 0.003-in.-thick titanium, form an oval pattern 2.2 × 1.7 in. in the expanded condition. In the fully deployed condition, the booms are extended to 33 ft. The deployment and retraction rate is 1.7 in./s.

4. Substrate Materials

During the materials selection effort of this program, the contractors evaluated various candidate substrate materials; the two most promising were fiberglass and Dupont Kapton H-film. Kapton H-film was selected for its greater tear resistance, greater flexible strength, and strength-to-weight advantage.

Because of the two-boom system selected, Ryan can manufacture the substrates of 0.001-in.-thick H-film. General Electric, with a substrate tension of 4.0 lb, has selected 0.002-in. H-film. Fairchild–Hiller, having the highest substrate tension of 10 lb, is using 0.003-in. H-film.

5. Storage Drum

Each contractor will use a cylindrical drum upon which the solar cell substrate will be stored. The drums vary in diameter from 5.0 to 12.0 in.

Choice of drum material by the contractor is optional inasmuch as the materials chosen can be used interchangeably between the three designs without a significant weight penalty. With the exception of the graphite/epoxy composite material used by Fairchild–Hiller, all materials are available and considered state-of-the-art. Composite materials have limited acceptance in proven applications. General Electric has specified beryllium; Ryan has specified either beryllium or titanium as drum materials.

F. Planetary Solar Array Development,

W. A. Hasbach

Trade-off studies of weight, power capabilities, and structural integrity versus exposure to the Martian environment have resulted in the selection of a preferred solar array design for a soft lander capsule. Although studies have confirmed three possible approaches which have the potential of meeting the program objective, one is felt to be more worthy of further investigation. The non-oriented conical truncated cone (Fig. 14) was selected as the most feasible design because of its higher reliability factor, best compromise in power versus weight, absence of electrical motor gear drives, no power requirement necessary for operation, and its design growth potential; this array, once released from its locked, launched, and flight position, will require no power from the lander capsule for deployment or continuous operation for the mission life of 1 yr. A summary of earlier studies of the three possible approaches appears in SPS 37-51, Vol. III, pp. 37–41.

The non-oriented conical truncated cone array, as recognized initially in its concept, will not meet the desired goal of 20 W/lb (1 AU) and under worst-case conditions will attain less than the minimum power requirement of 200 W of electrical power at solar noon. On the other hand, Figs. 15 and 16 show that in the large majority of cases the power output exceeds the minimum requirement of 200 W. The worst-case minimum power is 5% low at 190 W. The best-case condition is 35% high at 256 W. The average noon power outputs of the limiting conditions shown are 17% high at 223 W. At the higher solar intensities that occur during the spring and fall seasons, the power level is above 200 W for all conditions.

The power-to-weight ratio varies with the power output of the array at noon at a specific Martian location. The range of power outputs for the first day of summer (lowest solar intensity) are shown for the noon conditions in Figs. 15 and 16. The array weight breakdown is:

Mechanical (solar panel, adhesives, supports, frames, deployment mechanism, etc.).	33.07 lb
Electrical (solar panel and adhesives).	23.37 lb
Total system weight	56.44 lb

The specific power output is based on the equivalent power at 1 AU. Taking the power output at the worst-case condition of 46 mW/cm² (summer), the following limits are obtained:

$$\text{Summer (maximum)} = 256 \text{ W}$$

$$\text{Summer (minimum)} = 190 \text{ W}$$

Converting to 1 AU by the ratio of $46/140 = 0.328$,

$$256/0.328 = 780 \text{ W (1 AU)}$$

$$190/0.328 = 580 \text{ W (1 AU)}$$

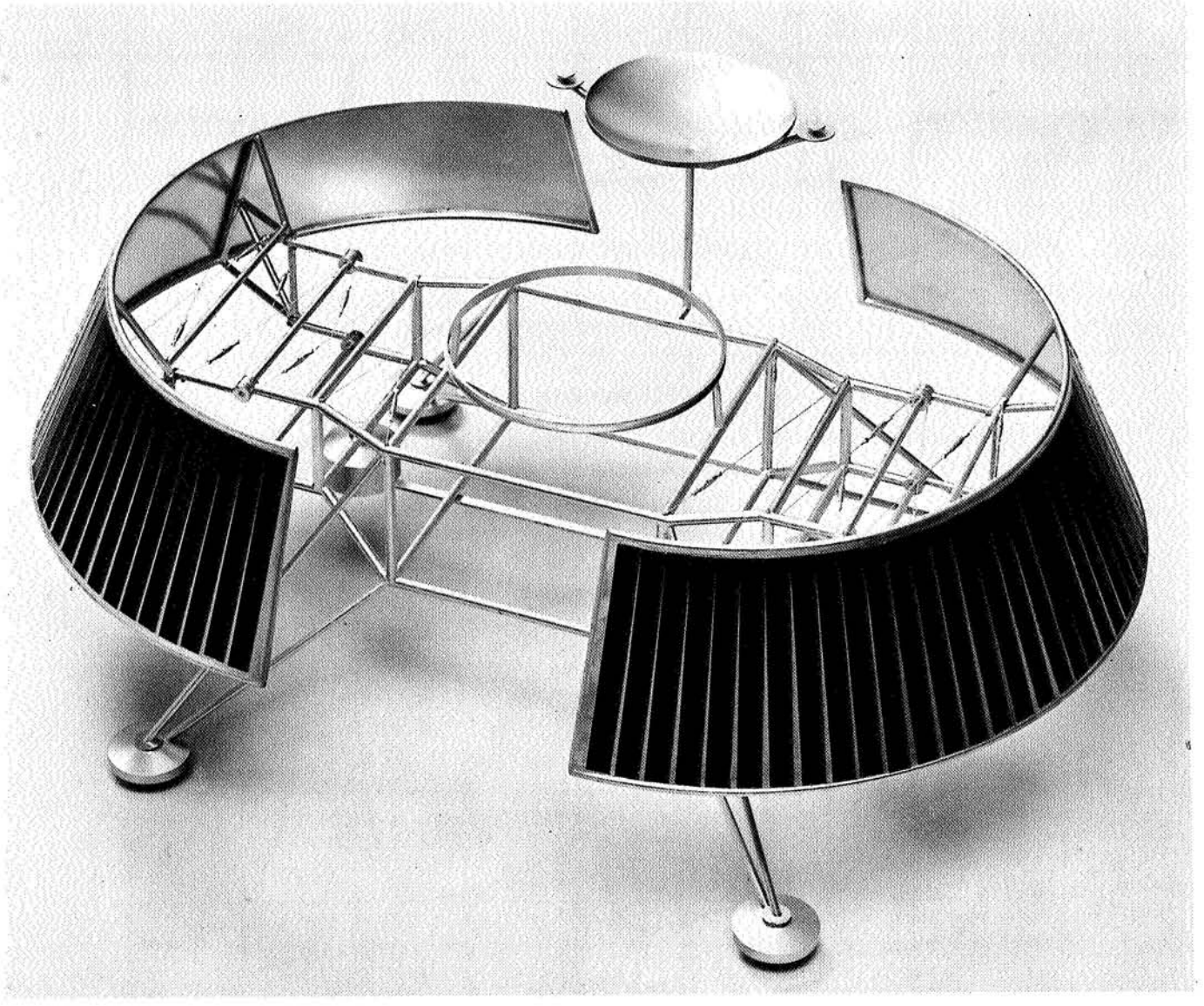


Fig. 14. Planetary solar array (non-tracking deployable system)

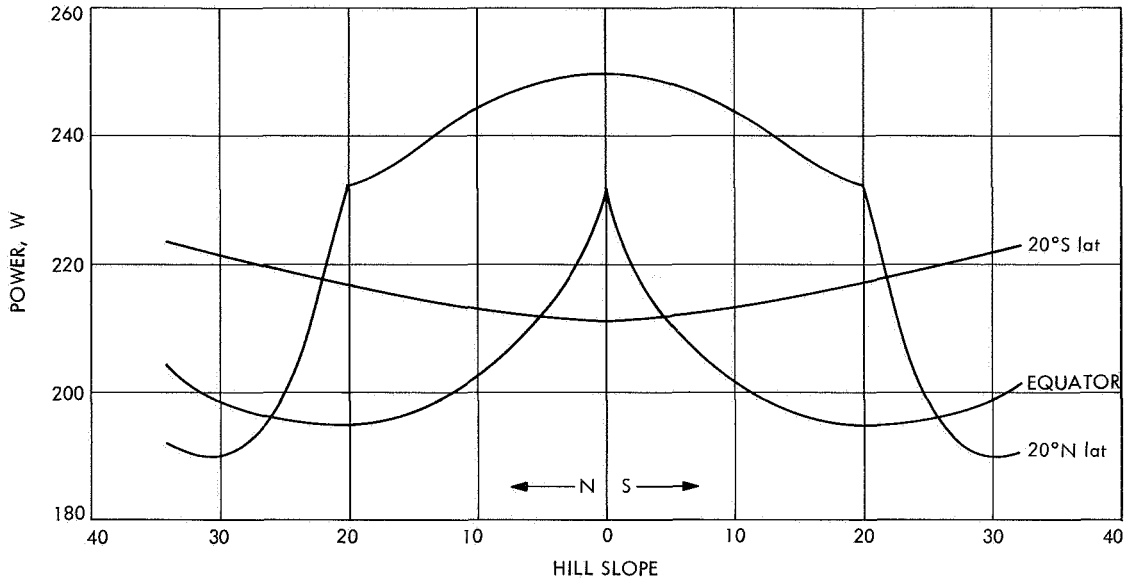


Fig. 15. Power output at noon for 20°N lat, equator, and 20°S lat vs north–south hill slope for the first day of summer

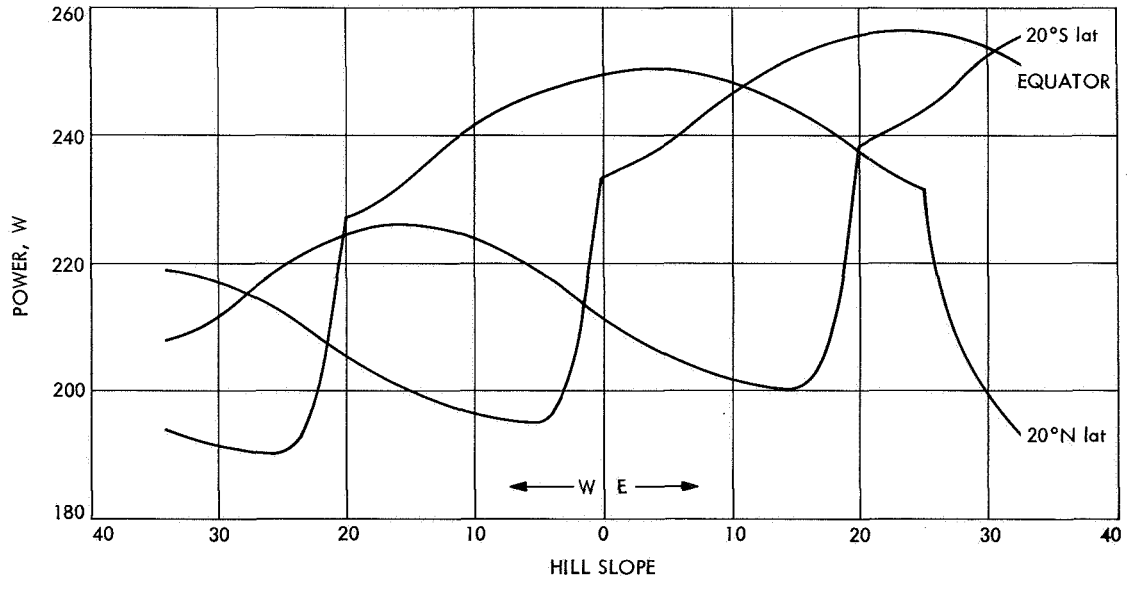


Fig. 16. Power output at noon for 20°N lat, equator, and 20°S lat vs east–west hill slope for the first day of summer

Therefore, the specific power would lie between the range of

$$780/56.44 = 13.8 \text{ W/lb}$$

$$580/56.44 = 10.3 \text{ W/lb}$$

G. Electrolytic Determination of the Effective Surface Area of the Silver Electrode, Part II,

G. L. Juvinall

1. Introduction

A major objective of the continuing study of the reaction geometry of alkaline battery electrodes is the development of new and better methods of measurement of the effective electrolytic surface area of a working electrode. This study is being performed at Brigham Young University under JPL contract; Dr. Eliot Butler is the principal investigator. Earlier results of the surface area studies were reported in SPS 37-39, Vol. IV, pp. 19–21. A new coulometric potentiostatic method is reported here.

2. Measurement Method

The new method of electrode area determination is based upon the charge-acceptance per unit area of smooth standard electrodes. The charge-acceptance can be directly related to the thickness of the oxide layer formed during the oxidation of the silver electrode in alkaline solution. At present, the relationship between the charge-acceptance per unit area and the applied potential in a constant potential oxidation is under study.

Earlier work on the aluminum–aluminum oxide electrode has shown that the charge-acceptance per unit area is the same at identical applied potentials for electrodes of different surface roughness (Ref. 1). Thus, the following equation may be used to calculate surface area:

$$a_{\text{unknown}} = a_{\text{standard}} \frac{q_{\text{unknown}}}{q_{\text{standard}}} \quad (1)$$

where a = surface area in cm^2 and q = total charge in coulombs. Two oxidation runs are required; then, if the area of one electrode is known, the area of the other can be calculated. The total charge is obtained from the integral of the voltage–time curve. A diagram of the potentiostat and integrator circuit is shown in Fig. 17. Figure 18 is a diagram of the oxidation cell, showing the location of the electrodes. This electrode arrangement minimizes the I–R drop between the reference and working electrodes. The cell was thermostatted at

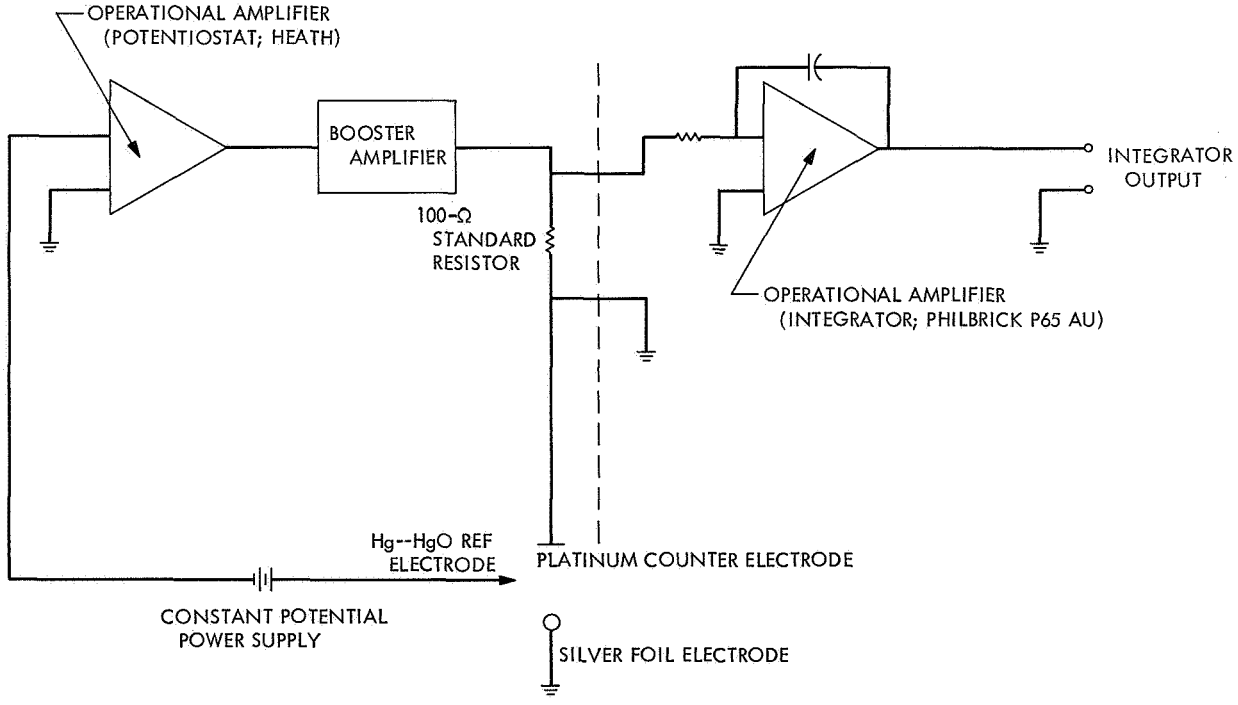


Fig. 17. Potentiostat and integrator circuit

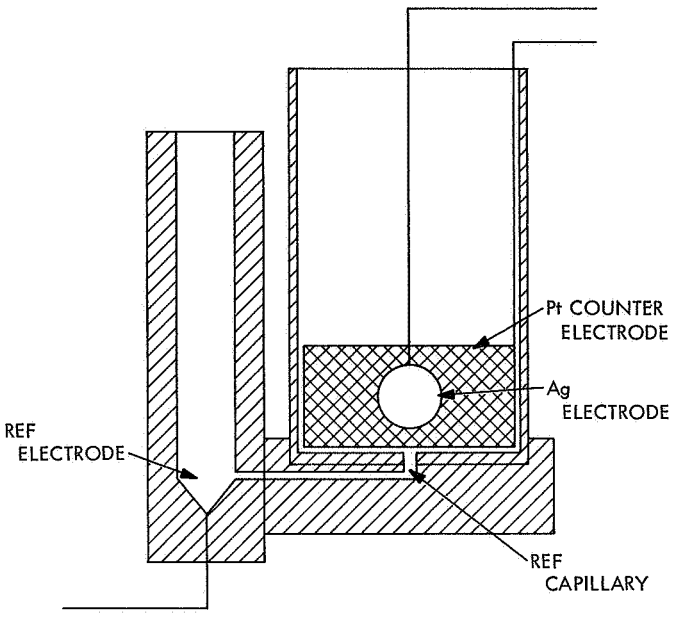


Fig. 18. Cross-sectional view of oxidation cell, showing location of working electrodes

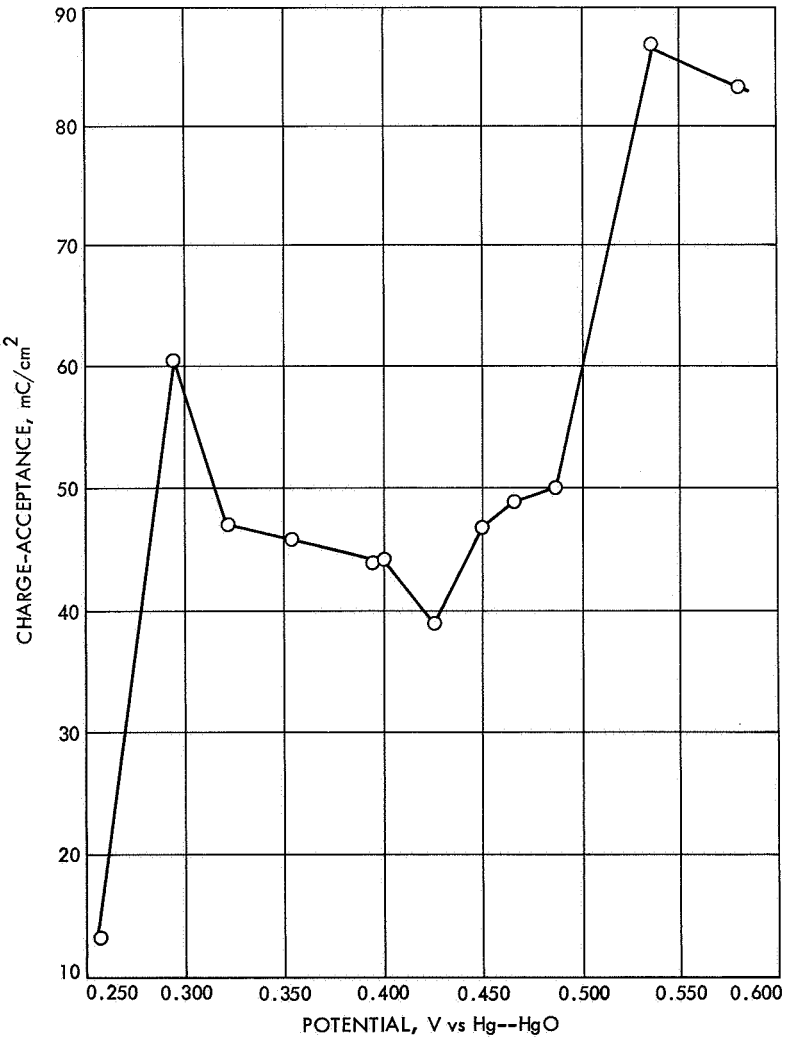


Fig. 19. Plot of charge-acceptance per unit area of smooth electrodes vs applied potential

$20.0 \pm 0.1°C$ during all oxidation runs. Standard electrodes were prepared by the vapor deposition of metallic silver on smooth glass discs.

3. Results and Conclusions

A plot of charge-acceptance per unit area of smooth standard electrodes versus applied potential is shown in Fig. 19. The purpose of this curve is to indicate acceptable ranges of potential for use in the constant potential estimation of surface area. For example, the region from 0.275 to 0.315 V versus the Hg–HgO reference is not acceptable; there are large changes in charge-acceptance coupled with small changes in applied potential. Conversely, the region from 0.325 to 0.400 V is desirable because of its flatness. Here, small errors in potentiostatic control cause only small variations in charge-acceptance.

Comparison runs were made on silver foil discs of two different geometric areas. Four runs on electrodes of each size were made utilizing the previously reported constant current method (SPS 37-39, Vol. IV), giving a reproducibility in surface area of $\pm 3\%$. Four runs were also made potentiostatically on electrodes of each size at an applied potential of 0.400 V versus Hg–HgO. The reproducibility in total charge passed was $\pm 7\%$. These results are in agreement within experimental accuracy.

The effective electrolytic surface areas of the small discs were then calculated, using Eq. (1). The electrolytic area of the large disc as determined by the constant current method was taken as the area of the standard electrode for use in the calculation. The results of the calculations are given in Table 4. The method apparently is a very promising one for the electrolytic determination of the effective electrode surface area.

Table 4. Comparison of foil electrodes in constant current and potentiostatic surface area estimations

Parameter	Foil electrodes cleaned by electropolishing	
	2.53 cm^2	0.688 cm^2
Effective electrolytic surface area at constant current	2.84 $cm^2 \pm 3\%$	0.775 $cm^2 \pm 3\%$
Total charge passed in oxidation at 0.400 V vs Hg–HgO	100 mC $\pm 7\%$	27 mC $\pm 7\%$
Effective electrolytic surface area at constant potential	2.84 cm^2†	0.770 $cm^2 \pm 7\%$‡

†Assumed to be the same as the constant current value. That is, this electrode was used as the standard in this comparison.

‡Value calculated by using Eq. (1).

Reference

1. Plumb, R. C., *J. Electrochem. Soc.*, Vol. 105, p. 502, 1958.

H. X-ray Radiography of *Mariner*-Type Battery Cells, *S. Krause*

1. Introduction

A series of x-ray radiographic studies of *Mariner*-type battery cells has recently been completed at the U.S. Naval Ordnance Laboratory. These studies represent part of a continuing effort to improve the manufacturing processes used in fabricating flight batteries.

It is important to improve the uniformity of the cells in order to achieve higher in-flight reliability as well as more accurate interpretation of laboratory test data. The use of x-ray radiography as a quality-control tool can help to achieve this goal.

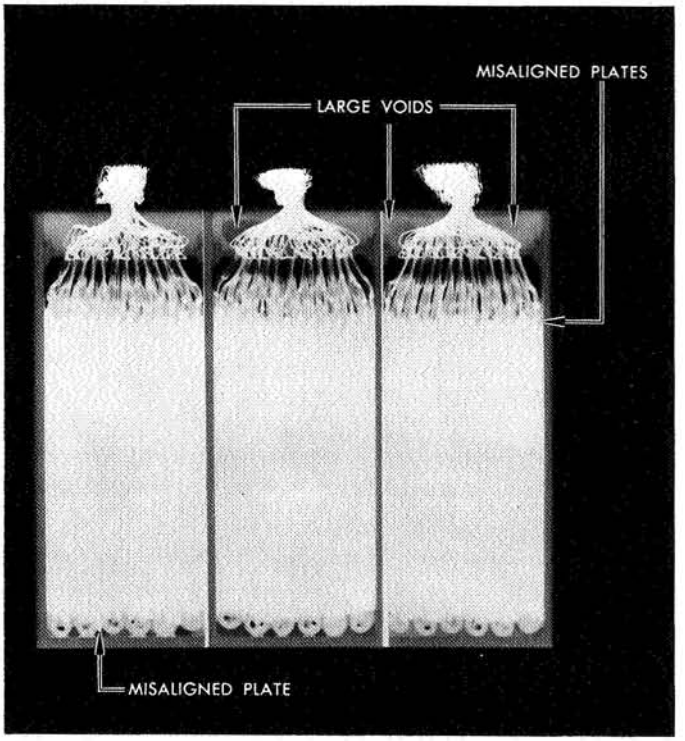


Fig. 20. X-ray side-view of a three-cell monoblock, showing large voids and misaligned plates

The two preliminary problem areas that were examined in the course of this effort were cell seals and plate alignment. Both can affect cell performance.

A 250 kV General Electric x-ray camera has been used to x-ray over 50 monoblocks (150 cells).

2. Cell Seals

Plate lead wires are positioned through a subcover and an upper cover. Between these two covers is a cavity which is filled with an opaque potting material to seal around the 25 lead wire bundles. Voids and defects in the areas around the wires allow cell leakage to occur more rapidly during environmental or cycling tests, and thus affect the test results adversely.

An example of the x-ray side-view of a three-cell monoblock may be seen in Fig. 20. This x-ray shows what appear to be voids in the upper cavity around the plate lead wires. In a sealed silver–zinc cell (the *Mariner* cell is this type) it is necessary to prevent electrolyte leakage resulting from the electrolyte "wicking up" the plate lead wires to the external environment. This condition results in degraded performance and ultimate failure. Encapsulation around the plate lead wires prevents such electrolyte leakage. Voids in the potting material around the plate lead wires substantially reduce the impeded leakage path to the outside of the cell. Variations in the size, number, and location of these voids could cause cell failures due to leakage that would occur at different rates and under different conditions. Defects of this nature could influence the evaluation of a design for a flight battery.

A monoblock was progressively dissected in a serial cross-section manner, parallel to the plane of the x-ray view seen in Fig. 20. The results of this sectioning showed the presence of large voids, as seen in one section in Fig. 21. The size, shape, and location of the voids were accurately predicted by the x-ray radiographs.

When the existence of voids of this nature was confirmed, a number of changes in the potting process were

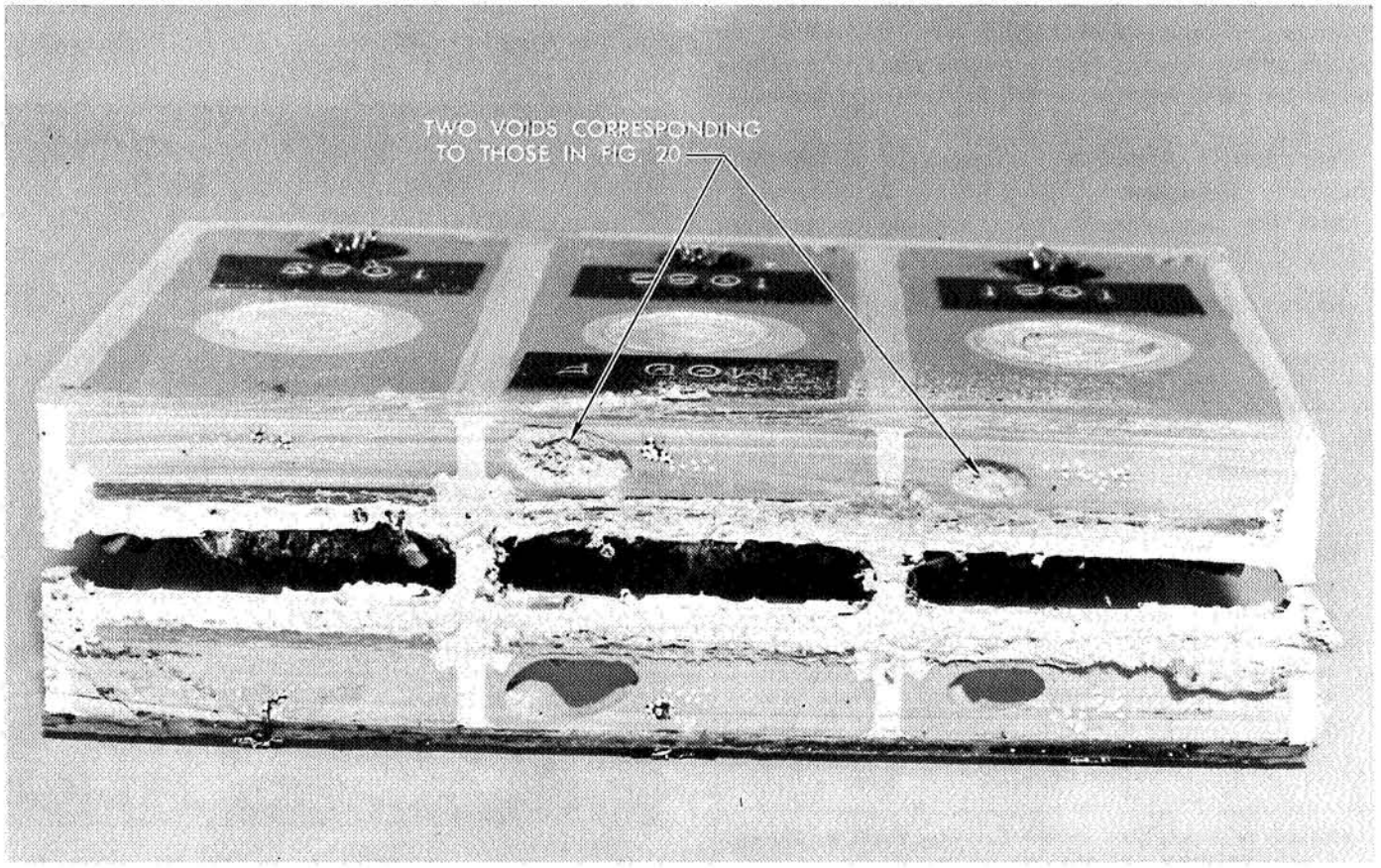


Fig. 21. Results of sectioning three-cell monoblock, showing large voids

instituted. Subsequent production and x-ray analysis of more cells resulted in the type of upper cavity potting shown in Fig. 22. The process changes virtually eliminated all voids, so that in all the newly fabricated cells, the seals in this area are as alike as possible.

3. Plate Alignment

Although the cell case is visually inspected after plate insertion, considerable plate-pack misalignment can result because of process variations and inspection limitations. Subsequent cycle data can be affected by misaligned plates.

The misalignment of cell plate packs can affect cell capacity if portions of the active electrode surface areas are not fully reacted. Figure 20 shows the three-cell monoblock with plates misaligned in the vertical direction. Note the large variation between the tops of some adjacent plates. It is quite possible that these cells would not perform as well as others with better plate alignment, particularly after several cycles. This condition would then cause considerable spread in capacity data and hinder the completion of an accurate design evaluation.

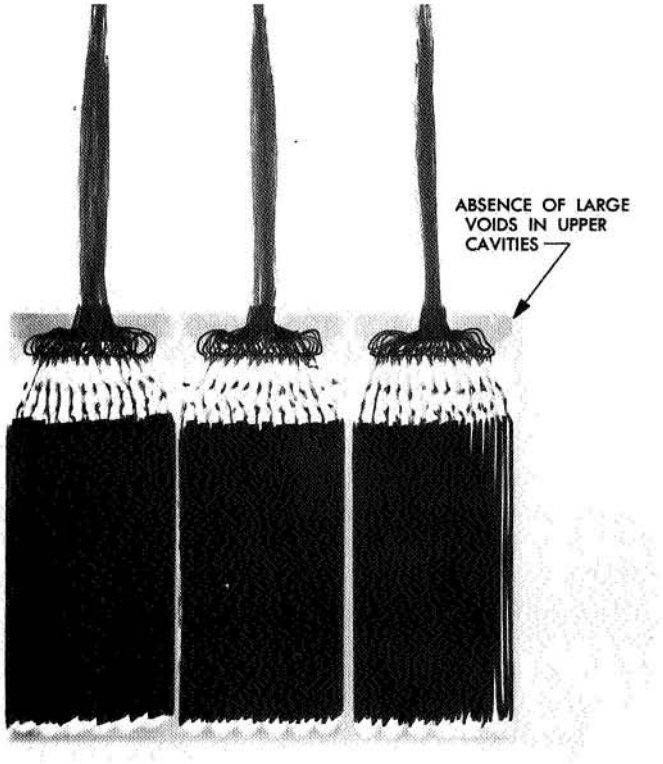


Fig. 22. X-ray side-view of a three-cell monoblock, showing absence of large voids after changes in potting process

4. Conclusions

Although at present in-process x-ray of plate alignment is not possible, future production procedures might be modified to include this type of inspection and allow a realignment step. Certainly, such a procedure would make the cells as alike as possible, allowing better comparison of test results. If the realignment step proves to be impractical, then at least the x-ray catalogue of every cell used in a test program would aid in pointing out those differences in performance or failure rates that were due to structural inconsistencies rather than to actual design deficiencies.

I. Calorimetric Measurements on the *Surveyor* Main Battery, *W. L. Long*

1. Introduction

The mission requirements of future JPL space programs may be expected to impose increasingly severe operating constraints on the spacecraft power system. A battery system designed for a long, sophisticated mission, such as a multiplanet probe, will necessarily be operating much closer to the performance limits than ever before. In addition, more complicated environmental and component compatibility problems arise.

Calorimetric data on spacecraft components are always required for proper thermal control. Heat generation by most electrical components is obtained from input–output efficiencies, and, for a constant power level, the heat generated is fairly constant. However, the heat generated by a silver–zinc battery varies with state of charge as well as with power level, and accurate calorimetric data will be necessary for future spacecraft design studies. Preliminary work in this area has been previously reported (Ref. 1). This work is being continued by Hughes Aircraft Co. under JPL contract.

2. Battery Description

The *Surveyor* main battery, as manufactured by ESB Co., is a 150 A-h, 14-cell, sealed silver–zinc battery. The cell cases are constructed of polystyrene; the battery case is magnesium. The total weight of the flight battery is 46.5 lb.

3. Measurement Technique

Calorimetric measurements have thus far been performed at discharge rates of 5 and 18 A. All measurements to date have been performed at 118°F. The calorimeter is isothermal, operating at a boiling point

of a liquid which surrounds the battery and fills the calorimeter. The liquid is heated by automatic precalibrated electric heaters, and thus maintained at the boiling point automatically. When heat is given off or absorbed by the battery, the change in power furnished to the heater is automatically recorded. When steady-state operation is reached, the change in power is the heat generated (or absorbed) by the battery. Provisions are included for maintaining constant pressure, constant rate of vaporization, and for the return of condensed vapor at a constant temperature. Freon F-113 was used for these tests.

4. Results

The results of the preliminary measurements show that the *Surveyor* battery produces heat equivalent to 5 W of power during a 5-A discharge. The battery produces heat equivalent to 80 W of power during an 18-A discharge. These results are based on steady-state operation. It is apparent that the heat evolution increases very sharply with an increase in discharge rate.

Further studies will extend the measurements to different discharge rates and temperatures, as well as define the effects of state of charge on heat generation by the battery.

Reference

1. Rowlette, J. J., "Heat Generation in the *Surveyor* Main Battery", Paper No. 47, The Electrochemical Society Fall Meeting, Oct. 1967.

J. Six-Converter Solar Thermionic Generator,

O. S. Merrill

1. Introduction

This is a summary of the work performed by Thermo Electron Corp., Waltham, Mass., under JPL contract during the period from Jan. 10, 1967 through Mar. 31, 1968 (Ref. 1). The work reported includes the design and fabrication of a six-converter solar thermionic generator designated as JG-4, and the design, fabrication, and performance testing of twelve identical converters, six of which were incorporated into the generator. The generator is to operate in a solar-energy concentrating system consisting of a parabolic mirror of 57-in. rim radius and a 69-in. focal length. The mirror generates a solar image in the form of a circular ellipsoid which at the focal plane of the mirror has a cross-sectional area of about 0.885 in.^{2} and an approximate maximum energy of 5000 W at 1 AU. The design of the six converters is similar to that of the series VIII converters (used in previous solar generators; SPS 37-40, Vol. IV, pp. 1–14) but has been modified to be compatible with a six-converter system. The converters have planar electrodes with a Re emitter and Mo collector. The emitter area is 2 cm^2. The converters are designed to operate at an emitter temperature of 2000°K at an interelectrode spacing of 2 mils.

2. Generator

A detailed review and evaluation was made of the original generator design.[1] This was necessitated by tests conducted at JPL subsequent to the original design which showed that solid Re emitters and Re sleeves were more reliable for extended operation than the Ta substrate pressure-bonded Re emitters and Ta sleeves proposed in the original design. However, incorporating solid Re emitters into the converters of the generator required redesign of the generator cavity because of the lower thermal conductivity of Re. The new design has the rear surfaces of the six emitters (Re) forming a cylindrical cavity of 0.61-in. radius and 0.658-in. length. The front opening of the cavity is to be placed at a distance of about 0.4 in. behind the mirror's focal plane, i.e., towards the sun. This produces optimum impingement and absorption of the solar energy on the cavity wall.

A tungsten cone with a 1-in. diameter opening protects the cavity walls from adverse effects caused by misalignment of the generator and the mirror. The rear of the cavity is formed by a highly reflective electropolished W surface in the form of an inverted, doubly truncated cone (back-piece), so designed as to direct reflected solar energy uniformly to the cavity wall (the emitters). This back-piece reflector is thermally isolated from the converters. The energy absorbed by the reflector is dissipated by a large Cr_2O_3-coated Mo radiator which is brazed to the W piece by a high-temperature braze. During solar operation, the W back-piece is designed to operate at a temperature less than 1200°C. With this solar image–cavity arrangement, which is expected to result in a near-optimum generator performance, approximately 4500 W of solar energy enter the cavity; 200 W are absorbed by the front piece; 2300 W are absorbed by the six Re emitters; 1400 W are absorbed by the W back-piece; the remaining 600 W are reflected and/or re-radiated and escape through the front opening of the cavity.

[1]JG-4, discussed in Thermo Electron Corp. Report TE 18-66, JPL Contract 951230.

Considerable effort was devoted to the generator assembly, which required the fabrication of special tools for aligning to critical tolerances the converters and the back-piece which form the cavity. Prior to constructing the Mo block, to which the various parts of the generator were mounted, an Al model was fabricated and checked for feasibility of the overall block design.

Other work performed in association with the JG-4 included the evaluation of the thermal transfer characteristics of the 0.65 Pd–0.35 Co braze selected for joining the W and Mo parts of the cavity back-piece. For this purpose, two samples, identical in geometry, were prepared and tested. One sample consisted of a W and a Mo disk joined together with the 0.65 Pd–0.35 Co braze; the other sample was a solid Mo disk. Both samples were tested under identical conditions, and comparison of the test results indicated that the rate of heat transfer in the W–braze–Mo sample was equal to or slightly higher than that measured in the all-Mo sample.

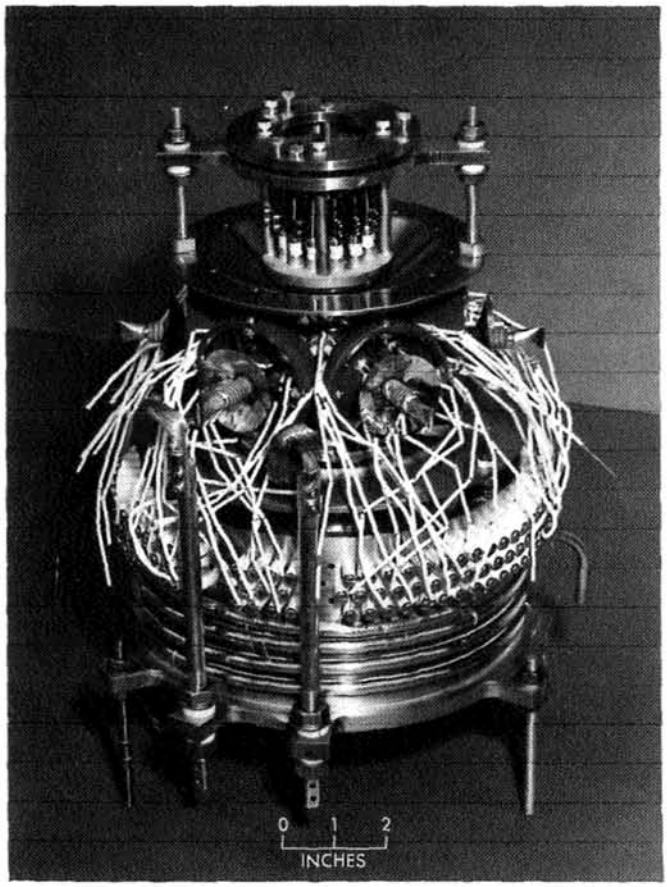


Fig. 23. Complete JG-4 generator with electron-bombardment unit

The completed generator (JG-4), with the electron bombardment unit attached for electrically heating the generator during laboratory testing, is shown in Fig. 23. Figure 24 shows a close-up of the generator cavity.

3. Converters

Twelve identical thermionic converters (Fig. 25) were fabricated and individually tested; six of these converters were incorporated into the generator. During the test of each converter, the output current was measured at different output voltages and at given emitter temperatures, with the cesium temperature optimized for maximum output. All twelve converters generated nearly identical data. The current–voltage data indicate an average power output of 37.5 W from each converter, or a total of 225 W from the six converters used. The same power output was obtained at a lower emitter temperature but also at a lower output voltage. The performance characteristics of the six converters used in the generator are shown in Fig. 26.

Considerable effort was expended in the preparation of the Re sleeves, which developed vacuum leaks during

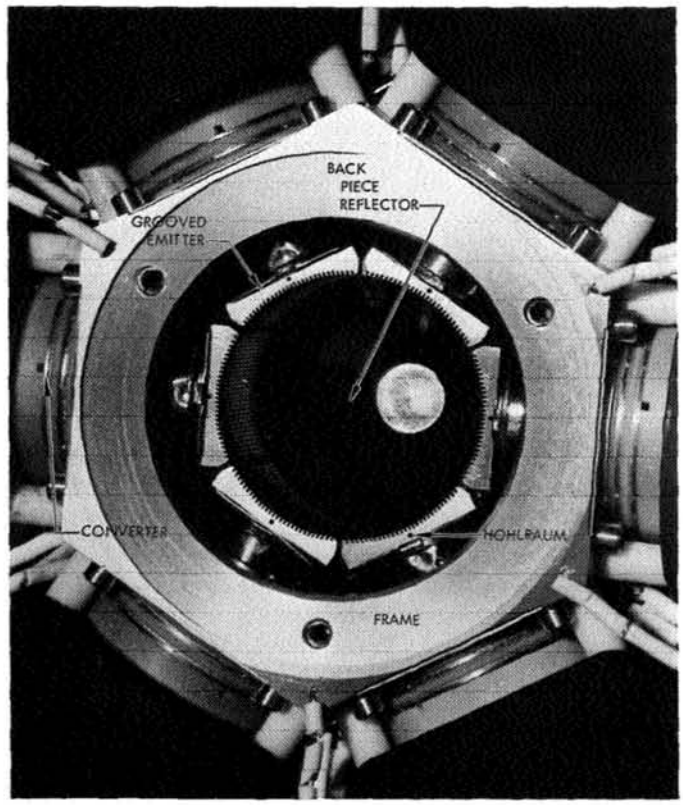


Fig. 24. JG-4 generator cavity

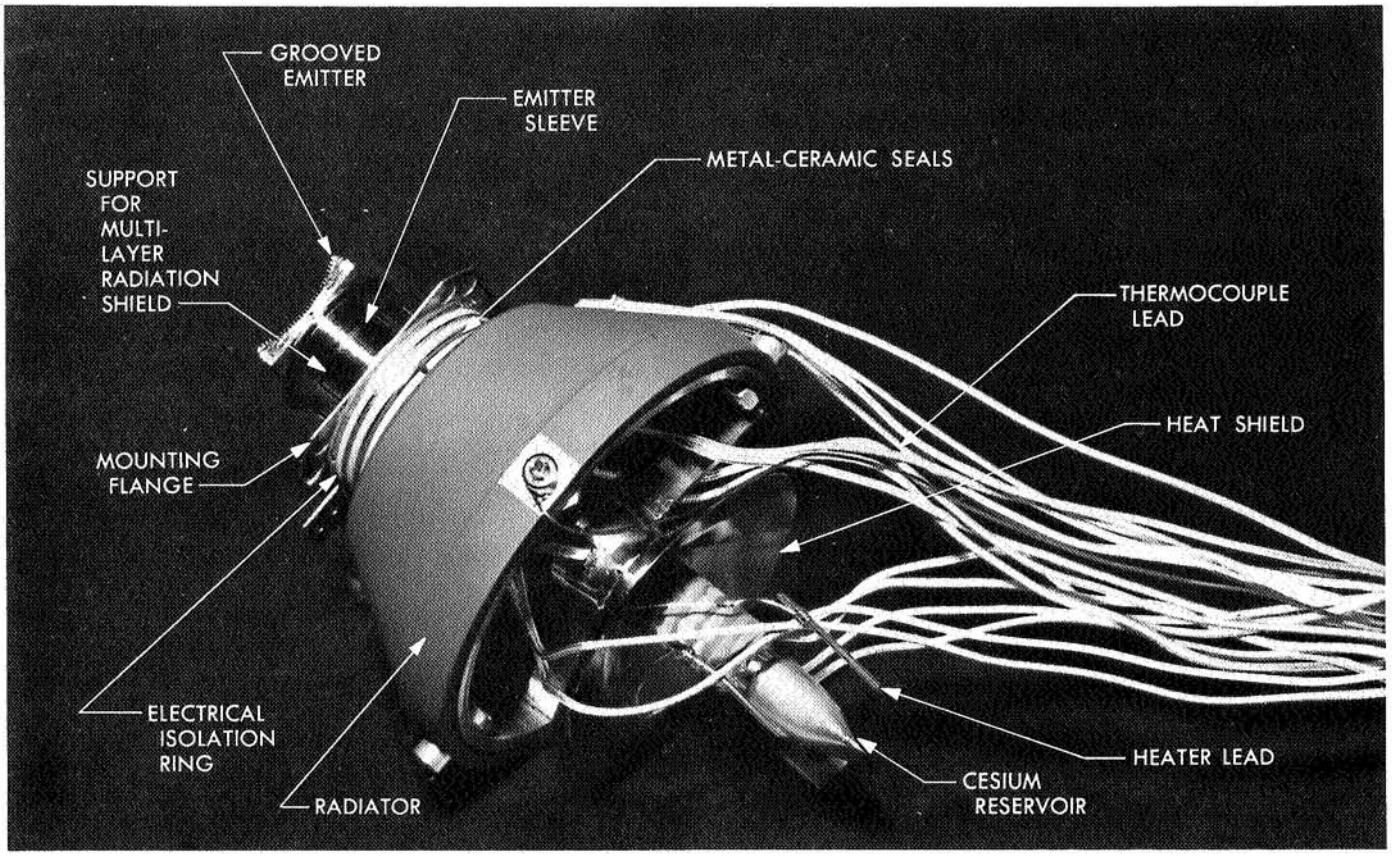


Fig. 25. Thermionic converter used in JG-4 generator

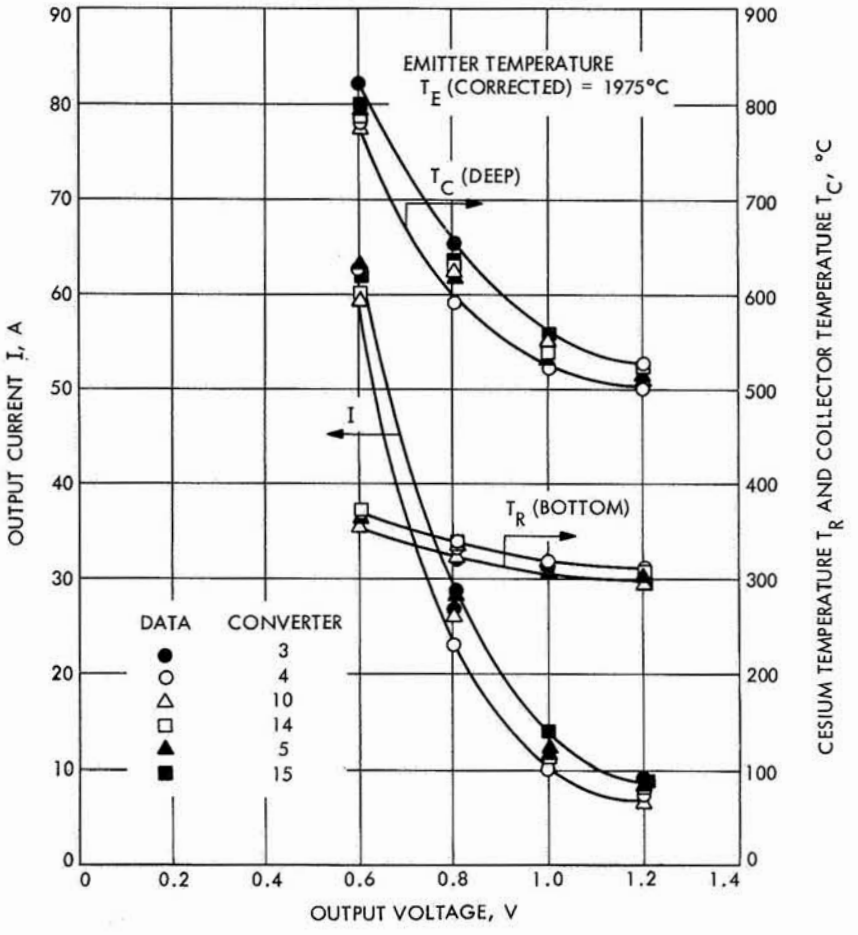


Fig. 26. Performance characteristics of six converters used in JG-4 generator

machining or thermal cycling. These leaks resulted from voids left in the seam of the Re tubing during the heliarc-welding process conducted by the vendor. This problem was eliminated after the Re tubing was purchased from the vendor in the "rolled only" state and the seam was electron-beam-welded by TECO. Substantial effort was also expended in the fabrication of the emitters which, due to their complex geometry and extremely close tolerances, required special preparatory techniques, particularly during electron discharge machining and subsequent processing.

4. Electron Bombardment Unit

For laboratory tests of the JG-4, an electron bombardment unit (Fig. 27) was fabricated and tested. This unit consists of six hairpin tungsten filaments arranged to form a cylindrical unit suitable for insertion into and heating of the generator cavity. The W filaments can be connected either in parallel, and controlled as a single unit, or individually, and controlled as six separate units. The unit was tested inside a cylindrical Mo block having approximately the same geometry as the generator cavity.

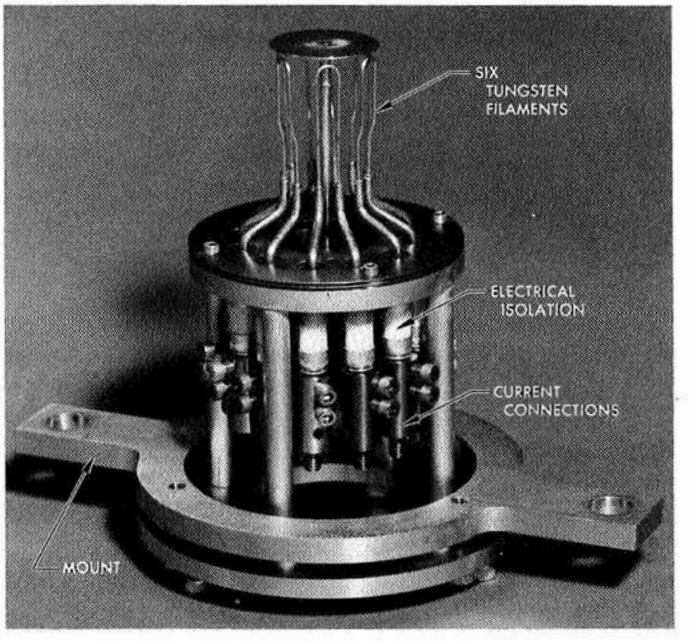


Fig. 27. Electron-bombardment gun used for laboratory test of JG-4 generator

The test results indicated that for a temperature of about 2000°K on the Mo surface facing the filaments, a total output power of 2200 W was required from the filament assembly.

Reference

1. Athanis, T., Shefsiek, P., and Lazaridis, L., *Final Report, Six-Converter Solar Thermionic Generator*, Report TE4073-146-68. JPL Contract 951770, Thermo Electron Corp., Waltham, Mass., June 1968.

K. Power Conversion Circuit Development,

D. J. Hopper

1. Introduction

The objective of the power conversion circuit development task is to develop advanced technology power conditioning components, i.e., regulators, inverters, battery chargers, etc., capable of meeting JPL advanced mission requirements.

The most recent activity on this task has been to develop a boost regulator with characteristics that are significantly better than the characteristics presently available in the *Mariner* spacecraft boost regulators. In particular, efficiency, transient response, and a reduced number of component parts were all goals of this regulator development program.

The work described below was performed under JPL contract by Wilorco, Inc., Long Beach, Calif.

2. Regulator Requirements and Description

The design requirement goals for the boost regulator are shown in Table 5. Three power levels were of interest: 100, 200, and 400 W. Figure 28 shows the functional block diagram for the regulator design that was developed.

When Q1 is turned on, the current in the left side of T1 will increase, and, due to transformer action, the voltage on the right side of T1 will also increase. When the voltage reaches a set level (56 V), the error amplifier signals the drive amplifier. The drive amplifier then turns Q1 off. The voltage out of T1 will then start to decrease. When the voltage decreases sufficiently, the error amplifier signals the drive amplifier to again turn Q1 on. The cycle then repeats. The input and output filters are inductance–capacitance type filters.

The error amplifier is a differential amplifier using a zener diode reference. As can be seen in Fig. 28, the

Table 5. Boost regulator design requirements

Parameter	Specification characteristic
Input	
Input voltage, V	25 to 50
Reverse polarity protection	Yes
Environmental	
Operating temperature range, °C	−10 to 75
Survival temperature range, °C	−55 to 145
Degraded performance range, °C	−40 to 100
Output	
Efficiency (min), %	90
Regulation (line, load, and temperature variation), %	±½
Output voltage, Vdc	56
Output ripple, mV rms	less than 50

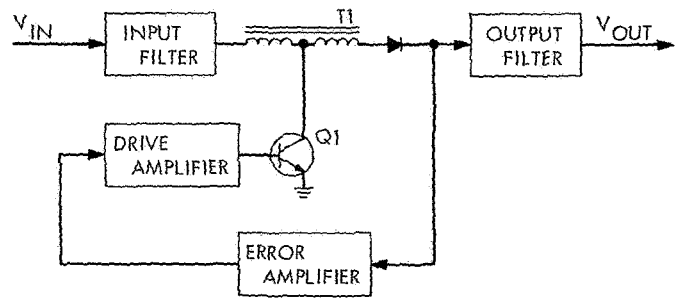


Fig. 28. Boost regulator block diagram

Q1, T1 combination only has to supply voltage in excess of the input voltage, $V_{out} - V_{in}$. This keeps the power switched by Q1 to a minimum. Since Q1 is switching relatively small quantities of power, the efficiency of this type of circuit can be quite high.

The performance characteristics of the 200-W regulator, which are typical of those obtained for the other power levels, are shown in Table 6 along with the characteristics of the *Mariner* Mars 1969 boost regulator. The *Mariner* Mars 1969 boost regulator uses magnetic amplifier control, while the developed boost regulator uses solid-state control. This means that the transient response of the developed boost regulator is better than that of the *Mariner* Mars 1969 boost regulator. There are 66 parts in the new regulator; the *Mariner* regulator has 74 parts.

3. Conclusion

Because of the increase in efficiency, regulation, and transient response and the reduction in parts, the newly

Table 6. Boost regulator performance characteristics

Parameter	Actual characteristics	
	New 200-W regulator	*Mariner* Mars 1969 regulator
Efficiency, %	93 to 97	87 to 93
Regulation, %	±0.5	±1
Voltage, V	56	56
Output power, W	200	250

developed boost regulator is a significant improvement over the existing *Mariner* design.

L. Electric Propulsion Power Conditioning,

E. Costogue

1. Introduction

The electric propulsion power conditioning effort is directed towards the design and procurement of hardware for two programs—SEPST II and III.[2]

The SEPST II program will utilize one breadboard power conditioner unit. This breadboard was built for the SERT II program[3] and is being modified to present power requirements. Hughes Aircraft Co. has been contracted to modify the unit and the test console which will be used to qualify the power conditioner.

The SEPST III program will utilize one breadboard and two experimental power conditioner units. Hughes Aircraft Co. has been contracted to design, develop, fabricate, and qualify-test the units.

[2]SEPST = solar electric propulsion system test.
[3]SERT = space electric rocket test.

2. Power Requirements and Characteristics

The power requirements for the power conditioner unit are shown in Table 7. There are two groups of power supply—Group I, low-voltage supply, and Group II, high-voltage supply.

Figure 29 shows the preliminary power conditioner unit block diagram, identifying the modules necessary to generate the voltages and currents required by the thrusters.

A line regulator is provided to generate 35-V regulated power from the 40- to 80-V line, which in turn drives the 5-kHz heater inverter. The output of the heater inverter is fed to the following modules: neutralizer heater, neutralizer keeper, vaporizer heater, and magnet.

The cathode heater power is supplied by a 5-kHz inverter. A standby inverter is provided to supply the power when a failure is detected. The output of the operating inverter is fed to the cathode filter module for filtering and control.

A master oscillator and phase shift module are utilized to provide base drive to all screen inverters. Phase shift is required to stagger the outputs of the screen supply inverters. The screen power is supplied by eight 12.5-kHz inverters. The output of the inverters is fed to the screen filter module for filtering and control.

The arc power is supplied by a 12.5-kHz inverter. A standby inverter is provided to supply power when a failure is detected. The output of the operating inverter is fed to the arc filter module for filtering and control.

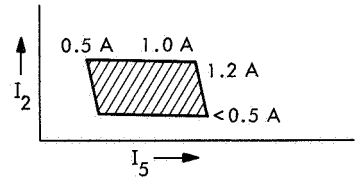
Similarly, the accelerator power is supplied by a 12.5-kHz inverter and a standby inverter. The output of the operating-accelerator inverter is fed to the accelerator filter module for filtering and control.

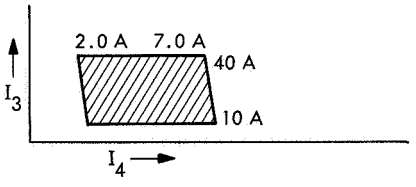
Table 7. Power conditioner unit requirements

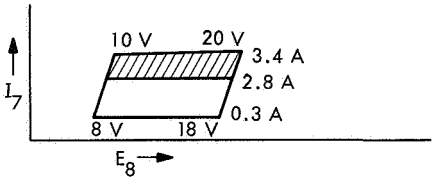
Group	Power supply	Type	Output	Maximum ratings			Nominal ratings					Range of control,[b] A	Frequency, kHz	
				Energy, V	Current, A	Current limit,[a] A	Energy, V	Current, A	Power, W	Regulator, %	Peak ripple, %		Output	Ripple
I	Magnet manifold heater	dc	Fixed	19	0.85	0.9	15	0.67	10.5	1.0 (current)	5	—	—	10
	Cathode heater	ac	Variable	5	40	45	4.5	35	160	Loop	—	10–40	5	—
	Neutralizer heater	ac	Variable	12	3.4	4	12	2.8	35	Loop	—	0.3–3.4	5	—
	Neutralizer keeper	dc	Fixed	300 V at 5 mA	0.02 at 30 V	0.55	10	0.50	5	1.0 (energy)	2 at 30 V 5 at 10 V	0.02–0.5	—	10
II	Vaporizer	ac	Variable	10	2	2.05	5.5	1.1	6	Loop	—	0.5–1.2	5	—
	Arc	dc	Variable	150 V[c] at 20 mA	7 at 36 V	8	34.5	6	210	1.0 (energy)	2	2–7	—	30
	Beam	dc	Variable	2050	1.0	1.05	2000	1.0	2000	1.0 (energy)	5	0.5–1.0	—	30
	Accelerator	dc	Variable	2050	0.1[d]	0.105	2000	0.01	20	1.0 (energy)	5 at 0.1 A	—	—	30

[a]Exact values to be specified by the manufacturer.

[b]Current varies as function of engine loop control:







[c]Starting characteristics: 150 V to 36 V at 20 mA.

[d]Current stays at this level for less than 10 min at very low repetition rate.

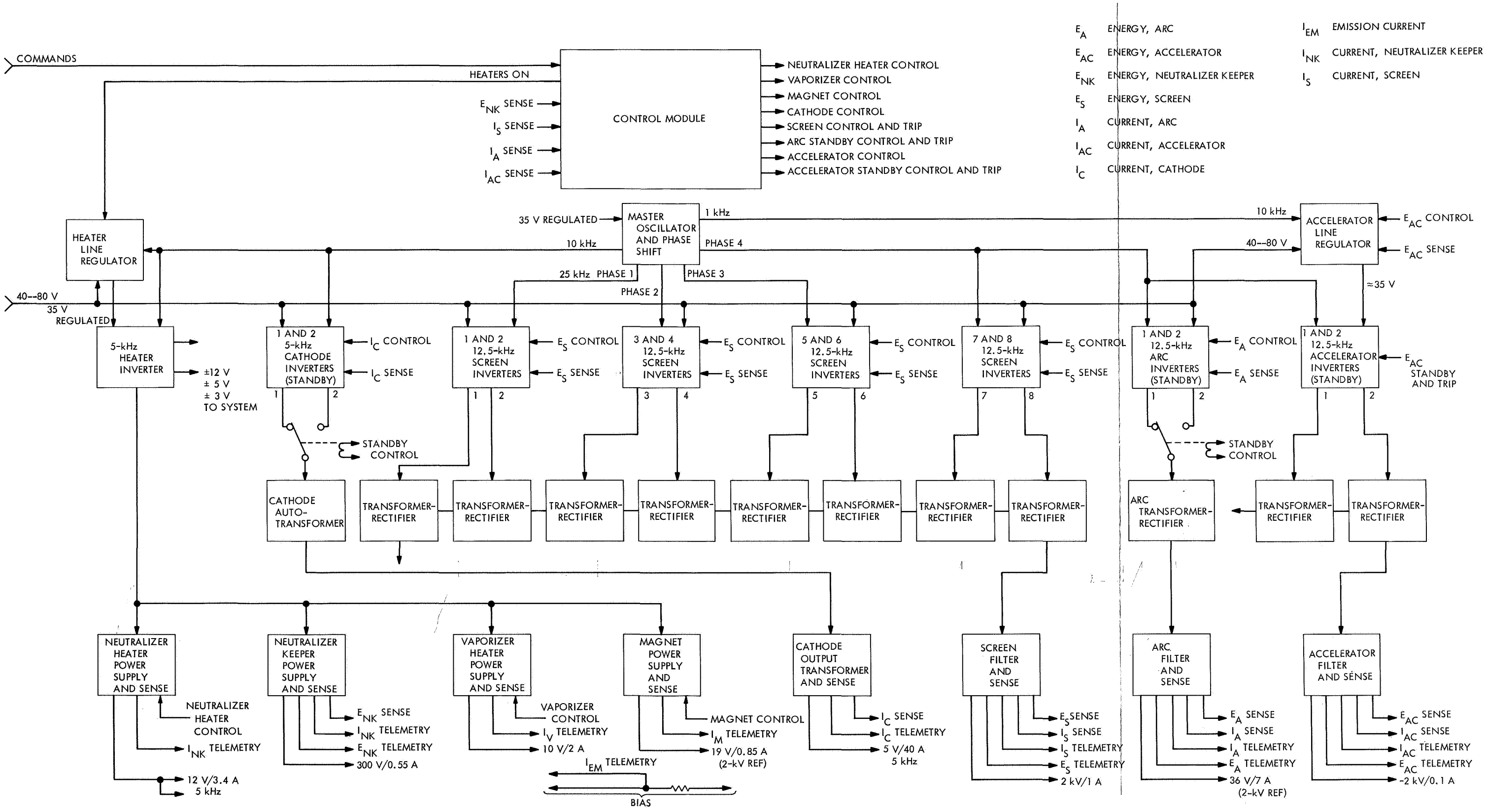


Fig. 29. Power conditioner unit block diagram

V. Spacecraft Control

GUIDANCE AND CONTROL DIVISION

A. Partial Inertial System Integration Test, *G. Paine*

1. Introduction

The LAB 1 Alert program was exercised during integration of the strapdown electrostatic aerospace navigator (SEAN) system from August 19 through 28, 1968. The system included the Alert computer,[1] the computer adapter and display (CAD), and the inertial measurement unit (IMU) with digital velocity meters (DVMs) but without the electrostatic gyros (ESGs). Tests were conducted to determine that: (1) the basic inertial integrations had been correctly mechanized, (2) data were being correctly received from the CAD, (3) the schemes for using altimeter data to damp the position and velocity errors were correct, and (4) the proper computations were being performed to convert from inertial to local coordinates. These tests were successfully completed after several minor errors were discovered and corrected. As the LAB 1 program[2] constitutes about one-fourth of the final flight navigation program, the successful completion of the tests signifies a major milestone in the development of a flight program.

[1]Manufactured by Honeywell, Inc., Minneapolis, Minn.

[2]Paine, G., *Preliminary SEAN Navigation Program (LAB 1)—Request for Programming*, Mar. 25, 1968; Markiewicz, B. R., *SEAN Navigation Equations for the Alert Computer*, Apr. 30, 1968 (JPL internal documents).

The short length of time required for the integration proves the value of the complete system simulations performed in generating the equations for LAB 1 and the value of program checkout employing a computer simulator. These simulation efforts have been described previously in SPS 37-52, Vol. III, pp. 52–55.

2. Test Setup

No attempt was made to align the IMU with the local geodetic vertical. Instead, the DVM biases were set to zero in the computer program and the IMU was tilted until the level accelerometers were indicating less than 50-μg output (less than 1 pulse/45 s), thus aligning the IMU to astronomic vertical, offset by the real DVM biases and the residual DVM outputs.

To replace the gyros, LAB 1 computes a transformation matrix based on time and geodetic latitude to convert between geodetic and inertial coordinates. Consequently, the outputs from the three DVMs are processed as though the IMU were aligned to geodetic vertical and there were no local gravitational anomalies.

3. Test Results

The results of two of the test runs, a 52.266-min test without altimeter damping and a 26.133-min test with altimeter damping, are presented in Table 1. A 14-min

Table 1. LAB 1 test results of observed position, computed equivalent bias, and measured equivalent bias errors

Time, min	Cross-range error			Down-range error			Altitude error	
	Observed position, ft	Computed equivalent bias, μg	Measured equivalent bias, μg	Observed position, ft	Computed equivalent bias, μg	Measured equivalent bias, μg	Observed position, ft	Computed equivalent bias, μg
52.266-min run (without altitude control)								
11.2	−249	−36	−22	236	34	44	−403	−49
46.67	−1800	−43	−22	2875	48	44	−29000	−40
52.27	−1750	−48	−22	3689	24	44	−50000	−39
26.133-min run (with altitude control)[a]								
26.13	−1459	−51	−22	1344	47	44	−60	—
26.13[b]	−1078	−38	−22	1307	44	44	−174	—

[a]CV = 1/32, CR = 1/16, corrections every 14 s, dominant time constant 15.3 min.

[b]Based on 52.266-min run with corrections for addition of altitude control.

test was also run with altimeter damping and a large vertical accelerometer bias. The residual errors (the difference between expected and observed) were small. The data in each case were analyzed to provide equivalent DVM bias errors. Gain constants *CR* and *CV* were used to control the position and velocity feedback during altitude control.

These biases are actually the combination of IMU misalignments with DVM biases, and do not reflect the accuracy of DVM calibration. The misalignments could have been reduced, but this was not done in order to provide more data on system performance. In addition, the measured equivalent values of the down-range error (DRE) and cross-range error (CRE) were obtained by examining the level DVM outputs directly. No such estimate was obtained from the vertical (ALT) DVM. The computed equivalent bias shifts of ± 10 μg seen in the data can be attributed to a wide variety of sources (computation errors, real bias shifts, etc.). Since the magnitude is so small, this shift is considered secondary.

The 52.266-min run data show good agreement between the measured values of the DVM bias and the values needed to produce the position errors. The difference between the value of level acceleration observed on the DVM outputs directly and that needed to produce the CRE is probably caused by an attitude misalignment in the body-to-inertial transformation matrix. The offset of about 22 μg is equivalent to an attitude error of only 5 arc sec. The value of CRE decreases after the 48-min value because it is almost a pure Schuler error. The vertical DVM bias remained relatively constant throughout the test period. The large altitude error (ALTE) of 50,000 ft is to be expected since the altitude channel is unstable and no altimeter damping was employed.

The 26.133-min run with altimeter damping followed the conclusion of the previous run. The observed CRE did not match, as well as desired, the measured value based on the 52.266-min run when corrected for the addition of altitude control. However, there was excellent agreement between the measured and computed values of CRE, DRE, and ALTE. At the end of the test period, the ALTE was drifting upwards so that, if data had been taken over an extended period, still better agreement could have been expected.

The 14-min run with altimeter damping (CV = 1/16, CR = 1/8, corrections every 14 s) was made with a large vertical DVM bias introduced (1140 μg) to make the comparison easier between the observed results and those predicted from theory. The IMU was not carefully leveled at this point, so no comparisons of DRE or CRE could be made. There was an excellent comparison between the altitude error observed (805 ft) and those predicted (848 ft). In addition, the predicted dominant time constant (304 s) was in good agreement with the observed value (270 s).

B. Automatic Lens Design Program, *L. F. Schmidt*

The objective of this study is to further develop an automatic lens design program into a practical design tool, which will be flexible enough to handle most optical designs and yet not be cumbersome for designing simple optical systems.

The lens design program was rewritten into Fortran IV language and submitted to the COSMIC Computer Center[3] in March 1968. (A detailed explanation of the program is provided in a three-volume document.) Prior to its submission to COSMIC, 10 organizations were supplied with copies of the program. This increased use of the program is possible since several different computer systems can handle the Fortran version. A new computer system is being planned that is not compatible with the machine language version of the program; however, no problems are anticipated in converting the Fortran version to the new computer.

The Fortran version contains a newly developed option that provides a graphical plot of the optical system configuration. This eliminates the necessity for a manually drawn optical system to determine if a design is progressing in a satisfactory manner or needs new input data to direct its progress.

Several improvements to the code have been undertaken since it was submitted to COSMIC.

An improved method of handling a curved image plane has been devised. The general technique used is to shift the position of the flat image plane for each object and corresponding image height. This position is calculated to cause the image plane to intersect the desired curve at the expected image height. In the old method, the image plane shift was calculated manually and entered. As the image height changed during design operations, the image plane shifts were updated in a trial-and-error fashion throughout the design phase.

In the new method, the image plane shift is automatically recalculated by the code after each design iteration. In this way the time necessary for manual calculation is saved as well as the extra computer time that was expended during the trial-and-error process of adjusting the image plane.

An option has been developed that will generate a punched card output for use on another program (PAGOS), which examines optical systems in terms of modulation transfer function (MTF). This provides a fast method of MTF analysis for image-forming optical systems.

A method of controlling the contrast of the image has been developed. Weights can be entered that place more importance on the central image rays, compared to the outer ones. The program assigns weights for each ray in a gaussian fashion when this option is exercised.

These program modifications are currently being incorporated into the three-volume document.

An attempt was made to develop a method of controlling the ratio of spot sizes relative to each object height. This was discontinued when the initial efforts were unsuccessful and it was concluded that such an option was not worth the effort required to get it operating properly.

C. Vibration and Shock Analysis: Strapdown Electrostatic Aerospace Navigator, *G. T. Starks*

The purpose of the vibration and shock analysis was to (1) determine the g level that could be transmitted through various combinations of natural frequency and damping ratios (ζ) of a mechanical isolation system, which would be detrimental to the survival of the electrostatic gyroscope (ESG), and (2) determine the minimum parameters that would isolate the destructive g input.

Part of the strapdown electrostatic aerospace navigator (SEAN) system consists of an inertial measuring unit (IMU) housing the two ESGs, three accelerometers, and their associated electronic circuits in a rigid body structure weighing approximately 88 lb.

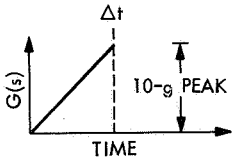
The g capability of the ESG to be used in the first developmental navigational test is 4 g, of which 1 g is allotted to the static g environment. Exceeding the 3-g capability above the 1-g static environment of the ESG will result in destruction of the gyroscope.

The environment specified by MIL STD 810A that the ESG is expected to survive is:

Shock, 10-g terminal sawtooth	0.011 s
Vibration:	
Military aircraft	
0.10-in. double amplitude (DA)	5–14 Hz
1 g	14–23 Hz
0.036-in. DA	23–74 Hz
10 g	74–500 Hz
Van (truck), smooth roads	
1.00-in. DA	0–5 Hz
1.3 g	5–500 Hz

[3]Computer Software Management and Information Center, Computer Center, University of Georgia, Athens, Georgia.

The transfer function for a shock input of a 10-g, 0.011-s terminal sawtooth driving function



is

$$G(s) = \frac{10g}{\Delta t}\left(\frac{1}{s^2} - \frac{e^{-\Delta t s}}{s^2} - \frac{\Delta t\, e^{-\Delta t s}}{s}\right) \quad (1)$$

where the duration of input force $\Delta t = 0.011$ s.

The response of the IMU (Fig. 1) is

$$M\ddot{X} + K(\dot{X} - \dot{Y}) + D(\dot{X} - \dot{Y}) = \text{forcing function } F(s) \quad (2)$$

where

D = damping of isolator

K = equivalent spring constant of isolator

M = W/g of IMU

with

M = mass of IMU

W = weight of IMU

The transfer function describing the response of the IMU in g to a shock input is

$$G(s)_{IMU} = \frac{10g}{\Delta t}\left(\frac{1 + \frac{2\zeta}{\omega_n}s}{\frac{s^2}{\omega_n^2} + \frac{2\zeta s}{\omega_n} + 1}\right)\left(\frac{1}{s^2} - \frac{e^{-\Delta t s}}{s^2} - \frac{\Delta t\, e^{-\Delta t s}}{s}\right) \quad (3)$$

The inverse of $G(s)$ is

$$G(t)_{IMU} = \frac{10g}{\Delta t}\left(\left[\Delta t + \frac{\sin}{e^{\alpha\Delta t}\beta}\left(\beta t + 2\tan^{-1}\frac{\beta}{\alpha} + \tan^{-1}\frac{\beta}{a_0 - \alpha}\right)\right]U(t)\right.$$
$$- \left\{t - \Delta t + \frac{\sin}{\beta e^{\alpha(t-\Delta t)}}\left[\beta(t-\Delta t) + 2\tan^{-1}\frac{\beta}{\alpha} + \tan^{-1}\frac{\beta}{a_0 - \alpha}\right]\right\}U(t - \Delta t)$$
$$\left. - \Delta t\left\{1 - \frac{\omega_n \sin}{\beta e^{\alpha(t-\Delta t)}}\left[\beta(t-\Delta t) + \tan^{-1}\frac{\beta}{\alpha} + \tan^{-1}\frac{\beta}{a_0 - \alpha}\right]\right\}U(t - \Delta t)\right) \quad (4)$$

where

$U(t)$ = unit step function equal to unity to the right of $t = 0$ and is zero to the left of $t = 0$

$U(t - \Delta t)$ = unit step function equal to zero to the left of $t = \Delta t$ and is unity to the right of $t = \Delta t$

$\Delta t = 0.011$ s

ω_n = natural frequency of isolator design

$\beta = (\omega_n^2 - \zeta^2\omega_n^2)^{1/2}$

$a_0 = \frac{\omega_n}{2\zeta}$

$\alpha = \zeta\,\omega_n$

The values for various natural frequency isolators given in Table 2 were obtained using Eq. (4) at damping ratios 0.3 and 0.2.

The gain or transmissibility of a second-order system to a sinusoidal input forcing function (configuration shown in Fig. 1) is

$$\text{Transmissibility} = \frac{\left[1 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}}{\left[\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\zeta\frac{\omega}{\omega_n}\right)^2\right]^{1/2}} \quad (5)$$

Equation (5) was used to obtain the transmissibility factors presented in Table 3 for damping ratios equal to 0.2 and 0.3.

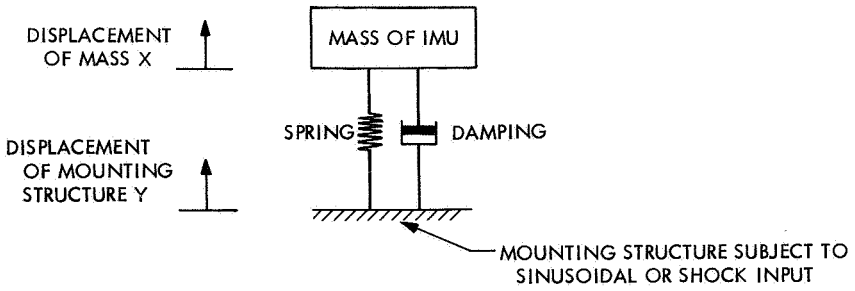


Fig. 1. Shock mount spring mass configuration

The terminal or peak acceleration in g for a sinusoidal input can be computed from

$$\frac{a_n}{g} = \frac{r\omega^2}{12g} \tag{6}$$

where

a_n = peak acceleration, ft/s²

r = ½ double amplitude of driving function, ft

ω = angular frequency, rad

Table 4 lists the values of acceleration using Eq. (6). The computations were based on the environment specified for the military aircraft portion of MIL STD 810A. Using data from Tables 3 and 4, the values of acceleration transmitted to the IMU are shown in Table 5 for some representative frequencies.

Assuming a simultaneous input from both shock (Table 2) and sinusoidal frequency acceleration (Table 5), the total possible peak accelerations that are transmitted to the IMU are given in Table 6 for both 6- and 8-Hz shock mounts.

Table 2. Acceleration response of IMU to 10-g input

Frequency, Hz	Damping ratio	Acceleration level at 0.011 s, g	Peak acceleration, g
18	0.3	4.64	4.95
12	0.3	3.1	3.4
10	0.3	2.47	2.8
	0.2	1.94	2.8
8	0.3	1.93	2.24
	0.2	1.47	2.24
6	0.3	1.44	1.7
	0.2	0.95	1.7

Table 3. Frequency transmissibility values for damping ratios 0.3 and 0.2

Frequency, Hz	Transmissibility factor	
	Damping ratio 0.3	Damping ratio 0.2
8	1.995 (peak)	2.734 (peak)
10	1.333	1.486
20	0.330	0.265
30	0.186	0.137
40	0.131	0.093
50	0.1015	0.071
60	0.0832	0.0572
74	0.0666	0.0453

Table 4. Acceleration values versus frequency input

Frequency, Hz	Peak acceleration, g
6	0.184
8	0.304
14	1.000
30	1.6535
40	2.9395
50	4.5930
60	6.6139
70	9.000
74	10.000

Table 5. Acceleration levels transmitted to IMU[a]

Frequency, Hz	Acceleration response of IMU, g
8	0.832
30	0.23
40	0.273
50	0.324
60	0.378
74	0.453

[a]Shock mount of $\zeta = 0.2$, natural frequency = 8 Hz.

Table 6. Possible total peak acceleration transmitted through 6- and 8-Hz shock mounts

Frequency, Hz	Peak acceleration, g			
	6-Hz shock mount		8-Hz shock mount	
	$\zeta = 0.2$	$\zeta = 0.3$	$\zeta = 0.2$	$\zeta = 0.3$
6	2.2	2.1	—	—
8	—	—	3.07	2.85
74	2.14	2.4	2.69	2.71

The 6-Hz shock mount would provide about 0.6- to 0.8-g isolation to the IMU for the expected aircraft environment and combined 10-g shock. This is shown graphically in Fig. 2, where transmissibility is decreasing prior to the 14-Hz, 1-g point.

However, in reference to the expected van environment as shown in Fig. 2, the 1.3-g environment extends back to 5 Hz inside of the peak transmissibility point. At resonance the 6-Hz ($\zeta = 0.3$) shock mount could transmit approximately 2.6 g independent of the shock spectrum. An approximately 4-Hz isolator would be required to reduce the expected van environment to a level compatible with a shock environment such that the total expected input does not exceed 3 g.

Typical response curves of the IMU using an 8-Hz isolator to a 10-g shock input are shown in Fig. 3.

A van to be used in performing navigational tests has been procured for the SEAN system. The van has been instrumented and operated over terrain that is expected to be encountered during the navigational tests.

Data are being reduced to determine the criteria for the best design to isolate the IMU.

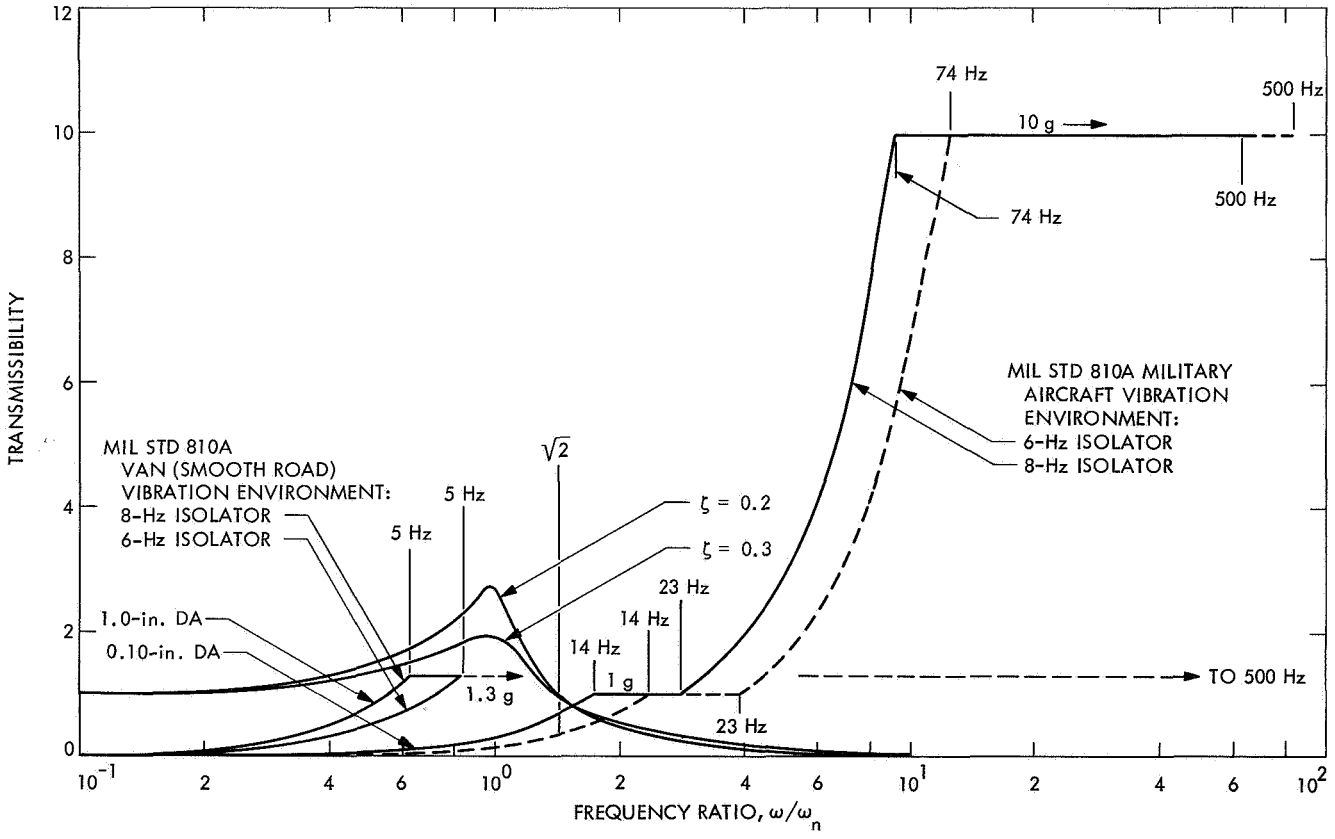


Fig. 2. 6- and 8-Hz transmissibility plots related to expected environments

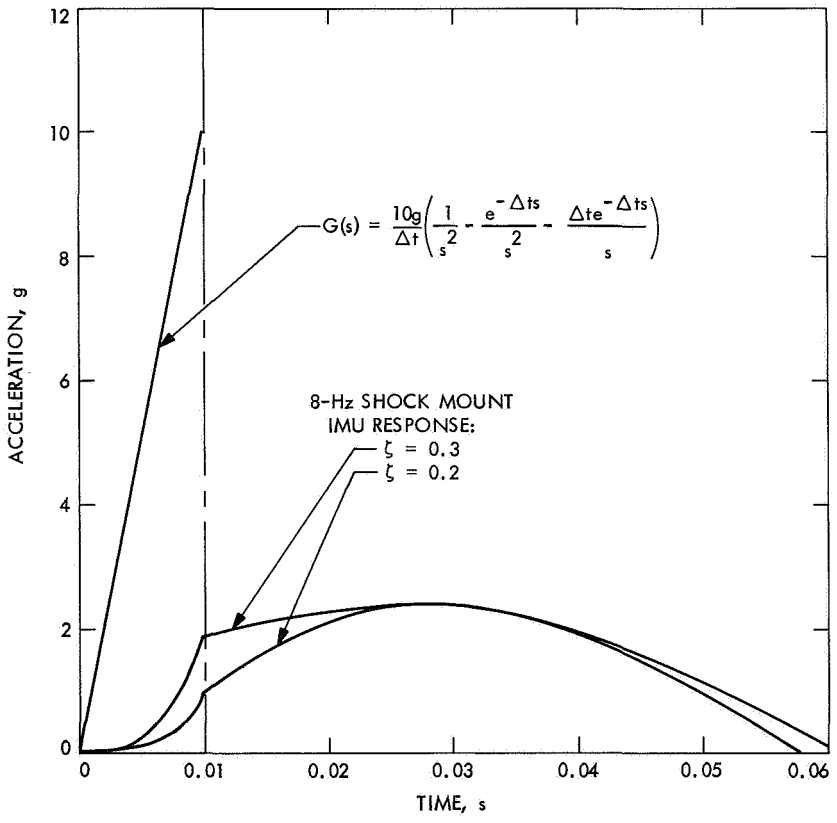


Fig. 3. Typical response curves of IMU using an 8-Hz isolator to a shock input

VI. Guidance and Control Research

GUIDANCE AND CONTROL DIVISION

A. Effective Work Function of Metal Contacts to Vacuum-Cleaved Photoconducting CdS,

R. J. Stirn

1. Introduction

In a previous article (SPS 37-51, Vol. III, pp. 78–82), photoconductive gains greater than unity have been reported for CdS, even when one of the metal contacts on the sample has a work function ϕ_m greater than the electron affinity E_A of CdS, i.e., when the contact should have blocking characteristics. Similar results have now been found at JPL for metals deposited on crystals of photoconducting CdS that have been cleaved in a vacuum, thus eliminating possible interfacial effects. Such results imply that some mechanism is operating that allows transport of electrons through the contact in quantities, *above* those expected for thermionic emission over the barrier, at least for barrier heights for these same metals measured on more conducting CdS (Refs. 1 and 2). Thus, the barriers appear to be *lower* on photoconducting CdS than those measured in Ref. 2. However, electron transport *through* the barrier by tunneling via trap levels within the forbidden gap has not been completely ruled out.

The electron concentration at the boundary of the metal contact to vacuum-cleaved surfaces of photoconducting CdS is also being determined at various temperatures and light intensities by an analysis of stationary high-field domains in the range of negative differential conductivity. The preliminary data yield barrier heights that are lower than those published in Refs. 1 and 2, and are relatively insensitive to the metal; in addition, their values are strongly dependent on temperature and moderately dependent on light intensity. None of these phenomena is observed in nonphotoconducting CdS.

In the following, some preliminary data obtained at JPL from the high-field domain analysis are presented, as well as current published values of barrier heights for CdS with various metal contacts obtained by other techniques.

2. Effective Barrier Heights in Photoconducting CdS

In CdS crystals with an N-shaped negative differential conductivity range (due to field-enhanced freeing of holes from traps and, hence, enhanced recombination), stationary high-field domains attached to the cathode are

observed above a critical applied voltage (Ref. 3). Above that voltage, the current saturates and the steplike domains increase in width linearly with voltage, eventually filling up the entire crystal. Such domains can be conveniently observed by the Franz–Keldysh effect using band-gap light, because of the increased optical absorption in the region of higher field. Details of the theory and experimental setup will be presented in a future SPS article.

A boundary-value carrier concentration n_{II} can be determined uniquely from the values of the saturation current, the mobility, and the electric field within the domain. The latter value is obtained from the domain width–voltage relationship and crystal length. This boundary value n_{II} can be shown to be nearly equal to the carrier concentration at the interface n_c. Given this value of n_c, which is found to vary with temperature and light intensity, an effective barrier height can be derived from the expression

$$n_c = N_c \exp\left[-\phi_B/(kT)\right] \tag{1}$$

where

$$N_c = 2\,(2\pi m^* kT/h^2)^{3/2}$$

$$= 2.3 \times 10^{18}\,(T/300)^{3/2}$$

is the effective density of states for CdS. The barrier height at the surface is denoted by ϕ_B. Actually, Eq. (1) is strictly valid only for thermodynamic equilibrium. Thus, in a photoconductor where generation of electrons and holes is taking place, this relationship (Eq. 1), is not exact. However, it is phenomenologically meaningful in that it helps in comparing different metal contacts.

The values of n_{II} (assumed to be equal to n_c) obtained from the analysis on vacuum-cleaved crystals of CdS for a photon flux density of $5 \times 10^{15}/\text{cm}^2/\text{s}$ ranged from 10^8 to $10^{10}/\text{cm}^3$. (The method of analysis will be discussed in a future SPS article.) The values of the saturation current density are typically 10^{-2} to 10^{-3} A/cm^2 and the electric field within the domain varies between 3.0×10^4 and 1.20×10^5 V/cm, depending on the metal and the light intensity. The electron mobility was assumed to be constant up to 4.0×10^4 V/cm and then drop as the inverse of the electric field above that value.[1]

[1]Böer, K. W., and Bogus, K., "Electron Mobility of CdS at High Electric Fields," to be published in *Phys. Rev.*

Calculated values of the effective barrier height are shown in Table 1 for the above photon flux density. The domains are very hard to see at the higher temperatures because of the decreased Franz–Keldysh effect. It can be seen that there is little difference between the metals, unlike those reported in Ref. 2 (Table 2). Also, there is a marked temperature dependence.

The light intensity dependence of ϕ_B for several samples is shown in Table 3. The light used was monochromatic and ranged from 512 to 490 nm, depending on the temperature (band gap). The range of intensity (flux

Table 1. Temperature dependence of the effective barrier height measured on photoconducting CdS

Temperature, °K	Effective barrier height, eV					
	Gold	Platinum	Silver	Nickel	Copper	Tin
155	0.30	0.27	—	—	0.27	0.26
180	0.35	0.30	0.34	0.31	0.32	0.29
220	0.42	0.36	—	0.37	—	0.35
255	0.47	—	0.47	—	0.44	—
295	0.53	0.50	0.53	—	—	—

Table 2. Summary of barrier heights on CdS (taken from Ref. 2)

Metal	Barrier height from photoresponse, eV	Barrier height from capacitance data, eV	Metal work function, eV
Platinum	0.85 ± 0.03	0.86 ± 0.02	5.0
Gold	0.78 ± 0.03	0.80 ± 0.05	4.7 (5.2)
Silver	0.56 ± 0.02	0.58 ± 0.03	4.4
Nickel	0.45	—	4.7
Copper	0.36	0.35 ± 0.03	4.5
Aluminum	Ohmic contact	—	4.2

Table 3. Light dependence of the effective barrier height measured on photoconducting CdS

Relative light intensity, %	Effective barrier height, eV		
	Gold (155° K)	Nickel (180° K)	Silver (295° K)
100.0	0.28	0.29	0.48
40.0	0.29	0.30	0.51
20.0	0.30	0.31	0.53
10.0	0.30	0.31	0.54
4.0	0.31	0.32	0.55
2.0	0.32	0.33	0.57
1.0	0.33	0.34	—

density) was varied by 100 with the highest value[2] $I_0 = 5 \times 10^{16}/\text{cm}^2/\text{s}$.

One can see that the effective barrier heights increase with decreasing light intensity. An interesting point is that at room temperature the values of ϕ_B (which are about the same for all six metals) tend to a zero light-intensity value, which is about equal to the energy of the bulk Fermi level ϕ_n (in the dark) as measured from the conduction-band edge. The Fermi level in the bulk of the crystal is obtained from the dark resistivity ρ_0 and the expressions

$$\left.\begin{aligned} n &= N_c \exp\left[-\phi_n/(kT)\right] \\ \rho_0 &= (q\, n\, \mu)^{-1} \end{aligned}\right\} \qquad (2)$$

where n is the bulk free-carrier concentration and q is the electronic charge. The mobility μ can be assumed to follow a T^{-2} law quite accurately between 100 and 300°K (Ref. 4), and to have a value of 300 cm²/V-s at 300°K. Then,

$$\begin{aligned} \phi_n &= kT \ln\left[N_c\,(q\,\mu\,\rho_0)\right] \\ &= kT \ln\left[2.5 \times 10^{18} \left(\frac{T}{300}\right)^{3/2} q \cdot 300 \left(\frac{T}{300}\right)^{-2} \rho_0\right] \\ &= kT\left[4.75 + \tfrac{1}{2} \ln\left(\frac{300}{T}\right) + \ln \rho_0\right] \end{aligned} \qquad (3)$$

A room temperature value of about 0.60 eV was obtained for the crystals used in this report, and a value of 0.34 eV at 155°K.

It would be very desirable to measure the barrier heights by the technique of high-field domains on crystals having different resistivities, i.e., Fermi-level positions, and look for any possible correlation. However, it has not been possible up to this time to dope the CdS crystals in order to get substantially different resistivities and still obtain the required stationary domains.[3]

3. Photoconductive Gains

As noted in the *Introduction*, gains greater than unity are observed in these crystals even though the published barrier heights for the metal–CdS (see *Subsection 4*) are of such magnitude that the contact should be completely blocking. The concept of gain in a photoconductor was reviewed in the previous article (SPS 37-51, Vol. III). The gain factor, expressed as the ratio of the saturation current to the photon flux density absorbed times the charge of an electron, has been measured to be at least 4 or 5 in these samples. The actual gain at the contact (in the domain) is at least 10 times higher because a stable domain can be formed with less than a tenth of the sample length (~ 0.5 to 1.0 mm). It is this length (volume) which should be considered in the light absorption since it is the domain region that is controlling the gain factor. The contribution of the rest of the crystal only reduces the gain factor because of the presence of recombination. In fact, the measured gains are lower than what might be because of considerable thermal quenching in these samples.[4] For this reason, no formal table of gain factors versus metal contact or temperature, for example, is given. The bulk characteristics and sample geometry considerably limit the gain possible with these contacts.

4. Published Values of Barrier Heights on Semiconducting CdS

A review of the three types of measurements used for barrier-height studies (photoresponse, diode forward characteristic, and differential capacitance) was given in SPS 37-51, Vol. III, as well as a description of barrier parameters. A large number of vacuum-cleaved semiconductors with various metal contacts were investigated by Spitzer and Mead (Refs. 1 and 2). Of a total of 14 different semiconductor materials of the diamond group IV or zincblende III–V types, with band gap covering a factor of 40 in energy, all but 3 (InAs, InP, and GaSb) exhibit a barrier height nearly equal to one-third of the band gap as measured from the valence-band edge. This value was nearly independent of the metal work function, which indicates that surface states play a major role in these materials in determining the barrier height. Another exception (in the wurzite II–VI class) was CdS, where the barrier height did depend in some way on the metal work function. The barrier heights found in their samples of CdS, which had carrier densities ranging from 10^{15} to 10^{17}, are listed in Table 2. Data from both photoresponse and capacitance–voltage measurements are given. The second figure in each column (±) gives the variation found from sample to sample. Substantial difference can be seen between the values in Tables 2 and 3.

[2]This value of photon flux density drops slightly to $4 \times 10^{16}/\text{cm}^2/\text{s}$ at the lowest temperatures because of the shift in band gap (wavelength) with temperature and the nonlinear monochromator output.

[3]The crystals used in this investigation, purchased from the University of Delaware, are slightly doped with silver and aluminum.

[4]Thermal quenching in a photoconductor is the reduction of photocurrent because of increased recombination due to thermal releasing of holes from traps. A similar effect is seen for infrared quenching.

Many other investigators have published data on barrier heights of metal contacts evaporated or plated onto CdS thin films or crystals, which have been exposed to air or etched. The values are higher for some metals and lower for others. These values and results obtained at JPL on air-cleaved crystals of CdS, using the analysis presented in *Subsections 2* and *3*, will be discussed in a future SPS article.

The particular values of ϕ_B obtained for lower resistivity CdS from photoresponse measurements, for example, are not seriously in question here. However, the preliminary results of this work show that the interpretation of barrier heights on photoconductors, if indeed one can define a barrier height in this nonequilibrium situation, needs to be thoroughly investigated. In fact, much more work is needed in the case of the relatively few semiconducting materials that show differences in contact properties from metal to metal. The simple relationship

$$\phi_B = \phi_m - E_A \tag{4}$$

which is then used in these cases, can give misleading results. This is due to the fact that ϕ_B is the difference between two numbers nearly equal in magnitude. The value of the metal work function ϕ_m, in turn, is an average of many determinations made by different experimenters whose results have been strongly influenced by the surface purity of the metal. In addition, metal thin films probably have a different work function than bulk crystals. Most serious is the apparent fact that the value of ϕ_m for thin films depends strongly on the substrate which is used. This effect has been demonstrated for gold in an elegant experiment (Ref. 5), where the values of ϕ_m varied from 5.08 to 5.40 to 5.59 eV, when the substrate was changed from CdTe to polished stainless steel to CdS, respectively. In addition, these values are well above the commonly used values of 4.7 eV. Good agreement with the higher values has been reported recently (5.2 eV) by two other investigators (Refs. 6 and 7), who demonstrated that mercury contamination was the cause of the lower value of 4.7 eV. Thus, the values of ϕ_B in Table 2 are *not* in order of ϕ_m, especially if the higher value of 5.6 eV for ϕ_m is used for gold on CdS. (Also note the position of nickel.)

It should be pointed out that Swank (Ref. 5) did obtain a barrier height of 0.80 eV for gold on *semiconducting* CdS in agreement with Ref. 2. This required a revision of the value of the electron affinity for CdS. The new value reported by Swank is 4.8 eV, as compared to the previously accepted value of 3.9–4.0 eV.

5. Conclusions

The results of *Subsections 2* and *3*, if the analysis using high-field domains is correct, show that the concepts of contact barriers on photoconductors are not useful when relatively high currents are flowing. For the latter situation, it appears that the barrier height between the metal and CdS is effectively lowered, allowing much larger gain factors than would normally be expected.

Additional investigations on different metals contacted to CdS, with varied amounts of doping, would be desirable. Extension of the measurements to light intensities far lower than those required to actually see the domain is also needed in order to observe the effect of decreasing current densities.

References

1. Spitzer, W. G., and Mead, C. A., "Barrier Height Studies on Metal-Semiconductor Systems," *J. Appl. Phys.*, Vol. 34, p. 3061, 1963.
2. Mead, C. A., and Spitzer, W. G., "Fermi Level Position at Metal-Semiconductor Interfaces," *Phys. Rev.*, Vol. 134, p. A713, 1964.
3. Böer, K. W., and Voss, P., "Stationary High-Field Domains in the Range of Negative Differential Conductivity in CdS Single Crystals," *Phys. Rev,*. Vol. 171, p. 899, 1968.
4. *Physics and Chemistry of II-VI Compounds*, p. 581. Edited by M. Aven and J. S. Prener. North-Holland Publishing Co., Amsterdam, 1967.
5. Swank, R. K., "Surface Properties of II-VI Compounds," *Phys. Rev.*, Vol. 153, p. 844, 1967.
6. Huber, E. E., "The Effect of Mercury Contamination on the Work Function of Gold," *Appl. Phys. Lett.*, Vol. 8, p. 169, 1966.
7. Riviere, J. C., "The Work Function of Gold," *Appl. Phys. Lett.*, Vol. 8, p. 172, 1966.

B. Preliminary Results From Switching Experiments on MnBi Films, *G. W. Lewicki and J. E. Guisinger*

1. Introduction

The role of Curie-point switching in a proposed high-density magneto-optic memory utilizing MnBi films has been described in SPS 37-42, Vol. IV, pp. 59–61; theoretical aspects of this switching have been considered in SPS 37-46, Vol. IV, pp. 84–87. This article describes the experimental apparatus used for normal and Curie-point-switching experiments, and discusses the significance of some preliminary results with respect to memory and recording applications.

2. Experimental Apparatus

A block diagram of the apparatus used in the switching experiments is given in Fig. 1. For visual observation of domain structure, an MnBi film is situated within an electromagnet with hollow pole pieces, illuminated with light from a xenon arc lamp passed through polarizer 1, and viewed through polarizer 2. Polarizer 2 can be set so that areas having opposite average magnetizations appear as light and dark regions. This can be done because the rotation of plane-polarized light passing through the film is proportional to the average magnetization within the film. The transmission by the film–polarizer 2 combination of plane-polarized light represents a measure of the average magnetization within the film. The electromagnet allows the generation of a magnetic field perpendicular to the plane of the film. The strength of this field is monitored by a Hall probe.

For the purpose of electronically measuring average magnetization, a small amount of laser radiation (0.1 mW) is allowed to pass through the Pockel's cell, polarizer 1, the film, and polarizer 2, and is collected by a photodetector. The output of the photodetector is used as a measure of the average magnetization within the film.

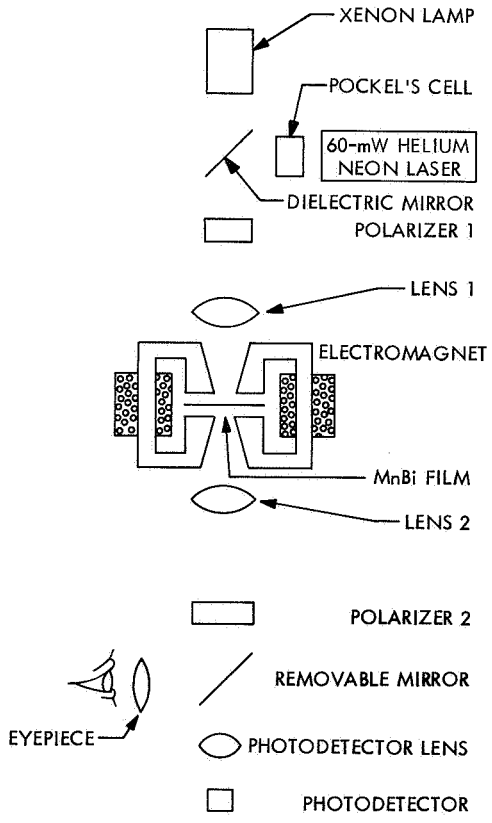


Fig. 1. Block diagram of experimental apparatus used for switching experiments

For Curie-point switching, a high-voltage pulse is fed into the Pockel's cell to allow some 15 mW of laser radiation to be focused onto a 2-μm spot on the film for approximately 1 μs. This amount of incident optical energy is sufficient to heat a 1-μm spot on the film past its Curie temperature. The spots, upon cooling, acquire an average magnetization dependent on the value of magnetic field applied during the switch.

3. Normal Switching Characteristics of 700-Å MnBi Films

Normal switching refers to changing the average magnetization within a film with an applied magnetic field without heating the film by the laser beam. Typical major and minor loop plots of the average magnetization as a function of applied field are shown in Fig. 2.

The average magnetization versus applied-field loop that is generated by a varying field, which does not reverse itself until the film becomes magnetically saturated, is called a major loop. A minor loop is generated by a varying field that reverses itself before the film is magnetically saturated.

For a major loop, a very high field ($H > 3.5$ kOe) is applied so that the film becomes magnetically saturated (the film is completely magnetized in the direction of the applied field with an average magnetization equal to the saturation magnetization). When the field is reduced to zero value, the film remains magnetically saturated. Only when the field is taken to a negative value coercive force ($-H_c$) does the average magnetization respond by suddenly dropping. This drop corresponds to the sudden appearance of domain structure or areas of opposite magnetization having the appearance of tree branches with doughnuts for leaves. This structure is shown in Fig. 3a; the size of the doughnuts is on the order of 5 to 10 μm. As the field is taken to still larger negative values, more and more of the tree-like structure appears suddenly (Fig. 3b) to give the film a negative average magnetization. As the field is taken to very large negative values, the film becomes magnetically saturated with its magnetization in the direction of the applied field. When the field is taken back to very high positive values, a similar process occurs.

The salient feature of the experimentally observed major loop is that 700-Å films have a large coercive force H_c ($H_c \approx 1$ kOe). Consequently, once a film is saturated,

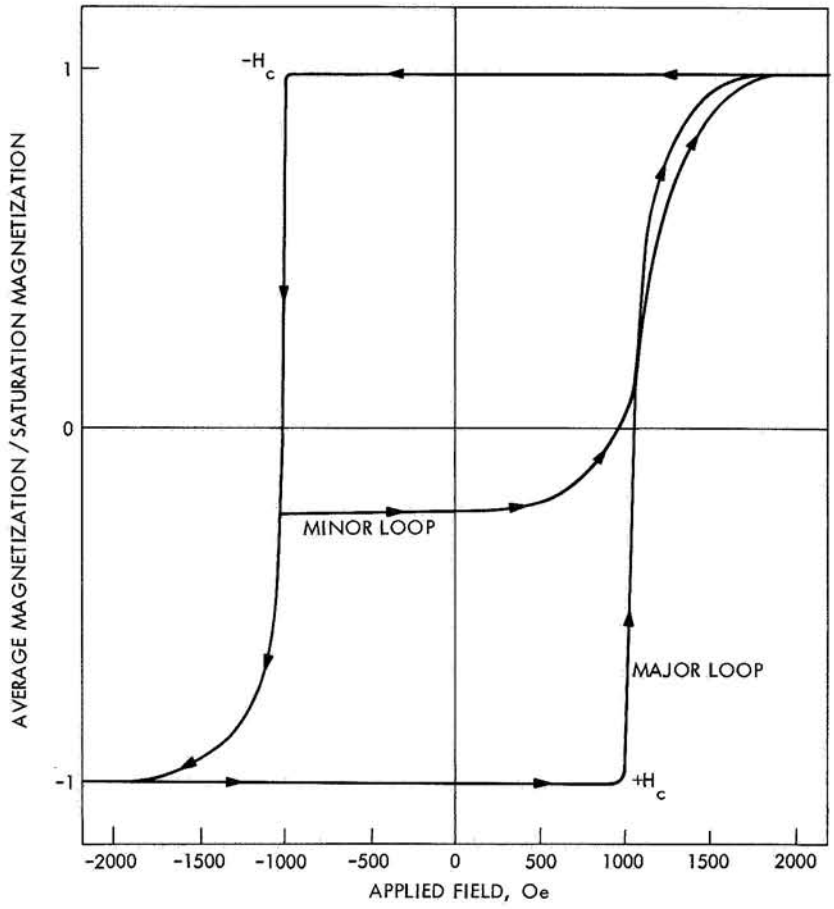


Fig. 2. Major and minor loops of the ratio of average magnetization to saturation magnetization as a function of applied field for 700-A MnBi film

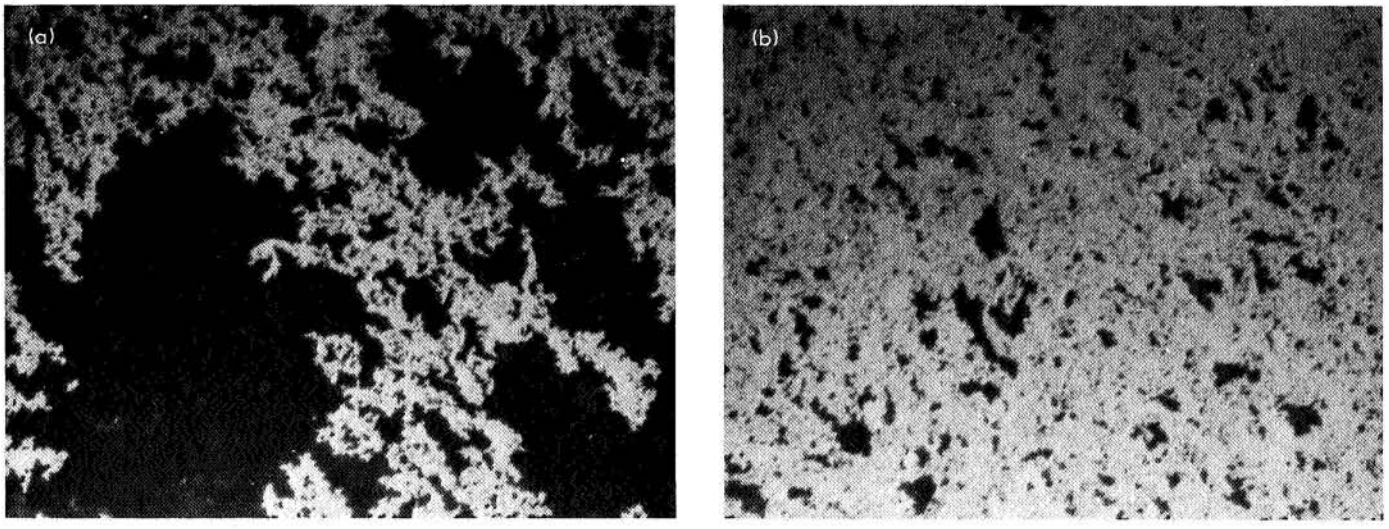


Fig. 3. Domain structure during normal switching of MnBi film: (a) partially switched, (b) further switched

large magnetic fields $H < H_c$ with directions opposite to that of the magnetization do not affect the film. During Curie-point switching of saturated films, fields opposed to the magnetization $H < H_c$ can be applied to influence the areas being Curie-point-switched without affecting other areas of the film. It has been observed that the coercive force decreases as the film thickness increases, becoming zero at a thickness of approximately 2500 Å. Lack of absolute reproducibility of coercive force from film to film has prevented the measurement of this dependence.

An interesting phenomenon was discovered upon visual observation of films being taken through minor loops such as the one shown in Fig. 2. Starting with the domain structure shown in Fig. 3a, when the field was varied in the direction to make the structure disappear, rather than make more of it appear (see the minor loop shown in Fig. 2), the structure did not suddenly disappear at a critical field as suddenly as it had appeared. Instead, it slowly turned from white to gray to black until the film was again magnetically saturated. Even with optics of high numerical aperture, the boundaries between light and dark regions could not be observed to move.

This observation indicates that the tree-like structures are not areas completely magnetized in one direction, but blocks of oppositely magnetized domains smaller than the wavelength of light, the preponderance of fine domains magnetized in one direction over fine domains magnetized in the other direction determining the shade of gray of the block. Thus, a 1-μm discrete spot of 700-Å-thick MnBi film will probably not be a single domain, and, thus, will not be a magnetically bistable element for Curie-point switching, as discussed in SPS 37-46, Vol. IV.

4. Curie-Point Switching Characteristics of 700-Å MnBi Films

Small areas of a magnetically saturated MnBi film (surface dimension on the order of 1 μm) were Curie-point-switched with different values of magnetic field applied during the switch. The average value of the magnetization within the areas following the Curie-point switching was sensed with a photodetector. A plot of the resulting average magnetization as a function of the field applied during the switch is shown as a solid line in Fig. 4. The offset of the curve is explained by the fact that an area being Curie-point-switched is acted upon not only by the applied field but also by the demagnetizing field of the surrounding film (SPS 37-42, Vol. IV). The dotted line represents the normal switching characteristic of the film.

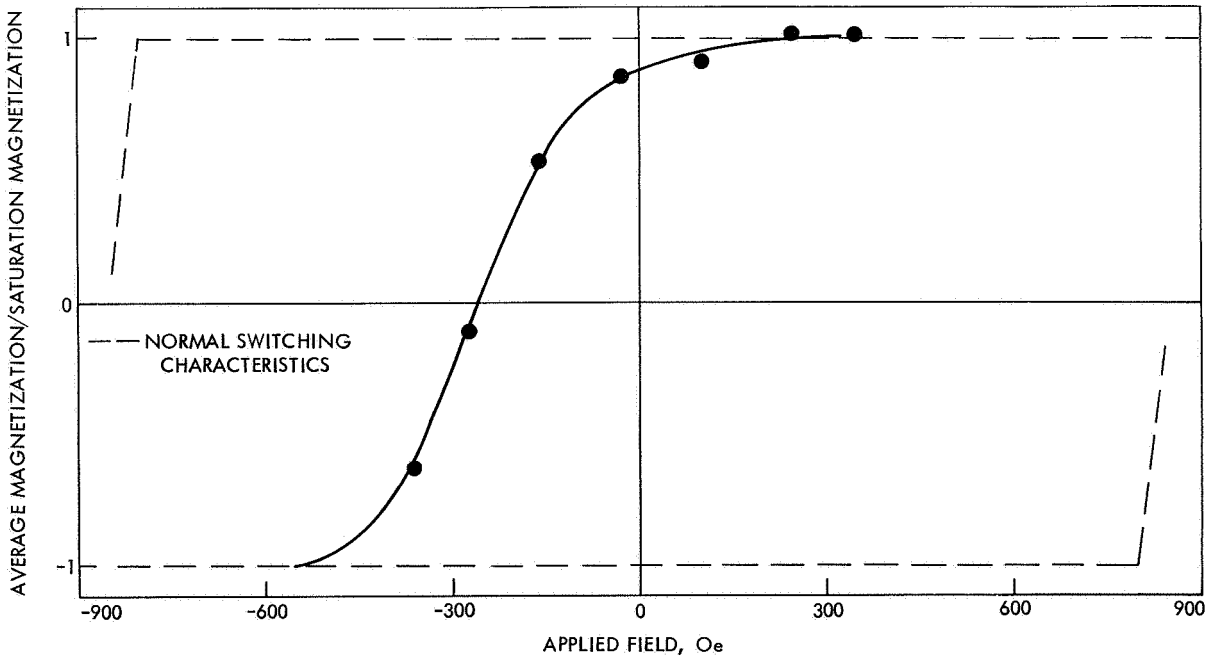


Fig. 4. Ratio of average magnetization to saturation magnetization of Curie-point-switched 1-μm spot on magnetically saturated MnBi film as a function of field applied during switching

The salient feature of this result is that the switching is not binary. Any value of average magnetization can be attained simply by varying the value of applied field during Curie-point switching. The result reaffirms the view that the basic domain size within a film 700 Å thick is much smaller than the wavelength of light.

Although not completely verified experimentally, it appears that there does exist a wide range of applied fields (-350 Oe $< H < -150$ Oe) in Fig. 4, which does not affect the average magnetization within an area that has been Curie-point-switched. This phenomenon is important for any recording application utilizing Curie-point switching in MnBi films.

5. Memory and Recording Applications

Application of MnBi films to the magneto-optic mass-memory scheme (SPS 37-42, Vol. IV) requires discrete areas of film that can be switched with reasonable values of applied field, i.e., no larger than tens of oersteds. The result shown in Fig. 4 implies that some 175 Oe above that required to compensate for demagnetizing fields would be required to switch such an area.

Theoretical considerations given in SPS 37-46, Vol. IV, suggest that a spot heated past its Curie temperature is completely switched by the smallest field when cooled to just below its Curie temperature. The relevant parameter in this situation is the ratio of field to saturation magnetization H/M_s; this ratio is very large at that temperature even for very small H since M approaches zero at the Curie temperature. However, upon further cooling, M_s increases and the completely switched spot can revert to multi-domain configuration, i.e., become unswitched, if both of the following conditions are satisfied: (1) the single-domain spot does not represent a lower energy state as compared to a multi-domain spot, and (2) there does not exist a sufficiently high energy barrier separating the single-domain spot configuration and the multi-domain spot configuration.

The data given in Fig. 4 show that a 1-μm discrete spot would favor a multi-domain configuration. However, they do not show the absence of a sufficiently high energy barrier separating the single-domain and the multi-domain states for a discrete 1-μm spot cooling from its Curie temperature because the data shown correspond to a continuous film.

Experiments are currently being performed to determine the Curie-point-switching fields for discrete areas. If discrete-area experiments yield a result similar to that given in Fig. 4, another avenue of approach can be taken. Theory suggests that the basic domain size becomes large for very thin films, and also for thick films. One-μm domains have been observed in films having a thickness of 1 μm. A 1-μm discrete spot of this thickness film should favor a single-domain state, and thus be amenable to Curie-point switching with very small fields. Work in this direction is also being pursued.

The results obtained from Curie-point-switching experiments on continuous 700-Å MnBi films that seemingly have negative value for memory applications have positive value for recording applications. They suggest the following recording technique.

A high-intensity focused laser beam would be scanned across a magnetically saturated MnBi film to Curie-point-switch a track. The average magnetization along the track would be controlled by the magnetic field present during Curie-point switching of that portion of the track. This magnetic field would be limited to a range where it would not affect anything previously recorded on the track. By controlling the applied magnetic field during Curie-point switching, a time-varying signal could be recorded as a spatial variation of the average magnetization along the tracks of an MnBi film. The recording could then be read out by rescanning the track with a low-intensity laser beam and sensing the average magnetization along the track with a polarizer and photodetector. A laser-beam scanning system is being assembled to further determine the physics relevant to such an application.

C. Fabrication of Small-Area $p^+\pi p^+$ Solid-State Diodes for Noise Measurements, *A. Shumka*

A program was established to investigate noise in germanium solid-state diodes.[5] Design criteria were established for optimizing these structures for noise measurements. The following guidelines were chosen:

(1) A $p^+\pi p^+$ structure to be used instead of the usual $n^+\pi n^+$ structure (SPS 37-39, Vol. IV, pp. 49–51) because it has an ohmic region in its I–V characteristic that can be effectively used for calibrating the noise analyzer.

(2) The input impedance of the solid-state diode to be of the order of 10 kΩ for operation within the maximum sensitivity range of the noise analyzer.

[5]In collaboration with N.-A. Nicolet, California Institute of Technology, Pasadena, Calif.

(3) The dc operation of the solid-state diode to extend well into the space-charge-limited (SCL) region before heating effects became important.

(4) Contacts to be alloyed because it was anticipated that they would not exhibit a large $1/f$ component of noise.

For these requirements to be satisfied, the solid-state diode was to have two p^+ contacts of 0.125-mm diameter alloyed on a π-type germanium wafer (acceptor doping density $\sim 10^{12}$ cm^{-3}). The separation between the contacts was to be 20 to 30 μm.

The contacts of 0.125-mm diameter specified for the $p^+\pi p^+$ solid-state diodes are much smaller than those of 0.5 and 0.7-mm diameter previously fabricated. The large reduction in contact area precluded the use of the masking and aligning techniques reported in SPS 37-32, Vol. IV, pp. 64–67. An optical system was constructed for defining and aligning the contacts through the use of photo-resist techniques.

Definition of the contact was obtained by placing a 0.125-mm circle, unexposed to light, on a thin layer of positive-working photo-resist, coated on a germanium wafer. This was accomplished by projecting onto the wafer a shadow from a 1.3-mm opaque disk that was placed in a beam of light from a mercury-arc lamp. A 7× microscope objective was used for focusing and reducing the image to required dimensions. The sensitized layer of photo-resist was dissolved in a developer. Positioning of the contact was performed with a special holding fixture that had a rotatable vacuum chuck for the wafer mounted on a three-axis micropositioner.

The photo-masking procedures were as follows: a thin and uniform layer of Shipley AZ1350 photo-resist was brushed on a wafer. Prior to mounting the wafer in the vacuum chuck, an optical filter (Schott-GG14) was inserted, which permitted the exposure of the specimen without sensitizing it. The wafer was rotated until it was aligned with the vertical and horizontal axes of the positioner and translated until an image of the disk was focused and centered. A beam splitter was used to monitor these operations with a microscope. After removal of the filter, an exposure time of about 4 s was sufficient to sensitize the photo-resist. The photo-resist mask took 2 min to develop. A SiO film was deposited on the masked surface. The photo-resist mask was subsequently dissolved in acetone, leaving the contact surrounded by the deposited SiO film. An identical procedure was then followed for the opposite side of the wafer. By this method the contact diameters could be controlled to within 5 μm and aligned to within 15 μm. The alloying techniques discussed in SPS 37-32, Vol. IV, could now be applied.

The germanium wafers were 2.5 mm square with a nominal thickness of 45 μm. To satisfy the requirement that the contacts be 20 to 30 μm apart, the penetration depths of the indium-alloyed contacts were controlled. Calculations based on the germanium–indium phase diagram were used to determine the amounts of indium necessary for alloying. There was a tendency for alloy pits to form because of the small contact areas, which made it difficult to control the penetration depths within 5 μm. Despite these alloy pits, which were larger in area than the contacts, it was found from metallurgical cross sections that the recrystallized p^+ contacts were defined by the mask only.

An I–V characteristic is shown in Fig. 5 for one of the $p^+\pi p^+$ solid-state diodes at dry-ice temperature. Three straight lines are drawn to depict the ohmic region and two SCL current regions, the uppermost region being related to the saturation of hole drift velocity. This solid-state diode can operate at a dc current as high as 8 mA, which is well within the SCL current region before any heating effects are observed. The I–V characteristics and the noise in the $p^+\pi p^+$ solid-state diode will be reported in a future SPS.

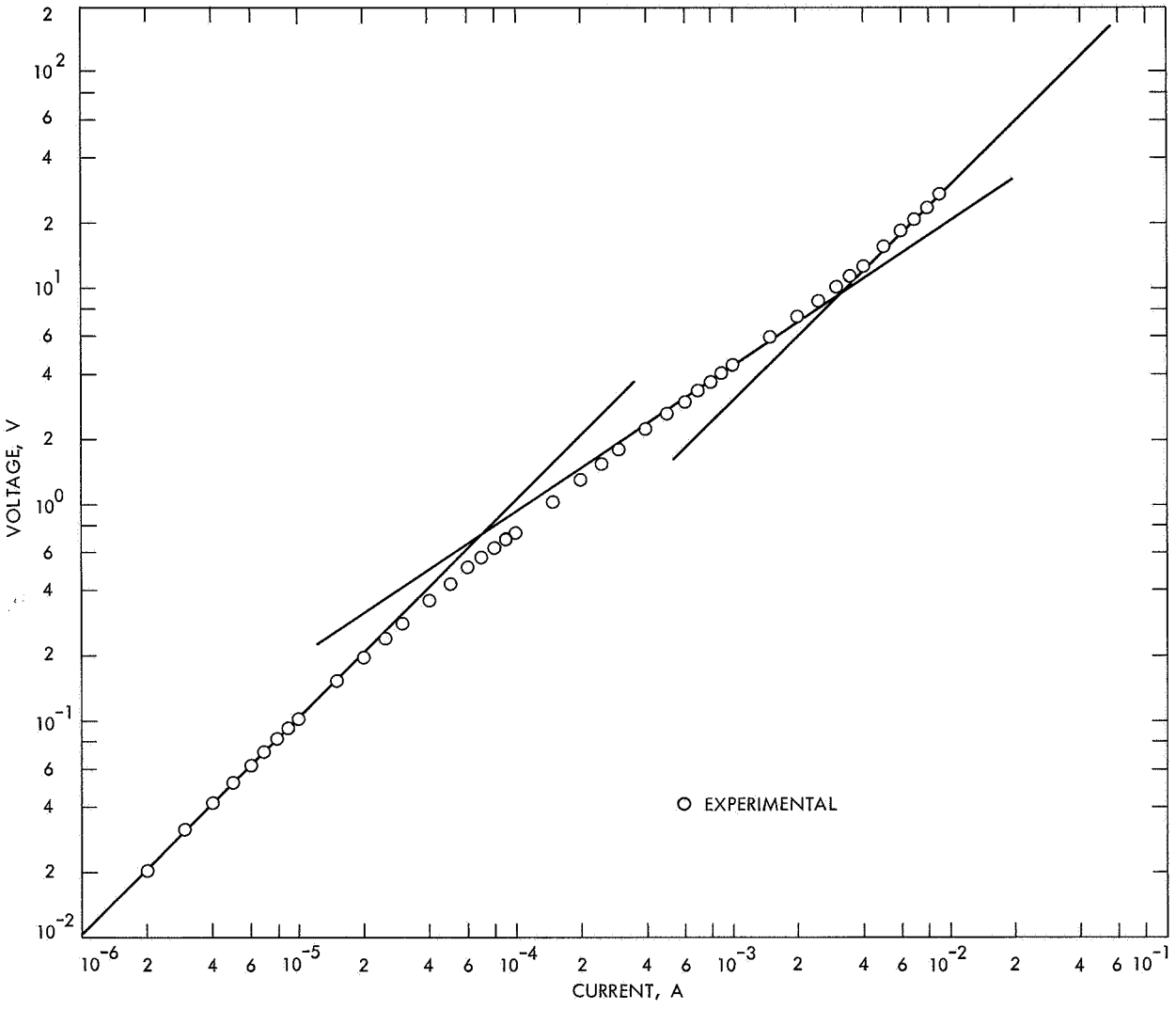


Fig. 5. *I–V* characteristic of a $p^+\pi p^+$ solid-state diode at $T = 195°K$

VII. Applied Mechanics

ENGINEERING MECHANICS DIVISION

A. Simulation of Venus Atmospheric Entry by Earth Reentry,[1] *J. M. Spiegel, F. Wolf, and D. W. Zeh*

1. Introduction

Each time a new unmanned entry mission to Mars or Venus is considered, the question of the value of an earth reentry test invariably arises. That is, can it be shown that the simulation of entry dynamics, heating, and heat shield response is adequate? This question arises from the well-known fact that the atmospheric composition and effective scale heights of the inner planets differ from that of earth.

Although the atmospheric compositions of Mars and Venus are now considered to be similar (mostly carbon dioxide), the lower velocities for Mars entry place the environment of heat shield materials of test specimen size within the reach of ground-based test facilities, whereas similar tests for direct Venus entry are, at present, marginal at best (SPS 37-49, Vol. III, pp. 141–152). Since analytical methods are still somewhat uncertain for predicting mass loss rates at Venus entry conditions (SPS 37-49, Vol. III, pp. 141–152), even one flight qualification test for this subsystem alone would be significant if it can be shown that the earth environment is comparable in severity and sufficiently similar to Venus entry.

Past work on the subject of planetary entry simulation by earth reentry (Refs. 1, 2, and 3) has emphasized many aspects of the problem, but little attention has been directed toward identifying the specific differences in radiative heat transfer and heat shield response, both of which might be expected to be particularly sensitive to chemical differences in atmospheric composition. It is concluded by H. Kennet (Ref. 1) that simultaneous simulation of all entry environments in a single flight test is not feasible, but that selected simulation can be achieved (Refs. 1 and 3). However, no conclusions were obtained regarding the response of an ablating heat shield.

In the present study, it was accepted at the outset that complete simulation is unlikely, but it was also postulated that the lack of simulation might be of an acceptable magnitude for proof test purposes. Time histories of acceleration, angle-of-attack envelope, entry heating, and heat shield ablator response were calculated at two locations on a spherically-blunted, 60-deg half-angle conical body to determine the best match of path angles and the degree of simulation attainable for the four factors specified.

[1]This article is a condensation of a paper of the same title to be presented at the AIAA Entry Vehicle Systems and Technology Conference, Williamsburg, Va., December 3–5, 1968.

2. Methods of Approach

a. Entry conditions. The entry-body configuration and other related factors used in this study are shown on Fig. 1. This configuration was chosen as a representative case although there are various reasons why a larger nose radius or a smaller cone angle might be more desirable. The entry angle-of-attack was taken as 50 deg; the initial pitch, yaw, and roll rates were taken as zero. At station 0.8r, the shock stand-off distance and wave angle were obtained from available flow-field solutions.

All conditions specified on Fig. 1 were used throughout the study for both earth and Venus entries.

b. Trajectory and heating. Entry trajectories, dynamic motion, and heat transfer were calculated with a modified version of a computer program (designated 1880)

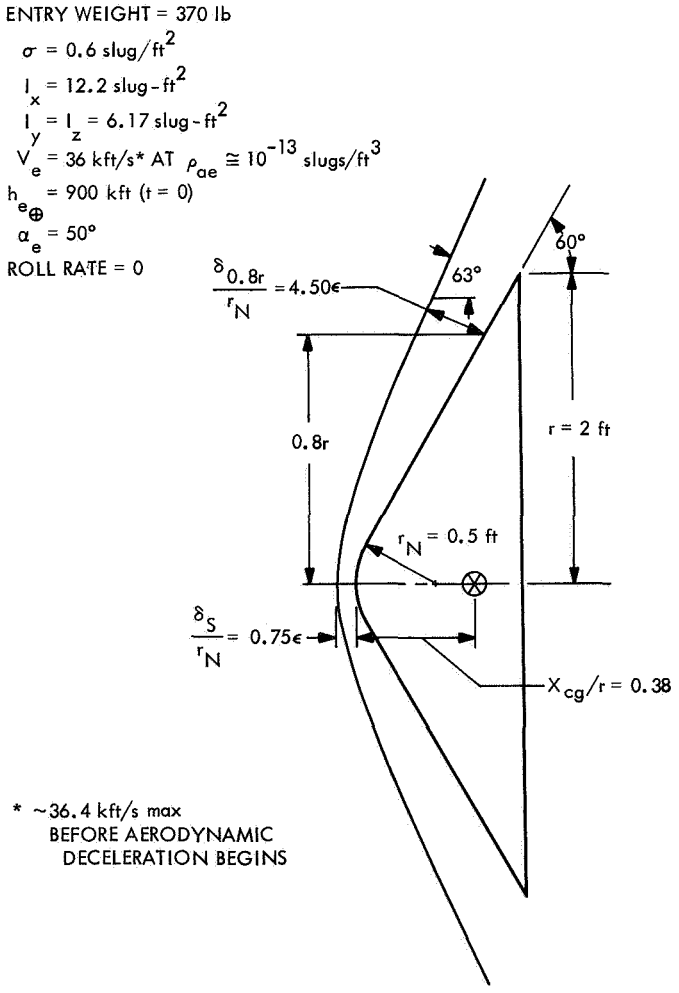


Fig. 1. Entry body configuration and flight conditions

originally developed by the AVCO Corporation for JPL and described in Refs. 4 and 5. Density and temperature profiles for earth and Venus were taken from Refs. 6 and 7, respectively. Reference 7 presents two basic model atmospheres (obtained from the *Mariner* Venus 67 and *Venera 4* probes) of which MV-3 was selected because the presence of N_2 in addition to CO_2 was expected to yield the highest radiative heat transfer. The mixture was approximated as 90% CO_2 and 10% N_2.

Stagnation point convective heat transfer was computed from a density, velocity, and molecular weight correlation originally built into the trajectory program and modified to agree with the results given in Refs. 8 and 9. Convective heating at a 0.8r station location on the conical region of the entry body was taken as 0.30 of the stagnation value based on available published and unpublished information. For local Reynolds numbers above 300,000, based on wetted distance from the stagnation point to the body edge, turbulent heating was assumed at the 0.8r station location, as computed by an appropriate expression built into program 1880.

Radiative heating was computed for both the stagnation point and a point on the cone (0.8r) (Fig. 1) using the slab approximation. Molecular bands, atomic lines, and continuum radiation sources are included. The air calculations were obtained by combining shock layer data from the trajectory program 1880 with radiation data from Ref. 10. The nonair calculations were obtained from program 1880 by a newly added routine based on the Kivel–Bailey method (Ref. 11) above a wavelength of 2000 Å, supplemented by vacuum ultra-violet contributions of C, N, and O lines as well as the CO(4+) band systems. Gaseous self-absorption is accounted for in an approximate manner, but radiation cooling and ablation product radiation interactions are not included. For comparative purposes, and for the velocity ranges considered, this omission is judged to be acceptable. The nonequilibrium contribution was treated in the manner described in Ref. 12 as incorporated in program 1880.

c. Heat shield response. The in-depth response of a charring ablator heat shield for the earth and Venus entries considered herein was calculated using the Equilibrium Surface Thermochemistry (EST) (Ref. 13) and Charring Material Ablation (CMA) (Refs. 14 and 15) computer programs developed by the Aerotherm Corporation for the NASA Manned Spacecraft Center. The CMA program gives a realistic description of chemical interaction between ambient and heat shield species and, therefore, met the requirements of the present study.

These requirements were to determine the effects of differing ambient gas composition on heat shield response during an earth flight simulation of Venus entry.

The heat shield was assumed to be made of high-density phenolic nylon (75 lb/ft^3, 50% phenolic resin and 50% nylon resin by weight) of sufficient thickness to approximate a semi-infinite body. Although materials forming stronger chars may be preferable for the actual heat shield, phenolic nylon was chosen both because it adequately characterizes the chemical response of many other charring ablators and because of the availability of input data necessary for the CMA program. Thermophysical properties were taken from Ref. 16; kinetic data was taken from Ref. 17.

Assumptions and uncertainties in the analysis that may affect final heat shield design, but are not likely to affect the qualitative results of the present study, include the following:

(1) All diffusion coefficients are assumed equal in the boundary layer.

(2) The heat transfer coefficient and mass transfer coefficient are assumed equal.

(3) Substantial uncertainties exist for kinetic and thermophysical data for the heat shield material.

(4) Internal reactions between pyrolysis gases and char have been ignored, as has char shrinkage.

(5) Mechanical char removal has been ignored; this could be most important, as noted later.

(6) Equilibrium conditions exist at the heat shield surface between the char, pyrolysis gases, and ambient species.

3. Results and Discussion

a. Entry path angle for simulation. An examination of approximate scaling rules and atmospheric density profiles for earth and Venus in the region of maximum heating and deceleration led to the preliminary finding that 45 deg is the steepest Venus entry that could be simulated by earth reentry. The final selection of an entry angle (γ_e) for a Venus trajectory that would be best simulated by a vertical earth entry was made by fixing all entry conditions as shown in Fig. 1 while varying γ_e and investigating the resulting simulation of deceleration and heating rates, integrated heating, and heat shield mass loss. On this basis, a 40-deg Venus entry was selected, and all subsequent discussion will relate to the simulation of this case by a vertical earth entry.

b. Trajectory and heat transfer simulation. Time histories of deceleration and angle-of-attack envelope were calculated to be quite similar for the 40-deg Venus and vertical earth entries, except that maximum deceleration is predicted to be about 18% higher for earth entry.

Time histories of convective heat transfer are presented in Figs. 2(a1) and 2(a2). The simulation is observed to be good for both the stagnation point and the 0.8r case. Transition to turbulent flow at 0.8r is predicted to occur at about the time of maximum stagnation point heating.

Time histories of radiative heat transfer are presented in Figs. 2(b1) and 2(b2). The stagnation point heating pulses appear quite similar in both atmospheres, except for the relatively later onset of the equilibrium radiation during Venus entry. The stagnation point maximum radiative transfer for air occurs at a velocity of 34,000 ft/s and at a shock layer temperature (T_{SL}) $\cong$ 11,000°K where the radiation is primarily atomic in nature. At the corresponding point in the Venus entry trajectory, the temperature is more than 1000°K lower, thereby depressing atomic sources, but still above the level at which substantial quantities of CO molecules are formed. The peak Venus radiation comes primarily from the CO(4+) molecular band system at wavelengths < 0.2 μ, and occurs at a velocity of 30,000 ft/s where $T_{SL} \cong 8400$°K. Therefore, the delay in the occurrence of peak radiative transfer for Venus relative to earth entry is attributable to the differences in chemical composition of the respective atmospheres.

At the 0.8r station of the entry body, the radiative heat pulses for Venus and earth are quite different in shape and magnitude as seen in Fig. 2(b2). There, shock layer temperatures at maximum radiative heating ($V \cong 30{,}000$ ft/s) are around 7700°K for both entry cases. At this temperature, air has no radiating species comparable in intensity to CO and CN. This explains the large peak for Venus entry, compared to the almost constant air radiation, and the large difference in the maximum radiative rates. This breakdown of similarity of radiative transfer distribution around the body, as for the stagnation point, also stems from the chemical difference of the two atmospheres and probably is not adjustable either by modification of entry conditions or any other factors.

c. Heat shield response simulation. Comparison of heat shield response for the 90-deg earth entry and 40-deg Venus entry is provided in Figs. 2(c1), 2(c2), 2(d1), and 2(d2) for both the stagnation point and the 0.8r location.

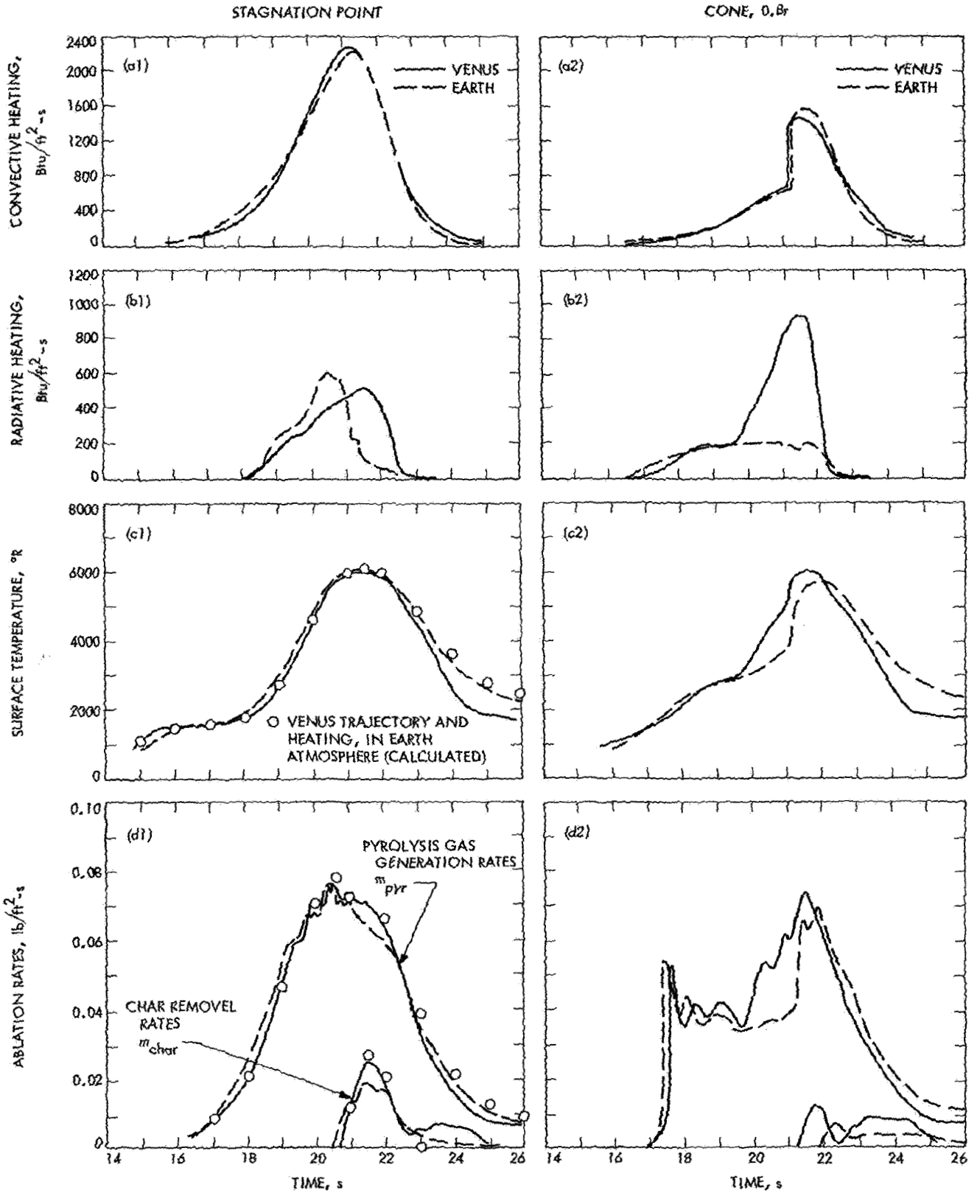


Fig. 2. Comparison of earth and Venus entry

Although the applied heating pulses are quite smooth, the curves representing ablation rates are irregular. Such irregularities or oscillations are inherent in the computational procedure and are not representative of any real physical process; they are discussed in some detail in Ref. 14 and shown to be characteristic only of materials, such as nylon, that decompose rapidly. Most important, it was found that the oscillations had relatively little effect on overall ablation response. In the present case, the oscillations in the predicted pyrolysis gas generation rates appear to be sufficiently small that even the detailed nature of ablation response is relatively unmasked.

Time histories of applied heating, surface temperature and ablation rates, have been vertically aligned in Fig. 2 to simplify visualization of ablation response. The first obvious conclusion upon inspection of these figures is that, within the limitations of the present analysis, simulation of the heat shield response for the 40-deg Venus entry by the 90-deg earth entry is remarkably good during the entire heating pulse at both the stagnation point and the 0.8r location. This, in the face of somewhat different radiative heating pulses, particularly at the 0.8r location, is indicative of the flexible, self-regulating, "heat-absorption" mechanisms of a charring ablator.

The most important of these mechanisms are convective blockage by ablation gases, reradiation at high surface temperatures, and endothermic char removal (sublimation and reaction with hydrogen in pyrolysis gases). It is important to note that the latter two mechanisms are critically dependent upon the presence of the char and are largely responsible for the comparable ablation response for the Venus and earth entries in spite of somewhat different applied heating rates. If the char were lost by thermomechanical means, pyrolysis rates would increase markedly and would be much more sensitive to applied heating rates.

It is apparent from the close agreement in ablation response for the Venus and earth entries that direct chemical effects due to ambient gas composition are negligible. To demonstrate this even more clearly, the CMA program was run using the 40-deg Venus entry conditions but assuming an air atmosphere. These results, shown as circles in Figs. 2(c1) and 2(d1), follow the true Venus calculations very closely up to about 22 s, and indicate that the assumed ambient gas composition is unimportant. After this time, the ambient gas effects become noticeable and the calculations follow the earth results, as would be expected since an air atmosphere was assumed. The surface temperatures and pyrolysis rates are somewhat higher for the earth entry in this region since the chemical reactions between ambient species and ablation products are relatively more exothermic in air than in the assumed Venus atmosphere. In any case, these chemical effects are clearly not important in determining overall heat shield response.

It should be emphasized that the above remarks on heat shield response are valid only within the framework of the ablation model used. It is equally important to note, however, that an actual flight test is, at present, the only means of determining whether other ablation mechanisms may be active and/or controlling during Venus entry since the entry conditions cannot be simulated fully with existing ground facilities. It has been shown that the heating, pressure, and, hence, shear histories of a 40-deg Venus entry can be closely simulated by a 90-deg earth entry, and that the ambient gas composition has negligible direct effect on ablation response. Hence, any unidentified ablation mechanisms such as thermomechanical char removal that might occur in the Venus entry should also appear in the earth flight test.

d. Launch vehicle considerations. The need for a near-vertical earth atmospheric entry at about 36,000 ft/s to simulate a Venus entry at path angles up to 40 or 50 deg for about a 4-ft diameter, 400-lb vehicle places stringent requirements on the launch system. Existing applicable systems are both costly ($2–$4 $\times$ 10^6) and, in their current state, flight tested to operate only at shallow (0–15 deg) entry path angles.

One means of alleviating the size and weight requirement, with a possibly acceptable compromise in simulation, is to essentially truncate a full-scale entry configuration at a location that permits full-scale simulation of the stagnation region. Under these conditions, the best choice at this time for a launch vehicle of modest cost would appear to be the (prospective) Athena Super H, of which the three upper stages are in existence and have been flown.

4. Conclusions

For the Venus entry mission considered herein, a full-scale vertical earth reentry flight test at the same ballistic coefficient and entry velocity should provide an acceptable simulation of deceleration, dynamic motion, and heat shield response at a Venus path angle of about 45 deg despite significant differences in radiative heating.

Since a full-scale test imposes stringent requirements on the launch system for such a test, a compromise test

configuration is possible in which a full-scale configuration is truncated at a location that permits full-scale simulation of the nose region only.

When some compromises are accepted, it appears that a meaningful earth reentry test of a simulated full-scale aeroshell (or part thereof) for the Venus case considered herein could be made. As a minimum, the earth reentry test can place a Venus capsule in a flight environment comparable in severity to actual Venus entry. This environment is difficult to obtain in ground facilities for any significant piece of flight hardware.

References

1. Kennet, H. and Taylor, R. A. "Earth Reentry Simulation of Planetary Entry Environment," *J. Spacecraft Rockets,* Vol. 3, pp. 504–512, 1966.
2. Beuf, F. G., Katz, G. D. and Kern, R. J., "Earth Entry Flight Test of Mars Entry Vehicles," *J. Spacecraft Rockets,* Vol. 3, pp. 498–503, 1966.
3. Stimpson, L. D., "Earth Simulation of Planetary Entry," Paper 65-442, presented at the AIAA Second Annual Meeting, San Francisco, Calif., July 26–29, 1965.
4. *Mars-Venus Capsule Parameter Study,* Technical Report 64-1, Vols. I, II, III. AVCO Corporation, Research and Advanced Development Division, Wilmington, Mass., Jan.–Mar. 1964.
5. *JPL Entry Vehicle Design Computer Program Users Manual,* Scientific and Technical Aerospace Report N67-13141. AVCO Corporation, Space Systems Division, Lowell, Mass., Dec. 1, 1966.
6. *U.S. Standard Atmosphere, 1962,* U.S. Government Printing Office, Washington, Dec. 1962.
7. Schiffer, R. A., "Engineering Models of the Venus Atmosphere Based on an Interpretation of Recent Space Observations of Venus," submitted for presentation at AIAA Seventh Aerospace Sciences Meeting, New York, Jan. 20–22, 1969.
8. Hoshizaki, H., "Heat Transfer in Planetary Atmospheres at Super-Satellite Speeds," *ARS J.,* Vol. 32, pp. 1544–52, Oct. 1962.
9. Marvin, J. G., and Deiwert, G. S., *Convective Heat Transfer in Planetary Atmospheres,* NASA TR R-224. National Aeronautics and Space Administration, Washington, 1965.
10. Page, W. A., *et al.,* "Radiative Transport in Inviscid Non-Adiabatic Stagnation-Region Shock Layers," Paper 68-784, presented at the AIAA Third Thermophysics Conference, Los Angeles, Calif., June 1968.
11. Kivel, B., and Bailey, K., *Tables of Radiation from High Temperature Air,* AVCO–Everett Research Laboratory Research Report 21. AVCO Corporation, Everett, Mass., 1957.
12. Wolf, F., and Spiegel, J. M., "Status of Basic Shock-Layer Radiation Information for Inner-Planet Atmospheric Entry, *J. Spacecraft Rockets,* Vol. 4, pp. 1166–1173, 1967.
13. *Aerotherm Equilibrium Surface Thermochemistry Program, Version 2, User's Manual,* Aerotherm Corporation, Palo Alto, Calif., June 1966.
14. Moyer, C. B., and Rindal, R. A., *An Analysis of the Coupled Chemically Reacting Boundary-Layer and Charring Ablator: Part II, Finite Difference Solution for the In-Depth Response of Charring Materials Considering Surface Chemical and Energy Balances,* Final Report 66-7. Prepared for the NASA Manned Spacecraft Center under contract NAS 9-4599. Aerotherm Corporation, Palo Alto, Calif., Mar. 14, 1967. Also available as NASA CR-1061, National Aeronautics and Space Administration, Washington, June 1968.
15. *Aerotherm Charring Material Ablation Program, Version 2, User's Manual,* Aerotherm Corp., Palo Alto, Calif., Jan. 1966.
16. Wilson, R. G., *Thermophysical Properties of Six Charring Ablators from 140 to 700°K and Two Chars from 800 to 3000°K,* NASA TN D-2991, National Aeronautics and Space Administration, Washington, Oct. 1965.
17. Rindal, R. A., and Kratsch, K. M., *Prediction of the Ablative Material Performance on a Scout Entry Vehicle,* Final Report 66-4. Prepared for the NASA Ames Research Center under contract NAS 2-3587. Aerotherm Corporation, Palo Alto, Calif., July 1966.

B. Mobility and Wheel–Soil Interaction: Study and Tests, *I. Kloc*

1. Introduction

This article presents an outline of the basic theoretical concepts and analysis being used to determine the bearing capacity and the pressure–sinkage relationships of soft soil surfaces when subjected to uniform vertical or inclined loads. This study is being implemented by an exploratory testing program. The results of both analysis and tests will be used to evaluate mobility performance of planetary roving vehicles.

The need for this analysis arises due to the practical impossibility of obtaining bearing capacity values and pressure–sinkage relationships by direct tests as done on earth over potentially known and accessible environments. Faced with this problem, the objectives of this study are as follows:

(1) Theoretical determination of the ultimate pressure load of horizontal or sloping lunar surfaces.

(2) Determination of soft soil surface pressure–sinkage relationships to evaluate lunar vehicle performance on horizontal and sloping surfaces.

(3) Testing to see if the pressure–sinkage relationships can be predicted by using the ultimate pressure load formulation as a function of depth below the surface.

This study relates mainly to the safety and performance of vehicles operating over horizontal soil surfaces or when climbing, descending, or traversing a sloping ground. For testing purposes, and to verify the theory, use is made of a cohesionless soil (sand).

2. Soil Bearing Capacity Problem

Currently, there is no general and reliable theoretical formulation, based upon tests, that presents the influence of terrain slope on the soil bearing capacity. Available solutions provide only the ultimate bearing load applied uniformly on an infinitely long strip over a horizontal terrain. A solution of this problem, applying the method of characteristics, was obtained by V. V. Sokolovsky (Ref. 1) who also considered inclined loads. An approximate solution to the same problem was obtained by K. Terzaghi (Ref. 2) and improved by G. G. Mayerhof (Ref. 3) who expanded the results (Ref. 4) to include oblique loads over horizontal terrains. This author also studied the case of vertical loads applied on slopes (Ref. 5) considering either purely cohesive or cohesionless soils only, a limitation which restricts its general and practical application. Recently, L. L. Karafiath (Ref. 6), following Terzaghi's basic concepts, formulated the ultimate (vertical or oblique) load on the slopes. No load optimization or tests were made to better define and verify the results. Since the problem at hand is of non-linear character and the methods of approximation resort to the principle of linear superposition, it is impossible to state the degree of approximation obtained.

As an attempt to solve the problem of ultimate load on sloping surfaces, it is considered that the application of limit load analysis of the theory of plasticity to soil mechanics will permit a satisfactory solution. In this context, the theorems of collapse load of limit analysis define upper and lower bound loads between which lies the ultimate load. A proper selection of both velocity and stress field pattern permits narrowing the interval of these bounds, and a better definition of the ultimate load is obtained. The identity of both the upper and lower bounds is a sufficient condition to define the true maximum load.

Plastic limit analysis theory was used in the study of the vertical punch indentation problem of soils by R. T. Shield (Ref. 7). A rather large difference between bounds was obtained for friction angles between 30 and 40 deg, of particular interest to lunar soils. Furthermore, the soil was considered weightless although this factor bears significantly on the ultimate load value. To obviate these limitations, the present analysis considers the soil weight and attempts to narrow the interval between the limiting loads.

To this effect, a theoretical study was done and numerical results have been obtained that define the upper limit load of the lunar surface when subjected to a quasi-static vertical or inclined load applied on an infinitely long strip of known width resting either on a horizontal surface or along a sloping terrain (Fig. 3). The solution accounts for the influence of footing depth, soil weight, friction, and cohesion. The load is minimized by optimizing the failure angle ψ.

It is assumed that the lunar soil behaves as a rigid plastic material that follows the Coulomb–Mohr failure condition for soils and its associated flow rule. This criterion satisfies reasonably well the expected failure modes and the character of the lunar soil as described by its mechanical properties, which are approximately as follows: the angle of internal friction, 37 ± 2 deg; cohesion, 0.06 psi; and unit weight in the lunar gravitational field, 15 ± 2 lb/ft^2. A computer program was developed that accounts for all these factors. The general form of the ultimate pressure load[2] is

$$p = c(N_c)_{\alpha,\delta} + \gamma z(N_q)_{\alpha,\delta} + \tfrac{1}{2}\bar{B}\gamma(N\gamma)_{\alpha,\delta}$$

where

p = ultimate pressure

c = soil cohesion

z = depth below surface

$\bar{B}$ = pad width

N_c, N_q, N_δ = bearing capacity coefficients functions of ϕ, α, and δ

ϕ = soil friction

α = terrain slope

δ = load inclination with reference to local vertical

γ = soil unit weight

This study, now in process, attempts to define the lower bound load that results from an appropriate selection of a stress field pattern. It is futher emphasized that the infinite strip direction is horizontal across the slope and

[2]Kloc, I., *Load Bearing Capacity Bounds of Sloping Soil Surfaces* (to be published).

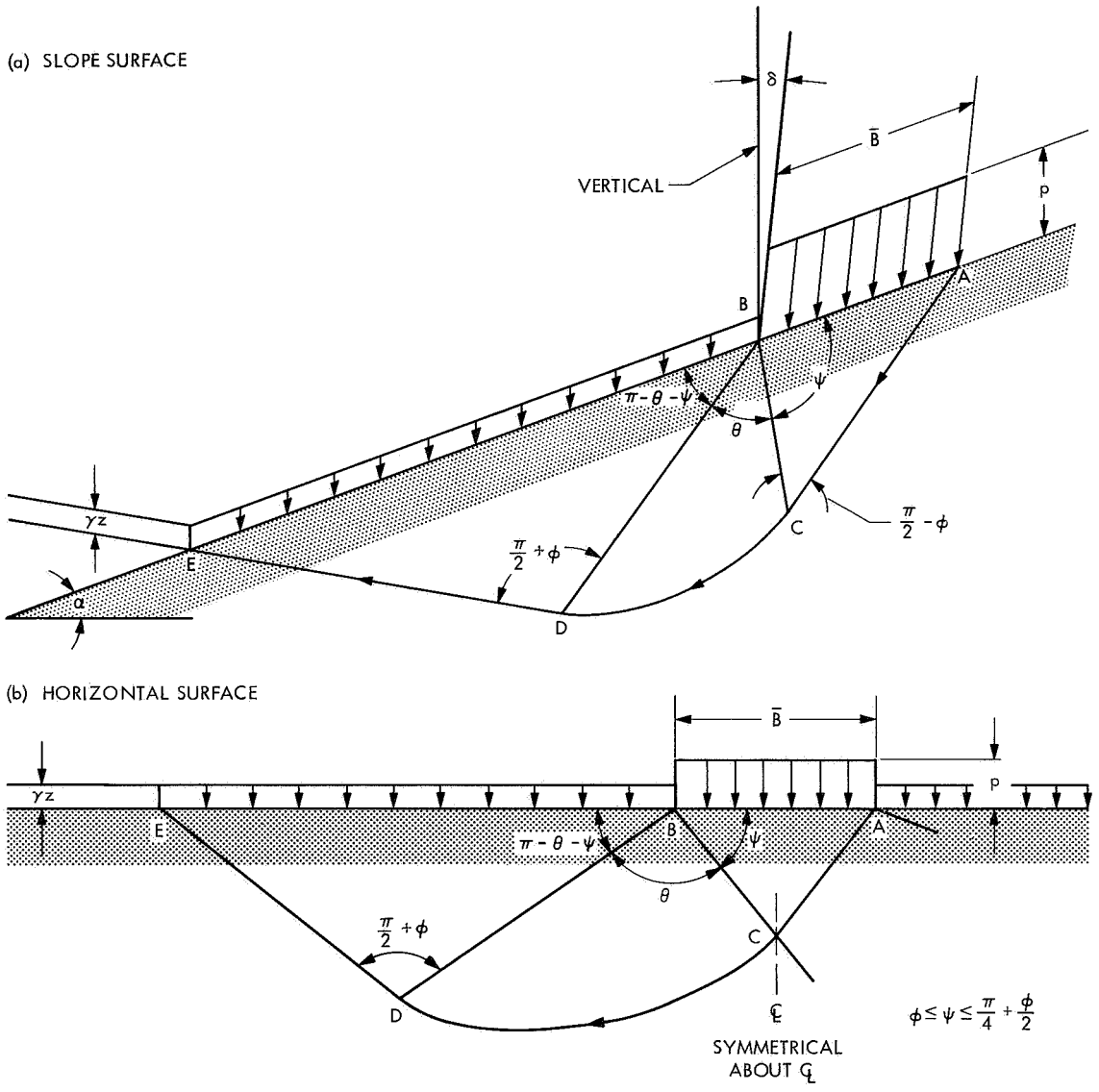


Fig. 3. Soil failure mechanisms

not up, down, or at the top of the slope (rim of a lunar crater). Each of these cases requires a separate study which can be done based on the same method of solution.

In the application of the numerical data, extreme care should be exercised when trying to extend the results of an infinite strip to a finite circular, rectangular, or elliptical load pattern. These factors should be interpreted in the light of the three-dimensional criteria of soil failure which, in the present state of theoretical soil mechanics, are still under discussion and experimentation. Nevertheless, an attempt is being made to better define and improve on the currently available concepts.

3. Extension of Analysis

The importance of the various stability conditions to which the planetary roving vehicle will be subjected as it traverses sloped surfaces cannot be over-emphasized. The sorting out, and overriding of obstacles can always be managed on a go/no-go basis as long as the vehicle has the required power and discrimination capability to operate accordingly. When faced with the maximum support that the ground can provide, the soil mechanical properties and lunar topography, slopes, and regions of slope changes play an important role in disclosing the potential risks of the vehicle becoming immobilized. This eventuality may be caused either by excessive sink-

age or the existence of weakening tension cracks around the periphery of a crater edge. In the latter case, a state of failure could result in a pronounced tilting of the vehicle accompanied by a loss of stability. Within the same analytical framework, an extension of this study could cover the ultimate load a lunar crater edge can support.

4. Testing Program

The following exploratory testing program describes the basic information, methods, and procedures required to corroborate the load-bearing capacity values and the pressure–sinkage relationships to be used in connection with lunar surface vehicle mobility studies. Specific information is given with reference to equipment handling, instrumentation, soil preparation, and generation of testing data. The purpose of these tests is to disclose the soil failure character and response (load versus displacement) of the following two principal soil surface types:

(1) Horizontal soil surface.

(2) Sloping soil surface.

In both cases, the applied loads are uniform and vertical. In the case of horizontal surfaces, the displacements are restricted to develop along the vertical direction only. For sloping surfaces, the bearing plates are guided and displacements may occur simultaneously in the vertical and horizontal directions.

Plate–soil contact starts either on or at a specified depth below the surface. In all cases, during the penetration process, the bearing plates are maintained parallel to the original soil surface.

The objectives of the tests are as follows:

(1) To verify the analytical approach that formulates the bearing capacity of horizontal and sloping terrain surfaces subjected to quasi-static uniform vertical loads.

(2) To find out if the bearing capacity of cohesionless soil surface slopes can be predicted up to slopes equivalent to the soil angle-of-repose.

(3) To qualify the semiempirical procedure designed to estimate the pressure–sinkage relationships whereby a single continuous load penetration test may be approximated by a series of load bearing capacities as functions of the depth below the terrain surface.

(4) To compare the influences of plate size and geometry (rectangular, square, circular, and wheel shapes) on the soil bearing capacity and pressure–sinkage relations.

a. Test box and soil hopper. A plywood box reinforced with peripheral aluminum corner angles is provided to contain the soil material (Fig. 4). The inside box dimensions are $28 \times 38 \times 30$ in. deep. Box dimensions are established to minimize the wall boundary effects for the selected bearing plates. Special aluminum guides are provided to support a metal straight edge that smooths the soil surface to prescribed heights and slopes. An overhead crane picks up the box by means of four steel cables attached to the corners and carries it to the load testing machine.

A specially designed wood hopper ($26 \times 11.5 \times 13$ in. high) permits the soil to flow down into the test box at a controlled rate and height while it displaces horizontally and/or vertically (Fig. 4). The bottom of the hopper has an adjustable gap that permits control of the sand flow. The soil is deposited in the test box under similar conditions of flow rate and height. This generates an artificial sand bed of homogeneous and reproducible granular structure over a wide range of densities.

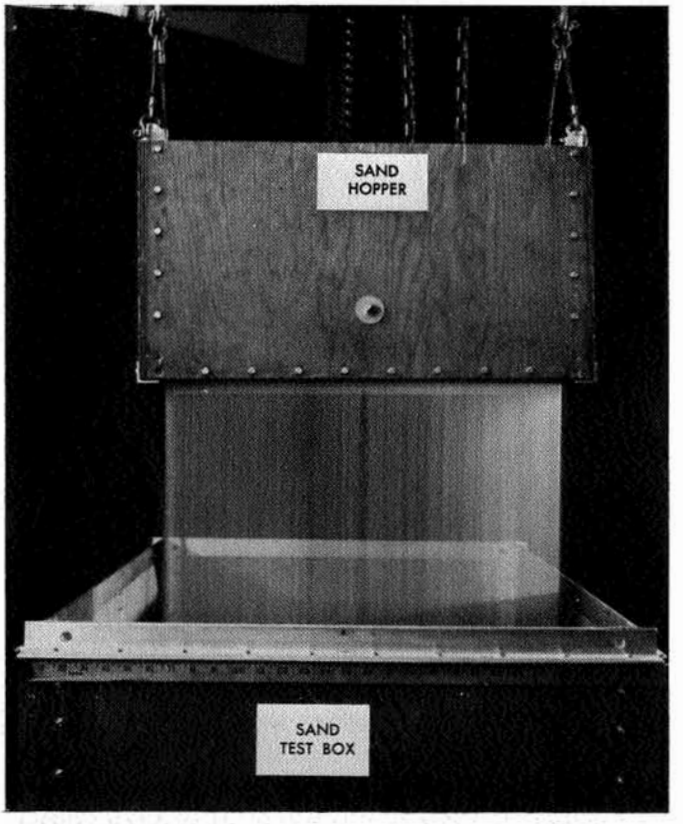


Fig. 4. Control of sand density

b. Soil material. The soil material selected is an air-dried, coarse, medium-to-fine white silica sand. Penetration tests are done on a compact sand in which a shear-type failure can be produced by bearing plates.

The following tests will be made to evaluate the sand properties:

(1) Sieve analysis to define the grain size distribution.

(2) Maximum and minimum sand density to determine the parametric limits of soil behavior.

(3) Direct shear tests of sand for dense and loose conditions to determine the angle of internal friction as a function of relative density.

(4) Angle-of-repose of compacted sand to define the limiting slope and stability conditions of the sand surface.

c. Sand bed preparation. The most important controlling factors of soil placement relate to the uniformity and repeatability of soil structure formation to set the test results on an equal basis. To this end, a simple procedure using the soil hopper permits control of the sand density at all levels of soil bed preparation (Fig. 4). The sand is compacted by its own weight falling through the hopper.

The main factors controlling density of sand deposition are the rate (weight deposited per unit time) and height of the flow. For a constant flow height, an increase of the rate of flow will decrease the density of the sand. The rate of flow has a more predominant influence in the soil density outcome than the variation of flow height. For instance, a height of 32 ± 2 in. produces less than 1% variation in density. Dust conditions have been practically eliminated by selecting a sand mixture with less than 1% passing sieve No. 200 by weight and a maximum fall height of 32 in.

After the box is filled to the prescribed level, its surface is smoothed out. Control weights are taken using a load cell and an SR-4 gage.

d. Bearing plates and loading ram. The bearing plates are made of aluminum. A variety of shapes and sizes can be used limited by the boundary influences of the testing box walls. Rectangular plate aspect ratio is 1/5 to simulate a uniformly loaded long strip in plane deformation.

A common fixture supports all bearing plates and a carriage permits unrestricted horizontal displacement to occur simultaneously with the vertical displacement. The horizontal carriage may be locked to produce only vertical displacements (Fig. 5).

5. Pressure–Sinkage Test

The sand test box is set on the loading machine verifying that the bearing plate and soil surface are parallel. The compressive vertical load is applied at a penetration rate of 0.5 in./min and unloaded at 0.05 in./min. Loads, pressures, and displacements are automatically and continuously recorded by two x–y plotters that record load versus horizontal and vertical displacement.

6. Testing Plan

Four groups of bearing plate tests are planned. These are identified by their configuration as rectangular, square, circular, and wheel tests. At least two load penetration tests are required on different sizes of the same plate configuration in order to determine significant pressure–sinkage relations. The minimum number of tests

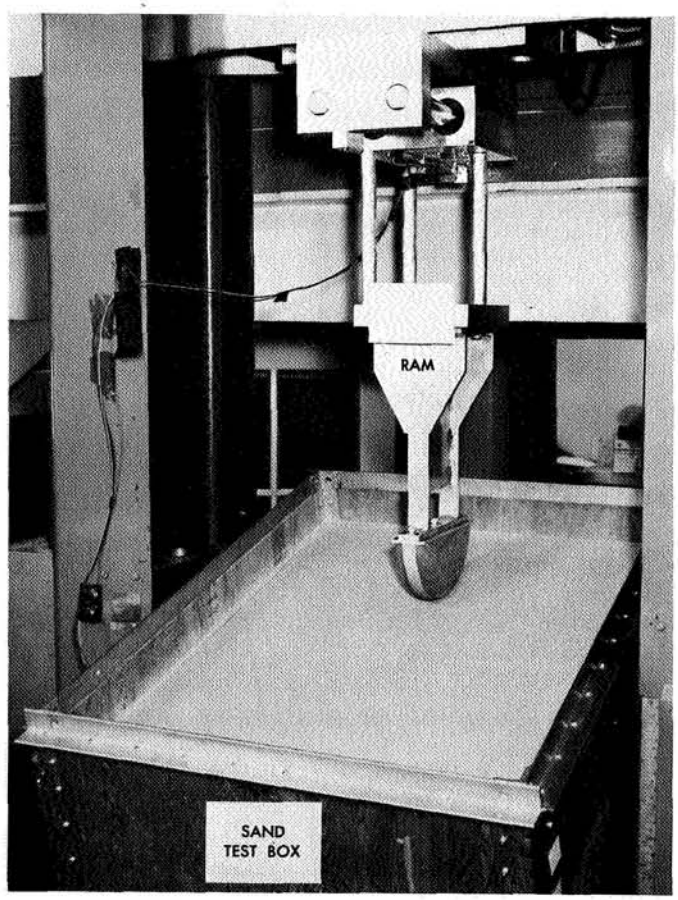


Fig. 5. Fifteen-degree-slope surface test (wheel up–down slope) with wheel raised

required results from the combination of plate shape, size, slope angle and initial plate level below the soil surface (Table 1).

7. Initial Exploratory Tests

Some initial soil tests were made to check out the experimental equipment. These clearly demonstrate the effect of soil slope on the pressure–sinkage relationship and bearing capacity of loaded wheels and flat plates on frictional-type soils. The test soil used was coarse medium-to-fine silica sand with a density of 105 lb/ft^3.

Segments of a wheel, 10 in. diam × 2.5 in. wide, were attached to the loading ram (Fig. 5). Load penetration tests were made for loads applied vertically at a constant penetration rate of 0.5 in./min on both horizontal as well as 15-deg sloped soil surfaces. Vertical, as well as horizontal, displacements were recorded using a special ball-bearing supported loading carriage that permitted free horizontal displacement of the loading fixture. Thus, the wheel segment follows the soil surface failure pattern simulating a free-loaded, non-rotating wheel on a slope.

Results of these early tests for the wheels oriented across the 15-deg slope side, as well as up the slope face, are shown in Fig. 6. With this limited information, it is clear the slope effect on soil properties is considerable even at this low slope angle (15 deg) relative to the angle-of-repose (30 deg). Preliminary tests were also made with rectangular plates (2.0 × 10.0 in., Fig. 7). The results point out the same controlling slope influence on

Table 1. Test plan for plate configuration

Plate shape	Size, in.	Slope, deg	Depth below soil surface, in.	Number of tests
Rectangular	2.00 × 10.00 3.00 × 15.00	0 15 30	0 1 1.5	18
Square	3.54 5.00	0	0 1 1.5	6
Circular (diam)	4.00 5.64	0 15 30	0 1 1.41	18
Rigid wheels (wood) (diam × width)	10 × 2.5 14 × 3.5	0 15 30	— — —	18

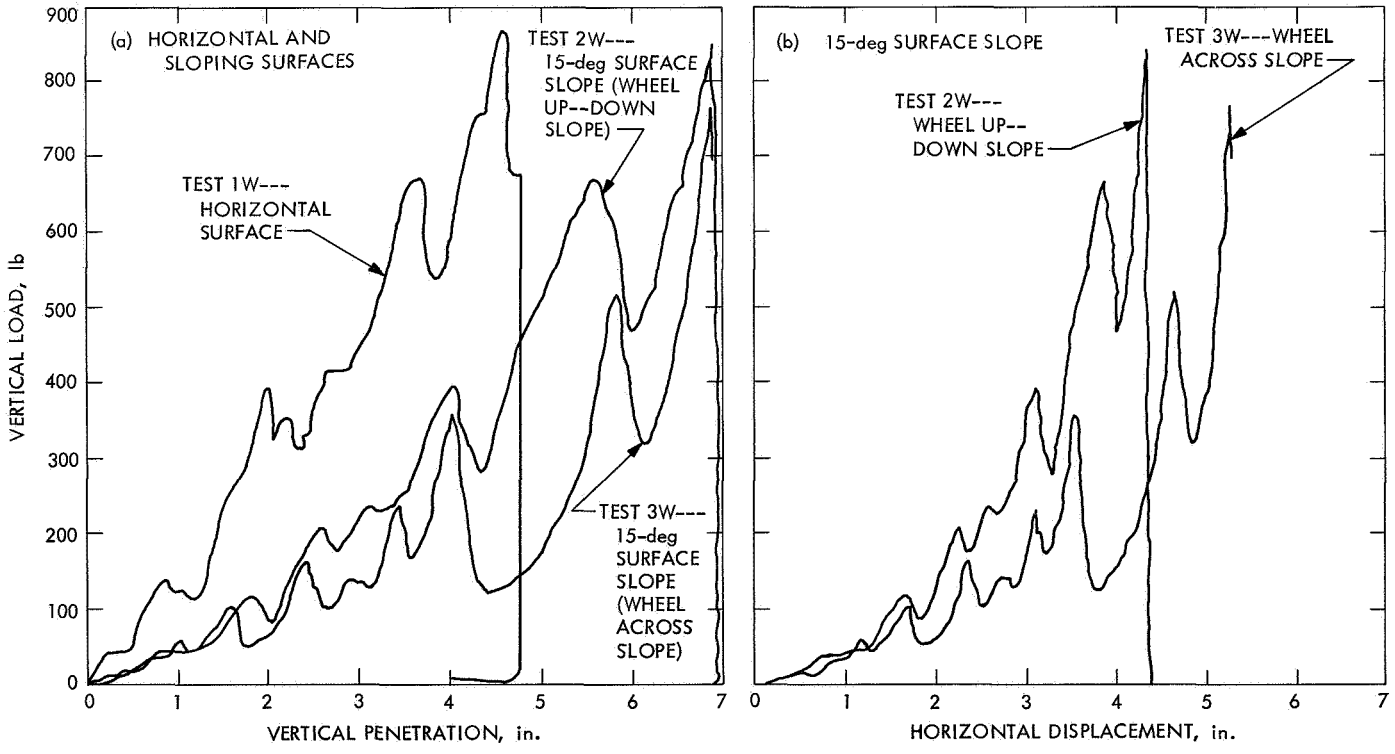


Fig. 6. Wheel–load sinkage test results (10.0-in.-diam × 2.5-in.-width wheel)

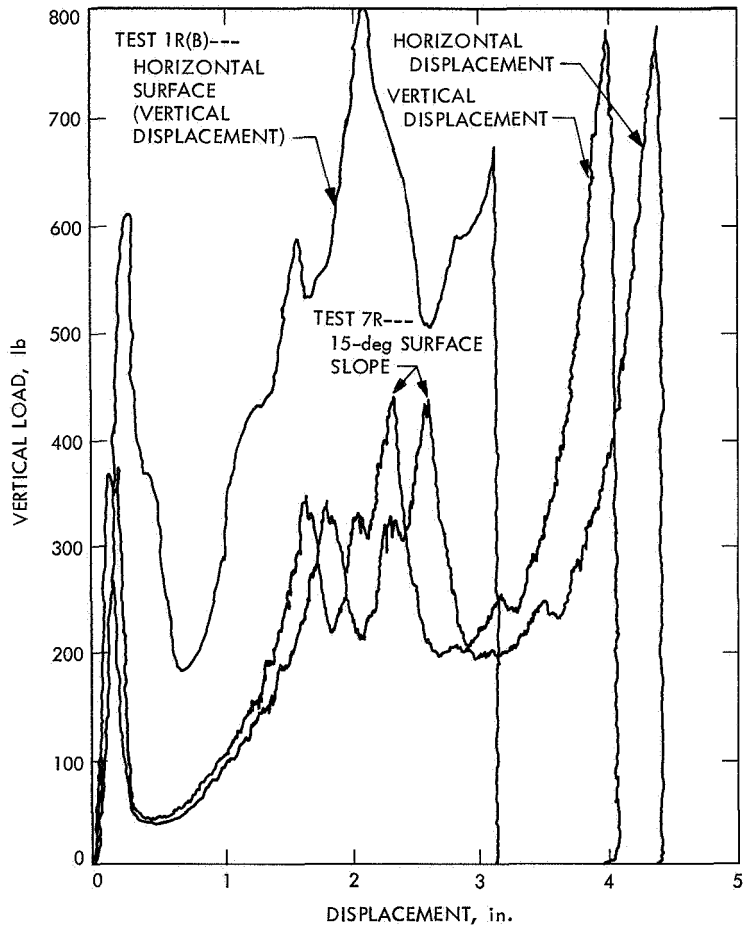


Fig. 7. Preliminary tests made with rectangular (2- X 10-in.) plates

bearing capacity and displacements. The principal conclusions of these exploratory tests are as follows:

(1) Significant degradation of bearing capacity due to slope influence occurs even at relatively low surface slopes compared to the soil angle-of-repose.

(2) Reduction of load support capabilities due to slope effect is accompanied by increased displacements. Horizontal and vertical displacements are of comparable magnitudes and importance in sloped surface tests, whereas the vertical displacements predominate on horizontal surfaces. The total vector displacement of a sloping surface does not coincide with the load direction.

(3) Current mobility concepts refer only to horizontal terrains and these concepts will have to be re-evaluated and extended to include slope effects. In particular, new concepts of soil thrust, mobility resistance, and slip on sloping terrains must be investigated to properly model and evaluate planetary roving vehicle mobility performance.

(4) Vehicle operational safety on slopes is largely impaired due to excessive sinkage. Since vehicle design performance is highly dependent on slopes, further analysis and tests must be done to disclose preferred relative wheel–slope orientation to obtain optimum performance.

References

1. Sokolovsky, V. V., *Statics of Soil Media,* London, Butterworth, 1956.
2. Terzaghi, K., *Theoretical Soil Mechanics,* John Wiley & Sons, 1943.
3. Mayerhof, G. G., "The Ultimate Bearing of Foundations," *Geotech.*, Vol. 2, p. 301, 1951.
4. Mayerhof, G. G., "The Bearing Capacity of Footings Under Eccentric and Inclined Loads," in *Proceedings of the Third International Conference on Soil Mechanics and Foundation Engineering,* Vol. 1, p. 440, 1953.
5. Mayerhof, G. G., "The Ultimate Bearing Capacity of Foundations on Slopes," in *Proceedings of the Fourth International Conference on Soil Mechanics and Foundation Engineering,* Vol. 1, p. 384, 1957.
6. Karafiath, L. L., and Nowatzky, E. A., *A Study of the Effect of Sloping Ground on Bearing Strength and the Landing Performance of Space Vehicles,* Grumman Research Department Memorandum RM 407. Mar. 1968.
7. Shield, R. T., "On Coulomb's Law of Failure in Soils," *J. Mech. Phys. Solids,* Vol. 4, pp. 10–16, 1955.

VIII. Solid Propellant Engineering

PROPULSION DIVISION

A. Surface Temperature Relationships for Ignition Material Deflagration Onset, *O. K. Heiney*

1. Introduction

For many electroexplosive applications it is necessary to know or be able to predict the temperature at which deflagration onset will occur in ignition mix materials. The element of the explosive train of interest here is the ignition bead chemicals immediately in contact with and ignited by a heated bridgewire. These materials are normally a heavy metal-oxidizer mixture bonded to the bridgewire with a nitrocellulose paste. The purpose of the experimental effort is to ignite materials whose thermal and chemical properties are carefully controlled with a bridgewire at varying known energy input rates. With proper analysis, the conventional power versus time-to-fire curve would then give an ignition temperature–time relationship, providing, hopefully, a constant temperature ignition criterion or a regular temperature–time-to-fire dependence.

2. Analysis and Procedure

The electroexplosive system used is as illustrated in Fig. 1. The ignition temperature is that at the mix wire interface at the time the ignition explosion destroys the wire.

Thermal energy input to the system is, of course, provided by large metered pulses of electrical current flow in the highly resistant wire. With the known electrical and thermal properties of nichrome and the ignition material, an interface temperature expression may be formulated from a solution provided for this type problem by Ref. 1, Sec. 13.8.

For a heat generation rate in region A of Q_0, the quoted solution as a function of time is then

$$T - T_0 = \frac{4Q_0K_2k_2}{\pi^2 a}\int_0^\infty \frac{[1 - \exp(-k_1u^2t)]\,J_0(ur)\,J_1(ua)\,du}{u^4[\phi^2(u) + \psi^2(u)]} \tag{1}$$

with u being a dummy variable of integration

$$K = (K_1/K_2)^{1/2}$$

and

$$\psi(u) = K_1k_2^{1/2}\,J_1(au)\,J_0(Kau) - K_2k_1^{1/2}\,J_0(au)\,J_1(Kau)$$

$$\phi(u) = K_1k_2^{1/2}\,J_1(au)\,Y_0(Kau) - K_2k_1\,J_0(au)\,Y_1(Kau)$$

J_0 and J_1 are zero and first-order Bessel functions of the first kind, and Y_0 and Y_1 are zero and first-order Bessel

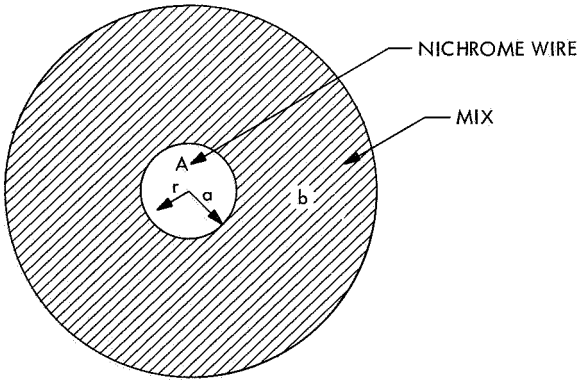


Fig. 1. Bridgewire ignition bead system

functions of the second kind, respectively. The remainder of the notation is as listed in Table 1.

To determine the temperature at the interface of the wire and mix for a given firing it is necessary to set $r = a$, input the appropriate values of Q_0 and t, and numerically integrate (by computer) Eq. (1) on the dummy variable.

Accomplishment of the above requires, however, detailed knowledge of the thermal properties of both nichrome wire and the ignition materials of interest. The data on the nichrome wire are available in the literature (Refs. 2 and 3). The data on the specific heat and heat conductivity for the ignition mixes had to be experimentally derived. The mixes were designated X-26 for a zirconium-ammonium perchlorate formulation and X-29 for a zirconium-barium chromate type.

Specific heats for both were determined at various temperatures on a differential scanning calorimeter by means of comparison with a synthetic sapphire (Al_2O_3) reference

Table 1. Definition of terms

a	= radius of wire
E_A	= lumped activation energy of ignition reaction
k_1	= thermal diffusivity of wire
k_2	= thermal diffusivity of ignition mix
K_1	= thermal conductivity of wire
K_2	= thermal conductivity of ignition mix
Q_0	= energy generation rate per unit volume in wire
r	= arbitrary radius reference
T_0	= ambient temperature
T	= wire mix interface temperature
u	= dummy variable of integration
τ	= specific time to fire

whose specific heat at various temperatures is precisely known. Results are as tabulated on Table 2.

Table 2. Specific heats of X-26 and X-29

Temperature, °C	Specific heat, cal/g/°C	
	X-26	X-29
50	0.142	0.129
100	0.150	0.137
150	0.161	0.149
175	0.170	0.154

The heat conductivity of the materials was determined by a guarded hot plate method very similar to that described in ASTM procedure C-177 (Ref. 4). These results are given on Table 3.

Table 3. Heat conductivities of X-26 and X-29

Temperature, °C	Heat conductivity, cal/s/cm²/°C	
	X-26	X-29
50	7.2×10^{-4}	8.3×10^{-4}
100	7.4×10^{-4}	8.4×10^{-4}
150	7.6×10^{-4}	8.6×10^{-4}
175	7.9×10^{-4}	8.9×10^{-4}

These data with the densities of the materials (X-26: 1.84 g/cm³; X-29: 2.10 g/cm³) provide the necessary constants for the evaluation of Eq. (1) for each firing. The interface temperatures computed are described below.

3. Results

The experimental curves for power input versus time to fire are shown in Fig. 2 for X-26 and X-29. They are typical squib firing plots in that they can be divided into a linear or constant energy regime, transition region, and finally a constant power line. Using Eq. (1) and the techniques described in the analysis section, these data may be transformed into an ignition temperature versus time plot as shown in Fig. 3. These figures graphically illustrate the fact that at short ignition-time intervals it is incorrect to assume or prescribe a constant ignition temperature to a given ignition material. The temperatures at ignition range from 2200 to 250°C for X-26 and 1350 to 240°C for X-29, dependent upon rate of heat input. The lower range temperatures correlate quite well with crucible furnace ignition data which indicates mass cook-off temperatures of 215°C for X-26 and 225°C for X-29.

In light of the marked dependence of ignition temperature on time to fire, it is plausible to anticipate an

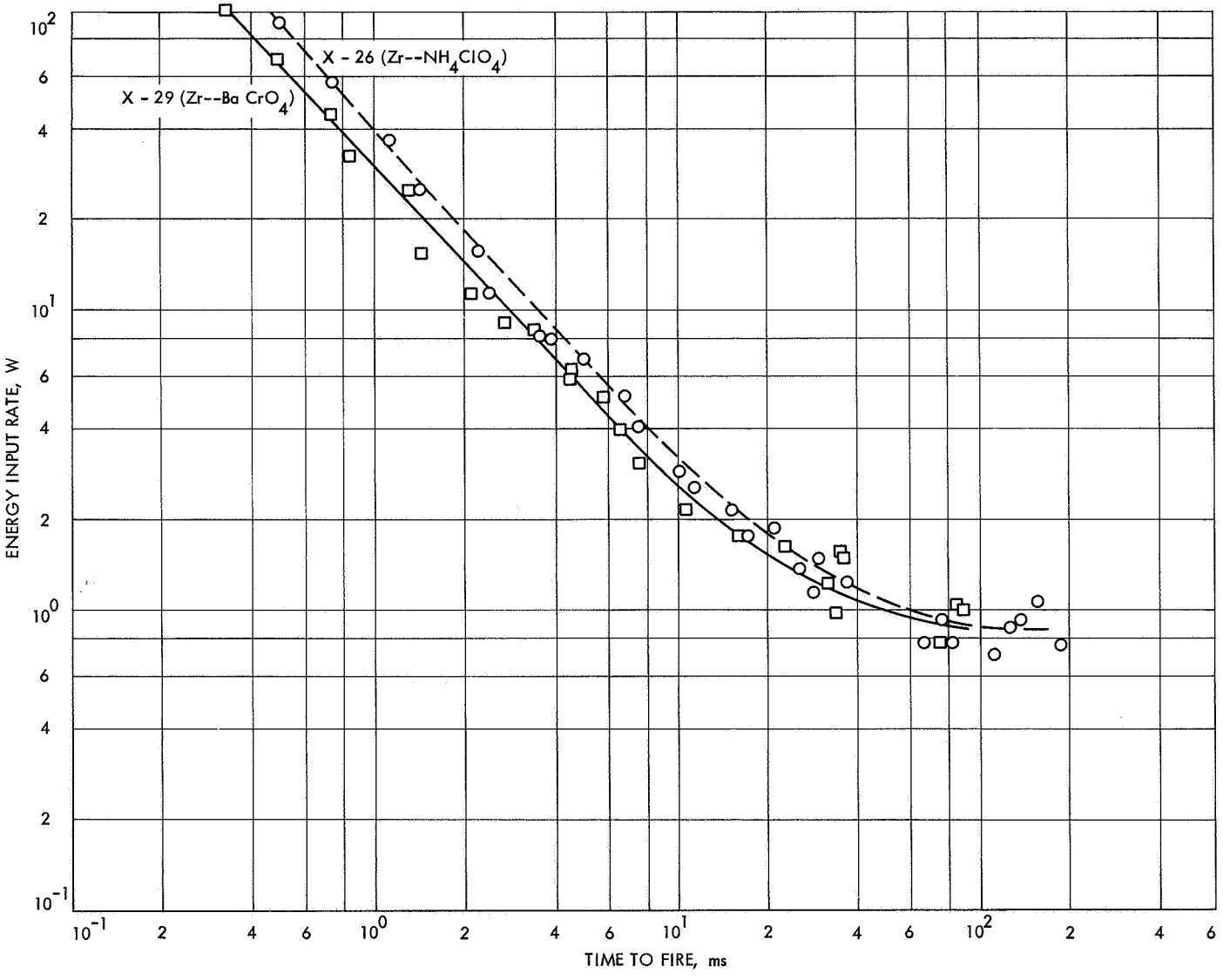


Fig. 2. Power versus time to fire Zr–NH_4ClO_4 (X-26) and Zr–$BaCrO_4$ (X-29)

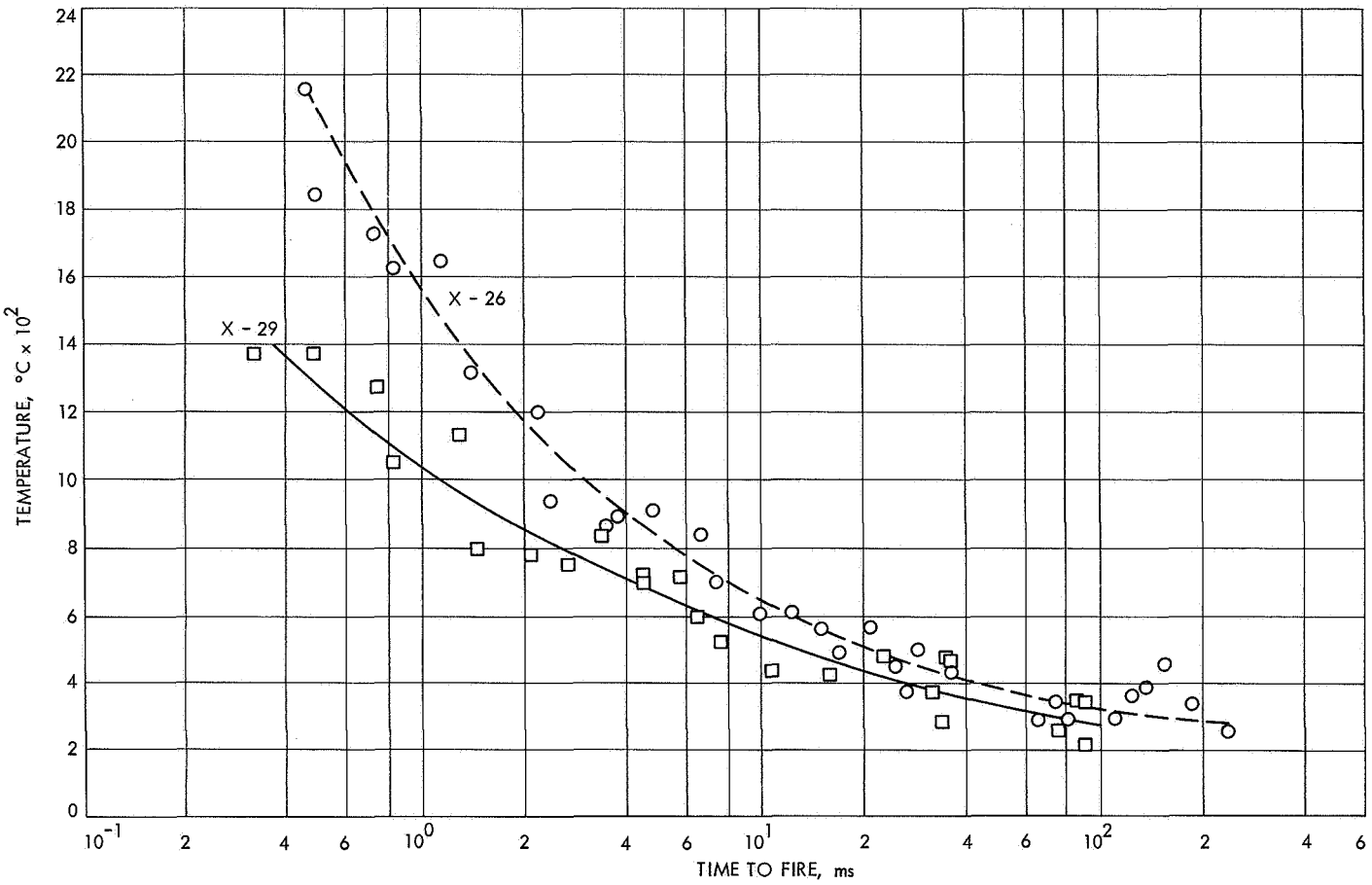


Fig. 3. Temperature at ignition versus time to fire Zr–NH_4ClO_4 (X-26) and Zr–$BaCrO_4$ (X-29)

Arrhenius-type time dependence for the ignition reaction; that is

$$\frac{dc}{dt} = r = Ae\frac{-E_a}{RT}$$

$$r = 1/\tau$$

Then it follows that

$$\ln \tau \sim \frac{1}{T}$$

With the exception of the monotonically increasing time-variant temperature, this argument is similar to that used to predict thermal explosion times for various reactive gas mixtures.

Figure 4 shows plots of ln τ versus the reciprocal temperature, each of which would be expected to display a straight line if the above argument were valid. It is seen that the points tend to be somewhat skewed, and no such simple relationship is identifiable.

4. Conclusions

The conclusions of this analytic and experimental effort are largely negative. No single ignition temperature was found; neither was there an explicit, simple, fundamental relationship between ignition temperature and time to fire. It should be realized that the conductivity and diffusivities are functions of temperature, and that the form of Eq. (1) requires that mean values be used. The effects in the numerator and denominator tend to stabilize the results, but the absolute temperature values of the very short firing time cases are inherently less accurate than the lower temperature cases. This fact does not, however, affect the soundness of the basic conclusion; which is that the prescribing of a single ignition temperature for short thermal induction times is neither accurate nor reasonable.

References

1. Carslaw, H. S., and Jaeger, J. C., *Conduction of Heat in Solids*, Oxford University Press, 1959.

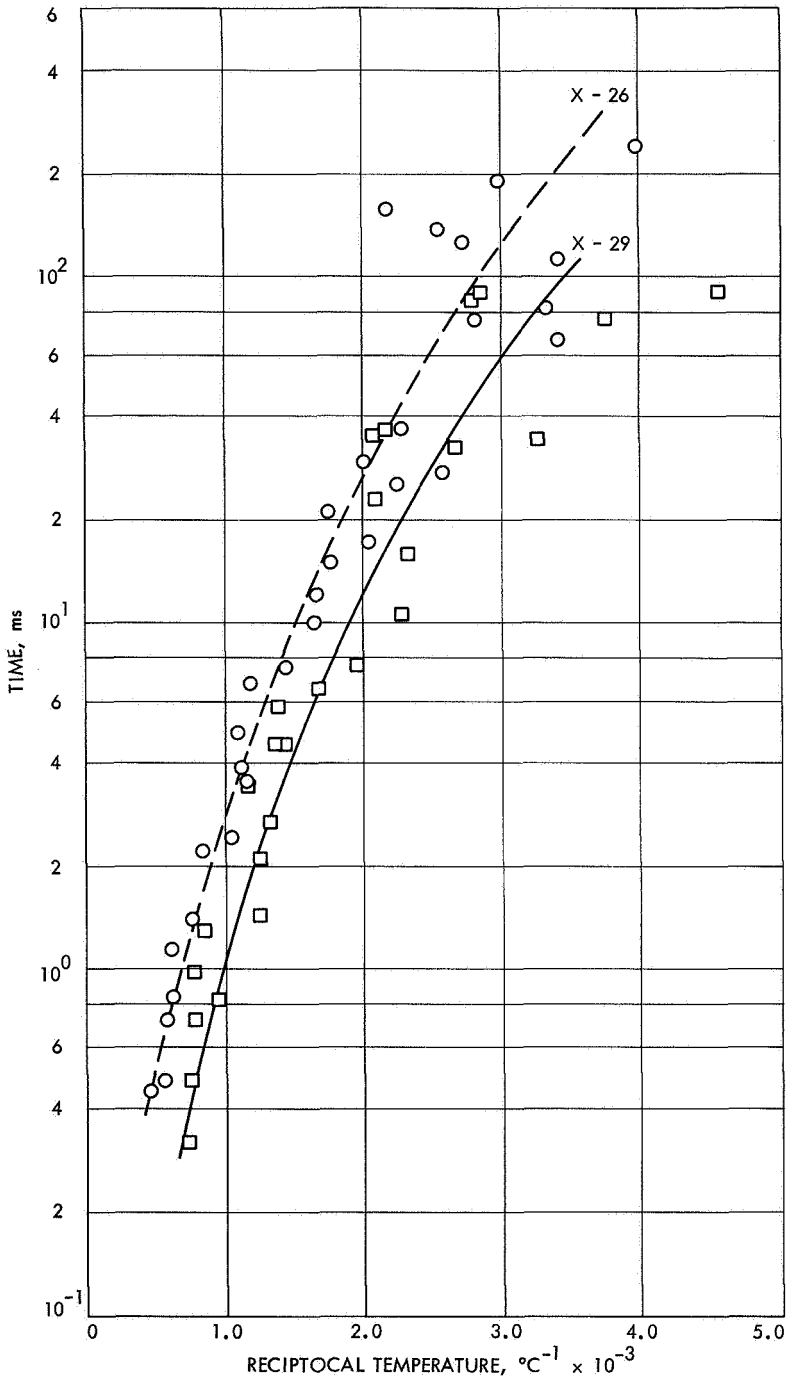


Fig. 4. Time to fire versus reciprocal of ignition temperature

2. Silverman, L., "Thermal Conductivity Data," J. Metals, Vol. 5, pp. 631–632, 1953.

3. Douglas, T. B., and Dever, J. L., *J. Res. NBS*, Vol. 54, No. 1, pp. 15–19, 1955.

4. *Thermal Conductivity of Materials by Means of the Guarded Hot Plate*, American Society of Testing Materials, ASTM Standard C177, 1963.

B. Applications Technology Satellite Motor Development, *R. G. Anderson and R. A. Grippi*

1. Introduction

Previous reports of progress on the development of the ATS motor have been published in SPS 37-20 to 37-33, Vol. V, SPS 37-34 to 37-45, Vol. IV, and SPS 37-47 to 37-49, Vol. III. As of this date the formal motor development and qualification phases are complete.

2. Program Status

Three flight units, processed in September 1966, were used to support the successfully launched ATS-B (December 1966) and ATS-C (November 1967) satellites. The spare flight unit (code Z-1) has been returned to JPL–Edwards Test Station (JPL–ETS) and placed into ambient storage. This spare unit will remain at ambient temperature conditions until September 1969, at which time it will be inspected, X-rayed, and static-fired. The testing of unit Z-1, 3 yr after it was cast, will extend the storage characteristics of the ATS apogee unit by 1 yr. The original storage program (SPS 37-48, Vol. III) verified motor integrity for a period of 2 yr by the testing of three units, one each after 16, 20, and 24-mo aging periods.

During March and April 1968, the last four ATS apogee flight units were cast. Unit Z-7 was processed to flight standards and static-tested at the JPL–ETS as a confirmation unit for the three remaining flight units (Z-4, Z-5 and Z-6). The fourth ATS satellite (ATS-D) was unsuccessfully launched on August 10, 1968. The remaining ATS satellite (ATS-E) is presently scheduled for an April–May 1969 launch.

3. Static Test of Motor Z-7

Apogee unit Z-7 was static-tested on May 9, 1968. This unit was flight quality and was processed to established flight-loading procedures. Prior to static testing, the unit was subjected to a temperature cycle of 10, 110, and 10°F and returned to ambient temperature for visual and radiographic inspection. The unit was static-fired with a grain temperature of 10°F while spinning at 150 rev/min. Postfire inspection of the motor hardware indicated normal performance and operation during its 43-s run. Table 4 lists the static test summary for this unit.

Table 4. ATS Apogee motor flight test motor Z-7 static test summary

Test conditions	
Type	Atmospheric–spin
Location	JPL–ETS
Date	May 9, 1968
Run No.	E-884
Grain temperature, °F	10
Propellant weight, lb	760.5
Pressure data	
Characteristic velocity, W*, ft/s	4959
Chamber pressure integral, psia-s	8905
Igniter peak pressure, psia (ms)	2032 (18)
Chamber ignition peak pressure, psia (ms)	259 (28)
Chamber starting pressure, psia (s)	103 (0.18)
Chamber run peak pressure, psia (s)	251 (33.5)
Time	
Ignition delay, ms	6
Run time, s	43.44
Nozzle dimensions	
Throat diameter, in.	
initial	4.085
final	4.106
average	4.096
Throat erosion area, %	1.03

4. ATS-D Launch

Two apogee flight units (Z-4 and Z-5) were shipped by ground transportation to the Air Force Eastern Test Range (AFETR) for support of the ATS-D launch.

After three weeks of flight preparation, the units received visual inspection, propellant grain X-ray, and motor assembly pressure testing. After these initial inspections unit Z-4 was returned to storage, and unit Z-5 was completed for flight use. Flight completion includes the mating of the igniter to the safe-and-arm device, the installation and orientation of the igniter to apogee motor, the installation of the apogee unit to the spacecraft, and the installation of the apogee motor's thermal blankets. When the apogee unit is secured to the spacecraft, JPL's

flight support functions are essentially complete. Table 5 lists the weight data for the Z-4 and Z-5 apogee units.

The ATS-D was launched on August 10, 1968. A malfunction in the *Centaur* engine prevented a second engine burn, leaving the spacecraft/apogee motor/*Centaur* in a 100- by 400-nm parking orbit. The Goddard Space Flight Center indicates that the orbiting package will reenter the earth's atmosphere approximately 2 mo after launch.

The ATS-D and the ATS-E (final satellite in the ATS program) are both synchronous-altitude gravity-gradient-stabilized satellites. The ATS-E is tentatively scheduled for April–May of next year. JPL has two apogee units to support this final launch. The Z-4 apogee unit is presently in storage at AFETR. The Z-6 apogee unit has been cast and is in storage at JPL–ETS. This unit will be inspected, assembled, and shipped to AFETR during the first quarter of 1969.

Table 5. Weight data for ATS-D apogee units

Part	Weight, lb	
	Unit Z-4[a] Chamber T-19 Nozzle F-33	Unit Z-5[b] Chamber T-21 Nozzle F-48
Chamber[c]	36.69	37.36
Nozzle[d]	38.10	37.15
Miscellaneous[e]	0.37	0.37
Igniter	1.00	1.00
Safe and arm	5.00	5.00
Total inerts	81.16	80.88
Propellant	759.50	760.00
Total motor weight	840.66	840.88

[a]Flight preference: backup for ATS-D and E.

[b]Flight preference: prime for ATS-D.

[c]Chamber includes chamber insulation, curing agent, and lead balance weight.

[d]Nozzle includes diaphragm and balance weight.

[e]Miscellaneous includes 36 screws and washers, O-ring, and nozzle lock wire.

IX. Polymer Research

PROPULSION DIVISION

A. Estimation of Solubility Parameters From Refractive Index Data, *D. D. Lawson and J. D. Ingham*

1. Introduction

For the evaluation of polymeric materials for propellant expulsion bladders, it is often desirable to have a quick and convenient method for estimating solubility parameters. The solubility parameter, or cohesive energy density, of a substance is an extremely fundamental physical constant. Inasmuch as it is a measure of intramolecular forces, it defines many characteristics besides mere solubility. In the case of cryogenic expulsion bladder materials, solubility parameters can be useful in the estimation of glass temperatures (T_g), and also may be used in the study of diffusion processes in elastomers (Refs. 1 and 2). It was suggested by J. Hildebrand and R. Scott (Ref. 3), and later by G. Scatchard (Ref. 4), that solubility parameters could be calculated from optical data such as refractive index measurements. In this article, some semi-empirical correlations between solubility parameters and refractive indices for a series of model compounds and polymers will be discussed.

2. Solubility Parameter Relationships

The potential energy of a mole of material (E) is $E = N_v$, where N is Avogadro's number and the potential energy of a molecule is ν. The cohesive-energy density is thus numerically equal to the negative potential energy of one cubic centimeter of the material ($-E/V$) where V is the molar volume. When solute–solvent systems are to be studied it is convenient to define the square root of the cohesive-energy density as the solubility parameter (δ).

$$\delta^2 = -\frac{E}{V} = -\frac{N\nu}{V} \tag{1}$$

The vaporization of a material can be imagined as a process involving the transport of all molecules from their equilibrium distance, where they have an equilibrium cohesive-energy density, to an effectively infinite distance relative to each other so that the potential energy of each molecule is reduced to zero. The heat of vaporization per mole (Hv) is thus the term used to compensate both for the potential energy per mole (E) and for the volume

work, which for a vapor phase obeying ideal gas laws is RT (where R is the molar gas constant and T is the absolute temperature). That is, $\Delta Hv = -E + RT$. It follows that the cohesive-energy density can then be obtained from Hv and V.

$$\delta^2 = \frac{Hv - RT}{V} \tag{2}$$

This is only true for materials that can be vaporized. In the case of polymers, other means can be used to reliably evaluate $-E/V$. In 1910, P. Walden (Ref. 5) derived an empirical relationship that relates the latent heat of vaporization to refractive index (n). This relationship is

$$lH \cong \frac{310}{M}\left(\frac{n^2 - 1}{n^2 + 2} V\right) \tag{3}$$

where lH is the latent heat of vaporization in cal/g and M is the molecular weight of the material being vaporized. This then can be rewritten so that

$$\Delta Hv \cong 310\left(\frac{n^2 - 1}{n^2 + 2} V\right) \tag{4}$$

Then, by substitution in Eq. (2),

$$\delta \cong \left[310\left(\frac{n^2 - 1}{n^2 + 2}\right) - \frac{RT}{V}\right]^{1/2} \tag{5}$$

By inspection of some data for many model compounds, it was obvious that each chemical class has a value somewhat different from 310. Using these compounds, a series of constants (C) were calculated from

$$C = \frac{Hv}{\left(\frac{n^2 - 1}{n^2 + 2} V\right)} \tag{6}$$

Equation (5) can then be changed to the more general form

$$\delta \cong \left[C\left(\frac{n^2 - 1}{n^2 + 2}\right) - \frac{RT}{V}\right]^{1/2} \tag{7}$$

Table 1 gives average values of C for seven different chemical classes using 57 model compounds. The average value for all 57 compounds is 304.5 which is remarkably close to that of 310 obtained by Walden.

Table 1. Values of C for different chemical types

Chemical type	C calculated from Eq. (6)	Number of materials of each type	Comments and data source
Normal aliphatic hydrocarbons	254.1 ± 2.1	10	This includes 4 alkenes (Refs. 7, 8, 13)
Branched aliphatic hydrocarbons	230.9 ± 9.3	6	Straight chain with methyl groups (Refs. 7, 8, 13)
Aliphatic ethers	279.2 ± 25.0	4	Refs. 7, 8, 13
Aliphatic esters	353.3 ± 30.9	6	Refs. 7, 8, 13
Normal Aliphatic fluoro-carbons	205.1 ± 9.5	4	Refs. 7, 11
Chloro-carbons	330.8 ± 53.1	12	3 methane derivatives and the chloroethanes and chloroethenes (Refs. 7, 12)
Aromatics	287.6 ± 10.4	15	2 aromatic fluorocarbons and 3 cycloalkane +3 cycloalkenes (Refs. 7, 8, 13)

3. Direct Estimation of the Solubility Parameter From Refractive Index Data

The simplest correlation of refractive index and δ is that obtained by a least-square curve fit of the δ of selected organic compounds and the Lorentz–Lorenz function (Ref. 6). By use of 18 compounds, a straight-line fit was obtained that passed through the origin and had a slope of 30.3. The corresponding point-slope-intercept equation is

$$\delta \cong 30.3\left(\frac{n^2 - 1}{n^2 + 2}\right) \tag{8}$$

In Eq. (7), the RT term is of the order of $\sim$600 cal and small when compared to the Hv term ($<10\%$). Thus, the second term can be dropped to give

$$\delta \cong \left[C\frac{n^2 - 1}{n^2 + 2}\right]^{1/2} \tag{9}$$

Equations (8) and (9) permit estimation of solubility parameters from refractive indices.

4. Solubility Parameters of Polymers Calculated From Refractive Index Data

Table 2 shows results for several polymers. The δ were calculated from Eqs. (8) and (9). It can be seen that δ from Eq. (8) do not agree very well with the literature values, but that Eq. (9) gives very reasonable values of δ. Since the second term of Eq. (7) is subtractive, a refined form of Eq. (9) could include a negative constant of the order of a few tenths of a cal/cm^3; however, the effect would be to increase the deviation from other δ values for a few polymers. A relatively arbitrary value of C used for the acrylates (304.5) was the average of the values of Table 1. This value was used primarily because the value determined from aliphatic esters did not result in as good an agreement. It can be concluded that Eq. (9) can be used to obtain satisfactory values of δ from the refractive index for most polymers if a reasonable estimate of C is available.

Table 2. Solubility parameters (δ) estimated from refractive index data

Polymer	Refractive index	δ^a, cal/cm^3	C value used	δ^b, cal/cm^3	δ, cal/cm^3 literature values
Poly(ethylene)	1.51	9.06	254.1	8.72	7.87–8.10 (Ref. 10)
Poly(isobutylene)	1.5089	9.04	230.9	8.31	7.80–8.05 (Ref. 7)
Natural rubber	1.5191	9.20	230.9	8.37	7.90–8.35 (Ref. 10)
Poly(butadiene)	1.5160	9.15	230.9	8.35	8.32–8.60 (Ref. 10)
Poly(styrene)	1.595	10.29	287.6	9.89	8.56–9.70 (Ref. 10)
Poly(methyl acrylate)	1.4725	8.50	304.5	9.24	9.8–10.4 (Ref. 7)
Poly(ethyl acrylate)			304.5	9.44	9.2–9.70 (Ref. 7)
Poly(propyl acrylate)			304.5	9.14	9.0–9.05 (Ref. 7)
Poly(butyl acrylate)			304.5	9.14	8.50–9.10 (Ref. 7)
Poly(vinylidene chloride)	1.63	10.78	330.8	10.39	12.2 (Ref. 7)
Poly(tetrafluoroethylene)	1.35	6.52	205.1	6.49	6.2 (Ref. 10)
Nitroso rubber	1.3170	5.96	205.1	6.36	5.2 (Ref. 9)

[a]Calculated from $\delta = 30.3\ [(n^2 - 1)/(n^2 + 2)]$.

[b]Calculated from $\delta = \{C[n^2 - 1)/(n^2 + 2)]\}^{1/2}$; the δ values for which refractive indices are not given were obtained from calculated molar refractivities divided by specific volumes given in Ref. 6.

References

1. Hayer, R. A., *J. App. Polymer Sci.*, Vol. 15, p. 318, 1961.
2. Van Amerongen, G. J., *Rubber Rev.*, Vol. 37, p. 1092, 1964.
3. Hildebrand, J., and Scott, R., *The Solubility of Nonelectrolytes*, Rheinhold Publishing Corporation, New York, 1948.
4. Scatchard, G., *Chem. Rev.*, Vol. 44, p. 24, 1949.
5. Walden, P., Z. *Phys. Chem.*, Vol. 70, p. 587, 1910.
6. Sewell, J. H., RAE TR 66185, Ministry of Aviation, Farnborough Harts, England, June, 1966.
7. Burrell, H., and Immergut, B., *Polymer Handbook*, Part IV, p. 341, Interscience Publishers division of John Wiley & Sons, Inc., New York, 1966.
8. Riddick, J. A., and Toaps, E. E., *Technique of Organic Chemistry: Volume VII. Organic Solvents*, pp. 43–258, Interscience Publishers division of John Wiley & Sons, Inc., New York, 1955.
9. Crawford, G. H., Rice, D. E., and Sandrum, B. F., *J. Polymer Sci.*, Part A-1, p. 565, 1963.
10. Sheeham C. J., and Bisio, A. L., *Rub. Chem. Tech.*, Vol. 149, 1966.
11. Lovelace, A. M., Rausch, D. A., and Postelnek, W., *Aliphatic Fluorine Compounds*, ACS Monograph Series No. 138, 1958.
12. Huntress, E. H., *Organic Chlorine Compounds*, John Wiley & Sons, New York, 1948.
13. Huntress, E. H., and Mulliken, S. P., *Identification of Pure Organic Compounds*, John Wiley & Sons, New York, 1941.

B. Cationic Crosslinking Agents—Potential Solid Propellant Binders, *A. Rembaum, A. M. Hermann, and H. Keyzer*

1. Introduction

The compounds formed by means of the reaction between an α-ω-dihaloalkane and a tertiary amine containing a double bond yield tetrafunctional monomers (SPS 37-50, Vol. III, pp. 161–165). The latter have positive nitrogens in their structure and therefore may be classified as cationic crosslinking agents. Viscoelastic materials are formed at room temperature with high molecular weight α-ω-dihaloalkanes. This process occurs equally well in presence of large amounts of oxidizers. These materials, characterized by a glass transition temperature of about $-80°C$, show promise as solid propellant binders since they form rubbery products containing positively charged nitrogen atoms.

Low molecular weight α-ω-dihaloalkanes serve as model compounds and are examined here. The investigations

of the viscoelastic materials of more direct interest will be described at a later date.

The synthesis and some characterization details including preliminary nuclear magnetic resonance (NMR) data of cationic crosslinking agents were discussed in SPS 37-50, Vol. III. The present article contains additional analytical NMR results as well as an electron spin resonance (ESR) study of cobalt γ-irradiated crosslinking agents. The radicals obtained during irradiation were identified by means of ESR and the high stability of these materials under the influence of cobalt γ radiation was established.

The irradiation of cationic crosslinking agents in the solid state by means of the cobalt γ source yields a crosslinked polymer, the conversion of monomer to polymer increasing with the radiation dose. The free radical concentration is found to increase initially with the radiation dose, but, in contrast with the percent conversion, it decays rather sharply after reaching a maximum at a dose of about 24 Mrad. The ESR spectra obtained during polymerization under irradiation were found to be identical to those of the crosslinked material formed by polymerization of the monomer using sodium bisulfite and ammonium persulfate as initiator. Thus, the radicals formed during irradiation of monomer are the same as those observed in the irradiation of polymer.

The ESR spectra of methacrylate and acrylate crosslinking agents were very similar to those previously discussed (SPS 37-50, Vol. III) for irradiated methacrylate and acrylate polymers of different structure. This permits a conclusive identification of radicals presently observed.

The crosslinking agent containing an allyl group also yielded an ESR spectrum (after a dose of 50 Mrad) in spite of the fact that no polymerization occurred. This is consistent with the well known general behavior of allyl monomers and with the type of radical identified by means of ESR. In order to gain some insight into the nature of radical disappearance during irradiation, the rate of decay was studied as a function of temperature.

2. Experimental

The NMR spectra were taken at room temperature using the Varian A60 spectrometer. The ESR spectra were taken with the Varian 4502 spectrometer (X band) using 100-kHz modulation. Irradiation doses at room temperature (for the identification of the hyperfine spectra and decay kinetics) were of the order of 2 Mrad; the samples were transferred into quartz tubes after irradiation in pyrex containers.

3. Results

a. NMR analysis. The NMR spectra of a series of crosslinking agents prepared from dimethylaminoethyl acrylate (DA) and methacrylate (DMA) are shown in Fig. 1. The bromine analysis (SPS 37-50, Vol. III) and the good agreement between the integrated intensity of proton absorption (Fig. 1) confirms the postulated structures. It should be noted that the reaction between dibromomethane and DMA leads to a difunctional compound containing a non-ionic bromine atom (Fig. 1f). In this case, only one molecule of DMA reacts because the formation of a cationic crosslinking agent would require presence of two quaternary nitrogens separated by only one CH_2 group. The repulsive interaction of ionic charges is evidently responsible for the formation of a difunctional instead of a tetrafunctional compound. Diallyl crosslinking agents were prepared by a modification of the previous method (SPS 37-50, Vol. III). Allyl bromide was reacted with tetramethylaminoethane and tetramethylaminohexane. The bromine analysis (44.6 and 38.6% theoretical, and 44.3 and 38.3% actual, respectively), as well as the proton integration of the NMR spectra, was found to be in excellent agreement with the theoretical structures (Fig. 2).

b. ESR data. Figure 3 shows the percent conversion and the corresponding ESR intensity with irradiation dose. Figure 4 records the solid state ESR spectra and structure of cationic crosslinking agents. Figure 5 shows the first-order decay of the ESR signal of the dimethacrylate compound (see structure in Fig. 1a) at room temperature. The rate constant calculated from the slope of the figure is 4.8×10^{-4} s^{-1}. A similar study was carried out at 110°C both in air and in vacuum. In both cases, the decay was more rapid than at room temperature and obeyed second-order kinetics (Fig. 6).

The solution of the standard second-order differential equation describing such decay can be cast into the form (Ref. 1)

$$\log\left(1 - \frac{a-b}{N_s}\right) = \frac{b-a}{2.303}\,kt + \log\frac{b}{a}$$

where N_s represents the concentration of unpaired spins at any time t, a is the initial concentration of unpaired spins, b is the initial concentration of the other reactant (presumably oxygen), and k is the rate constant. The

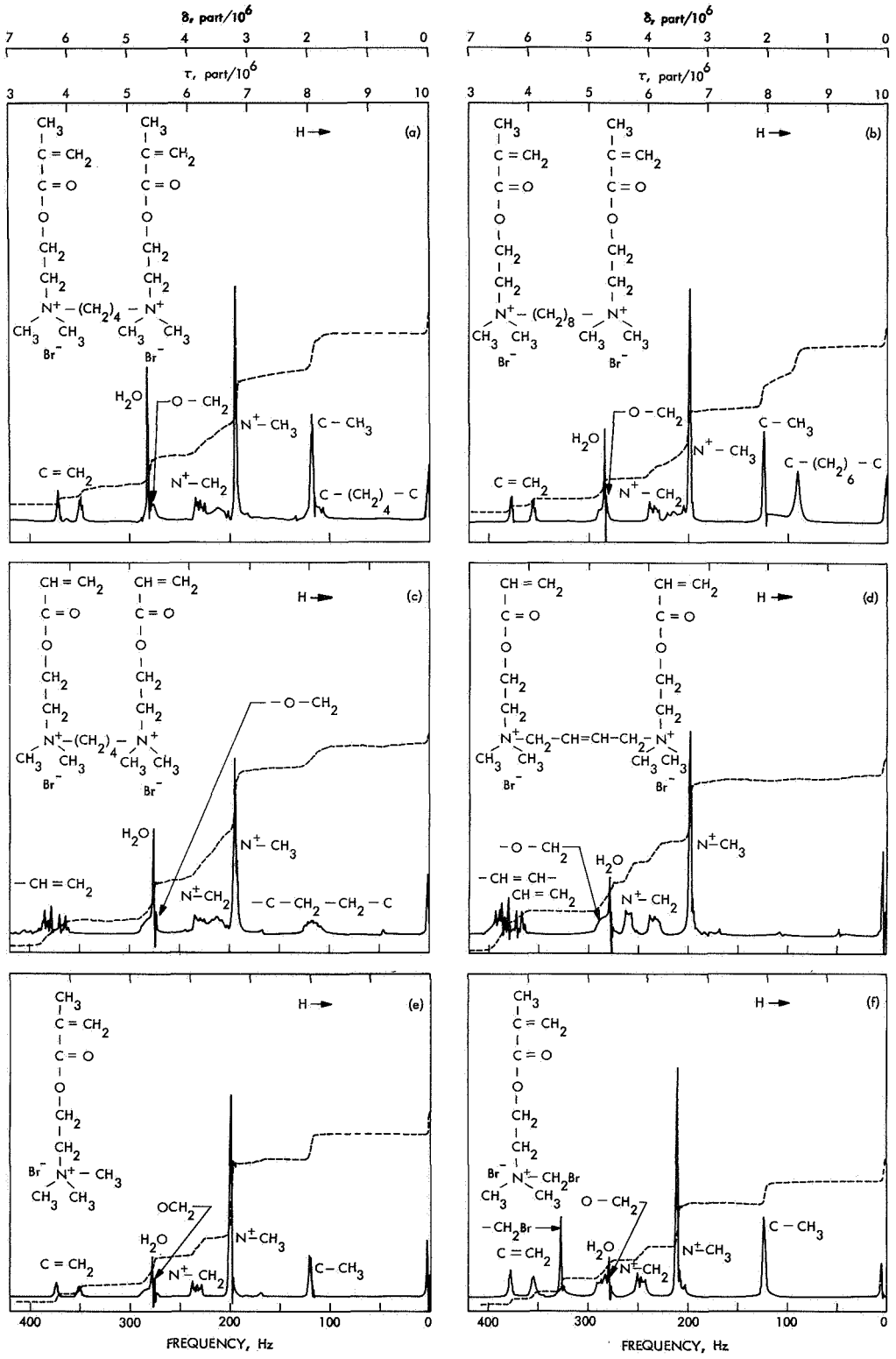


Fig. 1. NMR spectra of methacrylate and acrylate crosslinking agents

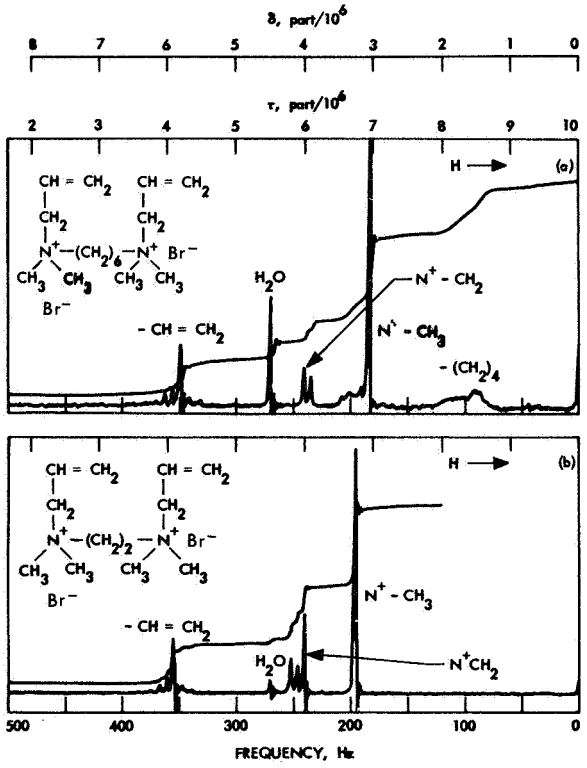


Fig. 2. NMR spectra of allyl crosslinking agents

values of a, N_s, and t are easily measurable. That value of b is chosen to give the best straight line fit on a semi-log plot of $[1 - (a - b)/N_s]$ versus t. In the case of the open-to-air ESR tube (Fig. 6, case A), the value of b required for a straight-line fit is greater than that of a (consistent with excess oxygen), and hence the positive slope. In the case of the evacuated ESR tube (Fig. 6, case B), the value of b required for a straight-line plot is less than that of a (consistent with excess free radicals), and hence the negative slope. Furthermore, it should be mentioned that the rate constant for case A, 128 liters mole^{-1} s^{-1}, is far greater than case B, 3.8 liters mole^{-1} s^{-1}. In terms of raw data, in 1 hr the ESR signal decays to about 2% of its initial value in case A, and to about 50% of its initial value in case B.

4. Discussion

a. ESR spectrum of a methacrylate crosslinking agent. The principal features of the ESR spectrum (see structure in Fig. 1a) are five major lines of approximate intensity ratios 1:4:6:4:1, each line separated from the next by 23.5 G (Fig. 4a). Also observable are two smaller peaks on either side of the central line. While the five major hyperfine lines would suggest interaction with four equivalent protons, one must have further evidence to

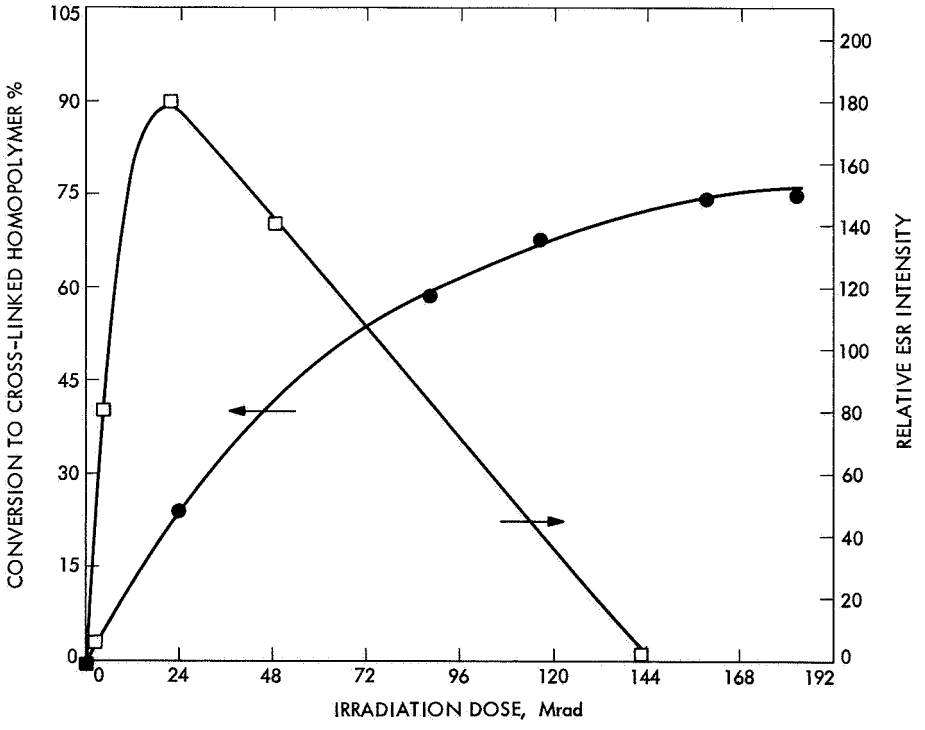


Fig. 3. Percent conversion and relative ESR intensity as a function of γ-irradiation dose for a crosslinking agent (Structure Fig. 1a)

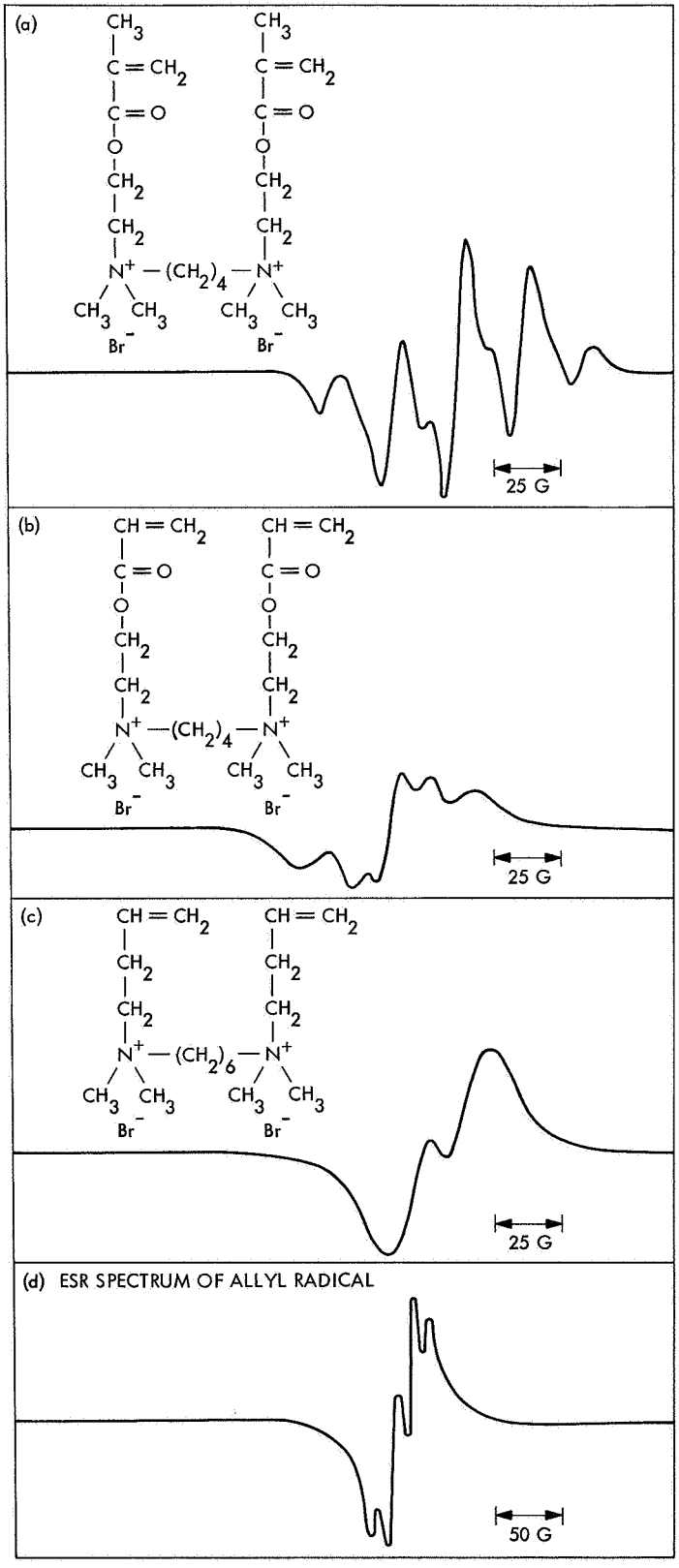


Fig. 4. ESR spectra of γ-irradiated crosslinking agents

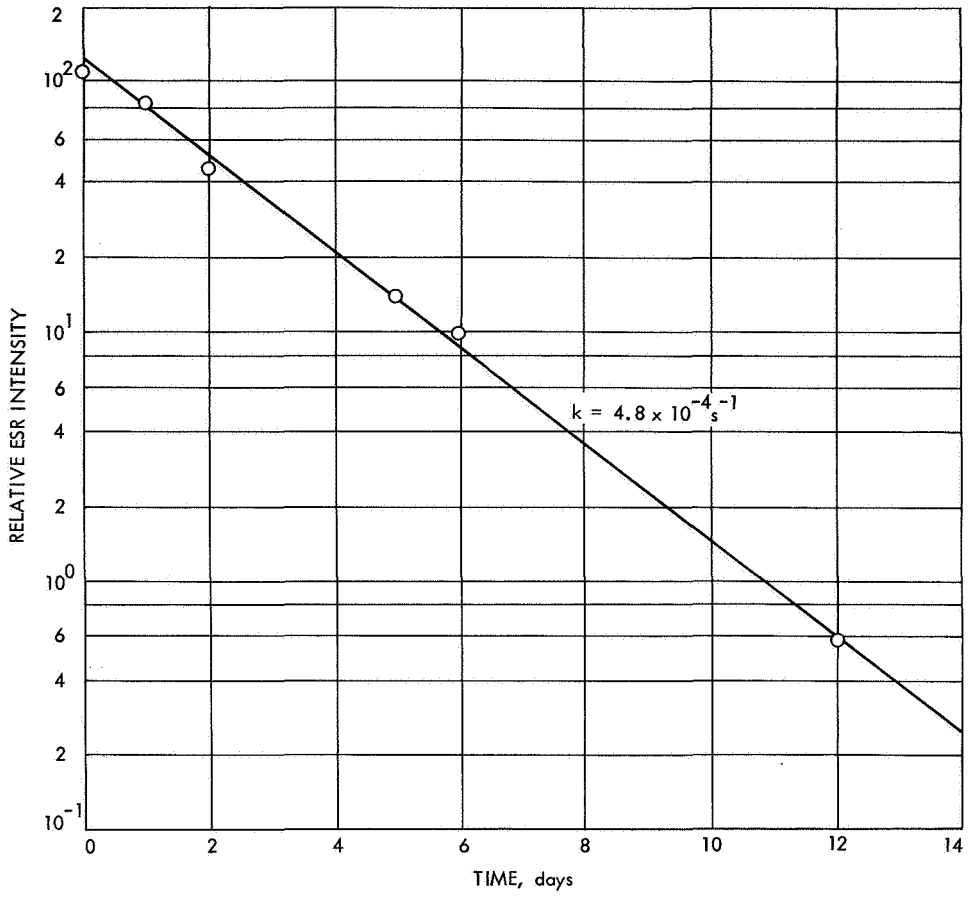
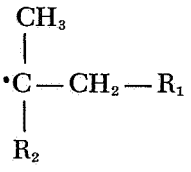


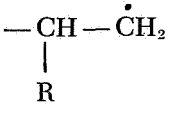
Fig. 5. First-order decay of γ-irradiated crosslinking agents (Structure Fig. 1a)

correctly interpret the unresolved structure. A nearly identical spectrum of irradiated solid methacrylic acid has been reported by H. Fischer (Ref. 2). Fischer has also reported the spectrum of the polymerization radicals of methacrylic acid taken in aqueous solution, thereby affording excellent resolution of each of 16 hyperfine lines giving conclusive evidence for the radical identification. On the basis of Fischer's studies, as well as those of M. C. R. Symons (Ref. 3), the radical whose ESR spectrum is shown (Fig. 4a) is identified as



Therefore, it is identical to the radicals previously detected during free radical polymerization of methacrylic monomers.

b. ESR spectrum of an acrylate crosslinking agent. The principal features of the spectrum (see structure in Fig. 1b) are three lines of approximate intensity ratios 1:2:1 separated by about 28 G (Fig. 4b). In addition, there are two other resolved lines on either side of the central peak. This spectrum appears quite similar to that found by R. J. Abraham and D. H. Whiffen (Ref. 4) for γ-irradiated solid polyacrylic acid. On this basis, the radical is identified as



c. ESR spectrum of an allyl crosslinking agent. The principal feature of the spectrum (see structure in Fig. 1c) appears to be two equally intense lines of 21-G separation, but previous evidence suggests this to be superposition of at least four lines (Ref. 5). On the basis of previous studies of the allyl radical (Ref. 5 and Fig. 4d),

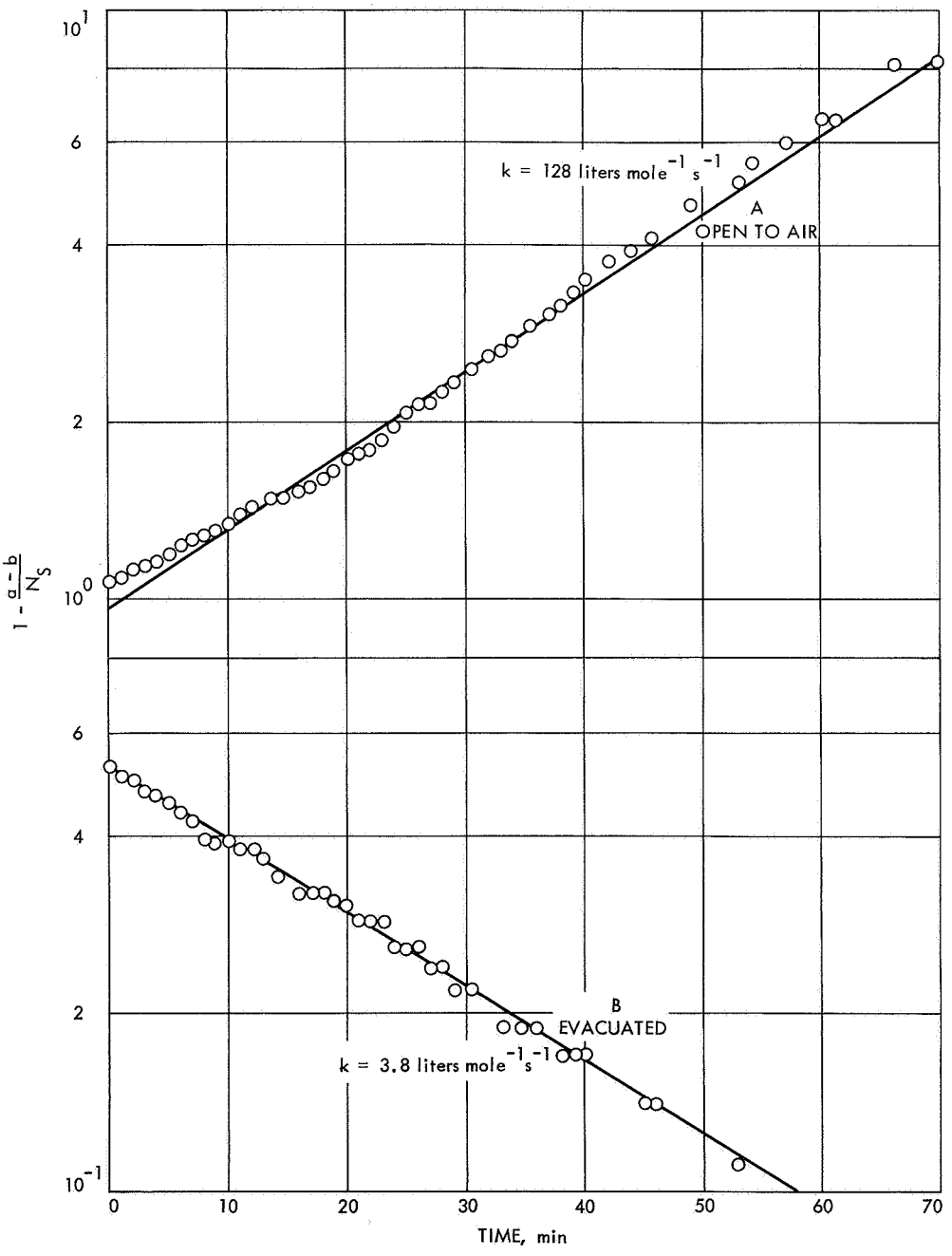


Fig. 6. Second-order decay of γ-irradiated crosslinking agent (Structure Fig. 1a at 110°C)

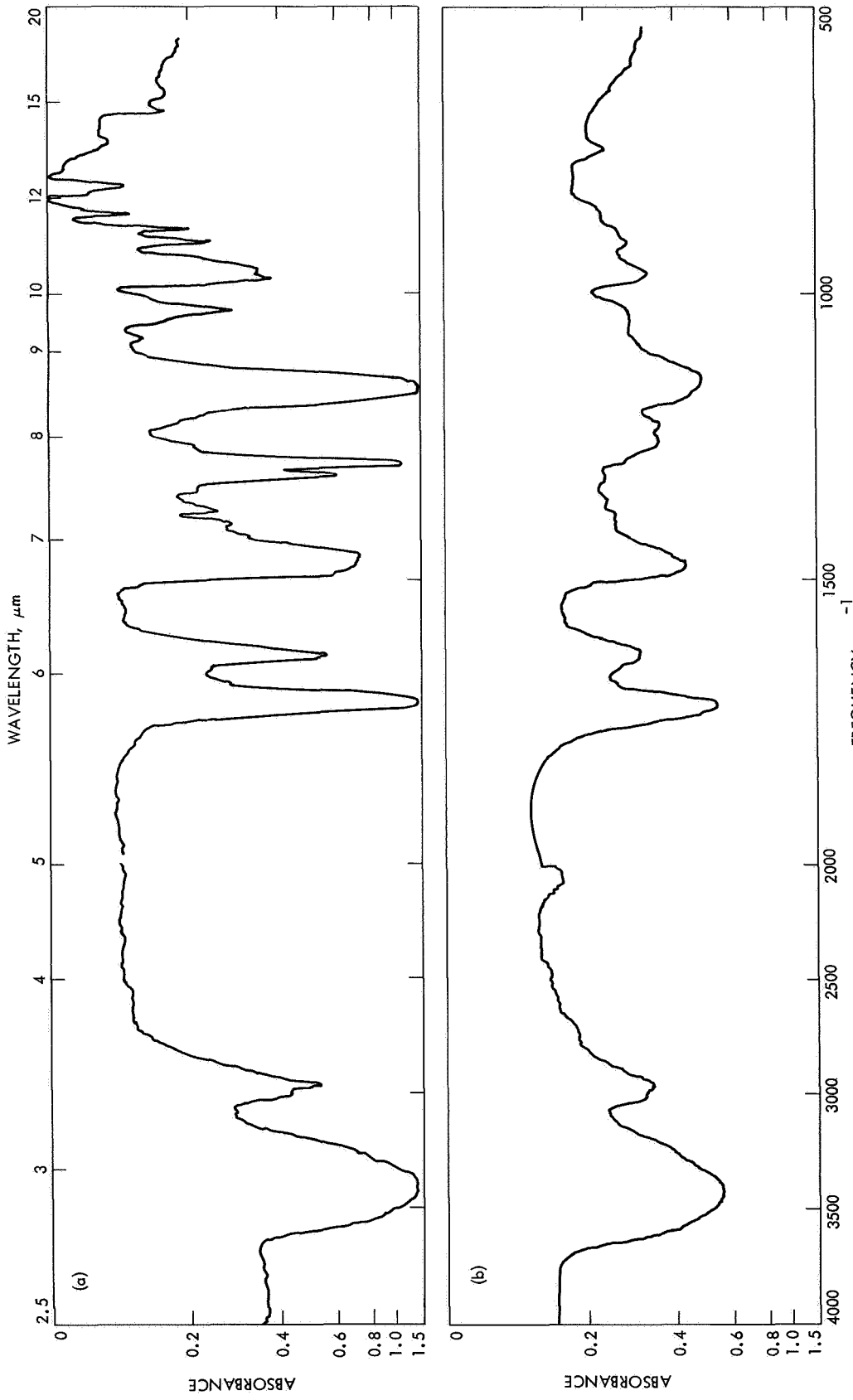
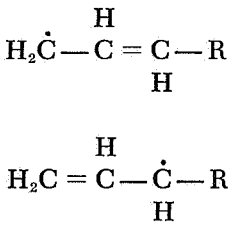


Fig. 7. IR spectra: (a) monomer (structure Fig. 1a) and (b) γ-irradiated polymer

the ESR spectrum (Fig. 4c) is interpreted as being due to a radical whose structure is



It is to be noted that the allyl radical formed by hydrogen abstraction is stabilized by resonance, a feature responsible for the lack of homopolymerization of the diallyl crosslinking agent.

The relative stability of these crosslinked polymers under cobalt radiation became apparent during this study. Solvent extraction and liquid gas chromatography showed an absence of volatile products after a 140-Mrad dose. The IR spectrum after this radiation treatment (Fig. 7) was identical to that obtained by redox polymerization of the monomer.

References

1. Daniels, F., *Physical Chemistry*, John Wiley & Sons, Inc., New York, 1948.
2. Fiścher, H., *J. Polymer Sci.*, Part B, Vol. 2, p. 529, 1964.
3. Symons, M. C. R., *J. Chem. Soc., (London)*, p. 1186, Feb. 1963.
4. Abraham, R. J., and Whiffen, D. H., *Trans. Faraday Soc.*, Vol. 54, p. 1291, 1958.
5. Fujimoto, M., and Ingram, D. J. E., *Trans. Faraday Soc.*, Vol. 54, p. 1304, 1958.

C. Evidence for Activated Carrier Mobility in Organic Solids, *F. Gutmann, A. M. Hermann, and A. Rembaum*

Steady-state space charge limited currents (SCLC) were obtained at different temperatures for a series of dipyridylium model compounds (Fig. 8) into which one TCNQ molecule (tetracyanoquinodimethane), in form of a radical ion, was introduced. Some compounds were produced containing two molecules of TCNQ, one associated with each ring. Polymers of these compounds were also produced and SCLC measured. The details of preparation will appear in a forthcoming issue of the *Journal of Physical Chemistry*.

The conductivity measurements were made on 0.5-in. diam cylindrical pellets. Pellets prepared under pressures

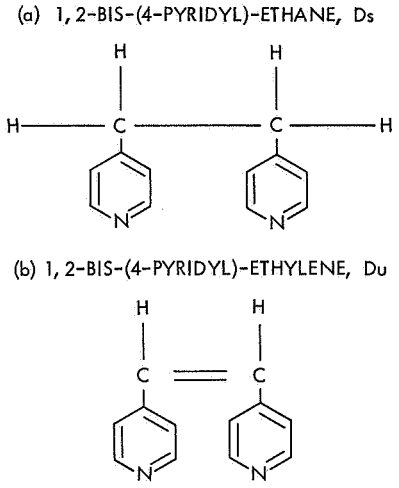


Fig. 8. Dipyridylium model compound chemical structure

in vacuum between 20,000 and 100,000 lb/in. had essentially identical conductivities. Electrical contact was made with vacuum-deposited gold electrodes or, in some cases, contacts were applied by covering the top and bottom surfaces with a thin layer of gold powder followed by recompression in the hydraulic press. Both processes resulted in firmly adherent, cohesive contacts ohmic at voltages below those at which appreciable charge injection occurred. In only one case was it possible to carry out measurements on a single crystal; the resulting activation energy (0.129 eV) is close to that obtained with the compactions (0.103 eV).

The conductance measurements were carried out in an evacuated glass cell (containing a thermocouple) immersed in a dewar vessel containing the temperature bath. Voltages up to 550 V were obtained from a Hewlett–Packard regulated power supply and those higher from a rectified power supply capable of delivering 5 mA. Currents and voltages were measured by means of Hewlett–Packard vacuum-tube voltmeters and Keithley electrometers.

The concentrations of free carriers at thermal equilibrium, n_{co}, were obtained (Ref. 1) from the transition voltage, V_{tr}, from ohmic to parabolic voltage dependence evaluated graphically using

$$n_{co} = \frac{2\epsilon\epsilon_0 V_{tr}}{et^2} \tag{1}$$

where e is the charge of the electron, ϵ_0 the permittivity of free space, ϵ the relative permittivity, and t is the inter-electrode spacing. With knowledge of the carrier concentration and the conductivity σ, the mobility may

then be obtained from

$$\mu = \frac{\sigma}{en_{co}}, \qquad \text{cm}^2/\text{V-s} \tag{2}$$

The results are summarized in Fig. 9. It is seen that the carrier concentration remains substantially constant to plus or minus an order of magnitude, at about 10^{11} cm^{-3}, over a temperature range in which the resistivities, some of which are also shown, change by up to eight orders of magnitude. In view of the uncertainties involved in the graphical location of V_{tr} and the probable changes in the effective permittivity at low temperatures, the values of n_{co} are estimated to be accurate only to within a factor of five. The largest apparent change in n_{co} was found in the unsaturated Du 1-TCNQ compound, which appeared to drop from 4.5×10^{11} at $-187°$C to 1.2×10^{10} at $-78°$C. In that temperature interval, the conductivity increased by five orders of magnitude.

While the temperature dependence of the mobility deduced from SCLC data could be fitted to a shallow-trap model (Ref. 2), this would require mobilities substantially larger than those shown in Fig. 9. These would

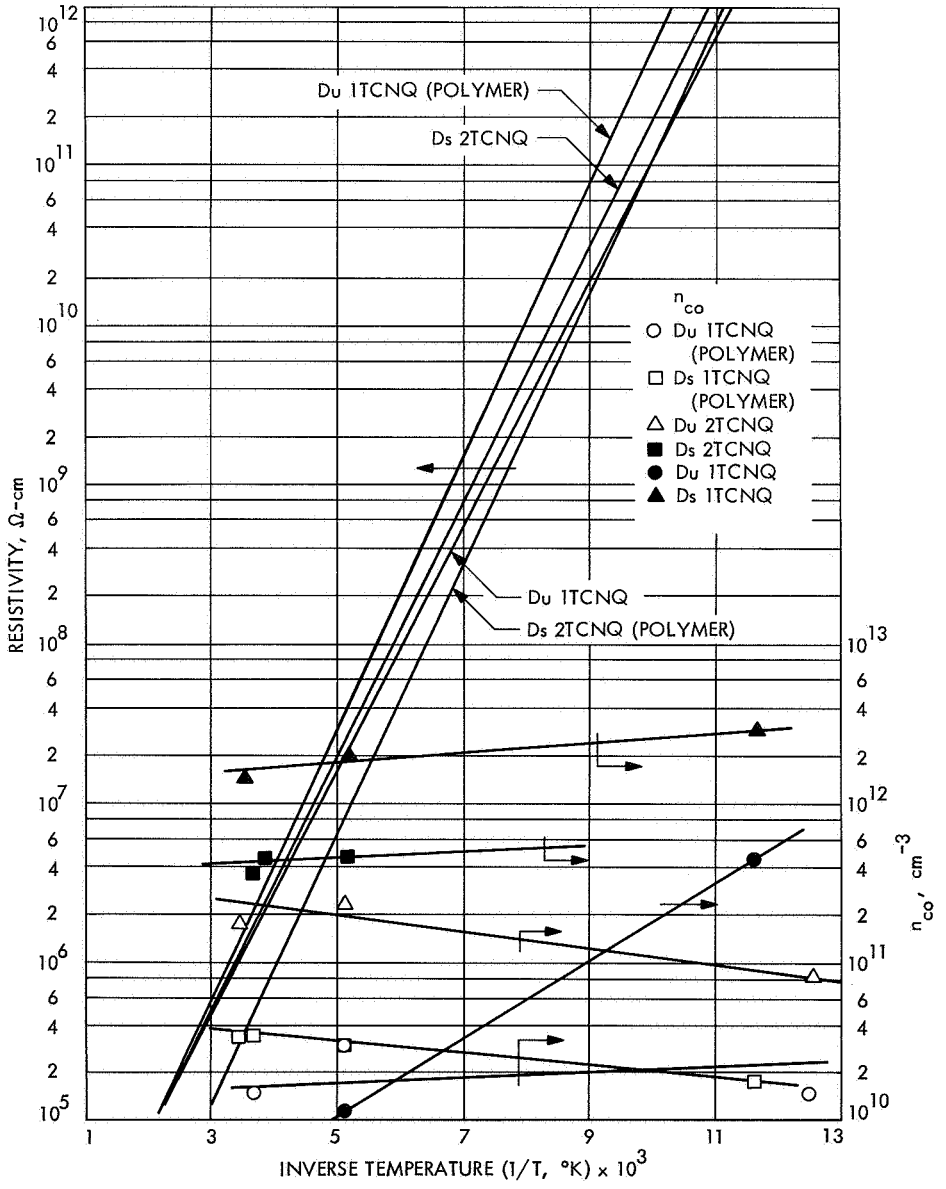


Fig. 9. Temperature dependence of resistivity (ρ) and concentration of carriers (n_{co})

not only be in greater discrepancy with Hall mobility measurements, but would require mobilities ($\gtrsim$ 40,000 cm²/V-s larger than those found in most single-crystal inorganic semiconductors; mobilities this large are extremely unlikely in systems lacking long-range order.

Thus, the conclusion is unavoidable that the observed conductivity changes are due to mobility changes. Such a thermally-activated mobility in organic materials has been proposed repeatedly (Refs. 3–8), but, to our knowledge, this work presents the first evidence for its existence.

It is of interest that, again to plus or minus an order of magnitude, the carrier concentration values appear to be invariant with respect to saturation versus unsaturation, introduction of a second TCNQ molecule, and polymerization. In fact, carrier concentration of the order discussed herein has also been observed by other researchers (Ref. 9).

References

1. Lampert, M. A., Rose, A., and Smith, R. W., *J. Phys. Chem. Solids*, Vol. 8, p. 464, 1959.
2. Rose, A., *Phys. Rev.*, Vol. 97, p. 1538, 1955.
3. Frolich, H., and Sewell, G. L., *Proc. Phys. Soc. London*, Vol. 74, p. 643, 1959.
4. Tredgold, R. H., *Proc. Phys. Soc. London*, Vol. 80, p. 807, 1962.
5. Pohl, H. A., and Opp, D. A., *J. Phys. Chem.*, Vol. 66, p. 2121, 1962.
6. Hadek, V., Ulbert, K., *Coll. Czech. Chem. Comm.*, Vol. 32, p. 1118, 1967.
7. Hermann, A. M., and Rembaum, A., *J. Polymer Sci.*, Part C, Vol. 17, p. 120, 1967.
8. Cherry, R. J., *Quart. Rev.*, Vol. 22, p. 162, 1968.
9. Gutmann, F., and Lyons, L. E., *Organic Semiconductors*, John Wiley & Sons, Inc., New York, 1967.

D. The Ethylene Oxide–Freon 12 Decontamination Procedure: The Control and the Determination of the Moisture Content of the Chamber,

R. H. Silver and S. H. Kalfayan

1. Introduction

The evaluation of several commercially available humidity sensing and controlling instruments for use in an ethylene oxide (ETO)–Freon 12 atmosphere was discussed in SPS 37-52, Vol. III, pp. 101–105. Instruments evaluated and discussed included the electrical-resistance, electrical-impedance, and optical–cold mirror types. None of the sensors were wholly satisfactory for service in the ETO–Freon 12 environment.

Since the last article, another optical–cold mirror type instrument was evaluated. The sensor of this instrument proved to be the most satisfactory tested so far. The results of these tests are described herein.

2. Experimental

The sensor of the test instrument, obtained from Cambridge Systems, Inc., is shown in Fig. 10. The general class of this type of humidity sensor is described in SPS 37-52, Vol. III, and Refs. 1 and 2.

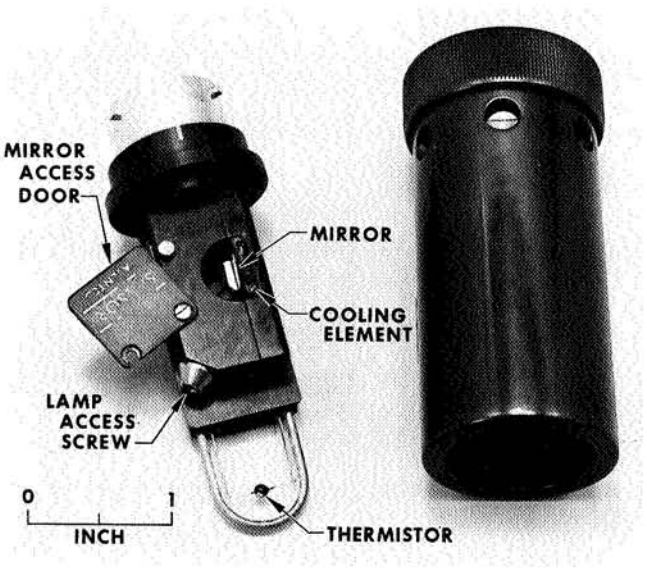


Fig. 10. Optical–cold mirror type sensor — Cambridge Systems, Inc., Model 137-S1-TH

The test procedure consisted of measuring the dew point in air at room conditions both before and after exposure of the sensing element to the ETO–Freon 12 decontamination chamber. Measurements were made simultaneously or within the minute. The computed percent relative humidity (RH) values were compared with those obtained by the manual dewpointer (Alnor Instrument Company, Model 7000 U), which served as a reference standard and is considered to be capable of measuring RH with an accuracy of $\pm 2\%$.

3. Results and Discussion

Because the reference standard is not suitable for use in the ETO–Freon 12 chamber, the desirable performance evaluation of the test instrument *in situ* could not be made. Consequently, the readings taken in air after the test sensor was exposed to ETO–Freon 12 were only a measure of the effects of the ETO–Freon 12 exposure

on the sensor. However, dew point or RH measurement in ETO–Freon 12 at the decontamination temperature of 50°C were made, and the results compared with those obtained by the electrical-resistance type hygrometer (SPS 37-52, Vol. III).

Table 3 gives the differences between the RH values obtained by the test sensor and the reference standard, $\Delta_{RH} = (RH_{test\ sensor} - RH_{standard})$. The Δ_{RH} values for the electrical-resistance type sensor are also included. The latter type is commonly used for RH determinations in the ETO decontamination chambers. The Δ_{RH} values obtained for the optical–cold mirror type sensor were low, indicating a close agreement with the readings of the reference standard. The Δ_{RH} values were generally positive. The Δ_{RH} values obtained for the electrical-resistance type sensor, except for the unexposed condition, were relatively high and consistently negative.

Table 3. Difference in RH values obtained by test sensors and reference standard

ETO–Freon 12 chamber exposure, h	Δ_{RH} electrical-resistance type sensor	Δ_{RH} optical–cold mirror type sensor	RH of air during measurements
0 (unexposed)	2.6 ± 1.4	1.6 ± 0.9	43.0 – 46.0
60	8.2 ± 0.6	1.0 ± 0.05	52.0 – 53.0
180	12.6 ± 0.8	1.3 ± 0.5	32.5 – 35.0
360	10.2 ± 0.01	1.3 ± 0.3	47.0

The percent RH readings from the electrical-resistance type sensor were 10–15 points lower than those from the optical–cold mirror type when both were placed in the ETO decontamination chamber operating at 50°C. Table 3 shows that this kind of difference in range was also obtained when readings were made in air, at room conditions, after exposure to ETO–Freon 12.

4. Conclusions

(1) The sensitivity of the optical–cold mirror type sensor was not impaired after long exposure to the ETO–Freon 12 environment.

(2) The relative humidity readings made by this instrument agreed closely with those made by the reference standard.

(3) The sensitivity of the electrical-resistance type sensor was changed by exposure to the ETO–Freon 12 atmosphere, and the RH readings thereafter were consistently lower by 8–13 points than those of the reference standard.

(4) There was no indication, however, that a continuous deterioration of the sensor took place by prolonged exposure to the ETO–Freon 12 decontamination atmosphere.

References

1. *General Catalog,* Cambridge Systems, Inc., Newton, Mass., 1968.
2. Paine, L. C., and Farrah, H. R., "Design and Application of High Performance Dewpoint Hygrometers," in *Humidity and Moisture, Vol. I,* pp. 174–188. Edited by A. Wexler, Reinhold Publishing Company, New York, 1965.

E. Dependence of Relative Volume on Strain for an SBR Vulcanizate, *R. F. Fedors and R. F. Landel*

1. Introduction

Although the importance of direct measurements of the volume change of an elastomer on stretching is readily recognized, very little data (especially at large strains) appear in the literature. The data that have been published are almost exclusively limited to either natural rubber gum vulcanizates cured with peroxides, which are known to undergo significant stress-induced crystallization at sufficiently high strains, or to noncrystallizable elastomers which, however, contained active fillers such as carbon black (Refs. 1–4).

Both the occurrence of crystallization and the presence of filler make at least some portions of the volume-extension response difficult to interpret. For example, L. Mullins and N. R. Tobin (Ref. 2) working at extension ratios of up to 5 find that for a natural rubber vulcanizate cured with peroxide, the relative volume first increases with strain, then passes through a very broad region containing a maximum, and finally goes through zero to become negative as the strain is further increased. Near the maximum, the volume increase due to the hydrostatic component of the stress becomes comparable to the volume decrease accompanying stress-induced crystallization.

Beyond the maximum, which for the data of Mullins and Tobin occurs at an extension ratio of about 3.5, the volume decrease due to crystallization swamps out the pure hydrostatic contribution. It is likely that crystallization occurs at extension ratios less than that at which the maximum appears, and a legitimate question to ask is: What is the strain interval for which stress-induced crystallization is negligible and hence can be safely ignored?

Since the hydrostatic component of stress gives rise to relative volume changes of the order of 10^{-4}, even small extents of crystallization can contribute significantly to the observed volume change. For example, assuming the density of the crystalline phase to be 0.97 g/cm³ and the density of the amorphous phase to be 0.91 g/cm³ (Ref. 5), only about 0.16% of the material would have to undergo stress-induced crystallization in order to produce a relative volume change (decrease) of 10^{-4}. Such small extents of crystallization would be difficult to detect, especially if the gross stress–strain behavior of the material were used as the sole criterion.

Using more sensitive X-ray methods, S. G. Nyburg (Ref. 6) reports, for a natural rubber-peroxide vulcanizate, a threshold extension ratio of about 3.5 below which no stress-induced crystallinity was observed. However, the usual X-ray determination of crystallinity is not particularly sensitive to very small extents of crystallization. Using low-angle light scattering, W. Yau and R. S. Stein (Ref. 7), working with a sulfur-cured natural rubber vulcanizate, report the presence of heterogenities that parallel the development of stress-induced crystallization at extension ratios of 2 to 3. This range is below the extension ratio at which crystallinity is first detected by X-rays.

A few years ago, volume–extension data were reported for a noncrystallizable styrene–butadiene copolymer (SBR) vulcanizate (SPS 37-18, Vol. IV, pp. 113–120) and, apparently, these still represent the only published data for such a system taken to large strains. Since the data were not fully discussed in the preliminary account, the purpose of this article is to more adequately describe and interpret the significance of the data.

2. Experimental Part

a. Material and characterization. The material studied was a gum vulcanizate based on a noncrystallizable SBR-1500. The gum contained 0.5 parts of dicumyl peroxide per 100 parts of rubber and was vulcanized for 0.5 h at 150°C. The measured density is 0.951 g/cm³. The stress–strain response in uniaxial tension was measured at 25°C on a ring-shaped specimen at a strain rate of 1.16/min. These data, represented by the filled circles, are shown in Fig. 11 as a plot of the stress as a function of the extension ratio.

In an attempt to find an analytic form for these stress–strain results, several of the more commonly used one- and two-parameter stress–strain relationships were evaluated. As might be anticipated, the two-parameter expressions were superior to those containing only one parameter.

The dashed curve shown in Fig. 11 is the fit provided by the two-parameter Mooney–Rivlin equation (Refs. 8 and 9) which has the form

$$\sigma = \left(\lambda - \frac{1}{\lambda^2}\right)\left(2C_1 + \frac{2C_2}{\lambda}\right) \tag{1}$$

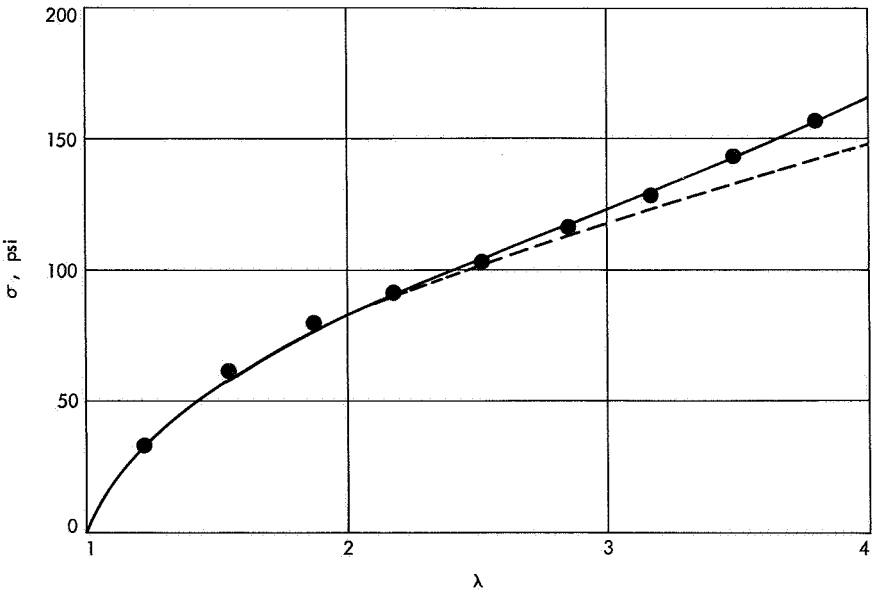


Fig. 11. Stress–strain response for an SBR peroxide vulcanizate

where σ is the stress based on the original cross-sectional area, λ is the extension ratio, and $2C_1$ and $2C_2$ are constants. For the fit shown, $2C_1 = 28$ psi and $2C_2 = 38$ psi. It is evident that Eq. (1) provides a good fit to the experimental data at least for λ values up to about 3. For larger λ, the experimental data fall above the response predicted by Eq. (1).

The solid curve shown in Fig. 11 is the fit provided by the Martin–Roth–Stiehler (MRS) equation (Ref. 10) which has the form

$$\sigma = E\left(\frac{1}{\lambda} - \frac{1}{\lambda^2}\right)\exp\left[A\left(\lambda - \frac{1}{\lambda}\right)\right] \qquad (2)$$

where E is Young's modulus and A is a parameter that depends on both the degree of crosslinking and the time scale. For the fit shown, $E = 174$ psi and $A = 0.43$. As can be seen, this expression provides a good representation for the experimental data over the entire range of λ. Thus, for the present data, the MRS equation provides a better fit than does the Mooney–Rivlin equation.

b. Volume changes on extension. The relative volume changes on extension are of the order of 10^{-4} and, hence, too small to be determined accurately by any simple technique that relies on measurement of linear specimen dimensions. Therefore, a direct measure of the volume change itself is preferred and simple hydrostatic weighing was employed. A rig was made up similar to that employed by Mullins and Tobin whereby four ring specimens 1.375-in. ID × 1.625-in. OD and about 0.1-in. thick could be stretched under water in 0.5-in. increments. The totally-immersed rig and either relaxed or extended specimens were suspended by a 10-mil platinum wire from a Mettler single-arm analytical balance mounted on an essentially vibration-proof table. The water temperature was maintained at 23 ±0.05°C.

Since rubber vulcanizates absorb water to some extent, the specimens were relaxed and reweighed between each successively increasing strain increment in order to follow the absorption of water. The results are shown in Fig. 12 where the weight of the immersed rig and specimens is shown as a function of the immersion time and extension ratio. After about 50 min, the immersed weight of the relaxed specimens increased essentially linearly with time. The rate of increase in weight was 7.72×10^{-5} g H_2O/min which corresponds to a rate of 1.93×10^{-5} g H_2O/g rubber/h for specimens with a surface-to-volume ratio of about 14 cm^{-1}.

Ideally, one would like to have a system in which the relaxed specimens undergo no weight change with time. The rate of change in weight can be reduced by using more rubber, especially if the surface-to-volume ratio is also decreased. At the conclusion of the run, i.e., after about 150 min, the water adhering to the surface of the broken rings was removed by blotting and the weight of the rings in air was measured. The difference between the

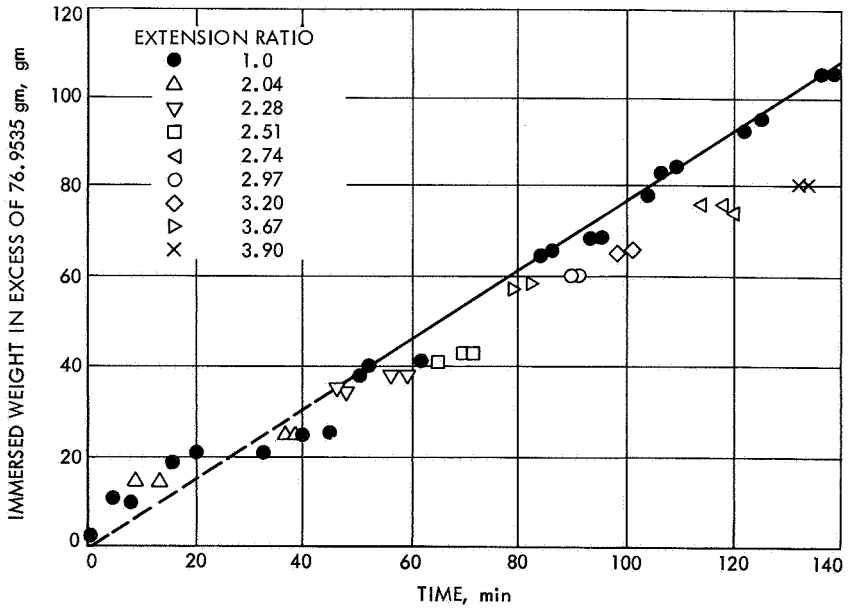


Fig. 12. Immersed weight of rig and specimens as a function of time

final and initial weight in air yields a value of 0.0124 g for the weight increase. This corresponds to a water pickup of 3.10×10^{-3} g H_2O/g rubber if it is assumed that the weight change is due solely to water absorption. The volume of water absorbed by the specimens thus exceeds by about an order of magnitude the relative volume change expected to be produced by the hydrostatic component of stress.

The change in weight on extension was taken as the difference in weight of the extended specimens and the weight of the relaxed specimens, the latter as interpolated from the linear portion of the curve. The data for times less than about 50 min were not used.

The average value of the relative volume change, $\Delta V/V_0$, as a function of the extension ratio is shown in Fig. 13 as the filled circles. The length of the bar associated with each point represents the scatter observed for $\Delta V/V_0$. The scatter appears to be typical of this type of measurement and is comparable to that reported by F. G. Hewitt and R. L. Anthony (Ref. 3). The solid curve shown in Fig. 13 represents the prediction of Eq. (14) using the experimental stress–strain curve.

3. Discussion

In this subsection, it will be pointed out that, for the present SBR vulcanizate, the observed increase in weight of the relaxed specimens with time, as indicated in Fig. 12, can be simply interpreted as due to water entering and filling up voids and cavities pre-existing in the rubber without, at the same time, causing the volume to increase. That is, the specimen volume is constant and

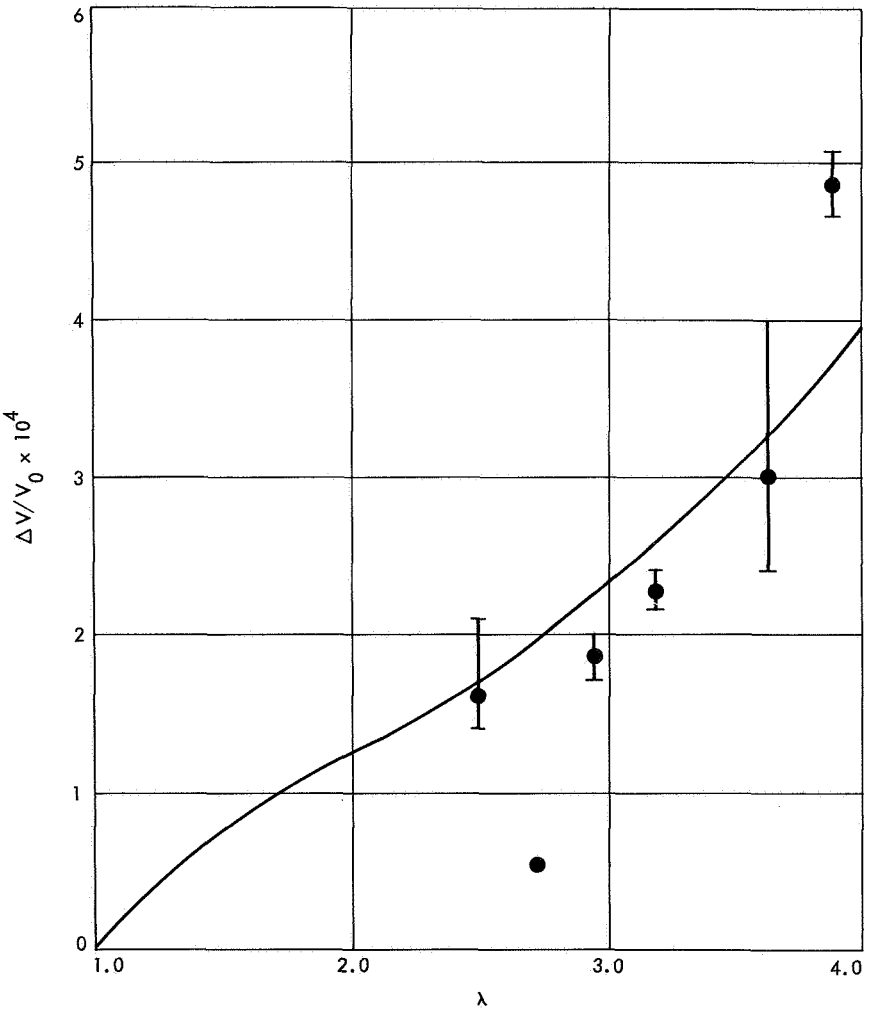


Fig. 13. Dependence of relative volume increase on extension ratio

independent of time. Considering this to be the primary cause of the observed weight increase with time, a simple argument will show that the deformation of these cavities will not contribute an appreciable component to the volume–extension response.

a. Water absorption. The immersed weight of the rig and specimens as a function of time can be written as

$$W_{imm} = V_{rig}(\rho_{rig} - \rho_w) + V_{ru}(t)[\rho_{ru}(t) - \rho_w] \quad (3)$$

where V_{rig} and ρ_{rig} are the volume and density of the rig, respectively; $V_{ru}(t)$ and $\rho_{ru}(t)$ are the time dependent volume and density of the specimens, respectively; and p_w is the density of water. The equation assumes that the observed time dependence is due entirely to the effect of water on the rubber.

The rate of change of the immersed weight with time can be obtained from Eq. (3) as

$$\frac{dW_{imm}}{dt} = \frac{dV_{ru}(t)}{dt}[\rho_{ru}(t) - \rho_w] + V_{ru}(t)\frac{d\rho_{ru}(t)}{dt} \quad (4)$$

Various assumptions concerning the effect of water absorption on both $V_{ru}(t)$ and $\rho_{ru}(t)$ can now be made in order to determine which model most closely predicts the experimental data of Fig. 12.

For example, if we assume that the absorbed water increases both the specimen weight and volume, and further that the total volume is the sum of the volume of rubber and absorbed water, then it is easy to show that

$$V_{ru}(t) = V_{ru}(0)\left[\frac{\rho_w + x(t)\rho_{ru}(0)}{\rho_w}\right] \quad (5)$$

where $x(t)$ is the weight fraction of water absorbed at time t. Similarly, the density can be written as

$$\rho_{ru}(t) = \frac{\rho_w\rho_{ru}(0)[1 + x(t)]}{\rho_w + x(t)\rho_{ru}(0)} \quad (6)$$

Substituting Eqs. (5) and (6) into Eq. (4) yields the result $dW_{imm}/dt = 0$. Thus, the assumption that the absorbed water merely increases the volume of the rubber by simple additivity leads to the prediction that the immersed weight will be time independent. The data in Fig. 12 show that this assumption must be rejected.

On the other hand, we can assume that the absorbed water increases the weight of the specimen without producing any concomitant volume change. This assumption would imply that the water fills up pre-existing voids and cavities present in the rubber. The following expressions are easily derived:

$$V_{ru}(t) = V_{ru}(0) \quad (7)$$

and

$$\rho_{ru}(t) = \rho_{ru}(0)[1 + x(t)] \quad (8)$$

Substituting Eqs. (7) and (8) into Eq. (4), we obtain

$$\frac{dW_{imm}}{dt} = V_{ru}(0)\rho_{ru}(0)\frac{dx}{dt} \quad (9)$$

Since x is an increasing function of time, Eq. (9) predicts a positive slope for dW_{imm}/dt. Also, since the slope in Fig. 12 is constant, x is a linear function of time: $x = 1.93 \times 10^{-5}t$. Knowing $x(t)$ explicitly as well as the duration of the run, we can now calculate the total water absorption and compare it to the measured value which is 3.10×10^{-3} g H_2O/g rubber. The value predicted using $x = 1.93 \times 10^{-5}t$ is 2.90×10^{-3} g H_2O/g rubber. Thus, the close correspondence between the predicted and measured water absorption lends support to the notion that water fills pre-existing voids and cavities.

An alternative way to explain the data of Fig. 12 is to assume that the water leaches out water-soluble materials, present in the rubber, whose density is less than that of water. The water can then replace the material lost. The difference between the two approaches thus relates to whether the voids and cavities pre-exist in the rubber or are formed by a leaching-out process. Unfortunately, the *dried* weight of the specimens was not measured after the test and so it is not known if any material was leached out during the run.

b. Effect of deformation of cavities on volume change. If the absorbed water is indeed present in the form of filled cavities, it is easy to demonstrate that the deformation of such cavities will not contribute appreciably to the volume–strain response. For simplicity, assume that the water present fills a single spherical cavity. Uniaxial deformation of the rubber will deform the sphere into a prolate spheroid. Further, since the water present can easily accommodate a deformation, no dewetting in the sense of a separation of the rubber from the water should occur. If V_w is the volume of water present per cubic

centimeter of rubber, then the relationship between V_w and the radius, r_0, of the spherical cavity is

$$V_w = \frac{4}{3}\pi r_0^3 \tag{10}$$

As the specimen is deformed to an extension ratio λ_1, the major radius of the deformed sphere, r_1, will be $r_1 = \lambda_1 r_0$. The minor radii r_2 and r_3 will be equal and will be taken as $r_2 = \lambda_2 r_0$. Since

$$\lambda_1\lambda_2^2 = [1 + (\Delta V/V_0)]$$

the minor radius is

$$r_2 = \frac{r_0\left(1 + \frac{\Delta V}{V_0}\right)^{1/2}}{\lambda_1^{1/2}} \tag{11}$$

The volume of the deformed sphere is

$$\frac{4}{3}\pi r_0^3\left(1 + \frac{\Delta V}{V_0}\right) \tag{12}$$

while the volume of the original sphere is $(4/3)\pi r_0^3$. Thus, the relative volume change contributed by the deformation of the sphere alone is $V_w(\Delta V/V_0)$. Since V_w was measured to be 2.95×10^{-3} cm^3 H_2O/cm^3 rubber, it is obvious that we can neglect volume changes attending deformation of the cavity. The water was assumed to be present in a single spherical cavity; however, variation in the number of cavities or in the shape should not alter significantly the results obtained on the assumption of a single sphere.

c. Temperature effects. The water bath employed for hydrostatic weighing was maintained at a constant temperature to within about 0.1°C. It is, therefore, important to know the effect of such a temperature variation on the observed weight data. It is easy to demonstrate that the change in weight accompanying a temperature change of ΔT is given by

$$\Delta W_{imm} = \frac{\rho_w \Delta T}{1 + \beta_w \Delta T}\left[V_{rig}(\beta_w - \beta_{rig}) + V_{ru}(\beta_w + \beta_{ru})\right] \tag{13}$$

where β_w, β_{rig}, and β_{ru} are the cubical expansion coefficients of water, rig, and specimens, respectively. The change in immersed weight is

$$\Delta W_{imm} \approx 0.1\left[11.4\,(1.6 \times 10^{-4}) - 4.2\,(4 \times 10^{-4})\right]$$

where

$$\Delta T = 0.1°\text{C}$$

$$\beta_w = 2 \times 10^{-4}$$

$$\beta_{rig} = 0.4 \times 10^{-4}$$

$$\beta_{ru} = 6 \times 10^{-4}$$

$$V_{rig} = 11.4\ \text{cm}^3$$

$$V_{ru} = 4.2\ \text{cm}^3$$

or

$$\Delta W_{imm} \approx 1.4 \times 10^{-5}\ \text{g}$$

and, hence, weight changes caused by temperature fluctuations of about 0.1°C can be ignored.

d. Form of the volume–strain response. Several expressions have been proposed to relate volume change to extension. For example, based on thermodynamic arguments, G. Gee (Ref. 1) derived the expression

$$\frac{\Delta V}{V_0} = \frac{1}{3B}\int_1^{\lambda} \lambda\left(\frac{\partial \sigma}{\partial \lambda}\right)d\lambda \tag{14}$$

where B is the bulk modulus. This expression assumes the material to remain isotropic in the stretched state. If the actual stress–strain curve were at hand, or if an explicit analytic form were available, Eq. (14) could be employed to predict the volume–extension response. It is of interest to consider how the volume–extension response depends on the form of the stress–strain response.

One-parameter equation. If it is assumed that the kinetic theory expression is an adequate representation of the stress–strain response, then Eq. (14) can be integrated directly to yield

$$\frac{\Delta V}{V_0} = \frac{E}{9B}\left[\frac{\lambda^2 - 1}{2} + 2\left(1 - \frac{1}{\lambda}\right)\right] \tag{15}$$

This expression was shown by Hewitt and Anthony to be in good accord with their experimental volume–strain response for λ values up to about 1.5. However, this approach presents some conceptual difficulties, for the kinetic theory expression is derived under the restriction that there is *no* volume change on stretching. The apparent contradiction is resolved by assuming that the

volume change makes such a negligibly small contribution to the stress–strain curve that the kinetic theory expression is still adequate.

Another one-parameter stress–strain relationship of interest was proposed by G. M. Bartenev (Ref. 11), who established that the stress based on the deformed cross-sectional area is proportional to strain for strains up to 100–200%. This has been confirmed by T. L. Smith (Ref. 12) who finds, however, that deviations from strict linearity occur at strains of about 100%. Hence, assuming the stress strain response to be

$$\sigma\lambda = E\,\epsilon \tag{16}$$

the volume–strain response predicted by Eq. (14) becomes

$$\frac{\Delta V}{V_0} = \frac{E}{3B}\ln\lambda \tag{17}$$

It is noteworthy that a similar equation was found by Smith (Ref. 13) to afford a good representation to the volume–strain behavior of polyvinyl chloride glass bead composites.

An additional one-parameter expression that seems to have wide applicability was proposed by K. Valanis and R. F. Landel (Ref. 14):

$$\sigma = \frac{2}{3}\,E(\ln\lambda)\left(1 + \frac{1}{2\lambda^{3/2}}\right) \tag{18}$$

This equation was shown to provide a very good fit to both uniaxial *and* biaxial stress–strain data for λ in the range $1 < \lambda < 2$. In addition, Eq. (18) reproduces the general form of the stress–strain behavior of a Mooney–Rivlin material. Using Eq. (18), Gee's expression yields

$$\frac{\Delta V}{V_0} = \frac{2E}{9B}\left[\left(\lambda + \frac{2}{\lambda^{1/2}}\right) + \frac{3}{2}\,\frac{\ln\lambda}{\lambda^{1/2}} - 3\right] \tag{19}$$

Two-parameter equations. Perhaps the simplest of the two-parameter equations is the Mooney–Rivlin form. As might be expected, this generally provides a better representation of the observed stress–strain response than the one-parameter equation. Substituting Eq. (1) into Eq. (14), the volume–strain response for a Mooney–Rivlin material is

$$\frac{\Delta V}{V_0} = \frac{2C_1}{3B}\left[\frac{\lambda^2}{2} - \frac{2}{\lambda} - \frac{3}{2}\,\frac{C_2}{C_1\lambda} + \frac{3}{2}\left(1 + \frac{C_2}{C_1}\right)\right] \tag{20}$$

On the other hand, if the MRS relationship (Eq. 2) is substituted into Eq. (14), the resulting volume–extension response is not obtained in closed form, i.e., the following integral equation is obtained:

$$\frac{\Delta V}{V_0} = \frac{E}{3B}\int_1^\lambda \frac{[A\lambda^3 - (A+1)\lambda^2 + (A+2)\,\lambda - A]}{\lambda^4} \times \exp\left[A\left(\lambda - \frac{1}{\lambda}\right)\right]d\lambda \tag{21}$$

Once the value of A has been determined, the integral can be evaluated numerically.

Equations for the volume–extension response based on the use of Eq. (14), which have been derived above, implicitly assume that rubber remains isotropic in the stretched state. It was pointed out by T. N. Khasanovich (Ref. 15) that an elastomer would not be expected to remain isotropic at large strains, and, hence, that the integrand in Eq. (14) should be multiplied by a factor μ that takes into account the anisotropy of linear compressibility for a stretched material. If the elastomer obeys the kinetic theory stress–strain law derived by James and Guth, then Khasanovich shows that

$$\mu = \frac{3}{(\lambda^3 + 2)} \tag{22}$$

Using this result, Eq. (14) becomes

$$\frac{\Delta V}{V_0} = \frac{E}{3B}\left(1 - \frac{1}{\lambda}\right) \tag{23}$$

Figure 14 shows a comparison of the form of several of the $\Delta V/V_0$, λ expressions using the reduced relative volume change, $(\Delta V/V_0)\,(3B/E)$, as the variable on the ordinate. Curve 1 corresponds to the kinetic theory expression (Eq. 15); curve 2 corresponds to the Valanis–Landel expression (Eq. 19); curve 3 corresponds to the prediction of Eq. (14) obtained by actual integration of the experimentally obtained stress–strain curve taking B as 3.1×10^5 psi (a value close to that reported in the literature). This same curve also corresponds to the MRS expression (Eq. 21), using $E = 174$ psi and $A = 0.43$. Curve 4 corresponds to the Bartenev expression (Eq. 17); curve 5 corresponds to the Mooney–Rivlin expression (Eq. 20), using $2C_1 = 28$ psi and $2C_2 = 38$ psi; and curve 6 corresponds to the Khasanovich expression (Eq.

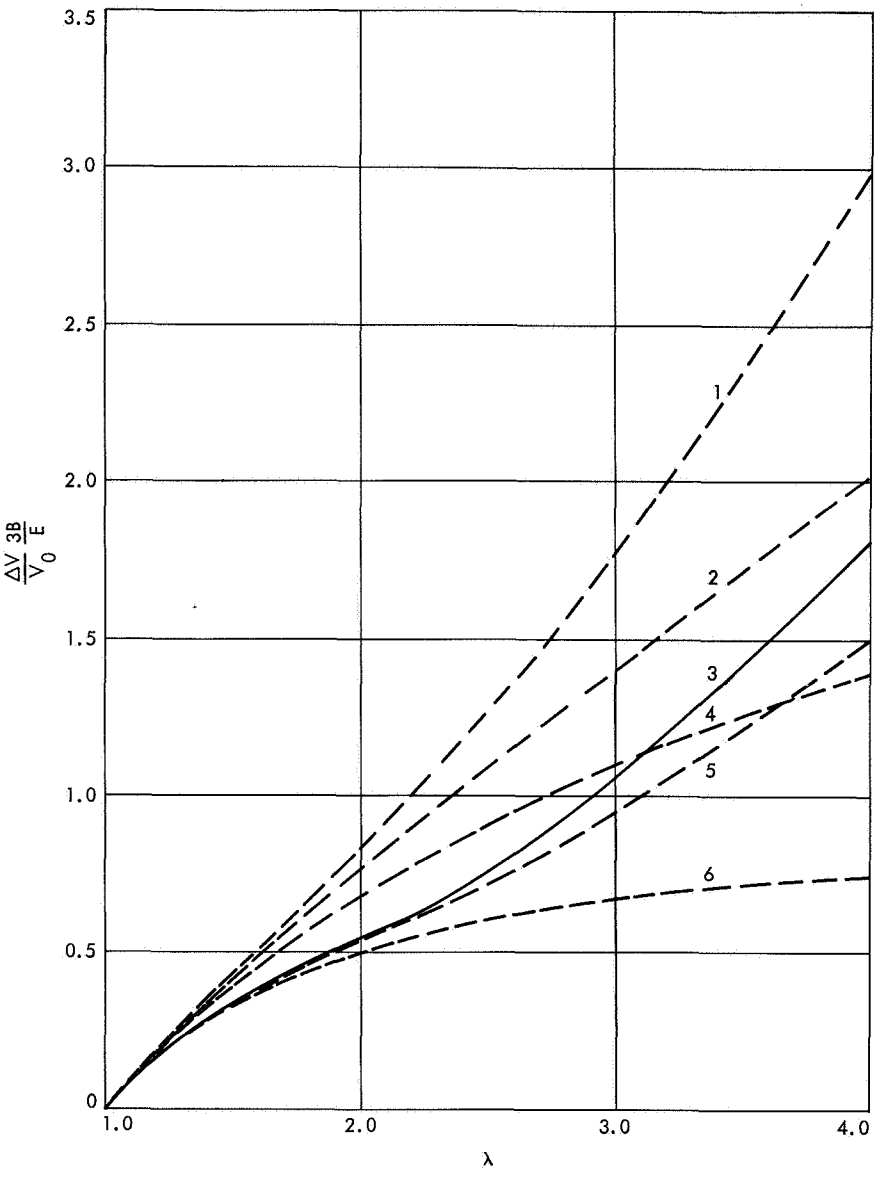


Fig. 14. Dependence of reduced relative volume increase on extension ratio

23). Of the various expressions considered, the upper limit to $\Delta V/V_0$ is apparently predicted by the kinetic theory expression, while the lower limit to $\Delta V/V_0$ is provided by the equation proposed by Khasanovich.

Figure 14 clearly demonstrates that, for small λ values, the predictions of the various proposals rapidly converge, so that, in this region, it will be difficult if not impossible to determine the degree of conformance to a given expression, especially in view of the scatter typical of $\Delta V/V_0$, λ data. At large strains, however, the predictions diverge so that it would be much easier to test for conformance in this region.

In Fig. 13, the solid curve corresponds to the predicted $\Delta V/V_0$, λ response using Gee's Eq. (14) with the integrated form of the experimental stress–strain curve. [This same curve is generated using the MRS expression (Eq. 21)]. The fit shown is considered satisfactory in view of the preliminary nature of these direct volume change measurements.

References

1. Gee, G., Stern, J., and Treloar, L. R. G., *Trans. Faraday Soc.*, Vol. 46, p. 1101, 1950.
2. Mullins, L., and Tobin, N. R., *Trans. Inst. Rubber Ind.*, Vol. 33, p. 2, 1957.

3. Hewitt, F. G., and Anthony, R. L., *J. Appl. Phys.*, Vol. 29, p. 1411, 1958.

4. Allen, G., Bianchi, U., and Price, C., *Trans. Faraday Soc.*, Vol. 59, p. 2493, 1963.

5. Miller, R. L., *Polymer Handbook*, p. 111-1. Edited by J. Brandrup, and E. H. Immergut. Interscience Publishers division of John Wiley & Sons, Inc., New York, 1966.

6. Nyburg, S. C., *Brit. J. Appl. Phys.*, Vol. 5, p. 321, 1954.

7. Yau, W., and Stein, R. S., *Polym. Lett.*, Vol. 1, p. 231, 1964.

8. Mooney, M., *J. Appl. Phys.*, Vol. 11, p. 582, 1940.

9. Rivlin, R. S., *Phil. Trans.*, Vol. A240, p. 459, 1948.

10. Martin, G. M., Roth, F. L., and Stiehler, R. D., *Trans. Inst. Rubber Ind.*, Vol. 32, p. 189, 1956.

11. Bartenev, G. M., *Kolloidn Zh.*, Vol. 11, p. 2, 1949.

12. Smith, T. L., *Trans. Soc. Rheol.*, Vol. 6, p. 61, 1962.

13. Smith, T. L., *Trans. Soc. Rheol.*, Vol. 3, p. 113, 1959.

14. Valanis, K., and Landel, R. F., *J. Appl. Phys.*, Vol. 38, p. 2997, 1968.

15. Khasanovich, T. N., *J. Appl. Phys.*, Vol. 30, p. 948, 1959.

X. Research and Advanced Concepts

PROPULSION DIVISION

A. Hollow Cathode Operation in the SE-20C Thruster, *T. D. Masek and E. V. Pawlik*

1. Introduction

Operation of the SE-20C mercury ion thruster using an oxide cathode has been reported previously (Refs. 1 and 2). Interest in the hollow cathode, as discussed in SPS 37-48, Vol. III, pp. 119–125 and SPS 37-49, Vol. III, pp. 207–211, has resulted in adapting this cathode to the SE-20C. This article describes initial test results using an adjustable cathode pole piece and baffle assembly. The objectives of this work were the evaluation of the effect of

(1) The pole piece side slot and baffle open area on discharge losses.

(2) Introducing all the propellant flow through the cathode.

(3) Total flowrate on discharge losses.

2. Experimental Setup

The hollow cathode thruster was one of four thrusters mounted in a basic test array (Refs. 1 and 2). Except for the use of the SE-20C thruster and a pole piece assembly to be described later, the setup was similar to that reported in Ref. 1. A new cathode with a tip diameter of 0.42 cm, a tip thickness of 0.10 cm, and an orifice diameter of 0.05 cm was used.

The cathode pole piece and baffle assembly is shown in Fig. 1. The pole piece used in oxide cathode thruster designs forms the basic structure. The side of the pole piece was slotted in eight places (2.54 cm long, including circular ends, and 1.59 cm wide). A similarly slotted sleeve was fitted inside. Rotation of the sleeve relative to the pole piece varied the slot area from 0 to 28 cm^2. The baffle was constructed from two disks mounted on the same axis, each having four 1.59-cm-diameter holes. Rotation of the outer disk relative to the other disk fixed to the pole piece varied the baffle area from 0 to 7.9 cm^2. As shown in Fig. 1, both the sleeve and baffle were adjusted using gear sections. Shafts through the thruster backplate allowed these adjustments to be made from outside the vacuum chamber while the thruster was operating.

3. Test Results

Operation at a single cathode flowrate (3.35 g/h) was sufficient to obtain the data of interest. Using this

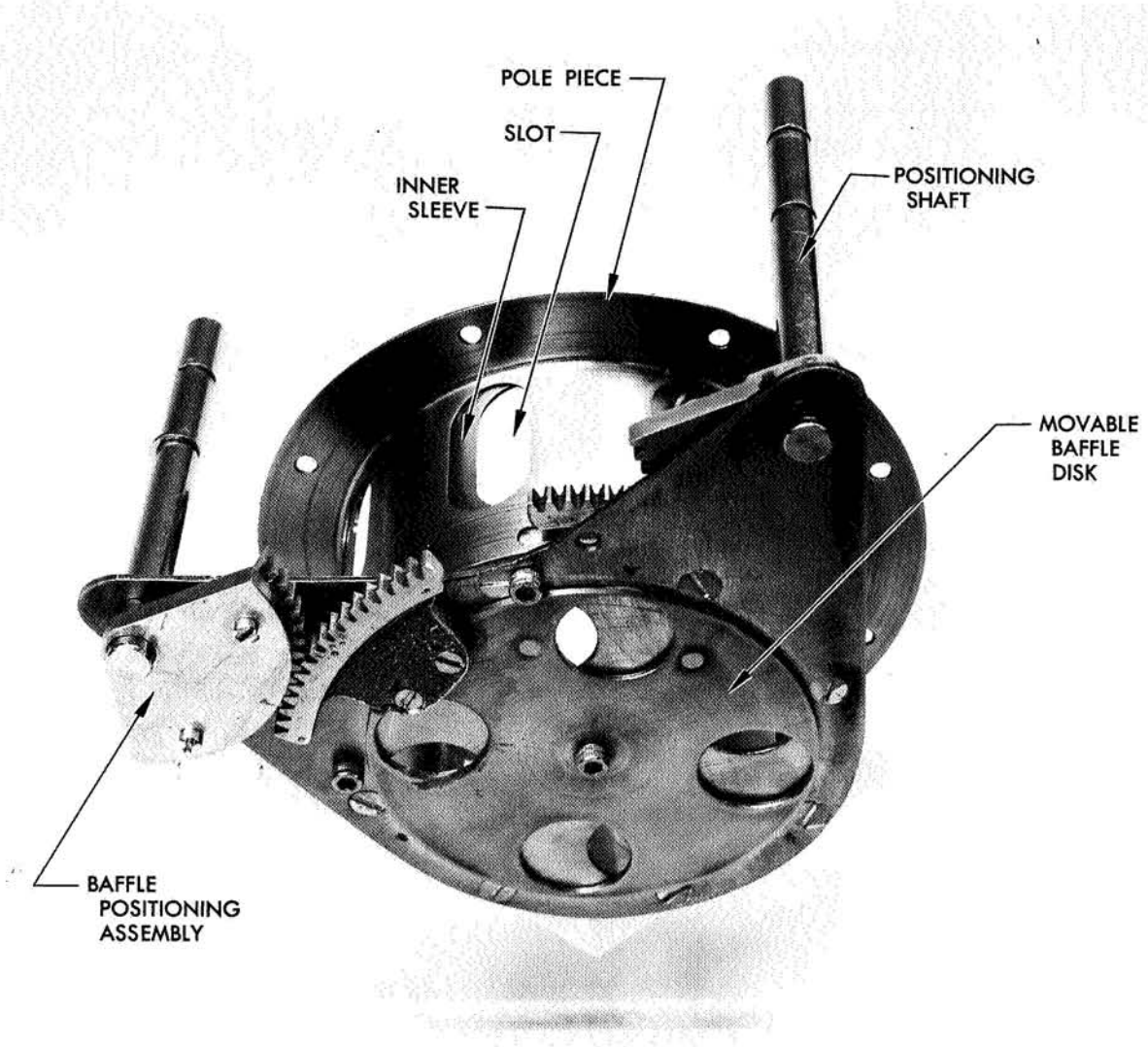


Fig. 1. Variable-area cathode baffle

cathode flow, the thruster was tested with other main flowrates. Because of limited test time, the pole piece slot was adjusted for minimum discharge losses at the beginning (only cathode flow) and remained fixed. This slot area was 9.5 cm².

Data showing the effect of baffle open area on discharge losses and utilization efficiency are shown in Fig. 2 for two flowrate conditions. The data show the best trade-off of discharge losses against utilization occurs at about 1.5 cm². Smaller areas increase the discharge losses with little or no change in utilization. This condition is similar for cathode flow alone and with main flow. Previous hollow cathode tests (SPS 37-48 and 37-49, Vol. III) with the SE-20B thruster, without pole piece slots, showed a baffle open area of 2.5 cm² was near optimum.

The effect of introducing all of the propellant through the cathode is also illustrated in Fig. 2. With only cathode flow, the discharge losses were approximately 760 eV/ion. With the addition of main flow, the losses were reduced 300 eV/ion. This result is consistent with physical reasoning and previous results. With only cathode flow, neutrals must be ionized within or near the pole piece. If not ionized in this region, the neutral flux is collimated by the pole piece baffle and can escape without a collision. Similar tests (Ref. 1) also indicated a fraction of

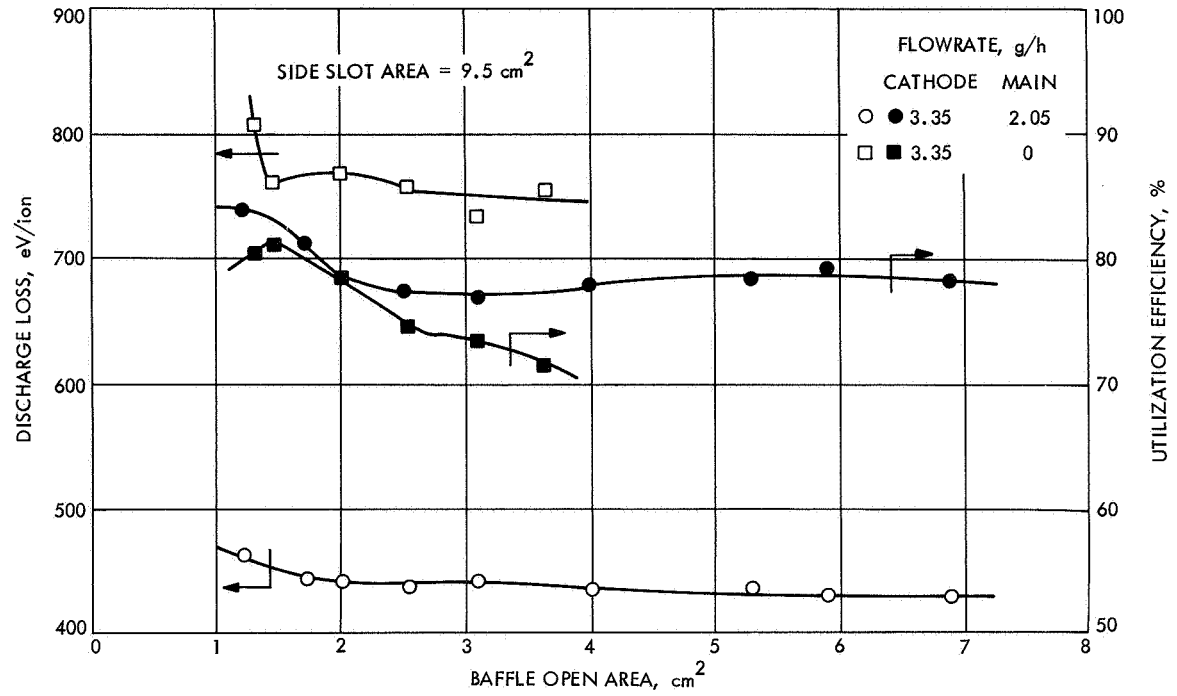


Fig. 2. Discharge loss and utilization efficiency as a function of baffle open area

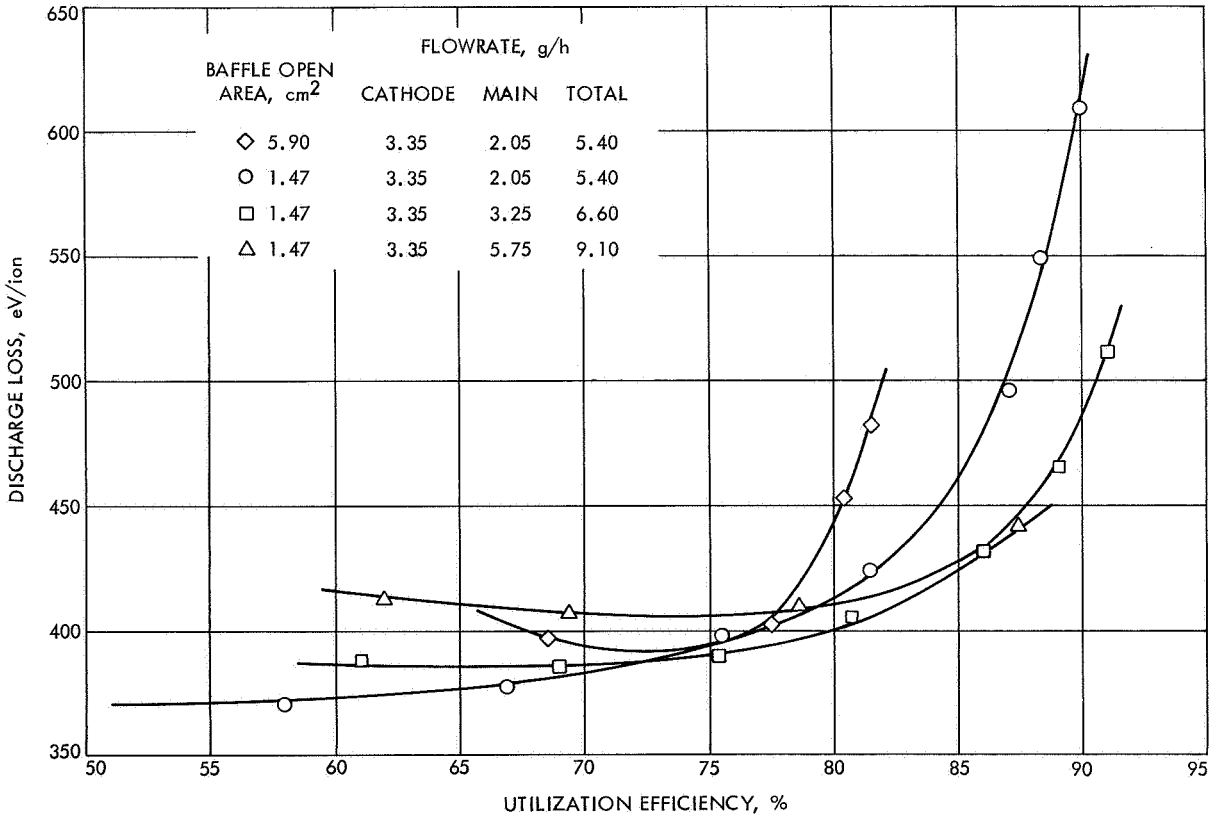


Fig. 3. Discharge loss variation for different baffle open areas and different total flowrates

the neutrals is lost directly when fed in from the rear. Raising the ionization fraction in the pole piece to prevent the loss of neutrals forces both the wall losses and basic ion production cost up. The addition of main flow makes the cathode flow relatively less important. The cathode density can then be reduced, decreasing cathode wall losses and ion production costs, and allowing more "primary" electrons to reach the main discharge chamber to ionize the main flow. This is a more efficient mode because the "reverse feed" main flow does not allow neutrals to escape without a collision.

The influence of main flow is illustrated further in Fig. 3. These data show a number of interesting effects. First, the effect of pole piece open area on the ability to obtain high utilization is shown. For the same total flowrate of 5.40 g/h, the curves are shifted by about 5% in utilization with different baffle open areas (1.47 and 5.90 cm^2). The more open condition apparently allows a higher fraction of neutrals to escape directly. Second, with the same baffle open area (1.47 cm^2), the data show that minimum discharge losses occur with a flowrate of about 6–7 g/h. This is also consistent with other results. When the cathode flow is as small as in Ref. 2, the curves shift consistently upward with increased flow. However, this "normal" trend is offset by the improvement provided by increasing the fraction of main flow. Note that the curves continue to shift toward higher utilization as the main flow fraction increases.

4. Conclusions

The results indicate that introducing all or a large fraction of the propellant through the cathode may limit utilization and increase discharge losses. Other configurations might avoid this difficulty. A good pole piece configuration was shown to be established rapidly (in one test) with the assembly used. Additional tests should be performed with the cathode in other positions and with other cathode flowrates. The change of discharge loss characteristics with flowrate supports physical arguments in describing discharge operation.

References

1. Masek, T. D., *Experimental Studies With a Mercury Bombardment Thrustor System*, Technical Report 32-1280. Jet Propulsion Laboratory, Pasadena, Calif., July 15, 1968.
2. Masek, T. D. and Pawlik, E. V., "Thrust System Technology for Solar Electric Propulsion," Paper 68-541, presented at the AIAA Fourth Propulsion Joint Specialist Meeting, Cleveland, Ohio, June 10, 1968.

B. Plasma Investigation in the SE-20C Thruster,

T. D. Masek

1. Introduction

Thruster efficiency improvements in the past 2 yr have resulted from configuration changes (Refs. 1–3). In addition, the influence of operating parameters (propellant flowrate, discharge voltage, magnetic field strength, and ion accelerating voltage) on efficiency has been studied (Refs. 2 and 3). Since thruster efficiency depends directly on the characteristics of the discharge plasma, a study of the plasma in an improved thruster (SE-20C) is of interest. Such a study is required to evaluate the efficiency improvements, determine efficiency limits, and understand the influence of the operating parameters. This article describes the test setup and a portion of the preliminary results of a study with these goals. A conventional Langmuir probe was used to measure plasma properties.

2. Experimental Setup

In order to obtain probe measurements throughout the thruster rapidly, a motor-driven probe positioner was built (Fig. 4). The drive assembly, mounted near the thruster in the chamber, provided radial and axial positioning with positions determined by potentiometers. The potentiometer readings were displayed on an X–Y recorder. A layout of possible probe positions and mechanical interference was used on the recorder to simplify positioning. A set of probe traces, taken on an X–Y–Y recorder, for about 70 positions throughout the thruster, could be obtained in 1 h.

3. Test Results

The plasma properties (electron energy distribution function ion density and plasma potential) of the improved thruster are basically similar to those reported previously (Ref. 4). That is, the electron energy distribution is still composed of primary and maxwellian groups and the density and potentials are of the same order of magnitude as before. The present article will discuss typical data obtained with the improved thruster and give a brief comparison with older thruster data. Analysis of the non-maxwellian characteristics was accomplished as discussed in Ref. 5.

Data taken with a flowrate of 8.8 g/h and 90% utilization efficiency is shown in Figs. 5 and 6. Equivalue curves are estimated in these figures for ion density and plasma potential. Plots for a flowrate of 5.7 g/h and

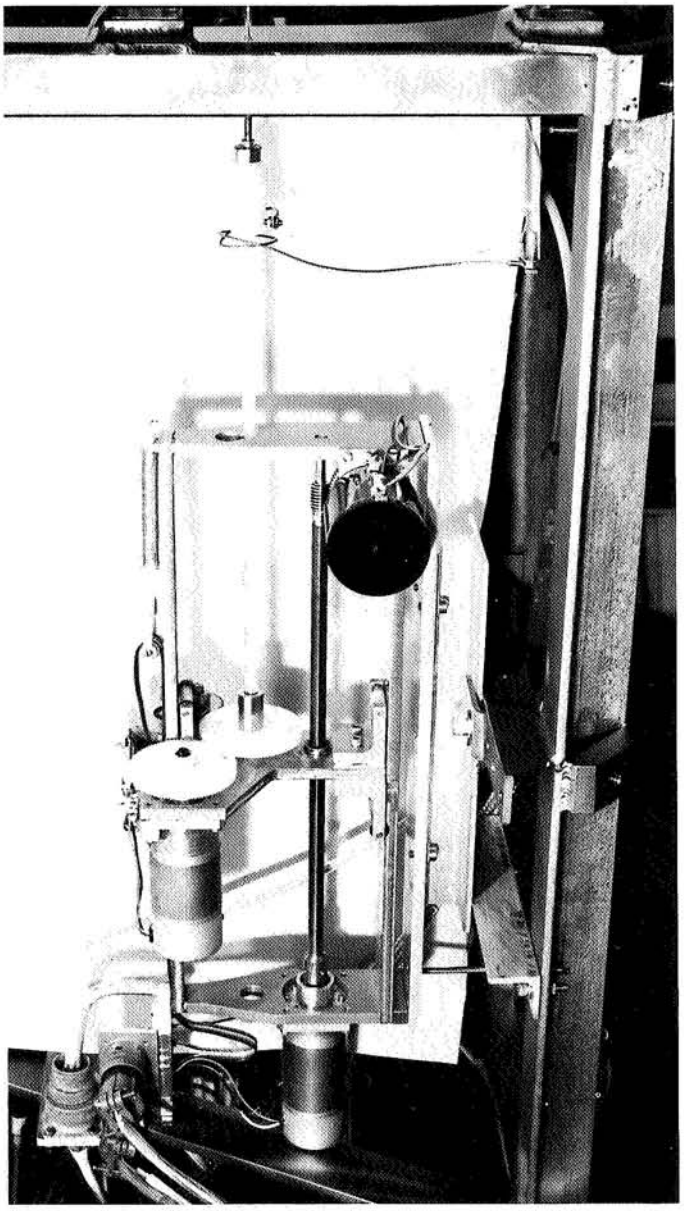


Fig. 4. Probe drive assembly

90% utilization resulted in similarly shaped curves but with somewhat different values. The measured values of each property are indicated at each probe location. Also shown on the figures are the discharge chamber surfaces (cathode, anode screen grid, and housing).

The ion density at the grid (Fig. 5) is seen to vary by only a factor of 3 from the center to 0.9 of the radius. This variation in unimproved thrusters (Ref. 1) was on the order of 10. In addition, the axial variation of density, from maximum to the grid, on the center line is about 2.2, compared with about 1.4 in the older thruster. These differences, in part, result in higher ion axial drift velocity in the plasma and a more uniform ion beam for the improved thruster.

The plasma potential distribution (Fig. 6) plays a principal role in directing ions through the plasma to the grid. The maximum axial variation of potential on the center line is seen to be 2.4 V (compared with less than 1 V maximum in the unimproved thruster). This axial variation at other radial locations is 1.9 V or above. The radial change in potential, from the center line to 0.9 of the radius, varies from 1.4 to 2.0 V over most of the chamber. In addition, about 80% of the locations on the grid side of axial position 4.0 have higher axial potential gradients than radial gradients. This tends to accelerate a large fraction of the ions preferentially toward the grid. The unimproved thruster plasma potential peak was located 0.3 of the chamber length from the grid and had higher radial than axial potential gradient throughout. The density and potential distributions providing higher axial and lower radial ion velocities indicate the manner in which the newer thruster has improved.

The remaining plasma properties also play an important role in thruster operation. However, the differences between these data and those of older thrusters do not appear to contribute greatly to the thruster improvements. The variation of these properties, as well as the potential and density, must be considered in evaluating performance variations with operating conditions. This subject will be discussed in future articles.

References

1. Bechtel, R. T., "Discharge Chamber Optimization of the SERT II Thruster," Paper 67-668, presented at the AIAA Electric Propulsion and Plasmadynamics Conference, Colorado Springs, Colo., Sept. 1967.
2. Masek, T. D. and Pawlik, E. V., "Thrust System Technology for Solar Electric Propulsion," Paper 68-541, presented at the AIAA Fourth Propulsion Joint Specialist Meeting, Cleveland, Ohio, June 10, 1968.
3. Masek, T. D., *Experimental Studies With a Mercury Bombardment Thruster System*, Technical Report 32-1280. Jet Propulsion Laboratory, Pasadena, Calif., July 15, 1968.
4. Masek, T. D., *Plasma Characteristics of the Electron Bombardment Ion Engine*, Technical Report 32-1271. Jet Propulsion Laboratory, Pasadena, Calif., Apr. 15, 1968.
5. Strickfaden, W. B., and Geiler, K. L., "Probe Measurements of the Discharge in an Operating Electron Bombardment Thruster," *AIAA J.*, Vol. I, pp. 1815–1823, 1963.

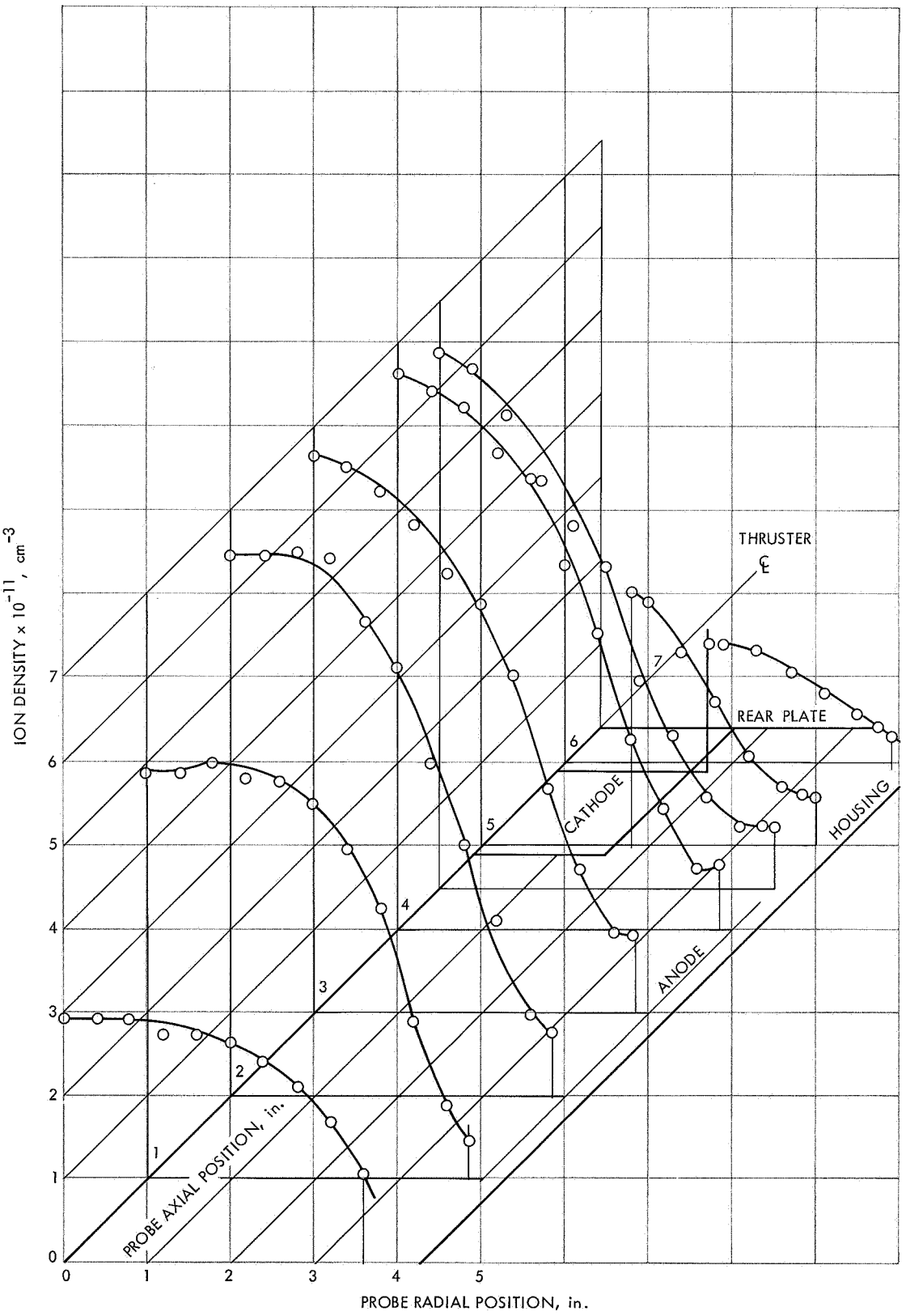


Fig. 5. Ion density distribution in the SE-20C thruster (flowrate = 8.8 g/h, utilization = 90%)

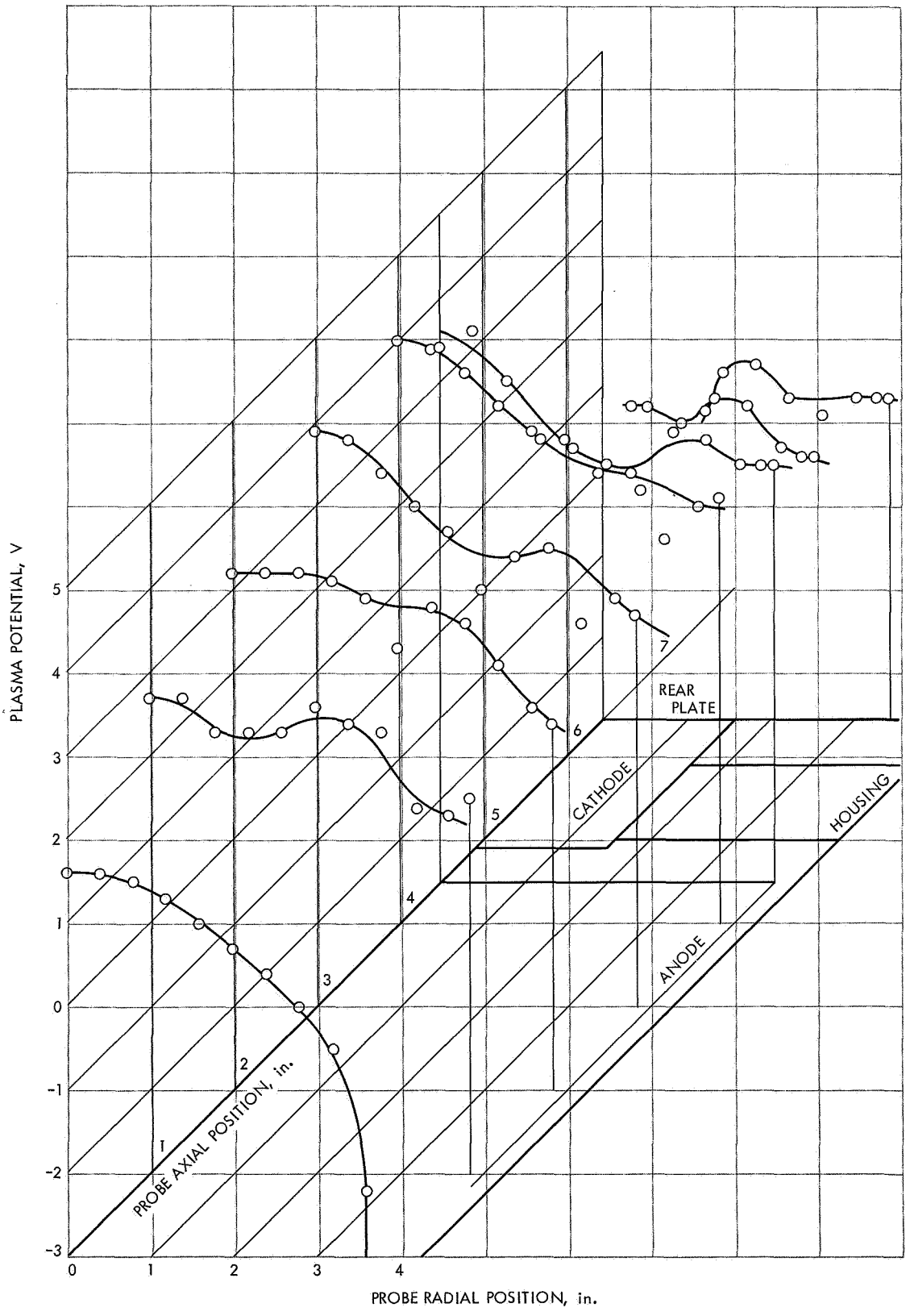


Fig. 6. Plasma potential distribution in the SE-20C thruster (flowrate = 8.8 g/h, utilization = 90%)

C. Liquid–Metal MHD Power Conversion,

D. J. Cerini

1. Introduction

Liquid–metal magnetohydrodynamic (MHD) power conversion is being investigated as a power source for nuclear–electric propulsion. A liquid–metal MHD system has no moving mechanical parts and operates at heat-source temperatures between 1600 and 2000°F. Thus, the system has the potential of high reliability and long lifetime using readily available containment materials such as Nb–1%Zr.

In the particular MHD cycle being investigated, liquid lithium would be (1) heated at about 150 psia in the reactor or reactor-loop heat exchanger; (2) mixed with liquid cesium at the inlet of a two-phase nozzle, causing the cesium to vaporize; (3) accelerated by the cesium to about 500 ft/s at 15 psia; (4) separated from the cesium; (5) decelerated in an alternating-current MHD generator; and (6) returned through a diffuser to the heat source. The cesium would be condensed in a radiator or radiator-loop heat exchanger and returned to the nozzle by an MHD pump.

2. Generator Tests

The ac generator for the NaK–nitrogen conversion system (SPS 37-51, Vol. III, pp. 120–124) is undergoing empty-channel electrical tests. The empty-channel generator tests have the fourfold purpose of (1) providing an operational checkout of the facility instrumentation and electrical control systems; (2) determining the operating characteristics of the generator, with regard to balancing the eight generator phase circuits, in order to obtain a uniform traveling magnetic field with proper upstream and downstream compensating-pole fields; (3) determining the power losses due to induced eddy currents in the structural components of the generator assembly, which includes the laminated stator blocks, the stator clamps (H-frames), stator to H-frame bolts, copper side bars, stator slot plugs (laminated and solid), downstream diffuser, upstream and downstream compensating-pole vanes; and (4) determining the winding loss of the actual coils as a function of traveling-wave magnetic field and frequency.

The generator tested consists of the stator assembly described in SPS 37-50, Vol. III, pp. 182–186, with Litz wire coils as shown in Fig. 7. Shown also is the coolant tank which is raised while testing to immerse the generator in the Freon TF coolant; the gaussmeter probe is used to map the channel magnetic field, and also provides a calibration of the search coils, which are individual wires inserted in each of the stator teeth. By comparison of the voltage induced in any two search-coil wires during NaK–nitrogen system tests, this calibration will permit the evaluation of the magnetic-field amplitude and wave speed in the channel between the two wires.

The search coils and gaussmeter probe were used as an aid in setting the proper phase currents and capacitance values to achieve the desired channel field. A variable capacitor bank on each of the eight generator phases provides the reactive power to the coils, with a five-phase 40-kV-A motor generator set supplying the real power to the two compensating-pole phases and to three of the traveling-wave phases; the remaining three traveling-wave phases are excited by transformer coupling to the three driven phases to produce a six-phase system.

To achieve a balanced operating condition at any current or magnetic field level, the six traveling-wave phase currents are set equal; the compensating-pole currents are set approximately at twice the traveling-wave current; the magnetic field in the upstream and downstream compensating poles is set by rotating the compensating-pole current phase angles to produce maximum downward flux at $\omega t = 0$ and 180 deg, respectively, where $\omega t = 0$ and 90 deg are the times at which the magnetic field has a positive sine wave and negative cosine wave shape, respectively, in the traveling-wave region. Further adjustment of the compensating-pole current amplitudes and phase angle may then be needed to make the search coils at the inlet, center, and exit of the traveling wave equal in phase and magnitude, indicating that the field is symmetrical about the channel midpoint and the flux in both compensating poles is equal to one half the amount of flux in the traveling-wave region. Equal search coil magnitudes and 30-deg phase spacing between adjacent search coils indicate a uniform amplitude and velocity of the traveling wave. Shown in Fig. 8 is a typical magnetic-field survey obtained with the aid of a waveform analyzer which outputs the magnetic-field amplitude at the preset value of ωt to an X–Y plotter.

None of the generator structural components mentioned above had a significant effect on the field shape except for the aluminum downstream diffuser, which fits into generator gap in the downstream compensating pole; this had sufficiently high induced eddy currents and associated magnetic field to completely cancel the desired

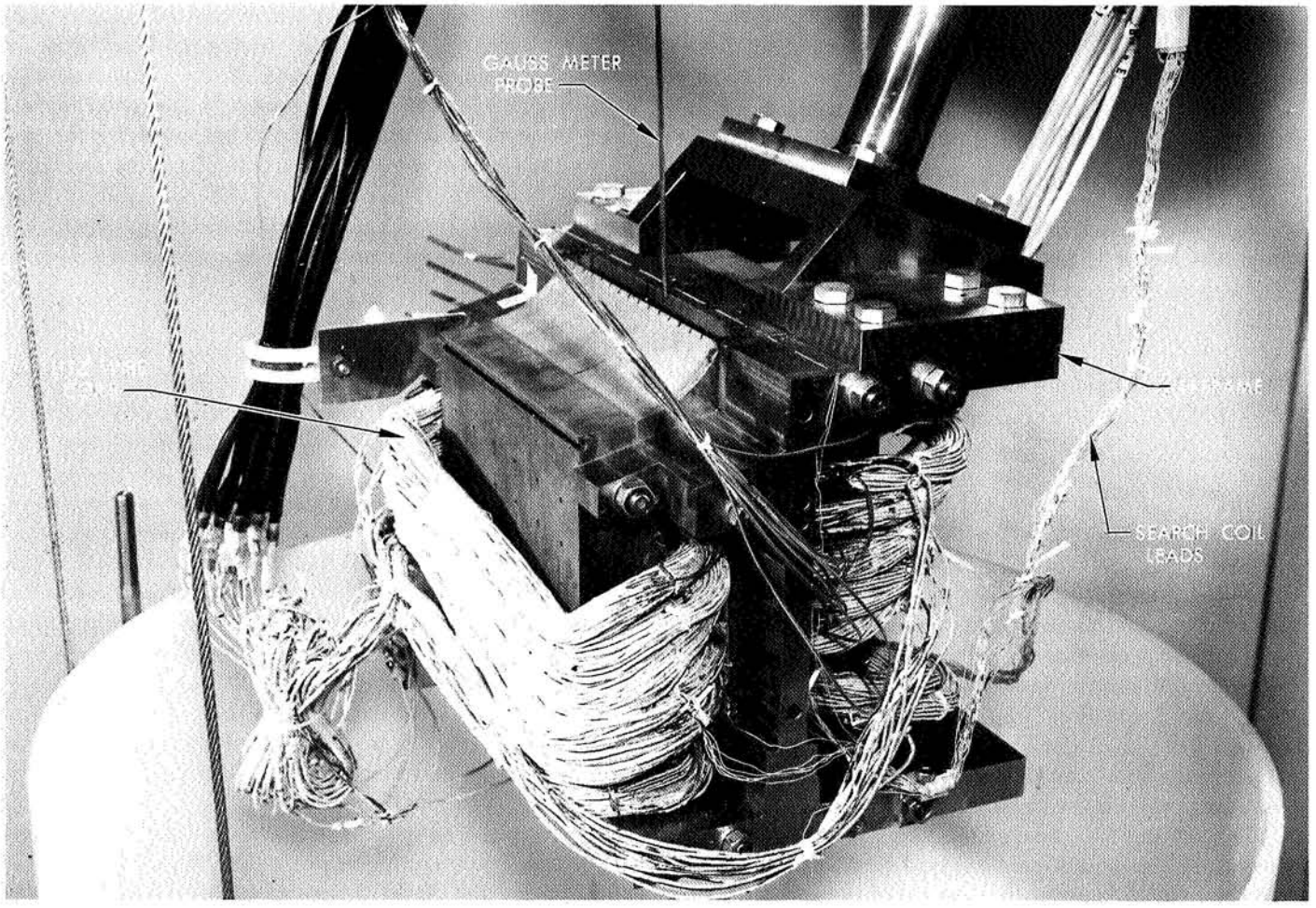


Fig. 7. Test arrangement for measuring generator core losses and magnetic-field profiles

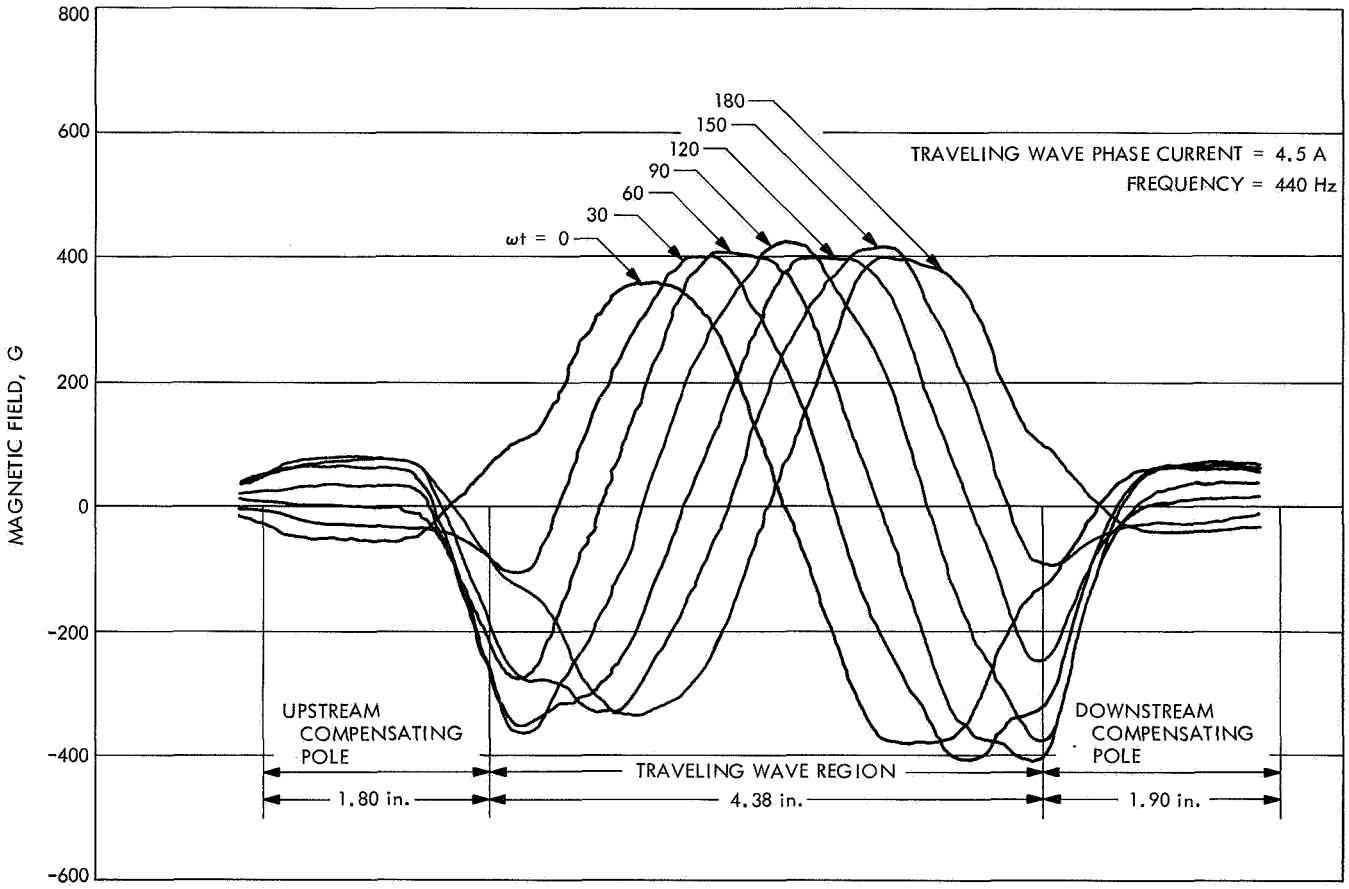


Fig. 8. Measured ac magnetic-field profiles in NaK-nitrogen conversion system generator

compensating-pole field. This result will require the use of a non-metallic diffuser inlet, which was fabricated previously out of Vespel.

Subsequent power-loss tests indicated that other structural components required modification because of excessive power loss even though they had no effect on the field shape. With only the stators and coils assembled, the power loss in the stator blocks due to eddy currents was evaluated as the total input power to the eight phases less the winding power, which is the product of the Litz coil dc resistance and the square of the rms current summed for the eight phases. Extrapolating the power loss to the nominal operating condition indicates an excessive core loss of about 6 kW, which is apparently due to burrs shorting the laminations as evidenced by the low ohmic resistance of the block of about 0.1 Ω.

Subsequent disassembly and acid etch have returned the stator resistance to greater than 6 Ω. The copper side bars, H-frame bolts, and H-frames were added to the stator assembly with a power-loss measurement taken after each was added. The results indicate the copper side bars and bolts will have an acceptable 1-kW power loss, while the H-frames will have an excessive 6-kW loss attributed to their high permeability produced by heat treating. Annealing is expected to reduce the permeability and thus the power loss to an acceptable level without appreciably reducing the mechanical strength. The generator is being reassembled to verify the power-loss reduction in the stators and H-frames and to complete the tests with the addition of the slot plugs and vanes.

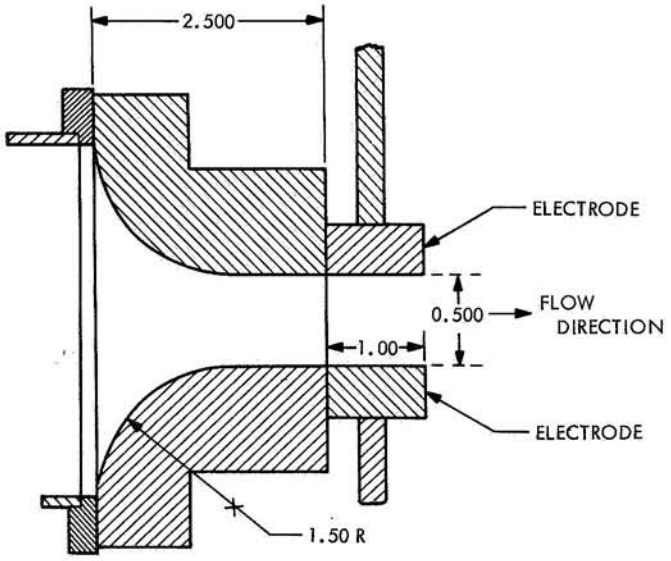
D. Potential Distribution Associated With a Glow Discharge Influenced by a Transverse Gas Flow, *J. A. Gardner and M. B. Noel*

1. Introduction

Preliminary experimental investigations concerning flow visualization studies as related to the effects of a transverse gas velocity on a glow discharge were reported

in SPS 37-45, Vol. IV, pp. 162–167. The objective of those initial experiments was to determine in a qualitative sense the influence of a transverse velocity on a glow discharge in the absence of an applied magnetic field.

Additional experiments have now been conducted in which the potential distribution within the electrical discharge region has been obtained by means of a probe. The same parallel-plate copper electrodes that were used previously (Fig. 9) were employed in this investigation. The final objective of these investigations is to achieve a better understanding of the many combinations of processes that occur in the electrical discharge region at and near the electrode surfaces of electrical propulsion and power generation devices.



DIMENSIONS IN INCHES

Fig. 9. Experimental apparatus, showing two-dimensional subsonic nozzle and electrode configuration

The effect of blowing on the visual appearance of the discharge was discussed in SPS 37-45, Vol. IV, and is shown in Fig. 10. These photographs indicated that there is most likely a considerable change associated with the electric-field distribution between the static and dynamic conditions. The electric-field distribution for a plasma column between flat parallel-plate electrodes without blowing has been reported by several investigators, e.g., Ref. 1. The anode and cathode fall regions located near the electrode surfaces account for a large fraction of the potential drop. The region between the cathode and

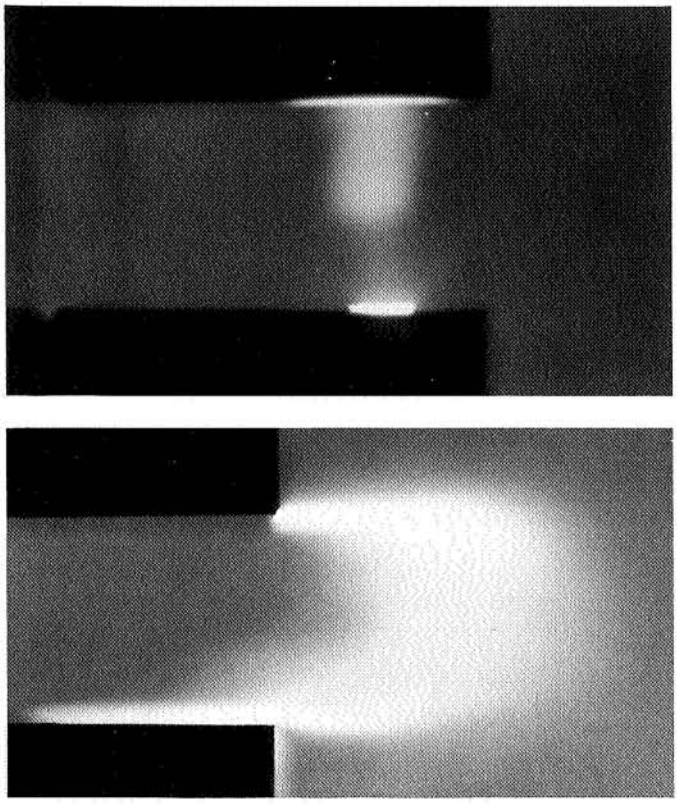


Fig. 10. Glow discharge in argon for a ½-in. electrode separation (15-mA discharge current, 100-torr static pressure): (a) no flow, (b) 127-ft/s flow

anode fall regions is the area where the electric-field gradient is essentially constant. However, in the presence of a transverse gas flow the electric-field distribution becomes quite distorted.

2. Experimental Apparatus

The apparatus used for these experiments (Fig. 9) consisted of a pair of flat parallel copper electrodes enclosed in a six-port glass chamber. Argon gas entered the region between the electrodes at ambient temperature and at a pressure of 100 mm Hg from a two-dimensional convergent nozzle having throat dimensions of 0.5×2.0 in. From the electrode region, the gas flowed into the exhaust duct and the vacuum system. The experiments were conducted at a gas velocity of 127 ft/s, which was computed from the measured total mass flow rate, the pressure and temperature of the gas, and the cross sectional area of the nozzle throat. In prior experiments, pitot tube traverses were performed in the absence of a discharge to verify the uniformity of velocity at the nozzle exit.

3. Procedure and Results

The flow was first established at the desired conditions and then the discharge was initiated and the current level set at 15 mA. Steady state (as determined by visual observation) occurred quickly and remained for an extended period. The local potential was then measured using a 0.010-in.-diam wire enclosed in a needle-shaped glass probe (Fig. 11). The other end of the wire was connected to a 100-MΩ voltmeter and referenced to the cathode (lower electrode in photograph). This electric probe was then traversed in three normal directions to map out the potential distribution.

A portion of this data is presented in Fig. 12. This represents the potential measurement that would correspond to the view seen in the dynamic case (Fig. 10b) taken in the plane of the discharge. The plasma column visually appeared to be approximately 1/16-in. thick in the plane of the photograph and was essentially planar. The shape of the potential curve normal to the velocity vector and normal to the plane of the photograph was symmetrical with respect to the plasma column.

As observed previously, the transverse flow of gas caused the discharge to distort into a U-shape and attach to the trailing edge of the anode. The attachment

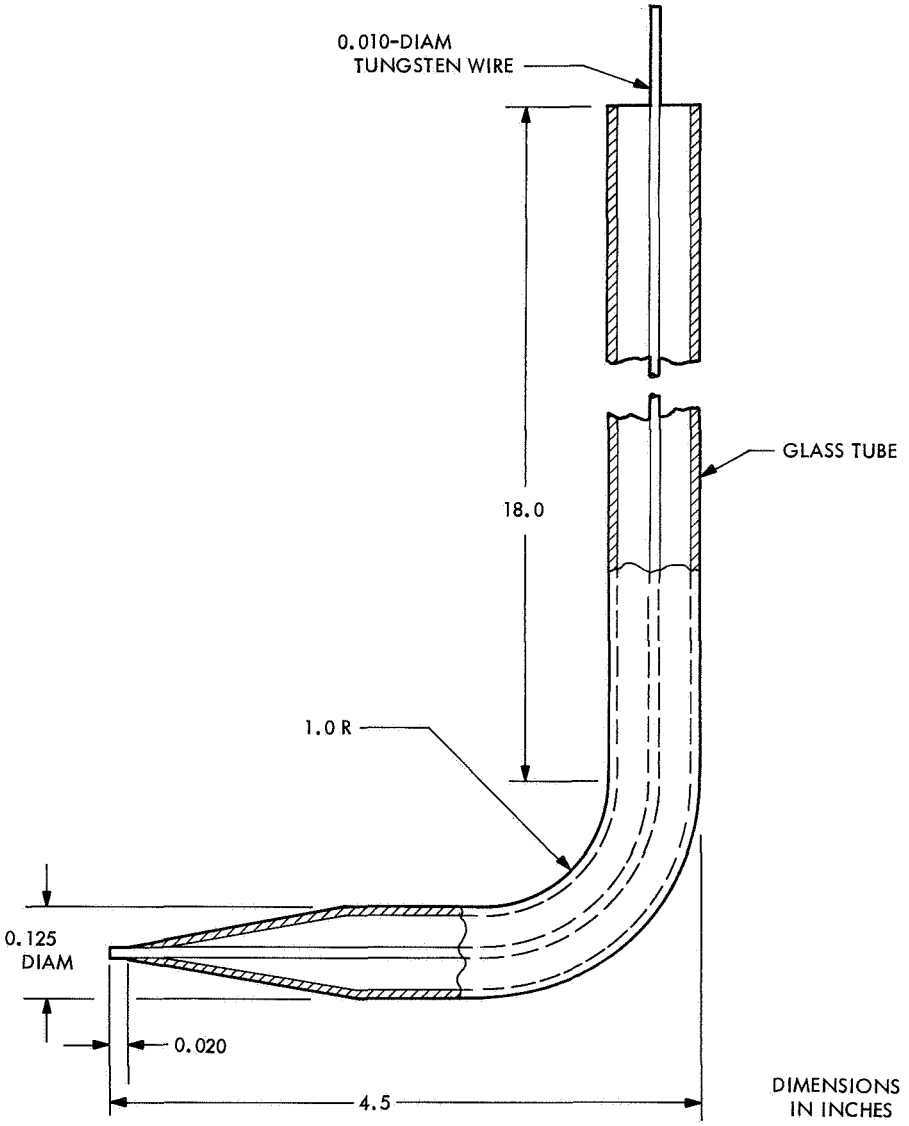


Fig. 11. Glass tube containing tungsten wire

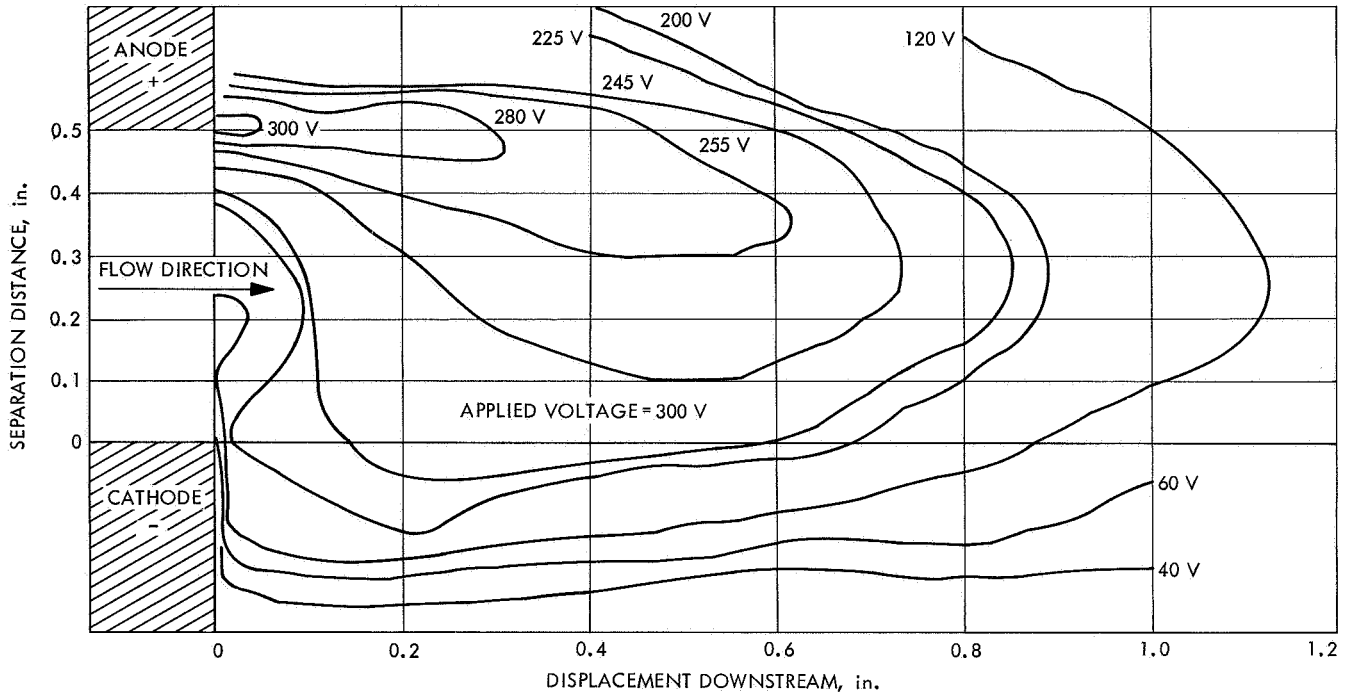


Fig. 12. Potential distribution in plane of plasma discharge affected by 127-ft/s transverse gas flow, 100-torr pressure, and 15-mA current

at the cathode was larger in size than at the anode and extended around the trailing edge. The potential distribution (Fig. 12) in the plane of the discharge is highly distorted in comparison to the static case and appears to correspond with that anticipated from the visual appearance of the plasma column. Constant potential lines shown in Fig. 12 are time-average steady-state values. The applied voltage between the electrodes was 300 V. Assessment of the fall regions cannot be accomplished realistically from Fig. 12, particularly since the data shown is for one plane only. It seems evident, however, that the largest gradient occurred at the trailing edge of the cathode.

4. Conclusions

Severe distortions of the potential distribution, and hence current distribution, caused by a transverse flow field introduce significant complexity into theoretical approaches that might be attempted for computing heat transfer to the electrodes. The flow regions both outside and within the boundary layer of a downstream electrode in a device containing segmented electrodes can be greatly influenced by the effect of blowing on the electrical discharges upstream. For the analysis to be realistic, free-stream conditions at the edge of the boundary layer may require specification of current distributions that are both longitudinal and transverse to the flow direction. Furthermore, these currents may be associated with upstream electrodes.

Reference

1. Thiene, P., "Convective Flexure of a Plasma Conductor," *Phys. Fluids*, Vol. 6, p. 1319, 1963.

XI. Lunar and Planetary Instruments

SPACE SCIENCES DIVISION

A. A Folding Rotating Cup Anemometer,

J. B. Wellman

1. Introduction

A rotating cup anemometer has been designed, fabricated, and tested as a part of the science payload of the capsule system advanced development (CSAD) hard-landing Mars probe. The instrument is designed to survive the rigors of space flight and hard landing and to make meaningful measurements of Martian wind speed. The effort has been limited to developing the mechanical system but space has been reserved for an optical system for sensing rotation, utilizing a light-emitting diode and a phototransistor. Magnetic and capacitive pickups have also been considered.

From the early trade-off studies, the rotating cup anemometer emerged as the most likely candidate for the CSAD wind-speed measurement because of its broad range of sensitivity, linearity, ease of calibration, and simplicity of data processing. The goal has been to design a rotating cup anemometer that would be sensitive enough to satisfy the science goals and also satisfy the sterilization, impact, and geometry requirements of the CSAD lander. The resulting anemometer configuration is not only applicable to CSAD but may be worthy of consideration in other programs that impose severe volume restrictions or require survival under impact.

2. Design Criteria

The criteria that must be met by the anemometer design are of two types: performance (range and threshold sensitivity of the anemometer in a Mars atmosphere) and compatibility (size, weight, sterilizability, and survival under impact).

The performance criteria are determined to a large extent by the nature of the Mars winds. The surface wind speed is expected to vary from 0 to 200 ft/s with gusts as high as perhaps 450 ft/s. The surface pressure should be in the range of 5 to 20 mbars.

A threshold sensitivity requirement for the anemometer has been chosen to be between 5- and 10-ft/s wind speed. Above the threshold the instrument should be linear to 15% of the reading up to 200 ft/s and should survive gusts as high as 450 ft/s.

To be compatible with CSAD the instrument must survive sterilization and must be capable of surviving 2500 g of impact acceleration. The last constraint (and perhaps the most demanding) is that the anemometer fit within a cylindrical volume 1 in. in height and 1.5 in. in diameter.

3. Analytical Model

In order to determine the effects of anemometer geometry on its performance, an analytical model was investigated. The performance of a rotating cup anemometer is characterized by the ratio of the cup tangential velocity to the free-stream wind velocity. This performance factor can be calculated from experimental data by taking measurements of lift and drag forces on an anemometer cup as a function of angle of attack. The lift and drag forces are then resolved normal to the plane of the cup rim and a normal force coefficient is calculated (Ref. 1). The performance factor k is given by

$$k = \frac{\dfrac{C_1 - C_2}{C_1 + C_2} - \epsilon}{3 - \epsilon^2}$$

where C_1 and C_2 are the average normal force coefficients for the backward and forward directions of the cup and $(\pi/2) - \epsilon$ is the angle of attack at which the normal force coefficient passes through zero (Ref. 2). Since the lift and drag coefficients vary with Reynolds number, the performance factor will also, in general, vary with Reynolds number.

Calculations were made for a variety of cup shapes. The hemispherical cup and the conical cup exhibited performance factors several times greater than those of other geometries, such as vanes formed as sectors of a cylindrical shell. Furthermore, the conical cup exhibited a lesser dependence on Reynolds number than did the hemispherical cup. The conical cup with performance factor of 0.27 to 0.30 was considered the best cup geometry.

The threshold wind velocity is that for which the aerodynamic torque exerted on the cups is equal to the starting torque of the bearings. From the torque balance equation, it can be shown that the threshold velocity varies as the $-3/2$ power of the linear dimensions of the anemometer. In terms of the cup area A, the radius of the anemometer r, and the number of cups N, the threshold velocity V_s satisfies the proportionality

$$V_s \propto A^{-1/2} r^{-1/2} N^{-1/2}$$

The conclusion of the analytical study was that the anemometer should have three conical cups of 45-deg half-angle and that the dimensions of the anemometer should be as large as possible within the specific constraints prescribed.

4. Evolution of the Design

It was apparent from the analytical study that in order to satisfy the performance criteria the anemometer would need to be larger than the 1.5-in.-diam cylindrical volume in the lander; thus, a collapsible anemometer would be required.

The largest cone of 45-deg half-angle that could be fitted within the prescribed cylindrical volume would be 1.5 in. in diameter. Three of these cones could be included by nesting them within one another and locating them coaxially within the cylinder. The remaining space within and below the cups would be allocated to the hub and arm assembly. The maximum arm length could be achieved if the arms were nearly 1.5 in. long and hinged at one end to the hub so that they could rotate 180 deg to extend nearly to their full length outside the 1.5-in.-diam volume. This would require that the arms overlap in the folded configuration. In order to move the three arms with their attached cups to the open position with the arms 120 deg apart and the axes of the cups horizontal, a set of spring-loaded hinges would be necessary.

A system of three orthogonal hinge axes for each arm was considered first. The mechanical complexity of this type of arrangement was considered a significant drawback. A method of reducing the number of hinges from nine to six was envisioned. The inner hinge which attaches the arm to the hub would allow the arm to rotate outward 180 deg in the horizontal plane. The skewed axis that attaches the cup to the arm would allow the cup to rotate from its vertical folded position to the horizontal unfolded position with minimal interference among the cups. A preliminary model was built to test this type of hinge arrangement. The unfolding mechanism proved workable and a design incorporating shock resistance was accomplished.

The anemometer expands from its folded configuration (Fig. 1) of 1.5 in. in diameter to a deployed configuration (Fig. 2) of 5.5 in. in diameter. The weight of the anemometer alone is 0.95 oz.

Several aspects of the design can be visualized in Fig. 3. In the folded configuration the hub rests directly

on the base, thereby relieving the load from the miniature ball bearings. The cups are supported by the cap, which transmits the load to the hub and also encloses the upper bearing to prevent contamination by particulate matter. The base is designed to fit congruently into the hub both to relieve the load during impact and to provide a baffle for the lower bearing. When the anemometer is permitted to deploy, the hub is raised a short distance from the base by a spring-loaded bushing. Rotation sensors can be incorporated into the base by machining the required cavity.

Fig. 1. Anemometer in folded configuration

5. Preliminary Testing

The unfolding process was tested by repeated operations and a time study was made using high-speed motion pictures. It was observed that the motion about the inner and outer hinge joints took place concurrently. The unfolding reached completion in less than 80 ms.

When the reliability of the unfolding mechanism had been demonstrated, the anemometer was mated to the collapsible boom. The assembly was collapsed into the boom retainer, and the combined package was subjected to a number of shock tests up to a maximum of 4000 *g*. In each case the anemometer survived with no damage.

In the next stage of evaluation the anemometer and boom were installed in the CSAD lander. The lander was

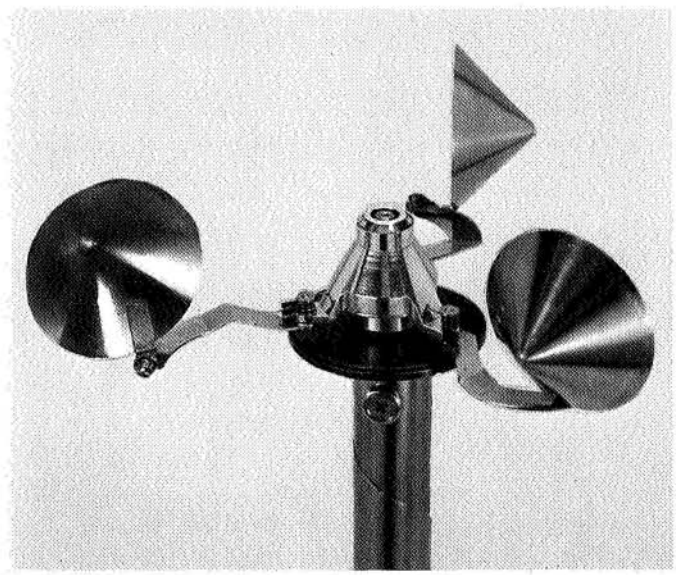


Fig. 2. Anemometer in deployed configuration

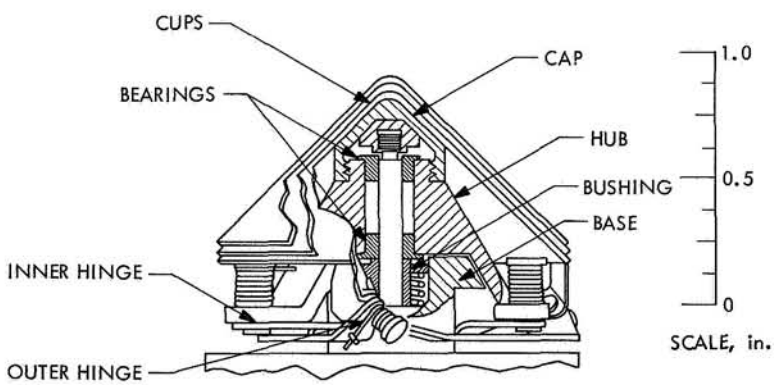


Fig. 3. Cut-away view of anemometer in folded configuration

then sterilized at 125°C for approximately 16 h. Two drop tests of the CSAD lander were made from 250 ft into the dry lake bed at Goldstone, Calif. The capsule velocity at impact was 120 ft/s. In both cases the boom extended on command and the anemometer deployed and began rotating in the wind (Fig. 4).

The compatibility of the anemometer with the CSAD lander was demonstrated by these tests. However, it was observed that the anemometer interfered somewhat with the radio transmission from the lander; the full import of this observation is not yet known. Some changes in the materials used in the cups may be necessary to overcome this problem.

Fig. 4. Anemometer and CSAD capsule after drop test

A brief experiment in a Mars atmosphere flow system was conducted to determine the threshold sensitivity of the anemometer. A value of 10 ft/s at a pressure corresponding to 7 mbars of Mars atmosphere was observed. Although the threshold measurements are incomplete, it appears that the present design is capable of satisfying the threshold criterion. More extensive threshold and linearity measurements will be made.

References

1. Brevoort, M. J., and Joyner, U. T., *Experimental Investigation of the Robinson-Type Cup Anemometer*, Report 513. National Advisory Committee for Aeronautics, Washington, D.C., 1935.
2. Corcoran, J. W., and Esau, D. L., *Comparison of a Theoretical Model for Anemometer Cups with Experimental Data*. Beckman and Whitley, Inc., Mountain View, Calif., Oct. 1964.

B. Selection of Wind Measurement Instruments for a Martian Lander, *J. M. Conley*

1. Introduction

Selection of instrument types for the measurement of wind velocity at the surface of Mars has received considerable study during the past several years (Refs. 1–3). Unfortunately, very little experimental work has been done toward determining the parameters necessary for arriving at rational selections of instrument types. This preliminary parametric study of all known instruments will be followed by experimental measurements under Mars surface conditions in order to more thoroughly investigate those instruments that appear to be promising.

Any complete trade-off study must, of course, be mission-dependent; i.e., the instrument selection process is a function of the mission characteristics, such as launch and interplanetary environments, landing shock, lander size, and orientation capability. During the progress of this work, the probable configuration of the first Mars lander has varied from the large *Voyager* soft lander to small, short-life hard landers and, for this reason, these studies have been broadened. The primary emphasis has been on determining the present state of development of the various instrument types and identifying those parameters for which more information is required.

A small wind tunnel has been completed to make the required measurements under Mars surface conditions and will be described in a subsequent SPS article.

2. Wind Instrument Constraints and Characteristics

Theoretical studies (Ref. 4) and observations of the Martian yellow clouds indicate that the expected near-surface wind speeds are in the range of 0 to 60 m/s and that maximum continuous wind speeds up to 140 m/s can be expected. For the JPL series of VM-1 to VM-10 model atmospheres, it is assumed that the dynamic pressure is approximately constant for the various densities (actually $\rho^{1.48}V^2$ = constant). The presently accepted Mars surface atmospheric density falls in the range of 1×10^{-5} to 3.5×10^{-5} g/cm^3. An atmospheric temperature range of approximately 145 to 320°K must be accommodated. Other constraints and instrument properties that must be considered are discussed in general terms.

a. Quantity measured. The functional relation between the instrument output and the atmospheric variables affecting it is described under this heading. It varies from a linear proportionality to wind speed for tracer-type instruments to a complex function of gas thermal conductivity, viscosity, pressure, temperature, and wind speed for the thermal transport (hot wire)-type instruments.

b. Component measured. The geometry of some sensors makes them sensitive to particular wind components. For example, the rotating cup and hot wire anemometers measure the component perpendicular to the instrument axis, and no amount of manipulation of the signal will allow the vector wind to be completely specified from the basic instruments. However, since mission constraints on size, weight, and data transmission may be severe for early Mars landers, thorough investigation of the simpler devices is desirable.

c. Accuracy. The accuracy of an anemometer is usually specified in terms of percent of full scale, but since relatively crude measurements should be acceptable in the present application and since the wind-speed range of the models is so great, it is desirable to consider the accuracy as a percent of reading. Thus, a logarithmic response has merit. Stability of the instrument zero and sensitivity and lack of hysteresis are the determining factors for this parameter since fixed nonlinearities can be readily compensated.

d. Range. All of the instruments considered are potentially capable of operation over the full range of postulated Mars wind speeds, but design trade-offs inflict some severe penalties on several of the instruments in return for a large dynamic range.

e. Distance constant or frequency response. The frequency response requirement has not been specified for any proposed mission but is given for each instrument. Unless some on-board harmonic analysis is done, it is not likely to be of importance. Available response varies from the kHz region for thin hot wires to 0.02 Hz for a rotating (cup or propeller) anemometer at low wind speeds. In some cases, the distance constant is a more appropriate parameter. This is the length of a column of moving air required to produce a 63% response, or alternately, the product of the time constant and the wind speed.

f. Atmospheric temperature effect. Many of the anemometer types suffer from a strong sensitivity to variations in atmospheric temperature. Additionally, the sensor itself will be exposed to the Martian atmospheric temperature and deployed electronics will need to operate at 145°K or be heated.

g. Blown dust effect. The erosion and contamination effects of Martian dust storms may severely affect some instruments. Examples are possible breaking of a thin hot wire and contamination of the bearings of a rotating device.

h. Sterilization. Both chemical and thermal sterilization may be required for instruments landed on Mars. Chemical sterilization may be necessary only in case of a high-spore population prior to the heat cycle and would consist of immersion in an atmosphere of 12% ethylene oxide and 88% Freon 12. The terminal heat sterilization would be attained by heating to approximately 125°C for 24 h.

i. Shock and vibration. Landing shocks for a hard lander are expected to be in the 1200- to 2500-*g* range with a duration of 1 to 3 ms. Type-approval shock levels may be as great as 5000 *g*. Mars atmospheric entry accelerations will be of the order of 250 *g* with pulse half-width of several seconds. The launch vibration environment is not so readily specified since the vehicle is unknown, but sinusoidal accelerations of 15 *g* rms in the 100- to 2000-Hz range are typical.

j. Deployment considerations. The mission constraints will probably require that the anemometer be stowed within or near the lander during entry and landing. Thus, the anemometer and mast must be stowed until some time after landing and then deployed. A somewhat arbitrary deployed height of 1 m has been selected. (Greater heights are certainly desirable and under many circumstances easily obtainable.)

The difficulty of obtaining a stiff deployable mast under the severe weight restrictions of a small lander makes it highly desirable that the cross section presented to the wind by the anemometer be small. A bending load of 0.03 lb/in.^{2} of mast and instrument surface is produced by a 140-m/s wind. For a mast 1 in. in diameter by 6 ft in height, this may be an appreciable force, adequate to cause oscillation of the mast, malfunction of some instruments, or erroneous data. A small, lightweight sensor is thus highly desirable. Also, some proposed masts have very little capability for routing wires to the sensor. Thus, a minimum number of the smallest possible lead wires is desirable. Radiation from other deployed sensors may necessitate shielded wires or a very low-impedance sensor.

An early lander will probably not be leveled; therefore, the anemometer should be designed to operate when the lander is resting at a large angle with respect to the horizontal. The *Voyager* constraints specified a 35-deg angle, which would produce an error of 18% for a cosine law response to wind velocity. This could in some cases be corrected if the lander orientation were known.

k. RF interference. RF reflectors should be kept as far as possible from the telemetry antennas. This may require that the anemometer be kept lower than is desirable if the antenna is deployed on the same or another mast. A large, high-gain antenna (or solar cell panel) may produce severe interference with the wind flow in this case. Ideally, the anemometer sensor and mast would be fabricated of RF transparent material and thus could be deployed far enough above the antenna to be out of the region of influence (perhaps 10 antenna diameters).

A more subtle consideration is the presence of rapidly moving metal parts (e.g., rotating anemometers) on the lander. Under some conditions, the radio signal reflected from such parts may be of great enough amplitude to produce severe multipath effects, and the doppler frequency may be great enough to cause a narrow-band phase-locked receiver to lose lock. It is therefore desirable (or perhaps necessary) that any "rapidly" moving parts be fabricated of RF transparent material.

l. Thermal vacuum. Cold welding during the interplanetary cruise must be considered for any instrument employing moving parts. However, this problem must also be solved for a deployable mast and many other instruments. If necessary, a low-pressure atmosphere could be provided in the instrument compartment. The Mars atmosphere obviates this problem during operation.

In addition to the mission constraints and instrument characteristics described, other factors that must be considered are radioactive thermal generator environment, space radiation, deployment shock, launch pressure profile, total weight, deployed sensor weight, volume, power, and cost.

3. Anemometer Types

All of the known anemometers have been classified according to the physical principle of operation (Table 1). In some cases a further division is convenient and has been made according to whether the measurement is made locally (immediate vicinity of the lander) or remotely (tracking a balloon, etc.). Since the limited missions that are likely for the immediate future will probably preclude the use of the remote technique, it will not be considered at this time. Only the significant characteristics are discussed for each instrument type; none of the other constraints is expected to be critical.

a. Thermal transport. These instruments, typified by the hot wire anemometer, operate on the principle of forced convective cooling of a heated element by the wind. They possess several distinct advantages, particularly high sensitivity, rapid response, and the easy deployment associated with their small size. Unfortunately, the stability of the sensitivity and zero are quite poor due to contamination problems and the output is sensitive to gas composition, pressure, and temperature.

A wide variety of instrument forms has been reported, varying from the Kata thermometer (a heated, large-bulb, alcohol thermometer) to fragile platinum wires 20 μin.

Table 1. Anemometer types and typical examples

Anemometer type	Typical instrument examples
Thermal transport	Hot wire, hot film, heated thermistor
Dynamic pressure	
Pitot	Servoed pitot, multiport pitot devices
Drag device	Drag bodies with strain gages or other pickups
Rotating	Rotating cup, propeller, wind vane
Sonic	
Remote	Rocket grenade experiment
Local	Pulse or continuous wave transmission over fixed baseline
Tracer	
Remote	Radar- or laser-tracked chaff, aerosols, shock wave
Local	Ion or thermal gradient tracer
Vortex frequency	Vortex shedding cylinder, hot wire detector

in diameter by 0.005 in. in length. The most suitable type for the present application would seem to be a quartz-coated cylindrical hot film. The sheathed or quartz-coated wires and films are much less susceptible to contamination than the thin hot wires, although at the expense of sensitivity and frequency response. The problem of sensitivity to gas composition, pressure, and temperature is more difficult. It may be possible to reduce the data after these parameters are known but the experiment would then be compromised by dependence on other instruments. Another possibility is that the instrument be calibrated on-site by means of pressure measurements or that the zero be determined by means of a mechanical device, which would intermittently shield the sensor from the wind. Temperature compensation is also possible. None of these techniques can be fully evaluated until the actual magnitude of the zero and sensitivity variations of a specific instrument under mission conditions are determined for the extremes of Martian model atmospheres.

Quantity measured. The relationship between flow velocity and heat transfer rate is given by King's law:

$$H = k\theta + (2\pi k\, c_v\, \rho d_s V_s)^{1/2}\, \theta$$

where

θ = difference between wire and fluid temperature

k = thermal conductivity of fluid

c_v = specific heat at constant volume

ρ = density

d_s = sensor diameter

V_s = fluid velocity normal to sensor

Since the heat transfer rate is proportional to the power, we have, for operation in the constant resistance mode and suppressed zero,

$$E_0 \propto V_s^{1/4}$$

where E_0 is the instrument output voltage.

Component measured. The magnitude of the wind component normal to the axis is measured. The wind azimuth determination has a fourfold degeneracy when two orthogonal horizontal sensors are used.

Accuracy. The accuracy of this type instrument is presently unknown under mission conditions. One percent or better can be achieved after calibration under laboratory conditions.

Range. Operational range is approximately 3 to greater than 140 m/s.

Frequency response. Frequency response is very high (greater than 10 kHz).

Atmospheric temperature effect. Effect of atmospheric temperature is severe, but can be compensated to some degree.

Blown dust effect. Limited measurements reported in Ref. 3 indicate that a shielded sensor is not abraded under simulated Martian conditions.

Shock and vibration. The sensors are normally considered to be delicate since they are small, but should readily support their own weight under acceleration. Resonant frequencies are probably higher than any shock or vibration components likely to be transmitted to the sensors.

Deployment considerations. The sensor itself can be deployed very readily. Deployed calibration devices could add great complexity.

b. Dynamic pressure. The instruments of this class can be conveniently divided into those which utilize a pressure gage to determine the dynamic head produced by the wind at ports (pitot devices) and those which measure the drag force on an object placed in the wind stream (drag devices). All operate on the principle of determining the dynamic pressure associated with the wind motion and given by $\frac{1}{2}\rho V^2$, where V is the wind speed. They have the common disadvantages of low sensitivity at low speeds and a dependence on atmospheric density. The dynamic pressure due to a 3-m/s wind at an atmospheric density of 2×10^{-5} g/cm^3 is 9×10^{-4} mbars, and that for a wind of 140 m/s in a 3.5×10^{-5}-g/cm^3 atmosphere is 3.5 mbars. Thus, a dynamic range of 3800 is required. Small pressure gages reputed to meet these specifications have been built. The drag-type instruments require either precise leveling, very stiff masts, or heavy counterbalances. (One type actually measures the drag force on the mast itself.) If the dynamic pressure method is used, the deployment considerations strongly favor the pitot device.

Several pitot devices have been proposed. A pitot-static tube servoed to point into the wind, multiport

devices that utilize the ratio of heads at ports symmetrically located around a vertical cylinder, and a horizontal tube mounted on a radial arm and rotated rapidly around a vertical axis are typical. The rotating tube reduces the required dynamic range but at the expense of considerable mechanical complexity. The multiport schemes would almost certainly require leveling to approximately 5 deg. The following considerations are applicable to any of the pitot-type devices.

Quantity measured. The pitot-static tube yields $\frac{1}{2}\rho V^2$, whereas a simple pitot gives $p + \frac{1}{2}\rho V^2$, where p is the ambient atmospheric pressure. This holds within 5% for yaw angles less than 15 deg and Reynolds numbers greater than 30. Gas compressibility effects must be considered for Mach numbers greater than ~ 0.4.

Component measured. The magnitude and direction of the horizontal component are obtained.

Potential accuracy. Five percent or better accuracy is estimated with leveling, unknown without.

Range. Operational range is 3 to 140 m/s.

Frequency response. Frequency response requirement is approximately 10 Hz for a sensitive gage. Long-pressure tubes will slow the response.

Deployment considerations. Severe difficulties with some masts exist due to pressure tubes or large deployed gages.

c. Rotating. This category includes not only the rotating cup- and propeller-type instruments but also the wind vane, which yields direction only. They are the most popular earth anemometers because of their relatively low cost, high accuracy, and simplicity. They are also relatively insensitive to atmospheric density. All operate on the principle of aerodynamic lift and, therefore, require that the Reynolds number be adequate (~ 100) to establish good circulation. This is satisfied by a 1-in. chord at 3 m/s in the least-dense Mars atmosphere. The operation threshold is then established when the torque produced by the lifting surface exceeds the bearing starting torque. Theory and experiment indicate that thresholds in the range of 1 to 3 m/s can be achieved under Mars conditions of gravity (3/8 earth g) and atmospheric density. The devices are nonlinear at low wind speeds due to the increasing bearing loads and to Reynolds number effects. Development of bearings suitable for space flight would require considerable effort.

Quantity measured. For wind speeds above threshold the response is directly proportional to wind speed except for the low-speed nonlinearity mentioned above. Wind-tunnel calibrations of specific instruments under mission conditions are needed to establish the magnitude of these effects.

Component measured. The rotating cup measures the magnitude, not the angle, of the component normal to its axis. The propeller measures the component parallel to its axis; some propeller designs yield a nearly cosine response. Three of these would give the vector wind. The vane may be used either to point a propeller into the wind or to measure only wind azimuth or elevation.

Accuracy. A 1% magnitude accuracy or 5-deg direction accuracy should be achievable under Mars conditions. Without leveling this would be reduced to about 20%. Wind-tunnel tests under mission conditions are needed.

Range. Operation from 3 to 60 m/s should be readily achieved. The instruments will survive to 140 m/s but their operation at high speeds requires further investigation.

Distance constant. This parameter will fall in the range of 50 to 300 m. Wind-tunnel measurements are needed.

Atmospheric temperature effect. The rotation sensor must either operate at a temperature of 145°K or be heated. Sensitivity of the anemometer calibration to gas temperature is small.

Blown dust effect. The bearings must be shielded and a baffle or labyrinth seal should be provided.

Sterilization. The bearing lubricant must not be adversely affected.

Shock and vibration. Small instrument bearings can survive the 5000-g shock when very lightly loaded. A light load will also be required during vibration to prevent chatter.

Deployment considerations. The rotating cup anemometer can be folded into a compact form and readily deployed as described by Wellman (see *Section A*). Deployment of a two- or three-axis propeller or a vane would require considerably more space.

RF interference. Two possible methods of solving this problem are apparent. The cups and arms or vanes can be fabricated of a dielectric material, or the instrument can be electrostatically shielded by a large disk at the base. The second method may influence the wind-flow pattern excessively.

d. Sonic. Sonic anemometers are based on the fact that sound waves are propagated at a fixed speed with respect to the medium. Thus, the apparent speed with respect to a stationary observer is modulated by the wind. All of the instruments in this category utilize this effect to determine wind speed. The speed of sound c in a gas is given by

$$c = \left(\frac{\gamma RT}{M}\right)^{1/2}$$

where

γ = specific heat ratio
R = universal gas constant
T = absolute temperature
M = mean molecular weight

Thus, the instrument is quite sensitive to both gas composition and temperature.

The most promising sonic anemometer presently available is a pulse-type device which, using six transducers at opposite ends of three baselines, measures the three orthogonal components of wind speed. The instrument is sensitive to the speed of sound and this may be a major barrier to its use; it would also be large and difficult to deploy. However, it does possess one very important feature. If nearly calm conditions should prevail during any measurement period (approximately 30 ms), and an independent measurement of the air temperature is available, then a very precise measure of the speed of sound, and thus γ/M, will be obtained. An additional measurement of the received signal amplitude would yield the acoustic impedance ρc, and thus M and γ. In this manner not only the wind speed but also the important thermodynamic properties of the atmosphere would be determined.

Quantity measured. The difference in propagation time in opposite directions is measured. This time difference Δt is given by

$$\Delta t = \frac{2L}{c^2} \frac{V_b}{1 - V_b^2/c^2}$$

where L is the baseline length and V_b is the speed of the wind component parallel to the baseline.

Component measured. One pair of transducers yields the component parallel to the line joining them; three orthogonal sets give the vector wind.

Accuracy. A very sophisticated earth atmosphere instrument is capable of $\pm 3\%$ with temperature compensation and known atmospheric composition. A detailed study as well as experimental measurements would be required to estimate the instrument performance under Mars conditions. The nonlinearity should be of no consequence.

Range. Theoretically, the device can not operate beyond the speed of sound. In actual practice turbulence around the sensing heads will probably set a much lower limit, perhaps about 60 m/s. This effect will be accentuated by the large sensors required because of the low acoustic impedance of the Mars atmosphere.

Distance constant. This is several times the baseline length, which should be about ⅓ m.

Atmospheric temperature effect. Besides the effect on the speed of sound, the transducers must either operate at 145°K or be heated. This would require a minimum of several watts per head and any insulation would aggravate the turbulence problem.

Blown dust effect. Impact of dust on the transducers would necessitate either higher signal levels or a coded pulse, such as a chirp.

Shock and vibration. Articulated arms may be used to deploy the sensors and multiple caging mechanisms may be required for restraint. Transducer crystal breakage may be a problem.

Deployment considerations. The minimum baseline length would be about ⅓ m. Deployment of a two- or three-axis instrument from a folded configuration may be difficult. Also, the sensors would need to be some distance from sound reflectors to prevent multipath effects. Wind-drag loads on the boom would be high.

RF interference. The matrix of arms used to support the transducers may seriously perturb antenna patterns.

Volume. Stowed volume would be high.

e. Tracer. The tracer category includes all anemometers that record the motion of an individual element of fluid. This may be accomplished by injecting and tracking smoke puffs, ion clouds, temperature gradients, balloons, radar chaff, and shock waves. This is the only physical principle for which the true fluid velocity would seem to be measured directly. The output of the sensor may be in terms of sequential positions of the tracked elements (theodolite-tracked balloon), the traverse time for a fixed baseline (ion tracer anemometer), or a velocity (doppler radar or laser beam tracking aerosols). Thus, although the velocity seems to be measured directly, the effects of transducers, diffusion of the tracer element, turbulence, and other factors must be considered.

The local tracer instruments consist of a source of ions or heat located at the center of a circular detector which may consist of either a series of integral detectors, a loop of wire, or a spherical screen. The source is pulsed and the transit time of the ions or thermal gradients is determined. One severe disadvantage of this device is that wind components normal to the plane of the detector may cause the ions to miss the detector. The spherical screen detector obviates this problem but complicates data interpretation.

Quantity measured. The transit time of the tracer is measured and is related to the wind speed by $V = \ell/t$, where ℓ is the distance from source to detector and t is the transit time. The source can be retriggered by the detector, in which case the pulse frequency is proportional to wind speed.

Component measured. The magnitude of the component in the source-detector plane is measured. Direction can also be determined by means of schemes such as discrete detectors at azimuth angle intervals of 360/m deg, where m is any integer.

Accuracy. Accuracy is determined by the finite size of the ion cloud, which is influenced by diffusion and turbulence. Accuracies of 5% have been quoted. Wind tunnel tests are required.

Range. Range is dependent on the size of the instrument. Five to 50 m/s has been quoted and a greater range should be feasible.

Distance constant. The distance constant is equal to the detector diameter, of the order of ⅓ m or less.

Deployment considerations. The detector diameter must be determined before specific data can be given. However, even a spherical detector should be readily foldable.

RF interference. No moving parts are involved but a large detector may perturb antenna patterns.

f. Vortex frequency. When a cylinder is immersed in a moving fluid, it sheds vortices alternately from the two sides. Close behind the cylinder these vortices are well defined between Reynolds numbers of about 40 to 2000. They persist to some degree up to Reynolds numbers of 10^5 or greater, but are not as well defined. The frequency of shedding n is approximately proportional to the wind speed and is given by $n = S(V/d)$, where S is the Strouhal number and d is the diameter of the cylinder. The Strouhal number is approximately constant and equal to 0.21. More precise expressions, involving the Reynolds number, have been developed empirically. The technique of measuring the vortex frequency by placing a hot wire anemometer in the wake of a cylinder has been widely used in wind-tunnel work. Maintaining the detector downstream of the cylinder would, however, be difficult. It might be possible to construct the cylinder of piezoelectric material and detect the pressure changes associated with the vortex shedding. The same method might be used with a cylindrical hot film anemometer.

Quantity measured. A frequency proportional to the wind speed is determined. There is a dependence on kinematic viscosity at low speeds.

Component measured. The magnitude, but not the direction, of the component normal to the cylinder axis is measured.

Accuracy. Low-speed accuracy will be dependent on a knowledge of the kinematic viscosity of the atmosphere. At high speeds, accuracies of the order of 2% are obtained.

Range. The dynamic range for a given cylinder diameter is 50:1 for a relatively clean signal. The use of a phase-locked tracking filter may extend this range to 1000:1 if lock can be acquired and maintained.

Frequency response. The lowest vortex frequency would be of the order of 5 Hz for a 1-in. cylinder and a 1-m/s wind speed.

Atmospheric temperature effect. This effect would be small and readily corrected.

Blown dust effect. An acoustic transducer would suffer from dust storms but a quartz-sheathed hot film should be resistant.

Deployment considerations. A very small and rugged sensor should be possible.

4. Discussion

Preliminary studies indicate that the only relatively compact and simple instrument that yields both magnitude and direction is the tracer type. It would be highly desirable to fabricate such an instrument and test it under simulated mission conditions. The thermal transport and dynamic pressure instruments do not presently seem to be promising but they will be used as laboratory references in testing the others and may thus be further evaluated. The folding rotating cup anemometer (*Section A*) performed well in the capsule system advanced development program and its performance characteristics at low atmospheric density will be measured. It would also be desirable to determine whether or not the vortex shedding frequency of a single or multiple cylinder can be extracted from the noise with a narrow-band phase-locked tracking filter. All of these experiments can be performed in the JPL Mars wind tunnel.

Further evaluation of the sonic anemometer will require considerably different facilities. However, if one of the simpler anemometers proves adequate, the combination of one of these and a sonic densitometer would perform the function with greater economy of space, weight, and electronics complexity.

The Mars wind tunnel has been used for preliminary tests of the rotating cup anemometer and a threshold of approximately 3 m/s has been measured. Further work awaits evaluation of the low-speed flow profile of the tunnel.

References

1. *Selection of Instruments for Atmospheric Measurements,* Report ED-22-6-106. Martin-Marietta Corp., Denver, Colo., Sept. 1967.
2. *Mars Probe,* Final Report, Vol. 5, Book 4, p. 375, Contract NAS1-5224. AVCO Corp., Space Systems Div., Lowell, Mass., May 11, 1966.
3. *Voyager Phase B Capsule,* Final Report, Vol. 3, Part B, pp. 5.9-1 to 5.9-110, Report F694. McDonnell Astronautics, St. Louis, Mo., Aug. 31, 1967.
4. Anderson, A.D., "Spherical Particle Terminal Velocities in the Martian Daytime Atmosphere From 0 to 50 Kilometers," *J. Geophys. Res.,* Vol. 72, No. 7, pp. 1951–1958, Apr. 1, 1967.

XII. Science Data Systems

SPACE SCIENCES DIVISION

A. High-g Testing Multilayer Laminate Packaging, *J. H. Shepherd*

1. Introduction

Shock testing was recently performed on two subassemblies of the *Mariner* Venus 67 prototype data automation subsystem (DAS). Its purpose was to provide information to the JPL Capsule System Advanced Development Project on the ability of the multilayer board packaging of these subassemblies to withstand shock levels beyond those for which it had been designed. This type of packaging has possible application in the design of a future hard-landed capsule due to features such as a reasonable form factor and a construction which facilitates repair and modifications. This potential usage, plus the opportunity to conduct the tests at very little cost, made these tests desirable at this time.

These subassemblies contained Signetics 400 series integrated circuits mounted in glass packages using gold-wire bonding. One subassembly (20A2) has primarily discrete components with a few integrated circuits on two-sided boards using plated-through holes for interconnections; the other subassembly (20A6) has integrated circuits installed on nine-layer laminates using plated-through holes for interconnections.

2. Testing

Prior to shock testing, the subassemblies were conformal-coated with solithane 113/300. For all but one test the subassemblies were positioned so that the side shown in Figs. 1 and 2 faced the direction in which the unit and fixture traveled. The shock, therefore, was perpendicular to the surface of the board to simulate a worst-case environment. The other test was in shear, with the subassembly connectors facing away from the shock.

Subassemblies were installed in the *Mariner* Venus 67 DAS prototype case, and subsystem tests were performed before and after each shock test.

The first series of tests utilized the JPL 50-ft drop tower, which consists of a carriage (the specimen mounting fixture) that rides down on two guide cables and impacts on lead pads. The size, shape, and thickness of the lead pads varies the shock level.

Table 1 shows the conditions of the first series of tests. Subassemblies showed no visual damage after each test and functioned properly when tested as a subsystem.

For the next test the subassemblies were stacked one on top of the other (Figs. 1 and 2 sides down) in a

Table 1. Drop-test conditions

Subassembly	Drop test	Peak g	Duration, ms	Velocity, ft/s
20A2[a]	2A1 side down	1700	1.25	57.2
20A6[b]	6A1 side down	1700	1.25	57.4
20A2	2A1 side down	2700	1.0	57.5
20A6	6A1 side down	2900	1.0	57.5
20A2	Shear	2950	1.0	57.5
20A6	Shear	2850	1.0	57.5

[a]Fig. 1.
[b]Fig. 2.

dummy capsule and dropped from an altitude of 250 ft on a dry lake bed. The dummy capsule weighed 54 lb and made a crater approximately 4 in. deep on one side and 2 in. deep on the other. Estimated shock for this test was calculated at 1100 *gs*.

The edges of the chassis walls showed minor damage where a mounting shim caused some misalignment. There was no damage to the components, and the units functioned properly when tested as a subsystem.

In order to attain the desired higher *g* levels the JPL slingshot facility was used for the next shock test. The slingshot uses large bungee cords to propel a fixture and specimen along guide rails into an abutment. The shock level is determined by three factors: the impact tool diameter, thickness and type of material of the target mounted in the abutment, and the distance traveled or impact velocity.

Subassemblies were mounted in the fixture with the 2A1 and 6A1 sides facing the abutment. Table 2 gives the conditions of this test.

Table 2. Slingshot test conditions

Sub-assembly	Timer reading, ms/6 in.	Impact velocity, ft/s	Penetration depth, in.	Calculated Average g
20A2	3.973	125.5	0.670	4,400
20A6	3.987	125.0	0.560	4,525

Weight of specimen and fixture, 14.6 lb
Distance from target, 10 ft
Impact tool diameter, 7/8 in.
Target material, copper
Target size, 1.5 × 3 in.
Target thickness, 1.5 in.

3. Damage to Chassis and Components

Visual examination after the slingshot test showed the chassis were bent on all four walls (Figs. 1 and 2); subassemblies webs were bowed toward impact sides 2A1 and 6A1; connectors J-1 through J-4 shells were bent; and pin holding blocks were broken.

Subassembly 20A2 lost the lid from one integrated circuit, and one side of a lid was loose on another. The glass case of the integrated circuit with the lid off was cracked on the inside of the package around the pad to which the die is attached. Due to the web flexing, five of the pulse transformers' cases were crushed, and ten transistors' leads bent against the transipads by coming in contact with the tooling fixture. Fourteen glass diodes' cases cracked on the 2A1 side, and one cracked on the far side (2A2) of the subassembly.

Subassembly 20A2 did not function properly until nine of the fifteen damaged diodes and three integrated circuits were replaced. One of the integrated circuits replaced was on the far side of this assembly; the gold bond wires internal to the integrated circuits had sagged down against the edge of the die, and lead eight was open where it had shorted to ground, melting the wire (Fig. 3a).

Subassembly 20A6 had the case cracked on one filter capacitor, and the lids came off of eleven integrated circuits on the 6A1 side. The glass around the pad to which the integrated circuit die is attached was cracked on these eleven integrated circuits, and the pad raised on some of them. This subassembly functioned properly without changing any components, and the prototype passed a complete subsystems test without malfunctions.

4. Conclusions

Two changes are suggested for the subassembly structure to enable it to withstand shock at the 4,525 *g* level:

(1) A thicker or reinforced web for the subchassis.

(2) A conformal coating thick enough to cover the components entirely.

There were no problems with the multilayer boards either from the shock or the subassembly flexure.

Integrated circuits for use on a possible *Mariner* Mars 1971 flight will most likely have a strengthened ceramic package with a gold-plated molybdenum–manganese die

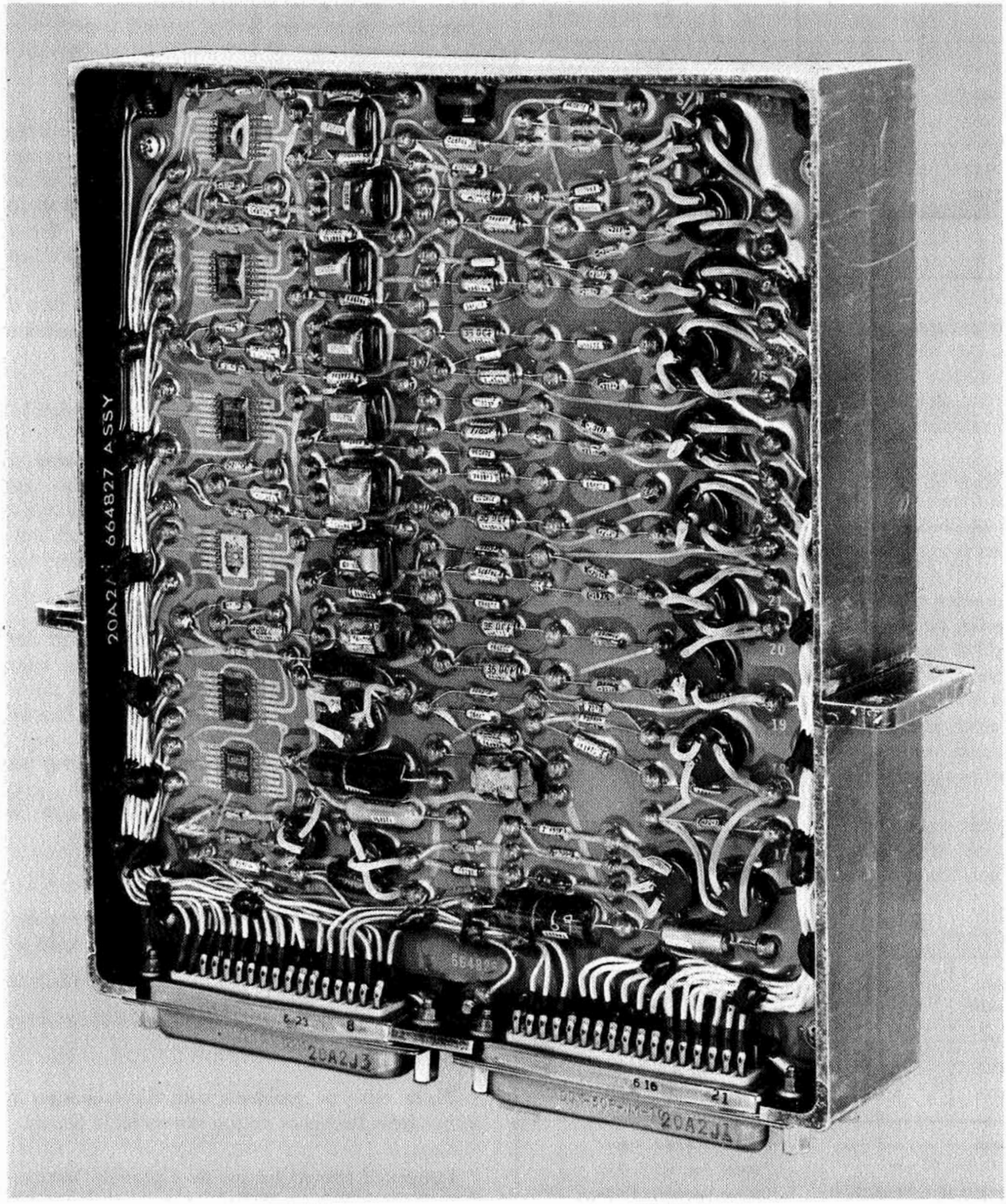


Fig. 1. Damaged DAS discrete assembly 2A1

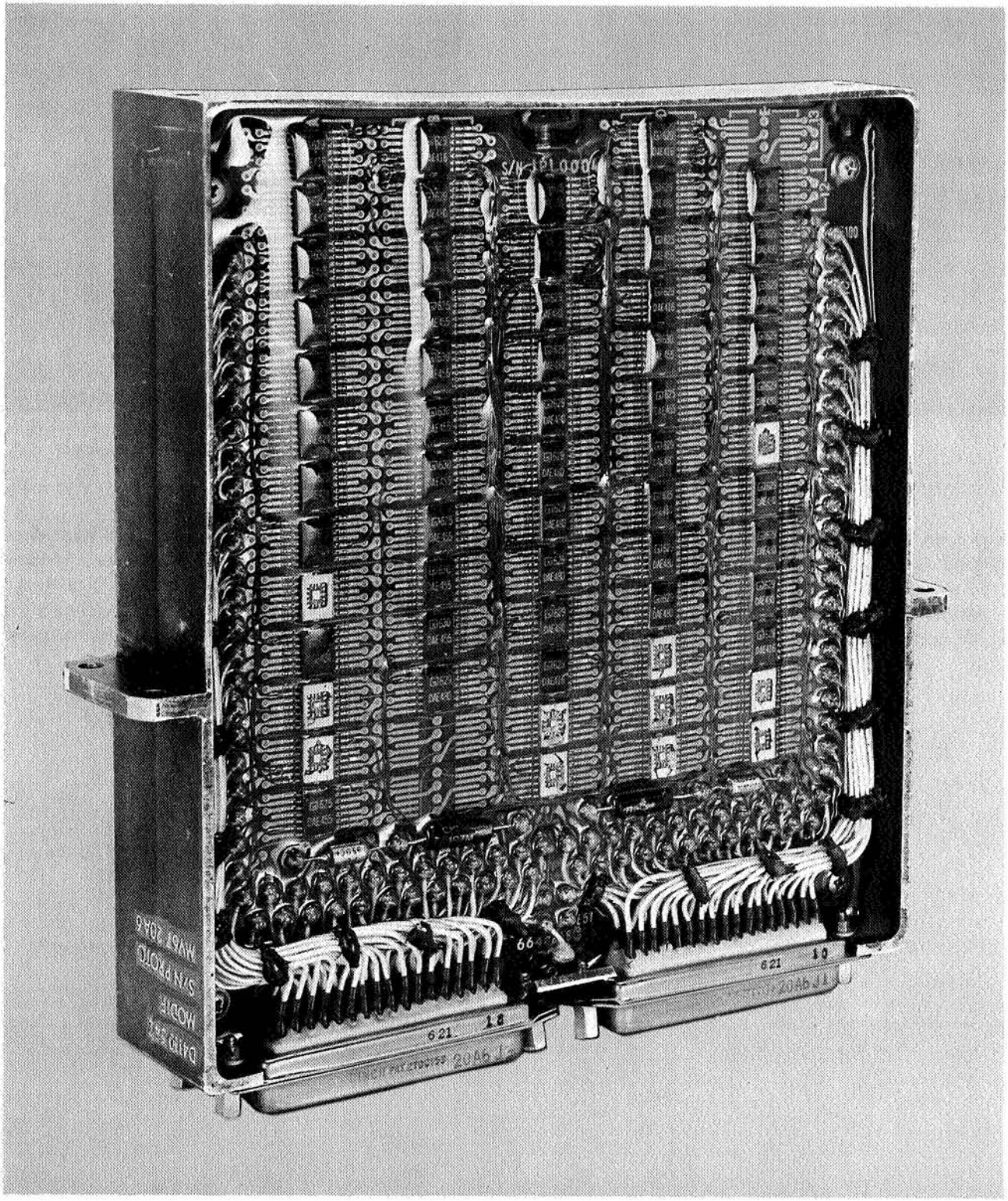


Fig. 2. Damaged DAS logic assembly

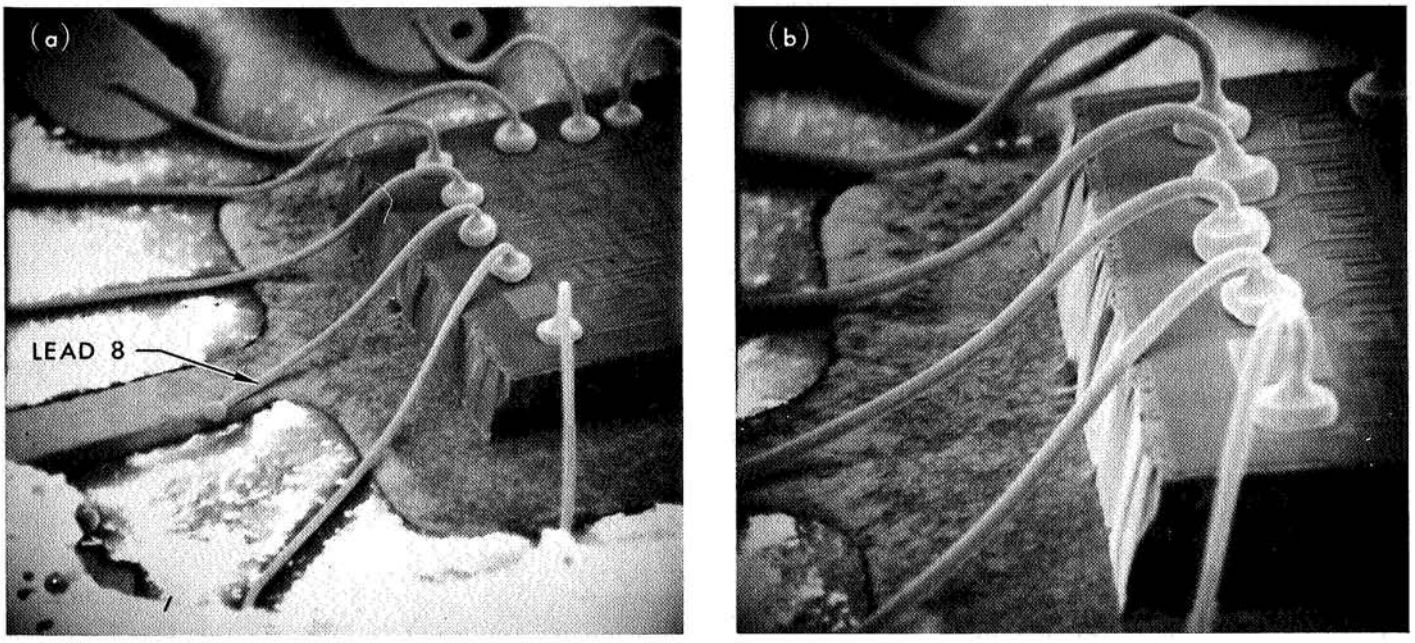


Fig. 3. Integrated circuit photographed by scanning electron microscope showing wire sag: (a) magnified 50 times, (b) magnified 100 times

attach area instead of the gold-plated kovar die attach pad used on the *Mariner* Venus 67 system. This change should prevent the die attach area from separating from the bottom of the package. Aluminum wire will probably be used instead of gold, since it has less mass and does not have a loop in it as the gold wire does. This should solve the sagging wire problem. Stringent environmental and mechanical tests are in process at the integrated circuit manufacturer at this time to qualify the aluminum wire interconnect and ceramic package.

XIII. Physics

SPACE SCIENCES DIVISION

A. An Ion Cyclotron Resonance Study of the Energy Dependence of the Ion-Molecule Reaction in Gaseous HD, *D. D. Elleman, J. King, Jr., and M. T. Bowers*

The ion cyclotron resonance (ICR) spectrometer is a relatively new instrument used in the study of ion-molecule reactions (Refs. 1–4). Whenever possible, it is important to compare these early ICR results with both theoretical predictions and results obtained by more conventional techniques. The following ion-molecule reactions in gaseous hydrogen were the subjects of an ICR study at JPL:

$$H_2^+ + H_2 \longrightarrow H_3^+ + H \qquad \text{(I)}$$

$$D_2^+ + D_2 \longrightarrow D_3^+ + D \qquad \text{(II)}$$

$$HD^+ + HD \begin{cases} \longrightarrow H_2D^+ + D & \text{(IIIa)} \\ \longrightarrow HD_2^+ + H & \text{(IIIb)} \end{cases}$$

These reactions were chosen because of the rather extensive theoretical work performed on the reaction cross sections (Ref. 5). Comparisons can also be made between the ICR data and the results obtained by the more conventional techniques, i.e., high-pressure mass spectroscopy (Refs. 6 and 7), tandem mass spectroscopy (Refs. 8 and 9), and merged beam spectroscopy (Ref. 10).

Both ICR and ion cyclotron double resonance (ICDR) techniques have been used to study the energy dependence of the rate constants of reactions I and II (Ref. 11). In these studies, it became apparent that a more careful determination of the reactant ion energy would be needed. This article describes how the average reactant ion energy in an ICDR experiment can be calibrated using the ratio H_2D^+/HD_2^+ versus the energy data of J. H. Futrell and F. P. Abramson (Ref. 9). The value of the ratio observed in ICDR experiments at various settings of the irradiating field strength $E = E_0 \sin \omega t$ is then compared with the ratio given in Ref. 9.

The ICDR ratio was obtained in the following manner: The HD gas was placed in the *ICR* cell [2.5 cm (trapping field) × 2.1 cm (irradiating field)] and was slightly ionized with a 30-eV electron beam. The H_2D^+ and HD_2^+ resonance lines resulting from reactions IIIa and IIIb were observed with the pulsed-drift-voltage mode of operation (Ref. 3), while the HD^+ resonance line was irradiated

with the second RF oscillator. The intensity of the HD_2^+ line was corrected for its different extent of reaction by multiplying its intensity by 4/5 (Ref. 12). Corrections for the line intensities had to be made since the second RF irradiating electric field heated the ion in the observing section of the cell only, and not in the source region. The observing section of the cell was 5.08 cm long, and the source region where reaction could also take place was 2.54 cm long. Therefore, ⅓ of the intensity of the secondary line was due to thermal reactions that occurred in the source region of the cell. The corrected intensity of the H_2D^+ ion is then given by

$$^{H_2D^+}I_c = {}^{H_2D^+}I_m - \tfrac{1}{3}\,{}^{H_2D^+}I_m^0 \qquad (1)$$

where $^{H_2D^+}I_m$ is the uncorrected measured intensity of the H_2D^+ ion when the HD^+ ion is heated by double resonance, and $^{H_2D^+}I_m^0$ is the measured intensity of the H_2D^+ ion when it is produced by thermal HD^+ ions only. A similar expression gives the corrected intensity for the HD_2^+ ion:

$$^{HD_2^+}I_c = \tfrac{4}{5}\,{}^{HD_2^+}I_m - \tfrac{1}{3}\,{}^{HD_2^+}I_m^0 \qquad (2)$$

where the ⅘ factor takes into account the difference in the extent of reaction for H_2D^+ and HD_2^+ ions. This correction factor is the ratio of the H_2D^+ and HD_2^+ masses, or the ratio of the magnetic field values that satisfies the resonance conditions in the field-sweep mode of operation. These corrected intensities were then used to give the ratio H_2D^+/HD_2^+ in Table 1. Also given in Table 1 are the electric field strength of the second RF oscillator used to heat the primary ions and the laboratory kinetic energy of the ions obtained by comparing the ratio H_2D^+/HD_2^+ to the data of Ref. 9. The fields reported are 80% of those at the plates in the resonance region. (A complete discussion of the electric field profiles can be found in Ref. 13.)

As a further comparison, the energy of the primary ion was determined by two additional methods. For the first, it was assumed that, if the ion energy is limited by collisions, the average ion energy can be estimated by measuring the linewidth at a known value of the electric field strength of the observing oscillator. The average ion energy can then be calculated using (Refs. 14 and 15)

$$E_{\text{ion}} = \tfrac{3}{2}KT + \frac{q^2E^2(m_p + M)}{\delta\xi_p^2 M_p} \qquad (3)$$

where M is the mass of the neutral species relaxing the ion momentum, m_p is the mass of the primary ion, E is the irradiating field strength, ξ_p is the collision frequency related to the line width, and $\frac{3}{2}KT$ is the thermal energy of the ion. The fourth column in Table 1 gives the results obtained using this method of analysis.

Table 1. Observed H_2D^+/HD_2^+ ratio at various *ICDR* irradiating field strengths

H_2D^+/HD_2^+	E_0, V/m	Laboratory kinetic energy, eV		
		Calibration data[a]	Linewidth data[b]	Drift time data[c]
0.73	0	~0	—	—
0.83	1.1	0.2	0.12	0.6
0.89	1.6	0.4	0.20	1.3
0.93	2.0	0.5	0.27	2.0
0.95	2.6	0.6	0.30	3.5
0.95	3.2	0.6	0.36	5.3
0.98	4.0	0.8	0.48	8.3
1.01	5.0	0.9	0.58	13.2
1.02	6.2	1.0	0.68	20.4
1.05	8.0	1.2	0.87	33.2
1.09	10.0	1.4	1.12	51.0
1.16	12.4	1.6	1.33	80.0
1.17	16.0	1.7	1.46	133.0
1.18	20.0	1.7	1.75	215.0
1.19	25.8	1.8	2.16	320.0
1.23	31.1	2.0	2.60	505.0

[a] Extracted from Fig. 8 in Ref. 9, using the ratios in column 1. The error in the extraction is of the order of ±10%.

[b] Calculated from Eq. (19) in Ref. 12 and the experimental *ICDR* linewidths.

[c] Calculated from Eq. (20) in Ref. 12, with the drift time τ calculated from the resonance region voltage. A value of $\tau/2$ was used to calculate the energies reported here.

For the second additional comparison, it was assumed that the ion energy is limited by the amount of time the ion is exposed to the irradiating field. In the low collision limit, the ion is exposed to the irradiating field during the time, τ, that it takes the ion to drift through the observing region of the *ICR* cell. Under these conditions, the energy of the ion is given by (Refs. 14 and 15)

$$E_{\text{ion}} = \frac{q^2E^2\tau^2}{\delta m_p} \qquad (4)$$

The fifth column in Table 1 gives the results of using this approach. It is evident that the collision-limited energies calculated from ICDR linewidths are of the proper order of magnitude, while the drift time energy estimates are grossly exaggerated. The zero irradiating field value of the *ICR* ratio H_2D^+/HD_2^+ corresponds to the thermal value of the tandem mass spectroscopy given in Ref. 9. This observation verifies that the average energy of the reactant ions in an *ICR* cell in the single-resonance mode is thermal, as has been suggested in earlier work (Ref. 12).

References

1. Anders, L. R., et al., *J. Chem. Phys.*, Vol. 45, No. 1062, 1966.
2. Beauchamp, J. L., Anders, L. R., and Baldeschwieler, J. D., *J. Am. Chem. Soc.*, Vol. 89, No. 4569, 1967.
3. Baldeschwieler, J. D., *Science*, Vol. 159, No. 263, 1968.
4. Fluegge, R. A., "Symmetrical Charge Exchange and Ion Atom Reactions," Cornell Aeronautical Laboratory Report UA-1854-P-1. U. S. Department of Commerce Clearing House for Federal Scientific and Technical Information, Washington, D.C.
5. Gioumousis, G., and Stevenson, D. P., *J. Chem. Phys.*, Vol. 29, No. 294, 1958.
6. Stevenson, D. P., and Schissler, D. O., *J. Chem. Phys.*, Vol. 29, No. 282, 1958.
7. Reuben, B. G., and Friedman, L., *J. Chem. Phys.*, Vol. 37, No. 1636, 1962.
8. Giese, C. F., and Maier, W. B., II, *J. Chem. Phys.*, Vol. 39, No. 739, 1963.
9. Futrell, J. H., and Abramson, F. P., "Ion Molecule Reactions in the Gas Phase," *Advan. Chem.*, Series 58, 1966.
10. Neynaber, R. H., and Trujillo, S. M., *Phys. Rev.*, Vol. 170, No. 0000, 1968.
11. Bowers, M. T., Elleman, D. D., and King, J., Jr., *J. Chem. Phys.*, Vol. 49, No. 0000, 1968.
12. Bowers, M. T., Elleman, D. D., and Beauchamp, J. L., *J. Phys. Chem.*, Vol. 72, No. 3599, 1968.
13. Anders, L. R., Ph.D thesis. Harvard University, Cambridge, Mass., 1966.
14. Beauchamp, J. L., *J. Chem. Phys.*, Vol. 46, No. 1231, 1967.
15. Beauchamp, J. L., and Buttrill, S. E., Jr., *J. Chem. Phys.*, Vol. 48, No. 1783, 1968.

B. Observation of Fluorine-19 Isotopic NMR Chemical Shifts Due to Chlorine-35 and Chlorine-37 Isotopes, *E. A. Cohen and S. L. Manatt*

1. Introduction

Isotopic substitution on an atom possessing a nuclear spin has been known for some time to cause changes of nuclear-magnetic-resonance (NMR) shieldings or chemical shifts (Ref. 1). The NMR shielding or chemical shift of a nucleus is defined as the extent to which the NMR field is increased by coupling between the nuclei and the electronic circulation induced by the applied magnetic field. The values of the chemical shift for a particular isotope in different chemical environments can vary tremendously and can be used to characterize very subtle features of local electronic structure.

Usually the NMR shielding of a nucleus in a particular chemical environment is highest when substituted with the heavier isotope of an isotopic pair (Ref. 1). Many examples of this effect on proton and fluorine shifts arising from substitution of a deuterium atom for a proton have been described. Also, numerous examples of the proton and fluorine shifts caused by replacement of ^{12}C by ^{13}C have been reported (Ref. 1). Only a few examples of shifts have been observed for other heavier isotopes, such as ^{28}Si—^{29}Si on ^{19}F, ^{32}S—^{33}S on ^{19}F, ^{32}S—^{34}S on ^{19}F, ^{80}Se—^{76}Se—^{77}Se—^{78}Se—^{82}Se on ^{19}F, ^{12}C—^{13}C on ^{59}Co, and ^{14}N—^{15}N on ^{59}Co (Ref. 1).

Only very approximate theories now exist for this effect in the case of molecules more complicated than the isotopic species of the hydrogen molecule (Ref. 1). Isotopic shifts most certainly arise from small differences in the average bond distances due to differences of molecular zero-point vibrational energies and anharmonic contributions to potential functions. A complete quantitative treatment of these effects would require accurate molecular wave functions, which are not yet available, and rather detailed knowledge of how these functions are affected by the various vibrational degrees of freedom of the isotopic species of a molecule. In general, the isotope chemical shift of a particular nucleus is proportional to the number of atoms in the molecule that have been isotopically substituted (Ref. 1). This article presents the first examples of chlorine isotope shifts on the fluorine-19 resonances in three molecules.

2. Fluorotrichloromethane Molecule

The fluorine-19 chemical shift of fluorotrichloromethane (CCl_3F) has for some time been used as the accepted chemical shift reference compound for fluorine-19 NMR (Ref. 2). Reference 3 states that, "A convenience reference signal is one which is sharp and well separated from other signals in the spectrum." It is common practice, and quite convenient at present, to utilize an internal reference (usually added to a sample subjected to NMR studies) as the control signal for locking the field and frequency of an NMR spectrometer. Such a control system is described in Ref. 4.

In Fig. 1a, the NMR spectrum of CCl_3F is exhibited along with that of hexafluorobenzene (C_6F_6) for a resolution reference. It is clearly evident that the signal from CCl_3F is a doublet, each member of which is significantly broader than the signal from C_6F_6. The doublet nature of the CCl_3F signal explains why a noisy lock signal and poor resolution were obtained when, some time ago, an attempt was made to utilize the CCl_3F resonance as a control signal for field-frequency control of an NMR spectrometer having a magnet exhibiting very high field homogeneity. Apparently, rapid magnetic-field fluctuations and the close proximity of the two axis crossing

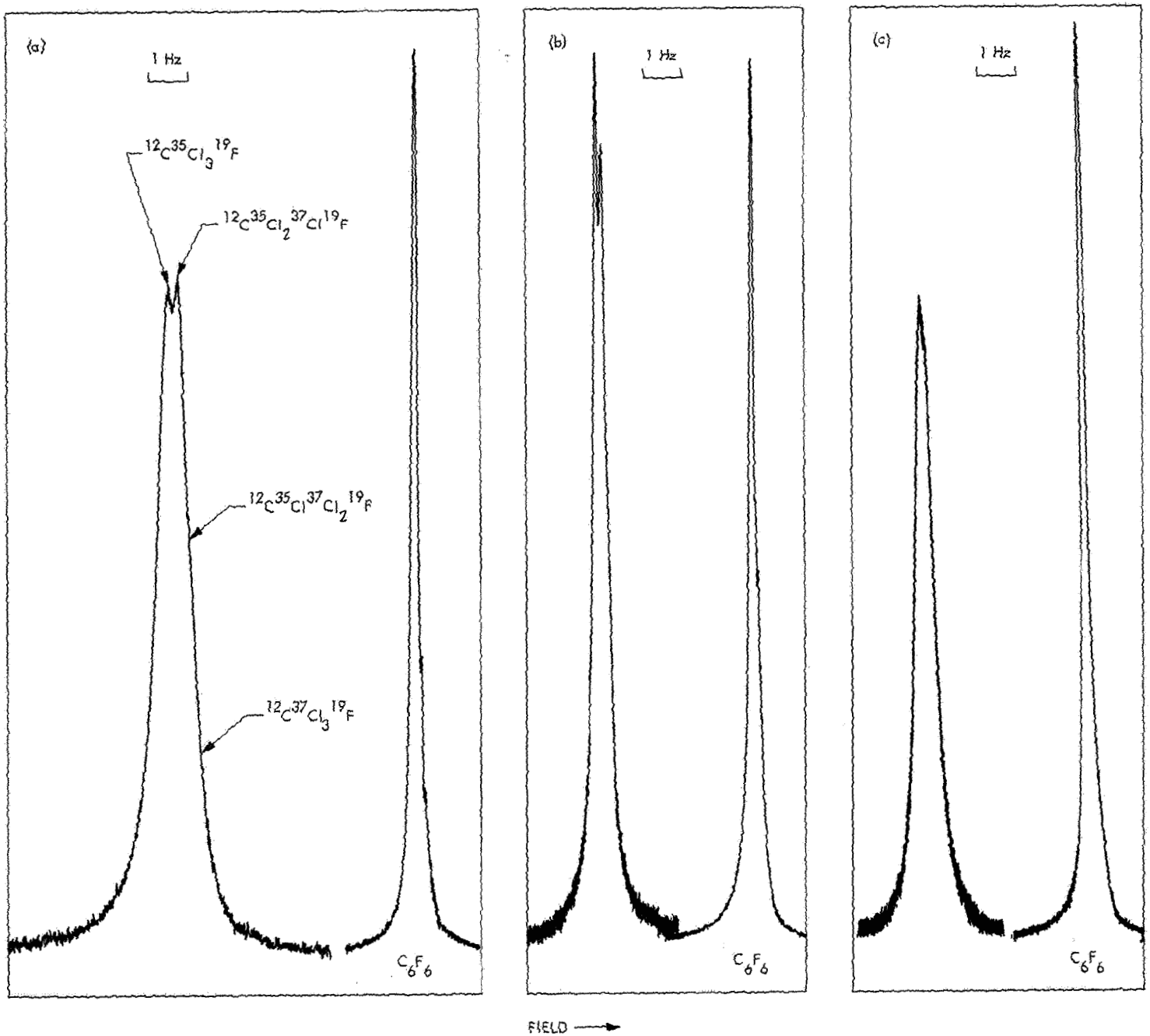


Fig. 1. Fluorine-19 *spectra*: (a) fluorotrichloromethane, (b) *cis*-1,2-difluorodichloroethylene, (c) *trans*-1,2-difluorodichloroethylene

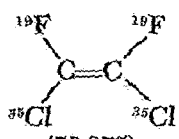
points of the derivative-like signal (from which the control signal to the magnet is derived) resulted in rapid jumping between the two possible stable control points.

The spectrum in Fig. 1a can be interpreted as exhibiting an isotope shift of 0.30 Hz at 56.4 MHz between the species $^{12}C^{35}Cl_3{}^{19}F$ and $^{12}C^{35}Cl_2{}^{37}Cl^{19}F$. Besides these two isotopic species, which should be present in 42.87% and 41.95%, respectively, the molecules $^{12}C^{35}Cl^{37}Cl_2{}^{19}F$ and $^{12}C^{37}Cl_3{}^{19}F$ are present to the extents of 13.69% and 1.49%, respectively. The low field peak of CCl_3F is interpreted as due to the $^{12}C^{35}Cl_3{}^{19}F$ species and the high field peak as due to the $^{12}C^{35}Cl_2{}^{37}Cl^{19}F$ species. The larger apparent intensity of the latter peak stems from the presence of the weaker unresolved peaks of $^{12}C^{35}Cl^{37}Cl_2{}^{19}F$ and $^{12}C^{37}Cl_3{}^{19}F$ species under the high field skirt of the $^{12}C^{35}Cl_2{}^{37}Cl^{19}F$ line.

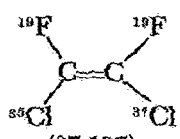
The difference of linewidths between the CCl_3F species and C_6F_6 most probably arises because chlorine-35 and chlorine-37 have nuclear spins greater than ½. Such nuclei have electric quadrupole moments whose fast relaxation usually effectively decouples spin $>½$ nuclei from spin $= ½$ nuclei when the electric-field gradients are large in the vicinity of the former nuclei (Ref. 5). The broadness of the CCl_3F signals suggests that, in this molecule, some residual spin–spin interaction exists between the chlorine nuclei and the fluorine because of relatively small electric-field gradients at the chlorine nuclei.

3. *cis*-1,2-Difluorodichloroethylene Molecule

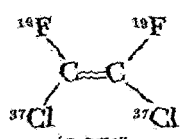
In Fig. 1b, the fluorine-19 spectrum of *cis*-1,2-difluorodichloroethylene is shown. In this case, the following isotopic species are present in the abundances indicated:



(56.85%)



(37.18%)



(6.05%)

The spectrum of this compound is interpreted as exhibiting a low field peak of relative intensity 0.57 and a higher field peak of intensity 0.37, with a peak intensity of 0.06 up-field and not resolved from the latter. In this molecule, the fluorines on a carbon atom with a chlorine-35 will all have the same chemical shift because of the effect of a chlorine-37 atom three bonds away (in the $^{19}F^{35}Cl^{12}C{=}{}^{12}C^{19}F^{37}Cl$ molecules) will be negligible.

For this compound one might expect to see two peaks with intensities of 0.754 and 0.246. However, the situation is slightly more complicated because, in the molecule $^{19}F^{35}Cl^{12}C{=}{}^{12}C^{37}Cl^{19}F$, the two fluorines have slightly different chemical shifts and the nuclear spin–spin coupling of 37.5 Hz between the two fluorines gives rise to a spectrum characteristic of a highly coupled AB spin system (Ref. 6). The separation between the two resolved peaks in Fig. 1b is 0.15 Hz. A calculation of an AB system with a coupling constant of 37.5 Hz and chemical shift of about 0.3 Hz, which was observed between $^{12}C^{35}Cl_3{}^{19}F$ and $^{12}C^{35}Cl_2{}^{37}Cl^{19}F$, indicates that the two center lines of such an AB spin system would only be 0.0024 Hz apart or, for all practical purposes, a singlet halfway between the chemical shift positions of the two nuclei. The observed 0.15-Hz separation then corresponds to one-half the isotope shift between the $^{19}F{-}{}^{12}C{-}{}^{35}Cl$ and $^{19}F{-}{}^{12}C{-}{}^{37}Cl$ fragments present in the species $^{19}F^{35}Cl^{12}C{=}{}^{12}C^{35}Cl^{19}F$ and $^{19}F^{35}Cl^{12}C{=}{}^{12}C^{37}Cl^{19}F$.

The spectral linewidths for *cis*-1,2-difluorodichloroethylene are significantly narrower than those for CCl_3F. This suggests that the local electric-field gradients must be significantly greater than in the former; thus, the chlorine isotopes' nuclear spins are more effectively decoupled from the fluorine nuclear spins by fast nuclear quadrupole relaxation.

4. *trans*-1,2-Difluorodichloroethylene Molecule

In the fluorine-19 spectrum of *trans*-1,2-difluorodichloroethylene shown in Fig. 1c, two things are evident: (1) The linewidths are significantly broader than those in the *cis*-isomer; and (2) the lineshape is indicative of unresolved multiplet structure.

One would expect that the isotope shift in this compound is similar in magnitude and nature to that observed in the *cis*-compound. The observed spectrum results from a single low field line from the all chlorine-35 species, a single line from a strongly coupled AB system ($J_{AA} \approx -129.6$ Hz) from the chlorine-35–chlorine-37 species one half the isotope shift up-field from the first line,

and a much weaker highest field line from the all chlorine-37 species. A complete lineshape analysis and studies at a higher magnetic field should confirm this. The broader linewidths in the *trans*-compound indicate that the electric field gradients experienced by the chlorine nuclei are significantly less (and much smaller) than those existing in the *cis*-compound.

5. Concluding Remarks

For the first time, the effect of chlorine isotopic substitution on fluorine-19 chemical shifts has been observed in three molecules. Also, from the fluorine linewidths in these three compounds, the relative magnitudes of electric-field gradients at the chlorine nuclei could be inferred. Because of its significant linewidth and apparent doublet structure, the unsatisfactory nature of CCl_3F as an internal reference signal for obtaining a nuclear stabilization signal for controlling NMR spectrometers and for referencing fluorine-19 NMR chemical shifts has been pointed out.

References

1. Batiz-Hernandez, H., and Bernheim, R. A., "The Isotope Shift," *Progress in Nuclear Magnetic Resonance Spectroscopy*, Vol. 3, p. 63. Edited by J. W. Emsley, J. Feeney, and L. H. Sutcliffe. Pergamon Press, London, 1967.
2. Filipovich, G., and Tiers, G. V. D., "Fluorine N.S.R. Spectroscopy. I. Reliable Shielding Values, ϕ, By Use of CCl_3F as Solvent and Internal Reference," *J. Phys. Chem.*, Vol. 63, No. 761, 1959.
3. Pople, J. A., Schneider, W. G., and Bernstein, H. J., *High-Resolution Nuclear Magnetic Resonance*, p. 78. McGraw-Hill Book Company, Inc., New York, 1959.
4. Elleman, D. D., Manatt, S. L., and Pearce, C. D., "Relative Signs of the Nuclear Spin Coupling Constants in Propylene Oxide and Indene Oxide," *J. Chem. Phys.*, Vol. 42, No. 650, 1965.
5. Abragam, A., *The Principles of Nuclear Magnetism*, Chap. VII, VIII, and XI. Oxford University Press, London.
6. Corio, P. L., "The Analysis of Nuclear Magnetic Resonance Spectra," *Chem. Rev.*, Vol. 60, No. 363, 1960.

C. An Energy-Level Iterative NMR Method for Sets of Magnetically Nonequivalent, Chemical Shift Equivalent Nuclei, *S. L. Manatt, M. T. Bowers, and T. I. Chapman*

1. Introduction

Nuclear magnetic resonance (NMR) spectroscopy has developed into one of the most useful tools for the chemist's study of both dynamic properties, such as relaxation processes and conformational changes, and steady-state properties, such as stereochemistry and molecular structure. To extract the maximum information from the NMR spectrum of a system, it is usually necessary to accurately analyze complex NMR multiplet patterns. The eigenvalues of the high-resolution spin Hamiltonian

$$\mathbf{H} = \sum_{i=1}^{n} h_i I_{zi} + \sum_{i<j}^{n} J_{ij} \mathbf{I}_i \cdot \mathbf{I}_j \qquad (1)$$

must be obtained, thus yielding the chemical shifts h_i and the spin–spin coupling constants J_{ij}. In Eq. (1), $\mathbf{I}_i$ is the total spin angular momentum of nucleus i and I_{zi} is the z-component of $\mathbf{I}_i$. The important parameters for the chemist that are obtained from NMR spectral analyses are h_i and J_{ij}. These parameters can be related directly to molecular configuration, molecular wave functions, intramolecular fields, and intermolecular fields.

The summations in Eq. (1) extend over all n nuclei and all unique pairs of nuclei in the first and second terms. If the number of nuclei is small, certain analytical solutions exist for the secular equations resulting from Eq. (1). From these closed expressions, the values of the h_i and J_{ij} may usually be obtained from a straightforward analysis of the experimental spectrum. The easily obtained analytical solutions are discussed in detail in Refs. 1–3.

As is most often the case, the experimental spectrum may result from a large number ($\geqq 4$) of strongly coupled nuclei, and in this situation a simple analytical solution is not possible. The first attempts at analyzing these complex spectra utilized the trial-and-error method. Trial h_i and J_{ij} were guessed, and the resulting Hamiltonian matrix was diagonalized to yield an approximate set of energy levels and transition frequencies. This trial-and-error approach is laborious and inefficient and may result in spectral parameters of questionable accuracy.

When high-capacity, high-speed electronic computers became available, several investigators wrote iterative-type computer programs to reduce both the inaccuracy and the labor of retrieval of NMR parameters extracted from complex spectra. The programs most widely used are primarily the result of the initial work of J. D. Swalen and C. A. Reilly (Refs. 4 and 5) and R. A. Hoffman (Ref. 6). Both types of programs are reviewed here to facilitate understanding of the structure and utility of the method developed in the present work.

2. Swalen and Reilly (SR) Method

In the SR method, an approximate Hamiltonian matrix $\mathbf{H}^0$ is calculated from Eq. (1), using the spin product functions as a basis set and assumed values of the h_i and J_{ij}. The matrix elements are calculated by a straightforward application of the coupling properties of angular momentum (Refs. 1–3). The expectation value of the total spin component in the z-direction for a particular submatrix of energy levels is

$$\mathbf{F}_z = \sum_i^n I_{zi} \tag{2}$$

which commutes with $\mathbf{H}$. The matrix $\mathbf{H}^0$ is block diagonal in $\mathbf{F}_z$. The various $\mathbf{F}_z$ blocks are then diagonalized using

$$\mathbf{S}^{-1}\,\mathbf{H}^0\,\mathbf{S} = \Lambda^0 \tag{3}$$

with the columns of the transformation matrix $\mathbf{S}$ corresponding to the individual eigenvalues of $\mathbf{H}^0$. From the Λ_i^0 in Eq. (3), approximate allowed transition frequencies are obtained ($\Delta\mathbf{F}_z = \pm 1$) by

$$\nu_{ij}^0 = \Lambda_i^0 - \Lambda_j^0 \tag{4}$$

and the corresponding intensities are obtained in the usual manner (Refs. 1–3).

Using the approximate frequencies in Eq. (4), the experimental spectrum is assigned, and the iterative portion of the SR method is then used. From the observed transitions, the equations containing the n experimental energy levels can be written as

$$E_i - E_j = \nu_{ij}, \qquad i = 1, 2, \cdots, n-1, \qquad i < j \tag{5}$$

If all n energy levels are connected by transitions, Eq. (5) can be solved for the E_i using the trace relationship

$$\sum_i^n E_i = 0 \tag{6}$$

Since it is usually possible to overdetermine the E_i from Eqs. (5) and (6), a least-squares method is used to solve for the E_i. The larger the number of transitions that can be assigned, the better the value of the E_i obtained. The program written by Swalen and Reilly for the above energy level analysis is called NMREN.

In certain cases, the rank of the energy level matrix resulting from the analysis of Eq. (5) is less than $n-1$, and conditions in addition to Eq. (6) are needed to remove the resultant singularity. These singularities arise any time the Hamiltonian matrix is partitioned into two or more groups with no connecting off-diagonal elements and no allowed transitions between the groups. This is almost always the case when chemical shift equivalent, magnetically nonequivalent nuclei are present in the molecule to be analyzed.

Swalen and Reilly realize this singularity exists and propose that additional trace relationships similar to Eq. (6) could be included to correct the problem (Refs. 4 and 5). However, different trace relationships are needed for each distinct partitioning of the Hamiltonian matrix, and often the Hamiltonian matrix contains three or four such noncoupled subgroups. In addition, to avoid singularities it is necessary to assign enough transitions to couple all of the energy levels within each subgroup. Often this is very difficult or impossible. Again this lack of energy level coupling would require specific trace relationships for each such case encountered. The presentation of a general method for avoiding such singularities without using any trace relationship other than Eq. (6) is the main purpose of this article. A detailed account of the procedure is given in *Subsection 4.*

After the set of energy levels, termed Λ_{obs}, is obtained from the least-squares analysis of Eqs. (5) and (6), the approximate eigenvectors $\mathbf{S}$ are used in a reverse transformation to obtain improved diagonal elements:

$$H_{ii} = (\mathbf{S}\,\Lambda_{\text{obs}}\,\mathbf{S}^{-1})_{ii} = \sum_k S_{ik}^2\,E_k \tag{7}$$

The quantities

$$(H_{ii}^{(1)})^2 = (H_{ii} - \Lambda_i^0)^2 \tag{8}$$

are minimized, where the Λ_i^0 are the diagonal elements of the approximate Hamiltonian matrix $\mathbf{H}^0$ (derived in Eq. 3). Corrections to the parameters, Δh_i and ΔJ_{ij}, are calculated from the least-squares solution of Eq. (8), and the entire process is repeated. The iterative process is stopped after a set number of iterations or when the solution converges or diverges to a preset limit. Standard error analysis on the energy levels and final parameters is performed in the usual way (Refs. 4 and 5). The program that calculates $\mathbf{H}^0$, diagonalizes $\mathbf{H}^0$, and performs the iterative least-squares solution for the h_i and J_{ij} is called NMRIT. When used only to generate trial spectrum, it is designated NMRIT(0).

An extension of the SR method to include symmetry factoring of the Hamiltonian due to the presence of both chemically and magnetically equivalent nuclei has been carried out by R. C. Ferguson and D. W. Marquardt (Ref. 7). While this extension is very useful in handling large-spin systems containing many equivalent nuclei, it does not allow solution of the problems that arise from the chemically equivalent, magnetically nonequivalent nuclei mentioned above or those cases where not enough spectral lines can be assigned to relate all the energy levels within each symmetry subgroup of energy levels.

3. Hoffman Method

The Hoffman method (Ref. 6) is quite similar to the SR method. Hoffman decomposes the Hamiltonian matrix

$$\mathbf{H} = \mathbf{H}^0 + \mathbf{H}^{(1)} \tag{9}$$

into a zero-order component $\mathbf{H}^0$, which is identical to that in Eq. (3), and a correction term $\mathbf{H}^{(1)}$. Diagonalization of $\mathbf{H}^0$ is performed as in Eq. (3), and the resulting transformation matrix is used to transform $\mathbf{H}^{(1)}$. The observed line frequencies ν_{ij} are related to ν^0_{ij}, calculated from $\mathbf{H}^0$, and $\nu^{(1)}_{ij}$ by

$$\nu^{(1)}_{ij} = \nu_{ij} - \nu^{(0)}_{ij} \tag{10}$$

where

$$\nu_{ij} \cong (\mathbf{S}^{-1}\,\mathbf{H}^{(1)}\,\mathbf{S})_{ii} - (\mathbf{S}^{-1}\,\mathbf{H}^{(1)}\,\mathbf{S})_{jj} \tag{11}$$

From Eq. (11), a set of corrections, Δh_i and ΔJ_{ij}, to the h^0_i and J^0_{ij} can be obtained, and the entire process is repeated until a fit is obtained.

A method formally identical to that of Hoffman has been extensively developed by S. Castellano and A. A. Bothner-By (Ref. 8). With their method, Eq. (11) is solved using a least-squares technique. The structure of this program is such that the problem of chemical shift equivalent, magnetically nonequivalent sets of nuclei is treated by specifying the sets of spectral parameters which should be equally and synchronously varied.[1] This aspect of the method had not been succinctly stated when the present work was initiated.

4. Computational Method

The formalism of the method reported here closely follows that of the SR method given above. However, modifications were made to eliminate the necessity for submitting at least two separate programs for a single attempted fit of the observed spectrum, i.e., NMREN followed by NMRIT. If the energy level matrix is singular, for reasons discussed above, each sub-block of levels must be run separately as if it were a complete set of energy levels. Considerable hand calculation must then be performed to correct for the error in assuming Eq. (6) is valid for each sub-block. For a six-spin system, the hand calculations can require several hours for each run; for larger-spin systems, the time required becomes formidable. As mentioned above, an array of trace relationships could, in theory, be included to cover all possible contingencies. This would considerably complicate NMREN and would still leave NMREN isolated from NMRIT.

In the method described here, once an assignment is made, the experimental transitions are separated into connected groups and each group is treated automatically by NMREN-2. The partitioning into groups is arbitrary, and the initial number of groups has proven to have no effect on the values of the final parameters, in most cases, within experimental error. The only requirement on a group is that all energy levels within it be connected. The program then proceeds as in NMREN-2 and calculates a set of energy levels for each group of transitions.

Although the groups of levels are internally consistent, they are not zeroed properly relative to each other since Eq. (6) was assumed valid for each group. Correction to the proper zero is made as follows: The observed energy levels are stored on magnetic tape, and a set of approximate energy levels Λ^0_i is calculated from assumed parameters and Eq. (3). The observed energy levels are recalled, and

$$\Delta^k_i = E^k_i - \Lambda^{k0}_i \tag{12}$$

is calculated. The indices i and k are the energy level and group index, respectively. The Δ values are then averaged:

$$\langle\Delta^k\rangle = \frac{1}{N}\sum_{i=1}^{N}\Delta^k_i \tag{13}$$

where N is the number of energy levels in group k. The $\langle\Delta^k\rangle$ are closely approximated by

$$\langle\Delta^k\rangle \cong \frac{1}{N}\sum_{i=1}^{N}E^k_i \tag{14}$$

[1]Private communication with S. Castellano of Carnegie-Mellon University, Pittsburg, Pa.

and hence yield the correction factor necessary to compensate for the assumption of the validity of Eq. (6) for each group.

This program, termed NMRENIT, then proceeds to calculate the corrected set of observed energy levels

$$E_i^{k\prime} = E_i^k \pm |\langle \Delta^k \rangle| \tag{15}$$

where the plus sign holds if $\langle \Delta^k \rangle$, calculated from Eq. (13), is negative and the minus sign holds if $\langle \Delta^k \rangle$ is positive. The program then orders the $E_i^{k\prime}$ to correspond to the Λ_i^{k0}, and an iterative procedure identical to NMRIT is followed. After each iteration, the above correction procedure is repeated to adjust for any errors in Eq. (14) and to assure that the $E_i^{k\prime}$ remain properly labeled relative to the Λ_i^{k0}. Limitation of the number of iterations and calculation of the errors in the parameters are identical to those for the NMRIT case.

The operations of the NMRENIT program can be summarized by the following seven steps, using Fig. 2, which is similar to the schematic diagram presented in Ref. 4:

(1) The energy level matrix Λ_{obs}^k for each of the k groups is calculated from the assigned $(\nu_{ij}^k)_{obs}$.

(2) $\mathbf{H}_{calc}$ is calculated from the trial parameters h_i and J_{ij}.

(3) $\mathbf{H}_{calc}$ is diagonalized to give the diagonal matrix Λ_{calc} and the corresponding approximate eigenvector matrix $\mathbf{S}$.

(4) Λ_{obs} is obtained from the Λ_{obs}^k using Λ_{calc} and the averaging and ordering technique discussed above.

(5) Λ_{obs} is reverse-transformed using the approximate eigenvectors $\mathbf{S}$ to yield an approximate observed Hamiltonian $\mathbf{H}_{obs}$.

(6) The diagonal elements of $\mathbf{H}_{obs}$ are used along with Λ_{calc} to calculate changes in the trial h_i and J_{ij} by the least-squares technique indicated above. [Steps (2–6) are repeated until a convergent solution is reached or until a preset number of iterations has been completed.]

(7) The ν_{ij} and the corresponding intensities are computed from the final Λ_{calc} and $\mathbf{S}$. The ν_{ij} are compared with the $(\nu_{ij}^k)_{obs}$ with an output format identical to that of NMRIT.

Thus, two separate programs, NMREN-2 and NMRIT, are no longer necessary when using the SR method

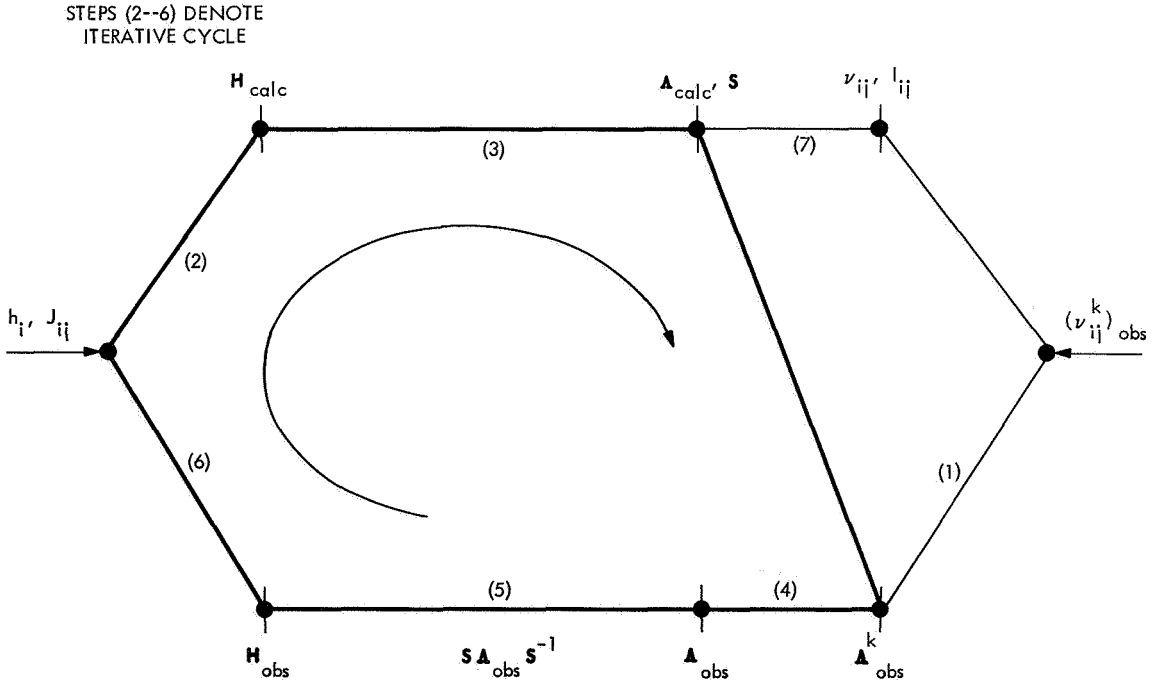


Fig. 2. Operations of the NMRENIT program

(Refs. 4 and 5); instead, programs similar to NMREN-2 and NMRIT are used as subroutines in a master program termed NMRENIT. Hand calculations and data transfer of NMREN-2 results are no longer required. It is no longer necessary to run a separate NMREN program for each subgroup of connected energy levels; instead, a single input carries the problem to an iterative solution. The present program greatly facilitates the energy level iterative analyses of spin systems possessing sets of magnetically nonequivalent, chemical shift equivalent nuclei. The methods that accomplish the latter also allow for the first-time attainment of convergent iterative solutions based on fewer spectral transitions than are necessary to link all the energy levels in a particular symmetry group (provided each energy level is represented at least once). Naturally, the more transitions that can be assigned, the better the value of the E_i^k obtained, but adequate fits have been obtained from a minimum number.

Finally, the methods of the present program, although limited for the moment to eight-spin systems, were conceived allowing for their possible extension to more complex spin systems, perhaps through incorporation of the features of the equivalence factoring approach (Ref. 7) and/or with computers possessing larger storage. The program NMRENIT was fit into an IBM 7094 computer using overlay techniques. The running time of NMRENIT using the JPL IBM 7094 computer system is 2 to 3 min for 10 iterations on a six-spin AA'BB'CC' system.

References

1. Pople, J. A., Schneider, W. G., and Bernstein, H. J., *High Resolution Nuclear Magnetic Resonance.* McGraw-Hill Book Company, Inc., New York, 1959.
2. Emsley, J. W., Feeney, J., and Sutcliffe, L. H., *High Resolution Nuclear Magnetic Resonance Spectroscopy.* Pergamon Press, New York, 1965.
3. Corio, P. L., *Structure of High-Resolution NMR Spectra.* Academic Press, New York, 1966.
4. Swalen, J. D., and Reilly, C. A., "Analysis of Complex NMR Spectra. An Iterative Method," *J. Chem. Phys.*, Vol. 37, No. 21, 1962.
5. Reilly, C. A., and Swalen, J. D., "Nuclear Magnetic Resonance of Some Simple Epoxides," *J. Chem. Phys.*, Vol. 32, No. 1378, 1960.
6. Hoffman, R. A., "Analysis of High-Resolution NMR Spectra by Iterative Methods," *J. Chem. Phys.*, Vol. 33, No. 1256, 1960.
7. Ferguson, R. C., and Marquardt, D. W., "Computer Analysis of NMR Spectra: Magnetic Eigenvalence Factoring," *J. Chem. Phys.*, Vol. 41, No. 2087, 1964.
8. Castellano, S., and Bothner-By, A. A., "Analysis of NMR Spectra by Least Squares," *J. Chem. Phys.*, Vol. 41, No. 3863, 1964.

D. Exterior Forms and General Relativity,

F. B. Estabrook and T. W. J. Unti

When a physical theory is formulated as a coupled set of first-order partial differential equations, the existence of "general solutions" of the set can immediately be discussed using the results of E. Cartan (Ref. 1). When other, so-called "particular" classes of solutions can be found, they too may similarly, but separately, be discussed and classified.

The method consists of recasting the problem to that of finding certain integral surfaces of closed systems of exterior differential forms in a space of n dimensions, where $n - p$ is the number of dependent variables and p the number of independent variables of the original partial differential set. The integral surfaces sought are submanifolds of dimensions p or less; in the p-dimensional integral manifolds, the independent variables may vary freely. The general solutions of the original set are those integral manifolds of dimension p that contain, at every point, integral elements of dimension $p - 1$, which in turn contain integral elements of dimension $p - 2$, etc. The existence criteria for such general solutions have been derived by Cartan through systematic use of the Cauchy–Kowalewski theorem; i.e., these criteria may, in principle, be obtained by integration from suitably set boundary conditions on manifolds of dimension $p - 1$.

Cartan gives a systematic procedure for first calculating a set of "reduced characters," integers $s_0', s_1', \cdots, s_{p-1}'$, and then comparing the integer

$$ps_0' + (p-1)s_1' + \cdots + s_{p-1}'$$

to the number of constraints, h, that relate the differentials of the dependent variables in the p-dimensional integral manifolds. (In the present case, h is just the original number of first-order partial differential equations.) If these are equal, the set is said to be "in involution," and the general solution exists; in a Cauchy integration from conditions on an arbitrary $(p-1)$-dimensional boundary, precisely

$$s_p' \equiv n - p - s_0' - s_1' - \cdots - s_{p-1}'$$

functional relationships (between the independent and dependent variables) may still be arbitrarily imposed to achieve a unique solution.

Here, the property of being "in involution" is regarded as a requirement for a well-set physical theory. As a first example, Maxwell's equations *in vacuo* become two 3-form equations:

$$[dE_x dx\,dt] + [dE_y\,dy\,dt] + [dE_z\,dz\,dt] + [dB_x\,dy\,dz] + [dB_y\,dz\,dx] + [dB_z\,dx\,dy] = 0$$

$$[dB_x\,dx\,dt] + [dB_y\,dy\,dt] + [dB_z\,dz\,dt] - [dE_x\,dy\,dz] - [dE_y\,dz\,dx] - [dE_z\,dx\,dy] = 0$$

with $n = 10$ and $p = 4$. The reduced characters are $s'_0 = 0$, $s'_1 = 0$, $s'_2 = 2$, $s'_3 = 4$, $h = 8$, and the system is in involution. Since $s'_4 = 0$, suitably set boundary conditions on a general bounding 3-dimensional manifold such as $t =$ constant serve uniquely to determine a solution through some range of t. (Characteristic 3-surfaces may also be treated by Cartan's methods.)

The dyadic equations for general relativity (Ref. 2) may be similarly analyzed. They have been found to be expressible as six 2-form equations and six 3-form equations. Only the terms in these involving the differentials of the dependent variables are included here; the terms involving forms that are products of the basis forms of independent variables (denoted by $\bar{\omega}_1$, $\bar{\omega}_2$, $\bar{\omega}_3$, $\bar{\omega}_4$) all have coefficients that are quadratic functions of the dependent variables, and can, when necessary, be read off from the original dyadic formulation. (Knowing already that the set is closed, only the structure of the terms involving differentials of the dependent variables is required for calculation of the reduced characters.) The 2-forms $d\bar{\omega}_1$, $d\bar{\omega}_2$, $d\bar{\omega}_3$, $d\bar{\omega}_4$ can similarly be obtained from the dyadic commutation relationships for $\mathbf{D}$ and differentiations with respect to time:

$$[dN_{1i}\,\bar{\omega}_1] + [dN_{2i}\,\bar{\omega}_2] + [dN_{3i}\,\bar{\omega}_3] - [d\omega_i\,\bar{\omega}_4] = \cdots$$

$$[d\bar{S}_{1i}\,\bar{\omega}_1] + [d\bar{S}_{2i}\,\bar{\omega}_2] + [d\bar{S}_{3i}\,\bar{\omega}_3] + [da_i\,\bar{\omega}_4] = \cdots$$

$$[dB^*_{i1}\,\bar{\omega}_1\,\bar{\omega}_4] + [dB^*_{i2}\,\bar{\omega}_2\,\bar{\omega}_4] + [dB^*_{i3}\,\bar{\omega}_3\,\bar{\omega}_4] + [dP_{1i}\,\bar{\omega}_2\,\bar{\omega}_3] + [dP_{2i}\,\bar{\omega}_3\,\bar{\omega}_1] + [dP_{3i}\,\bar{\omega}_1\,\bar{\omega}_2] = \cdots$$

$$[dQ_{i1}\,\bar{\omega}_1\,\bar{\omega}_4] + [dQ_{i2}\,\bar{\omega}_2\,\bar{\omega}_4] + [Q_{i3}\,\bar{\omega}_3\,\bar{\omega}_4] - [dB^*_{1i}\,\bar{\omega}_2\,\bar{\omega}_3] - [dB^*_{2i}\,\bar{\omega}_3\,\bar{\omega}_1] - [dB^*_{3i}\,\bar{\omega}_1\,\bar{\omega}_2] = \cdots$$

where

$$\bar{\mathbf{S}} = \mathbf{S} - \Omega \times \mathbf{I}$$

$$i = 1, 2, 3$$

with $n = 48$ and $p = 4$, $s'_0 = 0$, $s'_1 = 6$, $s'_2 = 12$, $s'_3 = 14$, and $h = 8$. The system is in involution; $s'_4 = 12$ arbitrary functions may be chosen to determine a unique general solution.

Particular solutions are especially important in nonlinear problems, since they may provide simpler sub-cases for study. For these, adding a number of further algebraic requirements does not reduce s'_4 by an equal number. For example, adding to the above the requirements for vacuum, which are 10 additional constraints ($\mathbf{T} = 0$, $\mathbf{t} = 0$, $\rho = 0$), gives a new involutory set for which $s'_4 = 6$. Adding further the nine requirements for a gaussian reference frame ($\mathbf{a} = 0$, $\Omega = 0$, $\omega = 0$) gives again an involutory set, for which $s'_4 = 0$. This is the discovery upon which so-called canonical formulations of general relativity are based.

This method has been used to analyze the dyadic equations for a case where the following are imposed: (1) vacuum, (2) axial symmetry, and (3) the compatible assumptions that $\dot{\mathbf{N}} + \mathbf{S}^* \cdot \mathbf{N} = 0$, $\mathbf{E} = 0$, and $\Omega = 0$. Since *no* stationary or static condition has been assumed, there are three independent variables described by forms $\bar{\omega}_1$, $\bar{\omega}_2$, $\bar{\omega}_3$.

In fact, with

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta & \delta \\ 0 & \delta & \gamma \end{pmatrix}$$

$$\mathbf{N} = \begin{pmatrix} 0 & 0 & 0 \\ -\phi & 0 & 0 \\ \chi & 0 & 0 \end{pmatrix}$$

$$\mathbf{a} = (0, r, s)$$

$$\omega = (p, 0, 0)$$

the following set of forms (or the equivalent set of partial differential equations) is shown, by lengthy calculation, to be *closed* (i.e., to have no further integrability

conditions):

$$\beta\gamma - \delta^2 = 0$$

$$[d\phi\,\bar{\omega}_1] - [d\chi\,\bar{\omega}_2] + (\phi^2 + \chi^2)\,[\bar{\omega}_1\,\bar{\omega}_2] + (-\beta\phi + \delta\chi - p\chi)\,[\bar{\omega}_1\,\bar{\omega}_3] + (\gamma\chi - \delta\phi - p\phi)\,[\bar{\omega}_2\,\bar{\omega}_3] = 0$$

$$[d\beta\,\bar{\omega}_1] + [d\delta\,\bar{\omega}_2] + [dr\,\bar{\omega}_3] + (\beta\phi - \gamma\phi - 2\delta\chi)\,[\bar{\omega}_1\,\bar{\omega}_2] + (-s\phi + r^2 + 2p\delta - \theta\beta)\,[\bar{\omega}_1\,\bar{\omega}_3] + (s\chi + rs - p\beta + p\gamma - \theta\delta)\,[\bar{\omega}_2\,\bar{\omega}_3] = 0$$

$$[d\delta\,\bar{\omega}_1] + [d\gamma\,\bar{\omega}_2] + [ds\,\bar{\omega}_3] + (-\gamma\chi + \beta\chi + 2\delta\phi)\,[\bar{\omega}_1\,\bar{\omega}_2] + (r\phi + rs + p\gamma - p\beta - \theta\delta)\,[\bar{\omega}_1\,\bar{\omega}_3] + (-r\chi + s^2 - 2p\delta - \theta\gamma)\,[\bar{\omega}_2\,\bar{\omega}_3] = 0$$

$$[dp\,\bar{\omega}_3] + (rp + r\delta - \beta s)\,[\bar{\omega}_1\,\bar{\omega}_3] + (sp - s\delta + \gamma r)\,[\bar{\omega}_2\,\bar{\omega}_3] = 0$$

where $\theta = \beta + \gamma$. There is one 0-form equation, three 1-form equations derived from it (not included here), and four 2-form equations. With $n = 11$ and $p = 3$, $s_0' = 1$, $s_1' = 4$, $s_2' = 3$, $h = 14$, the system is in involution, and $s_3' = 0$.

The forms spanning the space of independent variables fulfill

$$d\bar{\omega}_1 = -\phi\,[\bar{\omega}_1\,\bar{\omega}_2] + (p + \delta)\,[\bar{\omega}_2\,\bar{\omega}_3] - \beta\,[\bar{\omega}_3\,\bar{\omega}_1]$$

$$d\bar{\omega}_2 = \chi\,[\bar{\omega}_1\,\bar{\omega}_2] + \gamma\,[\bar{\omega}_2\,\bar{\omega}_3] + (p - \delta)\,[\bar{\omega}_3\,\bar{\omega}_1]$$

$$d\bar{\omega}_3 = -s\,[\bar{\omega}_2\,\bar{\omega}_3] + r\,[\bar{\omega}_3\,\bar{\omega}_1]$$

These are, in form language, the commutation relationships of the usual dyadic operators, ∇ and $(\cdot)$, applied to scalars; $\bar{\omega}_1$ is radial, $\bar{\omega}_2$ expresses a co-latitude direction, and $\bar{\omega}_3$ is timelike. The unique solution is, in principle, now a matter of straightforward integration. Whether it can be found in closed form, however, is not yet known. The hope is that any such axial, non-static, vacuum solution can be of importance in the study and understanding of gravitational radiation. The assumptions of the present case would make the physical interpretation of any solutions especially easy, inasmuch as there is always present a one-parameter family of imbedded flat, Euclidean 3-spaces.

References

1. Cartan, É., *Les Systèmes Différentiels Extérieurs et Leurs Applications Géométriques*. Hermann et Cie., Paris, 1945.
2. Estabrook, F. B., and Wahlquist, H. D., Dyadic Analysis of Space-Time Congruences, *J. Math. Phys.*, Vol. 5, pp. 1629–1644, 1964.

XIV. Communications Elements Research

TELECOMMUNICATIONS DIVISION

A. Spacecraft Antenna Research: High-Power 400-MHz Coaxial Cavity Radiators, *K. Woo*

1. Introduction

In SPS 37-48, Vol. III, pp. 238–240, and SPS 37-51, Vol. III, pp. 307–309, the design of a 400-MHz coaxial cavity radiator was described. The power-handling capability of the antenna at very low pressures was found to be 76 W in dry air and 62 W in 100% CO_2. In this article, the design of two new models of the antenna, capable of handling much higher power, is presented. One of the models being described can handle as high as 251 W in dry air, 213 W in 100% CO_2, and 186 W in the mixture of 50% CO_2 and 50% argon.

2. Antenna Design

The two new models are: (1) a flared-aperture model, a modification of the original antenna by flaring its aperture (Figs. 1a and 2a) and (2) a wide-cavity model, an antenna having a cavity width approximately twice that of the original antenna (Figs. 1b and 2b). The flared-aperture model is intended to demonstrate that the power-handling capability of a coaxial cavity radiator can be improved upon reducing the high field at and near the cavity aperture by flaring the aperture. The wide-cavity model is intended to demonstrate that the power-handling capability of a coaxial cavity radiator can be greatly improved upon reducing the overall field strength throughout the cavity by widening the cavity.

The cavity of each new model is excited by two orthogonally located metallic feed probes. To prevent breakdown between the cavity walls and the probes, each probe is surrounded completely by teflon insulator. The probes of each antenna are fed with equal power in time quadrature by a 3-dB hybrid of the incoming line, connected to the input terminals of the antenna. When energized, each antenna radiates circularly polarized waves.

3. Test Results

The radiation patterns of the right-hand and left-hand circularly polarized components at 400 MHz of each antenna are shown in Fig. 3. The gain of each antenna is lower than it should be because of the relatively high voltage standing-wave ratio (2.8 for the flared-aperture model, 2.5 for the wide-cavity model) looking into each input terminal of each antenna. When the matching is improved, the gain will be higher.

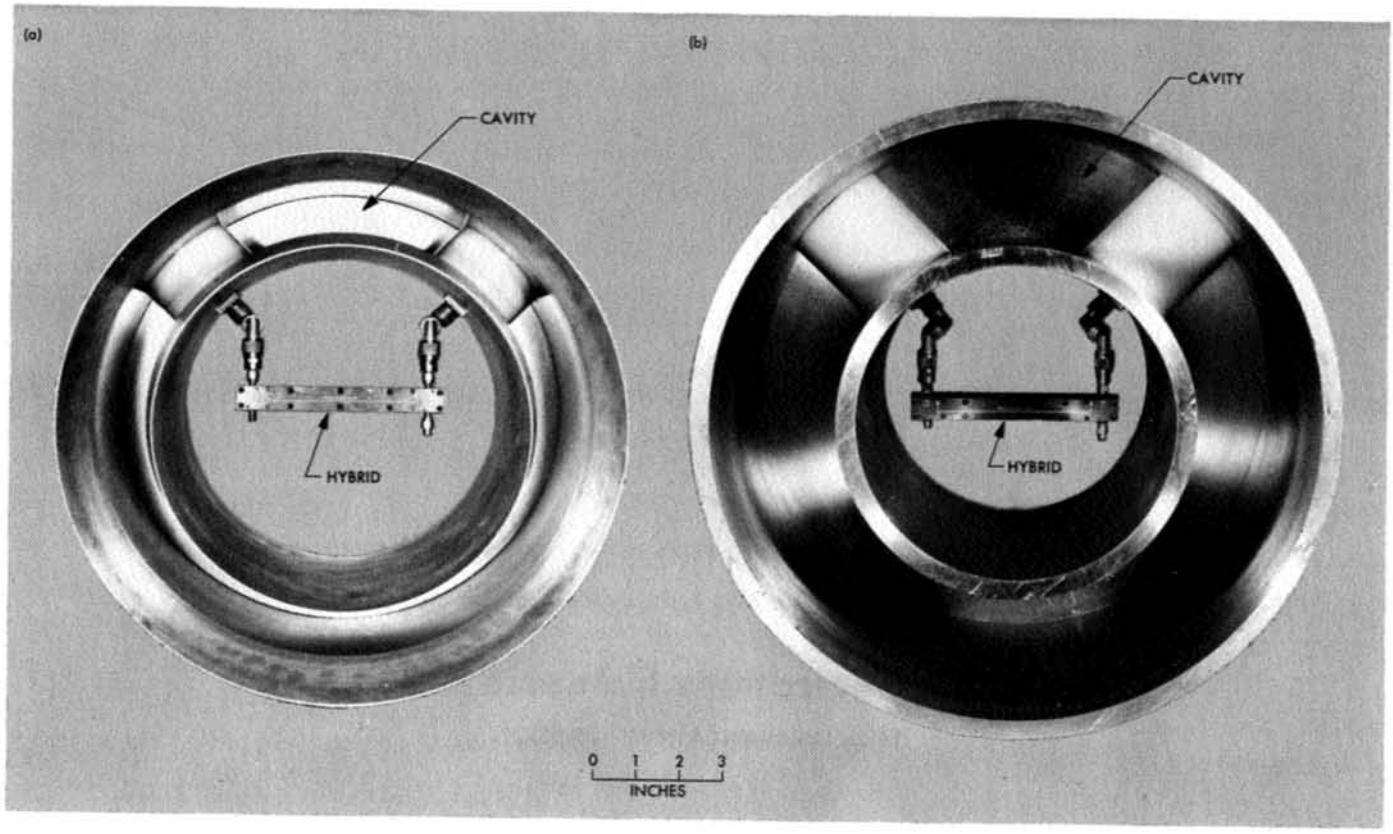
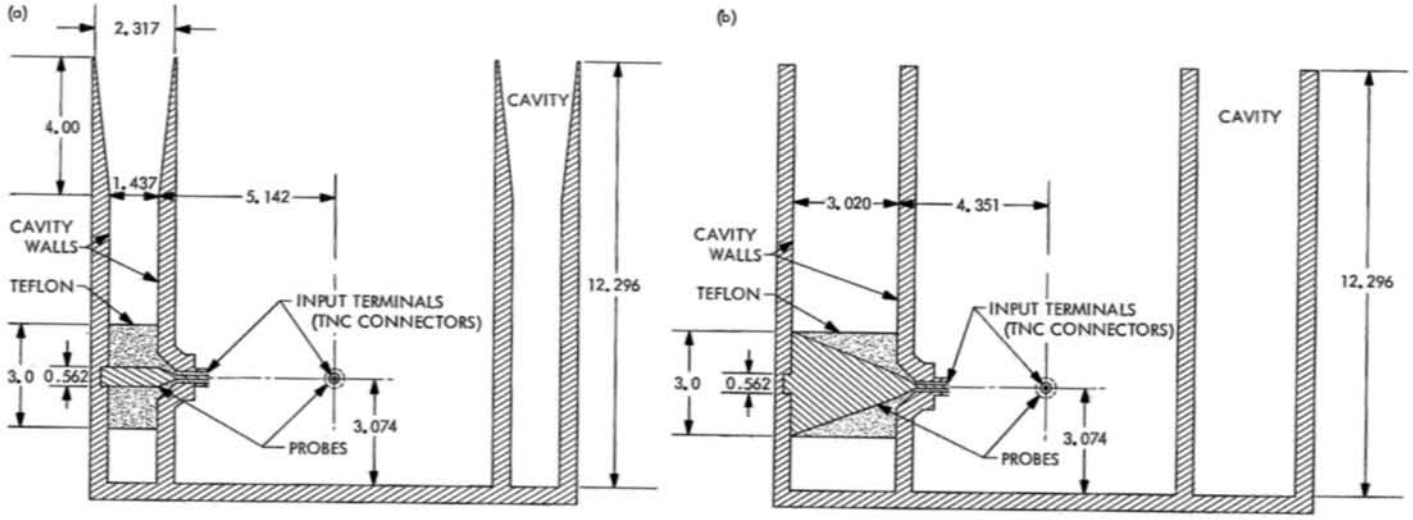


Fig. 1. Antenna model: (a) flared-aperture, (b) wide-cavity



DIMENSIONS IN INCHES

Fig. 2. Cavity and feed configuration: (a) flared-aperture model, (b) wide-cavity model

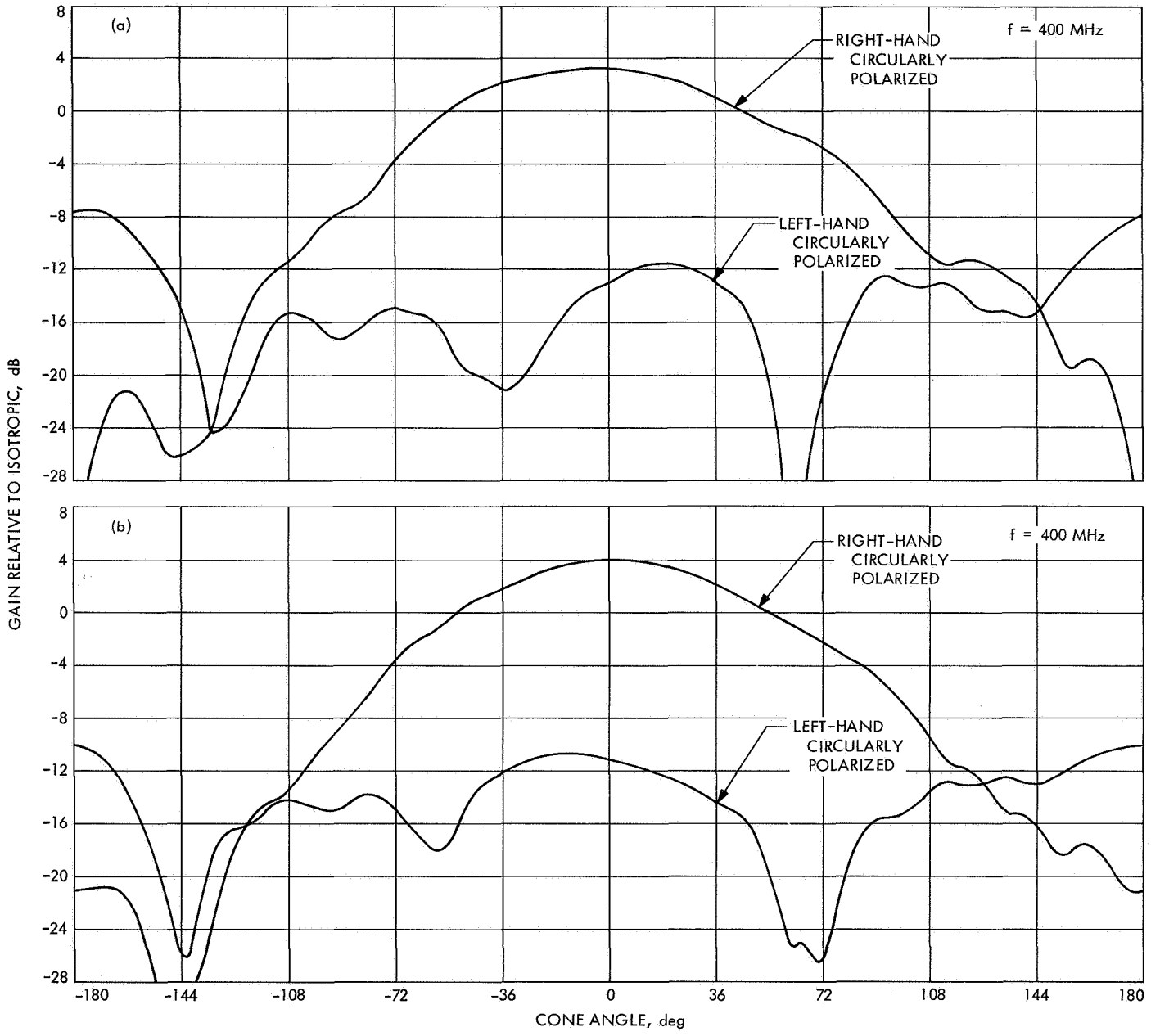


Fig. 3. Radiation patterns: (a) flared-aperture model, (b) wide-cavity model

The power-handling capabilities of the new models at very low pressures were determined in the JPL voltage breakdown facility. The antennas were each tested in the vacuum chamber with dry air, 100% CO_2, and the mixture of 50% CO_2 and 50% argon.

Figure 4 shows the ionization breakdown power level at 400 MHz of each antenna as a function of pressure near and at the point where the power-handling capability of the antenna is a minimum. These levels represent the power that each antenna cavity actually received, i.e., the power fed into the input terminals of each antenna minus the power reflected back to the hybrid as the result of mismatch.

The minimum ionization breakdown power of the flared-aperture model is 118 W (at 0.22 torr) in dry air, 97 W (at 0.28 torr) in 100% CO_2, and 84 W (at 0.33 torr) in the mixture of 50% CO_2 and 50% argon. The minimum ionization breakdown power of the wide-cavity model is 251 W (at 0.20 torr) in dry air, 213 W (at 0.23 torr) in 100% CO_2, and 186 W (at 0.28 torr) in the mixture of 50% CO_2 and 50% argon. In all cases, the breakdown took place around the probes as well as in the middle of the cavity.

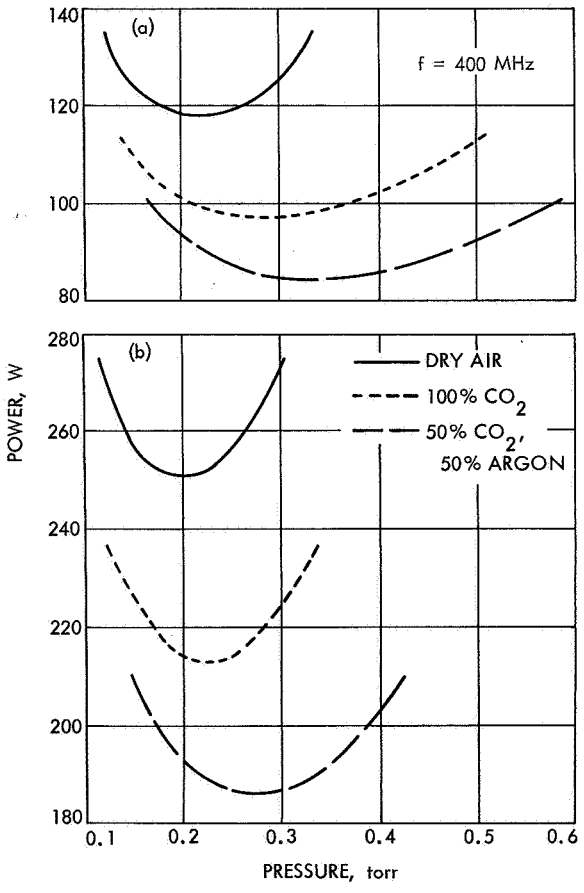


Fig. 4. Ionization breakdown characteristics: (a) flared-aperture model, (b) wide-cavity model

The multipacting breakdown (tested around 8×10^{-5} torr) was not observed at 400 MHz up to a power level of 124 W for the flared-aperture model and 320 W for the wide-cavity model. These values also represent the power that each antenna cavity actually received. The multipacting tests were not carried to a higher power level in each case due to the power limitation of the feeding hybrid used.

4. Conclusion

Based on the test results, it can be said that the power-handling capability of a coaxial cavity type radiator can be improved by flaring its aperture, and can be greatly improved by widening the overall cavity width within the limits of practicality. Further work on the antennas should improve the input voltage standing-wave ratio and the ellipticity.

B. Spacecraft Antenna Research: Sterilizable High-Impact Square-Cup Radiator, Part II, K. Woo

1. Introduction

The sterilizable high-impact square-cup radiator reported in SPS 37-49, Vol. III, pp. 345–347, was potted with Eccofoam PT. This foam has satisfactory thermal, mechanical, and electrical properties except that it is marginal in compressive strength in comparison with that of the balsa-wood impact limiter. Under unfavorable conditions, the foam might not be able to resist the force transmitted by the balsa wood sufficiently to prevent the balsa wood from crushing into the cup during impact. In order to provide a margin of safety, a high-strength foam, Stafoam AA 630,[1] has been under investigation for possible replacement of Eccofoam PT.

2. Test Results

The square-cup radiator potted with Stafoam AA 630 is shown in Fig. 5. The foam has a density of 30 lb/ft^3 and is sterilizable. It has electrical properties similar to those of Eccofoam PT but has a considerably higher compressive strength. A compression test of the foam when potted

[1]Distributed by Olin Chemicals.

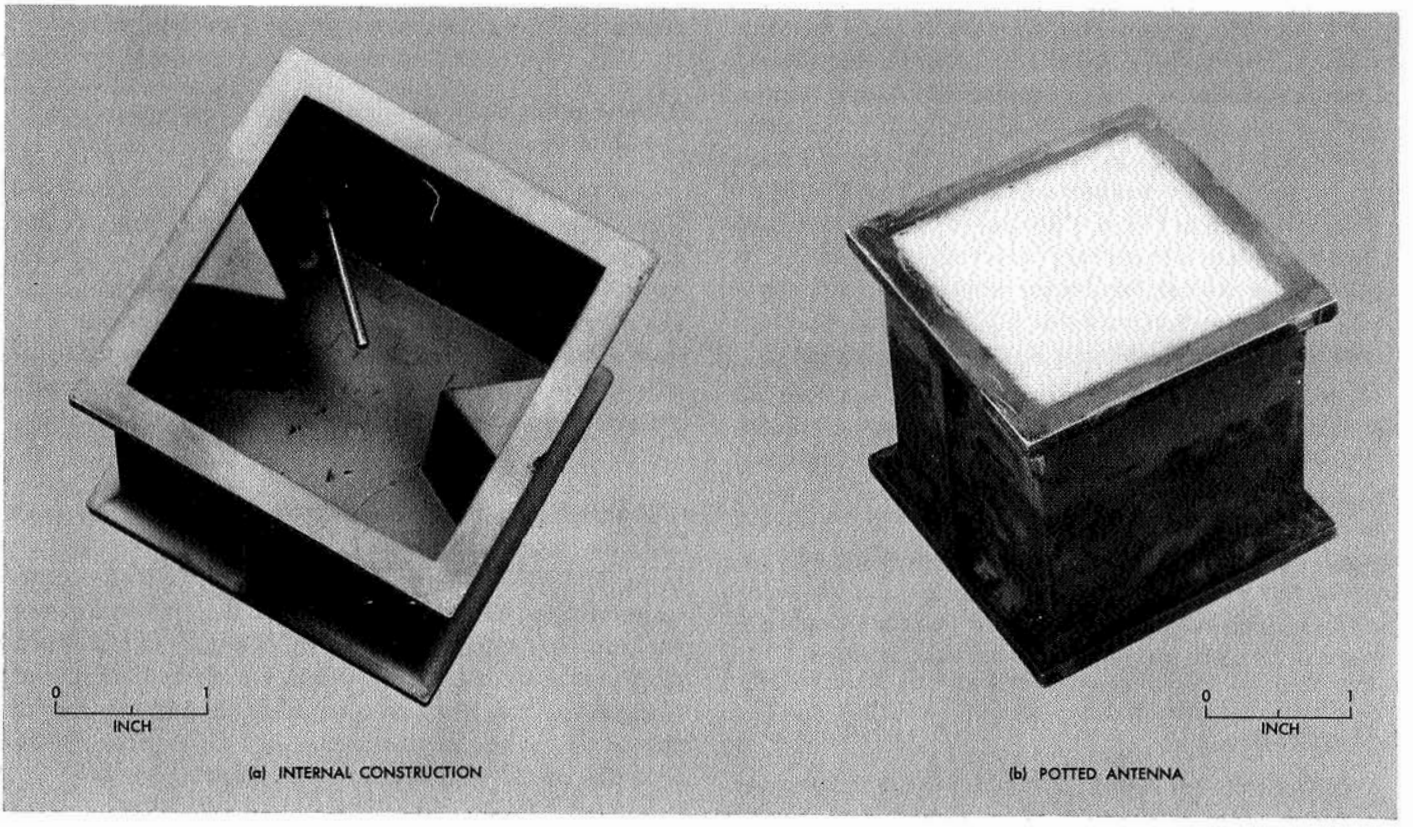


Fig. 5. Sterilizable high-impact square-cup radiator

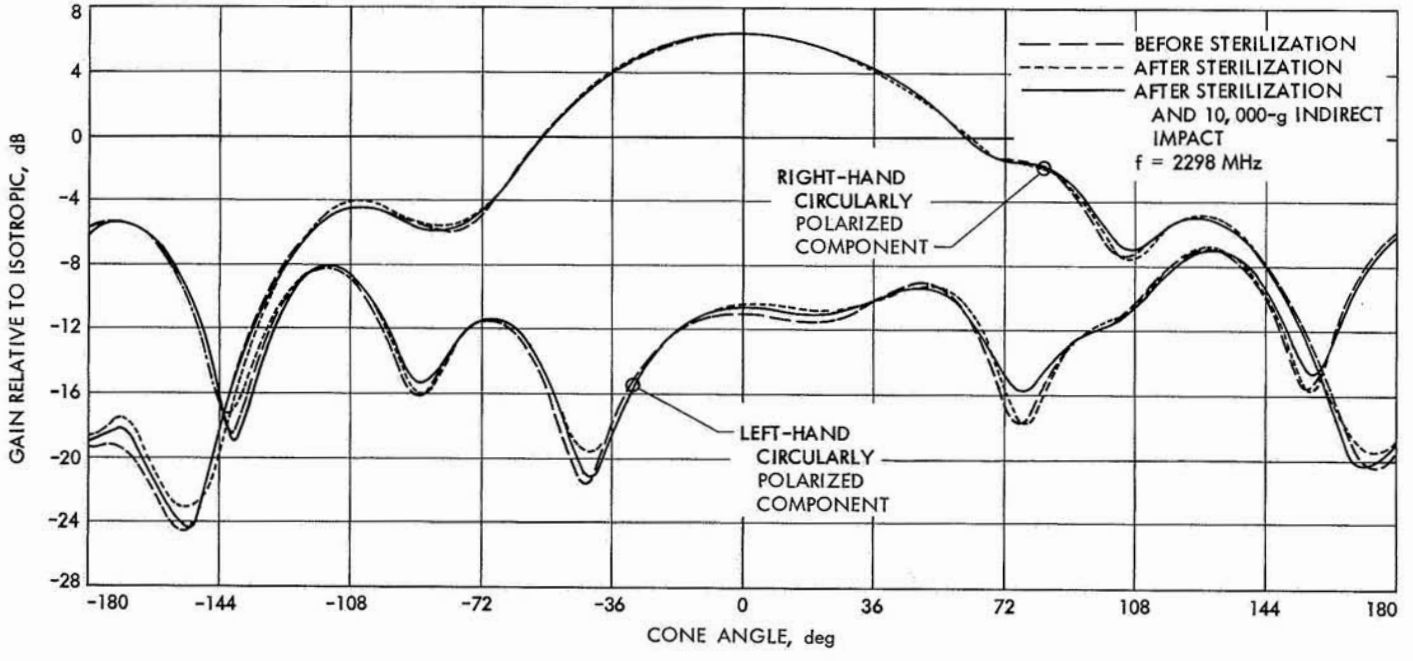


Fig. 6. Radiation patterns

in the square cup shows a compressive strength over 2,000 lb/in.[2] This is more than sufficient to resist the force transmitted by the balsa wood to prevent the balsa-wood impact limiter from crushing into the cup during impact, since the compressive strength of sterilized balsa wood is only around 1,400 lb/in.[2] on the average.[2] The electrical performance of the antenna when potted with Stafoam AA 630 is shown in Fig. 6. The dashed, dotted, and solid curves represent, respectively, the radiation patterns at 2298 MHz of the antenna before sterilization, after sterilization (70-h duration), and after 10,000-g indirect impact.

These radiation patterns show that there has been no significant change in the electrical performance of the antenna as the result of sterilization and impact. The input voltage standing-wave ratios of the antenna before sterilization, after sterilization, and after impact were, respectively, 1.60, 1.50, and 1.48. An examination of the antenna after impact revealed a few small fractures in the foam. However, the fractures were so minor that they would not influence the electrical performance of the antenna.

3. Conclusion

Based on the test results, Stafoam AA 630 is adequate for meeting the requirements of sterilizable high-impact antennas. It should be adapted for general use where high compressive strength is important. The thermal breakdown properties of the foam are presently being investigated.

C. Spacecraft Antenna Research: Large Aperture Antennas, *R. M. Dickinson*

1. Introduction

The object of this study is to determine the RF performance of erectable spacecraft antennas. For large-diameter spacecraft antennas, the resulting narrow antenna beamwidth will make it desirable to have some means of pointing the beam to earth. One method of steering the beam, that could possibly be used to remove small pointing uncertainties (such as attitude-control deadbands), consists of mechanically displacing the antenna feed at an angle to the reflector focal axis. Although the long-term reliability of a constantly moving mechanical device may be questionable, this detail implementation will not be discussed here. In any case, to cause the beam to move with respect to the reflector, the feed phase-center must be moved laterally from the focus point.

2. Mechanical Beam Steering Performance in a Model Erectable Antenna

The model erectable antenna used in the experiment (Fig. 7) consists of a 6-ft-diameter radial parabolic rib antenna of 8 ribs. The ribs have a focal-length-to-diameter ratio (f/D) of 0.35. The gores or the singly curved reflecting material between the ribs are aluminized mylar. The feed consists of a column-supported circular-cupped turnstile for operation at 2297.6 MHz. Feed phase-center displacement was effected by hinging the feed column at the vertex of the reflector.

Figure 8 shows the scanning performance of the feed reflector combination. The upper abscissa is the number of beamwidths scanned. The lower abscissa is the corresponding feed tilt angle in degrees. The top curve shows the scan loss in decibels. The data of Fig. 8 is for the feed displaced along a gore center line. For the feed displaced along a rib line, the performance is, in general, similar except for slightly more scan loss (0.4 dB more at 20 deg feed tilt) but lower sidelobes (-20 dB) at 0 deg tilt.

The scan loss for the effective f/D of 0.333 is greater than existing theory (Ref. 1) predicts. This is thought to be due to the greater phase errors in the reflector surface due to modeling the paraboloid by an erectable surface. Phase errors would be expected to increase more rapidly with scan angle in the erectable antenna.

The second curve from the top in Fig. 8 shows the main lobe beamwidth as a function of scan angle. The beamwidth variation is less (Ref. 1) than normally predicted. The reason for less beamwidth changes with scan angle is at present unknown.

The center curve shows the feed voltage standing-wave ratio change with scan angle. The changes are small. The next curve down shows the sidelobe performance relative to the peak of the beam. The lower curve plots main beam angular position versus feed angular position. The almost constant scope of the curve (0.8) is the beam factor. The 0.8 beam factor agrees with theory for the effective f/D of 0.333.

Reference

1. Kelleher, K. S., and Coleman, H. P., *Off-Axis Characteristics of the Paraboloidal Reflector*, NRL Report 4088. Navy Research Laboratory, Washington, D. C., Dec. 31, 1952.

[2]Sorkin, A. B., *Effects of Sterilization on the Energy Dissipating Properties of Balsa Wood*, Technical Report 32-1295. Jet Propulsion Laboratory, Pasadena, Calif. (to be published).

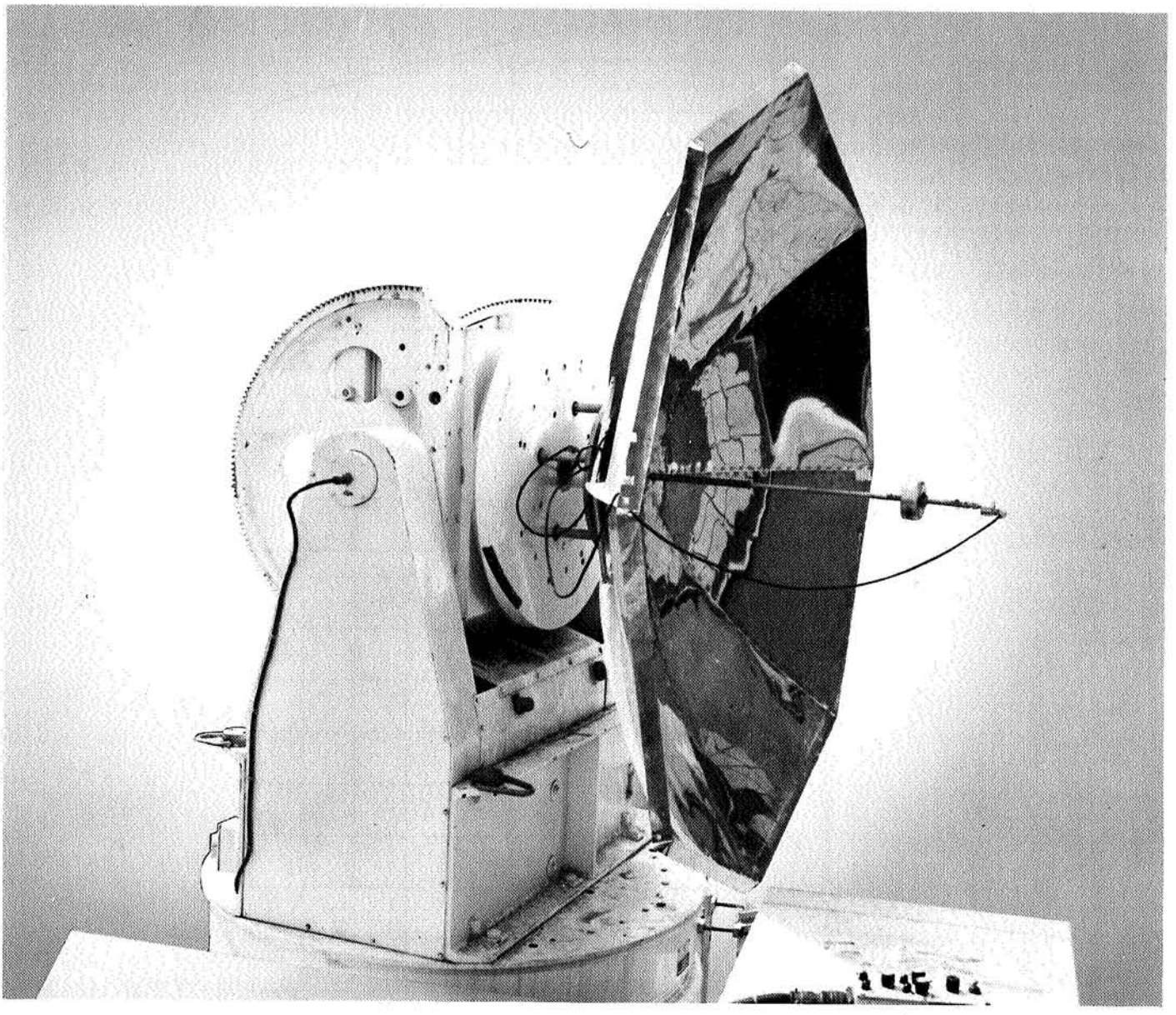


Fig. 7. 6-ft-diameter, 8-rib erectable antenna model

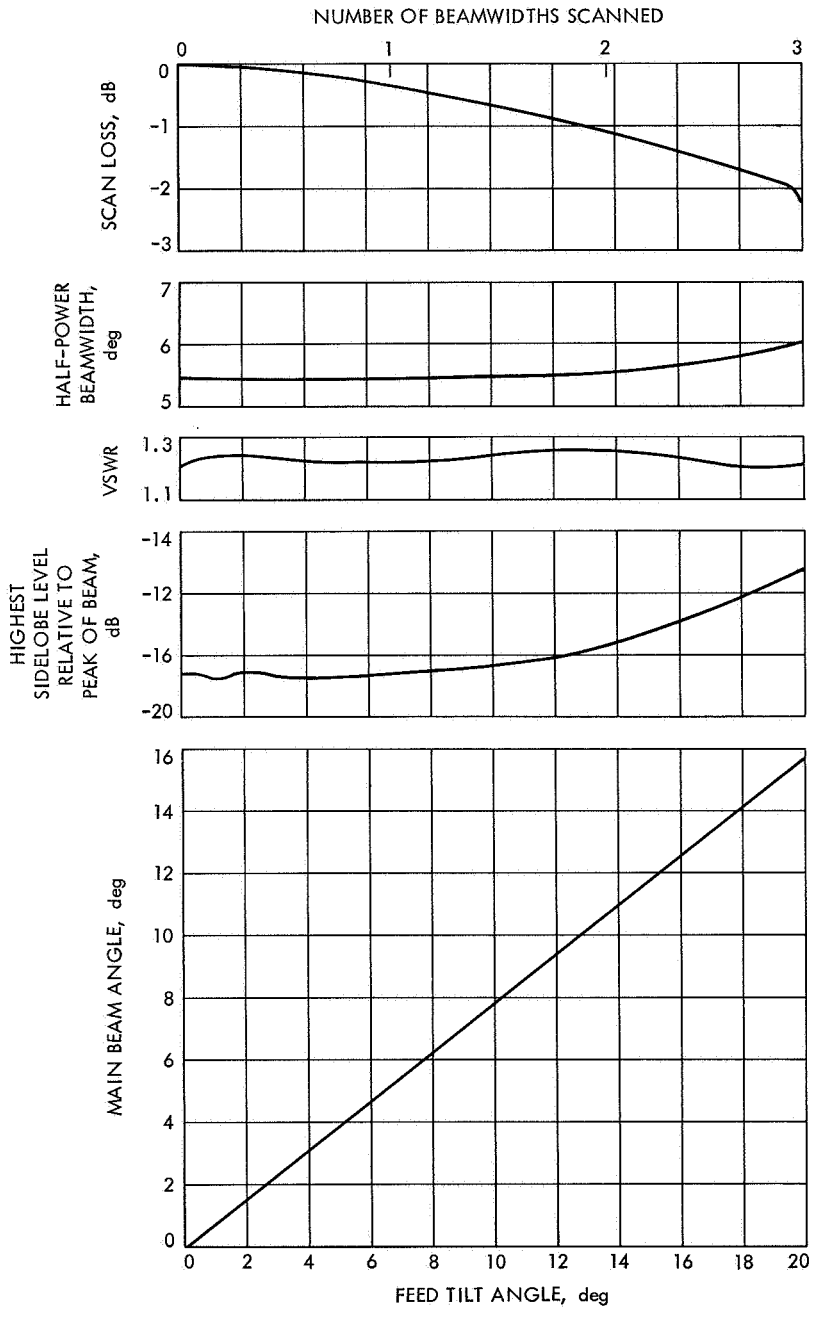


Fig. 8. Erectable reflector scanning performance

XV. Spacecraft Radio

TELECOMMUNICATIONS DIVISION

A. Spacecraft Power Amplifier, *L. J. Derr*

1. Introduction

The electrostatically focused amplifier (ESFA) project is a portion of JPL's advanced development program for S-band (2295 MHz) spaceborne transmitter tubes. The work is being performed by the Klystron Department of EIMAC, Division of Varian Associates, under JPL Contract 951105 (SPS 37-37, Vol. IV, pp. 258–259; SPS 37-48, Vol. III, pp. 278–280).

2. Mechanical and Electrical Design

A fifth developmental tube has been fabricated and tested. This was the first of the experimental ESFAs to use a radiation-cooled collector system. The electrical design of tube 5 was similar in most respects to its water-cooled predecessor, tube 4, which had reached nearly all of the electrical design goals. The mechanical design, however, was changed significantly to accept the radiating cooling system and to conform to the intended final packaging design.

a. Radiating cooling system. The kinetic energy remaining in the spent beam of a klystron is converted to thermal energy in the collector element of the tube. Normally, this unwanted heat is conducted away by an external cooling system that may be either active or passive. The ESFA project supported the development of a collector-cooling system that efficiently radiates this heat directly through the vacuum envelope of the tube into its outside environment, whether atmospheric or vacuum, thus relieving the spacecraft cooling system of this significant thermal load.

The basic design is a collector element that is a thin tungsten egg-shell shape. One end is opened to admit the electron beam that, in turn, heats the tungsten shell to extremely high temperatures. The shell radiates this heat through a sapphire, infrared transparent window with the aid of a tantalum reflector. Prototypes of this collector system were tested earlier in the program and demonstrated radiating efficiencies of 80%.

Tube 5 was the first operating ESFA to employ the radiating collector and successfully demonstrated that this technique is usable. Figure 1 shows tube 5 under undriven conditions where its collector is operating at 1600°C. At this temperature, the collector is directly radiating 214 W of heat. Vendor measurements show that simultaneously 32 W is conducted back to the tube body. Also heating the body is the beam interception current, amounting to 18 W, and the heater power which is 4.3 W.

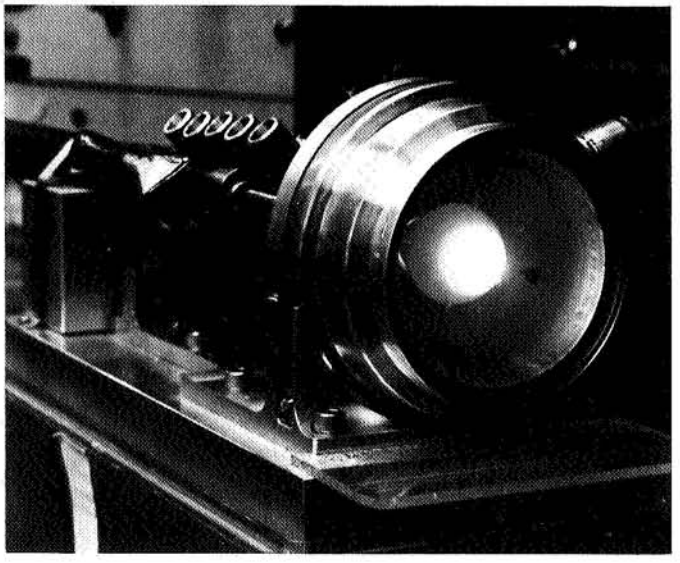


Fig. 1. Tube 5 operating in an RF undriven condition

This brings the total body heat to 54.3 W. Thus, the total tube, in this condition of operation, radiates its own generated heat with an efficiency of 75%.

Under RF driven conditions (100-W RF output), the beam is spread by RF fields and heats the collector more uniformly. The RF power output of the tube subtracts energy from the beam, and the collector temperature lowers to 1300°C. The intercepted body currents increase with the beam spreading, and the radiation efficiency of the collector decreases because of the lower collector temperature. Accurate thermal measurements have not yet been made in this condition of operation.

b. Electrostatic lenses. It is intended that all of the focusing lenses be connected to the cathode potential. In prior experimental tubes, each lens was brought out separately to evaluate its influence on confining the beam. The sizes of the lenses were refined in tube 5 so that it was possible to connect the first 5 (total of 8) directly to the cathode and yet obtain proper focusing fields. No RF feedback through the lens system was observed.

c. Helical resonators. The loaded Q of resonators 2, 3, and 4 was lowered from 270 to 150 to produce a smoother bandpass characteristic of the tube. This was accomplished by plating a thin coating of iron on a portion of each of the helical circuits. This successfully removed the high-Q ripples in the output bandpass of tube 5.

d. Electron gun. Dispenser cathodes have been used in all prior experimental models of the ESFA because of their ability to be reactivated after a tube has temporarily lost its vacuum. These cathodes require 12 W of heater power and do not possess the life capability for the specified 20,000 h. A long-life oxide cathode was designed for the final units, and the first one was installed in tube 5. It provided full beam power and required only 4.3 W of heater power.

3. General Tube Performance

The electron gun in tube 5 was misaligned during its assembly. The beam was thereby directed out of the focusing axis, which resulted in heavy interception on the helical circuits of resonators 5, 6, and 7. The resulting heat that was created evaporated the helix brazing material which redeposited on the surrounding insulator rods. After locating the beam by thermal profile measurements, it was centered in the tube by external magnets. Subsequent RF tests made at the 100-W output level indicated the gain to be 40 dB and the bandwidth at 30 MHz. The damaged resonators, however, lowered the electronic efficiency to 34%, whereas prior tubes had produced 45% at the same operating point.

4. Future Tasks

Tube 5 will be rebuilt, using new helical resonators and a new electron gun. Its output cavity and radiating collector will be used again, since no damage occurred in these elements. Heliarc seals will be used in place of brazed joints in several areas to ensure better alignment of the component parts. Tube 5 is scheduled to finish its refurbishing cycle late in November 1968.

XVI. Communications Systems Research: Sequential Decoding

TELECOMMUNICATIONS DIVISION

A. Performance of *Pioneer*-Type Sequential Decoding Communications Systems With Noisy Oscillators, *J. A. Heller*

1. Introduction

Convolutional coding with sequential decoding will be used as an engineering experiment (Ref. 1) aboard the *Pioneer D* to be launched in August 1968. The information rates to be used are 512, 256, 64 and 16 bits/s. As the spacecraft moves away from the earth, the bit rate will be lowered whenever the overall error probability rises above 10^{-3}. This will continue until the lowest rate (16 bits/s) is used.

The convolutional code that will be used is a rate 1/2 systematic code with constraint length 25. Data will be encoded into blocks of 210 information bits, after which a 14-bit fixed sequence will be inserted in the coder for the purpose of decoder resynchronization. Incoming data at the receiver will be quantized into eight levels prior to inputting it to the decoder.

Pioneer D will operate at a maximum rate of 512 bits/s. A sequential decoder, being designed and built at JPL (SPS 37-50, Vol. II, pp. 71–78), will be available for experimentation on a portion of the *Pioneer D* mission. At rate 1/2, it will have the capability of performing one computation per microsecond and a memory capable of storing about 10^4 branches. This decoder will be so fast, compared to the *Pioneer D* bit rate, that it will be possible to operate at an energy-per-bit $E_b = (E_b)_{\min}$ with a negligible ($<10^{-8}$) erasure probability (Ref. 2).

2. Decoder Error Probability

Theoretically, it has been shown that decoder error probability decreases exponentially with constraint length independent of decoder speed and memory size for rates less than capacity (Ref. 3 and SPS 37-50, Vol. III, pp. 241–248). However, probability of decoder memory overflow or block erasure depends on these two factors and on the (energy-per-bit)-to-noise ratio, E_b/N_0. Since the average decoder computation per bit decoded is unbounded for rates above a certain rate, R_{comp}, sequential decoding is limited to operate at rates less than R_{comp}. That is, there is a minimum E_b defined by

$$(E_b)_{\min} \triangleq \frac{E}{R_{\text{comp}}} \frac{\text{energy/code symbol}}{\text{bits/code symbol}} \tag{1}$$

that must be used in order to induce a stable computational behavior in the decoder. Depending on the decoder

speed and memory size, it is usually necessary to operate with an E_b somewhat above $(E_b)_{\min}$ to meet a specified erasure probability (Ref. 2).

3. Noisy Phase Reference

The phase reference used for demodulating the incoming data from *Pioneer* is derived from a phase-locked loop that tracks the unmodulated carrier component of the signal. The received signal is essentially composed of the carrier with power P_C and the phase-shift-keyed (PSK) data with power P_D. The total power is $P_T = P_C + P_D$ and the modulation index, m^2, is the fraction of power in the carrier ($m^2 = P_D/P_T$). Since the *Pioneer* communication system operates at low rates, P_D/N_0 is small. In order to have sufficient carrier signal-to-noise ratio (SNR) to develop a reliable phase reference, the optimum m^2 is typically measured in tenths rather than hundredths as in high-rate systems. Even at these high modulation indices, the effects of an inaccurate phase reference make the required E_b much larger than that needed in a perfectly coherent system.

In addition to the phase uncertainty caused by tracking a carrier in the presence of thermal noise, the spacecraft oscillator itself is noisy to some degree. Four oscillators with differing degrees of noisiness will be considered. Figure 1 shows the rms phase error incurred when the output of these oscillators is tracked by a phase-locked loop as a function of the bandwidth of the tracking loop. The four oscillator characteristics are generated by letting $n = 1$–4 (Fig. 1).

4. System Performance in the Presence of a Noisy Phase Reference

It has been shown (SPS 37-48, Vol. III, pp. 181–187, and Ref. 4) that for an infinite bandwidth unquantized channel with constant but unknown phase, the use of a decision-directed method of obtaining a phase reference from the information signal itself results in an R_{comp} for sequential decoding given by

$$R_{\text{comp}} = \max_{0<\eta<1} \frac{E}{N_0 \ln 2} \left(\frac{\eta}{1+\eta} - \frac{1}{2P_T \tau / N_0} \ln \frac{1}{1-\eta^2} \right) \tag{2}$$

in bits per code symbol, where P_T is the power in the received signal, τ is the time over which the phase reference is formed, and E/N_0 is the SNR per code symbol. In the case at hand, the phase reference is derived from a separate carrier—not from the data signal. In can be argued that the measurement of phase by means independent of the data will improve the performance slightly; however, Eq. (2) will still be approximately true with P_C in place of P_T; W_L, the loop bandwidth, in place

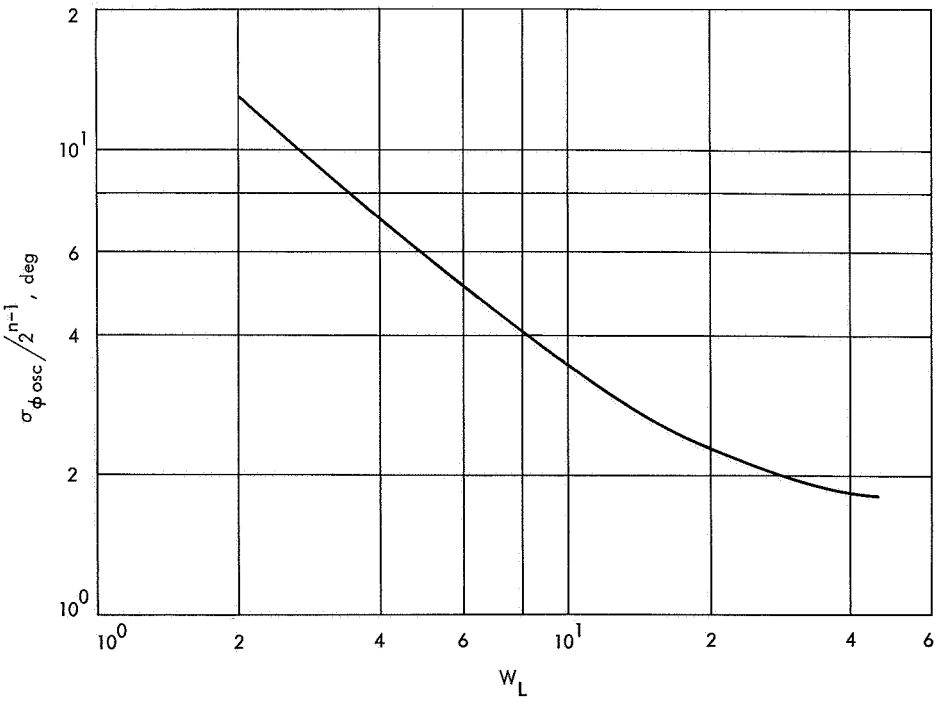


Fig. 1. Rms phase error vs W_L for four oscillators

of $1/\tau$; and E_D, the energy per code symbol in the data signal, in place of E:

$$R_{\text{comp}} \cong \max_{0<\eta<1} \frac{E_D}{N_0 \ln 2}\left(\frac{\eta}{1+\eta} - \frac{1}{2P_C/N_0W_L}\ln\frac{1}{1-\eta^2}\right) \tag{3}$$

As it stands, this result holds only for a constant phase channel and the effects of a noisy oscillator have not yet been taken into account. When a phase-locked loop is operating in its linear region, the phase-error distribution due to thermal noise is approximately gaussian with variance (Ref. 5):

$$\sigma_\phi^2 = \frac{1}{2P_C/N_0W_L} \tag{4}$$

Thus, Eq. (3) can be rewritten

$$R_{\text{comp}} \cong \max_{0<\eta<1} \frac{E_D}{N_0 \ln 2}\left(\frac{\eta}{1+\eta} - \sigma_\phi^2 \ln\frac{1}{1-\eta^2}\right) \tag{4}$$

Now, however, any phase jitter due to a noisy oscillator is independent of jitter due to thermal noise. Hence, the total phase-error variance when the oscillator is noisy will be

$$\sigma_\phi^2 = \frac{1}{2P_C/N_0W_L} + \sigma_{\phi\text{osc}}^2(n, W_L) \tag{5}$$

If we accept the fact that the effect on R_{comp} of a given phase-error variance due to oscillator jitter is the same as that due to a like variance due to thermal noise (a justifiable assumption when σ_ϕ^2 is small), then Eq. (5) may be substituted directly into Eq. (4).

It is now desired to put Eq. (4) into a form that shows its dependence on E_b/N_0; the rate in bits per second, R; and m^2:

$$\frac{2P_C}{N_0W_L} = \frac{2m^2P_T}{N_0W_L} = \frac{2m^2P_TT_bR}{N_0W_L} = 2m^2\frac{E_b}{N_0}\frac{R}{W_L} \tag{6}$$

where $R = 1/T_b$ and $E_b \triangleq P_TT_b/N_0$. Using the fact that $E_D = (1-m^2)E$, and combining Eqs. (4) and (6), Eq. (1) becomes

$$\frac{E_b}{N_0} \cong \min_{0<\eta<1}\left[\frac{1-m^2}{\ln 2}\left(\frac{\eta}{1+\eta} - \sigma_\phi^2\ln\frac{1}{1-\eta^2}\right)\right]^{-1} \tag{7}$$

where σ_ϕ^2 is given by Eqs. (5) and (6). Since σ_ϕ^2 is a function of E_b/N_0, Eq. (7) may be solved explicitly for E_b/N_0:

$$\frac{E_b}{N_0} \cong \min_{0\leq\eta<1}\left(\frac{\dfrac{\ln 2}{1-m^2} + \dfrac{W_L}{2m^2R}\ln\dfrac{1}{1-\eta^2}}{\dfrac{\eta}{1+\eta} - \sigma_{\phi\text{osc}}^2\ln\dfrac{1}{1+\eta^2}}\right) \tag{8}$$

where $\sigma_{\phi\text{osc}}$ depends on n and W_L (Fig. 1).

The expression for R_{comp} in Eq. (2) was obtained for an unquantized, infinite bandwidth channel; thus, Eq. (8) holds only for that channel. It has been shown (Ref. 6) that, in going from a coherent, infinite bandwidth (zero rate), unquantized channel to a rate 1/2, 3-bit quantized channel, an additional 1.25 dB in E_b/N_0 are required. Equation (8) has been numerically minimized with respect to η and m^2. Figure 2 shows the E_b/N_0 resulting from this minimization for W_L of 3, 6, 12 and 24 Hz as a function of R. In each graph, there is one curve for each noisy oscillator ($n = 1$ to 4) and an additional curve for a perfect oscillator ($\sigma_{\phi\text{osc}} = 0$) that indicates the performance when phase jitter is due to thermal noise alone. Figures 2a–2d include the additional 1.25 dB due to rate 1/2 and 3-bit receiver quantization. These figures also include the value of m^2 that optimizes Eq. (8) versus R.

5. Conclusions and Interpretations

The steps leading up to Eq. (8) depended heavily on the assumption that the linear model of the tracking loop was appropriate. For this reason, it is meaningless to extrapolate the results to very low rates where the phase-error variance of Eqs. (5) and (6) is so large as to clearly indicate non-linear loop operation. To reflect this, the curves in Fig. 2 are truncated at that rate at which $\sigma_\phi = 0.5$ rad. This, for instance, completely eliminates the curves for two oscillators in Fig. 2.

At low-rate operation in a *Pioneer*-type system (16 bits/s), the required E_b/N_0 is several dB larger than that required for a perfectly coherent system. The E_b/N_0 needed for a perfectly coherent system can be obtained from Eq. (8) by letting $\sigma_{\phi\text{osc}}^2 \to 0$, $W_L \to 0$, $m \to 0$ and $\eta \to 1$. Thus, for this case, $E_b/N_0 \to 2\ln 2$. Adding the 1.25 dB for rate 1/2 and 3-bit quantization yields

$$\left(\frac{E_b}{N_0}\right)_{\text{coherent}} = 2.65\,\text{dB} \tag{9}$$

The curves for the perfect oscillators in Fig. 2 asymptotically approach this value at high rates.

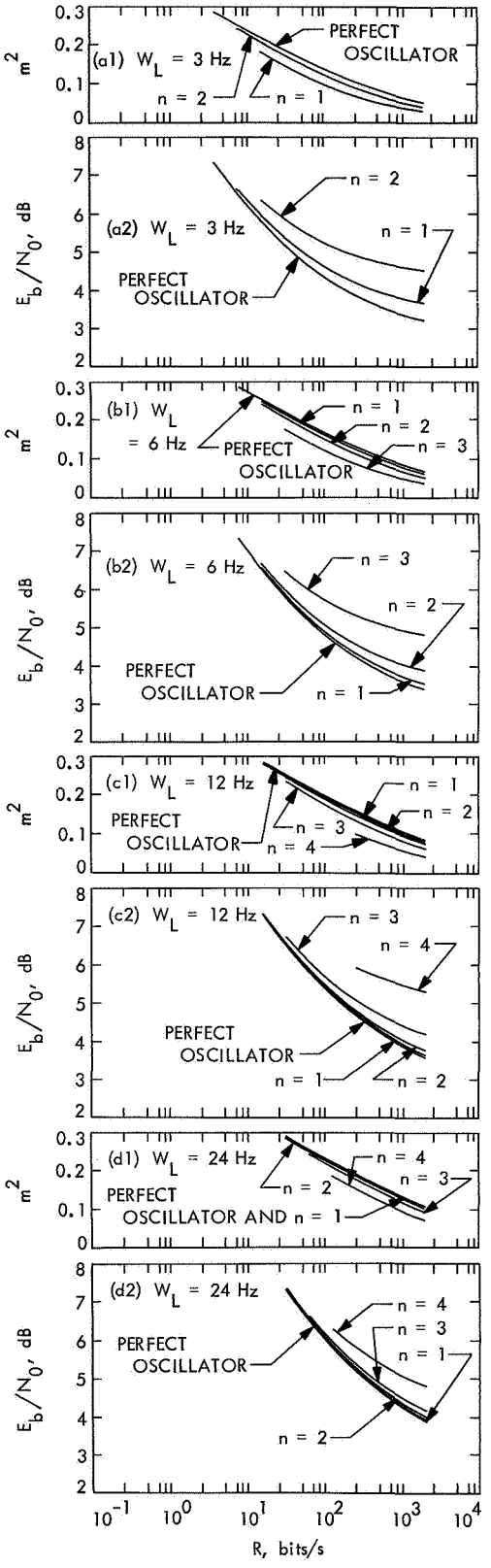


Fig. 2. Minimized E_b/N_0 as a function of R

There is an optimum tracking W_L for any given spacecraft oscillator and rate. This is true because thermal noise considerations dictate the use of the narrowest possible loop while a wide-band loop is best able to track the phase fluctuations of a noisy oscillator. The net result is that at high rates, where the phase-error variance due to thermal noise is small even with a low m^2, noisy-oscillator effects dominate. At low rates, the opposite is true.

The relative looseness of the arguments leading to the E_b/N_0 curves indicates that equipment design should not be based solely on them. It is reasonable to assume, however, that within the linear region of phase-locked loop operation, the curves should be accurate. On this assumption, a comparison with results for uncoded PSK systems (Ref. 6) indicates that, even at rates on the order of 16 bits/s, the sequential decoding system offers about a 2 dB advantage.

References

1. Lumb, D. R., and Hofman, L. B., *An Efficient Coding System for Deep Space Probes with Specific Application to Pioneer Missions,* NASA TN D-4105, National Aeronautics and Space Administration, Washington, August 1967.
2. Heller, J. A., and Lindsey, W. C., "Improved Modulation and Coding Methods for Space Communication Systems," paper presented at the IEEE International Conference on Communications, Philadelphia, Pa., June 12–14, 1968.
3. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Inform. Theory,* IT-13, April 1967.
4. Heller, J. A., *Sequential Decoding for Channels with Time Varying Phase,* Ph.D. Thesis. The Massachusetts Institute of Technology, Cambridge, Mass., Sept. 1967.
5. Lindsey, W. C., "Optimal Design of One-Way and Two-Way Coherent Communication Links," *IEEE Trans. Commun. Technol.,* Vol. 14, No. 4, August 1966.
6. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering,* John Wiley and Sons, New York, 1965.

XVII. Communications Systems Research: Coding and Synchronization Studies

TELECOMMUNICATIONS DIVISION

A. Performance of a Low-Rate Command Data Link, *S. Farber*

1. Introduction

This article gives the performance of an orthogonal signal frequency-shift-keyed command link. This ground-to-spacecraft link will code low-rate binary information into one of two frequency-modulated tones modulated onto the carrier. The purpose of this article is to examine the error performance capabilities of such a scheme under the assumptions (1) that the phase error out of the spacecraft tracking loop does not vary significantly over a bit time and (2) that the phase error out of the spacecraft tracking loop does vary significantly over a bit time.

2. System Model

The transmitted signal is assumed to be of the form

$$s(t) = (2P)^{1/2} \cos [\omega_c t + k \sin (\omega t + \theta) + \Psi], \qquad \omega = \omega_0, \omega_1$$

where ω_c is the carrier frequency, k the index of modulation, and $\omega = \omega_0$ represents a *zero* being transmitted while $\omega = \omega_1$ represents a *one* being transmitted. The angles Ψ and θ are assumed to be uniformly distributed random variables defined on the interval $-\pi, \pi$ radians.

The signal $s(t)$ can be expanded in a Fourier series about ω_c to yield

$$\begin{aligned} s(t) = {} & (2P)^{1/2} J_0(k) \cos (\omega_c t + \Psi) \\ & - (2P)^{1/2} J_1(k) \{\cos [\omega_c t - (\omega t + \theta) + \Psi] - \cos [\omega_c t + (\omega t + \theta) + \Psi]\} \\ & + (2P)^{1/2} J_2(k) \{\cos [\omega_c t - 2(\omega t + \theta) + \Psi] + \cos [\omega_c t + 2(\omega t + \theta) + \Psi]\} \\ & - (2P)^{1/2} J_3(k) \{\cos [\omega_c t - 3(\omega t + \theta) + \Psi] - \cos [\omega_c t + 3(\omega t + \theta) + \Psi]\} \\ & + (2P)^{1/2} J_4(k) \{\cos [\omega_c t - 4(\omega t + \theta) + \Psi] + \cos [\omega_c t + 4(\omega t + \theta) + \Psi]\} \\ & - \cdots \end{aligned}$$

where J_k is the Bessel function of order k. An indication of the behavior of $J_i(k)$, $i = 0, 1, 2$, can be seen in Fig. 1 for values of k satisfying $0 \leq k \leq 2$.

If the tracking loop works on the fundamental component, it will form an estimate $\hat{\Psi}(t)$ of $\Psi(t)$ so that the received signal mixed with $(2)^{1/2} \sin [\omega_c t + \hat{\Psi}(t)]$ and filtered will yield

$$r_1(t) = (P)^{1/2} \sin [k \sin (\omega t + \theta) + \phi(t)] + n_1(t)$$

while the received signal mixed with $(2)^{1/2} \cos [\omega_c t + \hat{\Psi}(t)]$ and filtered will yield

$$r_2(t) = (P)^{1/2} \cos [k \sin (\omega t + \theta) + \phi(t)] + n_2(t)$$

where $\phi(t) = \hat{\Psi}(t) - \Psi$ and $n_1(t)$ and $n_2(t)$ represent independent white gaussian noise of single-sided spectral density N_0 (Ref. 1).

If the tracking loop is a phase-locked loop preceded by a bandpass limiter, then the distribution on ϕ as given by Lindsey (Ref. 2) using DSN parameters is

$$p(\phi) = \frac{\exp(\rho_L \cos \phi)}{2\pi I_0(\rho_L)}, \qquad -\pi < \phi \leq \pi$$

where

$$\rho_L = \frac{3z}{\Gamma\left(1 + \frac{2}{\mu}\right)}, \qquad \Gamma = \frac{1 + 0.345\,zy}{0.862 + 0.690\,zy}$$

$$z = \frac{P_c}{N_0 b_{L0}}, \qquad y = \frac{1}{800}$$

$$\mu = \frac{(\gamma_0)^{1/2} \exp\left(-\frac{\gamma_0 y}{2}\right)\left[I_0\left(\frac{\gamma_0 y}{2}\right) + I_1\left(\frac{\gamma_0 y}{2}\right)\right]}{(z)^{1/2} \exp\left(-\frac{zy}{2}\right)\left[I_0\left(\frac{zy}{2}\right) + I_1\left(\frac{zy}{2}\right)\right]}$$

and $\gamma_0 = 4$. (It should be noted that the usual DSN parameters are $y = 1/400$ and $\gamma_0 = 2$.) I_k is the modified Bessel function of order k. P_c represents the power in the carrier, b_{L0} the loop design bandwidth, and N_0 the single-sided spectral density of the noise. For the above signal, we find $P_c = PJ_0^2(k)$.

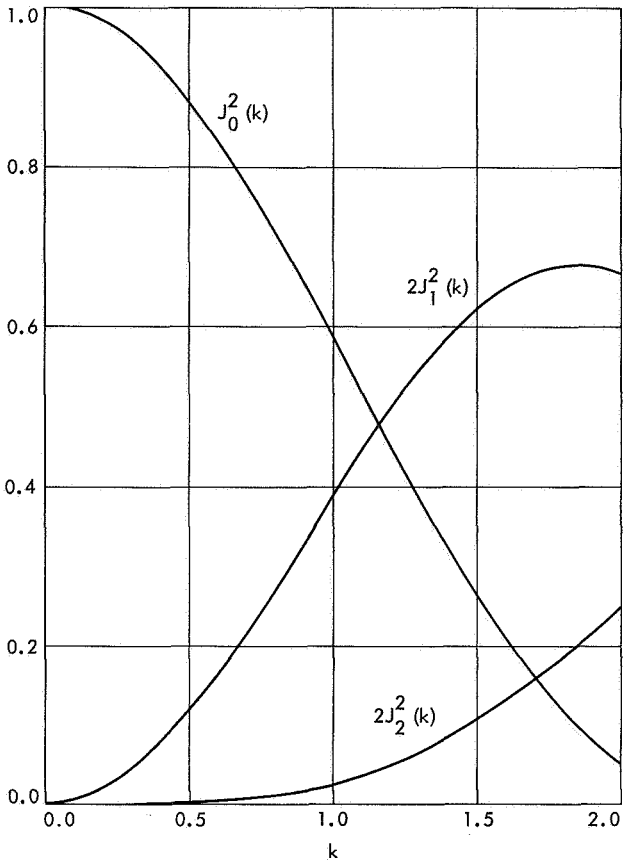


Fig. 1. Plot of J_0^2, $2J_1^2$, and $2J_2^2$, showing division of power between fundamental and other components

3. Error Rates for Various Detectors

For convenience, let us define the random variables

$$k_\phi = \frac{1}{T_b} \int_0^{T_b} \cos \phi(t)\, dt$$

$$\lambda_\phi = \frac{1}{T_b} \int_0^{T_b} \sin \phi(t)\, dt$$

where T_b is the time per bit.

If the data is extracted using only the component of $r_1(t)$ at frequency ω, namely,

$$(P)^{1/2} \cos \phi(t)\, 2J_1(k) \sin (\omega t + \theta) + n_1(t), \qquad 0 < t \leq T_b;\ \omega = \omega_0, \omega_1$$

then the problem is essentially to decide which of two signals is present. Hence, an incoherent phase receiver using orthogonal signals can be used to obtain a bit probability of error (Ref. 3) of

$$P_E^I = E\left\{\frac{1}{2}\exp\left[-\frac{1}{2}k_\phi^2 R\right]\right\}$$

where E is the expectation operation, $R = ST_b/N_0$, and $S = 2J_1^2(k)P$ is the power in the data.

If the data is extracted from the fundamental components of both $r_1(t)$ and $r_2(t)$, namely,

$$(P)^{1/2}\cos\phi(t)\,2J_1(k)\sin(\omega t+\theta)+n_1(t), \qquad 0<t<T_b$$

and

$$(P)^{1/2}\sin\phi(t)\,2J_1(k)\sin(\omega t+\theta)+n_2(t), \qquad \omega=\omega_0,\omega_1$$

then by using the doubly incoherent receiver discussed in *Subsection 8*, it is possible to obtain a probability of error of

$$P_E^D = \min_{0\le\beta\le 1} E\{\tfrac{1}{2}c(\beta)\exp[-\tfrac{1}{2}(k_\phi^2+\lambda_\phi^2)R]\}$$

where

$$c(\beta)=\frac{\exp\left[\frac{1}{2}\left(\frac{1-\beta}{1+\beta}\right)\lambda_\phi^2 R\right]-\beta^2\exp\left[-\frac{1}{2}\left(\frac{1-\beta}{1+\beta}\right)k_\phi^2 R\right]}{1-\beta^2}$$

and β is an arbitrary gain factor, $0\le\beta\le 1$.

We note that when $\beta = 0$, the doubly incoherent receiver degenerates to the incoherent receiver so that we always have $P_E^D \le P_E^I$ for a given index of modulation k. Since, as the index of modulation increases from zero, the amount of power in the data will increase, causing R to increase, while the amount of power in the carrier will decrease, causing ρ_L to decrease, there will be an optimum value of k, corresponding to an optimum division of power. In particular, ρ_L depends on

$$z=\frac{P_c}{N_0 b_{L0}}=\frac{PT_b}{N_0}\cdot\frac{1}{b_{L0}T_b}\cdot J_0^2(k)$$

where

$$R=\frac{ST_b}{N_0}=\frac{PT_b}{N_0}\cdot 2J_1^2(k)$$

By letting $\delta = 1/2b_{L0}T_b$ and $\mathcal{R} = PT_b/N_0$, we can write

$$z = 2\mathcal{R}\,\delta\,J_0^2(k)$$

$$R = 2\mathcal{R}J_1^2(k)$$

[It should be noted that Lindsey (Ref. 2) uses $\delta = 1/b_{L0}T_b$.]

4. Extremely Low Data Rates

When the data rate *is* extremely low, corresponding to $\delta << 1$, it is appropriate (Ref. 1) to use the approximations

$$k_\phi=\frac{1}{T_b}\int_0^{T_b}\cos\phi(t)\,dt\simeq E\{\cos\phi\}$$

and

$$\lambda_\phi=\frac{1}{T_b}\int_0^{T_b}\sin\phi(t)\,dt\simeq E\{\sin\phi\}$$

Using Lindsey's model for the density on ϕ as given above, we find

$$k_\phi\simeq\eta=\frac{I_1(\rho_L)}{I_0(\rho_L)}$$

$$\lambda_\phi\simeq 0$$

The optimum value of β for the doubly incoherent receiver then occurs at $\beta = 0$ so that the doubly incoherent receiver reduces to the incoherent receiver with probability of error given by

$$P_E^{DI} = P_E^I = \tfrac{1}{2}\exp(-\tfrac{1}{2}\eta^2 R)$$

The resulting minimum value of the probability of error is plotted in Fig. 2a versus $\mathcal{R}$ for several values of δ, while the optimum values of k are plotted in Fig. 2b and the resulting values of ρ_L are plotted in Fig. 2c. In order that the tracking loop acquire frequency lock, it is necessary to require that $\rho_L \geqq 6$.

5. Moderate Data Rates

Moderate data rates occur when the phase does not vary significantly over a bit time so that $\delta \cong 1$ and the approximations

$$k_\phi = \frac{1}{T_b}\int_0^{T_b} \cos\phi\,(t)\,dt \cong \cos\phi$$

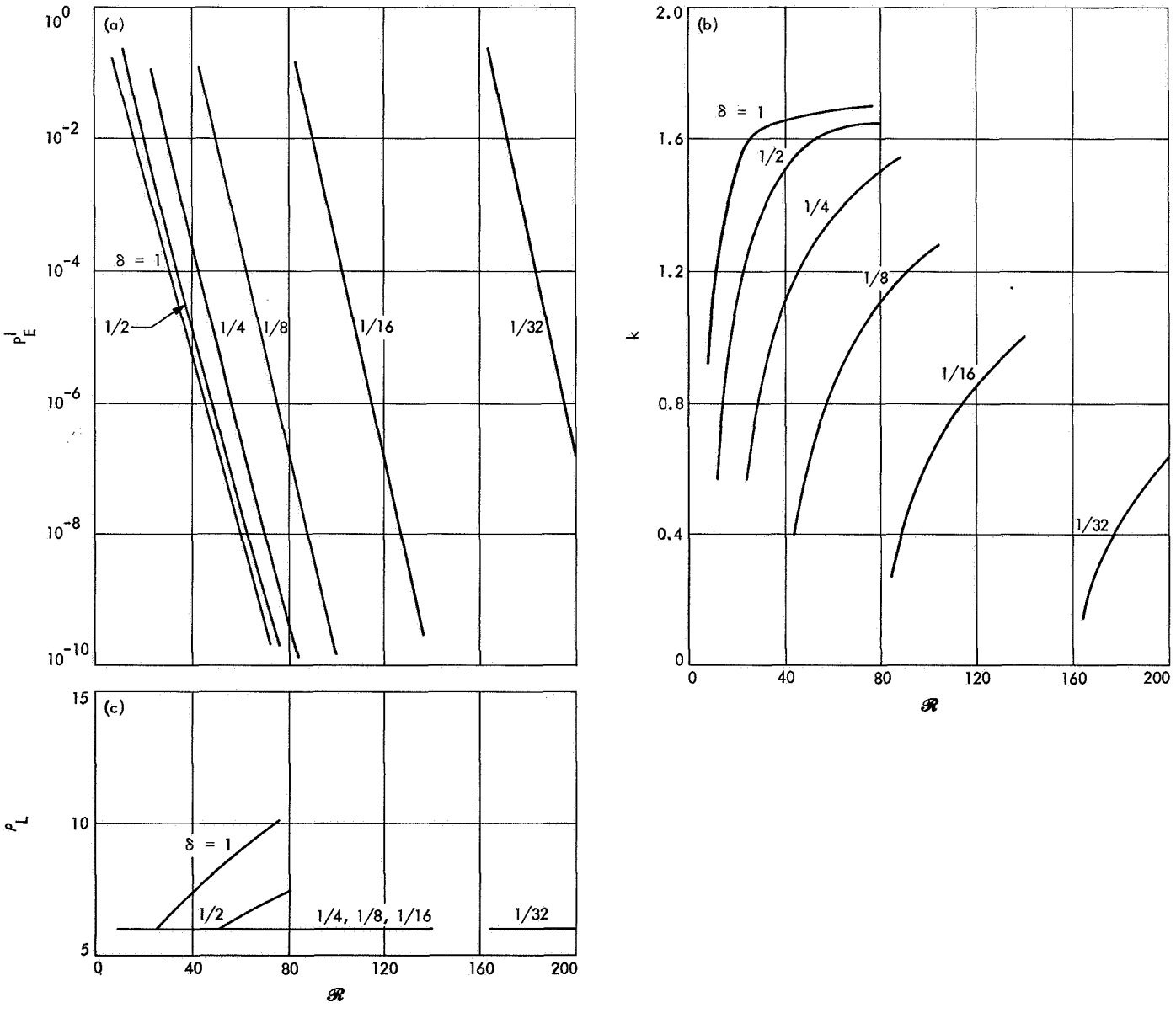


Fig. 2. Plots of behavior of incoherent receiver under the assumption of non-constant phase, showing (a) probability of error, (b) optimal value of modulation index k, and (c) resulting value of ρ_L

and

$$\lambda_\phi = \frac{1}{T_b}\int_0^{T_b} \sin\phi\,(t)\,dt \cong \sin\phi$$

are valid.

Under these circumstances, we find the probability of error for the incoherent receiver is

$$P_E^I = \int_{-\pi}^{\pi} \tfrac{1}{2}\exp\left(-\tfrac{1}{2}R\cos^2\phi\right)\exp\left(\rho_L\cos\phi\right)\frac{d\phi}{2\pi I_0\left(\rho_L\right)}$$

The minimum value of P_E^I is plotted in Fig. 3a, the optimum value of k to give this value of P_E^I is plotted in Fig. 3b, and the resulting value of ρ_L is plotted in Fig. 3c.

The probability of error for the doubly incoherent receiver is

$$P_E^D = \min_{0\le\beta\le 1}\int_{-\pi}^{\pi} c\,(\beta)\frac{d\phi}{2\pi I_0\left(\rho_L\right)}\,\tfrac{1}{2}\exp\left(-\tfrac{1}{2}R\right)$$

where

$$c\,(\beta) = \frac{\exp\left[\frac{1}{2}\left(\frac{1-\beta}{1+\beta}\right)R\sin^2\phi\right] - \beta^2\exp\left[-\frac{1}{2}\left(\frac{1-\beta}{1+\beta}\right)R\cos^2\phi\right]}{1-\beta^2}$$

We note that, when $\beta \to 1$, we can evaluate $c\,(\beta)$ by L'Hospital's rule to find

$$c\,(1) = 1 + \frac{R}{8}$$

which is independent of ϕ. Combining this with the fact that the performance of the doubly incoherent detector cannot be better than the performance of the incoherent detector with $\phi = 0$, we find

$$\frac{1}{2}\exp\left(-\frac{1}{2}R\right) \le P_E^D \le \frac{1}{2}\left(1+\frac{R}{8}\right)\exp\left(-\frac{1}{2}R\right)$$

so that P_E^D must be exponentially asymptotic to the optimum receiver performance for the given signaling scheme.

The minimum value of P_E^D is plotted in Fig. 4a, the optimum value of k to give this value of P_E^D is plotted in Fig. 4b, the resulting value of ρ_L is plotted in Fig. 4c, and the optimum value of β is plotted in Fig. 4d.

6. In Between Rates

For rates between those discussed in *Subsections 3, 4,* and 5, we expect the probabilities of error to fall somewhere in between the probabilities of error obtained above.

This would imply that as the rate increases from extremely low to moderate, the probability of error for the incoherent detector would increase from the values in Fig. 2a to the much larger values in Fig. 3a. The probability of error for the doubly incoherent detector, however, would decrease from the values in Fig. 2a to the slightly lower values in Fig. 4a. The desirability of using a doubly incoherent receiver, which is somewhat more complicated to implement as opposed to the simpler incoherent receiver, would, of course, depend on the exact probability of error for the rate under consideration.

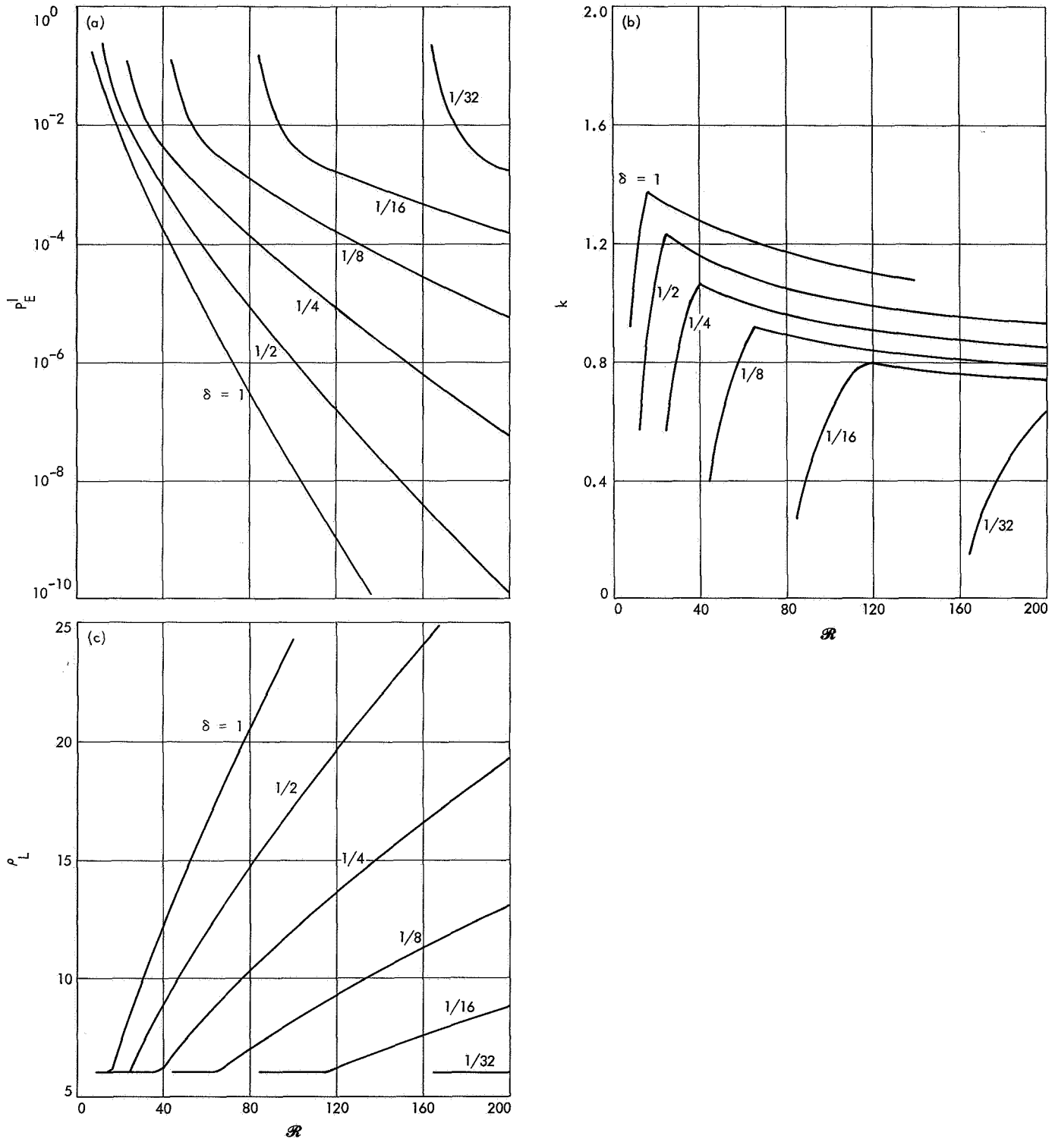


Fig. 3. Plots of behavior of incoherent receiver under the assumption of constant phase, showing (a) probability of error, (b) optimal value of modulation index k, and (c) resulting value of ρ_L

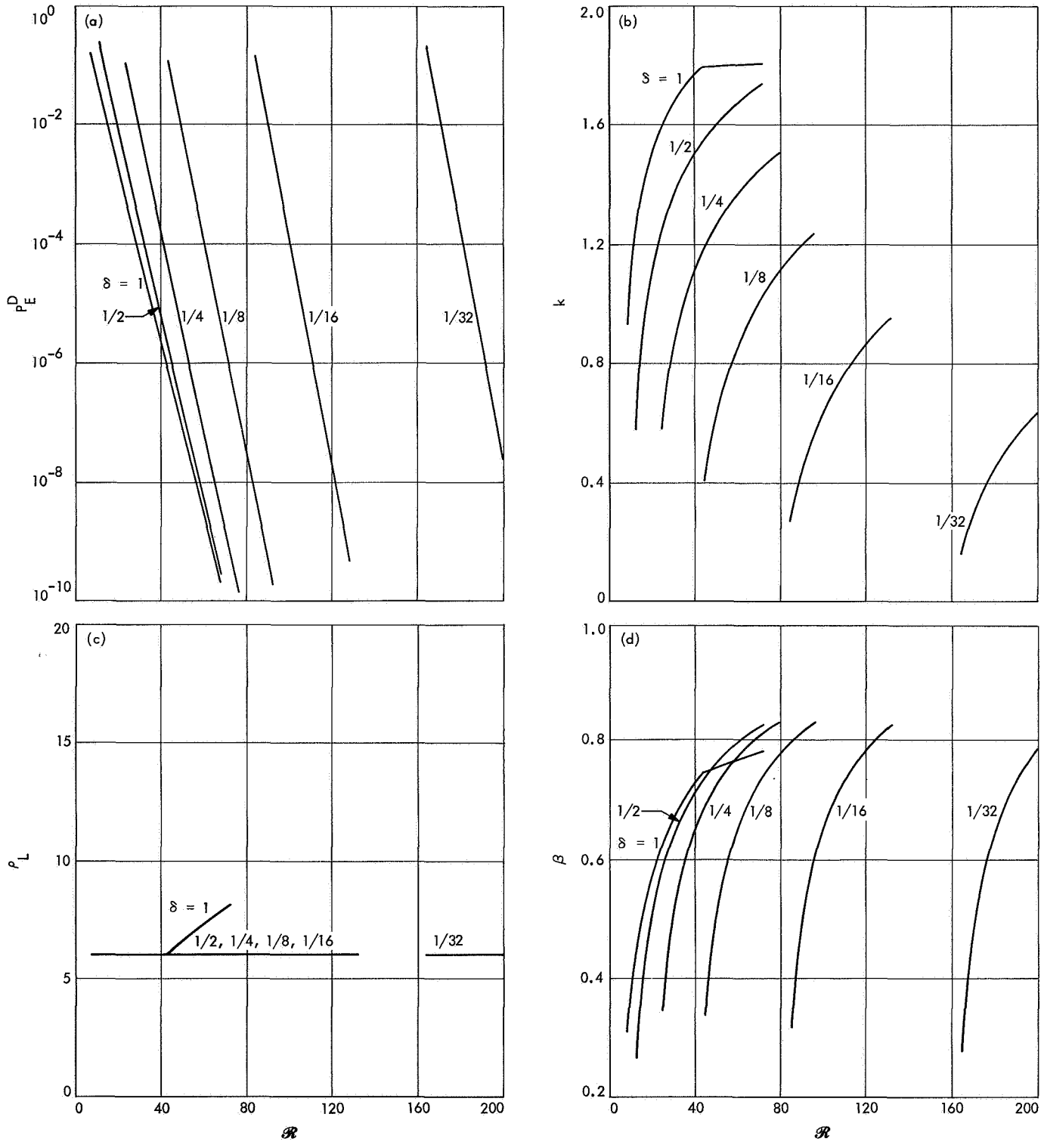


Fig. 4. Plots of behavior of doubly incoherent receiver under the assumption of constant phase, showing (a) probability of error, (b) optimal value of modulation under k, (c) resulting value of ρ_L, and (d) optimal value of β

7. Using the Second Harmonic

The doubly incoherent detector may also be used to extract information about the data from the first harmonic of $r_1(t)$ and the second harmonic of $r_2(t)$; namely,

$$(P)^{1/2} \cos\phi(t)\, 2J_1(k) \sin(\omega t + \theta) + n_1(t), \qquad \omega = \omega_0, \omega_1$$

and

$$(P)^{1/2} \cos\phi(t)\, J_2(k) \cos 2(\omega t + \theta) + n_2(t), \qquad 0 < t \leq T_b$$

This would yield a probability of error of

$$P_E^D = \min_{0 \leq \beta \leq 1} E\{\tfrac{1}{2} c(\beta) \exp[-\tfrac{1}{2} k_\phi^2 (R_1 + R_2)]\}$$

where

$$c(\beta) = \frac{\exp\left[\frac{1}{2}\left(\frac{1-\beta}{1+\beta}\right) k_\phi^2 R_2\right] - \beta^2 \exp\left[-\frac{1}{2}\left(\frac{1-\beta}{1+\beta}\right) k_\phi^2 R_1\right]}{1-\beta^2}$$

and

$$R_1 = 2J_1^2(k)\,\mathscr{R}$$

$$R_2 = 2J_2^2(k)\,\mathscr{R}$$

This yields an improvement in signal-to-noise ratio of about $J_2^2(k)/J_1^2(k)$ over the incoherent receiver. An indication of this ratio can be obtained from Fig. 1. For values of k near 1, the improvement is about 10%.

8. Description and Analysis of the Doubly Incoherent Receiver

We assume two signals of the form

$$r_1(t) = (P)^{1/2} \sin[k \sin(\omega t + \theta_1) + \phi] + n_1(t)$$

and

$$r_2(t) = (P)^{1/2} \cos[k \sin(\omega t + \theta_2) + \phi] + n_2(t), \qquad 0 < t \leq T;\ \omega = \omega_0, \omega_1$$

where the angles θ_1 and θ_2 are arbitrary and $n_1(t)$ and $n_2(t)$ are the independent white gaussian noise process of the one-sided spectral density N_0.

The doubly incoherent receiver then consists of two sections of the form shown in Fig. 5, one with $\omega = \omega_0$ and one with $\omega = \omega_1$. The variable β is an arbitrary gain factor which is to be chosen to minimize the probability of error. If we define the random variables

$$k_\phi = \frac{1}{T}\int_0^T \cos\phi(t)\, dt$$

and

$$\lambda_\phi = \frac{1}{T}\int_0^T \sin\phi(t)\, dt$$

then the output of the ω_0 section when $\sin\omega_0 t$ was transmitted is

$$Q_0 = u_1^2 + u_1^2 + u_2^2 + \beta(u_3^2 + u_4^2)$$

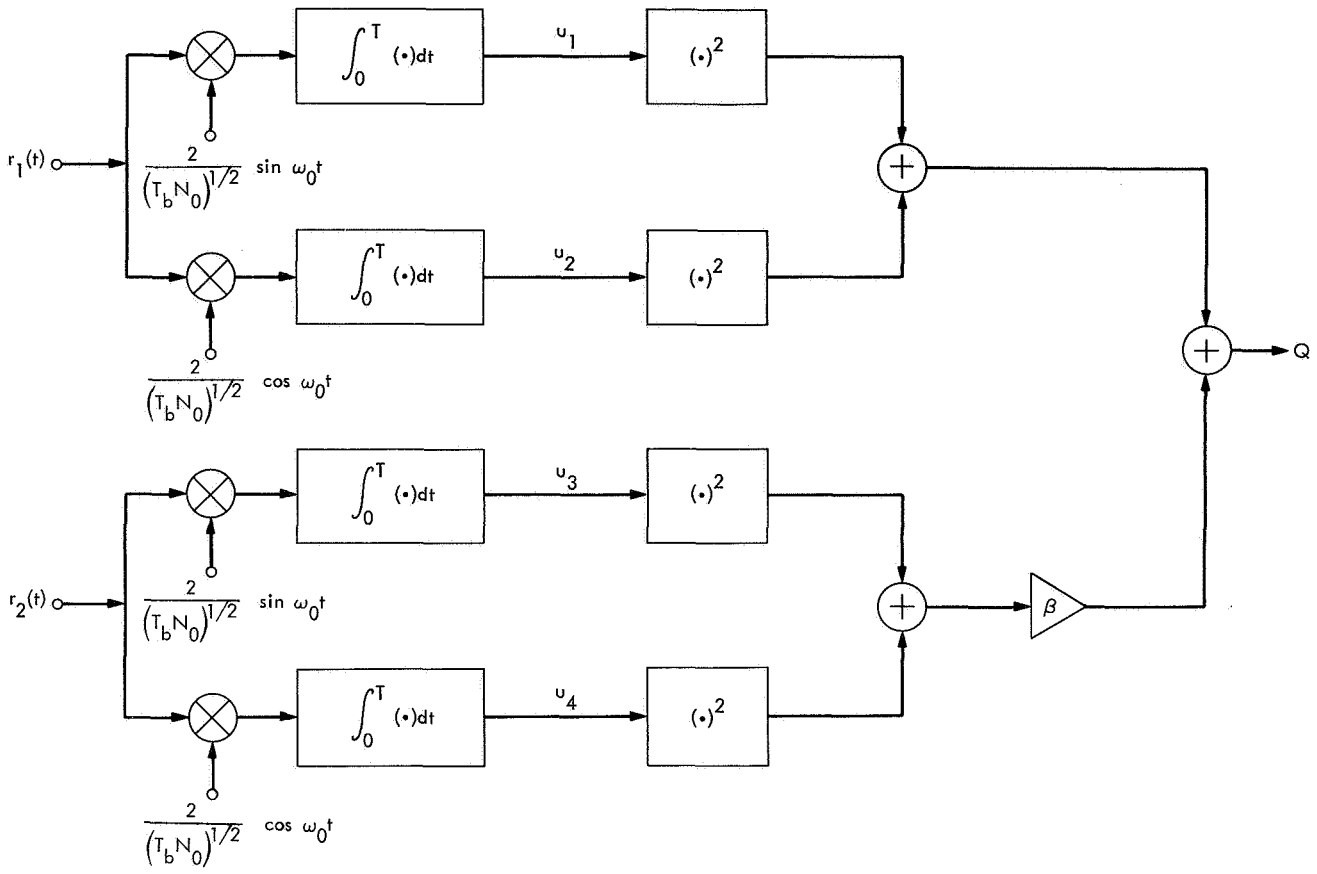


Fig. 5. Diagram of one section of the doubly incoherent receiver

where

$$u_1 = \left(\frac{2E}{N_0}\right)^{1/2} k_\phi \cos\theta_1 + n_1, \qquad u_2 = \left(\frac{2E}{N_0}\right)^{1/2} k_\phi \sin\theta_1 + n_2$$

$$u_3 = \left(\frac{2E}{N_0}\right)^{1/2} \lambda_\phi \cos\theta_2 + n_3, \qquad u_4 = \left(\frac{2E}{N_0}\right)^{1/2} \lambda_\phi \sin\theta_2 + n_4$$

and the output of the ω_1 section when $\sin \omega_0 t$ was transmitted is

$$Q_1 = v_1^2 + v_2^2 + \beta (v_3^2 + v_4^2)$$

where

$$v_1 = m_1, \qquad v_2 = m_2, \qquad v_3 = m_3, \qquad v_4 = m_4$$

The noises n_i and m_i, $i = 1$ to 4, are mutually independent gaussian random variables of unit variance. Similar variables are defined in a symmetrical way when $\sin \omega_1 t$ was transmitted.

The estimate of which value of ω was sent is taken to correspond to the section with the largest output. Thus, the probability of error is given by

$$P_E^D = \min_{0 \leq \beta \leq 1} P_r (Q_0 < Q_1 | \omega = \omega_0)$$

assuming that prob $(\omega = \omega_0)$ = prob $(\omega = \omega_1)$.

By first conditioning on Q_0, we readily find

$$P_r(Q_0 < Q_1 | \omega = \omega_0, k_\phi, \lambda_\phi, Q_0) = \frac{1}{1-\beta} \exp\left(-\frac{1}{2} Q_0^2\right) - \frac{1}{1-\beta} \exp\left(-\frac{1}{2} Q_0^2\right)$$

But we have that

$$P_E^D = \min_{0 \le \beta \le 1} E\{P_r(Q_0 < Q_1 | \omega = \omega_0, k_\phi, \lambda_\phi, Q_0)\}$$

where the expectation is taken over the variables n_1, n_2, n_3, and n_4 and the functionals k_ϕ and λ_ϕ. A straightforward integration yields

$$P_E^D = \min_{0 \le \beta \le 1} \frac{1}{2} E \frac{\exp\left[-\left(\frac{\beta}{1+\beta}\lambda_\phi^2 + \frac{1}{2}k_\phi^2\right)\frac{PT_b}{N_0}\right] - \beta^2 \exp\left[-\left(\frac{1}{2}\lambda_\phi^2 + \frac{1}{1+\beta}k_\phi^2\right)\frac{PT_b}{N_0}\right]}{1-\beta^2}$$

where the expectation is now only over k_ϕ and λ_ϕ. This is the expression used in *Subsection 3.*

9. Conclusion

For the schemes discussed, it can be seen that the optimum value of the index of modulation k for low rates is almost always given by the constraint $\rho_L = 6$.

Also, for certain rates the doubly incoherent receiver gives considerably better error performance than the incoherent receiver. Just how much better for a given rate, however, remains an open question which can perhaps best be answered by simulation.

References

1. Viterbi, A. J., *Optimum Detection and Signal Selection for Partially Coherent Binary Communication*, Wescon 13.1. Western Electronic Manufacturers' Association, Los Angeles, Calif., 1964.
2. Lindsey, W. C., "Performance of Phase-Coherent Receivers Preceded by Bandpass Limiters," *IEEE Trans. on Commun. Technol.*, April 1968.
3. Wozencraft and Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, Inc., New York, 1965.

B. Analysis of a Serial Orthogonal Decoder,

R. R. Green

1. Introduction

This article presents a more straightforward mathematical analysis of the decoder discussed in SPS 37-39, Vol. IV, pp. 247–252. As before, the problem is to perform the matrix vector product $y = H_n x$, where x is a real vector with 2^n components. H_n is the code matrix, or dictionary, defined inductively by $H_n = H_{n-1} \otimes H_1$, with

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and $\otimes$ denotes the Kronecker product.

2. Notation

Subscripts are used to denote the size of matrices in the following way: A_m implies that A_m is a 2^m by 2^m square matrix. I_m denotes the 2^m by 2^m identity matrix.

The Kronecker product of two matrices, say A and B, is defined by $A \otimes B = (a_{ij} B)$. This product is associative, i.e.,

$$(A \otimes B) \otimes C = A \otimes (B \otimes C)$$

and, if the dimensions are correct for the necessary ordinary matrix products to be defined, we have (Ref. 1)

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

From the foregoing, we have the following useful relations:

$$I_m \otimes I_n = I_{m+n}$$

$$(I_m \otimes A_n)(I_m \otimes B_n) = I_m \otimes A_n B_n$$

$$(A_n \otimes I_m)(B_n \otimes I_m) = A_n B_n \otimes I_m$$

$$(A_n \otimes I_m)(I_n \otimes B_m) = A_n \otimes B_m = (I_n \otimes B_m)(A_n \otimes I_m)$$

3. Motivation

The difficulty in evaluating $H_n x$ directly, on a term-by-term basis, is the size of H_n. Since every element in H_n is either 1 or -1, direct evaluation would involve $2^n(2^n - 1)$ additions or subtractions. This difficulty can be relieved by factoring H_n into the matrix product of n different 2^n by 2^n matrices, which will be denoted $M_n^{(1)}, M_n^{(2)}, \cdots, M_n^{(n)}$. Each matrix $M_n^{(i)}$ has only two non-zero elements per row, thus only $n2^n$ additions or subtractions are involved.

Furthermore, the structure of each matrix $M_n^{(i)}$ is such that it can be easily implemented with special-purpose digital equipment. Thus, we can construct a set of decoder stages, the first realizing $M_n^{(1)}$, the second $M_n^{(2)}$, etc. If these stages are then connected together serially, the input to stage 1 being x, the output of stage 1 being the input to stage 2, etc., the output of stage n will be y. Due to this serial structure of the decoder, n additions or subtractions are being done simultaneously. Thus, the digital equipment need only be fast enough to perform 2^n additions or subtractions per code word time, or one addition or subtraction per symbol time.

It is an interesting and somewhat surprising result that the stages of the decoder may be connected in an arbitrary order and the output of the last stage will still be the desired vector y.

---

4. Analysis

The following analysis is a special case of a more general result involving a code matrix which is the Kronecker product of n arbitrary matrices. Since the general case provides no particular additional insight into the decoder under consideration, the results have been particularized to this special case. It should be noted, however, that in the general case the factor matrices have the same form, the same commutivity result holds, and a somewhat more general product theorem can be proved.

Define

$$M_n^{(i)} = I_{n-i} \otimes H_1 \otimes I_{i-1}, \qquad \text{for } 1 \leq i \leq n$$

Theorem 1

$$M_n^{(i)} M_n^{(j)} = M_n^{(j)} M_n^{(i)}$$

Proof. Assume $i > j$ (if $i = j$, the result is trivial) then

$$\begin{aligned} M_n^{(i)} M_n^{(j)} &= (I_{n-i} \otimes H_1 \otimes I_{i-1})(I_{n-j} \otimes H_1 \otimes I_{j-1}) \\ &= [(I_{n-i} \otimes H_1 \otimes I_{i-j-1}) \otimes I_j][I_{n-j} \otimes (H_1 \otimes I_{j-1})] \\ &= I_{n-i} \otimes H_1 \otimes I_{i-j-1} \otimes H_1 \otimes I_{j-1} \\ &= [I_{n-i+1} \otimes (I_{i-j-1} \otimes H_1 \otimes I_{j-1})][(I_{n-i} \otimes H_1) \otimes I_{i-1}] \\ &= (I_{n-j} \otimes H_1 \otimes I_{j-1})(I_{n-i} \otimes H_1 \otimes I_{i-1}) \\ &= M_n^{(j)} M_n^{(i)} \end{aligned}$$

Thus, Theorem 1 shows that the order of any two successive stages may be interchanged, and thus any possible permutation of the stages may be realized, without changing the final output. Also, the commutivity shown implies that we need not keep track of order when discussing matrix products of the $M_n^{(i)}$.

Theorem 2

$$\prod_{i=1}^{m} M_n^{(i)} = I_{n-m} \otimes H_m, \qquad 1 \leq m \leq n$$

Proof. For $m = 1$, we have

$$\prod_{i=1}^{m} M_n^{(i)} = M_n^{(1)} = I_{n-1} \otimes H_1$$

Assume the result is true for m, then prove for $m + 1$:

$$\prod_{i=1}^{m+1} M_n^{(i)} = M_n^{(m+1)} \prod_{i=1}^{m} M_n^{(i)} = (I_{n-m-1} \otimes H_1 \otimes I_m)(I_{n-m} \otimes H_m) = I_{n-m-1} \otimes H_1 \otimes H_m = I_{n-m-1} \otimes H_{m+1}$$

Thus, by induction, the result is true for any m between 1 and n.

In particular, we see from Theorem 2, letting $m = n$, that

$$\prod_{i=1}^{n} M_n^{(i)} = I_{n-n} \otimes H_n = I_0 \otimes H_n = H_n$$

Also, as in the previous article on this decoder, it can be shown that

$$M_n^{(i)} = P_n^i R_n (P_n^{i-1})^T$$

where P_n and R_n are defined inductively for $n \geq 1$ by

$$P_{n+1} = (I_1 \otimes P_n)(P_2 \otimes I_{n-1})$$

and

$$R_{n+1} = (P_2 \otimes I_{n-1})(I_1 \otimes R_n)$$

with $P_1 = I_1$ and $P_2 = (P_{ij})$. Here

$$P_{11} = P_{23} = P_{32} = P_{44} = 1$$

and otherwise $P_{ij} = 0$; $R_1 = H_1$. Thus, we see that connecting the n decoder stages $M_n^{(1)}$ through $M_n^{(n)}$ in any order whatever performs the operation $H_n x$. Furthermore, if the stages are connected in numerical order, $M_n^{(1)}$ first, $M_n^{(2)}$ second, etc., the output at any intermediate stage, say the jth stage, provides a decoder for H_j. Thus, the algorithm has multiple-mission capability.

Reference

1. Bellman, Richard, *Introduction to Matrix Analysis.* McGraw-Hill Book Co., Inc., New York, 1960.

C. Optimal Codes and a Strong Converse for Transmission Over Very Noisy Memoryless Channels, A. J. Viterbi[1]

1. Introduction

Wyner (Ref. 1) has obtained the following lower bounds on the asymptotic performance of the optimal codes for the additive white gaussian channel, where T is the message duration, R is the rate in nats/s, and C is the channel capacity:

$$P_E > \exp\{-T[E(R) + o(T)]\}$$

where

$$E(R) = \frac{C}{2} - R, \qquad 0 \leq R \leq \frac{C}{4}$$

$$= [(C)^{1/2} - (R)^{1/2}]^2, \qquad \frac{C}{4} \leq R < C \qquad (1)$$

and

$$1 - P_E < \exp\{-T[E^*(R) + o(T)]\}$$

where

$$E^*(R) = [(R)^{1/2} - (C)^{1/2}]^2, \qquad R > C \qquad (2)$$

The second bound on the probability of correct decision for rates above capacity is generally referred to as a "strong converse."

It is well known (Ref. 2) that equal-energy orthogonal signals are asymptotically optimum because they achieve the error probability

$$P_E < \exp[-T E(R)] \qquad (3)$$

[1]Consultant, University of California at Los Angeles.

where $E(R)$ is given by Eq. (1). We begin by showing, through an application of extreme value theory (Ref. 3), that for rates above capacity orthogonal signals yield

$$1 - P_E > \exp\left\{-T\left[E^*(R) + O\left(\ln\frac{T}{T}\right)\right]\right\}, \qquad R > C \tag{4}$$

which proves their asymptotic optimality[2] above as well as below capacity.

We extend these results by showing the essential equivalence of all memoryless input-discrete very noisy channels to the white gaussian channel and thus extend the strong converse to this wider class of channels.

2. The Additive White Gaussian Channel

Balakrishnan (Ref. 4) has shown that for any equal-energy, *a priori* equiprobable set of M signals used on the white gaussian channel, the probability of correct decision using the optimum (maximum likelihood) decision rule is

$$1 - P_E = M^{-1} e^{-\lambda} E\left\{\exp\left[(2\lambda)^{1/2} \max_{1 \le m \le M} z_m\right]\right\} \tag{5}$$

where $\lambda = CT$, $C = S/N_0$, the ratio of received signal power to one-sided noise spectral density, while $\{z_m\}$ is a set of M zero-mean, unit-variance, gaussian random variables with a covariance matrix whose elements are the normalized integral inner products among signals.

Let $R = \ln M/T$ and restrict to orthogonal signals; Eq. (5) becomes

$$1 - P_E = \exp\left[-T(C+R)\right] \times E\left\{\exp\left[(2CT)^{1/2} \max_{1 \le m \le e^{RT}} z_m\right]\right\} \tag{6}$$

where $\{z_m\}$ are independent normalized gaussian variables, since the covariance matrix for orthogonal signals is the identity matrix.

Equation (6) can be rewritten as

$$1 - P_E = \exp\left[-T(C+R)\right] \times \int_{-\infty}^{\infty} \exp\left[(2CT)^{1/2} x\right] \frac{d}{dx}\left[F(x)\right]^{e^{RT}} dx$$

$$= \exp(-TC) \int_{-\infty}^{\infty} \exp\left[(2CT)^{1/2} x\right] \left[F(x)\right]^{e^{RT}-1} d\left[F(x)\right] \tag{7}$$

where

$$F(x) \triangleq \int_{-\infty}^{x} e^{-y^2/2} \frac{dy}{(2\pi)^{1/2}}$$

is the (cumulative) gaussian distribution.

We proceed to evaluate Eq. (7) by applying a technique from extreme value theory due to Cramér (Ref. 3). Consider the transformation

$$1 - \xi e^{-RT} = F(x) \tag{8}$$

which has the inverse (Ref. 3)

$$x = F^{-1}(1 - \xi e^{RT}) = (2RT)^{1/2} - \frac{\ln 4\pi RT}{2(2RT)^{1/2}} - \frac{\ln \xi}{(2RT)^{1/2}} + O\left(\frac{1}{RT}\right) \tag{9}$$

Substituting Eqs. (8) and (9) into Eq. (7), we obtain

$$1 - P_E = \exp\left[-T(C+R)\right] \int_0^{e^{RT}} \exp\left[(2CT)^{1/2} F^{-1}(1 - \xi e^{RT})\right] (1 - \xi e^{-RT})^{e^{RT}-1} d\xi$$

$$= \exp\left\{-T\left[C + R - 2(RC)^{1/2} + \frac{\ln 4\pi RT}{2\left(2\frac{R}{C}\right)^{1/2} T} - O(T^{-3/2})\right]\right\} \int_0^{e^{RT}} \xi^{-(C/R)^{1/2}} (1 - \xi e^{-RT})^{e^{RT}-1} d\xi \tag{10}$$

[2]It has long been conjectured that regular simplex signals are globally optimum for all rates on the white gaussian channel, and this obviously implies the asymptotic optimality of orthogonal signals. However, only the local first- and second-order conditions of optimality of regular simplex signals have been shown (Ref. 4) and all attempts at proving global optimality at all rates have met with failure.

The last integral is bounded from below by

$$\left\{e\left[1-\left(\frac{C}{R}\right)^{1/2}\right]\right\}^{-1}$$

for $R > C$. Thus, it follows that for orthogonal signals on the white gaussian channel

$$1 - P_E > \exp\left(-T\left\{[(R)^{1/2} - (C)^{1/2}]^2 + O\left(\frac{\ln T}{T}\right)\right\}\right), \quad R > C \tag{11}$$

which proves Inequality (4).

3. Input-Discrete Very Noisy Memoryless Channels

The error probability expression for the white gaussian channel, Eq. (5), can be generalized to any memoryless finite-dimensional (or time-discrete) channel. For any set of M equally likely messages and a maximum likelihood decision rule, for any set of N-dimensional channel input sequences $\{\mathbf{x}^{(j)}; j = 1, 2, \cdots, M\}$, and for $\mathbf{y}$, an N-dimensional output sequence, we have

$$1 - P_E = M^{-1} \sum_{j=1}^{M} \int_{D_j} p(\mathbf{y}|\mathbf{x}^{(j)})\, d\mathbf{y} \tag{12}$$

where

$$D_j = \{\mathbf{y} : p(\mathbf{y}|\mathbf{x}^{(j)}) = \max_m p(\mathbf{y}|\mathbf{x}^{(m)})\} \tag{13}$$

Then, since

$$\bigcup_{j=1}^{M} D_j = Y_N$$

the N-dimensional output space, we can rewrite Eq. (12) as

$$\begin{aligned} 1 - P_E &= M^{-1} \int_{Y_N} \max_m p(\mathbf{y}|\mathbf{x}^{(m)})\, d\mathbf{y} \\ &= M^{-1} \int_{Y_N} q(\mathbf{y}) \max_m \frac{p(\mathbf{y}|\mathbf{x}^{(m)})}{q(\mathbf{y})}\, d\mathbf{y} \\ &= M^{-1} E_{\mathbf{y}}\left[\max_m \frac{p(\mathbf{y}|\mathbf{x}^{(m)})}{q(\mathbf{y})}\right] \end{aligned} \tag{14}$$

where $q(\mathbf{y})$ is an arbitrary probability measure on the output space and $E_{\mathbf{y}}$ is the expectation with respect to this measure. Substitution of the appropriate likelihood functions for the white gaussian channel and for the $q(\mathbf{y})$ corresponding to the likelihood function for a zero-signal hypothesis reduces Eq. (14) to Eq. (5) (cf Helstrom, Ref. 5).

For memoryless time-discrete channels,

$$p(\mathbf{y}|\mathbf{x}^{(m)}) = \prod_{n=1}^{N} p(y_n|x_n^{(m)})$$

and specializing to the independent output measure,

$$q(\mathbf{y}) = \prod_{n=1}^{N} q(y_n)$$

Eq. (14) becomes

$$1 - P_E = M^{-1} E_{\mathbf{y}}\{\exp[\max_m z_m(\mathbf{y})]\} \tag{15}$$

where

$$z_m(\mathbf{y}) = \sum_{n=1}^{N} \ln\left[\frac{p(y_n|x_n^{(m)})}{q(y_n)}\right] \tag{16}$$

Such channel is said to be *very noisy* if

$$p(y_n|x_m^{(n)}) = q(y_n)\,[1 + \epsilon(x_n^{(m)}, y_n)] \tag{17}$$

where $\epsilon(x, y) \to 0$ uniformly in x and y, and $q(y_n)$ is an arbitrary probability density or distribution. It follows that for any n and m

$$0 = \int_Y p(y|x)\, dy - 1 = \int_Y q(y)\, \epsilon(x, y)\, dy \tag{18}$$

We now restrict attention to a discrete input alphabet of K symbols, so that each $x_n^{(m)}$ is taken from the set $\{x_1, x_2, \cdots, x_K\}$. For this class of channels, we need only consider the class of *fixed composition* codes, which are characterized by the property that each code word is some permutation of the same sequence of N symbols, since Shannon, Gallager, and Berlekamp (Ref. 6) have shown that the asymptotic performance of the best code in the restricted subclass is the same as for the best code in the unrestricted class. Thus, given that the relative frequency of the symbol x_k in each code word of the fixed composition code is

$$p_k \triangleq \frac{\text{number of occurrences of } x_k}{N}, \qquad k = 1, 2, \cdots, K$$

we have that the means of the random variables $z_m(\mathbf{y})$ relative to the output measure

$$\prod_{n=1}^{N} q(y_n)$$

are all equal to

$$\left.\begin{aligned} E_{\mathbf{y}}[z_m(\mathbf{y})] &= N\sum_{k=1}^{K}\rho_k\int_Y q(y)\left[\ln\frac{p(y|x_k)}{q(y)}\right]dy \\ &\approx -N\sum_{k=1}^{K}\rho_k\int_Y q(y)\left[\frac{\epsilon^2(x_k,y)}{2}\right]dy \end{aligned}\right\} \quad (19)$$

where we have used Eq. (18) and also Condition (17) to neglect all terms above quadratic in ϵ.

Similarly, since the channel is memoryless,

$$\mathrm{var}_{\mathbf{y}}[z_m(\mathbf{y})] = N\sum_{l=1}^{K}\rho_k\,\mathrm{var}_y\left[\ln\frac{p(y|x_k)}{q(y)}\right] \quad (20)$$

But again neglecting terms above quadratic in ϵ,

$$\mathrm{var}_y\left[\ln\frac{p(y|x_k)}{q(y)}\right] \approx \int_Y q(y)\,\epsilon^2(x_k,y)\,dy \quad (21)$$

Also, the capacity of a very noisy input-discrete memoryless channel is given by

$$\begin{aligned} C &= \max_{\{p_k\}}\sum_{k=1}^{K}p_k\int_Y p(y|x_k)\ln\frac{p(y|x_k)}{q(y)}\,dy \\ &\approx \max_{\{p_k\}}\sum_{k=1}^{K}p_k\int_Y q(y)\,[1+\epsilon(x_k,y)] \\ &\quad\times\left[\epsilon(x_k,y)-\frac{\epsilon^2(x_k,y)}{2}\right]dy \\ &\approx \max_{\{p_k\}}\sum_{k=1}^{K}p_k\int_Y q(y)\frac{\epsilon^2(x_k,y)}{2}\,dy \end{aligned} \quad (22)$$

Thus, choosing the relative frequencies $\{\rho_k\}$ corresponding to the maximizing distribution for capacity, we have from Eqs. (19), (20), and (21)

$$E_{\mathbf{y}}[z_m(\mathbf{y})] \approx -NC \quad (23)$$

$$\mathrm{var}_{\mathbf{y}}[z_m(\mathbf{y})] = 2NC \quad (24)$$

Furthermore, since $z_m(\mathbf{y})$ is the sum of N independent random variables, by the central limit theorem it must be asymptotically gaussian. In fact, if we normalize by letting

$$v_m(\mathbf{y}) \triangleq \frac{z_m(\mathbf{y})+NC}{(2NC)^{1/2}} \quad (25)$$

it follows from Eqs. (23) and (24) that $v_m(\mathbf{y})$ is a zero-mean, unit-variance, random variable and by the Berry–Esseen theorem (cf Loève, Ref. 7) we have that $P_v(x)$, the distribution function of the normalized variable v_m, differs from the normalized gaussian distribution $F(x)$ by no more than

$$\begin{aligned} |P_v(x)-F(x)| &\leq \frac{\theta E(|z_m|^3)}{(\mathrm{var}\,z_m)^{3/2}} \\ &\approx \frac{\theta N\sum_{k=1}^{K}\rho_k\int_Y q(y)|\epsilon(x_k,y)|^3\,dy}{(2NC)^{3/2}} \\ &\approx 0 \end{aligned}$$

when we neglect all terms in ϵ of order higher than quadratic.

Thus, all the variables v_m are asymptotically gaussian with zero means and unit variances. Applying Eq. (25) to Eq. (15) and letting $R' = (\ln M)/N$ nats/symbol,

$$\begin{aligned} 1-P_E &= \exp[-N(R'+C)] \\ &\quad\times E_{\mathbf{y}}\{\exp[(2NC)^{1/2}\max_{1\leq m\leq e^{NR'}}v_m(y)]\} \end{aligned} \quad (26)$$

This formula is identical to the form of Eq. (6) for orthogonal signals on white gaussian channels, except that the variables v_m are not necessarily independent. However, for rates below capacity it is well known (Ref. 6) that the error probability for the best code on memoryless very noisy input-discrete channels behaves asymptotically exactly as that for orthogonal signals in the white gaussian channel [i.e., Expressions (1) and (3) hold with T replaced by N and R replaced by R']. For

this to be the case, the best code on memoryless very noisy input-discrete memoryless channels must asymptotically lead to independent $v_m(y)$ in Eq. (26), since any other covariance matrix would lead asymptotically to a greater P_E below capacity. Thus, Eq. (26) reduces to Eq. (6), and the asymptotic behavior above capacity given by Expressions (2) and (4) must hold also for the best code on this class of channels.

References

1. Wyner, A. D., "On the Probability of Error for Communication in White Gaussian Noise," *IEEE Trans. on Inform. Theory*, Vol. IT-13, pp. 86–90, Jan. 1967.
2. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*. John Wiley & Sons, Inc., New York, 1965.
3. Cramér, H., *Mathematical Methods of Statistics*, pp. 374–376. Princeton University Press, Princeton, N. J., 1946.
4. Balakrishnan, A. V., "A Contribution to the Sphere-Packing Problem of Communication Theory," *J. Math. Analysis and Appl.*, Vol. 3, No. 3, pp. 485–506, Dec. 1961.
5. Helstrom, C. W., Editor's Note, *IEEE Trans. Inform. Theory*, Vol. IT-14, No. 2, p. 311, Mar. 1968.
6. Shannon, C. E., Gallager, R. G., and Berlekamp, E. R., "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels, I," *Inform. Contr.*, Vol. 10, No. 1, pp. 81–83, Jan. 1967.
7. Loève, M., *Probability Theory*, p. 288. Van Nostrand, New York, 1965.

XVIII. Communications Systems Research: Combinatorial Communication

TELECOMMUNICATIONS DIVISION

A. Cross-Correlations of Reverse Maximal-Length Shift Register Sequences,

T. A. Dowling[1] *and R. McEliece*

1. Introduction

Maximal-length binary shift register sequences with uniformly low cross-correlations are used for synchronization in spread spectrum communication systems (Ref. 1). In this article, a result on exponential sums in finite fields is applied to bound the cross-correlation of a maximal-length sequence with any phase shift of the reverse sequence. The result can also be applied to bound the non-zero weights of a certain $(2^k - 1, 2k)$ cyclic code.

2. A Bound on Cross-Correlations

If $a = (a_0, a_1, \cdots, a_{n-1})$ and $b = (b_0, b_1, \cdots, b_{n-1})$ are two sequences of length n over $GF(2)$, the cross-correlation function $C(\tau)$ with respect to a and b is defined by

$$C(\tau) = \sum_{i=0}^{n-1} s(a_i)\, s(b_{i+\tau})$$

where the subscripts are reduced modulo n and $s(x) = (-1)^x, x \in GF(2)$.

Consider the case where a is a maximal-length shift register sequence of length $2^k - 1$ and b is the reverse sequence, i.e., $b_i = a_{2^k-2-i}$, $i = 0, 1, \cdots, 2^k - 2$. Regard a as a non-null code word of the $(2^k - 1, k)$ cyclic code A generated by linear recursion (Ref. 2) by a primitive polynomial $f(x)$ of degree k over $GF(2)$. Then, the elements of a may be characterized by the Mattson–Solomon polynomial

$$g_a(x) = \mathrm{Tr}(cx)$$

where $c \in GF(2^k)$ and

$$\mathrm{Tr}(x) = \sum_{i=0}^{k-1} x^{2^i}$$

is the trace of $GF(2^k)/GF(2)$. The polynomial $g_a(x)$ satisfies $g_a(\alpha^i) = a_i, i = 0, 1, \cdots, 2^k - 2$, where α is a root of $f(x)$. With no loss of generality, a is assumed to be phase-shifted so that $c = 1$. Then,

$$b_i = \mathrm{Tr}(\alpha^{-(i+1)})$$

Thus,

$$C(\tau) = \sum_{i=0}^{2^k-2} (-1)^{\mathrm{Tr}(\alpha^i + \alpha^{-(\tau+1+i)})}$$

since $\mathrm{Tr}(x + y) = \mathrm{Tr}(x) + \mathrm{Tr}(y)$ for $x, y \in GF(2^k)$.

[1]NASA Summer Faculty Fellow, Dept. of Statistics, Univ. of No. Carolina.

Now, a result on exponential sums in finite fields is applied. This result was first proved by A. Weil (Ref. 3) and later, using different methods, by L. Carlitz and S. Uchiyama (Ref. 4). Let $K = GF(q)$, where $q = p^k$, and let $\mathrm{Tr}(x) = x + x^p + \cdots + x^{p^{k-1}}$; then, for any $c \in K$,

$$\left| \sum_{\substack{x \in k \\ x \neq 0}} \exp\left[\frac{2\pi i \,\mathrm{Tr}(x + cx^{-1})}{p}\right] \right| \leq 2q^{1/2} \tag{1}$$

Setting $p = 2$ and $c = \alpha^{-(\tau+1)}$ and expressing x as α^i, Eq. (1) becomes

$$|C(\tau)| \leq 2^{(k+2)/2} \tag{2}$$

Calculations indicate that this bound is tight. It can be shown that $C(\tau) \equiv -1 \pmod 4$ for any τ. The extremal values, both maximum and minimum, satisfying Eq. (2) and this condition are attained for the cases computed ($k \leq 8$).

3. Application to Coding Theory

The bound given by Eq. (1) can also be applied to bound the non-zero weights of the $(2^k - 1, 2k)$ cyclic code generated by $f(x)f^*(x)$ by linear recursion, where $f^*(x) = x^k f(1/x)$ is the reciprocal polynomial of $f(x)$. The Mattson–Solomon polynomial for this code is $g_a(x) = \mathrm{Tr}(cx) + \mathrm{Tr}(dx^{-1})$, where $c, d \in GF(2^k)$. If $w(c, d)$ denotes the weight of the code word with this polynomial, then $w(c, d) = w(1, cd)$ if $c \neq 0$, since the pairs (c, d) and $(1, cd)$ correspond to different cyclic shifts of the same code word. Hence, we can take $c = 1$ if $c \neq 0$. Then,

$$w(1, d) = \frac{2^k - 1 - \rho(d)}{2} \tag{3}$$

where

$$\rho(d) = \sum_{\substack{x \in GF(2^k) \\ x \neq 0}} (-1)^{\mathrm{Tr}(x + dx^{-1})}$$

Thus, by Eq. (1), $|\rho(d)| \leq 2^{(k+2)/2}$. Since $w(0, d) = 2^{k-1}$ if $d \neq 0$, we have, using Eq. (3),

$$2^{k-1} - 2^{k/2} - 1/2 \leq w \leq 2^{k-1} + 2^{k/2} - 1/2$$

where w is the weight of any non-zero code word of A.

References

1. Gold, R., "Optimal Binary Sequences for Spread Spectrum Multiplexing," (Correspondence), *IEEE Trans. Inform. Theory*, Vol. IT-13, pp. 619–621, 1967.
2. Mattson, H. F., and Solomon, G., "A New Treatment of Bose-Chaudhuri Codes," *J. Soc. Ind. Appl. Math.*, Vol. 9, pp. 654–669, 1961.
3. Weil, A., "On Some Exponential Sums," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 34, pp. 204–207, 1948.
4. Carlitz, L., and Uchiyama, S., "Bounds for Exponential Sums," *Duke Math. J.*, Vol. 24, pp. 37–41, 1957.

XIX. Communications Systems Research: Propagation Studies

TELECOMMUNICATIONS DIVISION

A. Two Stochastic Approximation Procedures for Identifying Linear Systems, *J. K. Holmes*

1. Introduction

Many important problems in communications research can be posed as a problem of system identification. That is, given an input signal record and an output signal record, find an equivalent system that fits these data. An important example occurs in the following communication problem where it is desired to study the effects of the atmosphere on signal transmission. A transmitter transmits directly to a tracking station via cables or a microwave link, and to a spacecraft which retransmits to the tracking station. The signal, as it passes through, for example, the Solar Corona to and from the spacecraft, undergoes distortion. One method of studying the distortion is to characterize it as a finite memory linear system. The identification could be derived from the input (the directly transmitted component) with the proper time delay, and the output, which is the retransmitted, distorted signal. Naturally, noise would, in general, corrupt both the input and output signal measurements.

This article, then, considers the basic problem of identifying linear systems from noisy input–output measurements. Related work is listed in Refs. 1–5.

Let N denote the set of natural numbers and let n be in N. Denote by $\mathcal{S}_L(\ell)$ the class of linear, finite memory, time invariant, time discrete systems. Then, for systems $s \in \mathcal{S}_L(\ell)$, we can relate the input and output random sequences by

$$v_n = \sum_{j=0}^{\ell-1} \phi_j u_{n-j}, \qquad \ell < \infty \tag{1}$$

The vector $\Phi = (\phi_0, \cdots, \phi_{\ell-1})$ defines the system when it is in the class $\mathcal{S}_L(\ell)$. Further, we assume that the observable sequences v'_n and u'_n are defined as $v'_n = v_n + \delta_n$ and $u'_n = u_n + \epsilon_n$, where ϵ_n and δ_n are mutually and individually independent noise sequences with zero means and respective variances σ_ϵ^2 and σ_δ^2. Also, u_n is assumed to be independent of ϵ_n and δ_n. This article, then, is concerned with the following problem: From the noise-corrupted input and output measurements, estimate the unknown system (i.e., Φ) via nonparametric methods. See Fig. 1 for a block diagram.

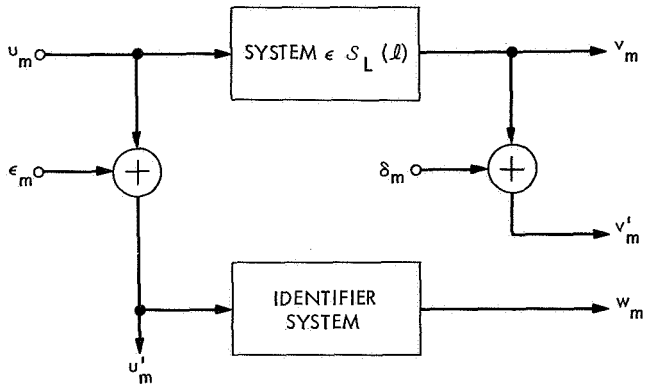


Fig. 1. Block diagram of available measurements u'_m and v'_m and conceptual location of identifier system

2. Development of the First Algorithm

Let the estimator of the sequence $v_m(\varphi)$ be

$$w_m(\mathbf{z}) = \sum_{j=0}^{\ell-1} z_j u'_{m-j}$$

where $m = \ell n$ and the row vector $\mathbf{z} \in E^\ell$. Now, we shall derive a specific sequential estimate of φ denoted by $\mathbf{x}^m$. Define an error measure in the following way:

$$M(\mathbf{z}) = E[(w_m(\mathbf{z}) - v'_m)^2 | \mathbf{z}] \tag{2}$$

The procedure to be used to estimate φ will be to recursively determine $\mathbf{z}$ in such a manner that $M(\mathbf{z})$ is a minimum. Specifically, we consider a Kiefer–Wolfowitz stochastic approximation method.

Let

$$\rho_m = (w_m(\mathbf{x}^m) - v'_m)^2$$

and define

$$D_i^m = \rho_m(\mathbf{x}^m + c_m \mathbf{e}_i) - \rho_m(\mathbf{x}^m - c_m \mathbf{e}_i),$$
$$i = 0, \cdots, \ell - 1$$

where c_m satisfies

$$c_m \geqq 0 \qquad \text{and} \qquad \lim_{m\to\infty} c_m = 0$$

and the $\mathbf{e}_i$ are the orthonormal unit vectors $\in E^\ell$, viz.,

$$\mathbf{e}_0 = (1, 0, \cdots, 0) \qquad \text{and} \qquad \mathbf{e}_{\ell-1} = (0, 0, \cdots, 1)$$

Then, recursively define $\mathbf{x}^m$ by

$$\mathbf{x}^{m+\ell} = \mathbf{x}^m - a_m \frac{\mathbf{D}^m}{c_m} \text{ for } m = \ell, 2\ell, 3\ell, \cdots \tag{3}$$

where

$$\mathbf{D}^m = (D_0^m, D_1^m, \cdots, D_{\ell-1}^m)$$

and a_m satisfies

$$a_m \geqq 0 \qquad \text{and} \qquad \lim_{m\to\infty} a_m = 0$$

By using the definition of $\mathbf{D}^m$ and ρ_m, one obtains the following scalar algorithm for each component of the estimate:

$$x_i^{m+\ell} = x_i^m - 4a_m u'_{m-i}\left[\sum_{j=0}^{\ell-1} x_j^m u'_{m-j} - v'_m\right],$$
$$i = 0, \cdots, \ell - 1$$

or in vector form, designating $(\mathbf{u}'_m)^T$ as the transpose of $\mathbf{u}'_m$, we have

$$\mathbf{x}^{m+\ell} = \mathbf{x}^m - 4a_m \mathbf{x}^m (\mathbf{u}'_m)^T \mathbf{u}'_m + 4a_m v'_m \mathbf{u}'_m \tag{4}$$

which is independent of c_m, and where

$$\mathbf{u}'_m = (u'_m, u'_{m-1}, \cdots, u'_{m-\ell+1})$$

3. The Estimate Error

Based on minimizing the mean square error between the sequence $v_m(\varphi)$ and $w_m(\mathbf{x}^m)$, where $\mathbf{x}^m$ is defined by Eq. (3), we show that $\mathbf{x}^m$ does not converge to φ unless the input noise sequence ϵ_n is identically zero. More precisely, we have our first result.

Theorem 1. If the conditions

(a) $a_m \geqq 0$, $\sum_1^\infty a_m = \infty$, and $\sum_1^\infty a_m^2 < \infty$.

(b) $\mathbf{y}R(\mathbf{y})^T > 0$, $\forall \mathbf{y} \neq 0$, and $E[\mathbf{u}_m^T(\mathbf{u}_m)|\mathbf{x}^m] = \mathrm{R}$.

(c) $E[(u'_m)^4|\mathbf{x}^m] < c_1 < \infty$, and $E[u_m^2 \delta_m^2|\mathbf{x}^m] < c_2 < \infty$.

(d) The random sequence u_m and both noise sequences δ_m and ϵ_m are time-stationary.

(e) $E[\delta_m] = E[\epsilon_m] = 0, E[\delta_m \delta_n] = \sigma_\delta^2 \delta_{mn}$, and $E[\epsilon_m \epsilon_n] = \sigma_\epsilon^2 \delta_{mn}$.

(f) $E[\delta_m \epsilon_n] = E[u_m \delta_n] = E[\delta_m \epsilon_n] = 0$.

(g) Unknown system $s \in \mathcal{S}_L(\ell)$.

are met, then the sequence $\mathbf{x}^m$, defined by Eq. (3), converges in mean square to the vector

$$\theta = [R + \sigma_\epsilon^2 I]^{-1} R\phi$$

Proof. Equation (4) may be expanded to

$$\mathbf{x}^{m+\ell} = \mathbf{x}^m - 4a_m \mathbf{u}'_m (\mathbf{u}'_m)^T \mathbf{x}^m + 4a_m [\delta_m \mathbf{u}'_m + \mathbf{u}'_m (\mathbf{u}_m)^T \phi] \tag{5}$$

where in Eq. (5) and throughout the rest of this proof $\mathbf{x}^m$, $\mathbf{u}'_m$, $\mathbf{u}_m$, and ϕ are now defined as column vectors instead of row vectors. Using $E[(\cdot)] = EE[(\cdot)|\mathbf{x}^n]$ on both sides of Eq. (5) results in

$$E[\mathbf{x}^{m+\ell}] = E[\mathbf{x}^m] - 4a_m [R + \sigma_\epsilon^2 I] E[\mathbf{x}^m] + 4a_m R\phi \tag{6}$$

Clearly, a solution to Eq. (6) is given by

$$E[\mathbf{x}^m] = [R + \sigma_\epsilon^2 I]^{-1} R\phi \tag{7}$$

We shall now show that

$$\lim_{m \to \infty} E[\|\mathbf{x}^m - \theta\|^2] = 0$$

Subtracting θ from both sides of Eq. (5) and rearranging yields

$$(\mathbf{x}^{m+\ell} - \theta) = (\mathbf{x}^m - \theta) - 4a_m \mathbf{u}'_m (\mathbf{u}'_m)^T (\mathbf{x}^m - \theta) + 4a_m [\delta_m \mathbf{u}'_m + \mathbf{u}'_m \mathbf{u}_m^T \phi - \mathbf{u}'_m (\mathbf{u}'_m)^T \theta]$$

Forming the norm squares of both sides and averaging, we obtain

$$\begin{aligned} b_{m+\ell} = b_m &- 8a_m E\langle \mathbf{x}^m - \theta, \mathbf{u}'_m (\mathbf{u}'_m)^T (\mathbf{x}^m - \theta)\rangle \\ &+ 16a_m^2 E[\|\mathbf{u}'_m (\mathbf{u}'_m)^T (\mathbf{x}^m - \theta)\|^2] \\ &+ 16a_m^2 E[\|\delta_m \mathbf{u}'_m + \mathbf{u}'_m \mathbf{u}_m^T \phi - \mathbf{u}'_m (\mathbf{u}'_m)^T \theta\|^2] \\ &+ 8a_m E\langle \mathbf{x}^m - \theta, \delta_m \mathbf{u}'_m + \mathbf{u}'_m \mathbf{u}_m^T \phi - \mathbf{u}'_m (\mathbf{u}'_m)^T \theta\rangle \\ &- 32a_m^2 E\langle \mathbf{u}'_m (\mathbf{u}'_m)^T (\mathbf{x}^m - \theta), \delta_m \mathbf{u}'_m \\ &+ \mathbf{u}'_m \mathbf{u}_m^T \phi - \mathbf{u}'_m (\mathbf{u}'_m)^T \theta\rangle \end{aligned} \tag{8}$$

where we have let

$$b_m = E[\|\mathbf{x}^m - \theta\|^2]$$

By Condition (b) we have for the second term in Eq. (8)

$$\sigma_\epsilon^2 b_m < \lambda_1 b_m \leq \langle \mathbf{x}^m - \theta, [R + \sigma_\epsilon^2 I](\mathbf{x}^m - \theta)\rangle \tag{9}$$

where λ_1 is the minimum eigenvalue of $[R]$. The third term can be bounded, using the Schwartz inequality, by

$$16a_m^2 b_m k_1 \tag{10}$$

where $k_1 < \infty$ by Condition (c). The fourth term can be bounded by the following quantity

$$32a_m^2 E[\|\delta_m \mathbf{u}'_m\|^2] + 32a_m^2 E[\|\mathbf{u}'_m \mathbf{u}_m^T \phi\|^2] + 32a_m^2 E[\|\mathbf{u}'_m (\mathbf{u}'_m)^T \theta\|^2]$$

All these terms are similarly bounded by

$$k_2 a_m^2 \tag{11}$$

where $k_2 < \infty$. Now, the fifth term can be reduced to

$$8a_m E\langle x^m - \theta, R\phi - [R + \sigma_\epsilon^2 I]\theta\rangle = 0$$

since $\theta = [R + \sigma_\epsilon^2 I]^{-1} R\phi$. The last term is bounded by

$$\begin{aligned} 32a_m^2 \{E[\|\mathbf{u}'_m (\mathbf{u}'_m)^T (\mathbf{x}^m - \theta)\|^2] E[\|\mathbf{u}'_m \mathbf{u}_m^T \phi - \mathbf{u}'_m (\mathbf{u}'_m)^T \theta\|^2]\}^{1/2} \\ \leq 32a_m^2 \{E[\|\mathbf{x}^m - \theta\|^2 E[\|\mathbf{u}'_m (\mathbf{u}'_m)^T\|^2 | \mathbf{x}^m]] E[\|\mathbf{u}'_m \mathbf{u}_m^T \phi - \mathbf{u}'_m (\mathbf{u}'_m)^T \theta\|^2]\}^{1/2} \end{aligned} \tag{12}$$

where $\|A\|$, with A a square matrix, is the usual euclidean ℓ_2 norm; i.e.,

$$\|A\| = \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$$

Equation (12) can be shown to be bounded by

$$32a_m^2 (b_m k_3)^{1/2} \leq a_m^2 (1 + b_m) k_4 \tag{13}$$

for $k_4 < \infty$. Hence, Eq. (8) yields

$$b_{m+\ell} \leqq b_m [1 - 8a_m(\lambda_1 - k_5 a_m)] + a_m^2 k_6 \leqq b_m(1 - 4a_m \lambda_1) + a_m^2 k_6 \tag{14}$$

for m sufficiently large and k_5 and k_6 finite. By applying an interesting application of Kronecker's theorem (e.g., Ref. 5) or Lemma I of Ref. (3), we have that $b_m \to 0$. The theorem follows directly.

To conclude the statement made in the first paragraph of this subsection, we have:

Lemma 1. If the conditions of Theorem 1 are satisfied and if $\phi \neq 0$, then $\mathbf{x}^m$ is a consistent estimator of ϕ if and only if $\sigma_\epsilon^2 = 0$.

Proof. The Lemma follows from Theorem 1 and the fact that $[R + \sigma_\epsilon^2 I]^{-1} R\phi$ is equal to ϕ if and only if $\sigma_\epsilon^2 = 0$, since by Condition (b) R is positive definite.

The fact that $\mathbf{x}^m$ converges to $\theta \neq \phi$, based on minimizing $M(\mathbf{x}^m)$, is not so surprising since Eq. (2) can be written as

$$M(\mathbf{z}) = E\left[\left(\sum_{j=0}^{\ell-1} (z_j - \phi_j) u_{m-j}\right)^2 | \mathbf{z}\right] + \sigma_\epsilon^2 \|\mathbf{z}\|^2 + \sigma_\delta^2 \tag{15}$$

which reduces to

$$M(z) = (\mathbf{z} - \theta)[R + \sigma_\epsilon^2 I](\mathbf{z} - \theta)^T + (\theta - \phi) R (\theta - \phi)^T + \sigma_\epsilon^2 \|\theta\|^2 + \sigma_\delta^2 \tag{16}$$

where

$$\theta = [R + \sigma_\epsilon^2 I]^{-1} R\phi$$

Since R is positive definite, we have that the minimum occurs at $\mathbf{z} = \theta$. Hence, we *cannot* exactly identify an unknown system, when the input noise variance is non-zero, without additional knowledge of the input noise statistics.

4. Second Algorithm

A modification of the original algorithm, defining a new sequence $\mathbf{y}^m$, has the property that if σ_ϵ^2 is known then the modified algorithm leads to the result that $\mathbf{y}_m \to \phi$ in mean square.

Theorem 2. If the conditions of Theorem 1 are fulfilled and if $\mathbf{y}^m$ is defined recursively by

$$\mathbf{y}^{m+\ell} = \mathbf{y}^m - 4a_m [\mathbf{u}'_m (\mathbf{u}'_m)^T - \sigma_\epsilon^2 I] \mathbf{y}^m + 4a_m \mathbf{u}'_m v'_m \tag{17}$$

then $\mathbf{y}^m$ converges in mean square to ϕ.

Proof. The proof follows the lines of Theorem 1 and will not be indicated here.

5. Convergence to a Subspace

Up to this point, the condition that R be positive definite was required when $\sigma_\epsilon^2 = 0$ to prove convergence (in mean square) to a unique point. It is to be noted that if $\sigma_\epsilon^2 > 0$ then it is not necessary to assume that R is positive definite. However, if it turns out that R is of rank $r < \ell$, then it can be proven that $\mathbf{x}^n$ converges to a subspace of E^ℓ. Obviously, this is a very weak form of convergence. However, looking at it philosophically, we cannot hope to do any better since if $x^n \epsilon \eta$, the null space of R, $M(z)$ will be minimum and that is the best we can do using the mean square error criterion. This then leads us to the statement of Theorem 3.

Theorem 3. If the Conditions (a), (c)–(g), $0 <$ rank $R = r < \ell$, and $\sigma_\epsilon^2 = 0$ are satisfied, then $R(\mathbf{x}^n - \phi)$ converges in mean square to zero.

Proof. Starting with Eq. (5), setting ϵ_m equal to zero, subtracting ϕ from both sides of the equation, and premultiplying by R yields

$$R(\mathbf{x}^{m+\ell} - \phi) = R(x^m - \phi) - 4a_m R\mathbf{u}_m \mathbf{u}_m^T (\mathbf{x}^m - \phi) + 4a_m \delta_m R\mathbf{u}_m$$

Forming the norm square of both sides and averaging, we obtain

$$\begin{aligned} d_{m+\ell} = d_m &- 8a_m E\langle R(\mathbf{x}^m - \phi), R\mathbf{u}_m \mathbf{u}_m^T (\mathbf{x}^m - \phi)\rangle \\ &+ 16a_m^2 E[\|R\mathbf{u}_m \mathbf{u}_m^T (\mathbf{x}^m - \phi)\|^2] \\ &+ 16a_m^2 E[\delta_m^2 \|R\mathbf{u}_m\|^2] \\ &+ 8a_m E\langle R(\mathbf{x}^m - \phi), \delta_m R\mathbf{u}_m\rangle \\ &- 32a_m^2 E\langle R\mathbf{u}_m \mathbf{u}_m^T (\mathbf{x}^m - \phi), \delta_m R\mathbf{u}_m\rangle \end{aligned} \tag{18}$$

where we have let

$$d_m = E[\|R(\mathbf{x}^m - \phi)\|^2]$$

Now, even though it is no longer true that

$$E\langle \mathbf{x}^m - \boldsymbol{\phi}, R(\mathbf{x}^m - \boldsymbol{\phi})\rangle \geqq \lambda_1 E[\|\mathbf{x}^n - \boldsymbol{\phi}\|^2]$$

it is true that

$$E\langle R(\mathbf{x}^m - \boldsymbol{\phi}), RR(\mathbf{x}^m - \boldsymbol{\phi})\rangle \geqq \lambda_\rho E[\|R(\mathbf{x}^m - \boldsymbol{\phi})\|^2] = \lambda_\rho d_m \tag{19}$$

where λ_ρ is the least nonzero eigenvalue of R.

The third term can be bounded by

$$16a_m^2 d_m K_7 \tag{20}$$

where $K_7 < \infty$ by Condition (c). The fourth term can be bounded by

$$16K_8 a_m^2 \tag{21}$$

The fifth and sixth terms are zero since $E[\delta_m] = 0$.

Hence, we have that

$$d_{m+1} \leqq d_m[1 - 8a_m(\lambda_\rho - K_7 a_m)] + 16K_8 a_m^2$$

or

$$d_{m+1} \leqq d_m(1 - 4a_m \lambda_\rho) + 16K_8 a_m^2 \tag{22}$$

for m sufficiently large and K_7 and K_8 finite. As before, by application of Kronecker's theorem to Eq. (22), we have that $d_m \to 0$. The theorem follows directly. It is not hard to show from Eq. (15) that $M(\mathbf{z})$ is minimized for any $\mathbf{z} \in \eta$.

6. An Example

A simple example was devised to compare the first with the second algorithm. The "unknown system" was programmed to be of the form

$$v_m = \sum_{J=0}^{9} \phi_j u_{m-j}, \qquad \phi_j = \exp\left(-\frac{j}{2}\right); j = 0, \cdots, 9 \tag{23}$$

and the identifier system was programmed to be of the form

$$w_m = \sum_{J=0}^{9} x_j^m u'_{m-j}$$

The relative mean square error was used to measure the performance of the algorithm and, for our example, is defined by

$$\text{rmse} = \frac{\sum_{i=0}^{9} (x_i - \phi_i)^2}{\sum_{i=0}^{9} \phi_i^2} \tag{24}$$

where the true system values are given by ϕ_i and the estimate of the system by x_i.

The respective elements of the correlation matrix were simulated to satisfy

$$E[u_m u_n] = \delta_{mn}, \qquad E[\epsilon_m \epsilon_n] = K\delta_{mn}, \qquad E[\delta_m \delta_n] = 0 \tag{25}$$

where u_n and ϵ_n were programmed to simulate independent gaussian random sequences. In Eq. (25), δ_{mn} is the Kronecker delta and K was chosen to be either (a) $K = 1$ or (b) $K = \frac{1}{4}$, corresponding to an observational signal-to-noise ratio ($\sigma_u^2/\sigma_\epsilon^2$) of 0 and 6 dB, respectively. Forty thousand samples were used to obtain the system estimates.

Figure 2 illustrates the plot of the rmse as a function of the time index m. As can be seen from Fig. 2, the first algorithm, for both signal-to-noise ratios, approached the theoretical limit defined by the result of Theorem 1 and Eq. (24). Furthermore, the second algorithm continued to decrease, on the average, as m increased, as claimed by Theorem 2.

7. Conclusions

Two sequential identification procedures were presented for the identification of a time invariant, linear system in which no knowledge of the dynamics of the system were known prior to the identification, with the exception that the memory of the system was required to be finite.

The first algorithm for identification required only a mild condition on the covariance function of the input random process and no knowledge of the input or output measurement noise statistics other than that they have finite variances. It was shown that the estimate of the system converged in mean square; however, a bias developed that was due to the input noise and consequently prevented the estimate from being consistent. This error or lack of complete identification, without further knowledge of the statistics of the noises, is an example of what has been called the "structural regression paradox" in the statistical literature.

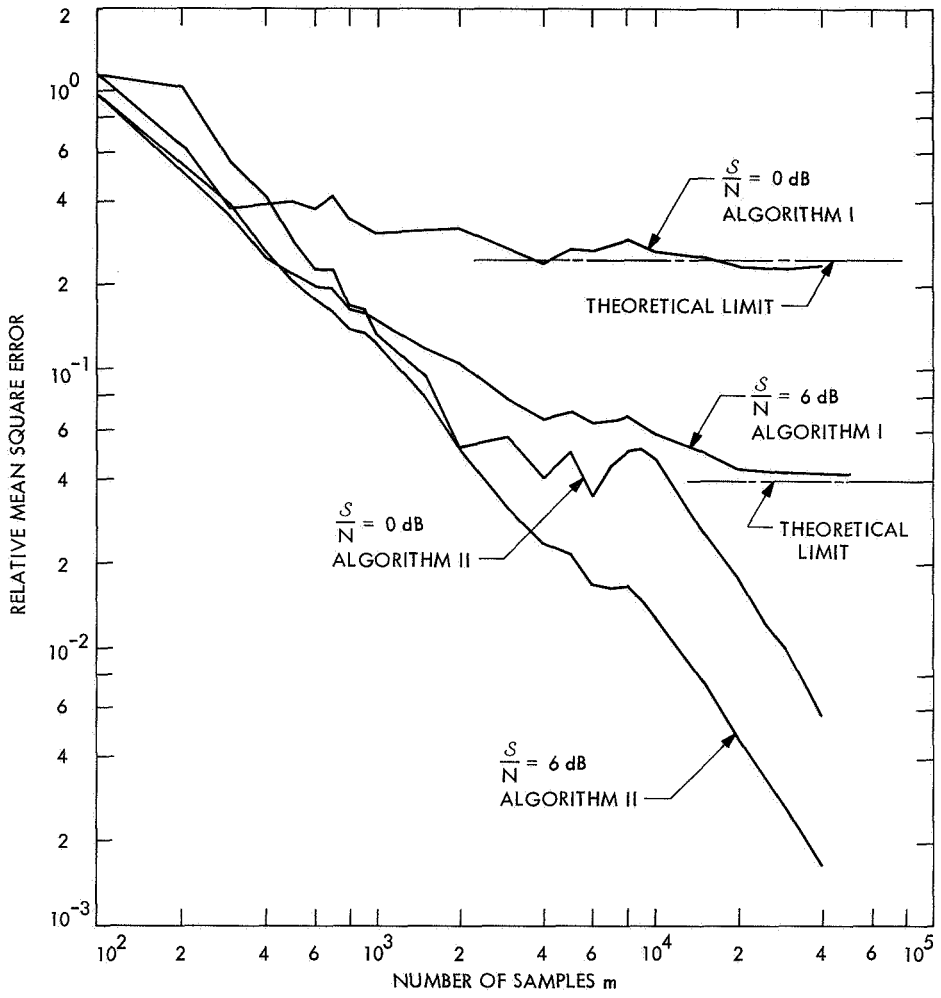


Fig. 2. Relative mean square error as a function of number of samples m and theoretical limit of error for first algorithm

The second algorithm was derived based on the additional assumption that the input measurement noise variance was known. With this additional knowledge, it was shown that the estimate of the unknown system converged in mean square to the unknown system. Hence, if the input noise variance is known, this algorithm can be used to obtain a consistent estimate of the unknown system.

Theorem 3 showed that as long as $r = \text{rank } R$ was greater than zero the first algorithm would reduce $M(\mathbf{z})$ to a minimum; however, there could be an uncountable number of values of z that achieved this minimum.

The computer simulations agreed very well with the theory, indicating that the algorithms are useful and practical methods for identification of linear systems.

References

1. Kushner, H. J., "A Simple Iterative Procedure for the Identification of the Unknown Parameters of a Linear Time Varying Discrete System," *J. Basic Eng.*, pp. 227–235, June 1963.
2. Papers presented at the IFAC Symposium on Identification in Automatic Control Systems, Prague, Czechoslovakia, 1967.
3. Saridis, G. N., and Stein, G., "Stochastic Approximation Algorithms for Linear Discrete-Time System Identification," *National Electronics Conference*, pp. 45–50, 1967.
4. Sakrison, D. J., "Application of Stochastic Approximation Methods to System Optimization," Technical Report 391. Massachusetts Institute of Technology Research Laboratory of Electronics, Cambridge, Mass., 1962.
5. Balakrishnan, A. V., "Determination of Non Linear Systems from Input-Output Data," paper presented at the 54th Meeting of the Princeton University Conference on Identification Problems in Communication and Control Systems, Princeton, N.J., 1963.

XX. Communications Systems Research: Communications Systems Development

TELECOMMUNICATIONS DIVISION

A. On Estimating the Phase of a Square Wave in White Noise, *S. Butman*

1. Introduction

Square waves are to be used in the JPL sequential ranging system for locating distant spacecraft such as *Mariner* Mars 1969 (SPS 37-53, Vol. II, Chapter III-A). The system operates by transmitting and receiving, in succession, square-wave components whose frequencies are successively halved. The first, or highest-frequency, component provides the most precise range estimate within an unknown integer multiple of the component wavelength or period. However, each succeeding component removes half of the ambiguity left by its predecessors. The process terminates when the balance of the range ambiguity becomes discernible from other considerations.

Range measurements are obtained by estimating the phase or time delay of the received noise-corrupted target return relative to a locally generated noiseless replica of the square wave. Specifically, the received signal is correlated with two square-wave replicas spaced one-quarter period apart, with analogy to the optimum estimator for the phase of a sine wave (Ref. 1). The two correlator outputs are then combined (in a nonlinear manner) to give the required phase estimate. This is the optimum method for determining the range through tracking.

The purpose here is to determine the functional form of the optimum (maximum-likelihood) processing of the outputs of the two correlators and the accuracy of the resulting estimate. One measure of accuracy is given by the signal-to-noise ratio (SNR) out of the correlators. The sum of the output SNRs is a function of the unknown phase of the received signal, ranging from a high equal to the theoretical maximum to a low of one half of the theoretically maximum SNR, or -3 dB. This amounts to an average SNR which is 1.8 dB below the theoretical maximum, where the average is taken with respect to a uniform *a priori* phase distribution between 0 and 2π. Such an *a priori* distribution is justifiable when there is no *a priori* phase information, as would be the case during acquisition. This raises the question of whether there may not be a better choice of the two correlator functions.

A general two-correlator estimation scheme is, therefore, considered from the point of view of maximizing the average SNR during acquisition. It is found that the

best two correlators for this purpose are the sine and cosine waves, even though the received signal is not sinusoidal. The sum of the SNRs is then phase-independent and is only 1.0 dB below the theoretical maximum when the received signal is a square wave. Moreover, the processing of the above correlator outputs to give the maximum-likelihood phase estimate is also independent of the structure of the ranging signal, being of the same form for all signals that it is for the sine-wave phase estimator.

2. Formulation

Let $s(t-\tau)$ denote a square wave of unit amplitude and period T that has been delayed by an amount τ, $-T/2 < \tau \leq T/2$, and observed in the presence of additive gaussian white noise $n(t)$ of one-sided spectral density N_0, in watts/hertz, as

$$z(t) = s(t-\tau) + n(t), \qquad 0 \leq t \leq MT \tag{1}$$

where MT is the length of the observation time which, for convenience, is taken to be an integral number of periods. It is assumed that $s(t-\tau)$ is present during the entire observation time, starting on or before $t = 0$ and extending to $t = MT$ or beyond. It is also assumed that the *a priori* probability density $p(\tau)$ is uniform on $(-T/2, T/2]$ and that the amplitude of the signal or, equivalently, the value of N_0 is known exactly.

3. Estimation of τ Using the Outputs of Two Square-Wave Correlators Separated by One-Quarter Period

When $z(t)$ is correlated with the locally generated square waves $s(t)$ and $s[t+(T/4)]$, the correlator outputs will be

$$x = \frac{1}{MT}\int_0^{MT} z(t)\, s(t)\, dt \tag{2}$$

$$y = \frac{1}{MT}\int_0^{MT} z(t)\, s[t+(T/4)]\, dt \tag{3}$$

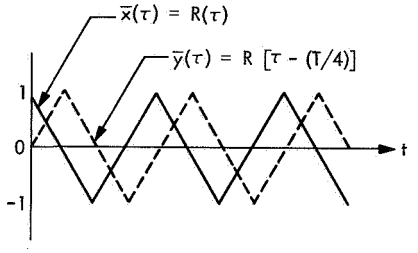
Substitution of Eq. (1) into Eqs. (2) and (3) immediately shows that

$$x = \bar{x}(\tau) + n_x \tag{4}$$

$$y = \bar{y}(\tau) + n_y \tag{5}$$

where, as shown in the following sketch, $\bar{x}(\tau) = R(\tau)$, and $\bar{y}(\tau) = R[\tau - (T/4)]$, with $R(\tau)$ being the autocorrelation function of $s(t)$ defined by

$$R(\tau) = \frac{1}{T}\int_0^T s(t)\, s(t+\tau)\, dt \tag{6}$$



Also,

$$n_x = \frac{1}{MT}\int_0^{MT} n(t)\, s(t)\, dt \tag{7}$$

$$n_y = \frac{1}{MT}\int_0^{MT} n(t)\, s[t+(T/4)]\, dt$$

are zero-mean gaussian random variables of variance $E[n_x^2] = E[n_y^2] = \sigma^2 = N_0/2MT$, where E is the expectation or averaging operator. They are statistically independent because they have zero cross covariance, $E[n_x n_y] = 0$.

In vector notation, we have $\mathbf{z} = \mathrm{col}(x, y)$, $\bar{\mathbf{z}}(\tau) = \mathrm{col}[\bar{x}(\tau), \bar{y}(\tau)]$, and $\mathbf{n} = \mathrm{col}(n_x, n_y)$, where $E[\mathbf{n}\mathbf{n}^T] = \sigma^2 \mathbf{I}$ is the covariance matrix of the noise, $\mathbf{I}$ is the two-dimensional identity matrix, and the superscript T denotes transpose. Now, $\mathbf{z}$ is conditionally normal with conditional mean $E[\mathbf{z}|\tau] = \bar{\mathbf{z}}(\tau)$ and covariance matrix $E\{[\mathbf{z} - \bar{\mathbf{z}}(\tau)][\mathbf{z} - \bar{\mathbf{z}}(\tau)]^T|\tau\} = \sigma^2 \mathbf{I}$. Consequently, the conditional probability density $p(\mathbf{z}|\tau) = p[\mathbf{z}|\bar{\mathbf{z}}(\tau]$ is

$$p(\mathbf{z}|\tau) = (2\pi\sigma^2)^{-1} \exp\left[-\frac{\|\mathbf{z} - \bar{\mathbf{z}}(\tau)\|^2}{2\sigma^2}\right] \tag{8}$$

or

$$p(x, y|\tau) = (2\pi\sigma^2)^{-1} \exp\left\{-\frac{[x - \bar{x}(\tau)]^2 + [y - \bar{y}(\tau)]^2}{2\sigma^2}\right\} \tag{9}$$

where $\|\cdot\|$ denotes the Euclidian norm.

The *a posteriori* probability density, as given by Bayes' rule, is

$$p(\tau|\mathbf{z}) = \frac{p(\mathbf{z}|\tau)}{Tp(\mathbf{z})} \tag{10}$$

where $p(\tau) = 1/T$ is the assumed *a priori* density. It is obvious from Eq. (10) that the most probable *a posteriori* estimate $\hat{\tau}$ that maximizes $p(\tau|\mathbf{z})$ over $-T/2 < \tau \leq T/2$ also maximizes $p(\mathbf{z}|\tau)$ and is, therefore, identical to the maximum-likelihood estimate, given z. However, from Eq. (8) or (9) it is clear that $p(\mathbf{z}|\tau)$ is greatest when $\|\mathbf{z} - \bar{\mathbf{z}}(\tau)\| = \{[x - \bar{x}(\tau)]^2 + [y - \bar{y}(\tau)]^2\}^{1/2}$ is least.

Geometrically, this implies that $\hat{\tau}$ must be selected such that $\hat{\mathbf{z}} = \bar{\mathbf{z}}(\hat{\tau})$ is the closest point, from the set of possible points $\bar{Z} = \mathbf{z}(\tau), \tau \in (-T/2, T/2]$, to the observed point z. To determine $\hat{\mathbf{z}}$ analytically would be difficult, since it would be necessary to minimize $\|\mathbf{z} - \bar{\mathbf{z}}\|$ over $\bar{\mathbf{z}} \in \bar{Z}$, where $\bar{Z}$ is the locus of points described parametrically by

$$\bar{x} = R(\tau) = 1 - (4|\tau|/T), \qquad |\tau| \leq T/2 \tag{11}$$

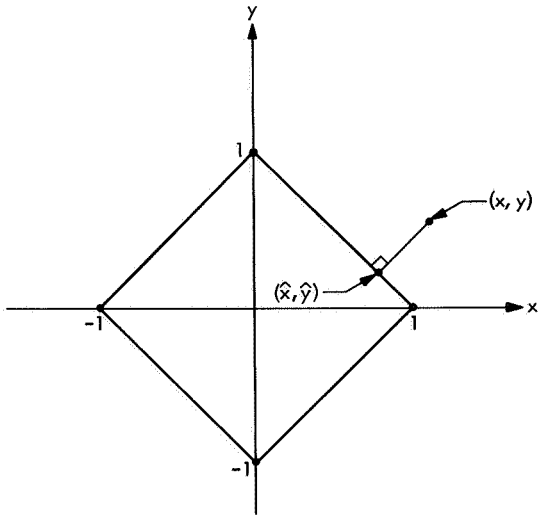
$$\bar{y} = R[\tau - (T/4)]$$

$$= \begin{cases} 4\tau/T & |\tau| \leq T/4 \\ 2\,\mathrm{sgn}\,\tau - (4|\tau|/T), & T/4 < |\tau| \leq T/2 \end{cases} \tag{12}$$

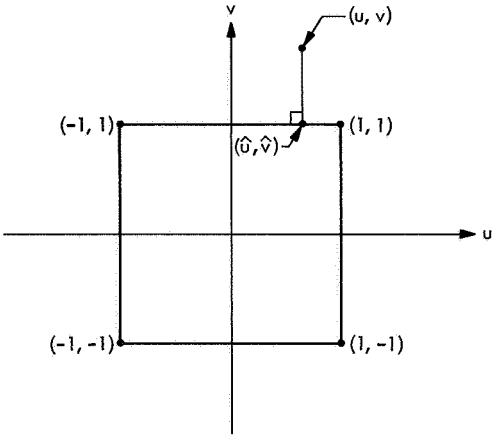
which can be combined into the simpler, but not analytic, constraint equation

$$|\bar{x}| + |\bar{y}| = 1 \tag{13}$$

Equation (13) describes the two-dimensional square of side $2^{1/2}$ shown below:



The following geometry results when the axes are rotated 45 deg using the transformation $u = 2^{-1/2}(x - y)$, $v = 2^{-1/2}(x + y)$:



In vector notation, we have $\mathbf{w} = \mathbf{U}\mathbf{z}$, where

$$\left.\begin{aligned} \mathbf{U} &= 2^{-1/2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \\ \mathbf{U}^{-1} &= 2^{-1/2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \end{aligned}\right\} \tag{14}$$

is the orthogonal matrix defining the rotation, and $\mathbf{w} = \mathrm{col}(u, v)$ denotes a vector in the new coordinates. Referring to the above sketch, it is easy to see that the point $\hat{\mathbf{w}} = \bar{\mathbf{w}}(\hat{\tau})$ that is nearest to the received point w is given by

$$\hat{u} = 2^{-1/2}\,\mathrm{sgn}\,2^{1/2}u, \qquad \hat{v} = 2^{-1/2}\,\mathrm{sat}\,2^{1/2}v, \qquad |u| > |v|$$

$$\hat{u} = 2^{-1/2}\,\mathrm{sat}\,2^{1/2}u, \qquad \hat{v} = 2^{-1/2}\,\mathrm{sgn}\,2^{1/2}v, \qquad |u| < |v|$$

where $\mathrm{sat}\,u = u$ for $|u| < 1$, $\mathrm{sat}\,u = \mathrm{sgn}\,u$ for $|u| \geq 1$, $\mathrm{sgn}\,u = 1$ for $u \geq 0$, and $\mathrm{sgn}\,u = -1$ for $u < 0$.

The region $|u| \geq |v|$ corresponds to the region $\mathrm{sgn}\,x = -\mathrm{sgn}\,y$; similarly, $|u| < |v|$ corresponds to the region $\mathrm{sgn}\,x = \mathrm{sgn}\,y$. Therefore,

$$\hat{x} = 2^{-1/2}(\hat{u} + \hat{v})$$

$$= \begin{cases} \tfrac{1}{2}(\mathrm{sgn}\,2^{1/2}u + \mathrm{sat}\,2^{1/2}v), & |u| > |v| \\ \tfrac{1}{2}(\mathrm{sgn}\,2^{1/2}v + \mathrm{sat}\,2^{1/2}u), & |u| < |v| \end{cases}$$

becomes

$$\hat{x} = \begin{cases} \tfrac{1}{2}[\operatorname{sgn}(x-y) + \operatorname{sat}(x+y)], & \operatorname{sgn} x = -\operatorname{sgn} y \\ \tfrac{1}{2}[\operatorname{sgn}(x+y) + \operatorname{sat}(x-y)], & \operatorname{sgn} x = \operatorname{sgn} y \end{cases}$$

$$= \tfrac{1}{2}[1 + \operatorname{sat}(|x| - |y|)] \operatorname{sgn} x \tag{15}$$

Next, from Eq. (11) we have $|\hat{\tau}| = (1 - \hat{x})\, T/4$, and from Eq. (12) it is evident that $\operatorname{sgn} \hat{\tau} = \operatorname{sgn} \hat{y}$. Also, $\operatorname{sgn} \hat{y} = \operatorname{sgn} y$. Therefore,

$$\frac{\hat{\tau}}{T} = \tfrac{1}{4}\{1 - \tfrac{1}{2}[1 + \operatorname{sat}(|x| - |y|)] \operatorname{sgn} x\} \operatorname{sgn} y \tag{16}$$

$$= \tfrac{1}{8}[2 \operatorname{sgn} y - \operatorname{sgn} x \operatorname{sgn} y - \operatorname{sat}(x \operatorname{sgn} y - y \operatorname{sgn} x)] \tag{17}$$

The last expression for $\hat{\tau}$ can be implemented easily in either digital or analog fashion. The pieces of analog equipment required, in addition to the correlators, are three multipliers, three adders, two hard limiters, and one soft limiter. The complete mechanization is shown in Fig. 1.

The distribution of the true value of the delay τ about the maximum-likelihood estimate $\hat{\tau}$ is given by the *a posteriori* probability density $p(\tau|\hat{\tau})$, which is related to the conditional probability density $p(\hat{\tau}|\tau)$ through Bayes' formula:

$$p(\tau|\hat{\tau}) = \frac{p(\hat{\tau}|\tau)\, p(\tau)}{\int_{-T/2}^{T/2} p(\hat{\tau}|\tau)\, p(\tau)\, d\tau} \tag{18}$$

$$= \frac{p(\hat{\tau}|\tau)}{\int_{-T/2}^{T/2} p(\hat{\tau}|\tau)\, d\tau}$$

since $p(\tau) = 1/T$. The conditional probability density $p(\hat{\tau}|\tau)$, on the other hand, determines the distribution of the maximum-likelihood estimate $\hat{\tau}$ about the true value τ. It is clear from Bayes' formula that the plot of $p(\tau|\hat{\tau})$ versus τ is of the same shape as that of $p(\hat{\tau}|\tau)$ versus τ [but not the same shape as that of $p(\hat{\tau}|\tau)$ versus $\hat{\tau}$].

It is a straightforward matter to determine $p(\hat{\tau}|\tau)$ analytically. Thus, the probability of obtaining $\hat{\tau} = 0$ is equal to the probability of decoding $\mathrm{w} = \operatorname{col}(u, v)$ as the corner $\hat{\mathrm{w}} = (2^{-1/2}, 2^{-1/2})$. This happens if, and only if, $u > 2^{-1/2}$

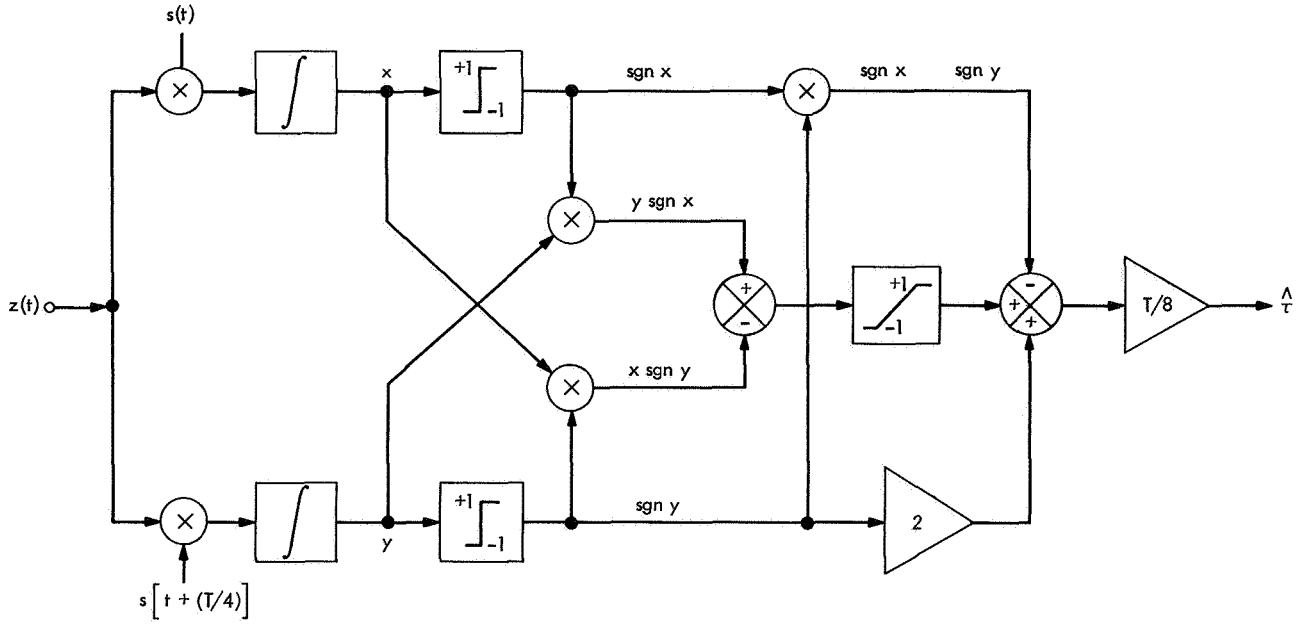


Fig. 1. Range estimator mechanization

and $v > 2^{-1/2}$; hence,

$$\begin{aligned}\Pr\{\hat{\tau}=0|\tau\} &= \Pr\{u>2^{-1/2}, v>2^{-1/2}|\hat{u}(\tau),\hat{v}(\tau)\}\\ &= \int_{2^{-1/2}}^{\infty}\int_{2^{-1/2}}^{\infty}\exp\left[-\frac{(u-\bar{u})^2+(v+\bar{v})^2}{2\sigma^2}\right]du\,dv\\ &= Q[2^{-1/2}-\bar{u}(\tau)]\,Q[2^{-1/2}-\bar{v}(\tau)]\end{aligned} \tag{19}$$

where

$$Q(a) = (2\pi\sigma^2)^{-1/2}\int_a^{\infty}\exp(-b^2/2\sigma^2)\,db \tag{20}$$

and $\Pr\{\cdot\}$ denotes probability (as opposed to probability density). In the present context, $p(\hat{\tau}=0|\tau) = \delta(\hat{\tau})\Pr\{\hat{\tau}=0|\tau\}$,

$$\begin{aligned}\bar{u}(\tau) &= 2^{-1/2}[\bar{x}(\tau)-\bar{y}(\tau)]\\ &= 2^{-1/2}\left[1-4\left|\frac{\tau}{T}\right|-\left(1-\left|1-4\left|\frac{\tau}{T}\right|\right|\right)\operatorname{sgn}\tau\right]\end{aligned} \tag{21}$$

$$\begin{aligned}\bar{v}(\tau) &= 2^{-1/2}[\bar{x}(\tau)+\bar{y}(\tau)]\\ &= 2^{-1/2}\left[1-4\left|\frac{\tau}{T}\right|+\left(1-\left|1-4\left|\frac{\tau}{T}\right|\right|\right)\operatorname{sgn}\tau\right]\end{aligned} \tag{22}$$

Similarly,

$$\Pr\{\hat{\tau}=(T/4)|\tau\} = Q[2^{-1/2}+\bar{u}(\tau)]\,Q[2^{-1/2}-\bar{v}(\tau)]$$
$$\Pr\{\hat{\tau}=(T/2)|\tau\} = Q[2^{-1/2}+\bar{u}(\tau)]\,Q[2^{-1/2}+\bar{v}(\tau)]$$
$$\Pr\{\hat{\tau}=-(T/4)|\tau\} = Q[2^{-1/2}-\bar{u}(\tau)]\,Q[2^{-1/2}+\bar{v}(\tau)]$$

Next, we determine $p(\hat{\tau}|\tau)$ for $-(T/4)<\tau<0$. In this region, $\hat{u}=2^{-1/2}$ and $|\hat{v}|<2^{-1/2}$ with $\hat{v}=2^{-1/2}[1+(8\hat{\tau}/T)]$. Therefore,

$$p(\hat{\tau}|\tau) = 2^{-1/2}\frac{8}{T}p\{\hat{u}=2^{-1/2},\hat{v}=2^{-1/2}[1+(8\hat{\tau}/T)]|\tau\}$$

However, $\hat{u}=2^{-1/2}$ and $\hat{v}=v$ if, and only if, $u>|v|=|\hat{v}|$; hence,

$$\begin{aligned}p(2^{-1/2},v) &= Q(|\hat{v}|-\bar{u})\\ &\quad\times(2\pi\sigma^2)^{-1/2}\exp[-(\hat{v}-\bar{v})^2/2\sigma^2]\end{aligned}$$

and

$$\begin{aligned}p(\hat{\tau}|\tau) &= (4\pi\sigma^2)^{-1/2}\frac{8}{T}Q[2^{-1/2}|1+(8\hat{\tau}/T)|-\bar{u}]\\ &\quad\times\exp\left[-\frac{1-(8\hat{\tau}/T)-2^{1/2}\bar{v}}{4\sigma^2}\right]\end{aligned}$$

A similar procedure is used for the remaining regions. The end result is

$$\begin{aligned}p(\hat{\tau}|\tau) &= \delta(\hat{\tau})\,Q[2^{-1/2}-\bar{u}(\tau)]\,Q[2^{-1/2}-\bar{v}(\tau)] + \delta[\hat{\tau}-(T/4)]\,Q[2^{-1/2}+\bar{u}(\tau)]\,Q[2^{-1/2}-\bar{v}(\tau)]\\ &\quad+\delta[\hat{\tau}+(T/4)]\,Q[2^{-1/2}-\bar{u}(\tau)]\,Q[2^{-1/2}+\bar{v}(\tau)] + \delta[\hat{\tau}-(T/2)]\,Q[2^{-1/2}+\bar{u}(\tau)]\,Q[2^{-1/2}+\bar{v}(\tau)]\\ &\quad+(4\pi\sigma^2)^{-1/2}\frac{8}{T}\begin{cases}Q\left[2^{-1/2}\left|1+\frac{8\hat{\tau}}{T}\right|-\bar{u}(\tau)\right]\exp\left\{-\frac{[1+(8\hat{\tau}/T)-2^{1/2}\bar{v}(\tau)]^2}{4\sigma^2}\right., & \hat{\tau}\in\left(-\frac{T}{4},0\right)\\ Q\left[2^{-1/2}\left|1+\frac{8\hat{\tau}}{T}\right|-\bar{v}(\tau)\right]\exp\left\{-\frac{[1-(8\hat{\tau}/T)-2^{1/2}\bar{u}(\tau)]^2}{4\sigma^2}\right., & \hat{\tau}\in\left(0,\frac{T}{4}\right)\\ Q\left[2^{-1/2}\left|3-\frac{8\hat{\tau}}{T}\right|+\bar{u}(\tau)\right]\exp\left\{-\frac{[3-8\hat{\tau}/T)-2^{1/2}\bar{v}(\tau)]^2}{4\sigma^2}\right., & \hat{\tau}\in\left(\frac{T}{4},\frac{T}{2}\right)\\ Q\left[2^{-1/2}\left|3-\frac{8\hat{\tau}}{T}\right|+\bar{v}(\tau)\right]\exp\left\{-\frac{[3+(8\hat{\tau}/T)-2^{1/2}\bar{u}(\tau)]^2}{4\sigma^2}\right., & \hat{\tau}\in\left(-\frac{T}{2},-\frac{T}{4}\right)\end{cases}\end{aligned} \tag{23}$$

The conditional probability density $p(\hat{\tau}|\tau)$ is plotted in Fig. 2 for several values of σ and $\tau \in [0, T/8]$. The plots for $\tau \in [T/8, T/4]$ are then obtained by reflecting the original set of graphs about the axis $\hat{\tau} = T/8$. Similarly, $p(\hat{\tau}|\tau)$ for $\tau \in [T/4, 3T/8]$ is a reflection about $\hat{\tau} = T/4$ of the plot of $p(\hat{\tau}|\tau)$ for $\tau \in [T/8, T/4]$, etc.

The *a posteriori* probability density $p(\tau|\hat{\tau})$ is plotted in Fig. 3 and satisfies the same conditions of symmetry as $p(\hat{\tau}|\tau)$. However, it should be observed that $p(\tau|\hat{\tau})$ has no delta functions even though $p(\hat{\tau}|\tau)$ does. Since $p(\hat{\tau})$ has delta functions at the same places as $p(\hat{\tau}|\tau)$, the delta functions cancel in $p(\tau|\hat{\tau}) = p(\hat{\tau}|\tau)/Tp(\hat{\tau})$.

4. The General Two-Correlator Problem

The preceding discussion was concerned with making the best estimate, given the outputs x and y of the two orthogonal square-wave correlators. It is logical to inquire now whether there may not exist a better choice of correlators (assuming, of course, that the correlator outputs can always be processed in an optimum manner).

One measure of the performance of a correlator is the SNR:

$$\rho_x(\tau) = [\bar{x}(\tau)/\sigma]^2 \tag{24}$$

$$\rho_y(\tau) = [\bar{y}(\tau)/\sigma]^2 \tag{25}$$

The sum of the two SNRs is then

$$\rho_z(\tau) = [\|\bar{\mathbf{z}}(\tau)\|/\sigma]^2 \tag{26}$$

For the square-wave correlators, $\bar{\mathbf{z}}(\tau)$ is on the square of side $2^{1/2}$ and $\rho_z(\tau)$ varies from a maximum of $1/\sigma^2$ to a minimum of $1/2\sigma^2$ with periodicity $T/4$. The average value of $\rho_z(\tau)$ when τ is uniform on $(-T/2, T/2]$ is, therefore,

$$\bar{\rho}_z = \frac{1}{T}\int_{-T/2}^{T/2} \rho_z(\tau)\, d\tau \tag{27}$$

$$= 2/3\sigma^2 \tag{28}$$

This average value would be obtained during acquisition, when there is no knowledge of τ other than that it is equi-probable on $(-T/2, T/2]$. However, once an estimate $\hat{\tau}$ has been obtained, the receiver can readjust the local zero reference to $\hat{\tau}$ and obtain $\rho_z(\tau - \hat{\tau})$ with an *a posteriori* average of

$$\bar{\rho}_z(\tau) = \int_{\hat{\tau}-(T/2)}^{\hat{\tau}+(T/2)} \rho_z(\tau - \hat{\tau})\, \rho(\tau|\hat{\tau})\, d\tau \tag{29}$$

The continual updating of the local zero reference to the latest estimate $\hat{\tau}$ is known as *tracking* and can be implemented with a phase-locked loop. As the estimate $\hat{\tau}$ improves, $p(\tau|\hat{\tau})$ approaches $\delta(\tau - \hat{\tau})$ and

$$\bar{\rho}_z(\hat{\tau}) \to \rho_z(0) = 1/\sigma^2$$

which is the theoretical maximum. Therefore, the use of square-wave correlators is optimum during tracking. However, it is still necessary to determine the best two correlators for acquisition purposes.

5. Optimum Correlators for Acquisition

Suppose that $r(t)$ is correlated with some pair of periodic, orthonormal, but otherwise arbitrary, time functions $f(t)$ and $h(t)$. Then, the correlator outputs are

$$f = \frac{1}{MT}\int_0^{MT} r(t) f(t)\, dt \tag{30}$$

$$= \frac{1}{MT}\int_0^{MT} s(t-\tau) f(t) + \frac{1}{T}\int_0^{T} n(t) f(t)$$

or

$$f = \phi_{fs}(\tau) + n_f \tag{31}$$

and

$$h = \phi_{hs}(\tau) + n_h \tag{32}$$

where $\phi_{fs}(\tau)$ and $\phi_{hs}(\tau)$ are cross-correlation functions, while n_f and n_h are independent zero-mean gaussian random variables of variance σ^2. Consequently,

$$\begin{aligned} \rho_f(\tau) &= \phi_{fs}^2(\tau)/\sigma^2 \\ \rho_h(\tau) &= \phi_{hs}^2(\tau)/\sigma^2 \\ \rho &= \rho_f + \rho_h \end{aligned} \tag{33}$$

$$\bar{\rho} = \frac{1}{T}\int_{-T/2}^{T/2} \rho(\tau)\, d\tau$$

$$= \bar{\rho}_f + \bar{\rho}_h \tag{34}$$

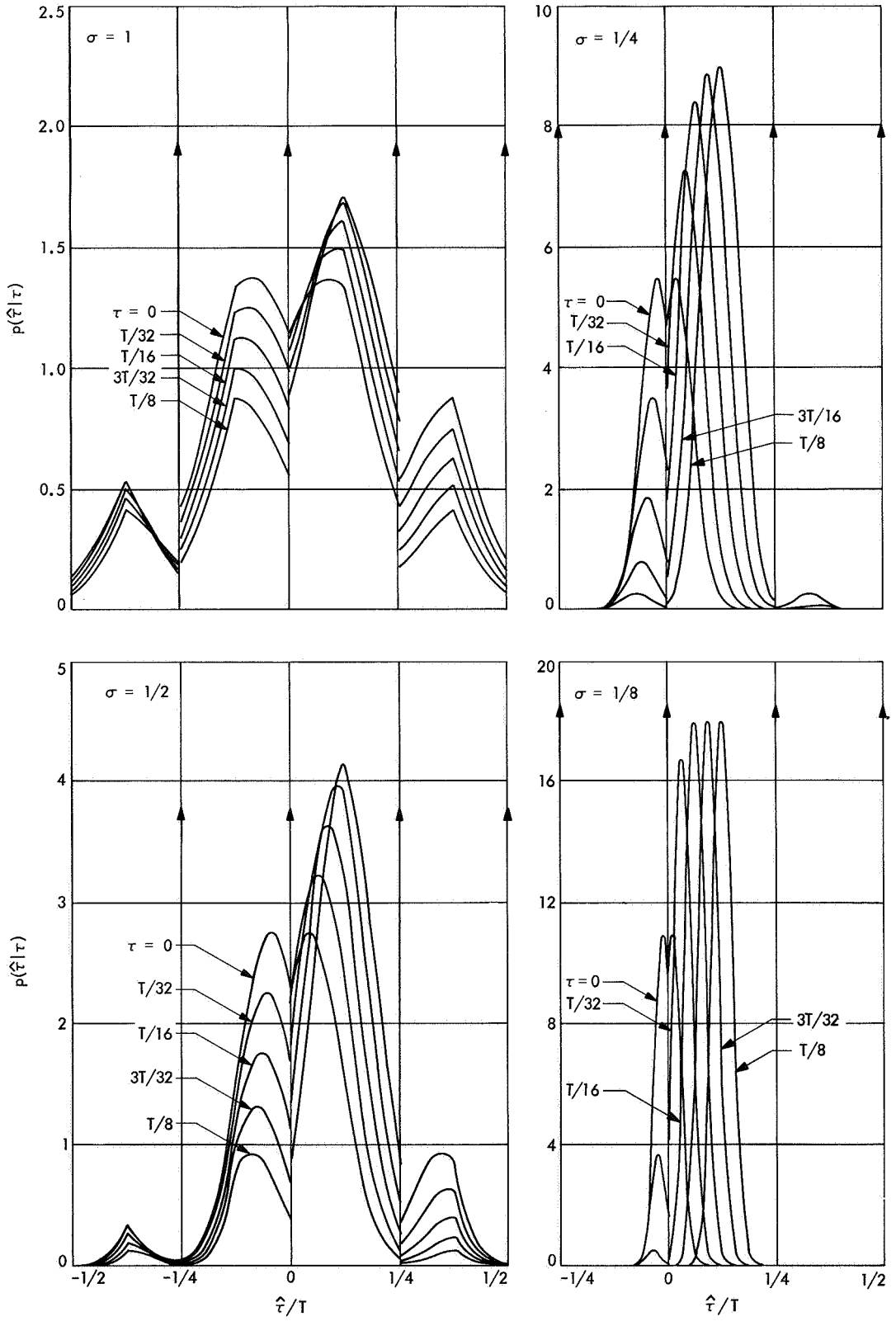


Fig. 2. Conditional probability density $p(\hat{\tau}|\tau)$ vs $\hat{\tau}/T$ for various values of σ

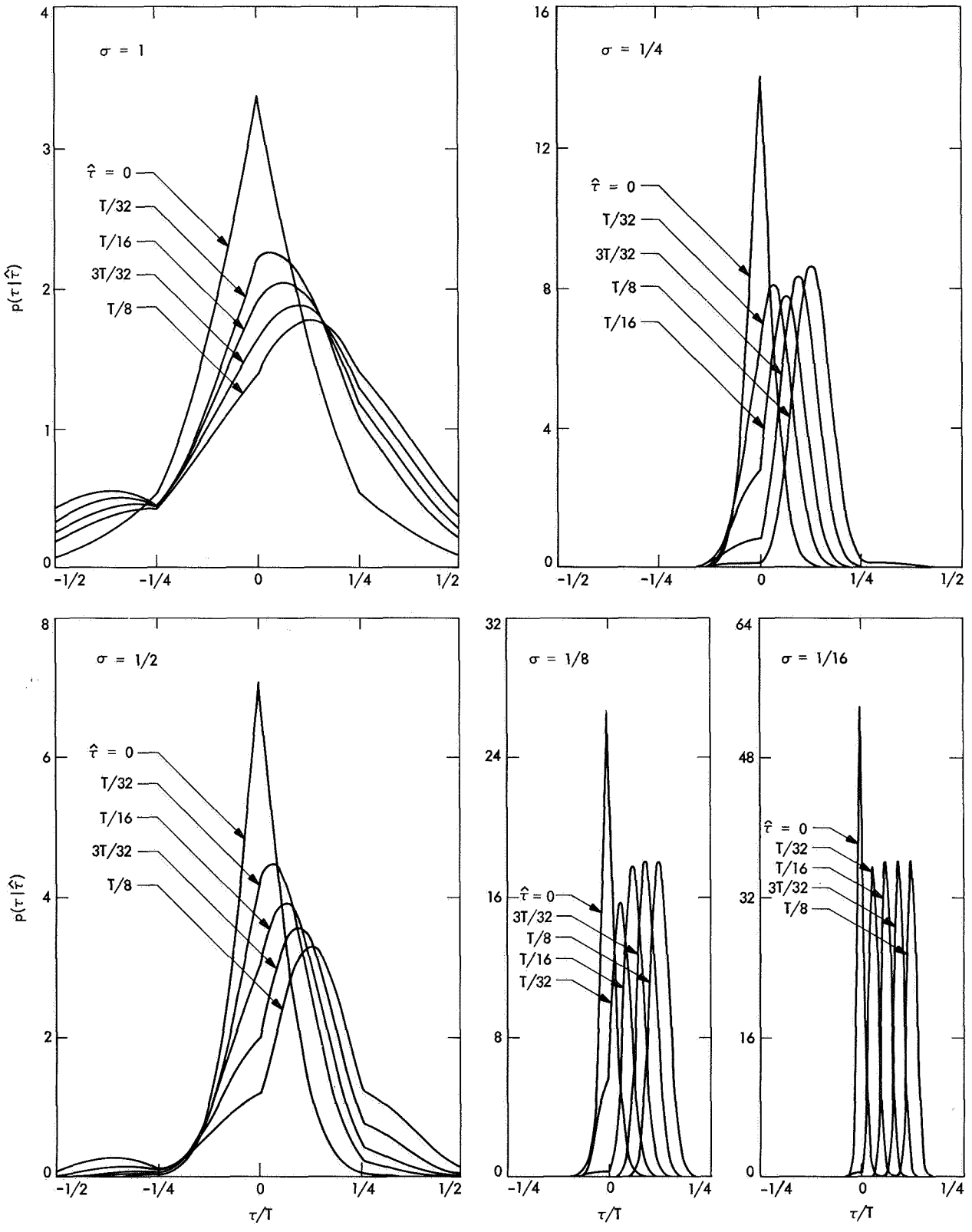


Fig. 3. *A posteriori* probability density $p(\tau|\hat{\tau})$ vs $\hat{\tau}/T$ for various values of σ

Now,

$$\bar{\rho}_f = \frac{1}{\sigma^2 M^2 T^3}\int_{-T/2}^{T/2}\int_0^{MT}\int_0^{MT} f(t)\,s(t-\tau)\,s(u-\tau)\,f(u)\,dt\,du\,d\tau$$

$$= \frac{1}{\sigma^2}\left(\frac{1}{MT}\right)^2\int_0^{MT}\int_0^{MT} f(t)\,R(t-u)\,f(u)\,du\,dt \tag{35}$$

Similarly,

$$\bar{\rho}_h = \frac{1}{\sigma^2}\left(\frac{1}{MT}\right)^2\int_0^{MT}\int_0^{MT} h(t)\,R(t-u)\,h(u)\,dt\,du \tag{36}$$

However, $R(t)$ is an even periodic function of period T with a positive spectrum:

$$R(t-u) = \sum_k s_k^2 \cos\frac{2\pi k}{T}(t-u) \tag{37}$$

Also, $f(t)$ and $h(t)$ are periodic functions of the general form

$$f(t) = \sum_{k=1}^{\infty}\left(a_k \sin\frac{2\pi k}{T}t + d_k\cos\frac{2\pi k}{T}t\right) \tag{38}$$

$$h(t) = \sum_{k=1}^{\infty}\left(c_k \sin\frac{2\pi k}{T}t + d_k\cos\frac{2\pi k}{T}t\right) \tag{39}$$

with

$$\frac{1}{MT}\int_0^{MT} f^2(t)\,dt = \sum_{k=1}^{\infty}(a_k^2 + b_k^2) = 1 \tag{40}$$

$$\frac{1}{MT}\int_0^{MT} h^2(t)\,dt = \sum_{k=1}^{\infty}(c_k^2 + d_k^2) = 1 \tag{41}$$

Therefore,

$$\bar{\rho}_f = \frac{1}{\sigma^2}\sum_k s_k^2 \frac{1}{MT}\int_0^{MT}\int_0^{MT} f(u)\,f(t)\cos 2\pi k(t-u)\,dt\,du$$

$$= \frac{1}{2\sigma^2}\sum_k s_k^2(a_k^2 + b_k^2)$$

$$\leq \frac{1}{2\sigma^2}s_{\max}^2 \tag{42}$$

with equality if, and only if,

$$f(t) = a\sin\frac{2\pi m}{T}t + b\cos\frac{2\pi m}{T}t$$

where m is the subscript of $s_{\max}$. Consequently, the optimum choices for $f(t)$ and $h(t)$ are

$$f(t) = 2^{1/2}\cos\left(\frac{2\pi m}{T}t + \theta\right) \tag{43}$$

$$h(t) = 2^{1/2}\sin\left(\frac{2\pi m}{T}t + \theta\right) \tag{44}$$

where θ is arbitrary and $\bar{\rho} = s_m^2/\sigma^2$. In addition, $\rho_f(\tau)$ is now independent of τ; hence,

$$\rho(\tau) = \bar{\rho} = s_m^2/\sigma^2 \tag{45}$$

In the case of a square wave, $s_m^2 = s_1^2 = 8/\pi^2$, and we obtain

$$\rho(\tau) = \bar{\rho} = 8/\pi^2\sigma^2 \tag{46}$$

which shows that the theoretical maximum of $1/\sigma^2$ is unachievable in the square-wave case.

For this choice of correlators, the maximum likelihood-estimate is well-known to be (Ref. 1)

$$\hat{\tau} = \frac{T}{2\pi}\arctan(y/x) + \frac{T}{2}(1 - \operatorname{sgn} x) \tag{47}$$

where

$$x = \frac{1}{MT}\int_0^{MT} r(t)\cos(2\pi t/T)\,dt$$

$$y = \frac{1}{MT}\int_0^{MT} r(t)\sin(2\pi t/T)\,dt$$

are the correlator outputs.

6. Mean-Square Error

The probability densities $p(\tau|\hat{\tau})$ and $p(\hat{\tau}|\tau)$ contain all of the statistical information about the performance of the estimator. Of particular interest, however, is the *a posteriori* mean-square error

$$E\left[\left(\frac{\tau-\hat{\tau}}{T}\right)^2\middle|\hat{\tau}\right]=\int_{\hat{\tau}-(T/2)}^{\hat{\tau}+(T/2)}\left(\frac{\tau-\hat{\tau}}{T}\right)^2 p(\tau|\hat{\tau})\,d\tau$$

since it gives the scatter of the true value of the delay τ about the maximum-likelihood estimate $\hat{\tau}$. It is easy to show that this *a posteriori* mean-square error approaches $\sigma^2/32$ as σ^2 goes to zero, provided $\hat{\tau}\neq\pm kT/4$, $k=0,1,2,3$. This is because $p(\tau|\hat{\tau})$ tends to a gaussian density of mean $\hat{\tau}/T$ and variance $\sigma^2/32$. At $\hat{\tau}=\pm kT/4$, however, $p(\tau|\hat{\tau})$ is proportional to $Q(4\cdot 2^{1/2}|\tau-\hat{\tau}|/T)$. Thus, if $\sigma^2\to 0$ and $x=(\tau-\hat{\tau})/T$, we can write

$$E\left[\left(\frac{\tau-\hat{\tau}}{T}\right)^2\middle|\hat{\tau}=\pm kT/4\right]=\frac{\int_{-\infty}^{\infty}x^2\,Q(4\cdot 2^{1/2}|x|)\,dx}{\int_{-\infty}^{\infty}Q(4\cdot 2^{1/2}|x|)\,dx}$$

$$=\frac{\int_{0}^{\infty}x^2\,Q(4\cdot 2^{1/2}x)\,dx}{\int_{0}^{\infty}Q(4\cdot 2^{1/2}x)\,dx}$$

Integrating by parts and noting that

$$Q'(4\cdot 2^{1/2}x)=(2\pi\sigma^2)^{-1/2}\exp(-32x^2/\sigma^2)$$

we obtain

$$E\left[\left(\frac{\tau-\hat{\tau}}{T}\right)^2\middle|\hat{\tau}=\pm kT/4\right]=\frac{2}{3}\frac{\int_{0}^{\infty}x^3\exp(-32x^2/\sigma^2)\,dx}{\int_{0}^{\infty}x\exp(-32x^2)\,dx}$$

$$=\sigma^2/48$$

The above analytical results have also been verified numerically on a general-purpose digital computer. Numerical values of the *a posteriori* mean-square error in units of $\sigma^2/32$ are tabulated in Table 1 for $\hat{\tau}=kT/32$, $k=0,\cdots,4$, and $\sigma=2^{-k}$, $k=0,\cdots,4$.

Reference

1. Viterbi, A. J., *Principles of Coherent Communication*, pp. 129. McGraw Hill Book Company, Inc., New York, 1966.

Table 1. *A posteriori* mean-square errors

$\hat{\tau}/T$	*A posteriori* mean-square error for indicated σ				
	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$
0	0.865	0.703	0.667	0.667	0.667
1/32	1.300	1.250	1.050	0.905	0.910
1/16	1.430	1.600	0.910	0.910	0.995
3/32	1.560	2.300	0.900	0.995	0.995
1/8	1.390	3.700	2.000	1.000	1.000

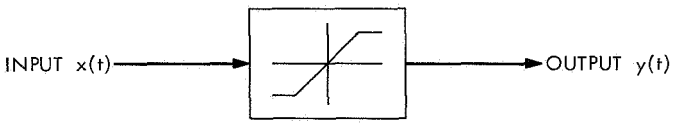
B. Analysis of Narrow-Band Signals Through the Band-Pass Soft Limiter, *R. C. Tausworthe*

1. Introduction

Several authors (Refs. 1–3) have examined the output SNR characteristics of the so-called "soft" limiter, giving several approximations for the output signal and noise terms as functions of the input parameters. The ensuing article illustrates that, under a widely accepted model of the soft limiter, the output signal power and signal suppression can be found exactly in terms of the hard-limiter signal-suppression function. The output noise is correspondingly then well approximated.

2. Limiter Suppression Factor

In the discussion below, we shall assume that the following device input



is a narrow-band waveform consisting of a signal immersed in gaussian noise of variance σ_N^2:

$$x(t)=V(t)\sin[\omega_0 t+\theta(t)]+n(t)$$
$$=V\sin\phi+n(t) \qquad (1)$$

where

$$V=V(t)$$
$$\phi=\omega_0 t+\theta(t)$$

It has been shown (SPS 37-44, Vol. IV, pp. 303–307) that the portion of the limiter output due to input signal is

$$G(V\sin\phi)=E[y(x)|V\sin\phi]$$
$$=c_1\sin\phi+c_2\sin 2\phi+\cdots \qquad (2)$$

in which the coefficient

$$c_k = \frac{1}{\pi}\int_{-\pi}^{\pi} G(V\sin\phi)\sin k\phi\, d\phi \tag{3}$$

represents the amplitude of the signal in the kth harmonic zone. For the hard limiter, the c_k have been evaluated as

$$c_k = L\left(\frac{2}{\pi}\right)^{1/2}\left(\frac{v}{k\pi}\right)\int_{-\pi}^{\pi}\cos\phi\cos k\phi\exp\left\{-\frac{1}{2}v^2\sin^2\phi\right\}d\phi$$

$$= L\left(\frac{2}{\pi}\right)^{1/2}\left(\frac{v}{k}\right)\exp\frac{-v^2}{4}\left[I_{(k-1)/2}\left(\frac{v^2}{4}\right)+I_{(k+1)/2}\left(\frac{v^2}{4}\right)\right] \tag{4}$$

for odd k, in terms of the parameter $v = V(t)/\sigma_N$ and the modified Bessel functions of the first kind (Ref. 4).

With an input SNR of ρ the *hard-limiter suppression factor* $\alpha^2(\rho)$ is defined as the ratio of the fundamental signal output power to what it would be if noise were absent. When $V(t)$ is a constant amplitude,

$$\rho = \frac{1}{2}v^2,$$

so

$$\alpha^2(\rho) = \frac{\pi}{4}\rho e^{-\rho}\left[I_0\left(\frac{\rho}{2}\right)+I_1\left(\frac{\rho}{2}\right)\right]^2 \tag{5}$$

An excellent approximation for α^2 (Ref. 5) is

$$\alpha^2(\rho) = \frac{0.7854\rho + 0.4768\rho^2}{1 + 1.024\rho + 0.4768\rho^2} \tag{6}$$

If, however, $V(t)$ is time varying, then the input SNR is

$$\rho = E\left(\frac{1}{2}v^2\right)$$

and

$$\alpha^2(\rho) = E\left[\alpha^2\left(\frac{1}{2}v^2\right)\right] \tag{7}$$

Suppression is probably computed with least difficulty in this case through the approximation given in Eq. (6).

3. Soft Limiter Model

We shall take as the model of the soft limiter the function plotted in Fig. 4:

$$y = L\,\mathrm{erf}\left[\left(\frac{K\pi^{1/2}}{2L}\right)x\right] = L\,\mathrm{erf}(Bx) \tag{8}$$

where $\mathrm{erf}(x)$ is the well-known error function (Ref. 4)

$$\mathrm{erf}\,x = \frac{2}{\pi^{1/2}}\int_0^x \exp(-t^2)\,dt \tag{9}$$

and $B = K\pi^{1/2}/2L$.

Our model is thus seen to possess the following characteristics: For values of x much less than $2L/K\pi^{1/2}$, the device acts as a linear amplifier with voltage gain K. For inputs x much larger than $2L/K\pi^{1/2}$, signal limiting occurs, with the limit level L. Further, as $K\to\infty$ for fixed L, the device becomes a hard limiter, and as $L\to\infty$ for fixed K, the device becomes a linear amplifier. The soft limiter model we have chosen thus degenerates to previously analyzed devices in limiting cases.

Evaluation of the limiter performance thus now depends only upon finding $G(V\sin\phi)$ and its Fourier coefficients for the assumed characteristic. In the present case $G(V\sin\phi)$ takes the form

$$G(V\sin\phi) = \frac{L}{\sigma_N(2\pi)^{1/2}}\int_{-\infty}^{+\infty}\mathrm{erf}\,B(V\sin\phi+n)$$

$$\times\exp\left\{\frac{-n^2}{2\sigma_N^2}\right\}dn \tag{10}$$

Although the results to follow are quite general, we shall evaluate only the behavior in the fundamental output zone:

$$c_1 = \left(\frac{L}{\pi}\right)\frac{1}{\sigma_N(2\pi)^{1/2}}\int_{-\infty}^{+\infty}\exp\left\{\frac{-n^2}{2\sigma_N^2}\right\}dn$$

$$\times\int_{-\pi}^{\pi}\sin\phi\,\mathrm{erf}\,[B(V\sin\phi+n)]\,d\phi \tag{11}$$

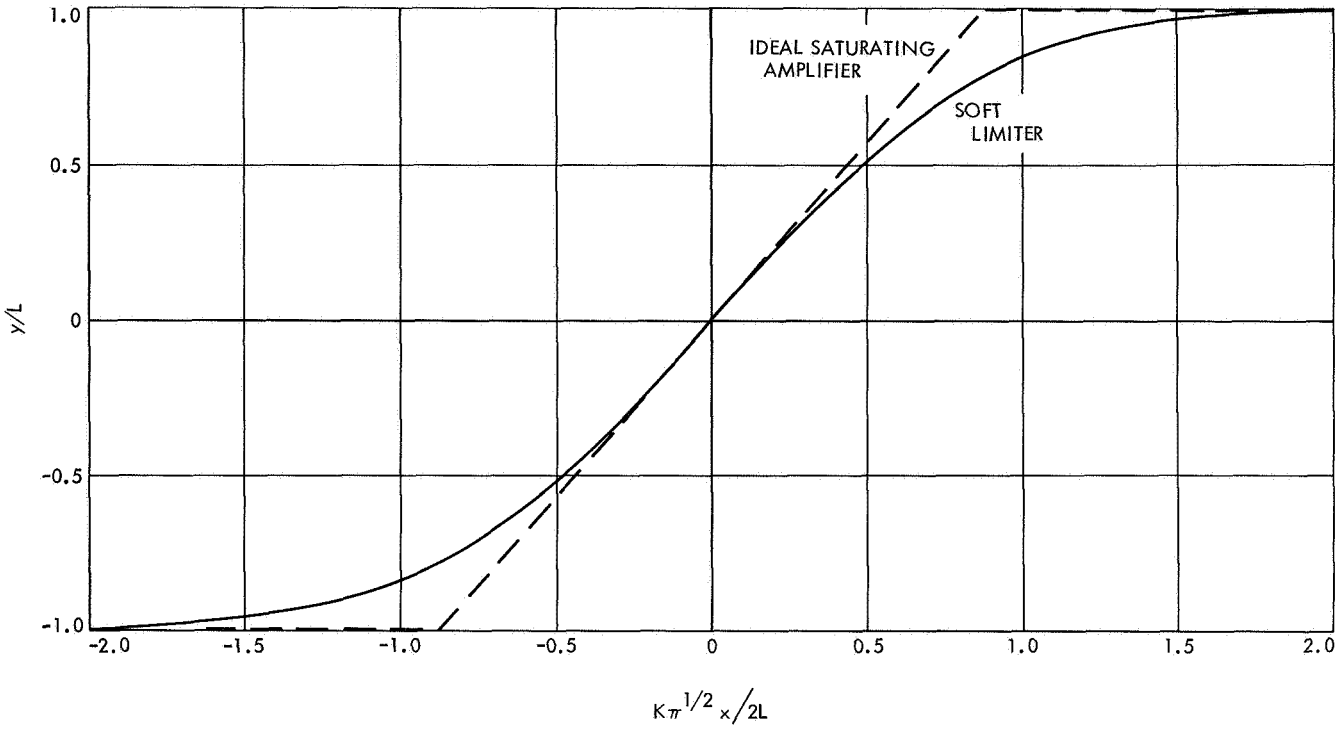


Fig. 4. Soft limiter model characteristics

The inner integral can be integrated by parts to give

$$F = \frac{2BV}{\pi^{1/2}} \int_{-\pi}^{\pi} \cos^2 \phi \exp\{-B^2 (V \sin\phi + n)^2\}\, d\phi \tag{12}$$

which, when inserted back into the expression for c, produces the relation

$$c_1 = \frac{2^{1/2} BLV}{\sigma_N \pi^2} \int_{-\pi}^{\pi} \cos^2\phi \int_{-\infty}^{+\infty} \exp\left\{-\left[\left(B^2 + \frac{1}{2\sigma_N^2}\right) n^2 + 2B^2 V n \sin\phi + B^2 V^2 \sin^2\phi\right]\right\} dn\, d\phi \tag{13}$$

The inner integral is tabulated (Ref. 4):

$$\int_{-\infty}^{+\infty} \exp\{-(at^2 + 2bt + c)\}\, dt = \left(\frac{\pi}{a}\right)^{1/2} \exp\left(\frac{b^2 - ac}{a}\right) \tag{14}$$

Mere substitution thus provides

$$c_1 = L\left(\frac{2}{\pi}\right)^{1/2} \frac{v}{\pi} \int_{-\infty}^{+\infty} \cos^2\phi \exp\left\{-\frac{1}{2} v^2 \sin^2\phi\right\} d\phi \tag{15}$$

in terms of the parameter ratio

$$v^2 = \frac{2B^2V^2}{1 + 2B^2\sigma_N^2} = \frac{V^2}{\sigma_N^2 + \frac{2}{\pi}\left(\frac{L}{K}\right)^2} \tag{16}$$

But the form of c_1 is now recognized to involve the same integral as that of the hard limiter, except with a different v. As a consequence, the results for a soft limiter are expressible in terms of the hard limiter suppression factor α^2. For example, the device output power P_s, considering $V(t) = V$ as a constant, is

$$P_s = \frac{8}{\pi^2} L^2 \alpha^2 \left[\frac{\rho}{1 + \left(\frac{2L}{\pi^{1/2} VK}\right)^2 \rho}\right] \tag{17}$$

and, if $V(t)$ is time varying, P_s is

$$P_s = \frac{8}{\pi^2} L^2 E\left\{\alpha^2 \left[\frac{\frac{V^2(t)}{2\sigma_N^2}}{1 + \frac{2}{\pi}\left(\frac{L}{K\sigma_N}\right)^2}\right]\right\} \tag{18}$$

Here again we can define a signal suppression factor α_s^2 as the ratio of signal output powers with and without

noise. Because of the last equation above, we see this can be written as

$$\alpha_s^2 = \frac{\alpha^2\left[\dfrac{\rho}{1+\left(\dfrac{2L}{\pi^{1/2}VK}\right)^2\rho}\right]}{\alpha^2\left[\left(\dfrac{\pi^{1/2}VK}{2L}\right)^2\right]} \qquad (19)$$

in the simpler, constant-V case. This function appears plotted in Fig. 5 for various values of VK/L.

Note in the limiting case

$$\left(\frac{VK}{L}\right) \to \infty \qquad \text{(approaching a hard limiter)}$$

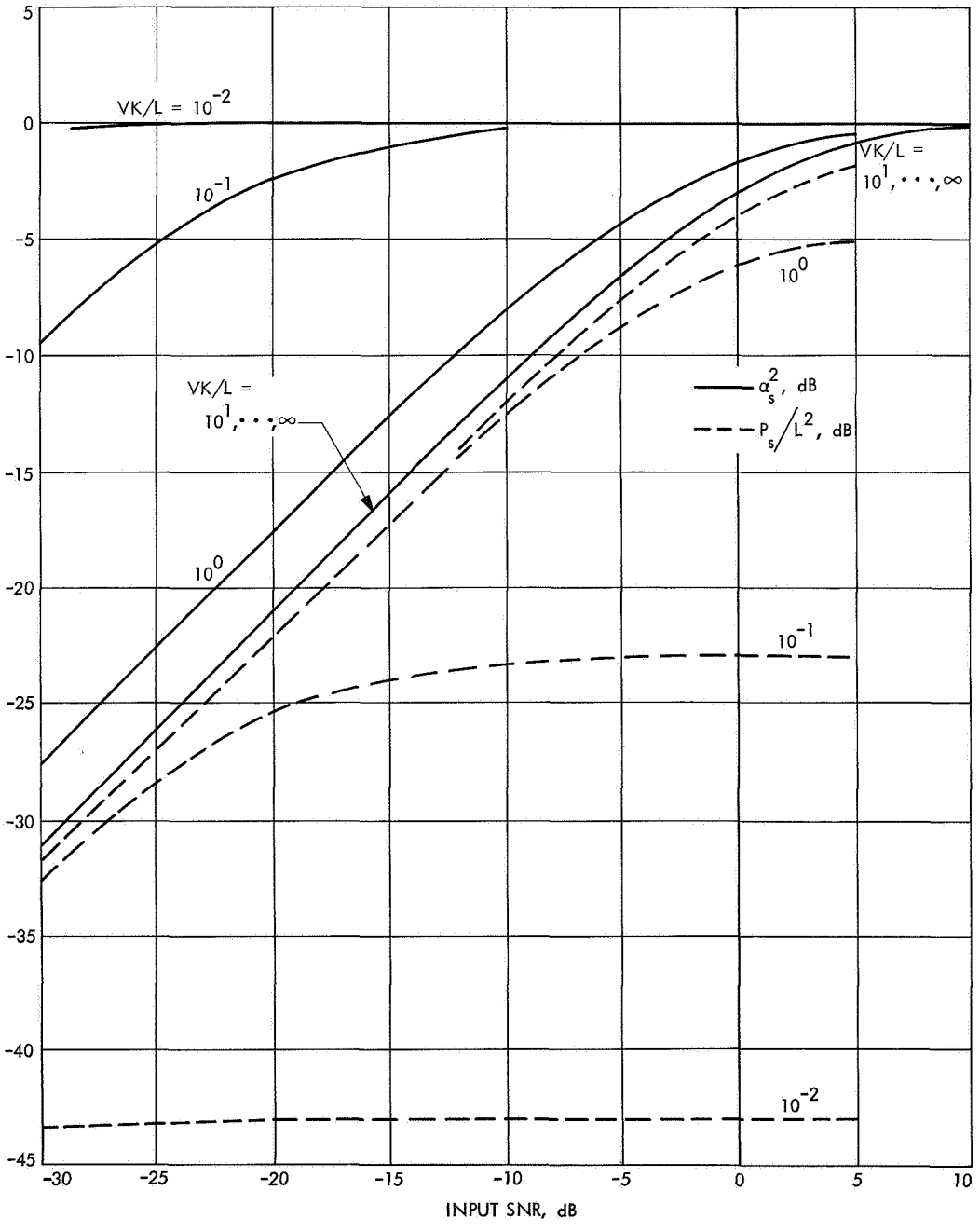


Fig. 5. Suppression factor α_s^2 and normalized output signal power P_s/L^2 as functions of input SNR

that

$$\alpha_s^2 \to \alpha^2 \tag{20}$$

as it should, and as

$$\left(\frac{VK}{L}\right) \to 0 \qquad \text{(approaching a linear amplifier)}$$

for a fixed K, that

$$P_s \to E\left[\frac{K^2V^2(t)}{2}\right] = K^2P_{\text{sig}} \tag{21}$$

as it should.

4. Noise Output Power

In the linear region, the output SNR equals that input to the device

$$\rho_s \to \rho \qquad \text{(linear region)} \tag{22}$$

whereas, when severe clipping is taking place,

$$\rho_s \to \rho_l = \frac{\alpha^2(\rho)}{1-\alpha^2(\rho)} \qquad \text{(limiting region)} \tag{23}$$

(considering now only the constant-V case). The cross-over between these two conditions begins at the point when the input begins to saturate.

Considering that the noise may be decomposed into independent in-phase and quadrature-phase terms

$$n(t) = n_c \cos\phi + n_s \sin\phi \tag{24}$$

in which $\sigma_c^2 = \sigma_s^2 = \sigma_N^2$, then it is immediate that $x(t)$ takes the form

$$x(t) = V_{eq}(t) \sin\phi_{eq} \tag{25}$$

with the amplitude function

$$V_{eq}^2(t) = (V + n_s)^2 + n_c^2 \tag{26}$$

The amplitude of the output fundamental term is

$$a_1 = \frac{1}{\pi}\int_{\pi}^{\pi} \text{erf}\,(BV_{eq}\sin\phi)\sin\phi\, d\phi \tag{27}$$

which can be integrated by parts to produce

$$a_1 = \left(\frac{2}{\pi}\right)^{1/2}(2^{1/2}BV_{eq})\exp\left(\frac{-B^2V_{eq}^2}{2}\right) \times\left[I_0\left(\frac{B^2V_{eq}}{2}\right) + I_1\left(\frac{B^2V_{eq}}{2}\right)\right] \tag{28}$$

This expression is the same as c_1 for the signal output portion only, except for the substitution $v = 2^{1/2}BV_{eq}$. The total limiter output power is

$$P_{s+n} = \frac{1}{2}E(a_1^2) = \frac{8}{\pi^2}E\left[\alpha^2(B^2V_{eq}^2)\right] \tag{29}$$

Asymptotically for very large and very small values of $B^2V_{eq}^2$, the value of P_{s+n} behaves as

$$P_{s+n} \sim \frac{8}{\pi^2}L^2\alpha^2\left[B^2E(V_{eq}^2)\right] = \frac{8L^2}{\pi^2}\alpha^2\left[\left(\frac{\pi^{1/2}KV}{2L}\right)^2\frac{(1+\rho)}{\rho}\right] \tag{30}$$

Thus an asymptotically correct approximate expression for the output SNR of the device is

$$\rho_l = \frac{\alpha^2\left[\dfrac{\rho}{1+\left(\dfrac{2L}{\pi^{1/2}VK}\right)^2\rho}\right]}{\alpha^2\left[\dfrac{(1+\rho)}{\left(\dfrac{2L}{\pi^{1/2}VK}\right)^2\rho}\right] - \alpha^2\left[\dfrac{\rho}{1+\left(\dfrac{2L}{\pi^{1/2}VK}\right)^2\rho}\right]} \tag{31}$$

Finally, of interest is the ratio Γ_s of the input and output signal-to-noise densities at the fundamental frequency; this function is needed when the limiter output filter is considerably narrower than the bandwidth of the input process. It is clear that in the linear region, the SNR is preserved so that

$$\frac{N_l}{P_s} = \frac{N_0}{P_{\text{sig}}} \qquad \text{(linear region)} \tag{32}$$

i.e., $\Gamma_s = 1$. At the other extreme, it has been shown (Ref. 5) that

$$\frac{N_0}{P_s} = \frac{\Gamma(\rho)N_0}{P_{\text{sig}}} \qquad \text{(limiting region)} \tag{33}$$

where $\Gamma(\rho)$ is approximately

$$\Gamma(\rho) = \frac{1+\rho}{0.862+\rho} \tag{34}$$

In the transition region, Γ_s lies somewhere between 1 and Γ. Thus a simple asymptotic approximation to the true behavior can be expressed in the form

$$\Gamma_s = \frac{1 + aP_{\text{in}}\Gamma}{1 + aP_{\text{in}}} \tag{35}$$

in which the parameter a can be chosen to make a good fit in the transition region. To match the same type of crossover that we notice between P_s and P_{s+n}, we can take

$$a = \left(\frac{\pi^{1/2}K}{2L}\right)^2$$

to provide

$$\Gamma_s = \frac{\rho + \left(\frac{\pi^{1/2}KV}{2L}\right)^2 (1+\rho)\Gamma(\rho)}{\rho + \left(\frac{\pi^{1/2}KV}{2L}\right)^2 (1+\rho)} \tag{36}$$

5. Conclusions

Depending on the parameter VK/L, the soft limiter performs in varying degrees between the characteristics of a linear amplifier and a hard limiter. The performance parameters are furthermore expressible in terms of the hard-limiter suppression function under a change of variables. Such parameters include the output signal and noise powers, signal suppression factor, output SNR, and output signal-noise-density ratio.

References

1. Galejs, J., "Signal-to-Noise Ratios in Smooth Limiters," *IEEE Trans. Inf. Theory*, Vol. IT-1, June 1959, pp. 79–85.
2. Baum, R. F., "The Correlation Function of Smoothly Limited Gaussian Noise," *IEEE Trans. Inf. Theory*, Vol. IT-3, Sept. 1959, pp. 193–197.
3. Deutch, R., *Nonlinear Transformations of Random Processes*, Prentiss Hall, Inc., 1962, pp. 24–25.
4. "Handbook of Mathematical Functions," National Bureau of Standards, *Appl. Math Ser. 55*, June 1964.
5. Tausworthe, R. C., *Theory and Practical Design of Phase-Locked Receivers*, Technical Report 32-819, Jet Propulsion Laboratory, Pasadena, Calif., Feb. 16, 1966.

XXI. Communications Systems Research: Information Processing

TELECOMMUNICATIONS DIVISION

A. Digital Filtering of Random Sequences,

G. Jennings

1. Introduction

This article reports recently established results concerning the output of a digital filter when the input is a random sequence. The results are an improvement on those reported in SPS 37-48, Vol. III, pp. 213–220. The stability of the numerical method discussed therein is established under weaker conditions for recurrences of general length. A conjecture is used that has been proved in special cases.

In SPS 37-48, Vol. III, we considered filtering a random sequence $\{x_1, x_2, \cdots\}$ to form another sequence $\{y_1, y_2, \cdots\}$ by a linear recurrence of the form

$$y_n = \sum_{j=1}^{K} a_j y_{n-j} + x_n \tag{1}$$

When the values of $\{x_n\}$ are arbitrary real numbers, Eq. (1) can be solved only approximately. The sequence that is actually found satisfies

$$y_n = \sum_{j=1}^{K} a_j y_{n-j} + x_n + \delta_n \tag{2}$$

where δ_n is the error involved in evaluating the right side of Eq. (1).

We consider the type of error that occurs if we try to solve Eq. (1) on a digital computer. The details of the approximation procedure were specified in the previous article. Recall that

$$P = \{0, \pm 2\delta, \pm 4\delta, \cdots, 2k_o\delta = \pm M\}$$

is the set of possible values of each y_i. We had established bounds for the mean square of the difference between the solutions of Eqs. (1) and (2) under the condition that

$$\sum |a_j| < 1$$

This condition is weakened for recurrences of length 2 to the condition that the equation

$$x^2 - a_1 x - a_2 = 0$$

has only roots of modulus less than one. A lemma, proved for $K = 2$, is conjectured to hold for any K. Assuming this, the same generalization can be made for $K > 2$.

We denote the solution of Eq. (1) by $\tilde{y}_1, \tilde{y}_2, \cdots$, rewriting that equation as

$$\tilde{y}_n = \sum_{j=1}^{K} a_j \tilde{y}_{n-j} + x_n \tag{3}$$

We want to compare the solution of Eq. (2) with the solution of Eq. (3) when the same initial values are used for each sequence:

$$y_n = \tilde{y}_n = 0, \qquad n = 0,1, \cdots, K-1$$

Moreover, we shall require that each of the roots of the polynomial

$$t^k - \sum_{j=1}^{K} a_j t^{K-j} \tag{4}$$

be less than one in modulus.

2. The Conjecture

It is convenient to reformulate the problem as a matrix equation. We note the identity

$$\mathbf{Y}_n = A\mathbf{Y}_{n-1} + \mathbf{X}_n + \mathbf{D}_n \tag{5}$$

where

$$\mathbf{Y}_n = \begin{pmatrix} \mathbf{y}_{n-k+1} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{y}_n \end{pmatrix}, \qquad \mathbf{X}_n = \begin{pmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_n \end{pmatrix}, \qquad \mathbf{D}_n = \begin{pmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ \delta_n \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & 1 & 0 & & & & \\ 0 & 0 & 1 & & & & \\ & \cdot & & \cdot & & & \\ & & \cdot & & \cdot & & \\ & & & \cdot & & \cdot & \\ & & & & 0 & 0 & 1 \\ \mathbf{a}_j & \mathbf{a}_{j-1} & & \cdot & \cdot & \cdot & \mathbf{a}_1 \end{pmatrix}$$

Equation (5) provides a formulation of the problem which is equivalent to Eq. (2) in an obvious way. A minimal polynomial of A is Eq. (4), and therefore, the spectral radius of A, $\rho(A)$, is strictly less than one because of the restrictions on the roots of Eq. (4). Hence, by a fundamental theorem of spectral theory, there exists a norm on R^K, depending on ϵ, such that the induced matrix norm on A is less than $\rho(A) + \epsilon$, where ϵ may be chosen to be any positive number. This follows from the fact that a matrix is similar to a matrix of the form

$$\begin{pmatrix} \Lambda_1 & & \\ & \cdot & \\ & & \cdot \\ & & \Lambda_n \end{pmatrix}$$

where each Λ_i is a block of the form

$$\begin{pmatrix} \lambda & \epsilon & 0 & & & \\ 0 & \lambda & \epsilon & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & 0 & \lambda & \epsilon \\ & & & & 0 & \lambda \end{pmatrix}$$

where ϵ is an arbitrary real number (Ref. 1).

We denote such a norm by $||\cdot||_\epsilon$. We shall require in the following that $\rho(A) + \epsilon$ be less than one.

For such an ϵ, it immediately follows that A carries the set

$$B_{\eta,\epsilon} = \{\mathbf{y} \in R^K : ||\mathbf{y}||_\epsilon \leq \eta\}$$

into itself, where η is taken to be any positive number. As any two norms on a finite dimensional Euclidean space are equivalent (Ref. 2), there exists an $\eta > 0$, given $\epsilon > 0$, such that $B_{\eta,\epsilon}$ is contained in the unit hypercube

$$H_1^K = \left\{ \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{y}_K \end{pmatrix} : |\mathbf{y}_i| \leq 1, i = 1, \cdots, K \right\}$$

We assume that such an η and ϵ are chosen and fixed.

We shall now formulate the conjecture. We view the rounding process as being composed of two parts written symbolically as R_M and R_δ. R_M applied to a vector in R^K changes any component of that vector that exceeds M in absolute value to M, if that component is positive, to $-M$, otherwise; recall that M is positive. R_M leaves unchanged any component less than M in absolute value. R_δ rounds each component of a vector to that element of P to which it is closest in absolute value. With these conventions the calculation of the right-hand side of Eq. (4) can be viewed as applying A to $\mathbf{Y}_{n-1}$, adding $\mathbf{X}_n$, and then applying R_M followed by R_δ. That is,

$$\mathbf{Y}_n = R_\delta R_M(A\mathbf{Y}_{n-1} + \mathbf{X}_n)$$

We then have the following conjecture:

Conjecture. For a matrix A,

$$A = \begin{pmatrix} 0 & 1 & & & & & \\ & 0 & 1 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & & & 0 & 1 \\ a_K & a_{K-1} & & & & \cdot & a_1 \end{pmatrix}$$

where the polynomial

$$t^K - \sum_{j=1}^{K} a_j t^{K-j}$$

has only roots of modulus less than one, there exists a positive integer N, such that $(R_1A)^N\mathbf{Y}_0$ is contained in $B_{\eta/3,\epsilon}$, if $\mathbf{Y}_0$ is any vector contained in the unit hypercube. (N may depend on A, η, and ϵ but must be independent of $\mathbf{Y}_0$.) ϵ is chosen so that $\rho(A) + \epsilon < 1$. η is fixed so that $B_{\eta,\epsilon}$ is contained in H_1^K.

3. Estimates Using the Conjecture

We now proceed to derive estimates for $\int (y_i - \tilde{y}_i)^2 \, d\mu$ with the conjecture as an hypothesis. Let μ denote the probability measure that is induced when the x_i's are independent random variables with a gaussian distribution. (We do not try for the best possible estimates.) Assuming that $M = 1$, we apply the conjecture to deduce that $(R_1A)^N\mathbf{Y}_{i-N-1}$ is contained in $B_{\eta/3,\epsilon}$ regardless of the value of $\mathbf{Y}_{i-N-1}$, since $\mathbf{Y}_{i-N-1}$ must lie in the unit hypercube. With

$$\hat{\mathbf{Y}}_n = R_1 (A\hat{\mathbf{Y}}_{n-1} + \mathbf{X}_n)$$

neglecting to apply R_δ, the process of determining the $\hat{\mathbf{Y}}_K$'s is a continuous process of the $\mathbf{X}_K$'s. Therefore, if each x_j is restricted to be sufficiently small for $j = i - N, \cdots, i - 1$, $\hat{\mathbf{Y}}_{i-1}$ must be contanied in $B_{2\eta/3,\epsilon}$ where $\hat{\mathbf{Y}}_i$ has been determined from the equation

$$\hat{\mathbf{Y}}_K = R_1 (A\hat{\mathbf{Y}}_{K-1} + \mathbf{X}_K)$$

Let $a > 0$ be such a bound on the absolute value of the x_j's, $j = i - N, \cdots, i - 1$. We can increase the scale of allowed values of $\hat{\mathbf{Y}}_i$ by a factor of M, rounding to the hypercube each of whose edges is the length $2M$ and requiring

$$| x_j | \leq aM, \qquad \text{for } j = i - N + 1, \cdots, i - 1$$

to obtain a similar result. More precisely, define

$$C_i^M = \{\{x_j\}_{j=K}^{\infty} \,||\, x_j | \leq aM, \text{ for } j = i - N - 1, \cdots, i - 1\}$$

Then for $\{x_j\}$ in C_i^M, $\hat{Y}_{i-1}$ $\{x_j\}$ is contained in $B_{(2\eta/3)\,(M,\epsilon)}$. In this instance $\hat{\mathbf{Y}}_K$ satisfies the recursion

$$\hat{\mathbf{Y}}_K = R_M(A\hat{\mathbf{Y}}_{K-1} + \mathbf{X}_K)$$

Noting that

$$\mathbf{Y}_K = R_\delta R_M(A\mathbf{Y}_{K-1} + \mathbf{X}_K)$$

we can deduce that

$$\mathbf{Y}_K - \hat{\mathbf{Y}}_K = A(\mathbf{Y}_{K-1} - \hat{\mathbf{Y}}_{K-1}) + E_K$$

where each component of $\mathbf{E}_K$ is less than δ in magnitude. Noting that $\mathbf{Y}_K$ and $\hat{\mathbf{Y}}_K$ have the same value at the index $i - N - 1$, and that the solution of a linear recurrence depends continuously on the inhomogeneous term, we can deduce that, for $M \geq M_0 > 0$, there exists a δ_0 depending on M_0 but independent of M such that $\delta \leq \delta_0$, $M \geq M_0$ implies that $\mathbf{Y}_{i-1}(\{x_j\})$ is contained in $B_{\eta M,\epsilon}$, for $\{x_j\}$ in C_i^M. We thus obtain the condition that $A\mathbf{Y}_{i-1}$ is contained in $B_{(\rho(A)+\epsilon)\eta M,\epsilon}$, and therefore,under the additional constraint that the modulus of x_i be less than $(1 - \rho(A) - \epsilon)M$, the absolute value of δ_i must be less than δ. Let

$$S = C_i^M \cap \{\{x_j\} \,||\, x_i | \leq (1 - \rho(A) - \epsilon)M\}$$

and let S^c be its complement, and integrate with respect to probability measure:

$$\int \delta_i^2 d\mu \leq \int_c \delta_i^2 d\mu + \int_{S^c} \delta_i^2 d\mu$$

$$\leq \delta^2 + \int_{S^c} \delta_i^2 d\mu$$

The integrand of the last integral can be bounded using

$$| \delta_i | \leq \left(1 + \sum_j | a_j | \right) M + | x_i |$$

It is a straightforward estimate that the integral of the right-hand side of the above inequality over the complement of S goes to zero as δ and M go to zero and ∞,

respectively. The proof follows techniques used in SPS 37-48, Vol. III. We have, therefore, the following result:

Lemma

$$\int \delta_i^2 d\mu$$

goes to zero as δ goes to zero and M goes to infinity, uniformly in i.

Similarly, using techniques from SPS 37-48, Vol. III, we obtain:

Theorem 1

$$\int (y_i - \tilde{y}_i)^2 \, d\mu$$

goes to zero, uniformly in i, as δ and M approach zero and infinity, respectively.

4. The Conjecture for $K = 2$

We now proceed to prove the conjecture for recurrences of length 2 subject to the condition that the roots of Eq. (4) are less than one in modulus. First, we establish the notation that $\bar{\lambda}$ is the complex conjugate of λ, which is a complex number. If

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix}$$

and

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix}$$

are column vectors over the complex numbers, we define the inner product of the two $(\mathbf{v},\mathbf{w})$ as

$$v_1\bar{w}_1 + v_2\bar{w}_2$$

Let λ be a root of Eq. (4); then a calculation verifies that $\binom{1}{\lambda}$ is an eigenvector of the matrix

$$A = \begin{pmatrix} 0 & 1 \\ a_2 & a_1 \end{pmatrix}$$

with eigenvalue λ.

$$\mathbf{v}_1 = \frac{1}{(1 + |\lambda|^2)^{1/2}} \cdot \begin{pmatrix} 1 \\ \lambda \end{pmatrix}$$

and

$$\mathbf{v}_2 = \frac{1}{(1 + |\lambda|^2)^{1/2}} \cdot \begin{pmatrix} -\bar{\lambda} \\ 1 \end{pmatrix}$$

form an orthonormal basis for the two-dimensional vector space of column vectors over the complex numbers. As $\mathbf{v}_1$ is an eigenvector of A, the matrix of A in the basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ is upper triangular. Hence, $(A\mathbf{v}_2, \mathbf{v}_2)$ is an eigenvalue of A and must be less than one in magnitude.

Let

$$\mathbf{w} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

be a vector with real entries each of which is less than one in absolute value. As $\mathbf{v}_1$ and $\mathbf{v}_2$ form an orthonormal basis, $\mathbf{w}$ may be written in the form

$$\mathbf{w} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2$$

A computation yields that

$$\alpha_2 = (\mathbf{w}, \mathbf{v}_2) = (-\lambda\mathbf{y}_1 + \mathbf{y}_2)/(1 + |\lambda|^2)^{1/2}$$

Recall that y_1 and y_2 are real. The square of the modulus of α_2 is

$$[\mathrm{Re}\,\lambda y_1 + y_2) + (\mathrm{Im}\,\lambda)^2 y_1^2]/(1 + |\lambda|^2)$$

Multiplying by $1 + |\lambda|^2$ and calculating the partial derivative with respect to y_2 we achieve

$$\frac{1}{2}\frac{\partial}{\partial y_2}(1 + |\lambda|^2)\,|\alpha_2|^2 + \mathrm{Re}\,\lambda y_1 + y_2$$

We note two easily derived inequalities when $|y_1| \leq 1$:

$$\left.\begin{aligned} &\mathrm{Re}\,\lambda y_1 + y_2 \geq y_2 - |\lambda| \\ \text{and} \quad &\\ &\mathrm{Re}\,\lambda y_1 + y_2 \leq y_2 + |\lambda| \end{aligned}\right\} \qquad (6)$$

Recalling Eq. (5), specialized to the case at hand, we see that the image of a vector $\mathbf{w}$ in the unit square has first component less than or equal to one. If the second component, y_2, is greater than or equal to one, rounding will decrease the second component of the vector. Selecting the first of the two inequalities in Eq. (6), we see

$$1/2\,\frac{\partial}{\partial y_2}(1 + |\lambda|^2)\,|\alpha_2|^2 \geq 1 - |\lambda| \geq 0$$

from which we obtain

$$\frac{\partial}{\partial y_2} | \alpha_2 |^2 \geq 0$$

when $y_2 \geq 1$. In this case, the effect of rounding is to decrease the modulus of α_2. Similarly, if y_2 is less than -1, rounding increases y_2. The second inequality of Eq. (6) yields

$$\frac{\partial}{\partial y_2} | \alpha_2 |^2 \leq (-1 + | \lambda |) \frac{2}{1 + | \lambda |^2} \leq 0$$

when $y_2 \leq -1$. It follows that the modulus of α_2 is also decreased in this case.

We thus obtain the condition that the effect of applying R_1A to a vector in the unit cube

$$\alpha_1^{(0)} \mathbf{v}_1 + \alpha_2^{(0)} \mathbf{v}_2$$

produces another vector in the unit cube

$$\alpha_1^{(1)} \mathbf{v}_1 + \alpha_2^{(1)} \mathbf{v}_2$$

where

$$| \alpha_2^{(1)} | \leq \rho(A) | \alpha_2^{(0)} |$$

We see that successive application of R_1A yields a vector in the unit cube

$$\alpha_1^{(n)} \mathbf{v}_1 + \alpha_2^{(n)} \mathbf{v}_2$$

with

$$| \alpha_2^{(n)} | \leq \rho(A)^n | \alpha_2 |$$

and

$$\alpha_1^{(n)} = \lambda \alpha_1^{(n-1)} + \gamma \alpha_2^{(n-1)}$$

where $\gamma = (A\mathbf{v}_2, \mathbf{v}_1)$. This last equation defines $x_1^{(n)}$ as the solution of a nonhomogeneous linear recurrence.

The solution for $\gamma_1^{(n)}$ can be written in the closed form

$$\alpha_1^{(n+j)} = \lambda^j \alpha_1^{(n)} + \gamma \sum_{s=0}^{j} \lambda^s \alpha_2^{(n+j-s-1)}$$

If

$$| \alpha_2^{(r)} | \leq b \text{ for } r = n - 1$$

we obtain

$$| \alpha_1^{(n+j)} | \leq | \lambda |^j | \alpha_1^{(n)} | + \frac{| \gamma | b}{1 - \rho(A)^2}$$

As b may be taken arbitrarily small, $\alpha_1^{(n+j)}$ may be made arbitrarily small. As

$$| \alpha_2^{(n+j)} | \leq \rho(A)^n | \alpha_2^{(0)} |$$

$\alpha_2^{(n)}$ goes to zero, and hence, $(R_1A)^n$ shrinks the unit square uniformly to a point. Hence, the conjecture is established for $K = 2$.

References

1. John, F., *Notes on Ordinary Differential Equations*, p. 101, Courant Institute of Mathematical Sciences, New York, N.Y., 1964–1965.

2. Taylor, A. E., *Introduction to Functional Analysis*, p. 95, John Wiley & Sons, Inc., New York, N.Y., 1958.

B. Maximum Likelihood Symbol Synchronization for Binary Systems With Coherent Subcarrier-Symbol Rate, *W. J. Hurd*

1. Introduction

This article considers the symbol synchronization problem for binary systems in which the subcarrier frequency and the symbol rate are coherent; i.e., they are derived from the same frequency reference. In such systems, with the additional assumption that the subcarrier phase is known at the receiver, there are only a finite number of possible phases for the symbol clock. Typically there are an integral number, e.g., N, of subcarrier half cycles in each symbol time, so that the symbol phase can occur at N positions. Normally the subcarrier phase is tracked by a phase-locked loop.

The analysis and results are also applicable to systems in which there are an infinite number of possible phases for the subcarrier clock. In these cases, however, one must assume a finite number of possible phases, and accept phase errors smaller than the difference between the assumed phase position candidates. There is, however, the additional requirement that the symbol repetition rate be known exactly, a requirement that is automatically

satisfied when the subcarrier is tracked and the symbol timing is derived from the same clock.

The basic symbol synchronization problem is to find the optimum decision rule for estimation of the correct symbol timing based on observations of the received noisy data for a fixed length of time. Other related problems are to evaluate the performance of the optimum and near-optimum decision rules as functions of the symbol signal-to-noise ratio (SNR) and the observation time, and to find the required observation times to achieve given error probabilities at a given SNR. Stiffler[1] has derived the maximum likelihood decision rule for general (n-ary) amplitude-modulation (AM) systems, of which binary AM and biphase modulation systems are special cases. Here we present a different derivation for the binary case, and a more concise expression for the final result. We also examine two methods that approximate the maximum likelihood rule, one at high SNRs and one at low SNRs, and present numerical results for these approximations. The two approximations were suggested by Stiffler, but numerical results were not given, although numerical comparison of the two methods has been given by Stiffler for a problem that somewhat resembles the synchronization problem (SPS 37-29, Vol. 14, pp. 285–290).

2. Derivation of Maximum Likelihood Rule

Suppose the received signal waveform $r(t) = s(t) + n(t)$ is observed for $0 \leq t \leq (M+1)T$ sec, where $s(t)$ is the signal, $n(t)$ is white gaussian noise with two-sided spectral density $N_0/2$, T is the duration of one symbol, and $M+1$ is the number of symbol times observed. The signal $s(t)$ is constant at either $+A$ or $-A$ over each symbol time, so that the symbol energy is A^2T, but it is not known at which points in time the symbols start. The N possible candidates for the starting time of the first full symbol observed are $0, T/N, 2T/N, \cdots, (N-1)T/N$, and if the actual starting time is kT/N, successive symbols start $kT/N, (kT/N)+T, (kT/N)+2T, \cdots$. We denote the first M symbols by the vector $\mathbf{A} = (A_1, A_2, \cdots, A_M)$. The probability that each symbol is $+A$ or $-A$ is one half, and each symbol is independent of all others.

To make a maximum likelihood decision as to correct symbol timing, we must compute the *a posteriori* probability of each candidate, given that $r(t)$ is received, and choose the candidate for which this probability is maximized. The first step is to convert the problem from one involving random functions to a finite dimensional vector problem. The signal component of any possible received waveform can then be expressed as

$$s(t) = \sum_{i=1}^{N(M+1)} s_i \, \phi_i(t) \tag{1}$$

where the $\phi_i(t)$ are the orthonormal functions

$$\phi_i(t) = \begin{cases} (N/T)^{1/2}, & \text{for } (i-1)T/N \leq t < iT/N \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

and

$$s_i = \int_0^{(M+1)T} s(t) \, \phi_i(t) \, dt = \pm A(T/N)^{1/2} \tag{3}$$

The vector $\mathbf{s} = (s_1, s_2, \cdots, s_{N(M+1)})$ completely defines $s(t)$. Similarly, we define the noise vector

$$\mathbf{n} = (n_1, n_2, \cdots, n_{N(M+1)}) \tag{4}$$

where

$$n_i = \int_0^{(M+1)T} n(t) \, \phi_i(t) dt \tag{5}$$

The noise components n_i are independent zero mean gaussian random variables with variances $N_0/2$.

We also define the received signal plus noise vector

$$\mathbf{r} = \mathbf{s} + \mathbf{n} \tag{6}$$

With this notation, it can be shown (Ref. 1) that the vector $\mathbf{r}$ is a sufficient statistic, and contains all of the data that is relevant to determining the received signal waveform; i.e., the *a posteriori* probability of the transmitted signal vector conditioned on $r(t)$ is the same as that conditioned on $\mathbf{r}$.

$$p(\mathbf{s} \mid r(t)) = p(\mathbf{s} \mid \mathbf{r}) \tag{7}$$

The *a posteriori* probability that the correct timing occurs at position k (i.e., at $t = kT/N$), is

$$p(k \mid r(t)) = p(k \mid \mathbf{r}) = \frac{p(\mathbf{r} \mid k) \, p(k)}{p(\mathbf{r})} \tag{8}$$

[1]Stiffler, J. J., *The Synchronization in Communication Systems*, to be published by Prentiss Hall Pub. Co., Englewood Cliffs, N.J., in 1969.

Since $p(\mathbf{r})$ is not a function of k, and since $p(k) = 1/N$ for all k, the best estimate of k can be made by computing $p(\mathbf{r} \mid k)$ for each k and choosing the maximum.

It is now convenient to neglect the data for $t < kT/N$ and for $t \geq MT + kT/N$, i.e., to neglect r_i for $i \leq k$ and for $i > MN + k$, $k = 0, 1, \cdots, N-1$. This will result in negligible degradation for reasonably large M, and the more exact result can easily be obtained if desired. Since the n_i and the A_m are all independent, and the signal components s_i are all the same for $mN - N + k < i \leq mN + k$, the probability density of $\mathbf{r}$, conditioned on phase position k, is

$$p(\mathbf{r} \mid k) = \prod_{m=1}^{M} p(r_{mN-N+k+1}, \cdots, r_{mN+k} \mid k) \tag{9}$$

Conditioned on A_m, the r_i for $mN - N + k < i \leq mN + k$ are independent. Furthermore, the s_i are all $+A(T/N)^{1/2}$ or all $-A(T/N)^{1/2}$ with equal probability, so

$$p(r_{mN-N+k+1}, \cdots, r_{nM+k} \mid k) = \frac{1}{2} \prod_{i=mN-N+k+1}^{mN+k} p(r_i \mid k, s_i = +A(T/N)^{1/2}) + \frac{1}{2} \prod_{i=mN-N+k+1}^{mN+k} p(r_i \mid k, s_i = -A(T/N)^{1/2}) \tag{10}$$

Using the gaussian densities with means $\pm A(T/N)^{1/2}$ and variances $N_0/2$ for the conditional densities in Eq. (10), substituting into Eq. (9), and simplifying, we get

$$p(\mathbf{r}|k) \approx (\pi N_0)^{-NM/2} \exp\left\{ - (MA^2T/N_0) - N_0^{-1} \sum_{i=k+1}^{NM+k} r_i^2 \right\} \prod_{m=1}^{M} \cosh\left(2A(T/N)^{1/2} N_0^{-1} \sum_{i=mN-N+k+1}^{mN+k} r_i \right) \tag{11}$$

But, neglecting end effects,

$$\sum_{i=k+1}^{NM+k} r_i^2$$

is approximately the same for all k, so the exponential terms in Eq. (11) can be dropped.

Finally, the maximum likelihood decision rule, neglecting only negligible end effects, is to choose the k which maximizes the function

$$L(k) = \prod_{m=1}^{M} \cosh\left(2A(T/N)^{1/2} N_0^{-1} \sum_{i=mN-N+k+1}^{mN+k} r_i \right) \tag{12}$$

which, using the defining relation for the r_i, can be written as

$$L(k) = \prod_{m=1}^{M} \cosh\left(2AN_0^{-1} \int_{(m-1)T+kT/N}^{mT+kT/N} r(t)\, dt \right) \tag{13}$$

3. Approximation Methods

The maximum likelihood rule is impractical to use because of the product of the hyperbolic cosines, and because the expression depends on knowledge of the signal amplitude A and the noise spectral density. However, $L(k)$ can be approximated using one expression for high SNRs and another for low SNRs.

Suppose we define

$$x_m(k) = \int_{(m-1)T+kT/N}^{mT+kT/N} r(t)\, dt \tag{14}$$

Then, conditioned on the mth symbol, the mean and variance of the argument $2AN_0^{-1}\, x_m(k)$ of the hyperbolic cosine are

$$E\left\{ 2AN_0^{-1}\, x_m(k) \mid A_m = \pm A \right\} = \pm 2A^2T/N_0 \tag{15}$$

$$\text{var}\left\{ 2AN_0^{-1}\, x_m(k) \mid A_m = \pm A \right\} = 2A^2T/N_0 \tag{16}$$

Since the SNR is equal to the square of the conditional mean divided by the conditional variance, the quantity $2A^2T/N_0$ is the SNR. Hence, on the average, the argument of the hyperbolic cosine is small for low SNRs and large for high SNRs.

a. Squaring method. For low SNRs, the product of hyperbolic cosines can be expanded into a product of Taylor series, and all but the first terms can be dropped.

$$\prod_{m=1}^{M} \cosh\left(x_m(k)\right) \approx 1 + \sum_{m=1}^{M} \left(2AN_0^{-1}\, x_m(k)\right)^2 \tag{17}$$

In this case, synchronization is determined by measuring

$$L_S(k) = \sum_{m=1}^{M} x_m^2\,(k) \tag{18}$$

for each k and choosing the largest. This is called the *squaring method.*

b. Absolute value method. For high SNRs, the hyperbolic cosine is approximately exponential, so

$$\prod_{m=1}^{M} \cosh\left(2AN_0^{-1}\, x_m(k)\right) \approx \prod_{m=1}^{M} \frac{1}{2} \exp\left(\left|\, 2AN_0^{-1}\, x_m(k)\,\right|\right) \tag{19}$$

It suffices to measure

$$L_A(k) = \sum_{m=1}^{M} \left|\, x_m(k)\,\right| \tag{20}$$

for each k and to choose the largest. This is called the *absolute value method.*

c. Performance evaluation. By the central limit theorem, the $L_S(k)$, $1 \leq k \leq N-1$, and the $L_A(k)$, $1 \leq k \leq N-1$, are approximately jointly gaussian for sufficiently large M. Let us assume that the $k = 0$ position is the actual transition point, and define the normalized variables

$$M^{1/2} A_k = \frac{L_A(0) - L_A(k)}{\left(\text{var}\left\{L_A(0) - L_A(k)\right\}\right)^{1/2}} \tag{21}$$

and

$$M^{1/2} S_k = \frac{L_S(0) - L_S(k)}{\left(\text{var}\left\{L_S(0) - L_S(k)\right\}\right)^{1/2}}$$

The normalization is chosen so that the statistics of A_k and S_k are independent of M. A correct decision is made whenever all of the A_k (or S_k) are greater than zero for $k = 1, 2, \cdots, N-1$.

Since $M^{1/2}A_k$ and $M^{1/2}S_k$ are linear combinations of approximately gaussian random variables, they are also approximately gaussian. Furthermore, their variances are unity by the normalization in Eq. (22), so

$$\Pr\left\{A_k < 0\right\} = (2\pi)^{-1/2} \int_{-\infty}^{0} \exp\left\{-\frac{1}{2}\left(x - M^{1/2}\, E\{A_k\}\right)^2\right\} dx$$

$$= \frac{1}{2} \operatorname{erfc}\left(2^{-1/2}\, M^{1/2}\, E\{A_k\}\right) \tag{23}$$

and

$$\Pr\left\{S_k < 0\right\} = \frac{1}{2} \operatorname{erfc}\left(2^{-1/2}\; M^{1/2}\, E\{S_k\}\right) \tag{24}$$

where $\operatorname{erfc}(a) = 1 - \operatorname{erf}(a)$ is the complementary error function and

$$\operatorname{erf}(a) = 2\pi^{-1/2} \int_0^a \exp\left\{-x^2\right\} dx \tag{25}$$

For the absolute value method, for example, the probability of error is at least as great as the probability that $A_1 \leq 0$ and by a "union bound" does not exceed

$$\sum_{k=1}^{N-1} \Pr\left\{A_k < 0\right\}$$

Hence for the absolute value method, the error probability is bounded below by

$$P_E \geq \frac{1}{2} \operatorname{erfc}\left(\left(\frac{M}{2}\right)^{1/2} E\{A_1\}\right) \tag{26}$$

and above by

$$P_E \leq \frac{1}{2} \sum_{k=1}^{N-1} \operatorname{erfc}\left(\left(\frac{M}{2}\right)^{1/2} E\{A_k\}\right) \tag{27}$$

Similar expressions for the squaring method are obtained by substituting $E\{S_k\}$ for $E\{A_k\}$ throughout. Note that by symmetry $E\{A_k\} = E\{A_{N-k}\}$ so that the k and $N-k$ terms in Eq. (27) are equal.

d. Numerical results. The expected values of A_k and S_k are derived in the following *Subsection 4.* The results are shown as a function of symbol SNR, $R = 2A^2T/N_0$, in Fig. 1 for fractional timing errors $k/N = 1/2,\ 1/4,\ 1/8, \cdots, 1/1024$. As expected, these curves show that the squaring method is better at low symbol signal-to-noise ratios and that the absolute value method is better at higher SNRs. At SNRs of interest for coded systems, around $R = 1$, the two methods are approximately equally good. The squaring method may be preferred because

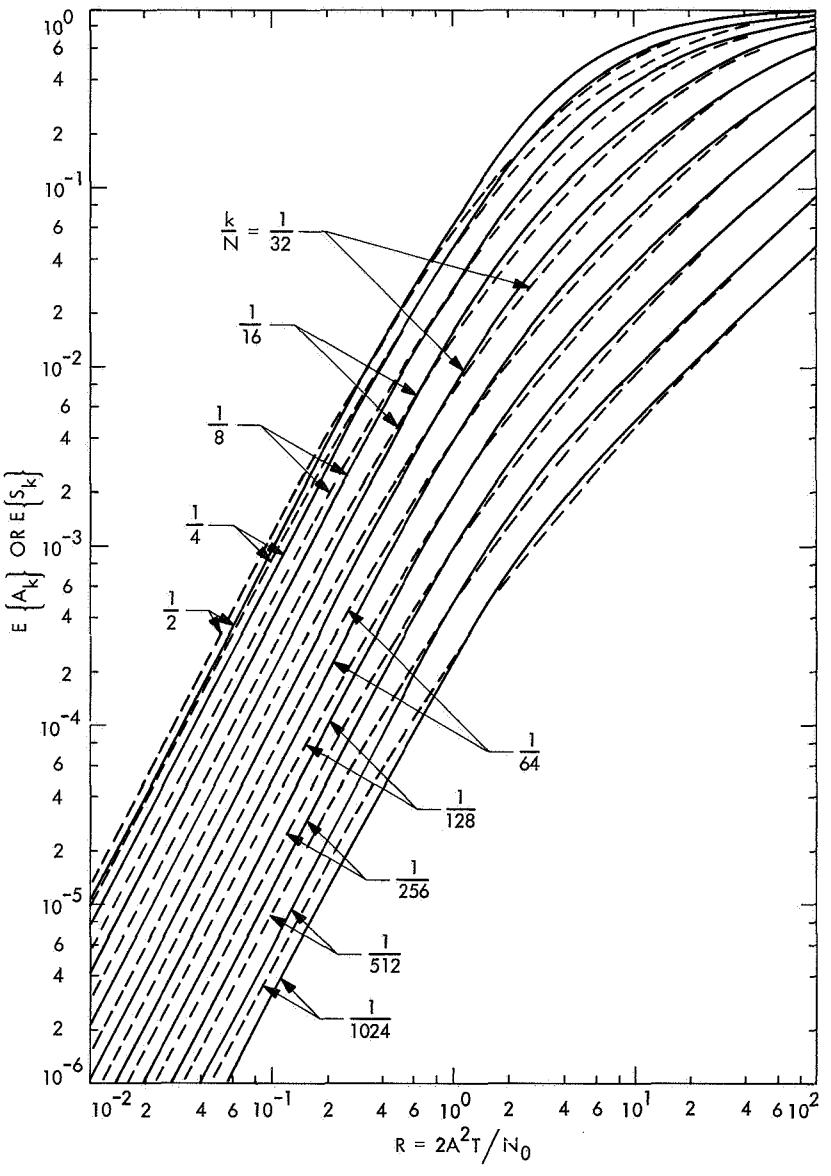


Fig. 1. A_k and S_k as functions of symbol SNR

it is about 1.5 dB better for low SNRs, and the absolute-value method is not much better at higher SNRs. On the other hand, the absolute-value method may be easier to implement in some situations.

e. Simplified upper bound on P_E. The results shown in Fig. 1 can be used to simplify the expression for the upper bound on the probability of incorrect synchronization. The basic idea is that synchronizing incorrectly at position $k = 1$ or $k = N - 1$ is much more likely than at any other position, so that all terms in Eq. (27) can be dropped except the $k - 1$ and $k = N - 1$ terms.

To reduce the lower bound of Eq. (26) on probability of synchronization error to a reasonably low level, we must choose M large enough so that

$$0.5 \operatorname{erfc}\left(\left(\frac{M}{2}\right)^{1/2} E\{A_1\}\right)$$

for the absolute value method, is small. Since $E\{A_2\}$ for fixed N is typically 1.5 to 2 times $E\{A_1\}$ for that N,

$$\operatorname{erfc}\left(\left(\frac{M}{2}\right)^{1/2} E\{A_2\}\right)$$

will be negligible compared to

$$\operatorname{erfc}\left(\left(\frac{M}{2}\right)^{1/2} E\{A_1\}\right)$$

whenever the latter is small, because erfc(z) decreases approximately as $\exp\{-z^2\}/z$. Similarly, all terms in Eq. (27) will be small except the $k-1$ and $k=N-1$ terms, which are equal by symmetry, so that the upper bound becomes approximately

$$P_E \lesssim \operatorname{erfc}\left(\left(\frac{M}{2}\right)^{1/2} E\{A_1\}\right) \tag{28}$$

for the absolute value method, and similarly for the squaring method.

4. Calculations for Approximate Methods

The mean values of A_k and S_k can be expressed in terms of the relevant statistics for the $x_m(k)$, given that $k=0$ is the correct position. Normalizing to unit noise energy per symbol at the detector (integrator, matched filter) output, we define

$$\sigma^2 = \frac{N_0 T}{2} = 1 = \text{noise energy per symbol}$$

$$\mu = |AT|$$

$$R = \frac{2A^2T}{N_0} = \frac{\mu^2}{\sigma^2} = \mu^2$$

$$= \text{SNR at detector output}$$

$$\rho_k = \frac{N-k}{N} = \text{fraction of } m\text{th symbol in assumed position of the } m\text{th symbol for given } k$$

$$\alpha_k^2 = \rho_k^2 + (1-\rho_k)^2$$

$$\beta_k = \rho_k(1-\rho_k)$$

and

$$\gamma_k = 2\rho_k - 1$$

With this notation, we can write $x_m(k)$ as

$$x_m(k) = \int_{(m-1)T+kT/N}^{mT+kT/N} n(t)\,dt + \rho_k A_m T + (1-\rho_k) A_{m+1} T \tag{29}$$

a. Squaring method. To calculate the statistics of the $L_S(k)$, we note that $x_m(k)$ and $x_n(k)$ are independent for $|m-n|>1$, and that $x_m(0)$ and $x_n(k)$ are independent except for $n=m$ and $n=m-1$. In other cases, no symbol affects both $x_m(k)$ and $x_n(k)$ or both $x_m(0)$ and $x_n(k)$, and the noise components are independent because the noise is white. Hence

$$E\{L_S(k)\} = ME\{x_m^2(k)\} \tag{30}$$

$$\begin{aligned} E\{L_2^S(k)\} &= E\sum_{m=1}^{M}\sum_{n=1}^{M} x_m^2(k)\,x_n^2(k) \\ &= ME\{x_m^4(k)\} + 2(M-1)E\{x_m^2(k)\,x_{m+1}^2(k)\} \\ &\quad + (M^2-3M+2)E^2\{x_m^2(k)\} \end{aligned} \tag{31}$$

$$\begin{aligned} E\{L_S(0)L_S(k)\} &= E\sum_{m=1}^{M}\sum_{n=1}^{M} x_m^2(0)x_n^2(k) \\ &= ME\{x_m^2(0)x_m^2(k)\} \\ &\quad + (M-1)E\{x_m^2(0)\,x_{m-1}^2(k)\} \\ &\quad + (M^2-2M+1)E\{x_m^2(0)\,x_m^2(k)\} \end{aligned} \tag{32}$$

The expectations in the above equations are obtained by using Eq. (29), expanding, and taking expectations term by term:

$$E\{x_m^2(k)\} = 1 + \alpha_k^2 R \tag{33}$$

$$E\{x_m^4(k)\} = 3 + 6\alpha_k^2 R + (\alpha_k^4 + 4\beta_k^2)R^2 \tag{34}$$

$$E\{x_m^2(k)\,x_{m+1}^2(k)\} = 1 + 2\alpha_k^2 R + \alpha_k^4 R^2 = E^2\{x_m^2(k)\} \tag{35}$$

$$E\{x_m^2(0)x_m^2(k)\} = 1 + 2\rho_k^2 + (2 - 2\rho_k + 6\rho_k^2)R + \alpha_k^2 R^2 \tag{36}$$

$$\begin{aligned} E\{x_m^2(0)x_{m-1}^2(k)\} &= 3 - 4\rho_k + 2\rho_k^2 \\ &\quad + (6 - 10\rho_k + 6\rho_k^2)R + \alpha_k^2 R^2 \end{aligned} \tag{37}$$

Substituting Eqs. (33) to (37) into Eqs. (30) to (32), combining terms, and assuming $M \gg 1$ so that terms not depending on M are negligible, the mean and variance of $L_S(0) - L_S(k)$ are

$$E\{L_S(0) - L_S(k)\} = 2M\beta_k R \tag{38}$$

and

$$\operatorname{var}\{L_S(0) - L_S(k)\} = 8M\beta_k\left(1 + R + \frac{\beta_k R^2}{2}\right) \tag{39}$$

As expected, these expressions are linear in M and symmetric in k and $N-k$. Finally, the mean of S_k is

$$E\{S_k\} = R\left(\frac{\beta_k}{2(1+R)+\frac{\beta_k^2}{2}}\right)^{1/2} \tag{40}$$

b. Absolute value method. Following the same procedure as above but replacing squares by absolute values, we get

$$\frac{1}{M}E\{L_A(0) - L_A(k)\} = E\{|\,x_m(0)\,|\} - E\{|\,x_m(k)\,|\} \tag{41}$$

and

$$\begin{aligned}\frac{1}{M}\operatorname{var}\{L_A(0) - L_A(k)\} = {} & E\{x_m^2(0)\} - E^2\{|\,x_m(0)\,|\} \\ & + E\{x_m^2(k)\} \\ & + 2E\{|\,x_m(k)x_{m+1}(k)\,|\} \\ & - 3E^2\{|\,x_m(k)\,|\} \\ & - 2E\{|\,x_m(0)x_m(k)\,|\} \\ & - 2E\{|\,x_m(0)x_{m-1}(k)\,|\} \\ & + 4E\{|\,x_m(0)\,|\}\,E\{|\,x_m(k)\,|\}\end{aligned} \tag{42}$$

The absolute moments in Eq. (42) must now be evaluated. Conditioned on the received symbols, $x_m(0)$ and $x_n(k)$ are jointly gaussian, and, by symmetry, we can always assume that the first symbol affecting the desired statistic is equal to $+A$. Hence, for one variate,

$$E\{|\,x_m(k)\,|\} = \frac{1}{2}E\left\{|\,x_m(k)\,||\,A_m = A, A_{m+1} = A\right\} + \frac{1}{2}E\left\{|\,x_m(k)\,|\,A_m = A, A_{m+1} = -A\right\} \tag{43}$$

Since

$$E\{x_m(k)\,|\,A_m = A, A_{m+1} = \pm A\} = (1 - (1-\rho_k) \pm (1-\rho_k))\mu$$

and

$$\operatorname{var}\{x_m(k)\,|\,A_m, A_{m+1}\} = \sigma^2 = 1$$

$$\begin{aligned}E\{|\,x_m(k)\,|\,|A_m = A, A_{m+1} = \pm A\} &= (2\pi)^{-1/2}\left[\int_0^\infty - \int_{-\infty}^0\right] x\exp\left\{-\frac{1}{2}\left(x - [1-(1-\rho_k)\pm(1-\rho_k)]\mu\right)^2\right\}dx \\ &= [1-(1-\rho_k)\pm(1-\rho_k)]\,\mu\,\operatorname{erf}\left(2^{-1/2}\,[1-(1-\rho_k)\pm(1-\rho_k)]\,\mu\right) \\ &\quad + 2(2\pi)^{-1/2}\exp\left\{-\frac{[1-(1-p_k)\pm(1-p_k)]^2\,\mu^2}{2}\right\}\end{aligned} \tag{44}$$

and

$$E\{|\,x_m(k)\,|\} = \frac{\mu}{2}\operatorname{erf}(2^{-1/2}\mu) + \frac{\gamma_k\mu}{2}\operatorname{erf}(2^{-1/2}\gamma_k\mu) + (2\pi)^{-1/2}\exp\left\{-\frac{\mu^2}{2}\right\} + (2\pi)^{-1/2}\exp\left\{-\frac{\gamma_k^2\mu^2}{2}\right\} \tag{45}$$

To evaluate the joint absolute moments, we first define the function g by

$$g(m_x,m_y,\rho) = E\{|\, xy \,|\} \tag{46}$$

where x and y are jointly gaussian with means m_x and m_y, unit variances, and covariance ρ. Then

$$E\{|\, x_m(k)x_{m+1}(k) \,|\} = \frac{1}{4}\left[g(\mu,\mu,0) + g(\mu,\gamma_k\mu,0) + g(\gamma_k\mu, -\gamma_k\mu,0) + g(\gamma_k\mu, -\mu,0)\right] \tag{47}$$

where the first two arguments of g reflect the four possible combinations of A_{m+1} and A_{m+2}. Similarly

$$E\{|\, x_m(0)x_m(k) \,|\} = \frac{1}{2}\left[g(\mu,\mu,\rho_k) + g(\mu,\gamma_k\mu\,\rho_k)\right] \tag{48}$$

and

$$E\{|\, x_m(0)x_{m-1}(k) \,|\} = \frac{1}{2}\left[g(\mu,\mu,1-\rho_k) + g(-\gamma_k\mu,\mu,1-\rho_k)\right] \tag{49}$$

The function g is

$$g(m_x, m_y,\rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\, xy \,| \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{m})^T \mathbf{R}^{-1}(\mathbf{x} - \mathbf{m})\right\} dx\, dy \tag{50}$$

where

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{51}$$

$$\mathbf{m} = \begin{pmatrix} m_x \\ m_y \end{pmatrix} \tag{52}$$

and $\mathbf{R}$ is the covariance matrix of $\mathbf{x}$. We now perform a transformation to polar coordinates, letting

$$\mathbf{B} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{53}$$

so

$$\mathbf{x} = r\,\mathbf{B} \tag{54}$$

Then g becomes

$$g(m_x,m_y,\rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}}\int_0^{2\pi} d\theta \int_0^{\infty} dr\, r^3 \,|\cos\theta \sin\theta\,|\, x \exp\left\{-\frac{1}{2}(r\,\mathbf{B} - \mathbf{m})^T \mathbf{R}^{-1}(r\mathbf{B} - \mathbf{m})\right\} \tag{55}$$

The infinite integral can now be integrated in closed form yielding

$$g(m_x,m_y,\rho) = \frac{1}{2\pi(1-\rho^2)^{1/2}}\int_0^{2\pi} d\theta \,|\cos\theta \sin\theta\,| \exp\{-(a_3 - a_2^2/a_1)/2\}$$
$$\times\left[(2a_1 + a_2^2)\, a_1^{-3} \exp\{-a_2^2/(2a_1)\} + \pi^{1/2}\, 2^{-1/2}\,(a_2^3 + 3a_1a_2)\, a_1^{-7/2} \operatorname{erfc}(-a_2(2a_1)^{1/2})\right] \tag{56}$$

where

$$a_1 = \mathbf{B}^T \mathbf{R}^{-1} \mathbf{B} \tag{57}$$

$$a_2 = \mathbf{B}^T \mathbf{R}^{-1} \mathbf{m} \tag{58}$$

and

$$a_3 = \mathbf{m}^T \mathbf{R}^{-1} \mathbf{m} \tag{59}$$

Equation (56) was integrated numerically to obtain the curves in Fig. 1.

5. Summary

Although the maximum likelihood method for symbol synchronization derived in *Subsection 2* is impractical to implement, it is closely approximated by the squaring method at low SNRs and by the absolute value method at high SNRs. The probability that synchronization does not occur at exactly the correct place is bounded by

$$\frac{1}{2}\operatorname{erfc}\left((M/2)^{1/2} E\{A_1\}\right) \leq P_E \lesssim \operatorname{erfc}\left((M/2)^{1/2} E\{A_1\}\right) \tag{60}$$

for the absolute value, and by

$$\frac{1}{2}\operatorname{erfc}\left((M/2)^{1/2} E\{S_1\}\right) \leq P_E \lesssim \operatorname{erfc}\left((M/2)^{1/2} E\{S_1\}\right) \tag{61}$$

for the squaring method. In these expressions, M is the number of symbols used in the estimation, and A_1 and S_1 are given as a function of SNR in Fig. 1. The parameter N in Fig. 1 is the number of places at which synchronization might occur, which is typically the number of subcarrier half cycles in one symbol time.

Reference

1. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, Chap. 4, John Wiley and Sons, New York, 1965.

XXII. Communications Systems Research: Data Compression Techniques

TELECOMMUNICATIONS DIVISION

A. Estimating the Proportions in a Mixture of Two Normal Distributions Using Quantiles, Part II,

I. Eisenberger

1. Introduction

The density function, $g(x)$, of a mixture of two normal distributions with proportions p and $1-p$, is given by

$$g(x) = \frac{p}{\sigma_1 (2\pi)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 \right] + \frac{1-p}{\sigma_2 (2\pi)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 \right]$$

The problem of estimating p using a small number of sample quantiles, when the parameters of the normal distributions are known and the sample sizes are large, was considered in SPS 37-32, Vol. IV, pp. 263–268, where an estimator of p using four sample quantiles was proposed. Further study indicates, however, that it is possible to construct estimators for p which, in general, will be more efficient than that proposed previously. In particular, when $\mu_1 \neq \mu_2$, we will give one estimator based on a single quantile and another based on a linear combination of six quantiles. It will also be shown that, in some cases, combining the two estimators gives the best results. For the special case $\mu_1 = \mu_2$, estimators using two symmetric quantiles give results comparable to that achieved using the four quantile estimators.

For $\mu_1 \neq \mu_2$, an investigation was made of 28 cases involving four sets of the parameters μ_1, μ_2, σ_1, and σ_2, with values of p ranging from 0.05 to 0.95. Columns 2–6 of Table 1 give the parameter values of each case. For $\mu_1 = \mu_2$, 14 cases were considered. These values are given in columns 2–6 of Table 2.

Since estimation by means of sample quantiles is usable for on-board data compression in deep-space probes, the analysis will be given for each of the following conditions:

(1) The orders of the quantiles must be fixed in advance.

(2) The orders can be changed by signals from earth.

2. Review of Quantiles

To define a quantile, consider n independent sample values, $x_1, x_2, \cdots, x_n$, taken from a distribution of a continuous type with distribution function $H(x)$ and density function $h(x)$. The sth quantile, or the quantile of order s of the distribution or population, denoted by $\zeta(s)$, is defined as the root of the equation $H(\zeta) = s$.

Table 1. Variances of several estimators of the proportions in a mixture of two normal distributions for $\mu_1 \neq \mu_2$

Case	p	μ_1	μ_2	σ_1	σ_2	opt s	opt s n Var (p_1^*)	$s = 0.5$ n Var ($\hat{p}_1$)	$s = 0.332$ n Var ($\tilde{p}_1$)	6 quantiles		7 quantiles		n Var ($\tilde{p}_4$)	ML n Var (p^*)
										n Var ($\hat{p}_6$)	E ($\hat{p}_6$)	n Var ($\hat{p}_7$)	n Var ($\tilde{p}_7$)		
1	0.05	0	1	1	0.5	0.0237	0.1613	1.999	1.0269	0.3503	0.0425	0.2878	0.1135	6.950	0.1347
2	0.10	0	1	1	0.5	0.0532	0.2752	1.860	0.9687	0.4453	0.0947	0.2621	0.1052	6.487	0.2325
3	0.30	0	1	1	0.5	0.1874	0.5990	1.374	0.7944	0.7481	0.2963	0.1777	0.0801	4.906	0.5137
4	0.50	0	1	1	0.5	0.3320	0.7848	1.020	0.7848	0.9261	0.4975	0.1289	0.0698	3.709	0.6716
5	0.70	0	1	1	0.5	0.4810	0.8547	0.8569	1.084	1.043	0.7001	0.1181	0.0854	2.873	0.7218
6	0.90	0	1	1	0.5	0.6314	0.8129	0.9596	1.667	1.069	0.8994	0.1237	0.1463	2.390	0.6658
7	0.95	0	1	1	0.5	0.6688	0.7848	1.023	1.846	1.066	0.9495	0.1281	0.1688	2.326	0.6346
									$s = 0.4987$						
8	0.05	0	4	1	0.5	0.0497	0.0484	0.9026	0.8980	0.0356	0.0500	0.1416	0.1409	1.344	0.0482
9	0.10	0	4	1	0.5	0.0997	0.0914	0.8103	0.8059	0.0620	0.0763	0.1722	0.1716	1.223	0.0911
10	0.30	0	4	1	0.5	0.2989	0.2128	0.4902	0.4877	0.1342	0.2608	0.2808	0.2806	0.8762	0.2123
11	0.50	0	4	1	0.5	0.4987	0.2536	0.2539	0.2536	0.1170	0.5115	0.2592	0.2673	0.7901	0.2529
12	0.70	0	4	1	0.5	0.6987	0.2139	0.4900	0.4926	0.3697	0.7252	0.9376	0.9374	0.8774	0.2130
13	0.90	0	4	1	0.5	0.8992	0.0931	0.8100	0.8142	0.2070	0.9107	0.4969	0.4970	1.223	0.0924
14	0.95	0	4	1	0.5	0.9404	0.0500	0.9025	0.9072	0.0871	0.9514	0.1624	0.1625	1.344	0.0494
									$s = 0.7946$						
15	0.05	0	1	1	3	0.5973	1.796	2.350	3.505	9.148	0.0478	1.1442	1.239	16.17	0.7900
16	0.10	0	1	1	3	0.6190	1.766	2.611	3.152	8.833	0.0969	1.0797	1.182	18.45	0.8126
17	0.30	0	1	1	3	0.7065	1.587	4.164	1.973	7.210	0.2991	0.8763	0.9692	29.81	0.8359
18	0.50	0	1	1	3	0.7946	1.307	6.415	1.307	5.559	0.4966	0.9606	0.7550	63.19	0.7536
19	0.70	0	1	1	3	0.8821	0.9157	9.279	1.235	3.830	0.7101	1.3154	0.5111	44.52	0.5628
20	0.90	0	1	1	3	0.9655	0.3832	12.73	1.453	2.097	0.9071	2.008	0.3129	85.18	0.2472
21	0.95	0	1	1	3	0.9844	0.2150	13.68	1.529	1.357	0.9641	2.201	0.2345	91.38	0.1404
									$s = 0.6003$						
22	0.05	0	4	1	3	0.2089	0.3303	0.9028	1.355	0.6172	0.0503	1.561	1.391	1.577	0.2658
23	0.10	0	4	1	3	0.2528	0.3682	0.8105	1.217	0.6445	0.1007	1.672	1.497	1.546	0.3067
24	0.30	0	4	1	3	0.4281	0.4561	0.5095	0.7369	0.5948	0.3007	1.669	1.487	1.661	0.4055
25	0.50	0	4	1	3	0.6003	0.4481	0.5364	0.4481	0.7331	0.4972	2.140	2.088	2.115	0.4092
26	0.70	0	4	1	3	0.7678	0.3463	0.8336	0.5671	0.3820	0.7014	1.000	0.9986	2.913	0.3207
27	0.90	0	4	1	3	0.9276	0.1470	1.250	0.8247	0.1675	0.9078	0.4023	0.3746	4.076	0.1371
28	0.95	0	4	1	3	0.9653	0.0796	1.370	0.9010	0.1921	0.9540	0.5330	0.4964	4.437	0.0744

Table 2. Variances of several estimators of the proportions in a mixture of two normal distributions for $\mu_1 = \mu_2$

Case	p	μ_1	μ_2	σ_1	σ_2	Using opt s			Using opt s of $p = 0.5$		n Var $(\tilde{p}_4)$	ML n Var (p^*)
						opt s	n Var $(\hat{p}_1)$	n Var $(\hat{p}_2)$	n Var $(\tilde{p}_1)$	n Var $(\tilde{p}_2)$		
29	0.05	0	0	1	0.5	0.0063 (0.9937)	1.050	0.5218	3.470	1.542	2.158	0.3688
30	0.10	0	0	1	0.5	0.0155 (0.9845)	1.607	0.7906	3.468	1.541	2.101	0.5796
31	0.30	0	0	1	0.5	0.0574 (0.9426)	3.207	1.506	3.680	1.635	1.956	1.171
32	0.50	0	0	1	0.5	0.1003 (0.8997)	4.438	1.972	4.438	1.972	1.999	1.563
33	0.70	0	0	1	0.5	0.1425 (0.8575)	5.473	2.281	5.990	2.661	2.251	1.812
34	0.90	0	0	1	0.5	0.1839 (0.8161)	6.361	2.464	8.436	3.748	2.702	1.937
35	0.95	0	0	1	0.5	0.1941 (0.8059)	6.563	2.491	9.185	4.081	2.846	1.949
36	0.05	0	0	1	3	0.2624 (0.7376)	3.407	1.097	5.858	2.474	2.133	0.8724
37	0.10	0	0	1	3	0.2485 (0.7515)	3.305	1.106	5.280	2.229	1.966	0.8923
38	0.30	0	0	1	3	0.1918 (0.8082)	2.830	1.079	3.361	1.419	1.438	0.9045
39	0.50	0	0	1	3	0.1346 (0.8654)	2.247	0.9488	2.247	0.9488	1.096	0.8094
40	0.70	0	0	1	3	0.0773 (0.9227)	1.538	0.7043	1.988	0.8394	0.9432	0.6034
41	0.90	0	0	1	3	0.0223 (0.9777)	0.6453	0.3153	2.187	0.9233	1.048	0.2672
42	0.95	0	0	1	3	0.0099 (0.9901)	0.3677	0.1820	2.273	0.9596	1.100	0.1532

That is,

$$s = \int_{-\infty}^{\zeta(s)} dH(x) = \int_{-\infty}^{\zeta(s)} h(x)\,dx$$

The corresponding sample quantile, $z(s)$, is defined as follows: If the sample values are arranged in non-decreasing order of magnitudes

$$x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(n)}$$

then $x_{(i)}$ is called the ith *order statistic*, and

$$z(s) = x_{([ns]+1)}$$

where $[ns]$ denotes the greatest integer $\leqq ns$.

Reference 1 shows that if $h(x)$ is differentiable in some neighborhood of each quantile value considered, the joint distribution of any number of quantiles is asymptotically normal as $n \to \infty$ and that, asymptotically,

$$E(z(s)) = \zeta(s)$$

$$\text{Var}(z(s)) = \frac{s(1-s)}{nh^2(\zeta(s))}$$

$$\rho_{12} = \left[\frac{s_1(1-s_2)}{s_2(1-s_1)}\right]^{1/2}$$

where ρ_{12} is the correlation between $z(s_1)$ and $z(s_2)$, $s_1 < s_2$.

Throughout the remainder of this article, $F(x)$ and $f(x) = F'(x)$ will denote the distribution function and density function, respectively, of the standard distribution. That is,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

where

$$f(x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right)$$

Thus, the density function of a mixture of two normal distributions can be written as

$$g(x) = \frac{p}{\sigma_1} f\left(\frac{x-\mu_1}{\sigma_1}\right) + \frac{(1-p)}{\sigma_2} f\left(\frac{x-\mu_2}{\sigma_2}\right) \qquad (1)$$

and the population quantile $\zeta(s)$ can be defined as

$$s = pF\left[\frac{\zeta(s)-\mu_1}{\sigma_1}\right] + (1-p)F\left[\frac{\zeta(s)-\mu_2}{\sigma_2}\right] \qquad (2)$$

Since we are assuming a large sample size, the asymptotic distribution of the sample quantiles will be assumed.

3. Estimators of p Using Quantiles for $\mu_1 \neq \mu_2$

In Eq. (2), $\zeta(s)$ is defined uniquely for a fixed value of s. This relationship provides a simple estimator for p using

one quantile. Replacing $\zeta(s)$ in Eq. (2) by the corresponding sample quantile $z(s)$ and solving for p, one obtains

$$\hat{p}_1 = \frac{s - F\left[\frac{z(s) - \mu_2}{\sigma_2}\right]}{F\left[\frac{z(s) - \mu_1}{\sigma_1}\right] - F\left[\frac{z(s) - \mu_2}{\sigma_2}\right]} \tag{3}$$

which is easy to compute from a table of the standard normal distribution.

The estimator $\hat{p}_1$ is asymptotically unbiased and its asymptotic variance is given by

$$\mathrm{Var}(\hat{p}_1) = \left[\frac{\partial \hat{p}_1}{\partial \zeta(s)}\right]^2 \mathrm{Var}(z(s))$$

where $\partial \hat{p}_1 / \partial \zeta(s)$ denotes the partial derivative $\partial \hat{p}_1 / \partial z(s)$ evaluated at $z(s) = E(z(s)) = \zeta(s)$. Since $\mathrm{Var}(\hat{p}_1)$ depends upon the value of p as well as upon the parameters of both the normal distributions, the optimum value of s [i.e., the value of s that minimizes $\mathrm{Var}(\hat{p}_1)$] cannot be determined if one has no knowledge of p_1. However, the optimum s can be determined with little difficulty once p is known. Column 7 of Table 1 gives this optimum value for all cases considered. Column 8 of Table 1 gives $\mathrm{Var}(p_1^*)$ when this value is used.

If the order of the single quantile to be used in estimating p must be specified in advance, a reasonable choice is to set $s = 0.5$. Column 9 of Table 1 gives the variances of the estimator $\hat{p}_1$ for this choice of s. If, however, one can choose s after the parameters are known, a generally better procedure is to use the value of s that gives optimum results for $p = 0.5$. Column 10 of Table 1 gives $\mathrm{Var}(\tilde{p}_1)$ when this procedure is adopted.

The mean of a mixture of two normal distributions is given by

$$\mu = E(x) = p\mu_1 + (1 - p)\mu_2 \tag{4}$$

Solving for p in Eq. (4), one has

$$p = \frac{\mu - \mu_2}{\mu_1 - \mu_2}$$

If one now estimates μ using quantiles (obtaining $\hat{\mu}$), an estimator for p using quantiles is given by

$$\hat{p} = \frac{\hat{\mu} - \mu_2}{\mu_1 - \mu_2} \tag{5}$$

Optimum unbiased quantile estimators of the mean and standard deviations of a normal distribution are derived under various conditions in Ref. 2. The estimators of the mean, which are linear combinations of pairs of symmetric quantiles, are relatively insensitive to deviations from normality in the sense that they are unbiased when used to estimate the mean of any distribution whatever with a density function symmetric about its mean. In fact, for asymmetric distributions with the type of density function given by $g(x)$, the bias is small if at least several pairs of quantiles are used. In particular, a suboptimum estimator of the mean using six quantiles is derived in Ref. 2 where the orders of the quantiles are chosen for the purpose of estimating the mean and standard deviation using the same quantiles. This estimator is given by

$$\begin{aligned}\hat{\mu}_6 = {} & 0.0497\,[z(0.0231) + z(0.9769)] \\ & + 0.1550\,[z(0.1180) + z(0.8820)] \\ & + 0.2953\,[z(0.3369) + z(0.6631)]\end{aligned}$$

Using $\hat{\mu}_6$ in Eq. (4) gives the estimator $\hat{p}_6$. The expected variances and values of $\hat{p}_6$, given by

$$\mathrm{Var}(\hat{p}_6) = \frac{\mathrm{Var}(\hat{\mu}_6)}{(\mu_1 - \mu_2)^2}$$

$$E(\hat{p}_6) = \frac{E(\hat{\mu}_6) - \mu_2}{\mu_1 - \mu_2}$$

were computed for all cases and are shown in columns 11 and 12, respectively, of Table 1. It can be seen from column 12 that, except for case 9, the bias is not excessive. From column 11, it can also be seen that, in most cases, the variance of $\hat{p}_6$ is less than those of the two previous one-quantile estimators.

In some cases, a better estimate can be obtained by averaging the one- and six-quantile estimates obtaining either

$$\hat{p}_7 = \tfrac{1}{2}(\hat{p}_1 + \hat{p}_6)$$

or

$$\tilde{p}_7 = \tfrac{1}{2}(\tilde{p}_1 + \tilde{p}_6)$$

The asymptotic variances of $\hat{p}_7$ and $\tilde{p}_7$ were computed and are given in columns 13 and 14 of Table 1, respectively. It can be seen that, in almost all cases, either the six- or seven-quantile estimator, has a smaller variance than the corresponding one-quantile estimator. Thus, it remains to be decided when to use a seven- rather than a six-quantile estimator. If one divides the 28 cases into

four blocks, as shown in Table 1, it is readily seen that, for fixed values of σ_1 and σ_2, if $\mu_1 - \mu_2$ is sufficiently small, the seven-quantile estimator should be used. On the other hand, if $\mu_1 - \mu_2$ is sufficiently large, the six-quantile estimator should be used. It is then reasonable to infer that, for some range of values of $\mu_1 - \mu_2$, it makes very little difference, practically speaking, which estimator is used.

An estimator using m quantiles can be constructed as a linear combination of one-quantile estimators as follows:

$$\tilde{p}_m = \sum_{i=1}^{m} \alpha_i \hat{p}_i \tag{6}$$

where

$$\sum_{i=1}^{m} \alpha_i = 1$$

and $\hat{p}_i$ denotes the one-quantile estimator using the quantile of order s_i. For a given value of p, one can determine, theoretically, the α_i and s_i that will minimize Var $(\tilde{p}_m)$. Increasing m will decrease this minimum variance. However, in practical situations where one can, at best, optimize with respect to only one value of p, say $\bar{p}$, and then use the resulting estimator no matter what p is, the results for values of p other than $\bar{p}$ may be very poor. In the event that the order of the quantiles must be specified in advance, the probability of getting poor estimates increases sharply. Moreover, in this case, increasing the number of quantiles almost ensures one of getting poor results. The estimator proposed in SPS 37-32, Vol. IV is of the type given in Eq. (6) with $m = 4$ and $\alpha_i = 1/4$ $(i = 1, 2, 3, 4)$. The variances of these estimators were computed for all cases and are shown in column 15 of Table 1.

The asymptotic variance of the maximum-likelihood (ML) estimator, denoted by p^*, is given by

$$\text{Var}\,(p^*) = -\frac{1}{nE\left[\dfrac{\partial^2}{\partial p^2} \ln g\,(x)\right]}$$

In order to show how "good" the quantile estimators are compared to the best possible asymptotically-unbiased estimator using all the sample values, the Var (p^*) were computed for all cases and are given in column 16 of Table 1. It is interesting to note that some of the biased quantile estimators have smaller variances than the corresponding ML estimators (but larger square errors).

4. Estimators of p Using Quantiles for $\mu_1 = \mu_2$

If $\mu_1 = \mu_2 = \mu$, it can be seen from Eq. (1) that

$$g\,(\mu + x) = g\,(\mu - x)$$

so that $g\,(x)$ is symmetric about $x = \mu$, and $E\,(x) = \mu$, independent of p. Moreover, in the estimator using one quantile given by Eq. (3); namely,

$$\hat{p}_1 = \frac{s - F\left[\dfrac{z\,(s) - \mu_2}{\sigma_2}\right]}{F\left[\dfrac{z\,(s) - \mu_1}{\sigma_1}\right] - F\left[\dfrac{z\,(s) - \mu_2}{\sigma_2}\right]}$$

$s = 0.5$ cannot be used since $\zeta\,(0.5) = \mu$. However, due to symmetry, if for a given value of p, s_0 is optimum, then $1 - s_0$ is also optimum. Column 7 of Table 2 gives the two values of opt s for all cases, column 8 gives the variances of the estimators using one of the optimum values of s, and column 9 gives the variances when each is used and the results averaged. Columns 10 and 11 give the variances of the one- and two-quantile estimators, respectively, if one uses for each case the optimum values of s for $p = 0.5$. The same procedure holds that was suggested in the case $\mu_1 \neq \mu_2$ if the orders of the quantiles can be chosen after one knows the values of the parameters.

In order to assist in making a decision as to the specification of the orders of the quantiles when this decision must be made in advance, a study was made of the behavior of the optimum values of s for $p = 0.5$ as the ratio of the standard deviation varies, since these optimum values, and the variance of the estimators based on them, depend only on this ratio. Figure 1 is a plot of the larger of the two values of opt s as $\sigma_2/\sigma_1\,(\sigma_1/\sigma_2)$ increases from unity.

Column 12 of Table 2 gives the variances of the four-quantile estimators proposed in SPS 37-32, Vol IV; column 13 gives the variances of the ML estimators. It should be observed that, no matter which estimator is used, if the computed estimate of p is negative or greater than one, the estimate should be taken as zero or one, respectively. These end effects were not taken into account in the above analysis since, for large sample sizes, they would be significant only for values of p close to zero or one (the estimators would be biased but have smaller variances).

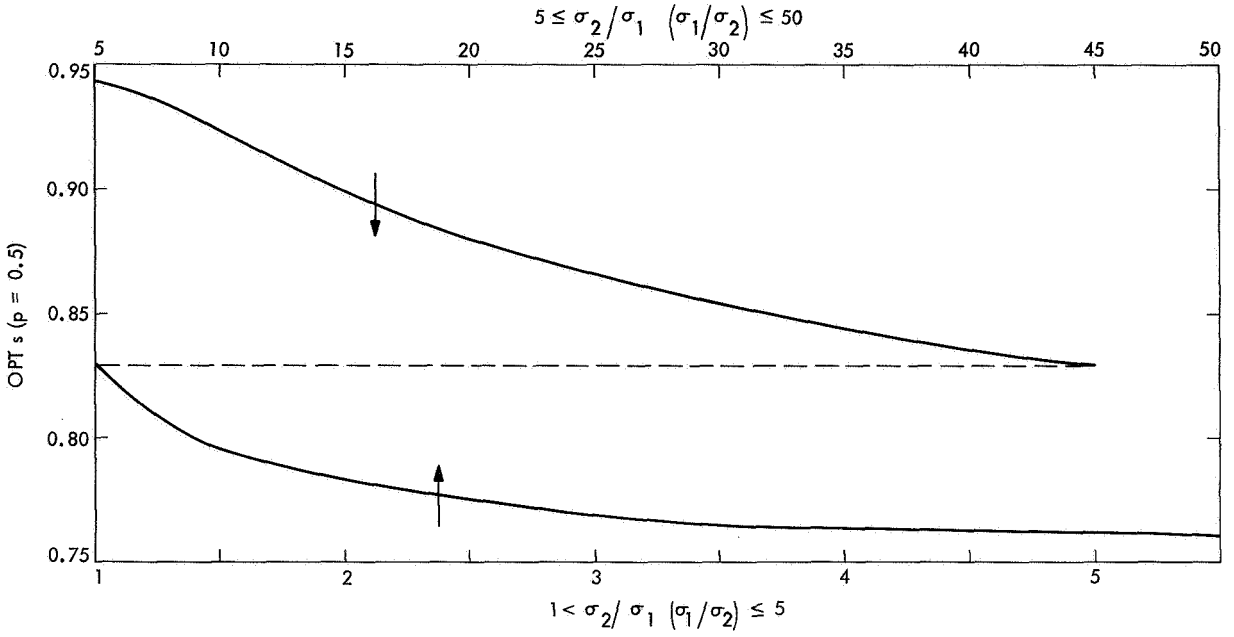


Fig. 1. Larger of two values of opt s for $p = 0.5$, $\mu_1 = \mu_2$

5. Estimating p From Real Data Using Quantiles

A table of random digits can be used to obtain a sample quantile $z(s)$ of order s from a sample of size n drawn from a population with distribution function $G(x)$. A set of n k-digit numbers is drawn from the table and the sample quantile (v/s) of order s is determined from the sample. The desired sample quantile $z(s)$ of $G(x)$ is obtained by solving for $z(s)$ in the equation

$$[v(s) + 0.5]\, 10^{-k} = G[z(s)]$$

This procedure was adopted with $n = 256$ in order to obtain sample quantiles necessary for estimating p for cases 2, 10, 18, 26, 31, and 40. The results for each case are as follows:

(1) For case 2 with $p = 0.1$:

$$\hat{p}_1 = 0.0137 \qquad \hat{p}_6 = 0.1178$$
$$\tilde{p}_1 = 0.1022 \qquad \hat{p}_7 = 0.0658$$
$$\tilde{p}_4 = 0.2391 \qquad \tilde{p}_7 = 0.1100$$

(2) For case 10 with $p = 0.3$:

$$\hat{p}_1 = 0.3275 \qquad \hat{p}_6 = 0.2765$$
$$\tilde{p}_1 = 0.3271 \qquad \hat{p}_7 = 0.3020$$
$$\tilde{p}_4 = 0.3269 \qquad \tilde{p}_7 = 0.3013$$

(3) For case 18 with $p = 0.5$:

$$\hat{p}_1 = 0.5068 \qquad \hat{p}_6 = 0.4274$$
$$\tilde{p}_1 = 0.5172 \qquad \hat{p}_7 = 0.4671$$
$$\tilde{p}_4 = 0.2733 \qquad \tilde{p}_7 = 0.4723$$

(4) For case 26 with $p = 0.7$:

$$\hat{p}_1 = 0.6988 \qquad \hat{p}_6 = 0.7317$$
$$\tilde{p}_1 = 0.7309 \qquad \hat{p}_7 = 0.7153$$
$$\tilde{p}_4 = 0.7827 \qquad \tilde{p}_7 = 0.7313$$

(5) For case 31 with $p = 0.3$ and $\mu_1 = \mu_2 = 0$:

$$\tilde{p}_2 = 0.2315 \qquad \tilde{p}_4 = 0.2868$$

(6) For case 40 with $p = 0.7$ and $\mu_1 = \mu_2 = 0$:

$$\tilde{p}_2 = 0.7678 \qquad \tilde{p}_4 = 0.7368$$

References

1. Cramer, H., *Mathematical Methods of Statistics*, pp. 367–370, Princeton University Press, Princeton, N.J., 1946.

2. Eisenberger, I., and Posner, E. C., "Systematic Statistics Used for Data Compression of Space Telemetry," *J. Am. Stat. Assoc.*, Vol. 60, pp. 97–133, Mar. 1965. Also available as Technical Report 32-510, Jet Propulsion Laboratory, Pasadena, Calif., Oct. 1, 1963.

B. Epsilon Entropy of Gaussian Processes,

E. C. Posner, E. R. Rodemich, and H. Rumsey, Jr.

1. Introduction

This article shows that the epsilon entropy of any mean-continuous gaussian process on $L_2[0,1]$ is finite for all positive ϵ. The epsilon entropy of such a process is defined as the infimum of the entropies of all partitions of $L_2[0,1]$ by measurable sets of diameter at most ϵ, where the probability measure on L_2 is the one induced by the process. Fairly tight upper and lower bounds are found for the epsilon entropy as $\epsilon \to 0$ in terms of the eigenvalues of the process. The full article on this subject has been submitted to the *Annals of Mathematical Statistics*; proofs are omitted in this summary.

Let $x(t)$ be a mean-continuous gaussian process with mean zero on the unit interval. Its covariance function $R(s,t)$ is then a continuous function on the unit square and its eigenfunction expansion

$$R(s,t) = \sum_{n=1}^{\infty} \lambda_n \phi_n(s)\phi_n(t)$$

converges uniformly (Ref. 1, p. 478). The eigenvalues $\lambda_n = \sigma_n^2$ are non-negative numbers with $\sum \lambda_n < \infty$. The eigenfunctions $\{\phi_n(t)\}$ are continuous and form an orthonormal system in $L_2[0,1]$.

If we assume the process is measurable (Ref. 1, p. 502), then the paths are functions in $L_2[0,1]$ and we can take $L_2[0,1]$ as the probability space. This gives a measure on the Borel sets of $L_2[0,1]$, which is uniquely determined by the covariance function.

One way of determining this measure is to take our process to be the sum of the Karhünen–Loéve series

$$x(t) = \sum_{n=1}^{\infty} x_n \phi_n(t),$$

where the $\{x_n\}$ are independent gaussian random variables, with

$$Ex_n = 0, \qquad Ex_n^2 = \lambda_n.$$

If we take Ω_0 to be the product space of the x_n, this series converges in

$$L_2\{[0,1] \times \Omega_0\}.$$

The subset Ω of Ω_0, on which $\sum x_n^2 < \infty$, has probability 1 and is a Hilbert space under the norm

$$\|\{x_n\}\|^2 = \Sigma x_n^2.$$

The map $\{x_n\} \to x(t)$ is an isometry of Ω onto the subspace Ω^* of $L_2[0,1]$ generated by the eigenfunctions. This mapping induces a measure in L_2 that is concentrated on the subspace Ω^*.

For $\epsilon > 0$, we define an *ϵ-partition* of $X = L_2[0,1]$ (with the given probability measure) to be a finite or denumerable collection of disjoint ϵ-sets (Borel sets of diameter $\leqq \epsilon$) that cover a subset of L_2 of measure 1. More generally, an *$\epsilon;\delta$-partition* is such a collection of sets that omits a subset of L_2 with measure no greater than δ. Let such a partition U consist of sets U_i of measures

$$p_i = \mu(U_i), \qquad \Sigma p_i = 1.$$

Then the *entropy* of U is defined as the entropy of the discrete distribution $p_1, p_2, \cdots$:

$$H(U) = \sum p_i \log \frac{1}{p_i}.$$

(We use logarithms to the base e for convenience.)

The *ϵ-entropy* of X, $H_\epsilon(X)$, is the infimum of $H(U)$ over all ϵ-partitions U of X. The *$\epsilon;\delta$-entropy* $H_{\epsilon;\delta}(X)$ is defined similarly as the infimum over all $\epsilon;\delta$-partitions. If $U = \{U_i\}$ is an $\epsilon;\delta$-partition with

$$\mu(U_i) = p_i, \qquad \Sigma p_i = m \geqq 1 - \delta,$$

then

$$H(U) = \sum \frac{p_i}{m} \log \frac{m}{p_i}.$$

These concepts were introduced in a more general setting in Ref. 2. It was shown there that $H_{\epsilon;\delta}(X)$ is finite for $\delta > 0$.

Note that any partition U can be restricted to the subspace Ω^* of $L_2[0,1]$ on which the measure is concentrated. This subspace can be identified with the Hilbert space Ω of sequences $\{x_n\}$ where the coordinates are independent gaussian random variables. Thus, the ϵ-entropy of the process depends only on the measure

on Ω, and not on how Ω is embedded in $L_2[0,1]$. That is, the ϵ-entropy is a function only of the eigenvalues $\{\lambda_n\}$.

The purpose of these definitions is to make precise the notion of data compression. Thus, $H_\epsilon(X)$ is the channel capacity needed to describe sample functions of X to within ϵ in L_2-norm with probability 1.[1] Reference 2 showed that for mean-continuous, but not necessarily gaussian, processes X on the unit interval, the following holds:

(1) $H_\epsilon(X)$ is finite for every $\epsilon > 0$, provided the eigenvalues λ_n of X (written, as usual, in non-increasing order) satisfy

$$\Sigma n\lambda_n < \infty.$$

(2) If, on the other hand,

$$\Sigma n\lambda_n = \infty,$$

then there exists a mean-continuous process X on the unit interval such that, for every $\epsilon > 0$ no matter how large, $H_\epsilon(X)$ is infinite.

One of the principal results of this article is that, if X is a gaussian process, $H_\epsilon(X)$ is finite for every positive ϵ no matter how small and no matter how slowly the eigenvalues λ_n approach 0 (as long, of course, as $\sum \lambda_n < \infty$). Another is that $H_\epsilon(X)$ is a continuous function of ϵ for a fixed mean-continuous gaussian process X on the unit interval. We also find upper and lower bounds for $H_\epsilon(X)$ that are reasonably tight as $\epsilon \to 0$. These bounds are given in terms of the eigenvalues of the process.

If the only partitions of $L_2[0,1]$ allowed are products of partitions of each eigenfunction axis, the resulting entropy, called *product ϵ-entropy*, need not be finite.[2] In fact, a necessary and sufficient condition that product ϵ-entropy be finite for one (or all) positive epsilon is that the "entropy of the eigenvalues"

$$\Sigma \lambda_n \log \frac{1}{\lambda_n}$$

be finite. The reason that $H_\epsilon(X)$ is always finite for a gaussian process when $\epsilon > 0$ is that the partitions used to show finiteness of $H_\epsilon(X)$ involve finite-dimensional subspaces of $L_2[0,1]$ generated by an arbitrarily large finite number of eigenfunctions. As we shall see, the partitions used on these subspaces differ essentially from products of one-dimensional partitions.

2. Continuity of $H_\epsilon(X)$

In this subsection, it will be shown that if X is a mean-continuous gaussian process and $\epsilon > 0$, then $H_\epsilon(X)$ is continuous in ϵ; we shall assume the result, to be proved later in the article, that $H_\epsilon(X)$ is finite for every positive ϵ. Since the continuity of H_ϵ in ϵ is not used subsequently, there is no loss in the assumption.

Reference 2 shows that if the measure μ on X has no atoms, then

$$H_\epsilon(X) \to \infty \text{ as } \epsilon \to 0.$$

Since X has at least one positive eigenvalue (because we assumed that $R(s,t)$ is not identically 0), μ is non-atomic. Thus, if $H_0(X)$ is interpreted as $+\infty$, $H_\epsilon(X)$ is continuous even at 0.

Continuity from above in ϵ was proved in Ref. 2. Thus, the only thing that remains to be shown here is that $H_\epsilon(X)$ is continuous from below (for $\epsilon > 0$). This is proved in Theorem 1 in a more general context: the $\epsilon;\delta$-entropy $H_{\epsilon;\delta}(X)$ is continuous from below in ϵ for $\delta \geqq 0$. The following required lemma is of interest in its own right.

Lemma 1. If X is the Hilbert space of a mean-continuous gaussian process on the unit interval, the set of extreme points of any convex set in X has measure zero.

We can now state Theorem 1.

Theorem 1. The $\epsilon;\delta$-entropy of a gaussian process on $L_2[0,1]$ is continuous from below in ϵ for fixed δ.

3. Lower Bounds for $H_\epsilon(X)$

In this subsection, we derive some lower bounds for the ϵ-entropy of a mean-continuous gaussian process on the unit interval.

[1] Posner, E. C., and Rodemich, E. R., "Epsilon Entropy and Data Compression" (in preparation).

[2] Posner, E. C., Rodemich, E. R., and Rumsey, H, Jr., "Product Entropy of Gaussian Distributions," (submitted to *Ann. Math. Statist.*).

First note that for any ϵ-partition $U = \{U_j\}$ of X, if $U(x)$ denotes the set U_j containing x, we have

$$H(U) = E \log \left\{ \frac{1}{\mu} [U(x)] \right\}. \tag{1}$$

This expression is decreased if we replace $U(x)$ by the sphere of radius ϵ about x. It follows that

$$H_\epsilon(X) \geqq E_y \log \left[\frac{1}{\mu} \{x | d(x, y) \leqq \epsilon\} \right], \tag{2}$$

where d denotes the metric in X and E_y indicates that the expectation is to be taken with respect to y. The first lower bound to be derived is a lower bound for the right side of Ineq. (2).

First, we need the upper bound for

$$\mu \{x | d(x, y) \leqq \epsilon\}$$

obtained from Lemma 2.

Lemma 2. If Z is a non-negative random variable with characteristic function f, then for a and $b \geqq 0$,

$$\Pr \{Z \leqq a\} \leqq \exp(ba) f(ib).$$

The next lemma gives an upper bound for the probability of the ϵ-sphere about a fixed point y.

Lemma 3. Let a mean-continuous gaussian process X have eigenvalues $\{\lambda_n\}$. Then, in the L_2 norm d, for any fixed $y \in X$, we have

$$\mu \{x | d(x, y) \leqq \epsilon\} \leqq \inf_{b \geqq 0} \frac{\exp(b\epsilon^2)}{[\prod_n (1 + 2b\lambda_n)]^{1/2}} \exp\left[- \sum \frac{b y_n^2}{1 + 2b\lambda_n} \right].$$

Using the estimate of Lemma 3 in Eq. (2), we arrive at the lower bound

$$H_\epsilon(X) \geqq E_y \sup_{b \geqq 0} \left\{ - b\epsilon^2 - \frac{1}{2} \sum \log(1 + 2b\lambda_n) + \sum \frac{b y_n^2}{1 + 2b\lambda_n} \right\}. \tag{3}$$

The disadvantage of this estimate is that a set of diameter ϵ containing y has been replaced by a sphere of diameter 2ϵ. Another lower bound will be derived that does not have this disadvantage. We first prove that the sphere of radius $\epsilon/2$ about the origin has at least as much probability as any set of diameter ϵ in X, a result of independent interest. Actually, strict inequality can be proved but is not needed.

Lemma 4. Let X be the Hilbert space of a gaussian process, and V any measurable set in X of diameter at most ϵ. Then

$$\mu(V) \leqq \mu [S_{\epsilon/2}(0)],$$

where $S_{\epsilon/2}(0)$ is the sphere of radius $\epsilon/2$ about the origin.

Applying Lemma 4 to Eq. (1), we get

$$H_\epsilon(X) \geqq \log \left\{ \frac{1}{\mu} [S_{\epsilon/2}(0)] \right\}. \tag{4}$$

The following theorem presents two lower bounds: $L_\epsilon(X)$, derived from Eq. (3), and $M_\epsilon(X)$, derived from Eq. (4). Note that $L_\epsilon(X)$ is always weaker. It is of interest mainly because of Theorem 4 (*Subsection 4*),which bounds $H_\epsilon(X)$ from above in terms of $L_\epsilon(X)$.

Theorem 2. Let X be a mean-continuous gaussian process with eigenvalues $\{\lambda_n\}$. Define $b = b(\epsilon) \geqq 0$ by

$$\left.\begin{aligned} \sum \frac{\lambda_n}{1 + b\lambda_n} = \epsilon^2, \qquad & \Sigma \lambda_n > \epsilon^2 \\ b = 0, \qquad & \Sigma \lambda_n \leqq \epsilon^2 \end{aligned}\right\} \tag{5}$$

Put

$$L_\epsilon(X) = \frac{1}{2} \Sigma \log [1 + \lambda_n b(\epsilon)] \tag{6}$$

and

$$M_\epsilon(X) = \frac{1}{2} \sum \log \left[1 + \lambda_n b\left(\frac{\epsilon}{2}\right) \right] - \frac{1}{8} \epsilon^2 b\left(\frac{\epsilon}{2}\right). \tag{7}$$

Then

$$H_\epsilon(X) \geqq M_\epsilon(X) \geqq L_\epsilon(X).$$

Next, we give an improvement on the lower bound $M_\epsilon(X)$ that is difficult to use in general, but will be evaluated for special processes in *Subsection 4.* This is based on the following lemma.

Lemma 5. Let $x_1, \cdots, x_n$ be independent gaussian random variables with

$$Ex_j = 0, \qquad Ex_j^2 = \lambda_j > 0, \qquad j = 1, \cdots, n.$$

Consider the n-dimensional probability space X of $x_1, \cdots, x_n$ under the euclidian metric d. Let

$$a = (a_1, \cdots, a_n)$$

be a fixed point of X with $d(a, 0) > \epsilon$ and $S_\epsilon(a)$ be the set of points x with $d(x, a) \leqq \epsilon$. There is a translation

$$x \to x' = x + b$$

such that, for any x in $S_\epsilon(a)$, the probability density $p(x)$ satisfies the inequality

$$\frac{p(x')}{p(x)} \geqq \exp\left[\frac{1}{2}\sum_{k=1}^{n}\frac{\lambda_k a_k^2 q^2}{(\epsilon + \lambda_k q)^2}\right], \tag{8}$$

where q is the unique positive solution of

$$\sum_{k=1}^{n}\frac{a_k^2}{(\epsilon + \lambda_k q)^2} = 1. \tag{9}$$

The improvement to the lower bound $M_\epsilon(X)$ can now be given.

Theorem 3. Let X be the Hilbert space of a mean-continuous gaussian process on $[0, 1]$. Define the non-negative random variable $q = q(x)$ by

$$q = 0, \qquad \|x\| \leqq \epsilon,$$

and, for $\|x\| > \epsilon$, by

$$\sum\frac{x_k^2}{(\epsilon + \lambda_k q)^2} = 1, \tag{10}$$

where $\{\lambda_k\}$ are the eigenvalues of the process. Then

$$H_\epsilon(X) \geqq M_\epsilon(X) + \frac{1}{2}\sum E\frac{\lambda_k x_k^2 q^2}{(\epsilon + \lambda_k q)^2}. \tag{11}$$

A result of A. N. Kolmogorov's [Ref. 3, Eq. (12)] implies that the ϵ-entropy has a lower bound

$$H_\epsilon(X) \geqq Y_\epsilon(X) = \frac{1}{2}\sum_{n=1}^{N}\log\frac{\lambda_n}{\theta^2},$$

where N and θ are defined (for $\epsilon^2 \leqq \Sigma\lambda_n$) by the equation

$$\epsilon^2 = \sum\min(\theta^2, \lambda_n) \equiv N\theta^2 + \sum_{n \geqq N+1}\lambda_n.$$

A simple, but lengthy, variational argument shows that

$$L_\epsilon(X) \geqq Y_\epsilon(X)$$

with equality only in the case where $\lambda_1 = \lambda_2 = \cdots = \lambda_n$ and $\lambda_n = 0$ for $n > N$. (Kolmogorov's bound is actually a bound for the problem of communicating X holding the expected square error to within ϵ^2.) In the finite-dimensional case, a result in Footnote 1 gives an even more precise lower bound for $H_\epsilon(X)$. Hence, we do not have to use Kolmogorov's bound.

4. An Upper Bound for $H_\epsilon(X)$

In Theorem 4, we bound the ϵ-entropy of a gaussian process from above asymptotically in terms of the quantity $L_\epsilon(X)$ introduced in Theorem 2. The method of proof uses a special partition of X. To estimate its entropy, we need some preliminary lemmas which give bounds on the entropy of a finite dimensional gaussian distribution. The first of these lemmas bounds the probability of being outside a spherical shell centered on the sphere of radius $n^{1/2}$ for the joint distribution of n independent unit normal variables.

Lemma 6. Let X be the n-dimensional euclidian space of n independent normal random variables of mean zero and variance 1. Let S be the spherical shell

$$|n^{1/2} - (\Sigma x_j^2)^{1/2}| < d,$$

where $0 < d < n^{1/2}$, and

$$\nu(n, d) = 1 - \mu(S).$$

Then there is a universal constant C_1 such that

$$\nu(n, d) < \frac{C_1 \exp(-d^2)}{d}.$$

The next lemma bounds the ϵ-entropy of the unit $(n-1)$-sphere with the uniform probability distribution.

Lemma 7. Let X be the unit sphere in n-dimensional euclidian space with a uniform probability distribution. If β and γ are positive numbers, then for $\epsilon > 0$,

$$H_\epsilon(X) < (1+\beta)\, n \log^+ \frac{2+\gamma}{\epsilon} + C_4(\beta, \gamma),$$

where C_4 depends only on β and γ.

The next lemma bounds the ϵ-entropy of euclidian n-space under the joint distribution of n independent gaussian random variables.

Lemma 8. Let X be the n-dimensional euclidian space of n independent normal random variables of mean zero and variances $\lambda_1, \cdots, \lambda_n$. Let α be a number between 0 and 1, and for

$$0 < (1-\alpha)\epsilon < 2(n\lambda)^{1/2}$$

set

$$\nu = \nu\{n, (1-\alpha)\epsilon/[2(\lambda)^{1/2}]\},$$

where λ is the maximum of $\lambda_1, \cdots, \lambda_n$. Then, there is a universal constant C_2 such that

$$H_\epsilon(X) < (1+\beta)\, n \log^+ \frac{(2+\gamma)(n\lambda)^{1/2}}{\alpha\epsilon} + n\nu \log^+ \frac{(n\lambda)^{1/2}}{\epsilon} + C_4(\beta, \gamma) + C_2(1+n\nu)$$

if β, γ are any positive numbers and $C_4(\beta, \gamma)$ is the constant of Lemma 7.

An alternate upper bound is obtained in Lemma 9. The bounds of both Lemmas 8 and 9 are needed in Theorem 4.

Lemma 9. Let X be the n-dimensional euclidian space of n independent normal random variables of mean zero with variances $\lambda_1, \cdots, \lambda_n$, and $\lambda = \max(\lambda_1, \cdots, \lambda_n)$. There is a universal constant C_3 such that, if $\epsilon > 2(n\lambda)^{1/2}$,

$$H_\epsilon(X) < C_3 n^{3/2} \left[g \exp\left(\frac{1-g^2}{2}\right)\right]^n,$$

where $g = \epsilon/[2(n\lambda)^{1/2}]$.

Now we are ready to state the upper bound of Theorem 4.

Theorem 4. Let m be any positive number less than ½. Then

$$H_\epsilon(X) \leqq L_{m\epsilon}(X)\,[1 + o(1)]$$

as $\epsilon \to 0$. In particular, $H_\epsilon(X)$ is finite for X a mean-continuous gaussian process on the unit interval and $\epsilon > 0$.

The idea of the proof is as follows: For any $\delta > 0$, X will be broken up as the product of a sequence of finite-dimensional spaces $\{X_k\}$ in a way that depends on δ as well as on ϵ, so that, for the optimum product partition U,

$$H(U) \leqq (1+\delta)\, L_{m\epsilon}(X)\,[1 + o(1)].$$

The meshes $\{\epsilon_k\}$ of the component partitions are suggested by Definition (5). The most natural product partitions to try are one-dimensional product partitions, where we take

$$\epsilon_k^2 = \frac{A^2 \lambda_k}{1 + b\lambda_k} \tag{12}$$

for the partition of the kth coordinate. It turns out that this does not always work. In fact, if the eigenvalues decrease slowly enough, there are no one-dimensional product ϵ-partitions with finite entropy (Footnote 2) even if $\Sigma \lambda_k$ is finite. However, for small ϵ, this is the best way to handle the large eigenvalues, and there is a first range of k in which one-dimensional subspaces are used. Beyond this point, the dimensions of the subspaces are consecutive integers beginning with 1. This sequence of subspaces is also split up into two ranges; up to a certain point, the entropy of the subspace is estimated by Lemma 8. Beyond this point, Lemma 9 is applied.

5. Entropy of Special Processes—the Wiener Process

By the Wiener process, we mean that gaussian process on $[0, 1]$ that has covariance function $R(s, t) = \min(s, t)$,

and

$$\lambda_n = \frac{1}{\pi^2\left(n-\frac{1}{2}\right)^2}, \qquad n = 1, 2, \cdots. \tag{13}$$

This can be treated as a special case of a more general process, such as the solutions of finite-order stochastic differential equations; in such cases, we have

$$\lambda_n \approx An^{-p}, \qquad p > 1. \tag{14}$$

First, we estimate $L_\epsilon(X)$ and $M_\epsilon(X)$ for such processes to get the upper and lower bounds of Theorems 2 and 4. Then, we use the lower bound of Theorem 3 to obtain the best known bounds for this class of processes.

We need to find the asymptotic behavior of b as a function of ϵ, given Ineq. (14) and

$$\sum \frac{\lambda_n}{1+b\lambda_n} = \epsilon^2. \tag{15}$$

Note that $b \to \infty$ as $\epsilon \to 0$. If A_1 is any number greater than A, $\lambda_n \leqq A_1 n^{-p}$ except for a finite number of values of n. Hence,

$$\epsilon^2 < \sum_{n=1}^{\infty} \frac{A_1 n^{-p}}{1+bA_1 n^{-p}} + O(b^{-1}).$$

It is easily shown that, as $b \to \infty$,

$$\sum_{n=1}^{\infty} \frac{A_1 n^{-p}}{1+bA_1 n^{-p}} \sim \int_0^{\infty} \frac{A_1 t^{-p}\,dt}{1+bA_1 t^{-p}} = A_1^{1/p}\, b^{(1/p)-1} \frac{\pi}{p \sin(\pi/p)}.$$

Hence,

$$\epsilon^2 \lesssim A_1^{1/p}\, b^{(1/p)-1} \frac{\pi}{p \sin(\pi/p)} [1 + o(1)].$$

Similarly, if $A_1 < A$, the reverse inequality holds. It follows that

$$\epsilon^2 \sim A^{1/p} b^{(1/p)-1} \frac{\pi}{p \sin(\pi/p)},$$

or

$$b(\epsilon) \sim A^{1/(p-1)} \left(\frac{\pi}{p \sin(\pi/p)}\right)^{p/(p-1)} \epsilon^{-[2p/(p-1)]}. \tag{16}$$

The same type of reasoning applies to the series for $L_\epsilon(X)$. We have by Eq. (6)

$$L_\epsilon(X) = \frac{1}{2}\Sigma \log(1+b\lambda_n) \sim \frac{1}{2}\int_0^{\infty} \log(1+bAt^{-p})\,dt = (bA)^{1/p} \frac{\pi}{2\sin(\pi/p)}.$$

Using Ineq. (16),

$$L_\epsilon(X) \sim B_1 \epsilon^{-[2/(p-1)]},$$

where

$$B_1 = \frac{1}{2} pA^{1/(p-1)} \left(\frac{\pi}{p \sin(\pi/p)}\right)^{p/(p-1)} \tag{17}$$

In applying Theorem 4, the growth rate of $L_\epsilon(X)$ is sufficiently small that we can put $m = 1/2$. Thus, Theorem 4 gives us

$$H_\epsilon(X) \lesssim 2^{2/(p-1)} B_1 \epsilon^{-[2/(p-1)]}. \tag{18}$$

Now $M_\epsilon(X)$ can be quickly evaluated. From Eqs. (6) and (7) and Ineq. (16),

$$M_\epsilon(X) = L_{\epsilon/2}(X) - \frac{1}{8}\epsilon^2 b\left(\frac{\epsilon}{2}\right) \sim L_{\epsilon/2}(X) - \frac{1}{p} 2^{2/(p-1)} B_1 \epsilon^{-[2/(p-1)]},$$

and

$$H_\epsilon(X) \geqq M_\epsilon(X) \sim \frac{p-1}{p} 2^{2/(p-1)} B_1 \epsilon^{-[2/(p-1)]}. \tag{19}$$

In examining the lower bound of Theorem 3, we first state a general lemma that applies to any gaussian process for which the eigenvalues do not decrease too rapidly. It states that, in some sense, the random variable q behaves like the deterministic function $r = r(\epsilon)$, which is the positive solution of

$$\sum \frac{\lambda_n}{(\epsilon + \lambda_n r)^2} = 1, \tag{20}$$

when $\epsilon^2 < \Sigma \lambda_n$. This can be made precise when the eigenvalues satisfy Ineq. (14).

Lemma 10. Let the eigenvalues $\{\lambda_n\}$ (in non-increasing order) of a mean-continuous gaussian process X have the

following property: There is a sequence $n_1 < n_2 < \cdots$ such that

$$\frac{n_{k+1} - n_k}{\log k} \to \infty \tag{21}$$

and

$$\frac{\lambda_{n_{k+1}}}{\lambda_{n_k}} \to 1 \tag{22}$$

as $k \to \infty$. Let δ be given with $0 < \delta < 1$. Then for ϵ sufficiently small, and q as defined in Theorem 3 (Eq. 10), we have

$$\left| \frac{\sum \frac{x_k^2}{(\epsilon + \lambda_k q)^2}}{\sum \frac{\lambda_k}{(\epsilon + \lambda_k q)^2}} - 1 \right| < \delta \tag{23}$$

and

$$\left| \frac{\sum \frac{\lambda_k x_k^2 q^2}{(\epsilon + \lambda_k q)^2}}{\sum \frac{\lambda_k^2 q^2}{(\epsilon + \lambda_k q)^2}} - 1 \right| < \delta \tag{24}$$

except on a set of x of probability less than δ.

Now we shall apply this lemma and Theorem 3 to processes satisfying Ineq. (14).

Theorem 5. If a mean-continuous gaussian process X has eigenvalues

$$\lambda_n \sim An^{-p}, \qquad p > 1,$$

then

$$H_\epsilon(X) \gtrsim A^{1/(p-1)} \left(\frac{\pi}{p \sin(\pi/p)} \right)^{p/(p-1)} \frac{p-1}{2} \{ 2^{2/(p-1)} + p^{-[p/(p-1)]} \} \epsilon^{-[2/(p-1)]}. \tag{25}$$

Proof. First, we use Lemma 10 to estimate the last term of Ineq. (11). On a set of measure $1 - \delta$, we have, for ϵ sufficiently small

$$\sum \frac{\lambda_k}{(\epsilon + \lambda_k q)^2} < (1 - \delta)^{-1}.$$

This sum is asymptotically equal to an integral as $q/\epsilon \to \infty$:

$$\sum \frac{\lambda_k}{(\epsilon + \lambda_k q)^2} \sim \int_0^\infty \frac{At^{-p}\, dt}{(\epsilon + Aqt^{-p})^2} = A^{1/p} q^{(1/p)-1} \epsilon^{-[(1/p)+1]} \frac{\pi}{p^2 \sin(\pi/p)}.$$

Hence,

$$q \gtrsim A^{1/(p-1)} \left[\frac{\pi(1-\delta)}{p^2 \sin(\pi/p)} \right]^{p/(p-1)} \epsilon^{-[(p+1)/(p-1)]}.$$

Also, we have

$$\sum \frac{\lambda_k^2 q^2}{(\epsilon + \lambda_k q)^2} \sim \int_0^\infty \frac{A^2 q^2 t^{-2p}\, dt}{(\epsilon + Aqt^{-p})^2} = \left(\frac{Aq}{\epsilon} \right)^{1/p} \frac{(p-1)\pi}{p^2 \sin(\pi/p)} \gtrsim [A(1-\delta)]^{1/(p-1)} (p-1) \times \left[\frac{\pi}{p^2 \sin(\pi/p)} \right]^{p/(p-1)} \epsilon^{-[2/(p-1)]},$$

off the exceptional set. Then by Ineq. (24),

$$\sum \frac{\lambda_k x_k^2 q^2}{(\epsilon + \lambda_k q)^2} > (1 - \delta) \sum \frac{\lambda_k^2 q^2}{(\epsilon + \lambda_k q)^2} \gtrsim (1 - \delta)^{p/(p-1)} B_2 \epsilon^{-[2/(p-1)]},$$

where

$$B_2 = A^{1/(p-1)} (p-1) \left[\frac{\pi}{p^2 \sin(\pi/p)} \right]^{p/(p-1)}.$$

This asymptotic inequality holds uniformly on a set of measure at least $1 - \delta$. Hence,

$$E \sum \frac{\lambda_k x_k^2 q^2}{(\epsilon + \lambda_k q)^2} \gtrsim (1 - \delta)^{1+[p/(p-1)]} B_2 \epsilon^{-[2/(p-1)]},$$

and letting $\delta \to 0$,

$$\frac{1}{2} E \sum \frac{\lambda_k x_k^2 q^2}{(\epsilon + \lambda_k q)^2} \gtrsim \frac{1}{2} B_2 \epsilon^{-[2/(p-1)]}.$$

Using this estimate for the last term of Ineq. (11), together with the asymptotic form (Ineq. 19) of $M_\epsilon(X)$, we obtain Ineq. (25) and prove Theorem 5.

Corollary. For the Wiener process,

$$\frac{17}{32\,\epsilon^2} \lesssim H_\epsilon(X) \lesssim \frac{1}{\epsilon^2}.$$

Proof. The lower bound results from putting $p = 2$, $A = \pi^{-2}$ in Ineq. (25). The upper bound is Ineq. (18) for this special case. This proves the corollary.

There is no gaussian process X for which we know that $L_{\epsilon/2}(X)$ is not asymptotic to $H_\epsilon(X)$ as $\epsilon \to 0$. Resolution of this question would be extremely interesting.

References

1. Loéve, Michel, *Probability Theory*, Van Nostrand, New York, 1955.
2. Posner, Edward C., Rodemich, Eugene, and Rumsey, Howard, Jr., "Epsilon Entropy of Stochastic Processes," *Ann. Math. Statist.*, pp. 1000–1020, Vol. 38, 1967.
3. Kolmogorov, Andrei N., "The Shannon Theory of Information Transmission in the Case of Continuous Signals," *IRE Trans. Inform. Theory*, Vol. IT-2, pp. 102–108, 1956.