1+24

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

25

Space Programs Summary 37-51, Vol. III

Supporting Research and Advanced Development

For the Period April 1 to May 31, 1968

GPO PRICE $ ____________

CSFTI PRICE(S) $ ____________

Hard copy (HC) 3.00

Microfiche (MF) .65

ff 653 July 65

N68-37397 – N68-37421

(ACCESSION NUMBER)	(THRU)
499	1
(PAGES)	(CODE)
CR-97120	11
(NASA CR OR TMX OR AD NUMBER)	(CATEGORY)

FACILITY FORM 602

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

June 30, 1968

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Space Programs Summary 37-51, Vol. III

Supporting Research and Advanced Development

For the Period April 1 to May 31, 1968

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

June 30, 1968

SPACE PROGRAMS SUMMARY 37-51, VOL. III

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

Preface

The Space Programs Summary is a bimonthly publication that presents a review of engineering and scientific work performed, or managed, by the Jet Propulsion Laboratory for the National Aeronautics and Space Administration during a two-month period. Beginning with the 37-47 series, the Space Programs Summary is 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)

Approved by:

W. H. Pickering

W. H. Pickering, Director
Jet Propulsion Laboratory

Contents

SYSTEMS DIVISION

GUIDANCE AND CONTROL DIVISION

Contents (contd)

Contents (contd)

ENGINEERING MECHANICS DIVISION

ENVIRONMENTAL SCIENCES DIVISION

PROPULSION DIVISION

Contents (contd)

SPACE SCIENCES DIVISION

Contents (contd)

TELECOMMUNICATIONS DIVISION

Contents (contd)

ADVANCED STUDIES

I. Systems Analysis Research

SYSTEMS DIVISION

A. Shadow Equation for a Satellite, *J. Lorell*

1. Introduction

This article discusses computation procedures for finding the shadow entry and exit angles for an artificial satellite of the moon. The results are also applicable to satellites of earth or the planets. To determine whether a given position in space is in sunlight or in shadow is relatively simple; however, the edges of the shadow, i.e., intersection points of an elliptic orbit with a cylindrical shadow, are not so directly computable.

The difficulty lies in the fact that a fourth-degree algebraic equation must be solved. The roots of such equations may be written down immediately using Ferrari's (Cardan's) formula, but the result involves the cube roots of complex numbers—even when the solutions are real.

R. P. Yeremenko (Ref. 1) solves the problem using Ferrari's formula, in spite of the inconvenience of the complex numbers. Another approach, taken by A. A. Karytevn (Ref. 2), first solves the problem for a circular orbit, and then treats the low eccentricity orbit as a perturbation.

In this article, a third approach is presented, viz., the use of an iterative, or search procedure. This method is particularly useful when shadow conditions are required for each of a sequence of orbits. The fact that orbit precession and shadow rotation produce only slowly changing values of the entry and exit angles is used to advantage.

2. Shadow Geometry

Consider the geometry associated with a lunar satellite and its intersections with the moon's shadow. In Fig. 1, the x–y plane is the plane of the satellite orbit, labelled SAT, which is assumed to be an ellipse with one focus at the center of the moon, 0.

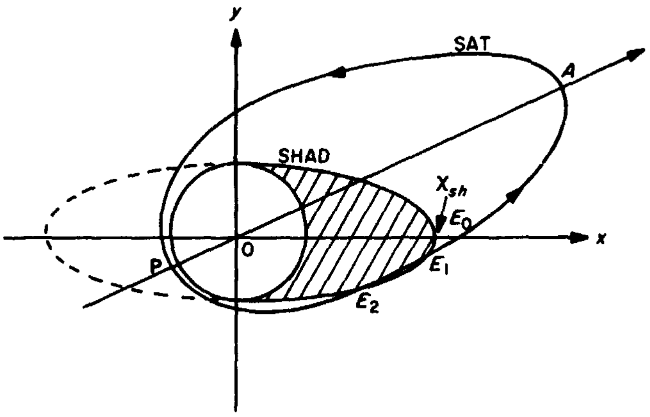


Fig. 1. Configuration in plane of satellite orbit

The orbit plane must intersect the moon's shadow (assumed to be bounded by a half-circular-cylinder emanating from the moon) in a semi-ellipse with center at 0 and major axis along x. This shadow ellipse (labelled SHAD) is also shown in Fig. 1. Only the shaded portion represents shadow. Note that point 0 is simultaneously the center of the shadow ellipse and the focus of the orbit ellipse.

The SAT may intersect SHAD at as many as four points, although only two of these, at most, can be on the shadow side. Let these be labelled E_1 and E_2, such that the satellite exits the shadow at E_1 and enters at E_2. Of course, E_1 and E_2 may either coincide (tangency of ellipses) or there may be no intersection on the shadow side (satellite always in the sun). The latter case is of no concern to the present discussion.

If we let E_0 be the point of orbit crossing the shadow side of the x-axis, there are several possibilities which may be listed as follows:

(1) E_0 is in shadow. In this case, E_1 and E_2 exist and are on opposite sides of E_0.

(2) E_0 is in sun and there are no intersection points, E_1 and E_2. In this case, the satellite is always in the sun.

(3) E_0 is in sun and there is one intersection point, $E_2 = E_1$. Here, also, the satellite is always in sun.

(4) E_0 is on the shadow–sun boundary, and hence coincides with either E_1 or E_2 or both.

(5) E_0 is in the sun and there are two intersection points, E_1 and E_2. This is the case illustrated in Fig. 1. Here, both E_1 and E_2 are on the same side of E_0.

The problem is to specify an algorithm, appropriate for computer use, to determine E_1 and E_2.

3. Derivation of Shadow Equation

R. W. Bryant (Ref. 3) introduces the shadow equation in terms of eccentric anomaly, E

$$F(E) = \mathbf{P}\cdot\mathbf{u}\,(\cos E - e) + \mathbf{Q}\cdot\mathbf{u}\,(1 - e^2)^{1/2}\sin E + \left[(1 - e\cos E)^2 - \rho^2/a^2\right]^{1/2} = 0 \qquad (1)$$

The values of E satisfying Eq. (1) correspond to positions of the satellite on the shadow–sun boundary. When $F(E) > 0$, the satellite is in sun; when $F(E) < 0$, the satellite is in shade.

As shown in Fig. 2, the derivation of the shadow equation is straightforward. In Fig. 2, coordinate axes are shown in the plane parallel to the moon–sun line. The unit vector in the direction of pericenter is **P**, **Q** is a unit vector in the direction with 90 deg true anomaly, and **R** (not shown) is an out-of-plane vector completing a right-handed system. The unit vector **u** is in the moon–sun direction, while **r** is the radius vector from the moon-center to the satellite.

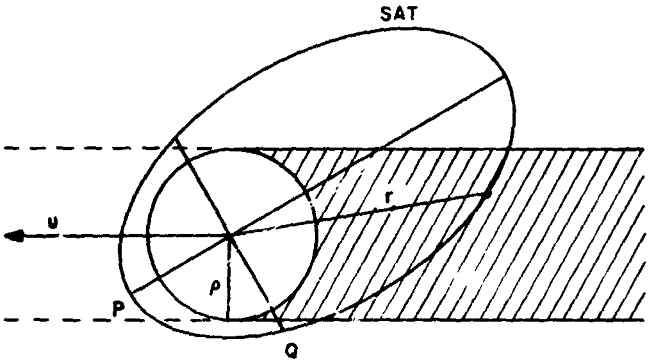


Fig. 2. Coordinate system for shadow equation

Consider an arbitrary satellite position **r**, on the shadow border. The projection of **r** in the shadow direction (along **u**) is given by $(\mathbf{u}\cdot\mathbf{r})\mathbf{u}$. On the other hand, this same quantity may be obtained geometrically as $-(r^2 - \rho^2)^{1/2}\,\mathbf{u}$ (see Fig. 2). The negative sign is required since we have specified shadow side. Hence, it follows that

$$\mathbf{u}\cdot\mathbf{r} + (r^2 - \rho^2)^{1/2} = 0 \qquad (2)$$

Then Eq. (1) follows from Eq. (2) and the standard relations for an elliptic orbit

$$\mathbf{r}/a = (\cos E - e)\,\mathbf{P} + (1 - e^2)^{1/2}\sin E\,\mathbf{Q} \qquad (3)$$

and

$$r/a = 1 - e\cos E \qquad (4)$$

The significance of $F(E)$ is easily inferred from Fig. 3, which is an edge-on view of the orbit. Consider any point $\mathcal{S}$ on the orbit in the sun, and pass a circle, center 0, through $\mathcal{S}$ intersecting the shadow at $\mathcal{S}'$. Then

$$(r^2 - \rho^2)^{1/2} = b\mathcal{S}' > -\mathbf{u}\cdot\mathbf{r} \qquad (5)$$

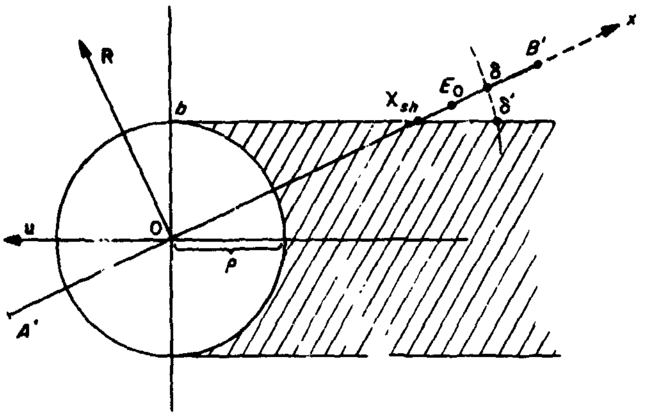


Fig. 3. Configuration in plane ›erp›endicular to satellite orbit and containing moon–sun line

Hence, $F(E) > 0$ when the satellite is in the sun [see Eq. (2)] and $F(E) < 0$ when it is in the shade.

4. Shadow Computation Algorithm

The computation starts with E_0 [the value of the eccentric anomaly on the $+x$-axis crossing (Fig. 1)] and proceeds as a search using small increments in E. The search may require many trials for the first orbit; however, for successive orbits, the number of trials is minimal since the search can start with the previous value of E_1 or E_2 instead of with E_0.

It is convenient to consider three regimes as follows:

(1) $F(E_0) < 0$.

(2) $0 \leq F(E_0) < K$.

(3) $K \leq F(E_0)$.

where K is a constant to be computed by Eq. (6).

In regime (1), the satellite is in shadow at E_0, and the search procedure is followed as given. In regime (3), the satellite is always in the sun, and no search is needed. If regime (2) occurs, the satellite may or may not be always in the sun. In either case, a search for E_1 and E_2 must be followed. However, it need only proceed in one direction from E_0 since, if they exist, both E_1 and E_2 are on the same side of E_0. To determine the direction, note whether E_0 is less than or greater than 180 deg and:

(1) If $0 < E_0 < 180$ deg, then both $0 \leq E_1 < E_0$ and $0 \leq E_2 < E_0$.

(2) If 180 deg $< E_0 <$ 360 deg, then both $E_0 < E_1 \leq$ 360 deg and $E_0 < E_2 \leq$ 360 deg.

(3) If $E_0 = 0$ deg or $E_0 = 180$ deg, then the satellite is always in sun.[1]

The search can be limited by noting (1) that it need not be pursued past pericenter and (2), since the change in true anomaly from E_0 to E_1 or E_2 can not exceed 90 deg, the search on E can be limited to a span of slightly more than 90 deg (say 100 deg) for practical purposes.

It remains only to determine a value for K.

5. Value of the Constant K

We shall show that when K is appropriately defined [see Eq. (12)], then K may be computed by the formula

$$K = \frac{\chi_{sh}}{a}\left[1 - \frac{1}{(1 - R_M^2/r_\perp^2)^{1/2}}\right] \tag{6}$$

where

$$\chi_{sh} = R_M / |\mathbf{u} \cdot \mathbf{R}|$$

$$|r_\perp| = a(1 - e\,|\sin E_0|)$$

$$R_M = \text{radius of moon}$$

This value of K corresponds to tangency of SHAD with the line connecting E_0 and the closest point of intersection of the orbit and the y-axis. In Fig. 4, the points $r_\perp$, χ_{sh}, and the distance aK between χ_{sh} and E_0, are identified. To derive Eq. (6), it is sufficient to solve

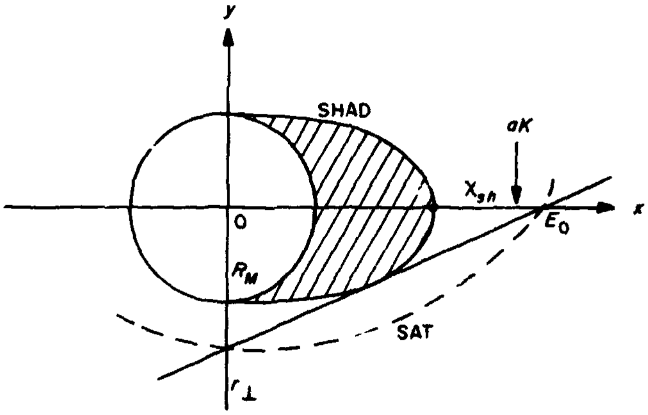


Fig. 4. Satellite orbit satisfying sufficiency criterion for satellite always in sun

[1] Implicit in our argument is the assumption that only one shadow region can occur. This is intuitively obvious, but not too easily shown mathematically.

algebraically for the intersection of the line $r_{\perp} E_0$ and SHAD, and then require tangency. Thus[2]

$$r_{\perp} E_0: \qquad \frac{\chi}{E_0} - \frac{y}{r_{\perp}} = 1 \tag{7}$$

$$\text{SHAD:} \qquad \frac{\chi^2}{\chi_{sh}^2} + \frac{y^2}{R_M^2} = 1 \tag{8}$$

Eliminate χ to obtain the quadratic in y

$$\left(\frac{E_0^2}{r_{\perp}^2} + \frac{\chi_{sh}^2}{R_M^2}\right) y^2 + 2\frac{E_0^2}{r_{\perp}} y + E_0^2 - \chi_{sh}^2 = 0 \tag{9}$$

whose discriminant is

$$E_0^2 \left(\frac{\chi_{sh}^2}{R_M^2} - \frac{\chi_{sh}^2}{r_{\perp}^2}\right) - \frac{\chi_{sh}^4}{R_M^2} \tag{10}$$

which must vanish for tangency. Solving for E_{0T} (value E_0 for tangency)

$$E_{0T} = \frac{\chi_{sh}}{(1 - R_M^2/r_{\perp}^2)^{1/2}} \tag{11}$$

Then Eq. (6) follows from Eq. (11), and the definition of K is

$$K = \frac{1}{a} (E_{0T} - \chi_{sh}) \tag{12}$$

[2] Using the symbol E_0 to represent the length of $0E_0$.

References

1. Yeremenko, R. P., "Exact Solution of the Shadow Equation," *Inst. Teor. Astron.*, Vol. 10, No. 6, pp. 446–449, 1965 (in Russian).

2. Karytov, A. A., "Determination of the Time in Which an Artificial Earth Satellite is Illuminated by the Sun," *Kosm. Issled.*, Vol. 5, No. 2, pp. 298–301, 1967.

3. Bryant, R. W., *"The Effect of Solar Radiation Pressure,"* NASA TN D-1063. National Aeronautics and Space Administration, Washington, Sept. 1961.

B. A Consistent Ephemeris of the Major Planets in the Solar System,

W. G. Melbourne and D. A. O'Handley

1. Introduction

The system of computer programs known as the solar system data-processing system (SSDPS) has been used to compute a consistent ephemeris of the major planets that has been fit in a weighted least-squares sense to both optical and radar-time-delay observations of the planets. The SSDPS has been described fully in SPS 37-49, Vol. III, pp. 1–14. This ephemeris has been adopted for the planetary ephemerides contained in developmental ephemeris (DE) 40. Although the developmental ephemerides are continually being updated by the processing of new or refined data, or by the improvement of the mathematical model used in the data processing, DE 40, nevertheless, represents something of a milestone in the ephemeris development activity. For this reason, a brief summary of the data processing, and the resulting ephemeris, is presented here.

Until 1967, the planetary ephemeris tape system at JPL was obtained from least-squares fits to source ephemerides based on planetary theories fit to meridian circle observations of the sun and the planets (Refs. 1 and 2). In early 1967, ephemerides of Venus and the earth–moon barycenter were produced that had been fit to both 1950–1966 U.S. Naval Observatory meridian-circle observations and planetary radar range and doppler observations of Venus taken over the period 1961 to 1966. The best example of this series is DE 24[3] which was used in the *Mariner* V operations. These ephemerides were obtained with the "phase I" system of programs. These included an orbit determination system used in early work on the determination of the astronomical unit (AU) and the radius of Venus (Ref. 3), but modified to include optical data. The path generation for the phase I system was the PLOD II system (Ref. 4). Although intended to be valid only over a relatively short arc, DE 24, nevertheless, represented an improvement of between one and two orders of magnitude in accuracy over previous ephemerides. The phase II program development activity, begun in late 1966, has led to the current version of the SSDPS.

2. Data Set

The optical data set used in DE 40 is presented in Table 1. These are all the meridian observations from the 6-in. transit circle of the U.S. Naval Observatory over the interval 1949–1967. This set of observations differs from those reported in SPS 37-48, Vol. III, pp. 8–9 primarily by the data taken between 1966–1967.

The planetary radar data have been taken since 1961. Initially, the data type was doppler, and, beginning in

[3] Lawson, C. L., *Announcement of JPL Developmental Ephemeris No. 24*, Apr. 1967 (JPL internal document).

Table 1. DE 40 optical data set observations

Planet	No. of observations
Sun	2136
Mercury	553
Venus	1165
Mars	243[a]
Jupiter	348
Saturn	338
Uranus	330
Neptune	325

[a] This number does not include the set of observations compiled by Clemence for his theory of Mars.

1964, both doppler and range were obtained. The four sources of this planetary range data have been Arecibo Ionospheric Observatory in Puerto Rico, Haystack and Millstone Hill sites, and the Venus DSS (SPS 37-48, Vol. III, pp. 8–9). The usable data (in the sense of accuracy), cover the period from 1964 onwards. This set of planetary range data is given in Table 2.

Table 2. Planetary range data

Planet	No. of observations	Source	Period
Mercury	151	Arecibo	Apr. 1964–Aug. 1967
Venus	181	Arecibo	Mar. 1964–Oct. 1967
	35	Haystack	July 1967–Sept. 1967
	99	Millstone	Aug. 1967–Oct. 1967
	281	Venus DSS	May 1964–Oct. 1967
Mars	39	Arecibo	Nov. 1964–June 1965
	10	Haystack	Apr. 1967–June 1967

An additional discussion of these data appears in SPS 37-48, Vol. III, pp. 8–9. The total of the Venus DSS set is radically changed from that of Table 1. The values given in Table 1 referred to the uncompressed data. The full discussion of these data appears in Ref. 5. In addition, the total includes 15 time-delay measurements of Venus obtained by D. A. O'Handley[4] at the Venus DSS during July–October 1967 inferior conjunction.

Because of advances in radar technology involving larger antennas, increased transmitter power levels, and improved data reduction techniques, the precision of the time-delay measurements has improved by an order of magnitude over the 1964–1967 period, i.e., a typical standard deviation of a 1964 Venus time-delay measurement is in the 20–50 μs range, while a 1967 inferior conjunction measurement lies in the 3–5 μs range. Current Mercury observations are precise to about 10 μs and the 10 normal points for Mars, based on the 1967 Haystack observations, are of similar quality. On the other hand, the precision of a radar doppler measurement is about 1 Hz. A simple calculation will show that for a typical orbital parameter, a precision in doppler of 1 Hz is equivalent to a precision of about 10^3 μs in a time-delay measurement. Further, doppler does not provide information about planetary radii. For these reasons, doppler information, although extremely valuable in the radar data reduction process and in the study of planetary topography and surface characteristics, is not presently used in ephemeris development.

Special mention should be made of the 10 high-precision, time-delay normal points of Mars taken during the April–June, 1967 period at the Haystack facility. Each point corresponds to the observations taken in one night. The 10 observation nights are spread over the 2-month period at weekly intervals. During an observation session, the planet rotates under the radar beam, and the half-power width of the return beam covers about 200 km on the Martian surface. Consequently, topographic features on Mars are observed to move through the return radar beam giving variations in time delay with a magnitude of up to 100 μs. The regions on Mars observed on successive nights partially overlap, and, during the 2-month period, a strip covering the entire 360 deg of longitude was observed. Because of this overlap, it is possible to determine the relative altitude, on every observation night, of any point on this strip. A reference point was chosen that was close to representing a mean altitude with respect to the topographic variations; it is the range to this reference point that is given in the data set for the 10 observing sessions.[5]

3. Parameter Set

The conditional equations were formed from the residuals constructed from the observations and the predicted observations (observed minus computed) based on DE 35. The DE 35 was generated from the N-body integrator in SSDPS using an up-to-date set of planetary masses (SPS 37-45, Vol. IV, p. 17) that incorporates the mass determinations by radio tracking data from spacecraft. The initial conditions of DE 35 were based on a least-squares fit to an earlier JPL ephemeris (DE 26) in order to minimize the secular effects resulting from adopting a new set of planetary masses significantly different, in some cases, from the IAU set used previously. The planetary masses in DE 40 are the same as in DE 35.

[4] O'Handley, D. A., *Reconstruction of JPL Radar-Range of Venus—29 July, 1967 to 27 October, 1967*, (JPL internal document).

[5] Private communication from G. H. Pettengill (Apr. 2, 1968).

The orbital coefficients of the conditional equations constructed from DE 35 are basically the osculating Set III elements of D. Brouwer and G. M. Clemence (Ref. 6) at the epoch JD 2440800.5.

The simultaneous incorporation of optical and range observations in a single solution for all the planets, with the exception of Pluto, has not been accomplished previously. It therefore became necessary to examine the parameters that could be solved for in light of the limited data set currently available.

With range data alone, a 21-parameter solution for Mercury, Venus, and Mars gave a solution in which the parameters were reasonably determined (see Table 3). The first 6 rows of Table 3 correspond to the Set III orbital parameters in Ref. 6.

Table 3. Parameter determination using range data

Mercury	Venus	Earth–Moon	Mars
ΔL	ΔL	—	ΔL
Δp	Δp	—	—
Δq	Δq	—	—
$e\Delta r$	$e\Delta r$	$e\Delta r$	$e\Delta r$
Δe	Δe	Δe	Δe
$\Delta a/a$	—	$\Delta a/a$	—
Radius	Radius	—	Radius
AU	—	—	—

Several comments should be made with regard to this set of parameters. With radar data only, it is necessary to limit the parameter set to those parameters that are sensitive to time-delay measurements. Consequently, the parameters defining the orientation of the orbit of the earth relative to the astronomical right ascension and declination coordinate system were not adjusted. Even with a well-distributed data set, solving for the semimajor axis of the orbits of these planets simultaneously leads, in the pure radar solution, to a near singular normal matrix. The dominant signature in the time-delay observable resulting from adjusting the semimajor axis is due to the change in the mean motion of the planet rather than the direct effect of the change in the semimajor axis itself. The orbits of Venus and the earth are nearly coplanar and circular; therefore a change in the mean motion of Venus is almost indistinguishable from a corresponding negative change in the mean motion of the earth. In the radar-only solution, the semimajor axis of the earth–moon barycenter is used because it gives a slightly better fit in the least-squares sense, and because it has the weight of all the range observations.

The radar data for Mars are too scant and not well enough distributed to give good determination of the quantities Δp, Δq, and a/a. The quantities Δp and Δq are rotations of the orbit plane about orthogonal axes embedded in the orbit plane, and cause displacements of the planet perpendicular to its orbit plane. Since the inclination of the orbit plane of Mars to the ecliptic is only 1.9 deg, these two out-of-plane quantities are in excess of an order of magnitude more difficult to determine than the in-plane quantities, even with an optimally distributed data set. The high-precision Haystack points, coupled with the relatively low-precision 1964 Arecibo data, are not sufficient to obtain a definitive value for the mean motion quantity $\Delta a/a$.

With the inclusion of optical data for all the major planets, with the exception of Pluto, an expanded parameter set is used. This set consists of 56 unknowns as follows:

(1) Six elements of 7 planets.

(2) Six elements of the earth–moon barycenter.

(3) Four limb corrections, right ascension, and declination of Mercury and Venus.

(4) Three radii (Mercury, Venus, and Mars).

(5) One AU.

The 18-yr span of this data does not permit a definitive set of corrections for the outer planets. Including the $\Delta a/a$ parameters of the outer planets is somewhat ambitious for this data set; however, this parameter set gave corrections for each planet that diminished or removed the secular trends in the residuals published by the U.S. Naval Observatory from transit circle observations.

4. Solutions

Two solutions were made in order to arrive at the current ephemeris. Initially, a solution that utilized both the optical and radar-range conditional equations was made. The motivation here was to allow the optical data set to determine those quantities that are sensitive only to the optical data, but simultaneously using the range data to anchor the range sensitive parameters. The rank 52 solution from an Eigenvalue–Eigenvector analysis[6] of the 56-parameter set was chosen because the correction to Δp and Δq of the earth stabilized at a value which is in agreement with the known error in these quantities. For solutions of rank greater than 52, the normal matrix

[6] Lawson, C. L., *Eigenvalue–Eigenvector Analysis for SSDPS*, Jan. 17, 1968 (JPL internal document).

is too near singular and causes significant instability in these and other parameters. The resulting ephemeris is DE 39.

At this point, the optical data have provided the reference frame to which the relative measurements of range can be evaluated. The radar data was felt to be a much more accurate source of information for those parameters best solved for by this type of data. It was suspected that this data type would be degraded when used simultaneously with optical data. For this reason, an iteration was made on this solution using the range observations alone. The range data were compared against DE 39 and corrections to this ephemeris were calculated based upon the 21-parameter radar set described in Subsection 3. None of the 21 corrections obtained was statistically significant when compared to its formal standard deviation; nevertheless, they were applied for reasons of consistency. The ephemeris generated by applying these corrections is called DE 40.

The values of the constants to be used with DE 40 are as follows[7]:

(1) AU = 149,597,895.8.

(2) Radius of Mercury = 2437.3.

(3) Radius of Venus = 6055.8.

(4) Radius of Mars = 3375.3.

The values of the AU and the radii of Mercury and Venus given here are essentially in agreement with those found by the MIT group (Ref. 7). The value of the radius of Mars, however, is weakly determined (see *Subsection 5*) because of the poor distribution of radar points; the best value available at this time is the *Mariner IV* occultation experiment value of 3393 ± 4 (SPS 37-43, Vol. IV, p. 7).

5. Standard Deviations

The subject of the relative sigmas of each data type present in the solution was considered. The optical data were given the following sigmas:

(1) Right ascension = $1''.0/\cos\delta$.

(2) Declination = $1''.0$.

The range data were given the standard deviation assigned by the respective observers.

The covariance matrix resulting from the optical and range data may be used to obtain both formal standard deviations of the estimated parameters and the correlations among them. Table 4 gives the formal standard deviations of the 24 orbital parameters of the inner planets, the three planetary radii, and the astronomical unit. The units of the standard deviations are arc seconds except for the radii and the AU which are in kilometers.

Table 4. Standard deviations of orbital parameters, planetary radii, and AU

Data type	Mercury	Venus	Earth–Moon	Mars
ΔL	0.031	0.031	0.031	0.032
Δp	0.023	0.019	0.018	0.030
Δq	0.022	0.019	0.019	0.031
$e\Delta r$	0.005	0.0006	0.0007	0.004
Δe	0.002	0.0005	0.0004	0.007
$\Delta a/a$	0.00007	0.0001	0.0002	0.0006
Radius	1.0	0.2	—	7.0
AU	0.27	—	—	—

The formal standard deviations exhibit their usual degree of optimism. The reader, therefore, should be aware that they do not account for either possible systematic error factors in the data or unmodelled parameters in the mathematical model. The correlation matrix, although not shown here, verifies that high correlations exist among the mean longitude parameters (ΔL), and the mean motion parameters ($\Delta a/a$). With this exception, the problem is well-conditioned.

In spite of known biases in the optical data related to limb corrections, there is some encouraging evidence of consistency between the optical and radar data. For example, the corrections from an optical solution alone to the orientation of the orbit plane of Venus relative to the ecliptic are found to agree with the values obtained in a pure radar solution. The radar data also exhibit a degree of internal consistency. For example, the corrections to $e\Delta r$ and Δe of the earth from processing Mercury range data alone are the same as those obtained when only Venus ranging data are processed.

The standard deviations for ΔL, Δp, Δq, and $\Delta a/a$ are an order of magnitude smaller in the 21-parameter pure radar solution. This is due to the precision of the radar measurements and the fact that these parameters become *relative* quantities for which radar obtains extremely powerful solutions. The reader can easily verify with a simple model consisting of circular coplanar orbits, that a set of one-hundred 10-μs quality range points, well distributed, enables one to determine the longitude of

[7] In converting from "light-seconds" to kilometers, the velocity of light is taken to be exactly the IAU value of 299,792.5 km/s.

Ve· · s relative to the longitude of the earth to about 0.″002. Furthermore, additional error analyses show that with only 3 years of ranging to Venus, the mean motion of Venus relative to the mean motion of the earth is determined with a precision (formal) of 0.″1/100 yr. It has been known for several years that the relative longitude of Venus required a correction ranging between +0.″5 and +1.″0. This is, most likely, an accumulated effect due to an error in relative mean motion; the current analysis gives a correction to the relative mean motion of Venus of +1.″2/100 yr.

6. Residuals

The range residuals for Mercury, Venus, and Mars are shown in Figs 5–9. The residuals of Mercury are shown in Fig. 5 based on an ephemeris (contained in DE 35) which closely matches the Newcomb ephemeris (Ref. 2).

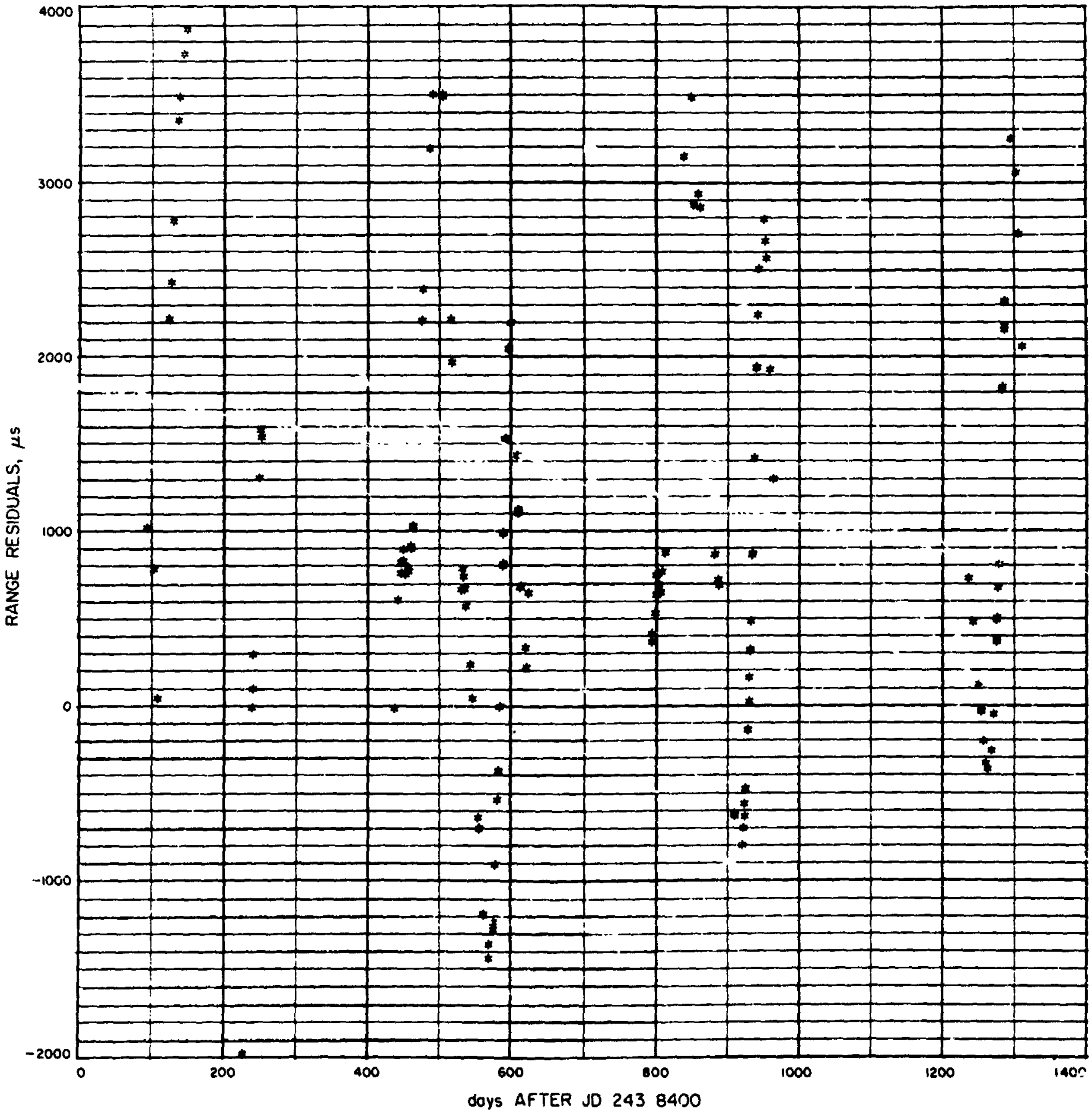


Fig. 5. Mercury residuals based on DE 35 ephemeris and DE 40 AU and radius

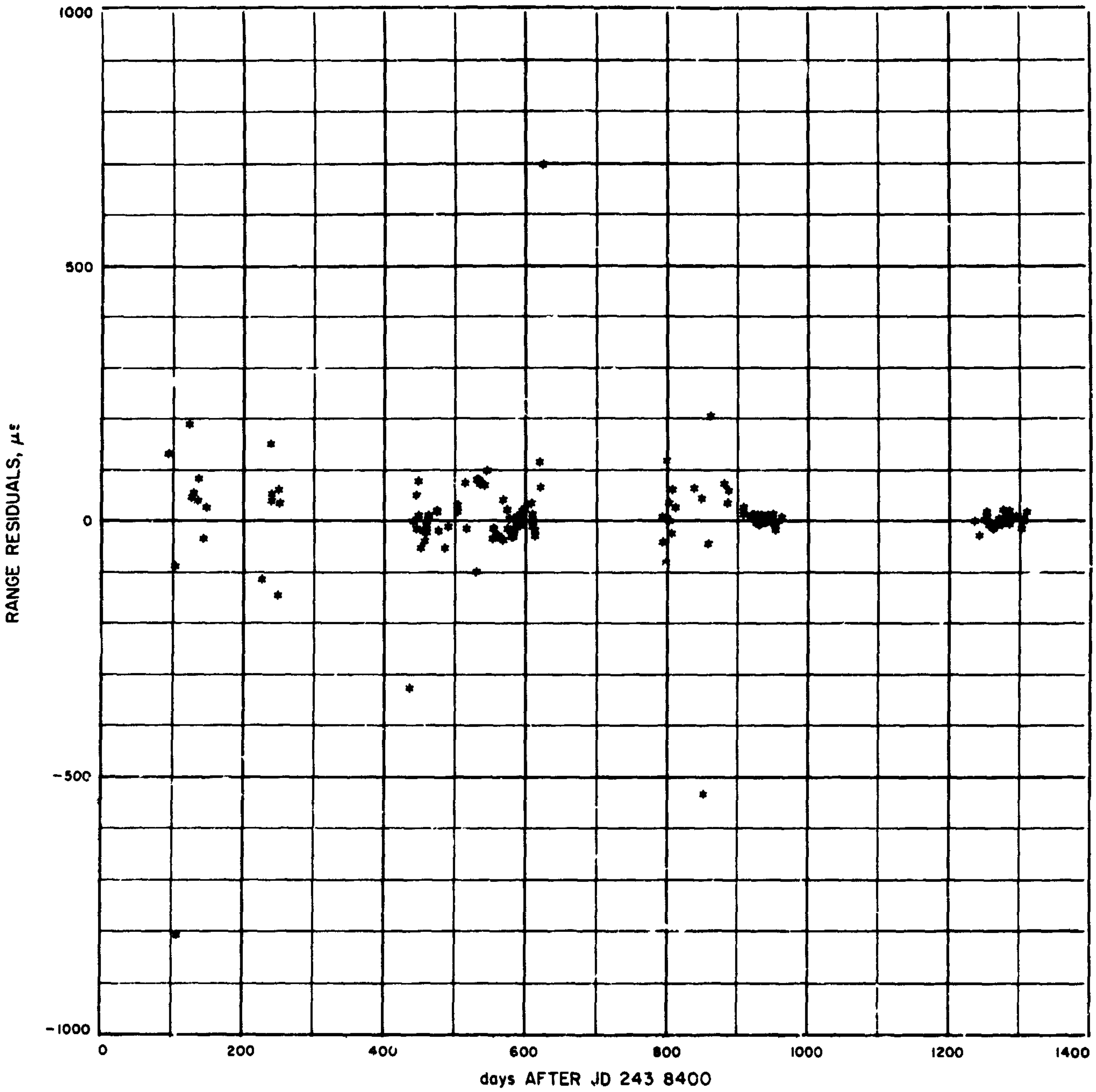


Fig. 6. Improvement in Mercury residuals resulting from DE 40

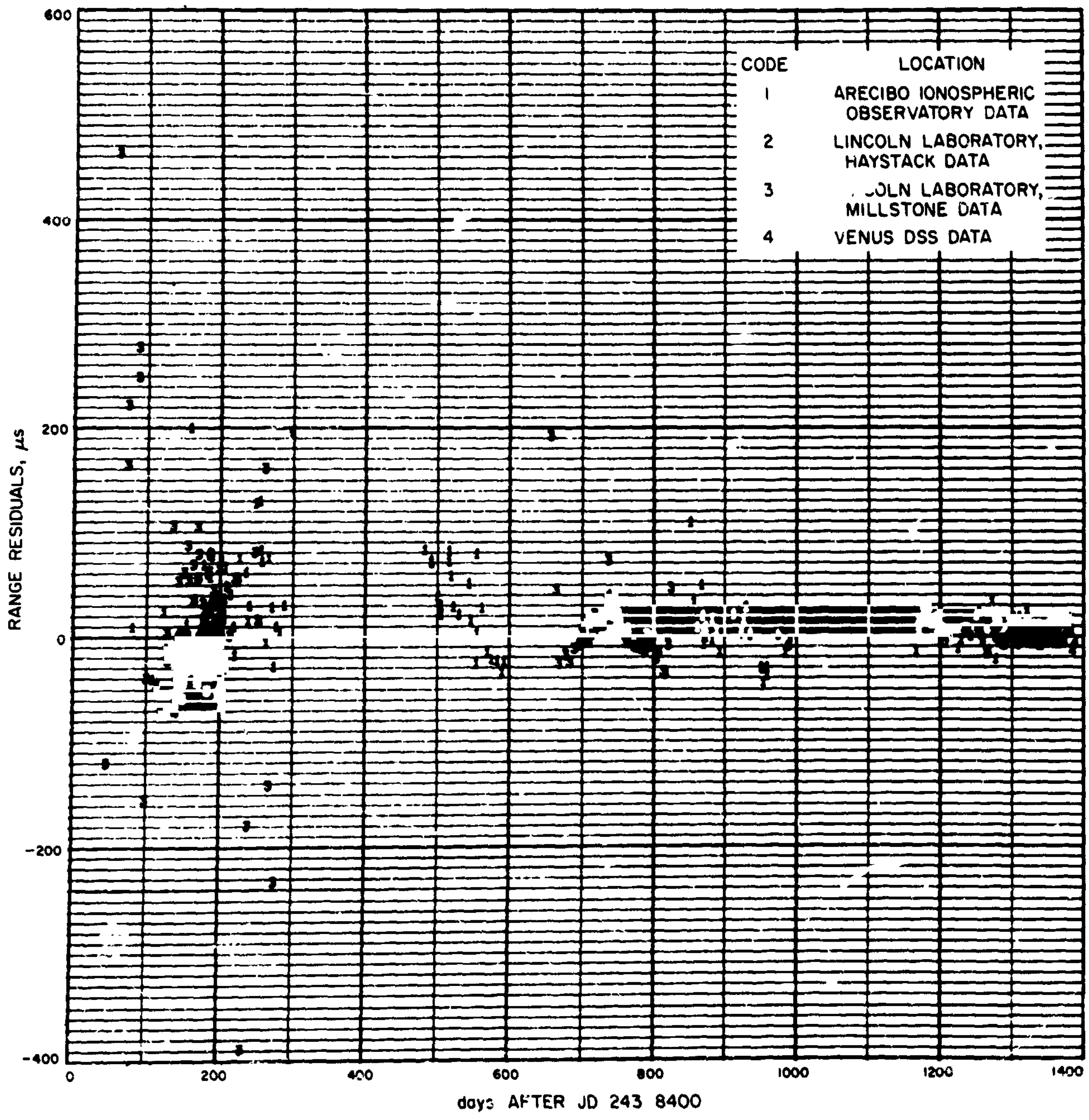


Fig. 7. Residuals of all available Venus ranging data obtained from DE 40

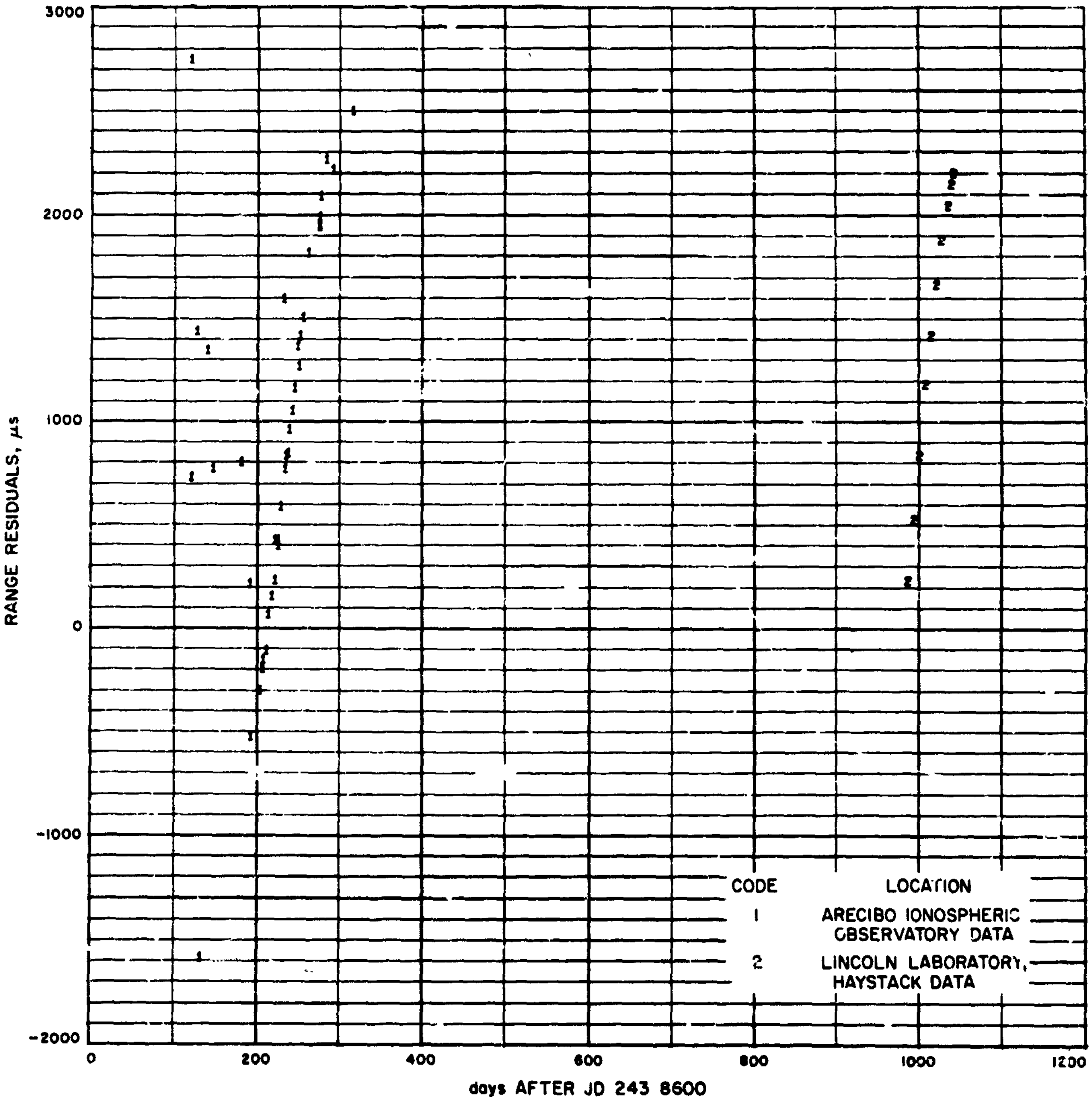


Fig. 8. Mars residuals compared to DE 35

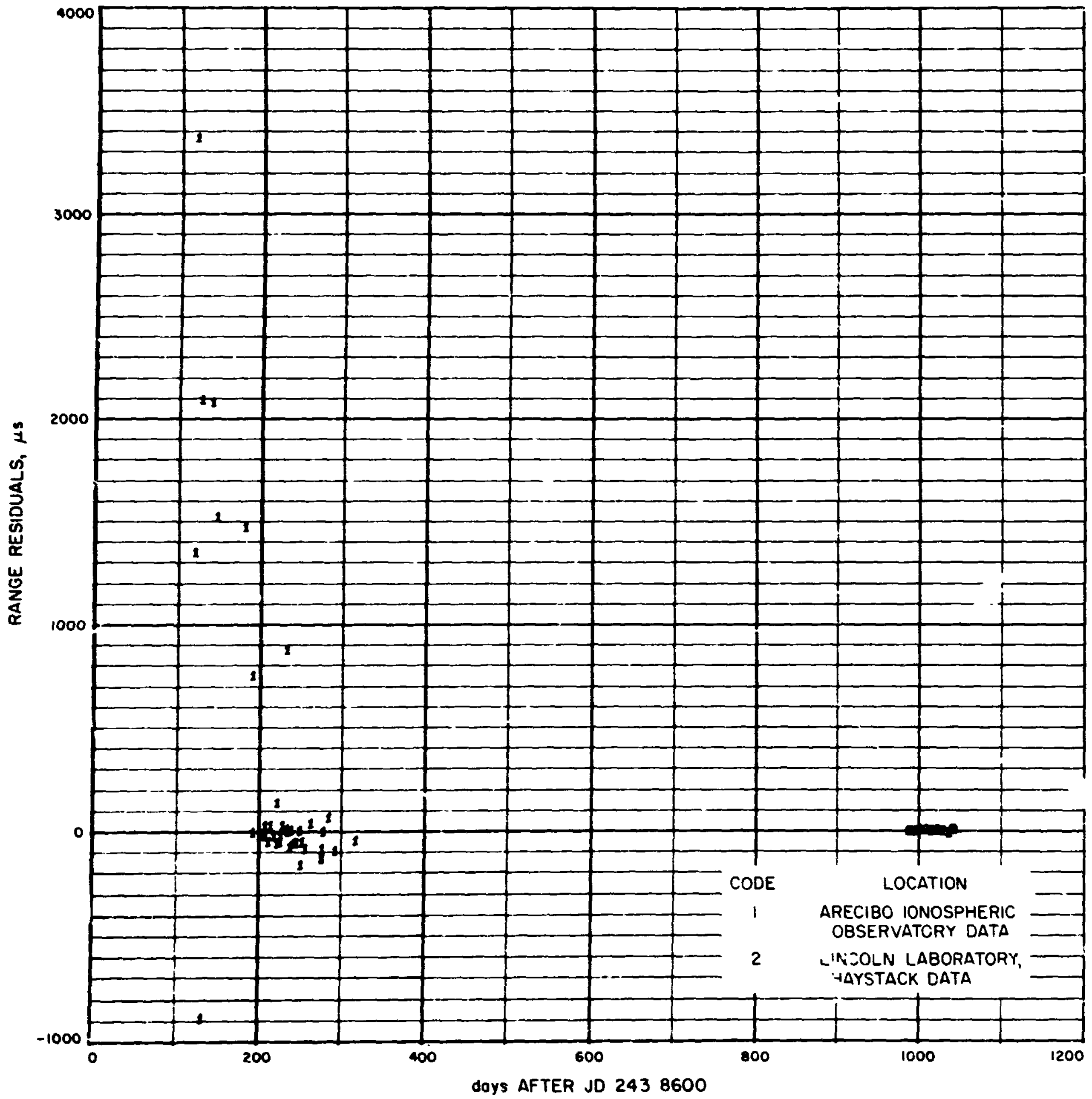


Fig. 9. DE 40 residuals for Mars

In this figure, the AU and radius from DE 40 were used. The improvement in residuals resulting from DE 40 is shown in Fig. 6. All of the range data shown here were taken at the Arecibo Ionospheric Observatory. The residuals of all available Venus ranging data obtained from DE 40 are shown in Fig. 7.

The tremendous improvement of radar techniques over the period 1964–1967 is shown in all of the figures after solution. An as yet unexplained anomaly in the residuals of the 1965–1966 ranging period is shown in Fig. 7. The fact that the radar-range residuals from both JPL and Millstone show this anomaly independently establishes that it is not due to an instrumentation effect. Current conjecture is that it is due to second-order effects of fixed parameters.

The residuals from ranging Mars, when compared to DE 35, are shown (Fig. 8) to have very large trends. The Mars ephemeris in DE 35 closely fits Clemence's second-order theory of Mars used as a source ephemeris for DE 19 (Ref. 2). The DE 40 residuals for Mars are shown in Fig. 9.

DE 40 should not be considered the final "best" ephemeris. The lunar ephemeris incorporated into this ephemeris is LE 4. There is a new version DE 43 which has LE 6 on it. Certain problems with the 1967 Venus radar-range data, from the Arecibo Ionospheric Observatory, lead to the conclusion that another solution should be made. There is, at present, new data on Mercury and Venus, and some revised data over other periods to be added. A few data points should be edited. Finally, the SSDPS is a rather complex and evolving system containing over 150 subroutines and about 200,000 words of machine-level instructions. The possibility of subtle errors in this system is not unlikely, and efforts are continuing to validate the current working version.

References

1. Peabody, P. R., Scott, J. F., and Orozco, E. G., *Users' Description of JPL Ephemeris Tapes,* Technical Report 32-580. Jet Propulsion Laboratory, Pasadena, Calif., Mar. 2, 1964.
2. Devine, C. J., *JPL Development Ephemeris Number 19,* Technical Report 32-1181. Jet Propulsion Laboratory, Pasadena, Calif., Nov. 15, 1967.
3. Muhleman, D. O., and Holdridge, D. A., and Block, N., "The Astronomical Unit Determined by Radar Reflections from Venus," *Astron. J.,* Vol. 67, p. 191, 1962.
4. Devine, C. J., *PLOD II: Planetary Orbit Determination Program for the IBM 7094 Computer,* Technical Memorandum 33-188. Jet Propulsion Laboratory, Pasadena, Calif., Apr. 15, 1965.
5. Muhleman, D. O., O'Handley, D. A., Lawson, C. L., and Holdridge, D. B., *JPL Radar Range and Doppler Observations of Venus 1961–1966,* Technical Report 32-1123. Jet Propulsion Laboratory, Pasadena, Calif., 1967.
6. Brouwer, D., and Clemence, G. M., *Methods of Celestial Mechanics,* p. 241, Academic Press, New York, 1961.
7. Ash, M. E., Shapiro, I. I., and Smith, W. B., "Astronomical Constants and Planetary Ephemerides Deduced from Radar and Optical Observations," *Astron. J.,* Vol. 72, p. 338, 1967.

C. Correction of the Lunar Orbit Using Analytic Partial Derivatives, *J. D. Mulholland*

As reported in an earlier article (SPS 37-49, Vol. III, pp. 21–23), work is underway on the numerical integration of the lunar ephemeris. The primary difficulty in such an undertaking lies in the formulation of the differential correction process—not a trivial process for such a highly perturbed object.

In order for a differential correction process to work, it is necessary that the vector $\rho\ (\eta_i)$, whose first variation is represented by the left-hand side of the conditional equation (SPS 37-50, Vol. III, pp. 50–53), be a reasonably close approximation to the real motion over the correction arc.[8] As a result of recent efforts, it is now known that Keplerian or Hansen-type approximations are not adequate for the correction of the lunar orbit for arcs of 5 years. What is required is some formulation of the conditional equations that conforms rather closely with the motion that is being used as the "observations," in this case lunar ephemeris (LE) 6. Three ways in which this might be accomplished are as follows:

(1) Integration of the variational equations.

(2) Construction of finite difference quotients.

(3) Derivation of high-accuracy analytic partial derivatives.

All three means are being investigated and compared.

Integration of the variational equations represents the most accurate and the most rigorously correct of the possible approaches. If done properly, the conditional equations would represent the correct first variation of the computed state vector. This process, however, requires large amounts of computer time and, for this reason, does not seem promising.

Finite difference quotients are approximations of the integrals of the variational equations. They are formed

[8] This is an intentionally ambiguous statement, because this qualitatively true statement can only be given a quantitative meaning in terms of the specific problem of interest.

by making a series of computations, varying one element at a time, and computing the differential effects $\Delta\rho/\Delta\eta_i$. Thus, 13 orbit integrations are required, rather than one. Again, this is an expensive process.

The use of analytic partials would appear to be very desirable if they can be made to provide an adequate representation of the perturbed motion. This will be assured if they are derived directly from the Lunar Theory; they will then represent the correct first variation of the observed motion—the Lunar Theory itself. Unfortunately, there is no simple correspondence between the parameters of the Lunar Theory and the set of elements to be corrected, the Brouwer and Clemence Set III parameters (Ref. 1). Define the following sets of parameters:

$$\delta\kappa\text{:}\ \{\Delta\epsilon, \Delta i, \Delta\Omega, \Delta\omega, \Delta a/a, \Delta e\}$$

$$\delta\text{III:}\ \{\Delta g_0 + \Delta r_x, \Delta p, \Delta q, e\Delta r_x, \Delta a/a, \Delta e\}$$

$$\hat{s}\text{:}\ \{\lambda, \beta, r, \dot{\lambda}, \dot{\beta}, \dot{r}\}$$

$$\hat{r}\text{:}\ \{x, y, z, \dot{x}, \dot{y}, \dot{z}\}\quad \text{(ecliptic)}$$

where $\epsilon = \Omega + \omega + g_0$, and all other symbols have their usual meanings.

The difficulty lies in the circumstance that the PLOD II differential correction treats the orbit elements κ_1 osculating at the epoch, while the Lunar Theory is developed in terms of the mean elements κ_0. Thus, it is necessary to form the conditional equations according to the matrix relation

$$\left[\frac{\partial\hat{s}}{\partial\text{III}}\right] = \left[\frac{\partial\hat{s}}{\partial\kappa_0}\right]\left[\frac{\partial\kappa_0}{\partial\kappa_1}\right]\left[\frac{\partial\kappa_1}{\partial\text{III}}\right]$$

The factor $[\partial\hat{s}/\partial\kappa_0]$ is obtainable directly from the theory, while the factor $[\partial\kappa_1/\partial\text{III}]$ is strictly geometric and is readily shown to be the matrix

$$\begin{bmatrix} 1 & \dfrac{\sin\omega}{\sin i}(1-\cos i) & \dfrac{\cos\omega}{\sin i}(1-\cos i) & 0 & 0 & 0 \\ 0 & \cos\omega & -\sin\omega & 0 & 0 & 0 \\ 0 & \dfrac{\sin\omega}{\sin i} & \dfrac{\cos\omega}{\sin i} & 0 & 0 & 0 \\ -1 & -\dfrac{\sin\omega}{\tan i} & -\dfrac{\cos\omega}{\tan i} & \dfrac{2}{e} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

What, however, is the relationship between the mean elements and the elements osculating at epoch? Recalling that the $[\partial\hat{s}/\partial\kappa_0]$ are available, one may write

$$\left[\frac{\partial\kappa_1}{\partial\kappa_0}\right]=\left[\frac{\partial\kappa_1}{\partial\hat{s}_1}\right]\left[\frac{\partial\hat{s}_1}{\partial\kappa_0}\right]$$

where $\hat{s}$ evaluated at the epoch is denoted $\hat{s}_1$. Since κ_0 and κ_1 are each sets of 6 linearly independent parameters, then the inverse exists and

$$\left[\frac{\partial\kappa_0}{\partial\kappa_1}\right]=\left[\frac{\partial\kappa_1}{\partial\kappa_0}\right]^{-1}$$

To find the formulation of $[\partial\kappa_1/\partial\hat{s}_1]$, one may write

$$\left[\frac{\partial\kappa_1}{\partial\hat{s}_1}\right]=\left[\frac{\partial\kappa_1}{\partial\hat{r}_1}\right]\left[\frac{\partial\hat{r}_1}{\partial\hat{s}_1}\right]$$

The matrix $[\partial\hat{r}_1/\partial\hat{s}_1]$ is readily found from geometric relations and will not be given here. The problem finally comes down to the computation of $[\partial\kappa_1/\partial\hat{r}_1] = [\partial\hat{r}_1/\partial\kappa_1]^{-1}$. A relatively simple approach to this is to define the rectangular state vector $\hat{\rho}$ in orbit-fixed coordinates

$$\{\hat{\rho}\} = \left\{\begin{array}{l} a(\cos E - e) \\ a(1-e^2)^{1/2}\sin E \\ \quad 0 \\ -\,na\sin E/(1-e\cos E) \\ +\,na(1-e^2)^{1/2}\cos E/(1-e\cos E) \\ \quad 0 \end{array}\right\}$$

If the matrix $A_k(\alpha)$ is used to effect a rotation about the k-axis through the angle α, then

$$\{\hat{r}\} = A_3(\Omega)\,A_1(i)\,A_3(\omega)\,\{\hat{\rho}\}$$

Define the matrix

$$B_k(\alpha) = d[A_k(\alpha)]/d\alpha$$

Then the columns of $[\partial\hat{r}/\partial\kappa_1]$ are given by

$$\left\{\frac{\partial\hat{r}}{\partial\epsilon}\right\}=\frac{1}{n}\left\{\frac{d\hat{r}}{dt}\right\}$$

$$\left\{\frac{\partial\hat{r}}{\partial i}\right\}=A_3(\Omega)\,B_1(i)\,A_3(\omega)\,\{\hat{\rho}\}$$

$$\left\{\frac{\partial\hat{r}}{\partial\Omega}\right\}=B_3(\Omega)\,A_1(i)\,A_3(\omega)\,\{\hat{\rho}\}$$

$$\left\{\frac{\partial\hat{r}}{\partial\omega}\right\}=A_3(\Omega)\,A_1(i)\,B_3(\omega)\,\{\hat{\rho}\}$$

$$a\left\{\frac{\partial\hat{r}}{\partial a}\right\}=\{\hat{r}\}-\frac{3}{2}\left\{\frac{d\hat{r}}{dt}\right\}\Delta t$$

$$\left\{\frac{\partial\hat{r}}{\partial e}\right\}=H\,\{\hat{r}\}+K\left\{\frac{d\hat{r}}{dt}\right\}$$

where H and K are as defined previously (SPS 37-50, Vol. III, pp. 50–53).

The application of these relations would be as follows: At the beginning of the differential correction process, it is necessary to form the matrix

$$[C]=\left\{\left[\frac{\partial\hat{r}_1}{\partial\kappa_1}\right]^{-1}\left[\frac{\partial\hat{r}_1}{\partial\hat{s}_1}\right]\left[\frac{\partial\hat{s}_1}{\partial\kappa_0}\right]\right\}^{-1}\left[\frac{\partial\kappa_1}{\partial\mathrm{III}}\right]$$

At every subsequent time point at which conditional equations are required, one need only form the matrix product

$$\Delta\hat{s}=\left[\frac{\partial\hat{s}}{\partial\kappa_0}\right][C]\,\{\delta\mathrm{III}\}$$

where $\{\delta\mathrm{III}\}$ is the vector of unknown increments that the solution is expected to determine.

It is expected that this approach will be more economical of computer time by a factor of 3 to 6 over the other methods of computing accurate partial derivatives.

Reference

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

D. Bayesian Estimation Based on the Gram–Charlier Expansion, *W. Kizner*

1. Introduction

In previous articles (SPS 37-49, Vol. III, pp. 23–31, and SPS 37-50, Vol. III, pp. 20–22), the author discussed a method that uses a numerical approximation to find the coefficients of a Hermite series expansion (used in nonlinear estimation). This article shows that an approximation based on the Gram–Charlier expansion may be optimal in cases where all that is desired is the conditional mean of the distribution, and the distribution is, approximately, gaussian. As is well known, this is desirable with quadratic loss functions.

2. A Numerical Approximation of the Gram–Charlier Expansion

Let x be a scalar and assume that all the moments of the probability density function $p(x)$ exist. Then $p(x)$ may be represented by the Gram–Charlier expansion

$$p(x) \approx \frac{1}{(2\pi)^{1/2}} \exp\left(\frac{-x^2}{2}\right)\left[\sum_{n=0}^{\infty} a_n He_n(x)\right] \qquad (1a)$$

$$a_n = \frac{1}{n!}\int_{-\infty}^{\infty} p(x)\, He_n(x)\, dx \qquad (1b)$$

Here $He_n(x)$ is the Hermite polynomial of nth degree and can be defined by

$$He_0(x) \equiv 1, He_1(x) \equiv x$$
$$He_{n+1}(x) \equiv x\, He_n(x) - nHe_{n-1}(x) \qquad (2)$$

It is known that these polynomials are mutually orthogonal with respect to the weight function $\exp(-x^2/2)$. The fact that the area of $p(x)$ is one implies that a_0 is one.

To find a numerical approximation for a_n without having to evaluate the integral in Eq. (1b) analytically, one proceeds as follows:

Let $\pi_n(\xi)$ be a polynomial of degree n or less. Applying the theorem of Gauss and Jacobi

$$\int_{-\infty}^{\infty} \pi_n(\xi) \exp(-\xi^2)\, d\xi = \sum_{i=1}^{m} W_i^m \pi_n(\xi_i^m) \qquad (3)$$

where $2m - 1 \leq n$, and the W_i^m and ξ_i^m are weights and nodes for gaussian quadrature. They are tabulated for the weight function $\exp(-\xi^2)$; the nodes correspond to the zeroes of $H_m(x)$. Let $\xi = x/2^{1/2}$. Then Eq. (3) becomes equal to

$$\frac{1}{2^{1/2}}\int_{-\infty}^{\infty} \pi_n\left(\frac{x}{2^{1/2}}\right)\exp\left(\frac{-x^2}{2}\right)dx$$

Define a new nth degree polynomial by

$$\phi_n(x) = \pi_n \frac{x}{2^{1/2}}$$

Then

$$\int_{-\infty}^{\infty} \phi_n(x)\exp\left(\frac{-x^2}{2}\right)dx = 2^{1/2}\sum_{i=1}^{m} W_i^m \phi(2^{1/2}\xi_i^m) \qquad (4)$$

If we approximate $p(x)$ by

$$p(x) \cong \exp\left(\frac{-x^2}{2}\right)\pi_k(x) \qquad (5)$$

Then substituting Eq. (5) into Eq. (1b), and using Eq. (4), we arrive at

$$a_n \cong a^m = \frac{2^{1/2}}{n!}\sum_{i=1}^{m} W_i^m p(2^{1/2}\xi_i^m)\exp[(\xi_i^m)^2]\, He_m(2^{1/2}\xi_i^m) \qquad (6)$$

Let

$$p^m(x) = \frac{1}{(2\pi)^{1/2}}\exp\left(\frac{-x^2}{2}\right)\left[\sum_{n=0}^{m-1} a_n^m He_n(x)\right] \qquad (7)$$

Then it can be shown (as previously) that $p^m(x)$ coincides with $p(x)$ at the m points $2^{1/2}\xi_i^m$, $i = 1,2, \cdots, m$ which are the zeroes of $He_m(x)$. Also as before, we are led to believe that whenever Eq. (1b) exists as a Riemann integral

$$\lim_{m\to\infty} a_n^m = a_n \qquad (8)$$

For a k dimensional distribution, the procedure is similar to the case for the Hermite functions.

3. The Convergence of the Gram–Charlier Expansion

The reason for employing this expansion is as follows:

Theorem. Assume that $p(x)$ is given exactly as a combination

$$\exp\left(\frac{-x^2}{2}\right)\left[\sum_{i=0}^{n} a_i He_i(x)\right]$$

Then the approximation given in Eq. (7), using the values of $p(x)$ at n points, is exact as far as the area and moments up to the $(n-1)$th order.

Proof. Since this method is an interpolation using the zeroes of $He_n(x)$, the result will be exact as far as the first n coefficients go $(a_0, a_1, \cdots, a_{n-1})$. These determine the area (if the distribution is not normalized) and the first $n - 1$ moments.

Thus, if the distribution can be accurately approximated by an expression of the form Eq. (5), then this method should allow one to calculate the moments with great accuracy.

Checks on the convergence of this method are given in Table 5. These may be compared with the results in SPS 37-49, Vol. III, pp. 23–31, using the Hermite expansion. It will be seen that the first 2 moments (when they exist) are given more accurately by this procedure, but the approximation does not generally converge uniformly or in the mean-square sense.

Table 5. Convergence of Gram–Charlier approximation

Name of distribution	Scale factor	No. of interpolation points	Area	Mean	Variance	L_2 norm of error	L_∞ norm of error
Unknown phase angle	1	72	0.89464	0.03834	1.00607	0.00000	0.00000
		2	0.89371	0.01270	not defined	0.00637	0.00436
		3	0.89420	0.03808	1.00046	0.00926	0.00724
		4	0.89464	0.03818	1.00405	0.00073	0.00052
		5	0.89464	0.03830	1.00600	0.00026	0.00020
		6	0.89464	0.03834	1.00604	0.00056	0.00049
		7	0.89464	0.03834	1.00606	0.00012	0.00010
		8	0.89464	0.03834	1.00607	0.00003	0.00003
		9	0.89464	0.03834	1.00607	0.00005	0.00005
		10	0.89464	0.03834	1.00607	0.00002	0.00002
		12	0.89464	0.03834	1.00607	0.00001	0.00001
		14	0.89464	0.03834	1.00607	0.00000	0.00000
Cauchy distribution	1	2	0.65774	—	—	0.15407	0.09913
		3	0.82991	—	—	0.14425	0.09250
		4	0.78861	—	—	0.13492	0.09066
		5	0.85464	—	—	0.08738	0.05555
		6	0.84093	—	—	0.12698	0.11324
		7	0.87276	—	—	0.10223	0.06921
		8	0.86864	—	—	0.31294	0.23133
		9	0.88631	—	—	0.12124	0.09026
		10	0.88590	—	—	0.86881	0.76146
		12	0.89780	—	—	3.06090	2.59333
		14	0.90661	—	—	11.80041	10.21137
		16	0.91347	—	—	49.40664	42.85682
		20	0.92359	—	—	0.102×10^4	0.895×10^3
		40	0.94733	—	—	0.204×10^{11}	0.183×10^{11}
		48	0.95215	—	—	0.250×10^{14}	0.22×10^{14}
Normalized student t distribution, $\nu = 3$	$3^{1/2}$	2	0.65774	0	not defined	0.20799	0.23768
		3	1.21284	0	0.36854	0.12301	0.11172
		4	0.84759	0	0.86082	0.18303	0.20991
		5	1.09702	0	0.55811	0.08198	0.06813
		6	0.92305	0	0.82206	0.11766	0.14354
		7	1.04953	0	0.66315	0.06892	0.05695
		8	0.95777	0	0.81389	0.12469	0.14404
		9	1.02707	0	0.72590	0.04485	0.03975
		10	0.97536	0	0.81655	0.03654	0.04593
		12	0.98491	0	0.82311	0.22953	0.22499
		14	0.99038	0	0.83084	0.50863	0.41463
		16	0.99365	0	0.83861	2.02287	1.77571
		48	0.99979	0	0.90469	0.294×10^{13}	0.263×10^{12}
Normalized srudent t distribution, $\nu = 20$	1.0540	2	0.97285	0	not defined	0.01803	0.01636
		3	1.00220	0	0.92257	0.00974	0.00657

Table 5 (contd)

Name of distribution	Scale factor	No. of inter-polation points	Area	Mean	Variance	L_2 norm of error	L_∞ norm of error
Normalized student t distribution, $\nu = 20$	1.0540	4	0.99810	0	0.99545	0.01714	0.01545
		5	1.00014	0	0.99057	0.00117	0.00091
		6	0.99979	0	0.99850	0.00123	0.00135
		7	1.00000	0	0.99838	0.00126	0.00090
		8	0.99997	0	0.99955	0.00352	0.00317
		9	1.00000	0	0.99963	0.00015	0.00011
		10	0.99999	0	0.99985	0.00130	0.00104
		12	1.00000	0	0.99995	0.00244	0.00213
		14	1.00000	0	0.99998	0.00285	0.00245
		20	1.00000	0	1.00000	0.01690	0.01478
		36	1.00000	0	1.00000	0.597×10^{2}	0.533×10^{2}
		56	1.00000	0	1.00000	0.533×10^{9}	0.479×10^{8}
Extreme value	1	2	0.81657	0.36418	not defined	0.10223	0.09394
		3	1.02431	0.39141	0.53758	0.08334	0.07552
		4	0.98532	0.35252	0.95907	0.10672	0.09395
		5	0.98253	0.47214	0.77955	0.04019	0.03835
		6	1.01088	0.39468	0.96166	0.04312	0.04719
		7	0.98518	0.46418	0.91741	0.07504	0.07114
		8	1.00778	0.42780	0.95980	0.05766	0.05532
		9	0.99343	0.45136	0.97471	0.02371	0.02370
		10	1.00239	0.44489	0.96709	0.06557	0.06054
		12	0.99969	0.45080	0.97845	0.13009	0.10453
		20	0.99991	0.44975	0.99928	6.16576	5.34279
		32	0.99999	0.45007	0.99980	0.922×10^{4}	0.818×10^{4}
		40	1.00000	0.45005	0.99998	0.212×10^{7}	0.190×10^{7}

II. Systems Analysis

SYSTEMS DIVISION

A. A Proposed Venus Coordinate System,

F. M. Sturms, Jr.

1. Radar Studies of Venus

During 1964 and 1966, radar studies of Venus (Refs. 1–3) have produced solutions for the radius, axis and rotation period, and also identified several surface features. This knowledge permits, for the first time, specification of coordinate systems associated with the equatorial plane of Venus. Selection of such a coordinate system is complicated somewhat by the fact that Venus rotation is retrograde.

From Ref. 2, the best solutions for the rotation (or angular momentum) vector and period are as follows:

(1) Right ascension (α_0) = 98 ±5 deg.

(2) Declination (δ_0) = −89 ±2 deg.

(3) Period = 242.6 ±0.6 days.

From Refs. 1 and 2, the prime meridian, or zero aphrodiographic longitude, is chosen to pass through a prominent narrow feature denoted as F or α. However, the coordinate system proposed in this report is based on a choice of north pole opposite that used in Refs. 1 and 2. This article discusses the reasons for this choice.

2. Coordinate System Geometry

In 1964, R. Richard[1] presented arguments for standardizing the method of choosing the north pole and the direction for measuring positive longitude. The advantages described include a reduced possibility of confusion, due to the proposed analogy to terrestrial conventions, and a single set of formulas for expressing rotations, angles, and oblateness perturbations. Accordingly, the following conventions are adopted for Venus:

(1) The north pole is that end of the rotational axis in the direction of the angular momentum vector (right-hand rule).

(2) Body-fixed longitude is measured positive in the direction of rotation, i.e., with convention (1), to the east.

Convention (1) is opposite to that given in Refs. 1 and 2, and convention (2) is opposite to that generally used in Refs. 4 and 5.

a. Adopted pole and rates. By the above convention, the north pole of Venus has the right ascension and

[1] Richard, R. J., *On a Standardized Method of Reckoning Longitude on the Various Celestial Bodies*, June 16, 1964 (JPL internal document).

declination given in *Subsection 1*. Because of the fairly large uncertainty in the values, the epoch associated with these values is not tightly constrained and shall be taken as 1964.5 (a convenient value near the Venus conjunction of that year). Also, the values shall be taken as being with respect to the mean equator and equinox of date.

The values of the pole location will change with time due to the precession of both the earth and Venus equators. At the present time, since no estimate of the oblateness of Venus is available, the precession of Venus is taken as zero. Therefore, due to the precession of earth

$$\frac{d\alpha_0}{dt} = m + n \sin \alpha_0 \tan \delta_0$$

$$\frac{d\delta_0}{dt} = n \cos \alpha_0$$

Using values of the annual general precession in right ascension, *m*, and the annual general precession in declination, *n* (Ref. 5, p. 38), the resulting pole location is

$$\alpha_0 = \quad 98 - 0.0015551\,(t - 1964.5)\ \mathrm{deg}$$

$$\delta_0 = -69 - 0.0007748\,(t - 1964.5)\ \mathrm{deg}$$

where $(t - 1964.5)$ is in tropical years.

b. Equator and orbit angles. Given the location of the pole of Venus, several useful angles describing the orientation of the equator and orbit of Venus may be computed. Using the formulas on p. 332 of Ref. 5, and the values of the mean orbital elements of Venus from p. 113 of Ref. 5, the results for the epoch 1964.5 are as follows:

☊ = angle from mean equinox along ecliptic to ascending node of the Venus mean orbit = 76.360 deg

i = inclination of the Venus mean orbit to ecliptic = 3.394 deg

Ω = angle from node, ☊, along the Venus orbit to *descending* node of orbit on equator (Venus autumnal equinox) = 290.878 deg

I = inclination of Venus orbit to Venus equator (Venus obliquity) = 176.545 deg

Δ = angle from ascending node of Venus equator on earth *mean* equator along Venus equator to autumnal equinox = 180.075 deg

Note that for Venus, the *vernal* equinox is analogous to that of earth, i.e., the point where the sun crosses from the southern hemisphere to the northern hemisphere (beginning of northern spring). Because of the north-pole convention used, the Venus obliquity is greater than 90 deg. The proper quadrants for these angles follow unambiguously from the equations in Ref. 5.

The obliquity of the Venus equator is very nearly 180 deg and, consequently, the seasons are not very different from one another in terms of the incidence of the sun's rays and the maximum elevation of the sun at noon. Coupled with the nearly circular orbit of Venus, this results in a day–night cycle that is nearly constant.

Finally, it is interesting to note that the Venus equinoxes lie very nearly in a plane parallel to the earth equatorial plane.

3. Venus Rotation

a. The Venus day. The Venus sidereal day is, as given in *Subsection 1,* 242.6 ephemeris days, and the sidereal rotation rate is, correspondingly, 1.484 deg/day. With the adopted north-pole convention, the apparent star motion is from east to west.

The mean orbital motion is 1.602 deg/day, and the sun appears to move from east to west in right ascension against the star background, which is opposite to that seen from earth.

These motions combine to form a solar day that is shorter than the sidereal day, contrary to that of earth. The mean rotation rate, with respect to the sun, is the sum of the above rates (3.086 deg/day), and the Venus mean solar day is, therefore, 116.7 ephemeris days. The apparent solar motion is from east to west.

b. The prime meridian and central meridian. In Refs. 1 and 2, the Venus prime meridian is chosen to pass through a narrow feature identified as *F* for α. This choice is also made here. The method for establishing the prime meridian is to define the sub-earth longitude or central meridian[2] at some epoch. Thus, following Ref. 1, the apparent aphrodiographic longitude of the earth at 0^h ephemeris time (ET) on June 20, 1964 (JD 243 8566.5) is +40 deg. (Note that the epoch has been arbitrarily changed to 0^h ET, rather than 0^h UT, in order to simplify the computations below.) Because of the reversed pole, the values in Table 1 of Refs. 1 and 2 should be changed

[2] The longitude at the apparent center of Venus as seen from earth.

Table 1. Phenomena for proposed Venus coordinate system

Same as earth	Opposite to earth
Dependent on north-pole convention, i.e., reverses if convention is reversed	
1. Apparent star motion east to west	1. Venus obliquity greater than 90 deg. Sun moves east to west in RA
2. Right ascension (RA) positive in direction of rotation	
3. Sun rises in east, sets in west	
4. Hour angle of equinox opposite rotation, increases with time	
5. Effect of Venus precession is to increase Venus RA	
Independent of north-pole convention	
1. Longitude positive east (longitude of central meridian reverses with convention)	1. Sun moves opposite rotation
2. Definition of vernal equinox (identity of given intersection reverses with convention)	2. Solar day shorter than sidereal day
3. Hour angle positive west	3. Effect of Venus precession is to decrease Venus celestial longitude
4. RA positive east	

by reversing the signs on the latitudes and longitudes of the features.

From the discussion on pp. 335 and 336 of Ref. 5, the longitude of the central meridian, λ, is given by

$$\lambda = A_E - V + \frac{R\omega}{\tau}$$

(note reversed signs to account for reversed convention for positive longitude)

where

A_E = Venus right ascension of apparent earth

V = hour angle of Venus vernal equinox from prime meridian

and the third term is the rotation during the light time, where

R = earth–Venus distance (AU)

τ = light time for 1 AU = 499.012 s

ω = sidereal rotation rate

From the equations on p. 334 of Ref 5, A_E is computed in terms of the right ascension and declination of the Venus pole (α_0, δ_0) and the apparent Venus coordinates (α, δ). Note that α, δ, α_0, δ_0 and Δ must be consistently given with respect to either the mean or true earth equator and equinox. The quantities A_E and D_E are independent of the choice, and P will be measured from a mean or true declination circle, respectively (D_E = aphrodiocentric latitude of earth; P = position angle of Venus' north pole from earth declination circle).

From Ref. 4 for 0^h ET on June 20, 1964, with respect to the true equator and equinox of date

$$\alpha = 5^h\,53^m\,59^s\!.71 = 88.4988 \text{ deg}$$

$$\delta = 21^\circ\,36'\,28''\!.3 = 21.6079 \text{ deg}$$

$$R = 0.2395 \text{ AU}$$

Converting to mean equinox and equator

$$\alpha = 88.5040 \text{ deg}$$

$$\delta = 21.6081 \text{ deg}$$

and

$$A_E = 278.75 \text{ deg}$$

$$D_E = 0.87 \text{ deg}$$

$$P = 176.61 \text{ deg}$$

Then, the reference value of V is (light time correction is negligible to significance retained)

$$V_0 = 278.75 - 40 + 0.002 = 238.75 \text{ deg}$$

and subsequently

$$V = 238.75 + 1.483924\,(\text{JD} - 243\,8566.5) \text{ deg}$$

$$\lambda = A_E - V + 0.0086\,R$$

Finally, it should be noted that the Venus right ascension and declination of the earth, A_E and D_E, are measured positive east (in direction of rotation) and north, from the Venus vernal equinox and equator, respectively, and the hour angle of the equinox, V, is measured positive west from the prime meridian to the equinox. These are analogous to the measurement conventions on earth.

4. Coordinate Transformations

The mean earth equator and equinox of 1950.0 is a standard non-rotating coordinate system in common use.

The cartesian position transformations to Venus coordinate systems involve the following rotations:

(1) Rotate to mean earth equator and equinox of date (precession matrix A).

(2) Rotate $\alpha_0 + 90$ deg about Z axis (matrix S_1).

(3) Rotate $90 - \delta_0$ deg about X axis (matrix S_2).

(4) Rotate $\Delta + 180$ deg about Z axis (matrix S_3).

At this point, the coordinates are with respect to the Venus equator and equinox of date. Two options are as follows:

(1) Rotate I about X axis (matrix E). This yields coordinates with respect to Venus mean orbit and equinox of date.

(2) Rotate V about Z axis (matrix H). This yields coordinates with respect to Venus equator and prime meridian (aphrodiographic).

Then, in summary

Venus equator and equinox: $(X) = S_3\, S_2\, S_1\, A\, (X)_{1950.0}$

Venus orbit and equinox: $(X) = E\, S_3\, S_2\, S_1\, A\, (X)_{1950.0}$

Aphrodiographic: $(X) = H\, S_3\, S_2\, S_1\, A\, (X)_{1950.0}$

where

A = precession matrix (Ref. 5)

$$S_1 = \begin{pmatrix} -\sin\alpha_0 & \cos\alpha_0 & 0 \\ -\cos\alpha_0 & -\sin\alpha_0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sin\delta_0 & \cos\delta_0 \\ 0 & -\cos\delta_0 & \sin\delta_0 \end{pmatrix}$$

$$S_3 = \begin{pmatrix} -\cos\Delta & -\sin\Delta & 0 \\ \sin\Delta & -\cos\Delta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos I & \sin I \\ 0 & -\sin I & \cos I \end{pmatrix}$$

$$H = \begin{pmatrix} \cos V & \sin V & 0 \\ -\sin V & \cos V & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The velocity rotations are obtained from differentiation of the above matrix equations.

5. Discussion

The proposed Venus coordinate system is based on conventions for defining the north pole and the direction of positive longitude. The convention for measuring longitude positive east has been adopted by the International Astronomical Union (IAU) (Ref. 6, p. 174) in conjunction with gravitational potential expressions, and is undoubtedly the best choice. The choice of north-pole convention is not so clear, however. In making the choice, it was desired to retain as much analogy and consistency with earth as possible. Accordingly, Table 1 presents a list of items pertaining to the proposed Venus coordinate system. The table is a useful aid in visualizing phenomena as they appear relative to a Venus observer.

The adoption of the proposed coordinate system leads to the question of how improved values are incorporated. The following procedures are based on historical precedent.

Improved values of the pole location should be stated in terms of the 1964.5 values. This can be done by mapping a solution for a current epoch backward, or by solving directly in terms of the 1964.5 value and mapping forward to compute current observations. The improved location should be included in the rates due to the earth precession.

When information on the figure of Venus has been obtained, the precession of the Venus equator can be determined (Ref. 5, p. 327). This can then be incorporated into the computation of the pole location rates. If the rate of precession of the Venus equator on the Venus orbit is denoted by μ, the contribution to the rates is (Ref. 7).

$$\frac{d\alpha_0}{dt} = \mu \sin I \cos\Delta \sec\delta_0$$

$$\frac{d\delta_0}{dt} = \mu \sin I \sin\Delta$$

(Note: for Venus, μ is positive, i.e., in the direction of *orbital* motion, whereas it is normally negative for other planets.)

A more precise value for the rotation period of Venus will directly update the rate term in the expression for V. The lengths of the solar and sidereal days are easily corrected.

The leading term in the expression for V must be rederived for improved values of α_0 and δ_0. Changes will enter through a different value of A_E and also the inclusion of the light time term, if it is significant. In this step, the longitude of the central meridian at the reference epoch is unchanged, i.e., it is fixed at a value of 40 deg. This procedure is similar to that followed for the physical ephemeris of Mars, where initially the longitude of the central meridian is computed to place the prime meridian through a prominent feature. Subsequently, however, the longitude of the central meridian at the reference epoch is held constant, and the longitudes of the prominent feature, as well as all other features, will vary slightly.

References

1. Carpenter, R. L., "Study of Venus by CW Radar—1964 Results," *Astron. J.*, Vol. 71, No. 2, Mar. 1966. Also available as Technical Report 32-963, Jet Propulsion Laboratory, Pasadena, Calif.
2. Goldstein, R. M., "Radar Studies of Venus," *Moon and Planets*, North-Holland Publishing Co., Amsterdam, 1967. Also available as Technical Report 32-1081, Jet Propulsion Laboratory, Pasadena, Calif.
3. Ash, M. E., Shapiro, I. I., and Smith, W. B., "Astronomical Constants and Planetary Ephemerides Deduced from Radar and Optical Observations," *Astron. J.*, Vol. 72, No. 3, Apr. 1967.
4. *American Ephemeris and Nautical Almanac, 1964*. United States Government Printing Office, Washington, 1962.
5. *Explanatory Supplement to the Ephemeris*. Her Majesty's Stationery Office, London, 1961.
6. "Proceedings of the Eleventh General Assembly, Berkeley, Calif., 1961," *Trans. IAU*, Vol. XIB. Academic Press, New York, 1962.
7. de Vaucouleurs, G., "The Physical Ephemeris of Mars," *Icarus*, Vol. 3, 1964.

III. Computation and Analysis

SYSTEMS DIVISION

A. Orthonormal Transformations for Linear Algebraic Computations, *C. L. Lawson*

1. Introduction

The basic step in many methods for solving systems of linear equations, or computing eigenvalues or singular values of a matrix, may be interpreted as premultiplication of a matrix A by a matrix T, where T is chosen so that certain elements of TA are zero. Methods in which T is orthonormal are stable with respect to growth of rounding errors and are particularly appropriate in least-squares computations because of their property of preserving the euclidean length of vectors.

In this article, we review the properties of two orthonormal transformations that are well known in numerical analysis, and introduce a third orthonormal transformation that combines certain features of the first two.

We denote the transpose of an m-vector $\mathbf{v}$ by $\mathbf{v}^T$ and its euclidean norm by

$$||\mathbf{v}|| = \mathbf{v}^T\mathbf{v} = \sum_{i=1}^{m} \mathbf{v}_{(i)}^2$$

The identity matrix of order m is denoted by I_m.

All of these transformations can be discussed in the following setting:

Problem. Given an m-vector $\mathbf{v}$, find an $m \times m$ orthonormal matrix Q such that components 2 through m of $Q\mathbf{v}$ are zero.

Since only the first element of $Q\mathbf{v}$ is permitted to be nonzero, and since $||Q\mathbf{v}|| = ||\mathbf{v}||$, it follows that the first component of $Q\mathbf{v}$ must be either $||\mathbf{v}||$ or $-||\mathbf{v}||$.

2. The Jacobi Transformation

A single Jacobi transformation alters only two components of a vector, one of which will be transformed to zero if the transformation matrix is appropriately chosen. Thus, the *Problem*, above, can be solved by a sequence of $m-1$ Jacobi transformations.

A Jacobi transformation matrix can be denoted by $B_{i,j,\theta}$ and is identical with the $m \times m$ identity matrix I_m with the exception of the four elements

$$b_{ii} = b_{jj} = c = \cos\theta$$

$$b_{ij} = -b_{ji} = s = \sin\theta$$

Let $\tilde{\mathbf{v}} = \mathrm{B}_{i,j,\theta}\,\mathbf{v}$, and suppose θ is to be chosen so that $\tilde{\mathbf{v}}_{(j)} = 0$. This is accomplished by computing

$$d = (\mathbf{v}^2_{(i)} + \mathbf{v}^2_{(j)})^{1/2}$$

$$c = \begin{cases} \mathbf{v}_{(i)}/d & \text{if } d \neq 0 \\ 1 & \text{if } d = 0 \end{cases}$$

$$s = \begin{cases} \mathbf{v}_{(j)}/d & \text{if } d \neq 0 \\ 0 & \text{if } d = 0 \end{cases}$$

Then

$$\tilde{\mathbf{v}}_{(i)} = d$$

$$\tilde{\mathbf{v}}_{(j)} = 0$$

$$\tilde{\mathbf{v}}_{(k)} = \mathbf{v}_{(k)} \textit{ for } k \neq i \text{ and } k \neq j$$

A geometric interpretation of the Jacobi transformation is given in Fig. 1.

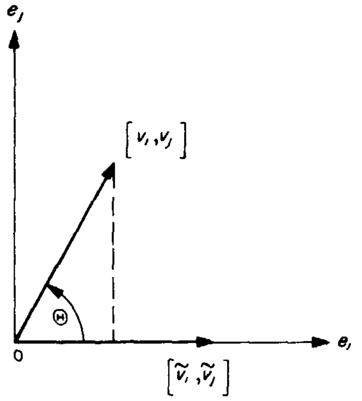


Fig. 1. Geometric interpretation of a Jacobi transformation

The multiplication $\tilde{\mathbf{x}} = B\mathbf{x}$, where $\mathbf{x}$ is an arbitrary m-vector, can be done as

$$\tilde{\mathbf{x}}_{(i)} = c\mathbf{x}_{(i)} + s\mathbf{x}_{(j)}$$

$$\tilde{\mathbf{x}}_{(j)} = -s\mathbf{x}_{(i)} + c\mathbf{x}_{(j)}$$

$$\tilde{\mathbf{x}}_{(k)} = \tilde{\mathbf{x}}_{(k)} \qquad \text{for } k \neq i \text{ and } k \neq j$$

3. The Householder Transformation

An $m \times m$ Householder transformation matrix, $H_\mathbf{w}$, may be parameterized by an m-vector $\mathbf{w}$, where either $\mathbf{w} = 0$ or $||\mathbf{w}|| = 1$. If $\mathbf{w} = 0$, we define $H_\mathbf{w} = I_m$.

If $||\mathbf{w}|| = 1$, the matrix $H_\mathbf{w}$ is a reflection matrix characterized by the fact that it transforms $\mathbf{w}$ to $-\mathbf{w}$ and acts as an identity on the $(m - 1)$-dimensional subspace, $\mathcal{S}$, orthogonal to $\mathbf{w}$. These properties completely characterize the eigenvalue–eigenvector structure of $H_\mathbf{w}$ and thus permit an explicit construction of $H_\mathbf{w}$ as follows:

Let $\mathbf{p}_2, \cdots, \mathbf{p}_m$ be an orthonormal basis for $\mathcal{S}$ and let $P = [\mathbf{w}, \mathbf{p}_2, \cdots, \mathbf{p}_m]$. Then P is an $m \times m$ orthonormal matrix and

$$H_\mathbf{w} P = PD$$

where $D = \mathrm{diag}\,(-1, 1, 1, \cdots, 1)$.

Let E_{11} denote an $m \times m$ matrix whose only nonzero element is a one in the $(1, 1)$ position. Then

$$H_\mathbf{w} = PDP^T = P(I - 2E_{11})P^T$$

$$= I - 2PE_{11}P^T = I - 2\mathbf{w}\mathbf{w}^T$$

Now consider the *Problem* presented in *Subsection 1*. Let the m-vector $\mathbf{v}$ be given. Define

$$\sigma = \mathrm{sgn}\,(\mathbf{v}_{(1)}) = \begin{cases} +1 & \text{if } \mathbf{v}_{(1)} \geq 0 \\ -1 & \text{if } \mathbf{v}_{(1)} < 0 \end{cases}$$

Define the m-vector $\mathbf{e}_1$ by $\mathbf{e}_1 = (1, 0, \cdots, 0)^T$. Let $\mathbf{w}$ be the unit vector bisecting the angle between $\mathbf{v}$ and $\sigma||\mathbf{v}||\mathbf{e}_1$; explicitly

$$\mathbf{u} = \mathbf{v} + \sigma||\mathbf{v}||\mathbf{e}_1$$

$$\mathbf{w} = \begin{cases} \mathbf{u}/||\mathbf{u}|| & \text{if } \mathbf{u} \neq 0 \\ 0 & \text{if } \mathbf{u} = 0 \end{cases}$$

The matrix $H_\mathbf{w} = I_m - 2\mathbf{w}\mathbf{w}^T$ solves the *Problem* since

$$\tilde{\mathbf{v}} = H_\mathbf{w}\mathbf{v} = -\sigma||\mathbf{v}||\mathbf{e}_1$$

A geometric interpretation of a Householder transformation is given in Fig. 2.

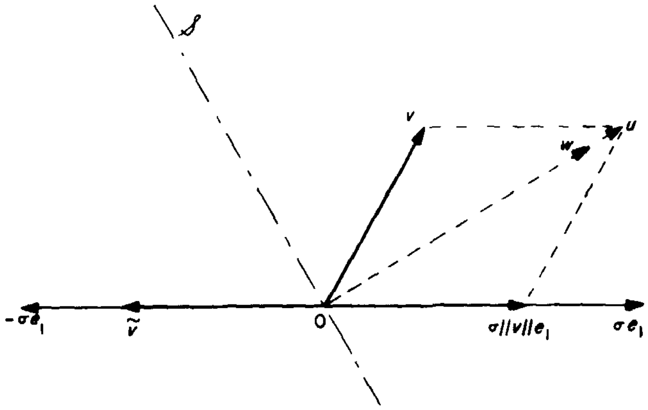


Fig. 2. Geometric interpretation of a Householder transformation

In a computer program, this computation is commonly organized so that the vector **w** is not explicitly computed. We may write

$$H_w = I_m + b^{-1}\,\mathbf{u}\mathbf{u}^T \tag{1}$$

where

$$\begin{aligned} -b &= \|\mathbf{u}\|^2/2 \\ &= (\mathbf{v} + \sigma\|\mathbf{v}\|\mathbf{e}_1)^T(\mathbf{v} + \sigma\|\mathbf{v}\|\mathbf{e}_1)/2 \\ &= \|\mathbf{v}\|^2 + \|\mathbf{v}\|\cdot|\mathbf{v}_{(1)}| \\ &= \sigma\|\mathbf{v}\|(\sigma\|\mathbf{v}\| + \mathbf{v}_{(1)}) \\ &= -\tilde{\mathbf{v}}_{(1)}(-\tilde{\mathbf{v}}_{(1)} + \mathbf{v}_{(1)}) = -\tilde{\mathbf{v}}_{(1)}\mathbf{u}_{(1)} \end{aligned} \tag{2}$$

Note that since **u** differs from **v** only in the first component, the construction of H_w, as given in Eq. (1), requires only the computation of $\mathbf{u}_{(1)}$ and b. Furthermore, the only nonzero element of $\tilde{\mathbf{v}}$ is $\tilde{\mathbf{v}}_{(1)}$. The computation of $\mathbf{u}_{(1)}$, b, and $\tilde{\mathbf{v}}_{(1)}$ can be organized as

$$\tilde{\mathbf{v}}_{(1)} = -\left(\sum_{i=1}^{m} \mathbf{v}_{(i)}^2\right)^{1/2} \operatorname{sgn}(\mathbf{v}_{(1)}) \tag{3}$$

$$\mathbf{u}_{(1)} = \mathbf{v}_{(1)} - \tilde{\mathbf{v}}_{(1)} \tag{4}$$

$$b = \tilde{\mathbf{v}}_{(1)}\,\mathbf{u}_{(1)} \tag{5}$$

If the matrix H_w is to be saved, it suffices to save the m-vector **u** and the scalar b. If $\tilde{\mathbf{v}}_{(1)}$ is also being saved, then one need not save b as it can be recomputed when needed using Eq. (5).

The multiplication $\tilde{\mathbf{x}} = H_w\mathbf{x}$ for an arbitrary m-vector **x** proceeds as

$$c = \left(\sum_{i=1}^{m} \mathbf{u}_{(i)}\,\mathbf{x}_{(i)}\right)\Big/ b$$

$$\tilde{\mathbf{x}}_{(i)} = \mathbf{x}_{(i)} + c\mathbf{u}_{(i)} \qquad i = 1, \cdots, m$$

4. The RSP Transformation

The RSP (Rotation in a Selected Plane) transformation will combine the Householder-like ability to transform $m-1$ elements of an m-vector to zero in a single transformation with the Jacobi-like property of using a plane rotation instead of a reflection.

Let $\mathcal{S}$ denote a two-dimensional subspace of m-space with orthonormal basis vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. We wish to construct an orthonormal matrix R that will act as a rotation in $\mathcal{S}$, rotating $\mathbf{w}_2$ through an angle θ toward $\mathbf{w}_1$, and act as an identity on $\mathcal{S}^{\perp}$, the orthogonal complement of $\mathcal{S}$.

Let $\mathbf{w}_3, \cdots, \mathbf{w}_m$ be an orthonormal basis for $\mathcal{S}^{\perp}$. Define

$$\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_m]$$

$$c = \cos\theta$$

$$s = \sin\theta$$

$$B = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

Then

$$\left.\begin{aligned} R\mathbf{W} &= \mathbf{W}\begin{bmatrix} B & 0 \\ 0 & I_{m-2} \end{bmatrix} \\ R &= \mathbf{W}\begin{bmatrix} B & 0 \\ 0 & I_{m-2} \end{bmatrix}\mathbf{W}^T = I_m \\ &\quad + [\mathbf{w}_1, \mathbf{w}_2]\,(B - I_2)\begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \end{bmatrix} \end{aligned}\right\} \tag{6}$$

Consider the *Problem* given in *Subsection 1.* Let **v** be a given m-vector, and again let

$$\sigma = \operatorname{sgn}\mathbf{v}_{(1)} = \begin{cases} +1 & \text{if } \mathbf{v}_{(1)} \geq 0 \\ -1 & \text{if } \mathbf{v}_{(1)} < 0 \end{cases}$$

$$\mathbf{e}_1 = [1, 0, \cdots, 0]^T \qquad (m\text{-dimensional})$$

We seek orthonoımal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ and an angle θ such that the matrix R, defined by Eq. (6), satisfies

$$\tilde{\mathbf{v}} = R\,\mathbf{v} = \sigma\,||\,\mathbf{v}\,||\,\mathbf{e}_1 \qquad (7)$$

Define

$$\left.\begin{aligned}
\mathbf{w}_1 &= \sigma\mathbf{e}_1 \\
\mathbf{u} &= \mathbf{v} - \mathbf{v}_{\{1\}}\mathbf{e}_1 = [0, \mathbf{v}_{\{2\}}, \mathbf{v}_{\{3\}}, \cdots, \mathbf{v}_{\{m\}}]^T \\
\mathbf{w}_2 &= \begin{cases} \mathbf{u}/||\,\mathbf{u}\,|| & \text{if } \mathbf{u} \neq 0 \\ [0, 1, 0, \cdots, 0]^T & \text{if } \mathbf{u} = 0 \end{cases} \\
c &= \begin{cases} |\,\mathbf{v}_{\{1\}}\,|/||\,\mathbf{v}\,|| & \text{if } \mathbf{v} \neq 0 \\ 1 & \text{if } \mathbf{v} = 0 \end{cases} \\
s &= \begin{cases} ||\,\mathbf{u}\,||/||\,\mathbf{v}\,|| & \text{if } \mathbf{v} \neq 0 \\ 0 & \text{if } \mathbf{v} = 0 \end{cases}
\end{aligned}\right\} \qquad (8)$$

It can be verified by substitution into Eq. (6) that these values of $\mathbf{w}_1$, $\mathbf{w}_2$, c, and s provide a matrix R that satisfies Eq. (7).

A geometric interpretation of these quantities is provided in Fig. 3.

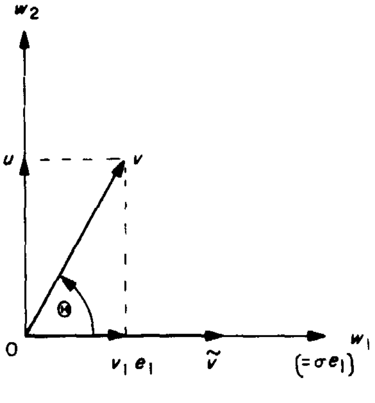


Fig. 3. Geometric interpretation of an RSP transformation

We now consider computational details. When $\mathbf{u} = 0$, we have $R = I_m$, and this case can be given special treatment. We thus consider only the case of $\mathbf{u} \neq 0$, which of course implies $\mathbf{v} \neq 0$.

It is possible to rewrite Eq. (6) as

$$R = I_m + [\mathbf{e}_1, \mathbf{u}]\,F \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{u}^T \end{bmatrix} \qquad (9)$$

where the elements of the 2×2 matrix F are

$$\begin{aligned}
f_{12} &= \sigma s\,||\,\mathbf{u}\,||^{-1} = \sigma\,||\,\mathbf{v}\,||^{-1} = \tilde{\mathbf{v}}^{-1}_{\{\cdot\}} \\
f_{21} &= -f_{12} \\
f_{22} &= (c - 1)\,||\,\mathbf{u}\,||^{-2} = -s^2\,(c + 1)^{-1}\,||\,\mathbf{u}\,||^{-2} \\
&= -\,[||\,\mathbf{v}\,||\cdot|\,\mathbf{v}_{\{1\}}\,| + ||\,\mathbf{v}\,||^2]^{-1} \\
&= -[\tilde{\mathbf{v}}_{\{1\}}\mathbf{v}_{\{1\}} + \tilde{\mathbf{v}}^2_{\{1\}}]^{-1} = -\tilde{\mathbf{v}}^{-1}_{\{1\}}(\mathbf{v}_{\{1\}} + \tilde{\mathbf{v}}_{\{1\}})^{-1} \\
f_{11} &= c - 1 = ||\,\mathbf{u}\,||^2 f_{22}
\end{aligned}$$

The computation of these quantities could proceed as

$$\begin{aligned}
||\,\mathbf{u}\,||^2 &= \sum_{i=2}^{m} \mathbf{v}^2_{\{i\}} \qquad (10) \\
\tilde{\mathbf{v}}^2_{\{1\}} &= \mathbf{v}^2_{\{1\}} + ||\,\mathbf{u}\,||^2 \\
\tilde{\mathbf{v}}_{\{1\}} &= (\tilde{\mathbf{v}}^2_{\{1\}})^{1/2}\,\mathrm{sgn}\,(\mathbf{v}_{\{1\}}) \\
f_{22} &= -[\tilde{\mathbf{v}}_{\{1\}}\mathbf{v}_{\{1\}} + \tilde{\mathbf{v}}^2_{\{1\}}]^{-1} \\
f_{11} &= ||\,\mathbf{u}\,||^2 f_{22} \\
f_{12} &= \tilde{\mathbf{v}}^{-1}_{\{1\}} \\
f_{21} &= -f_{12}
\end{aligned}$$

Saving the matrix R would require space for the $m - 1$ nonzero elements of $\mathbf{u}$ plus f_{22}, f_{11}, and f_{12}, i.e., a total of $m + 2$ locations. If $\tilde{\mathbf{v}}_{\{1\}}$ is also being saved, then f_{12} need not be saved.

To compute $\tilde{\mathbf{x}} = R\mathbf{x}$ for an arbitrary m-vector $\mathbf{x}$

$$\begin{aligned}
g &= \sum_{i=2}^{m} \mathbf{u}_{\{i\}}\mathbf{x}_{\{i\}} \\
h &= -f_{12}\mathbf{x}_{\{1\}} + f_{22}g \\
\tilde{\mathbf{x}}_{\{1\}} &= \mathbf{x}_{\{1\}} + (f_{11}\mathbf{x}_{\{1\}} + f_{12}g) \\
\tilde{\mathbf{x}}_{\{i\}} &= \mathbf{x}_{\{i\}} + h\mathbf{u}_{\{i\}} \qquad i = 2, \cdots, m
\end{aligned}$$

5. Conclusion

The Jacobi transformation is used primarily in cases in which the pattern of elements to be zeroed is somewhat irregular. When a number of elements in one column are to be zeroed, it is more economical and more accurate to use the Householder transformation. The RSP transformation is nearly as economical as the Householder transformation and could reasonably be used in the same circumstances.

Although the Householder transformation is very stable with regard to roundoff error propagation (see Ref. 1, p. 101), the RSP transformation may be even slightly more stable. The Householder transformation is applied in the form $H = I + G$ where, using the spectral matrix norm, $||G|| = 2$ for all $\mathbf{w}$ except for the special case of $\mathbf{w} = 0$ (which we will henceforth exclude). Similarly, the RSP transformation is applied in the form $R = I + K$, but

$$||K|| = [2(1 - c)]^{1/2} \qquad (11)$$

where c is defined by Eq. (8). In particular $0 \leq c \leq 1$, and thus

$$0 \leq ||K|| \leq 2^{1/2} \qquad (12)$$

and, consequently

$$||K|| \leq 0.71 \cdot ||G|| \qquad (13)$$

The relative roundoff error, ϵ_H, in computing $\tilde{\mathbf{x}}_{(H)} = \mathbf{x} + G\mathbf{x}$ using arithmetic having relative precision α, is bounded by

$$||\epsilon_H|| \leq \alpha\left(1 + \frac{||G\mathbf{x}||}{||\mathbf{x}||}\right) \leq \alpha(1 + ||G||) = 3\alpha \qquad (14)$$

with a similar bound of

$$||\epsilon_R|| \leq \alpha\left(1 + \frac{||K\mathbf{x}||}{||\mathbf{x}||}\right) \leq \alpha(1 + ||\mathbf{K}||) =$$

$$\alpha\{1 + [2(1 - c)]^{1/2}\} \leq 2.42\alpha \qquad (15)$$

for the computation of $\tilde{\mathbf{x}}_{(R)} = \mathbf{x} + K\mathbf{x}$. Since G and K are of rank 1 and 2, respectively, the ratios $||G\mathbf{x}||/||\mathbf{x}||$ and $||K\mathbf{x}||/||\mathbf{x}||$ are usually not close to their respective upper bounds, $||G||$ and $||K||$ (say, averaging over all $||\mathbf{x}|| = 1$). Thus, comparison of average behavior cannot be based on Eqs. (14) and (15). Further investigation will be made of the relative merits of the Householder and the RSP transformations.

Reference

1. Ralston, A., and Wilf, H., *Mathematical Methods for Digital Computers: Volume II*, John Wiley & Sons, Inc., New York, 1967.

IV. Spacecraft Power

GUIDANCE AND CONTROL DIVISION

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

1. Introduction

A project was initiated by JPL in 1962 to improve the accuracy of predicting solar array performance in space. High-altitude balloon flights have been used to achieve the near-zero air-mass conditions required for calibrating the solar cell standards.

Balloon-calibrated solar cells in modular form were recovered and mounted on a temperature-controlled housing (Fig. 1) and used as intensity reference standards during performance testing of solar arrays under terrestrial sunlight conditions. It has been shown by Ritchie (Ref. 1, pp. 6 and 7) that, if the standard solar cell and the cells used for array fabrication have the same spectral response, the space short-circuit current output of the solar array can be predicted with an accuracy of better than 2%. Since 1962, high-altitude balloon flights for solar cell standardization have been conducted at the rate of three or four flights per year. Cooperative efforts between JPL and other NASA and government agencies have provided standard solar cells at minimum expense for a variety of space projects and advanced development work.

2. 1968 Balloon Flight Project

Three 80,000-ft balloon flights are scheduled for July and August 1968. Fabrication and testing of standard

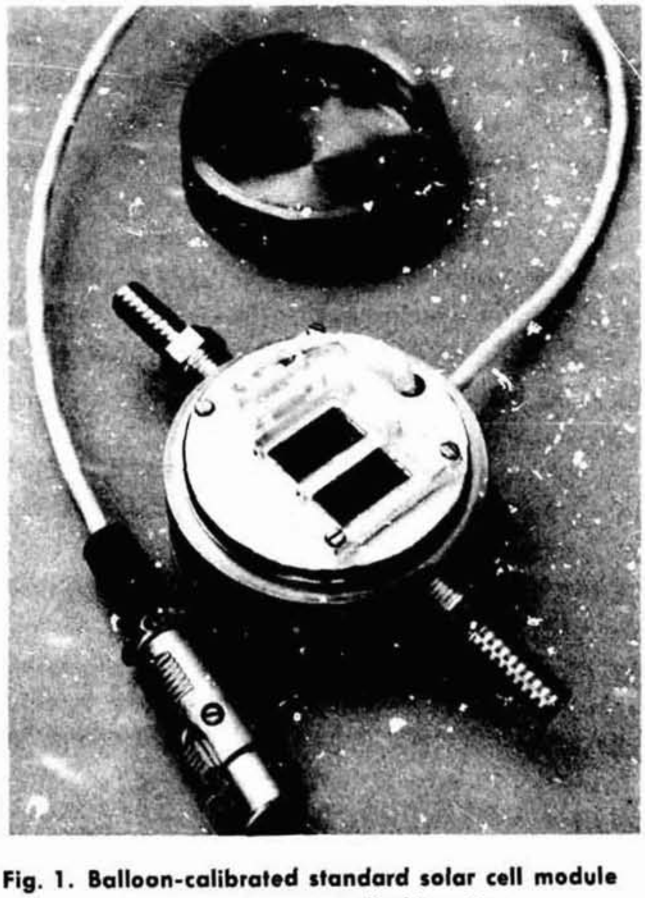


Fig. 1. Balloon-calibrated standard solar cell module on temperature-controlled housing

solar cell modules are in progress. The cooperative effort between JPL and other government agencies is continuing this year with the Air Force Aero Propulsion Laboratory, the Johns Hopkins University, the NASA Langley Research Center, and the NASA Goddard Space Flight Center supplying standard solar cell modules for calibration.

Improvements to the balloon flight system are currently in progress. Design modifications of the solar tracker having an increased payload capacity have been completed, and actual modification has begun. Figure 2 shows the old solar tracker configuration. The modified solar tracker will provide for 26 solar cell calibration channels, an increase of 12 channels. This will be accomplished by replacing the old 24-position stepping switch with a new 36-position stepping switch. At the same time, the solar cell module mounting area will be increased to accommodate the added module capability.

Due to the increased amount of data returned per flight as a result of increased payload capacity, improved methods of data reduction are required. To meet this problem, flight data will be supplied by the balloon flight project contractor on IBM punched cards. A JPL computer program is in the process of being updated, which will be compatible with the contractor-supplied data. The computer program will reduce, average, and correct the solar cell data for intensity and temperature. A summary sheet will give solar cell descriptive information along with calibration data at a standard intensity and temperature. It is expected that, through improved data handling and processing methods, calibration data will be available within a few days following a balloon flight series.

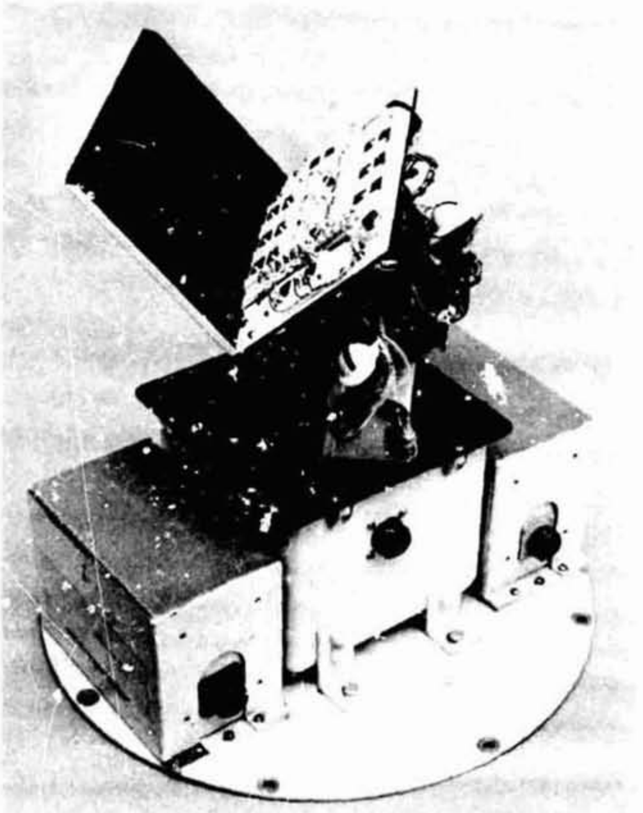


Fig. 2. Present balloon apex-mounted solar tracker

Reference

1. Ritchie, D. W., *Development of Photovoltaic Standard Cells for NASA*, Technical Report 32-634. Jet Propulsion Laboratory, Pasadena, Calif., June 1, 1964.

B. Solar Power System Definition Studies,

H. M. Wick

1. Introduction

The overall objective of this effort is to investigate the problems associated with developing spacecraft power systems for unmanned planetary missions. The effort stresses development of the technology required to solve system design problems associated with meeting JPL mission requirements. One task which is presently being undertaken is the investigation and development of computer programs for power system design, integration and analysis.

2. Power Profile Computer Program for a Mars 1971 Mission Study

The successful development of a spacecraft power system requires a compromise between user power requirements and available power limitations. Due to the multitude of changes in the user system power requirements during the spacecraft design phases, continuous monitoring by the spacecraft power design team becomes an absolute necessity. Data processing methods provide both an effective and accurate method for maintaining an up-to-date status of the spacecraft power requirements.

A computer program was recently developed to assist in determining the spacecraft electrical power requirements for power system sizing and spacecraft power management for a Mars 1971 mission.

A functional block diagram of the power system with the spacecraft loads is shown in Fig. 3, which represents

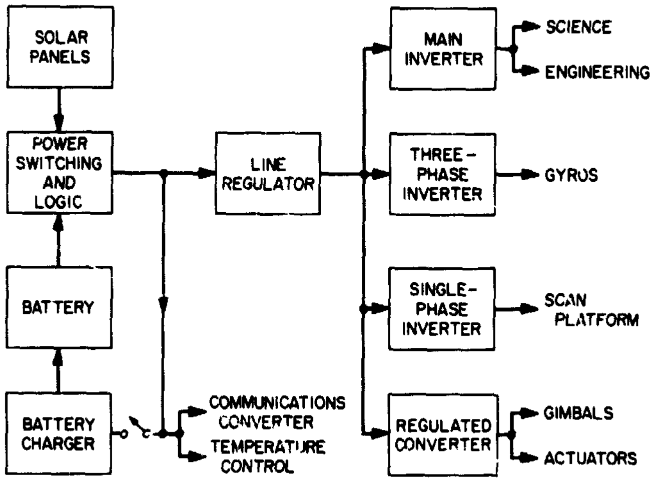


Fig. 3. Power system functional block diagram

the power system model used by the power profile computer program. Power is derived from photovoltaic solar panels and a secondary battery. The power switch and logic (PS&L) distributes raw power to the line regulator, battery charger, communications converter, and temperature control system.

The first two pages of the computer printout lists all input data for reference. The spacecraft systems and their power requirements for each of the mission flight phases are tabulated on the first page. On the second page is a listing of power and efficiency data points for each of the inverters and the line regulator. These data are used by an interpolation subroutine to define the operating efficiency as a function of power output. The program then calculates and prints out the power output, efficiency, and power input for each of the inverters of the power system along with their respective user system requirements. The line regulator power output, efficiency, and power input is then determined and listed. The total power demand of the spacecraft power system is obtained by summing the PS&L loads and dividing by the PS&L efficiency. This process continues until all power system operating modes and mission flight phases have been considered. This computer program is an extension of the programming work done to support *Mariner* Mars 1969. The program has been written in Fortran IV for the IBM 7094.

3. Battery Cell Data Reduction Program

A Fortran II program was written for the IBM 1620 data processing system. The program appropriately reduces raw battery cell data in addition to providing plots of the cell discharge curves.

4. Shepherd's Equation Battery Discharge Programs

The BATT3 and BATT4 battery discharge programs have been verified. These programs are now considered operational; however, additional checkout runs are planned as soon as more detailed battery discharge data become available.

C. Development of Improved Solar Cell Contacts, *P. Berman*

1. Introduction

Silicon solar cells are presently the most reliable direct energy converters for space applications, and it appears that this will continue to be the case for some time to come. Over the past years, there have been significant increases in cell conversion efficiency, as well as reduction in cell size and manufacturing costs; however, improvements of the same magnitude have not been made in the area of solar cell contacts and solar cell interconnection techniques. The environmental limitations imposed on the solar cell contacts to avoid mechanical and electrical degradation have remained the same for many years, and in some cases have even become more restrictive. Therefore, solar panels are environmentally limited in many cases as a result of solar cell contact restrictions, and it can be expected that significant improvements on solar panel reliability will result from improvements in solar cell contacts and interconnection techniques.

The objective of this study is the development of silicon solar cell electrical contacts and interconnection techniques which are less susceptible to mechanical and electrical degradation resulting from exposure to extremes of earth- and space-type environments. A major objective is the development of cell electrical contacts and interconnection techniques which do not require the use of solder. There should be less degradation of contact strength and electrical characteristics after exposure to thermal shock, humidity–temperature, vacuum–temperature, high-temperature, and low-temperature environments. The solar cell contact and interconnection techniques are also to be optimized with respect to the (1) effects on solar cell current-voltage characteristics, (2) series and/or contact resistance, (3) stresses due to fabrication procedure, (4) compatibility with requirements for fabrication into submodules, (5) reliability, (6) handling and manufacturing characteristics and restraints, (7) repair or rework capability, (8) reproducibility, (9) production cost, (10) ease of production, (11) weight, (12) compatibility with large-area cells, (13) requirements for special equipment and tooling, and (14) compatibility with inorganic, integral protective coatings.

2. Development Activities

Development contracts have been awarded to Ion Physics Corporation and the Librascope Division of General Precision Systems, Inc., and work was initiated in January 1968. Ion Physics is presently utilizing its high-vacuum sputtering technique, and Librascope is utilizing its cold-substrate deposition process to deposit the contact materials onto the silicon.

The first material to be investigated by both organizations is aluminum. Ion Physics has produced and delivered to JPL sample solar cells having aluminum contacts. Cells have been fabricated having an efficiency of 9–10% at air mass zero (28°C). This compares quite favorably with the 11–11.5% efficiencies characteristic of state-of-the-art solar cells when one considers the developmental status of the former cells. Thus far, only adhesive tape-peel tests (utilizing Scotch tape) have been used to evaluate contact adherence. Several cells exhibited contact peeling as a result of these tests; however, most of the cells were capable of passing with no apparent peeling of the contacts. Two cells which did not exhibit peeling were placed in a humidity chamber at 60°C and 95% relative humidity for a period of 1 week. The cells showed no apparent deterioration in contact adhesion. Preliminary attempts to utilize parallel gap welding have not been successful, due to the oxide layer on the aluminum which inhibits constant current flow. The technique is still under investigation.

Librascope has deposited aluminum on low resistivity (approximately 0.001 Ω-cm) *n*-type silicon, which is representative of the diffused layer of an *n/p* solar cell. The first attempts gave rise to rectifying (non-ohmic) contacts which exhibited nonlinear current–voltage characteristics. Through a series of experiments it was found that the glow-discharge operation, which was utilized as a cleaning procedure prior to aluminum deposition, was a major reason for the non-ohmic behavior of the contact. Use of a field to ionize the aluminum and yield an average ion energy of 112 eV (200 eV maximum), in conjunction with the elimination of the glow-discharge operation, produced contacts which exhibited ohmic behavior. The same technique was then utilized on *p*-type wafers that are representative of the base region of an *n/p* solar cell, and ohmic contacts were also obtained. The contact resistances, especially in the latter case, appear to be extremely high, and will probably not result in high-efficiency solar cells.

3. Conclusions

Significant progress has been made in the use of aluminum as a contact material for silicon solar cells. It has been demonstrated that the high-vacuum sputtering process is capable of producing aluminum-contact cells with reasonable efficiencies. The series resistance of these cells was found to be of the order of 0.5 Ω for 2 × 2-cm cells, in comparison with state-of-the-art titanium-silver contact cells which exhibit series resistance of the order of 0.3 Ω for 2 × 2-cm cells.

At the present, problems seem to exist with the cold-substrate deposition process in achieving contact resistances low enough to yield high-efficiency solar cells, although it was possible to obtain contacts to the *n* and *p* layers which are ohmic in nature.

D. Capsule System Advanced Development: Power Subsystem, *R. G. Ivanoff and D. J. Hopper*

1. Introduction

The primary purpose of the Capsule System Advanced Development (CSAD) project is to obtain an improved understanding of planetary entry lander capsule system design and integration problems and to obtain experience in several critical and new technologies that relate to planetary capsule missions. To accomplish this objective, a specific Mars entry and hard-landing capsule system is being designed to obtain scientific information on the Martian atmosphere and surface conditions during capsule system entry and subsequent landing.

To support the CSAD activity, power subsystems capable of supplying electrical energy at discrete levels were developed for the entry and lander capsules. These subsystems were designed to survive the sterilization requirements of the capsule system, and for the lander capsule, impact survival was required.

Power subsystems to be incorporated into an entry and lander capsule, as part of the CSAD, have been fabricated, integrated into the capsule system, and are now undergoing system-level tests.

2. Power Subsystem Design

Each power subsystem consists of a sterilizable silver-zinc battery and a power control unit to provide switching, conditioning and distribution of several regulated voltages. The entry capsule power subsystem functional block diagram (Fig. 4) illustrates the power subsystem design and method of electrical power distribution.

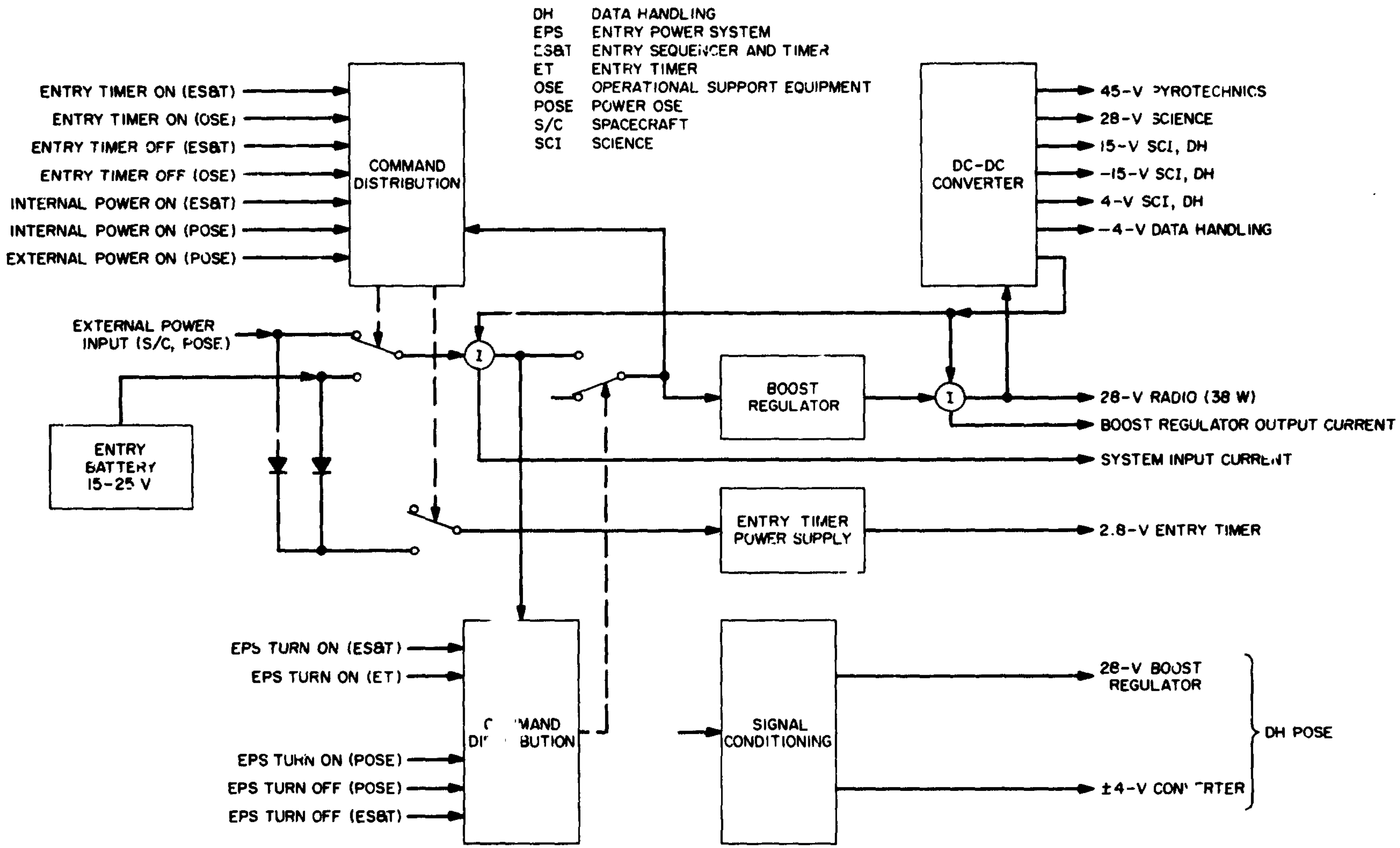


Fig. 4. Entry capsule power subsystem functional block diagram

Upon command from the entry sequencer and timer, the power subsystem battery is switched on-line providing power to the major capsule subsystems. The battery voltage of 15 to 25 V is boosted and regulated to 28 V by the boost regulator and used by the radio subsystem. This output is also distributed to a dc-dc converter. The converter, using voltage and current feedback, provides six regulated voltage outputs that are used by all subsystems except entry-capsule radio and entry timer. The entry timer is turned on by command well in advance of other subsystems and is, therefore, supplied power from a separate regulator to reduce losses inherent in the power conversion equipment. The regulator consists of a shunt zener control. The entry-capsule power control unit (Fig. 5) is capable of providing a maximum power of 50 W. Figure 6 illustrates how the power control unit and battery are combined prior to assembly in the entry capsule. The battery consists of 14 cells, each having a capacity of 5 A-h.

The lander-capsule power subsystem is similar in design to the entry-capsule power subsystem and performs identical functions. The major difference is the requirement for high-impact survival of the lander capsule and the ability to turn the lander capsule radio on and off independent of the other power loads.

3. Development Status

The procedure for the development of the power subsystems consists of design, fabrication, and testing of the individual units. The power subsystems are then integrated into the capsule system and functionally tested after being subjected to the selected environmental requirements.

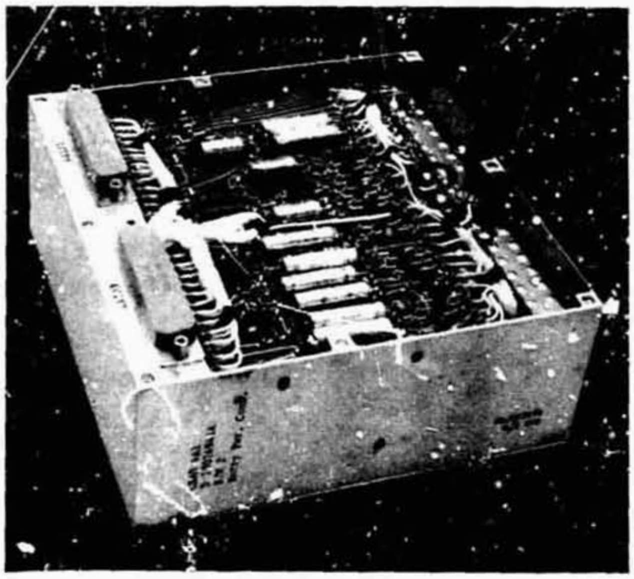


Fig. 5. Entry capsule power subsystem control unit

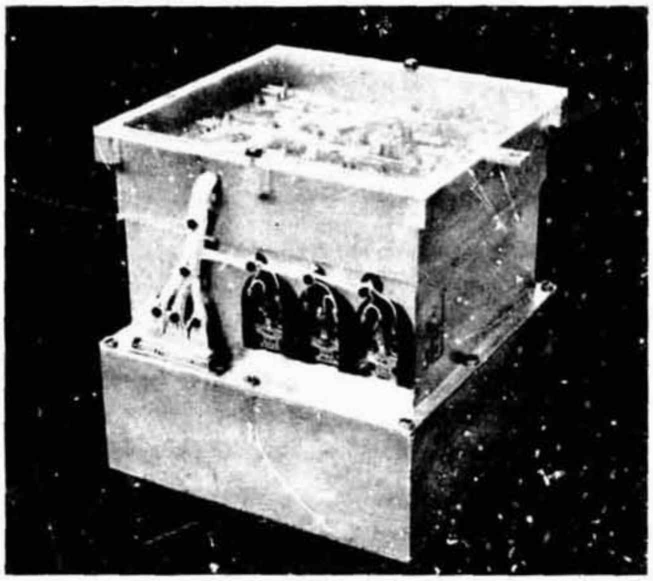


Fig. 6. Entry capsule power subsystem

Prototypes of the entry- and lander-capsule power subsystems have been tested at the subsystem level; in each case, results were within design limits. After integration of the power subsystem into the lander capsule, additional system-level tests were performed, including sterilization. The power subsystem was also integrated into the entry capsule. All system tests indicate nominal performance of the power subsystem.

The entry and lander capsules were combined and tested as a complete capsule system. Both entry- and lander-capsule power subsystems performed as expected.

The capsule system has undergone sterilization at 125°C. Post-sterilization tests using external power indicate no loss in performance in the entry- or the lander-capsule power control unit, which has now undergone two sterilization cycles and one impact test. Entry- and lander-capsule batteries are now being charged.

On May 4, 1968, the lander capsule was dropped from an altitude of 250 ft onto the dry lake bed at Goldstone, California. The capsule impacted after reaching a terminal velocity of 115 ft/s. The capsule then cycled through the nominal mission profile with no anomalies. Subsequent system tests indicated all subsystems were operational within design limits. At the conclusion of the drop test, the battery was monitored and found to have an open-circuit voltage of 17.5 V. This voltage indicates the battery was operating within design limits and had not been discharged more than expected. For the nominal mission profile, no more than 50% of the total battery capacity would be used.

A second drop test of the lander capsule was performed on May 28, 1968. The unit was dropped from an altitude of 250 ft onto a macadamized road to achieve a higher impact force than experienced on the Goldstone dry lake bed. The power subsystem performed all scheduled functions, with no apparent loss in capability.

E. Computer-Aided Circuit Analysis, *D. J. Hopper*

1. Introduction

The objective of this effort is to provide a generalized system of computer programs for analyzing electronic circuits. Computer programs are presently available to simulate circuits and to perform a steady-state, transient, or cyclic (AC) analysis on these simulated circuits depending upon which computer program is used. One of the advantages of being able to simulate a circuit is that an engineer can use components having "worst-case" values. Construction of an actual worst-case breadboard in the laboratory is a very difficult, if not an impossible, task. The major difficulty lies in obtaining components that have worst-case properties.

2. Simulation Problem

The computer simulation is accomplished by describing the circuit to the computer in an engineering-oriented computer language. Most of the programs can work with the simpler elements of a circuit, i.e., resistors, capacitors, inductors, mutual inductance, and ideal diodes. The rest of the circuit components must be described using these basic elements. This is where the difficulty lies. For example, one of the most common circuit elements, the transistor, has a small-signal model, a large-signal model, and a saturated model. The result obtained from the computer could be radically different from the expected result if the wrong model is used. The modeling of a transformer is another complex problem. In observing any magnetic induction versus field intensity (*B-H*) loop for magnetic materials, it is evident that *B* is a complex function of *H*, also the loop is dependent upon frequency.

3. Survey of Existing Computer Programs

A survey of the existing computer programs shows the number of programs available is extremely large, but most of the programs were written to solve specific problems instead of being general analysis programs. Several programs were studied, and it was found that a few programs could satisfy the total requirements.

With SCEPTRE, a program developed by IBM, one can perform both transient and steady-state analysis. Another feature of SCEPTRE is that it has a component-model library tape. Once a model has been derived, it can be stored on the library tape and used repeatedly. A copy of SCEPTRE was obtained on tape, and several sample circuits were analyzed. A few problems were uncovered, but the recent runs on SCEPTRE have been satisfactory.

ECAP is another useful program. With this program one can perform AC analysis. Unfortunately, it does not have the library tape feature possessed by SCEPTRE. This is an inconvenience, but the AC analysis feature is well worth the extra work involved.

4. Model Development

During this reporting period, several semiconductor models have been developed for transistors and diodes, but the major effort has been to develop a model of a transformer. Two methods of core modeling are currently under investigation. A piecewise linear model was obtained from IBM for use with SCEPTRE. It allows the entry of coordinate points corresponding to the magnetic induction versus field intensity (*B-H*) loop and provides a means of computing and storing values of *B* obtained as the transient solution proceeds. Operation within the *B-H* loop is simulated by taking the last value of *B* and a slope corresponding to that of the elastic region. Operation is otherwise constrained to points on the *B-H* loop. The other method being studied is the use of an exponential model. This model relates *B* and *H* with the use of exponential functions. The major difficulty with this model is that it is hard to simulate hysteresis.

F. Electric Propulsion Power Conditioning, *E. N. Costogue*

1. Introduction

The electric propulsion power conditioning project has two principal tasks. The first task, which is scheduled for completion in the early part of 1969, is to test a power conditioning unit with two ion engines. A switching module will be utilized to switch power to the engines by command. The second task, which is scheduled for completion in 1970, is to design, fabricate and test the complete power conditioning portion of an electric propulsion system.

The power conditioning unit will consist of (1) four or five units that will power five ion engines, (2) a switching unit that will switch power conditioning units to available ion engines as required, and (3) a maximum power point seeker unit that will examine the solar panel characteristics and verify the available maximum power of the source. Item (1), the power conditioning units, will be developed under contract. Items (2) and (3) will be developed at JPL.

2. Task 1 Power Conditioning Unit

Power conditioning hardware built for the SERT II program will be modified for the first task. The modified units are scheduled to be received from the contractor by November 1968. The power switching unit for the first task, which will switch power from one ion engine to the other by command, has been designed, and fabrication of the unit will be completed by August 1968.

The block diagram of the power switching unit is shown in Fig. 7. The major blocks of the unit are (1) the switch-position sense-logic circuit, (2) the switch driver, or stepper, and (3) the switch. The switch-position sense-logic circuit accepts the command for switching to the position requested and compares the position requested to the present position of the switch. When the signal received is satisfactory, indicating that the switch can move to the next position, the sense logic issues a signal drive to the switch driver. After the switch has moved to the position requested, a signal is generated to indicate the completion of the switching.

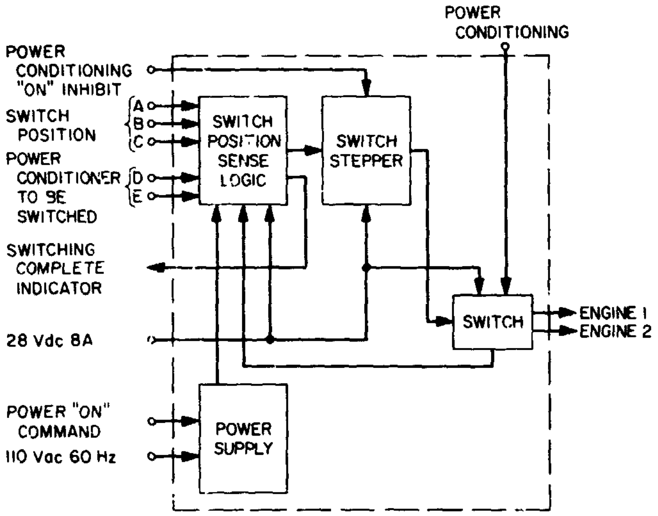


Fig. 7. Power switching unit block diagram

The switch driver (stepper) receives (1) the input for the switch-position sense-logic circuit, (2) verification that the power conditioner is functioning, and (3) verification that the switch is ready to switch. When all the signals received are satisfactory, the driver circuit generates the drive to move the switch. The switch is a heavy-duty, multiple-deck unit with high breakdown voltage.

3. Task 2 Design Studies

A study has been initiated to evaluate the merits of switching ion engines to power conditioning units versus providing a power conditioning unit per engine. The study will evaluate the reliability, weight, and cost of one system over the other.

Another study has been initiated to establish an efficient and acceptable method of determining the maximum power point of the solar panel source. The study will recommend (1) a design that will ensure safe operation of the engines throughout the mission, and (2) a means of identifying the available maximum power output of the panel.

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

In Phase I, General Electric Missile and Space Division and TRW Systems were selected to investigate and analyze various baseline power system configurations. In Phase II (FY 1969), JPL will select the best power system design and award a contract for the detail design and construction of a power system feasibility model.

2. General Electric Missile and Space Division

A contract[1] was awarded to the General Electric Missile and Space Division on January 28, 1968, for the Mars spacecraft power system development program. A detailed analysis of the load power profile and its effect

[1] JPL Contract 952150.

on power system sizing was performed; a partial-shunt regulation system was selected for analysis. The solar array/partial-shunt system integration investigation and the battery/battery charger interface study are continuing.

A distribution frequency optimization study indicated that a change from the presently used frequency of 2.4 kHz would not provide sufficient weight savings to warrant its consideration.

Reliability sensitivity studies indicated that fault-sensing and switchover to redundant devices should be considered only if their net reliability is equal to or greater than the reliability of the functions being protected. No distinct reliability advantage was determined for fault-sensing the regulator and inverter separately or as a pair. Fault criteria were identified for the principal power-conditioning units.

Power system reliability modeling was performed on the *Mariner* Mars 1969 system. A similar model is being programmed for the shunt system and a reliability comparison will be completed.

3. TRW Systems

TRW Systems began their investigation and analysis effort[3] for the Mars spacecraft power system development program on March 4, 1968. Mission and spacecraft requirements were reviewed and load power profile and power distribution/control requirements defined.

Five power system configurations were selected and are to be subjected to further detailed analysis to determine the optimum system. For these selections, a computer program was used to examine 70 baseline configurations. The selection criteria included weight and reliability assessments, maximization of solar array power margin, and minimum bus voltage excursion.

[3] JPL Contract 952151.

H. Planetary Solar Array Development,

W. A. Hasbach

1. Introduction

A report of the objectives and environmental design considerations were reported in SPS 37-49, Vol. III, pp. 112–114. Efforts to date have been in the conceptual design and analysis of three feasibility models capable of producing not less than 200 W of electrical power on the Martian surface (Refs. 1–3). Trade-off studies of weight and structural integrity versus exposure to the Martian environment have been conducted. Selection of materials, mechanisms, and solar cell panel configurations has confirmed three approaches that have the potential of meeting the goals of the program. The characteristics of each solar array system are summarized in Table 1.

Table 1. Characteristics of the solar array systems

Parameter	Solar array		
	Conical nontracking	Two-panel-oriented	Single-panel-oriented
Type of cell	0.010 in. thick, 2 × 2 cm	0.010 in. thick, 2 × 2 cm	0.010 in. thick, 2 × 2 cm
Cell output, at 485 mV, mW	58	58	58
Total number of cells	30,780	14,400	—
Number of cells/array	—	—	10,800
Number of deployment mechanisms	2	—	—
Number of orientation drives	—	3	2
Number of antenna orientation drives	—	—	2
Deployment mode	Torsion springs and snubbers	—	—
Orientation mode	—	Continuous motor gear drive with solar cell sun sensors	Continuous motor gear drive with solar cell sun sensors
Weight summary, lb			
Mechanical structure	33.07	19.23	20.94
Electrical components	23.37	10.95	8.24
Total	56.44	30.18	29.18

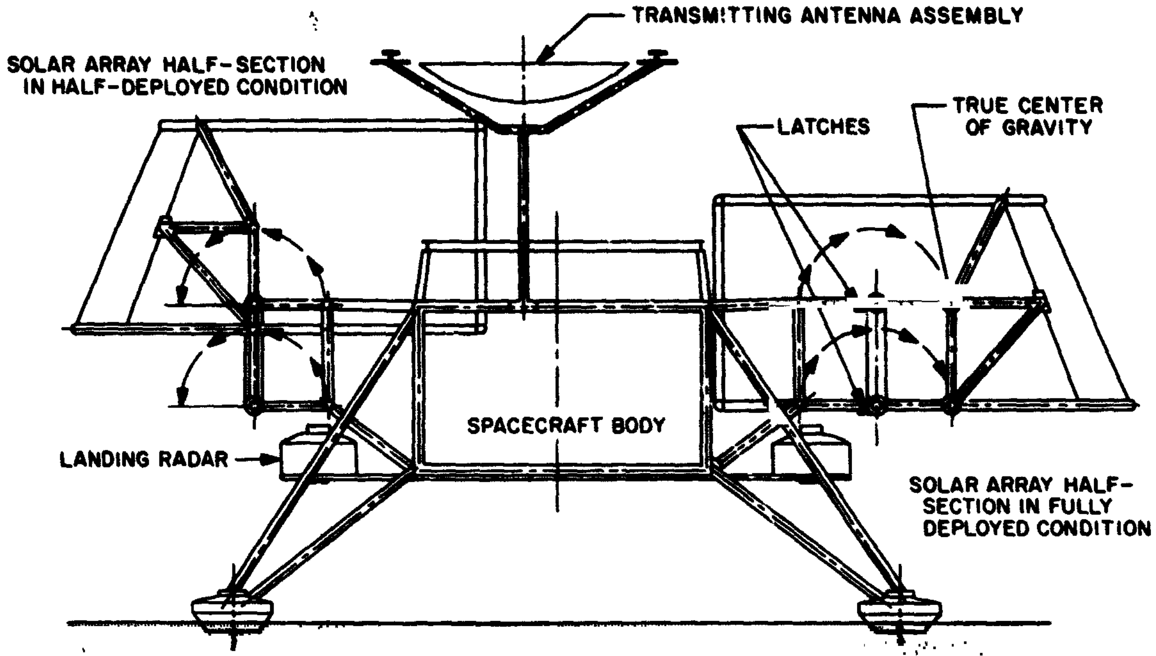

TRANSMITTING ANTENNA ASSEMBLY
SOLAR ARRAY HALF-SECTION IN HALF-DEPLOYED CONDITION
LATCHES
TRUE CENTER OF GRAVITY
SPACECRAFT BODY
LANDING RADAR
SOLAR ARRAY HALF-SECTION IN FULLY DEPLOYED CONDITION

Fig. 8. Conical nontracking array

2. Nontracking, Deployable, Conical Solar Array

The objective of the conical nontracking array (Fig. 8) is to produce power with a minimum of deployment mechanisms. The goal is to avoid complex mechanisms for latitude, slope, and position corrections and eliminate the need for a continuous tracking capability. This system, once released from its locked, launched, and flight positions, will not require power from the lander system for deployment or continuous operation for the mission life of 1 yr.

As recognized initially in its conception, this array will not meet the desired goal of 20 W/lb (1 AU) and under worst-case conditions will be under the minimum power requirements of 200 W of electrical power at solar noon. In the majority of the cases, the power output exceeds the minimum requirement of 200 W. The minimum power output at the worst-case condition of 46 mW/cm^2 (summer) is 5% low or 190 W, while the best-case condition is 35% high or 256 W. The average noon power output of the limiting conditions is 17% high (223 W). At higher solar intensities occurring in 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 specific power output is based on the equivalent power at 1 AU. Taking the power output at the worst-case condition of 46 mW/cm^2 and converting to 1 AU by the ratio of 46/140 mW/cm^2 = 0.328,

256 W/0.328 = 780 W
190 W/0.328 = 580 W

Thus, the specific power would lie between the range of

780 W/56.44 lb = 13.8 W/lb
580 W/56.44 lb = 10.3 W/lb

3. Two Solar Panels Having Sun Tracking Capabilities

The objective of the two-panel-oriented solar array (Fig. 9) is to provide a three-axis tracking capability. In this design, the solar panels are mounted on opposite sides of the spacecraft so that the other two spacecraft sides are always unobstructed, and there is no interference with the vehicle antenna system. The design is a trade-off against the antenna shadow problem in which the total array was sized at 10 circuits over the minimum of 30 circuits required.

This system will meet the desired goal of 20 W/lb at 1 AU and exceed the minimum power of 200 W at solar noon for worst-case conditions. The power output will vary, depending on the number of circuits that may possibly be shadowed at noon. For the lowest solar

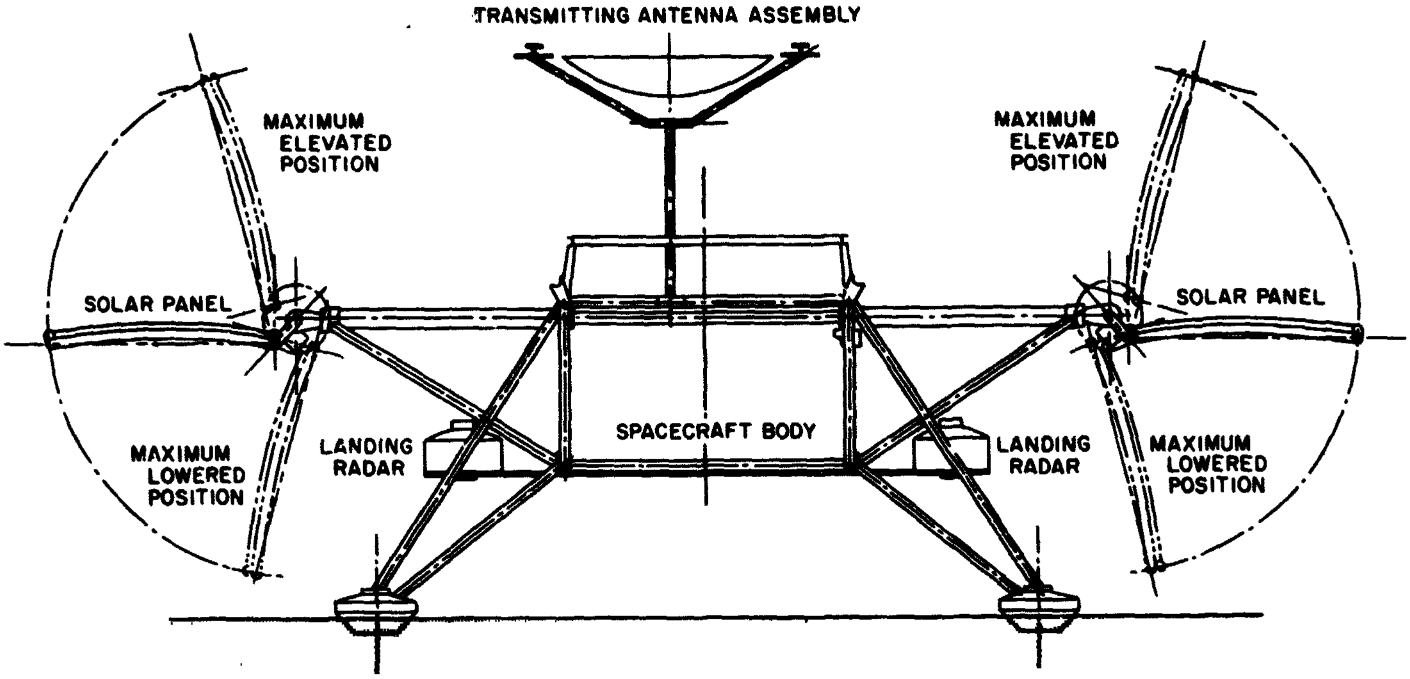


Fig. 9. Two solar panels having sun tracking capabilities

intensity (46 mW/cm^2) occurring at the first day of Martian summer, the power output limits are:

Maximum shadow (30 circuits) = 205.6 W
No shadow (40 circuits) = 274.2 W

For the higher solar intensities occurring in the spring and fall seasons, the power levels range from 234.4 to 312.5 W, considerably over the minimum requirement.

The power-to-weight ratio, based on 1 AU, varies with the power output of the panels as a function of shadowing. Taking the lowest output condition of noon at the summer solstice with a solar intensity on the Martian surface of 46 mW/cm^2 and converting to 1 AU by the ratio of 46/140 mW/cm^2 = 0.328,

205.6 W/0.328 = 626 W
274.2 W/0.328 = 837 W

Thus, the specific power at 1 AU would be

626 W/30.185 lb = 20.8 W/lb
837 W/30.185 lb = 28.6 W/lb

4. Solar Panel and Integrated Antenna System

The objective of the single-panel-oriented solar array (Fig. 10) is to provide a three-axis tracking capability. The deployment of the solar panel and antenna on a vertical boom eliminates the possibility of shadowing from the spacecraft body and antenna. This allows the minimum number of circuits (30) to be used to achieve the required power output of 200 W under worst-case conditions.

Combining the antenna and solar array mounting presents a problem in maintaining the point accuracy of the antenna when the system is buffeted by wind gusts. The vertical boom has been sized to minimize the deflection due to wind loads; however, other factors are present. The drive mechanisms will have to be designed to eliminate, as much as possible, any backlash in the gearing, and the latching mechanism of the vertical boom will have to be of a self-tightening design. Other factors, such as the stability of the spacecraft body and legs and the soil condition of the vehicle landing area, will affect the antenna point accuracy, but these are unanswerable at the present time.

The single-panel-oriented array of 30 circuits will meet the desired goal of 20 W/lb at 1 AU, and exceed the minimum power requirement of 200 W at solar noon for worst-case seasonal conditions. The power outputs for the limiting seasonal conditions are:

Summer noon = 205.6 W
Spring/fall noon = 234.4 W

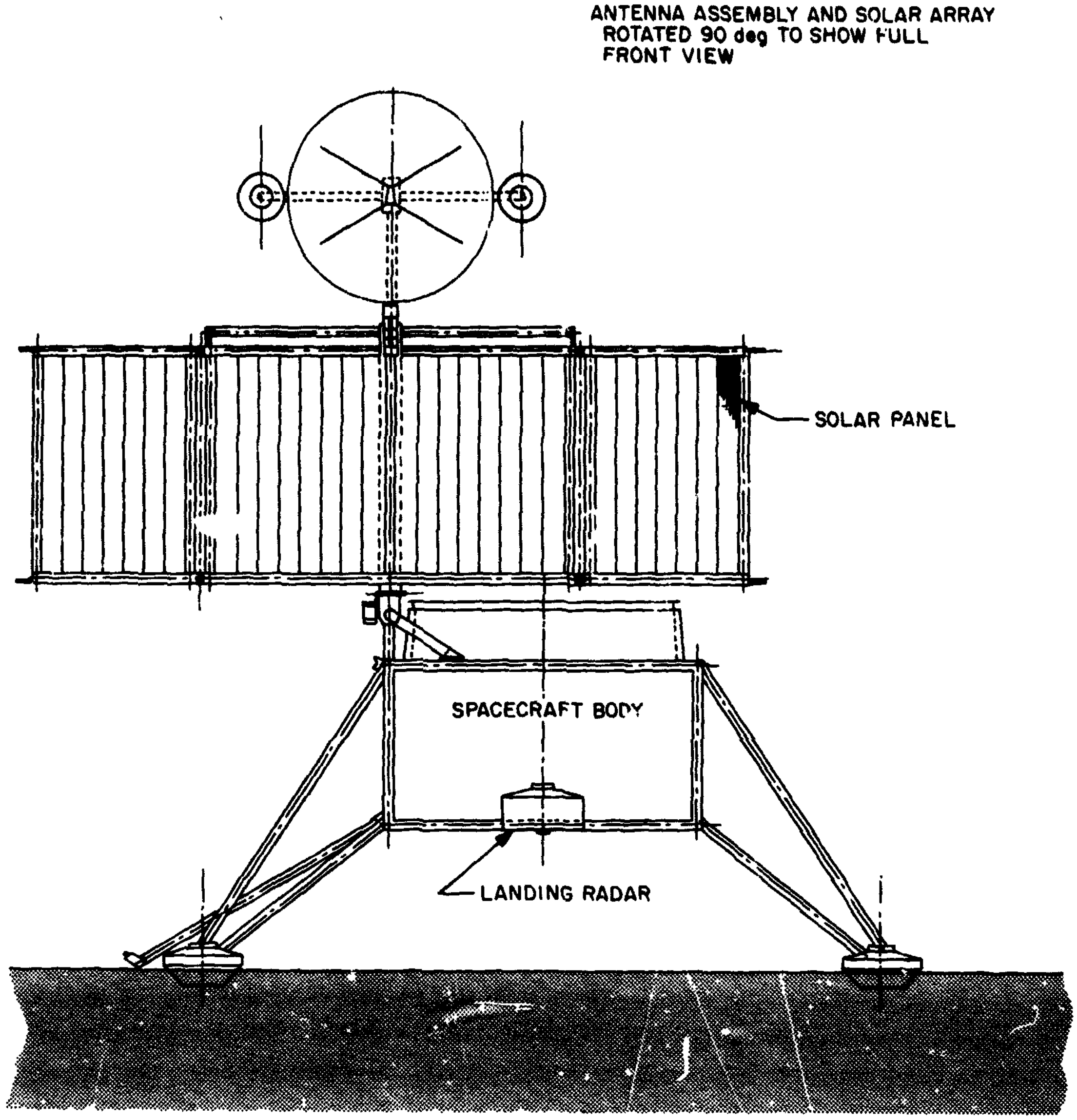

ANTENNA ASSEMBLY AND SOLAR ARRAY ROTATED 90 deg TO SHOW FULL FRONT VIEW
SOLAR PANEL
SPACECRAFT BODY
LANDING RADAR

Fig. 10. Solar panel and integrated antenna system

The power to weight ratio, based on 1 AU, for the output condition at noon at the summer solstice with a solar intensity on the Martian surface of 46 mW/cm^2 is 22.2 W/lb.

5. Solar Cell Covering

A solar cell power supply operating on the Martian surface has a problem that is unique and not found in space applications. Mars has a dust condition that is considered severe; the dust is assumed to be iron oxide and electrically conductive. The electrical shorting caused by the dust would be catastrophic to the solar array. One solution that was considered was to coat all electrically exposed areas of the solar cell circuit with filter adhesive. However, this method is tedious, virtually impossible to guarantee complete protection, and will add considerable weight to the solar array.

A second consideration was to eliminate the cover glass and use a semiorganic resin.[3] This coating has been

[3] Developed by B. Marks, Lockheed Missiles and Space Co., Palo Alto, Calif.

developed specifically as a solar cell coating and can be applied by spraying. Principal drawback to this coating technique is the inability to completely insulate all electrically conducting surfaces. Spray application of the coating would not insulate the back side of the connector tab, and dipping the total array is highly impractical.

The selected method for insulating the solar array is by encapsulating the cells with a continuous sheet of Tedlar film.[4] Tedlar film is essentially transparent to, and unaffected by, solar radiation in the near ultraviolet, visible, and near infrared regions of the spectrum.

The advantages of the film coating for the solar cell circuit are:

(1) Total insulation is obtained of all electrical conducting surfaces.

(2) Electrical loss is low due to coating.

(3) Installation is easily accomplished using a space-proven adhesive system.

(4) Finished coating eliminates gaps between cells, which would form a trap for dust accumulation.

(5) The coating protects the cells from low-energy proton radiation, as the cells will have no exposed areas common to typically filtered cells.

(6) The film has little weight.

(7) The flexible film, bonded with a resilient adhesive, should have better abrasion resistance to the "sandblasting effect" of the dust than the hard surface of glass or quartz filters.

References

1. Quarterly Report 7254-Q-1, Electro-Optical Systems, Inc., Pasadena, Calif., Oct. 13, 1967.

2. Quarterly Report 7254-Q-2, Electro-Optical Systems, Inc., Pasadena, Calif., Jan. 6, 1968.

3. Quarterly Report 7254-Q-3, Vols. I, II, and III, Electro-Optical Systems, Inc., Pasadena, Calif., Apr. 15, 1968.

I. Thermionic Research and Development,

O. S. Merrill

1. Introduction

A program to improve the output performance of cesium-vapor thermionic converters has been in progress for several years. The work reported in this article is a continuation of this program and was performed (Ref. 1) under contract to NASA, but with JPL technical direction, by Electro-Optical Systems, Inc. The first of three tasks of this effort was reported in SPS 37-50, Vol. III, pp. 82–92. The second and third tasks are reported here.

2. Variable-Spacing Test Vehicles

a. Design. Two variable-spacing test vehicles of the same basic design but incorporating different sets of electrode materials were fabricated. The test vehicles, drive mechanism, and supporting structure are of the same design as reported in SPS 37-39, Vol. IV, pp. 15–19. The first test vehicle had a polycrystalline rhenium emitter and a polycrystalline molybdenum collector; the second had a vapor-deposited rhenium emitter and collector.

b. Test results. Data typical of that taken in this project is shown in Fig. 11, where the voltage output at a constant current of 38 A is shown versus interelectrode spacing for both the rhenium–molybdenum and a previously tested polycrystalline rhenium–rhenium test vehicle

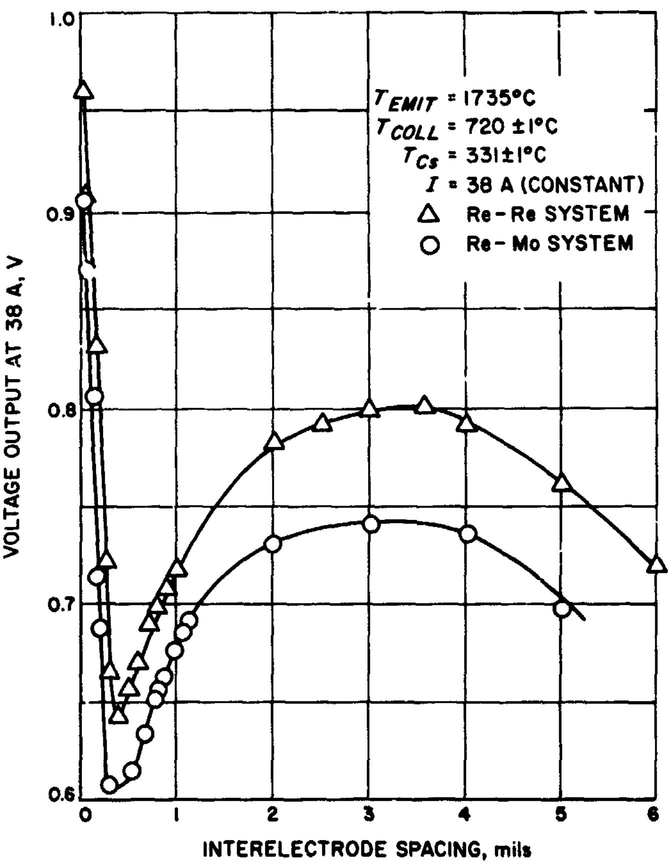


Fig. 11. Comparison of performance of rhenium–rhenium and rhenium–molybdenum variable-spacing test vehicles

[4] Polyvinylfluoride film manufactured by Film Dept., E. I. du Pont de Nemours and Co., Vernon, Calif.

(SPS 37-39, Vol. IV). The quantitative difference between these electrode systems is of the order of 60 mV at the optimum spacing of 3 to 4 mils. This figure demonstrates the central thesis of this program, namely, that increased thermionic converter performance results when a higher bare work function (lower cesiated work function) collector is used. The bare work function of vapor-deposited rhenium at 1735°C (2008°K) is 5.08 eV, as reported in SPS 37-50, Vol. III, while that of polycrystalline molybdenum is 4.2 eV.

Extensive data were taken with both test vehicles. When operated at high temperatures, the performance of the rhenium–molybdenum test vehicle approached and eventually reached that of the rhenium–rhenium test vehicle and remained comparable thereafter at both high and low temperatures. This was attributed to the deposition of rhenium onto the molybdenum collector, thus essentially changing this vehicle to a rhenium–rhenium test vehicle; however, subsequent long term, low emitter temperature operation of the device resulted in a return to the rhenium–molybdenum performance. It is postulated that the deposited rhenium diffused into the molybdenum collector substrate. although the device has not been disassembled and metallurgical tests performed to determine if this is indeed the explanation. To change the rhenium–molybdenum vehicle into a rhenium–rhenium vehicle, and to maintain it as such over desirable test periods, the rhenium emitter was periodically operated at temperatures as high as 2200°C for 5 h (with cesium reservoir heater turned off) to ensure a sufficient and stable rhenium coverage on the molybdenum collector (believed to be at least two or three monolayers thick). Data from this vehicle match those from the rhenium–rhenium test vehicle to within 2%.

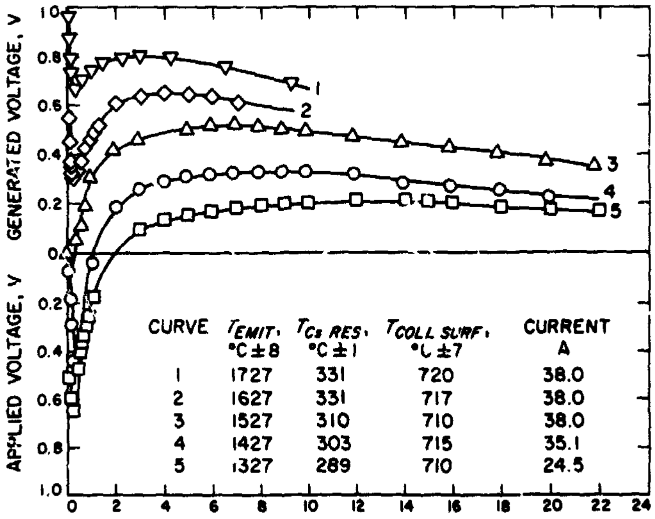


Fig. 12. Near-optimized voltage output vs interelectrode-spacing for a rhenium–rhenium electrode system at various constant operating conditions

Some of the most significant and useful data obtained from the test vehicles are shown in Fig. 12, showing, for the rhenium–rhenium electrode system, voltage output versus interelectrode spacing for a constant load current, constant emitter temperature, constant collector temperature, and constant cesium vapor pressure (as indicated by constant cesium reservoir temperature). When the current and temperatures are held constant to within experimental error and when the spacing between electrodes is precisely determined, the product of the cesium pressure *p* and the interelectrode-spacing *d* (i.e., *pd*) at the point of optimum voltage (and power) output is observed to be constant at a value of 16.0 ±0.8 mil-torr (Table 2). It can be noted from the curves and the tabulated data that the *pd* product is independent of emitter temperature. The optimum *pd* product also appears to be independent of the collector and emitter materials. It is further observed that at the lower emitter temperatures the optimum voltage for a given current is lower, and the interelectrode spacing is considerably larger and less critical; i.e., the optimum voltage is less sensitve to variations in the spacing.

The performance testing of the vapor-deposited rhenium–rhenium test vehicle was not successful due to a leak in one of the electron-beam welded flange joints. This was discovered after about 120 h of operation.

Table 2. Summary of pressure–distance data taken from interelectrode spacing versus voltage output curves

Emitter temperature, °C	Cesium reservoir temperature, °C	Cesium pressure *p*, torr	Inter-electrode–spacing *d*, mil	Product *pd*, mil-torr
1327	289	1.33	12.5	16.6
1427	291	1.43	11.0	15.7
1427	303	1.96	8.0	15.7
1527	310	2.35	7.1	16.8
1527	320	3.01	5.3	15.9
1527	331	4.02	3.9	15.7
1627	331	4.02	3.9	15.7
1735	331	4.02	3.9	15.7
1735	344	5.30	3.0	15.9
1735	350	6.06	2.7	16.3

By comparing data from this vehicle with those from the previous ‘est vehicles, its performance was observed to be inconsistent and considerably lower. A second set of parts has since been assembled into another test vehicle[5] and early tests on the second unit show it to be performing satisfactorily. Results of the tests will be reported in a future article of SPS, Vol. III.

3. Fixed-Spacing Vapor-Deposited Rhenium Converters

a. Design. Two thermionic converters of planar geometry employing vapor-deposited rhenium electrodes were fabricated. The two converters, designated SN-109 and SN-110, are identical to the SN-101 series converters (SPS 37-39, Vol. IV). The design criteria for these converters were to have bee.· based on the vapor-deposited rhenium–rhenium test vehicle data. Since those data were lacking, the design criteria were chosen based on data from the polycrystalline rhenium–rhenium test vehicle.

b. Test results. The test data from the polycrystalline rhenium–rhenium test vehicle and the data for the corresponding fixed-spacing converters SN-109 and SN-110 are compared in Table 3 and in the design temperature curves of Fig. 13. The agreement is very close and suggests that the electrode systems of vapor-deposited and polycrystalline rhenium yield nearly equivalent thermionic performance.

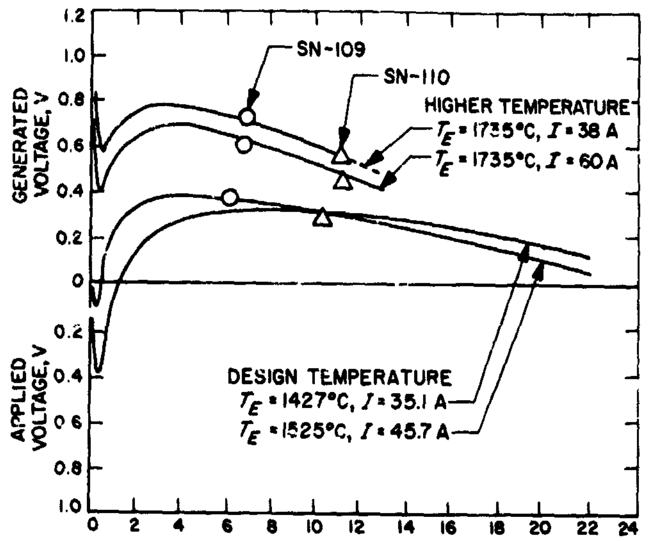


Fig. 13. Comparison of performance of fixed-spacing, vapor-deposited rhenium converters SN-109 and SN-110 with polycrystalline rhenium–rhenium variable-spacing test vehicle

Converters SN-109 and SN-110 were also tested at a higher emitter temperature for additional comparison to polycrystalline rhenium–rhenium performance. Their performance at an emitter temperature of 1735°C and at constant currents of 38 and 60 A (where optimum voltages were 0.8 and 0.7 V, respectively) was also found to be in excellent agreement with the variable-spacing test vehicle data for the same conditions. Their pcrformance was considerably off optimum, however, inasmuch as the optimum interelectrc·le spacing at this temperature is approximately 3.5 mils. The higher temperature curves of Fig. 13 also show this performance compaıisor. The interelectrode spacing is a few tenths of a mil larger at the higher temperature due to increased expansion of the emitter support sleeve and other converter components.

Table 3. Comparison of performance of converters SN-109 and SN-110 with polycrystalline rhenium–rhenium variable-spacing test vehicle data (at converter design point)

P·rameter	Test vehicle d·ta at SN-109 design point	SN-109 data	Test vehicle data at SN-110 design point	SN-110 data
Interelectrode spacing, mils	6 $^{+1}_{-0}$	6.2	10 $^{+1}_{-0}$	10.5
Emitter temperature, °C	1525	1524	1427	1425
Cesium reservoir temperature, °C	320	321	303	302
Collector surface temperature, °C	722	720	715	7C6
Voltage output, V	0.4	0.4	0.3	0.3
·urrent, A	45.7	45.2	35.1	35.1

4. Analysis of Test Vehicle Data

The primary objective of this task was to formulate a theoretical description of thermionic converter performance and to correlate it with an analysis of the parametric vehicle data. The effort proceeded sequentially in three parts, each part covering one of the regions of parametric vehicle operation as defined by Fig. 14. Region I is the *electron space charge region* and extends from zero interelectrode spacing to the minimum voltage identified as the plasma onset point. Region II is the *transition region* and extends from the plasma onset point to the optimum output point. Region III is the *positive column region* and extends from the optimum output point to the right margin of the ı.gure and beyond.

[5] Under JPL Contract 952217.

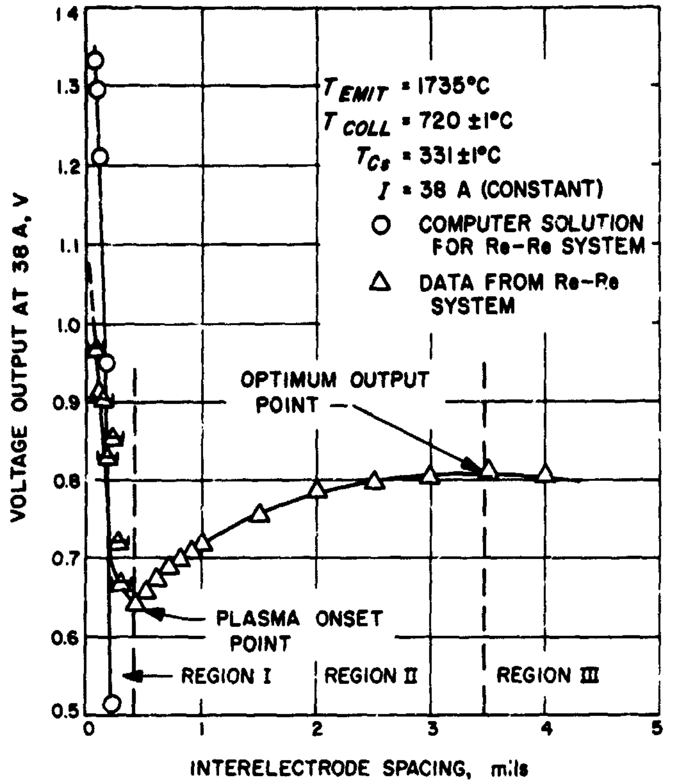


Fig. 14. Comparison of experimental results with computed results (region I only) for a typical voltage output vs interelectrode-spacing curve

This task was directed mainly toward the analysis of region I. The formulation of the problem and the analysis are given in Ref. 1, where the converter is viewed as a "double diode" described by Poisson's equation. A computer program was set up and solutions obtained. A comparison of the computer solution for region I and the test vehicle data is also shown in Fig. 14. The discrepancy in output voltage at low spacings is related to losses in lead resistance between the electrode surfaces and the point at which the potential was measured.

Reference

1. Campbell, A. E., and Jacobson, D. L., *Final Report, Thermionic Research and Development Program,* NASA Contract NAS 7-514, EOS Report 7118-Final. Electro-Optical Systems, Inc., Pasadena, Calif., Mar. 1, 1968.

J. Thermionic Converter Development, *P. Rouklove*

1. Introduction

The development of advanced technology thermionic converters is continuing at JPL. The series 9 planar converters, built by Thermo Electron Co., are still being used as test vehicles for technical improvements. The development of this type of converter was discussed previously in SPS 37-48, Vol. III, pp. 58–60.

2. Converter Designs and Test Results

Measurements performed on converter T-206 pointed out that any further improvement in power output was limited by the radiator geometry. This geometry, which was derived from the necessity to incorporate the converter into a 16-converter solar-heated generator, limited the cross-sectional area available for the collector–radiator heat flow and resulted in excessively high collector temperatures. The advantages of the application of the heat pipe as a collector heat rejection medium were presented in SPS 37-48, Vol. III, pp. 60–63.

Converter T-208 was assembled incorporating a niobium heat pipe as a collector–radiator structure (Fig. 15). The converter was assembled using a rhenium emitter and a rhenium collector, the latter consisting of a sheet of rhenium vanadium-brazed to the niobium pipe. Prior to the assembly, the emitter surface was electroetched and thermally stabilized in vacuum at approximately 2050°C for 2 h.

During the tests, it was observed that the performance of converter T-208 was inferior not only to that of thermal model T-3, which had a collector heat-pipe assembly, but also to that of T-206, which used a finned-type radiator. Both converters T-206 and T-208 used rhenium electrodes. The difference in the collector area (2.52 cm² for T-206 versus 2.34 cm² for T-208) did not account for the performance reduction. However, the collector surface in converter T-208 was further reduced to a net electrode area of 2.16 cm² by a groove cut in the collector for cesium vapor distribution and outgassing. The net ratio of collector areas of these converters was 0.86, or a 14% smaller area for T-208. Figure 16 indicates by dashed line the performance of converter T-206 reduced by 14% for comparison purposes.

Examination of the results implied that the interelectrode spacing in the two converters was different, being larger in the case of converter T-208. Comparison of the cesium conduction was made from test data and the following empirical formula was used to calculate the interelectrode spacing:

$$d = \left[\frac{0.0001475A(T_e - T_c)(T_e + T_c)}{\Delta Q/\Delta p}\right]^{0.5} - \frac{0.006(T_e + T_c)}{p}$$

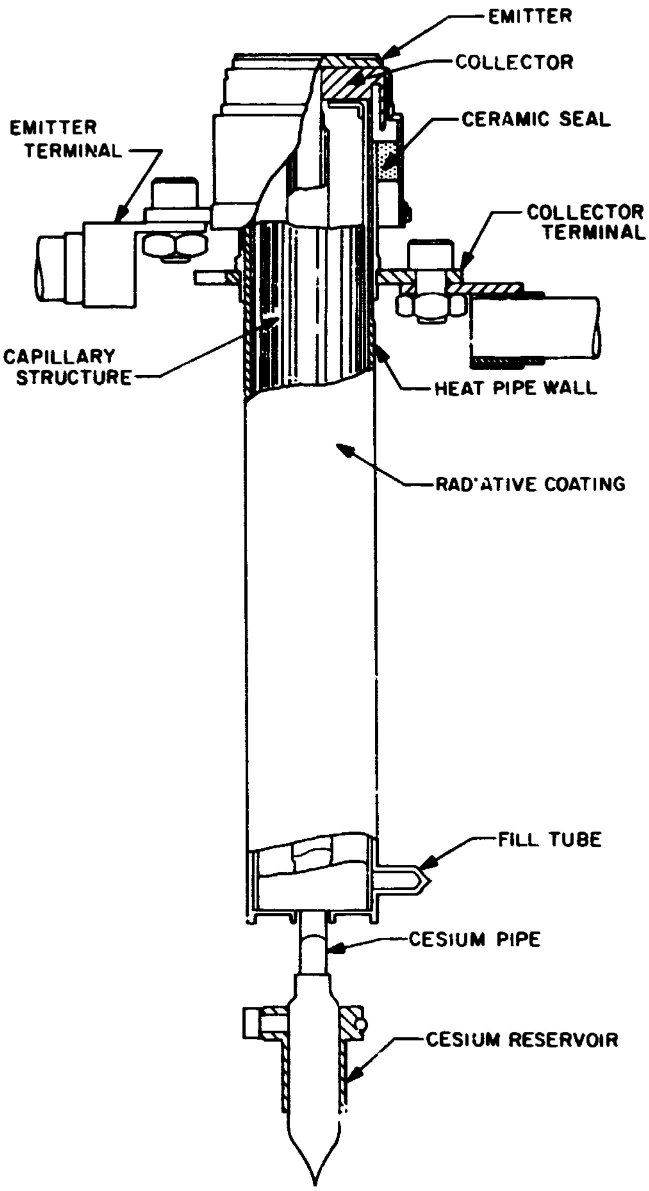


Fig. 15. Converter with collector heat pipe

where

d = interelectrode spacing, mils

A = interelectrode area, cm^2

T_e = emitter temperature, °K

T_c = collector temperature, °K

$\Delta Q/\Delta p$ = slope of cesium conduction curve

p = pressure, torr

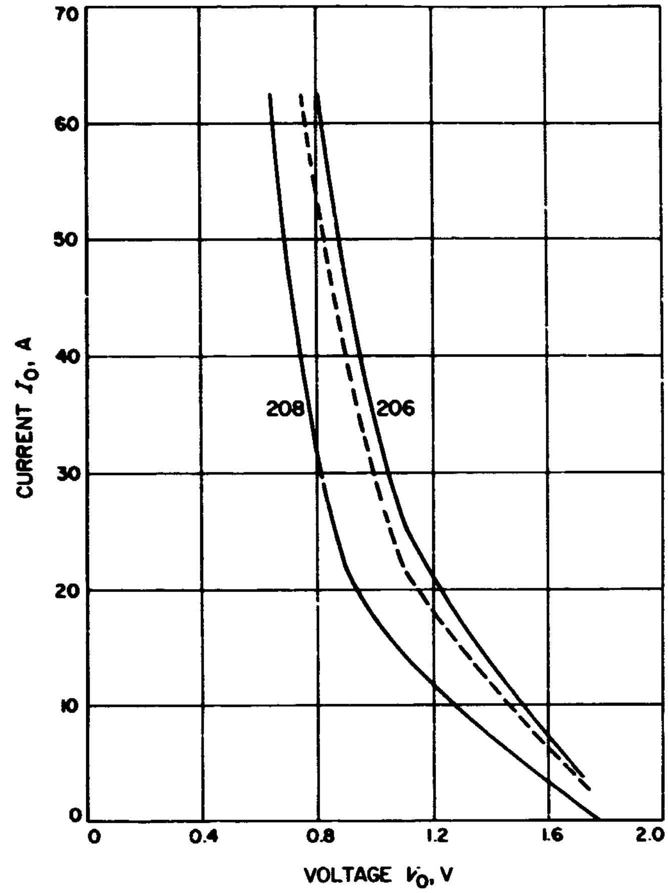


Fig. 16. Converter performance comparison

The results of the calculations are presented ir Table 4 for various cesium pressures. These data indicate a 65% difference in the interelectrode spacing between converters T-208 and T-206; the actual magnitude of the spacing should be larger because only the interelectrode areas were considered in the calculations, disregarding tne side effects.

Table 4. Comparison of converters T-206 and T-208 at various cesium pressures

Parameter	Converter T-206 at indicated pressure		Converter T-208 at indicated pressure	
	8 torr	12 torr	8 torr	12 torr
A, cm^2	2.52	2.52	2.16	2.16
Te, °K	1990	1990	2000	2000
Tc, °K	861	875	880	885
$\Omega Q/\Omega p$, W/torr	1.60	1.05	0.90	0.50
d, mils	1.29	1.38	2.06	2.33
d, mils (average)	1.33		2.20	

The lower performance of converter T-208 was also tentatively related to an overheating of the collector surface. Although no direct measure of the collector surface temperatures could be obtained, the inability to reproduce the dynamic curves in steady state pointed to an overheated collector. It was tentatively attributed to an excessive restriction in the vapor channels in the heat pipe at the heat receiving end near the collector. This could lead to an excessive temperature drop at the liquid–vapor interface and was estimated to be 60 to 80°C. Corrective measures have been taken in the assembly of converter T-209.

Converter T-207 was assembled using a rhenium emitter and a palladium-clad molybdenum collector. The configuration of converter T-207 was identical to that of converter T-206. This duplication was done to facilitate the comparison of experimental data and evaluate the performance of the palladium as a collector material. Tests were performed at emitter temperatures of 1800, 1900, and 2000°C. The converter configuration and the use of a finned radiator again did not allow proper collector cooling. The cesium conduction data corresponded to an interelectrode spacing of 2.54 mils. Some uncertainty exists as to the exact comparison between emitter surface temperatures of the two converters, due to a possible influence of the electron bombardment filament shape and to variations in the location of the hohlraum. An approximate 13% difference in the required power input, for otherwise similar test conditions, was observed between the two converters. This would correspond to a possible difference in emitter surface temperature of approximately 80°C.

The analysis of the test data indicated that the apparent cesiated work function of the palladium used as collector material in converter T-207 was higher by 0.037 eV than that of the rhenium utilized in converter T-206. This corresponds to a reduction in output current of between 3.9 and 6.1 A or a voltage shift between 0.030 and 0.044 V, with the current–voltage characteristics of converter T-207 to a lower output voltage (Fig. 17).

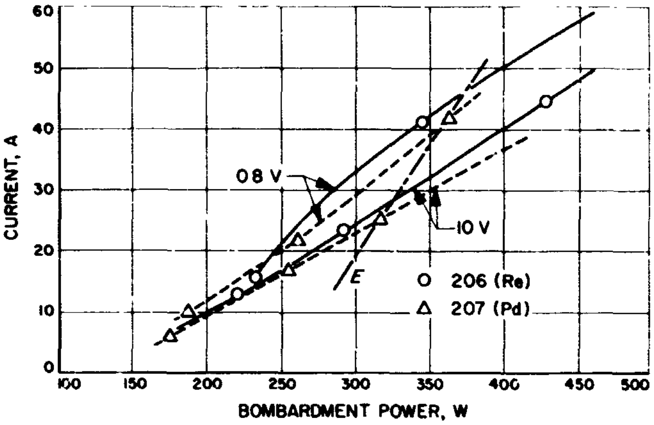


Fig 17. Converter work function comparison

3. Generator Design

Because of the necessity of using converters with heat pipe collector–radiators, the original design of the multiconverter generator had to be modified. The new assumptions for the generator design are a converter output of 70 A at 0.80 V and 28 A at 1.0 V, corresponding to a maximum power point power density of 20 W/cm^2. Two types of heating systems were considered for terrestrial tests: solar, using an 11.5-ft-diam mirror capable of a 6770-W thermal input into a 1.60-in. cavity aperture, and electron-bombardment heating. A study was made for comparison purposes using a 9.5-ft-diam mirror for cislunar application.

Parametric studies of optimum generator performance with varying converter complements as a function of

Table 5. Predicted generator performance data using solar and electro-bombardment heating systems

Parameter	Solar heating system		Electron-bombardment heating system	
	Ground (11.5-ft diam mirror)	Cislunar (9.5-ft diam mirror)	Case 1	Case 2
Cavity aperture diameter, in.	1.61	1.33	1.33	1.33
Cavity input, W	6770	7800	9070	8680
Available converter input, W	255	315	394	380
Converter output current, A	31.0	54.5	87.0	81.2
Converter output voltage, V	0.90	0.87	0.70	0.74
Converter output power, W	30 3	47.5	61.0	60.0
Generator output power, W	485	760	975	960
Absorber–generator efficiency, %	7.2	9.7	—	—
Generator efficiency, %	—	—	10.7	10.8

cavity aperture diameter indicated the desirability of selecting a 16-converter configuration composed of axial rows of 8 radially mounted converters. Calculations of the individual converter total power input requirements, including emitter support conduction, electrical output, cesium co..duction, interelectrode re-radiation losses, etc., were performed. These calculations lead to the predicted generator performance shown in Table 5.

In the case of the ground tests using the 11.5-ft-diam mirror, the following assumptions were made: incident flux 90 W/ft^2, reflectivity 88%, shadow factor loss due to the vacuum housing of the generator, 5%, and window loss 11%. For the 9.5-ft-diam mirror for cislunar application, an incident flux of 130 W/ft^2 was assumed, with a mirror reflectivity of 89% and a generator support shadow fa.tor loss of 2%.

V. Guidance and Control Analysis and Integration

GUIDANCE AND CONTROL DIVISION

A. Automation of Variational Techniques for the Solution of Optimum Control Problems,

H. Mack, Jr.

1. Introduction

Computer programs have been written to completely automate the solution of optimal control problems where the computation scheme assumes small scale variations about a nominal solution. This automation of variational techniques will enable the user to solve small as well as large scale optimal control problems with a minimum amount of programming for each specific problem. Since all of the variational equations are derived by a computer and compiled by Fortran IV with no intervening human action, the most time-consuming part as well as the greatest source of errors in the solution of variational problems has been eliminated.

2. Description of DEVNEC and QUASI Programs

The most widely used variational techniques are automated by two separate programs. The first program is called DEVNEC and is written in the IBM FORMAC language, which is currently available on the IBM 7094 as an extension of the Fortran IV compiler. DEVNEC uses the system equations and the boundary conditions as inputs to derive all of the necessary conditions for an optimum solution by use of the *maximum principle*. The maximum principle is automated in DEVNEC because it is one of the best methods for obtaining the solution to two-boundary-value problems that result from the formulation of the optimal control problem. This method has a significant advantage over the classical calculus of variations method and the dynamic programming method in that the maximum principle can be applied to problems where the control is constrained. The second program is called QUASI and automates a generalized Newton–Raphson method for the solution of two-point boundary-value problems. QUASI uses the variational equation derived by DEVNEC to obtain a numerical solution to the optimum control problem.

The flow chart in Fig. 1 shows the functions of each program in obtaining a solution to the optimal control problem. The process starts with the input of the system equations

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, \mathbf{u}, t)$$

and the boundary conditions at the initial and final times (t_0 and t_f)

$$\mathbf{D}(\mathbf{x}, \mathbf{u}, t_0) = 0$$

$$\mathbf{G}(\mathbf{x}, \mathbf{u}, t_f) = 0$$

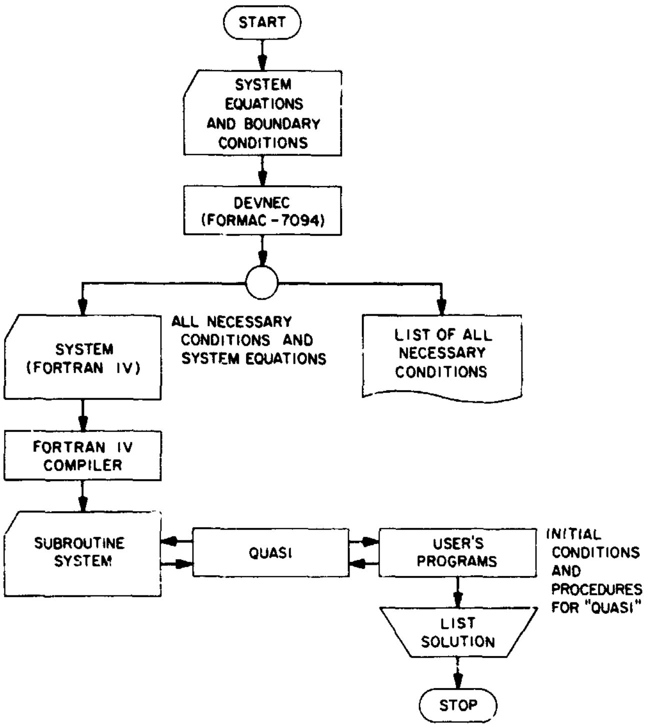

START
SYSTEM EQUATIONS AND BOUNDARY CONDITIONS
DEVNEC (FORMAC-7094)
SYSTEM (FORTRAN IV)
ALL NECESSARY CONDITIONS AND SYSTEM EQUATIONS
LIST OF ALL NECESSARY CONDITIONS
FORTRAN IV COMPILER
SUBROUTINE SYSTEM
QUASI
USER'S PROGRAMS
INITIAL CONDITIONS AND PROCEDURES FOR "QUASI"
LIST SOLUTION
STOP

Fig. 1. Computer program flow chart showing functions performed in solving optimal control problems

into the DEVNEC program. The numerical value for the dimensions of the state and control vectors ($\mathbf{x}$, $\mathbf{u}$) and the boundary conditions are also inputs to DEVNEC; these dimensions are as follows:

$$\dim(\mathbf{x}) = n$$

$$\dim(\mathbf{u}) = m$$

$$\dim(\mathbf{D}) \leq n$$

$$\dim(\mathbf{G}) \leq n$$

The quantity or performance index that is to be maximized or minimized is

$$J(\mathbf{u}) = <\mathbf{c}, \mathbf{x}(t_f)>$$

where $\mathbf{c}$ is a constant vector. The angular brackets $<>$ denote the inner-product operation, where

$$<\mathbf{c}, \mathbf{x}> = \sum_{i=1}^{n} c_i x_i$$

and c_i and x_i are elements of the $\mathbf{c}$ and $\mathbf{x}$ vectors, respectively.

Using these inputs and the mechanization of the maximum principle, DEVNEC computes the adjoint system of equations

$$\dot{\boldsymbol{\lambda}} = -\frac{\partial}{\partial \mathbf{x}} <\boldsymbol{\lambda}, \mathbf{F}(\mathbf{x}, \mathbf{u}, t)>$$

where $\boldsymbol{\lambda}$ is the adjoint vector, and then computes all partials of the state and adjoint equations, the Hamiltonian, and the boundary conditions. The state and adjoint equations and partials are output as equations in a subroutine that may be compiled directly by a Fortran IV compiler.

The DEVNEC output equations are compiled in a subroutine called SYSTEM. The SYSTEM subroutine is called by QUASI when the optimal control problem is solved numerically. QUASI is a mechanization of the quasilinearization or Newton–Raphson technique for solving two-point boundary-value problems. This technique solves the state and adjoint equations, which are usually nonlinear, by solving a sequence of linearized state and costate equations. The boundary, transversality, and optimality conditions are satisfied by constraining their variations to be zero. The quasilinearization process is initiated by a call statement in the user's program that contains an initial approximation to boundary conditions and some logic variables that specify optional procedures to be taken by QUASI.

3. Applications and Results

The programs DEVNEC and QUASI have been checked by applying them to the solution of several optimal control problems where the dimensions of the systems have varied from 4 to 10. The computer object run time for DEVNEC varied from 1.5 min for the 4-dimensional case to 5 min for the 10-dimensional case. The run time for QUASI was dependent on the linearity of the system equations and some of the options offered by QUASI. The run time for the 4-dimensional case was 25 s and for the 10-dimensional case it was 10 min.

The mechanization is being tested on trajectory problems of dimensions higher than 10, where QUASI attempts automatically to make trade-offs between run time and the amount of storage required. Procedures are also included so that QUASI will change its procedure automatically for solving a new optimum trajectory. These procedures will drastically reduce the necessary computations as the solution converges.

B. Optical Approach-Guidance Flight Feasibility Demonstration, *T. C. Duxbury*

1. Introduction

Studies indicate that improvement can be obtained in a Mars-encounter earth-based orbit estimate if information defining the direction from the spacecraft to Mars to an accuracy better than 1 mrad ($1\ \sigma$) is included in the estimate. An optical approach-guidance flight feasibility demonstration (SPS 39-42, Vol. IV, pp. 48–49) on the *Mariner* Mars 1969 mission was to use an on-board planet tracker. However, with the deletion of the planet tracker from the mission (as a result of budget constraints), a study was initiated to evaluate existing on-board sources of optical data that could be used for orbit determination.

2. Planet Tracker

The planet tracker was to give pointing angles to Mars during the 10 days before encounter. These angles along with spacecraft attitude information were to be ground-processed to produce a spacecraft trajectory estimate in near-real time.

A planet tracker was built, and data interfaces were established between the telemetry data stream in the Space Flight Operations Facilities and the optical data pre-processing software in the spacecraft performance analysis and command area. A draft description of a computer program for implementing each data interface was written. Equations were derived and documented that related the planet tracker measurements to the spacecraft trajectory parameters and to measurement errors. Orbit determination accuracy studies were performed to define measurement system accuracy requirements. A computer program simulating on-board system measurements was developed through a contracted effort (SPS 37-50, Vol. III, p. 104) to test and evaluate the ground processing software.

3. Alternate Sources of Optical Data

These sources include the far-encounter planet sensor, the scan platform, the attitude-control celestial sensors,

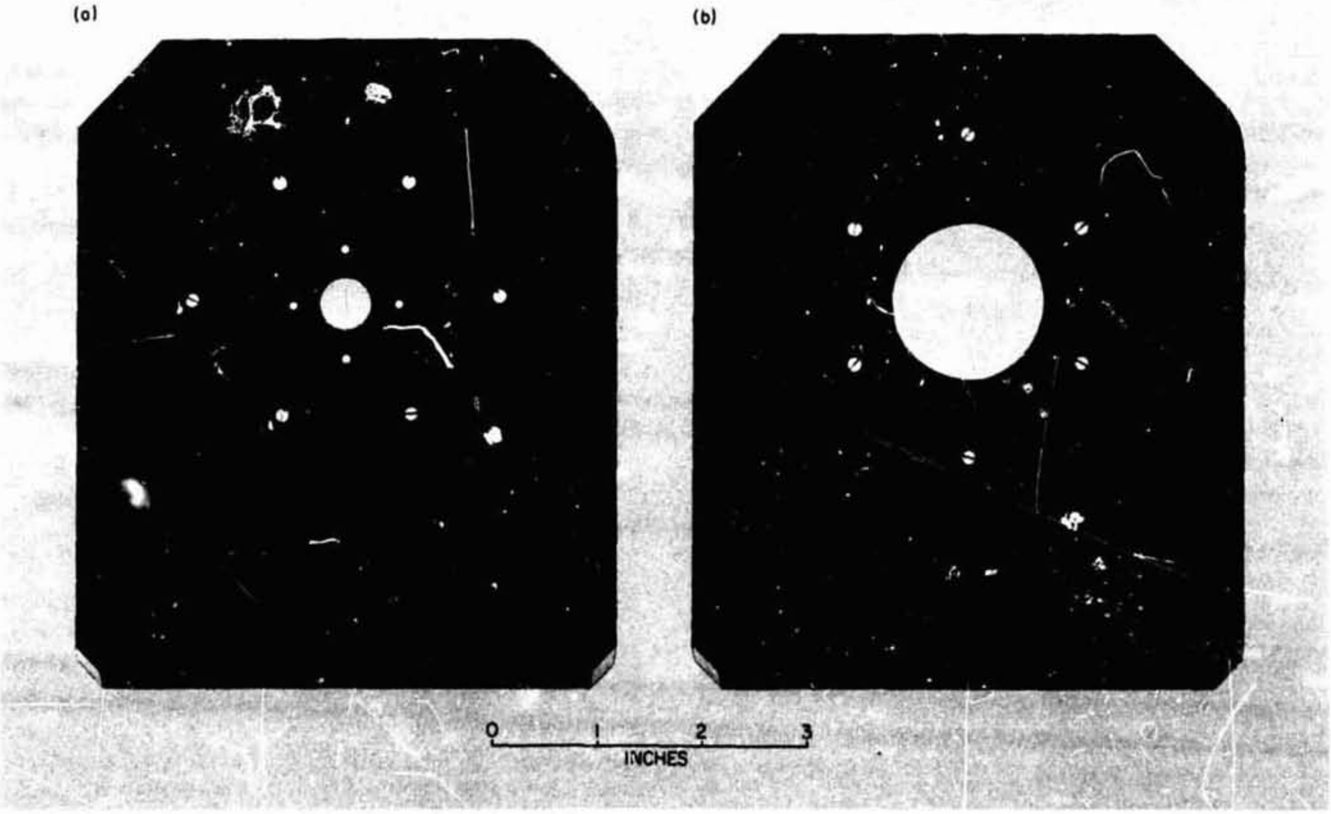


Fig. 2. Two TV calibration targets simulating Mars and eight stars: (a) 25-deg angle, (b) 75-deg angle

and the TV system. Sufficient accuracy can be obtained using data accumulated over a 24-h period.

Studies have also shown the usefulness of TV data in the orbit determination process when Mars is in a TV frame. The value of the TV data is greatly increased if a star is visible in the TV frame along with Mars. A star (ξ Serpentis) of 3.64 magnitude has a high probability of being in the TV field-of-view for trajectories having a launch data before March and an arrival date between July 31 and August 15. The TV was not specifically designed to photograph stars; therefore, it may not have the capability of detecting stars as dim as ξ Serpentis. To aid in determining if this capability exists, two test targets (Fig. 2) simulating Mars and eight stars ranging in magnitude from 1 to 5 will be included in the TV calibration schedule. The large hole in each test target simulates Mars, the cross-hair intersection designates the center of the large hole, and the eight apertures about the perimeter of the large hole simulate the stars. The geometric relationships between the simulated stars and planet have been measured to ascertain the accuracy with which the angle between a star and planet can be reconstructed from the TV data.

Selection of the set of optical data sources that provide the best orbit determination capability is under study; results will be reported in future articles of the SPS, Vol. III.

C. Development of Computer-Oriented Operational Support Equipment, *J. P. Perrill*

1. Objectives

The long-range objective is to develop the guidance and control operational support equipment (OSE) technology to meet the requirements of possible future planetary missions. Within this objective, the near-term goal is to develop an "OSE unified approach" concept. This concept is to be applied to the three guidance and control flight subsystems (electrical power, guidance and control, and central computer and sequencer) to provide an integrated approach to subsystem testing in the laboratory, manufacturing area, system test complex, and launch complex.

2. OSE Unified Approach

This concept will specify the use of the same basic OSE in all test areas where a flight subsystem exists as an assembled entity. Adaptors, buffers, or additional cabling will be added in areas where more test points are available and where more detailed tests are required.

The basic control element in all test areas is a small, commercially available general-purpose computer. A

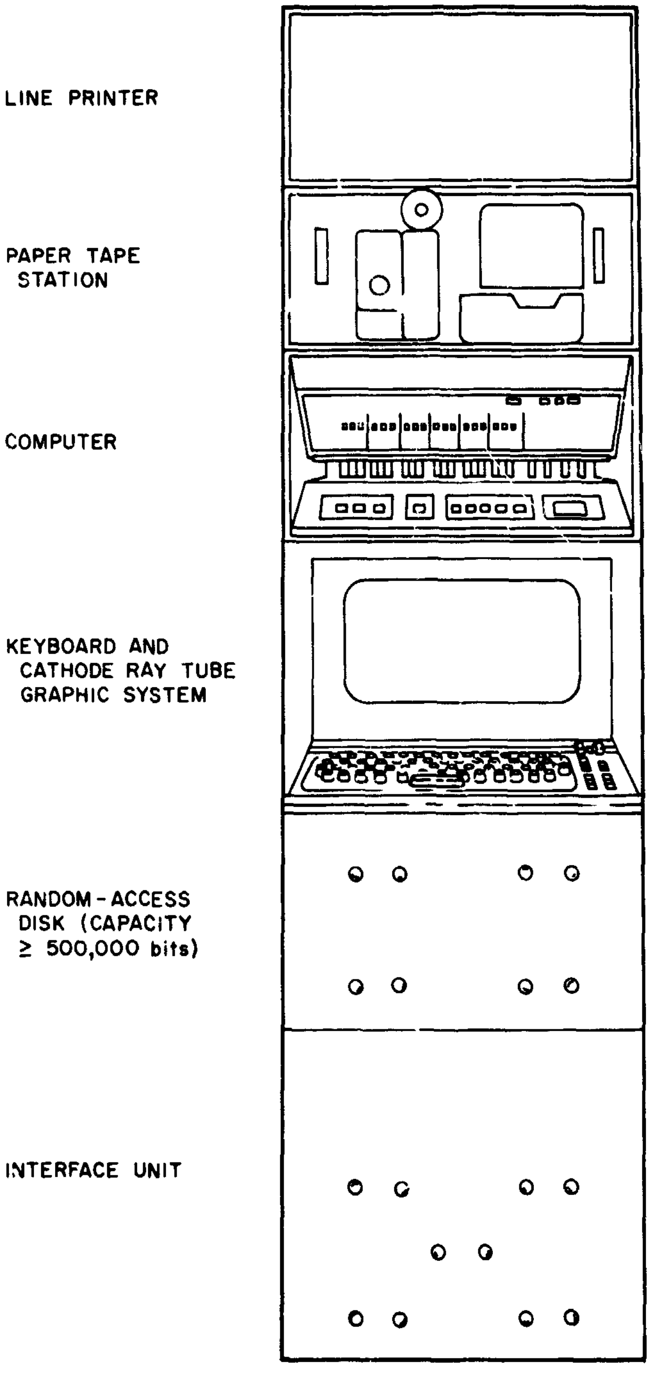

LINE PRINTER
PAPER TAPE STATION
COMPUTER
KEYBOARD AND CATHODE RAY TUBE GRAPHIC SYSTEM
RANDOM - ACCESS DISK (CAPACITY ≥ 500,000 bits)
INTERFACE UNIT

Fig. 3. Proposed hardware configuration for flight project operational support equipment

versatile man–machine interface is provided by a cathode ray tube graphic system, permitting instant input/output access to the user. A test language is provided that requires relatively little training and is based on user requirements rather than computer characteristics. The language has on-line response and presents the user with the capability of direct control access to the unit under test, with automatic checking of responses. Alterations from mission to mission are expected to be more economical than the present method of reworking existing OSE hardware.

The hardware interface unit provides special-purpose logic, signal conditioning, and buffering between the unit under test and the general-purpose input/output channels of the computer. The interface unit is designed as a "sample-and-hold" device in both directions, with control reserved and maintained by the computer. Figure 3 shows the envisioned OSE hardware that would form the computer-oriented test system.

The software functions as an interpreter between the test engineer and the unit under test. A primary function of the interpreter is to take a user-defined engineering language test program and develop the subsystem test program while maintaining an error monitoring and display capability.

The software is tailored to a hardware configuration of a central processing unit, display and/or typewriter, subsystem interface, and bulk storage. The software is modular in concept, real time in operation, and is classed as an interpreter. Among the attractive features planned for this interpreter are:

(1) A basic set of frequently used, thoroughly checked elementary programs that may be selected and run by the user with minimal effort.

(2) Orderly growth of programs obtained by user experience and the sequencing of elementary subprograms.

(3) A programming language consisting of engineering parameters rather than the mnemonics used by programmers.

(4) The ability to change, update, clear, and insert into the existing program on line; i.e., dynamic change of the checkout program without the aid of an off-line software assembly.

3. Progress

The preliminary design of the computer-oriented test system is essentially complete. A functional requirements document was generated in December 1967, describing a feasibility demonstration model to test a *Mariner* Mars 1969 central computer and sequencer (CC&S) spacecraft subsystem. This CC&S is considered to be representative of future flight project hardware. Preliminary software flow charts that incorporate the CC&S as the subsystem to be tested have been prepared up to the point of machine dependency, and will be completed when the particular computer system is selected and the computer procured.

VI. Spacecraft Control

GUIDANCE AND CONTROL DIVISION

A. Sterilizable Inertial Sensors: Gas Bearing Gyros, *P. J. Hand*

1. Introduction

The objective of this task is to perfect a complete family of miniature inertial sensors that will be capable of withstanding both thermal and gas sterilization without significant degradation of performance. Included in this family are long-life rate-integrating gas bearing gyros, subminiature ball bearing gyros, and high-performance linear accelerometers. These sensors have potential applications in both advanced spacecraft and entry capsule attitude control systems.

The gas bearing gyroscopes selected for this effort are the Honeywell, Inc. type GG159 and its wide-angle counterpart, GG334S. The gas bearing gyro does not demonstrate any wearout conditions during operation and can, therefore, be considered for application on very long missions.

2. Developmental Background

Experience with the GG159 gyro began at JPL in 1962 with the evaluation of a standard production version (GG159B1). Evaluation of the B1 version was followed in 1963 with a development program at Honeywell to improve the *g* capability of the gas bearing motor. The improvement program resulted in a motor design that was capable of passing the JPL shock requirement of 200 *g* peak. This environment is required for operation on all JPL-designed spacecraft.

In 1964 the development program was broadened to cover: (1) a gyro containing the 200-*g* motor (GG159C7), (2) a study to develop a gimbal suspension pump to operate at higher frequencies, and (3) a study to improve the gyro torquer efficiency as well as the first attempt at a thermally sterilizable gyro (GG159D1). Knowledge and experience obtained from these study programs and from the D1 gyro were used in the redesigned GG159D2 gyro. This gyro successfully passed seven sterilization cycles at 135°C without significant degradation of the important gyro drift parameters. Worst-case drifts were less than 0.5 deg/h.

During 1966, while JPL was evaluating the D2 gyro, Honeywell was developing a wide-angle gas bearing gyro (GG334A). Initial attempts to develop a low-power (4.0 W) spin motor were also started. In mid-1966 a contract for a sterilizable version of this instrument, to be known as GG334S, was released. Later in 1967 work was started for JPL on the GG159E. This gyro will contain all the improvements developed since 1962 plus the low-power spin motor which was brought to an advanced state of development in the GG334S program. Salient characteristics of the GG159E and GG334S designs are compared in Table 1.

Table 1. Comparison of gas bearing gyros

Parameter	Gyroscope GG159E	Gyroscope GG334S
Gyro gain (input to output), deg/deg	200	0.40
Diameter, in.	2.2	2.2
Length, in.	3.1	3.0
Weight, lb	1.1	1.1
Gimbal suspension	Pumped fluid	Dithered pivot and jewel
Gimbal freedom, deg	±0.5	±3.0
Operating temperature, °F	115	180
Motor power (at 26 V rms, 800 Hz), W	4.0	4.0
Motor speed, rev/min	24,000	24,000
Angular momentum, g-cm^2/s	100,000	100,000
Drift rates		
g-sensitive (spin axis), deg/h/g	±0.50	±0.50
g-sensitive (input axis), deg/h/g	±0.46	±0.46
g-insensitive, deg/h	±0.30	±0.46
Random drift (1 σ), deg/h	0.008	0.01
Elastic restraint, deg/h/mrad	0.06	0.06
Anisoelastic coefficient, deg/h/g^2	0.15	0.15

3. Gas Bearing Gyros Description

The Honeywell type GG159 is a miniature, high-gain, single-axis, floated, rate-integrating gyroscope, utilizing a hydrodynamic spin-motor bearing. This gyro was selected for sterilization development because it had the greatest potential for surviving the thermal environment. (The gas film bearing did not suffer from lubrication breakdown at the original sterilization temperature of 145°C).

The spin motor (Fig. 1) is designed to operate on 26-V rms, 800-Hz power and rotates at 24,000 rev/min, producing an angular momentum of 100,000 g-cm^2/s. As with all hydrodynamic bearings, the rotor is in contact with the journal at the start of motor operation, but lifts off within a few revolutions. The rotor is then carried on a gas film less than 100 μin. thick.

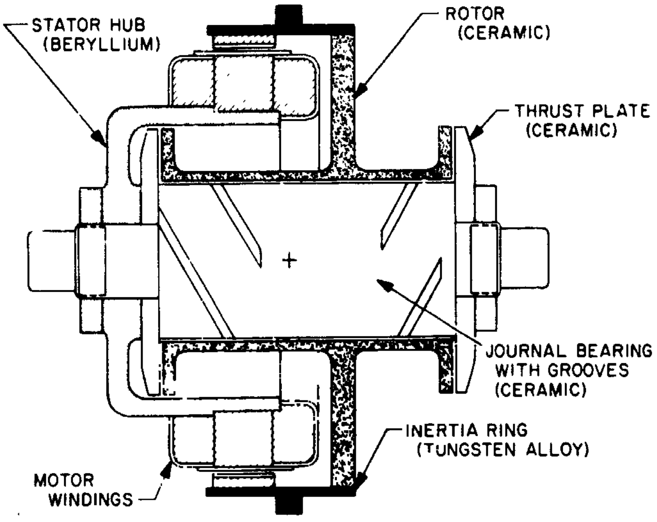


Fig. 1. Miniature gas bearing spin motor

The spin-motor construction materials are largely ceramic except for the magnetic parts and the inertia ring. To prevent scuffing or abrasion between rotor and journal during starts and stops, the ceramic materials are made very hard and are highly polished.

The gimbal structure, which carries the spin motor, is also made of ceramic to provide matching thermal expansion characteristics. This gimbal is floated at neutral buoyancy in a dense fluorolube fluid. The GG159 uses a very low viscosity fluid which allows a high input-to-output gain to be obtained at the gyro gimbal. A normal gain of 200 at 115°F operating temperature is developed. The floated gimbal of the GG159 is also suspended by pumping the same fluid through controlled orifices between the gimbal and the outer case.

The flotation fluid in the GG334S is very viscous and, therefore, supplies large damping forces to the gimbal. The GG334S gain is 0.40, allowing the gyro to store an input angle of ±7.5 deg. This flotation fluid is too viscous to allow pumping in the manner of the GG159. The gimbal suspension of the GG334 is more conventional in that pivots and jewels are used; however, the jewel is oscillated by a piezoelectric dither plate to eliminate static friction from the suspension. In both gyros, the outer case is made of conventional aluminum alloy with an integral heater and temperature sensor attached.

The GG159E is the culmination of the effort to perfect a gas bearing gyro for spacecraft operation. The GG334S will contain the same improvements as the GG159E but will be capable of storing up to ±7.5 deg of input angle information directly on the gyro gimbal without requiring the large integrating capacitors which the present *Mariner* spacecraft uses.

4. Status

Final fabrication of both the GG159E and the GG334S has been delayed due to a moisture contamination problem within the gimbal. This has been solved by redesign of the journal bearing to cause gas to flow through the bearing. Evaluation of both types of gyros will take place at both JPL and Honeywell during the latter half of 1968 and early 1969. Performance data will be presented in future editions of the SPS, Vol. III.

B. Analysis of Ion Thruster Control Loops,

P. A. Mueller and E. V. Pawlik

1. Introduction

Data on electric propulsion systems indicate ion thrusters to have several nonlinear properties that make the use of computer simulation and analysis quite attractive. Computer studies have been performed on controls for a thruster suitable for use as primary propulsion of a spacecraft for deep space missions (such as a Jupiter flyby[1]).

These computer studies have been performed for thruster controls and power matching that have been previously proposed for a thruster employing an oxide-coated cathode (Refs. 1 and 2). In the proposed control scheme, two control loops are utilized to maintain the thruster at a desired operating point (thrust) despite variations in cathode emission, vaporizer porosity, and thruster thermal emissivity during the thruster operating lifetime that may be as long as 10,000 h. In addition, the relationship between the two thruster control loops is used to indirectly specify the mercury propellant flow rate to the thruster.

The computer simulation was performed with Digital Simulation Language 90 on the IBM 7094 computer.

2. Computer Simulation Model Considerations

The thrust is approximately proportional to the product of the ion beam current I_B and the net acceleration voltage. Typical power conditioning and control loops for the 20-cm-diam thruster being simulated are shown in Fig. 2. For the present system the net acceleration voltage, the output of supply V5 (the high-voltage screen supply), is held constant and the ion beam current is commanded to operate over a two-to-one range corresponding to 0.5 to 1.0 A. Other fixed value supplies are V1, V4, and V6, which are the electromagnet supply, the arc or discharge voltage supply, and the high-voltage accelerator supply, respectively. These fixed values simplify the control loops to those presented in Fig. 3. Supplies V2 and V3 are the two controlled supplies, the vaporizer supply and the cathode supply, respectively.

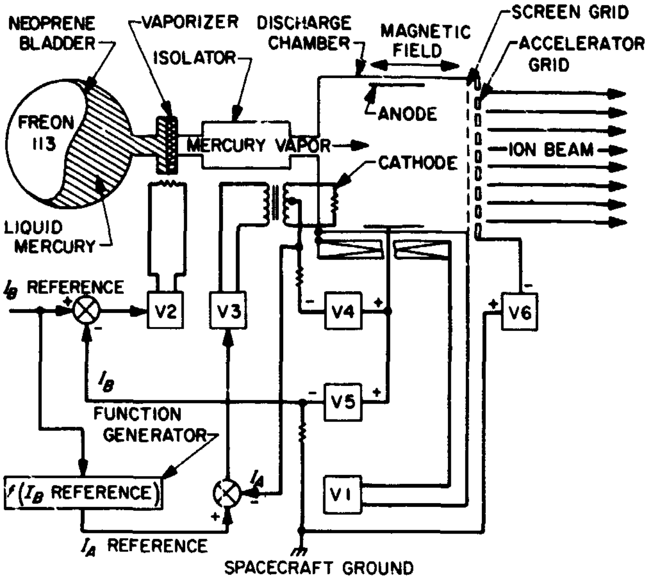


Fig. 2. Power conditioning and controls block diagram

Figure 4 presents the thruster ion chamber nonlinearity for a typical thruster. The straight lines for constant ion chamber or arc current I_A approximate the curved lines obtained from a characteristic mapping of a thruster which is not operating at maximum efficiency (Ref. 2). Characteristics of an optimized thruster have constant I_A lines that have negative slopes at low propellant utilization values. (This condition is also being studied.) For the nonoptimized thruster, the straight-line approximation is accurate for points where propellant utilization (the fraction of mercury propellant ionized and accelerated as the ion beam) is 0.8 or less. For higher values of utilization the constant I_A lines have greater slopes. Operation at a constant utilization value implies a unique combination of beam current and arc current (except where utilization is 1.0). This unique relationship is the function generated in the block, "Arc Current Reference." The particular function chosen depends on the propellant utilization value desired. All simulations to date have been at 0.80 utilization which is indicated by the heavy line in Fig. 4.

Without the controllers in the arc loop, the loop has one time lag of approximately 120 s due to the cathode. Nonlinearities cause the gain to vary by as much as a factor of 4, depending on the thruster conditions and the

[1] *1975 Jupiter Flyby Mission Using a Solar Electric Propulsion Spacecraft,* Mar. 1968 (JPL internal document).

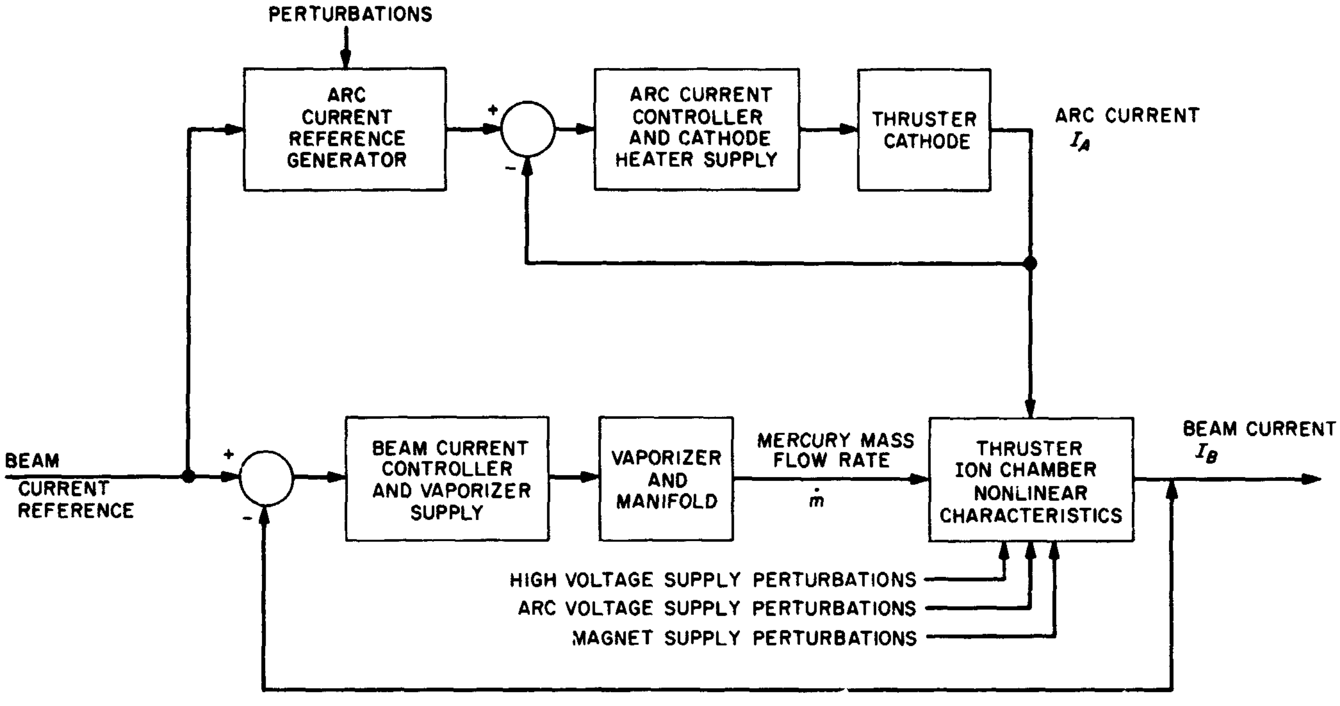

PERTURBATIONS
ARC CURRENT REFERENCE GENERATOR
+
−
ARC CURRENT CONTROLLER AND CATHODE HEATER SUPPLY
THRUSTER CATHODE
ARC CURRENT I_A
BEAM CURRENT REFERENCE
+
−
BEAM CURRENT CONTROLLER AND VAPORIZER SUPPLY
VAPORIZER AND MANIFOLD
MERCURY MASS FLOW RATE $\dot{m}$
THRUSTER ION CHAMBER NONLINEAR CHARACTERISTICS
BEAM CURRENT I_B
HIGH VOLTAGE SUPPLY PERTURBATIONS
ARC VOLTAGE SUPPLY PERTURBATIONS
MAGNET SUPPLY PERTURBATIONS

Fig. 3. Block diagram of two ion thruster control loops

degradation with use of the cathode. Without the controller in the beam loop, there are time lags of 0.02 s in the thruster manifold and from 120 to 600 s in the vaporizer. Gain variations attributable to component nonlinearities are on the order of 2.

3. Controllers

Several controllers have been studied for the two arc loops, including simple gain factors (proportional), type 1 or integral, and integral with lead compensation. Proportional and integral with lead appear most promising. Similar controllers are being considered for the beam loop. With ideal components used in the controllers, stable performance with 0.1% beam current error (difference between the specified and the actual beam currents) and 1% utilization error (difference between the desired and the actual utilizations) appears to be realistically feasible. Drift, offset, and other errors associated with the controllers' electronics, are being considered as the computer simulation becomes more complete.

4. Ion Chamber Perturbation Studies

The model of the thruster ion chamber nonlinear characteristics presented in Fig. 4 depends on other thruster parameters remaining constant. These parameters include the magnetic field, discharge voltage, screen voltage, and accelerator voltage. Deviations from these fixed values introduce variations in the ion chamber characteristics. The regulation of the power supplies, therefore, becomes an important consideration. Variations of the fixed outputs of the electromagnet supply, the high-voltage supplies for maintaining the voltage between the grids of the ion extraction system (the sum of the screen and accelerator voltages), and the discharge of arc voltage supply (V7, V5, V6, and V4, respectively) are the important perturbations. The predominant effect of the variation in power supply output is a translation effect in the characteristics. The dashed lines in Fig. 4 indicate such a shift due to a very small perturbation and are explained in the following paragraphs.

The computer was again used in determining the perturbation effects. A perturbation factor was calculated for each of the supplies using the experimental data obtained for large differences in supply outputs. These factors were introduced into the simulation program and computer runs were made for ±10% errors on the outputs of the power supplies.

While perturbations were introduced, the closed-loop system maintained the set point values of I_A and I_B. How-

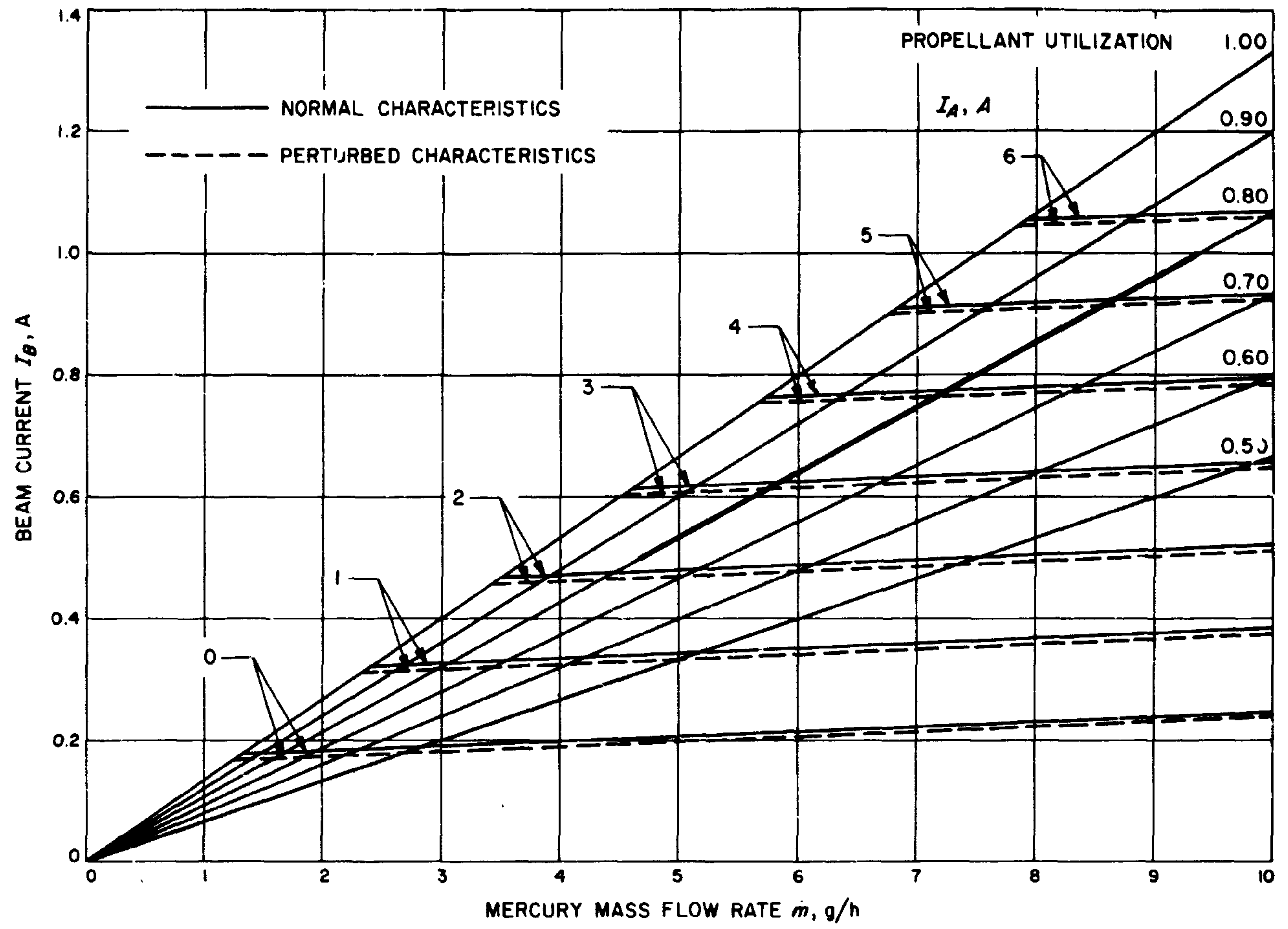


Fig. 4. Typical ion chamber nonlinear characteristics

ever, the set point values in this situation no longer defined the specified propellant utilization. Because of both the slight slopes in the ion chamber characteristics and the large perturbation factors, small variations in supply outputs resulted in large variations in propellant utilization. The worst-case variations obtained for the electromagnetic supply, high-voltage, and arc-voltage errors were 2, 1.5, and 0.5%, respectively, with propellant utilization errors of 0.05 in each case. All three supplies have the same effect so that the error can be additive; i.e., if the three supplies had the above errors, the utilization error would be 0.15. An error of 0.15 in the utilization when the set point is 0.80 is an error of 18.8%. The dashed lines in Fig. 4 are for this case.

To minimize the susceptibility of propellant utilization to such perturbations, greater characteristic slopes are desirable. One method of achieving this is to run at a higher value of utilization since, as previously mentioned, the true thruster ion chamber characteristics have a greater slope by a factor of 3 or 4 in the utilization range of 0.9 to 0.95. This was not considered in the computer model since 0.80 utilization was the designated set point.

5. Arc Reference Perturbations

The arc current reference function generator is also a critical component in determining the tolerances in propellant utilization. An error in the reference has the same effect as an error in the arc current itself. Computer simulation resulted in a worst-case error of 7.5% in utilization corresponding to a 1% error in the arc reference. Since the arc reference is a critical factor, it must be maintained with much better stability than 1% if the propellant utilization errors are to be kept within a realistic range of a few percent.

6. Conclusions

Several conclusions can be drawn from this study of throttled ion thruster controls. Stable ion thruster performance is feasible with a beam current error limit.tion of 0.1% and a propellant utilization error limitation of 1%. Errors of a few percent in the fixed output power supplies may cause utilization errors 10 times as large. Arc reference errors of 1% may yield propellant utilization errors of 7.5%. Thruster performance in the utilization range of 0.8 to 0.95 merits further investigation.

References

1. Pawlik, E. V., *Power Matching of an Ion Thruster to Solar Cell Power Output*, Technical Memorandum 33-392. Jet Propulsion Laboratory, Pasadena, Calif. (in press).
2. Maser, T. D., and Pawlik, E. V., "Thrust System Technology for Solar Propulsion," Sections III and IV, Paper 68-541, to be presented at the AIAA Fourth Propulsion Joint Specialists Conference, Cleveland, Ohio, June 1968.

C. Powered Flight Control Systems,

R. J. Mankovitz

1. Introduction

The original objective of this task was to develop nonlinear digital computer programs for various powered flight control systems, and to utilize these programs to perform parametric trade-off studies that could be used to select optimum systems for given requirements. After completion of a six-degree-of-freedom program for a gimballed-engine (chemical propulsion) autopilot system, the objective was revised, due to budgetary constraints, to the study of the attitude control of an electric-propulsion-powered (ion engine) vehicle, during the powered fligh^t phase. This work was directly applicable to an Advanced Technical Studies task related to a solar electric-powered spacecraft mission to Jupiter.

2. Basic Considerations

Trade-off studies have been conducted for the attitude control of an electric-propulsion-powered vehicle. A complete six-degree-of-freedom digital computer simulation has been developed and used to evaluate the following basic concepts.

(1) Three-axis cold gas control.

(2) Two-axis engine translation with third-axis cold gas control.

(3) Two-axis engine translation with engine gimballing for third-axis control.

In addition to the basic concepts, a hot gas system (resistojets) was considered in place of the cold gas system. Solar pressure control augmentation was also considered by rotating the solar panels (panel trim) to obtain solar torques. As a result, the third alternative (3) was chosen as the baseline configuration. This configuration reduces the gas usage to zero for powered flight control, and only requires a total of 20 lb of cold gas during the nonpowered phase. The hot gas and panel trim alternatives were rejected on the basis that the significant increase in complexity does not result in a significant reduction in stored-gas weight.

A basic control law has been developed and analyzed for the chosen configuration and has been demonstrated to p'·· ·ide stable operation.

3. Functional Description

During the powered flight portion of a solar electric mission, the spacecraft must remain sun–Canopus-oriented and have the ability to point the ion engine thrust vector over a 180-deg angle in the ecliptic plane, even when out of the ecliptic plane by as much as ±3 deg. A total thrust vector pointing error (in celestial coordinates) of less than 1 deg is desired.

a. Three-axis attitude control. The method selected for providing three-axis attitude control during that portion of powered flight when more than one ion engine is operating consists of a two-axis, bi-directional engine-translation system with third-axis control (thrust vector axis) provided by gimballing the outermost ion thrusters on the engine array (Fig. 5). The ion thruster gimbals are single-degree-of-freedom with opposite thruster gimbals slaved to each other. If one of the outermost thrusters should fail, control is switched to the other opposite pair. The baseline control system requires ±10-deg gimbal excursions and ±12-in. translator travel. Both the translator and gimbals are stepper motor-controlled, with resolutions of 0.005 in./step on the translator and 0.1 mrad/step on the gimbals. Maximum stepping rate for all systems is 50 steps/second. All control loops are passively compensated and do not require gyro signals.

b. Cold-gas/ion-engine switchover.[2] Upon completion of the Canopus acquisition phase, the ion engines are activated. A 5-min duration is allowed to permit the engines to achieve full thrust. During this period, the translation and gimbal systems are inactive, and the cold gas

[2] See also: *Section E*, "Extended Mission Control Systems Development."

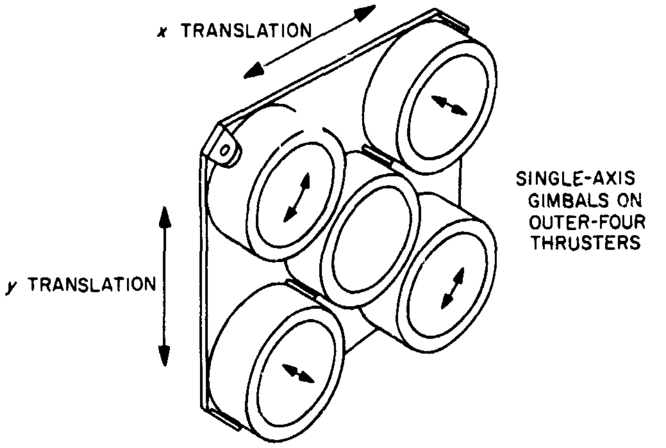


Fig. 5. Powered flight control using two-axis translation and third-axis gimballing

system maintains attitude control in the presence of thrust vector misalignments. At the end of this 5-min period, the engine control systems are activated, and the cold gas control system deadbands are increased from ±0.5 to ±3 deg. Under normal operating conditions, the engine translation and gimbal systems will be operating within the deadband of the gas system. If, however, a large disturbance should cause loss of acquisition, the ion engines will be deactivated and the celestial references reacquired using the cold gas system.

In the limit cycle mode, the engine control system stepper motors will pulse at a maximum rate of 1 pulse/ 32 s. The average deadband size is approximately ±1 mrad. The control system can recover from a 3.0-ft-lb-s torque impulse without losing celestial lock.

c. Orientation of the thrust vector.

Full thrust mode. Nominally, the spacecraft x–z plane is co-planar with the ecliptic plane. To meet the requirement of 180-deg angular freedom in pointing the thrust vector in the ecliptic plane, the ion engine array is mounted on a single-degree-of-freedcm platform which can rotate 180 deg in the spacecraft x–z plane. The position of this turret will be changed in increments (< 0.1 deg) determined by a central comp·iter and sequencer (CC&S)-stored thrust poinung program.

The ±3-deg out-of-plane pointing capability is mechanized by appropriately biasing the pitch sun sensor and the Canopus tracker error signals. Thus, the spacecraft itself is rotated about the pitch and roll axes to point the thrust vector. Utilizing sensors with ±8-deg linear fields of view enables the spacecraft to perform these turns without losing the celestial references. As in the case of the in-plane pointing angle, the out-of-plane angle is supplied by a stored CC&S program. This angle is resolved into sensor bias signals without the use of trigonometric functions.

A functional block diagram of the attitude control system during this phase is shown in Fig. 6. Since the engine control axes will, in general, not coincide with the spacecraft axes, a coordinate transformation is required to convert error signals from the spacecraft axes to the control axes. Since the engine pointing angle varies throughout the powered flight phase, the variable transformation mixing matrix is mechanized with resolvers The resolved error signals are sensed by the control systems, and translation and gimbal deflections of the thrust vector produce the three spacecraft body control torques. The control torques act through the spacecraft structural dynamics to counteract the disturbing torques and produce the error signals θ_x, θ_y, θ_z.

Reduced thrust mode. During the latter portion of the powered flight phase, only one ion engine is operating. Since this engine is centered about the spacecraft center of gravity by the translator, engine gimballing can no longer provide third-axis control. During this phase, the gimbal control system is deactivated, and the cold gas system is used to control the torques about the engine axis. The only disturbance torque generated by the single engine is due to swirl of the ion stream (engine misalignments are removed by the translator), so that only a small amount of cold gas is required during this phase.

4. Control System Analysis

Some of the powered flight attitude control mechanizations that were considered for this phase are:

(1) Three-axis cold gas control.

(2) System (1) with solar panel trim.

(3) Two-axis translator control plus cold gas third-axis control.

(4) System (3) with solar panel trim.

(5) Two-axis translator control plus gimbal third-axis control.

(6) System (5) with solar panel trim.

(7) Any of the above systems using heated N_2 (resisto jets).

The three basic systems are (1), (3), and (5). For those systems requiring cold gas for control, the options of heating

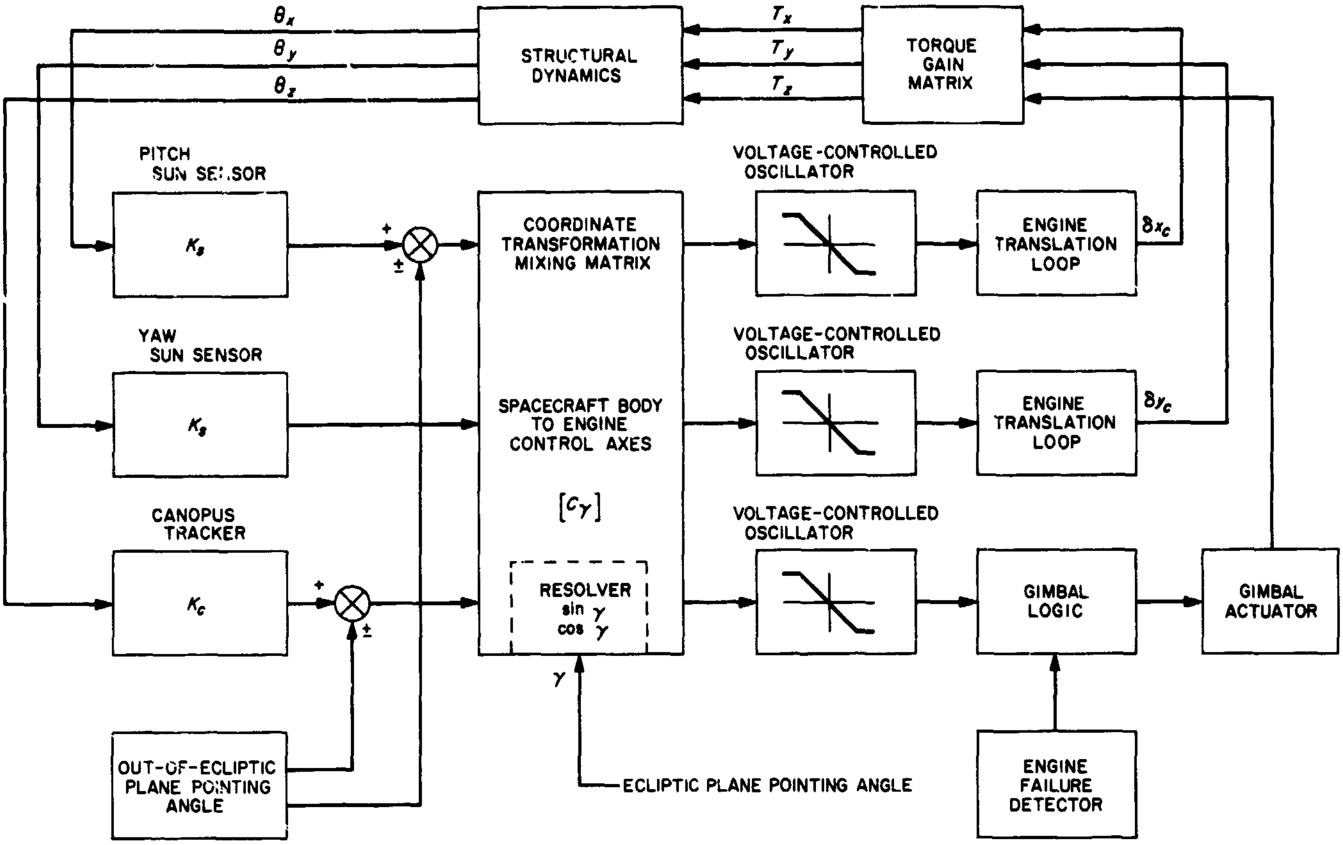


Fig. 6. Attitude control system during thrust phase

the gas to increase the $N_2 I_{sp}$ and tilting the solar panels to balance the disturbance torques were considered.

To permit an evaluation of these systems, a digital computer program was used to determine the attitude control gas storage requirements for each system during a 1200-day mission.

The mission gas storage requirements for each system is presented in Table 2. The use of a gas system for three-axis control. assuming a 0.1-ft engine array center of gravity offset and a 1-deg engine angular misalignment, was eliminated immediately due to excessive gas weight.

For the system using a two-axis translator plus gas system third-axis control (assuming a 1-deg engine angular misalignment), the only case that appeared feasible from a gas-weight standpoint required hot gas and solar panel trim capability. Considering the added weight, the decrease in reliability attendant with rotation of 46-ft-long solar arrays (as well as structural dynamics problems), and the lack of long term flight experience with hot gas systems, this mechanization was eliminated.

The third, baseline, mechanization, which requires a 20-lb gas weight, uses a two-axis translator with gimballed engine third-axis control. Cold gas is only used for acquisitions, cruise (nonpowered flight), and for third-axis control during that portion of the powered flight phase when only one engine is operating. Neither heated gas nor solar panel trim is required for the baseline mechanization.

The basic control loop for either the translator or gimbal system is shown in Fig. 7. The input is a position signal from a celestial sensor, referenced to spacecraft axes. An attitude bias, in the form of a dc voltage summed with the sensor output, may be present to orient the spacecraft out of the ecliptic plane for thrust vector pointing. The sensor signals are mixed in a transformation matrix to go from spacecraft axes to engine axes. The matrix is a function of the angle γ, which is the ecliptic plane engine pointing angle. The engine-referenced error signal is used to drive a voltage-controlled oscillator (VCO), yielding a variable frequency pulse train that is used to drive a stepper motor. Thus, the stepper motor rate is proportional to the error magnitude. The motor is used to drive either the translator platform or the engine gimbals to

Table 2. Attitude control gas storage requirements

Basic system	Cold gas, lb		Hot gas, lb	
	No panel trim	Panel trim	No panel trim	Panel trim
Three-axis cold gas	~1100	~800	~650	~500
Two-axis translator plus cold gas third axis	85	56	51.5	35.2
Two-axis translator plus gimbal third axis	20.4	20.4	15	15

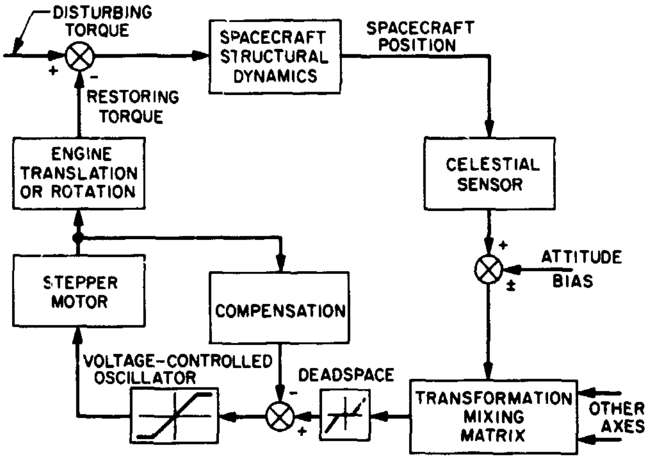


Fig. 7. Basic translation or gimbal control loop

produce restoring torques. These act on the structure, which includes the dynamics of the solar arrays.

Compensation networks are required to stabilize the loops, and since gyros cannot be used for extended durations, passive rate compensation must be employed.

The electronics required for the control systems are mechanized with linear and digital integrated circuits, employing triple modular redundancy for increased reliability. Redundant sun sensors are employed, and it also appears desirable to employ dual Canopus trackers which can be switched by ground command.

Since the translator and gimbal positions (and thus the restoring torques) are a discrete function of time, the steady-state behavior of the control loops will be a limit cycle of nominally ±1 step about the balanced torque point.

To analyze the control loops for the large signal mode, linear analysis methods were used to approximate the nonlinear loops. Considering the baseline configuration, it can be shown that the sampling rate (VCO output) of the actual loop is sufficiently high, in all modes except the steady-state limit cycle, to permit the use of linear analysis for preliminary investigations. Digital computer simulation programs were constructed to verify the analysis. Some of the major problems in mechanizing these loops are: interaction with the solar panel structural dynamics, passive rate compensation alone, and sensor noise.

The block diagram of a single-axis translator control system is shown in Fig. 8. Compensation (lag) is placed in the feedback loop, as opposed to lead compensation in the forward (sensor) loop, to minimize problems due to sensor noise. In addition, the sensor output is fed through a deadspace which is sufficiently wide to reject the ambient tracker noise at null, thus preventing stepper motor dither. A combination of positive and negative feedback is used to minimize steady-state error in spacecraft position. The structure (solar array) dynamics are modeled as a fourth-order polynomial, with coefficients chosen as a function of the spacecraft configuration. A first-order lag is associated with the sensor signals to model the effects of noise filters.

The gimbal control system block diagram is also shown in Fig. 8 and is analogous to the translator loop. The stepper motor output is proportional to gimbal angular position, which acts through the engine thrust F_G and moment arm L_G to produce restoring torque.

To optimize the compensation networks and determine the operating point for the system, a digital computer root locus program was used to analyze the open-loop transfer function.

The major parameters for the translator and gimbal loops are shown in Table 3. Many of the parameter values were dictated by hardware constraints, such as:

(1) A high stepper motor rate (SLEW) is desired. To achieve good dynamic response, 50 steps/s is considered a reasonable value for a magnetic detent stepper motor.

(2) It is desirable to minimize the step size to achieve accurate attitude control. The values chosen for K_m are within hardware capability.

(3) The translator and gimbal limits (δ_{max} and δ_{min}) were chosen as large as possible (to maximize restoring torque capability) within the structural limitations.

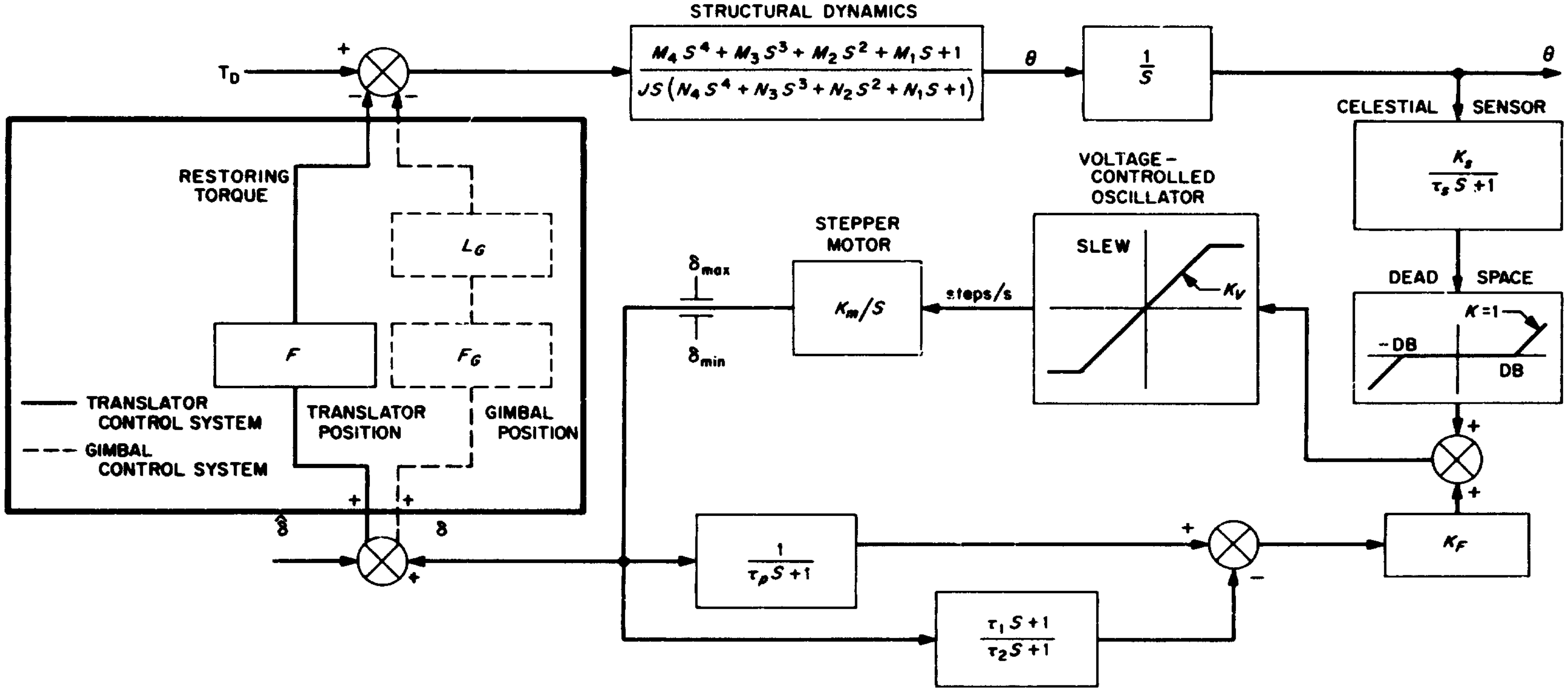


Fig. 8. Translator and gimbal control systems

Table 3. Translator and gimbal control system parameters

Parameter	Translator	Gimbal
Celestial sensor gain K_s, V/rad	134	134
Celestial sensor lag τ_s, s	0.5	0.5
VCO gain K_V, steps/s/V	287	1914
Maximum stepping rate SLEW, steps/s	50	50
Stepper motor gear train gain K_m, ft/step	4.167×10^{-4}	—
Stepper motor gear train gain K_m, rad/step	—	10^{-4}
Feedback gain K_F, V/ft	16.7	—
Feedback gain K_F, V/rad	—	10.45
Positive feedback lag τ_p, s	1000	1000
Negative feedback lead τ_1, s	50	50
Negative feedback lag τ_2, s	500	500
Maximum translator excursion δ_{max}, ft	1	—
Maximum gimbal excursion δ_{max}, deg	—	10
Minimum translator excursion δ_{min}, ft	−1	—
Minimum gimbal excursion δ_{min}, deg	—	−10
Ion engine thrust F, lb	0.01–0.06	—
Ion engine thrust F_G (2 engines), lb	—	0.02–0.03
Spacecraft inertia J, slug-ft^2	15,000–30,000	15,000–30,000
Coefficients of structural dynamics model M_1	6.34	6.34
Coefficients of structural dynamics model M_3	0.08	0.08
Coefficients of structural dynamics model M_2	5.04	5.04
Coefficients of structural dynamics model M_1	0.032	0.032
Coefficients of structural dynamics model N_4	0.73	0.73
Coefficients of structural dynamics model N_3	0.08	0.08
Coefficients of structural dynamics model N_2	2.8	2.8
Coefficients of structural dynamics model N_1	0.032	0.032
Distance from engine gimbal to array center of gravity L_G, ft	—	1.25
Celestial sensor deadband DB (equivalent to 1 mrad), V	$K_s \times 10^{-3}$	$K_s \times 10^{-3}$

(4) The sensor deadband (DB) is chosen sufficiently large, so that tracker noise will fall within its limits. Worst-case tracker noise, when acquired to Canopus, is estimated at 0.53 mrad peak to peak. The deadband width is chosen as ±1 mrad.

(5) The sensor filter time con tant (τ_s) is chosen to achieve the noise figure indicated above.

(6) The ion engine thrust range for the translator (F) covers the range from one to four engines operating from full to throttled-back thrust. The engine thrust range for the gimbal system (F_G) covers the throttling range of two engines.

(7) Since the engine bank can rotate 180 deg about the spacecraft yaw axis, both the translator and gimbal systems must be able to operate over the full range of spacecraft inertias (J).

(8) The torque moment arm (L_G) for the gimballed engines is determined by the engine diameter and engine mounting positions.

(9) The coefficients of the linear structural dynamics model (Ms and Ns) are calculated by a computer program. A solar array natural frequency of 0.1 Hz and a damping ratio of 0.005 were used, representing worst-case conditions.

From a closed-loop Bode plot (at a dc gain of 4×10^{-8}), the control bandwidth can be determined as 0.005 Hz. The effect of the panel dynamics occurs at 0.1 Hz.

To verify the performance and stability of the systems, a six-degree-of-freedom digital computer simulation program was constructed. Simulation results indicate stable operation over the gain variations anticipated (12:1 gain change due to thrust and moment of inertia variations).

The simulations also indicated that, with a ±5-deg sensor field of view and with minimum engine thrust, the control loops could maintain the spacecraft orientation when subjected to a 3-ft-lb second-torque impulse about all axes (corresponding to a step angular rate of ~0.1 mrad/s in all axes).

Further discussion of the results of the six-degree-of-freedom simulation will be presented in a future edition of the SPS, Vol. III.

D. Spacecraft Antenna Pointing for a Multiple-Planet Mission, *G. E. Fleischer*

Current preliminary studies of a gravity-assist mission (Grand Tour) to the Jovian planets (Jupiter, Saturn, Uranus, and Neptune) have included a rather broad look at the spacecraft high-gain antenna pointing problem. Several different pointing systems are being compared on

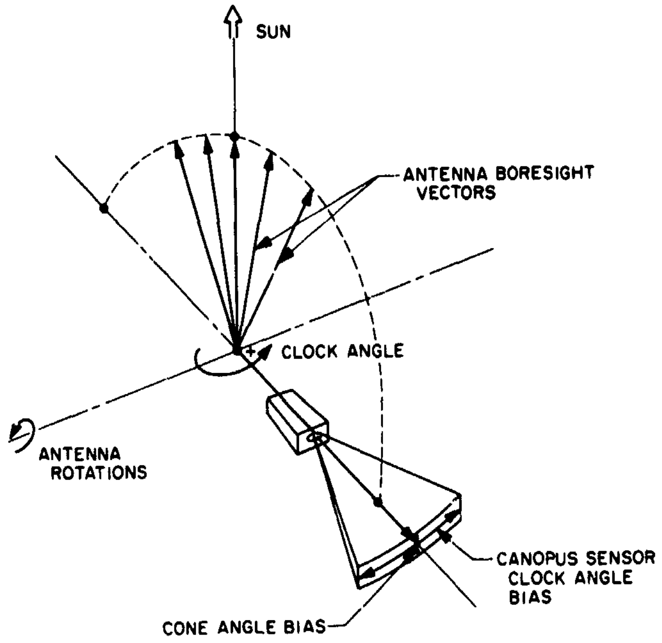
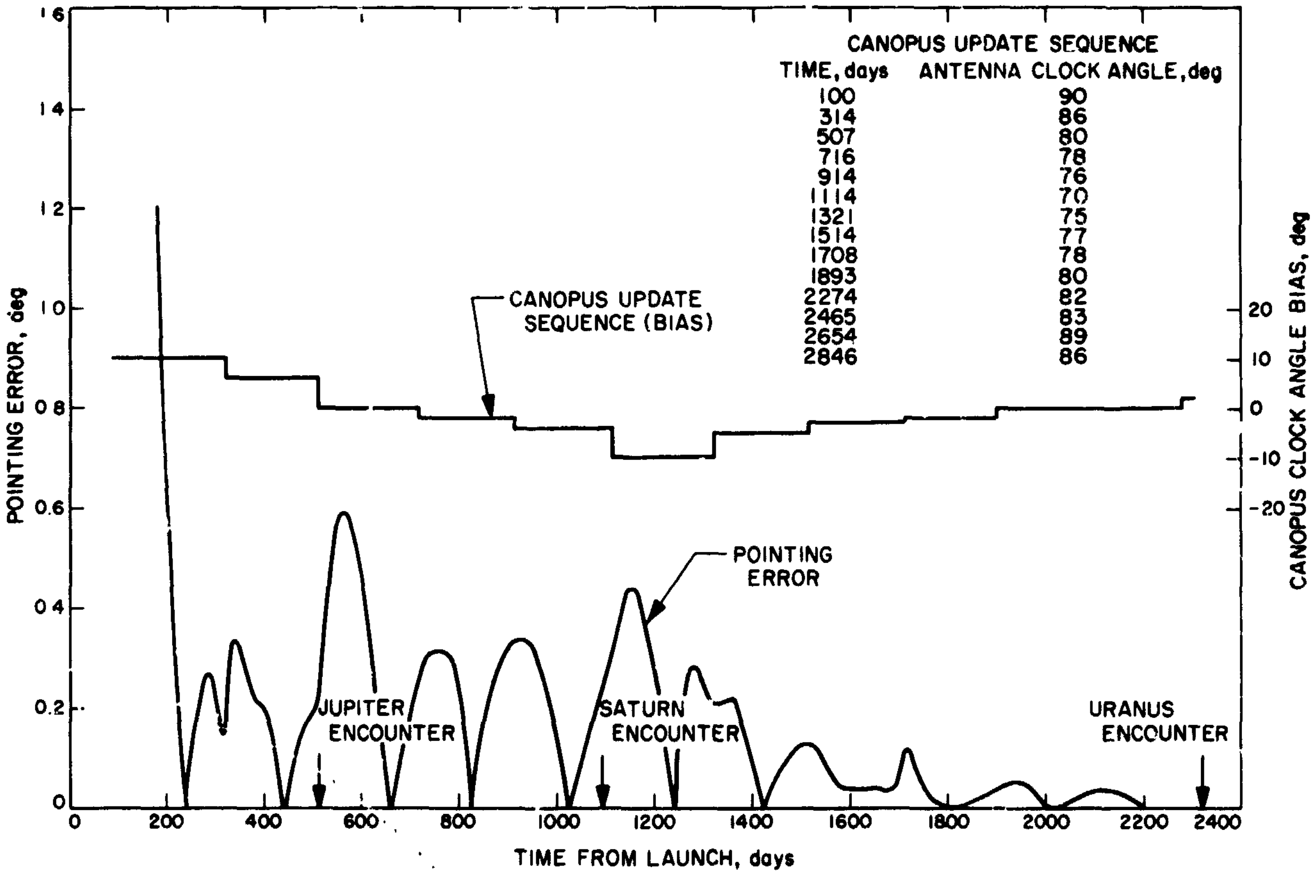


Fig. 9. Single-axis antenna with Canopus sensor clock bias



CANOPUS UPDATE SEQUENCE

TIME, days	ANTENNA CLOCK ANGLE, deg
100	90
314	86
507	80
716	78
914	76
1114	70
1321	75
1514	77
1708	78
1893	80
2274	82
2465	83
2654	89
2846	86

Fig. 10. Single-degree-of-freedom antenna articulation

the basis of their contributions to total communications system weight, power, and performance. A cone-clock type pointing system is briefly described here.

For the case in which the high-gain antenna is assumed to be articulated with respect to the spacecraft in one degree of freedom, two-degree-of-freedom earth pointing can effectively be achieved through the use of a spacecraft-fixed Canopus sensor whose field of view is electronically biased in clock angle (in addition to the present bias capability in cone angle). Thus, the clock angle degree of freedom is provided by the roll attitude control loop at relatively little cost in terms of weight and power. The antenna's degree of mechanical freedom quite naturally then becomes a rotation about an axis perpendicular to the spacecraft roll axis (Fig. 9). Antenna rotation is controlled by a stored program that generates the proper angular function of time (earth's cone angle). The result is a cone-clock type of pointing system with a limited capability (±15 deg) for biasing the Canopus sensor view in clock angle.

Due to this limitation in clock angle rotation of the spacecraft, an inherent pointing error occurs as the apparent earth track passes near the sun. Assuming all other pointing errors are negligible, that portion of the earth track outside of the region in which clock angle freedom is available for antenna pointing cannot be seen with zero error.

Assuming that very accurate pointing is not required during the earlier portions of the Grand Tour, the effect of a relatively coarse program of Canopus sensor clock bias angles was investigated and the result plotted in Fig. 10 Fourteen bias-angle updates of the clock angle are provided for the entire Grand Tour mission. Of course, the effects of spacecraft attitude errors, mechanical misalignments, cone angle program errors, etc., on total pointing error have not been included in Fig. 10; only the error resulting from a discrete and limited clock angle rotational capability is given.

E. Extended Mission Control Systems Development, *L. McGlinchey*

1. Introduction

The extended mission control systems development study is a new task for FY 1968. During the first quarter, a project was started to study the attitude control of vehicles utilizing electric propulsion systems. The scope of this work was directly applicable to an Advanced Technical Studies task related to a solar electric-powered spacecraft mission to Jupiter.

Providing attitude control for a solar electric spacecraft poses many new and unique configuration and design considerations not encountered previously. The mass and inertial properties of solar electric spacecraft pose unique problems with regard to sizing the control capability of the attitude control system, due to the constraints posed by dynamic interaction. The Jupiter spacecraft has inertias on the order of 15,000 slug-ft^2 about the pitch and yaw axes and 30,000 slug-ft^2 about the roll axis. These large inertias require a much higher control torque level to provide reasonable acquisition times and recovery from disturbances. In addition, the change in inertias (60:1) after solar array deployment requires that the attitude control system have a very large dynamic range.

The deployment of large solar arrays (1500 ft^2) can introduce disturbance torques that could cause such severe interaction with the attitude control system that the solar array structure and the deployment procedure would be adversely affected. Detailed structural analyses are required to evaluate this problem. At present, no detailed information is available regarding the structural properties of solar electric spacecraft. In this article the results of the attitude control system study are based on a linear lumped parameter model of the solar array structural dynamics, with the remainder of the spacecraft considered as a rigid body. On this basis, the baseline attitude control system was designed to be compatible with the structure. However, considerable analysis must be done to fully investigate and model all possible adverse structural resonance modes that can affect the attitude control of this type of spacecraft.

In addition to structural interaction, incident solar radiation on the large solar arrays can reduce significant disturbing torques on the spacecraft. Similarly, gravity gradient disturbance torques can be significant in the vicinity of the planet, especially a planet the size of Jupiter. In the case of the Jupiter mission, the attitude control system was configured for worst-case solar pressure and gravity gradient unbalance torques.

Several alternate attitude control configurations for the Jupiter spacecraft were examined during the course of the study. The following discussion describes the baseline attitude control system for the nonpowered flight portion of the mission. Attitude control during the powered flight phase is described in *Section C.*

2. Baseline Configuration Functional Description

a. General attitude control requirements. The basic requirements for the attitude control of the spacecraft are as follows:

(1) Provide initial rate removal and stabilization of the spacecraft following separation and solar panel deployment.

(2) Acquire celestial references (sun and Canopus).

(3) Provide thrust vector orientation and maintain a stable attitude during the thrust phase.

(4) Maintain a stable attitude during the cruise phase.

(5) Provide immediate reacquisition of the celestial references as required.

(6) Provide antenna and science instrument orientation as required.

The above requirements, with the exception of the third, do not pose serious constraints on the selection of an appropriate attitude control configuration. The third requirement presents unique problem areas because of (1) the duration of the thrust phase (470 days), and (2) a requirement for pointing the thrust vector out of the ecliptic plane (see *Section C*).

b. Nonpowered flight functional sequence and attitude control modes. During all phases of the mission, except the powered flight phase, attitude control and stabilization of the spacecraft is obtained by control torques provided by a N_2 cold gas mass expulsion system. The N_2 cold gas system was selected as the most feasible for the following reasons:

(1) Simplicity and inherent reliability.

(2) Space proven, particularly on *Mariner IV* where this type of system operated for over 1000 days.

(3) Minimum weight consistent with the attitude control requirements.

The basic operation of the cold gas system is as follows: error signals are measured by position and rate sensors and summed in their respective channels to operate gas-jet valve-switching amplifiers, which provide an *on–off* type control torque. A position limit cycle about each of the control axes is established by a switching amplifier deadband. A rate feedback signal provides the proper rate damping.

A description of the attitude control modes during each of the nonpowered flight phases of the mission sequence is given below.

Initial rate reduction and stabilization. Following separation from the launch vehicle, the structures supporting the gas jets are deployed. In the present configuration, the yaw jets are located on and are deployed with the low-gain antenna; the pitch and roll jets are located on a deployable boom (Fig. 11).

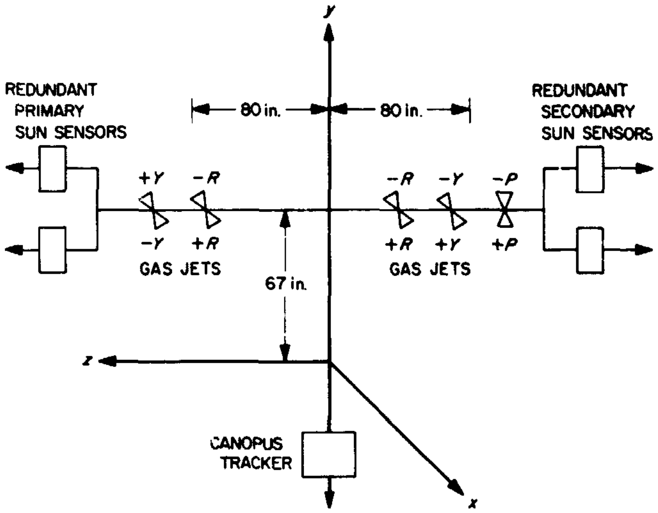


Fig. 11. Gas jet and celestial sensor locations

During this phase, the purpose of the attitude control system is to reduce the initial tumbling rates imparted to the spacecraft at separation to within a controlled rate deadband. The attitude control loop (single axis) during this mode is shown in Fig. 12. Three single-degree-of-freedom high-gain gyros operating in a caged configuration provide rate damping by sensing the components of spacecraft rate about each axis. After the initial rates have been removed, the solar arrays are deployed.

Acquisition. After reduction of the initial spacecraft tumbling rates and deployment of the solar arrays, sun acquisition will begin automatically. The sun sensors, which have a 4π-sr field of view, provide the pitch and yaw position error signals. Redundant sun sensors are employed to improve the reliability of this primary system. The controlled sun acquisition rate corresponds to the saturated output of the sun sensor. The pitch and yaw sun acquisition rates will nominally be 2.0 mrad/s. After acquiring the sun, Canopus acquisition will begin automatically. Upon receipt of the sun gate (sun acquisition signal), a calibrated command current is fed into the roll-

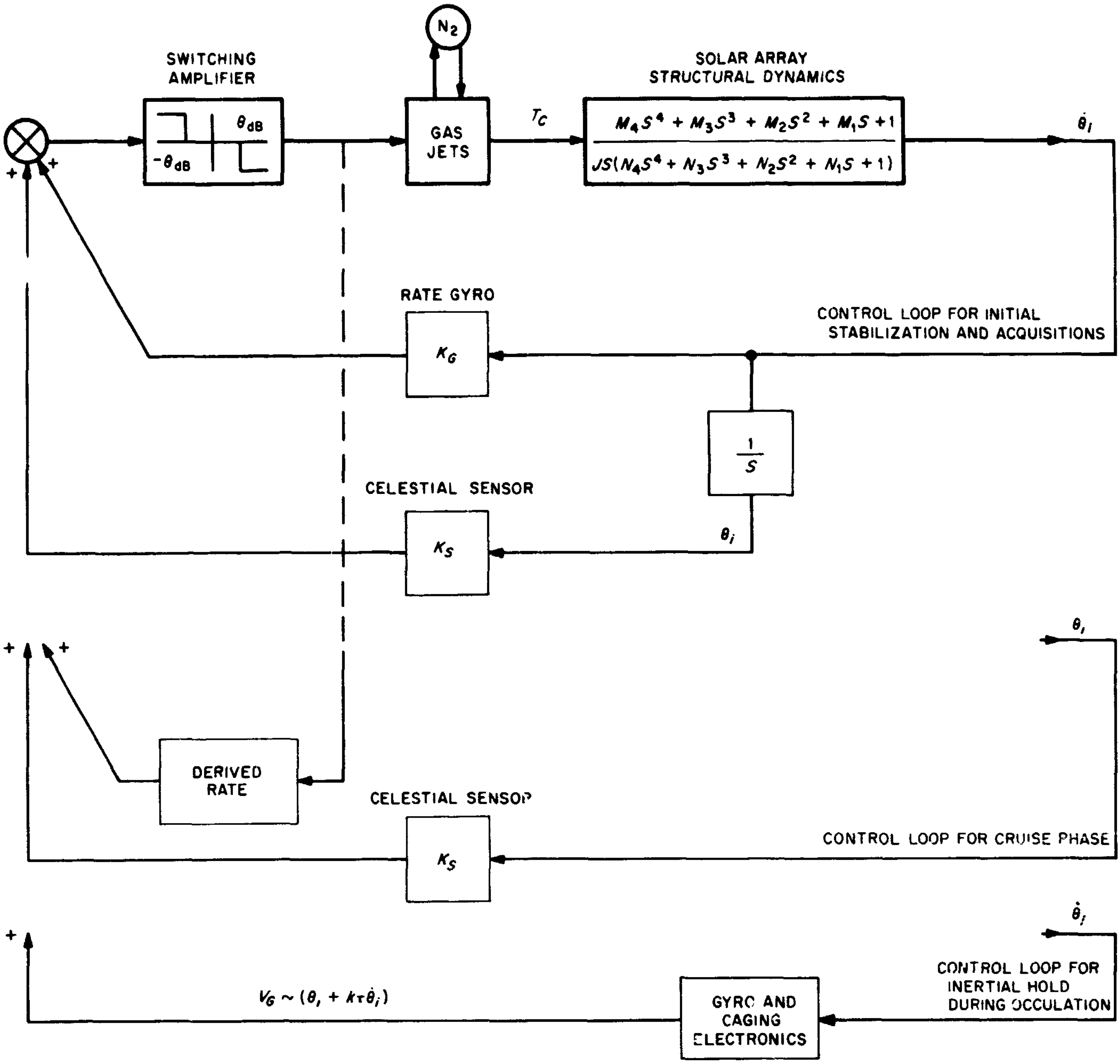

N2
SWITCHING AMPLIFIER
θ_{dB}
$-\theta_{dB}$
GAS JETS
T_C
SOLAR ARRAY STRUCTURAL DYNAMICS
$\frac{M_4S^4 + M_3S^3 + M_2S^2 + M_1S + 1}{JS(N_4S^4 + N_3S^3 + N_2S^2 + N_1S + 1)}$
$\dot{\theta}_I$
+
+
RATE GYRO
K_G
CONTROL LOOP FOR INITIAL STABILIZATION AND ACQUISITIONS
$\frac{1}{S}$
CELESTIAL SENSOR
K_S
θ_i
θ_I
+
+
DERIVED RATE
CELESTIAL SENSOP
K_S
CONTROL LOOP FOR CRUISE PHASE
$\dot{\theta}_I$
+
$V_G \sim (\theta_I + k\tau\dot{\theta}_i)$
GYRC AND CAGING ELECTRONICS
CONTROL LOOP FOR INERTIAL HOLD DURING OCCULATION

Fig. 12. Single-axis attitude control loops

switching amplifier. This signal causes the roll gas jets to fire and accelerate the spacecraft to a rate proportional to the magnitude of the command current. When the gyro feedback signal exactly balances the command current signal, the spacecraft is at roll search rate which is nominally 2 mrad/s. The basic control loop is the same as during separation rate reduction and is shown in Fig. 12. Nominally, acquisition of the sun and Canopus will require no more than 1.5 h.

Powered flight phase. Upon completion of Canopus acquisition, the powered flight phase begins. The attitude control system during this phase and its operation in conjunction with the N_2 cold gas system is described in *Section C.*

Cruise phase. Attitude control during the cruise phase is provided by the cold gas system. The basic system is identical to the *Mariner* system. A block diagram of the cruise attitude control system is shown in Fig. 12. During this phase of the mission, rate damping is provided by derived rate feedback around the switching amplifier. Passive derived rate compensation is used instead of the gyros, primarily to improve the reliability of the system. In addition to the derived rate circuitry, the switching amplifier will incorporate a minimum on-time circuit.

The operation of the circuitry is such that when the celestial sensor output reaches the deadband level of the switching amplifier, the amplifier is switched on for a time equal to the set minimum on-time. At the instant the amplifier turns on, the derived rate output builds up as a ramp function. At the end of the minimum on-time, the derived rate output voltage is large enough to turn the amplifier off and keep it turned off. The result is a stable and known controlled limit cycle. Deadbands of ±0.5 deg are used in all axes.

Periodically throughout the mission, the sun sensor scale factor will be updated (through central computer and sequencer update commands) to counteract the optical gain reduction caused by the decreasing solar energy.

Reacquisition. If loss of acquisition should occur, reacquisition of the celestial references will be performed by the cold gas attitude control system. The system configuration is the same as during initial acquisition and is shown in Fig. 12. When loss of acquisition is detected by the celestial sensor logic, the gyros automatically turn on and the acquisition sequence is initiated. If loss of acquisition occurs during the powered flight phase, the ion engines are first shut down and then the control is switched to the cold gas system.

Encounter. Approximately 10 days before closest approach, the ecliptic plane engine pointing control system (which serves the dual purpose of science platform pointing) is slewed to a nominal science platform pointing position to ensure that the planet will be within the planet tracker field of view. The platform tracks the planet in two axes until approximately 45 min before closest approach. At this time, the gyros are turned on in preparation for the sun occultation mode. The effects of the Jovian radiation belt on optical sensor performance were not evaluated in this study.

Occultation. During occultation, position reference cannot be maintained using the sun and Canopus. Attitude control during this phase will be accomplished by using the gas systems with the gyros in inertial hold. A block diagram of this system is shown in Fig. 12. This method is identical to that used on previous *Mariner* spacecraft during occultation. The *Mariner* gyros are high-gain, narrow-angle, rate-integrating gyros. Attitude control is accomplished by caging the gyro through a capacitor lead network. To ensure that the proper spacecraft attitude and antenna pointing accuracy is maintained, a drift compensation scheme may have to be incorporated in the control system. This, of course, depends on gyro drift and the length of time the spacecraft is in occultation. Upon completion of occultation, reacquisition of the celestial references, if required, is performed in the manner described previously.

Postencounter. The spacecraft remains on gyro inertial control until a sufficient postoccultation time period has elapsed to permit reacquisition of the celestial references without interference (stray light) from Jupiter. At this time, reacquisition of the sun and Canopus occurs in the manner described previously. Since the cold gas storage has been sized for a 1200-day mission, and nominal encounter time is 900 days, the attitude control system will continue operating in the cruise mode for an additional 300 days.

3. Cold Gas Attitude Control System Analysis and Description

During all phases of the mission, other than powered flight, attitude control is provided by the three-axis cold gas system. Figure 11 shows the location of the gas jets

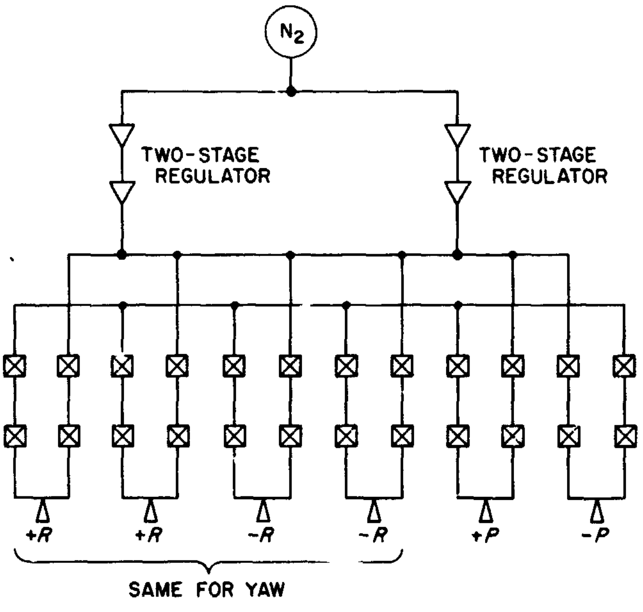


Fig. 13. Quad-redundant valve and gas jets

and their lever arms to the spacecraft center of gravity. The gas jets are located in these positions due to shroud packaging constraints and to eliminate gas impingement on the solar arrays. The gas valves for each control axis are connected in a quad-redundant fashion (Fig. 13). Connecting the valves in this manner provides high reliability and eliminates the requirement for storing additional gas to allow for a valve open failure. In addition, redundant sun sensors are employed to increase optical system reliability, and techniques employing triple redundancy will be incorporated to increase circuit reliability.

The attitude control position deadband θ_{DB}, consistent with attitude pointing accuracy requirements, is set at ±0.5 deg. The gas jet minimum on-time Δt_{on}, which assures a stable and predictable limit cycle, is set at 100 ms. Selection of this value is based on previous experience with this type of attitude control system. Conservative estimates of the spacecraft moments of inertia I_i were determined as

$$I_x = I_y = 14{,}216 \text{ slug-ft}^2$$

$$I_z = 29{,}653 \text{ slug-ft}^2$$

Determination of the gas system thrust level F_i is based on a trade-off between limit cycle behavior, acquisition time, recovery from disturbances, and interaction between the control system and the spacecraft structural dynamics. Ideally, the thrust level is set so that the minimum disturbing torque T_D will cause an ideal soft limit cycle resulting in lower gas consumption and less valve actuations. The minimum T_D is due to solar radiation pressure and will occur at the maximum distance from the sun. For zero rate at one side of the deadband,

$$\left(\frac{\Delta\dot{\theta}_i}{2}\right)^2 = 2\alpha_D(2\theta_{DB}) \qquad (1)$$

$$(\Delta\dot{\theta}_i)^2 = 16\frac{T_D}{I_i}\theta_{DB} \qquad (2)$$

where

$\Delta\dot{\theta}_i$ = minimum rate increment about ith control axis

α_D = angular acceleration due to disturbing torque

Also,

$$\Delta\dot{\theta}_i = \alpha_{ci}\Delta T_{on} \qquad (3)$$

$$= \frac{T_{ci}\Delta T_{on}}{I_i} \qquad (4)$$

$$T_{ci} = F_i L_i \qquad (5)$$

where

α_{ci} = gas jet angular acceleration constant (each axis), $i = x, y, z$

T_{ci} = gas jet control torque about ith axis, $i = x, y, z$

L_i = gas jet lever arm for ith control axis, $i = x, y, z$

Substituting,

$$\left(\frac{F_i L_i \Delta T_{on}}{I_i}\right)^2 = \frac{16 T_D \theta_{DB}}{I_i} \qquad (6)$$

Solving for the thrust level yields

$$F_i = \frac{4(T_D\theta_{DB}I_i)^{1/2}}{L_i\Delta T_{on}} \qquad (7)$$

The minimum disturbing torque due to solar radiation pressure is determined from

$$T_{D_{\min}} = A_p(1+\zeta_R)\left(\frac{P_0}{25}\right)L_D = 1.72\times10^{-6}\text{ ft-lb} \qquad (8)$$

where

A_p = area of one solar array = 380 ft^2

ζ_R = solar array reflectivity coefficient = 0.2

P_0 = solar radiation pressure at 1 AU = 9.72×10^{-8} lb/ft^2

L_D = disturbance torque effective lever arm = 1 ft

This disturbing torque would primarily influence the limit cycle behavior about the pitch and yaw axes since the solar arrays lie in the spacecraft x–y plane. Substituting $T_{D_{\min}}$ into Eq. (7) yields

$$F_x = F_y = 0.09 \text{ lb}$$

$$F_z = 0.15 \text{ lb}$$

The control angular acceleration constants are

$$\alpha_{cx} = \alpha_{cy} = \frac{F_x L_x}{I_x} = 4.3\times10^{-5}\text{ rad/s}^2$$

$$\alpha_{cz} = \frac{F_z L_z}{I_z} = 2.9\times10^{-5}\text{ rad/s}^2$$

Reduction of the separation rates is done prior to solar panel deployment. Due to the much smaller spacecraft inertias, the effective inertias are approximately 60 times greater than after solar panel deployment. For this reason, reduction of worst-case separation rates (3 deg/s) is accomplished in less than 1 min. A suitable value for the search rate $\dot{\theta}_s$ from which sun and Canopus acquisition will occur is 2 mrad/s. For sun acquisition, this corresponds to the saturated output of the pitch and yaw sun sensors. Acquisition of the sun in pitch and yaw will occur simultaneously and will take at the most

$$t = \frac{\pi}{2.0 \times 10^{-3}} = 1570 \text{ s}$$

For Canopus acquisition,

$$t = \frac{2\pi}{2.0 \times 10^{-3}} = 3140 \text{ s}$$

Control system damping during acquisition is determined by the rate to position gain k_T. This gain establishes the proper scaling between the gyro gain and celestial sensor gain. The criterion for determining this gain is as follows: when the celestial reference (sun or Canopus) enters its sensor field of view θ_v, the gyro rate signal must be sufficiently large, relative to the position error signal, to activate the jets having the polarity that will decelerate the spacecraft from the search rate. For example, if $\dot{\theta}_s$ is negative, then the positive gas jets should fire when $\theta = \theta_v$. Referring to Fig. 12, the gas jets fire when

$$\theta_{DB} = \theta_v - k_T \dot{\theta}_s \tag{9}$$

$$k_T = \frac{\theta_v - \theta_{DB}}{\dot{\theta}_s} \tag{10}$$

Substituting the nominal parameter values, $k_T = 40$ s.

Upon completion of the powered flight phase, control is switched back to the gas system in the manner described previously. During the cruise phase, the gyros are off and rate damping is provided by derived rate feedback as shown in Fig. 12. Selection of appropriate derived rate parameters is based on (1) providing a high degree of damping and thereby good reacquisition capability to rate disturbances, and (2) assuring stable minimum impulse limit cycle operation. The derived rate output, when the switching amplifier fires, is

$$\theta_{DR} = K_{DR}\alpha_c (1 - e^{t/\tau_c}) \tag{11}$$

where

K_{DR} = derived rate gain

τ_c = derived rate charge time constant

The derived rate damping factor to rate disturbances is defined as

$$\gamma = \frac{\dot{\theta}_0 - \dot{\theta}_R}{\dot{\theta}_0} \tag{12}$$

where

θ_0 = initial rate disturbance

θ_R = return rate after first excursion out of deadband

To provide good limit cycle performance and reacquisition capability, the derived rate time constants are usually different for the charge (τ_c) and discharge (τ_d) cycles, and the output is clamped at some level lower than the full scale output αK_{DR}. The following relationships can be derived for determining the derived rate parameters. The relation between the limited derived rate output θ_{DRL} and damping γ can be shown to be

$$\theta_{DRL} = (\theta_v - \theta_{DB})(2 - \gamma)\gamma \tag{13}$$

The derived rate gain can be determined from

$$\alpha_c K_{DR} = \frac{\theta_{DRL}}{1 - \exp\left[\left(\frac{-2\tau_d}{\tau_c}\right)\left(\frac{1-\gamma}{2-\gamma}\right)\right]} \tag{14}$$

A lower limit on the discharge time constant can be determined from

$$\tau_{d_{\min}} = \left[\frac{1}{2}\left(\frac{\theta_v}{\theta_{DB}} - 1\right)\right]^{1/2}\left(\frac{2-\gamma}{1-\gamma}\right)\gamma\,(\theta_{DB}/\alpha_c)^{1/2} \tag{15}$$

The above equations can be solved parametrically for different values of γ and τ_d/τ_c. This was done and the results verified through computer simulation. The selected values are

$$\gamma = 0.5$$

$$\theta_{DRL} = \alpha_c K_{DR} = 80 \text{ mrad}$$

$$\tau_c = 50 \text{ s}$$

$$\tau_d = 100 \text{ s}$$

To verify the control system analysis and to investigate the dynamic interaction between the control system and the spacecraft structure, a six-degree-of-freedom computer simulation program was written to assess the overall system performance. In addition, a digital computer program was written to facilitate the analysis of the attitude control gas storage requirements due to the many system iterations that were made in the course of determining the baseline system. The gas system for the baseline system is sized for a 1200-day mission. The required initial gas storage weight is 20 lb.

A future SPS, Vol. III, article will present (1) typical results from the six-degree-of-freedom computer simulation program and from the gas storage analysis computer program, and (2) a description of the various alternate system configurations that were examined during the course of the study.

VII. Guidance and Control Research

GUIDANCE AND CONTROL DIVISION

A. Josephson Junction Memory Elements, *P. V. Mason*

1. Introduction

A previously reported study (SPS 37-44, Vol. IV, p. 57, and SPS 37-46, Vol. IV, p. 97) led to the conclusion that the most useful application of superconducting phenomena on board spacecraft is probably in the area of high-density and/or high-speed computer memory and logic devices.

At present, the superconducting memory closest to actual application utilizes the cryotron, a device based on the superconducting-to-normal transition in a magnetic field. Several laboratories have devoted considerable effort to the development of such memories (Ref. 1), and it now appears that the obstacles to practical use are those ot economics and production rather than of fundamentals.

Another type of cryogenic memory, which should provide extremely short cycle times, is based on the Josephson effect, as described by J. Matisoo (Ref. 2). The primary reason for interest in such a device is the very high switching speed. Matisoo has shown the switching time to be less (probably much less) than 0.8 ns. Since such junctions also lend themselves to batch fabrication by microcircuit techniques, they seem to be very attractive devices for high-speed, high-density, low-cost-per-bit memory.

2. Functional Description

The basic element of such a memory is a junction formed of two superconductors separated by an insulating layer a few tens of angstroms thick. Such a junction can pass current in two different modes: (1) a zero voltage-drop superconducting tunneling mode, and (2) a finite voltage-drop normal tunneling mode. The current vs voltage (I vs V) characteristic of such a junction, taken from a junction made in the laboratory, is shown in Fig. 1. As the current increases from zero, the junction conducts without voltage drop until a critical junction current (I_j) is reached. There is an abrupt transition to the normal conduction mode, with voltage drop about equal to the superconducting energy gap (E_g) of the metal forming the junction. As the current is further increased, the voltage changes little until the line representing the ohmic drop of the normal junction is reached. If the current is now reduced, the junction follows a steep line of low resistance (about 1.4 Ω for the junction shown here) to a low current. A moderately abrupt transition to the superconducting mode then takes place. Reverse current yields identical behavior in the negative region.

The current I_j is, in theory, given by the equation

$$I_j = \pi \frac{E_g}{4R_n}$$

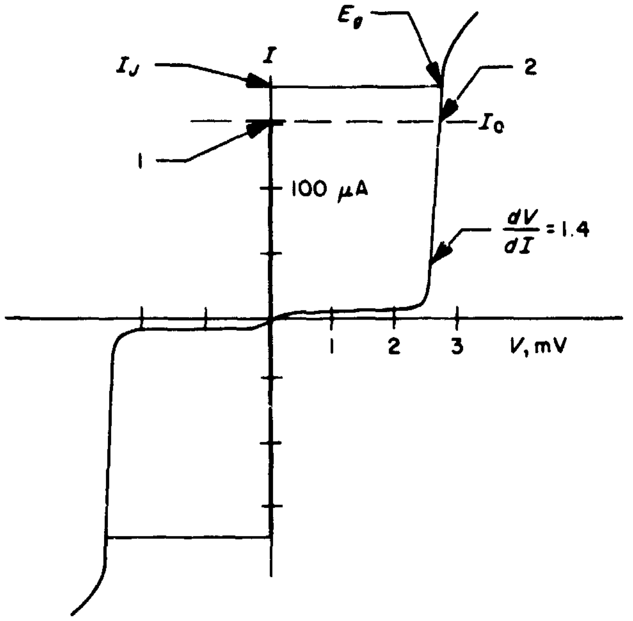


Fig. 1. Current vs v ltage for Josephson junction 72-1

where R_n is the normal resistance of the junction at the operating temperature but a lower value is usually obtained in practice. Also, I_j depends on magnetic field as shown in Fig. 2, which is taken for another of our diodes. The periodic minima occur at fields satisfying the condition

$$\Phi = BA = \frac{2e}{h} n = n\,2.1 \times 10^{-7}\,\mathrm{G\,cm^2}, n = 1, 2, 3, \cdots,$$

where Φ is total flux, B is magnetic field, and A is the cross-sectional area of the function normal to the field.

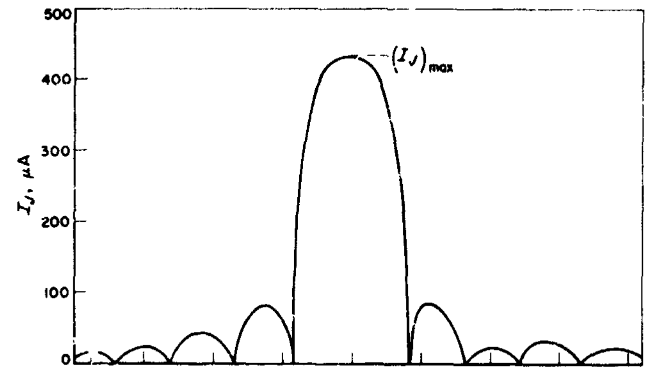


Fig. 2. Dependence of I_j on applied magnetic field—junction 28-1

It is now observed that, if a bias current I_0 that is less than I_j is applied, the junction may be in one of two stable states (indicated by 1 and 2 in Fig. 1). If we are at 1, we may switch to 2 by .pplying a switching pulse I_s such that $I_0 + I_s$ exceeds I_j, and we may switch back to 1 by reducing I_0 to zero as, for example, by applying a pulse $I_s > I_0$ in the reverse direction.

Another, and more useful, means exists to switch the junction from 1 to 2. By n 'ans of an external field (which will usually be generated by a current in a nearby wire), I_j may be reduced below its zero field value $I_{j\,max}$. If we reduce I_j below I_0, the junction must switch to 2. Thus, we have the necessary elements of a three-terminal, bistable device capable of serving as a memory element.

3. Experimental Program

a. Fundamental measurements. Several fundamental questions that arise are as follows:

(1) What is the actual switching speed?

(2) What physical process determines it?

(3) Are we process limited or circuit limited?

(Measurements so far are circuit limited, but there are indications that the fundamental limitation is less than the present measurement limitation of 0.1 ns).

In order to answer these questions, several test samples were fabricated with the junctions in the center of a thin-film superconducting transmission line (see Fig. 3a) in order to reduce effects of the circuit on the measured rise times. However, considerable difficulty was found in obtaining good Josephson characteristics. Because the fabrication in transmission-line form is complicated and therefore slow, it was decided to experiment with fabrication techniques in a simpler crossed-film form (see Fig. 3b). As fabrication methods improve, it will become feasible to return to the measurement of switching speed.

If switching speeds are slow enough to measure with present experimental techniques (about 0.1 ns), an attempt will be made to correlate them with theory. If they are faster, this, of course, will be impossible, but the device applications will be of even more interest.

b. Selection of materials and fabrication techniques. The selection of materials must be based on ease of fabrication in thin-film form, high superconducting transition temperature (in order to minimize cryogenic refrigerator power and weight), and reliability.

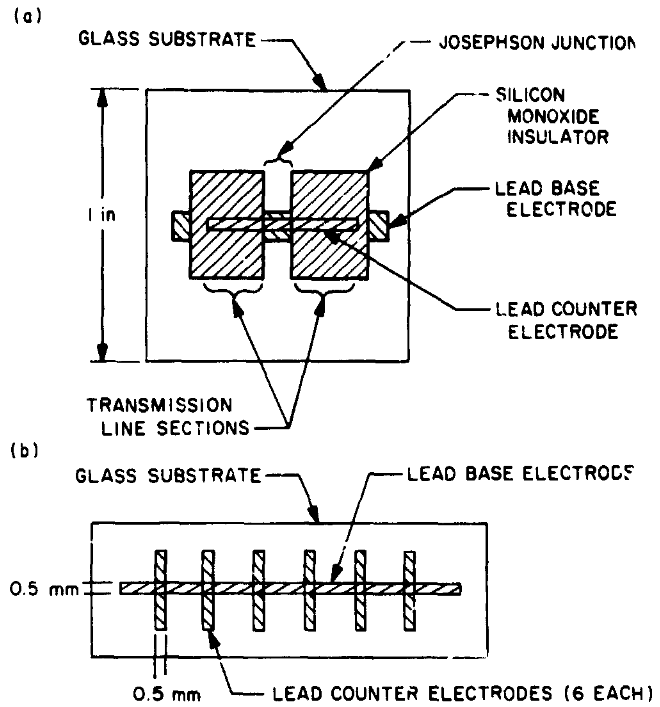


Fig. 3. Test sample configurations: (a) transmission-line test sample; (b) ladder-form test sample

We chose to begin our investigations using vacuum-evaporated lead. Lead has two major advantages. First, its critical temperature (T_c) is 7.2°K, second among the elements only to niobium, whose T_c is 9.2°K. Second, being a low-melting-point material, it is far easier to deposit than niobium. Lead can be easily evaporated from a resistively heated boat at temperatures of 800–900°C, while niobium must be evaporated from a high-power electron gun at 2500–2600°C. Furthermore, the deposition of niobium must be done either in an ultra-high vacuum or onto a heated substrate in order to obtain good superconducting properties. The high T_c of lead has another advantage; since the energy gap is directly proportional to T_c, junctions made of lead have a relatively large output signal (about 2.6 mV) in the normal conducting mode.

The crucial step in the fabrication process is the formation of the insulating film that separates the metal conductors. This film must be, uniformly, a few tens of angstroms thick, free of pinholes that would permit the formation of superconducting bridges between the metal electrodes, and must maintain constant electrical properties in storage and operating environments. Furthermore, the electrical properties must be quite uniform over a large number of samples, since wide variations over the large number of elements would make the memory inoperable.

A number of methods of forming thin insulating films are available. These include the following:

(1) Chemical reactions with the metal (e.g., oxidation or nitridization, both with and without voltage-induced reactions).

(2) Deposition of insulating materials (e.g., silicon monoxide by vacuum evaporation).

(3) Polymerization of organic materials on the surface (e.g., silicone pump oil by electron bombardment).

In general, the simplest method is the oxidation of the metal surface, which is the method most used by those investigating Josephson junctions. Therefore, for this study, the oxidation process was chosen to form the insulating layer. Since oxidation proceeds less rapidly as the oxide becomes thicker, the process tends to be self-healing and self-limiting with time. Thus, pinhole-free films of uniform thickness should be the end result of the process. It was also decided to begin with thermal oxidation, rather than anodization (i.e., voltage-induced oxidation) because of its relative simplicity, although it will probably be necessary to also experiment with anodization.

The process variables under our control are temperature, relative humidity, and time. Presumably, for reasonable ranges of time, the chemical and physical nature of the film should be fairly uniform, and the film thickness should simply increase monotonically. For the temperature range used (20 to 100°C), a logarithmic time dependence and a termination of growth at a few tens of angstroms were expected (Ref. 3).

Likewise, dependence on temperature might be expected to be straightforward, that is, faster rates and thicker films would be produced as the temperature increases. There are possible complications, however, in that lead has a number of oxides, and it is entirely possible that growth of one or the other could be favored at different temperatures.

The dependence on relative humidity is likewise complicated. The production of hydrates at high humidities could (and probably did) lead to poor insulators.

A number of films were made under various conditions of the parameters. Temperatures ranged from room temperature to 100°C with relative humidity from 10 to 80%; time ranged from 6 min to 24 h. It was found possible to form excellent films under nearly all conditions of temperature and humidity by adjusting the time, but reproducibility was very poor. Some general conclusions were: (1) Humidities above 70% usually give poor results, probably due to formation of lead-oxide hydrates. Humidities below 25% take an excessively long time to form an oxide. Between 30 and 50% gives highest yield. (2) Temperatures above 50°C seem to lower the yield, perhaps because of diffusion of lead into the oxide. In this connection, it should be remarked that we have found that it is necessary to store good diodes at liquid-nitrogen temperature in order to preserve their characteristics. This strongly suggests that a temperature-induced diffusion process is at work, and also that it would be very desirable to find an insulator with more stable properties. Anodization is known to produce stable films, especially on the harder materials such as niobium and tantalum. Thus far, room-temperature oxidation has not been explored except to note that it takes an inconveniently long time to form a film.

Films with excellent characteristics, with I_j ranging from a few microamperes to 1.7 mA, have been made. Attempts to produce higher I_j have invariably resulted in pcor characteristics, probably as a result of superconducting bridges across the insulating film.

References

1. Sass, A. R., Stewart, W. C., and Cosentino, L. S., "Cryogenic Random-Access Memories," *IEEE Spect.*, Vol. 4, p. 91, July 1967.
2. Matisoo, J., "The Tunneling Cryotron-A Superconductive Logic Element," *Proc. IEEE*, Vol. 55, p. 172, 1967.
3. Kubaschewski, O., and Hopkins, B. E., *Oxidation of Metals and Alloys*, Second Edition, p. 39, The Academic Press, New York, 1962.

B. Frequency Response of Thin-Film Thermal Detectors, *J. Maserjian*

1. Introduction

The response of a thermal detector to radiant energy is a two-fold process involving, first, the temperature rise resulting from the absorbed radiation and, second, the conversion of the temperature rise into a useful output signal by the active detector element. A new kind of thin-film thermal detector, discussed previously (SPS 37-41, Vol. IV, p. 115; SPS 37-47, Vol. III, p. 44), derives much of its sensitivity from the large temperature response possible in a thin-film structure. For a practical design, it is important to consider this temperature response in detail and, in particular, its dependence on the modulation frequencies of the incident radiation. This article summarizes our analysis of this problem.

2. The Thin-film Structure

The structure under consideration, which closely matches the experimental structure, is a thin film suspended across a hole in a supporting frame that is maintained at a fixed temperature T_0 (Fig. 4 inset). Radiant flux $Q_0 + Q \cos \omega t$, containing a harmonic component Q ($\leq Q_0$) with angular frequency ω, is absorbed over a disc of radius a less than or equal to the radius b of the hole. Only thermal conduction along the film to the frame is considered. The film is assumed to be in an evacuated chamber so that radiation is the only additional mechanism of heat loss; however, this becomes significant only in extreme cases which may then be considered separately. If the film is composed of layers of different materials, one is still free to use effective values for the thermal parameters to describe the composite film. We start with the diffusion equations

$$K\nabla^2 T = \rho c \frac{\partial T}{\partial t} \qquad \text{for } a < r < b$$

$$K\nabla^2 T + Q_0 + Q \cos \omega t = \rho c \frac{\partial T}{\partial t} \qquad \text{for } r < a \tag{1}$$

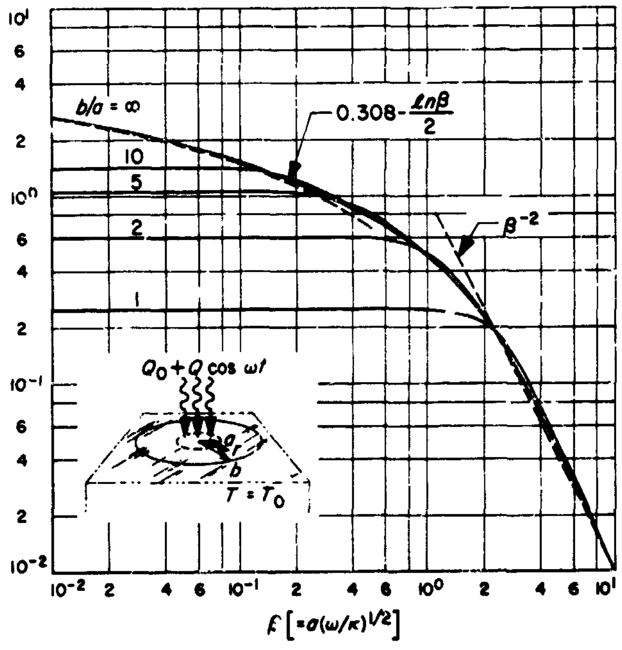


Fig. 4. Frequency response of thin-film thermal detectors

with the boundary condition $T(r = b) = T_0$, where K, ρ, and c are the effective values of the thermal conductivity, density, and specific heat, respectively. We seek the steady-state solution which is obtained at time → ∞ and consists of only the particular solution to the differential equations. The temperature may be assumed uniform throughout the thickness of the film τ, and with the radial symmetry, the Laplacian operator in Eq. (1) reduces to one dimension involving r in cylindrical coordinates. The general solution for arbitrary a and b has been obtained, and the amplitude of the harmonic component $|\Delta T|$ at $r = 0$ is plotted in dimensionless form for several ratios of b/a, where κ is the effective diffusivity of the film ($\kappa = K/\rho c$) and β is the ratio of the radius a to the diffusion length $[(\kappa/\omega)^{1/2}]$.

The general solution, expressed in terms of the tabulated Kelvin functions, is rather cumbersome and will not be reproduced here. However, the solution reduces to a much simpler form for $r = 0$ and $b/a = \infty$, the harmonic component being given by

$$\Delta T(0) = \frac{Q}{\pi K \tau} \frac{kei'\beta}{\beta \cos\theta} \cos(\omega t - \theta)$$

$$\theta = \tan^{-1}\left(\frac{1 + \beta\, ker'\beta}{\beta\, kei'\beta}\right), \qquad (2)$$

where kei' and ker', the first derivatives of the Kelvin functions kei and ker, are tabulated by H. B. Dwight (Ref. 1). The low- and high-frequency asymptotes are plotted as dashed curves in Fig. 4. The solutions for finite ratios of b/a are seen to follow a nearly constant plateau from their low-frequency limit until intersecting the above solution, after which they rapidly merge. The low-frequency limit may be readily calculated for arbitrary $r \leq a$ as follows:

$$\Delta T(r) = \frac{Q}{4\pi K \tau}\left[1 - \left(\frac{r}{a}\right)^2 + 2 \ln \frac{b}{a}\right] \qquad \text{for } \omega = 0 \qquad (3)$$

3. Observations and Conclusions

Some important observations can be made from these results. First of all, the solution is exact for the type of structure considered, and differs significantly from the approximation often made by assuming a fixed value for the thermal relaxation time. In such approximations, the relaxation time is calculated from the ratio of the value of the thermal capacitance of the irradiated region to the value of the low-frequency thermal conductance between this region and the heat sink—the calculation in this case giving $a^2(1 + 2 \ln b/a)4\kappa$. The dependence of the amplitude, in terms of β, then becomes

$$|\Delta T(0)| \approx \frac{Q}{\pi K \tau}\left[\left(\frac{4}{1 + 2 \ln b/a}\right)^2 + \beta^4\right]^{-1/2}$$

which is a fair approximation *only* for the particular case of $b/a = 2$. Thus, one cannot, in general, characterize the thermal response of such a structure by a relaxation time.

Secondly, the curve in Fig. 4, given by Eq. (2) for $b/a = \infty$, may be considered an envelope that essentially encompasses solutions for all finite values of b/a. Therefore, if one wishes to detect radiation modulated at a given frequency (or β), there exists a minimum ratio, b/a, above which one obtains aproximately the same response at this frequency. This minimum ratio b/a corresponds to that curve in which its low-frequency asymptote intersects the envelope at this response. Larger values of b/a add little to the detector's response at this frequency, but may increase the fragility and fabrication difficulties of the detector.

The effects of the constants K, κ, a, and τ are also noteworthy. The response is seen to be inversely proportional to the thickness τ *independent of frequency*. Thus, it is highly advantageous to make the composite film as thin as possible. The mechanical limitations the increase the importance of using the smallest radius b consistent with the operating frequency, as discussed above. The response also appears to be inversely proportional to the thermal conductivity K; however, this is actually true only at low frequencies. At high frequencies, the response approaches the β^{-2} asymptote which, when expressed in terms of the constants, gives

$$|\Delta T| \rightarrow \frac{Q}{\pi a^2 \rho c\, \tau\, \omega} \qquad \text{for } \omega >> a^2/\kappa$$

which is independent of K and depends instead on the product ρc. In this case, the harmonic component of heat is entirely contained in the irradiated region, and the temperature change is determined only by the thermal capacity of the region. Also, the response decreases according to a more rapid $1/f$ dependence in this range; however, this may still be a useful range, particularly when the signal bandwidth is of primary importance, or when excess noise of a $1/f$ dependence is present at low frequencies. The onset of this high-frequency limiting dependence is seen from Fig. 4 to occur at $\beta \approx 2$. If an area of 10^{-2} cm^2 is assumed, this value of β corresponds to frequencies ranging up to about 20 Hz for dielectrics and 200 Hz for metals. If the thermal response of the suspended thin-film detector in this high-frequency limit is compared with that of a thin-film detector in direct contact with an insulating substrate, the advantage is still maintained up to much higher frequencies where the thermal diffusion length becomes comparable to the film thickness ($\gtrsim 10^6$ Hz for 2000 Å film).

Reference

1. Dwight, H. B., *Tables of Integrals and Other Mathematical Data*, pp. 278–279, MacMillan Company, New York, 1957.

C. GaSe Schottky Barrier Gate, *S. Kurtin and C. A. Mead*[1]

The Schottky barrier gate (Ref. 1) is ideal for the construction of field-effect devices since it avoids the difficulties of p–n junction formation, particularly in wide-band-gap materials, and the Schottky barrier depletion layer is not affected by the presence of surface states. A properly-formed Schottky barrier has nearly theoretical reverse current and does not exhibit the drift and instability problems associated with metal-oxide semiconductor structures. Hencc, the Schottky barrier-gate technique can be employed to construct active devices from materials which cannot be otherwise utilized.

GaSe (Refs. 2 and 3) is a layer semiconductor having a 2-eV band gap. A recent study of surface barriers on GaSe (Ref. 4) indicates that the advantages of the Schottky barrier-gate technique will allow the construction of a field-effect device from this material.

Experimental devices were constructed from approximately 8-μm-thick cleaved layers of p-type ($p \sim 10^{14}$/cm^3) GaSe. A schematic cross section appears in the inset of Fig. 5. The source and drain ohmic contacts were alloyed

[1] Performing work supported by JPL at the California Institute of Technology.

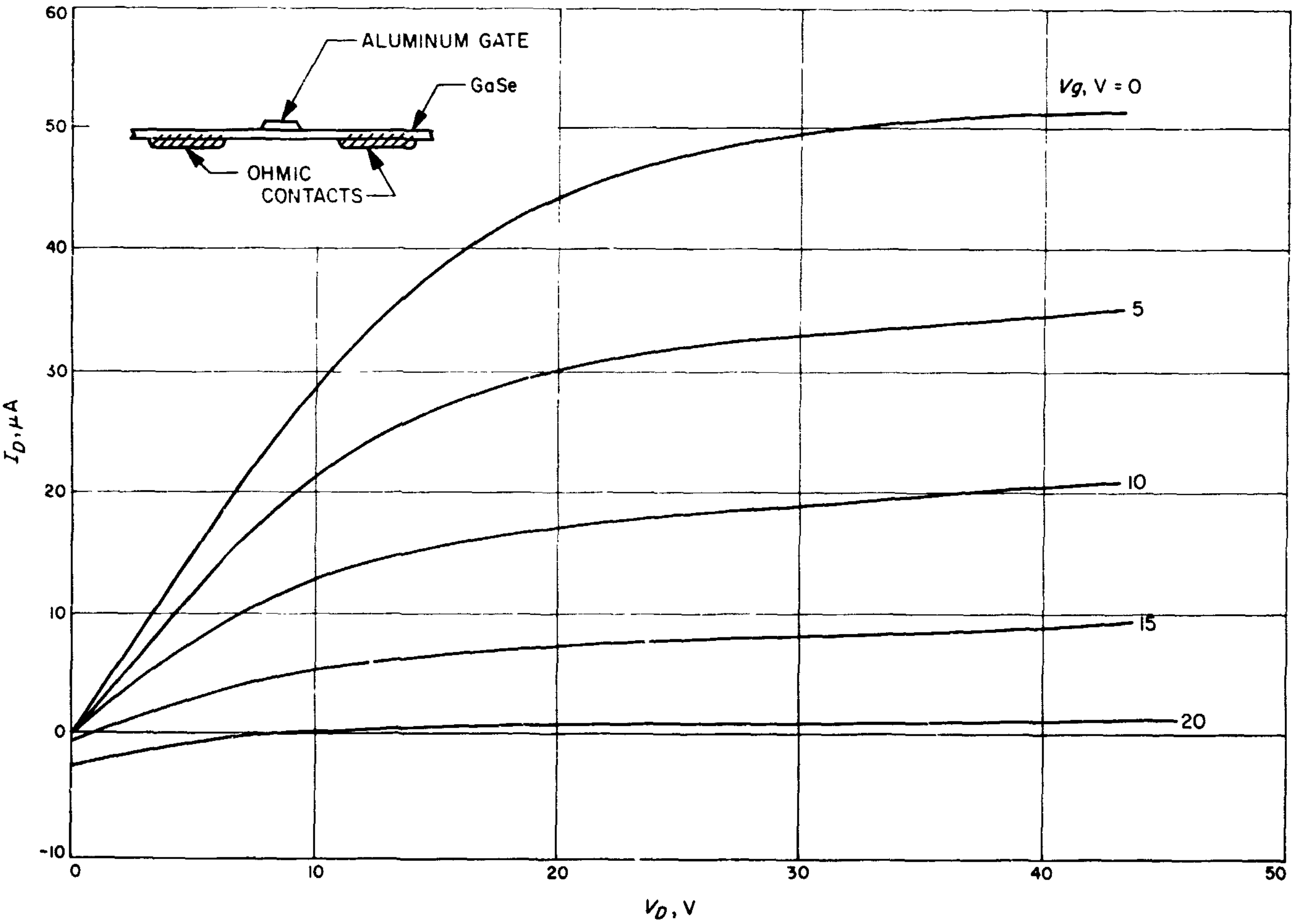


Fig. 5. Electrical characteristics of GaSe Schottky barrier gate

Zn–Au spaced 0.5 mm apart; the width of the device was 3 mm. An aluminum gate, 0.1 mm across, was evaporated directly onto the freshly cleaved surface. Open gate channel resistance was 300 kΩ. The I_{drain}–V_{drain} curves are shown in Fig. 5. Observed transconductance and pinch-off voltage agree well with those calculated for the materials and geometry employed. Channel depth was measured optically, and carrier concentration determined from the capacitance–voltage characteristic of the gate–channel barrier. Note that the zero-bias transconductance is equal to the channel conductance at small drain voltage.

References

1. Mead, C. A., "Schottky Barrier Gate Field Effect Transistor," *Proc. IEEE*, Vol. 54, p. 307, 1966.
2. Fisher, G., and Brebner, J. L., "Electrical Resistivity and Hall Effect of Single Crystals of GaTe and GaSe," *J. Phys. Chem. Solids*, Vol. 23, p. 1363, 1962.
3. Leung, P. C., Andermann, G., Spitzer, W. G., and Mead, C. A., "Dielectric Constants and Infrared Absorption of GaSe," *J. Phys. Chem. Solids*, Vol. 27, p. 849, 1966.
4. Kurtin, S., and Mead, C. A., "Surface Barrier on Layer Semiconductors: GaSe," *J. Phys. Chem. Solids* (in press).

D. Metal Contacts to Photoconductors, *R. J. Stirn*

1. Introduction

Recent developments in the physics of metal–semiconductor contacts indicate that current models of photoconductors may have to be re-evaluated. It has been found that photoconductive gains greater than unity are possible even when the contact is blocking, i.e., when the conduction electrons in the metal are separated from the photoconductor majority carriers by a potential barrier (depletion layer).

All photoelectric devices, in which the injection of carriers is controlled by ohmic or blocking contacts, depend on the metal contact properties. The injection of carriers may give rise to injection luminescence by visible radiative recombination, and, in the case of the illuminated metal–photoconductor contact, the photovoltaic effect shows promising application for large-area solar arrays. The performance of a photoconductor also depends critically on the type of metal contact. Since it now appears that the degree of blocking of "blocking contacts" to a photoconductor is dependent upon the illumination, as well as the photoconductor surface history, further investigations of metal contacts to photoconductors are being carried out at the Laboratory and by Prof. K. W. Böer and his group at the University of Delaware.

In this article, the concept of photoconductive gain and the general model of blocking contacts on lightly-doped semiconductors are briefly reviewed, plus methods for determining the potential barrier height of a metal–semiconductor contact. The results of these methods for various metals on cadmium sulfide (CdS) crystals will be presented in future articles, along with preliminary results from an analysis using stationary high-field domains in the range of negative differential conductivity. The CdS is the photoconductor of greatest interest because of its very high light-to-dark ratio of current, and its sensitivity in the visible region of the spectrum.

2. Photoconductive Gain

For a photoconductor of unit cross-section exposed to a uniform area excitation which generates free electrons at a total rate of F/s, the total number of electrons of charge q in the steady state is

$$N = F\tau \qquad (1)$$

where τ is the electron lifetime. The photocurrent I_p is

$$I_p = Nq/T_r \qquad (2)$$

where T_r is the transit time from the cathode to the anode.[2] For an electrode separation L, applied voltage V, and free carriers with thermal velocity v and mobility μ, the transit time is given by

$$T_r = \frac{L}{v} = \frac{L^2}{\mu V} \qquad (3)$$

Thus

$$I_p = qFG = \frac{qF\mu\tau}{L^2} V \qquad (4)$$

where the photoconductive gain G is defined by

$$G \equiv \frac{\tau}{T_r} = \frac{\mu\tau}{L^2} V \qquad (5)$$

It can be seen that, for a fixed geometry and voltage, the gain is affected by the material parameters τ and μ. Up to this point, it has been assumed that both contacts are ohmic, i.e., the carriers are free to leave and enter the crystal[1] without encountering any potential barrier (due

[1] Since it is CdS that is being considered, a material in which the hole mobility is very much lower than the electron mobility, any hole contribution to the photocurrent will be neglected.

to a high metal work function or to surface states on the photoconductor). In order to allow for possible contact effects, the gain G is now written in an equivalent but more operational sense: G defined by the ratio of the photocurrent to the total number of photons *absorbed* per second multiplied by the electron charge q. Thus

$$G = I_p/aqLW \qquad (6)$$

where W is the width of the crystal, and a is the number of photons absorbed per unit cross-section per second. Measured this way, gains much larger than unity have been reported for CdS with gold contacts,[3] though, as shall be seen, gold is considered to be highly blocking on CdS.

3. Blocking Contacts on Lightly-Doped Semiconductors

When a clean metal surface is brought closer and closer to a clean semiconductor surface while maintaining an electrical circuit between them, the electric field between them, due to the difference in the respective work functions of the materials, induces an electric charge on, or near, the two surfaces. In the semiconductor, this charge can be manifested by (1) a space-charge layer caused by ionized impurity (donor) atoms, and (2) a surface charge induced in surface states. The surface states arise from the termination of the crystal lattice (Tamm states), and possibly from an impurity interfacial layer. The role of these surface states in affecting the barrier energy depends upon their number and relative energy with respect to the Fermi level. In CdS, or other more ionic crystals, Tamm states appear to play a minor role (Ref. 1).

If the surface states are negligible, the barrier height ϕ_B, as seen from the metal side, obtained when the metal surface is in intimate contact with the semiconductor, should be simply

$$\phi_B = \phi_m - E_A \qquad (7)$$

where ϕ_m is the metal work function and E_A is the electron affinity of the semiconductor (Fig. 6). The total potential energy change that the carrier must have in passing through the *depletion* layer (formed when $E_A < \phi_m$) is $\phi_m - \phi_s$, where ϕ_s is the semiconductor work function. This rise in potential energy manifests itself in the semiconductor by the diffusion potential $q\ V_D = \phi_B - \phi_n$, where ϕ_n is the Fermi energy measured from the conduction band edge. The solution of Poisson's equation for the depletion layer (Ref. 2) yields the following relations for λ_0, the thickness of the barrier; E_s, the electric field at the contact; and $\psi_s(x)$, the potential energy in the barrier

$$\lambda_0 = \left[\frac{2\epsilon_0}{qN_D}(\phi_B - \phi_n)\right]^{1/2} \qquad (8)$$

$$E_s = \left[\frac{2qN_D}{\epsilon_0}(\phi_B - \phi_n)\right]^{1/2} \qquad (9)$$

and

$$\psi_r(x) = \frac{qN_D}{2\epsilon_0}(x - \lambda_0)^2 \qquad (10)$$

In these expressions, ϵ_0 is the static–semiconductor dielectric constant and N_D is the ionized donor density assumed to be equal to the total impurity density in the depletion layer. When bias V is applied to the semiconductor, qV is simply added in the parentheses in Eqs. (8) and (9).

An image force correction *lowers* the barrier height, as shown by the dashed line in Fig. 6. The change in barrier height, $\Delta\phi_B$, resulting from this correction and from the application of external voltages, is given by (Ref. 2)

$$\Delta\phi_B = \left[\frac{q^3N_D(\phi_B - \phi_n + qV - kT)}{8\pi^2\epsilon_0\epsilon_\infty^2}\right]^{1/4} \qquad (11)$$

where ϵ_∞ is the high-frequency semiconductor dielectric constant. The term kT arises from the contribution of the mobile carriers to the electric field.

Tunnel penetration of the top of the barrier can be expressed as an apparent lowering of the barrier given by (Ref. 2)

$$\Delta\phi_B = x_cE_s \qquad (12)$$

where E_s is the surface electric field given by Eq. (9) and x_c is a critical tunneling length of about 10^{-7} cm. This last expression gives only an approximate estimate of the actual importance of tunneling.

A correction factor, which would be important if surface states are present, has recently been suggested (Ref. 3) based on a surface-state model by Heine (Ref. 4).

[3] Böer, K. W., and Voss, P., "Light Dependence of an Effective Work Function of Gold Contacts on Photoconducting CdS," to be published.

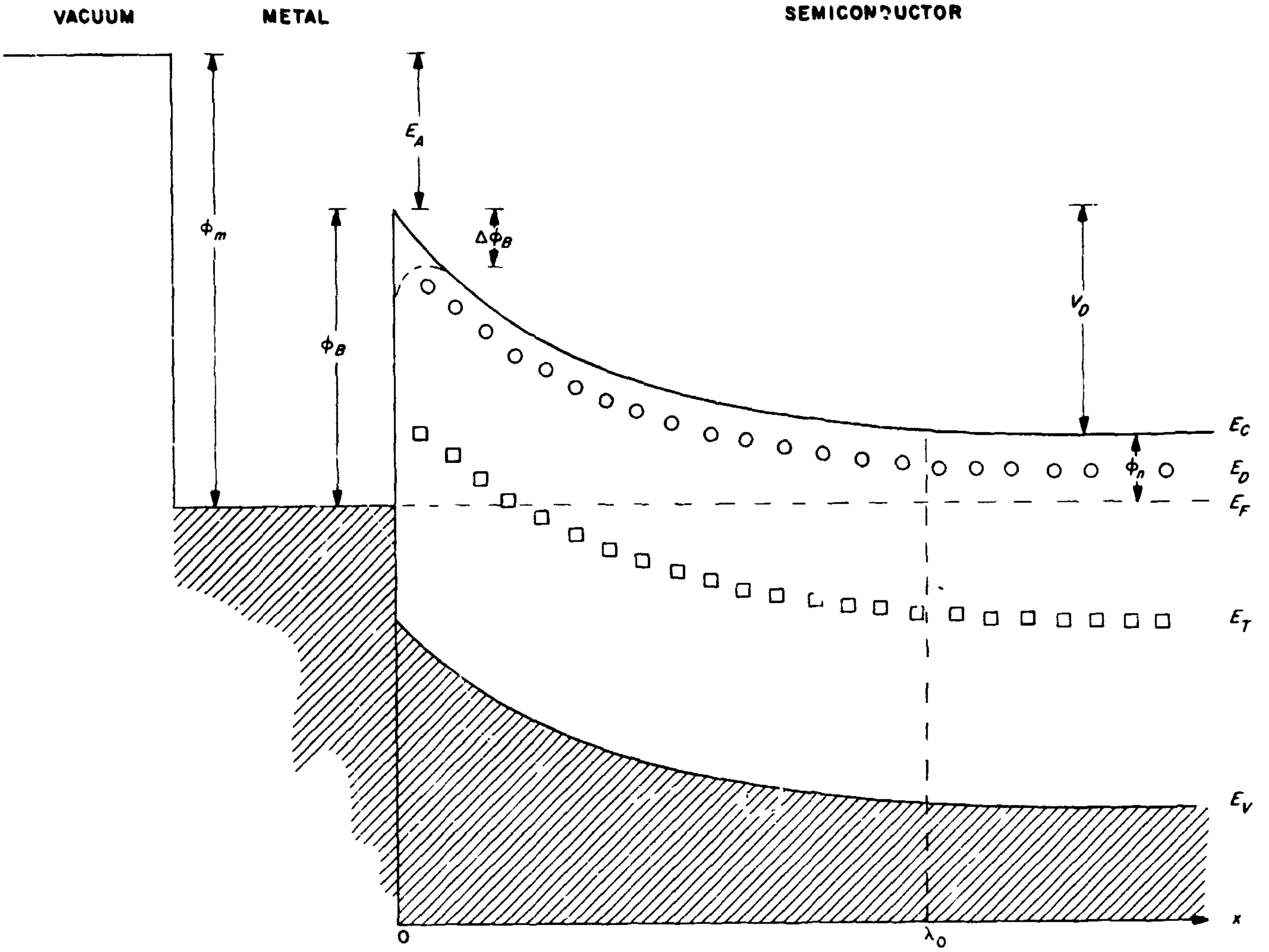


Fig. 6. Schottky model for metal—semiconductor contact with zero applied bias

The average volume charge density of these states can be written

$$\rho = \frac{qN_s}{d} \exp\left(-x/d\right) \qquad (13)$$

where N_s is the number of surface states per unit area, and d the penetration distance of the charge in these states (equal to about 5×10^{-8} cm). If this charge contribution is included in Poisson's equation, the lowering of the barrier height is calculated to be

$$\Delta\phi_B = dE_s \ln\left(qN_s/\epsilon_0 E_s\right) \qquad (14)$$

where E_s is given by Eq. (9).

As seen in Eq. (8), the barrier thickness decreases as the impurity concentration increases. For $N_D \gtrsim 10^{17}$ cm^{-3}, λ_0 is small enough that the barrier presents a finite transparency to electrons with energies lower than ϕ_B; i.e., the electrons can tunnel through at energies near the Fermi level (field emission) and at energies above the Fermi level at temperatures above absolute zero (thermionic field emission). Theories have been developed which satisfactorily account for the observed current–voltage characteristics (Ref. 5). However, in the case of more lightly-doped semiconductors, which includes photoconductors, Schottky barriers are not as well understood for reasons discussed in *Subsection 4*.

4. Techniques and Analyzing Barriers

One important technique used in analyzing barriers (principally GaAs and Si) is the measurement of the $j - V$ characteristic of the contact. It has been found empirically that the current density can be given by (Ref. 6)

$$j = A^*ST^2 \exp\left[-\frac{q\phi_B}{k(T+T_0)}\right]\left\{\exp\left[\frac{qV}{k(T+T_0)}\right]-1\right\} \tag{15}$$

In Eq. (15), the contact area is S, T_0 is a temperature-independent parameter that varies from contact to contact, and A^* is the Richardson constant for the semiconductor modified to take into account the temperature variation of the barrier height. This expression would be identical with the theoretical expression obtained from the so-called diode theory (Ref. 2) for an ideal Schottky barrier if T_0 were identically zero, and if A^* were equal to the Richardson constant $A = 4\pi q m^* k^2/h^3$ (where k is Boltzmann's constant and m^* is the effective mass of the electron). That the two expressions are not the same results from the following observations:

(1) At any voltage and temperature, the experimentally measured forward current (semiconductor positive) is higher than that predicted by the diode theory.

(2) The rate of increase in current with applied bias is smaller than the predicted rate.

(3) The difference between the experimentally measured and theoretically predicted currents becomes larger as the temperature and bias become lower.

(4) The experimental $j - V$ characteristics become independent of temperature at low temperatures.

These last points suggest that quantum mechanical tunneling is present to a much larger degree than expected for a large barrier thickness [derived from Eq. (8) for $N_D \leq 10^{17}$ cm^{-3}]. It is now thought that the actual barrier length λ_0 is much smaller than formerly believed. This could very well be due to the presence of deeper traps lying above the Fermi level in the vicinity of the contact and are thus ionized. These should be included in the value of N_D.[4] Such traps are indicated by the squares in Fig. 6. Since traps are so very important in the II–VI compounds, such as CdS, this point may be quite important in future work on CdS. The effects of these traps can be included, in principle, by capacitance measurements made at very low frequencies, i.e., with periods longer than the trap relaxation times.

The mention of this last type of measurement leads to a second means of investigating the barrier of a metal-semiconductor contact. Since the barrier width λ_0 changes with changing applied bias, the space–charge layer can be represented by an effective capacitance (per unit area) $C = \epsilon_0/\lambda_0 = \epsilon_0(\partial E_s/\partial V)$. From Eqs. (8) or (9), with an applied bias V, we obtain

$$C = \left[\frac{2N_D\epsilon_0}{2(\phi_B - \phi_n + qV)}\right]^{1/2} \tag{16}$$

Thus, a plot of $1/C^2$ vs reverse bias $(-V)$ will give the diffusion potential $qV_D = \phi_B - \phi_n$ as an intercept and the carrier concentration N_D from the slope.

A third, and perhaps the most useful, technique for measuring barrier heights is the measurement of the photoresponse of the barrier. When light is incident upon the contact, either entering from the semiconductor side (back-wall configuration) or through a semi-transparent metal contact, the following two distinct photoexcitation processes can occur:

(1) Photoemission of electrons in the metal over the barrier.

(2) Excitation of carriers in the semiconductor from either band-to-band transitions, or from impurity levels within the forbidden gap.

If one eliminates the possibility of process (2) by choosing wavelengths greater than that corresponding to the band gap, and reducing the amount of light entering the crystal by using the front-wall configuration and a fairly opaque metal contact (but not so thick that the hot electrons cannot reach the interface because of inelastic scattering), one can obtain the barrier height from process (1). For photon energies greater than a few kT above the barrier height, the photocurrent will be proportional to the square of the photon energy, and an extrapolation to zero response will give the energy of the barrier potential.

Using two of these three techniques will, in principal, give self-consistent information about the barrier. The use of these measurements on photoconducting (and, thus, highly insulating) semiconductors entails special problems besides those encountered with normal semiconductors and are not mentioned here. Future articles will go into more detail on these problems in regard to the CdS investigation.

5. Stationary High-Field Domain Analysis in CdS

In addition to one or two of the above techniques, it is hoped that investigations using stationary high-field

[4] F. A. Padovani, private communication, 1968.

domains in the range of negative differential conductivity will be useful for analyzing the metal–photoconductor interface in CdS. This technique, many aspects of which are currently being investigated at the University of Delaware, will be presented in the next article along with data obtained at metal contacts evaporated on air-cleaved CdS crystals. A review of the literature, with regard to experimental determinations of barrier heights on CdS for different metals, will also be presented.

Work is progressing to devise a system to vacuum-cleave crystals of CdS before depositing the metal. Metal contacts made in this manner should not have any interfacial layer, such as an oxide, and thus allow some comparisons to be made between the high-field domain analysis and any of the three techniques discussed in this article.

References

1. Mead, C. A., "Surface States on Semiconductor Crystals; Barriers on the CD(Se:S) System," *Appl. Phys. Lett.*, Vol. 6, p. 103, 1965.
2. Henisch, H. K., *Rectifying Semiconductor Contacts.* Clarendon Press, Oxford, England, 1957.
3. Parker, G. H., McGill, T. C., Mead, C. A., and Hoffman, D., "Electric Field Dependence of GaAs Schottky Barriers," *Solid State Electr.*, Vol. 11, p. 201, 1968.
4. Heine, V., "Theory of Surface States," *Phys. Rev.*, Vol. 138, p. A1689, 1965.
5. Padovani, F. A., and Stratton, R., "Field and Thermionic-Field Emission in Schottky Farriers," *Solid State Electr.*, Vol 9, p. 695, 1966.
6. Padovani, F. A., and Sumner, G. G., "Experimental Study of Gold–Gallium Arsenide Schottky Barriers," *J. Appl. Phys.*, Vol. 36, p. 3744, 1965.

E. Pre-ignition Characteristics of Cesium Thermionic Diodes: Part II, *K. Shimada*

1. Introduction

Pre-ignition volt–ampere curves of thermionic diodes can be divided into two regions: (1) the Boltzmann-type region, and (2) the apparent saturation region (Ref. 1). However, the current through a diode in the apparent saturation region usually does not assume a constant value; it increases slowly as the applied voltage increases. Two separate physical mechanisms are responsible for the increase; they are: (1) a surface effect, and (2) a cesium gas effect (SPS 37-50, Vol. III, pp. 122–125).

This article discusses the pre-ignition characteristics of a diode having an interelectrode distance of 0.0045 in. (compared with 0.028 in. for the diode previously tested). The results are qualitatively consistent with those previously discussed in that the functional dependence of the rate of current increase on emitter temperature and cesium reservoir temperature is similar. However, the rate of current increase in the avalanche region of the volt–ampere curve was noticeably different in the present diode from that of the previous diode. Such a difference seems reasonable since the increase of current in the avalanche region is governed by the volume ionization of cesium atoms, and, hence, by the cesium pressure and the interelectrode distance.

2. Test Diode

The cesium thermionic diode used for this experiment was the SN-107. The emitter and collector, fabricated of rhenium, were assembled in a manner determined to minimize the collection of spurious electrons emitted from the heat-choke area (Fig. 7). The area of the planar part of the emitter disc was 2.00 cm^2; the nominal interelectrode gap was 0.0045 in.

For subsequent analyses of the data, the actual emission area was assumed to be the dimensional area of 2 cm^2. This assumption was justifiable for this diode according to the result of a current measurement that agreed with one obtained from a test vehicle (with guard rings) whose emission area was accurately defined. It should be noted, however, that the agreement was obtained for the ignited mode, and that no data are available for the pre-ignition mode where spurious emission may contribute to the net current.

Currently, a theory is being developed that will enable a calculation to be made of the electron emission from the heat-choke area, which has a temperature gradient and corresponding variations of the work functions. The theory will be cross-checked against the measurements performed in a guard-ring research diode in the near future so that uncertainties due to spurious emission in the present results on the pre-ignition characteristics can be clarified.

No attempt has been made in this article to correct for any spurious emission that may have existed.

3. Pre-ignition Volt–Ampere Curves

The diode under test was operated at relatively low emitter temperatures T_E (1300°K–1600°K) and cesium reservoir temperatures T_{Cs} (453°K–553°K), with the ratios T_E/T_{Cs} being such that the emission was basically

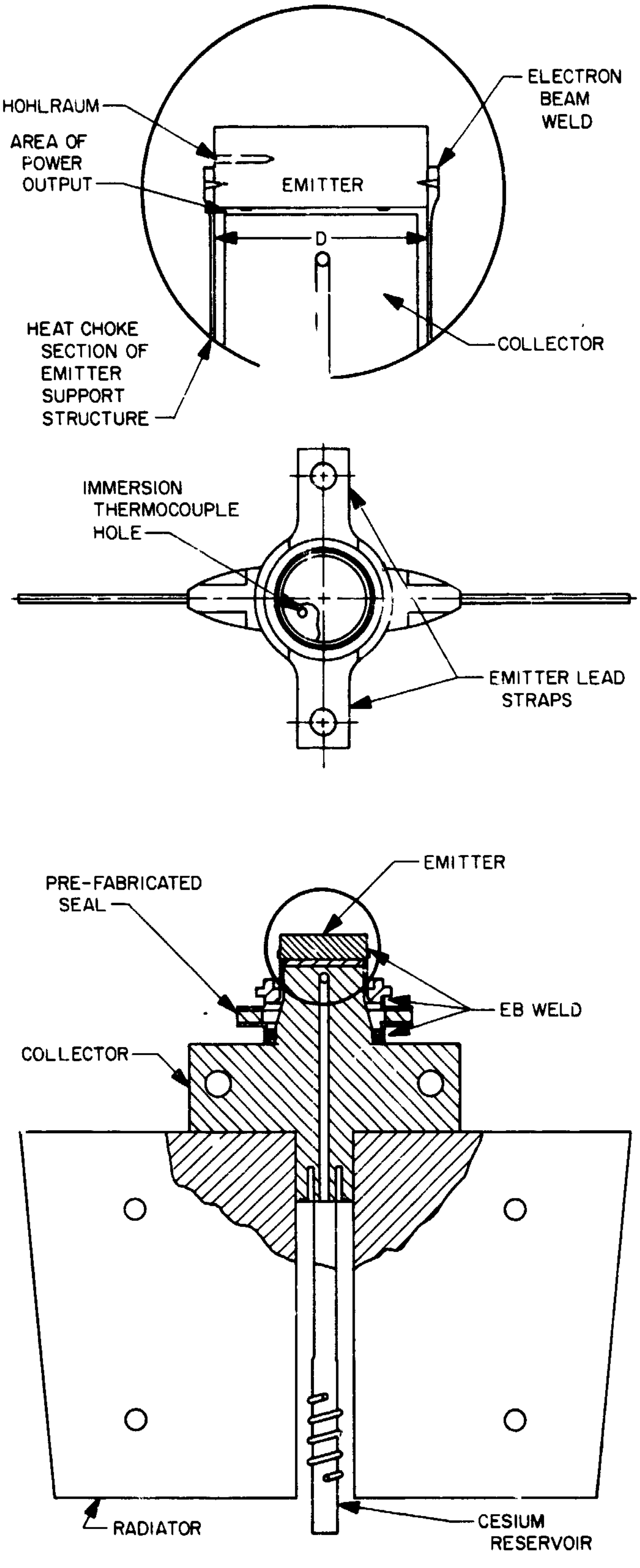


Fig. 7. Test diode

electron-rich (ion-richness ratio $\beta << 1$). Under such conditions, the current through the diode was limited by the electron-space-charge sheath at the emitter. Moreover, the diode was operating in a non-collision-dominated regime since the mean-free-path of electrons ranged between 10 and 0.3 times the interelectrode gap, depending on T_{Cs}.

The volt–ampere curves were obtained by a sampler (SPS 37-49, Vol. III, pp. 130–132), and displayed on linear and semi-log x–y plotters. Simultaneous acquisition of two x-y plots increased the accuracy of current measurements since the semi-log plot showed the Boltzmann region of the volt–ampere curve (where the current is small) in full detail. Typical results are shown in Figs. 8 and 9. In an output-voltage quadrant (negative-voltage part), the current increased sharply with voltage in a Boltzmann-like manner. The Boltzmann-like region is followed by the apparent saturation region in which two sub-regions, the Schottky-like and the avalanche regions, are observed. The rate of current increase in the Schottky-like region is nearly constant for a given emitter temperature, as shown in Fig. 9 where $T_E = 1400°K$. The current increases more rapidly in the avalanche region until the volume ionization in the diode causes ignition. The rate of current increase in the avalanche region depends on the cesium reservoir temperature. To demonstrate the logarithmic dependence of currents on voltages more clearly, the normalized currents I/I_0 (measured current/apparent saturation current) have been plotted against voltage corrected for the contact potential (measured voltage plus emitter work function minus collector work function) as shown in Fig. 10. Three noticeable features are as follows:

(1) All curves exhibit two distinct regions differentiated from each other by the rates of current increase.

(2) The rates of current increase in the Schottky-like region are the same, independent of the cesium reservoir temperature.

(3) All curves converge at the zero voltage (corrected).

Attempts were made to express the normalized currents as a function of voltage by the relation

$$I/I_0 = \exp\{k_1(V - V_1)\} + \exp\{k_2(V - V_2)\} \quad (1)$$

Here I is the measured current at a corrected voltage V, I_0 is the normalization (apparent saturation) current, and

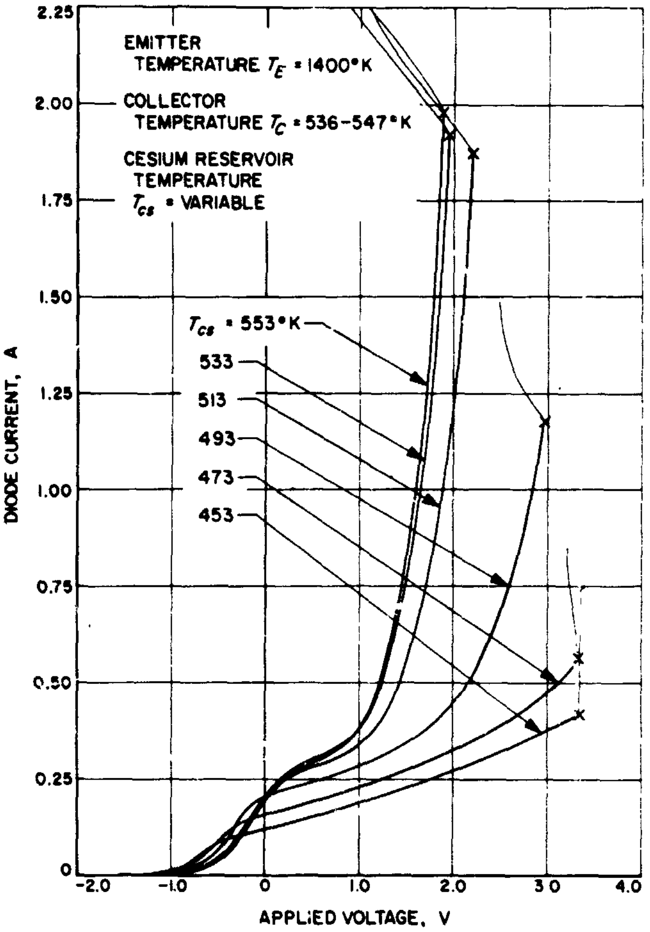


Fig. 8. Typical volt–ampere curves—linear plot

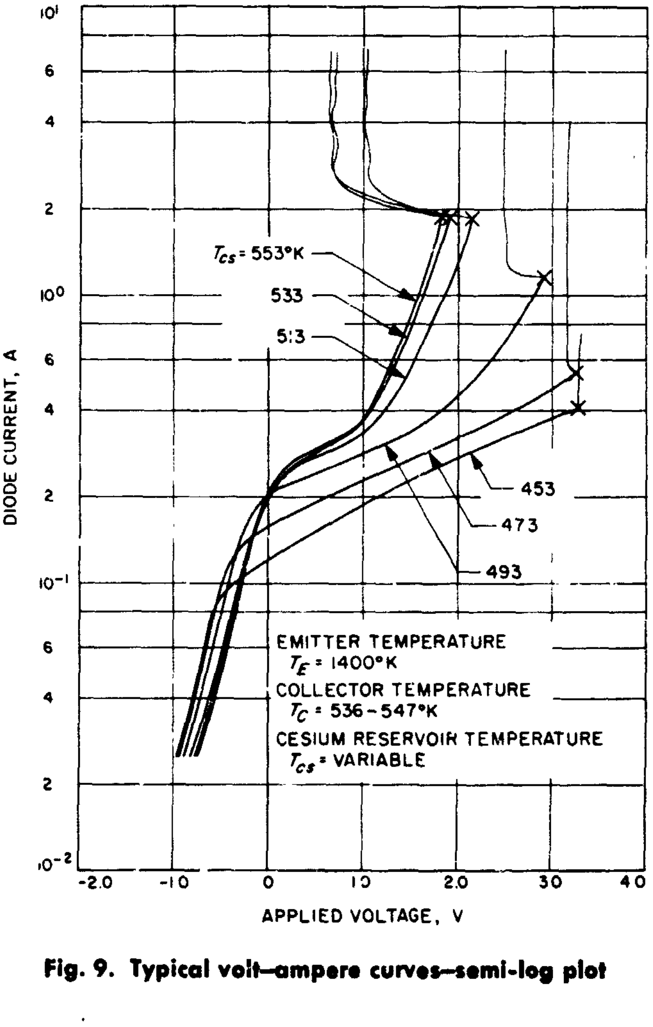


Fig. 9. Typical volt—ampere curves—semi-log plot

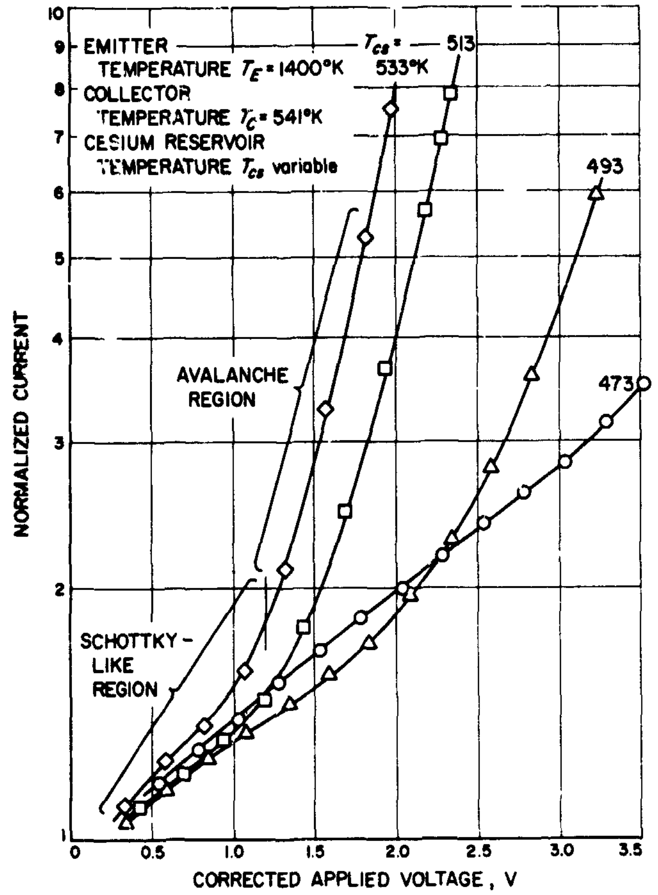


Fig. 10. Normalized current vs corrected applied voltage

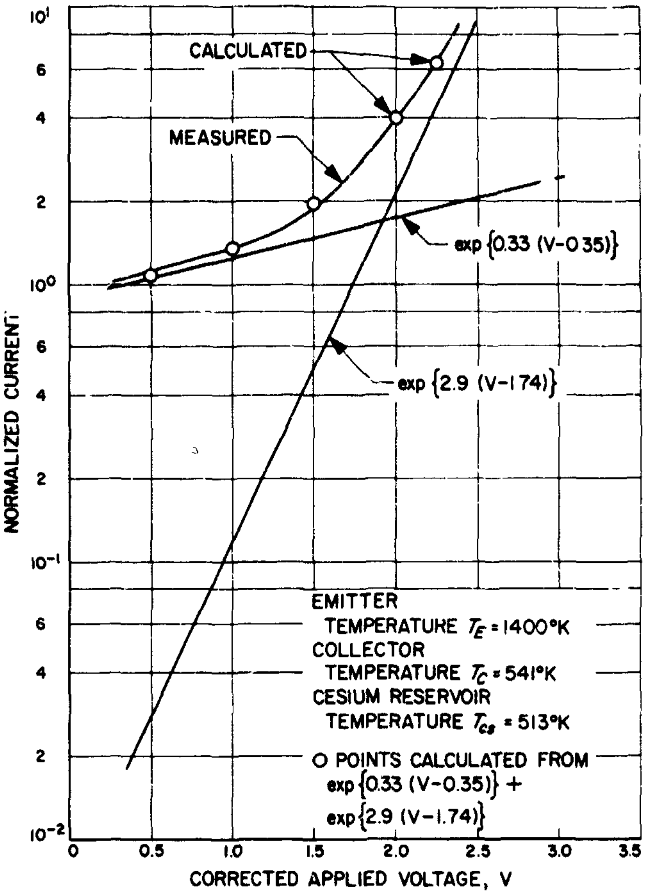


Fig. 11. Comparison of measured and calculated normalized current

I/I_0 is the normalized current. The first term on the right-hand side of Eq. (1) is the contribution by the Schottky-like current, and the second by the avalanche current. For example, at T_E = 1400°K and T_{cs} = 513°K, an empirical expression for the normalized current is

$$I/I_0 = \exp\{0.33\,(V - 0.35)\} + \exp\{2.9\,(V - 1.74)\} \tag{2}$$

The two terms on the right-hand side of Eq. (2), as well as I/I_0, are shown in Fig. 11. The calculated values, indicated by open circles, agreed excellently with the measured values. Matching of the measured I/I_0 with Eq. (1) is now being carried out for all temperatures, and the results will be reported as they become available. Preliminary analyses of k_1 and k_2 yield results consistent with those of previous analyses (SPS 37-50, Vol. III, pp. 122–125). The voltage coefficient k_1 increases from 0.1 to 0.6 as $10^3/T_E$ increases from 0.62 to 0.76, and k_2 is in the range between 3 and 4, but independent of T_E for a given T_{Cs}.

4. Conclusions

Pre-ignition volt–ampere curves for a cesium thermionic diode, operated at relatively low temperatures, exhibit Schottky-like and avalanche regions prior to ignition. The current increases exponentially with the diode voltage at different rates in the two regions. The rate in the Schottky-like region is determined by the emitter temperature and is nearly independent of both the cesium reservoir temperature and the interelectrode gap. On the other hand, the rate in the avalanche region is determined by the cesium reservoir temperature and by the interelectrode gap, but is independent of the emitter temperature. Therefore, it may be concluded that the current through the diode in the Schottky-like region is mainly controlled by the emitter surface effect, whereas the current in the avalanche region is controlled by the cesium gas effect.

These findings should be verified by using a cesium thermionic diode equipped with a guard ring to eliminate spurious currents.

Reference

1. Bullis, R. H., *et al.*, "The Plasma Physics of Thermionic Converters," *IEEE Report on the Thermionic Specialist Conference*, pp. 9–29, Oct. 1965.

F. Thermionic Diode Switch, *S. Luebbers*

1. Introduction

Certain unique characteristics of the thermionic diode allow its application to power switching. To investigate switching feasibility, a test circuit was designed and experiments were performed. The results of these experiments showed a dc-to-ac conversion efficiency in excess of 50%; however, values as high as 85% may be easily reached.

2. Diode Characteristics

The thermionic diode is conventionally employed as a high-temperature power source. Heat energy supplied to the electron-emitting surface (emitter) brings its temperature to incandescence causing it to serve as an efficient electron emitter. The emitted electrons traverse a cesium-vapor-filled interelectrode gap (typically 0.002 to 0.030 in.), and arrive at the collecting electrode (collector) at a higher potential energy than they initially possessed at the emitter. A load connected between the emitter and collector electrodes is supplied this potential energy and transforms it into useable electrical power. Under certain temperature conditions, the *power-generating* diode exhibits dual-mode properties that could be applied to switching; however, the efficiency of this switch would be extremely poor compared with that of a *power-consuming* diode, as will be discussed in this article. A typical volt–ampere characteristic for a power-generating diode is shown in Fig. 12 Two distinct modes of operation are evident—the ignited and unignited modes. If the diode was to be employed as a switch, a resistive load would be placed across the diode via transformer coupling. The optimum load resistance would be that for which the coupled load resistance matched the diode internal impedance. The converter can be switched by short-duration pulses between the ignited and unignited modes (between points A and B of Fig. 12). This change in voltage results in ac power output. The main factor limiting efficiency, with this scheme of power switching, is the small change in voltage incurred in going from the unignited to the ignited mode. (A change of less than 0.5 V is observed in Fig. 12.) In practice, this method of switching would yield efficiencies of the order of a few percent and be extremely sensitive to temperature conditions.

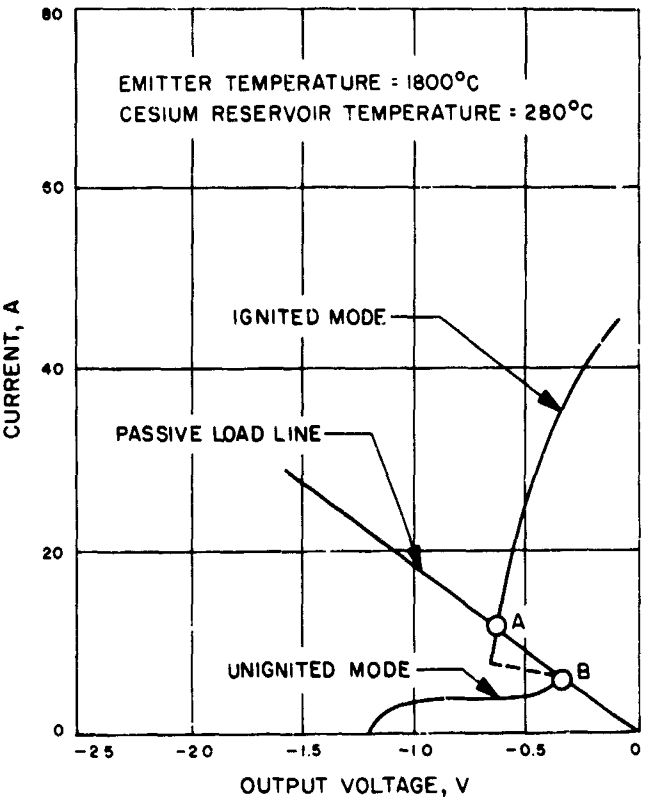


Fig. 12. Power-generating thermionic diode exhibiting dual-mode characteristics

The attractiveness of the thermionic diode for switching, as described herein, is not its capability to produce power, but to act as a passive switching element. When both the emitter and cesium-reservoir temperatures are lowered by approximately a factor of two from the power-producing temperature values, the volt–ampere characteristics shown in Fig. 13 result. The theoretical implications of these characteristics are discussed in SPS 37-44, Vol. IV, p. 59 and Ref. 1, and are shown here in contrast to the power-producing characteristics of Fig. 12. Once again, two distinct modes of operation are observed; however, under the low-temperature conditions, the voltage variation across the diode between the two modes has increased dramatically. This large voltage separation is desirable if an efficient switch is to result.

Figure 14 shows how a thermionic diode might act as a switch. The diode is connected in series with a power

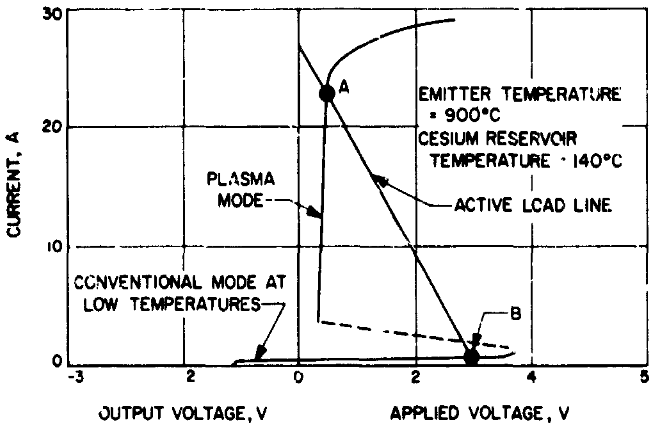


Fig. 13. Power-consuming thermionic diode exhibiting dual-mode properties

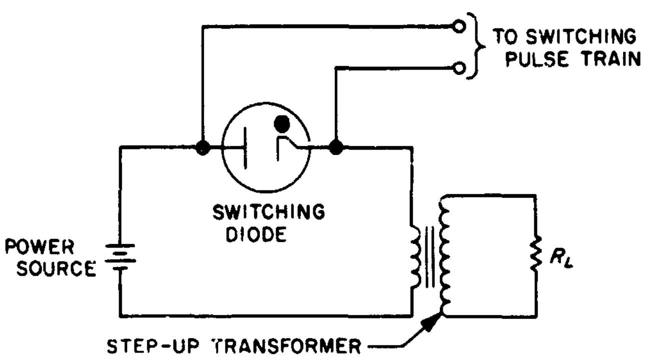


Fig. 14. Test circuit used in switching

source whose voltage is less than the ignition voltage of the passive diode switch, and a step-up transformer is connected to an appropriate load resistance, R_L. A load line for such a circuit arrangement is included in Fig. 13. By a sequence of short-duration positive and negative pulses, the diode operating point is alternately switched between points A and B (shown in Fig. 13), and the resulting variation in voltage appears across the load resistance. For the volt-ampere characteristics of Fig. 13, the maximum switching efficiency (ac power output/dc power available) may be calculated to be 85%. This relatively high efficiency makes the thermionic diode an attractive switching device for use in hostile environments where more conventional low-temperature devices would fail.

3. Design Considerations

The circuit used for the experimental portion of these tests is shown in Fig. 14. Because of its simplicity, only the transformer and the pulsing circuit portions of the circuit require careful design. Since the pulsing circuit operates into a nonlinear load, its optimum design is rather complicated This design was not considered particularly pertinent to the problem of proving switching capability, and, therefore, received little consideration. The pulsing circuit consisted of a free-running multivibrator, operating at 1000 Hz, and two channels of amplification to provide the necessary sequence of positive and negative on–off pulses. These pulses were then applied directly across the switching diode.

Since the ultimate performance of a thermionic diode switch will be greatly influenced by the series step-up transformer, the transformer design received careful attention. Only the high-lights of these considerations will be discussed here.

a. Transformer core selection. To fully utilize the high-temperature (900°C) characteristics of a thermionic diode switch, it should be used in conjunction with a high-temperature transformer. Present technology indicates that the iron–cobalt alloys offer the desired high-temperature capability. Reported Curie temperatures of 900°C, and saturation inductions as high as 23 kG, allow switching and voltage step-up to occur in the immediate vicinity of the power source.

The availability of a conventional (selectron) transformer core dictated its use rather than the high-temperature core specified above. Since the electrical performance of the experimental core was found to be comparable with the characteristics specified for iron–cobalt cores, this substitution of core materials will not significantly detract from the primary objective of proving switching feasibility.

b. Transformer specifications. From the discussion of diode characteristics (*Subsection* 2) and Fig. 13, one would expect the input to the transformer primary to be a square wave of approximately 2.5 V-peak-to-peak amplitude (i.e., the change in voltage in going from point A to B in Fig. 13). For this input voltage V, the product of the primary number turns, N_P, and the core cross-sectional area, A, may be easily calculated from Faraday's law

$$V = -N_P \frac{d\phi}{dt} = -N_P A \frac{dB}{dt} \tag{1}$$

or

$$N_P A \cong \left| \frac{V}{\Delta B/\Delta t} \right| \tag{2}$$

where ϕ is the core magnetic flux, B is the core flux density, and $\Delta t = 1/2$ (ac output frequency)$^{-1}$. If we assume a linearly increasing flux and a frequency of 1000 Hz, the product N_pA is found to be 25 turn-cm^2. The number of primary turns was set equal to 8, thus requiring a core cross-sectional area of approximately 3 cm^2. As a final check of these calculations, the flux build-up within the core was estimated and found to increase too rapidly and cause core saturation. To avoid this effect, the primary inductance was increased by enlarging the core cross-sectional area by a factor of 3. The final specifications for the experimental transformer were as follows:

(1) Core cross-sectional area = 9 cm^2.

(2) Core volume = 63 π cm^3.

(3) Primary turns = 8.

(4) Secondary turns = 88.

These specifications were met by a toroidal core wound with the appropriate gauge wire and number of turns.

Figure 15 shows the experimental transformer efficiency (power out/power in) versus frequency. For the design point of 1000 Hz, a transformer efficiency of 85% is observed in Fig. 15. This transformer efficiency reduces the ideal conversion efficiency from 85% (calculated from Fig. 13) to 72%. Considering the physical construction of the transformer, this efficiency is reasonable.

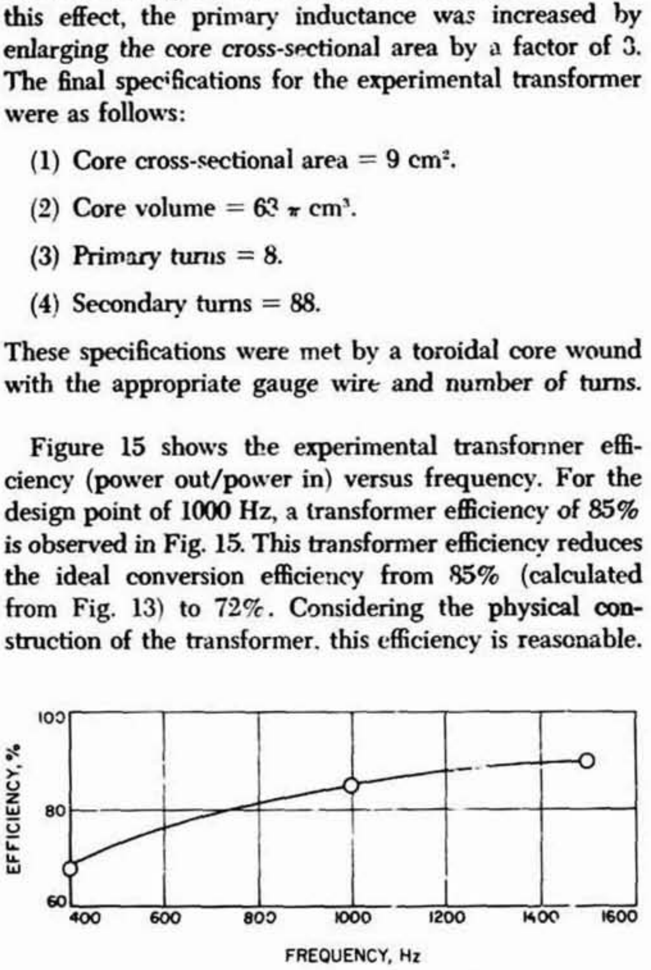


Fig. 15. Transformer efficiency vs frequency

4. Experimental Results

The experimental circuit is shown in Fig. 14 and discussed in *Subsection 3*. The experiments performed on the circuit consisted of the following:

(1) Measuring ac power output versus load resistance for the 400-, 1000-, and 1500-Hz frequencies at fixed emitter and cesium-reservoir temperatures.

(2) Measuring dc-to-ac conversion efficiency versus load resistance at the 400-, 1000-, and 1500-Hz frequencies for fixed emitter and cesium-reservoir temperatures.

(3) Measuring ac power output versus load resistance for several different cesium-reservoir and emitter temperatures (frequency = 1000 Hz).

Each of the above measurements is briefly discussed below.

a. Ac power output versus load resistance. To simulate the low internal resistance and low output voltage of a thermionic generator, a 1.5-V Ag–Zn battery was used as a power source. One would expect this low voltage power source to have poorer performance than the 3-V power source used in earlier calculations since the switching diode voltage drop represents a significant portion of the available 1.5 V. Figure 16 is a typical output-voltage waveform observed across a 30-Ω load resistance. The circuit parameters are also specified in Fig. 16. The dc power source (battery) was ac-modulated by the thermionic diode switch as indicated in Fig. 14. The droop in positive voltage curve, seen in Fig. 16, is caused by nonlinear effects experienced in the transformer core when used in the specified single-ended mode of operation (current passes through the transformer only in one direction). The negative voltage droop is the start of a resistance–inductance (R–L) decay exponential experienced when the thermionic diode is turned off.

A plot of the resulting ac power output versus load resistance is shown in Fig. 17. The 400-Hz data clearly illustrate transformer deficiency, and show that this is not a desirable operating frequency. The upper curves

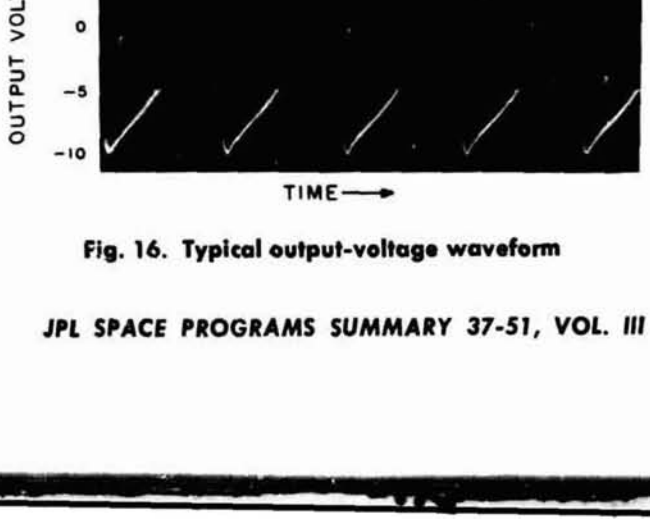


Fig. 16. Typical output-voltage waveform

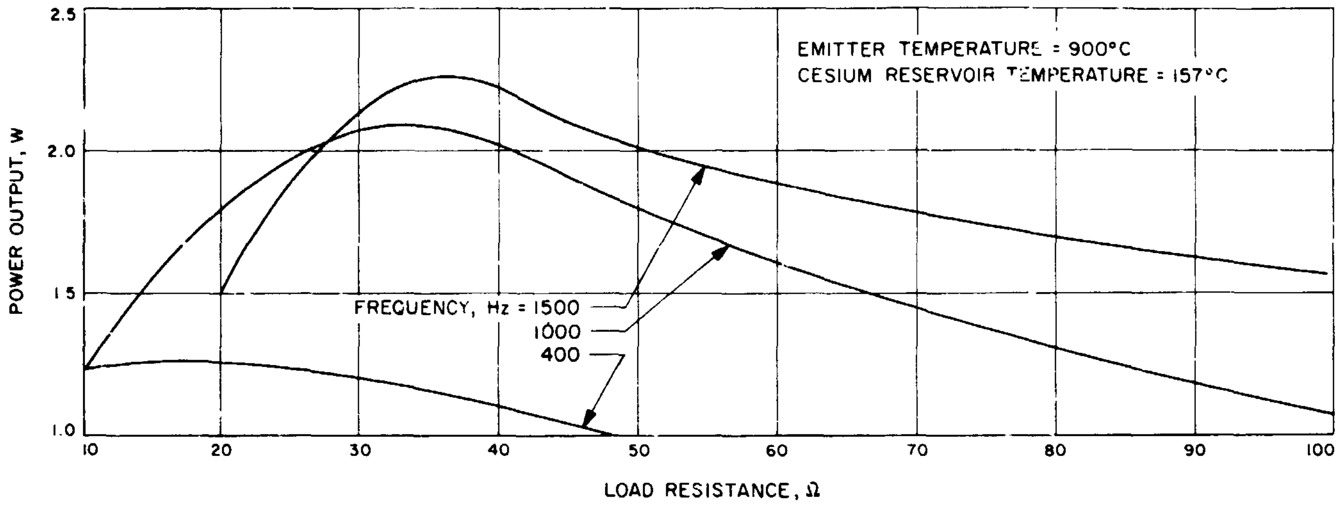


Fig. 17. Power output vs load resistance

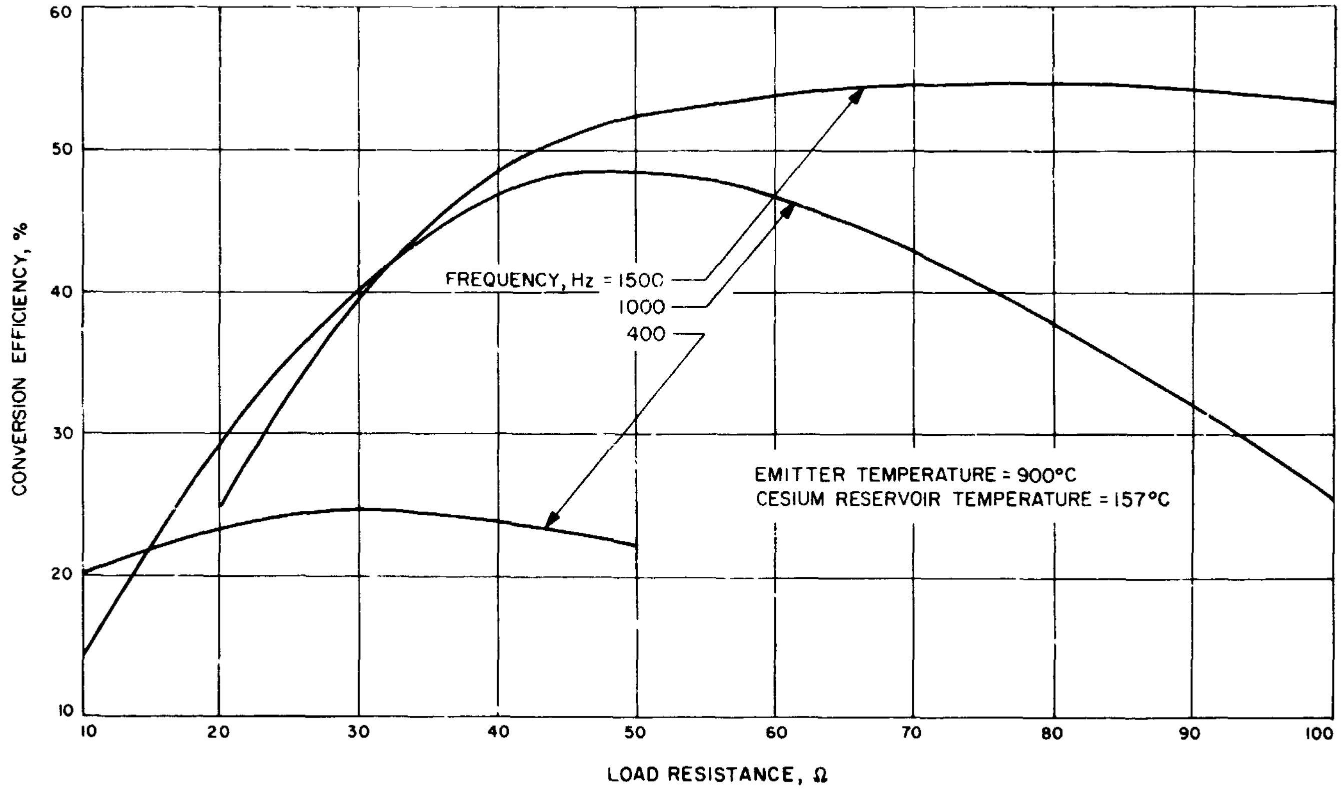


Fig. 18. Conversion efficiency vs load resistance

exhibit the anticipated behavior with falloff at both high and low values of load resistance. The low-resistance falloff may be attributed to resistance mismatch between the load and effective source internal resistance. The high resistance falloff includes both load mismatch and some transformer saturation. This falloff is common to all of the curves obtained in these experiments.

b. Dc-to-ac conversion efficiency. Figure 18 is a plot of the conversion efficiency versus load resistance for three different frequencies. Again, the 400-Hz data are low, as anticipated, and the 1000- and 1500-Hz data are comparable up to a load resistance of 50 Ω. At higher values of load resistance, transformer core saturation becomes evident in the 1000-Hz data. Conversion efficiencies of approximately 50% are observed. If a higher voltage power source were used, the efficiency would increase.

*c. **Ac power output versus load resistance.*** Ac power output versus load resistance, with the cesium-reservoir temperature as a parameter, is plotted in Fig. 19. In contrast to those of Fig. 17, these data were taken at a fixed frequency of 1000 Hz and an emitter temperature of 1100°C. The effect of increasing the cesium-reservoir temperature is that the diode's internal resistance decreases, and the output power increases. As would be expected, the shift to a lower optimum load resistance is also accompanied by an increased power output.

5. Summary

The thermionic diode has been successfully used to ac-modulate the power output from a 1.5-Vdc power source with a conversion efficiency in excess of 50%. A maximum efficiency of 85% is predicted for a 3-Vdc power source.

These relatively high efficiencies make the thermionic diode an attractive switching device for high-temperature applications as, for example, in a thermionic nuclear reactor. The high-temperature and radiation-resistant properties of the thermionic switching diode would permit its location in the immediate vicinity of the nuclear power source, thereby reducing power lost to the current-carrying conductors. Conventional semiconductor power components would have to be located at a remote position where temperature and radiation levels would be tolerable.

Reference

1. Shimada, K., and Luebbers, S., "Anomalous Electron and Ion Currents in Plasma-Mode Operation of a Thermionic Energy Converter," in *Advances in Energy Conversion Engineering*, ASME Conference, 1967.

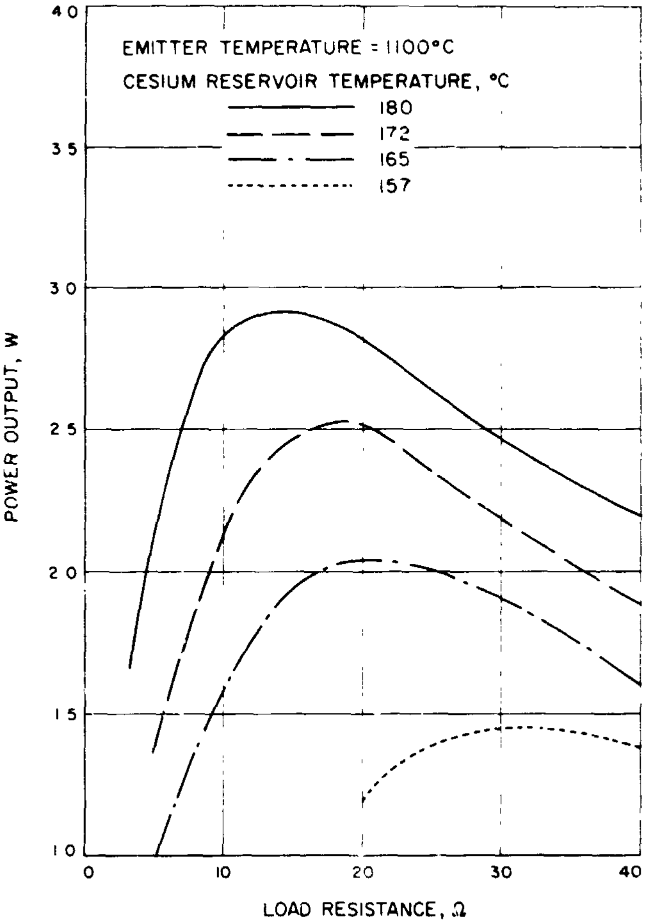


Fig. 19. Power output vs load resistance

VIII. Materials

ENGINEERING MECHANICS DIVISION

A. Effect of Notch Severity on Cross-Rolled Beryllium Sheet, R. Moss

1. Introduction

Brittle materials such as beryllium (Be) are considered notch sensitive. Presence of a sharp notch is believed to reduce the material strength and ductility so greatly that it is of questionable value in structural applications. Unfortunately, almost no work has been done to demonstrate quantitatively the effect of machined notches on the strength of cross-rolled Be sheet as a function of material variables, or notch severity; in particular, the effect of sharp notches has not been examined in any detail. Some data does exist on the effects of relatively dull notches in hot pressed block. Previous data on hot-pressed Be block showed an increase in notched/unnotched strength for a stress concentration factor (K_t) $\approx$ 3 to 4, and a reduction for K_t between 3 and 5 (Refs. 1–7). Some of the references give conflicting results for K_t between 3 and 4. Existing data on cross-rolled sheet show no notch strengthening even at $K_t < 2$ (Ref. 8). The series of tests reported here was intended to determine whether material other than hot-pressed block would show any notch strengthening at $K_t < 5$, and what the effects of different K_t, process history, and composition had on the transition from strengthening to weakening of Be.

2. Test Results and Discussion

This article presents preliminary data on the effect of sharp notches in cross-rolled ingot and powder sheet. Early results indicate that the expected severe reduction of notched/unnotched strength did not occur in the materials and sample configuration studied. Vendor analyses and properties of these materials are given in Table 1. Additional tests on a second grade of ingot sheet and two more grades of powder sheet are in progress.

Samples tested were double-edge-notched sheet tensile specimens 0.025 in. thick, ¼ in. wide, with a ¾-in. gauge length This represents a sheet thickness of interest in spacecraft applications. After rough blanking, 0.002 in. was etched from each specimen surface to remove microcracks and surface damage. Notches having a severity of K_t = 3.2 to 8.3 then were formed by electrical discharge machining. This K_t range spans the region in which notch sensitivity should be apparent. Both longitudinal

Table 1. Vendor-reported properties of beryllium

Property	Powder sheet HR-379	Ingot sheet IS-318
Composition, %		
BeO	1.58	0.32
C	0.100	0.064
Fe	0.092	0.114
Al	0.056	0.057
Mg	0.005	0.004
Si	0.048	0.064
Other metals	0.04 (max)	0.04 (max)
Be assay, %	98.48	99.41
Grain size, μm	60	60
Tensile strength (longitudinal), lb/in.^{2}	80,500	49,400
Tensile strength (transverse), lb/in.^{2}	79,300	63,700
Yield strength (longitudinal), 0.2%	55,000	41,200
Yield strength (transverse), 0.2%	55,700	47,500
Elongation (longitudinal), %	22.0	3.0
Elongatior (transverse), %	16.0	3.0

and transverse samples were tested. Tensile tests were run at a constant crosshead rate of 0.05 in./in./min, with results recorded directly on an x–y plotter.

Results are shown in Fig. 1. It is apparert that the expected severe loss of strength at $K_t \lesssim 4$ did not occur in powder sheet, or longitudinal samples of ingot sheet; indeed, a slight trend toward strengthening seems to be present at $K_t \leq 6$ for longitudinal samples. The trend for transverse ingot sheet samples was in the direction of reduced notched/unnotched strength ratio. Data scatter is too great to justify drawing simple curves, so scatter bands are shown. This scatter is to be expected for a brittle material such as Be; it is not unreasonable considering the normal scatter of unnotched cross-rolled powder sheet is ±4.5% (3-σ level, 90% probability, 95% confidence) (Ref. 9). Actual K_t was calculated for each sample, using measured dimensions and the standard nomographs (Refs. 10 and 11).

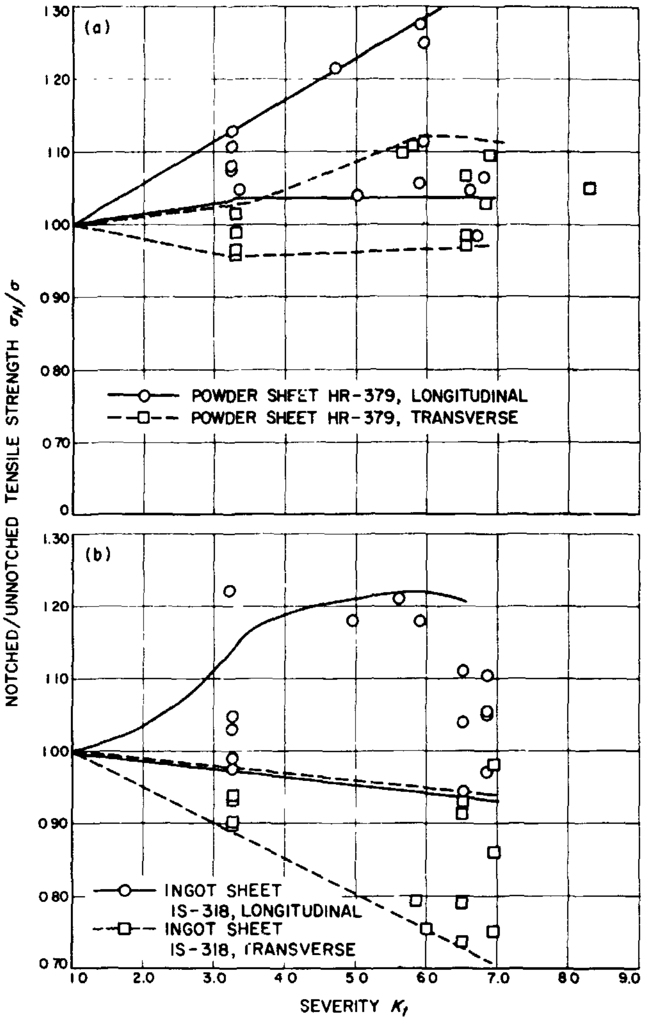


Fig. 1. Effect of notch severity and testing direction on the notched/unnotched tensile strength of Be: (a) powder sheet, (b) ingot sheet

There are several possible explanations for the differences between this data and previous notched/unnotched tensile test results. Most of the existing data was obtained from hot-pressed block. Sheet properties are significantly different from hot-pressed block in regard to strength, elongation, and anisotropy of mechanical properties. Another possible cause of reported notch sensitivity is the presence of machining damage on the sample surface. Some of the early work (Refs. 2 and 3) was done before the need for post-machining etching was established; other reports did not describe sample preparation (Refs. 5 and 6). It is likely that these samples contained microcracks, giving much higher effective K_t values than those measured and reported. Avoidance of surface cracks was a major objective of sample preparation in this program. Therefore, it is believed these results may be more representative of notch effects in samples without microcracks or twins. Sample geometry should be considered also. It is possible that another sample geometry would give different results.

The reason for the directionality of ingot sheet notch sensitivity is somewhat puzzling. One possible explanation is the well-known anisotropy of Be sheet mechanical properties. Reported grain size of this sheet is < 60 μm. A large amount of rolling is required in order to obtain fine grain sizes in ingot sheet. This might have introduced severe texturing, making the sheet more susceptible to crack propagation in one direction than the other. X-ray diffraction studies revealed appreciable texturing in the ingot sheet. Comparison with the powder sheet is in progress.

3. Conclusions

Although the test data obtained are preliminary and subject to further verification, results suggest that the presence of sharp notches in Be sheet need not cause catastrophic failure. Ingot sheet was weakened in the transverse direction, but was not weakened significantly in the longitudinal direction. It would be misleading to suggest that rolled Be sheet is not notch sensitive; however, in structures which use thin gauges of Be, there seems to be more tolerance for defects than generally anticipated. Similar resistance to crack propagation in 0.051-in. cross-rolled powder sheet was reported by others (Ref. 12).

References

1. Fellman, R. B., et al., *Final Report, Development of High Strength Beryllium Materials for Structural Applications*, Vol. 1, Report 675D519. General Electric Company, Re-entry Systems Division, Los Angeles, Calif., Feb. 1967.
2. Crawford, R. F., and Burns, A. B., *Strength, Efficiency, and Design Data for Beryllium Structures*, ASD-TR-61-692, AD290770. Lockheed Aircraft Corporation, Sunnyvale, Calif., Feb. 1962.
3. Hodge, W., *Beryllium for Structural Applications*, DMIC Report 168. Battelle Memorial Institute, Columbus, Ohio. May 18, 1962.
4. Kesterson, R. L., *The Cryogenic and Ambient Tensile and Compression Properties of Hot-Pressed Block Beryllium*, WANL-TME-1619, N68-13893. Westinghouse Astronuclear Laboratory, Large, Pa., June 1967.
5. *Beryllium Thermal Shock Testing*, preliminary report to NASA Research Advisory Committee. Westinghouse Astronuclear Laboratory, Large, Pa., Jan. 1967.
6. *Beryllium Fracture Mechanics*, preliminary report to NASA Research Advisory Committee. Westinghouse Astronuclear Laboratory, Large, Pa., Jan. 1967.
7. Campbell, J. E., *Mechanical Properties of Beryllium at Cryogenic Temperatures, Including Notch-Specimen Data*, DMIC Technical Note. Battelle Memorial Institute, Columbus, Ohio, Nov. 5, 1965.
8. Finn, J. M., Koch, L. C., and Muehlberger, D. E., *Design, Fabrication, and Test of an Aerospace Plane Beryllium Wing-Box*, AFFDL TR-67-38. McDonnell Douglas Corporation, St. Louis, Mo., Mar. 1967.
9. King, B., *New Grades of Beryllium—Their Meaning and Use*, paper presented at SAE Manufacturing Forum, Los Angeles, Oct. 2–6, 1967. Brush Beryllium Company, Cleveland, Ohio.
10. Peterson, R. E., *Stress Concentration Design Factors*. John Wiley & Sons, New York, N. Y., 1953.
11. Neuber, H., *Theory of Notch Stresses*. Translation published by J. W. Edwards Co., Ann Arbor, Mich., 1946.
12. Finn, J. M., Koch, L. C., and Muehlberger, D. E., *Design, Fabrication, and Ground Testing of the F-4 Beryllium Rudder*, AFFDL-TR-67-68. McDonnell Douglas Corporation, St. Louis, Mo., Apr. 1967.

IX. Aerodynamic Facilities

ENVIRONMENTAL SCIENCES DIVISION

A. Heat Transfer Study of 60-deg Half-Angle Cones, *M. F. Blair*

With the objective of determining the applicability of presently available theories of laminar convective heat transfer in planetary gases, a study of 60-deg half-angle cones is being carried out in the JPL 43-in. hypersonic shock tunnel and the JPL 12-in. free-piston shock tube. The bodies under investigation are three 60-deg half-angle blunted cones (Fig. 1) with various edge radii. All three cones have a bluntness ratio R_n/D of 0.10 while the shoulder radius/body diameter ratios R_s/D are 0.05, 0.025, and sharp. Heat transfer distributions are currently being measured and will eventually be compared to values predicted by using measured pressure distributions as input to a convective heat transfer computer program.

Measurement of the pressure distributions, carried out entirely in the JPL 43-in. hypersonic shock tunnel, has been completed. This tunnel is driven by a 3-in. inside diameter shock tube which is operated in the reflected mode (tailored interface). The shock tube driver gas for all cases was H_2 while the driven gas, and hence the tunnel working medium, was N_2. The working section Mach number was about 12.5 while the total enthalpy was approximately 1800 Btu/lbm. The flow Reynolds number was about 4.2×10^4/ft.

The pressure study consisted of measurements at 45-deg increments of roll angle for the following angles of attack; $\alpha = 0$, 5, 10, 15, 20, and 30 deg. Samples of the data obtained for the body of Fig. 1 at $\alpha = 0$ deg are shown in Fig. 2, and for the body pitched to $\alpha = 15$ deg in Fig. 3. The symbols represent the numerical average of data taken while the error bars show the extremes of data taken from all runs. Average points represent results from two to four tunnel runs. Also presented (Fig. 4) is a diagram of isobars that resulted from radially cross-plotting the curves of Fig. 3.

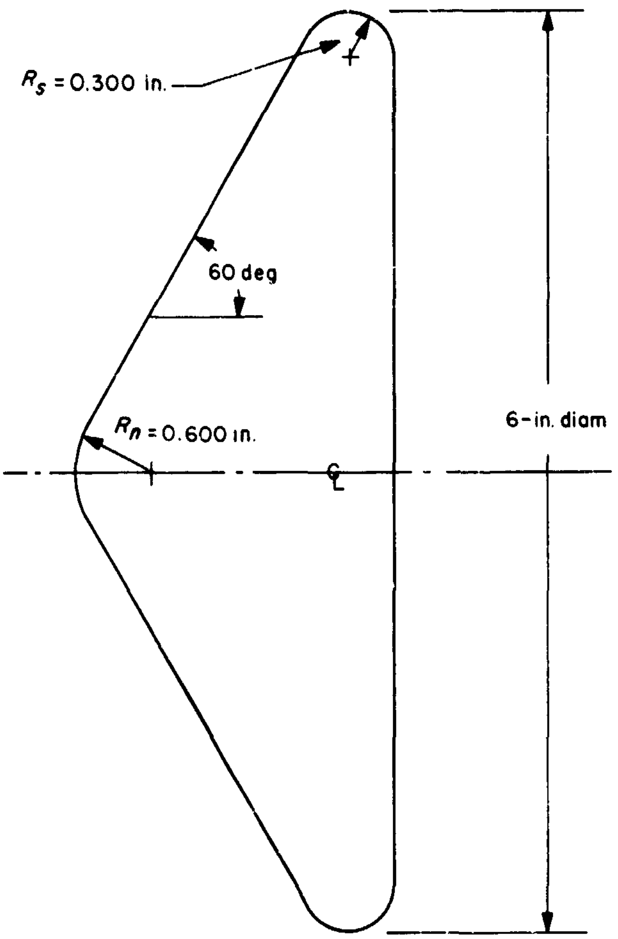


Fig. 1. Typical 60-deg half-angle blunted cone with 0.3-in. edge radius

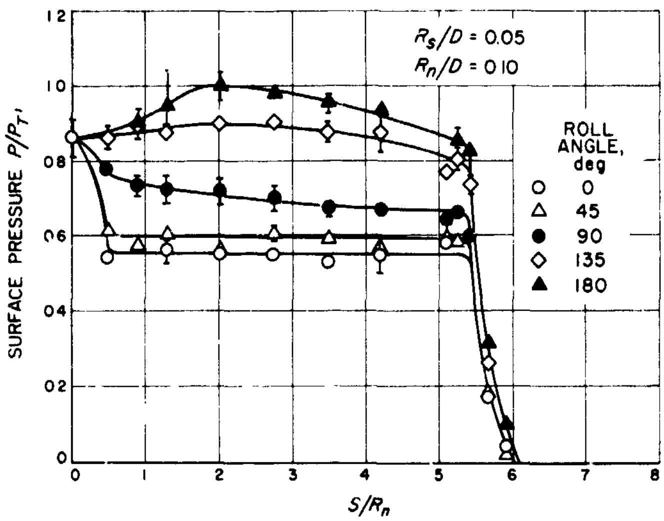


Fig. 3. Pressure distribution along 60-deg half-angle cone at 15-deg angle of attack

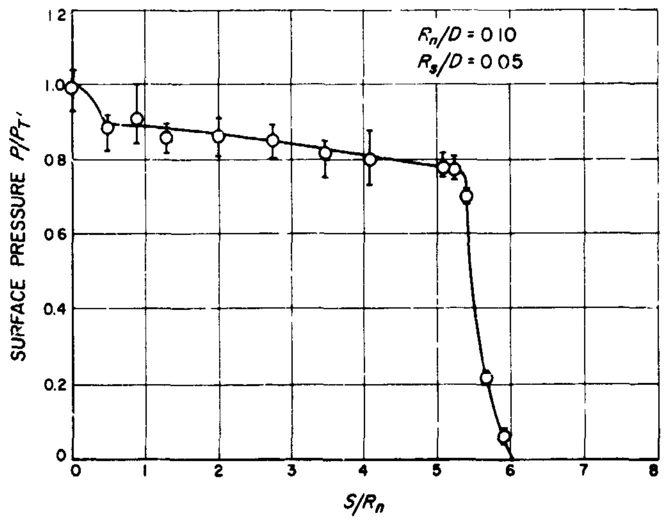


Fig. 2. Pressure distribution along 60-deg half-angle cone at 0-deg angle of attack

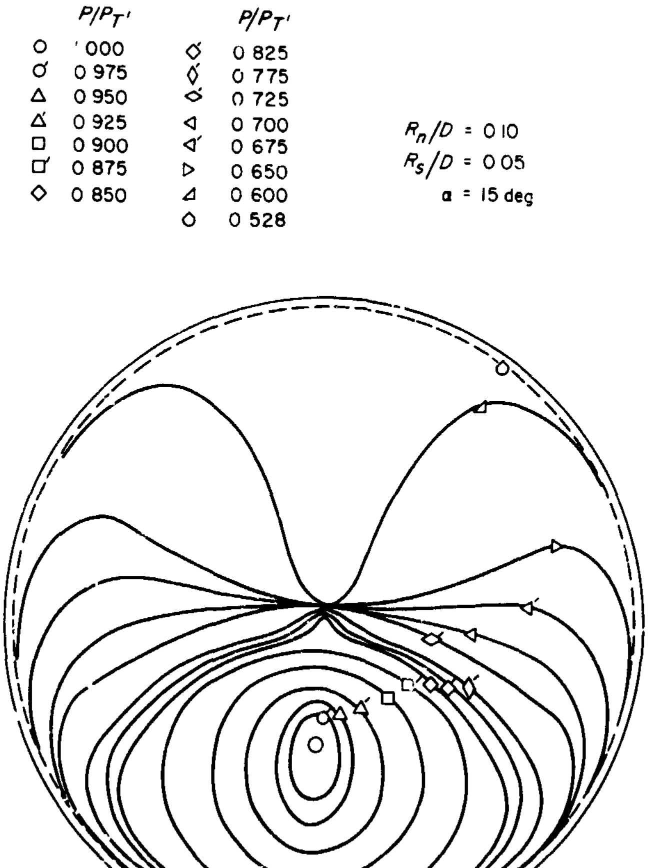


Fig. 4. Isobar diagram prepared from curves of Fig. 3

X. Environmental and Dynamic Testing

ENVIRONMENTAL SCIENCES DIVISION

A. Low-Frequency Plane-Wave Sound Generator and Impedance-Measuring Device, *C. D. Hayes*

1. Introduction

In the field of acoustic testing of spacecraft and subsystems, the production of correct sound power spectra is an important requirement for proper environmental qualification testing. At the present time, studies are being made to develop high-intensity sound generators with broader frequency response characteristics. Investigations of the response characteristics of acoustic horns will complement these sound generator studies (Refs. 1 and 2).

To empirically determine the response characteristics of any acoustic horn, a device providing plane-wave acoustic inputs over the frequency range of interest is needed. This device must also be capable of providing acoustic measurements within the horn to determine its response characteristics. The design should be such that the response characteristics of the acoustic source and of the termination can be mathematically eliminated, thus providing only the response characteristics of the horn. With this information, the precise contribution of a given horn to any acoustic system can be determined in advance of the actual system assembly.

A low-frequency plane-wave sound generator and impedance-measuring device (Fig. 1) was designed to fulfill these requirements (Ref. 3). This device (herein called impedance-measuring device) provides undistorted sinusoidal acoustic signals in the range from 10 to 400 Hz.

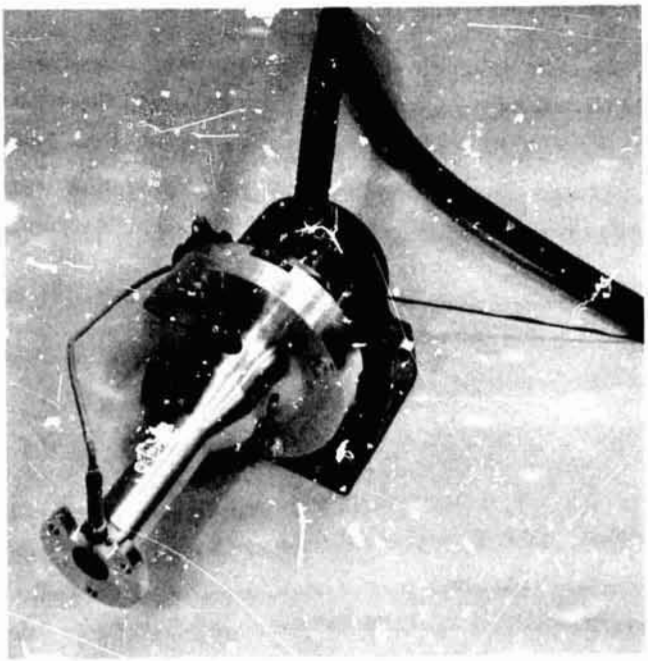


Fig. 1. Low-frequency plane-wave sound generator and impedance-measuring device, attached to vibration shaker

It establishes the acoustic pressures, particle velocities, and the phase relationship between the pressures and velocities at the input to the acoustic element under investigation. These same parameters are determined, either by measurement or by analytical prediction, at the output of the acoustic element. With these data, an accurate determination of the element impedance properties can be made at a given frequency and the element response characteristics determined. This process is repeated at enough frequencies within the frequency band of interest to provide adequate resolution.

2. Design

This device is designed to determine the acoustic impedance in terms of sound pressures and volume velocities as measured at the input of an attached acoustic element. Basically, the device requires only undistorted shaker output pressure signals, which are compatible with the sensitivities of the monitoring accelerometer and microphone, to provide accurate acoustic impedance information. (Mathematical derivations are described in detail in Ref. 3.)

3. Test Configuration

A typical test configuration (Fig. 2) consists of the following components:

(1) *Cylindrical tube*. This 1. [illegible] diam tube provides the volume area for the sound source over the length of the cylindrical tube and the alignment guide for the piston. The tube also provides the means for attaching a monitoring microphone for the throat pressure and an acoustic element or a variable-impedance tube.

(2) *Piston*. This unit is driven by a vibration shaker. The piston creates plane-wave sound fluctuations as it travels in the cylindrical tube under very close tolerance. An accelerometer is mounted, with its axis parallel to the piston motion axis, to measure the piston's acceleration. The face of the piston defines the "source" of the acoustic pressure fluctuations. A microphone can be installed in the face of the piston to measure the sound level and amount of distortion.

(3) *Vibration shaker*. This unit imparts oscillatory motion to the piston and is attached to the cylindrical tube with a mounting ring.

(4) *Acoustic element*. The acoustic element or horn to be analyzed is attached to the cylindrical tube so that the output of the tube becomes the input over its length for an output of the horn.

(5) *Variable-impedance tube*. For calibration purposes, a variable-impedance tube (blocked tube) is installed on the cylindrical tube. The diameter of the blocked tube is 1.375 in. with the inside diameter flared for a distance of 1.25 in. from the mating end to assure a smooth transition between the two tubes.

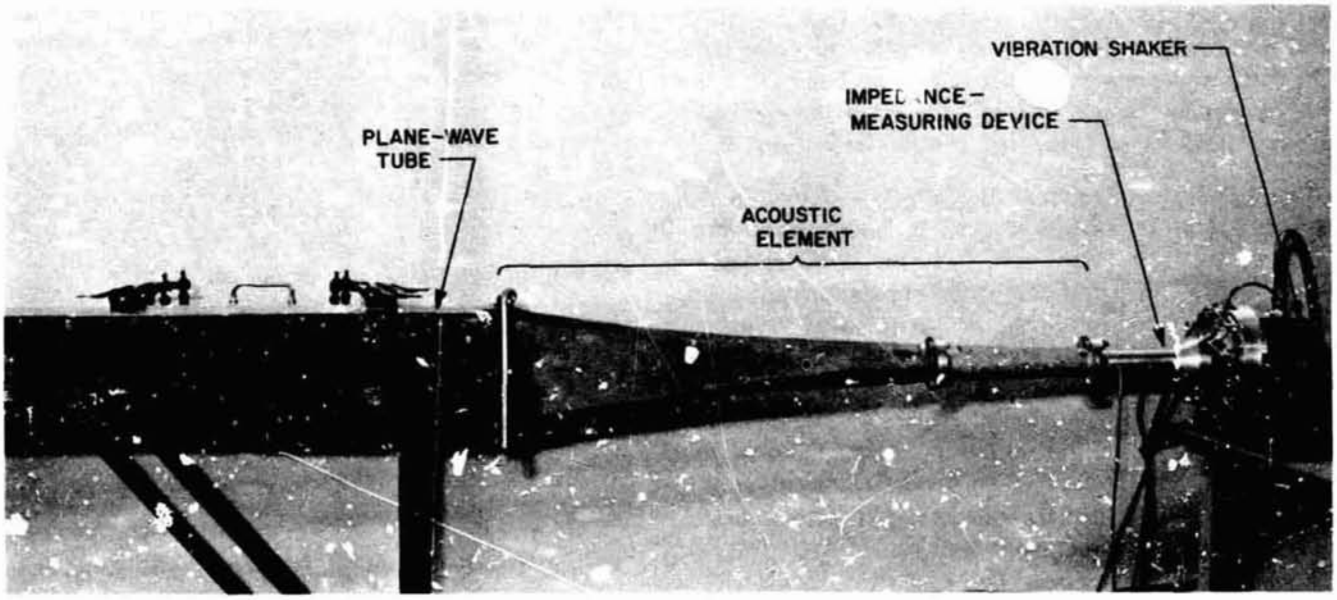


Fig. 2. Typical test configuration, using hyperbolic horn with ρc termination

(6) *Plane-wave tube*. For test purposes, a plane-wave tube may be attached to the mouth of the acoustic element to provide a ρc termination for the horn.

In a typical test setup (Fig. 2), the impedance-measuring device is attached to the throat of a (hyperbolic) horn. The mount of the horn is attached to a plane-wave tube, which is packed with absorbent material to provide a ρc termination for the horn. This configuration allows a comparison between the measured and the predicted (theoretical) response characteristics of a finite-length horn.

4. Calibration

The impedance-measuring device was calibrated using a variable-length blocked tube (Fig. 3) that provided a known loading impedance over a frequency range from 60 to 425 Hz. A computer program was written to implement the equations. The data indicate very good results over the frequency range of interest (60–425 Hz), and provide a verification of both the device and the calibration technique.

There are several possible sources of error in the calibration of the tube and the device.

(1) Variable-impedance tube.

(a) The tube w 'ls will resonate at particular frequencies within the frequency band of interest.

(b) The microphone head used to measure the throat pressure has a finite area rather than being a point. This source of error would increase with an increase in frequency.

(2) Impedance-measuring device.

(a) The microphone, which monitors the sound pressure levels at the output of the device, monitors pressures over a finite area (0.5-in.-diam circle); therefore, it is not a single point monitor as required by the accompanying mathematical theory.

(b) The phase angle between the accelerometer output voltage signal and the microphone output voltage signal is very critical.

(c) Small errors may be introduced by the manual readout of the accelerometer and the microphone output signal levels. (These outputs have not, as yet, been digitized.)

This calibration technique, which uses a blocked tube, provides a very wide dynamic range for calibrating the unit, since the tube will reflect impedances ranging from zero to infinity, depending on the value of cot KL, where

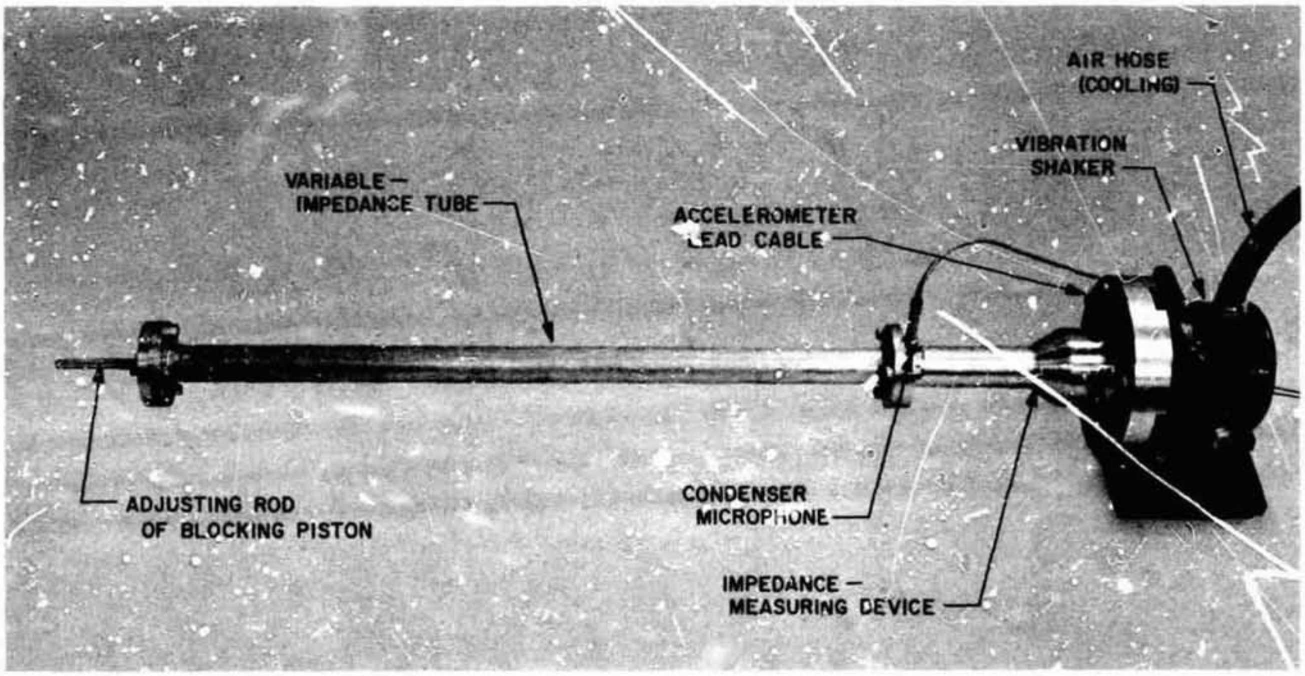


Fig. 3. Variable-impedance tube attached to device

wave number K, $\mathrm{cm}^{-1} \equiv 2\pi f/c$. This method provides a very accurate technique by selecting values of $KL = 0$ and $\pm(2n+1)\pi/2$, where $n = 0, 1, 2, \cdots$, since the value of $\cot KL$ changes very rapidly with KL. For these values of KL, any errors, such as described in *Paragraph (2-a)*, will greatly affect the value of the output impedance. The values of KL for these calibration runs were chosen such that $\cot KL$ was a fixed val e for each run and had, for all the runs, a range of absolute values of $0.303 \leqq \cot KL \leqq 1.000$.

5. Conclusions

The data obtained from the calibration runs indicate that accurate acoustic impedance information can be obtained. Calibration runs have verified the mechanical design of the device as well as the accompanying mathematical analysis. Use of this device requires accurate methods of data acquisition, such as digital readout of phase, acceleration, and acoustic pressure, and maintaining undistorted input signals at the output of the device.

References

1. Olson, H. R., *Acoustical Engineering*, pp. 103–114. D. Van Nostrand Co., Inc., Princeton, N. J., 1957.
2. Hayes, C. D., *Acoustic Spectrum Shaping Utilizing Finite Hyperbolic Horn Theory*, Technical Report 32-1141. Jet Propulsion Laboratory, Pasadena, Calif., Aug. 15, 1967.
3. Hayes, C. D., and Lamers, M. D., *Low-Frequency Plane-Wave Sound Generator and Impedance-Measuring Device*, Technical Memorandum 33-376. Jet Propulsion Laboratory, Pasadena, Calif., Mar. 1, 1968.

XI. Solid Propellant Engineering

PROPULSION DIVISION

A. Molecular Momentum Transfer From Regressing Solid Propellant Surfaces,

O. K. Heiney

1. Introduction

One of the more enduring suppositions of propellant deflagration is that of impulse propulsion. In essence, the hypothesis assumes a significant impulse pressure will be generated by an exchange of momentum between burning gas molecules and the surface of the propellant, from which these molecules were emitted. The following analysis briefly outlines the argument and development, largely on a molecular basis, that serves as justification for this effect, then considers more conventional gas-dynamics and ballistics which predict an effect of much lower magnitude. Finally, the experimental procedure used to adequately demonstrate that the lower predicted value of impulse pressure is the correct expression is described.

Symbols used in this article are defined in Table 1.

2. Analysis

a. Impulse pressure. Reference 1 is the generally quoted analysis for warranting the anticipation of this impulse effect. The development and assumptions presented below are those given in this reference:

(1) There is a 100% conversion of the heat of combustion of the propellant into the kinetic energy of the gas molecules.

(2) There is a directional equiprobability of molecular emission in the half hemisphere bounded by the propellant surface.

(3) A mean molecular emission velocity may be defined which is a function of the total combustion energy potential of this propellant.

The specific energy potential of the propellant is given as

$$u_s = \frac{F_p g}{\gamma - 1} = \frac{\sum\limits_{K=1}^{n} m_K n_K V_K^2}{2 \sum\limits_{K=1}^{n} m_K n_K}$$

The mean velocity $\bar{v}_e$ is

$$\bar{v}_e = \left(\frac{\sum\limits_{K=1}^{n} m_K n_K V_K^2}{\sum\limits_{K=1}^{n} m_K n_K} \right)^{1/2}$$

Then,

$$\bar{v}_e = \left(\frac{2F_p g}{\gamma - 1} \right)^{1/2}$$

which relates this postulated emission velocity to the impetus F_p of the propellant.

Table 1. Nomenclature

C^*	characteristic velocity
F_p	impetus of propellant
g	acceleration due to gravity
m_K	mass of K-type molecule
n	dimensionless burning rate exponent
n_K	number of K-type molecules
P_c	chamber pressure
P_ι	impulse pressure
R	gas constant
r	burning rate
S_B	burning surface
T_F	temperature of flame
u_s	specific energy potential of propellant
V_g	gas velocity
V_K	velocity of K-type molecule
$\bar{v}_e$	mean molecular ejection velocity
Γ	flow factor
γ	ratio of specific heat
ρ_g	gas density
ρ_p	propellant density

Using geometrical arguments, the development then states that the effective impulse pressure generated is equal to only one-fourth of the mass emitted at this velocity, giving

$$F = \frac{d(mV)}{dt} = \frac{\bar{v}_e}{4}\frac{dm}{dt} \qquad (1)$$

where

$$\frac{dm}{dt} = \frac{\rho_p S_B r}{g}$$

Then,

$$F = \frac{\rho_p S_B r}{4g}\left(\frac{2F_p g}{\gamma - 1}\right)^{1/2}$$

or for an end-burning configuration

$$P_\iota = \frac{\rho_p r}{4g}\left(\frac{2F_p g}{\gamma - 1}\right)^{1/2} \qquad (2)$$

which is the predicted impulse pressure with the given assumptions.

b. Conventional mass balance approach. The mass balance equation is given as (see Fig. 1)

$$\rho_p r S_B = \rho_g A V_g \qquad (3)$$

For end burner

$$V_g = \frac{\rho_p r}{\rho_g} \qquad (4)$$

$$P_c = \rho_g F_p \qquad (5)$$

Then, substituting Eq. (5) into Eq. (4) gives

$$V_g = \frac{\rho_r r F_p}{P_c}$$

for an impulse pressure of

$$P_\iota = \frac{\rho_p^2 r^2 F_p}{g P_c} \qquad (6)$$

Equations (2) and (6) are fundamentally different in both form and effect prediction. It can be seen, however, that for either equation this predicted impulse pressure is quite low. In fact, it is for most purposes a second order effect. Figure 2 gives a plot of the impulse pressure predicted by both equations as a function of chamber pressure. It can be seen that the conventional gas dynamic approach indicates a pressure 40 times lower than the molecular momentum transfer approach at low pressures.

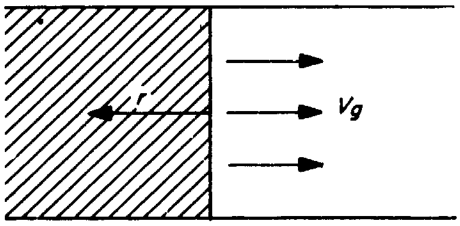


Fig. 1. Mass balance approach

While at higher pressure (e.g., 10,000 psia), the difference is well over three orders of magnitude. It must be understood that these figures are for a given propellant formulation and burning rate, as both Eqs. (2) and (6) are highly sensitive to the deflagration rate dependence on pressure. An end-burning configuration was also assumed. If charges are perforated to increase the burning area, Eq. (6) can be multiplied by the burning area to chamber area ratio. The analysis for Eq. (2) would completely fail, however, as the majority of the molecules would be "ejected" radially rather than longitudinally.

To determine which, if either, of the expressions is correct, an experimental program was undertaken to compare the thrust generated by a 3-in.-diam end-burning motor. This motor was fired at atmospheric pressure with a constant 7.07-in.^{2} burning surface.

As can be seen from Fig. 2, the molecular momentum exchange equation would predict a thrust of 1.98 lb, while conventional ballistics would predict a thrust of 0.044 lb.

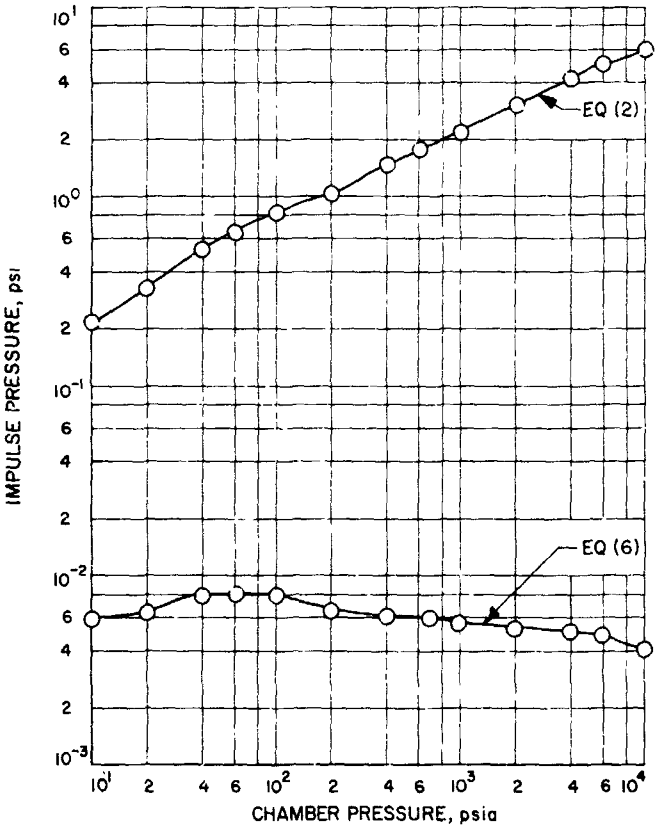


Fig. 2. Impulse pressure predictions as function of chamber pressure

The propellant used was of the aluminized composite rubber base type. An impetus F_p for the propellant was determined from the C^* value by the simple relationships (Ref. 2)

$$F_p = RT_0$$

$$C^{*2} = \frac{RT_0}{\Gamma^2}$$

$$F_p = C^{*2}\Gamma^2$$

The C^* of 4890 ft/s for the formulation gave an impetus of 312,400 ft-lb/lb, which is quite typical of average gun propellant impetus values. Other parameters of the propellant are:

$$\gamma = 1.14$$

$$r, \text{at } 1000 \text{ psia} = 0.37 \text{ in./s}$$

$$\rho_p = 0.065 \text{ lb/in.}^3$$

$$T_F = 2743°\text{K}$$

$$n = \sim 0.5$$

3. Experimental Procedure

The test configuration initially used is illustrated in Fig. 3. The load cell utilized had a maximum thrust capability of 2 lb and a resolution accuracy of ±0.002 lb. The wheeled suspension system was found to be too crude for the delicate thrust measurements. A suitably sensitive suspension system that was successfully utilized is illustrated in Fig. 4. The system was based on ballistic suspension of the motor and proved quite effective.

Figure 5 illustrates the plume developed from the 3-in. motor during a firing. The fiducial lines on the thrust stand are 1 ft apart. In general, the plume was quite impressive and one could legitimately suppose a sizable thrust was being generated. During firings for which data was developed, a ± 1-psid pressure gage with a resolution of ±0.01 psi indicated that chamber pressure and ambient pressure differentials were not measurable. Thrust measurements during the first firing showed a constant thrust of 0.046 lb for the 60-s firing duration while the second firing had a constant thrust of 0.042 lb for a like period. Within the limits of the load cell resolution, these values are as predicted by the ballistic analysis of Eq. (6). A shortened chamber pressure and thrust curve are shown in Fig. 6. It can be seen that the only noticeable pressure increment occurs at ignition and then falls to zero.

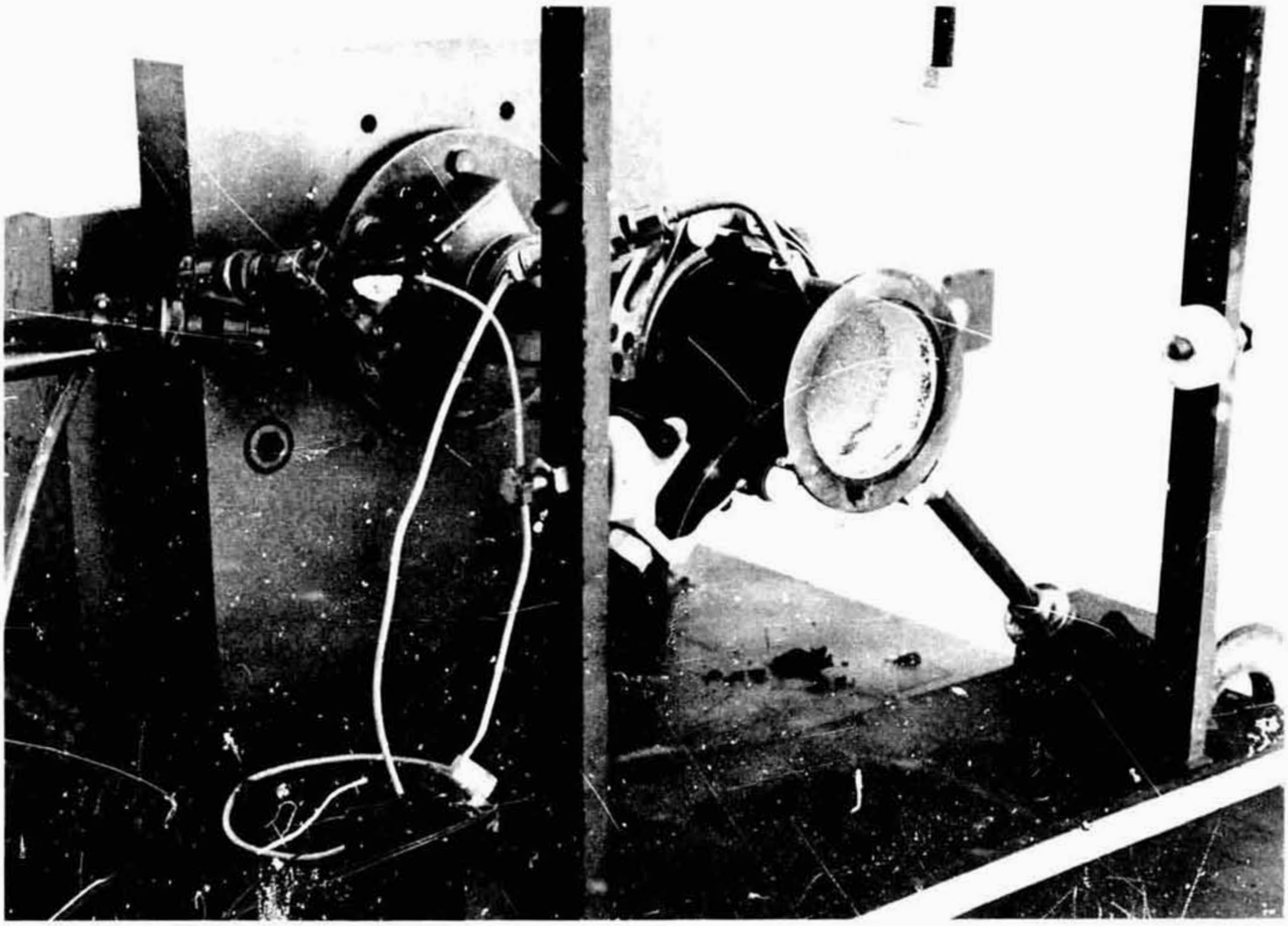


Fig. 3. Initial solid propellant motor test configuration using wheeled suspension system

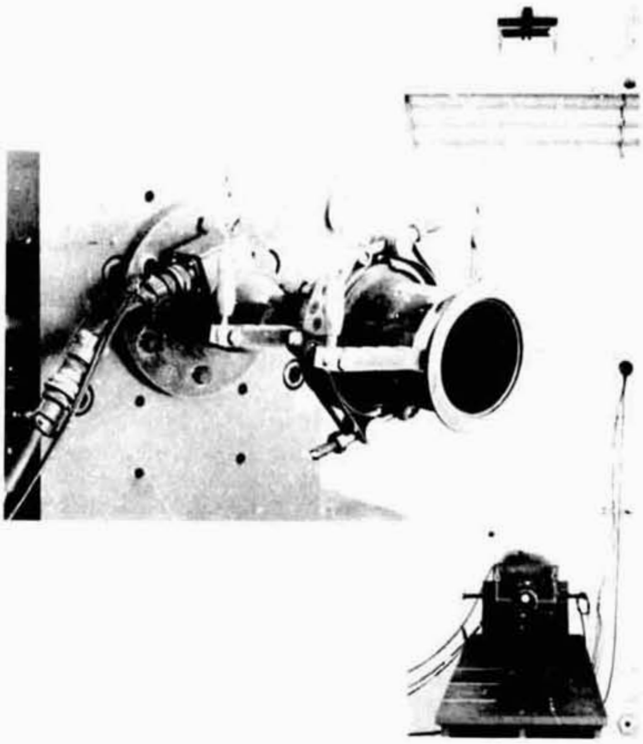


Fig. 4. Sensitive suspension system for motor thrust measurements

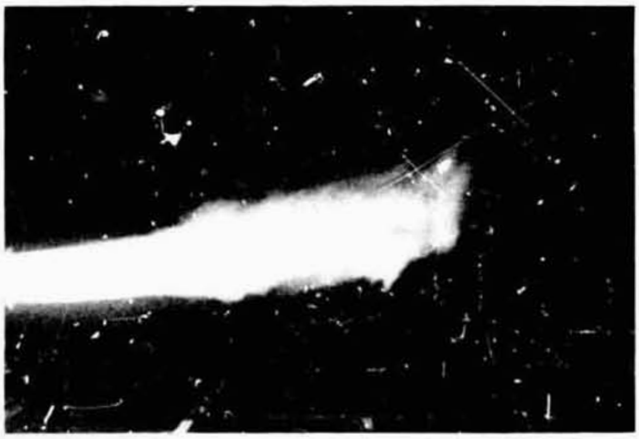


Fig. 5. Plume developed from firing a 3-in. solid propellant motor

4. Conclusion

The results of this study indicate that the molecular momentum exchange impulse pressure development is erroneous. This is primarily due to a fundamental physical misconception between the mean and net gas velocity, which is contained in the assumptions.

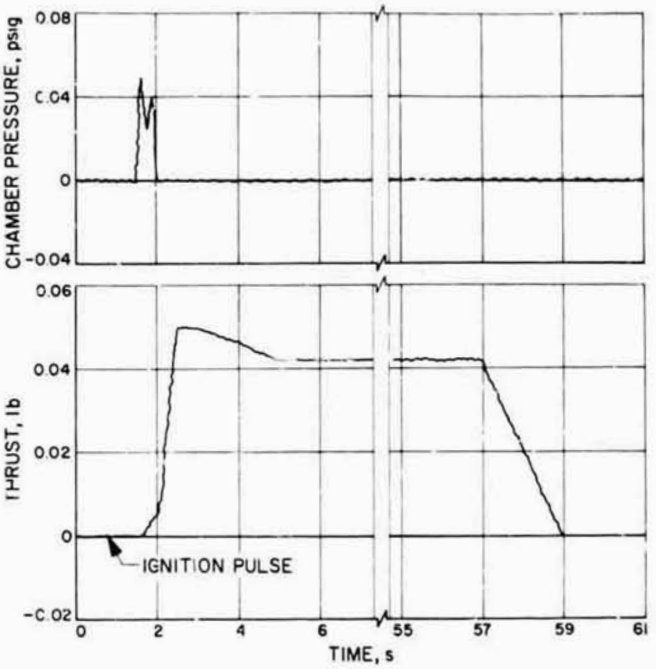


Fig. 6. Chamber pressure and total thrust curves for motor firing

Also implicit in these results is the fact that a significant ballistic effect is not obtainable from the impulse pressure concept. Considerable effort has been expended on the various "traveling charge" systems of gun ballistics in an attempt to utilize this impulse pressure phenomenon. These experiments usually failed due to propellant physical property considerations. If they had not, however, it would have been seen that fundamental physical misconceptions were present in the basic hypotheses.

References

1. Lee, L., and Laidler, K., "The Interior Ballistics of the Impulse Propulsion Gun," Catholic University of America, Washington, D.C., Aug. 1951.

2. Hugget, C., Bartley, C. E., and Mills, M. M., *Solid Propellant Rockets*, Princeton Aeronautical Paperbacks, Princeton, N. J., 1960.

B. T-Burner Studies, *E. H. Perry*[1]

1. Introduction

One of the primary objectives of the current T-burner studies at JPL is to gain a more thorough understanding of the burner itself. Experiments were conducted to measure the acoustic losses of a 1.5-in.-diam T-burner.

[1] California Institute of Technology, Pasadena, Calif.

Although the measurements were made under "cold" conditions in the burner, a basis is provided for understanding the losses observed during test firings.

2. Theoretical Acoustic Losses

The acoustic field within a T-burner during a firing consists of a standing wave of wavelength $2L$, where L is the length of the burner cavity. This field is maintained by the burning propellant at the ends of the cavity and accordingly decays after burnout of the propellant. The decay is approximately exponential in time with the time required for the acoustic pressure to drop by a factor of e, defined as the "decay time" of the burner. Usually, however, reference is made to the "decay constant," which is the reciprocal of the decay time.

It is well established that a sound wave traveling through a tube is attenuated at a rate proportional to the square root of the frequency and inversely proportional to the tube's radius. This decay is due to viscous and thermal dissipation near the tube wall. Reference 1 gives the following expression for the decay constant associated with these wall losses:

$$\alpha_w = (\pi)^{1/2}\left[(\nu)^{1/2} + (K)^{1/2}(\gamma - 1)\right]\frac{(f)^{1/2}}{R} \qquad (1)$$

where

ν = kinematic viscosity coefficient

K = thermal diffusivity coefficient

γ = specific heat ratio

f = frequency

R = tube radius

In addition, there are thermal losses associated with the reflection of the wave from the ends of the cavity. Through arguments similar to those used to derive Eq. (1), one can show that the decay constant for such end losses is given by:

$$\alpha_e = (4\pi K)^{1/2}(\gamma - 1)\frac{(f)^{1/2}}{L} \qquad (2)$$

Since the T-burner is a vented cavity, the possibility exists for acoustic radiation from the exhaust vent. However, the center of this vent is located precisely at the pressure node of the standing wave in the cavity. Therefore, if the diameter of the vent is small compared to the cavity length, any radiation losses from the vent can be expected to be very small. In the present experiments, the ratio of vent diameter to cavity length never exceeded 0.06.

Thus, it appears that the only losses in the "cold" T-burner should be those due to dissipation at the walls and ends of the cavity. If this is indeed the case, the decay constant of the burner should be the sum of the wall and end decay constants. That is, if α is the burner decay constant, then

$$\alpha = \alpha_w + \alpha_e \qquad (3)$$

where α_w and α_e are given above.

3. Experimental Procedures and Results

An acoustic environment simulating that encountered during a firing was provided within the cavity by a sound driver unit outside. An audio oscillator was used to drive this unit at the standing-wave frequency of the cavity. The sound introduced into the cavity through a small hole at one end was observed by a 0.25-in.-diam condenser microphone at the opposite end. Figure 7 illustrates the arrangement used.

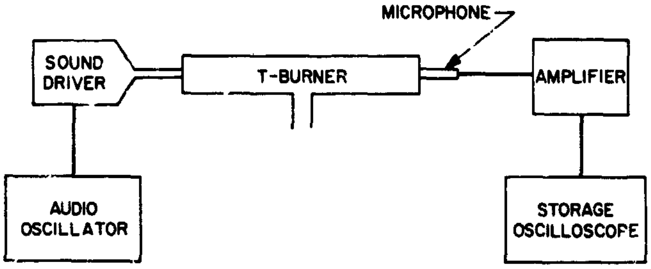


Fig. 7. Block diagram of experimental arrangement

By abruptly turning off the sound driver and observing the subsequent decay of the standing wave, the decay constant of the burner was determined. Burner lengths ranging from 7 to 42 in. were used to obtain a range of frequency.

Figure 8 illustrates the behavior of the decay constant as a function of frequency at atmospheric pressure. For the purpose of comparison, the values of the decay constant predicted by Eq. (3) are plotted along with the experimental values. The agreement is seen to be fairly good over the entire frequency range. The experimental values all lie above those given by the theory, which is to be expected since there are small losses associated with the sound lead-in and detection devices.

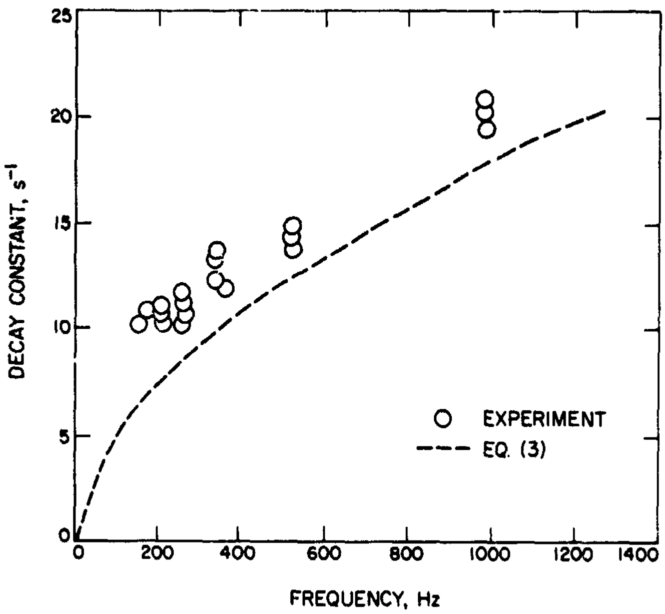


Fig. 8. Decay constant as a function of frequency at atmospheric pressure

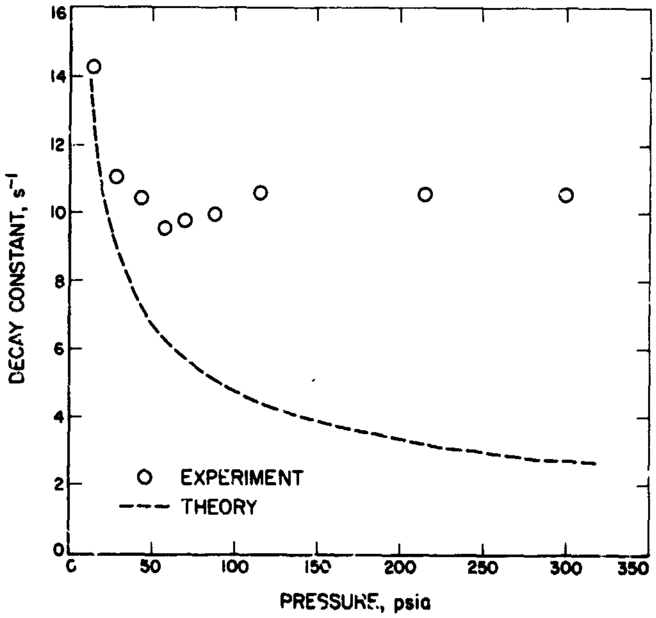


Fig. 9. Decay constant as a function of pressure

Figure 9 presents the experimental and theoretical values of the decay constant as a function of mean chamber pressure. To obtain these measurements, the apparatus was placed in a chamber pressurized with nitrogen. All of these measurements were made at a frequency of 530 Hz. As can be seen in the figure, the agreement between theory and experiment becomes progressively worse as the chamber pressure increases. The cause of this condition is not completely understood at present, although it might be due to losses associated with improper fitting of some of the burner sections. Small gaps between adjoining sections have been found to give rise to very large losses; possibly these losses increase with pressure, which would explain the above results.

The final phase of the experiments consisted of an attempt to measure the acoustic losses associated with the vent. A plug was made to fit into the vent so that the latter could be completely closed off, thereby eliminating the possibility of any acoustic radiation from the vent. Decay measurements obtained with the vent thus closed were compared with those obtained with it open. Any difference between the two sets of measurements was too small to be detected, which indicates the vent losses are indeed small as suggested above.

4. Application of Results

There is evidence that the above "cold" burner analysis applies also to the losses observed during actual T-burner firings. Figure 10 presents decay constant data reported in Ref. 2 for two similar composite propellants denoted as A-13 and A-14. The empirical curve through the data assumes the square-root dependence suggested by Eq. (3). The rather good fit suggests that the acoustic losses of

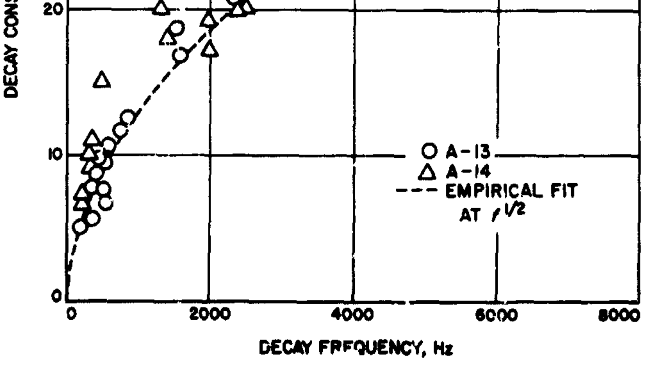


Fig. 10. Decay constant as reported in Ref. 2 for actual T-burner firings

the T-burner are described rather well for these two propellants by an equation similar to Eq. (3). It should be mentioned that other data of this reference exhibit a similar behavior. Future studies are expected to show, among other things, that the losses have the geometric dependence indicated in Eq. (3) as well as the frequency dependence discussed above.

References

1. Landau, L. D., and Lifschitz, E. M., *Fluid Mechanics*, p. 303. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1959.

2. Horton, M. D. *Testing the Dynamic Stability of Solid Propellants: Techniques and Data*, NAVWEPS Report 8596, NOTS TP 3510, pp. 34–35. U.S. Naval Ordnance Test Station, China Lake, Calif., Aug. 1964.

N68-37409

XII. Polymer Research

PROPULSION DIVISION

A. Investigation of the Transport Characteristics of an Ionene Membrane,

H. Y. Tom and J. Moacanin

1. Introduction

The battery separator material is one of the key factors that determine the lifetime of a silver–zinc battery. Ideally, the battery separator membrane should allow charge transfer to carriers such as OH^-, but should prevent silver and zinc ionic species from leaving their respective half-cells and thus avoid internal short circuits.

The objective of this work was to initiate a systematic study of the various transport characteristics of membranes to ascertain the chemical and morphological requirements that lead to desirable permselective properties. In free diffusion, the solvent and solute move relative to each other. Hence, only one transport coefficient would be required to relate flow and concentration. Imposing a membrane would require additional factors that must consider the interaction of the solute and solvent with the membrane. Another consideration that influences transport is the pore size in the membrane. Such membranes can then be experimentally tested for their permselectivity by the number of coefficients required to describe the transport of ions, using the formalism of irreversible thermodynamics, provided the process is just slightly off equilibrium (Refs. 1 and 2).

This portion of the study was performed on ionene membranes (Ref. 3). In ionenes, positive quaternary ammonium ionic groups are incorporated along the hydrocarbon backbone and their charge density can be varied in a systematic fashion to assess their effect on the transport coefficients. Although the current polyethylene–graft–acrylic-acid separator also has a hydrocarbon backbone, the acrylic acid branches are distributed at random; whereas, in ionenes, the charged groups are distributed in a uniform manner. This article covers the electrical properties of cells prepared with ionene membranes. When the concentration data that are presently awaiting analysis become available, an article demonstrating the presence or absence of preferred ionic transport will be presented.

2. Materials and Equipment

Materials procured for this study consisted of N,N,N',N'-tetramethylhexanediame, 1,6-dibromohexane, tetrachloro-o-benzoquinone (TCBQ), reagent grade potassium chloride, grade 72-51 polyvinyl alcohol (PVA), and battery separator membranes. De-ionized distilled water was used throughout the investigation.

Transport cells were fabricated from pyrex glass (Fig. 1). The glass joint holdıng the two half-cells has a grooved flange to permit the installation of an O-ring. The membrane is mounted on the flange of one cell with the O-ring pre-installed. The flange from the other cell is then brought in contact with the membrane and the assembly clamped.

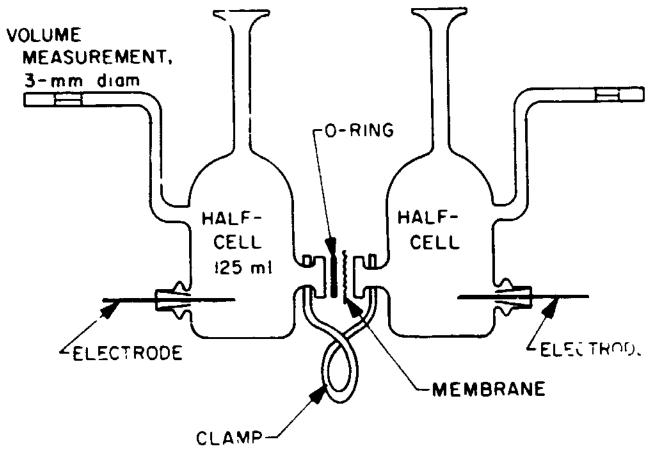


Fig. 1. Transport cells for ionene membrane tests

The horizontal arms with a 3-mm bore diameter are used for volume measurement. Since both arms are at the same height, flow of liquid across the membrane can occur without change in hydrostatic pressure. The volume measurement is good to ±14 μl with a volume of about 125 ml/cell. Glass joints were also included to permit the insertion of platinum electrodes.

Equipment required for electrical measurements consisted of a high-impedance dc millivoltmeter, an ac impedance bridge for resistance, a regulated dc power supply, and an electrical timer. Platinum blackened electrodes were obtained by electroplating 0.010-in.-diam platinum wire.

3. Membrane Fabrication

The membranes made for this study were prepared by combining N,N,N',N'-tetramethylhexanediame and 1,6-dibromohexane on a 1:1 gram molecular weight basis (Ref. 1). The synthesized copolymer designated as a 6,6-ionene was weighed and added to PVA and TCBQ in different proportions. The PVA and TCBQ weight ratio was maintained at 100:1. Water was added as needed. PVA was prepared as a solution by heating water to 100°C and adding PVA for supersaturation. Any insoluble PVA was removed by filtration.

The water mixture with the membrane ingredients was shaken, then cast the next day onto glass slides. The water was allowed to evaporate, and the films were later heat-treated at 100°C for 1 h. These membranes were then stored in petri dishes.

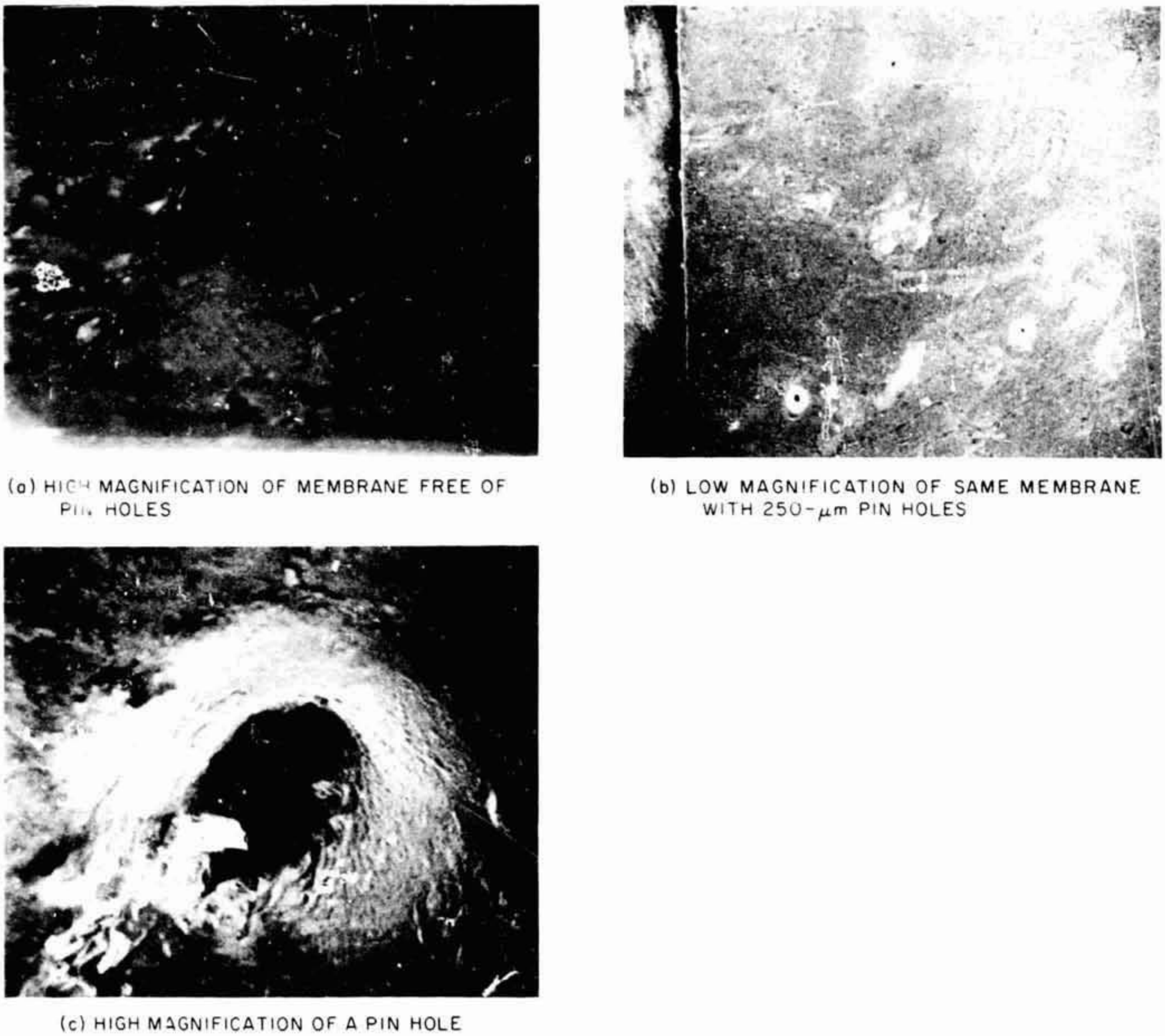
One membrane (50 wt % ionene) was inspected with the stereoscan electron microscope. The dry-mounted sample was found to be pinhole free; for comparison the same sample is shown with a puncture made with a 250-μm pin (Fig. 2). This result indicated that the fabrication procedure was satisfactory and that all the membranes should be free of pinholes. Further investigation on this point is being continued and will become a routine procedure for membrane characterization.

For chemical analysis (by Gulf General Atomic, San Diego, Calif.), the samples were first treated with neutron irradiation and then assayed in batches for potassium and chlorine as radioactive elements in a scintillation counter. A standard and a blank were always included with each batch.

4. Experimental Procedures

The membranes, water-prewetted or dry, were mounted first; the platinum electrodes were inserted next; then, the cells were filled with their respective bathing media. Once the media came in contact with the membrane, a timer was activated. The transport apparatus was then placed in an ultrasonic cleaner and vibrated for 2 min to remove any entrapped atmospheric gases. The cells were then transferred to a bench where an ac impedance bridge and a millivoltmeter were connected to each electrode. The high-impedance millivoltmeter which continuously monitored the potential difference across the membrane was assumed not to draw current from the system. The ac impedance bridge was activated only when the membrane resistance was measured. The bridge was energized by a 400-mV, 1000-Hz internal source. Aliquots of 50 μl were removed periodically from each cell.

Membrane thicknesses were measured to ±31 μm with the aid of a calibrated filar eyepiece and a stereomicroscope. A piece of the membrane was excised from the remaining stock material in the immediate area where a larger piece had been previously removed for the transport study. Its thickness was measured while dry, in water, and in salt solution. For the analysis of the transport experiments, the thickness was taken to be the average dimension in water and salt solution.



(a) HIGH MAGNIFICATION OF MEMBRANE FREE OF PIN HOLES

(b) LOW MAGNIFICATION OF SAME MEMBRANE WITH 250-μm PIN HOLES

(c) HIGH MAGNIFICATION OF A PIN HOLE

Fig. 2. Electron micrographs of a 50-wt-% ionene membrane and the detection of pinholes

5. Results

At different intervals of time, the volume difference of each cell, ac resistance at 1000 Hz, potential difference, and aliquots from each cell were obtained. The potential observed is that generated by the two cells, which act as concentration half-cells. Concentrations are not included at this time as the chemical analyses are incomplete. The volume changes are shown in Fig. 3.

During the first 3000 s, the ionene membranes (Fig. 3) exhibit fairly rapid volume changes. The initial phase is followed by a decrease in the rate of volume change with some indication that steady state is approached. This is best illustrated in Fig. 3f where the volume change for the 70-wt-% ionene content is essentially a straight line. The battery separator (Fig. 3g) also shows an initial rapid phase followed by a slower phase. For the battery separator, however, the initial phase takes only 1500 s. It is interesting that the initial phase, the incubation period, of the ionene membranes appears to be independent of its thickness. For the battery separator, this point could not be checked since only one thickness was available.

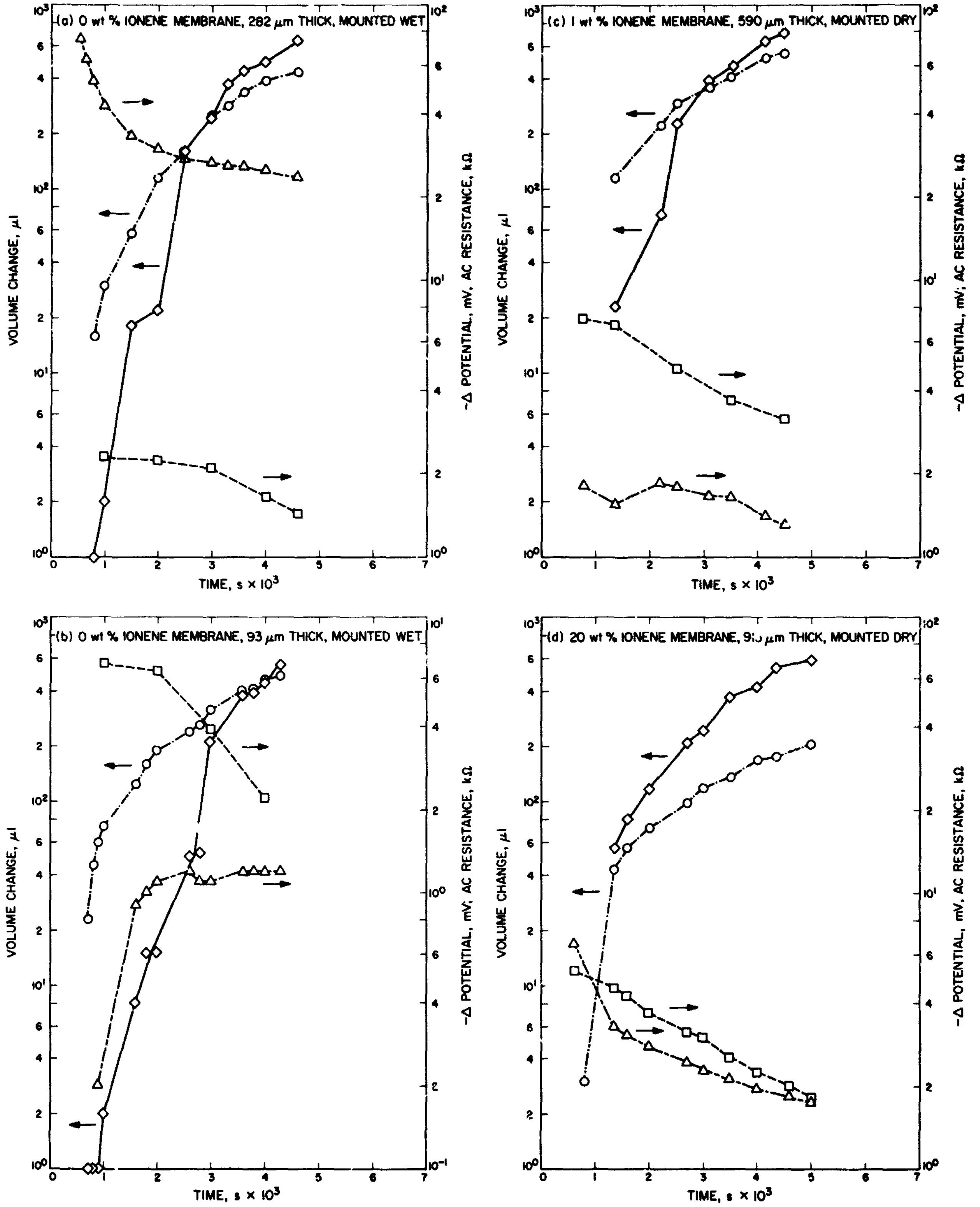


Fig. 3. Temporal responses

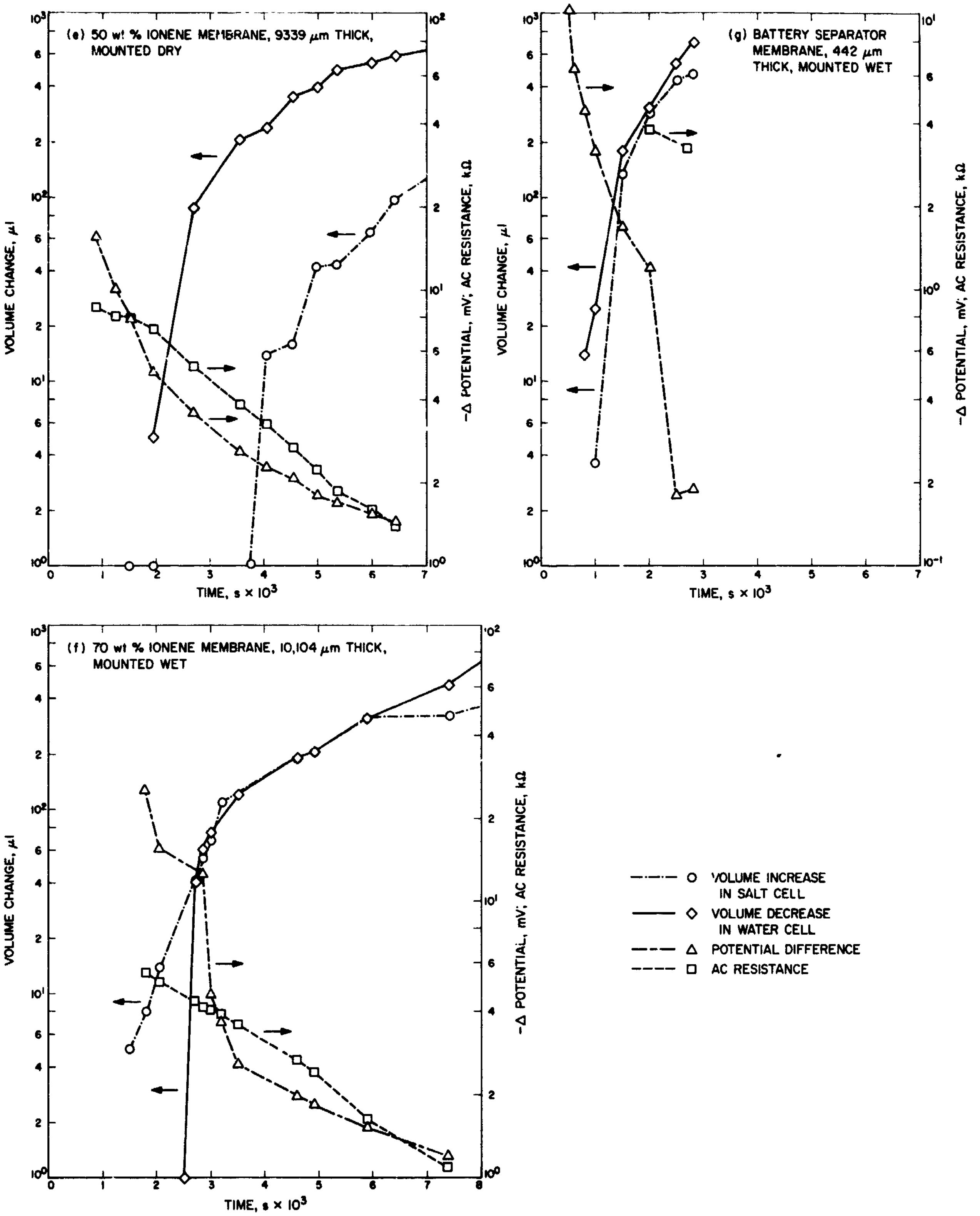
 (e) 50 wt % IONENE MEMBRANE, 9339 μm THICK, MOUNTED DRY
(g) BATTERY SEPARATOR MEMBRANE, 442 μm THICK, MOUNTED WET
(f) 70 wt % IONENE MEMBRANE, 10,104 μm THICK, MOUNTED WET
VOLUME CHANGE, μl
−Δ POTENTIAL, mV; AC RESISTANCE, kΩ
TIME, s × 10³
10⁰
10¹
10²
10³
10⁻¹
2
4
6
0
1
2
3
4
5
6
7
8
VOLUME INCREASE IN SALT CELL
VOLUME DECREASE IN WATER CELL
POTENTIAL DIFFERENCE
AC RESISTANCE

Fig. 3 (contd)

Once the slow phase begins, the sum of the volume differences may not be zero. However, where those sums are not zero, the membranes were mounted dry; those membranes whose values are about zero were mounted wet. To assess the swelling behavior, the membranes were bathed in both water and in potassium chloride solutions. Results in Table 1 show that the membrane thickness, including the separator material, is essentially unaffected by the media until the ionene content is 20 wt % or greater. When the ionene content exceeds 20 wt %, the membrane prewetted with water contracts markedly when plunged into a salt solution.

One variable sensitive to the volume differences is the potential difference (ψ) across the cell. The voltage in absolute units is used to generate the curves in Fig. 3. However, whether or not the potential difference is a measure of the ionic concentration across the membrane is still inconclusive as the aliquot concentrations have yet to be completed. Nevertheless, this measurement is certainly more sensitive to the volume changes than resistance and must necessarily be reflected in the ionic concentrations.

6. Conclusion

The results suggest that as the membrane absorbs water there ensues a decrease in the total volume in the cell (liquid plus membrane). One possible explanation is that when water is absorbed by the membrane, the hydration sphere around the quaternary ammonium ion reduces the specific volume of water in the sphere and thus leads to a negative volume of mixing. Volume contractions are well known for mixtures of salt solutions and water.

This volume decrease is evidently unrelated to the incubation period since it exists whether the membrane is mounted wet or dry. It may be argued that the initial phase is an artifact since the surface tension in the capillary of the water cell may be large enough to prevent flow. Flow begins at some later time when enough ions are transported across to reduce the surface tension.

Table 1. Temporal response of membrane to bathing media

Thickness of dry membrane,[a] μm	Time in 2-M potassium chloride solution, s	Thickness of membrane in solution, μm	Time in water, s	Thickness of membrane in water, μm
Control				
79	100	116	400	101
109	300	99	3500	93
115	600	92		
	1000	93		
Control				
273	170	261	400	223
	250	258	800	318
	425	225	1000	310
	1400	231	1350	290
	2000	248	6500	296
	2700	267		
1 wt % ionene				
455	800	599	100	545
	1500	714	500	621
	1800	574	800	636
	1950	672	1400	613
	2000	559	2900	621
20 wt % ionene				
827	460	586	1500	1247
	575	611		
	750	594		
	900	584		
	1800	583		
50 wt % ionene				
796	1070	—	670	17797
	1230	—	1030	17999
	1320	—	1460	16447
	1410	—	2490	17688
	1934	990		
70 wt % ionene				
45	600	1226	(prewetted)	
	1700	1188	∞	19162
	2080	1160		
	—	1069		
	—	1046		
Battery separator				
402	200	396	200	527
313	400	448	500	464
301	1400	438	900	403
341			3700	445
335				
335				
323				
313				
311				
304				

[a] Thicknesses of dry membranes measured at arbitrary times.

If this were so, the duration of the induction period would depend on the thickness; however, no correlation with thickness was observed. Moreover, even when some membranes were mounted prewetted, the incubation period was not reduced. Thus, the evidence seems to strongly indicate that the initial phase is real, although the mechanism is unknown.

References

1. De Groot, S. R., *Thermodynamics of Irreversible Processes*, North Holland, 1951.
2. Kedem, O., and Katchalsky, A., *Trans. Faraday Soc.*, Vol. 59, pp. 1918, 1931, and 1941, 1963.
3. Rembaum, A., Baumgartner, W., and Eisenberg, A., *J. Polym. Sci.*, Part B, Vol. 6, p. 159, 1968.

XIII. Research and Advanced Concepts

PROPULSION DIVISION

A. Laminarization in Nozzle Flow,

L. H. Back, R. F. Cuffel, and P. F. Massier

1. Introduction

Turbulent boundary layers under certain flow acceleration conditions can undergo reverse transition toward laminar boundary layers. This phenomenon offers the advantage of a reduction in convective heat transfer and is of considerable interest since it can sometimes be promoted in rocket nozzles. The reverse transition process, referred to as laminarization, has been found to occur when values of the parameter $K = (\nu_e/u_e^2)\ (du_e/dx)$ exceed about 2×10^{-6}. (Symbols used in this article are defined in Table 1.)

To better understand the conditions under which laminarization occurs and the effect of laminarization on the friction coefficient, an investigation of the structure of the boundary layer was undertaken in a nozzle.

2. Test Conditions and Apparatus

The nozzle used for the tests (Fig. 1) resembles a configuration used for rocket engines in which the combustion chamber is an integral part of the convergent portion of the nozzle. The conical half-angle of convergence was 10 deg, the inlet diameter 5.00 in., and the throat diameter 1.59 in. The nozzle was also instrumented so that heat transfer measurements could be made.

Boundary layer measurements upstream and within the nozzle were made at the stations noted in Fig. 1, where the free-stream Mach numbers were 0.066 and 0.19, respectively. Compressed air was used and data were obtained over a range of stagnation pressures between 15 and 150 psia and at a stagnation temperature approximately equal to the ambient temperature of 540°R. Consequently, the flow was essentially adiabatic in the boundary layer region where the measurements were made. The boundary layer was turbulent at the nozzle inlet with a thickness of about ½ the inlet radius of the nozzle.

Flattened pitot tubes 0.005 in. high were used to measure impact pressures; the tubes were moved mechanically normal to the wall by a micrometer lead screw.

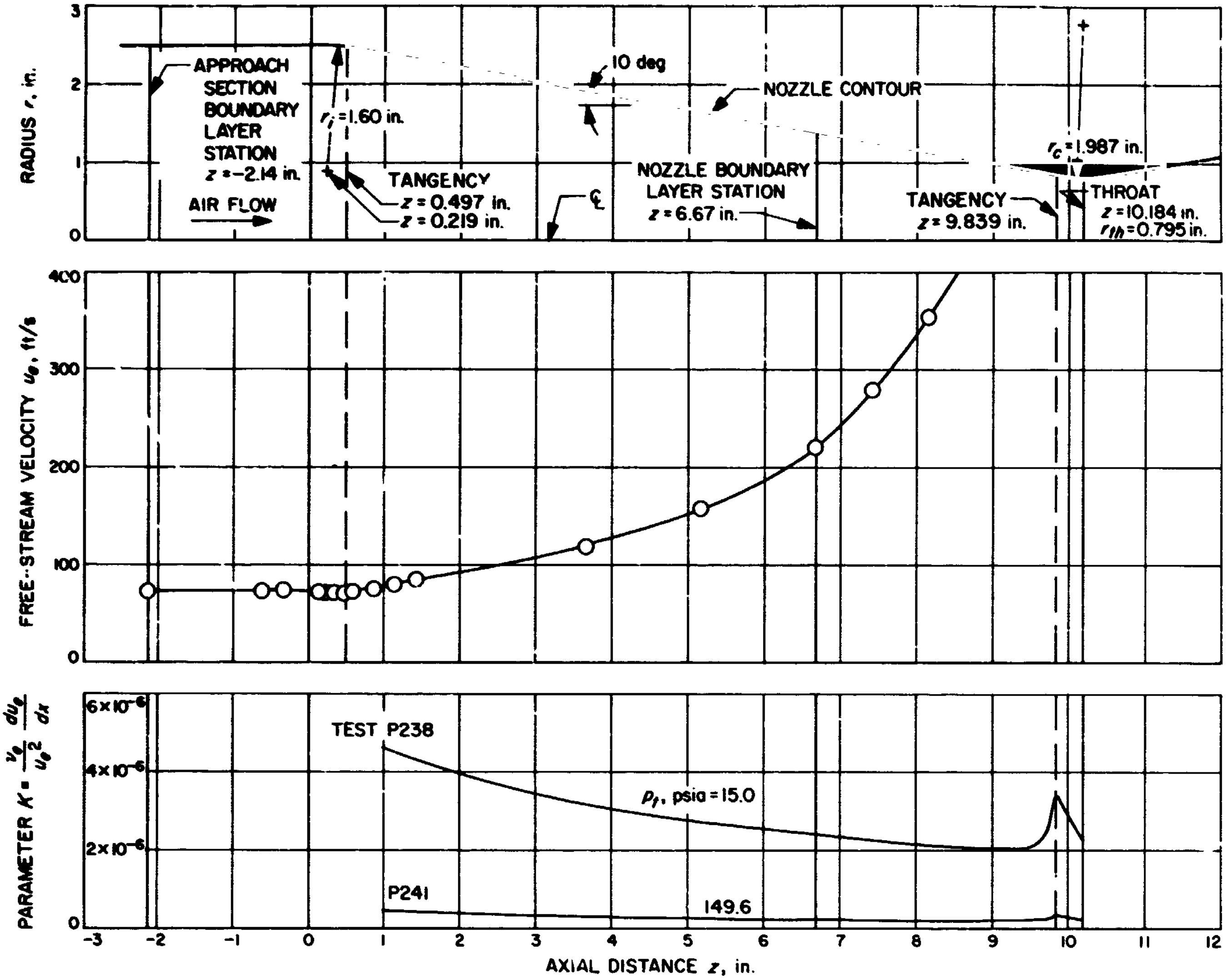


Fig. 1. Variation of flow variables

3. Experimental Results

The free-stream velocity variation obtained from the measured wall static pressures for isentropic core flow ($\gamma = 1.4$) is shown in Fig. 1. This distribution is essentially independent of stagnation pressure. The parameter K, indicated for two stagnation pressure tests, is highest in the inlet region of the nozzle. It then diminishes along the nozzle and is larger for the lower stagnation pressure test since $K \alpha (1/p_t)$ for nozzle flow.

At the approach section station, velocity profiles (Fig. 2) are seen to be typical of a turbulent boundary layer. In the representation of u^+ and y^+, the wall shear stress τ was determined in the approach section by fitting the profiles to the law of the wall which was taken in the form

$$u^+ = 5.5 + 2.5 \ln y^+, \qquad y^+ > 30 \qquad (1)$$

The velocity distribution is seen to agree well with the law of the wall relation. In the outer part of the boundary layer, the wake-like behavior found in many turbulent boundary layers is evident (Ref. 1).

The effect of flow acceleration on the velocity profiles is shown in the lower half of Fig. 2 at the nozzle station. At the higher stagnation pressure, the profile becomes relatively flat in the outer part of the layer. The wake-like behavior found upstream has disappeared and,

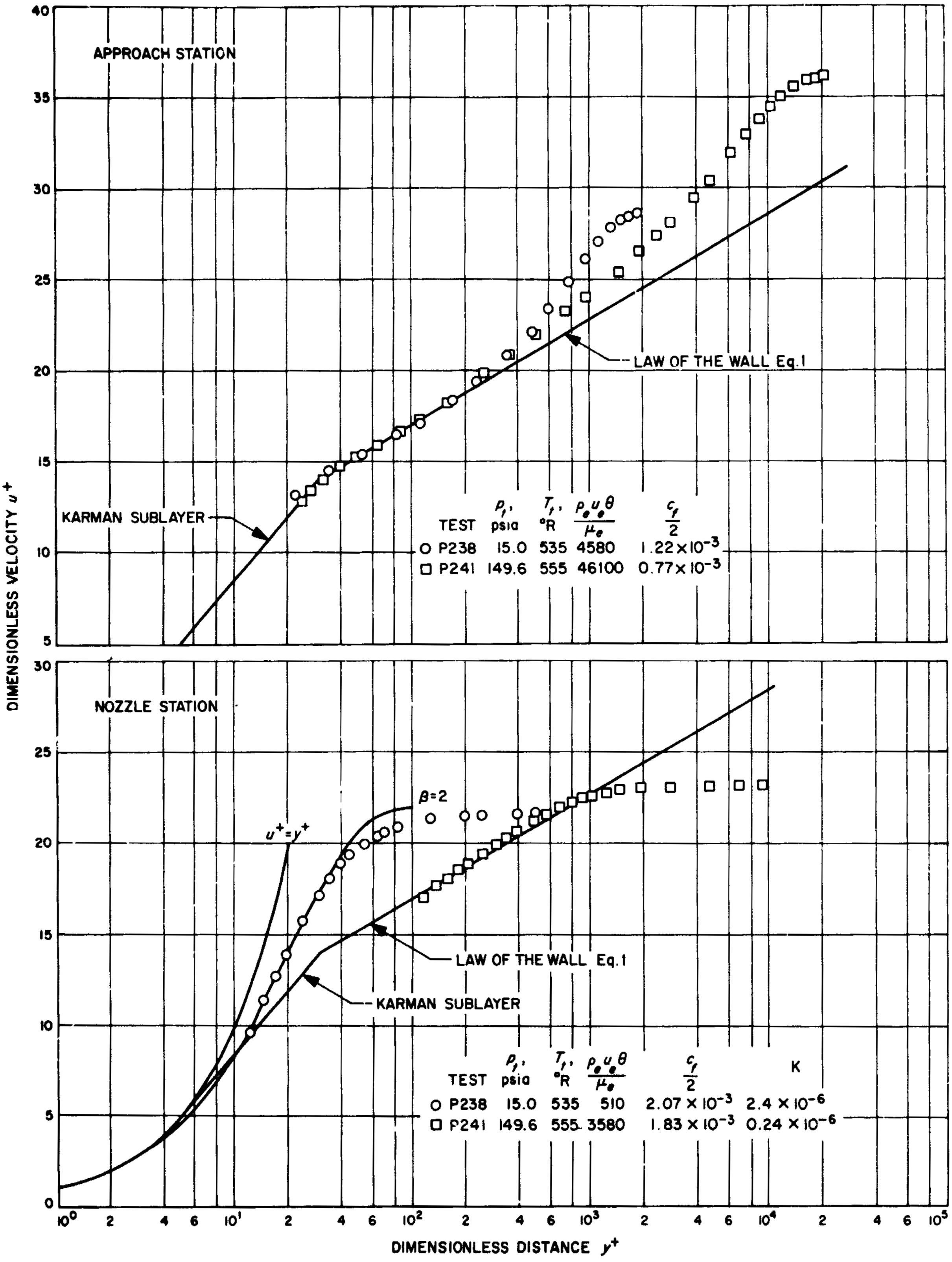


Fig. 2. Velocity profiles in the approach section and nozzle

although there is some curvature of the profile nearer the wall associated with the effect of acceleration, a fair fit is still found to the law of the wall. The friction coefficient of $c_f/2 = 1.83 \times 10^{-3}$ obtained from this fit is about 10% higher than the value that might be inferred from the Blasius turbulent boundary layer relation

$$\frac{c_f}{2} = \frac{0.0128}{\left(\frac{\rho_e u_e \theta}{\mu_e}\right)^{1/4}} \qquad (2)$$

The value of K corresponding to this higher pressure test is 0.24×10^{-6}.

A drastic change in the structure of the boundary layer in the nozzle occurred at the lower stagnation pressure where K is an order of magnitude higher (2.4×10^{-6}). The slope of the velocity profile (du/dy) is considerably reduced near the wall. In fact, the measurements near the wall can be linearly extrapolated to the wall, and the friction coefficient so deduced is $c_f/2 = 1.67 \times 10^{-3}$. This value is consistent with that obtained by fitting the Blasius flat plate laminar velocity profile $f'(\eta)$ (Ref. 2) to the measured values near the wall. The fit specifies η in terms of the experimental value of y/θ, and the friction coefficient is then determined from the slope of the exact solution f''_w at the wall:

$$\frac{c_f}{2} = \frac{\tau}{\rho_e u_e^2} = \frac{\mu_e}{\rho_e u_e^2}(u_e f''_w)\frac{d\eta}{d(y/\theta)}\frac{d(y/\theta)}{dy} = \frac{f''_w \dfrac{d\eta}{d(y/\theta)}}{\left(\dfrac{\rho_e u_e \theta}{\mu_e}\right)}$$

The Blasius profile, however, deviates from the measured profile at points away from the wall because the boundary layer that has apparently become laminar-like near the wall experiences flow acceleration. A better fit is afforded by the Hartree wedge flow profile for $\beta = 2$ (Ref. 3) shown in Fig. 2, and this profile yields a somewhat higher friction coefficient of $c_f/2 = 2.07 \times 10^{-3}$. Other accelerated laminar flow profiles for convergent channel or sink flow (Ref. 2), or perhaps more appropriately for conical channel or sink flow (Ref. 4), would fit the measured profile about as well as the Hartree profile for $\beta = 2$ and yield friction coefficients no more than 5% higher than that deduced from the Hartree profile.

To illustrate the reduction in the wall friction because of the apparent laminarization near the wall for the lower stagnation pressure test, the friction coefficient $c_f/2 = 2.07 \times 10^{-3}$ is about 25% below the value that might be inferred from the Blasius turbulent boundary layer relation Eq. (2). It is noteworthy that the reduction in wall friction occurred in a relatively high Reynolds number flow with the throat Reynolds number $[(\rho_e u_e D)/\mu_e]_{th} = 6.5 \times 10^5$ for the lower stagnation pressure test.

Table 1. Nomenclature

c_f	friction coefficient, $\frac{c_f}{2} = \frac{\tau}{\rho_e u_e^2}$
D	nozzle diameter
K	laminarization parameter, $\frac{\nu_e}{u_e^2}\frac{du_e}{dx}$
p_t	stagnation pressure
r	tube or nozzle radius
T_t	stagnation temperature
u	velocity component parallel to wall
u^+	dimensionless velocity, $\frac{u}{\left(\frac{\tau}{\rho_e}\right)^{1/2}}$
x	distance along the wall
y	distance normal to wall
y^+	dimensionless distance, $\frac{\rho_e\left(\frac{\tau}{\rho_e}\right)^{1/2} y}{\mu_e}$
z	axial distance
α	angle between wall and axis
γ	specific heat ratio
δ^*	displacement thickness $\delta^*\left(r - \frac{\delta^* \cos\alpha}{2}\right) = \int_0^\infty \left(1 - \frac{u}{u_e}\right)(r - y\cos\alpha)\,dy$
θ	momentum thickness $\theta\left(r - \frac{\delta^* \cos\alpha}{2}\right) = \int_0^\infty \frac{u}{u_e}\left(1 - \frac{u}{u_e}\right)(r - y\cos\alpha)\,dy$
μ	viscosity
ν	kinematic viscosity
ρ	density
τ	wall shear stress
Subscripts	
e	condition at free-stream edge of boundary layer

An indication of the region in which laminarization occurred near the wall in the nozzle at the lower stagnation pressure is shown in Fig. 2. Inference from the agreement with the Hartree profile for $\beta = 2$ suggests that the boundary layer was laminar-like out to a location where y^+ is about 30, a value associated with the viscous sublayer of a normal turbulent boundary layer and at which location laminar transport is small compared to turbulent transport. However, Launder (Ref. 5) still detected turbulent fluctuations close to the wall with his hot wire surveys in a similar laminarized boundary layer. Farther away from the wall, the velocity profile (Fig. 2) indicates that some turbulent transport still exists.

Thus, in the experiments discussed here, the velocity profiles measured upstream and within a conical axisymmetric nozzle revealed a strong effect of flow acceleration on the structure of an originally turbulent boundary layer. When values of the parameter K exceeded about 2×10^{-6}, the boundary layer became laminar-like near the wall because of flow acceleration, and the wall friction was correspondingly less than that associated with a turbulent boundary layer.

References

1. Coles, D., "The Law of the Wake in the Turbulent Boundary Layer," *J. Fluid Mech.*, Vol. 1, pp. 121–226, 1956.
2. Schlichting, H., *Boundary Layer Theory*, Sixth Edition. McGraw-Hill Book Co., Inc., New York, 1968.
3. Hartree, D. R., "On an Equation Occurring in Falkner and Skan's Approximate Treatment of the Equations of the Boundary Layer," *Proc. Cambridge Phil. Soc.*, Vol. 33, pp. 223–239, 1937.
4. Crabtree, L. F., Kuchemann, D., and Sowerby, L., "Three-Dimensional Boundary Layers," in *Laminar Boundary Layers*, p. 427. Edited by L. Rosenhead. Oxford University Press, New York, 1963.
5. Launder, B. E., *Laminarization of the Turbulent Boundary Layer by Acceleration*, Report No. 77. Gas Turbine Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1964.

B. Liquid-Metal MHD Power Conversion,

D. G. Elliott, L. G. Hays, and 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 MHD cycle being investigated, liquid lithium is (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 AC MHD generator; and (6) returned through a diffuser to the heat source. The cesium is condensed in a radiator or radiator-loop heat exchanger and returned to the nozzle by an MHD pump.

A 50-kW conversion system, which is to be operated with room-temperature NaK in place of lithium and nitrogen gas in place of cesium vapor, has undergone closed-loop tests with water and nitrogen. Cycle improvements have been studied and efficiencies of 8 to 11% were found to be theoretically possible through separator improvements or multistaging.

2. NaK-Nitrogen Conversion System

The conversion system was assembled without generator coils for water–nitrogen testing. Figure 3 shows the nozzle, separator, generator housing, diffuser, liquid return lines, nitrogen lines, and the starting and makup systems. The coils wrapped around the liquid return lines are heaters which will serve as the electrical load for the generator in the NaK-nitrogen tests and maintain the NaK at room temperature.

Twenty 5- to 10-min runs were made to determine the starting conditions and closed-loop operating limits. The system was started by turning on the nitrogen and then injecting water from the start tank at 140 psia and 120 lb/s while feeding 5 lb/s of water from the makeup-flow regulator which was set to maintain 150-psia nozzle inlet pressure. When the nozzle pressure exceeded 140 psia the start-tank flow stopped and back flow was prevented by check valves. The makeup regulator then continued to inject liquid until 150 psia was reached, after which the regulator continued to supply water to replace the 1.5 lb/s lost with the nitrogen. The start sequence required about 5 s. Various settings of the nitrogen flow rate and the start tank and makeup regulator pressures were tried in the first few runs until the smoothest pressure buildup was achieved. Pressure oscillations occurred with some settings and closed-loop operation was not sustained after the start-tank flow ceased.

After several runs the generator channel was inspected and it was found that the laminated vanes for eddy-current suppression at the generator inlet and three of the laminated slot plugs were missing. The tests were

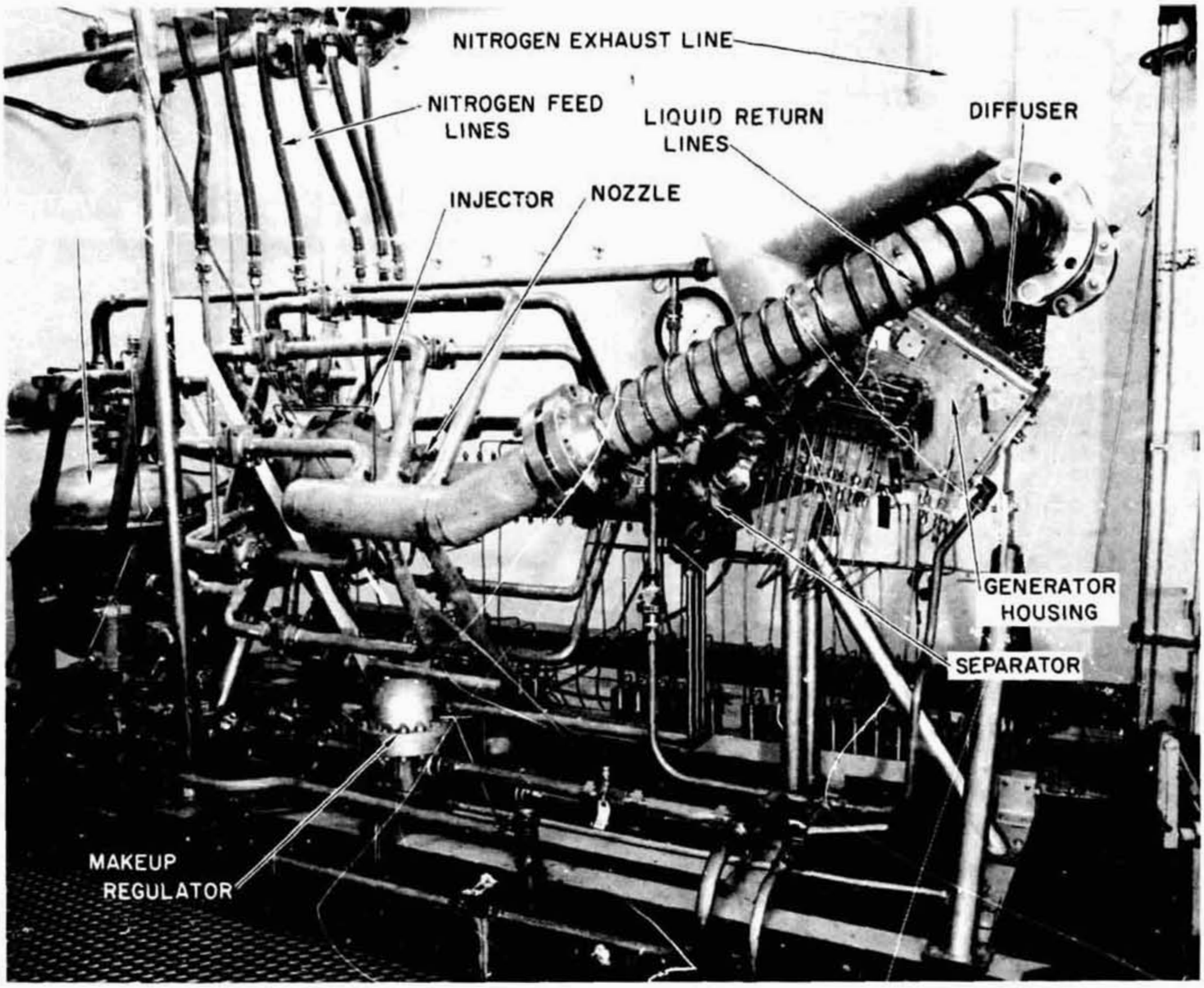


Fig. 3. Liquid-metal MHD reference system

continued and operation was obtained at several mixture ratios at each of the three nozzle pressures selected for the NaK tests: 150, 190, and 230 psia. The vanes and slot plugs were then replaced, and a second series of runs was made. Some vanes were again lost, and better anchoring techniques will be required in the NaK tests.

3. High Efficiency Cycles

Thermodynamically, liquid-metal MHD cycles using two components, such as cesium and lithium, and employing a regenerative heat exchanger between the cesium vapor and cesium condensate lines are limited only by the Carnot efficiency $1 - T_2/T_1$, since heat input and output are at essentially constant temperature. However, friction losses in the present design concept limit the efficiency to about 25% of the Carnot value at space powerplant conditions of $T_2/T_1 \cong 0.7$, or an efficiency of about 6% (half the efficiency of turbine and thermionic conversion systems).

The main friction losses are in the separator and generator and can be reduced in three ways: (1) decreasing the separator width to decrease the generator surface-to-volume ratio, (2) finding a method other than surface impingement for coalescing the flow, and (3) reducing the velocity of the liquid metal through multistage operation at reduced pressure ratio per stage.

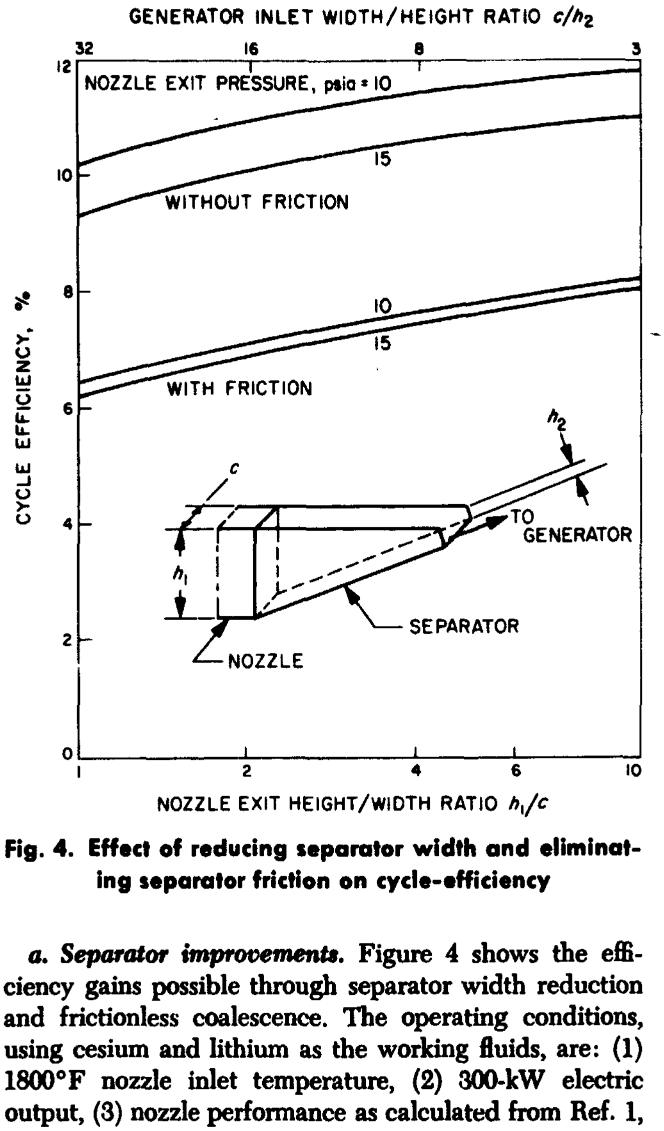


Fig. 4. Effect of reducing separator width and eliminating separator friction on cycle-efficiency

a. Separator improvements. Figure 4 shows the efficiency gains possible through separator width reduction and frictionless coalescence. The operating conditions, using cesium and lithium as the working fluids, are: (1) 1800°F nozzle inlet temperature, (2) 300-kW electric output, (3) nozzle performance as calculated from Ref. 1, (4) turbulent skin friction on the separator, (5) generator performance as given in Fig. 4 of Ref. 2 (compensated case), (6) 80% diffuser efficiency, and (7) 20% cesium pump efficiency. The lower curves show cycle efficiency as a function of height-to-width ratio h_1/c at the separator inlet for condensing pressures of 10 and 15 psia, the latter giving minimum radiator area. The cycle efficiency increases from 6.3% with a square inlet to 8.0% with a height-to-width ratio of 10. About 20% of the increase is due to the increased Reynolds number of the thicker liquid film on the separator surface, and the remainder is due to the increased width-to-height ratio of the generator channel c/h_2 (Fig. 4) which reduces the generator surface-to-volume ratio.

A separator with a square inlet, matching a circular nozzle exit, has been assumed in past cycle studies and has been the only type tested. Liquid impingement on the sidewalls of such a separator has been small and it may be possible to increase the height-to-width ratio to 4, for a 1-percentage-point efficiency gain, and even 10, for a 2-percentage-point gain, without excessive sidewall impingement. A nozzle with a ratio of 3.8 is being fabricated to investigate narrow separators.

The upper two curves in Fig. 4 show the efficiency attainable without separator friction. Even with a square nozzle exit, the efficiency is 9.3% at 15 psia condensing pressure and 10.2% at 10 psia. If, in addition, the frictionlessly coalesced liquid can be delivered to the generator at the more favorable aspect ratios, then cycle efficiencies could reach 11 to 12%. To determine to what extent such gains can be realized in practice, a pair of nozzles are being fabricated for impingement of the flows on each other instead of on a solid surface.

b. Multistage cycle. An efficiency improvement to the 9–11% range can be achieved without separator changes if power is extracted at intermediate stages of the expansion process. The improvement results from lower separator losses in stages at higher pressure and improved generator efficiency resulting from the lower liquid velocities.

A liquid-metal MHD cycle in which power is extracted in five stages is shown in Fig. 5. Lithium and cesium are expanded (at 1800°F) from 137 psia to a pressure of 88 psia in the first stage, producing a velocity of 245 ft/s. The two-phase mixture impinges on a separator where the lithium liquid is separated from the cesium vapor. The separated lithium enters an MHD generator at about 233 ft/s, and power is extracted at constant pressure, reducing the velocity to 50 ft/s. The lithium stream is then re-mixed with the separated cesium vapor in the second nozzle and the mixture is further expanded from 88 psia to 57 psia, giving a velocity of 240 ft/s. The process is continued through succeeding stages until the last stage is reached, where sufficient dynamic head is retained in the lithium at the generator exit to return the lithium through a diffuser to the heat source and first-stage nozzle. The cesium vapor from the last stage passes through a regenerative heat exchanger to the radiator (or other heat sink) where it is condensed. It is then returned by a pump through a regenerative heat exchanger to the first-stage nozzle.

An analysis was made of the performance of this cycle with 3, 5, and 7 stages for a few specific values of nozzle exit pressure and lithium-to-cesium mass ratio r_c. The

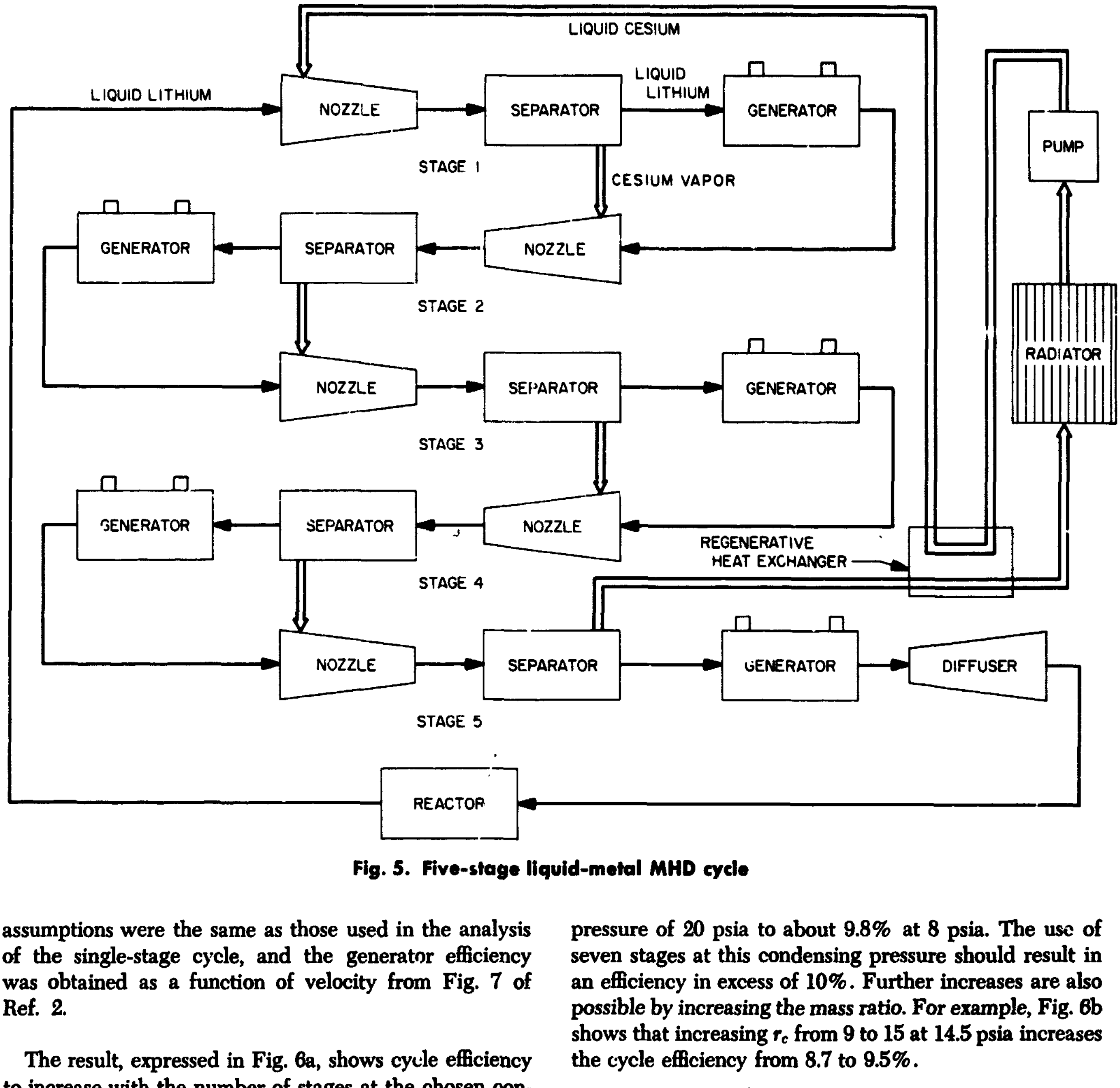


Fig. 5. Five-stage liquid-metal MHD cycle

assumptions were the same as those used in the analysis of the single-stage cycle, and the generator efficiency was obtained as a function of velocity from Fig. 7 of Ref. 2.

The result, expressed in Fig. 6a, shows cycle efficiency to increase with the number of stages at the chosen conditions of a mass ratio of 9 at a condensing pressure of 14.5 psia. The cycle efficiency increases from 6.3% for the single-stage reference design system to 9.3% for seven stages. For a condensing pressure of 9.5 psia, the efficiency rises from the single-stage value of 6.5% to 9.9% for seven stages. Further increases are attainable through variation of the condensing pressure and mass flow ratio, since the use of multiple stages lowers the frictional losses of the system. For five stages, Fig. 6b shows the efficiency to increase from 8.5% at a back pressure of 20 psia to about 9.8% at 8 psia. The use of seven stages at this condensing pressure should result in an efficiency in excess of 10%. Further increases are also possible by increasing the mass ratio. For example, Fig. 6b shows that increasing r_c from 9 to 15 at 14.5 psia increases the cycle efficiency from 8.7 to 9.5%.

Increasing the number of stages at constant condensing pressure reduces the specific radiator area in proportion to the increase in efficiency. For example, increasing the number of stages from one to seven decreases the isothermal, $\epsilon = 0.9$, radiator area from 3.8 to 2.4 ft^2/kWe. For a 300-kWe space power system, this would correspond to decreasing the isothermal radiator area from 1130 to 720 ft^2.

Multistage systems thus appear to offer considerable performance advantages over a single-stage system when

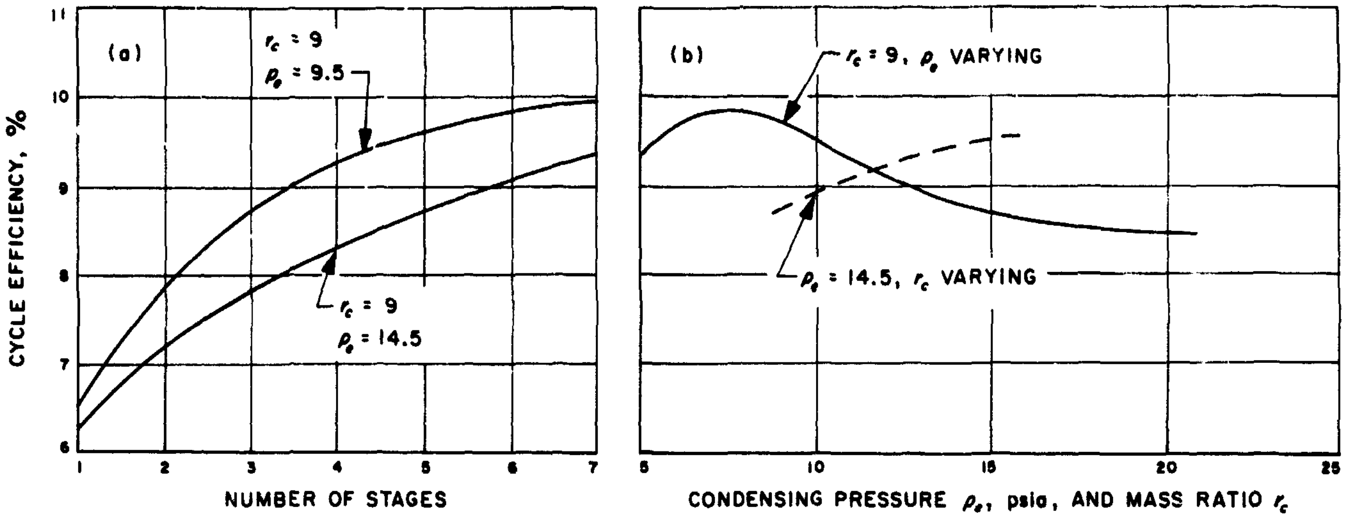


Fig. 6. Multistage cycle efficiency as a function of: (a) number of stages, and (b) condensing pressure and mass ratio

the major losses are taken into consideration. In addition to performance gains, increased reliability and operating life should be possible because of the lower liquid-metal velocities. For example, the reference single-stage system has a nozzle exit velocity of 513 ft/s while a five-stage system has an exit velocity of 245 ft/s. Furthermore, the nozzles in a five-stage system are subsonic so that further reduction in friction may be possible through reduction in separator area with the convergent flow.

References

1. Elliott, D. G., and Weinberg, E., *Acceleration of Liquids in Two-Phase Nozzles*, Technical Report 32-987. Jet Propulsion Laboratory, Pasadena, Calif. (in press).
2. Elliott, D. G., "Performance Capabilities of Liquid-Metal MHD Induction Generators," paper to be presented at the *Symposium on Magnetohydrodynamic Electrical Power Generation*, Warsaw, Poland, July 24–30, 1968.

C. Evaluation of the SE-20C Thruster Design,

T. D. Masek

1. Introduction

Improvements in thruster efficiency due to configuration changes have been reported in Refs. 1 and 2. These changes resulted in the SE-20B thruster design (solar electric 20-cm-diam thruster, modification B). Since many of the changes were made without complete thruster redesign, only minor consideration was given to weight, fabrication, and packaging. A new design (SE-20C) discussed in this article includes previous modifications but with variations required to reduce weight, provide strength, and ease assembly and mounting. Since these small variations in design might change thruster performance, the SE-20C thruster must be evaluated in detail. Thruster construction, weight, and performance are considered in this work.

2. Thruster Construction

The basic elements of the present 20-cm-diam thruster are shown in Fig. 7. The general size and shape of the ferromagnetic elements were established in Refs. 3 and 4. As in the initial design (Ref. 1), assembly ease and grid alignment were basic considerations. Use of previous grid designs was also required to allow interchangeability and to avoid the expense and time of new grid fabrication. Thus, the specific dimensions of the housing, anode, support rings and brackets were determined by the existing grid design.

Front and rear support rings mount the bar electromagnets and provide a magnetic flux path. Bar electromagnets were chosen (1) to provide for the possibility of using permanent magnets, (2) to allow the magnetic field to be adjusted in performance mapping, and (3) for low power since the magnetic flux is used more efficiently than with conventional solenoidal designs.

The mount assembly was designed to mate with the gimbal elements of the thrust vector alignment system (Ref. 5). High-voltage isolation is included in the mount assembly by four Alite insulators. Propellant is introduced, as in previous designs, through the side of the thruster at the center of the anode.

3. Weight Summary

A weight breakdown for the SE-20C thruster is presented in Table 2. The total weight of 4.06 kg (8.96 lb) includes a ground screen, connector halves, 10,000-h (estimated life) grids, and feed system up to the vaporizer.

Table 2. SE-20C thruster weight summary

Component	Weight, g
Housing	378
Screen grid pole piece	135
Support ring, forward	114
Support ring, aft	128
Anode	300
Rear plate	240
Cathode pole piece	70
Cathode (Hughes oxide)	135
Screen grid	106
Accelerator grid	675
Magnet (8 each)	640
Accelerator mount assembly (8 each)	168
Ground screen, forward assembly	128
Ground screen, aft assembly	246
Anode and ground screen insulators	40
Mount assembly (pad, insulators, and cover)	206
Connector halves	255
Feed system (vaporizer, isolator, and manifold)	100
Total	4064

The need for ferromagnetic parts places certain restrictions on thruster weight. Aluminum has been used in certain parts as indicated in Fig. 7 but cannot be used extensively. Additional weight reductions (approximately 10%) appear possible by reducing thicknesses. However, the effect of these reductions on the magnetic field shape and strength (or power) and on structural strength must be evaluated.

4. Test Results

a. Grid stability. Initial testing of the SE-20C thruster resulted in relative high discharge efficiency but showed high accelerator impingement for flow rates above 6 g/h. The impingement could be reduced bv increasing the total ion beam accelerating voltage (from 4.0 to 5.5 kV at 6 g/h). This indicated that the grid spacing, nominally 0.178 cm, had increased substantially.

Bench tests were conducted using dial indicators to measure screen and accelerator deflections. The grids were heated to simulate cathode and plasma radiation heating using lamps and a heat gun. The results of these tests are as follows:

(1) The accelerator deflected up to about 0.025 cm toward the screen when heated in the center region. As the housing and outer portion of the accelerator were heated, the deflection decreased.

(2) The direction of screen grid deflection depended upon its initial setting. When heated centrally, deflections up to 0.125 cm occurred in the direction of the initial bow. As with the accelerator, heating the housing reduced the screen deflection. Since fabrication always produces a slight bow, the initial assembly must force the screen to deflect toward the accelerator. This reduces the grid spacing with heating and is much more desirable than increased spacing.

As a result of these bench tests, a method for providing an initial positive deflection (toward the accelerator) was devised. The outer 0.475 cm of the housing side of the screen grid was chamfered at an angle of 1.5 deg. This slight chamfer produced an initial bow of about 0.05 cm at the center. Thruster operation with this configuration (with a 0.178-cm spacing at the outer edge) showed low impingement rates at all flow rates. However, the close spacing, probably as low as 0.05 cm during start-up or fast power level changes, caused sparking between the grids.

In addition to low impingement with the pre-bowed configuration, the thruster could be operated with lower accelerating voltages. A beam current of 1.0 A was obtained with a total voltage of 3.5 kV. This result verified the conclusion that the initial impingement difficulties and observations were caused by a large grid spacing.

Previous difficulties with high impingement rates have been attributed to magnetic field or plasma density distributions. Many of these problems may be resolved with the more controlled grid configuration.

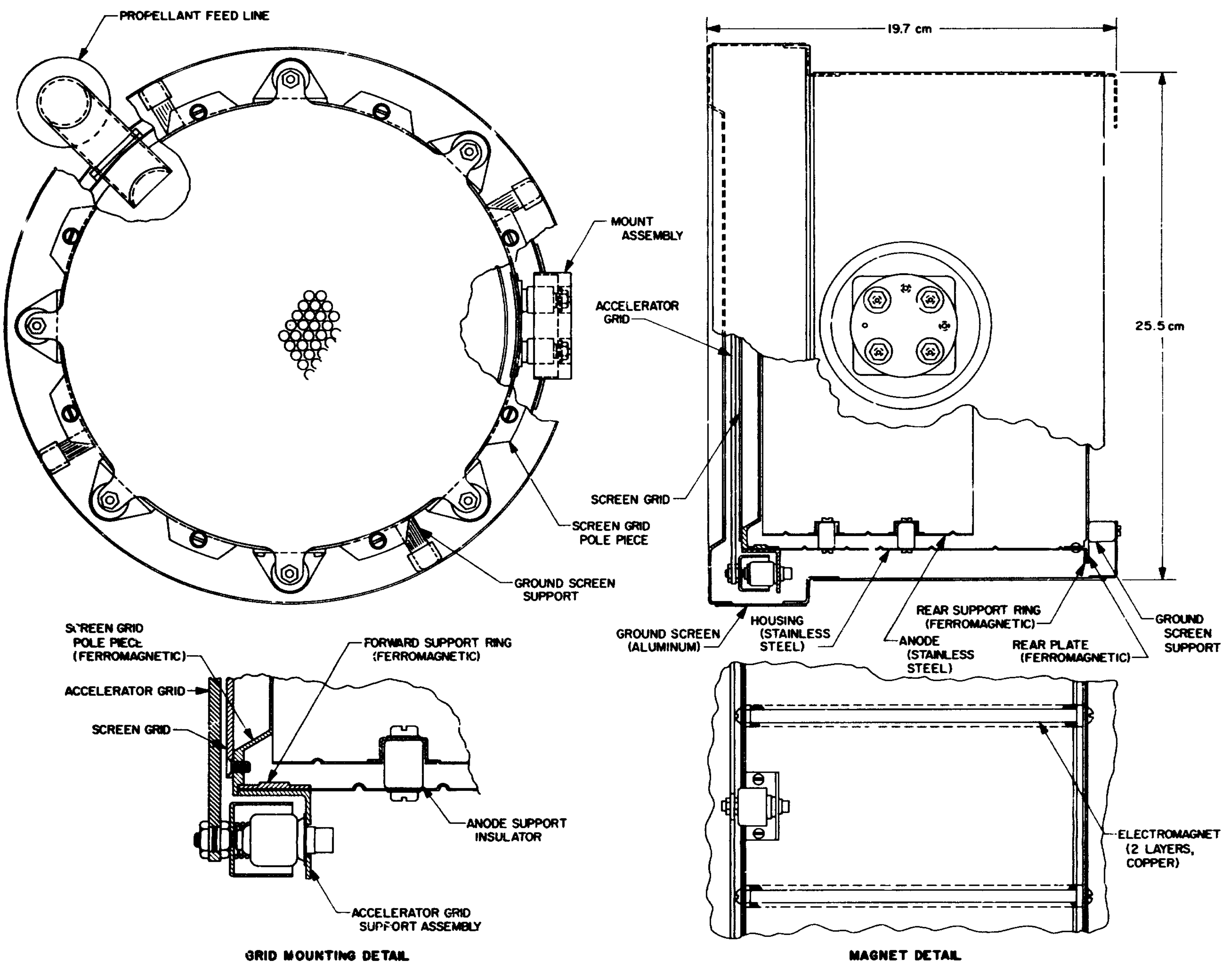


Fig. 7. Basic design of the SE-20C thruster

b. Performance. Thruster performance can be easily evaluated by considering only the discharge loss per beam ion. All other losses, although significant in determining thruster efficiency, are not important in comparing the SE-20B and SE-20C designs.

Discharge loss as a function of propellant utilization, propellant flow rate, and magnet current is presented in Fig. 8. A comparison of this data with that obtained in the SE-20B thruster (Ref. 2) is shown in Fig. 9 for a magnet current of 2.0 A. Higher losses (about 15 eV/ion at 80% utilization) and higher slopes are indicated for the SE-20C thruster. Since both magnet designs are nearly identical, the difference in performance is attributed to the minor differences in the ferromagnetic parts (thicknesses and construction).

The discharge losses of the SE-20C at slightly higher field are equivalent to the SE-20B thruster as shown in Fig. 8. The higher loss is attributed to a somewhat higher magnetic flux resistance in the new design due to thinner ferromagnetic elements. With the small differences noted, performance of the SE-20C is quite similar to the SE-20B design.

References

1. Masek, T. D., *Experimental Studies With a Mercury Bombardment Thruster System*, Technical Report 32-1280. Jet Propulsion Laboratory, Pasadena, Calif. (in press).

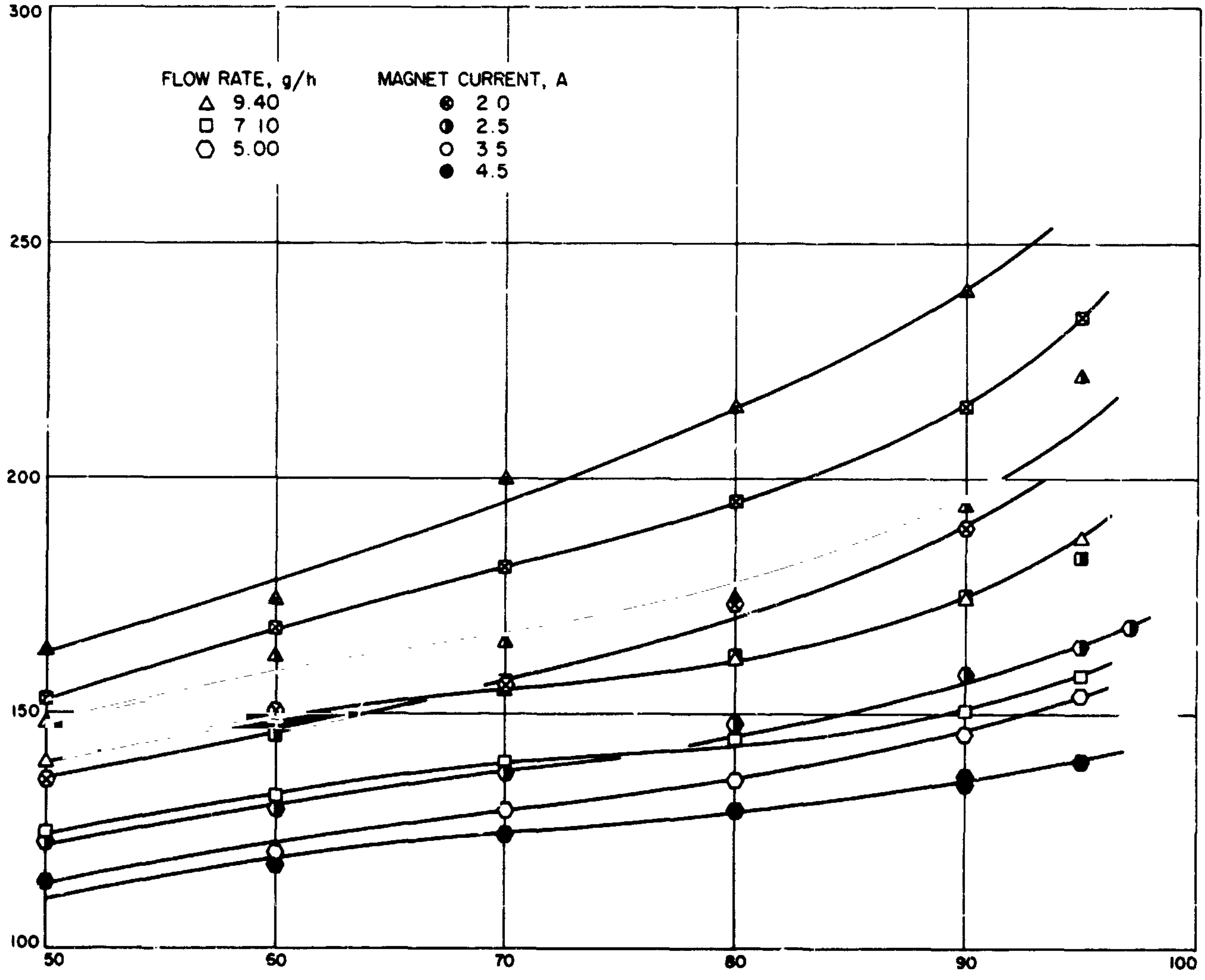


Fig. 8. SE-20C thruster performance data

2. Masek, T. D., and Pawlik, E. V., "Thrust System Technology for Solar Electric Propulsion," AIAA Paper 68-541, AIAA Fourth Propulsion Joint Specialists Conference, Cleveland, Ohio, June 10, 1968.

3. Bechtel, R. T., "Discharge Chamber Optimization of the Sert II Thruster," AIAA Paper 67-668, AIAA Electric Propulsion and Plasmadynamics Conference, Colorado Springs, Colo., Sept. 1967.

4. Pawlik, E. V., *Scaling of a High-Performance Ion Thruster*, Technical Memorandum 33-387. Jet Propulsion Laboratory, Pasadena, Calif., Apr. 1968.

5. Reader, P. D., and Mankovitz, R. J., "Attitude Control of an Electrically Propelled Spacecraft Using the Prime Thrust System," paper to be presented at the ASME 1968 Aviation and Space Conference, Los Angeles, Calif., June 1968.

D. Radial Distribution of Enthalpy in a High-Temperature Swirling Flow, *P. F. Massier*

1. Introduction

In arc heaters and plasma electrical propulsion devices, gas is sometimes injected tangentially upstream of the electrodes in order to introduce swirl into the flow. Although certain advantages may be gained from the swirl, one of the disadvantages may be an increase in the convective heat transfer rates as shown in SPS 37-24, Vol. IV, pp. 105–108, for flows through nozzles. Consequently, swirling flows are being investigated to acquire a better understanding of heat transfer to electrode and other surfaces so that improvements can be made in

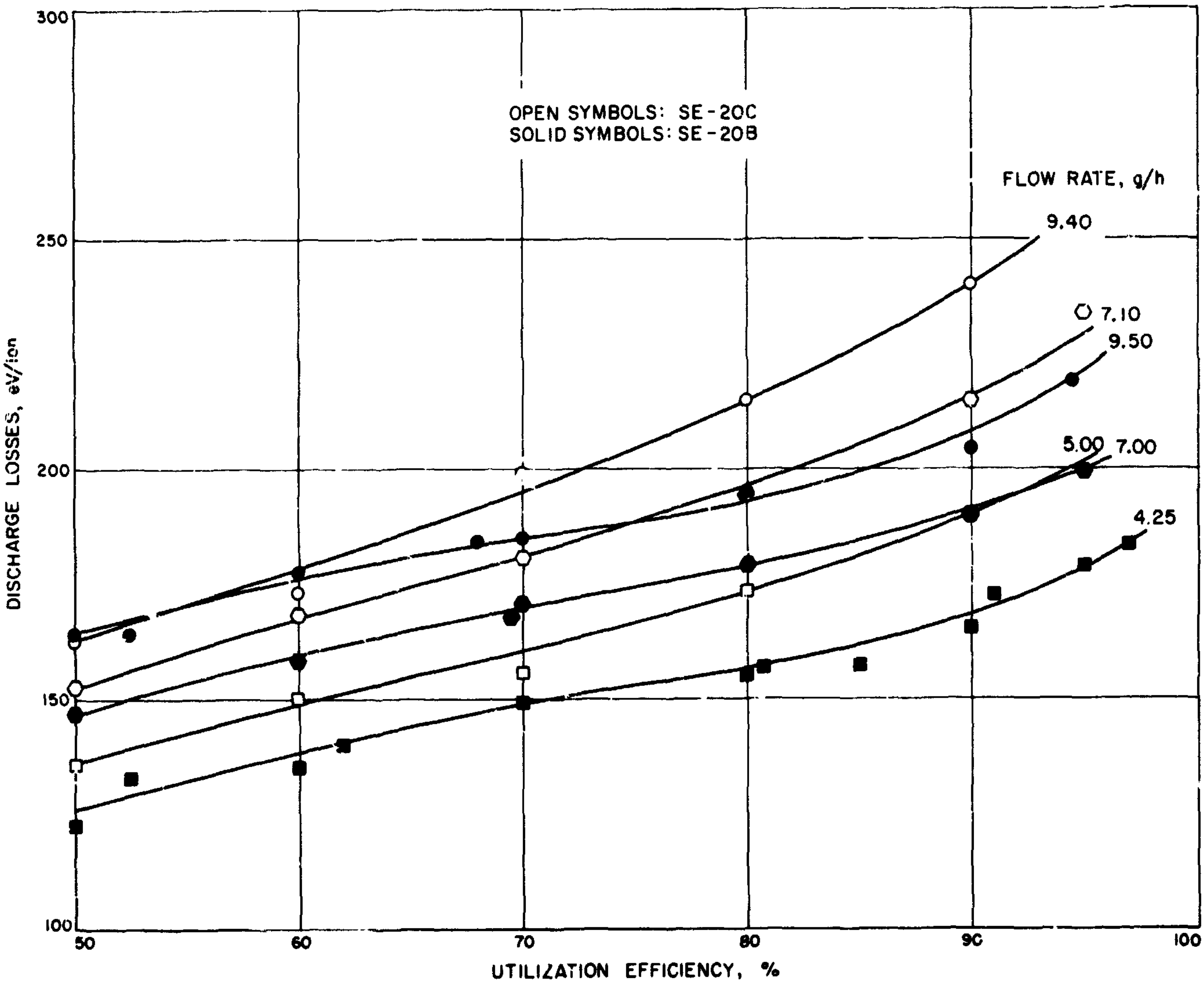


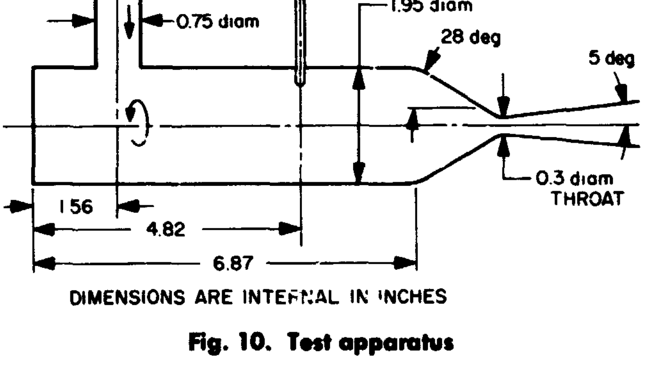
Fig. 9. Comparison of SE-20C and SE-20B thruster performance data

predicting the cooling requirements of plasma devices. Other effects also being investigated but not discussed here include severe wall cooling, acceleration, ionization, and applied magnetic and electric fields.

From a heat transfer viewpoint the important flow variable is the radial distribution of the enthalpy. When evaluated at the wall, the slope of the enthalpy is related to the wall heat flux. This distribution is generally dependent on many factors; in particular, for a swirling flow it depends upon the amount of swirl, i.e., the ratio of tangential to axial velocity. Consequently, a knowledge of the enthalpy distribution is essential for evaluating the theoretical methods now being advanced for predicting the convective heat transfer. The discussion in this article pertains primarily to the feasibility of using a calorimetric probe to determine the radial distribution of the enthalpy in a confined swirling flow of a high-temperature gas.

2. Experimental Apparatus

The experimental apparatus (Fig. 10) was fabricated to evaluate radial distributions of enthalpy and tangential velocity, and longitudinal distributions of wall heat flux in a constant-diameter duct. Arc-heated argon enters the duct through one port near the endwall. The gas then flows through the duct and discharges through a convergent–divergent nozzle attached to the other end. After leaving the nozzle the gas flows into a vacuum system. Several tests have been conducted in which the enthalpy distribution was obtained by radially traversing a calorimetric probe at the location shown in Fig. 10. Details of the probe and the associated data analysis procedure appear in SPS 37-46, Vol. IV, pp. 153–161. The probe used in the swirling flow investigation was straight with the tip pointing in the radial direction; hence, the local impact and static pressures were not measured.



DIMENSIONS ARE INTERNAL IN INCHES

Fig. 10. Test apparatus

The walls of the apparatus consisted of many individual circumferential coolant passages for determination of the wall heat flux distribution and the endwall contained numerous pressure taps for the purpose of evaluating the radial distribution of the tangential velocity. The velocity and the heat transfer results are not discussed here, however.

3. Results

The distribution of the enthalpy as determined by the calorimetric probe is shown in Fig. 11 for one test in which the pressure in the duct was 3.9 psia and the stagnation temperature was approximately 3000°R. Trends of the other tests are similar. At the probe location the Reynolds number of the main gas stream was 490 based on the average mass flux and duct diameter with viscosity evaluated at the average free-stream temperature. The Mach number based on the average axial velocity was 0.01 and the ratio of the maximum tangential to axial velocity was approximately 5.

Figure 11 indicates a symmetrical enthalpy distribution and shows that the edge of the thermal boundary layer was approximately 0.1 in. from the wall. The maximum values on either side of the centerline resemble the trends in stagnation temperature distributions observed in vortex flows near room temperature that are discussed in SPS 37-33, Vol. IV, pp. 133–141. It has been verified, however, that probes introduced into a swirling flow can have a significant effect on the flow field (Refs. 1 and 2). An influence of the probe on the enthalpy distribution shown in Fig. 11 appears to be evident when comparing the integrated average enthalpy based on the probe data with the average obtained by an energy balance which takes into account the applied electric power and heat transferred to the coolant. The probe average is about 17% lower than the enthalpy determined from energy balance. A comparison of the average enthalpies obtained by these two methods in nonswirling flows shows better agreement (SPS 37-47, Vol. III, pp. 103–116, and SPS 37-46, Vol. IV, pp. 153–161). The low probe readings

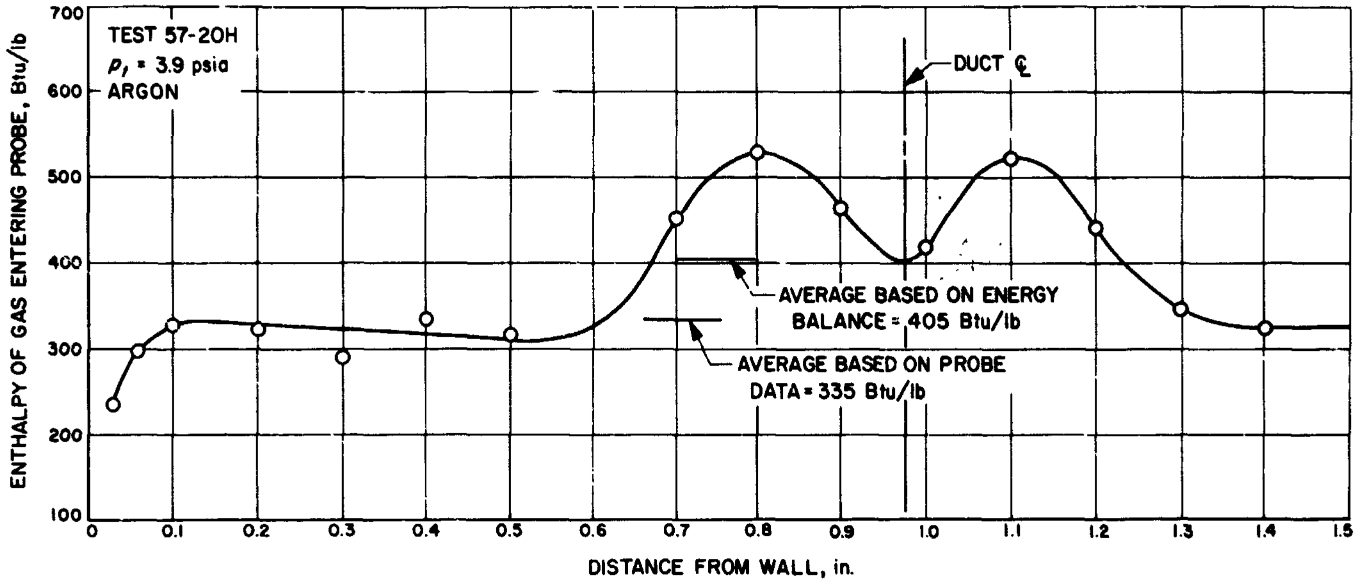


Fig. 11. Radial distribution of enthalpy

may have resulted from the integrated average enthalpy being based on $\int H_t\, dA$ instead of $\int \rho u\, H_t\, dA$. Available information was insufficient to determine the mass flux (ρu) distribution. It is also possible that the low probe average was caused by some of the cool gas in the boundary layer near the duct wall flowing radially inward along the outer wall of the probe tube and then entering the probe tube during sampling. Thus, the heat transfer measurements that are made on the sampled gas would indicate a lower enthalpy of the main gas stream at a particular radial position than would exist if the probe were not in the duct. The gas at the outer radii of the duct has a tendency to flow radially inward along the probe wall because of a reduction in tangential velocity caused by the boundary layer formed on the probe. Thus, locally, the radial pressure gradient $\partial p/\partial r$ is not balanced by the centripetal acceleration $\rho v^2/r$ maintained by the tangential velocity and, hence, radial flow occurs. Such radial flow can also occur in the wake of the probe.

4. Conclusions

Despite the apparent low average enthalpy determined from the probe data, the location of the edge of the thermal layer and the general distribution of the enthalpy are significant results. Near the duct wall the radial pressure gradient is comparatively small; hence, the radial flow there would not be large and the value of the enthalpy at the edge of the boundary layer is probably realistic.

References

1. Roschke, E. J., "Flow-Visualization Studies of a Confined, Jet-Driven Water Vortex," Technical Report 32-1004. Jet Propulsion Laboratory, Pasadena, Calif., Sept. 15, 1966.
2. Pivirotto, T. J., "An Experimental and Analytical Investigation of Concentration Ratio Distributions in a Binary Compressible Vortex Flow," Technical Report 32-808. Jet Propulsion Laboratory, Pasadena, Calif., Mar. 15, 1966.

E. Some Effects of an Applied, Transverse Magnetic Field on Heat Transfer With Swirling and Nonswirling Gas Flow, *E. J. Roschke*

1. Introduction

An apparatus for studying convective heat transfer from partially ionized gases in a transverse magnetic field was described in SPS 37-47, Vol. III, pp. 120–128. Modifications of this apparatus and some preliminary heat transfer results were discussed in SPS 37-49, Vol. III, pp. 199–201. This work is an initial step towards increasing the understanding of energy transfer processes that occur when a flow of ionized gas interacts with electric and magnetic fields. Such information is important in the prediction of electrode heat transfer and is also necessary for the design of magnetogasdynamic generators and propulsion devices such as magnetoplasmadynamic arcs. The purpose of this article is to present the effects of both magnitude and direction of an applied, transverse magnetic field on heat transfer from partially ionized argon that have been investigated in two tests, with and without swirl in the flow.

Symbols used in this article are defined in Table 3.

2. Description of Apparatus

The in-line arc configuration used for the present heat transfer experiments is shown in Fig. 12. The portion of the apparatus of immediate interest is the 2- × 2-in. square channel which is approximately 13 in. long. An inlet section is provided to promote adequate mixing and flow development of the high-temperature gas stream supplied by the electric arc heater. The test section (total length 4 in.) is the downstream portion of the channel; the four walls of each 1.0-in.-long segment are individually cooled so that heat transfer may be determined by calorimetry. [The walls are designated A, B, C, and D, clockwise, looking downstream (Fig. 12).] Flow is exhausted from the system by means of a 2.88-in.-diam circular duct approximately 19.3 in. long. All experiments are conducted at the short-circuit condition with zero load factor.

Table 3. Nomenclature

b	channel height, 2 in.
k	thermal conductivity of gas (Ref. 1)
$\dot{m}$	mass flow rate of gas
p	static pressure, absolute
q	heat flux
Q^*	non-dimensional heat flux
Q_0^*	non-dimensional heat flux at zero magnetic field
Re	Reynolds number based on mass flow rate, for square channel Re = $\dot{m}/\mu b$
T_i	inlet gas temperature, at center of first test-section segment
T_w	gas-side wall temperature
μ	gas viscosity (Ref. 1)

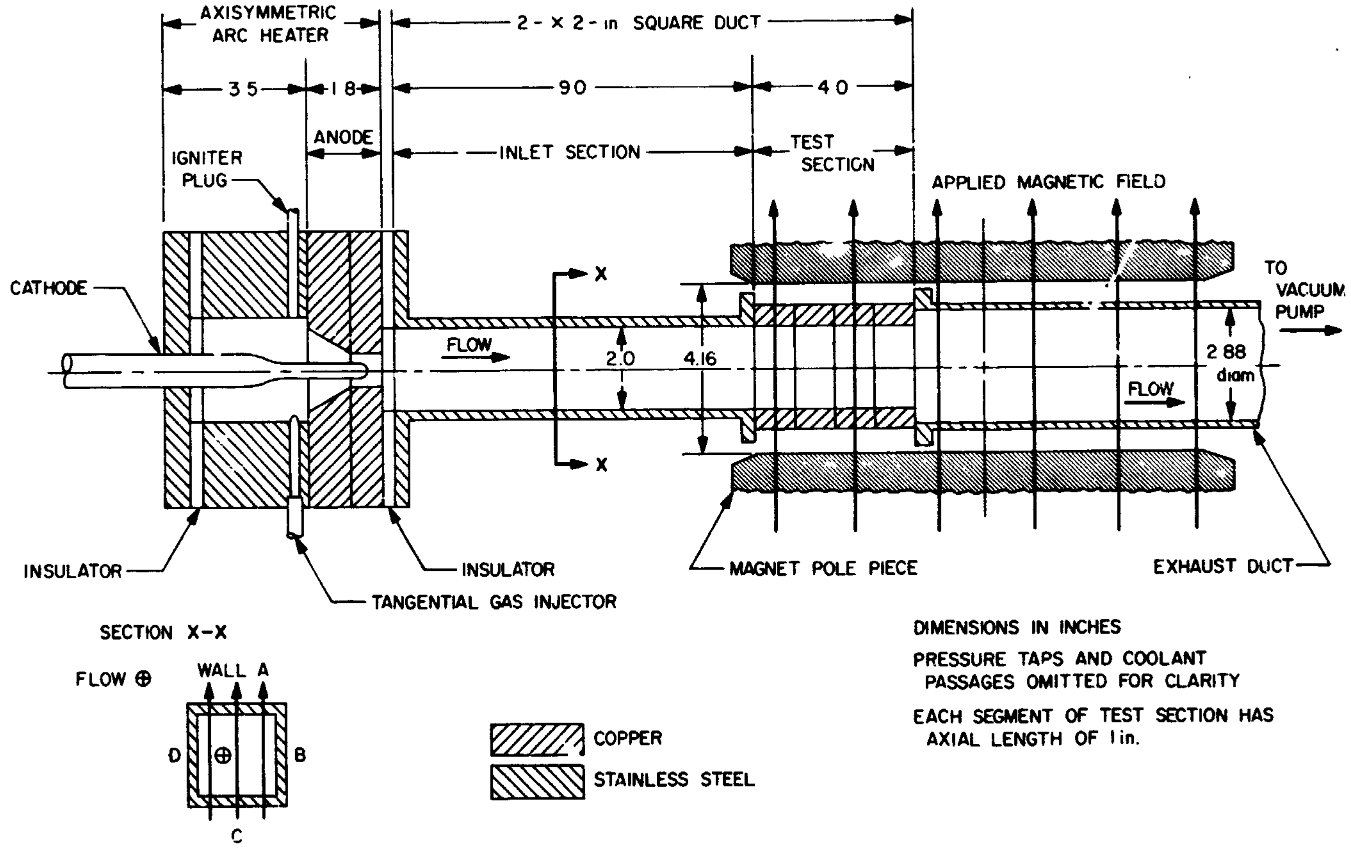


Fig. 12. In-line arc configuration for heat transfer experiments (side view)

The axial location of the test section with respect to the magnet pole pieces is also shown in Fig. 12. Heavy arrows indicate the normal direction of the applied magnetic field, termed "forward field." In the case of "reverse field," the arrows point in the opposite direction to that shown.

Experience has shown that tangential gas injection upstream of the anode (Fig. 12) is generally superior to radial injection because higher arc efficiencies are obtained for the same applied electric power. Thus, more heat may be added to the gas resulting in a higher temperature stream. In addition, it has been found that improved stability may be obtained at higher power levels. For this reason, most of the experiments have been obtained using tangential injection. (The direction of the gas injection in Fig. 12 would be clockwise, looking downstream.) However, the presence of swirl in the flow complicates the interpretation of data as well as any theoretical analysis that might be attempted; therefore, comparisons of data with and without swirl are desirable. Currently, a series of tests employing radial injection is being conducted. The first results are presented here.

3. Method of Presenting Data

To study the effect of the magnitude of the applied magnetic field on heat transfer, it is necessary to find some basis of comparison for a series of tests in which the only parameter varied deliberately is the magnitude of the applied field. The reason for this is that changes in the applied field produce internal changes in the gas which are often accompanied by changes in the voltage in the electric arc-heater. Thus, the initial energy content and temperature of the gas usually varies considerably with varying magnetic field. An approximate correction for this is obtained by using the non-dimensional heat flux defined by

$$Q^* = \frac{qb}{k(T_i - T_w)}$$

(Also see the theoretical analysis of Back, Ref. 2.) In the present application, T_i is obtained as a bulk or average value of temperature at the center of the first test-section segment by means of an energy balance applied to the system up to that axial location and by use of a mollier chart for argon. At that axial location, the magnitude of the magnetic field is 94% of the peak value. The thermal conductivity of the gas is obtained at T_i and the local wall pressure using the results given in Ref. 1 Gas-side wall temperatures are of the order of 100°F in these experiments. Heat transfer results are presented for the second segment of the test section, however, using T_i as discussed.

To isolate and clarify the effect of magnetic field still further, values of Q^* are normalized with respect to their values for each wall when the magnetic field is zero. Thus, the parameter used is Q^*/Q^*_0 which, in effect, reduces the results for the four walls to a comparable base value so that trends with varying magnetic field are more easily evaluated.

Changes in heat transfer brought about by the magnetic field through joule heating are independent of the direction of the induced current in the gas (only its magnitude, Ref. 2). Thus, the vertical orientation of the applied, transverse field is theoretically unimportant when the gas has axial motion alone and there are no Hall effects to cause transverse Lorentz forces. With swirl present, this would not be necessarily true. Two tests have been selected for presentation, one utilizing tangential injection and the other utilizing radial injection. Conditions for these tests are given in Table 4.

Table 4. Nominal test conditions at zero magnetic field

Parameter	Test 107-18H (Tangential) injector	Test 107-28H (Radial) injector
Applied power to electric arc, kW	49.5	34.8
Actual heat input to gas, Btu/s	31.2	18.1
Mass flow rate $\dot{m}$, lb/s	0.007	0.007
Inlet static pressure p_i, psia	0.85	0.83
Inlet static temperature T_i, °R	16,700	11,750
Inlet Reynolds number, Re	260	280

4. Experimental Results and Discussion

With the tangential injector, it has been generally found that the largest absolute changes in wall heat flux due to the applied magnetic field occur for reverse field. Also, the sidewalls generally experience relatively greater changes than the upper and lower walls. The non-dimensional heat-flux ratio Q^*/Q^*_0 for test 107-18H is shown in Fig. 13a. Although there is some scatter, the trends of the data are relatively clear. Heat transfer to the upper and lower walls tends to increase with increasing magnetic field regardless of the direction of the field.

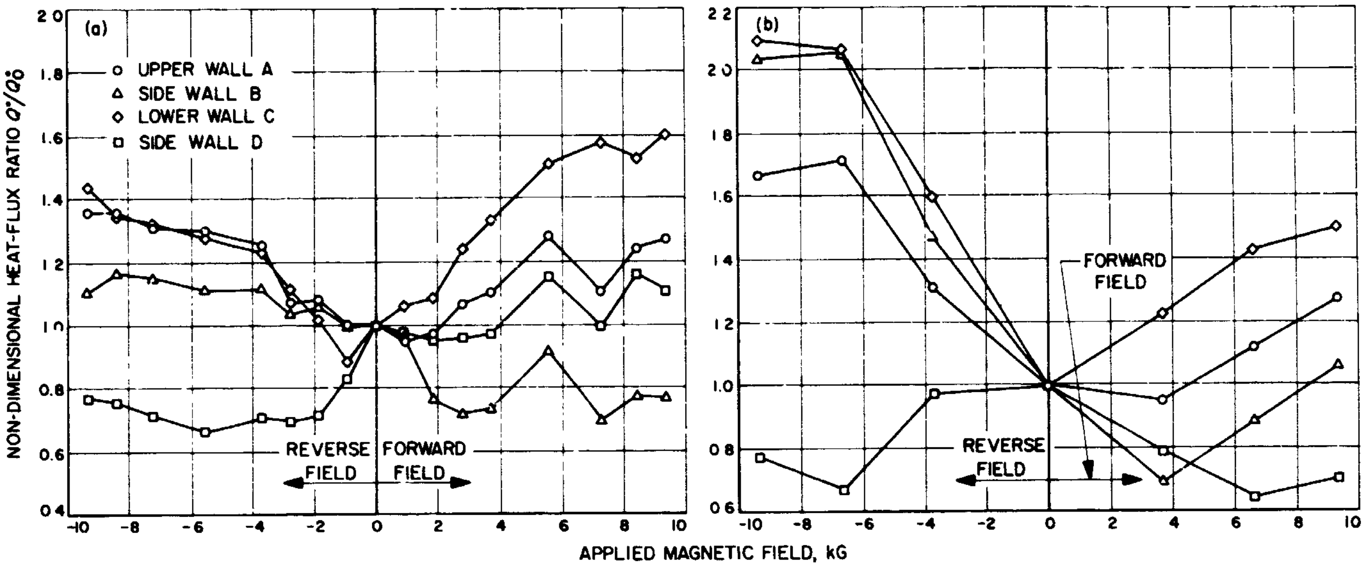


Fig. 13. Heat transfer test results using: (a) tangential injector, test 107-18H; (b) radial injector, test 107-28H

This trenu agrees with the predicted trend, (Ref. 2). The maximum effect measured was a 60% increase in heat transfer on lower wall C. Trends for sidewalls B and D are different; wall B experiences a marked decrease in heat transfer with increasing forward field but wall D experiences a marked decrease with increasing reverse field, and conversely. This behavior for the sidewalls is thought to be associated with a Hall effect and tends to agree with the lateral (side-to-side) deflection of the exhaust plume observed visually; i.e., an observed motion of the plume toward one wall coincides with an observed increase in heat transfer at that wall but a more significant decrease in the heat transfer at the opposite wall from which the plume was deflected. A prediction for the deflection of the gas stream due to a Hall effect is difficult to make in this case because of the consequences of swirl.

Comparable data using a radial injector at considerably lower power levels and gas temperatures are shown in Fig. 13b for test 107-28H. The trends of the curves for upper and lower walls agree with that obtained for the tangential injector, i.e., increasing applied magnetic field tends to increase the heat transfer at those surfaces. Results for the sidewalls are somewhat different; heat transfer to wall D was decreased regardless of direction of field, whereas wall B experienced an increased heat transfer for reversed field but a decrease for forward field. Visually observed deflections of the gas were not pronounced in this test although deflection towards walls B and C were noted for forward field. A noticeably stronger gas deflection towards wall B was detected with reverse field.

5. Conclusions

Based on the limited results obtained in this study, the following conclusions are made:

(1) Surfaces transverse to the applied magnetic field experience an increase in heat transfer with increasing field either with or without swirl in the flow regardless of the orientation of the field.

(2) The largest increases observed are 60% in the case of flow with swirl and 60 to 100% without swirl compared to results with zero magnetic field.

(3) Significant changes in heat transfer for walls parallel to the field occur and may be positive or negative depending on field orientation and the presence or absence of swirl. These observations are thought to be associated with Hall effects.

References

1. deVoto, R. S., *Argon Plasma Transport Properties*, Technical Report 217, Department of Aeronautics and Astronautics, Stanford University, Stanford, Calif., Feb. 1965. Also available in *Phys. Fluids*, Vol. 10, pp. 354–364, Feb. 1967.

2. Back, L. H., "Laminar Heat Transfer in Electrically Conducting Fluids Flowing Between Parallel Plates," paper accepted for publication in *Int. J. Heat Mass Transfer*.

F. Some Effects of an Applied, Transverse Magnetic Field on Wall Pressure in a Square Channel, *E. J. Roschke*

1. Introduction

Some heat transfer measurements for partially ionized argon flowing in a square channel with a transverse magnetic field were presented in *Section E.* The purpose of this article is to present wall pressure measurements indicating some of the effects produced by varying both magnitude and direction of an applied, transverse magnetic field on pressure within the channel. These are companion results to those presented in the previous article for test 107-18H and, therefore, apply for the case of swirl present in the flow.

2. Experimental Apparatus and Measurements

The arrangement of apparatus was identical to that shown in Fig. 12. (*Section E*); the designations of the four walls of the channel are retained. Static pressure taps were located at three axial positions of all four walls in the inlet section. Each 1-in.-long segment of the test section was provided with pressure taps at a mean axial position, but only on sidewalls B and D. Pressure was measured by means of oil manometers which could be read to a precision of better than 0.002 psia. The convention used for orientation of the magnetic field is the same as that of the previous article. Static pressure results given here were taken concurrently with the heat transfer data of test 107-18H; Table 4 of *Section E* listed the appropriate test conditions.

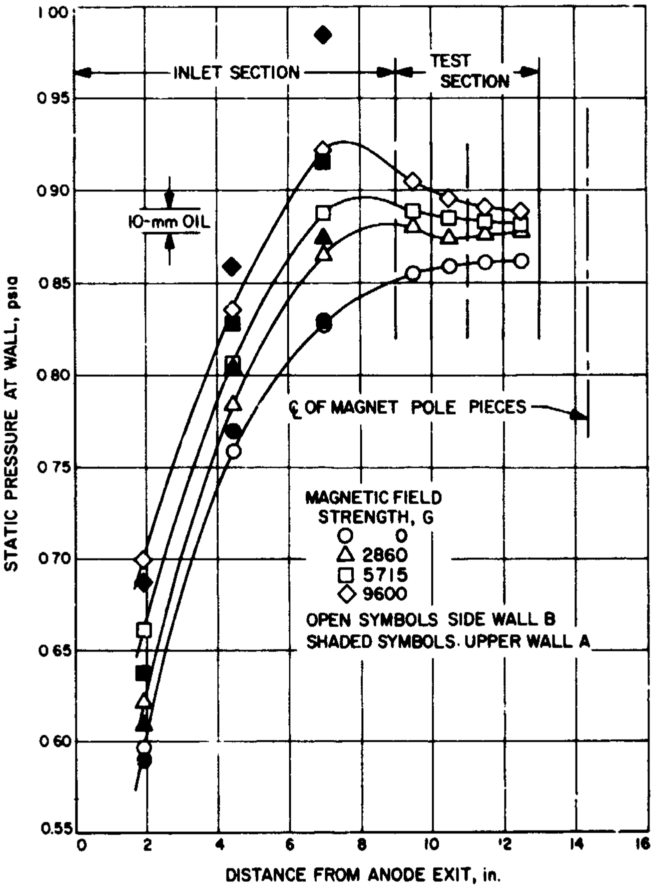


Fig. 14. Axial distribution of pressure along walls of square channel

3. Experimental Results

Axial distributions of static pressure were generally similar to those presented in SPS 37-49, Vol. III, pp. 199–201. The effect of the magnetic field was to increase the pressure throughout the channel for a constant mass flow rate and to cause a peak pressure to be reached near the downstream end of the inlet section (Fig. 14). Values of magnetic field listed in the figure are for the downstream end of the test section. Results are shown for forward field and for two walls, upper wall A and sidewall B. The relative differences observed between the two walls at zero field are not only preserved but increased with increasing field.

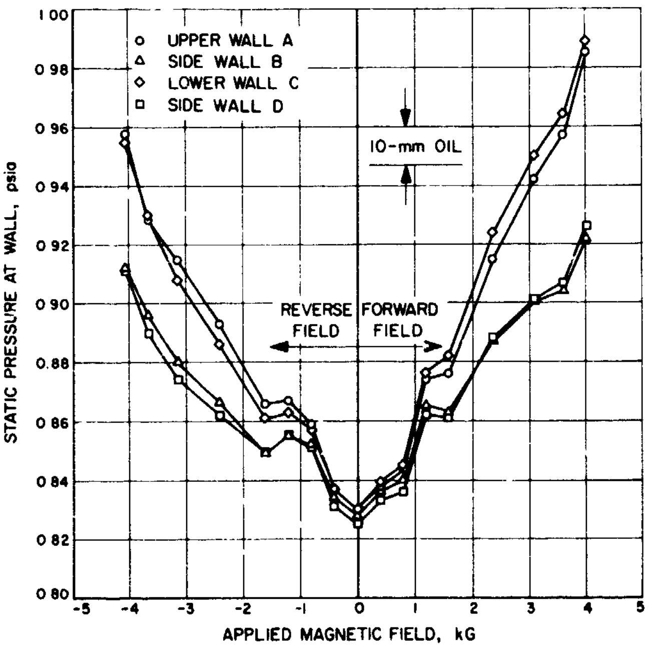


Fig. 15. Static pressure in inlet section at axial location 2 in. upstream of test section

The effect of magnetic field is examined in more detail in Fig. 15, where the static pressure for all four walls at one axial station has been plotted as a function of applied magnetic field at that axial location. The axial position selected corresponds to the third pressure tap location of the inlet section, i.e., at an axial distance of 7 in. from the anode exit (Fig. 12), which is the region of peak pressure (Fig. 14). Two results are apparent from Fig. 15: (1) the pressure increases with increasing field regardless of the direction of the field, and (2) the effect is much more pronounced at the upper and lower walls of the channel than at the sidewalls. It is also evident in Fig. 15 that, where wall C exhibits a higher pressure than wall A with forward field, the converse occurs with reverse field. Walls B and D exhibit similar trends. It is believed that this observation could result because of Hall effects; however, it could also be a consequence of swirl present in the flow.

In the regime of operation of the present experiment, theoretically high values of the Hall parameter are predicted (SPS 37-47, Vol. III, pp. 120–128). Since an induced electric field in the axial direction is unlikely because the four walls of the channel form a continuous electric conductor, a large axial current flow is possible when the Hall parameter $\omega\tau > 1$. Three experimental observations tend to indicate that Hall effects were present in this experiment: all three indicate the presence of a significant lateral (side) force, as well as transverse (vertical) component of force acting on the field. Firstly, an applied magnetic field had the effect of increasing heat transfer on one sidewall of the channel but decreasing the heat transfer on the opposite sidewall; when the direction of the magnetic field was reversed, the heat transfer effect also became reversed (see *Section E*). Secondly, the static wall pressure was slightly different comparing the two sidewalls, or comparing the upper and lower walls, and this effect also reversed when the field was reversed (Fig. 15). Thirdly, a visible effect was produced when the luminous core of the exhaust plume (in vacuum tank) was observed during a change in magnitude of the applied field. A significant lateral motion of the luminous core was observed with increasing magnetic field; when the field was reversed, the luminous core moved to the opposite side.

Thus, the effect of the magnetic field was to increase the static pressure throughout the channel regardless of the direction of the applied field. Walls transverse to the magnetic field experienced a greater increase in pressure than did the sidewalls which were parallel to the direction of the field. The presence of Hall effects during this experiment is considered likely although the magnitude of these effects has not yet been established.

XIV. Liquid Propulsion

PROPULSION DIVISION

A. Heat-Sterilization Compatibility of Ethylene-Propylene Rubber in N_2H_4, *O. F. Keller*

1. Introduction

This article presents the data covering the last part of a series of patch-type tests of an expulsion bladder material for the thermal sterilization compatibility study. The bladder material is ethylene-propylene rubber (EPR), Stillman Rubber compound SR 722-70. The results of the first three cycles at 275 and 300°F were previously reported in SPS 37-46, Vol. IV, pp. 167–173. The results of the last three cycles are reported herein and complete this phase of the program.

2. Test Procedure

Throughout the study, two test sample configurations have been used: (1) the circular-type, about 1.5-in. diam and 0.037 in. thick, and (2) the rectangular-type about 1.5 by 2.0 by 0.037 in. thick. These samples were cut from an existing diaphragm-type bladder as shown in Fig. 5 of Ref. 1. The circular samples have been designated a_1 and a_2; the rectangular samples, b_1 and b_2 (Table 1). Two circular samples and two rectangular samples have been tested in each of three stainless steel containers. The average total surface area of the samples exposed to hydrazine in each container was 21 in^2. The four samples in each container were separated from one another by a special stainless steel wire rack.

The propellant containers and the wire racks were made of AISI-type 347 stainless steel. The average volume of the containers with test samples removed was 502 ml. Each container was equipped with an inlet port near the bottom and a vent port near the top. A pressurizing port was included on the top of each container for adjusting the initial test pressure. The average container volume to the vent port was 305 ml.

Each container was filled with hydrazine up to the vent port. A fourth container, without patch-test samples, was used as a reference, or control, container. The average pretest ullage volume for each container, including lines and transducers, was 205 ml.

The four containers were mounted in a temperature control chamber, and heat-sterilization temperatures of 275 and 300°F were maintained. The length of time at heat-sterilization temperature was 60 h for each test cycle, and the maximum number of test cycles at each test temperature was six. Prior to testing, the containers were passivated, using dilute hydrazine at ambient temperature and pressure for a period of 20 h. Initial container pressure for the tests at 275°F was 40 psig and for the tests at 300°F was 50 psig.

3. Test Results

After the first heat-sterilization cycle, the liquid hydrazine was light brown in color and contained fine black material in suspension. This color did not change appreciably as a result of additional heat-sterilization cycles. The liquid hydrazine in the reference, or control, containers remained colorless after heat-sterilization cycling.

Following heat-sterilization testing with containers 2 and 4, a quantitative chemical analysis of the remaining hydrazine was made. The results of this analysis are shown in Table 2. The quantities of water, ammonia, aniline, and hydrazine v ere determined by gas chromatography. The ammcnia content was also confirmed by a titration technique. The volumes of residual hydrazine following heat-sterilization testing ranged from 140 to 270 ml. The quantity of residual hydrazine varied with the number of cycles and the quantities of hydrazine vapor lost during the venting operation between heat-sterilization cycles.

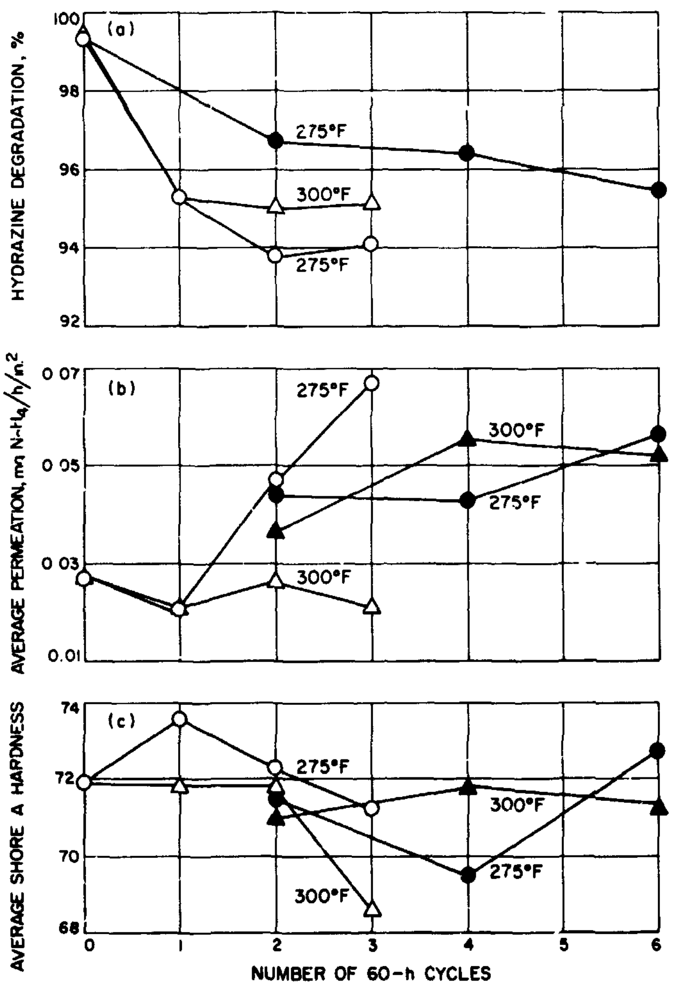


Fig. 1. Effect of hydrazine on EPR (SR 722-70) after heat sterilization cycling

This series of tests included determination of parameters for Shore A hardness of the patch-test samples, the permeation rate of the samples to hydrazine, and the degradation of the hydrazine resulting from heat-sterilization cycling. These parameters have been plotted as a function of the number of heat-sterilization cycles as shown in Fig. 1. The data indicate a slight increase in permeation rates with an increasing number of heat-sterilization cycles. Also an appreciable degradation of the hydrazine occurs during the first heat-sterilization cycle.

Previous test results, as reported in SPS 37-46, Vol. IV, p. 173, indicated that the average pressure rise in the reference containers (with hydrazine only) was greater than the average rise in the containers with both EPR patch-test samples and hydrazine (Fig. 2). To isolate the effects of the EPR/hydrazine reaction from the hydrazine reaction with the stainless steel containers, four type-347

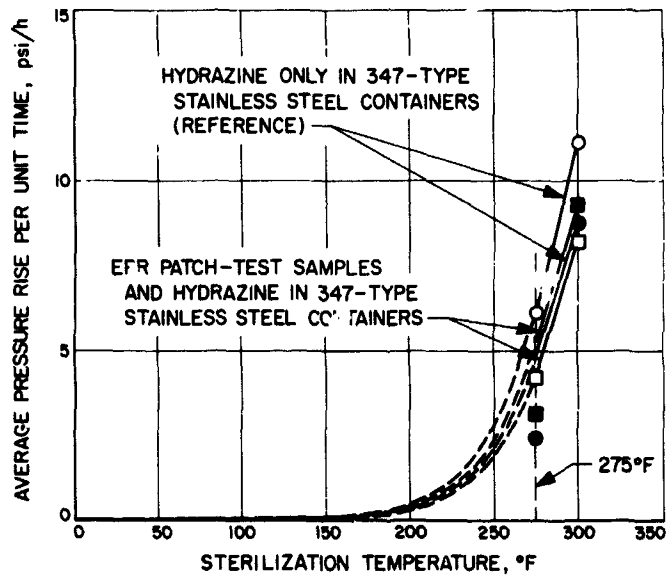


Fig. 2. Temperature sensitivity of EPR patch-test samples immersed in hydrazine

Item[a]	Sample No.	Initial thickness, in.	Final thickness, in.	Initial weight, g	Final weight, g	Net increase (decrease), g	Increase (decrease), %	Initial Shore A hardness ±2.0	Final Shore A hardness ±2.0	Shore A hardness after permeation test ±2.0
1	66 × 09201-1-a$_1$	0.034–0.036	—	1.3052	1.3163	0.0111	0.85	72	69	66
2	66 × 09201-1-a$_2$	0.035–0.037	—	1.3153	1.3248	0.0095	0.72	73	70	66
3	66 × 09201-1-b$_1$	0.038–0.041	—	3.3868	3.4076	0.0208	0.61	71	69	—
4	66 × 09201-1-b$_2$	0.039–0.041	—	3.3690	3.3930	0.0240	0.71	71	70	—
5	66 × 09201-2-a$_1$	0.034–0.036	—	1.2619	1.2597	(0.0022)	(0.17)	71	73	67
6	66 × 09201-2-a$_2$	0.035–0.037	—	1.3135	1.3146	0.0011	0.08	72	70	67
7	66 × 09201-2-b$_1$	0.037–0.041	—	3.3094	3.3211	0.0117	0.35	72	72	—
8	66 × 09201-2-b$_2$	0.038–0.042	—	3.5659	3.5777	0.0118	0.33	72	70	—
9	66 × 09201-3-a$_1$	0.036–0.038	—	1.3235	1.3137	(0.0098)	(0.74)	72	73	—
10	66 × 09201-3-a$_2$	0.034–0.037	—	1.2975	1.2991	0.0016	0.12	73	74	—
11	66 × 09201-3-b$_1$	0.039–0.042	—	3.4381	3.4360	(0.0021)	(0.06)	72	73	—
12	66 × 09201-3-b$_2$	0.036–0.040	—	3.2606	3.2566	(0.0040)	(0.12)	71	71	—
13	66 × 09201-4-a$_1$	0.034–0.036	0.033–0.035	1.2519	1.2468	(0.0051)	(0.41)	73	73	—
14	66 × 09201-4-a$_2$	0.033–0.037	0.033–0.037	1.2742	1.2757	0.0015	0.12	72	72	—
15	66 × 09201-4-b$_1$	0.038–0.041	0.037–0.041	3.3060	3.3020	(0.0040)	(0.12)	71	71	—
16	66 × 09201-4-b$_2$	0.038–0.041	0.038–0.041	3.3979	3.3914	(0.0065)	(0.19)	71	71	—
17	66 × 09201-5-a$_1$	0.035–0.038	0.035–0.037	1.3380	1.3368	(0.0012)	(0.09)	74	71	—
18	66 × 09201-5-a$_2$	0.033–0.034	0.032–0.034	1.2221	1.2196	(0.0025)	(0.20)	74	73	—
19	66 × 09201-5-b$_1$	0.035–0.039	0.035–0.039	3.1618	3.1499	(0.0119)	(0.38)	70	70	—
20	66 × 09201-5-b$_2$	0.036–0.040	0.036–0.040	3.2941	3.2845	(0.0096)	(0.29)	72	71	—
21	66 × 09201-6-a$_1$	0.035–0.037	0.035–0.037	1.3142	1.3176	0.0034	0.26	72	72	—
22	66 × 09201-6-a$_2$	0.035–0.038	0.035–0.037	1.3131	1.3188	0.0057	0.43	71	71	—
23	66 × 09201-6-b$_1$	0.035–0.039	0.035–0.038	3.0955	3.0970	0.0015	0.05	71	71	—
24	66 × 09201-6-b$_2$	0.037–0.041	0.037–0.041	3.4006	3.4011	0.0005	0.01	71	70	—
					Pressure generation reference (or control) samples					
25	66 × 09201-7-a$_1$	0.036–0.037	—	1.3228	—	—	—	72	—	71
26	66 × 09201-7-a$_2$	0.033–0.034	—	1.1773	—	—	—	73	—	73

[a] Compound SR 722-70.

[b] *Fluid test temperature 275°F (10 h at 315°F).*

[c] Fluid test temperature 300°F.

[d] Initial container pressure 40 psig.

[e] Initial container pressure 50 psig.

[f] *These data assumed low by a factor of 10.*

a$_1$ and a$_2$ are circular samples.

b$_1$ and b$_2$ are rectangular samples.

FOLDOUT FRAME 2

Table 1. Ethylene propylene patch-test samples heat-sterilization test data

Shore A hardness after permeation test ±2.0	Shore A dry hardness after vacuum ±2.0	Permeation mg N_2H_4/h/in.^{2}	Time of permeation test, h	No. of cycles	Total time at test ten perature, h	N_2H_4 concentrate, %	Container No.	Pressure designator	Pressure after two cycles, psig	Pressure after four cycles, psig	Pressure after six cycles, psig
66	69	0.045	108	4^b	250.5	96.4	2	P_1 ^{d}	310	190	—
66	70	0.041	108	4^b	250.5	96.4	2	P_1 ^{d}	310	190	—
—	73	—	—	4^b	250.5	96.4	2	P_1 ^{d}	310	190	—
—	72	—	—	4^b	250.5	96.4	2	P_1 ^{d}	310	190	—
67	71	0.057	120	2^b	130.5	96.7	4	P_2 ^{d}	276	—	—
67	70	0.030	48	2^b	130.5	96.7	4	P_2 ^{d}	276	—	—
—	73	—	—	2^b	130.5	96.7	4	P_2 ^{d}	276	—	—
—	72	—	—	2^b	130.5	96.7	4	P_2 ^{d}	276	—	—
—	71	0.049	92.5	6^b	369.5	95.5	5	P_3 ^{d}	226	210	174
—	72	0.059	92.5	6^b	369.5	95.5	5	P_3 ^{d}	226	210	174
—	74	—	—	6^b	369.5	95.5	5	P_3 ^{d}	226	210	174
—	72	—	—	6^b	369.5	95.5	5	P_3 ^{d}	226	210	174
—	70	0.0054[f]	92.5	4^c	240	—	1	P_1 ^{e}	588	529	—
—	70	0.0056[f]	92.5	4^c	240	—	1	P_1 ^{e}	588	529	—
—	71	—	—	4^c	240	—	1	P_1 ^{e}	588	529	—
—	71	—	—	4^c	240	—	1	P_1 ^{e}	588	529	—
—	69	0.0048[f]	92.5	6^c	360	—	8	P_2 ^{e}	694	625	536
—	70	0.0054[f]	92.5	6^c	360	—	8	P_2 ^{e}	694	625	536
—	71	—	—	6^c	360	—	8	P_2 ^{e}	694	625	536
—	71	—	—	6^c	360	—	8	P_2 ^{e}	694	625	536
—	69	0.0034[f]	70.5	2^c	120	—	9	P_3 ^{e}	684	—	—
—	69	0.0037[f]	70.5	2^c	120	—	9	P_3 ^{e}	684	—	—
—	71	—	—	2^c	120	—	9	P_3 ^{e}	684	—	—
—	70	—	—	2^c	120	—	9	P_3 ^{e}	684	—	—
				Reference (or control) containers							
71	72	0.026	40	3^b	179	96.1	6	P_4 ^{d}	270	160	136
73	73	0.028	40	3^c	180	—	10	P_4 ^{e}	572	595	567

Table 2. Chemical analysis of remaining hydrazine after heat-sterilization

Constituent	Container 2 %	Container 4 %
Hydrazine (N_2H_4)	96.4	96.7
Water (H_2O)	1.9	1.4
Aniline ($C_6H_5NH_2$)	0.3	0.3
Ammonia (NH_3)	1.2	1.4
Residue	0.06	0.09
Totals	99.86	99.89

stainless steel containers were filled with 305 ml of hydrazine and heat-cycled at 300°F for 60 h. Initial container pressure was approximately 50 psig. Passivation of the containers prior to testing was again accomplished using the same procedures. During the first 60-h heat-sterilization cycle at 300°F, one of the containers was vented because the pressure buildup exceeded 1500 psig—a tentative maximum safe test pressure based on previous test data, as shown in SPS 37-44, Vol. IV, p. 180, Table 7. The second 60-h heat-sterilization cycle at 300°F was terminated after 50.7 h when another test container pressure exceeded the 1500-psig limit.

The pressure rise per square-inch of .est sample surface area was also determined. The pressure rise data for all test containers after heat-sterilization cycling at 300°F in this series of tests were averaged and divided by the total patch-test surface area for a typical container. This calculation produced a value of 0.4-psi pressure rise per hour per square inch of test sample surface area. The test sample properties were determined before and after heat-sterilization cycling only. No attempt was made to take into account any changes occurring during the heat-sterilization cycle.

This ,eries of te.,ts concludes the patch-type testing. Further details concerning the AI.PS generant tank development program are described in Ref. 1.

4. Conclusion

Based on the results of this series of tests, it must be concluded that the ethylene propylene material is very marginal for expulsion use with hydrazine at a temperature level of 275°F.

Reference

1. Keller, O. F., and Toth, L. R., *ALPS Generant Tank and Cell Assembly*, Technical Report 32-865, Jet Propulsion Laboratory, Pasadena, Calif., Frb. 28, 1966.

XV. Lunar and Planetary Instruments

SPACE SCIENCES DIVISION

A. Atmospheric Entry Sampling System, *S. Rich*

1. Introduction

In order to analyze the composition of the Mars atmosphere with the JPL entry mass spectrometer (see *Section B*), uncontaminated samples of the atmosphere must be introduced into the ion source of the instrument under molecular flow conditions. To perform this type of analysis during the terminal descent phase of a Mars entry mission, the capability to continuously sample the atmosphere over the Mach No. range $0<M<9$ is required.

The method currently under consideration for obtaining uncontaminated atmospheric samples during terminal descent consists of inserting a sample tube through the entry capsule nose cap to sample the atmosphere behind the bow shock wave. To prevent sample contamination by the entry capsule, the sample tube inlet port must be located forward of the capsule boundary layer. For the VM-8 Mars model atmosphere and a 6.5-ft-diam 60-deg capsule with a ballistic coefficient of 0.12, the sample tube inlet port would have to be located approximately 0.5 in. in front of the nose cap.

In order to provide the required molecular flow into the mass spectrometer ion source, part of the atmosphere flowing into the sample tube must be converted to molecular flow and subsequently piped to the ion source. To accomplish this conversion, a variable conductance molecular leak is being developed. The rate of flow through a molecular leak is a function of the sample gas molecular weight, differential pressure across the leak, and the sample gas absolute temperature. Feedback control will be utilized to vary the conductance of the molecular leak. This provides a measure of adaptive flow control to compensate for atmospheric uncertainties which may affect sample inlet pressure and sample inlet temperature. By utilizing the mass spectrometer total ion current measurement as the feedback control signal, a uniform sample flow rate into the mass spectrometer can be maintained during the entire atmospheric sampling period. Maintaining an appropriate uniform flow rate permits mass spectrometer operation at maximum ion source pressure, which provides maximum measurement sensitivity during the entire atmospheric sampling period.

2. Sample Tube Configuration

Two alternate sample tube configurations under consideration are shown in Figs. 1 and 2. Both configurations utilize explosive actuators to deploy the sample tube in front of the entry capsule nose cap. The nose cap plug shown in Fig. 1 has a higher ballistic coefficient than the entry capsule; consequently, the plug falls free of the entry capsule after it is forced out of its hole by the sample tube.

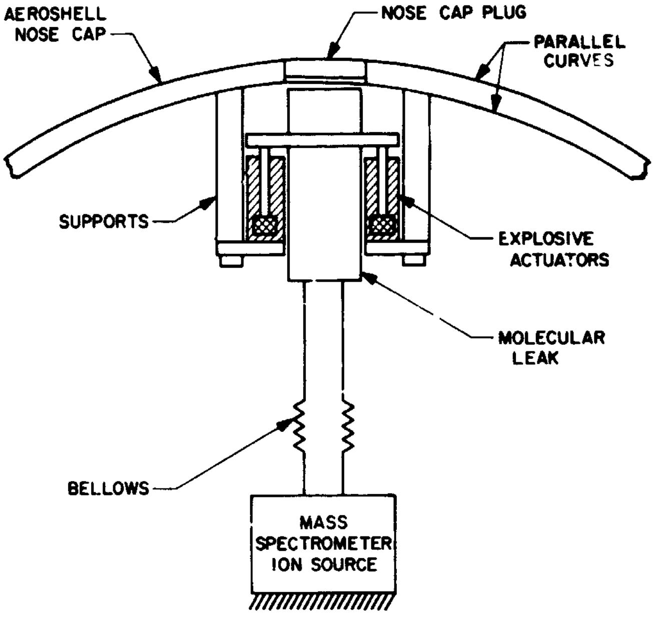


Fig. 1. Molecular leak deployment configuration

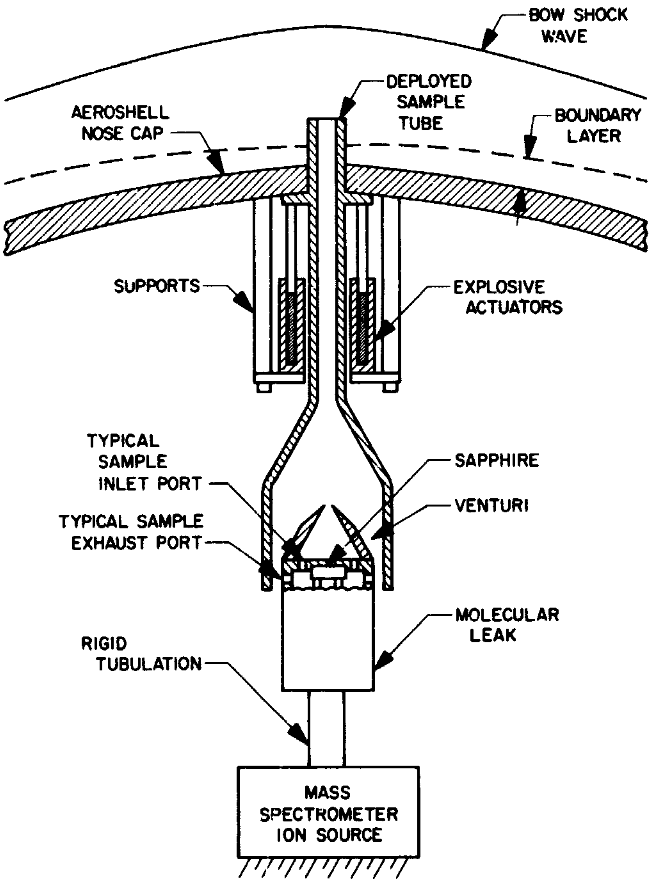


Fig. 2. Sample tube deployment configuration

Advantages and disadvantages of the two sample tube configurations under consideration are as follows:

(1) The configuration shown in Fig. 2 permits the use of a smaller diameter sample tube and will require a smaller diameter nose cap clearance hole and plug. Consequently, less force is required to eject the nose cap plug, and a smaller explosive actuator can be used.

(2) In the Fig. 1 configuration, the deployed molecular leak aperture is located in front of the nose cap, and the atmospheric sample flow does not enter the capsule. In the Fig. 2 configuration, the molecular leak aperture is located inside of the entry capsule, and the atmospheric sample flow enters the capsule. Entry of the sample flow into the capsule may cause a thermal control problem, and an additional sample exhaust duct may be required.

(3) The adaptive flow control problem is more complicated in the Fig. 1 configuration, since the feedback flow control system must compensate for atmospheric heating of the molecular leak.

(4) In the Fig. 1 configuration, a bellows is required to permit extension of the tubulation between the molecular leak and the ion source during sample tube deployment. Extension of the bellows may dislodge contaminants entrapped in the bellows wall.

For sample system simplicity, the configuration in Fig. 2 appears preferable; however, further study is required to determine the effect on capsule thermal control or the effect on capsule configuration if a sample exhaust duct should be required.

3. Variable Conductance Molecular Leak

A schematic diagram of a variable conductance molecular leak currently under development is shown in Fig. 3. The conductance of the leak is varied by applying current to the heating elements on the inner and outer shells. Heating the outer shell causes it to expand in length, opening the leak aperture to increase conductance. Similarly, heating the inner shell closes the leak aperture to decrease conductance.

The theoretical conductance of the molecular leak is given by the equation

$$F = 60.96 \frac{h^2 (T/M)^{1/2}}{\ln (d_o/d_i)} \qquad (1)$$

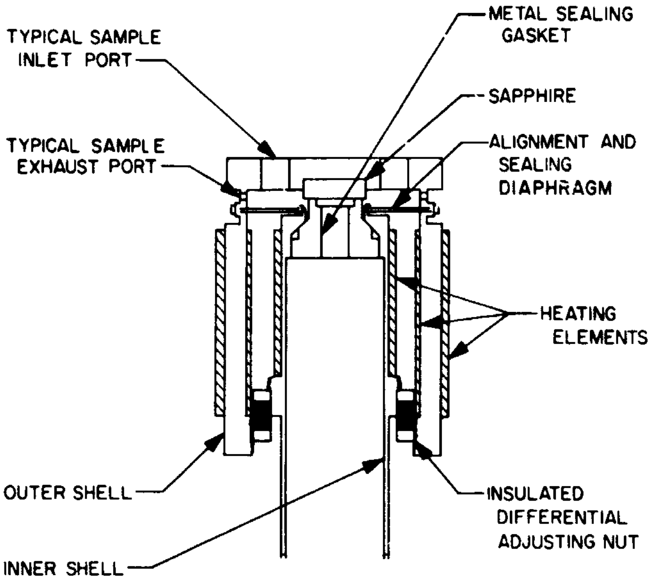

TYPICAL SAMPLE INLET PORT
TYPICAL SAMPLE EXHAUST PORT
METAL SEALING GASKET
SAPPHIRE
ALIGNMENT AND SEALING DIAPHRAGM
HEATING ELEMENTS
OUTER SHELL
INSULATED DIFFERENTIAL ADJUSTING NUT
INNER SHELL

Fig. 3. Variable conductance molecular leak

where

F = conductance, l/sec

h = the effective cylindrical aperture height between the optically flat sapphire and the circular metal-sealing surface, cm

d_o = the outside diameter of the circular metal-sealing surface, cm

d_i = the inside diameter of the circular metal-sealing surface, cm

T = the absolute temperature of the flowing gas, °K

M = the molecular weight of the gas

Rate of flow through the molecular leak is given by the equation

$$Q = F(P_s - P_i) \tag{2}$$

where

P_s = the pressure outside the leak (essentially the stagnation pressure behind the bow shock wave)

P_i = the pressure on the ion source side of the leak

Using Eqs. (1) and (2), the range of h required for a uniform flow rate of 10^{-7} torr-l/s was computed for the terminal descent phase of a Mars entry mission. A plot of the variation in h as a function of time to impact, altitude, and Mach No. is shown in Fig. 4.

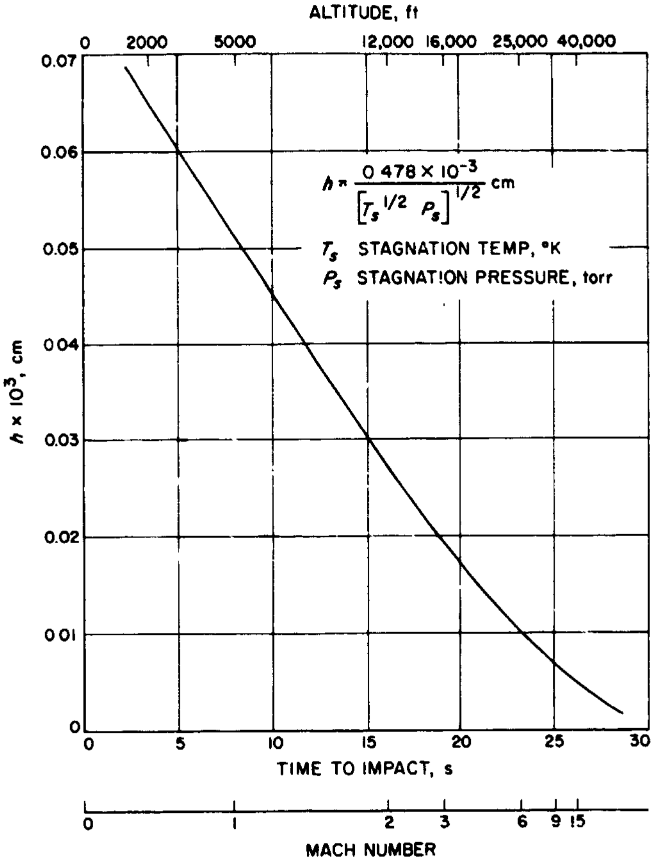


Fig. 4. Theoretical aperture height during terminal descent for 6.5-ft-diam 60-deg sphere/cone

The thermal energy required to produce an aperture height h by expansion of the outer shell (assuming no heat loss) is given by the equation

$$H = \frac{w c A h}{e} \tag{3}$$

where

H = the required thermal energy

w = the specific weight of the outer shell material

c = the specific heat of the outer shell material

A = the cross-sectional area of the outer shell

e = the coefficient of expansion of the outer shell material

For an outer shell constructed of 304 stainless steel, with a cross-sectional area of 0.15 in.^{2}, 0.02 Btu of thermal energy or an average power of approximately 0.84 W during the last 25 s prior to impact, is required to produce the maximum h (0.07×10^{-3} cm) shown in Fig. 4.

An estimate of the variable conductance molecular leak thermal actuation time constant for expansion of the outer shell (assuming no heat losses) is given by the equation

$$B = \frac{c\,w\,s\,i}{2k} \tag{4}$$

where

B = the thermal time constant

k = the thermal conductivity of the bonding material between the heating elements and the outer shell

c = the specific heat of the outer shell material

w = the specific weight of the outer shell material

s = the radial thickness of the outer shell wall

i = the bonding material thickness between the heating elements and the outer shell

For the outer shell constructed of 304 stainless steel, with a wall thickness of 0.05 in. and a 0.003-in. thickness of Delta Bond 152 cementing the heating elements to the outer shell, the computed thermal time constant is 0.174 s.

B. Prototype Mass Spectrometer for Planetary Atmospheric Analysis, *H. R. Mertz*

1. Introduction

One of the important tasks in planetary exploration is to determine the composition and density of the atmosphere of the planet. One way to obtain such information is with a flight-type atmospheric mass spectrometer which covers the desired mass range with the proper sensitivity. A first step in developing such an instrument is to construct, test, and make a flight evaluation of an engineering model. A contract was let in July of 1967 to design and construct an engineering model based upon the results of the science breadboard mass spectrometer design. The instrument was to be incorporated into the Capsule Systems Advanced Development (CSAD) program in the early Spring of 1968.

2. Instrument Operation

A mass spectrometer performs the compositional analysis of a gaseous sample by ionizing a portion of the gas being analyzed. The ions generated are separated according to their individual mass to charge (m/e) ratios. Once separated, the resulting ion currents are detected and amplified by an electron-multiplier-electrometer detection system, the output appearing in the form of discrete voltage peaks of different values of m/e. Relative abundance measurements are made by an intercomparison of the voltage levels of these peaks.

Mass spectrometers differ only in the method used to achieve m/e separation. The double-focusing magnetic sector instrument (Fig. 5) first accelerates ions through a radial electrostatic analyzer. The radius of curvature of the ion trajectories in this portion of the instrument is proportional to the energy of the ions, and the ions are focused accordingly. The ions are then directed through a magnetic field where the radius of curvature of the ion trajectories is proportional to the individual m/e value of each ion. With a constant magnetic field each variety of ion requires a different acceleration voltage (and, hence, electrostatic analyzer voltage) to traverse the two fixed curvatures of the instrument to be collected by the electron multiplier detector. By scanning the acceleration and electrostatic analyzer voltages cyclically between the proper limits, a mass spectrum is produced. Simultaneous correction of direction focusing and velocity focusing inhomogeneities in this instrument are achieved through the proper choice of the electrostatic and magnetic analyzer ion optical properties. Hence, high mass resolution and sensitivity are simultaneously obtained, a result not readily attainable in other types of mass spectrometers.

3. Instrument Description

The instrument described here is a double-focusing magnetic sector mass spectrometer. The critical specifications of this instrument are shown in Table 1.

Table 1. Specifications for double-focusing magnetic sector mass spectrometer

Electrostatic sector angle	90 deg
Electrostatic sector radius	2.480 in.
Magnetic sector angle	60 deg
Magnetic sector radius	2.003 in.
Ion source exit slit width	0.004 in.
Collector entrance slit width	0.008 in.
Source divergence angle α	1 deg
Mass resolution $(M/\Delta M)_{1\%}$	90
Dynamic range	1:5 × 10^6
Mass range	$M = 10$ to 90
Scan time (for one spectrum)	2.8 s

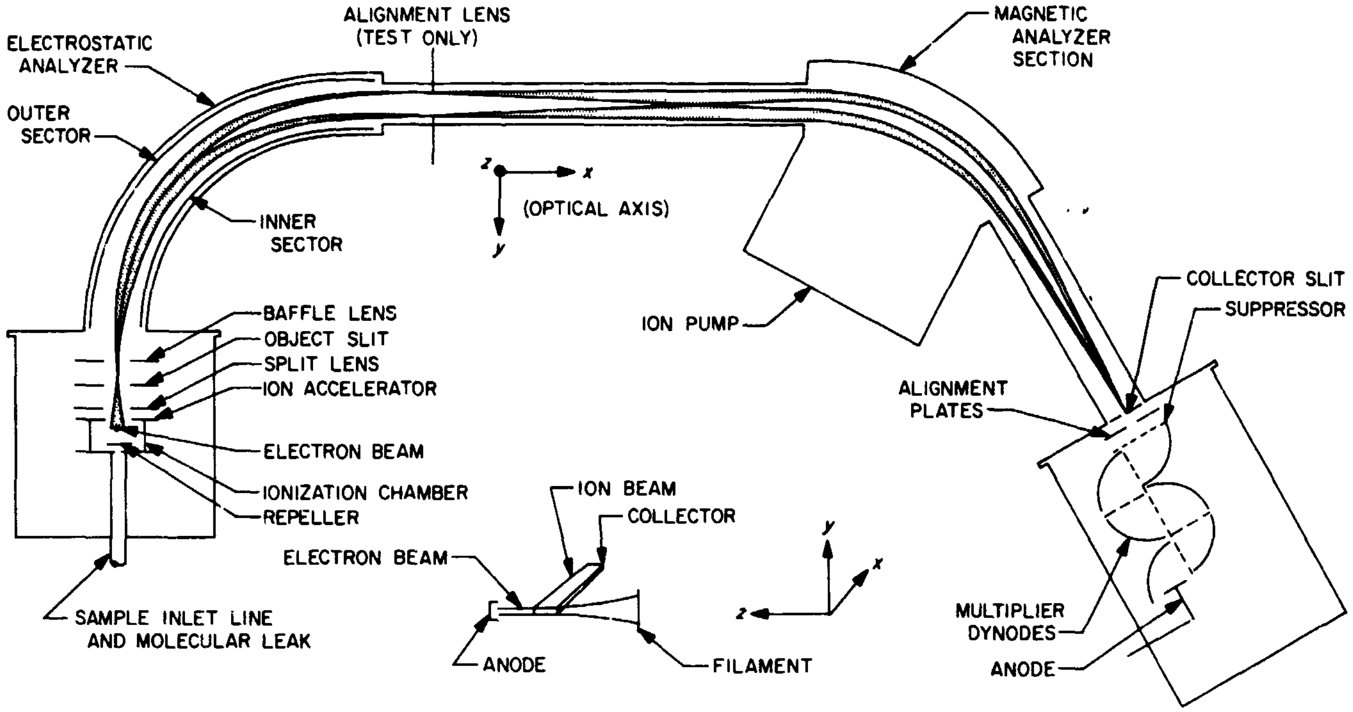


Fig. 5. Analyzer, ion trajectory, and optical system of mass spectrometer

The analyzer vacuum envelope consists of five individual components which enclose the ion source, the electric sector, the magnetic sector, and the electron multiplier. It is a thin-walled 304 stainless steel housing which is electrically accessible through several feedthroughs. Multiple pin feedthrough headers introduce voltages from the ion source electronics and feed them to the filament and to the various focusing electrodes; a single pin feedthrough transmits the ion current from the electron multiplier to the range-switching electrometer.

The internal vacuum necessary for operation of the analyzer is maintained by an ion pump which is made as an integral part of the magnetic sector.

In addition to the electrometer amplifier and ranging circuit, the support electronics consist of the following modules:

(1) The filament supply and emission regulator which maintains a constant ionizing electron current.

(2) The scanning electrode bias supply which provides the ion source accelerating potential and proportional electrostatic analyzer potentials.

(3) The low-voltage power supply for the various modules.

(4) High-voltage supplies for the ion pump and electron multiplier.

A more detailed description of the instrument components and design considerations are covered in the following sections.

4. Ion Source

A cross-sectional view of the ion source is shown in Fig. 6. A closed ion source design was used to obtain a minimum gas flow out of the source, allowing the source to be operated at a pressure higher than the rest of the analyzer. Electrons are admitted into the ionization region through a small aperture, and the ions are withdrawn through another small aperture. The structure of the electrodes is circular in shape so that they can be sealed and insulated from each other by ceramic rings. The close fit between the metallic lenses and the ceramic rings produces a very small gas conductance, which effectively seals the ion source.

The operation of the source at a pressure higher than that of the rest of the system has the following advantages toward reducing sample distortion:

(1) Outgassing from the hot filament and the surfaces of the system is pumped away, thereby minimizing entry of these species into the ionization chamber.

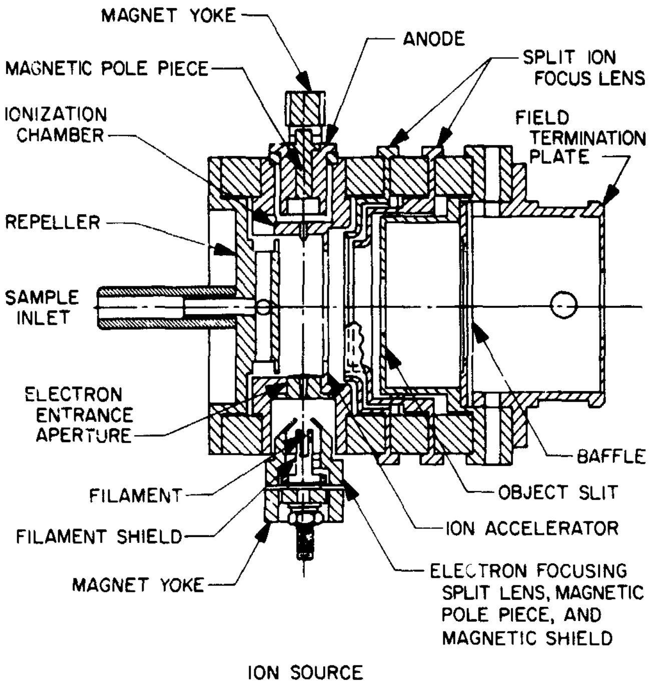


Fig. 6. Ion source cross section

(2) Variations in the ion pump speed have less influence on the source pressure.

(3) A source pressure that is too high for operation of either the filament or the electron multiplier is permitted, thereby improving the effective ion source sensitivity.

To obtain a high differential pressure between the ion source and analyzer, it was necessary to minimize the area of the electron aperture but still maintain a good electron transmission efficiency. Maximum transmission through a small aperture can be obtained by a well-focused beam. To obtain such a beam, a shield located on the sides of the filament plus an aperture lens located between the filament and the electron entrance slit were used. The aperture lens was split into two electrodes so that misalignment within the electron gun could be corrected by a differential voltage across these two electrodes. In addition to the electron gun, a magnetic field is employed in the ion source to align the beam for maximum stability. This gun configuration should provide a 50% transmission efficiency.

5. Electric Sector

The electric sector is used to compensate for the effects of velocity dispersion in the magnetic sector. It consists of the two cylindrical coaxial plates shown in Fig. 7. An electric potential is applied to each plate, establishing a force on the ions that balances their mean centrifugal force. The lip on the edge of the plates is used to compensate for the curved equipotential surfaces which the edge produces. These edge corrections are designed for a sharp 90-deg corner. The figure also shows a ground plane on the sides of the plate.

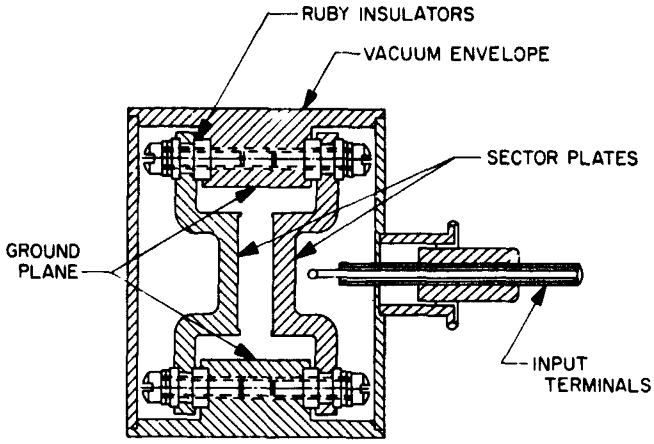


Fig. 7. Cross-sectional view of electric sector

Ruby washers are used to insulate the plate from the mounting points. Screws are used to hold the plates to the washers. The washers are contained in counterbores so that they will stay in place even if they are shocked to the point of fracture. The ruby washers can be shimmed to produce the required parallelism between the two plates. Because of the curvature of the plates, the assembled clearances can only be measured at the ends. Variations in the plate spacing results in either a beam spread or a beam displacement at the collector slit. A tolerance analysis was performed by the contractor to determine the alignment tolerances for the electric sector plates. These calculations indicate that the design will meet the instrument requirements.

6. Analyzer Tube and Ion Pump

The analyzer and ion pump sections were machined as a single part. An entrance and an exit tube were welded to the analyzer section. The ion pump housing forms an integral unit with the analyzer section. This unit is illustrated in Fig. 8.

The ion pump (Fig. 9) consists of two titanium plates with a titanium grid mounted between them. The plates are operated as a cathode, and the grid is operated as an anode. A basic design problem with an ion pump is the insulated mounting required for the anode. Since

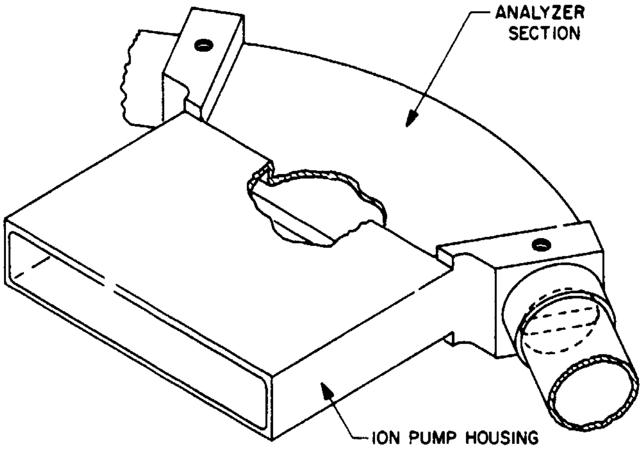


Fig. 8. Ion pump housing and analyzer section

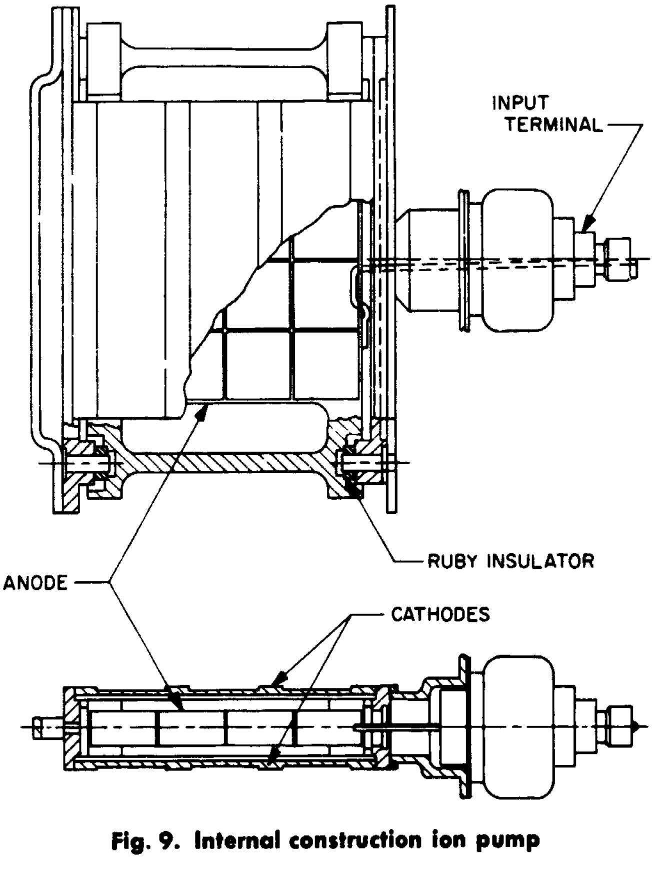


Fig. 9. Internal construction ion pump

the pumping action involves considerable sputtering, the insulating material used in the mounting can become coated with a film of sputtered metal. This metallic film would, in time, shortcircuit the ion pump power supply. To overcome this problem, the insulating material is surrounded by a shield held at the anode potential. The insulators used to support the grid structure are ruby washers. One side of the washer is in contact with a short boss which extends from the cathode. The other side of the washer is in contact with the anode. A cylindrical shield extends from the anode to surround the washer.

To understand the function of this shield, it is advantageous to review the nature of the gaseous discharge that occurs within the pump. Background radiation produces a small amount of electrons in any region where a gas is present. These electrons are accelerated by an electric field so that they will collide with neutral gas particles to produce ions and additional free electrons. One accelerated electron can produce many ions and additional free electrons if it is allowed to travel a long distance before it is collected by an electrode. The long electron path lengths are provided in a small container by causing the electrons to oscillate. They are accelerated toward the anode, which is a grid, but are not likely to be collected because of the grid geometry. After they pass the anode, they are decelerated by the cathode potential. The result is that they oscillate about the anode until they strike the grid. The number of cycles of oscillation can be increased if a longitudinal magnetic field is present. This field collimates the electrons as they oscillate about the anode.

The function of the shields around the ruby washers is to invert the electric potentials so that oscillations do not occur in the region around the washer. When electrons are produced by radiation and ionization in the region around the washers, they are accelerated and collected immediately by the shield. This arrangement establishes a short electron path, which greatly reduces the ion production and the amount of sputtering around the washer.

7. Electron Multiplier and Housing

The ion optical path in the mass spectrometer is terminated by an electron multiplier. A collector slit is located at the focal plane which blocks the entry of ion beams of other than the correct mass. The ion beam passes through the collector slit and strikes the first dynode of an electron multiplier and amplifies the ion current by secondary electron emission. The multiplier housing supports the electron multiplier and also provides a vacuum envelope.

8. Magnet Assembly

The magnetic assembly provides both the magnetic field necessary for mass separation and the field used by the ion pump. It consists of a yoke, a C-shaped structure of Armco Iron, permanent magnets of Alnico 5-7, and

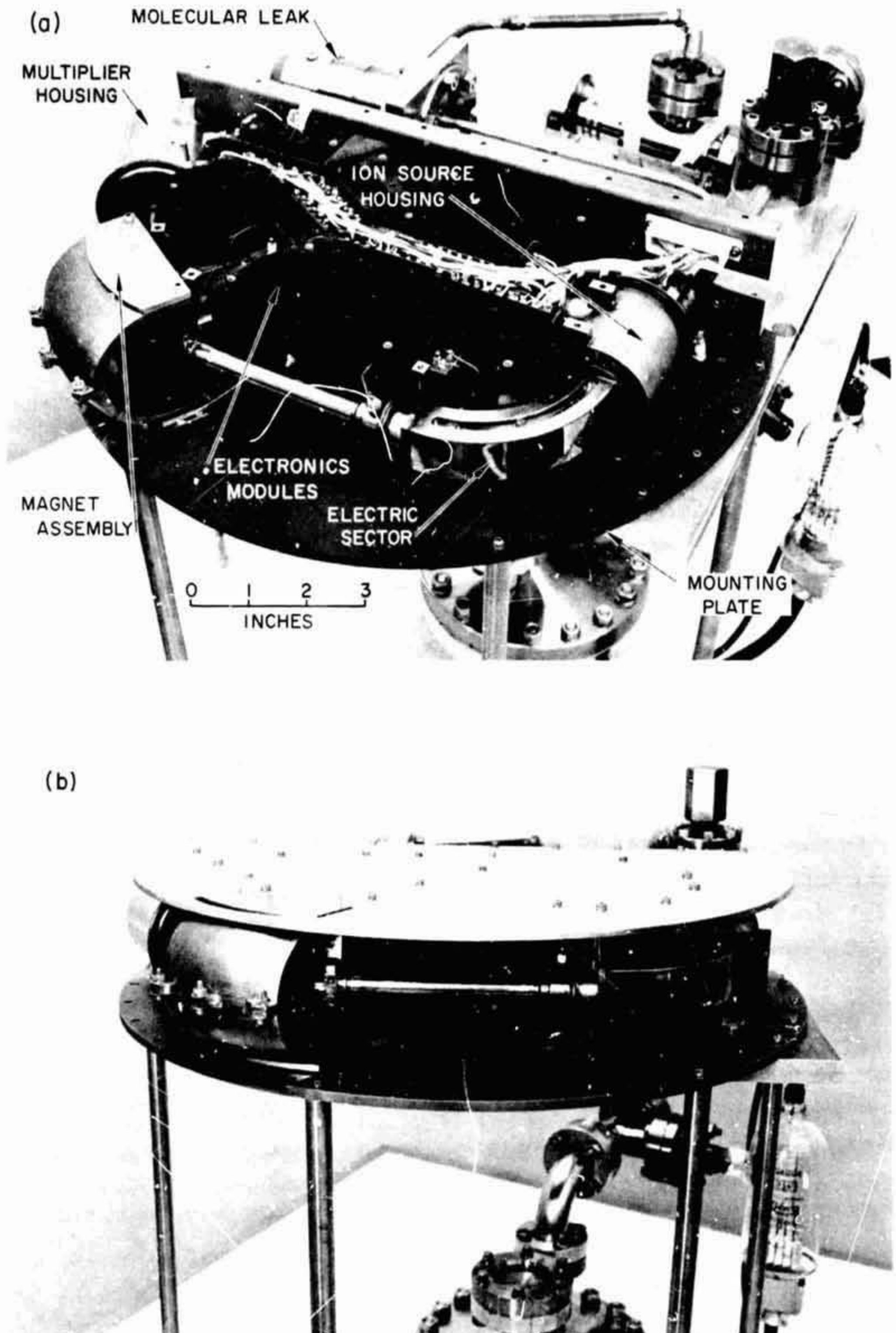


Fig. 10. Engineering model mass spectrometer: (a) without top mounting plate, (b) with top mounting plate

pole pieces of Armco Iron. The design was to give magnetic intensities of 4000 and 2000 G at the analyzer and ion pump sections, respectively.

9. Electronics Packaging

The electronics consist of 21 welded wire modules incorporated into 6 module assemblies. The modules were potted solid with Stycast 1090/11. Metal inserts were cast in the modules to provide for assembly to the mass spectrometer structure.

10. Structural Design

The structural support for the instrument consists of two semi-circular 321 stainless steel plates between which the analyzer and electronics modules are sandwiched. In addition, a stainless steel stiffener is also used. The bottom plate also provides the means for mounting the mass spectrometer to the CSAD nose cone. Figures 10a and b show the instrument with and without the top mounting plate.

11. Auxiliary Equipment

To facilitate the testing of the instrument a variable leak assembly was made an integral part of the instrument. The leak assembly is supported by the vertical stiffener. A valve was included so that the instrument could be connected to a commercial vacuum system to bake out the analyzer and could also be used for preliminary testing. The valve was subsequently removed from the instrument by pinching off at the interconnecting copper tubulation.

12. Preliminary Results

The analyzer assembly was completed during the third quarter of FY 1968. Preliminary tests were performed; electronic component selection was performed on the electronic modules; the modules were potted; final assembly was completed; and pinch-off performed.

There was not sufficient time to obtain quantitative measurements of the instrument performance; qualitative measurements showed, however, that the closed ion source provided an order of magnitude improvement in sensitivity over that exhibited by the science breadboard. The closed ion source design also allowed measurement of the oxygen peak. One area of ion source performance that was not up to expectation was electron beam efficiency. Rather than the predicted electron transmission of 50%, a value of about 10% was obtained. The electron gun design called for the filament shield to be at a slight negative voltage with respect to the filament. The design of the emission regulator circuit prevented the application of such a potential, so the shield was connected to the filament. The functional performance of the instrument as observed during the qualitative testing showed that the resolution was equal to that of the breadboard unit.

The testing of the unit revealed one major problem: The design of the shield, described in *Subsection 6*, proved inadequate, and a metallic film was deposited on the ruby washers that were used as the insulating material in the construction of the ion pump. This created a short across the ion pump supply, thereby shutting off the pump. In order to deliver a functioning unit to the CSAD program, a temporary adjustment was made, and the pump was able to maintain the system pressure at the proper level. New ruggedized supports have been designed and will be installed as soon as the unit is returned.

A sterilization cycle was performed on the instrument. No degradation in performance was noted. The instrument was delivered to the CSAD program for inclusion in the capsule system. The unit functioned properly during the subsequent subsystems and system tests and sterilization performed on the capsule system.

XVI. Space Instruments

SPACE SCIENCES DIVISION

A. A Pulse-Height Analyzer for Space Application,[1] *W. J. Schneider*[2]

1. Introduction

A number of scientific experiments performed from space vehicles make use of nuclear pulse spectrometry. The JPL program described here was designed to provide a pulse-height analyzer of sufficient precision and versatility to be suitable for any of a number of such spaceborne experiments. The analyzer may be commanded in flight to perform pulse-height analysis, time analysis, or multiparameter analysis. Instruction storage, data storage, and data readout, including a data compression option, are provided internally. (See Table 1 for the nomenclature used in this article.)

The analyzer incorporates the basic functional capabilities found in laboratory analyzers, with the exception of the linear amplifier and display sections (Table 2). The analog/digital converter (A/DC) has an input-pulse voltage range of 0.0 to 10.0 V with 19.5 mV resolution. Seven- and eight-bit resolutions are also available with corresponding memory subdivision. Coincidence and anti-coincidence modes are also provided, together with a live timer. Full-scale conversions are made in 128 μs, regardless of resolution.

The memory section of the analyzer stores 512 eighteen-bit words. The cycle time for "read-add/one-write" is slightly more than 5 μs. In pulse-height analysis, the first address contains live-time data, the last contains overflow or off-scale pulse count, and the remaining 125, 253, or 509 addresses contain spectral-density data.

The logic section of the analyzer accepts and stores externally generated instructions in its instruction register. Available instructions include: the analyzer modes; pulse-height analysis, time analysis, combined pulse-height and time analysis; two multiparameter modes; and a multiscaler mode. (A full description of the available instructions is given in Table 2.) Instructions stored in the register reorganize the analyzer's functional elements and control logic to fill the requirements of the commanded mode (Fig. 1). For example, in the pulse-height analysis mode, the receipt of a pulse for analysis causes an initiate-storage signal from the A/DC that, in turn, initiates a pulse sequence in the programmer. These pulses are routed by the programmer, under control of the instruction register, throughout the analyzer

[1] Work on the analyzer performed at Fabri-Tek Instrument Corp., Madison, Wis., for JPL (under Contract No. 951302) by R. Schumann, under the technical direction of the author.

[2] Member of JPL Technical Section 314.

as follows:

(1) Clear address register.

(2) Transfer contents of the pulse-height scaler to address register.

(3) Clear the data register

(4) Read the memory.

(5) Advance data register.

(6) Write the memory.

(7) Clear the pulse-height counter.

(8) Enable the A/DC.

Table 1. Nomenclature

A/DC	analog/digital converter
A/DCA	analog/digital converter advance
ADS	advanced data scaler
ANTI	anti-coincidence mode signal
APHS	advanced pulse-height scaler
CLOKF	1-MHz clock
COIN	coincidence mode signal
COINP	coincidence mode command
DISCH	discharge flip-flip
FETCH	externally produced pulse calling for new data during readout
FF	flip-flop
HISEN	high sensitivity threshold command
INDRN	initiate rundown signal
INITS	initiate-storage-cycle command
LGO	linear gate open
LIVEF	live-time flip-flop
MEASF	measurement mode flip-flop
MP	multiparameter
MSC	multiscaler clock
PCH0	pulse-height scaler zero
PHFF	pulse-height analysis mode flip-flop
PHSIG	pulse signal to be analyzed
P0	memory-busy flag
RDS	reset data scaler
REJF	reject flip-flop
RNDWN	rundown signal
RTFF	reset command for the T flip-flop
STA	start analysis
STOP	stop analysis
STR	start readout
T	synchronizing flip-flop set when an input signal is detected and cleared after the memory cycle
TMBS	telemetry bit sync

During readout, the instruction register is used to assemble the output data and shift it to telemetry. The first 18 bits shifted out contain all of the program instructions. Simplified instructions for the readout process are generated from the shift counter. These instructions cause reading of the addresses sequentially and the transfer of addresses and data into the instruction register.

Table 2. Pulse-height analyzer specifications

Item	Capability
Analog/digital converter accuracy	
Input range	0 to 10 V
Quantization	9 bits, 511 levels of 19.5 mV each
Zero stability	±0.04% from −5 to +45°C
Gain stability	±1% from −5 to +45°C
Linearity	±0.2% from best fit over upper 96% of scale
Slope	±4% from average over upper 96% of scale
Count-rate effect	±0.05% from 1 to 10^4 pulses/s
Analysis time	128 μs
Noise (uncertainty)	1 mV
Memory capacity	
Words	812
Bits/word	18
Access time	1.2 μs
Read-modify-write time	25 μs
Functional capability	
Pulse-height analysis	128, 256, or 512 with automatic gain change
Channels	
Coincidence	Non-coincidence, anticoincidence, coincidence
Threshold	High, low
Time analysis	
Channel width	1, 2, ···, 64, 128 μs
Multiscaler analysis	External clock is required
Multiparameter	MP9, 9 external bits, plus internal A/DC bits MP18, 18 external bits
Memory subdivision	Quadrant and half routing Overflow count in last channel of sector
Readout	
Normal: Full address and data	27 bits/address
Condensed: pulse-height analysis only	14 bits/address 3 bits of address 8 most-significant data bits 3 bits of data multiplied
Power	
Standby	7.15 W
20,000 events/s	12.2 W
Complexity (approximately)	500 IC flatpacks 400 discrete semiconductors

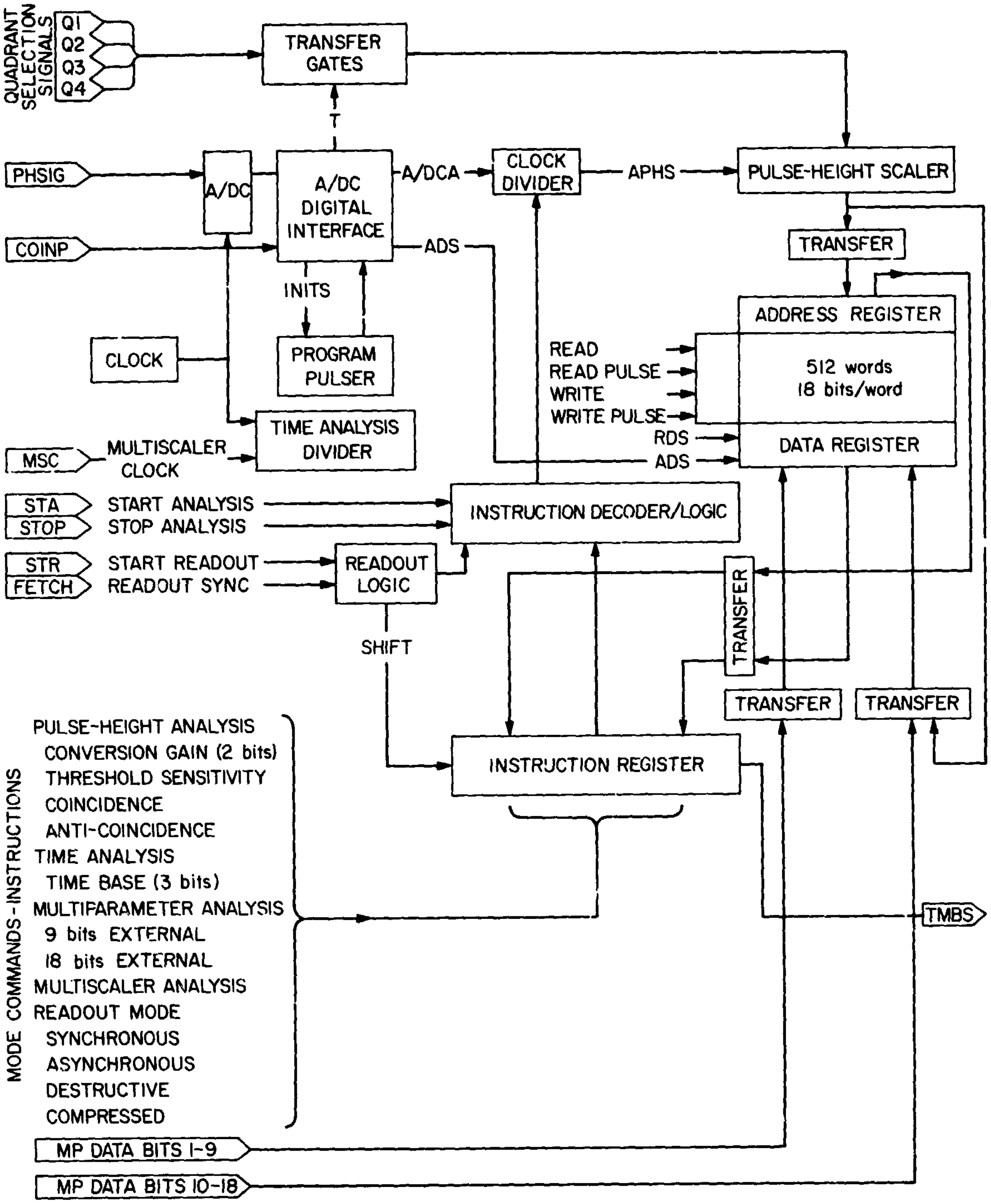


Fig. 1. Functional diagram of pulse-heigh' analyzer

2. Pulse-Height Analysis Mode

a. Analog/digital converter. The A/DC, shown schematically in Fig. 2, is a conventional Wilkenson type. The input circuit is a capacitively coupled emitter follower (Q1). Threshold control is provided (through Q2) by holding the emitter of Q1 above its base line value by the amount on the desired threshold. Thus, the peak value of the pulse is not altered by the threshold setting. Emitter follower Q1 drives the linear-gate emitter follower (Q4). The linear-gate shunt switches (Q5 and Q6) are required only to sink the current supplied from the load resistor (R8) of the linear-gate emitter follower.

The linear gate is normally open and is closed after a pulse peak has passed. The reference voltage for end of rundown is the output of the closed gate and, thus, is near ground potential and is independent of the threshold settings. Since neither the pulse peak nor the rundown reference is altered by the threshold setting, the position on pulses above the threshold are unaltered, while those pulses below the threshold are eliminated. The threshold may be altered during data accumulation without smearing the spectrum.

The stretcher amplifier (Q7 through Q10) acts to keep the voltage on the stretcher capacitor equal to the linear-gate output. Due to the unilateral nature of the charging diode (D5), this action is only possible when the linear-gate output is greater than the capacitor voltage. Accordingly, the amplifier is effective in charging the stretcher capacitor, while discharge is accomplished through the current-sink transistors (Q11 and Q12). The constant-current-sink transistor (Q11) is necessary to provide

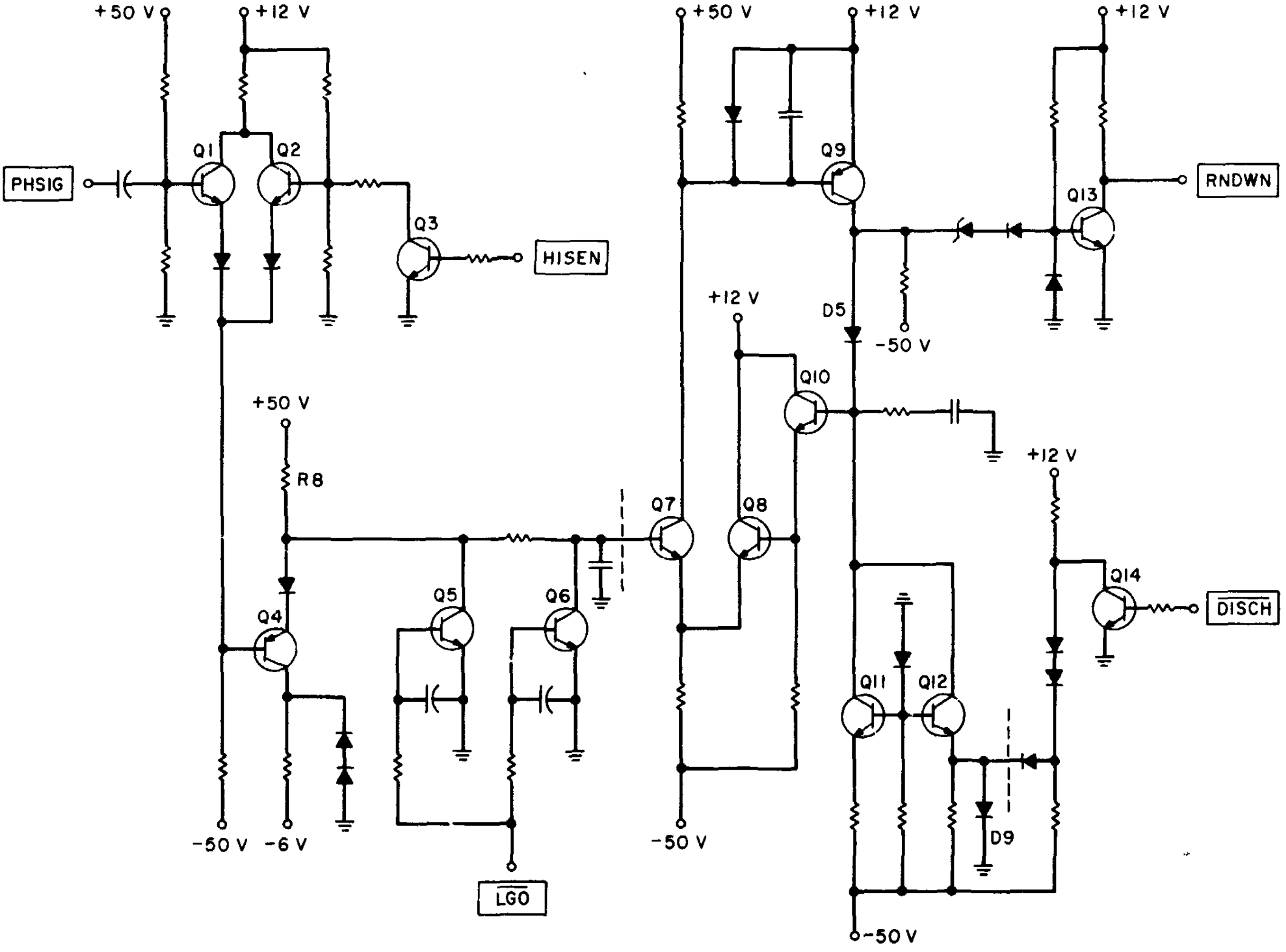


Fig. 2. Schematic diagram of analog/digital converter

negative corrections to the stretcher-capacitor voltage during base line keeping. The linear discharge to the reference is provided by the switched sink transistor (Q12).

The current-sink transistor is switched by diverting its emitter current through D9. The output of the stretcher amplifier (Q9) is a convenient source of a rundown signal ($\overline{\text{RNDWN}}$), since it goes negative as soon as the pulse peak has passed and remains so until the stretcher-capacitor voltage is again equal to the linear-gate output when the amplifier regains control.

The problems associated with pulse spectrometry are, in part, those of measurement precision and, in part, those associated with pulse-to-pulse interference brought on by the random nature and dynamic range of the nuclear phenomenon. In this pulse-height analyzer, the latter problems are handled by the digital interface between the A/DC and the balance of the analyzer functions. No means are incorporated to discriminate against nearly coincident pulses where their sum results in a monotonic increasing pulse. The decision not to incorporate pulse-shape discrimination was made primarily on the basis of the low event rates expected from spaceborne experiments.

*b. **Rundown control.*** The logic diagram of the analog/digital converter and its interface are shown in Fig. 3. The first indication that a pulse has been received is the occurrence of the rundown signal (RNDWN). The state of the discharge flip-flop (DISCH) follows $\overline{\text{RNDWN}}$ on the succeeding negative clock transitions. $\overline{\text{DISCH}}$ gates on the constant-current discharge of the stretcher capacitor and, together with the memory-busy flag (P0) and the pulse-height analyzer mode control signal, controls the analog/digital converter advance (A/DCA) pulses, which eventually advance the pulse-height scaler. DISCH also sets the T flip-flop that, together with RNDWN, causes the initiate-storage-cycle command (INITS).

*c. **Program pulser.*** The storage cycle for all modes of the analyzer is controlled by the program pulser. The pulser consists of a four-stage serial carry counter operating at a clock-derived 1-MHz frequency. When the counter is initiated, it generates 15 sequential 1-μs intervals (1P1 through 1P15) and locks up in the 16th state (P0 is true). Since the memory cycle for pulse-height analysis is completed during the 15 μs of $\overline{\text{P0}}$, it is convenient to use $\overline{\text{P0}}$ as a memory busy flag. The program pulser is initiated by INITS and sustained by $\overline{\text{P0}}$ through the completion of its cycle. INITS = T • RNDWN is an indication that a pulse has occurred, that a corresponding count has been accumulated in the pulse-height

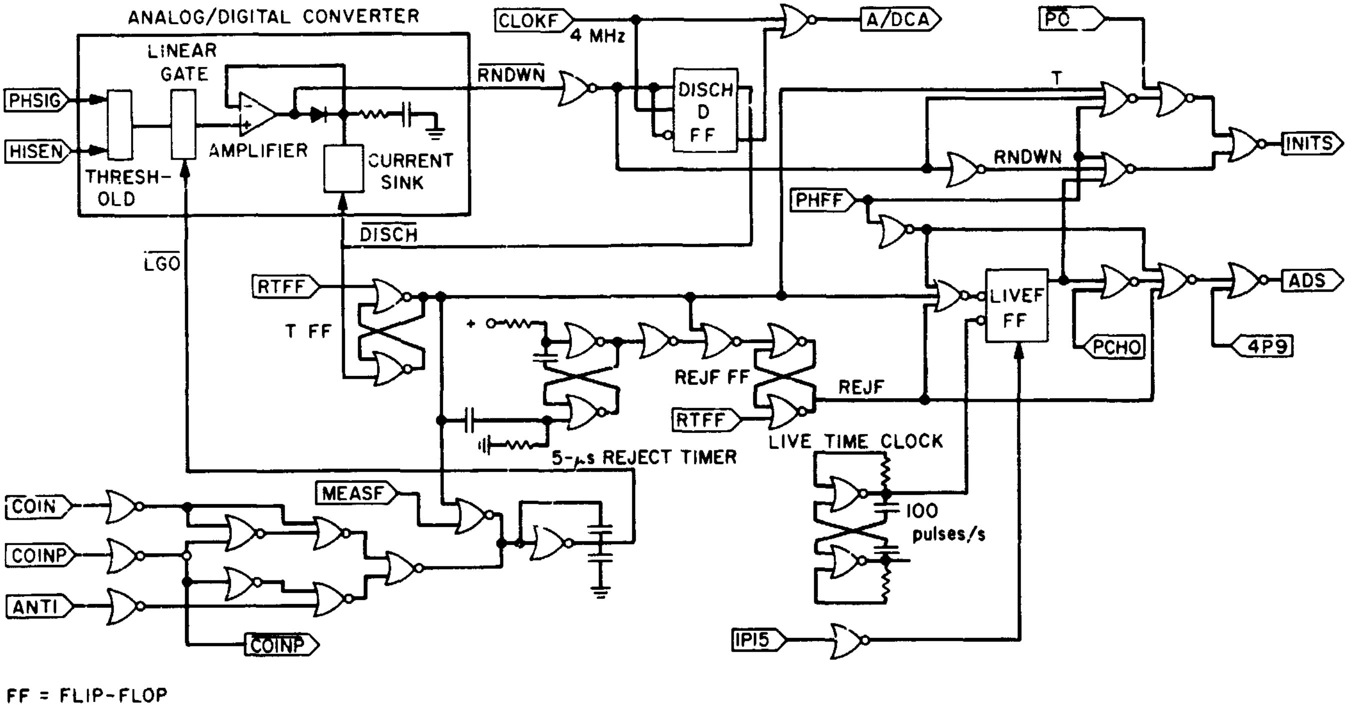


Fig. 3. Digital interface

scaler, and that the storage cycle should proceed. All storage-cycle pulses are decoded from the program pulser.

d. Linear-gate control. The linear gate is controlled in two ways; first, by the coincidence pulse in conjunction with the "coincidence mode" instruction, and second, by RNDWN or T, once a pulse has been detected. Thus, the gate will close during every coincidence pulse in the anti-coincidence mode. Even if there is no pulse-height signal received, an analysis will occur, since the threshold circuit output is greater than the output voltage of the closed gate. This is an undesirable mode of operation, since it has the effect of decreasing live time and of storing unwanted data in the address corresponding to the threshold voltage.

e. Reject circuits. At the end of the normal storage cycle, RTFF (at 1P14 time) causes reset of the T flip-flop, and the linear gate opens. Should this occur during the tail of a pre-existing pulse, an erroneous analysis would result. This condition is avoided by rejecting any analysis data that occurs within 5 μs of the linear-gate opening. This is accomplished by REJF, which prevents the advance of the data scaler during the storage cycle. Thus, rejection does not perturb the linear portions of the analyzer.

f. Memory subdivision and analog/digital converter conversion gain. Change in resolution or in conversion gain refers to the number of quantization levels used to measure pulse height. Conceptually, the discharge rate of the stretcher capacitor could be changed by altering the magnitude of the discharge generator current. It is preferable to leave the analog circuits of the A/DC unchanged and merely to alter the clock frequency. A change of the clock frequency prior to the A/DC gating circuits would result in an increase in the uncertainty of the stretcher capacitor discharge prior to INDRN. In the analyzer, a counter that can divide the A/DCA pulses by 1, 2, or 4 is provided between the A/DC and the pulse-height scaler.

g. Pulse-height scaler. The pulse-height scaler is a conventional ripple-carry counter, with provision for reset and for parallel output. It also has provisions for indicating pulse-height scaler full at a count of 511 and for indicating pulse-height scaler zero (PCHO). The former is used to prevent overflow and has the effect of indicating all overflow or off-scale pulses in address 511. The zero indicator has a special use during live-time determinations, as will be clear later. References made here to address 511 imply that the measurements were made at maximum resolution. The logic generating the full and zero signals is altered—as is the A/DCA divider—by the memory subdivision signals indicating quarters or halves, as is required. In such a case, the most significant stages of the pulse-height scaler are conditioned by externally generated quadrant-selection pulses.

h. Memory. The analyzer uses a conventional magnetic core memory containing storage for 512 words, each 18 bits in length. The cores are arranged in bit planes of 16 × 32 cores. Each particular core in the Nth plane then represents the Nth bit in one of the 512 words. When the memory is to be read, the address register is cleared, and the pulse-height scaler states are transferred in. This occurs in 4P4 time, 4 μs after the end of INDRN. The read signal, generated at 2P7 time, energizes the memory address decoding gates. These gates consist of both current sources and sinks which, in combination with routing diodes, route the read pulse half-select currents to one of the 16 Y wires and to one of the 32 X wires, simultaneously. The core at the intersection of the energized X and Y wires receives the full select current. Such an intersection exists once, and only once, on each of the 18-bit planes. The read pulse is generated at 1P9 time, starting 1 μs after the decoding gates are energized by the read signal. Both read and read pulse coexist for 1 μs, and during that time, the combined action of the X and Y half-select currents drive the selected cores to the reset state. The 511 cores in each bit plane that receive only one half-select current remain in their original state.

Each core of each bit plane is threaded by a single sense wire. If the selected core in the Nth bit plane is originally in a 1 state, a voltage will be generated in the sense wire as the core is reset. This sense voltage is amplified and is used to set the Nth flip-flop in the data register. In practice, the sense line contains a considerable amount of noise voltage induced by the leading and trailing edges of the read pulse. Time domain filtering is used to enhance the sense voltage signal-to-noise ratio.

The data register flip-flops are connected both for parallel entry from the sense amplifiers and for counter operation. During the pulse-height analysis read-add/one-write cycle, the data register is advanced at 4P9 time. Immediately, the write decoding gates are energized by the write signal at 2P13 time. The write pulse occurs at 1P14 time and causes half-select currents in the opposite direction to the read half-select currents to be generated.

Each core of each bit plane is threaded by a single fourth wire. This wire is energized with a half-select current in such a direction as to oppose the write half-select currents. The selected core in a bit plane where the fourth, or inhibit, wire is energized remains in the reset state. The selected core in an uninhibited bit plane is driven to the set state. The inhibit winding is energized with signals derived from the data scaler, itself, thus permitting rewriting of the modified register contents.

The time required for a read-add/one-write cycle is 5.6 μs. As a power conservation measure, both decoding gates and sense amplifiers are energized only during read and write pulses.

i. Live timer. The live timer provides a measure of the time during which the analyzer is available for the measuring of pulse heights. This measurement is accomplished by sampling the combined REJF and T functions with a 100 pulse/s clock. A coincidence of these signals sets the live-time flip-flop, initiates a storage cycle, and closes the linear gate. Since live-time data are to be accumulated in address 1, the address register is advanced from 0 to 1 at the 1P1 time of each storage cycle. The cycle proceeds to 4P9 time when the data scaler is to be advanced, indicating that the analyzer was interrogated and was found to be live. The advance is made conditionally on the state of the pulse-height scaler. A pulse-height scaler state other than 0, with the live-time flip-flop set, indicates a coincidence between a pulse-height signal and a live-time clock pulse. In such a case, the pulse-height signal must have been at least partially stored in the stretcher before the gate closed. When this condition is observed, both the live time and the signal pulse are lost.

j. Readout. The instruction register also serves as the output register for the analyzer during data readout. The readout function, itself, is controlled by the shift counter, the address register, and the control logic.

The first data shifted out of the analyzer is that stored in the instruction register, and tells the user exactly what the conditions of the analysis were. After the first 9 bits have been shifted out, the contents of the memory address register are transferred into the cleared positions. As shifting continues, the address register is advanced by 1, and the new address is read from the memory into the data register. After 18 additional shifts have occurred, the contents of the data register are transferred to the shift register. As shifting continues, the address register is advanced, transferred to the shift register, and the memory is read again. The data sequence is thus 18 bits of instruction data, followed by address 0, next data from address 1, followed by address 1, and so on, until all addresses have been read. Note that address 0 is never read out. Address 0 should have no data, since the address register has been advanced by 1 as a routine part of every storage operation.

3. Time-Analysis Mode

The analyzer has been designed to allow measurement of energy and die-away spectra of capture gamma rays, as might be obtained using a pulsed neutron source. When so instructed, the analyzer will, on command, begin a time measurement. The measurement consists of 126 intervals of 1 to 128 μs in duration, as instructed. When a pulse-height signal is received, a storage cycle is initiated, and the content of the address corresponding to the appropriate interval is advanced by one. Interval timing continues without interruption, but additional pulse-height signals do not initiate storage. When timing is complete, a second storage cycle is initiated, and the content of address 127 is advanced by 1 to indicate the total number of timing cycles that have been completed.

4. Combined Time and Pulse-Height Analysis

Pulse-height analysis on the first pulse-height signal proceeds concurrently with the time analysis described above. The result of the analysis is retained in the pulse-height scaler. When the time analysis and both storage cycles are complete, the pulse-height data are stored in the selected quadrant.

5. Multiparameter Modes

In the multiparameter mode the analyzer merely provides a means of recording 512 words of 18 bits each. On command, such data are transferred into the data register of the memory, written into the core, and the address register is advanced by 1. When the last address has been used, a memory full gate is set, and no further inputs are accepted.

Provision has been made to record 9 bits of multiparameter data, together with 9 bits of pulse-height analysis data in each address. In this case, a coincidence pulse is required to indicate the pulse to be analyzed and the data to be entered.

6. Multiscaler Mode

A conventional multiscaler mode has been provided. In this mode, interval pulses are provided externally. Pulses to be scaled are used to advance the data register of the memory. When the interval pulse arrives, a storage cycle is initiated, and the accumulated count in the data register is stored in the appropriate address. During the storage cycle of 16 μs, the scaler is inactive. When the last address has been used, a memory-full gate is set and no further inputs are accepted.

7. Conclusions

The versatility provided by this analyzer is greater than that provided in many laboratory instruments and, hence, it should be suitable for most space science experiments. Admittedly, this instrument may have an excess functional capability and be somewhat less efficient (in terms of power and weight) for specific applications. However, the excessive costs for developing a special instrument for a specific application favor the use of this analyzer design.

B. *Quantitative Use of Imaging Systems: An Electronic Camera System,*

A. T. Young and F. P. Landauer

1. Introduction

The purpose of this work is to develop an imaging astronomical photometer with both high photometric accuracy and high spatial resolution. Accurate, high-resolution photometric data are needed in a wide range of planetary and stellar investigations. For example, the problems of the clouds of Venus, the nature of seasonal changes on Mars, the dynamics and structure of Saturn's rings, and fundamental studies of stellar masses and evolution, all require such observations.

At the present time, low-resolution data of high photometric accuracy are obtained photoelectrically, moderate-resolution data of moderate accuracy are obtained photographically, and high-resolution data of low accuracy are obtained visually. In these conventional techniques, the "seeing" (image blurring produced by the Earth's atmosphere) is a major limitation, and has been regarded as an insuperable limitation. However, recent advances in understanding the "seeing" problem (Ref. 1) have shown that (1) the resolution advantage of visual or short-exposure photographic observations can be realized in longer exposures if the image motion is cancelled by an automatic guider, and (2) the remaining image blurring has the effect of attenuating high spatial frequencies in the image, which can then be "restored" by suitable image processing. Such image-restoration techniques have been developed by the image processing laboratory at JPL, and successfully applied to *Mariner IV* and *Surveyor* data.

The "restoration" of high spatial frequencies requires that the image be recorded by a *linear* process, and that the signal-to-noise ratio be so high that even the attenuated spatial frequencies are larger than the corresponding components of the noise. Conventional photographic recording is strongly nonlinear, and gives signal-to-noise ratios of about 30 to 50 at best. Furthermore, the detective quantum efficiency of photography is low, typically a few tenths of a per cent, so that telescope time is not used effectively.

These difficulties can be reduced by detecting and recording the image electronically. At the present time, the image isocon appears to be the most suitable detector, with good linearity, wide dynamic range, and excellent signal-to-noise ratio. With slow-scan readout and FM recording on magnetic tape, a detective quantum efficiency of a few per cent and a signal-to-noise ratio of at least 100 can be expected. An electronic camera has a considerable advantage over photography, since each picture element can be individually calibrated by exposure to a series of known light levels. The tape recording has the additional advantage of much faster conversion of data to digital form than can be achieved by scanning a photograph mechanically on a microdensitometer.

Figure 4 is a block diagram of the entire electronic camera system. From a systems point of view, the earth's atmosphere must be included with the telescope in determining the optical modulation transfer function. At the present time, we are concerned with the design of the portion of the system above the dashed line.

2. Electronic Camera System Design

a. Telescope and Earth's atmosphere. The telescope used will be primarily the 24-in.-diam telescope at Table Mountain, California. However, it may be desirable to use other telescopes (e.g., at McDonald or Kitt Peak), and provision should be made for mounting the equipment on other telescopes.

The work of Fried (Ref. 1) has shown that there is an optimum aperture for a given wavelength and "seeing"

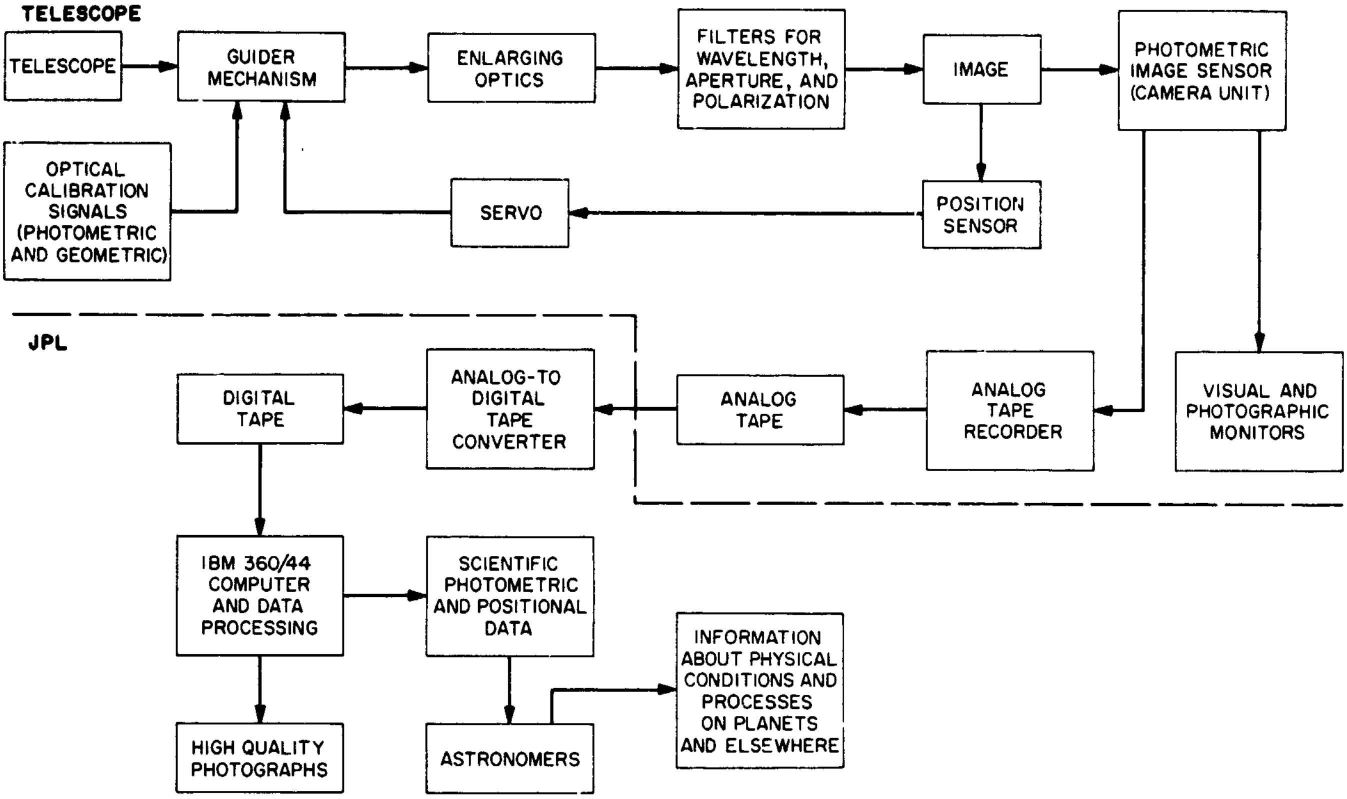

TELESCOPE
TELESCOPE
GUIDER MECHANISM
ENLARGING OPTICS
FILTERS FOR WAVELENGTH, APERTURE, AND POLARIZATION
IMAGE
PHOTOMETRIC IMAGE SENSOR (CAMERA UNIT)
OPTICAL CALIBRATION SIGNALS (PHOTOMETRIC AND GEOMETRIC)
SERVO
POSITION SENSOR
JPL
DIGITAL TAPE
ANALOG-TO DIGITAL TAPE CONVERTER
ANALOG TAPE
ANALOG TAPE RECORDER
VISUAL AND PHOTOGRAPHIC MONITORS
IBM 360/44 COMPUTER AND DATA PROCESSING
SCIENTIFIC PHOTOMETRIC AND POSITIONAL DATA
INFORMATION ABOUT PHYSICAL CONDITIONS AND PROCESSES ON PLANETS AND ELSEWHERE
HIGH QUALITY PHOTOGRAPHS
ASTRONOMERS

Fig. 4. Planetary photometry (electronic camera system)

quality. A review of current knowledge of the seeing problem will appear shortly in *Sky and Telescope*. Because the modulation transfer function of the telescope–atmosphere combination depends on the telescope aperture, Fried's work is being extended to annular apertures. Aperture filtering will be included in the optical head.

Because of the importance of atmospheric dispersion in high-resolution observations, relatively narrowband filters must be used. Partial compensation for dispersion is possible, but at the expense of more optical elements and two additional continuously varying degrees of freedom in an already complex system. Narrowband filters are desirable in planetary work, regardless of the dispersion problem.

b. Guider. Several alternative methods of sensing and controlling the position of the image are being investigated. Calculations indicate that adequate bandwidth will be available to guide on any naked-eye object in the blue and visible; for fainter objects, or in the infrared, the accuracy of motion compensation will be limited by photon noise.

c. Camera head and control system. The detailed electronic design of the camera system is essentially complete. A description of the electronic system follows.

Camera head.

Image detection. Recent data on the RCA type-C21093 image isocon indicate substantial superiority over image orthicons. With a close-spaced mesh and 10^{17} Ω-cm target, linearity and readout efficiency are high, and storage (integration) will be possible for several minutes. The bialkali photocathode is reported to be stable, uniform, and highly sensitive. With an expected detective quantum efficiency of a few per cent, the required exposure times to achieve 1% precision per picture element will be on the order of 0.2 s for Venus, 4 or 5 s for Mars and Jupiter, and 1 min for Saturn.

Preamplification. The output signal from the isocon will range from a few nanoamperes to several microamperes. This requires a gain adjustment of from 20 to 20,000 in the video preamplifier. It is anticipated that the existing nuvistor preamplifier will be augmented by programmed operational amplifiers to control gain and bandwidth.

Deflection. An operational amplifier-type deflection amplifier will be used but will not be an integral part of the camera head. Deflection waveforms and regulated focus and alignment currents will be remotely generated in the control console.

Control section.

Sweep and format generation. Horizontal line generation will be by means of binary division of a crystal-controlled clock. The sweep drive signals will be coupled to operational amplifier integrators operating through an integrate–reset bridge. The vertical drive will be derived by binary division of horizontal drive pulses. Thus, both the horizontal and vertical line numbers can be controlled by digital selection.

Because image orthicons and isocons are complex in operation, two sweep modes will be used: (1) a slow-scan mode limited by tape recorder bandwidth versus signal-to-noise ratio constraints, and (2) a fast-scan mode limited by the reset capabilities of the yoke drivers and yokes. The fast-scan mode is for visually monitoring the preliminary adjustment of the camera operating parameters; the slow-scan mode is for data recording. It will have constant sweep rate and video bandwidth, i.e., the raster size will change as the line number changes. This is consistent with the requirement for constant optical magnification and Nyquist sampling at the optical resolution limit. The reason for changing the line number at all is that if only a small field is required, then considerable time is wasted if an unnecessarily large target area is scanned. The formats to be accommodated are 256 × 256, 512 × 512, and 1024 × 1024 Nyquist samples. The fast system differs in that the active line time is constant, the video bandwidth varies, the line number changes, and the clock frequency and integration rate is eight times faster.

Exposure and Erasure. Exposure times will be generated in increments of powers of the square root of two by binary division of crystal clock. Automatic exposures from 1 ms to 45 s ($2^{15.5}$ ms) will be selectable.

Erasure will be available only during slow-scan operation. The sequence will consist of switching the system to fast scan, gating the target to a higher potential and increasing the beam current for one frame, then returning to slow-scan mode. Erasure scans will always be done in the 1024-line format. In slow-scan operation, the sweep will be disabled during exposure to eliminate the effects of a changing magnetic field on the image section of the camera tube. In fast scan, however, no hesitation will occur unless the selected exposure time exceeds the frame time; thus providing the fastest possible frame rate for setup.

Control of optical functions. Besides controlling the camera unit, the control unit must step the optical filters and calibration targets through an appropriate calibration sequence. The calibration procedure is complex; it provides adequate photometric and geometric calibration data for each combination of wavelength, polarization, and aperture filters selected. The control unit also sequences the filters and other optical and electronic adjustments during the observational cycle.

Visual monitor. The electrical characteristics of the display monitor will be similar to those of the film recorder. The format will be a minimum of 6 in.^{2} on a spherical faceplate. The resolution will be a minimum of 1000 lines using a dual-mode phosphor, i.e., different colors for phosphorescence and fluorescence so that when proper filters are used, the monitor will be suitable for both fast- and slow-scan rates.

A Tektronix RM 561 oscilloscope will be used for A-scope monitor. A type 3B3 time-base unit will be modified to accept external sweep from the control section, and to enable the delayed sweep gate to function as a line brightener on the monitor when the A-scope is used as a line selector.

Film recorder. A *Ranger* Block III film recorder will be used to provide 35-mm film output. Reliability and stability of the unit is being improved by the replacement of the high-voltage and focus power supplies with ultra-stable solid-state units. As in the camera head, the yoke driver will be a wideband, ultralinear, solid-state operational amplifier. It will receive its sweep waveforms from the control section. Resolution will be better than 1000 lines on 2.5 in. of the cathode-ray tube faceplate.

Tape recorder. Data will be recorded on magnetic tape using an Ampex FR1400 or equivalent operating at 120 in./s. The tape will be similar to *Ranger* and *Surveyor* analog tapes, i.e., frequency modulation by video data. Separate tracks will be used for vertical and horizontal sync signals because there is no reason to generate composite video. Sufficient telemetry data will be recorded to provide a record of the system operating parameters. A file number will also be recorded to aid in data extraction.

Reference

1. Fried, D. L., *J. Opt. Soc. Am.*, Vol. 56, p. 1372, 1966.

C. On the Slow-Scan Characteristics of the WX30691 SEC Vidicon, *K. J. Ando*

1. Introduction

A continuing task of the JPL image detector laboratory is the evaluation of new imaging devices for possible application in future interplanetary missions. One type is the secondary electron conduction (SEC) vidicon. The SEC vidicon is particularly suited for space applications due to its inherent simplicity and high sensitivity. SEC tubes have been selected for various future space systems, the most important ones being the *Apollo* missicn and the *Apollo* Telescope Mount program.

For any space application, the imaging device must be sufficiently rugged to withstand the severe environment of launch and a long-term flight. Further study and developmental work will be necessary to determine whether ruggedization of a SEC vidicon is feasible. Present results indicate that it may be. The SEC vidicon has already met many MIL specification shock and vibration requirements which specify typical levels encountered on airborne flights.

The present article discusses some results of the evaluation of the Westinghouse WX30691 SEC vidicon. The main purpose of this work was to determine the slow-scan capabilities of the WX30691 and provide information on the general characteristics of the SEC vidicon.

2. Brief Description of the SEC Vidicon

The SEC process and the SEC vidicon are described extensively in a series of papers (Ref. 1) by the Westinghouse group which developed it. Thus, only a brief description will be given here.

The SEC vidicon has many features and characteristics which are identical to those found in an image orthicon and a conventional vidicon. Figure 5 shows a simple schematic diagram of an SEC tube. Basically, the SEC vidicon consists of an image intensifier section coupled to a vidicon readout section. The tube has a fiber optic input window which couples light from the image plane to a hemispherical photocathode layer. The secondary electrons from the photocathode are accelerated and focused onto the SEC target. The resultant charge is stored on the surface of the target. Readout is accomplished by a reading beam in the conventional manner.

The unique feature of the SEC vidicon is the target, which is depicted schematically in Fig. 6. It consists of three layers. A thin layer of aluminum and a porous KCl layer are evaporated onto an alumina substrate. The KCl layer is evaporated in such a manner that it has an extremely low mass thickness of between 10 and 100 $\mu g/cm^2$. (For comparison, a solid KCl layer 20 μm thick would have a mass thickness of 1.984 $g/cm^3 \times 20 \times 10^{-4} = 4000\ \mu g/cm^2$.) In operation, the electrons from the photocathode bombard the target with an energy determined by the photocathode voltage. Typically, the electron energy is about 7 keV. The incident electrons pass through the alumina and the aluminum signal electrode layer and generate secondary electrons in the KCl layer.

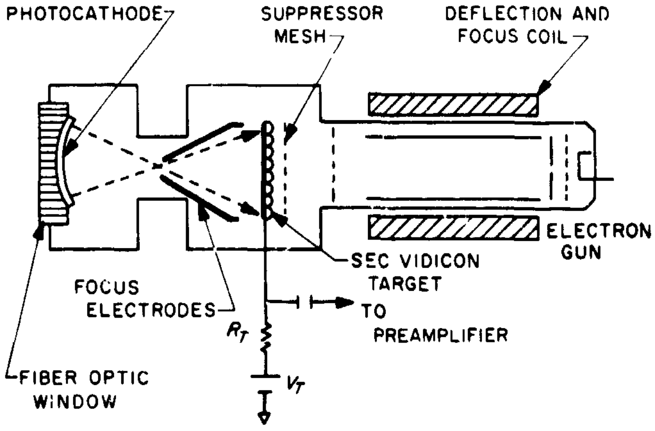


Fig. 5. Schematic diagram of the SEC vidicon

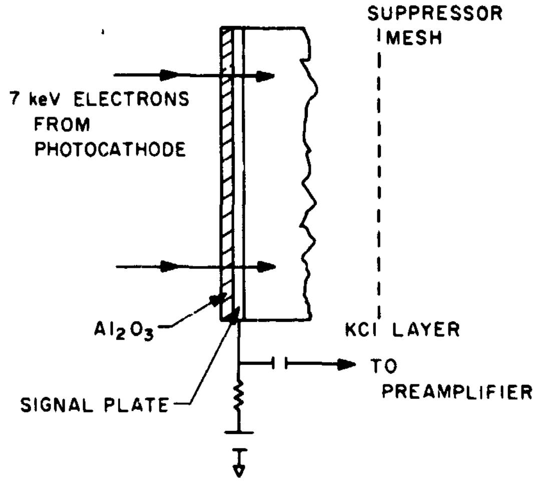


Fig. 6. SEC vidicon target

The alumina and the aluminum layer are sufficiently thin such that the transmittance for 7 keV electrons is high. The secondary electrons generated in the KCl layer are swept to the signal electrode by the reverse bias applied between the signal electrode and the KCl layer's surface which has been charged to cathode potential by the electron beam. The signal electrode is typically biased at between 10 and 20 V. The collection efficiency for the secondaries is largely due to the many voids in the target. Some of the secondary electrons will recombine with positive charge recombination centers, but most will reach the signal electrode. Typical electron gains between 100 to 200 have been achieved in SEC targets. The resulting positive image pattern is neutralized by the electrons from the scanning beam. The target current which discharges the KCl layer through the signal electrode during readout constitutes the video signal.

The most outstanding feature of the SEC target is probably its high resistivity, which permits charge integration and storage for extended periods of time. For low-light-level detection, time exposures up to several hours can be utilized with no reciprocity failure. In addition, the SEC target is capable of storing signals for several days without any significant image degradation.

3. Test Procedures

Most of the data were taken on two WX30691 tubes purchased from Westinghouse. A photograph of the WX30691 SEC vidicon is shown in Fig. 7. Additional data and experience were acquired from the evaluation of two *Apollo* SEC tubes during the previous year, and discussions with Westinghouse engineers at Elmira, New York, in connection with the SEC vidicon ruggedization proposal.

The WX30691 has a 25-mm photocathode. The input raster size is 0.6 × 0.8 in. although other raster sizes bounded by a 1-in. diameter can be used. The input format size is thus 60% larger in linear dimensions than a standard vidicon (½ × ⅜ in.) The read section of the WX30691 is identical to that of a conventional 1-in. all magnetic vidicon so that standard deflection yokes and focus coils can be utilized.

The WX30691s were evaluated in a camera head designed specifically for SEC vidicons (SPS 37-48, Vol. III, pp. 142–146) and the vidicon test facilities. The WX30691s were initially operated at EIA rates to optimize alignment currents, set size and centering, and optical–electrical focus. For slow-scan tests, line and frame rates were set to yield 600 noninterlaced television lines per picture height. A solenoid-driven Wollensak leaf shutter was used to shutter the WX30691 during slow-scan operation.

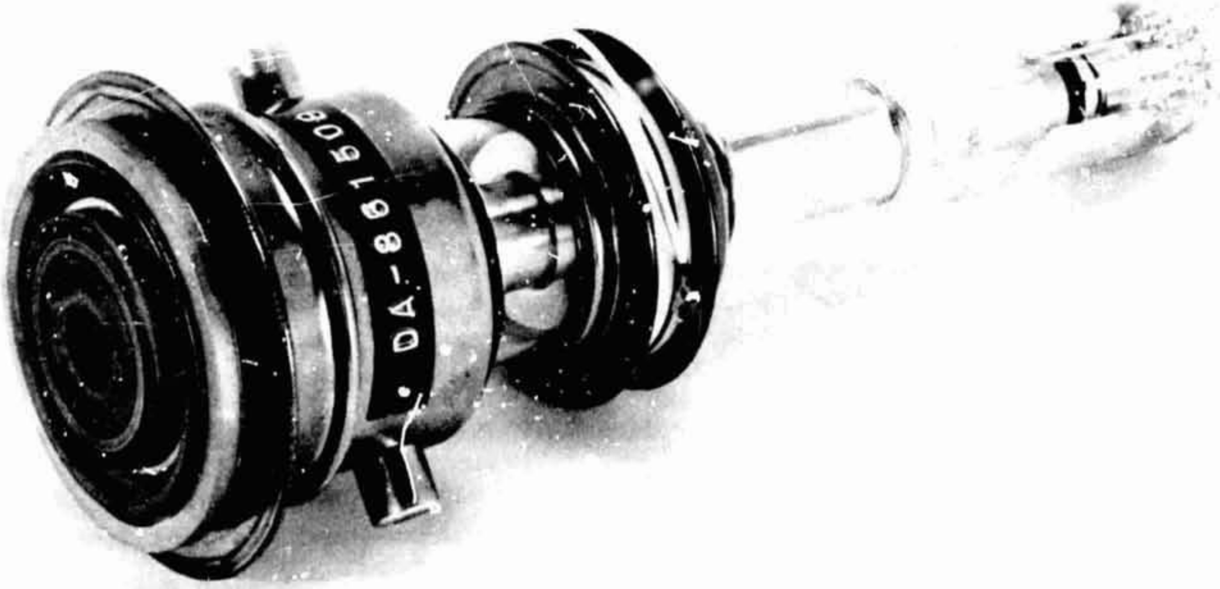


Fig. 7 WX30691 SEC vidicon

Since the sensitivity depends critically on the image area scanned, the fiber optic input faceplate was masked with a C.6 × 0.8-in. template. The proper raster size was set at all scan rates by adjusting size and centering controls, using the template as a reference. Static transfer and sensitivity measurements were made with a tungsten source (2875°K), a series of calibrated Iconal neutral density filters, and a Kietl.ley electrometer.

4. Results

The static transfer characteristics at EIA rates of the WX30691 for various target voltages V_T are shown in Fig. 8. The dynamic range extends typically over two orders of magnitude in illuminance. Since flat field illuminance was used, signal current as measured by the Kiethley can be expressed as a peak c· rrent if the blanking time is taken into consideration. The gamma typically varies from approximately unity at lower illuminance levels to about 0.5 at the saturation point. Beyond the maximum point indicat.d on each of the curves, the target is saturated at the suppressor mesh voltage. This provides a "knee" in the transfer curve. The "knee" region is not as extended as in an image orthicon, and operation in this region is not recommended due to image "burn in" and an abrupt change into a "crossed over" mode which produces "blacker than black" areas on the monitor.

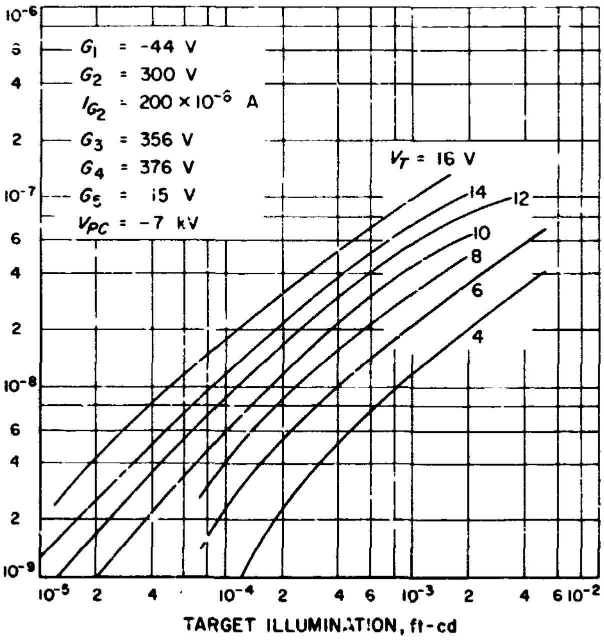


Fig. 8. Transfer characteristics of the WX30691 SEC vidicon at EIA rates

The operating range can be extended to higher illuminance levels by decreasing the target ga.n via a decrease in the photocathode voltage V_{PC}. This can be done in a dynamic fashion by an automatic gain control loop which samples the video level and varies the photocathode voltage accordingly. Such a system is incorporated in the *Apol. ›* lunar TV system. There is no perceptible degradation in resolution with photocathode voltages from 3 to 8 kV. The target gain dependence on photocathode voltage was determined at EIA rates by measuring the signal current as the photocathode voltage was varied with a constant target voltage. Figure 9 shows the resultant relative target gain versus photocathode voltage curve for the WX30691. Maximum target gain occurs at 7 kV. At higher photocathode voltages, the cross section of the target for primary electrons decreases and the gain accordingly drops off. At lower photocathode voltages, a substantial portion of the primary electrons are stopped by the Al and Al_2O_3 substrates with a resultant sharp decrease in gain.

Of particular importance for space applications are the slow-scan characteristics of the SEC vidicon. EIA rate operation requires a disproportionately large bandwidth and is not compatible nor feasible for *Mariner* class missions. It is anticipated that future missions will still

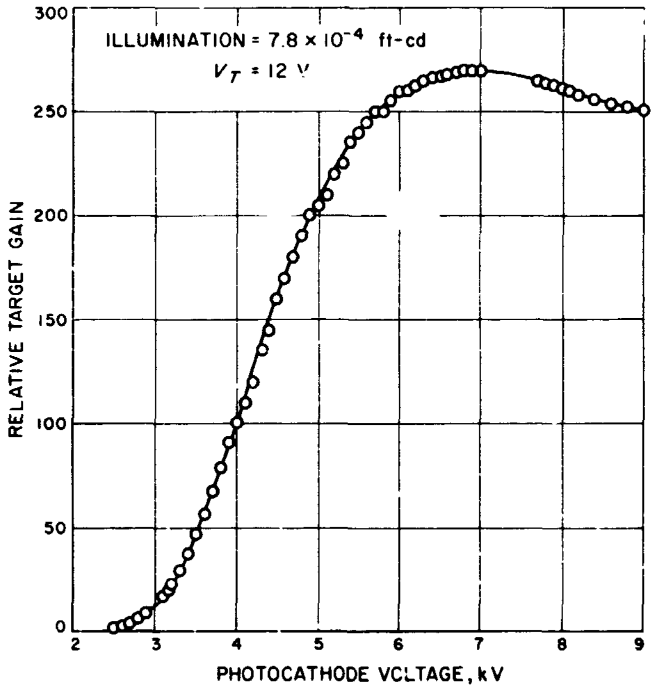


Fig. 9. Relative target gain vs photocathode voltage for the WX30691 SEC vidicon at EIA rates

utilize slow scanning as a means for bandwidth and data reduction. To my knowledge, there has not been any extensive evaluation or data on the slow-scan capabilities of the SEC vidicon due to its limited applicabil:ty. The *Apollo* lunar SEC camera utilizes frame times of 0.1 and 1.6 s, whereas the *Apollo* telescope mount SEC camera is an EIA system. Future *Mariner*-class missions will require operation of the SEC vidicon at longer frame times unless another means is developed to buffer down the data rate from the vidicon.

To determine the slow-scan capabilities of the SEC vidicon, the WX30691 was evaluated at several slow-scan rates. Figure 10 shows a typical transfer characteristic at a frame time of 1 s. Slow-scan operation requires some means of shuttering to stop motion since the long frame times do not permit an open shutter mode without image smear. In order to check reciprocity between shutter speeds and total light exposures, a series of transfer curves were obtained at various shutter speeds between 1 s and 20 ms. Figure 11 shows the superposition of the transfer curves taken with the different shutter speeds. The scatter in the data points is well within the measurement uncertainties. Therefore, within the limited range of shutter speeds utilized, it can be said that shutter speed-light reciprocity holds.

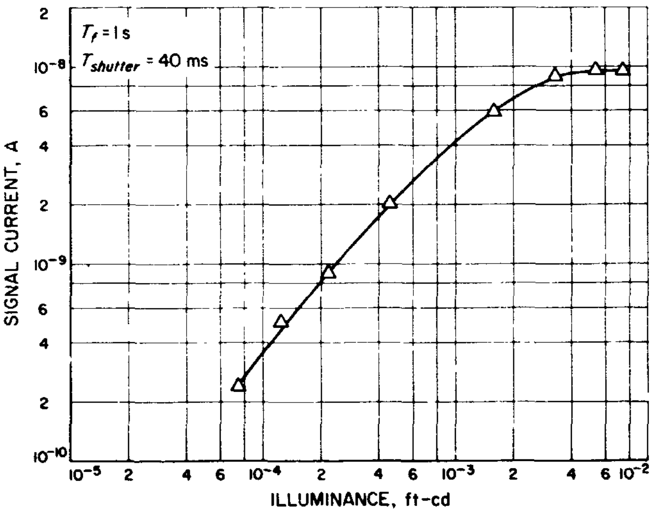


Fig. 10. Light transfer curve for a frame time of 1 s

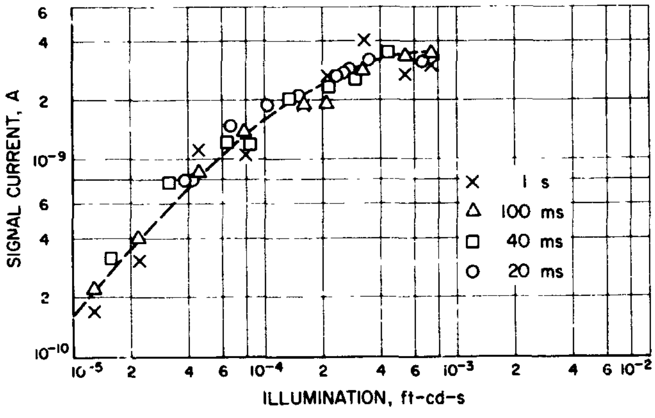


Fig. 11. Light transfer curves for different shutter speeds at a frame time of 4 s

Figure 12 shows the variation in signal current as a function of frame time for a number of constant light energy values within the normal operating region of the WX30691 at a target voltage of 16 V. As anticipated, the signal current drops off proportionately with scan rate. Although the signal current can be increased to some extent by higher target voltage operation with some loss in dynamic range, the WX30691 is limited to frame times that do not exceed 10 s if scanned in the conventional manner due to the degradation in signal-to-noise ratio arising from the decreased signal output. *Mariner* Mars 1969 type slow-scan vidicon, for example, has a typical signal current of 3 nA at 0.1 ft-cd-s at a 42-s frame time compared to an extrapolated maximum current of approximately 0.2 nA at the same frame time for the WX30691. The WX30691 can be operated at longer frame times, however, if special scanning techniques are utilized to achieve short beam dwell times under slow-scan conditions. Such techniques were used in the Uvicon SEC camera (*Orbiting Astronomical Observatory*

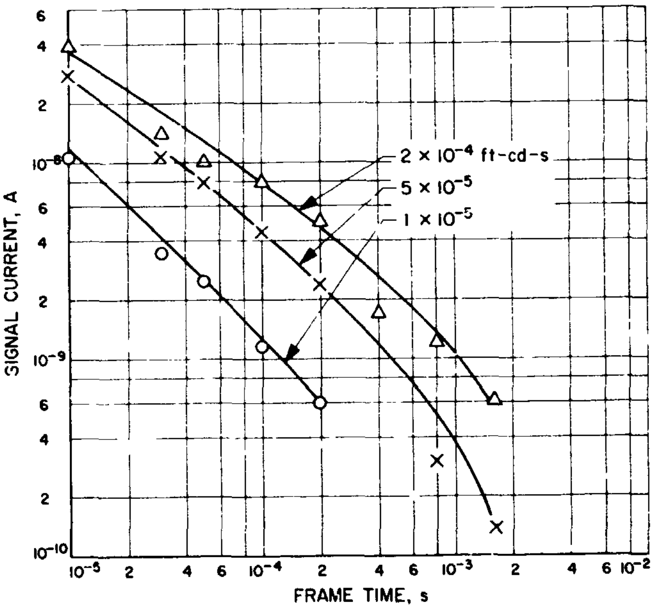


Fig. 12. Signal output vs frame time

satellite) and the *Mariner* Mars 1964 vidicon camera to maintain signal current output at slow-scan rates.

The higher signal current for the slow-scan vidicon is due to the higher electron charge density stored on the target for a given exposure time. The higher electron charge density for a *Mariner* Mars 1969 slow-scan v·dicon is the result of its higher target capacitance ($\approx$ 10,000 pF/cm^2 vs $\approx$ 200 pF/cm^2 for the WX30691). The lower capacitance of the SEC cannot be compensated for by a larger voltage excursion on the target since the maximum practical voltage excursion is limited to about 6 V. Larger voltage excursions result in a "beam pulling" phenomenon which lowers resolution. In addition, the operating region of the WX30691 is typically two orders of magnitude lower in illuminance. The combined photocathode and target gain of the WX30691 is not large enough tc offset this in terms of the total charge stored on the target.

Let us now consider the signal-to-noise ratio of the WX30691 at slow-scan rates in the light of its signal current output. As in all photocathode devices, the signal-to-noise ratio is shot noise limited by the less than unity quantum efficiency of the photocathode material. A typical figure quoted for the sensitivity of an S 20 photocathode is 150 A/lm. If we take the photo flux in 1 ft-c of illuminance as 1.1 × 10^{16}/s for a 2850°K source, 150 A/lm conesponds to a quantum efficiency of 9%. The shot noise in the photoelectron current is given by

$$I_{sn} = (2ei_s\,\Delta t)^{1/2} \qquad (1)$$

Utilizing Eq. (1), the shot-noise signal-limited signal-to-noise ratio can be written as

$$\mathrm{S/N} = \left(\frac{ASI}{ne}\right)^{1/2} \qquad (2)$$

where A = area of the photocathode in ft^2, S = sensitivity of the photocathode in μA/lm, I = illumination in ft-cd-s, n = number of pixels, and e = 1.6 × 10^{-19} A-s.

Figure 13 shows the limiting signal-to-noise ratio for the WX30691 for some typical resolution requirements. This signal-to-noise ratio is never achieved in practice because the preamplifier further degrades the signal-to-noise ratio, and the curves in Fig. 13 should be considered as upper limits.

The preamplifier noise level is not easuy determinable since it depends on the noise parameters for the specific input device and bias conditions. However, the preamplifier noise levels can be characterized by utilization of the concept of noise figure. Preamplifier noise is assumed to contribute additive "white" noise. This is a reasonable assumption since slow-scan operation requires only limited video basebands which can easily be shifted beyond the $1/f$ noise region of transistors by carrier or chopping techniques. Adding the vidicon shot-noise and load-resistor thermal noise in quadrature, the total equivalent input noise current can be expressed as

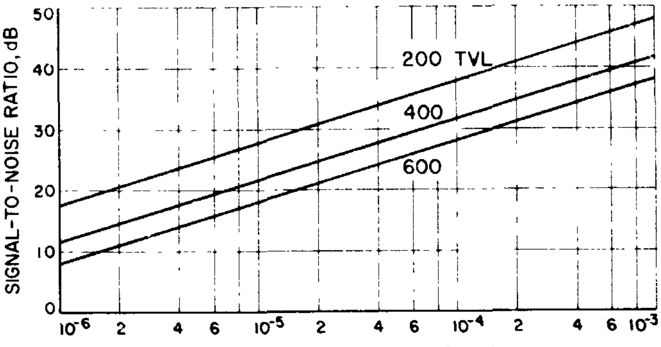


Fig. 13. Limiting signal-to-noise ratio for the WX30691 SEC vidicon for some typical resolution requirements

$$i_n = (4kT\,\Delta f F)^{1/2}\left(20\alpha i_s + \frac{1}{R_L}\right)^{1/2} \qquad (3)$$

where α is a parameter which characterizes the noise buildup in the stored charge, F = noise figure for the preamplifier, R_L = load resistor, i_s = signal current from vidicon, and Δf = video bandwidth.

In practice, $1/R_L \gg 20\alpha i_s$. Therefore,

$$i_n = \left(\frac{4kT\,\Delta f}{R_L}\right)^{1/2}(F)^{1/2} \qquad (4)$$

The signal-to-noise ratio is thus

$$S/N = i_s\left(\frac{R_L}{4kT\,\Delta f F}\right)^{1/2} \qquad (5)$$

For a given resolution requirement, the minimum bandwidth is given by

$$\Delta f = \frac{0.5\,N_x N_y b}{T_f} \qquad (6)$$

where N_x and N_y are the horizontal and vertical resolutions in TVL/picture height, b = blanking factor =

T_f/T_f (unblanked), X/Y = aspect ratio, and T_f = frame time.

Since the signal current output is approximately proportional to $1/T_f$ up to frame times of 10 s, from Eqs. (5) and (6) it follows that

$$S/N \approx (R_L \Delta f)^{1/2} \qquad (7)$$

Thus, the signal-to-noise ratio can be maintained at slow-scan rates up to 10 s by increasing R_L proportionately as the bandwidth is decreased. For typical values of 3 dB for the noise figure ($R_L = 500\,\Omega$, $T_f = 1$ s, and $N_x = N_v = 600$ TVL), the maximum highlight signal-to-noise ratio calculated for the WX30691 is 34 dB. A typical highlight signal-to-noise ratic for a *Mariner* Mars 1969 vidicon–preamplifier combination is approximately 50 dB. The higher signal-to-noise ratio is a result of the higher signal output of the slow-scan vidicon within its operating region, lower bandwidth due to the longer frame time, and a larger R_L.

5. Conclusions

The feasibility of operating the WX30691 at slow-scan rates up to 10 s was demonstrated. Longer frame time operation is not practical using conventional scanning techniques. Although there are other versions of the 25-mm SEC vidicon modified to favor specific system requirements, the performance characteristics of the WX30691 are typical of what can be expected from a 25-mm SEC vidicon. Further evaluation of the resolution, image storage, and spectral response of the WX30691 is in progress.

Reference

1. Goetze, G. W., et al., *Advances in Electronics and Electron Physics,* Report 22A, pp. 219–262. Westinghouse Electric Corporation, Elmira, N. Y, 1966.

XVII. Science Data Systems

SPACE SCIENCES DIVISION

A. Digital Techniques for Generating a Time-Dependent Acceleration Voltage for a Mass Spectrometer, *M. Perlman*

1. Introduction

A mass spectrometer can be used to determine the composition and relative abundance of the constituents of a planetary atmosphere. This article discusses a technique for the digital generation of the acceleration voltage of such an instrument.

The instrument first considered was a single-focusing mass spectrometer (Ref. 1), the essential components of which appear in Fig. 1. The instrument portion is shown in its mechanical configuration, whereas the support electronics are represented by functional blocks.

2. Instrument Operation

The gas to be analyzed is introduced into the ionization chamber, where a portion of it is ionized when bombarded by an electron beam that is parallel to the source exit slit. The high-voltage sweep produces an electrostatic field that accelerates the ions through the source exit slit with approximately uniform energy The resulting ion beam is deflected by the electromagnetic field of the analyzer (permanent) magnet such that, at a given value of v (high-voltage sweep), all ions with a particular mass-per-unit-charge are focused on the collector defining slit. The ion current is collected and fed into a sensitive operational amplifier called an electrometer. Automatic scale switching provides a large dynamic range.

A monotonically varying v is used to separate ions with different masses-per-unit-charge. A plot of the ion current versus time (resulting from a monotonically varying v) yields a spectrogram. The location of a peak in time identifies the associated mass-per-unit-charge, and the amplitude of the peak gives the relative abundance.

The instrument's resolution is an important parameter. The mass-per-unit-charge, M/q, is in atomic mass units (amu) where the isotope ${}^{16}_{8}O$ is taken to be 16. It differs slightly from the chemical scale of atomic weights (Ref. 2).

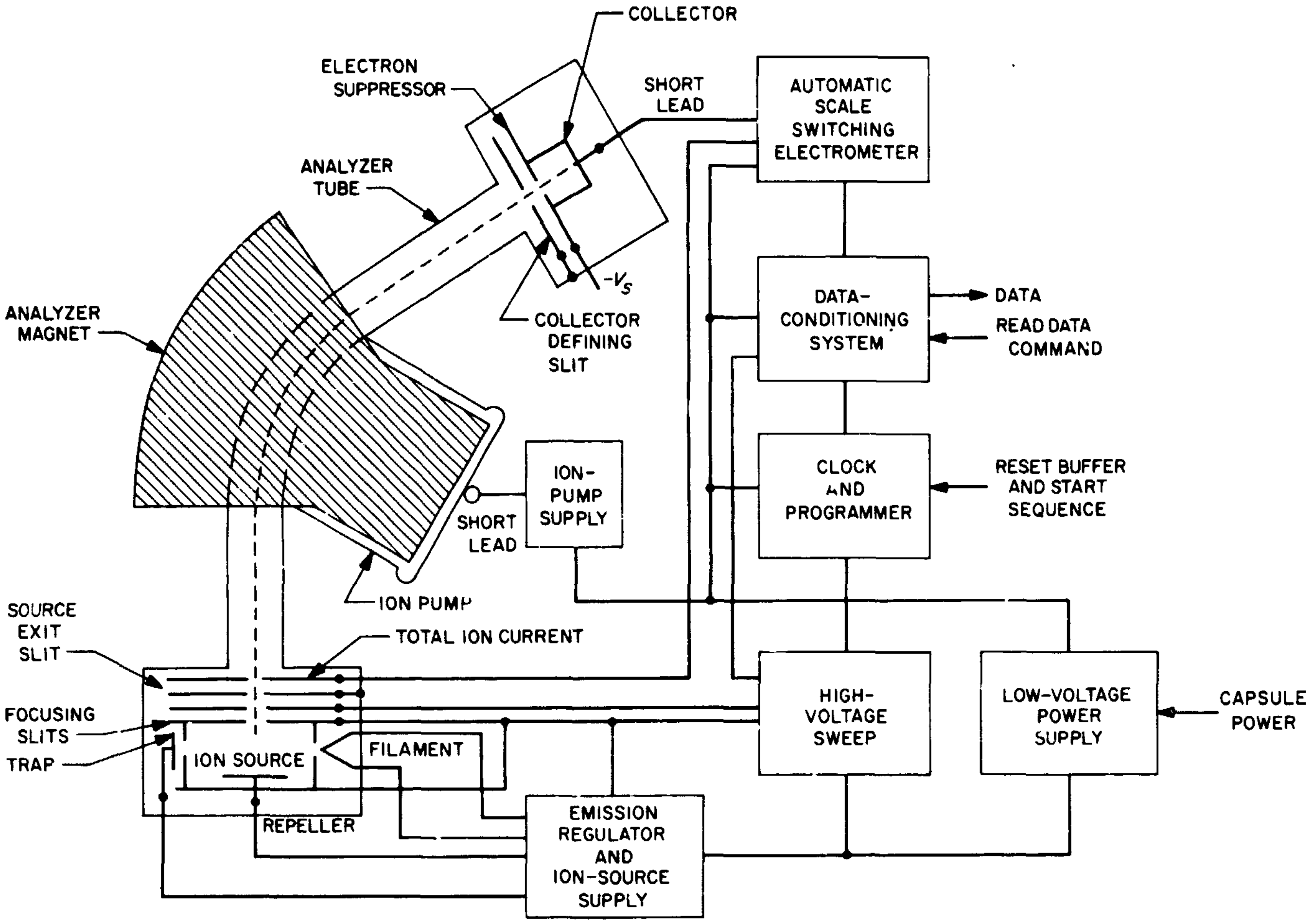
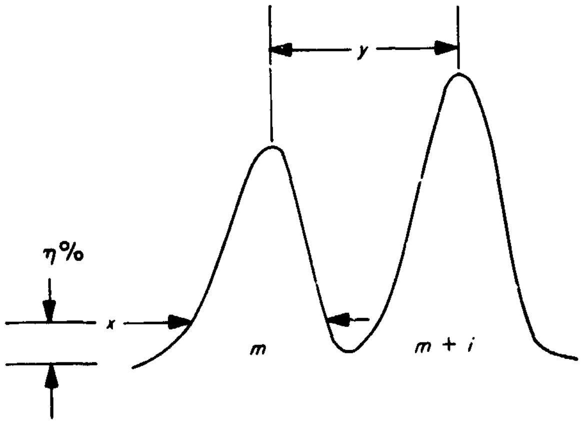


Fig. 1. Single-focusing mass spectrometer

Hereafter, the amu will be referred to as mass (m). The resolution of the instrument is defined at a particular m as



$$\left.\frac{m}{\Delta m}\right)_{\eta\%} = \frac{m}{(m+i)-m} = \mathbf{m}\left(\frac{y}{x}\right)_{\times 100\%}$$

where

$$\mathbf{m} = \frac{m + (m + i)}{2}$$

and x and y are time measurements. The resolution of the instrument described in this report is

$$\left.\frac{m}{\Delta m}\right)_{1\%} = 25 \qquad (1)$$

That is, at mass 25, the instrument can distinguish peaks differing by one unit.

3. Parameters for Determining the Acceleration Voltage Curve

a. Ion ballistics. The ion ballistics of the instrument in Fig. 1 are expressed

$$R = \frac{144}{B}\left[\left(\frac{M}{q}\right)v\right]^{1/2} \qquad (2)$$

where

$$R = 3.81 \text{ cm}$$

$$B = 3{,}780 \text{ G}$$

$$\frac{M}{q} = m \text{ is in amu}$$

and v is in volts.

Thus

$$m(t)v(t) = 10{,}000 \qquad (3)$$

At time t, the velocity (which is proportional to v) and the mass, M/q, of the ions determine its radius of deflection, which must be 3.81 cm, to be focused on the collector-defining stit. An accelerating voltage that decays exponentially can be approximated by the discharge of a capacitor through a resistor. The [illegible] width of the ion peaks over the entire mass range are nearly the same for the exponential accelerating voltage,

$$v(t) = v(0) \exp\left(-\frac{t}{\gamma}\right) \qquad (4)$$

Unfortunately, ion peaks will not appear linearly separated in time as indicated by

$$m(t) = \frac{10{,}000}{v(0)} \exp\left(\frac{t}{\gamma}\right) \qquad (5)$$

A linear separation of ion peaks, with respect to time, is desirable when interpreting a spectogram. The form required for $m(t)$ is

$$m(t) = at + m(0) \qquad (6)$$

Thus

$$v(t) = \frac{10{,}000}{at + m(0)} \qquad (7)$$

The hyperbolic (i.e., inverse) acceleration voltage expressed in Eq. (7) cannot be as readily generated by analog methods as the exponential.

Unlike the exponential case, the base width of the ion peaks varies directly with the amu interval.

b. Mass range. The mass range is 10 to 45 for the instrument in question. Thus, $v(t)$ must vary from 1000 to 222.22 V (a lower limit of 220 volts is actually used). This places the ion peak associated with mass 45 within the spectrum.

4. Hyperbolic Curve Generation Using Digital Techniques

a. The derivation of successive decremented dc voltage levels of fixed duration. The calculus of finite differences (Ref. 3) yields the discrete relationships

$$\left.\begin{aligned} mt(k) &= at(k) + mt(0) = ak + 10 \\ &= m(k) \\ vt(k) &= \frac{1000}{a't(k) + 1} = \frac{1000}{a'k + 1} = v(k) \\ t(k) &= k \text{ for } k = 0, 1, \cdots, 2^r - 1 \end{aligned}\right\} \qquad (8)$$

and r is an integer. From Eq. (8), where $v(2^r - 1) = 220$ V.

$$a' = \frac{39}{11(2^r - 1)} = \frac{\hat{a}}{2^r - 1}$$

The quantization required for v in quanta is

$$R = \left|\frac{v(0)}{\Delta v\,(2^r - 2)}\right| = \frac{[2^r(\hat{a} + 1) - (2\hat{a} + 1)]\,[\hat{a} + 1]}{\hat{a}}$$

Where $\Delta v(k)$ is the forward difference,

$$\Delta v(2^r - 2) = v(2^r - 1) - v(2^r - 2)$$

Note that $\Delta v(2^r - 2)$ is smallest change v undergoes.

$$\frac{\text{voltage quantization}}{\text{time quantization}} = \frac{R}{2^r} = \frac{\left[\hat{a} + 1 - \left(\frac{2\hat{a} + 1}{2^r}\right)\right]\left[\hat{a} + 1\right]}{\hat{a}} \qquad (9)$$

$$\frac{R}{2^r} \cong \frac{(\hat{a} + 1)^2}{\hat{a}} = 5.8 \text{ for } r \geq 5$$

Thus, if time is quantized with r bits ($r \geq 5$), voltage must be quantized to $r + 3$ bits to recognize $\Delta v(2^r - 2)$. (See Fig. 2 for an illustration of this method.)

Time is quantized by means of a feedback shift register (FSR) operating synchronously with a constant clock frequency. The 9-stage FSR is cycled through 512 internal states. The assertion outputs of the 9 stages represent a 9-bit non-weighted code. A 2-level diode

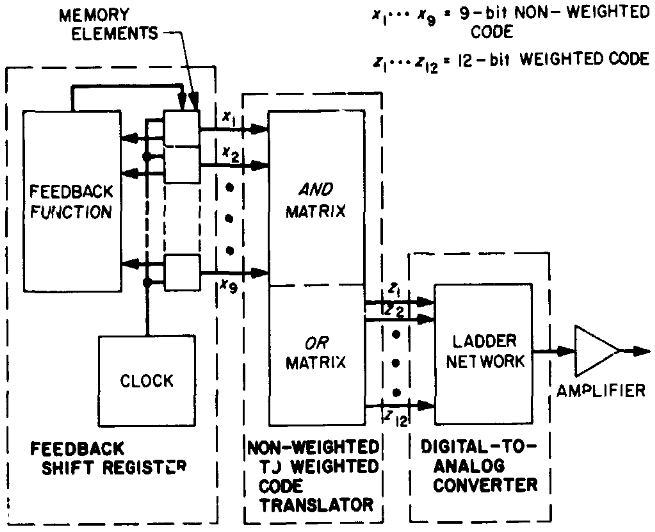

MEMORY ELEMENTS
$x_1 \cdots x_9$ = 9-bit NON-WEIGHTED CODE
$z_1 \cdots z_{12}$ = 12-bit WEIGHTED CODE
FEEDBACK FUNCTION
CLOCK
FEEDBACK SHIFT REGISTER
x_1
x_2
x_9
AND MATRIX
OR MATRIX
NON-WEIGHTED TO WEIGHTED CODE TRANSLATOR
z_1
z_2
z_{12}
LADDER NETWORK
DIGITAL-TO-ANALOG CONVERTER
AMPLIFIER

Fig. 2. Hyperbolic curve generator with time quantization

and–or matrix with 12 outputs translates the 9-bit nonweighted to a 12-bit weighted (positional) code. The 12-bit representation is converted to a dc voltage level proportional to the magnitude of a 12-bit binary number. This is the function of the digital-to-analog converter. The 1000- to 220-V hyperbolic sweep appears at the output of the high-voltage operational amplifier. Successive decremented levels of a fixed duration appear at the output of the digital-to-analog converter.

The number of diodes in the *and–or* matrix, which represents the 9-input/12-output truth table in disjunctive canonical form, is 4608 for *and*ing and 3054 for *or*ing, or a total of 7662 diodes. A silicon-on-sapphire microelectronic implementation of the diode *and–or* matrix is currently under test (see SPS 37-47, Vol. III, pp. 169–174).

A minimization program based on J. P. Roth's extraction algorithm (Refs. 4, 5, and 6), which is applicable to single-output Boolean functions, has been written for the IBM 7094 general-purpose computer. This program incorporates a transformation for handling multiple-output combination logic. An n-input/m-output problem is transformed into an imaginary $(m + n)$ input/single-output problem. The minimization of the single-output function yields the minimization of the simultaneous Boolean functions representing the original multiple-output problem in 2-level *and–or* form (see Ref. 7).

The minimization program was used to find an approximate minimum cover. A reduction of 738 diodes, or 9.6%, was realized in 4 h 12 m of computer running time. This program was the only one found that could handle the 12 Boolean functions of 9 variables. It has since been improved, particularly for the approximate minimum cover options. Further runs will be made with the improved program.

b. The derivation of successive and equally decremented dc voltage levels of varying duration. In this method, $v(k)$ is the independent variable. Thus

$$v(k) = \frac{1000}{\dfrac{\hat{a}t(k)}{2^r - 1} + 1}$$

$$\hat{t}(k) = \frac{t(k)}{2^r - 1} = \frac{1}{\hat{a}}\left(\frac{1000 - v(k)}{v(k)}\right) \qquad (10)$$

where

$$v(k) = 1000 + \frac{780k}{(2^r - 1)} \text{ for } k = 0, 1, \cdots, 2^r - 1,$$

and r is an integer. Therefore

$$\Delta v(k) = \frac{780}{2^r - 1}$$

$$\Delta t(k) = -\frac{1}{\hat{a}}\left(\frac{1000\,\Delta v(k)}{v(k)\,v(k + 1)}\right)$$

$$2\Delta\hat{t}(0) = \frac{22}{100}\,\frac{1}{(2^r - 1)}$$

Note that the quantum at $\Delta t(0)$ must be halved to ensure that two successive *one* levels are separated by a *zero* level. Thus

$$\Delta\hat{t}(0) = \frac{11}{100}\,\frac{1}{(2^r - 1)}$$

The required time quantization is

$$R_t = \frac{\hat{t}(2^r - 1)}{\Delta t(0)} = \frac{100(2^r - 1)}{11}$$

$$8 < \frac{R_t}{2^r} \leq 9.1 \text{ for } r \geq 4$$

Thus, if voltage is quantized with r bits (for $r \geq 4$), then time must be quantized with $r + 4$ bits to recognize $\Delta t(0)$. For 512 equal changes in voltage results, from deriving 512 unequally-spaced clock pulses to decrement a binary counter,

$$\frac{\hat{t}(512)}{\Delta t(0)} = 4638 \text{ quanta}$$

The 14-stage FSR in Fig. 3 cycles through 4639 states out of a possible 8192 states. The remaining states are treated as "don't cares." The 2-level diode *and–or* selection matrix converts 512 of the 4639 thirteen-bit representations of internal states to timing pulses. The timing pulses are properly spaced in time such that the binary counter, which they decrement, sequences through binary representations of a hyperbolic curve.

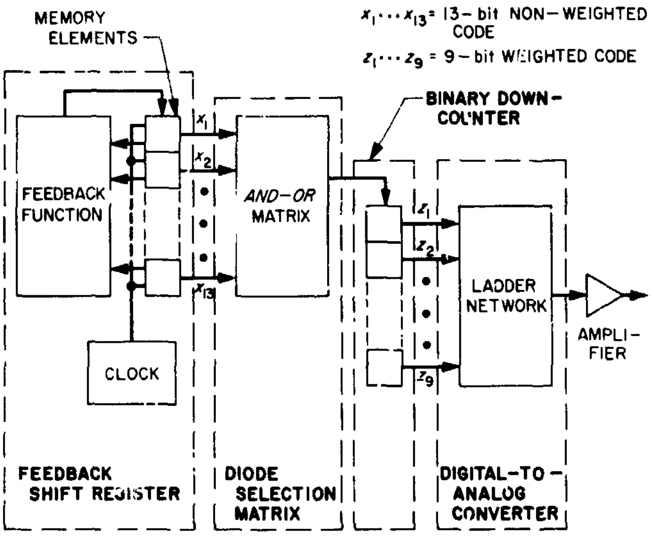


Fig. 3. Hyperbolic curve generator with amplitude quantization

The number of diodes in the *and–or* matrix, which represents a 13-input/single-output truth table in canonical form, is 6656 for *and*ing and 512 for *or*ing, or a total of 7168 diodes. This is 494 fewer diodes than needed in the (canonical) 9 × 12 matrix discussed previously.

The single Boolean function of 13 variables has not yet been subjected to minimization. A higher percentage of diode reduction than that for the 9 × 12 matrix is anticipated where Muller coding (i.e., multi-output to single-output transformation) introduces new prime impiicants in addition to expanding the number of inputs.

5. Examples of Hyperbolic Curve Generation with 2^5 Quanta

a. Successive decremented dc voltage levels of fixed duration. Since time is quantized with $r = 5$ bits, 8 bits are required to recognize $\Delta v(30)$. Thus

$$v(k) = \frac{255}{\frac{39}{11}\frac{k}{31} + 1} \quad \text{for } k = 0, 1, \cdots, 31$$

The largest 8-bit binary number, 255, is used to represent 1000 V. The feedback function for the 5-stage FSR is

$$a_k = a_{k-3} \oplus a_{k-5} \oplus a'_{k-1}\, a'_{k-2}\, a'_{k-3}\, a'_{k-4}\, a_{k-5}$$

where $\oplus$ denotes the *exclusive–or*, prime (′) denotes complementation and *and* is denoted by juxtaposition.

The Boolean variable a_{k-i} represents the state of the ith stage at clock-pulse interval (CPI) k. Successive inputs and outputs of a 5 × 8 matrix appears in Table 1. Note that a_{k-i} has been replaced by x_i. A plot of $Z = Z_1 Z_2 \cdots Z_8$ in decimal versus k appears in Fig. 4.

The 8 Boolean functions of 5 variables were minimized simultaneously under an approximate minimum cover

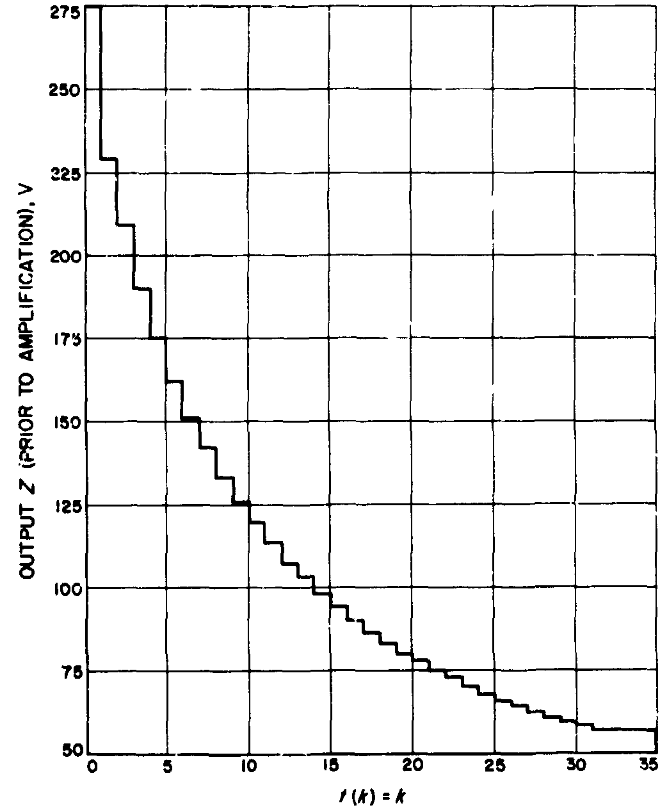


Fig. 4. Output Z (prior to amplification) in volts vs t (k) = k

Table 1. Nonweighted-to-weighted code translator

k	x_1	x_2	x_3	x_4	x_5	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Output Z (prior to amplification), V
0	1	0	0	0	0	1	1	1	1	1	1	1	1	255
1	0	1	0	0	0	1	1	1	0	0	1	0	1	229
2	0	0	1	0	0	1	1	0	1	0	0	0	0	208
3	1	0	0	1	0	1	0	1	1	1	1	1	0	190
4	0	1	0	0	1	1	0	1	0	1	1	1	1	175
5	1	0	1	0	0	1	0	1	0	0	0	1	0	162
6	1	1	0	1	0	1	0	0	1	0	1	1	1	151
7	0	1	1	0	1	1	0	0	0	1	1	1	0	142
8	0	0	1	1	0	1	0	0	0	0	1	0	1	133
9	1	0	0	1	1	0	1	1	1	1	1	1	0	126
10	1	1	0	0	1	0	1	1	1	0	1	1	1	119
11	1	1	1	0	0	0	1	1	1	0	0	0	1	113
12	1	1	1	1	0	0	1	1	0	1	0	1	1	107
13	1	1	1	1	1	0	1	1	0	0	1	1	1	103
14	0	1	1	1	1	0	1	1	0	0	0	1	0	98
15	0	0	1	1	1	0	1	0	1	1	1	1	0	94
16	0	0	0	1	1	0	1	0	1	1	0	1	0	90
17	1	0	0	0	1	0	1	0	1	0	1	1	1	87
18	1	1	0	0	0	0	1	0	1	0	0	1	1	83
19	0	1	1	0	0	0	1	0	1	0	0	0	0	80
20	1	0	1	1	0	0	1	0	0	1	1	1	0	78
21	1	1	0	1	1	0	1	0	0	1	0	1	1	75
22	1	1	1	0	1	0	1	0	0	1	0	0	1	73
23	0	1	1	1	0	0	1	0	0	0	1	1	0	70
24	1	0	1	1	1	0	1	0	0	0	1	0	0	68
25	0	1	0	1	1	0	1	0	0	0	0	1	0	66
26	1	0	1	0	1	0	1	0	0	0	0	0	0	64
27	0	1	0	1	0	0	0	1	1	1	1	1	0	62
28	0	0	1	0	1	0	0	1	1	1	1	0	1	61
29	0	0	0	1	0	0	0	1	1	1	0	1	1	59
30	0	0	0	0	1	0	0	1	1	1	0	1	0	58
31	0	0	0	0	0	0	0	1	1	1	0	0	0	56

option In Table 1, 10000 is the initial state and the singular state 00000 is the terminal state, which remains until the first stage is set (i.e., x_1 is made a *one*). This initial state yielded the best minimum cover of all the possible 32 initial states. The effect of using a different initial state is to cyclically permute the input states relative to the fixed output states. A total of 293 diodes is associated with each of 32 canonical truth tables. A reduction of 119 diodes or 40.6%, was realized with 1000 as an initial state. The initial state of 10101 yielded the smallest reduction (67 diodes, or 22.8%). Each of the minimization runs required less than 2 min of IBM 7094 computing time, including pre-processing, extraction, and post-processing time.

b. Successive and equally decremented dc voltage levels of varying duration. Since voltage is quantized with $r = 5$ bits, 9 bits are required to generate $\Delta\hat{t}(0)$. Thus

$$\hat{t}(k) = \frac{11}{39}\left(\frac{1000 - v(k)}{v(k)}\right)$$

where

$$v(k) = 1000 - \frac{780k}{31} \text{ for } k = 0, 1, \cdots, 31$$

$$2\Delta\hat{t}(0) = \frac{22}{3022}$$

$$\frac{\hat{t}(k)}{\Delta\hat{t}(0)} = 274.72$$

Thus, thirty-two 9-bit combinations are to be selected from a total of 276 successive states (or 275 time intervals) of an FSR.

$$\tilde{t}(k) = [274.72\, t(k) + 0.5]$$

represents time in quanta. The brackets denote the integer portion of $\tilde{t}(k)$.

The feedback function for the 9-stage FSR is

$$a_k = a_{k-5} \oplus a_{k-9} \oplus W$$

$$W = a_{k-1}\, a'_{k-2}\, a_{k-3}\, a'_{k-4}\, a_{k-5}\, a'_{k-6}\, a_{k-7}\, a_{k-8}\, a_{k-9}$$

The FSR will cycle through 276 of a possible 512 states. The remaining states are treated as "don't cares." The word detector W may also be used to inhibit the clock, thereby holding the FSR in state 276, corresponding to 220 V, or the end of the high-voltage sweep.

The thirty-two 9-bit combinations and the corresponding $t(k)$, for which a 9 × 1 matrix will furnish a time pulse to the binary down counter (Fig. 5), appear in Table 2. The output of the counter $Z = Z_1 Z_2 \cdots Z_5$ is represented decimally where 31 corresponds to 1000 V and 0 corresponds to 220 V.

The number of diodes in the *and–or* matrix, which represent the 9-input, single-output truth table of Table 2, is 288 for *and*ing and 32 for *or*ing, or a total of 320 diodes. Of the possible 276 states, 110101011 was used as an initial state in forming Table 2. A reduction of 144 diodes, or 45%, was obtained when minimized under an approximate minimum cover option. The total running time was 0.67 min.

Since this solution (176 diodes required) was comparable to the best solution found for the method in *Subsection 5-a* (174 diodes required), no other initial state was tested.

References

1. Duckworth, H. E., *Mass Spectroscopy*, Cambridge University Press, New York, 1958.
2. Leighton, R. B., *Principles of Modern Physics*, McGraw-Hill Book Company, Inc., New York, 1959.
3. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill Book Company., Inc., New York, 1962.
4. Roth, J. P., "Algebraic Topological Methods in Synthesis," *Proceedings of an International Symposium on the Theory of Switching*, April 1957, in *Annals of Computation Laboratory of Harvard University*, Vol. XXIX, pp. 57–73, 1959.
5. Roth, J. P., *Algebraic Topological Methods for the Synthesis of Switching Systems in n-variables*, ECP56-02, The Institute for Advanced Study, Princeton, New Jersey, April 1956.
6. Miller, R. E., *Switching Theory, Volume I: Combinational Circuits*, John Wiley & Sons, Inc., New York, 1965.
7. Muller, D. E., "Application of Boolean Algebra to Switching Circuit Design and to Error Detection," *IRE Trans.—Electronic Computers* Vol. EC-3, September 1954.

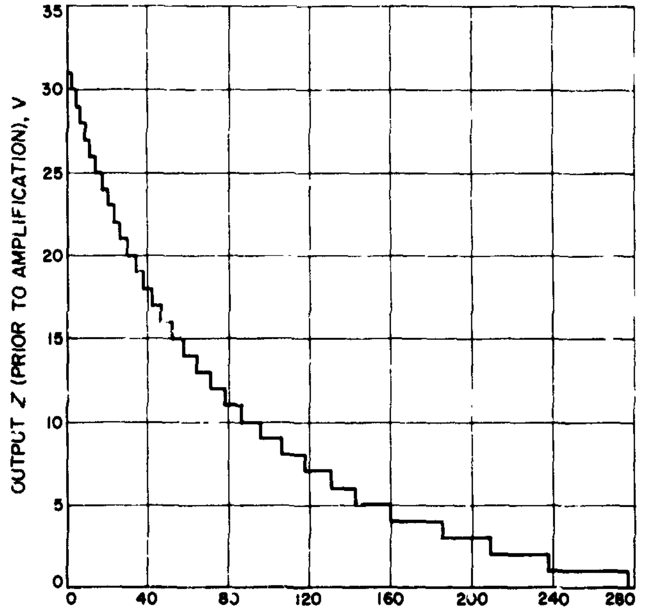


Fig. 5. Output Z (prior to amplification) in volts vs $\tilde{T}(k)$ in quanta

Table 2. An array for a 9 × 1 diode selection matrix

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	$\tilde{t}(k)$ in quanta	Output Z (prior to amplification), V
1	1	0	1	0	1	0	1	1	0	31
0	1	1	1	0	1	0	1	0	2	30
0	0	0	1	1	1	0	1	0	4	29
0	1	0	0	0	1	1	1	0	6	28
1	1	0	0	1	0	0	0	1	9	27
0	0	1	1	0	0	1	0	0	11	26
0	1	0	0	0	1	1	0	0	14	25
1	0	0	0	1	0	0	0	1	17	24
0	0	0	1	0	0	0	1	0	20	23
0	0	0	0	0	0	1	0	0	23	22
1	0	0	0	0	0	0	0	0	26	21
0	0	0	0	1	0	0	0	0	30	20
0	0	0	1	0	0	0	0	1	34	19
0	0	1	1	0	0	0	1	0	38	18
0	1	0	0	0	0	1	1	0	42	17
0	1	1	1	0	0	1	0	0	47	16
0	1	0	1	0	0	1	1	1	52	15
0	0	1	1	0	1	0	1	0	58	14
1	0	1	1	0	0	0	0	1	64	13
0	0	1	0	1	1	1	1	0	71	12
0	1	1	1	0	1	1	0	0	78	11
0	1	0	1	0	0	0	1	0	87	10
1	1	0	1	0	1	0	0	0	96	9
1	1	1	1	1	1	0	0	1	106	8
1	0	0	1	0	0	1	0	0	118	7
1	0	0	1	1	1	1	1	1	131	6
1	1	0	0	1	1	0	0	1	147	5
1	0	0	1	1	0	0	0	1	164	4
0	1	1	1	1	1	1	1	0	185	3
1	0	1	0	0	1	1	0	1	209	2
1	0	1	1	0	0	0	0	0	238	1
1	0	1	0	1	0	1	1	1	275	0

B. Capsule System Advanced Development Woven Plated-Wire Memory, *P. B. Whitehead*

1. Introduction

An 8,192-bit woven plated-wire memory has been developed for JPL by the Librascope Group of General Precision Systems, Inc., for use in the entry data system of the Capsule System Advanced Development (CSAD) program. The plated-wire stack was built in a flight configuration by Librascope, but the electronics for the memory were built in breadboard form by JPL using designs submitted by Librascope. Two stacks were manufactured: one for use with the breadboard electronics, and the other for testing the effects of sterilization and shock environments.

2. Background

Prior to the CSAD program, a program had been underway at JPL for the development of a low-power, non-destructive readout, plated-wire memory. A contract had been let to Librascope, a breadboard model produced, and a flight-qualified engineering model was to be delivered by the end of FY68. A description of this memory is given in SPS 37-45, Vol. IV, pp. 228–234. This memory was to have had a capacity of 20,480 bits, 1,024 words of 20 bits each. Data transfer into and out of the memory was serial, i.e., one bit at a time. Addressing, however, was random access by 20-bit word. An internal bit counter controlled which of the 20 bits was being selected.

When the requirements for the CSAD memory became known, it was decided to modify the existing contract and produce a memory compatible with this new project. The memory size was changed to 8,192 bits, with each bit random accessible, and the electronics were redesigned in order to more fully utilize integrated circuits. Power requirements were relaxed in order to make this utilization more feasible.

The functional characteristics of the memory are given in Table 3. The memory operates in the non-destructive readout (NDRO) mode, and can transfer data at 100,000 bits/s. The memory requires 200 mW of power on stand-by, 1 W when writing at 100,000 bits/s, and 800 mW when reading at 100,000 bits/s.

3. Woven Plated-Wire Stack

a. General description. The woven plated-wire stack is shown in Figs. 6 and 7. The stack consists of a single

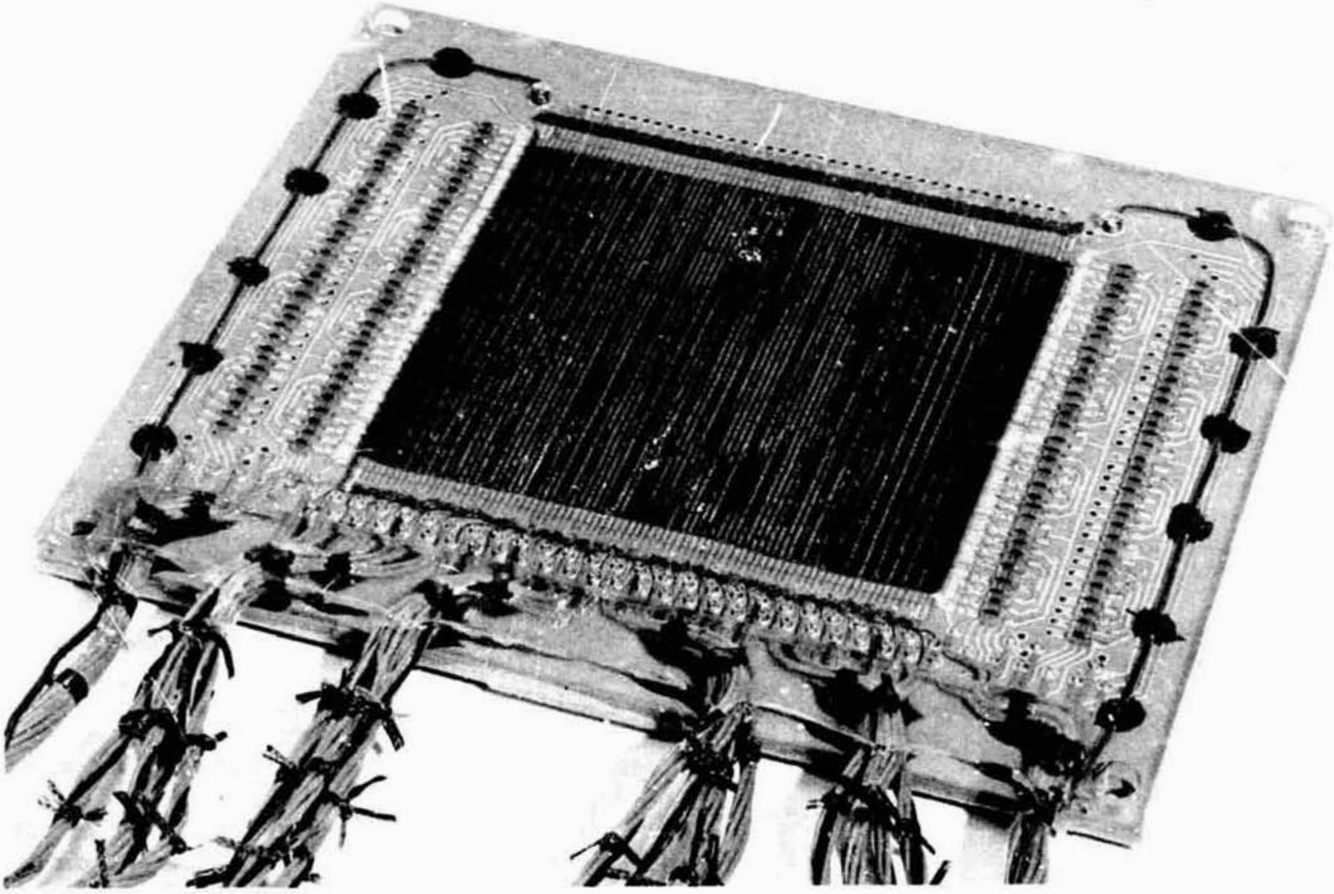


Fig. 6. Top of plated-wire stack

Table 3. Memory characteristics

Capacity	8,192 bits
Storage element	Plated wire
Addressing	Random access by bit
Data transfer mode	Bit serial, read and write
Data transfer rate	0 to 100,000 bits/s
Readout mode	Non-destructive
Volatility	Non-volatile
Input signals	Clock, read/write, address lines (1–13), data input
Output signals	Data output
Supply voltages	+15 V ± 10%, +5 V ± 7%, −3 V ± 7%
Power consumption	200 mW during standby 1000 mW during write at 100 K bits/s 800 mW during read at 100 K bits/s

8,192-bit plane. On either side of a 2-layer printed circuit board are two 4,096-bit mats. The printed circuit board measures 5⅛ × 6⅛ × ⅛ in., and the mats are 3¼ × 2⅞ in. The top of the plane is shown in Fig. 6. The 4,096-bit mat consists of 64 plated digit wires (shown running from front to rear in Fig. 6), and 64 word coils (running horizontally). The intersection of a digit wire and a word coil form a bit location. Woven along with the 64 plated wires are 16 unplated wires used as return lines—one return line for every 4 plated wires. The 4 rows of diodes on either side of the mat provide decoding for the word coils. The 2 rows closest to the mat provide decoding for the word coils in the mat on the top of the printed circuit board, and the 2 outer rows provide decoding for the word coils in the mat on the underside.

The stack is coated with a polyurethane resin, Solithane 113. This encapsulant provides the necessary adhesion to ensure that the mats are securely bonded to the printed circuit board, and yet provides sufficient flexibility to allow for contraction and expansion of the plated wires. Strain relief for the plated wires is provided by small coils of magnet wire located at the front of this mat. At the rear of the mat are plated-through holes that connect each plated wire to the corresponding plated wire in the mat on the underside. Figure 7 shows the

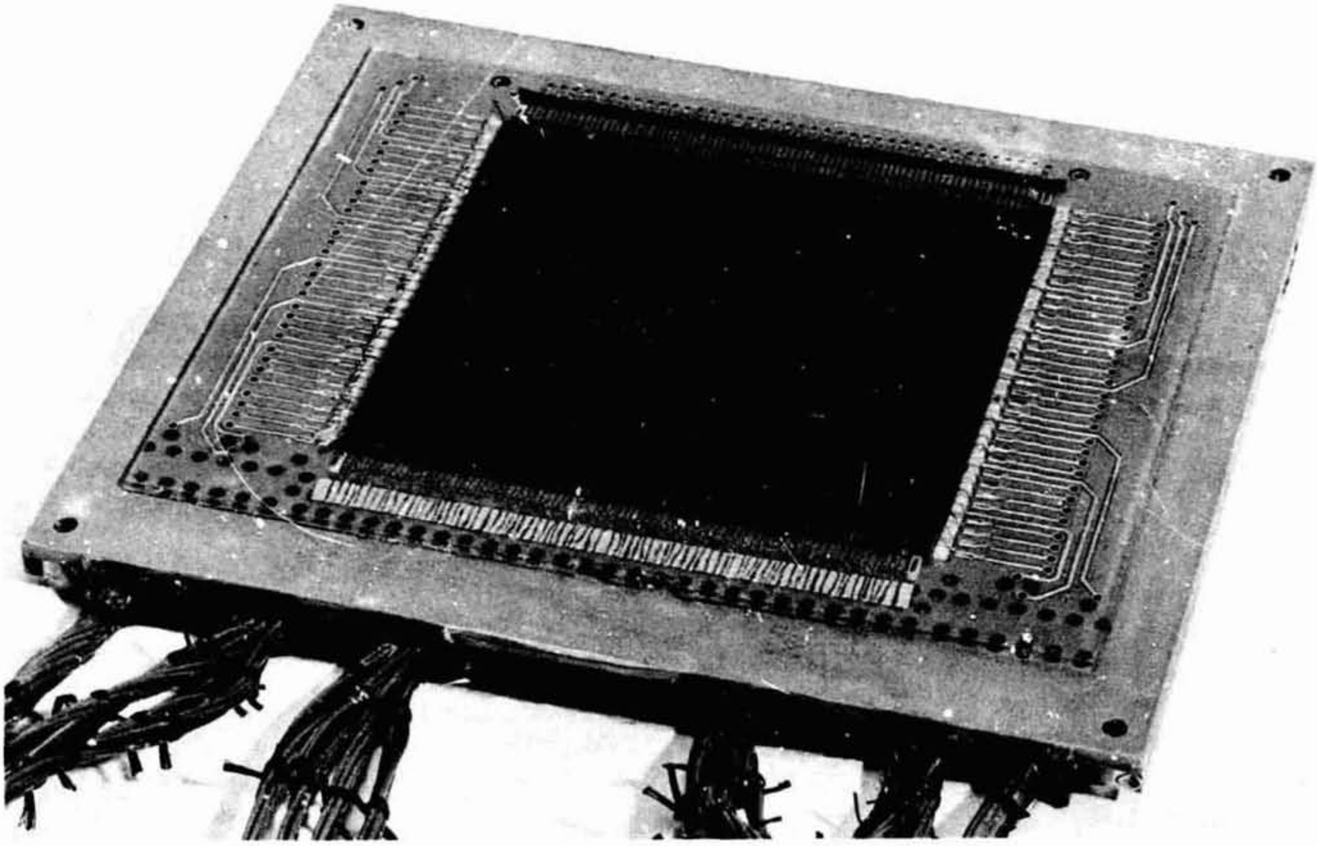


Fig. 7. Underside of plated-wire stack

underside of the plane. At the rear of the stack are the plated-through holes coming from the other side. At the front of the mat, the plated wires and unplated wires are bused together, and strain relief is provided. Along the edges of the printed circuit board, the Solithane 113 has been removed. In a flight configuration the stack would be bonded along these exposed surfaces to the web of a blivet. A recess would be provided in the center of the web to accept the mat.

b. Weave. The 4,096-bit mats are woven on a textile loom with 9-mil unplated wire as the warp and No. 41 AWG magnet wire as the woof. In the loom, the 9-mil unplated wires are held parallel (40 mils center-to-center). Alternate wires are raised, and the magnet wire is threaded through. One continuous strand of magnet wire forms a word coil. As shown in Fig. 8, a word coil consists of two loops of wire at each bit location. Word coils are separated by strands of magnet wire that form spacers. Alternate word coils are terminated on opposite sides of the mat.

Once the weaving process is completed, the mats are taken from the loom and hand-soldered to the printed-circuit board, at which time 64 of the 80 unplated wires are removed and replaced with 8 mil magnesium–copper wire plated with permalloy. The 64 plated wires become the digit lines, and the 16 remaining unplated wires become the return lines.

c. Operation. The magnetic action of the plated wire is described in SPS 37-45, Vol. IV, pp. 230–231 and will only be summarized here.

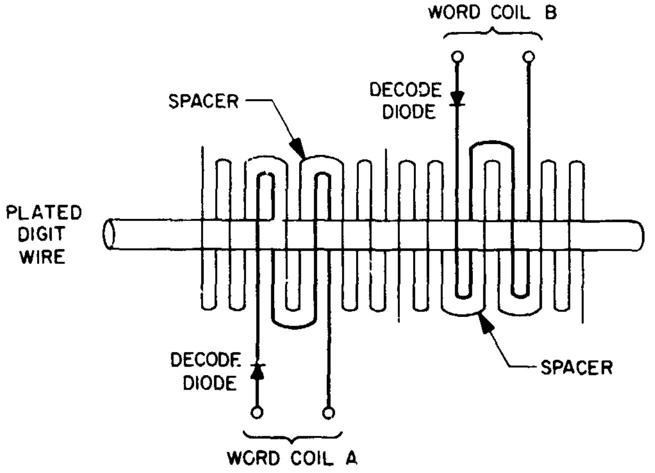


Fig. 8. Word coils

The plated wires are formed by electroplating a thin-film of permalloy onto 8 mil magnesium copper wire in the presence of a circumferential magnetic field. Thus, under quiescent conditions, the magnetization vectors along the wire lie in one of the two "easy" circumferential directions.

In order to write into the memory, a current is directed through one of the word coils. In that portion of the plated wire enclosed by the word coil, the magnetization vector is rotated until it is just short of being in the axial direction. A current through the plated wire then "tilts" the magnetization vector so that when the word current is removed, the vector will rotate back to the circumferential direction desired. Circumferential magnetization represents a *one* in one direction, and a *zero* in the oppodirection.

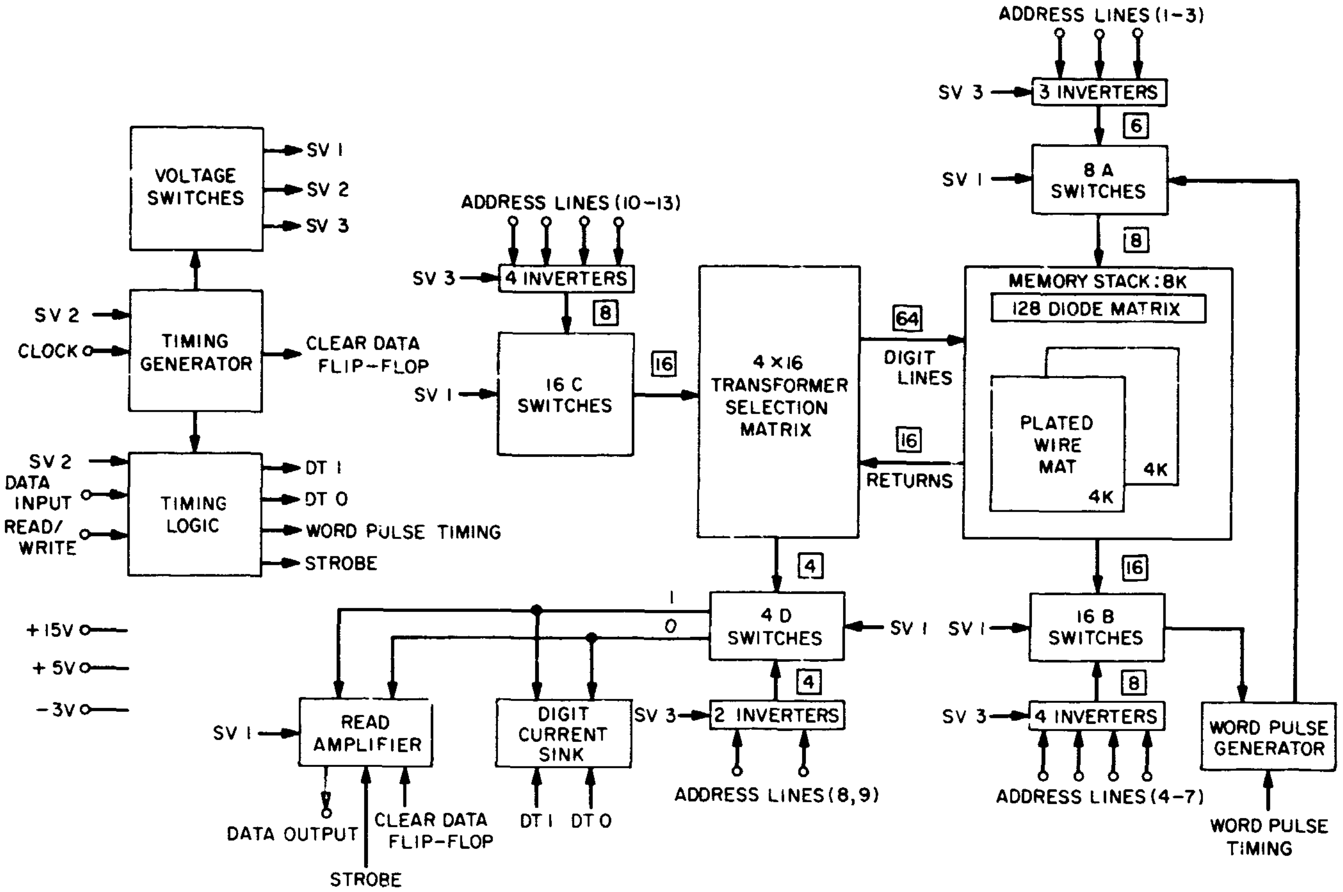
Data is read out of the memory by applying a word current only. This current causes the magnetization vector to again rotate just short of the axial direction. As the vector rotates, it causes a small voltage of a given polarity to appear at the ends of the digit line. Detection of the polarity of this signal determines whether a *one* or *zero* was stored in that bit location. When the current pulse is removed, the vector rotates back to its original position and the data is retained for future access.

4. System Design

a. Block diagram. The block diagram of the complete system, plated-wire stack plus associated electronics, is shown in Fig 9. The external signals are indicated by small circles in the figure and include the input and output signals, and the supply voltages, given in Table 3.

b. Initial operation. The operation of the memory begins with a 2 μs clock pulse sent to the memory from the data system. This clock pulse is received by the timing generator, initiating an 800 ns countdown. During the countdown, switched voltages (SVs) 1, 2, and 3 are turned on to provide power to various portions of the memory electronics. When the countdown is completed, the timing generator checks to see that the clock pulse is still being received. If so, the timing generator then initiates either a write cycle or a read cycle, depending on the condition of the read/write line into the memory. In either case the data output flip-flop in the read amplifier is reset to *zero.*

c. Read cycle. If the read/write line is high, then a read cycle is initiated after the 800 ns countdown. The



NUMBERS WITHIN BOXES REFER TO THE NUMBER OF SIGNAL LINES

Fig. 9. Memory system block diagram

word-pulse generator causes a 160 ns, 400 mA pulse to pass into one of the A switches, through a word coil in the stack, and out one of the B switches. Address lines 1 through 3 select which of the eight A switches is turned on, and address lines 4 through 7 select which of the sixteen B switches is turned on. The word pulse causes a "readback" voltage to appear on all 64 plated-wire digit lines. The transformer selection matrix determines which of these 64 signals reaches the read amplifier. This matrix consists of 64 transformers, one for each of the plated-wire digit lines. Address lines 10 through 13 select one of the C switches, and address lines 8 and 9 select one of the D switches. The combination of C and D switches select one of the transformers in the matrix. The transformer that is so selected allows the signal from its corresponding digit line to be passed through to the read amplifier. The read amplifier then takes this signal (typically 6 mV) and amplifies it. The timing generator and timing logic generate a strobe for the amplified readback signal. If the readback signal has a positive polarity at the time of the strobe, the data output flip-flop in the read amplifier is set to a *one*. If the readback signal has the opposite polarity, the data output flip-flop remains reset to *zero*. The output of this flip-flop becomes the data-output signal from the memory.

d. Write cycle. If the read/write signal is low, the memory initiates a write cycle, employing a bipolar write scheme, after the 800 ns countdown. For example, if a *one* is to be written into the memory, a *zero* is first written followed immediately by a *one*. This method ensures that an equal number of *ones* and *zeros* are written into every bit location in the memory, thus reducing the possibility of "creep" in the plated wire. (Creep is the enlargement of an area of magnetization caused by repeated writing of data of the same polarity into a given bit location.)

In order to write data into the memory, a 94-mA current pulse is drawn from one of the C switches, through the primary of one of the transformers in the selection matrix, through one of the D switches, and down to the digit current sink. A 94-mA current is then induced in the secondary of the selected transformer. The polarity of this current is determined by the data to be written into the memory. This current flows through the corresponding digit line, and its associated return wire, aı.d lasts for about 320 ns. During this time, the word pulse generator is activated for 160 ns. The coincidence of the word pulse and the digit current causes data to be written into the corresponding bit of the memory. Because of the bipolar write scheme used, however, the data that has been written is the complement of that desired. A digit current pulse of the opposite polarity is immediately initiated and the word pulse is repeated, writing the correct data into the memory.

*e. **Return to standby.*** After either the read or the write cycle is completed, the memory turns off SVs 1, 2, and 3, reducing power to standby mode. The timing generator then waits for the next clock pulse in order to re-initiate action.

5. Test Results

*a. **Stack tests.*** Before delivery from the contractor, the two stacks were tested using the following procedure for each bit in the plane:

(1) A *zero* was written into a bit location 1600 times, using a word current 5% higher than nominal and a digit current 10% higher than nominal.

(2) A *one* was written into the same bit location once, using a word current 5% less than nominal and a digit current 10% less than nominal.

(3) The digit currents used tr write the original *zero* were repeated 1600 times, then the word currents were repeated 1600 times. Since the digit current and word currents never coincided, the data should not have changed.

(4) The data was then read out of the bit location on an oscilloscope. If the polarity of the signal did not correspond to a *one*, or if the output signal was less than 2 mV, the bit was recorded as questionable.

(5) The procedure was repeated for data of the opposite polarity in the same bit location.

The first stack was tested only at 25°C. About 80 bits on the top and 60 bits on the underside of the plane were questionable—either low voltage or incorrect polarity. With the exception of two wires on the underside, the vast majority of the errors were in those wires near the edges of the mat. The nature and location of the errors indicated that the word currents flowing in the printed circuit board were interfering with the readout signal. The etched lines that carry the word currents are parallel to the digit lines, and only about ¼ in. from those lines on the edges.

The second stack was tested at −20, 25, and 90°C. The nominal currents used at the various temperatures are as follows:

Temperature, °C	Current, mA	
	Digit	Word
−20	116	460
25	94	400
90	83	329

The number of questionable bits remained much the same over the three temperatures—about 90 on the front and 30 on the back. As with the first stack, most of the

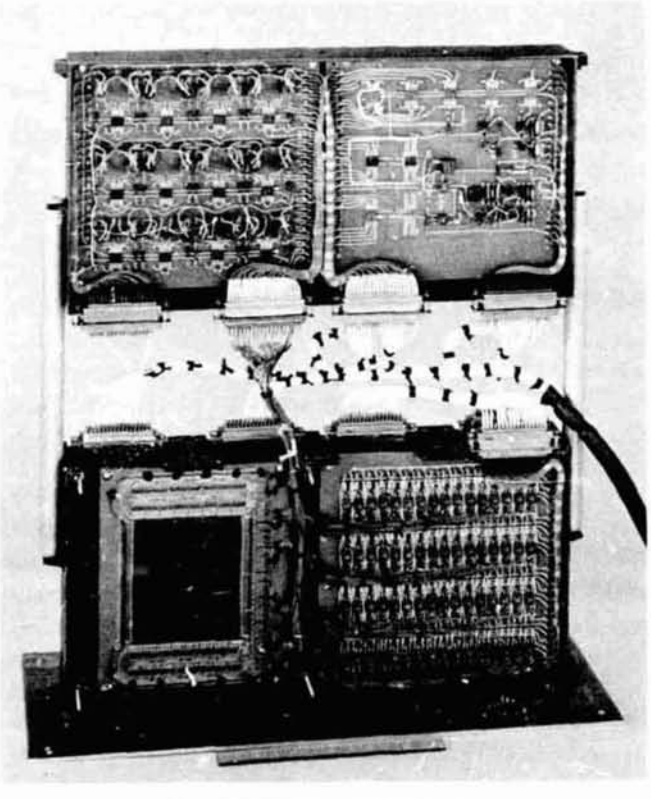


Fig. 10. System breadboard

errors occurred near the edges of the mats. The output voltage, however, did vary with temperature. At −20°C the average output was about 4.0 mV, at 25°C it was about 5.5 mV, and at 90°C it was up to about 7.0 mV.

The second stack will be subjected to an environmental test program that will include sterilization, shock, and vibration. After each test the electrical performance will be monitored to detect any degradation.

The errors in both stacks are well understood and could be eliminated by redesign of the printed-circuit boards. However, because of the limitations in funds and the pressures of the CSAD schedule, it was decided to accept the stacks without further modification. When the memory breadboard was operated as part of the Entry Data System of CSAD, those plated wires with bit errors were not used for storage. The remaining capacity of the memory was sufficient for CSAD requirements.

*b. **Breadboard tests.*** The plated-wire memory breadboard shown in Fig. 10 consists of the first stack (the lower left-hand corner) plus the associated electronics. The total parts count, including the diodes on the stack, is 717. The system was tested only at 25°C and, in general, operated properly. As expected, there were bit errors in the plated wires along the edges of the mats—about 30 on the front and 40 on the back.

The breadboard memory will be used for temperature-margin and extended-life tests.

XVIII. Lunar and Planetary Sciences

SPACE SCIENCES DIVISION

A. Scattering in the Twilight Atmosphere of Venus, *K. D. Abhyankar*

1. Introduction

Earlier computations of the scattering of light in the atmosphere of Venus made by Horak (Ref. 1) and Harris (Ref. 2) had shown that the observed visual brightness of Venus at phase angles greater than 120 deg exceeds the predicted theoretical brightness for isotropic and Rayleigh scattering phase functions (Fig. 1). The objective of this work was to test whether all or a part of this discrepancy could be caused by the neglected effect of sphericity of the Venus atmosphere as suggested by Harris.

2. Computational Factors and Results

By an appropriate geometrical consideration, which obviates the usual necessity of approximating each element of the spherical atmospheric shell by a plane parallel slab, it was possible to resolve the problem of scattering by the spherical twilight atmosphere into a series of separate problems that can be treated by the ordinary plane parallel technique. In this procedure, described in detail elsewhere, the disk of Venus is divided into four partly overlapping regions, each of which is illuminated in a different manner. The excess flux contributed by the twilight atmosphere is then easily computed by using the available tables of scattering functions for plane Rayleigh atmospheres of different optical thicknesses due to Coulson, Dave, and Sekera (Ref. 3) and Sekera and Kahle (Ref. 4). The Rayleigh scattering optical depths required for this purpose were derived from two models of the Venus atmosphere; one was the standard model of Kaplan (Ref. 5), and the other was a new extreme model quite similar to Kaplan's but more consistent with the recent data obtained from Venera 4 and *Mariner V* measurements. The latter model (Table 1) is cooler, denser, and more compact than Kaplan's model.

The total fluxes at various phase angles were computed for three wavelengths: V (5550 A), B (4550 A), and U (3700 A). The computed phase curves for the V wavelength (Fig. 1) show that Rayleigh scattering alone is not sufficient to account for the excess observed brightness

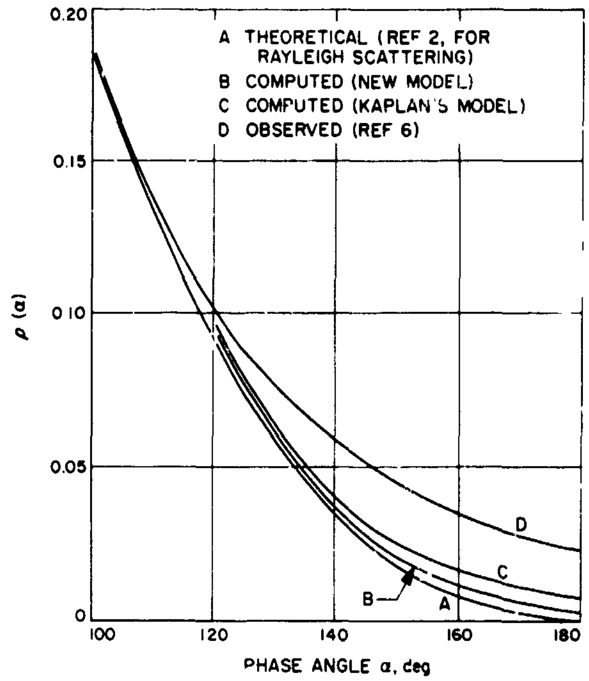


Fig. 1. Visual phase curves of Venus

Table 1. New extreme model of Venus atmosphere[a]

Height, km	Temperature, °K	Pressure, dyn/cm^2	Density, g/cm^3
0	550.0	2.20×10^7	2.00×10^{-2}
5	503.0	1.53×10^7	1.52×10^{-2}
10	456.0	1.01×10^7	1.11×10^{-2}
15	409.0	6.44×10^6	7.89×10^{-3}
20	362.0	3.87×10^6	5.36×10^{-3}
25	315.0	2.16×10^6	3.44×10^{-3}
30	268.0	1.08×10^6	2.03×10^{-3}
35	221.0	4.92×10^5	1.12×10^{-3}
40	219.0	1.85×10^5	4.25×10^{-4}
45	217.1	6.79×10^4	1.57×10^{-4}
50	215.2	2.42×10^4	5.64×10^{-5}
55	213.3	8.40×10^3	1.97×10^{-5}
60	211.4	2.91×10^3	6.90×10^{-6}
65	209.5	1.01×10^3	2.42×10^{-6}
70	207.6	3.50×10^2	8.45×10^{-7}
75	205.7	1.18×10^2	2.98×10^{-7}
80	203.8	3.89×10^1	9.57×10^{-8}
85	201.9	1.28×10^1	3.18×10^{-8}
90	200.0	4.21	1.05×10^{-8}
95	210.0	1.285	3.07×10^{-9}
100	220.0	0.421	9.59×10^{-10}
110	—	—	0

[a] Composition = 85% CO_2, 15% N_2 Mean molecular weight = 41.6.

at large phase angles. The main contribution to the observed brightness of Venus at inferior conjunction must be concluded to have come from particulate or condensate matter that scatters about one order of magnitude more efficiently in the forward direction than a Rayleigh scatterer.

To determine the possible nature of the scattering particles, the efficiency factors for V, B, and U wavelengths were obtained by combining the visual observations of Danjon (Ref. 6) and B–V, U–B colors measured by Knuckles, Sinton, and Sinton (Ref. 7), and comparing them with the computed brightness in the three colors at inferior conjunction. They were found to be 6, 10, and 17 for V, B, and U, respectively, in the case of the new model, and 2, 4, and 7, respectively, for Kaplan's model. The variation of the efficiency factor with color is caused partly by the variation of the phase function of the particles with wavelength and partly by the variation of their extinction coefficient with wavelength. From the scattering functions of water drops given by Deirmendjian (Ref. 8) for his haze model M, and from the extinction coefficients for dielectric particles tabulated by Penndorf (Ref. 9), it was found that the above efficiency factors were consistent with a haze model consisting of water drops of 0.1- to 1.0-μm radius, assuming a haze thickness of 20 km. For both models of the Venus atmosphere considered here, the total amount of water in a column of 1-cm^2 cross section above the lowest layers visible at inferior conjunction (above the height of 30–35 km) comes out to be close to 10^{-6} g/cm^2; i.e., about 0.01 μm of precipitable water above that level. This amount of water is too small to be detected by spectroscopic means. The total amount of water in the line of sight at inferior conjunction would be about 0.5 μm.

3. Atmospheric Contribution

The curves in Fig. 2 indicate the relative contributions I_t of the various layers of the atmosphere to the brightness of Venus at inferior conjunction. It is seen that the contribution to visible radiation (V, B, and U) comes mainly from the layers between 30 and 55 km in the new model and from 35 to 90 km in Kaplan's model. In both cases the densities in the effective layers range from 2×10^{-3} to 2×10^{-5} g/cm^3; the larger contribution in Kaplan's model is due mainly to the larger geometrical depth of the effective layers. It is also seen that in both models the V, B, and U radiations come from successively higher layers due to the increase of the scattering coefficient from V to U. The range in height between V and U is about 10 km in the new model and about 20 km in Kaplan's model. However, the geometrical thickness of the contributing layers is approximately the same for all the three colors, about 15 km in the new model and about 30 km for Kaplan's model. These values are in good agreement with the minimum thickness of 15 km derived by Schilling and Moore (Ref. 10) from the observed cusp extensions of Venus.

References

1. Horak, H. G., *Astrophys. J.*, Vol. 112, p. 445, 1950.
2. Harris, D., *Planets and Satellites*, Vol. III, Chap. 8 p. 311. Edited by G. P. Kuiper. University of Chicago Press, Chicago, Ill., 1961.
3. Coulson, K. L., Dave, J. V., and Sekera, Z., *Tables Related to Radiation Emerging From a Planetary Atmosphere With Rayleigh Scattering*. University of California Press, Berkeley and Los Angeles, Calif., 1960.
4. Sekera, Z., and Kahle, A. B., Rand Corporation Report R-452-PR, Santa Monica, Calif., 1966.
5. Kaplan, L. D., *A Preliminary Model of the Venus Atmosphere*. Technical Report 32-379. Jet Propulsion Laboratory, Pasadena, Calif., Dec. 12, 1962.
6. Danjon, A., *Bull. Astron.*, Vol. 14, p. 315, 1949.
7. Knuckles, C. F., Sinton, M. K., and Sinton, W. M., Lowell Observatory Bulletin 115, Vol. 5, No. 10, p. 153, 1961.
8. Deirmendjian, D., *Appl. Opt.*, Vol. 3, No. 2, p. 187, 1964.
9. Penndorf, R. B., *J. Opt. Soc. Am.*, Vol. 47, No. 11, p. 1010, 1957.
10. Schilling, G. F., and Moore, R. C., Rand Corporation Memorandum RM-5386-PR. Santa Monica, Calif., 1967.

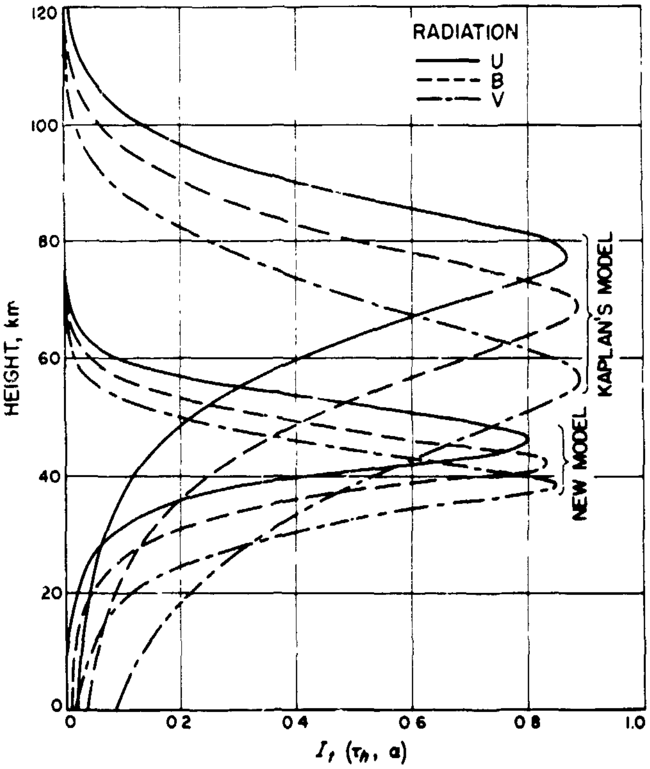


Fig. 2. Contribution of various atmospheric layers to brightness of Venus at inferior conjunction (α = 180 deg)

B. Water Vapor Variations on Venus, *R. A. Schorn, L. D. Gray, E. S. Barker*,[1] *and R. C. Moore*[2]

An extensive series of spectroscopic observations of Venus in the 8300-Å H_2O band was carried out during 1967. The purpose of this study was to try and reconcile the conflicting estimates of the water vapor abundance "above the clouds" of Venus by a homogeneous set of observations covering a large range of phase angles, a long period of time, and a variety of regions on the disk of the planet.

Early results in the near infrared (Refs. 1-4) gave values of w^* (the amount of precipitable H_2O in a vertical column through the atmosphere of Venus "above the clouds") ranging from 52 to 222 μm of precipitable H_2O. More recently Belton and Hunten (Ref. 5) and Spinrad and Shawl (Ref. 6) detected doppler-shifted Venus components to the 8189.272-Å H_2O line. Belton and Hunten observed a small region near the center of the disk and estimated the equivalent width of this weak feature as 20 mÅ, which corresponded to 317 μm of precipitable H_2O in the *total* path ($\eta w^* = 317$ μm, where η is the effective air mass). Spinrad and Shawl, using a spectrograph slit set parallel to the terminator, found the 8189-Å feature to have an equivalent width of about 15 mÅ at the center of the disk and less at the poles. They estimated $\eta w^* = 250$ μm and $w^* = 60$ μm at the center of the disk. A later discussion of the Kitt peak data (Ref. 7) gave an equivalent width of 15 mÅ, identical with the result of Spinrad and Shawl.

While Venus observations were being made at JPL, Owen (Ref. 8) observed the 8200-Å H_2O band on Venus and found no evidence of Cytherean H_2O. He set an upper limit of $w^* < 16$ μm and suggested that the faint 8189-Å "Venus" feature was a *solar* line. In addition, Connes, et al. (Ref. 9), set an upper limit of $w^* < 20$ μm from H_2O bands in the region $1 < \lambda < 2$ μm, while Kuiper (Ref. 10) set an upper limit of a few microns of H_2O from observations of the 1.4-μm H_2O band. The low H_2O limits at longer wavelengths do not necessarily contradict the larger H_2O abundances derived from the 8300-Å band (Ref. 11) but those of Owen clearly do. The results of observations at JPL in the 8200-Å region were negative from April 5 through June 23, 1967; however, observations in November and December 1967 gave positive results.

The methods of observation and reduction used in this study are the same as those used previously in a study

[1] University of Texas, Austin, Texas.

[2] Rand Corporation, Santa Monica, California.

of H_2O abundance and variability on Mars (Ref. 12). The spectra were taken with the 160-cm focal length camera of the 82-in. Struve Reflector Coudé spectrograph. All plates were ammonia-hypersensitized IV-N emulsions and utilized a projected slit width of 20 μm. The spectra used in this study are listed in Table 2.

About 30 uncontaminated H_2O lines (of varying *J*-value) were inspected in the 8200-Å band on each plate. Comparison with earlier work at JPL on Mars and examination of the visibility of weak solar lines of known equivalent width in the vicinity of strong terrestrial H_2O lines show that Cytherean lines with equivalent widths of >8 mÅ for the 4-Å/mm spectra and >4 mÅ for the 2-Å mm spectra should be detected.

None of the blue-shifted spectra showed any trace of Venus H_2O lines. Negative results on the 8176.975-Å H_2O line and the particular case of $\eta = 4$ will be used to compare these results with those of Spinrad and Shawl and Owen. According to Rank, et al. (Ref. 14), the intensity of the 8176.975-Å line is $S_0 = 0.077$ (cm-m-atm)$^{-1}$ (almost exactly the same as the intensity of 8189.272). This intensity leads to an *upper limit* of $w^* = 16$ μm for the 2-Å/mm plates and $w^* = 32$ μm for the 4-Å/mm spectra of this study. These upper limits are consistent with Owen's limited simultaneous observations and Kuiper's 1.4- and 1.9-μm results during the same period.

In contrast, all of the red-shifted spectra from November and December 1967 show positive evidence of Cytherean H_2O features, which appear weaker at the poles than at the equator. The Venus features appear only near the strongest lines of the 8200-Å band; i.e., 8164.54, 8169.995, 8189.272, 8197.704, 8226.962, and 8282.024 Å (all of which are low *J*-value lines). In fact, the visibility of the Venus lines is strictly proportional to the strength of the corresponding terrestrial lines (Ref. 15).

The 8189-, 8164-, and 8226-Å Venus H_2O lines have an estimated equivalent width of 8–10 mÅ on the plates used for this study (evidently the solar line near 8189 Å, suggested by Owen, did not affect these measurements). This compares with 15 mÅ for the 8189-Å feature estimated by Belton and Hunten in 1965–1966 and Spinrad and Shawl in 1964 and the *upper* limit of 4 mÅ set by Owen and JPL observers earlier in 1967. Evidently the water vapor "above the clouds" of Venus varies with time. If $\eta = 4$ is adopted for comparison purposes, it can be found that $w^* = 30$–40 μm of precipitable H_2O.

Table 2. Venus H_2O observations

Date, 1967	Grating[a]	*i*, deg[b]	$\Delta\lambda_{♀}$, Å[c]	Position of slit[d]
		Blue-shifted spectra		
Apr 5	III	52	−0.284	1
Apr 26	III	61	−0.324	4
Apr 27	III	61	−0.326	4
Apr 28	III	62	−0.328	4
Apr 30	III	63	−0.331	4
Apr 30	III	63	−0.331	4
May 1	III	63	−0.333	4
May 1	III	63	−0.333	1
May 1	III	63	−0.333	2
May 1	III	63	−0.333	3
May 1	III	63	−0.333	5
May 2	III	64	−0.334	1
May 23	III	74	−0.364	1
May 24	III	75	−0.365	1
May 24	III	75	−0.365	1
May 29	III	77	−0.372	1
Jun 18	I	89	−0.382	1
Jun 19	I	89	−0.382	1
Jun 19	I	89	−0.382	1
Jun 19	I	89	−0.382	1
Jun 22	III	91	−0.382	4
Jun 23	I	92	−0.382	1
		Red-shifted spectra		
Nov [illegible]	I	87	+0.352	2
N[illegible] 16	I	86	+0.352	1
Dec 11	I	73	+0.339	1
Dec 12	I	72	+0.338	1
Dec 12	I	72	+0.338	1
Dec 12	I	72	+0.338	4
Dec 17	I	70	+0.318	1
Dec 17	I	70	+0.318	1
Dec 18	I	69	+0.317	4
Dec 18	I	69	+0.317	2
Dec 18	I	69	+0.317	1
Dec 19	I	69	+0.316	2
Dec 19	I	69	+0.316	1
Dec 20	I	68	+0.315	1
Dec 20	I	68	+0.315	1

[a] Dispersion at 8200 Å. 4.1 Å/mm for grating I; 2.1 Å/mm for grating III.

[b] Planetocentric angle between sun and earth.

[c] Doppler shift according to Niehaus and Petrie (Ref. 13).

[d] Position of spectrograph slit on illuminated disk of Venus: 1 = pole to pole near terminator; 2 = parallel to 1, but near limb; 3 = parallel to equator near South Pole; 4 = parallel to equator through sub-earth point; 5 = parallel to equator near North Pole.

The modern observations of H_2O on Venus are compiled in Table 3, including the recent positive result of Kuiper. The evidence presented in Table 3 seems to argue strongly for a real variation of the observable Cytherean water vapor, although there is no recognizable pattern to the variations in the available data.

The confirmation of this variation, a study of the variation with time and phase angle if it is confirmed, and the confirmation of possible variations over the disk of the planet are obvious questions to be solved by further observations.

References

1. Dollfus, A., "Contribution au Colloque Caltech-JPL sur la Lune et les Planètes: Vénus," in *Proceedings of the Caltech-JPL Lunar and Planetary Conference, Sept. 13–18, 1965*, p. 187. California Institute of Technology and Jet Propulsion Laboratory, Pasadena, Calif., June 15, 1966.
2. Bottema, M., Plummer, W., and Strong, J., "Water Vapor in the Atmosphere of Venus," *Astrophys. J.*, Vol. 139, p. 1021, 1984.
3. Bottema, M., Plummer, W., and Strong, J., "A Quantitative Measurement of Water Vapor in the Atmosphere of Venus," *Ann. Astrophys.*, Vol. 28, p. 225, 1965.
4. Strong, J., "Balloon Telescope Studies of Venus," in *Proceedings of the Caltech-JPL Lunar and Planetary Conference, Sept. 13–18,1965*, p. 147. California Institute of Technology and Jet Propulsion Laboratory, Pasadena, Calif., June 15, 1966.
5. Belton, M., and Hunten, D., "Water Vapor in the Atmosphere of Venus," *Astrophys. J.*, Vol. 146, p. 307, 1966.
6. Spinrad, H., and Shawl, S., "A Search for Water Vapor on Venus–A Confirmation," *Astrophys. J.*, Vol. 146, p. 328, 1966.
7. Belton, M., Hunten, D., and Goody, R., *The Atmospheres of Venus and Mars*. Edited by J. Brandt and M. McElroy. Gordon and Breach, Science Publishers, Inc., New York (in press).
8. Owen, T., "Water Vapor on Venus–A Dissent and a Clarification," *Astrophys. J.*, Vol. 150, L121, 1967.
9. Connes, P., et al., "Traces of HCl and HF in the Atmosphere of Venus," *Astrophys. J.*, Vol. 147, p. 1230, 1967.
10. Kuiper, G., *Pub. Lunar Planet. Lab.* (in press).
11. Hunten, D., Belton, M., and Spinrad, H., *Astrophys. J.*, Vol. 150, L125, 1967.
12. Schorn, R., et al., "High-Dispersion Spectroscopic Observations of Mars II: The Water Vapor Variations," *Astrophys. J.*, Vol. 147, p. 743, 1967.
13. Niehaus, W., and Petrie, T., Tables of Stellar and Planetary Doppler Shifts from 1962 to 1982, Standard Oil Co. of Ohio, 1961.
14. Rank, P., et al., *Astrophys. J.*, Vol. 140, p. 366, 1964.
15. Moore, C., Minnaert, M., and Houtgast, J., *The Solar Spectrum 2935 A to 8770 A*, National Bureau of Standards Monograph 61, United States Government Printing Office, Washington, Dec. 1966.

Table 3. Estimates of H_2O abundance in a vertical column "above the clouds" of Venus

Date	i, deg[a]	Direction of shift	Wavelength of H_2O band(s), μm	w^*, μm[b]	Observers
Jun 22–23, 1959	90	Red	1.38	70	Dollfus (Ref. 1)
Feb 21, 1964	65	Blue	1.13	52–222	Strong (Ref. 4)
Apr 28, 1964	101	Blue	0.82	60	Spinrad and Shawl (Ref. 6)
Apr 29, 1964	102	Blue	0.82	60	Spinrad and Shawl (Ref. 6)
Nov 17, 1964	51	Red	0.82	60	Spinrad and Shawl (Ref. 6)
Nov 1965	~90	Blue	0.82	≤ 125	Belton and Hunten (Ref. 5)
May 1966	~70	Red	0.82	≤ 125	Belton and Hunten (Ref. 5)
Jun–Jul 1966	60–40	Red	$1 < \lambda < 2$	<20	Connes, et al. (Ref. 9)
Apr 1967	~55	Blue	0.82	<16	Owen (Ref. 8)
May 24, 1967	75	Blue	1.4, 1.9	~0	Kuiper (Ref. 10)
Jun 11, 1967	85	Blue	1.4, 1.9	~0	Kuiper (Ref. 10)
Apr–Jun 1967	52–92	Blue	0.82	<32, <16	This study
Nov–Dec 1967	87–68	Red	0.82	30–40	This study
Nov 1967	~80	Red	1.9, 2.7	~1	Kuiper (Ref. 10)

[a] Planetocentric angle between sun and earth.

[b] Amount of precipitable H_2O.

XIX. Physics

SPACE SCIENCES DIVISION

A. Auroral Arcs: Result of the Interaction of a Dynamic Magnetosphere With the Ionosphere,

G. Atkinson

1. Introduction

This article presents a theory to explain the occurrence of aurora in the form of arcs. The high-latitude auroral arcs are caused by electrons with energies of several thousand electron volts. These electrons travel down magnetic field lines from the outer magnetosphere until they collide with, and excite, particles in the atmosphere. The excited particles then emit light, thereby giving rise to what is called an aurora. The occurrence of aurora at high latitudes is believed to be the result of the structure and large scale properties of the magnetosphere. The most baffling feature has been their tendency to adopt the arc structure; i.e., thin parallel sheets of precipitating electrons, greater than 1000 km in east–west extent, a few hundred kilometers high, and yet less than 1 km thick in the north–south direction. The average separation between sheets is 30–40 km; the sheets lie along the magnetic field lines, which are nearly vertical at these high latitudes. The present theory explains this structure.

Two basic assumptions are made about the magnetosphere:

(1) There is a region in the outer magnetosphere capable of supplying electrons with the required energies.

(2) There are large scale electric fields in the magnetosphere, causing plasma flow.

Both of the assumptions are consistent with most of the current models of the magnetosphere and are supported by strong experimental evidence.

2. Structural Theory

Because of the high electrical conductivity parallel to magnetic field lines, the field lines approximate lines of equipotential. This may sometimes require that large electric currents (electron flows) occur parallel to the magnetic field lines. An auroral arc is such a current.

The auroral arc system is a regenerative or self-maintaining system. The current of precipitating electrons produces a region of intense ionization in the ionosphere as shown in Fig. 1a. Such a high conductivity region in

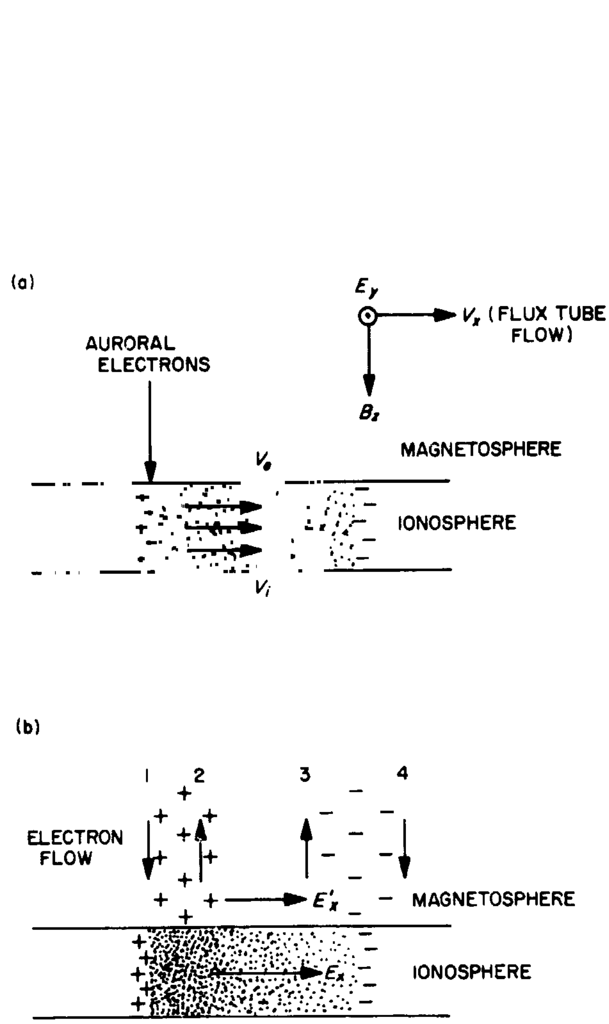


Fig. 1. Vertical section through the ionosphere and lower magnetosphere: (a) precipitating electrons producing polarization electric field E_x, (b) E_x mapping to the magnetosphere as E'_x

the ionosphere produces a polarization electric field E_x (Ref. 1).

If the magnetic field lines are to be lines of nearly constant voltage, then an electric field $E'_x \approx E_x$ must exist in the magnetosphere (Fig. 1b). This requires that there be regions of positive and negative space charges shown. All of the plasma in the magnetosphere is flowing in the x direction; the only way the region of space charge in the magnetosphere can remain stationary is for vertical electron flows (negative currents) to occur as shown. Arrow 1 is the auroral arc. Thus, the precipitating electrons cause the region of high electrical conductivity in the ionosphere, which in turn causes the precipitation of electrons.

The final downward flow of electrons (arrow 4) triggers the next arc, so that Fig. 1b is only one cell in a series of parallel arcs of great extent in the y direction.

It is possible, using a few simple assumptions, to predict the following: electron precipitation rates, ionosphere electron and ion densities, arc thicknesses, and distances between arcs. These predictions are in reasonable agreement with the observed values. In addition, the theo agrees with recent ionosphere measurements of electr fields and magnetic distortions.

3. Solution

A set of equations has been developed that describe the system, and a steady-state solution has been obtained for the special case $E_x = E'_x$; i.e., infinite conductivity along magnetic field lines. The solutions are shown in Fig. 2. The top curve shows electric field variation with distance; the second, height-integrated ionosphere current; the third, vertical current density; and the fourth, the height-integrated (in the ionosphere) number density of electrons or positive ions. Some of the quantities become infinite at the arcs because the conductivity has been assumed infinite.

The solution has three main uses:

(1) It shows the existence of an oscillatory solution.

(2) It predicts spacing between arcs.

(3) It allows a more detailed study of cause and effect.

One unexpected result was the requirement of a minimum average particle energy for auroral arcs to form (600 eV for the values used in this solution).

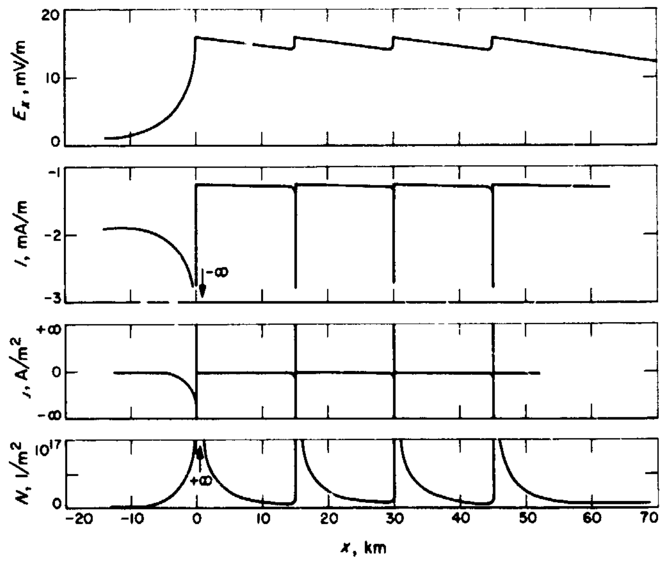


Fig. 2. A solution to the infinite parallel conductivity case

The plots are only quantitative in the region $0 < x < 50$ km. Outside of this, the curves are intended to be schematic.

Reference

1. Böstrom, R., "A Model of the Auroral Electrojets," *J. Geophys. Res.*, Vol. 69, pp. 4983–4999, 1964.

B. Rates and Mechanisms of the Gas Phase Ozonation of Ethylene and Acetylene,

W. B. DeMore

1. Introduction

Reactions of ozone with unsaturated hydrocarbons are key processes in air pollution, and also constitute an interesting class of molecule–molecule reactions that have not been studied in detail. This study describes gas phase rate measurements on the ozonation of C_2H_4 and C_2H_2. The results show that these two reactions, although formally similar, are fundamentally different with respect to detailed reaction mechanisms. Evidence has been found (1) that acetylene is inert to the ozonide-type reactions, which are characteristic of olefins; and (2) that acetylene reacts instead by a separate path, which has a higher collision efficiency and higher activation energy.

2. Experimental Methods

The reactions were carried out in a cylindrical metal cell coated on the inside with Kel-F grease. The cell temperature could be lowered to any desired point by flowing chilled N_2 gas through copper tubing wrapped around the cell. In this manner the temperature could be controlled to within ±0.2°C. To avoid temperature gradients due to self-heating, the gas mixtures were stirred vigorously with a small magnetically driven stirrer mounted in the cell. The O_3 concentrations were about 10^{-4} M, and the hydrocarbons were present in 2- to 25-fold excess. In most cases the mixtures were pressurized with argon to approximately 1 atm. For C_2H_4, the temperature range was −48 to −95°C, and for C_2H_2 the range was +10 to −30°C. The reaction rates were measured by following the decay of O_3 absorbance at 2537 Å, following rapid mixing of the reactants. In some of the experiments with C_2H_4, aerosol formation caused a transient baseline shift and this interfered with the spectrophotometric measurements. Fortunately, this effect could be minimized by effective stirring of the reaction mixture. Also, elimination of the argon pressurization reduced the aerosol interference. Little or no aerosol formation was observed with C_2H_2.

Most of the rates were measured under conditions where hydrocarbon excess was moderate and were plotted according to the equation

$$kt = \frac{1}{[S]^o - [O_3]^o} \ln \frac{[O_3]^o [S]^t}{[S]^o [O_3]^t} \tag{1}$$

where

$$[S] \equiv \text{concentration of } C_2H_4 \text{ or } C_2H_2$$

Since only the O_3 concentration was monitored, the hydrocarbon concentration at any time t was calculated on the assumption of a 1:1 reaction stoichiometry. The validity of this assump was borne out by the experimental results. In a few cases where the hydrocarbon excess was large, the following pseudo-first-order equation was used:

$$\ln [O_3]^t = \ln [O_3] - k't \tag{2}$$

where

$$k' \equiv k[S]$$

3. Results

a. Rate measurements. Figure 3 shows C_2H_4 data plotted according to Eq. (1), for those experiments in which the cell was pressurized to 1 atm with inert gas. In general, good straight lines were obtained, although in a few cases aerosol formation caused some error in determination of the initial O_3 concentration, which resulted in high intercepts. Figure 3 also shows the C_2H_4 data for experiments with no pressurization. The plots are excellent straight lines and show adherence to Eq. (1) for up to at least 90% completion of reaction. The rate data from Fig. 3 are summarized in Table 1 and are plotted in Arrhenius form in Fig. 4.

Data points from experiments with and without pressurization, and for various concentrations of O_3, all fall very nearly on a straight line (Fig. 4). The extrapolated Arrhenius line passes through the room temperature point of Hanst, et al. (Ref. 1). The rates of Bufalini and Altshuller (Ref. 2) are somewhat higher than those of this study. The following rate expression was derived from the slope and intercept of Fig. 4.

$$\log k_{C_2H_4} = 6.3 - 4.7/2.3\,RT \tag{3}$$

where k is expressed in $M^{-1}\,s^{-1}$.

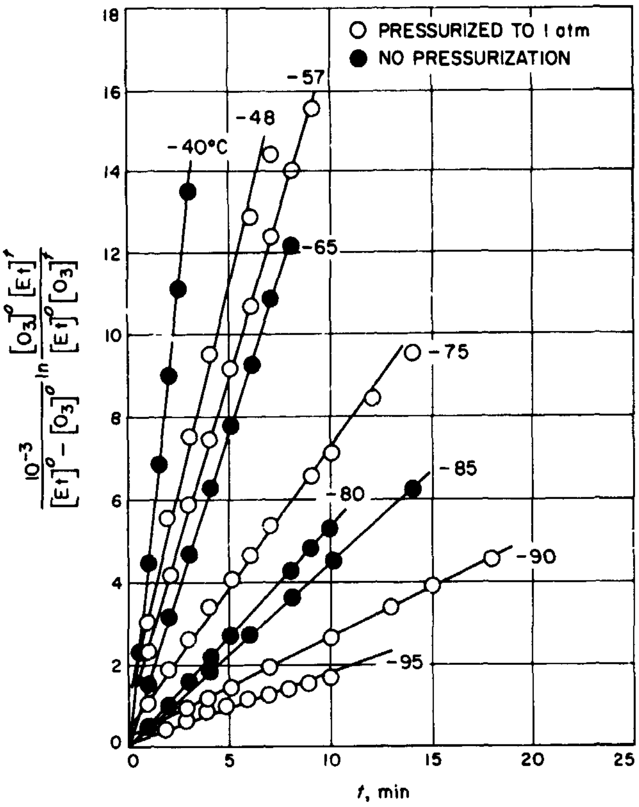


Fig. 3. Second-order plots for O_3–C_2H_4 reaction at various temperatures

Table 1. Summary of rate data for the O_3–C_2H_4 and O_3–C_2H_2 reactions

Initial concentrations,[a] $M \times 10^4$			Pressurizing gas[b]	Temperature, °C	k, $M^{-1}s^{-1}$
O_3	C_2H_4	O_2	O_3–C_2H_4 reaction		
0.716	1.963	14	None	−40	83.0
0.626	1.455	0	Argon	−48	44.0
0.892	1.980	5	Helium	−57	27.0
0.544	1.912	9	None	−65	26.0
1.057	3.420	0	Argon	−75	11.0
0.728	4.330	14	None	−80	9.0
0.564	2.909	9	None	−85	7.5
0.544	6.450	0	Argon	−85	5.3
0.434	4.493	5	Argon	−90	4.2
0.462	11.040	0	Argon	−95	3.0
O_3	C_2H_2	O_2	O_3–C_2H_2 reaction		
0.308	1.902	32	Argon	10	11.8
0.301	2.618	32	Argon	0	6.3
0.335	4.980	32	Argon	0	5.0
0.510	6.530	32	Argon	−15	1.4
0.355	2.141	32	Argon	−25	1.0
0.463	7.000	32	Argon	−25	0.8
0.435	26.800	32	Argon	−30	0.3

[a] Concentrations of O_2 are approximate.

[b] Pressure approximately 1 atm.

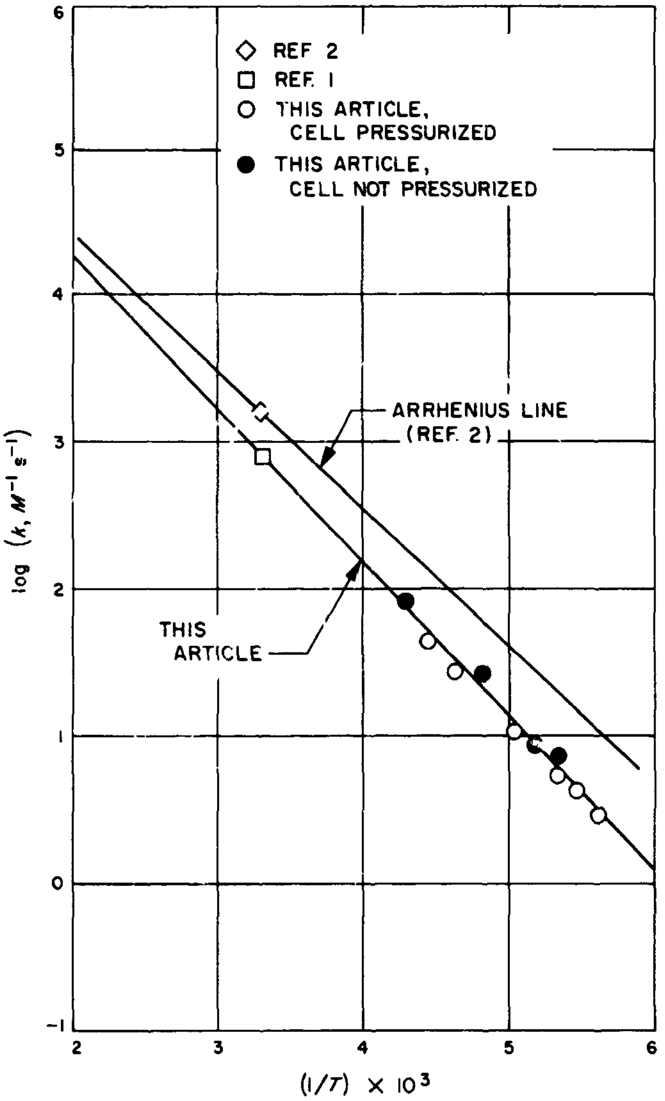


Fig. 4. Arrhenius plot of ethylene data

The rate data for C_2H_2 are shown in Fig. 5. In this case aerosol formation was not noted, and the rates gave good straight lines in every case. The rate data are also summarized in Table 1, and the rate constants are plotted in Arrhenius form in Fig. 6. The Arrhenius line from this study passes through the room temperature point of Cadle and Schadt (Ref. 3), but otherwise agrees very poorly with the rate parameters reported by them. From this study, the rate parameters are

$$\log k_{C_2H_2} = 9.5 - 10.8/2.3\,RT \tag{4}$$

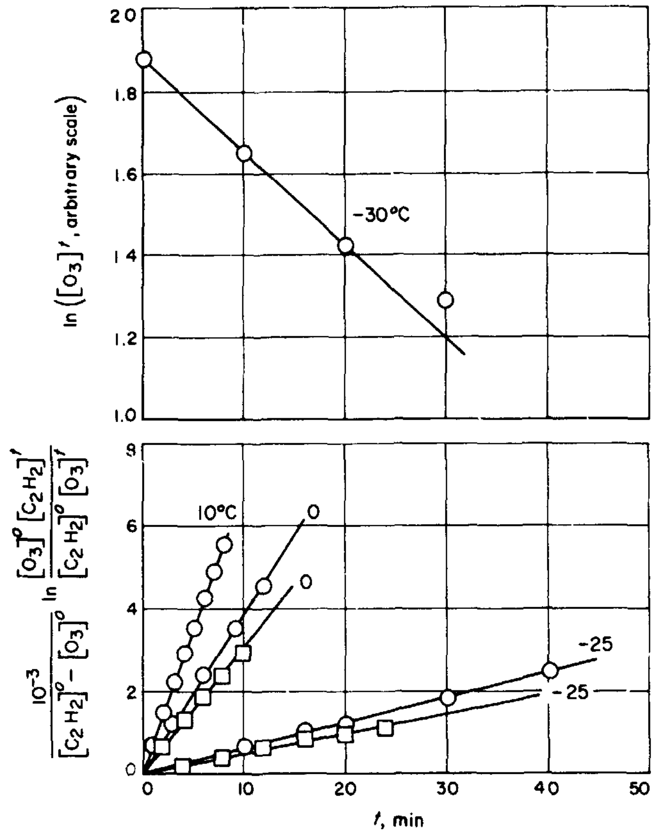


Fig. 5. Second-order and pseudo-first-order plots of acetylene data

b. Reaction stoichiometry. The straight line relationships obtained in the rate plots of Figs. 3 and 5 provide confirmatory evidence that the reaction stoichiometry was very nearly 1:1 because the latter assumption was used in the calculations. At high onversions the observed rates would have been fairly sensitive to any deviation from the assumed stoichiometry, particularly in cases where the hydrocarbon excess was not great. In addition, in several experiments the hydiocarbon loss was determined analytically after the reaction was complete. Within an experimental error of about 30%, the results agreed with the postulated 1:1 stoichiometry for both C_2H_4 and C_2H_2.

4. Discussion and Conclusions

The most surprising result of this work is the finding that the C_2H_2 ozonation reaction has a much higher activation energy and pre-exponential factor than the C_2H_4 reaction. As shown in the following paragraphs, this suggests very strongly that the two reactions are fundamentally dissimilar and do not both involve a 1,3 dipolar cycloaddition of O_3 to a π-bond of the hydrocarbons.

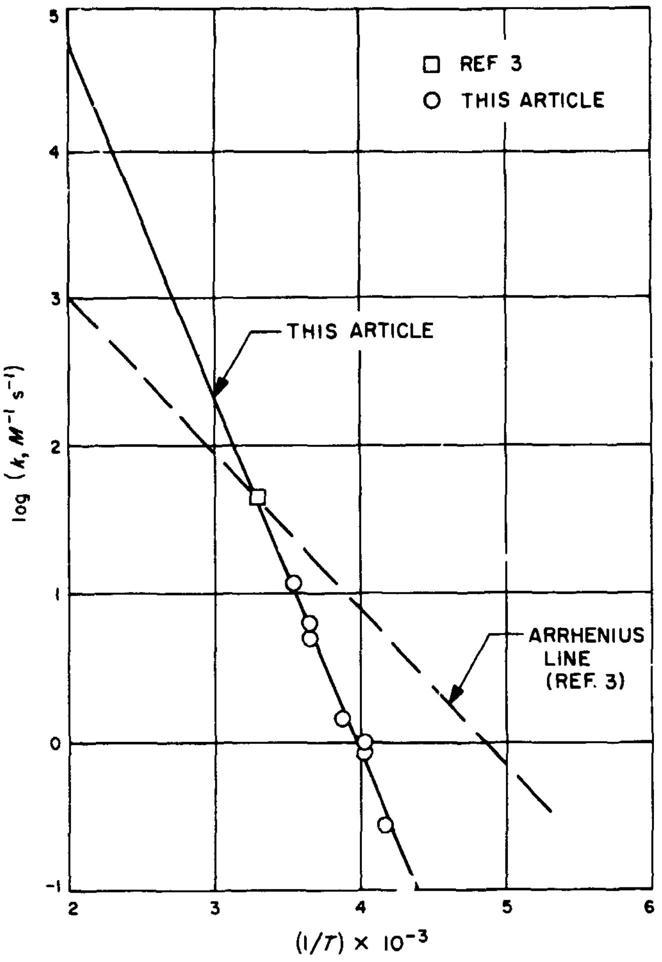


Fig. 6. Arrhenius plot for acetylene data

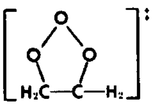
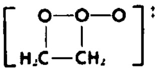
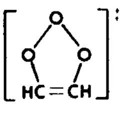
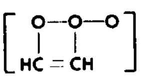
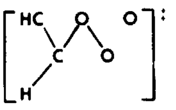
Rate measurements over a sufficiently wide range of temperatures provide an important clue to the nature of initial reaction structures because the pre-exponential factors derived from such measurements are related to activation entropies by the following equation from transition state theory (Ref. 4, p. 199):

$$A\,(M^{-1}s^{-1}) = \frac{e^2 kT}{h}(RT)\exp\left(\frac{\Delta S_p^\ddagger}{R}\right) \qquad (5)$$

for a reaction of molecularity 2. The activation entropy $\Delta S_p^\ddagger$ is in turn related to the structure of the transition state, so that in some cases a distinction can be made between possible structures which are widely different in entropy.

Table 2 shows some possible transition state structures for the reactions of O_3 with C_2H_4 and C_2H_2. The entropies of each were estimated by assuming that they are

Table 2. Possible transition state structures and corresponding estimated *A*-factors for ozonation of C_2H_4 and C_2H_2

Transition state equilibrium	Estimated entropy of transition state at 25°C, gibbs/mole	$A, M^{-1}s^{-1}$
$O_3 + C_2H_4 \rightleftarrows$ 	70[a]	$10^{6.4}$
$O_3 + C_2H_4 \rightleftarrows$ 	75[b]	$10^{8.5}$
$O_3 + C_2H_2 \rightleftarrows$ 	69.2[c]	$10^{7.2}$
$O_3 + C_2H_2 \rightleftarrows$ 	70.5[d]	$10^{7.5}$
$O_3 + C_2H_2 \rightleftarrows$ 	63.4[e]	$10^{10.3}$

[a] From S^0 of the hydrocarbon analog cyclopentane.

[b] From S^0 of the hydrocarbon analog methylcyclobutane, calculated by group additivity rules.

[c] From S^0 of the hydrocarbon analog cyclopentene.

[d] From S^0 of the hydrocarbon analog methylcyclobutene, calculated by group additivity rules.

[e] From S^0 of the hydrocarbon analog *n*-pentane.

equal to S^0 for the hydrocarbons of analogous structure, and the corresponding *A*-factors were calculated from Eq. (5).

Two results from Table 2 should be emphasized. First, the experimental *A*-factor of $10^{6.3}\,M^{-1}\,s^{-1}$ for C_2H_4 is in remarkably close agreement with the predicted value of $10^{6.4}\,M^{-1}\,s^{-1}$ for a cyclopentane-like transition state. This provides strong evidence that the initial adduct is indeed a five-membered ring, rather than a four-membered ring as has sometimes been suggested. The low collision efficiency of O_3–olefin reactions can be explained on a collision theory basis in terms of a strict steric requirement for ring formation.

Secondly, Table 2 shows that the transition state for the O_3–C_2H_2 reaction cannot have a five-membered ring structure because the predicted *A*-factor of $10^{7.2}\,M^{-1}\,s^{-1}$ is much lower than the observed $10^{9.5}\,M^{-1}\,s^{-1}$. Instead, a loose, open chain structure similar to *n*-pentane is required to explain an *A*-factor of the observed magnitude.

From a steric point of view, the O_3–C_2H_4 reaction has a collisional efficiency more than 1000 times higher than the O_3–C_2H_2 reaction. Nevertheless, the acetylene reaction is still much slower at ordinary temperatures because the activation energy is more than twice as high. The high activation energy of the acetylene reaction is consistent with the postulate of a different reaction mechanism for this reaction, and the magnitude of the activation energy suggests that both π-bonds are attacked.

The question remains as to why the low-energy reaction path of the O_3–C_2H_4 reaction is unavailable to the O_3–C_2H_2 reaction. The answer does not lie in ring strain because the strain energies of cyclopentane and cyclopentene are only slightly different. Neither can the answer be found in terms of a low-activation entropy for formation of the cyclopentene-like transition state because, as shown in Table 2, the *A*-factor for such a process should actually be higher than that of the O_3–C_2H_4 reaction.

Since the energy of a π-bond in C_2H_2 is almost identical to the π-bond energy in C_2H_4, there seems to be no way of escaping the fact that, from both an energy and an entropy point of view, the five-membered ring transition state should be as accessible in the O_3–C_2H_2 reaction as it is in the O_3–C_2H_4 reaction. Failure of the reaction to proceed in this manner can then only be attributed to insufficient energy release from the new bonds that are formed. The situation is illustrated schematically in Fig. 7. At point A in the O_3–C_2H_4 reaction, formation of the two new C–O bonds has provided more energy than the amount that was required to disrupt the original bonding in the reactants O_3 and C_2H_4, thus resulting in a change to a negative slope of the energy curve along the reaction coordinate. On the other hand, point A in the O_3–C_2H_2 reaction represents only a point of inflection, presumably involving rupture of both π-bonds in C_2H_2. It must be emphasized, of course, that Fig. 7 represents only relative energy requirements, and that the reaction does not have to pass through point A in order to reach point B.

The presently proposed mechanism for C_2H_2 ozonation is in disagreement with the commonly accepted assumption that ozonation of acetylenic compounds is analogous to the corresponding olefin reactions (Ref. 5). However,

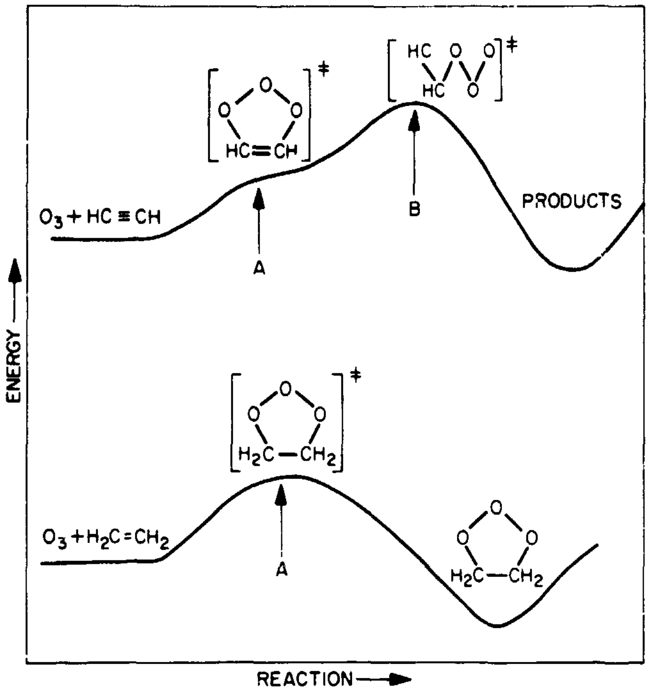


Fig. 7. Schematic representation of reactions of O_3 with acetylene and ethylene

relatively little work has been done on the acetylene reactions, and much of that was in the liquid phase. Also, the conclusions were based mainly on product analysis, which often is insensitive to detailed reaction mechanisms.

In an earlier report, the rate constant of the O_3-C_2H_4 reaction was measured in liquid argon at 87.5°K (SPS 37-49, Vol. IV, pp 273–278). The result was $\log k = -3.8$, and it was suggested on the basis of a semi-empirical treatment of the effect of solvent on reaction rates (Ref. 4, p. 409) that the gas phase rate at 87.5°K should be lower by a factor of $10^{1.6}$; i.e., $\log k\,(\text{gas phase}) = -5.4$. From Eq. (3), the extrapolated experimental gas phase value at 87.5°K would be $\log k = -5.5$, which is in very good agreement with the predicted value.

References

1. Hanst, P. L., et al., *Atmospheric Ozone-Olefin Reactions*. The Franklin Institute, Philadelphia, Pa., 1955.
2. Bufalini, J. J., and Altshuller, A. P *Can. J. Chem.*, Vol. 43, p. 2243, 1965.
3. Cadle, R. D., and Schadt, C., *J. Chem. Phys.*, Vol. 21, p. 163, 1953.
4. Glasstone, S., Laidler, K. J., and Eyring, H., *The Theory of Rate Processes*, pp. 199 and 409. McGraw-Hill Book Co., Inc., New York, 1941.
5. Bailey, P. S., *Chem. Rev.*, Vol. 58, p 956, 1958.

C. Prediction of OH Radical Microwave Lambda Doubling Transitions Below 120 GHz,

R. L. Poynter and R. A. Beaudet

1. Introduction

A number of anomalies has been observed in the 18-cm OH interstellar radio lines. These radio sources appear to vary widely in observed properties (Ref. 1). Of 50 or more radio sources that have been documented at this time, only two appear to be anywhere near "normal," as defined by the thermally expected absorption line intensities that would occur at the presumed temperatures in interstellar space. The remainder of the OH radio sources shows either or both emission and absorption features, frequently in all possible combinations. This observation indicates that the OH radio sources are generally not in a state of thermal equilibrium. Several mechanisms have been proposed (Refs. 2 and 3) to explain the observations. Each mechanism involves, in some way, an excitation process coupled with a cascade decay of the molecules into the ground rotational state. It has been proposed that if such a nonequilibrium distribution of OH molecules exists, there should be a finite population of OH in the higher rotational states, and that the lambda doubling transitions associated with these rotational states should be observable. Zuckerman, Palmer, and Penfield (Ref. 4) searched for the lambda doublets belonging to the lowest rotational state, $J = 1/2$, of the excited $^2\pi_{1/2}$ electronic state, which is 140 cm^{-1} higher in energy than the ground $^2\pi_{3/2}$ electronic state. Although these transitions had not been observed in the laboratory, their location in the frequency spectrum had been predicted from a set of molecular constants derived from the microwave spectroscopic studies of Dousmanis, Sanders, and Townes (Ref. 5). Unfortunately, these constants, based on relatively few observed lines in the spectrum, predicted a position that turns out to be 50 MHz removed from the correct value (Refs. 4 and 6). Zuckerman, et al., failed to observe these lines for this reason.

The present research does not resolve the anomalies that have been observed by the radio astronomers. It does define precisely the higher OH lambda doubling frequencies where further astronomical searches could be made for the purpose of studying the cascade decay processes.

2. Experimental Data

New measurements have been made of the OH microwave transitions in the range of 8.2 to 40 GHz. An accurate fit of these transitions has been achieved with a newly determined set of molecular constants. The analysis shows that there is a second complete set of detectable transitions belonging to the $^2\pi_{1/2}$ state. Two of these predicted transitions have been observed.

Because the low-frequency limit of the spectrometer at JPL is 8.2 GHz, the lambda doubling transitions below this frequency limit could not be measured directly. However, enough higher frequency transitions have been observed that, if the $J = 3/2, {}^2\pi_{3/2}$ transition as observed by Radford (Ref. 7) is included, a fairly complete analysis of the microwave spectrum can be obtained.

The calculated transition frequencies were obtained by exact diagonalization of the molecular Hamiltonian that was given by Dousmanis, Sanders, and Townes (Ref. 5); the molecular constants that were used in this study are essentially those defined by them. However, two centrifugal distortion constants are necessary to give a satisfactory fit of the experimental data. These are defined by the following two equations:

$$\langle\Sigma|B L_y|\Pi\rangle = \langle\Sigma|B_0 L_y|\Pi\rangle\,[1 - J(J+1)\,D/B_\Sigma]$$

$$\langle\Sigma|A L_y|\Pi\rangle = \langle\Sigma|A_0 L_y|\Pi\rangle\,[1 - J(J+1)\,\delta/B_\Sigma]$$

Here D represents the effect of centrifugal stretching on the internuclear distance and δ represents the effect of rotation on the electronic distribution. Of the eight molecular constants required for the lambda transitions, three were obtained from the optical OH studies of Dieke and Crosswhite (Ref. 8). The lambda doubling transitions were insensitive to these three parameters. The remainder of these constants were evaluated from the microwave spectra by the application of least squares methods. Four additional constants A, B, C, and D are required to describe the nuclear hyperfine splittings. Of the four constants, only one, D, is sensitive to the $\Delta F = 0$ transitions. The $\Delta F = \pm 1$ hyperfine transitions depend primarily on the other three constants.

3. Results and Discussion

A computer program has been written to perform the diagonalization and frequency calculations. This program has been modified to work with a least squares program for evaluating the molecular parameters. The program includes computation of Einstein A coefficients and intensities. The accuracy of the present analysis gives considerable confidence in predicting other low lying lambda doubling transition frequencies. The transitions that result from this analysis are given in Table 3, along with the Einstein A coefficients for the hyperfine components, and the intensities for an assumed temperature of 300°K, which represents normal laboratory conditions.

These frequencies differ by a significant amount from other values that have been reported. The differences result (1) from more accurate frequency measurements, (2) from least square fitting the new microwave constants that have been obtained using these frequencies, and (3) from the use of two centrifugal distortion constants.

The nuclear hyperfine constants obtained here are in excellent agreement with those that Radford (Ref 9) determined by electron spin resonance methods. In spite of this agreement, however, there remain some minor deviations between the calculated and observed hyperfine splittings. These deviations do not affect the general line predictions to any significant extent, because the absorption lines that are well-measured are fitted to high accuracy. The $\Delta F = 0$ transitions that deviate by ±1.0 MHz have not been measured in this work; some doubt exists about the accuracies of these frequencies. One suspects that the measurements of these lines may be off by as much as ±1.0 MHz, which would be consistent with the errors observed in the $^2\pi_{3/2}, J = 9/2$ transition frequencies. The minor deviations, ±0.4 MHz, that are observed in the $\Delta F = \pm 1$ components of the $^2\pi_{3/2}$, $J = 3/2$ and $J = 9/2$ transitions are caused by very small residual errors in the hyperfine coupling constants. This point (within the experimental error) has been verified at JPL by using the measured frequencies of the $^2\pi_{1/2}, J = 1/2$ and $^2\pi_{3/2}, J = 5/2$ transitions by Radford.[1] No changes are obtained in the lambda doubling molecular parameters.

The new lambda doubling constants are listed in Table 4. The nuclear hyperfine coupling constants are those given by Radford (Ref. 5). The predicted and observed line frequencies for *all observed* OH lambda doubling transitions in the microwave spectrum up to 40 GHz are given in Table 5.

Values of the Einstein spontaneous emission coefficients were calculated for a dipole moment of 1.66 ±0.01 D. The A coefficients for the $^2\pi_{3/2}, J = 3/2$ transitions agree with the values reported by Turner (Ref. 10), Carrington and Miller (Ref. 11), and Lide (Ref. 12).

Several additional comments may be made about Table 3. Laboratory measurements appear to be feasible

[1] H. E. Radford, private communication, Apr. 1968.

Table 3. Lambda doubling and hyperfine transitions[a]

J	F (F)	F (I)	Frequency, MHz	A (F, FP)	Intensity[b]	J	F (F)	F (I)	Frequency, MHz	A (F, FP)	Intensity[b]
${}^2\pi_{3/2}$ state						${}^2\pi_{1/2}$ state					
1.5	1.0	2.0	1611.844	1.29×10^{-11}	2.01×10^{-7}	0.5	0.0	1.0	4660.457	1.08×10^{-9}	9.23×10^{-6}
1.5	1.0	1.0	1665.403	7.11×10^{-11}	1.11×10^{-6}	0.5	1.0	1.0	4750.390	7.64×10^{-10}	6.52×10^{-6}
1.5	2.0	2.0	1667.349	7.71×10^{-11}	1.20×10^{-6}	0.5	1.0	0.0	4764.990	3.86×10^{-10}	3.29×10^{-6}
1.5	2.0	1.0	1720.908	9.42×10^{-12}	1.47×10^{-7}	1.5	1.0	2.0	7749.235	1.87×10^{-10}	1.19×10^{-6}
2.5	2.0	3.0	6016.520	1.09×10^{-10}	1.14×10^{-6}	1.5	1.0	1.0	7761.329	9.37×10^{-10}	5.97×10^{-6}
2.5	2.0	2.0	6030.731	1.53×10^{-9}	1.60×10^{-5}	1.5	2.0	2.0	7819.650	1.04×10^{-9}	6.59×10^{-6}
2.5	3.0	3.0	6035.059	1.57×10^{-9}	1.64×10^{-5}	1.5	2.0	1.0	7831.744	1.16×10^{-10}	7.36×10^{-7}
2.5	3.0	2.0	6049.270	7.90×10^{-11}	8.26×10^{-7}	2.5	2.0	3.0	8116.852	4.25×10^{-11}	1.67×10^{-7}
3.5	3.0	4.0	13441.927	3.40×10^{-10}	2.02×10^{-6}	2.5	2.0	2.0	8135.160	6.00×10^{-10}	2.35×10^{-6}
3.5	3.0	3.0	13434.605	9.17×10^{-9}	5.45×10^{-5}	2.5	3.0	3.0	8188.947	6.24×10^{-10}	2.45×10^{-6}
3.5	4.0	4.0	13441.374	9.26×10^{-9}	5.51×10^{-5}	2.5	3.0	2.0	8207.255	3.14×10^{-11}	1.23×10^{-7}
3.5	4.0	3.0	13434.051	2.64×10^{-10}	1.57×10^{-6}	3.5	3.0	4.0	5447.828	4.41×10^{-12}	8.84×10^{-9}
4.5	4.0	5.0	23838.799	7.09×10^{-10}	2.03×10^{-6}	3.5	3.0	3.0	5472.064	1.21×10^{-10}	2.42×10^{-7}
4.5	4.0	4.0	23817.616	3.11×10^{-8}	8.92×10^{-5}	3.5	4.0	4.0	5522.693	1.25×10^{-10}	2.51×10^{-7}
4.5	5.0	5.0	23826.634	3.13×10^{-8}	8.96×10^{-5}	3.5	4.0	3.0	5546.929	3.62×10^{-12}	7.26×10^{-9}
4.5	5.0	4.0	23805.451	5.78×10^{-10}	1.66×10^{-6}	4.5	5.0	4.0	194.888	9.06×10^{-17}	7.73×10^{-14}
5.5	5.0	6.0	37014.272	1.19×10^{-9}	1.39×10^{-6}	4.5	4.0	4.0	165.958	2.46×10^{-15}	2.10×10^{-12}
5.5	5.0	5.0	36983.501	7.71×10^{-8}	9.01×10^{-5}	4.5	5.0	5.0	117.905	8.86×10^{-16}	7.56×10^{-13}
5.5	6.0	6.0	36994.485	7.74×10^{-8}	9.04×10^{-5}	4.5	4.0	5.0	88.975	7.05×10^{-18}	6.07×10^{-15}
5.5	6.0	5.0	36963.714	1.00×10^{-9}	1.17×10^{-6}	5.5	6.0	5.0	8613.650	4.09×10^{-12}	1.24×10^{-9}
6.5	6.0	7.0	52759.426	1.75×10^{-9}	7.06×10^{-7}	5.5	5.0	5.0	8581.184	2.63×10^{-10}	8.00×10^{-8}
6.5	6.0	6.0	52721.719	1.57×10^{-7}	6.34×10^{-5}	5.5	6.0	6.0	8535.274	2.59×10^{-10}	7.89×10^{-8}
6.5	7.0	7.0	52734.387	1.57×10^{-7}	6.36×10^{-5}	5.5	5.0	6.0	8502.808	3.33×10^{-12}	1.01×10^{-9}
6.5	7.0	6.0	52696.680	1.51×10^{-9}	6.10×10^{-7}	6.5	7.0	6.0	19597.064	2.79×10^{-11}	2.54×10^{-9}
7.5	7.0	8.0	70886.167	2.36×10^{-9}	2.40×10^{-7}	6.5	6.0	6.0	19561.963	2.50×10^{-9}	2.28×10^{-7}
7.5	7.0	7.0	70843.272	2.81×10^{-7}	3.32×10^{-5}	6.5	7.0	7.0	19517.842	2.49×10^{-9}	2.26×10^{-7}
7.5	8.0	8.0	70857.368	2.81×10^{-7}	3.33×10^{-5}	6.5	6.0	7.0	19482.740	2.38×10^{-11}	2.17×10^{-9}
7.5	8.0	7.0	70814.457	2.08×10^{-9}	2.46×10^{-7}	7.5	8.0	7.0	32955.468	8.26×10^{-11}	1.90×10^{-9}
8.5	8.0	9.0	91229.499	3.01×10^{-9}	8.88×10^{-8}	7.5	7.0	7.0	32918.396	9.80×10^{-9}	2.25×10^{-7}
8.5	8.0	8.0	91182.618	4.57×10^{-7}	1.35×10^{-5}	7.5	8.0	8.0	32875.772	9.77×10^{-9}	2.25×10^{-7}
8.5	9.0	9.0	91197.927	4.58×10^{-7}	1.35×10^{-5}	7.5	7.0	8.0	32838.700	7.21×10^{-11}	1.66×10^{-9}
8.5	9.0	8.0	91151.046	2.69×10^{-9}	7.92×10^{-8}	8.5	9.0	8.0	48522.791	1.73×10^{-10}	8.52×10^{-10}
9.5	9.0	10.0	113640.690	3.68×10^{-9}	2.30×10^{-8}	8.5	8.0	8.0	48484.229	2.62×10^{-8}	1.29×10^{-7}
9.5	9.0	9.0	113590.676	6.95×10^{-7}	4.35×10^{-6}	8.5	9.0	9.0	48442.865	2.62×10^{-8}	1.29×10^{-7}
9.5	10.0	10.0	113607.018	6.96×10^{-7}	4.35×10^{-6}	8.5	8.0	9.0	48404.304	1.54×10^{-10}	7.57×10^{-10}
9.5	10.0	9.0	113557.003	3.32×10^{-9}	2.08×10^{-8}	9.5	10.0	9.0	66149.563	3.00×10^{-10}	2.69×10^{-10}
						9.5	9.0	9.0	66109.863	5.66×10^{-8}	5.07×10^{-8}
						9.5	10.0	10.0	66069.566	5.65×10^{-8}	5.06×10^{-8}
						9.5	9.0	10.0	66029.866	2.70×10^{-10}	2.42×10

[a] Einstein A coefficients, A (F, FP) $= A_{FF'}$, are given in s^{-1}. Values given for the ${}^2\pi_{3/2}$, $J = 3/2$ state are for comparison to transitions observed by radio astronomy.

[b] Intensities are for temperature of 300°K.

for OH lines with intensities larger than 10^{-7} cm^{-1}. However, lines this weak are at present marginally detectable and will require considerable care if they are to be observed. This results from the relatively low concentration (3% or less) of OH that can be generated by present methods, and from spectrometer sensitivity, which in this case has been measured as 10^{-9} cm^{-1}. The method of generation made use of the well-known H + NO_2 reaction.

In the ${}^2\pi_{1/2}$ state, the transitions are observed to rise to a maximum frequency, recede toward zero, and rise again. This effect is produced by an inversion in the lambda doubling energy levels that occurs between $J = 3.5$ and $J = 4.5$. As J approaches this inversion point, the lambda doubling energy splittings decrease. There are no restrictions on level symmetry, so that the transitions for J larger than 4.5 are allowed, although they are generally weaker than those of $J = 3.5$ and below. The net effect is to produce a sequence of transitions that have the appearance of Q, P, and R branches although they are not. The low-frequency lines exhibiting the analogous effect have been observed in the isotropic molecular species OD but apparently the effect was overlooked for neither comments nor explanation of this feature were reported (Ref. 5). The complete spectrum is plotted in Fig. 8, with appropriate identification of the two "branches" of the $^2\pi_{1/2}$ state, according to the direction in which absorption transitions would occur.

As can be seen from Table 3, the $J = 1/2$, $^2\pi_{1/2}$, $\Delta F = 0$, and $\Delta F = \pm 1$ hyperfine transition frequencies are about 33 and 42 MHz, respectively, from where Zuckerman, et al., attempted to search. Thus, it would seem that their conclusions about the upper limits of the intensities of these transitions must be invalid. Another search would be worthwhile for these and other low-frequency OH transitions in the interstellar medium.

Table 4. Molecular constants for OH, assuming $c = 2.997929 \times 10^{10}$ cm/s

Constant[a]	Value, MHz
Sigma state energy, $E_\Sigma - E_\pi$	979,798,100.0[b]
Rotational constant for sigma state, B_Σ	508,478.0[b]
Rotational constant for π state, B_π	555,066.0[b]
Spin orbit coupling constant, A_{so}	−4,163,508.0 ± 360
$\langle\Sigma\|BL_y\|\Pi\rangle$	377,382.2 ± 16
$\langle\Sigma\|(2B + A)L_y\|\Pi\rangle$	−1,531,211.0 ± 60
Centrifugal distortion constant, D	107.599 ± 0.27
δ	−44.539 ± 0.12
$\lambda = (A_{so}/B_\pi)$	−7.5009 ± 0.0001

[a] As derived from this work, λ agrees fairly well with both the optical (Ref. 8) and electron paramagnetic resonance (Ref. 9) results.

[b] Ref. 8.

References

1. Robinson, B. J., and McGee, R. X., *Annu. Rev. Astron. Astrophys.*, Vol. 5, pp. 183–212, 1967.
2. Cook, A. H., *Nature*, Vol. 210, p. 611, 1966.
3. Litvak, M. M., et al., *Phys. Rev. Lett.*, Vol. 17, p. 821, 1966
4. Zuckerman, B., Palmer, P., and Penfield, H., *Nature*, Vol. 213, p. 1217, 1967.
5. Dousmanis, G. C., Sanders, T. M., Jr., and Townes, C. H., *Phys. Rev.*, Vol. 100, p. 1735, 1955.
6. Barrett, A. H., *IEEE Trans. Mil. Electron.*, MIL-8, p. 156, 1964.
7. Radford, H. E., *Phys. Rev. Lett.*, Vol. 13, p. 534, 1964.
8. Dieke, G. H., and Crosswhite, H. M., *J. Quant. Spec. Rad. Transfer*, Vol. 2, p. 97, 1962.
9. Radford, H. E., *Phys Rev.*, Vol. 126, p. 1035, 1962.
10. Turner, B., *Nature*, Vol. 212, p. 184, 1966.
11. Carrington, A., and Miller, T. A., *Nature*, Vol 214, p. 998, 1967.
12. Lide, D. R., Jr., *Nature*, Vol. 213, p. 694, 1967.

Table 5. Comparison of observed and calculated frequencies in OH

Electronic state	J	$F_I \rightarrow F_F$	Frequency, MHz		Frequency difference (calculated − observed), MHz	Experimental error limits
			Calculated	Observed		
$^2\pi_{3/2}$	3/2	2 → 1	1611.844	1612.231[a]	−0.387	0.002[a]
		1 → 1	1665.403	1665.401[a]	+0.002	0.002[a]
		2 → 2	1667.349	1667.358[a]	−0.009	0.002[a]
		1 → 2	1720.908	1720.533[a]	+0.375	0.002[a]
$^2\pi_{1/2}$	3/2	1 → 1	7761.329	7760.36[b]	+0.97	1.0[b]
		2 → 2	7819.650	7819.92[b]	−0.27	1.0[b]
$^2\pi_{1/2}$	5/2	2 → 2	8135.160	8135.51[b]	−0.35	1.0[b]
		3 → 3	8188.947	8188.94[b]	+0.007	1.0[b]
$^2\pi_{3/2}$	7/2	3 → 3	13434.605	13434.62	−0.015	0.01
		4 → 4	13441.374	13441.36	+0.014	0.01
$^2\pi_{1/2}$	13/2	7 → 7	19517.868	19517.55	+0.32	0.3
		6 → 6	19561.932	19562.08	−0.15	0.3
$^2\pi_{3/2}$	9/2	4 → 5	23805.451	23805.13	+0.32	0.01
		4 → 4	23817.616	23817.64	−0.024	0.01
		5 → 5	23826.634	23826.62	+0.014	0.01
		5 → 4	23838.799	23838.46	+0.34	0.01
$^2\pi_{3/2}$	11/2	5 → 5	36983.501	36983.47	+0.031	0.03
		6 → 6	36994.485	36994.43	+0.055	0.05

[a] Observed by Radford (Ref. 7).

[b] Observed by Dousmanis, Sanders, and Townes (Ref. 5); error limits estimated to be much larger than they reported.

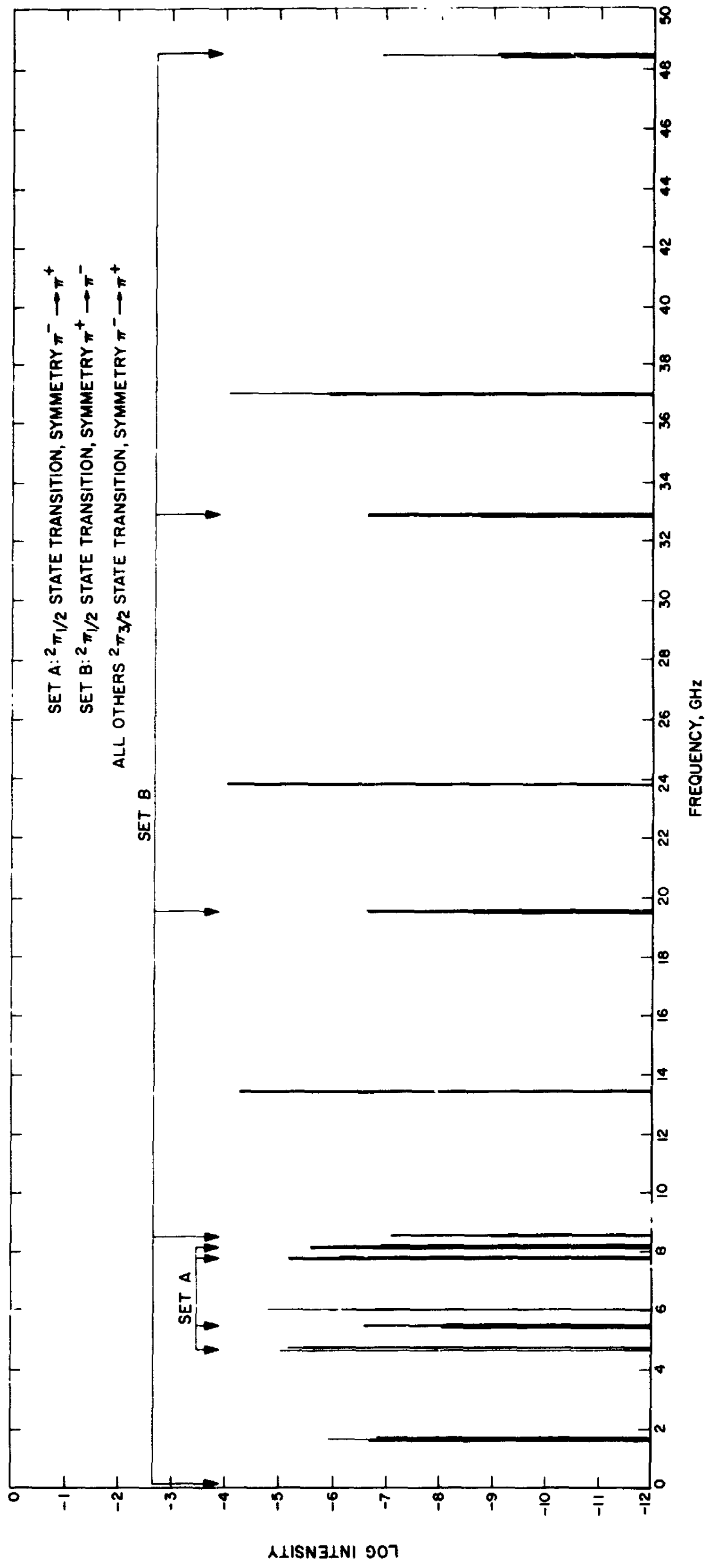


Fig. 8. OH free radical microwave transitions

D. An Ion Cyclotron Resonance Study of the Escape of Helium From the Earth's Atmosphere,

J. King, Jr., and D. D. Elleman

1. Introduction

Helium atoms are systematically being lost from the earth's atmosphere. This conclusion is based on many ion probe experiments, primarily by Hale (Ref. 1). For the He atoms to exist in steady-state concentrations in the atmosphere, they must be lost at a rate comparable to their rate of production by radioactive decay ($\sim 10^6$ atoms-cm^{-2}-s^{-1}). The fact that this rate is approximately the same as for He photo-ionization (Ref. 2) in the upper atmosphere suggests that its escape could be explained by an ion–molecule reaction mechanism which yielded He atoms with adequate kinetic energy. Generally, it has been assumed that a dissociative charge transfer reaction with N_2 predominates (Ref. 3). The reaction with N_2 is more probable than that with O_2 because of the greater abundance of the former in the upper atmosphere. Laboratory studies of the He^+–N_2 reaction have been made using a crossed-beam technique (Ref. 4). In those collisions that lead to N^+ production, the process is observed to have quasiresonant form. Because of these findings, the generally accepted mechanism is the accidental near-resonant charge transfer reaction

$$He^+(^2S) + N_2(^1\Sigma_g^+) \rightarrow He(^1S) + N_2^+(C\,^2\Sigma_u^+, v = 3, 4)$$

followed by predissociation

$$\rightarrow He(^1S) + N(^4S) + N^+(^3P) \qquad (1)$$

in accordance with the Franck–Condon principle.

The primary objection to this mechanism is that it does not produce He atoms with sufficient energy (2.4 eV) to escape the earth's gravitational field. This fact has led to an alternative mechanism in which the He^+, produced by solar photo-ionization, charge exchanges with O_2 instead of N_2 (Ref. 5). The He^+–O_2 reaction is exothermic by 5.8 eV which, if completely localized in the He fragment, gives it more than enough energy to escape.

The basic problem with this latter mechanism has been mentioned previously; i.e., N_2 is much more abundant in the upper atmosphere and any charge exchange is more likely to occur with N_2 than with O_2.

A more attractive mechanism is for He^+ to charge exchange with N_2 in a non-near-resonant process in which N_2^+ is produced in the ground state ($X\,^2\Sigma_g^+$) and the reaction is exothermic by 9 eV. A test for this mechanism is to look for N_2^+ as a stable product since the $C\,^2\Sigma_u^+$ state of Reaction (1) is known to predissociate in 10^{-8} s (Ref. 6).

It can be inferred from the spectroscopic studies of Inn (Ref. 6) that N_2^+ is produced as a stable ion in He^+–N_2 systems. A more direct study has recently been performed by Warneck (Ref. 7) using tandem mass spectrometers. He concluded that N_2^+ and N^+ are produced with about equal efficiency in the system.

The difficulty in unequivocally determining the N^+ and N_2^+ products in most mass spectroscopic experiments is that these ions are also produced initially by the same source used to ionize the He. Thus, the initial ions must, in some way, be differentiated from the product ions.

2. Experimental Procedure

The technique of ion cyclotron double resonance (ICDR) is ideally suited for selectively studying a particular ion–molecule reaction. This method, which has been described previously (SPS 37-46, Vol. IV, pp. 205–208), involves the simultaneous RF heating of one type of ion while a second type is being observed under cyclotron resonance conditions. When the first type of ion is heated with a strong RF electric field, $E_2(t)$ at ω_2, large changes should occur in the concentrations of the other types of ions, provided they are coupled with the first type through charge transfer. These changes are detected with a weak RF electric field, $E_1(t)$ at frequency ω_1, through changes in the intensity of the observed ion spectra. The amplitude of the field E_2 is modulated and the signal at ω_1 is detected with a phase detector referenced to the modulating frequency. With this setup only those additional ions produced by the RF heating are observed.

3. Results

The ICDR technique was used to study the production of N_2^+ and N^+ in the He^+–N_2 system when He^+ is subjected to RF heating. The ion production was studied as a function of He^+ energy. The He ion energy can be varied by varying the amplitude E of the irradiating field, $E_2(t) = E \sin \omega t$. The results in Fig. 9 show that both N_2^+ and N^+ are produced in the He^+–N_2 system. The ordinate denotes the amplitude of the double resonance signal that is proportional to the number density of N_2^+ or N^+ ions produced by RF heating of He^+ (SPS 37-50, Vol. III, pp. 231–236). The abscissa is the amplitude of the irradiating RF field and is proportional to

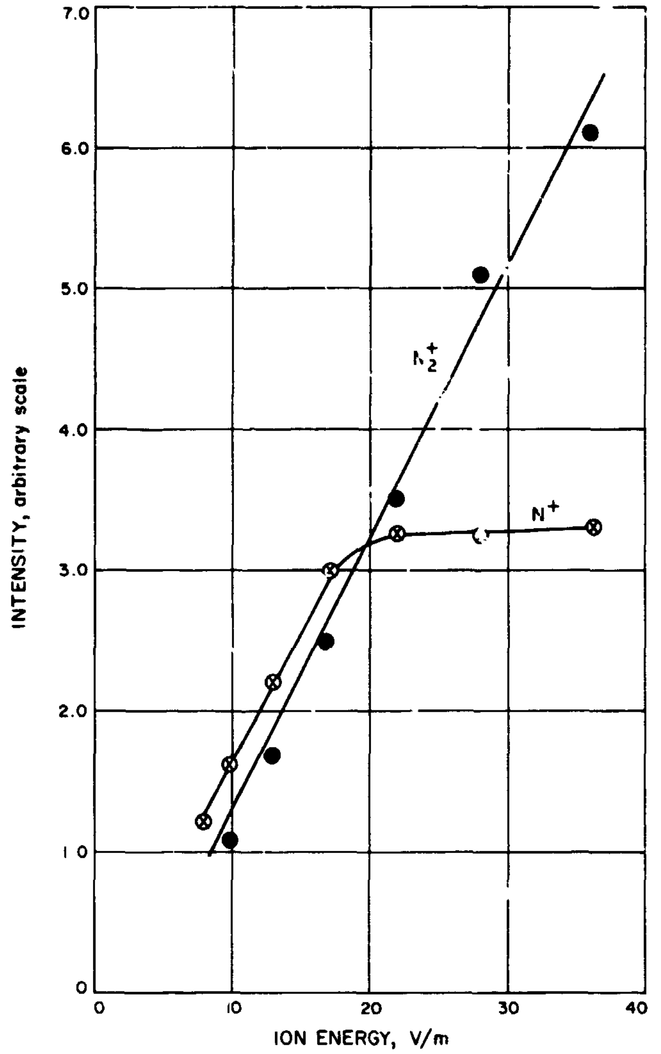


Fig. 9. Variation of N_2^+ and N⁺ production with energy of HE⁺ ions ($P = 7.5 \times 10^{-6}$)

the energy of the He⁺ ions. The surprising result in Fig. 9 is the leveling off of N⁺ production at high He⁺ energies. An explanation of this phenomenon is advanced in the following paragraphs.

4. Discussion

The results conclusively show that both N_2^+ and N⁺ are produced when He⁺ ions bombard neutral N_2. In order for the neutral He atom, produced in the charge exchange reaction, to have sufficient kinetic energy to escape the earth's gravitational pull, the N_2^+ can either be in the $B\,^2\Sigma_u^+$ state or the ground state ($X\,^2\Sigma_g^+$). The former state is 3.14 eV above the ground state (Ref. 8) and based on the difference in ionization potentials between He and N_2; the reaction leading to the $B\,^2\Sigma_u^+$ state is exothermic by approximately 6 eV. As noted earlier, this is more than enough to allow He to escape from the upper atmosphere.

The N⁺ production can be explained by Reaction (1). The N_2^+, initially formed in the $C\,^2\Sigma_u^+$ states, predissociates in 10^{-8} s to form N⁺ and N. Because of the limitations of the ion cyclotron resonance spectrometer, it is impossible to observe such a short-lived species. The plateau in the N⁺ curve in Fig. 9 shows that the formation of the $C\,^2\Sigma_u^+$ state does not continue to increase with increasing He⁺ ion energy. This could be caused by the fact that as the He⁺ velocity increases there is not sufficient time for it to form a complex with N_2 to produce the $C\,^2\Sigma_u^+$ state. Since the state is formed by the simultaneous ionization of one electron and excitation of another, the two species must be in close contact for a reasonable length of time. However, the production of N_2^+ in the ground state is not limited by this requirement since the electron can jump from the N_2 to the He⁺ over relatively large distances, similar to the modified stripping mechanism proposed by Herman, et al. (Ref. 9).

To understand why N⁺ production becomes constant rather than decreases, one must consider the details of the RF heating of the He⁺ ion. The ions are heated through power absorption from the irradiating RF field. The power absorption equation is (SPS 37-50, Vol. III)

$$A(\omega) = \frac{n^+ e^2 E^2}{4m} \cdot \frac{\nu_0}{(\omega - \omega_0)^2 + \nu_0^2} \qquad (2)$$

where

n^+ = ion density

e = charge on the electron

E = electric field strength

ω = oscillator frequency

ω_0 = cyclotron frequency of the ion

ν_0 = collision frequency for momentum transfer

As can be seen, maximum power absorption occurs at resonance when $\omega = \omega_0$. However, the experiments are performed by sweeping the frequency ω from off resonance, through resonance, and past resonance. When the frequency is off resonance, the He⁺ ions are absorbing less energy and their velocity is less. They can thus form complexes with N_2 and produce the $C\,^2\Sigma_u^+$ state. The number of these less energetic ions remains rather constant as E in Eq. (2) increases.

To test this hypothesis, experiments with argon (Ar^+) and neon (Ne^+) are being initiated.

References

1. Hale, L. C., "Ionospheric Measurements with a Multigrid Retarding Potential Analysis," Abstract, *J. Geophys. Res.*, Vol. 66, p. 1554, 1961.
2. Nicolet, M., "Helium, An Important Constituent in the Lower Exosphere," *J. Geophys. Res.*, Vol. 66, p. 2263, 1961.
3. Stebbings, R. F., Rutherford, J. A., and Turner, B. R., "Loss of He^+ Ions in the Upper Atmosphere," *Planet. Space Sci.*, Vol. 13, p. 1125, 1965.
4. St bbirgs R. F., Smith, A. C. A., and Ehrhart, H., "Dissociative Charge Transfer in He^+–O_2 and He^+–N_2 Collisions," *J. Chem. Phys.*, Vol. 39, p. 968, 1963.
5. Bates, D. R., and Patterson, T. H. L., "Helium Ions in the Upper Atmosphere" *Planet. Space Sci.*, Vol. 9, p. 599, 1962.
6. Inn, E. C. V., "Charge Transfer Between He^+ and N_2," *Planet. Space Sci.*, Vol. 15, p. 19, 1967.
7. Warneck, P., "Studies of Ion-Neutral Reactions by a Photoionization Mass-Spectrometer Technique. IV. Reactions of He^+ and N_2 and O_2," *J. Chem. Phys.*, Vol. 47, p. 4279, 1967.
8. Herzberg, G., *Molecular Spectra and Molecular Structure: Volume I, Spectra of Diatomic Molecules*, p. 554. D. Van Nostrand Co., Inc., Princeton, N. J., Feb. 1963.
9. Herman, Z., et al., "Crossed-Beam Studies of Ion-Molecule Reaction Mechanisms," *Discuss. Faraday Soc.* (to be published).

E. Shape of the Magnetosphere,

G. Atkinson and T. Unti

As the solar wind passes the earth, it confines the earth's magnetic field within a cavity called the magnetosphere. In recent years satellite data have shown the cavity to have a shape more complicated than had been anticipated. A neutral sheet has been observed, indicating that some of the field lines are dragged great distances downstream by the solar wind, strongly distorting the shape of the cavity and the magnetic field within it. While a number of attempts have been made to calculate the shape of the magnetosphere, the calculations have failed to include the effect of the neutral sheet satisfactorily. This article reports on the calculations that have been performed taking the neutral sheet into account. The calculations yield possible shapes for the magnetosphere that are illustrated in Fig. 10. Each of the shapes is determined by the amount of magnetic flux contained within the tail portion of the magnetosphere.

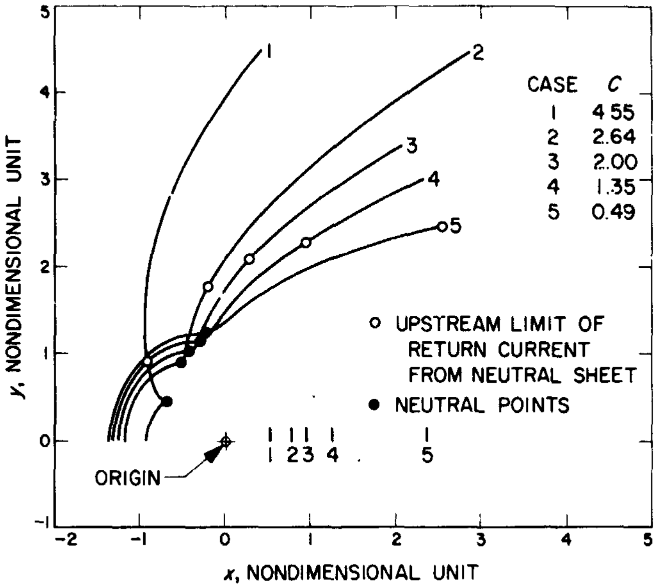


Fig. 10. Boundary of the two-dimensional magnetosphere

The calculation parallels a previous calculation made by Dungey (Ref. 1). The problem of calculating the shape of the magnetosphere regarding the solar wind as a particle gas is known as the Chapman–Ferraro problem, and was shown by Dungey to have an exact solution if certain simplifying assumptions were made. The most drastic simplification was to treat the problem in two dimensions only. The other assumptions are:

(1) The surface of the cavity is thin.

(2) The field is completely screened; i.e., plasma pressure and momentum in the interior are unimportant.

(3) Thermal velocities of the streaming particles are neglected.

(4) The particles are specularly reflected at the surface of the cavity.

Using Dungey's assumptions, a new exact solution to the Chapman–Ferraro problem was found in which the parameters of a neutral sheet are determined along with the shape of the field lines.

Since the problem is two-dimensional, the method of complex potentials can be applied. A scalar potential ϕ and vector potential Ψ are introduced, such that the magnetic field $H = \nabla\phi$, and Ψ is the stream function, constant on a magnetic line of force. The two-dimensional representation of the magnetosphere will be determined when Ψ is known as a function of x and y.

The free boundary, formed by magnetic field lines along which $\Psi = 0$, is not known as a function of $z = x + iy$. All that is known is that $\Phi = \phi + i\Psi$ must be an analytic function of z; but, then, z must also be an analytic function of Φ. Now, the free boundary in z space becomes a very simpie known boundary in potential space, $\Phi = \phi + i\Psi$. Therefore, it is only necessary to find that function $z(\Phi)$ which satisfies the Laplace equation and reduces to the proper boundary conditions in the potential plane. To find this function, a conformal transformation is made that maps the given boundary in Φ space onto the abscissa in $w = u + iv$ space. An application of Fourier transforms then yields the function $z(\Phi)$. Integrations were calculated on the IBM 7094 computer. The results, reduced to unit dipole, are shown in Fig. 10, in which the boundary of the two-dimensional magnetosphere is given for graded values of tail flux C.

Reference

1. Dungey, J. W., *J. Geophys. Res.*, Vol. 66, p. 1043, 1961.

XX. Communications Systems Research

TELECOMMUNICATIONS DIVISION

A. Coding and Synchronization Studies: A General Formulation of Linear Feedback Communications Systems With Solutions,

S. Butman

1. Introduction

A feedback communication system is a two-way system in which the state of a message at the receiver is made available to the transmitter. Although the benefits of feedback are greatest when the feedback link is noiseless, Shannon was able to prove (Ref. 1) that it cannot be used to exceed the capacity of a *memoryless* channel. It is possible, however, to *exceed* the capacity of a channel *with memory* (Ref. 2). Furthermore, feedback simplifies the coding and decoding effort and provides a lower error than could otherwise be achieved. These considerable advantages are obtained at the expense of the feedback link, which could be put to better use. Oftentimes, however, the return path is idle and should be used to benefit the forward link. In space applications, a relatively inexpensive high capacity *up-link* could be sacrificed for a more efficient exploitation of the *down-link* whose capacity is small due to weight restrictions required for take-off.

This article is concerned with linear feedback communication systems as originally studied by Elias (Refs. 3 and 4), later by Green (Ref. 5), and more recently by Schalkwijk and Kailath (Ref. 6), Schalkwijk (Refs. 7 and 8), Schalkwijk and Bluestein (Ref. 9), Omura (Ref. 10), and Butman (Ref. 11). The techniques used include the Robbins–Monro method of stochastic approximation (Ref. 12) used in Refs. 6 and 7, center-of-gravity (Ref. 8), Bellman's dynamic programming (Ref. 13) used in Ref. 10, directed graphs introduced by Elias (Ref. 4), and Kalman filtering (Ref. 14) in Refs. 10 and 11.

However, none of these techniques are adequate to handle the general linear feedback communication problem to be considered here. With the exception of Elias' work,[1] they fail to provide the correct approach to the noisy feedback problem even in the case of only one feedback iteration. Furthermore, in the case of a white gaussian noise channel with a noiseless feedback link, where all of these techniques have been successfully used, the results do not agree completely and the discrepancies are not adequately explained. In addition, the techniques are applied only after specific linear relationships are assumed to hold between the forward and feedback signals and between the feedback signals and the receiver's estimates of the message. These assumptions represent unnecessary constraints which confine the search for the

[1] The principle of optimality of dynamic programming used in Ref. 10 is not generally applicable to feedback systems. For counter examples see Chap. 10 of Ref. 15.

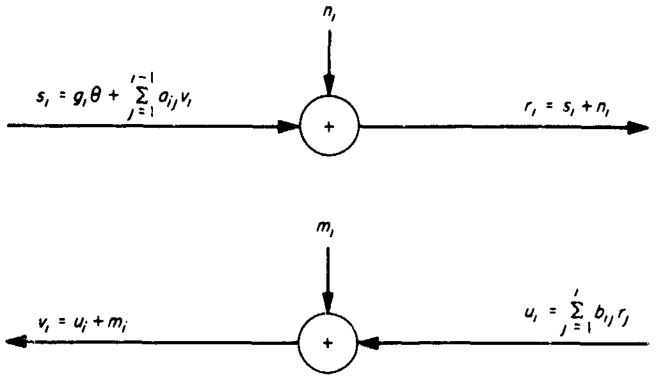
optimum to a subset of the class of all possible linear feedback codes and allow the possibility of the existence of better schemes.

A complete linear formulation in terms of arbitrary linear operations at the transmitting and receiving points is presented in *Subsection 2* for systems with additive noise in both the forward and feedback channels, including noise which is colored and correlated between channels. The optimum decision rule is derived in the case of gaussian noise, and the signal selection problem is stated for both the forward and feedback signal sets subject to an average power constraint on each. The gaussian assumption is a convenience since the problem is identical for any additive noise and a minimum mean-square error receiver.

Noiseless feedback is considered in *Subsection 3*, where the optimum sequential forms for the forward signals and the estimates at the receiver are derived. Also, a theorem is stated giving sufficient conditions for achieving channel capacity with a double-exponential decreasing error rate using partially optimum codes. There are more than a countable variety of such codes. The effect of noiseless feedback on a channel with memory is examined in the example of first-order Markov noise. The code used, although not optimum, achieves the theoretical capacity of the forward channel when the bandwidth is infinite and exceeds the theoretical capacity when the bandwidth is finite.

The noisy feedback problem for a system with independent white noise in each channel is treated in *Subsection 4*, where the optimum code for one feedback iteration is determined. Further penetration is algebraically unmanageable. However, successive iteration of the available result yields a better scheme than the iterative scheme suggested by Elias in Ref. 4. In addition, its asymptotic behavior is easily found in closed form, thereby determining a useful lower bound. This lower bound approaches the upper bound for noisy feedback for large signal-to-noise ratios in the forward link.

2. Formulation of the Problem



$s_i = g_i\theta + \sum_{j=1}^{i-1} a_{ij} v_i$

$u_i = \sum_{j=1}^{i} b_{ij} r_j$

Fig. 1. A linear feedback communication system

A linear feedback communication system using a sequence of N signals to transmit a message θ is illustrated in Fig. 1. Each signal is formed by amplitude modulating a basic pulse of unit energy and duration ½ W, where W is the bandwidth. The pulse is detected by a matched filter whose output is the amplitude corrupted by the additive noise in the channel. The sequence of amplitudes $s_1, s_2, \cdots, s_N$ is the code in the forward channel, and the sequence $r_1, r_2, \cdots, r_N$ is the set of noisy observations. Similarly, the feedback code is the sequence of feedback amplitudes $u_1, u_2, \cdots, u_{N-1}$ which are observed by the transmitter as $v_1, v_2, \cdots, v_{N-1}$. The process begins with $s_1 = g_1\theta$ being sent and r_1 being received. The first feedback signal is $u_1 = b_{11}r_1$, and it is observed at the transmitter as v_1. The second signal is now assumed to be a linear function of θ and v_1, thus, $s_2 = g_2\theta + a_{21}v_1$. In general, the ith signal and observation at each point is given by

$$s_i = g_i\theta + \sum_{j=1}^{i-1} a_{ij}v_j \tag{1}$$

$$r_i = s_i + n_i, \qquad i = 1, 2, \cdots, N \tag{2}$$

$$u_i = \sum_{j}^{i} b_{ij}r_j \tag{3}$$

$$v_i = u_i + m_i, \qquad i = 1, 2, \cdots, N-1 \tag{4}$$

where $n_1, n_2, \cdots, n_N$ and $m_1, m_2, \cdots, m_{N-1}$ are zero mean gaussian random variables representing the additive noise in the forward and feedback channels, respectively. The last feedback signal u_N is not used and is therefore not considered. Let **A** and **B** be $N \times N$ lower triangular matrices with the main diagonal of **A** and the last row of **B** identically zero, and let **g**, **m**, **n**, **r**, **s**, **u**, and **v** be N-dimensional column vectors or $N \times 1$ matrices.

Then

$$\mathbf{s} = \mathbf{g}\theta + \mathbf{A}\mathbf{v} \tag{5}$$

$$\mathbf{r} = \mathbf{s} + \mathbf{n} \tag{6}$$

$$\mathbf{u} = \mathbf{B}\mathbf{r} \tag{7}$$

$$\mathbf{v} = \mathbf{u} + \mathbf{m} \tag{8}$$

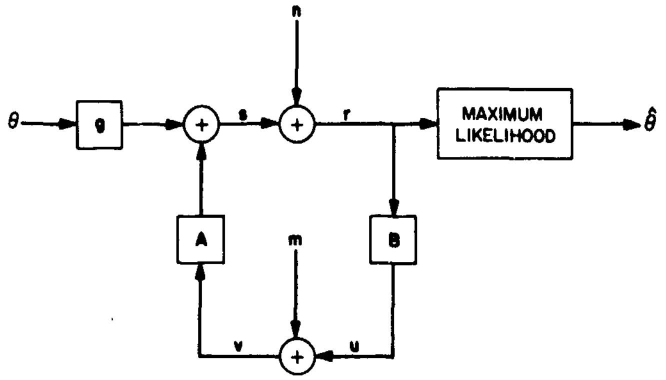


Fig. 2. Matrix formulation of the feedback communication process

and the system is equivalent to the vector feedback system of Fig. 2.

Equations (5) to (8) may be solved for **r**, **s**, and **u** as linear functions of the random noise vectors **m** and **n** and the random variable θ. Thus, substituting Eq. (6) into Eq. (7) into Eq. (8) and the result into Eq. (5) gives

$$\mathbf{s} = (\mathbf{I} - \mathbf{AB})^{-1}(\mathbf{g}\theta + \mathbf{Am} + \mathbf{ABn}) \qquad (9)$$

$$\mathbf{r} = (\mathbf{I} - \mathbf{AB})^{-1}(\mathbf{g}\theta + \mathbf{Am} + \mathbf{n}) \qquad (10)$$

and

$$\mathbf{u} = \mathbf{B}(\mathbf{I} - \mathbf{AB})^{-1}(\mathbf{g}\theta + \mathbf{Am} + \mathbf{n}) \qquad (11)$$

where **I** is the $N \times N$ identity matrix. Note that the inverse of $\mathbf{I} - \mathbf{AB}$ exists because the product **AB** is a lower triangular matrix with zeros along the main diagonal, whereupon $\mathbf{I} - \mathbf{AB}$ must be a lower triangular matrix with *ones* down the main diagonal and det $(\mathbf{I} - \mathbf{AB}) = 1$. The average energy transmitted in the forward and feedback directions is $E = E[\mathbf{s}^T\mathbf{s}]$ and $E' = \mathbf{E}[\mathbf{u}^T\mathbf{u}]$, respectively, where $\mathbf{E}[\,\cdot\,]$ is the expectation operator. The term $\mathbf{s}^T\mathbf{s} = \mathrm{tr}[\mathbf{s}^T\mathbf{s}] = \mathrm{tr}[\mathbf{s}\mathbf{s}^T]$, where $\mathrm{tr}[\,\cdot\,]$ is the trace operator which is invariant under cyclic permutation of the argument, and the superscript T denotes transpose. Since the expectation and trace operators commute, it follows that

$$E = \mathrm{tr}\,[(\mathbf{I} - \mathbf{AB})^{-1}(\sigma_\theta^2\,\mathbf{g}\mathbf{g}^T + \mathbf{AK}_m\mathbf{A}^T + \mathbf{AK}_{mn}\mathbf{B}^T\mathbf{A}^T + \mathbf{ABK}_{mn}^T\,\mathbf{A}^T + \mathbf{ABK}_n\mathbf{B}^T\mathbf{A}^T)(\mathbf{I} - \mathbf{B}^T\mathbf{A}^T)^{-1}] \qquad (12)$$

and

$$E' = \mathrm{tr}\,[\mathbf{B}(\mathbf{I} - \mathbf{AB})^{-1}(\sigma_\theta^2\mathbf{g}\mathbf{g}^T + \mathbf{K})(\mathbf{I} - \mathbf{B}^T\mathbf{A}^T)^{-1}\,\mathbf{B}^T] \qquad (13)$$

where θ is statistically independent of **m** and **n**, $\sigma_\theta^2 = \mathbf{E}[\theta^2]$, $\mathbf{K}_m = \mathbf{E}[\mathbf{m}\mathbf{m}^T]$, and $\mathbf{K}_n = \mathbf{E}[\mathbf{n}\mathbf{n}^T]$ are covariance matrices of the noise, $\mathbf{K}_{mn} = \mathbf{E}[\mathbf{m}\mathbf{n}^T]$ is the cross-covariance matrix, and

$$\mathbf{K} = \mathbf{K}_n + \mathbf{AK}_{mn} + \mathbf{K}_{mn}^T\mathbf{A}^T + \mathbf{AK}_m\mathbf{A}^T \qquad (14)$$

The average power used in the forward and feedback channels is then

$$P = E/T = 2WE/N \text{ and } P' = 2WE'/(N-1),$$

respectively, where $T = N/2W$ is the duration of the N forward signals and $(N-1)/2W$ is the duration of the $N - 1$ feedback signals.

The optimum decision rule for an equiprobable source. Given **A**, **B**, and **g**, the decision rule for minimum error is for the receiver to select the message for which the *a posteriori* probability $p(\theta\,|\,\mathbf{r})$ is a maximum over all the possible messages in the message set Θ. In general, this rule depends on the *a priori* probability $p(\theta)$ because by Bayes' rule

$$P(\theta\,|\,\mathbf{r}) = p(\mathbf{r}\,|\,\theta)\frac{p(\theta)}{p(\mathbf{r})} \qquad (15)$$

However, $p(\theta) = 1/M$ is independent of θ when Θ is a set of M equiprobable points, such as the M uniformly spaced points in the interval $[-L, L]$. In this case, it is equivalent for the receiver to select the message that maximizes $p(\mathbf{r}\,|\,\theta)$. The vector **r** is a sufficient statistic for estimating θ, another sufficient statistic is

$$\mathbf{y} = (\mathbf{I} - \mathbf{AB})\,\mathbf{r} \qquad (16)$$

$$= \mathbf{g}\theta + \mathbf{Am} + \mathbf{n} \qquad (17)$$

Since **y** is a linear function of the gaussian vectors **m** and **n**, it must be conditionally normal with conditional mean $\mathbf{E}[\mathbf{y}\,|\,\theta] = \mathbf{g}\theta$ and covariance matrix

$$\mathbf{E}[\mathbf{y} - \mathbf{g}\theta)(\mathbf{y} - \mathbf{g}\theta)^T\,|\,\theta] = \mathbf{K}$$

Thus,

$$p(\mathbf{y}|\theta) = [(2\pi)^N \det \mathbf{K}]^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{g}\theta)^T \mathbf{K}^{-1}(\mathbf{y} - \mathbf{g}\theta)\right] \tag{18}$$

Now, it is obvious that selecting θ to maximize $p(\mathbf{y}|\theta)$ is the same as minimizing the quadratic form

$$(\mathbf{y} - \mathbf{g}\theta)^T \mathbf{K}^{-1}(\mathbf{y} - \mathbf{g}\theta) = (\mathbf{y} - \mathbf{g}\hat{\theta}_N)^T \mathbf{K}^{-1}(\mathbf{y} - \mathbf{g}\hat{\theta}_N) + (\theta - \hat{\theta}_N)^2 \mathbf{g}^T\mathbf{K}^{-1}\mathbf{g} \tag{19}$$

where

$$\hat{\theta}_N = \frac{\mathbf{g}^T\mathbf{K}^{-1}\mathbf{y}}{\mathbf{g}^T\mathbf{K}^{-1}\mathbf{g}} \tag{20}$$

can be any point on the real line. Thus, the optimum decision procedure for the receiver is to select the *maximum-likelihood* estimate of θ as the point $\theta^* \epsilon \Theta$ which is closest to $\hat{\theta}_N$.

It can be verified easily that $\hat{\theta}_N$ is the *minimum-variance unbiased linear* estimate of θ given **r**. Note that $\hat{\theta}_N$ is also a sufficient statistic for estimating θ at the receiver and that $\hat{\theta}_N$ is conditionally normal with conditional mean $\mathbf{E}[\hat{\theta}_N|\theta] = \theta$ and variance $\mathbf{E}[\hat{\theta}_N - \theta)^2|\theta] = 1/\mathbf{g}^T\mathbf{K}^{-1}\mathbf{g}$. Therefore,

$$p(\hat{\theta}_N|\theta) = \left(\frac{2\pi\rho_{oN}}{\sigma_\theta^2}\right)^{1/2} \exp\left[-(\hat{\theta}_N - \theta)^2 \frac{\rho_{oN}}{2\sigma_\theta^2}\right] \tag{21}$$

where

$$\rho_{oN} = \sigma_\theta^2 \mathbf{g}^T\mathbf{K}^{-1}\mathbf{g} \tag{22}$$

is the signal-to-noise ratio $\mathbf{E}[\theta^2]/\mathbf{E}[\hat{\theta}_N - \theta)^2]$ at the receiver after N observations. A quantity closely related to $\hat{\theta}_N$ is

$$\tilde{\theta}_N = \frac{\hat{\theta}_N \rho_{oN}}{1 + \rho_{oN}} \tag{23}$$

which is the *minimum-variance (biased) linear* estimate of θ given the vector of observations **r**. Note that

$$\mathbf{E}[(\tilde{\theta}_N - \theta)^2] = \frac{\sigma_\theta^2}{1 + \rho_{oN}} < \mathbf{E}[(\hat{\theta}_N - \theta)^2]$$

and that $\tilde{\theta}_N$ is the value of θ that maximizes $p(\theta|\mathbf{r})$ when $p(\theta)$ is gaussian with zero mean and variance σ_θ^2, whereas $\hat{\theta}_N$ maximizes $p(\mathbf{r}|\theta)$ regardless of the distribution on Θ.

The probability of error given that θ was sent is the probability that $|\theta^* - \hat{\theta}_N| < |\theta - \hat{\theta}_N|$ for some $\theta^* \neq \theta$. Since the nearest neighbor distance is $2L/(M - 1)$, the condition for an error when θ is one of the $M - 2$ interior points of $[-L, L]$ is

$$|\hat{\theta}_N - \theta| \geqq \frac{L}{M - 1} = \left[\frac{3\sigma_\theta^2}{M^2 - 1}\right]^{1/2}$$

where $\sigma_\theta^2 = L^2(M + 1)/3(M - 1)$. Thus, the conditional probability of error is

$$P_e = \int_{|\theta_N - \theta| \geqq L/(m-1)} p(\hat{\theta}_N|\theta)\, d\theta$$

$$\doteq \operatorname{erfc}\left(\frac{15\,\rho_{oN}}{M^2 - 1}\right)^{1/2} \tag{24}$$

where

$$\operatorname{erfc}(x) = \frac{2}{(\pi)^{1/2}} \int_x^\infty \exp(-x^2)\, dx$$

When θ is one of the end points $\pm L$, the condition for an error becomes $\pm\theta \leqq L(M - 2)/(M - 1)$, respectively. As in this case the conditional error probability is negligibly lower; the average and conditional error probabilities are nearly equal.

From Eq. (24), it is clear that P_e decreases monotonically with ρ_{oN}. Consequently, **A**, **B**, and **g** should be chosen to maximize ρ_{oN} or $\ln(1 + \rho_{oN})$ and to satisfy the average energy constraints as given by Eqs. (10) and (13). Other constraints are not considered here. Conceptually, we can extremize the Hamiltonian

$$F = \ln(1 + \rho_{oN}) - \lambda E - \mu E' \tag{25}$$

where λ and μ are Lagrange multipliers, by setting the derivative of F with respect to each of the total of N^2 unknown elements in **A**, **B**, and **g** equal to zero and solving the resulting set of N^2 nonlinear equations. Practically, this is an extremely difficult, if not impossible, task for $N \geq 2$ unless the feedback channel is noiseless, that is, unless the constraint on the feedback energy is removed ($\mu = 0$).

3. Noiseless Feedback

The absence of feedback noise is indicated in the general formulation by the vanishing of $\mathbf{m}$, $\mathbf{K}_m$, and $\mathbf{K}_{mn}$. Therefore, $\mathbf{K} = \mathbf{K}_n$ is independent of $\mathbf{A}$ and $\mathbf{B}$, and the signal-to-noise ratio $\rho_{oN} = \sigma_\theta^2 \mathbf{g}^T \mathbf{K}^{-1} \mathbf{g}$ depends only on $\mathbf{g}$. The feedback energy E' is no longer a constraint, because it can be scaled down to $\epsilon^2 E'$ for $|\epsilon|$ arbitrarily small simply by scaling $\mathbf{B}$ to $\epsilon\mathbf{B}$ and $\mathbf{A}$ to $\mathbf{A}\epsilon^{-1}$. This leaves E, which now depends only on the product $\mathbf{AB}$, unaffected.

Define the lower triangular zero-main-diagonal matrix $\mathbf{C} = (\mathbf{I} - \mathbf{AB})^{-1}\mathbf{AB}$. Then

$$(\mathbf{I} - \mathbf{AB})^{-1} = \mathbf{I} + \mathbf{C}, \qquad \mathbf{AB} = (\mathbf{I} + \mathbf{C})^{-1}\mathbf{C}$$

$$\mathbf{s} = (\mathbf{I} + \mathbf{C})\,\mathbf{g}\theta + \mathbf{Cn} \tag{26}$$

and

$$E = \sigma_\theta^2 \,\|(\mathbf{I} + \mathbf{C})\,\mathbf{g}\|^2 + \mathrm{tr}\,[\mathbf{CKC}^T] \tag{27}$$

where $\|\cdot\|$ is the Euclidian norm, $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \mathrm{tr}\,[\mathbf{xx}^T]$. The $N(N-1)/2$ arbitrary elements of $\mathbf{C}$ can be chosen to minimize E independently of $\mathbf{g}$ and thus independently of ρ_{oN}. Let $\mathbf{Q}$ be the lower triangular nonsingular "whitening" matrix defined by the factorization $\mathbf{Q}^T\mathbf{Q} = \mathbf{K}^{-1}$, and let $\mathbf{f} = \sigma_\theta \mathbf{Qg}$, then $\rho_{oN} = \|f\|^2$.

Now, the result of the minimization of E with respect to $\mathbf{C}$ (Ref. 11) is the functional form of the optimum linear coding and decoding operations. Thus,

$$s_i = g_i\,(\theta - \tilde{\theta}_{i-1}) \tag{28}$$

where

$$\tilde{\theta}_i = \tilde{\theta}_{i-1} + \frac{\sigma_\theta f_i}{1 + \rho_{oi}} \sum_{j=1}^{i} q_{ij}\,[r_j - g_j\,(\tilde{\theta}_{i-1} - \tilde{\theta}_{j-1})] \tag{29}$$

is the minimum variance (biased) estimate of θ given $r_1, r_2, \cdots, r_i$ and

$$\rho_{oi} = f_1^2 + f_2^2 + \cdots + f_i^2 \tag{30}$$

is the signal-to-noise ratio associated with $\tilde{\theta}_i$,

$$1 + \rho_{oi} = \sigma_\theta^2 / E\,[(\theta - \tilde{\theta}_i)^2]$$

Equation (29) provides a recursive decoding procedure for the receiver, since θ^* is determined from

$$\hat{\theta}_N = \tilde{\theta}_N\,(1 + \rho_{oN})/\rho_{oN}$$

In addition, it gives a recursive procedure for generating the forward signals, because Eq. (28) provides

$$\frac{s_i}{g_i} - \frac{s_j}{g_j} = \tilde{\theta}_{j-1} - \tilde{\theta}_{i-1} \tag{31}$$

The order of the linear difference equation (Eq. 29) is determined by the order of the noise in the forward channel, which determines the elements q_{ij} of the matrix $\mathbf{Q}$. Thus, mth-order autoregressive noise is characterized by the vanishing of q_{ij} for $j < i - m$, which reduces Eq. (29) to a linear difference equation of order $m + 1$.

Now, the expected forward energy

$$E = \sum_{i=1}^{N} e_i$$

where from Eq. (28) and the fact that

$$E\,[(\theta - \theta_{i-1})^2] = \frac{\sigma_\theta^2}{1 + \rho_{o(i-1)}}$$

then

$$e_i = \sigma_\theta^2 \frac{g_i^2}{1 + \rho_{o(i-1)}} \tag{32}$$

is the energy in the ith signal, must be minimized over $\mathbf{g}$ subject to the constraint $\sigma_\theta^2 \mathbf{g}^T\mathbf{K}^{-1}\mathbf{g} = \rho_{oN} = \|\mathbf{f}\|^2$. This leads to the algebraic problem of solving the N nonlinear coupled equations

$$\frac{\partial}{\partial g_i}\,[E - \lambda \ln (1 + \rho_{oN})] = 0$$

where λ is a Lagrange multiplier. The transformation $\mathbf{f} = \sigma_\theta \mathbf{Qg}$ does not simplify the problem, and the equations are too difficult to solve in closed form except when the forward noise is white, or more generally, when $\mathbf{K} = \mathbf{K}_n$ is diagonal. A good choice of $\mathbf{g}$, which reduces to the optimum choice if $\mathbf{K}$ is diagonal, is obtained as follows.

Note that ρ_{oi} as given by Eq. (30) satisfies the identity

$$1 + \rho_{oi} \equiv \prod_{j=1}^{i} \left(1 + \frac{f_j^2}{1 + \rho_{o(j-1)}}\right) \qquad (33)$$

which from Eq. (32) is

$$= \prod_{j=1}^{1} \left(1 + \frac{e_j f_j^2}{\sigma_j^2 g_j^2}\right) \qquad (34)$$

where

$$\left(\frac{f_i}{\sigma_e g_i}\right)^2 = q_{ii}^2 \left(1 + \sum_{j=1}^{i-1} \frac{q_{ij} g_j}{q_{ii} g_i}\right)^2 \qquad (35)$$

Next, let $\sigma_i = 1/q_{ii}$, $\rho_i = e_i/\sigma_i^2$ and note from Eq. (32) that specifying the energies $e_1, e_2, \cdots, e_N$ determines only the magnitudes $|g_1|, |g_2|, \cdots, |g_N|$. Therefore, the sign of g_i can be chosen to give

$$\left(\frac{\sigma_i f_i}{\sigma_e g_i}\right)^2 = \left(1 + \left|\sum_{j=1}^{i-1} \frac{q_{ij} g_i}{q_{ii} g_i}\right|\right)^2 \qquad (36)$$

independently of the signal energies. Consequently, Eq. (34) becomes

$$1 + \rho_{oN} = \prod_{i=1}^{N} \left[1 + \rho_i \left(1 + \left|\sum_{j=1}^{i-1} \frac{q_{ij} g_j}{q_{ii} g_i}\right|\right)^2\right] \qquad (37)$$

$$\geq \prod_{i=1}^{N} (1 + \rho_i) \qquad (38)$$

with equality if and only if **K** is diagonal.

Now, the choice of signal energies that maximizes the lower bound (Eq. 38), and which is the optimum when **K** is diagonal, is found from

$$\frac{\partial}{\partial e_i} \sum_{j=1}^{N} \left[e_i - \lambda \ln\left(1 + \frac{e_j}{\sigma_j^2}\right)\right] = 0$$

Thus,

$$e_i = e + \sigma^2 - \sigma_i^2 \qquad (39)$$

where

$$\sigma^2 = \sum_{i=1}^{N} \frac{\sigma_i^2}{N}$$

and $e = E/N > \sigma_i^2 - \sigma^2$ for all i. Otherwise, it is necessary to omit the signal corresponding to the largest σ_i and to reduce N until the condition $e > \sigma_i^2 - \sigma^2$ holds. This will not be necessary if E is sufficiently large or if $\sigma_i^2 = \sigma^2$ for all i, in which case $e_i = e$ and $\rho_i = \rho = e/\sigma^2$ for all i.

Optimum code for the additive white gaussian noise channel. From Eq. (39) and the associated discussion and the fact that $\mathbf{K} = \sigma^2 \mathbf{I}$ is diagonal, where $\sigma^2 = N_0/2$ is the two-sided spectral power density of the white noise, it is clear that the optimum choice of signal energies is $e_i = e = P/2W$ so that $\rho_i = \rho = P/N_0W$. Therefore, Eq. (34) becomes

$$1 + \rho_{oi} = (1 + \rho)^i \qquad (40)$$

Eq. (32) gives

$$g_i = \frac{\sigma}{\sigma_e} [\rho (1 + \rho)^{i-1}]^{1/2} \qquad (41)$$

Eq. (29) reduces to

$$\tilde{\theta}_i = \tilde{\theta}_{i-1} + \frac{\sigma_e}{\sigma} \left(\frac{\rho}{(1 + \rho)^{i+1}}\right)^{1/2} r_i \qquad (42)$$

and Eq. (31) gives

$$s_i = (1 + \rho)^{1/2} \left(s_{i-1} - \frac{\rho}{1 + \rho} r_{i-1}\right) \qquad (43)$$

where the initial conditions are $s_1 = g_1 \theta$ and $\theta_0 = 0$.

The probability of error from Eq. (29) is exactly

$$P_e = \operatorname{erfc}\left[\frac{3}{2} \frac{(\exp 2CT) - 1}{(\exp 2RT) - 1}\right]^{1/2} \qquad (44)$$

where

$$C = \frac{[\ln (1 + \rho_{oN})]}{2T}$$

$$= W \ln\left(1 + \frac{P}{N_0 W}\right) \text{nats/s} \qquad (45)$$

is the theoretical capacity of the channel and $R = (\ln M)/T$ is the rate of the message source in nats/s. It then follows from the asymptotic expansion of the error function integral that P_e decreases to zero with increasing T for all $R < C$ as the doubly exponential function

$$P_e \sim \left(\frac{2}{3}\pi\right)^{1/2} \exp\{-(C - R)T - 1.5 \exp 2(C - R)T\} \tag{46}$$

From Eq. (44), it is also evident that $P_e = \operatorname{erfc}(3/2)^{1/2}$ for $R = C$ and $P_e \rightarrow 1$ as $T \rightarrow \infty$ for $R > C$.

Non-optimum codes. The choice of signal energies, and therefore the choice of g, is not critical for achieving the doubly exponential decrease of error or even channel capacity. Note that P_e, as given by Eq. (24), decreases to zero if and only if ρ_{oN} increases to infinity, which in turn requires that E increase to infinity, because $1 + E/\sigma^2 \leqq 1 + \rho_{oN} \leqq \exp E/\sigma^2$. Consequently, we define the critical rate cf a code, R_c, by the two conditions:

$$\frac{E}{\sigma^2} = \lim_{N \rightarrow \infty} \sum_{i=1}^{N} \rho_i = \infty \tag{47a}$$

$$\begin{aligned} R_c &= \lim_{N \rightarrow \infty} \frac{1}{2T} \ln(1 + \rho_{oN}) \\ &= \lim_{N \rightarrow \infty} \frac{P}{N_0} \frac{\sum_{i=1}^{N} \ln(1 + \rho_i)}{\sum_{i=1}^{N} \rho_i} \end{aligned} \tag{47b}$$

with equality for additive white gaussian noise, and prove the following theorem.

Theorem 1. If the sequence $\{\rho_i\}_{i=1}^{\infty}$ converges to a limit ρ, then the critical rate R_c is given by

$$R_c = \begin{cases} W \ln\left(1 + \dfrac{P}{N_0 W}\right), & \text{if } \rho = \dfrac{P}{N_0 W} \\ \dfrac{P}{N_0}, & \text{if } \rho = 0 \\ 0, & \text{if } \rho = \infty \end{cases} \tag{48}$$

with equality for additive white gaussian noise. The proof is in Ref. 11.

As an example of codes satisfying the conditions of Theorem 1, consider the class of codes in which $\rho_i = (1/i)^\gamma$, where $0 \leqq \gamma \leqq 1$. The optimum code is given by $\gamma = 0$, in which case $R_c = W \ln 2$. Otherwise, $R_c = P/N_0 = 1$ and the bandwidth is infinite. Although all these codes achieve the infinite bandwidth capacity limit, the growth of W with T is determined by γ. This is illustrated in Fig. 3.

First-order Markov channel. First-order Markov noise, or first-order autoregressive noise, is characterized by the first-order linear difference equation.

$$n_i = \alpha n_{i-1} + w_i, \qquad |\alpha| < | \tag{49}$$

where **w** is white noise with variance $\sigma_w^2 = N_0/2$. Stationarity implies that $E[n_i^2] = \sigma^2$ for all i. Therefore, $\sigma^2 = \alpha^2\sigma^2 + \sigma_w^2 = \sigma_w^2/(1 - \alpha^2)$.

The elements of **K** are $k_{ij} = \sigma^2 \alpha^{|i-j|}$, and the elements of **Q** are

$$q_{ij} = \begin{cases} \left(\dfrac{1 - \alpha^2}{\sigma_w}\right)^{1/2}, & \text{for } i = j = 1 \\ \dfrac{\delta_{ij} - \alpha\delta_{i(j-1)}}{\sigma_w}, & \text{otherwise} \end{cases}$$

where δ_{ij} is the Kroenecker delta.

Now, consider the following not necessarily optimum code: Take $e_1 = \sigma^2\rho_1$, $e_i = \sigma_w^2\rho$ for $i \geqq 2$ so that $P/N_0W = \rho$ as $N \rightarrow \infty$, and take $g_i/g_{i-1} = -x_i \operatorname{sgn} \alpha$, where

$$x_i = |g_i/g_{i-1}|$$

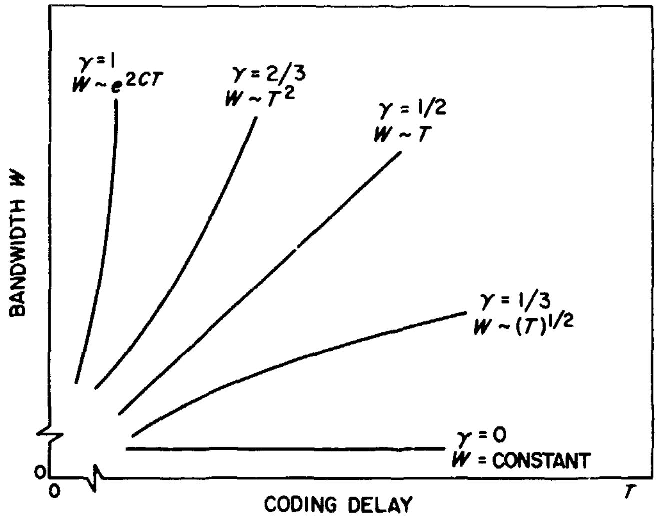


Fig. 3. Bandwidth vs coding delay for a class of codes

for $i \geqq 2$. With this choice of g, Eq. (35) gives

$$\left(\frac{\sigma_w f_i}{\sigma_e g_i}\right)^2 = \Bigg(1 + \left(\frac{|\alpha|}{x_i}\right)^2 \tag{50}$$

Eq. (32) gives

$$x_{i+1}^2 = \frac{1 + \rho_{0i}}{1 + \rho_{0(i-1)}}$$

which from Eqs. (37) and (50) becomes

$$x_{i+1}^2 = 1 + \rho\left(1 + \frac{|\alpha|}{x_i}\right)^2 \tag{51}$$

where the starting value as given by Eq. (32) is

$$x_2^2 = (1 - \alpha^2)\frac{\rho(1 + \rho_1)}{\rho_1} \tag{52}$$

We can select ρ_1 in Eq. (52) such that $x_2 = x$, where x is the only positive stationary point of Eq. (51), in order to obtain $x_i = x$ for all $i \geqq 2$. Thus, Eq. (37) becomes

$$1 + \rho_{oN} = (1 + \rho_1)x^{2(N-1)} \tag{53}$$

giving

$$R_c = W \ln x^2 \tag{54}$$

where, from Eq. (51) x is related to ρ by

$$\rho = \frac{x^2(x^2 - 1)}{(x + |\alpha|)^2} \tag{55}$$

and from Eq. (52)

$$1 + \rho_1 = \frac{(x + |\alpha|)^2}{(|\alpha| x + 1)^2}$$

and $1 + \rho \leqq x^2 \leqq 1 + \rho(1 + |\alpha|)^2$ with the lower bound holding for large values of ρ and the upper bound as ρ tends to zero ($W \rightarrow \infty$). Thus,

$$R_c = \begin{cases} W \ln x^2, & \text{for } W < \infty \ (\rho > 0) \\ (1 + |\alpha|)^2 \dfrac{P}{N_0}, & \text{for } W \rightarrow \infty \ (\rho \rightarrow 0) \end{cases} \tag{56}$$

For comparison, the one-way theoretical capacity of the first-order Markov channel is

$$C = \begin{cases} W \ln \left[\dfrac{1}{(1 + \alpha^2)} + \rho\right], & \text{for } \rho \geqq \dfrac{1}{(1 + |\alpha|)^2} \\ (1 + |\alpha|)^2 \dfrac{P}{N_0}, & \text{for } W = \infty \end{cases} \tag{57}$$

which shows that R_c exceeds the theoretical capacity of the forward link for $\rho \geqq 1/(1 - |\alpha|)^2$ since we obtain

$$\begin{aligned} R_c - C &= W \ln \left[\frac{1}{(1 - \alpha^2)x^2} + \frac{x^2 - 1}{(x + |\alpha|)^2}\right] \\ &= \frac{2|\alpha| W}{x}\left[1 - 0\left(\frac{1}{x}\right)\right] \\ &\geqq 0 \text{ for sufficiently large } x \end{aligned}$$

This does not violate Shannon's theorem (Ref. 1) because the channel has memory. In fact, knowledge of α is equivalent to having additional or side information at the transmitter. It is shown by Shannon (Ref. 2) that feedback can, in such cases, increase the capacity.

4. Noisy Feedback

The case of most practical interest is when all channels are corrupted, independently, by additive white noise. The optimum output signal-to-noise ratio for $N = 2$ in this case is

$$\rho_{o2} = \rho_1 + \rho_2 + \frac{\rho_1 \rho_2 \rho_1'}{(1 + \rho_1)(1 + \rho_2) + \rho_1'} \tag{58}$$

where ρ_1 and ρ_2 are signal-to-noise ratios of the two forward signals s_1 and s_2, and ρ_1' is the ratio of the feedback signal u_1. The optimum allocation of ρ_1 and ρ_2 subject to $\rho_1 + \rho_2 = 2\rho$ is $\rho_1 = \rho_2 = \rho$, hence

$$\rho_{o2} = 2\rho\frac{\rho^2 \rho'}{(1 + \rho)^2 + \rho'} \tag{59}$$

Unfortunately, a closed form expression for the optimum ρ_{oN} is unavailable for $N > 2$ because of algebraic difficulties. However, the upper bound

$$\rho_{oN} < N\rho + (N - 1)\rho' \tag{60}$$

where

$$\rho = \frac{1}{N}\sum_{i=1}^{N} \rho_i \text{ and } \rho' = \frac{1}{N-1}\sum_{i=1}^{N-1} \rho_i'$$

was recently proved by Elias (Ref. 4) by means of a rather complicated circuit theoretical argument. Equation (58) is also due originally to Elias (Refs. 3–4). A considerably simplified proof of Eqs. (58) and (59), using the matrices **A**, **B** and the vector **g**, follows:

Derivation of Eq. (58). There is no loss of generality in letting $\mathbf{K}_m = \mathbf{K}_n = \mathbf{I}$ and $\sigma_\theta^2 = 1$ since this converts signal energies to signal-to-noise ratios. In particular, for $N = 2$, $\rho_1 = g_1^2$, $\rho_1' = b^2(1 + g_1^2)$, $\rho_2 = (g_2 + abg_1)^2 + a^2(b^2 + 1)$ and $\rho_{o2} = g_1^2 + g_2^2/(1 + a^2)$. Next, let $abg_1 = kg_2$, where k will be determined shortly. Then

$$\rho_2 = g_2^2\left[(1 + k)^2 + (1 + \rho_1 + \rho_1')\frac{k^2}{\rho_1\rho_1'}\right]$$

and

$$\rho_{o2} = \rho_1 + \frac{\rho_1\rho_2\rho_1'}{\rho_1\rho_1'(1 + k)^2 + [(1 + \rho_1)(1 + \rho_2) + \rho_1']k^2}$$

The optimum choice of k is that which minimizes the quadratic denominator, thus

$$k_{opt} = \frac{-\rho_1\rho_1'}{(1 + \rho_1)(1 + \rho_1 + \rho_2)}$$

and

$$\rho_{o2} = \rho_1 + \rho_2 + \frac{\rho_1\rho_2\rho_1'}{(1 + \rho_1)(1 + \rho_2) + \rho_1'}$$

Proof of Elias' upper bound $\rho_{oN} \leqq N\rho + (N - 1)\rho'$. Since $\mathbf{K}_{mn} = 0$ when the noise is uncorrelated between channels, Eq. (14) becomes $\mathbf{K} = (\mathbf{I} + \mathbf{A}\mathbf{A}^T)$ and hence

$$\begin{aligned}\rho_{oN} &= \mathbf{g}^T(\mathbf{I} + \mathbf{A}\mathbf{A}^T)^{-1}\mathbf{g} \\ &= \|\mathbf{f}\|^2\end{aligned}$$

where $\mathbf{f} = (\mathbf{I} + \mathbf{A}\mathbf{A}^T)^{-1/2}\mathbf{g}$. From Eq. (12) and $(\mathbf{I} - \mathbf{A}\mathbf{B})^{-1} = (\mathbf{I} + \mathbf{C})$, $N\rho = E$ is

$$\begin{aligned}N\rho &= \|(\mathbf{I} + \mathbf{C})\mathbf{g}\|^2 + \mathrm{tr}[(\mathbf{I} + \mathbf{C})\mathbf{A}\mathbf{A}^T(\mathbf{I} + \mathbf{C})^T + \mathbf{C}\mathbf{C}^T] \\ &= \|(\mathbf{I} + \mathbf{C})\mathbf{g}\|^2 + \mathrm{tr}[(\mathbf{I} + \mathbf{C})(\mathbf{I} + \mathbf{A}\mathbf{A}^T)(\mathbf{I} + \mathbf{C}^T)] - N\end{aligned}$$

where the last line is obtained from the fact that $\mathrm{tr}[\mathbf{C}] \equiv 0$ and hence

$$\mathrm{tr}[(\mathbf{I} + \mathbf{C})(\mathbf{I} + \mathbf{C}^T)] = \mathrm{tr}[\mathbf{C}\mathbf{C}^T + \mathbf{C} + \mathbf{C}^T + \mathbf{I}] = \mathrm{tr}[\mathbf{C}\mathbf{C}^T] + N$$

From Eq. (13)

$$(N-1)\,\rho' = \|\mathbf{B}(\mathbf{I}+\mathbf{C})\,\mathbf{g}\,\|^2 + \mathrm{tr}\,[\mathbf{B}(\mathbf{I}+\mathbf{C})(\mathbf{I}+\mathbf{A}\mathbf{A}^T)(\mathbf{I}+\mathbf{C}^T)\,\mathbf{B}^T]$$

Now, define

$$\mathbf{M} = (\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{1/2}(\mathbf{I}-\mathbf{B}^T\mathbf{A}^T)^{-1}(\mathbf{I}+\mathbf{B}^T\mathbf{B})(\mathbf{I}-\mathbf{A}\mathbf{B})^{-1}(\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{1/2}$$

then it follows, after cyclic permutation of matrices under the trace operator when necessary, that

$$N\rho + (N-1)\,\rho' = \mathbf{f}^T\mathbf{M}\mathbf{f} + \mathrm{tr}\,\mathbf{M} - N$$

which is minimal with respect to **f** when **f** is the eigenvector corresponding to the smallest eigenvalue of **M**. Thus, without involving the constraint $\rho_{oN} = \|\mathbf{f}\|^2$, we have

$$N\rho + (N-1)\,\rho' = \lambda_1\rho_{oN} + \sum_{i=1}^{N}(\lambda_i - 1)$$

where $\lambda_1 \leqq \lambda_2 \leqq \cdots \lambda_N$ are the eigenvalues of M arranged in increasing order. Elias' result will follow if it can be shown that $\lambda_i \geqq 1$ and $\lambda_1 = 1$. This, in fact, has been proven by S. Farber of the California Institute of Technology. His proof is as follows:

$$\begin{aligned}(\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{1/2}\mathbf{M}^{-1}(\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{1/2} &= (\mathbf{I}-\mathbf{A}\mathbf{B})(\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{I}-\mathbf{B}^T\mathbf{A}^T)\\ &= (\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1} - \mathbf{A}\mathbf{B}(\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1} - (\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^T + \mathbf{A}\mathbf{B}(\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^T\end{aligned}$$

Next, apply the identities

$$\mathbf{B}(\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1} \equiv (\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}, \qquad (\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1} \equiv \mathbf{I} - \mathbf{B}^T(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$$

and

$$\mathbf{B}(\mathbf{I}+\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T \equiv \mathbf{I} - (\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}$$

to the appropriate terms on the right-hand side in order to obtain

$$\begin{aligned}(\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{1/2}\mathbf{M}^{-1}(\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{1/2} &= \mathbf{I} - \mathbf{B}^T(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} - \mathbf{A}(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} - \mathbf{B}^T(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{A}^T + \mathbf{A}[\mathbf{I}-(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}]\,\mathbf{A}^T\\ &= (\mathbf{I}+\mathbf{A}\mathbf{A}^T) - (\mathbf{A}+\mathbf{B}^T)(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}(\mathbf{A}^T+\mathbf{B})\end{aligned}$$

Therefore,

$$\mathbf{M}^{-1} = \mathbf{I} - \mathbf{H}$$

where

$$\mathbf{H} = (\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{-1/2}(\mathbf{A}+\mathbf{B}^T)(\mathbf{I}+\mathbf{B}\mathbf{B}^T)^{-1}(\mathbf{A}^T+\mathbf{B})(\mathbf{I}+\mathbf{A}\mathbf{A}^T)^{-1/2}$$

is obviously non-negative definite. This is sufficient to prove that $\lambda_i(\mathbf{M}) \geqq 1$. However, the rank of **H** is equal to the rank of $(\mathbf{A}^T + \mathbf{B})$ which is at most $N - 1$ because the last row is identically zero. Thus, at least one of the eigenvalues of **H** must be zero and, therefore, $\lambda_1(\mathbf{M}) = 1$.

A lower bound. A useful lower bound on the output signal-to-noise ratio can be obtained by applying Eq. (58) iteratively. Since the result of Eq. (58) is indistinguishable at the receiver from that of an equivalent single forward signal with ratio ρ_{o2}, we can apply Eq. (58) to ρ_{o2}, ρ_3, and ρ_2' to give

$$\rho_{o3} = \rho_{o2} + \rho_3 + \frac{\rho_{o2}\rho_3\rho_2'}{(1+\rho_{o2})(1+\rho_3)+\rho_2'}$$

continuing in this manner yields

$$\rho_{on} = \rho_{o(n-1)} + \rho_n + \frac{\rho_{o(n-1)}\,\rho_n\rho_{n-1}'}{(1+\rho_{o(n-1)})(1+\rho_n)+\rho_{n-1}'} \qquad (61)$$

Next, consider the asymptotic form of ρ_{on} when ρ_n and ρ_n' are constants ρ and ρ', respectively. ($\rho_n = \rho$ is the optimum allocation of forward ratios when the feedback link is noiseless, $\rho' = \infty$.) Equation (61) simplifies to

$$\rho_{on} - \rho_{o(n-1)} = \rho\left(1+\frac{\rho'}{1+\rho}\right)$$

$$\times\left[1-\frac{\rho'}{(1+\rho)(1+\rho_{o(n-1)})+\rho'}\right] \qquad (62)$$

hence

$$\rho < \rho_{on} - \rho_{o(n-1)} < \rho\left(1+\frac{\rho'}{1+\rho}\right)$$

$$n\rho < \rho_{on} < n\rho\left(1+\frac{\rho'}{1+\rho}\right) \qquad (63)$$

Substituting $n\rho$ for ρ_{on} in the right-hand side of Eq. (62) gives the inequality

$$\rho_{on} - \rho_{o(n-1)} > \rho\left(1+\frac{\rho'}{1+\rho}\right)$$

$$\times\left\{1-\frac{\rho'}{(1+\rho)[1+(n-1)\rho]+\rho'}\right\} \qquad (64)$$

Summing both sides from $n = 2$ to N gives

$$\rho_{oN} - \rho > \rho\left(1+\frac{\rho'}{1+\rho}\right)$$

$$\times\left[N-1-\frac{\rho'}{\rho(1+\rho)}\ln\left(\frac{\rho(1+\rho)(N-1)+1\rho+\rho'}{\rho(1+\rho)+1+\rho+\rho'}\right)\right]$$

Consequently,

$$\rho_{oN} > N\rho + \frac{(N-1)\,\rho'\rho}{1+\rho} - 0\,(\ln N) \qquad (65)$$

which equals 90% of the upper bound when $\rho = 10$, and only 10% when $\rho = 0.1$.

The iterative coding procedure represented by Eq. (61) gives a better result than the iterative scheme proposed by Elias (Ref. 4), in which $N = 2^K$ signals are coded in K stages (concatenated in a sense) via Eq. (59) to obtain

$$\rho_{on} = 2\rho_{o(\frac{1}{2}n)} + \frac{(\rho_{o(\frac{1}{2}n)})^2\;\rho_{(\frac{1}{2}n)-1}'}{(1+\rho_{o(\frac{1}{2}n)})^2 + \rho_{(\frac{1}{2}n)-1}'} \qquad (66)$$

The reason why Eq. (61) is better than Eq. (66) is because the feedback signals in Eq. (61) convey more information than they do in Eq. (66). This can also be verified numerically; for example, if $\rho = \rho' = 1$, then after $N = 2^{10}$ iterations Eq. (61) yields $\rho_{oN}/N = 1.496$ (max = 1.5) while Eq. (66) gives $\rho_{oN}/N = 1.400$.

There is cause to suspect that the upper bound Eq. (60) is too large for small values of ρ. For instance, if $\rho = 0$, then $\rho_{on} \equiv 0$ independently of ρ'. This suggests that there should be a term like $\rho/(1+\rho)$ multiplying ρ' in Eq. (60). It is, therefore, not unreasonable to conjecture that the results of Eqs. (61) to (65) differ only by a negligible amount from the truly optimum linear feedback code.

5. Conclusions

The utility of the complete formulation of linear feedback systems introduced in Subsection 2 has been demonstrated in Subsections 3 and 4, where new results and results previously obtained by others were derived from a unified approach. The derivations of Subsection 3 comprise a proof of the optimum linear noiseless feedback coding procedure not previously published. The formula

$$s_i = g_i\,(\theta - \tilde{\theta}_{i-1})$$

was previously obtained by Omura (Ref. 10) for a channel with white noise under the special assumption that $\tilde{\theta}_i$ satisfies a first-order linear difference equation and the receiver selects the message θ^* closes to $\tilde{\theta}_N$. This decision rule is not optimum when θ is uniformly distributed, although it is correct when θ is a gaussian random variable. The assumption that $\tilde{\theta}_i$ satisfies a difference equation is not necessary. It serves only to complicate the problem, and represents an additional *a priori* constraint on the signal set.

The feedback scheme described by Schalkwijk in Refs. 6–9 which uses signals that in our notation are given by

$$s_i = g_i(\theta - \hat{\theta}_{i-1})$$

is also linear because $\hat{\theta}_i$ is linear in $\mathbf{r}_i = \text{col}(r_1, \cdots, r_i)$. It can be easily derived from the general formulation, however, by including the additional linear constraint relationship $(\mathbf{I} - \mathbf{AB})^{-1}\,\mathbf{g} = \text{col}(g_1, 0, \cdots, 0)$ that must hold when $s_i = g_i(\theta - \hat{\theta}_{i-1})$ for then

$$\mathbf{E}[\mathbf{s}\,|\,\theta] = \text{col}(g_1\,\theta, 0, \cdots, 0)$$

(Ref. 11). Because of this constraint, the signal-to-noise ratio for the additive white gaussian noise channel is at most $(1 + \rho - 1/N)^N$, which is somewhat less than optimum linear result $(1 + \rho)^N$ and, as $N \rightarrow \infty$, the ratio

$$\frac{(1+\rho)^N}{\left(1 + \rho - \dfrac{1}{N}\right)^N} \rightarrow \exp\left(\frac{1}{1+\rho}\right)$$

Recently Schalkwijk and Bluestein (Ref. 9) pointed out that the rate distortion bound can be achieved in the case of a gaussianly distributed source by means of the noiseless feedback scheme $s_i = g_i(\theta - \tilde{\theta}_{i-1})$. For a uniformly distributed source, one would expect to achieve at least as good a signal-to-noise ratio as that of a gaussian source of equal variance, since the uncertainty (entropy) must be less for the uniformly distributed source. Schalkwijk and Bluestein suggest the inferior scheme $s_i = g_i(\theta - \hat{\theta}_{i-1})$ as "appropriate" for the reason that $\hat{\theta}_i$ is the maximum *a posteriori* probability (MAP) estimate of θ when θ is uniformly distributed (perhaps in analogy with the fact that $\tilde{\theta}_i$ is also the MAP estimate when θ is gaussianly distributed).

However, the MAP estimate is not the minimum variance (linear or nonlinear) estimate when θ is not gaussian. Moreover, the MAP estimate when θ is uniformly distributed on $[-L, L]$ is not $\hat{\theta}_i$ but the restriction of $\hat{\vartheta}_i$ to $[-L, L]$; that is, the MAP estimate is $\hat{\phi}_i = \hat{\theta}_i$, for $|\hat{\theta}_i| < L$, and $\hat{\phi}_i = L\,\text{sgn}\,\hat{\theta}$, for $|\hat{\theta}_i| > L$. The use of $\hat{\phi}_i$ as a feedback signal in $s_i = g_i(\theta - \hat{\phi}_i)$ takes us into the realm of nonlinear feedback, because ϕ_i is clearly a nonlinear function of $\mathbf{r}_i$. The best linear or nonlinear feedback signal with which to minimize the variance and hence the transmitted energy is well known (Ref. 16) to be the *conditional mean* $\mathbf{E}[\theta\,|\,r_i]$. Indeed, $\mathbf{E}[\theta\,|\,r]$ is the center-of-gravity proposed earlier by Schalkwijk in Ref. 8, but not used for $p(\theta)$ uniform.

Unfortunately, $\mathbf{E}[\theta\,|\,\mathbf{r}_i]$ is linear in $\mathbf{r}_i$ if and only if θ is gaussianly distributed. Thus, although it is possible to find $\mathbf{E}[\theta\,|\,\mathbf{r}_i]$ in closed form for θ uniform on $[-L, L]$, it is impossible to express $\mathbf{E}[(\theta - \mathbf{E}[\theta\,|\,\mathbf{r}_i])^2]$ in a workable manner. Nevertheless, since $\mathbf{E}[\theta\,|\,\mathbf{r}_i]$ must be in $[-L, L]$, it is reasonable to use $\hat{\phi}_i$ as an approximati .ı. By the same token, the truncated version of $\tilde{\theta}_1, \tilde{\phi} = \tilde{\theta}_i$ for $|\theta_i| < L$ and $\tilde{\phi}_i = L\,\text{sgn}\,\tilde{\theta}_i$ for $|\tilde{\theta}_i| > L$ can be used. It is then easy to show that

$$\mathbf{E}[(\theta - \tilde{\phi}_i)^2] < \mathbf{E}[(\theta - \tilde{\theta}_i)^2] - \mathbf{E}[(\tilde{\theta}_i - \tilde{\phi}_i)^2] < \frac{\sigma_\theta^2}{1 + \rho_{0i}}$$

Similarly, $\mathbf{E}[(\theta - \hat{\phi}_i)^2] < \mathbf{E}[(\theta - \hat{\theta}_i)^2 - \mathbf{E}[(\hat{\theta}_i - \hat{\theta}_i)^2]$. Although this author has not been able to establish an inequality between $\mathbf{E}[(\theta - \tilde{\phi}_i)^2]$ and $\mathbf{E}[(\theta - \hat{\phi}_i)^2]$, it is evident that the nonlinear feedback signals are better than the linear signals.

With colored noise in the forward channel, the intuitive suggestion of whitening the channel and using the white-noise code has been made (Ref. 10). This scheme would achieve capacity for the whitened (and hence also for the colored) channel, but it would not exceed the capacity as predicted by Shannon (Ref. 2) and explicitly verified in Subsection 3. Furthermore, pre-whitening is a "time consuming" operation which, theoretically, requires infinite delay and therefore gives no opportunity for feedback. Actually, the impossibility of a simultaneously time-limited and bandlimited signal (Ref. 17) implies the nonexistence of even a white-noise channel. This gives added importance to the colored-noise problem.

Round-trip signal delays, measured in units of pulse duration, are easily included. If there are k units of delay, the first k rows of the lower triangular matrix A vanish. The minimum delay, however, is 1 pulse. With k units of delay, time division multiplexing will give $1 + \rho_{0N} = k(1 + \rho/k)^N$ If the pulse duration is increased k-fold, there will be only N/k feedback iterations and hence $1 + \rho_{0N} = (1 + \rho)^{N/k} < k(1 + \rho/k)^N$ for all $k > 1$. However, with k separate multiplex channels it is possible to send k independent messages each having $1 + \rho_{0N} = (1 + \rho/k)^N$ for a total capacity of $kN\ln(1 + \rho/k) = W\ln(1 + P/N_0W)$.

References

1. Shannon, C. E., "The Zero-Error Capacity of a Noisy Channel," *IRE Trans. on Inform. Theory*, Vol. II-2, pp. 8–19, Sep. 1956.
2. Shannon, C. E., "Channels with Side Information at the Transmitter," *IBM Journal*, Vol. 2, pp. 289–293, Oct. 1958.

3. Elias, P., "Channel Capacity Without Coding," *Quarterly Progress Report, Research Laboratory of Electronics*, pp. 90–93. Massachusetts Institute of Technology, Cambridge, Mass., Oct. 15, 1956.

4. Elias, P., "Networks of Gaussian Channels with Applications to Feedback Systems," *IEEE Trans. on Inform. Theory*, Vol. IT-13, pp. 493–501, July 1967.

5. Green, P. E., "Feedback Communication Systems," in *Lectures on Communication System Theory*, pp. 345–366. Edited by Baghdady. McGraw-Hill Book Co., Inc., New York, 1961.

6. Schalkwijk, J. P. M., and Kailath, T., "A Coding Scheme for Additive Noise Channels with Feedback – Part I: No-Bandwidth Constraint," *IEEE Trans. on Inform. Theory*, Vol. IT-12, pp. 172–182, Apr. 1966.

7. Schalkwijk, J. P. M., "A Coding Scheme for Additive Noise Channels with Feedback–Part II: Band-Limited Signals," *IEEE Trans. on Inform. Theory*, Vol. IT-12, pp. 183–189, Apr. 1966.

8. Schalkwijk, J. P. M., *Center-of-Gravity Information Feedback*, Research Dept. 501. Applied Research Laboratory, Sylvania Electronic Systems, Waltham, Mass., May 1966.

9. Schalkwijk, J. P. M., and Bluestein, L. L., "Transmission of Analog Waveforms Through Channels with Feedback," *IEEE Trans. on Inform. Theory*, Vol. IT-13, pp. 617–618, Oct. 1967.

10. Omura, J. K, "Signal Optimization for Channels with Feedback," Report SEL-66-068. Stanford Electronics Laboratories, Stanford, Calif., Aug. 1966.

11. Butman, S., *Optimum Linear Coding for Additive Noise Systems Using Feedback*, Ph.D. Thesis. California Institute of Technology, Pasadena, Calif., May 1967.

12. Robbins, H., and Monro, S., "A Stochastic Approximation Method," *Ann. Math. Statist.*, Vol. 22, pp. 400–407, 1951.

13. Bellman, R., *Dynamic Programming*. Princeton University Press, Princeton, N. J., 1957

14. Kalman, R. E., and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," *Trans. ASME, Ser. D: J. Basic Eng.*, pp. 95–108, Mar. 1961.

15. Aris, R., *Discrete Dynamic Programming*, Blaisdell Publishing Company, New York, 1964.

16. Blake, I. F., and Thomas, J. B., "On a Class of Processes Arising in Linear Estimation Theory," *IEEE Trans. Inform. Theory*, Vol. IT-14, pp. 12–16, Jan. 1968.

17. Gabor, D., "Theory of Communication," *Proc. Inst. Elec. Eng.*, Vol. 93, pp. 429–441, 1946.

B. Combinatorial Communication: The Maximum Indices of Comma Freedom for the High-Data-Rate Telemetry Codes,

L. D. Baumert and H. C. Rumsey, Jr.

1. Introduction

The high-data-rate telemetry project (SPS 37-48, Vol. II, pp. 83–130) uses the three biorthogonal Reed–Muller codes with parameters (16,5), (32,6), and (64,7). [Parameters (n, k) indicate that the code consists of 2^k binary n-tuples.] Word synchronization for these codes is provided by modulo 2 adding a suitable fixed binary n-tuple to each code word before it is transmitted. This n-tuple is called the *comma free vector* and the set of transmitted words is a *coset* of the original Reed-Muller code. If this coset is such that all possible n-tuples, which could arise from erroneous synchronization of the data stream, differ in at least r symbols from every word of the coset, then the coset is said to be *comma free of index r*. The maximum values of r occurring for the high-data-rate telemetry codes are discussed below.

2. Previous Results

The maximum index of comma freedom for the Reed–Muller (16,5) code is 2. This fact has been known for some time and is due to Stiffler. Because of its importance for the high-data-rate telemetry project, the references are cited. In his thesis (Ref. 1, pp. 139–143), Stiffler shows that the maximum index for the Reed–Muller (16,4) code is 2; this implies that the Reed–Muller (16,5) code has maximum index ≤ 2. On the other hand, Stiffler (Ref. 2, p. 147) provides a comma free vector of index 2 for the (16,5) code.

The maximum index of comma freedom for the Reed–Muller (32,6) code is 7. In fact, all comma free vectors of index 7 are explicitly determined in SPS 37-46, Vol. IV, pp. 221–226.

3. The Reed–Muller (64,7) Code

The maximum index of comma freedom for the Reed–Muller (64,7) code (call it I_{64}) is unknown. Stiffler (Ref. 2, pp. 147–156) has established that $14 \leq I_{64} \leq 26$ and furnishes there a comma free vector of index 14. It is shown below that $16 \leq I_{64} \leq 22$ for this code.

Since there are 2^{57} $(= 2^{64-7} \sim 140{,}000{,}000{,}000{,}000{,}000)$ cosets for this code, in contrast with the 2^{26} $(= 2^{32-6} \sim 67{,}000{,}000)$ cosets possessed by the (32,6) code, it should be no surprise that the basically enumerative techniques used (SPS 37-46, Vol. II, pp. 221–226) for that code are of no value here. Instead, using Stiffler's comma free vector of index 14 as a starting point, a gradient-type computer search was made on an SDS 930 in the hope of finding comma free vectors of higher index. This search resulted in the determination of several hundred comma free vectors with index 16, but none of index 17 or higher.

(Of course, the search was far from exhaustive.) One such comma free vector of index 16 is

00001100 00000110 11111100 10010100
11011100 00010000 11000011 00111001.

4. The Upper Bound

Let $V = V_h$ be the h-dimensional vector space over $GF(2)$. Represent the vectors $w \epsilon V$ as h-bit "words" $w = \{w_i, i = 1, \cdots, h\}$. If S is a subset of V, write $D(S)$ for the minimum weight (= number of 1's in the vector) of the vectors in S; geometrically, $D(S)$ is the distance from S to the origin. Also write $D(w)$ for the weight of the vector w.

Let J be the linear operator on V defined by

$$(Jw)_1 = 0$$

$$(Jw)_j = w_{j-1}, \qquad j = 1, \cdots, h$$

That is, J shifts w one bit to the right and inserts a zero in the first bit position. The operator J^k shifts $w\,k$ bits to the right and inserts zeros in the first k-bit locations. J is clearly a singular operator (e.g., $J^h \equiv 0$), but by an abuse of the notation write J^{-k} for the operator which shifts k places to the *left* and inserts zeros in the *last* k places. Finally, let G_{2^n} represent the $(2^n, n + 1)$ bi-orthogonal code. The index of comma freedom I_{64} of the (64,7) code can be defined by

$$I_{64} = \max_{w \epsilon V_{64}} \min_{k=1,\cdots,63} D[w + G_{64} + J^k(w + G_{64}) + J^{k-64}(w + G_{64})] \qquad (1)$$

We shall prove that $I_{64} \leqq 22$ by considering the case $k = 33$. The proof proceeds by means of three simple lemmas.

Lemma 1. Let $w \epsilon V_{16}$, then

$$D(w + G_{16}) \leqq 6$$

and equality holds if and only if every element of $w + G_{16}$ has weight either 6 or 10.

Proof. This is an elementary consequence of the standard Chebyshev argument (Ref. 2, p. 154) which shows that

$$D(w + G_{16}) \leqq \frac{16 - (16)^{1/2}}{2} = 6$$

and equality can occur only if all the vectors in $w + G_{16}$ have weight 6 or its compliment $16 - 6$.

Lemma 2. Let $w \epsilon V_{32}$, then

$$D(w + G_{32}) \leqq 12$$

Proof. The Chebyshev argument shows that

$$D(w + G_{32}) \leqq \frac{32 - (32)^{1/2}}{2} = 13.1 \cdots$$

Hence, it is only necessary to show that $D(w + G_{32}) = 13$ is impossible. Assume that $D(w + G_{32})$ is odd, then $D(w)$ is odd since the vectors in G_{32} have even weight. Write $w = w_1 w_2$ where w_1 and w_2 are 16-bit words and assume (by symmetry) that w_1 has odd weight and that w_2 has even weight. There are two cases to consider. First, let $D(w_2 + G_{16}) = 6$. It follows from lemma 1 that $D(w_2 + g) = 6$ or 10 for any $g \epsilon G_{16}$. It also follows from lemma 1 and the fact that $D(w_1 + G_{16})$ is odd that

$$D(w_1 + G_{16}) \leqq 5$$

Hence, let $g \epsilon G_{16}$ be such that $D(w_1 + g) \leqq 5$. Both gg and $g\bar{g}$ are elements of G_{32} (where $\bar{g}$ is the compliment of g). Thus

$$\begin{aligned} D(w + G_{32}) &\leqq D(w + \{gg, g\bar{g}\}) \\ &\leqq D(w_1 + g) + D(w_2 + \{g, \bar{g}\}) \\ &\leqq 5 + 6 = 11 \end{aligned}$$

Similarly, if $D(w_2 + G_{16}) \leqq 4$ (the other case) let $g \epsilon G_{16}$ be such that $D(w_2 + g) \leqq 4$. Then

$$\begin{aligned} D(w + G_{32}) &\leqq D(w + \{gg, \bar{g}g\}) \\ &\leqq D(w_1 + \{g, \bar{g}\}) + D(w_2 + g) \\ &\leqq 7 + 4 = 11 \end{aligned}$$

This completes the proof of the lemma.

Lemma 3. Let G_{31} be any 31-bit code obtained from G_{32} by deleting one of its bit locations. Then for any $w \in V_{31}$

$$D(w + G_{31}) \leqq 11$$

Proof. The proof is a simple parity argument. Let $w' \in V_{32}$ be the vector of *odd* weight obtained by filling in the "missing" bit of w. It follows that

$$D(w + G_{31}) \leqq D(w' + G_{32}) \leqq 12$$

by lemma 2. But since $D(w' + G_{32})$ is odd, the lemma is proved.

Theorem. The index of comma freedom I_{64} of the (64,7) bi-orthogonal code is at most 22.

Proof. Let $k = 33$ in Eq. (1); then

$$I_{64} \leqq \max_{w \in V_{64}} D\left[w + G_{64} + J^{33}(w + G_{64}) + J^{-31}(w + G_{64})\right] \tag{2}$$

$$\leqq \max_{w \in V_{64}} D(w + G_{64} + J^{33}G_{64} + J^{-31}G_{64})$$

Let G_1 be the group generated by the vectors $g_1 = (100, \cdots, 00) \in V_{64}$ and $g_{33} = J^{32}g_1$. Let G be the group obtained from $J^{-31}G_{64}$ by setting the first and thirty-third bits of each vector in $J^{-31}G_{64}$ equal to zero. Finally, let $G' = J^{33}G_{64}$. Then

$$G_1 + G + G' \subset G_{64} + J^{33}G_{64} + J^{-31}G_{64}$$

since the vectors x, y are in J^{-31} and z is in G_{64}, where

$$\begin{array}{rcccccccccc}
 & 1 & 2 & 3 & \cdots & 32 & 33 & 34 & \cdots & 63 & 64 \\
x = & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
y = & 0 & 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 & 0 \\
z = & 1 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0
\end{array}$$

Thus, it follows from Inequality (2) that

$$I_{64} \leqq \max_{w \in V_{64}} D(w + G_1 + G + G')$$

$$\leqq \max_{w \in V_2} D(w + \{00, 01, 10, 11\})$$

$$+ \max_{w \in V_{31}} D(w + G'_{31}) + \max_{w \in V_{31}} D(w + G_{31}) \tag{3}$$

where G_{31} is obtained from G by suppressing the $1, 33, 34, \cdots, 64$ bit positions of G, and G'_{31} is obtained by suppressing the $1, 2, \cdots, 33$ bi' positions of G'. Since both G_{31} and G'_{31} are groups of the type defined in lemma 3, we have by that lemma and Inequality (3),

$$I_{64} \leqq 0 + 11 + 11 = 22$$

This completes the proof of the theorem.

It seems likely that 22 is the best upper bound for I_{64} that can be obtained by considering a single shift k. For example, the distance 22 is attained for $k = 17$, 47, 31, 33, 32 and other values of k. To obtain a smaller upper bound, it is presumably necessary to consider several shifts simultaneously.

References

1. Stiffler, J. J., *Self-Synchronizing Binary Telemetry Codes*, Ph.D. thesis. California Institute of Technology, Pasadena, Calif., 1962.
2. Golomb, S. W., et al., *Digital Communications with Space Applications*. Prentice-Hall, Inc., New York, 1964.

C. Propagation Studies: A Map of the Venus Feature β, *S. Zohar and R. Goldstein*

Radar studies of Venus have shown that there exist on its surface relatively permanent topographic prominences. These features rotate with the planet and return to radar view year after year. Because of the peculiar rotation period of Venus, the same features return very nearly to the same apparent position at the time of closest approach. The feature known as β is the "brightest" and hence most favorable to observe at these times. Several other features are brighter, but are on the other side of the disk and are not presented to view until the radar range is much larger.

The feature β, as well as the other prominent features, was first located by a technique which is sensitive to only one-dimension, radar doppler shift (Refs. 1, 2, and 3). It has been established that the reflectivity of these features at 12.5 cm is significantly stronger than that of the average regions of Venus. They also have the ability to depolarize microwaves; that is, if right circularly polarized waves are beamed toward Venus, the reflections from the features contain a much larger percentage of right circularly polarized energy than the surrounding areas. This indicates that the features are relatively rough to the scale of one wavelength (12.5 cm). However, it is not known whether the features are mountains or craters or fields of boulders or some other such rough formations.

In order to gain information about the actual size and nature of the region β, it has been studied with a two-dimensional technique utilizing both range and doppler shift (Ref. 4). The result is a two-dimensional radar map cf the area. It is a unique map except for a north–south ambiguity, i.e., there are tw੍ɔ points, symmetric about the doppler equator, which have the same values of doppler shift and range. The results of our earlier studies, taken over several conjunctions of Venus, demonstrate that the highly reflective areas are actually in the northern hemisphere.

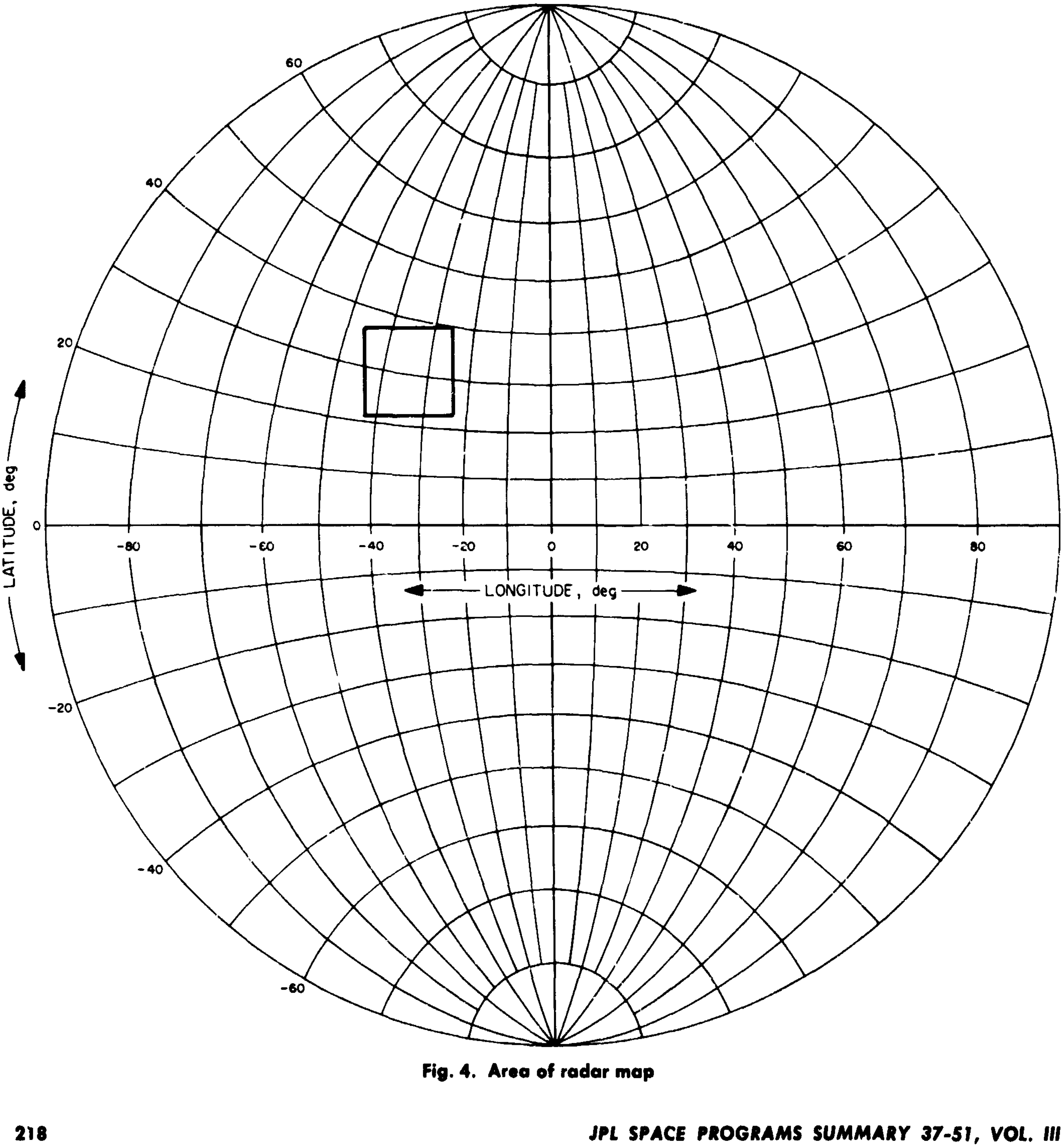


Fig. 4. Area of radar map

The location of the mapped region, in relation to the overall surface of Venus, is indicated in Fig. 4 by the rectangle in its upper left part. The grid of latitude and longitude circles shown here represents a frame of reference characterized as follows: The Venus rotation vector pierces the surface at latitude −90°. The zero meridian is the location of the sub-earth point at the 1967 inferior conjunction.

The map is shown in Fig. 5, where the darker regions represent areas of significantly higher than average reflectivity. This map was obtained as a weighted average of 17 probings of Venus, utilizing the radar capability of the Mars deep space station 210-ft antenna. These experiments were conducted between August 12 and September 11, 1967. Of the three distinct regions shown here, the one located at latitude 26° (β) was covered by 10 of 17 observations. The one at latitude 35°, previously identified as δ, was covered by six observations. The third region at longitude −40° was covered by three probings only and should thus be treated with some reserve, pending further experimental verification.

It is the nature of extended rough radar targets to show statistical variation. This is so because very small changes in aspect angle can cause large changes in reflected power. Hence, averages over many hours of observation are needed to produce reliable radar maps.

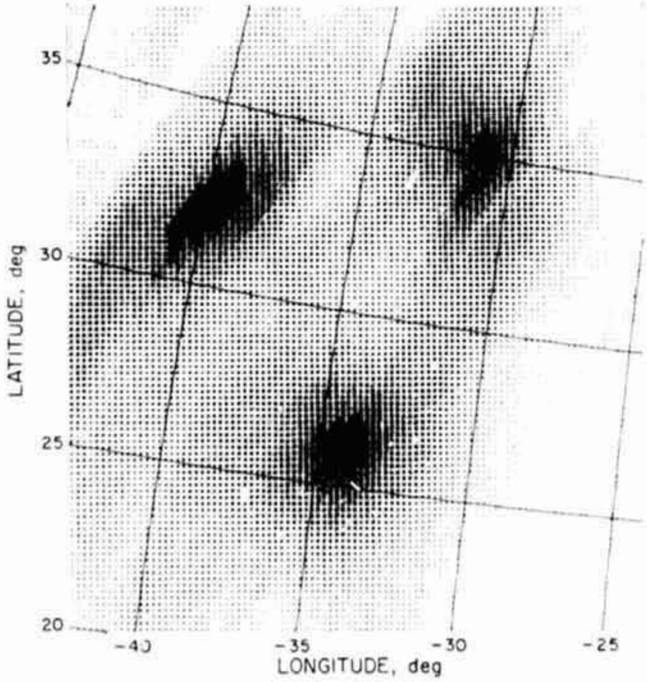


Fig. 5. Radar map of a Venus region

Some of the observations have shown a detailed structure for region β. However, the relatively high noise associated with these observations precludes their use in a single observation map.

References

1. Goldstein, R. M., "Preliminary Venus Radar Results," *Radio Science*, p. 1623, 1965.

2. Carpenter, R. L., *Astron. J.*, Vol. 71, p. 142, 1966.

3. Goldstein, R. M., *Moon and Planets*, pp. 126–131. Edited by Professor A. Dollfus. North-Holland Publishing Co., Amsterdam, The Netherlands, 1967.

4. Muhleman, D. O., Goldstein, R., and Carpenter, R., "A Review of Radar Astronomy," *IEEE Spectrum*, Oct., Nov. 1965.

D. Propagation Studies: The Variance of Scattering-Law Estimates, *D. G. Kelly*[2]

1. Introduction

If $\{x_j\}$ is a random process representing radar echoes from the surface of a planet, it is known (Ref. 1) that the power spectral density $P(f)$ of the process can be expressed in terms of the *backscatter function* $F(\theta)$ (the ability of the surface to reflect back to the observer a signal striking it at angle θ). The relation is

$$P(f) = \int_{\sin^{-1}(f/a)}^{\pi/2} F(\theta) \sin\theta \, (a^2 \sin^2\theta - f^2)^{-1/2} \, d\theta \qquad (1)$$

where $a\,(0 < a < 1)$ is the rotation constant, which can be defined here as the bandwidth of the spectrum, divided by the number of samples per second. Equation (1) has been inverted to yield

$$F(\theta) = -2a^2 \pi^{-1} \cos\theta \int_{a\sin\theta}^{a} P'(f)\,(f^2 - a^2 \sin^2\theta)^{-1/2} \, df \qquad (2)$$

Equation (2), in turn, makes it possible to estimate the backscatter function by expressing $P'(f)$ in terms of the covariances of the process and by using familiar estimates for the covariances.

[2] Resident Research Associate.

In this article, we derive asymptotic expressions and upper bounds for the variance of such an estimate. The estimate is

$$\hat{F}(\theta) = N^{-1} \cdot 4a^2 \cos\theta \sum_{j=1}^{K} w_j A_j \sum_{n=1}^{N} x_n x_{n+j} \tag{3}$$

where w_j and A_j are constants described in *Subsection* 2. The result is that as first N and then K tend to infinity,

$$\operatorname{var} \hat{F}(\theta) \sim CK^2 N^{-1} \cdot 16a^3 \cos^2\theta \, (\sin\theta)^{-1} P^2 (a \sin\theta) \tag{4}$$

for $0 < \theta \leq \pi/2$, and uniformly for θ bounded away from zero;

$$\operatorname{var} \hat{F}(0) \sim DK^3 N^{-1} \cdot 16a^4 \pi^2 P^2(0) \tag{5}$$

and finally,

$$\operatorname{var} \hat{F}(\theta) \leq DK^3 N^{-1} \cdot 16a^4 \pi^2 \cos^2\theta \cdot \max_{0 \leq x \leq 1} P^2(x) \tag{6}$$

uniformly for $0 \leq \theta \leq \pi/2$. [Here C and D are constants describing the asymptotic behavior of w_j; see Eqs. (13) and (14) below.]

2. Definitions and Assumptions

Let $\{x_j\}$ $(j = 1, 2, \cdots)$ be a real-valued stationary gaussian process, with $E(x_j) = 0$, $\operatorname{var}(x_j) = 1$. Let

$$r_j = r_{-j} = \operatorname{cov}(x_n, x_{n+j}) \tag{7}$$

We further assume that the spectral density $P(f)$, which is an even function of f, is continuous on the closed interval $(-1, 1)$.

In terms of the covariances,

$$P(f) = 1 + 2 \sum_{j=1}^{\infty} r_j \cos(\pi j f) \tag{8}$$

Differentiating and inserting in Eq. (2) yields

$$F(\theta) = 4a^2 \cos\theta \sum_{j=1}^{\infty} A_j r_j \tag{9}$$

where

$$A_j = j \int_{a \sin\theta}^{a} \sin(\pi j f) \cdot (f^2 - a^2 \sin^2\theta)^{-1/2} \, df \tag{10}$$

Inserting the estimates

$$\hat{r}_j = N^{-1} \sum_{n=1}^{N} x_n x_{n+j} \tag{11}$$

in Eq. (9) and truncating the sum in Eq. (9) leads to the estimate

$$\hat{F}(\theta) = 4a^2 \cos\theta \sum_{j=1}^{K} A_j \hat{r}_j \tag{12}$$

Equation (3) is more general than Eq. (12) because of the introduction of weight factors w_j. We shall regard $\{w_j\}$ as an infinite sequence of real numbers in which the terms may depend on K and are zero after the Kth term. We make four assumptions about the weight factors:

(1) $w_1 = 1$ for all K

(2) $\{w_1, \cdots, w_K\}$ is a nonincreasing sequence of nonnegative real numbers for each K

(3) $$\lim_{K \to \infty} K^{-2} \sum_{j=1}^{K} j w_j^2 = C \tag{13}$$

(4) $$\lim_{K \to \infty} K^{-3} \sum_{j=1}^{K} j^2 w_j^2 = D \tag{14}$$

Here C and D are positive constants.

Note, for example, that $w_1 = \cdots = w_K = 1$, and also $w_j = (K - j + 1)/K$, $(j = 1, \cdots, K)$, satisfy these assumptions; in the first case, Eq. (3) becomes Eq. (12), and in the second case, Eq. (3) is the arithmetic mean of the $\hat{r}_j$ and thus tends to the Cesàro sum of the $\hat{r}_j$.

3. Estimate of the Variance

In the language of Toeplitz matrices (Ref. 2),

$$\hat{F}(\theta) = (N + K)^{-1} \times Wx^T \tag{15}$$

where x represents the vector $(x_1, \cdots, x_{N+K})$ and W is the $(N + K) \times (N + K)$ Toeplitz matrix given by

$$W_{kl} = (N + K) N^{-1} \cdot 4a^2 \cos\theta \cdot w_{l-k} A_{l-k} \tag{16}$$

(Here we take $w_{-j} = w_j$ and $A_{-j} = A_j$.)

We can express W in terms of its "Toeplitz kernel" $w(\lambda)$ (see Ref. 2, pp. 16–19) as follows: If

$$w(\lambda) = (N + K) N^{-1} \cdot 4a^2 \cos\theta \sum_{k=-\infty}^{\infty} w_k A_k e^{ik\lambda} \tag{17}$$

then

$$W_{kl} = (2\pi)^{-1}\int_{-\pi}^{\pi} e^{i(l-k)\lambda}\, w(\lambda)\, d\lambda \tag{18}$$

Applying the results in Ref. 2, pp. 217–218, we obtain

$$\operatorname{var}\hat{F}(\theta) = 2\sum_{j=1}^{N+K} \lambda_j^2 \tag{19}$$

where $\lambda_1, \cdots, \lambda_{N+K}$ are the eigenvalues of $(N + K)^{-1} RW$, R being the covariance matrix given by $R_{kl} = r_{l-k}$.

Now using the results in Ref. 2, pp. 219–220, we find

$$(N+K)^{-1}\sum_{j=1}^{N+K} \lambda_j^2 \sim (2\pi)^{-1}(N+K)^{-2}\int_{-\pi}^{\pi}[r(x)\,w(x)]^2\,dx \tag{20}$$

as $N \to \infty$, where

$$r(x) = \sum_{k=-\infty}^{\infty} r_k e^{ikx} = P\frac{x}{\pi} \tag{21}$$

Combining Eqs. (19), (20), and (21) gives

$$\operatorname{var}\hat{F}(\theta) \sim [\pi(N+K)]^{-1}\int_{-\pi}^{\pi}\left[P\left(\frac{y}{\pi}\right)w(y)\right]^2 dy \tag{22}$$

If we define

$$A(x) = \sum_{j=1}^{\infty} w_j A_j \cos(\pi j x) = \sum_{j=1}^{K} w_j A_j \cos(\pi j x) \tag{23}$$

and replace y by πx in Eq. (22), we obtain

$$\operatorname{var}\hat{F}(\theta) \sim 64N^{-1}\cdot a^4\cos^2\theta\int_{-1}^{1} P^2(x)\,A^2(x)\,dx \tag{24}$$

or, since P and A are even functions,

$$\operatorname{var}\hat{F}(\theta) \sim 128N^{-1}a^4\cos^2\theta\int_{0}^{1} P^2(x)\,A^2(x)\,dx \tag{25}$$

The results of Eqs. (4), (5), and (6) will be obtained from Eq. (25) by asymptotic evaluation of the integral in that expression.

4. Proof of Eq. (4)

We suppose θ is a nonzero angle. Substituting $f = a \sin\theta \sec x$ in Eq. (10) gives

$$A_j = j\int_0^{(\pi/2)-\theta} \sin(\pi j a \sin\theta \sec x)\sec x\,dx \tag{26}$$

Denote the zeroth-order Bessel functions of the first and second kinds by $J_0(z)$ and $Y_0(z)$, respectively. Then Ref. 3, p. 30, Eq. (5) gives

$$J_0(z) + iY_0(z) = -\pi^{-1}\cdot 2i\int_0^{\pi/2} e^{iz\sec x}\sec x\,dx \tag{27}$$

From this, we get

$$A_j = j\left(\frac{\pi}{2}\right)J_0(\pi j a\sin\theta) - j\int_{(\pi/2)-\theta}^{\pi/2}\sin(\pi j a\sin\theta\sec x)\sec x\,dx \tag{28}$$

Integration by parts gives

$$\int_{(\pi/2)-\theta}^{\pi/2}\sin(\pi j a\sin\theta\sec x)\sec x\,dx = (\pi j a\cos\theta)^{-1}\cos(\pi j a) + (\pi j a\sin\theta)^{-1}\int_{(\pi/2)-\theta}^{\pi/2}\cos(\pi j a\sin\theta\sec y)\csc^2 y\,dy \tag{29}$$

Furthermore, as $k \to \infty$,

$$J_0(\pi k a\sin\theta)\cos(\pi k x) = \left(\frac{2}{\pi^2 ka\sin\theta}\right)^{1/2}\left[\cos(\pi k x)\cos\left(\pi k a\sin\theta - \frac{\pi}{4}\right) + O(k^{-1})\right] \tag{30}$$

[Ref. 4, p. 364, Eq. (9.2.1)]. Hence, since $A(x) \to \infty$ as $K \to \infty$, we can write

with

$$\left.\begin{aligned} A(x) &\sim f(x) + g(x) + h(x) \\ f(x) &= (2a\sin\theta)^{-1/2} \sum_{j=1}^{K} w_j\,(j)^{1/2} \cos(\pi j x) \cos\left(\pi j a \sin\theta - \frac{\pi}{4}\right) \\ g(x) &= -(\pi a\cos\theta)^{-1} \sum_{j=1}^{K} w_j \cos(\pi j x)\cos(\pi j a) \\ h(x) &= -(\pi a \sin\theta)^{-1} \sum_{j=1}^{K} w_j \cos(\pi j x) \int_{(\pi/2)-\theta}^{\pi/2} \cos(\pi j a \sin\theta \sec y)\csc^2 y\, dy \end{aligned}\right\} \qquad (31)$$

Now using

$$\int_0^1 \cos(\pi j x)\cos(\pi l x)\, dx = \frac{1}{2}\delta_{jl} \qquad (32)$$

we see that

$$\int_0^1 A^2(x)\, dx = (4a\sin\theta)^{-1} \sum_{j=1}^{K} j w_j^2 \cos^2\left(\pi j a \sin\theta - \frac{\pi}{4}\right) + o\left(\sum_{j=1}^{K} j w_j^2\right) \qquad (33)$$

And, since $\cos^2\alpha = (1 + \cos 2\alpha)/2$ and

$$\sum_{j=1}^{K} j w_j^2 \sim CK^2$$

we have

$$\int_0^1 A^2(x)\, dx \sim \frac{CK^2}{8a\sin\theta} \qquad (34)$$

Now Eq. (3) will follow from Eq. (25), Eq. (34), and

$$\int_0^1 P^2(x) A^2(x)\, dx \sim P^2(a\sin\theta) \int_0^1 A^2(x)\, dx \qquad (35)$$

On account of Eq. (34), we can prove Eq. (35) by showing

$$K^{-2} \int_0^1 A^2(x)\,[P^2(x) - P^2(a\sin\theta)]\, dx \to 0 \qquad (36)$$

Now define

$$\left.\begin{aligned} v_i &= w_i - w_{i+1}, \qquad i = 1, \cdots, K-1 \\ v_K &= w_K \end{aligned}\right. \qquad (37)$$

The v_i are non-negative, and

$$w_j = \sum_{i=j}^{K} v_i, \qquad j = 1, \cdots, K \qquad (38)$$

Inserting this into Eq. (23) gives

$$A(x) = \sum_{i=1}^{K} v_i A_i^*(x) \qquad (39)$$

where $A_i^*(x)$ is the same as $A(x)$, except that the w_j do not appear, and summation extends to i instead of to K.

Thus, Eq. (36), which is what we are trying to prove, can be rewritten

$$K^{-2}\sum_{j=1}^{K}\sum_{l=1}^{K} v_j v_l \int_0^1 A_j^*(x)\,A_l^*(x)\,[P^2(x) - P^2(a\sin\theta)]\,dx \rightarrow 0 \tag{40}$$

Schwarz' inequality implies that Eq. (40) will follow from

$$K^{-1}\sum_{j=1}^{K} v_j \left[\int_0^1 A_j^*(x)^2\,|P^2(x) - P^2(a\sin\theta)|\,dx\right]^{1/2} \rightarrow 0 \tag{41}$$

And for this, it is sufficient to prove

$$K^{-2}\int_0^1 A_K^*(x)^2\,|P^2(x) - P^2(a\sin\theta)|\,dx \rightarrow 0 \tag{42}$$

We thus complete the proof of Eq. (4) by showing Eq. (42).

We have the following four relations: for $0 \leq x \leq 1$,

$$\sum_{j=1}^{K}\cos(\pi j x) = \begin{cases} K, \text{if } x = 0 \\ O(1), \text{uniformly for } x \text{ bounded away from } 0 \end{cases} \tag{43}$$

$$\sum_{j=1}^{K}(j)^{1/2}\cos(\pi j x) = \begin{cases} \frac{2}{3}K^{3/2} + o(K^{3/2}), \text{if } x = 0 \\ O(K^{1/2}), \text{uniformly for } x \text{ bounded away from } 0 \end{cases} \tag{44}$$

$$\sum_{j=1}^{K}(j)^{1/2}\sin(\pi j x) = \begin{cases} 0, \text{if } x = 0 \\ O(K^{1/2}), \text{uniformly for } x \text{ bounded away from } 0 \end{cases} \tag{45}$$

$$\int_{(\pi/2)-\theta}^{\pi/2}\cos(\pi j a\sin\theta\sec y)\csc^2 y\,dy = O(j^{-1}) \tag{46}$$

(Eq. 46 may be seen using integration by parts.) From these, it follows that

$$\left.\begin{aligned} &A_K^*(a\sin\theta) \sim f(a\sin\theta) \sim \left(\frac{1}{6}\right)(a\sin\theta)^{-1/2}K^{3/2} \\ &A_K^*(a) \sim g(a) \sim (2\pi a\cos\theta)^{-1}K \\ &A_K^*(x) = O(K^{1/2}) \text{ uniformly for } x \text{ bounded away from} \\ &\qquad a\sin\theta \text{ and } a \end{aligned}\right\} \tag{47}$$

Now let $\epsilon > 0$ be arbitrary and choose $\delta > 0$ small enough that $|P^2(x) - P^2(a\sin\theta)| < \epsilon$ when $|x - a\sin\theta| < \delta$. Then write the integral in Eq. (42) as the sum of integrals over the regions

$$(0, a\sin\theta - \delta), \qquad (a\sin\theta - \delta, a\sin\theta + \delta), \qquad (a\sin\theta + \delta, a - \delta), \qquad (a - \delta, a + \delta), \qquad (a + \delta, 1)$$

Examination of each of the five integrals separately reveals that the limit of the left side of Eq. (42) is less than ϵ. This completes the proof of Eq. (4).

5. Proof of Eqs. (5) and (6)

When $\theta = 0$, Eq. (10) becomes

$$A_j = j\int_0^a f^{-1}\sin(\pi j f)\,df \tag{48}$$

Writing this as the integral from zero to infinity minus the integral from a to infinity, and using integration by parts on the latter integral, we obtain

$$A_j = \frac{j\pi}{2} + O(1) \tag{49}$$

Hence,

$$A(x) \sim \left(\frac{\pi}{2}\right)\sum_{j=1}^{K} jw_j\cos(\pi jx) \tag{50}$$

and

$$\int_0^1 A^2(x)\,dx = \left(\frac{1}{2}\right)\sum_{j=1}^{K} w_j^2 A_j^2\, DK^3\cdot\frac{\pi^2}{8} \tag{51}$$

So to prove Eq. (5), it suffices to show

$$K^{-3}\int_0^1 A^2(x)\,|P^2(x) - P^2(0)|\,dx \rightarrow 0 \tag{52}$$

Using the same argument as above to dispose of the w_j, we find that it suffices to prove

$$K^{-3}\int_0^1 A_K^*(x)^2\,|P^2(x) - P^2(0)|\,dx \rightarrow 0 \tag{53}$$

We have

$$A_K^*(x) \sim \left(\frac{\pi}{2}\right)\sum_{j=1}^{K} j\cos(\pi jx) = \begin{cases} \dfrac{\pi K^2}{4}, \text{if } x = 0 \\ O(K), \text{uniformly for } x \text{ bounded away from } 0 \end{cases} \tag{54}$$

Again let $\epsilon > 0$ be arbitrary, choose δ so that

$$|P^2(x) - P^2(0)| < \epsilon$$

when $|x| < \delta$, and examine Eq. (53) as the sum of integrals over $(0,\delta)$ and $(\delta,1)$. The limit of the left side of Eq. (53) is thus seen to be less than ϵ; this completes the proof of Eq. (5).

To prove Eq. (6), note that by Eq. (28) we have

$$A_j \leq \frac{j\pi}{2} \tag{55}$$

asymptotically and uniformly in θ, since the integral in Eq. (28) is $O(1)$. Hence,

$$\int_0^1 A^2(x)\,dx = \left(\frac{1}{2}\right)\sum_{j=1}^{K} w_j^2 A_j^2 \leq DK^3\cdot\frac{\pi^2}{8} \tag{56}$$

Thus,

$$\int_0^1 P^2(x)\,A^2(x)\,dx \leq \max_{0\leq x\leq 1} P^2(x)\cdot DK^3\frac{\pi^2}{8} \tag{57}$$

and Eq. (6) follows from Eqs. (57) and (25).

6. Example

To illustrate the estimates derived above, we use a Venus radar spectrogram obtained on September 30, 1967. On that date, the round-trip time of a radar signal from Venus was 398 s. Five round-trip runs were made at a sampling rate of 235.8 samples/s: a total of $N = 469{,}242$ observations. From these, $K = 64$ estimated correlations $\hat{r}_j$ were computed, and the backscatter function $F(\theta)$ was estimated for $0 \leq \theta \leq \pi/2$. The weight factors used were the so-called "hanning window"

$$\omega_j = \frac{1}{2} + \frac{1}{2}\cos\frac{j\pi}{N+1}$$

Evaluating Eqs. (13) and (14) for these ω_j gives

$$C = \frac{3}{16} - \frac{1}{\pi^2} \doteq 0.0862$$

$$D = \frac{3}{24} - \frac{1}{\pi^2} \doteq 0.0237$$

The bandwidth of the spectrum is 34 cycles/s; expressed in terms of the sampling rate, we get a rotation constant of $a = 34/235.8 \doteq 0.1442$.

The backscatter function was estimated for values of θ between 0 and $\pi/2$ in increments of $\pi/128$. We have computed the values of the estimates in Eqs. (4) and (6) for these values, using of course Eq. (5) instead of Eq. (4) for $\theta = 0$.

It was found that for $\theta = \pi/128$, the estimate in Eq. (6) is smaller than that of Eq. (4); for all other nonzero values considered, Eq. (4) is the better estimate.

Table 1 shows values of $\hat{F}(\theta)$ and $\sigma(\theta) = [\text{var}\,\hat{F}(\theta)]^{1/2}$ for some of the above-mentioned values of θ. The function $\sigma(\theta)$ is taken from Eq. (5) in the case of $\theta = 0$, and from Eq. (4) in all the other cases shown.

Figure 6 is a graph of the three functions $\hat{F}(\theta)$ and $\hat{F}(\theta) \pm \sigma(\theta)$ versus the angle θ.

References

1. Goldstein, R. M., *A Radar Study of Venus*, Technical Report 32-280. Jet Propulsion Laboratory, Pasadena, Calif., May 25, 1962.
2. Grenander, U., and Szego, G., *Toeplitz Forms and their Applications*. University of California Press, Berkeley, Calif., 1958.
3. Luke, Y., *Integrals of Bessel Functions*. McGraw-Hill Book Co., New York, 1962.
4. *Handbook of Mathematical Functions*. Edited by M. Abramowitz and I. A. Stegun, National Bureau of Standards, Washington, D.C., 1964.

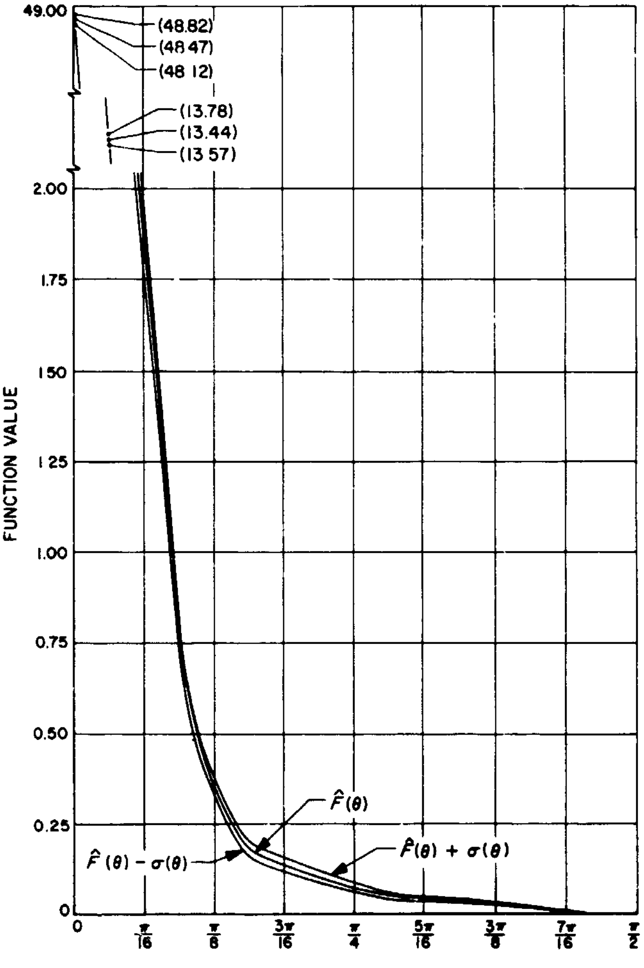


Fig. 6. Experimental backscatter functions vs angle θ

Table 1. Experimental backscatter function values for values of angle θ

θ, multiples of $\pi/32$	$\hat{F}(\theta)$	$\sigma(\theta)$	$\sigma(\theta)$, % of $\hat{F}(\theta)$
0	48.47	0.3495	0.72
1	13.78	0.2051	1.5
2	1.867	0.08534	4.6
3	0.8086	0.04295	5.3
4	0.3628	0.02899	8.0
5	0.1794	0.02206	12.3
6	0.1374	0.01752	12.8
7	0.1126	0.01419	12.6
8	0.07880	0.01160	14.7
9	0.05281	0.009507	18.0
10	0.04569	0.007764	17.0
11	0.03964	0.006252	15.8
12	0.03232	0.004886	15.1
13	0.02'02	0.003609	17.2
14	0.01074	0.002382	22.2
15	0.004188	0.001185	28.3
16	0.000000	0.000000	—

E. Communications Systems Development: Design of One- and Two-Way High-Rate Block-Coded Telemetry Systems, *W. C. Lindsey*

1. Introduction

Previous work (Refs. 1–4) has established performance characteristics and trends required for the design of one-way and two-way, phase coherent, uncoded communications systems. More recently, considerable interest has developed (SPS 37-48, Vol. II, pp. 83–91) in applying known techniques and theories, evolved over the past few years, to the mechanization of block-coded communications systems for deep space applications. Such words as "high-rate telemetry (HRT)," implying data rates in excess of a few thousand bits per second, and "system software" are becoming a part of the vocabulary of every communications design engineer faced with advancing the technology of deep space communications. For example, a major objective of the *Mariner* Mars 1969 missions is to obtain television pictures of Mars by applying the theory of block coding to the development of a 16,200-bit/s telemetry system. The HRT system is a modification of the basic digital telemetry system used on *Mariners IV* and V. The primary difference is that the data detection process is more efficient.

Discussed here is the performance of one-way and two-way phase-coherent communication systems which employ double-conversion superheterodyne phase-locked receivers preceded by a bandpass limiter to track the modulation. Such a setup is useful in testing, predicting performance, and evaluating the design of such systems prior to and after launch. The notation and terms used herein are those established in Refs. 1–4.

2. System Model

Before we proceed with the analysis, a functional description of the system illustrated in Figs. 7 and 8 will be given. Briefly, the data to be transmitted is assumed to be block-encoded into binary symbols. Each code word, say $x_l(t), l = 1, \cdots, N$, to be transmitted, is made comma-free (Ref. 5) by adding an appropriate comma-free vector

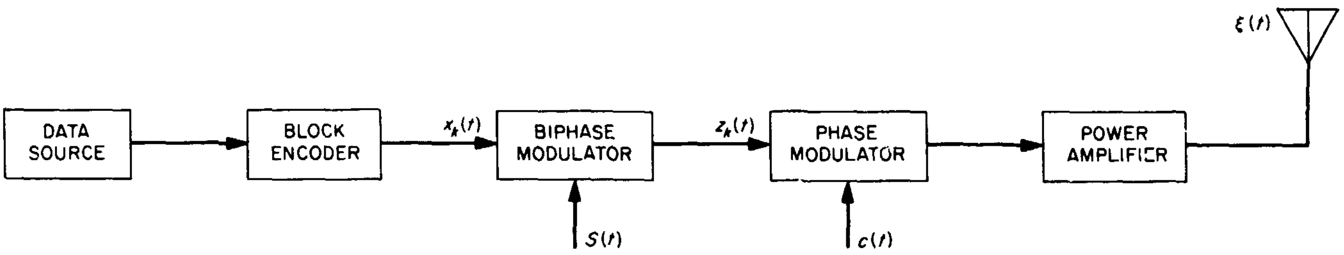


Fig. 7. Transmitter characterization

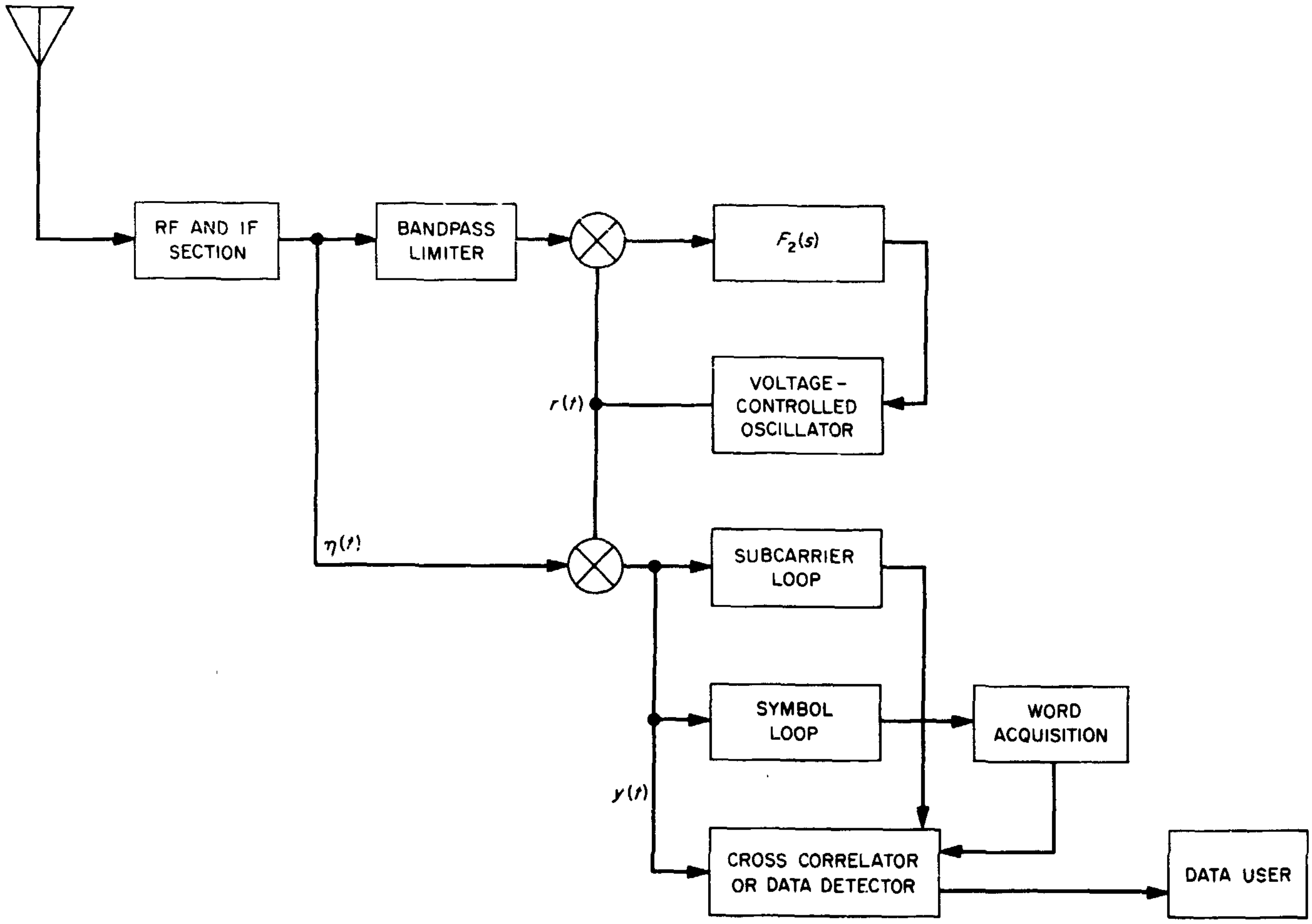


Fig. 8. Receiver characterization

to facilitate word synchronization at the receiver. The code symbols, appearing at the modulator in the form of a binary waveform, are used to biphase-modulate a square-wave data subcarrier (Ref. 3), say $S(t)$. The modulated data subcarrier (Refs. 1 and 2), in turn, phase-modulates the RF carrier $c(t)$, which is then amplified and radiated from the spacecraft or vehicle antenna (Ref. 3) as $\xi(t)$.

On the ground, a double-conversion superheterodyne phase tracking receiver is used to track the observed RF carrier component, thus providing a coherent reference for synchronously demodulating the subcarrier. The received signal is denoted by $\eta(t)$; see Fig. 8. Due to the fact that this reference is derived in the presence of white gaussian noise, a single-sided spectral density N_{02} watts per cycle per second, there will exist phase jitter due to the additive noise on the down-link (Refs. 1–3) and, if the system happens to be two-way locked (Refs. 1 and 2), the additive white noise, which is assumed to be white gaussian noise with single-sided spectral density of N_{01} watts per cycle per second on the up-link, also exerts another component of phase jitter.

In the following discussion, we shall be concerned with predicting system performance in both situations. The results are extremely useful in designing systems which must operate with narrow performance margins (margin denoting the number of decibels in excess of the sum of the negative tolerances in equipment performance). For deep space telecommunication links, the sum of the negative tolerances is typically 4 to 6 dB. Experience has shown that requiring the design to exceed the sum of the negative tolerances is slightly conservative; hence, reducing excess margin results in a much "tighter" or a less conservative design.

At the receiver (Fig. 8) a subcarrier tracking loop (Ref. 3) is assumed to exist for the purposes of providing subcarrier sync. In practice, phase jitter also exists on this reference; however, this phase jitter may usually be made negligibly small by designing a very narrowband subcarrier tracking loop (Ref. 3). Finally, word sync can be derived at the receiver by making use of the comma-free properties of the transmitted code (Ref. 5). Thus, the necessary timing information is provided for triggering the cross-correlation detector in Fig. 8. The output data is the recovered bit stream and may be recorded for the data user.

We assume that the code words, $x_l(t)$, $l = 1, 2, \cdots, N$ representing sequences of ±1's, occur with equal probability, contain equal energies, and exist for $T = kT_b = 2^kT_s$ seconds. Here, T_b is the time per bit, the reciprocal of the data rate $\mathcal{R}$, T_s is the time per code word symbol, and n is the number of kits per code word. Thus, the transmitted waveform may be represented by

$$\xi(t) = (2P)^{1/2} \sin[\omega t + (\cos^{-1} m) z_l(t)] \tag{1}$$

where P is the total radiated power, and m is the modulation factor which apportions the total power between the carrier component and modulation sidebands. In Eq. (1), the waveform $z_l(t) = x_l(t) S(t)$, $l = 1, 2, \cdots, N$, where $x_l(t)$ is the code word, in the form of a sequence of ±1's to be transmitted, and $S(t)$ is the unmodulated data subcarrier possessing unit power (Fig. 7). Since $S(t)$ is a sequence of ±1's, $z_l(t)$ is also a sequence of ±1's.

Assuming that the channel introduces an arbitrary (but unknown) phase shift θ to $\xi(t)$ and further disturbs $\xi(t)$ by additive white gaussian noise $n_2(t)$ of single-sided spectral density of N_{02} watts per cycle single-sided, one observes at the input to the receiver (Fig. 8)

$$\eta(t) = (2P)^{1/2} \sin[\omega t + (\cos^{-1} m) z_l(t) + \theta] + n_2(t) \tag{2}$$

when operating in a one-way locked condition (Ref. 1). If the receiver is operating in a two-way locked condition (Ref. 1), then the input to the receiver of Fig. 8 is taken to be

$$\eta(t) = (2P)^{1/2} \sin[\omega t + (\cos^{-1} m) z_l(t) + \hat{\theta}_1 + \theta] + n_2(t) \tag{3}$$

where $\hat{\theta}_1$ represents phase modulation due to the up-link additive noise (Ref. 1), i.e., noise introduced in the spacecraft transponder.

In either case, denote the output of the receiver's voltage-controlled oscillator by

$$r(t) = 2^{1/2} \cos[\omega t + \hat{\theta}_2] \tag{4}$$

where $\hat{\theta}_2$ is the estimate of the phase of the observed carrier component. Multiplying $\eta(t)$ by $r(t)$ and neglecting double frequency terms, it can be shown (Ref. 1) that the output $y(t)$ of the receiver's carrier tracking loop, which is the input to the data detector, is given by

$$y(t) = S^{1/2} z_l(t) \cos\phi + n_2'(t) \tag{5}$$

where $S = (1 - m^2)P$, $m^2 = P_c/P$, P_c is the power remaining in the carrier component at frequency $f = \omega/2\pi$, and ϕ is the receiver's phase error, i.e., $\phi = \theta - \hat{\theta}_2$ if one-way lock is assumed, and $\phi = \theta + \hat{\theta}_1 - \hat{\theta}_2$ if two-way lock is assumed. The probability distribution of the phase error ϕ is important in determining overall system performance. In the next two subsections, we present a model for this distribution when bandpass limiters precede the carrier tracking loop.

3. Probability Distribution for the Phase Error

a. One-way link. To characterize the distribution $p_1(\phi)$ requires considerable elaboration (beyond the scope of this article) on the response (signal plus noise) of a phase-locked loop preceded by a bandpass limiter. However, the distribution may be modeled on the basis of experimental and theoretical evidence given in Refs. 6–8. From these references, the distribution for $p_1(\phi)$ is approximated in the region of interest by

$$p_1(\phi) = \frac{\exp[\rho_L \cos\phi]}{2\pi I_0(\rho_L)}, \qquad |\phi| < \pi \tag{6}$$

where

$$\rho_L = \frac{2P_c}{N_0 w_{L0}} \cdot \frac{1}{\Gamma}\left(\frac{1 + r_0}{1 + \frac{r_0}{\mu}}\right) \tag{7}$$

and the parameters w_{L0}, r_0, and μ are defined from the closed-loop transfer function $H_2(s)$ of the carrier tracking loop,

$$H_2(s) = \frac{1 + \left(\frac{r_0 + 1}{2w_{L0}}\right)s}{1 + \left(\frac{r_0 + 1}{2w_{L0}}\right)s + \frac{\mu}{r_0}\left(\frac{r_0 + 1}{2w_{L0}}\right)^2 s^2} \tag{8}$$

Here, μ is taken to be the ratio of the limiter suppression factor α_0 at the loop's design point (threshold) to the limiter suppression, say α, at any other point, i.e., $\mu = \alpha_0/\alpha$. This assumes that the filter in the carrier tracking loop is of the form (Fig. 8)

$$F_2(s) = \frac{1 + \tau_1 s}{1 + \tau_2 s} \tag{9}$$

in which case

$$r_0 = \frac{\alpha_0 K \tau_2^2}{\tau_1} \tag{10}$$

and K is the equivalent simple-loop gain (Ref. 6). The subscripts 0 refer to the values of the parameters at the loop design point. The parameter w_{L0} is defined by

$$w_{L0} = \frac{1 + r_0}{2\tau_2\left(1 + \frac{\tau_2}{r_0 \tau_1}\right)} \tag{11}$$

The loop bandwidths are conveniently defined by w_L and b_L through the relationship

$$w_L = 2b_L = \frac{1}{2\pi j}\int_{-j\infty}^{j\infty} |H_2(s)|^2\, ds \tag{12}$$

Substitution of Eq. (8) into Eq. (12) yields

$$w_L = w_{L0}\left[\frac{1 + \frac{r_0}{\mu}}{1 + r_0}\right] = 2b_L \tag{13}$$

The relation $w_{L0} = 2b_{L0}$ can be defined in a similar way. Thus, Eq. (13) becomes

$$2b_L = (2b_{L0})\left[\frac{1 + \frac{r_0}{\mu}}{1 + r_0}\right] \tag{14}$$

This is the usual definition of loop bandwidth employed by practicing engineers. The factor Γ is approximated (Ref. 6) by

$$\Gamma = \frac{1 + 0.345\rho_H}{0.862 + 0.690\rho_H} \tag{15}$$

where ρ_H is the signal-to-noise ratio at the output of the receiver's IF amplifier, i.e.,

$$\rho_H = \frac{2P_c}{N_0 w_H} \tag{16}$$

The parameter w_H is the two-sided bandwidth of the second IF amplifier in the double-heterodyne receiver. In one-sided bandwidth notation, $w_H = 2b_H$ and

$$\rho_H = \frac{P_c}{N_0 b_H} \tag{17}$$

The parameter ρ_H is also the signal-to-noise ratio at the input to the bandpass limiter.

The remaining parameter to define is the factor $\mu = \alpha_0/\alpha$. It can be shown that limiter suppression α is given by

$$\alpha = \left(\frac{\pi}{2}\right)^{1/2}\left(\frac{\rho_H}{2}\right)^{1/2}\exp\left(-\frac{\rho_H}{2}\right)\left[I_0\left(\frac{\rho_H}{2}\right) + I_1\left(\frac{\rho_H}{2}\right)\right] \tag{18}$$

where $I_m(z)$, $m = 1, 2$, is the modified Bessel function of argument z and order. To specify α_0, the parameter ρ_H is rewritten as follows:

$$\rho_H = \frac{P_c}{N_0 b_H} \cdot \frac{b_{L0}}{b_{L0}} = \frac{P_c}{N_0 b_{L0}} \cdot \frac{b_{L0}}{b_H} = zy \tag{19}$$

where

$$\left.\begin{aligned} z &= \frac{P_c}{N_0 b_{L0}} \\ y &= \frac{b_{L0}}{b_H} \end{aligned}\right\} \tag{20}$$

In practice, the parameters of the carrier tracking loop are specified at the loop design point or threshold. If the design point is defined as $z_0 = \gamma_0 =$ constant, then the parameter α_0 is given by

$$\alpha_0 = \left(\frac{\pi}{2}\right)^{1/2} \left(\frac{\gamma_0 y}{2}\right)^{1/2} \exp\left(-\frac{\gamma_0 y}{2}\right)\left[I_0\left(\frac{\gamma_0 y}{2}\right) + I_1\left(\frac{\gamma_0 y}{2}\right)\right] \tag{21}$$

Therefore, it is clear that system performance depends upon the choice of γ_0. In the Deep Space Network, this choice is usually $\gamma_0 = 2$ so that

$$z_0 = \frac{P_{c0}}{(kT^\circ)(b_{L0})} = 2 \tag{22}$$

or, equivalently,

$$\frac{P_{c0}}{(kT^\circ)(2b_{L0})} = 1$$

at the design point. Here $N_0 = kT^\circ$, k is Boltzmann's constant, and T° equals the system temperature in degrees Kelvin.

b. Two-way link. In order to characterize the probability distribution $p_2(\phi)$ for the phase error in a two-way link, one must consider the up-link parameters and the mechanization of the transponder in the spacecraft (Ref. 1). As before, the characterization of $p_2(\phi)$ requires considerable elaboration (beyond the scope of this article) on the response (signal plus noise) of phase-locked loops in cascade. Certain theoretical and computer simulation results (Ref. 9) are available for explaining the nonlinear behavior of loops in cascade. The characterization which follows is predicated upon the work reported in Ref. 9 and that contained in Ref. 1. In the following discussion, we introduce the following notation: a subscript "1" refers to up-link parameters and constants associated with the spacecraft transponder mechanization, while a subscript "2" refers to down-link parameters and to constants associated with the mechanization of the ground receiver.

The generic form discussed in Refs. 1 and 9 for $p_2(\phi)$ is given by

$$p_2(\phi) = \frac{I_\infty\left[\,|\rho_1 + \rho_2 \exp(j\phi)|\,\right]}{2\pi I_0(\rho_1)\, I_0(\rho_2)}, \qquad |\phi| \leq \pi \tag{23}$$

where the definitions of ρ_1 and ρ_2 follow. The parameter ρ_2 equal to ρ_L in Eq. (7) becomes, in the new notation,

$$\rho_2 = \frac{2P_{c2}}{N_{02} w_{20}} \cdot \frac{1}{\Gamma_2}\left(\frac{1 + r_{20}}{1 + \dfrac{r_{20}}{\mu_2}}\right) \tag{24}$$

and

$$r_{20} = \frac{\alpha_{02} K_2 \tau_{22}^2}{\tau_{12}}$$

where the zero subscripts refer to the parameters at the loop design point. The parameter w_{20} replaces the design point loop bandwidth w_{L0} in Eq. (11) and is defined by

$$w_{20} = \frac{1 + r_{20}}{2\tau_{22}\left(1 + \dfrac{\tau_{22}}{r_{20}\tau_{12}}\right)} = 2b_{20} \tag{25}$$

when the loop filters are of the form as given in Eq. (9) with τ_1 replaced by τ_{12} and τ_2 by τ_{22}. The parameter Γ_2 is defined in Eq. (15) by adding the subscript "2" to all symbols. Likewise, Eqs. (18) and (21) define the limiter suppression α_{22} and α_{02}, respectively, by adding the subscript "2" to all symbols and

$$z_2 = \frac{P_{c2}}{N_{02} b_{20}}, \qquad y_2 = \frac{b_{20}}{b_{H2}} \tag{26}$$

In Eq. (26), we have dropped the "L" subscript on b_{L0} and replaced it by "2." The remaining parameter to define is the variable ρ_1, which is given by (Ref. 1)

$$\rho_1 = \frac{2P_{c1}}{N_{01} w_{10}} \cdot \frac{1}{G^2 \Gamma_1 K(k_1, k_2, \beta)} \tag{27}$$

where G, the static phase gain of the spacecraft transponder, is determined by the ratio of the output frequency to the input carrier frequency. The limiter performance factor is defined in Eq. (15); however, the parameter ρ_{H1} is now defined by

$$\rho_{H1} = \frac{2P_{c1}}{N_{01} w_{H1}} = \frac{P_{c1}}{N_{01} b_{H1}}$$

where w_{H1} is the two-sided bandwidth of the second IF amplifier in the spacecraft receiver, and b_{H1} is the one-sided bandwidth. The function $K(k_1, k_2, \beta)$ is given by

$$K(k_1, k_2, \beta) = \frac{1}{r_{10} + 1}\left[\frac{k_1(2 + k_1) + 2(k_1 + k_2 + 2)(\beta + \beta^2) + k_2(2 + k_2)\beta^3}{k_1^2 + 2k_1\beta + 2(k_1 + k_2 - k_1k_2)\beta^2 + 2k_2\beta^3 + k_2^2\beta^4}\right] \tag{28}$$

where

$$k_n = \frac{2\mu_n}{r_{n0}}, \qquad r_{n0} = \frac{\alpha_{n0}K_n\tau_{2n}^2}{\tau_{1n}}$$

$$\mu_n = \frac{\alpha_{nn}}{\alpha_{0n}}$$

$$\beta = \frac{w_{10}}{w_{20}}\left(\frac{r_{20} + 1}{r_{10} + 1}\right)$$

$$w_{n0} = 2b_{n0} = \frac{1 + r_{n0}}{2\tau_{22}\left(1 + \dfrac{\tau_{22}}{r_{n0}\tau_{12}}\right)}$$

$$\rho_{Hn} = \frac{2P_{cn}}{N_{0n}w_{n0}}$$

$$\alpha_{nn} = \frac{\pi}{2}\left(\frac{\rho_{Hn}}{2}\right)\exp\left(-\frac{\rho_{Hn}}{2}\right)\left[I_0\left(\frac{\rho_{Hn}}{2}\right) + I_1\left(\frac{\rho_{Hn}}{2}\right)\right]$$

with $n = 1, 2$. Now α_{0n} is defined by either the design point in the carrier tracking loops of the transponder, $n = 1$, or ground receiver, $n = 2$, through

$$\alpha_{0n} = \left(\frac{\pi}{2}\right)^{1/2}\left(\frac{\gamma_{0n}y_n}{2}\right)^{1/2} \times \exp\left(-\frac{\gamma_{0n}y_n}{2}\right)\left[I_0\left(\frac{\gamma_{0n}y_n}{2}\right) + I_1\left(\frac{\gamma_{0n}y_n}{2}\right)\right]$$

4. System Performance

a. Conditional word-error probability. The problem of evaluating system performance is described as follows: The output of the carrier tracking loop is given by Eq. (5). For k-bit orthogonal codes, the optimum decoder consists of 2^k cross-correlators whose outputs $C(j)$, $j = 1, 2 \cdots 2^k$, are

$$C(j) = \int_0^{kT_b} y(t)\,x_j(t)\,dt \tag{29}$$

where T_b is the transmission time per information bit. Once the set $\{C(j)\}$ has been determined, the most probable transmitted word corresponds to that $x_j(t)$ for which $C(j)$ is greatest. The output of the decoder will be those k bits which, if encoded, would produce this $x_j(t)$.

Since 2^k cross-correlators are required to decode a k-bit orthogonal code, the complexity of the decoder becomes impractical for k of about 8 or greater. Also, the complexity of the decoder and the maximum bit rate at which the decoder will operate are major factors in the design of the decoder. This article does not outline or investigate techniques for reducing the decoder complexity or for increasing the maximum bit rate at which the decoder will operate. The interested reader is referred to material contained in Koerner (SPS 37-17, Vol. IV, pp. 71–73) and Green (SPS 37-39, Vol. IV, pp. 247–252).

The conditional probability of correct word detection, $P_c(\phi)$, is shown (Ref. 3) to be given by

$$P_c(k, \phi) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right) dx \times \left[\int_{x + A_n}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{y^2}{2}\right) dy\right]^{2^k - 1} \tag{30}$$

where

$$\left.\begin{aligned} A_n &= (2kR_n)^{1/2}\cos\phi \\ R_n &= \frac{S_nT_{bn}}{N_{0n}} = \frac{S_n}{N_{0n}\mathcal{R}_n} \end{aligned}\right\} \tag{31}$$

and $\mathcal{R}_n = T_{bn}k$ = number of bits per code word. The subscript $n = 1$ is for one-way lock, while $n = 2$ implies two-way lock.

For biorthogonal codes of k bits per word, the probability of correct reception of a word, conditioned upon a particular phase error, is given by Ref. 3 as

$$P_c(k, \phi) = \int_{-\infty}^{\infty} \frac{\exp\left(-\dfrac{x^2}{2}\right) dx}{(2\pi)^{1/2}} \times \left[\int_{x + A_n}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{y^2}{2}\right) dy\right]^{2^k - 1} \tag{32}$$

where A_n is defined in Eq. (31). The probability of a word error, conditioned upon a fixed value of ϕ, is, of course,

$$P_E(k,\phi) - 1 \cdot P_c(k,\phi) \qquad (33)$$

For convenience, when $n = 1$ we will drop the subscript on A.

b. Average word- and bit-error probability. To obtain the average word-error probability $P_E(k)$, one averages Eq. (33) over the phase-error distribution. Thus,

$$P_E(k) = 1 - \int_{-\pi}^{\pi} p_n(\phi) P_c(k,\phi)\, d\phi\,, \qquad n = 1, 2 \qquad (34)$$

where $p_1(\phi)$ is given in Eq. (6) for one-way lock, and $p_2(\phi)$ is defined in Eq. (23) for two-way lock. Substitution of Eqs (6) or (23) and (30) or (32) into Eq. (34) yields integrals which generally cannot be evaluated analytically; however, numerical integration by an IBM 7090 computer is possible.

In certain cases of practical interest, the bit-error probability is of importance. For k-bit orthogonal codes, the bit-error probability is (Ref. 10)

$$P_B(k) = \frac{2^{k-1}}{2^k - 1} P_E(k)$$

while for k-bit biorthogonal codes the total bit-error probability is (Ref. 10)

$$P_B(k) = P_1(k) + \frac{(k-1)\,2^{k-2}}{k\,(2^{k-1}-1)} P_2(k)$$

where $P_1(k)$ is given in Eq. (34) for orthogonal codes, and $P_2(k)$ is given by Eq. (34) for biorthogonal codes.

5. Design Results

Since the integrals in Eq. (34) cannot be evaluated numerically, integration by an IBM 7090 computer yielded the results, for one-way lock, illustrated in Fig. 9 for code words containing $k = 6$ bits of information. These figures depict word-error rates versus the signal-to-noise ratio in the data for various values of the signal-to-noise ratio x in the design point bandwidth of the carrier tracking loop. Clearly, system performance depends upon the choice of a design point γ_0 in the carrier tracking loop. For purposes of presentation, the choice is taken to be that which corresponds to the design point in the Deep Space Network, i.e., $r_0 = 2$, $\gamma_0 = 2$, and $y = 1/400$. Clearly, as x approaches infinity, i.e., the case of perfect RF sync, the deleterious effects of a noisy phase reference disappear and perfect coherent detection is possible.

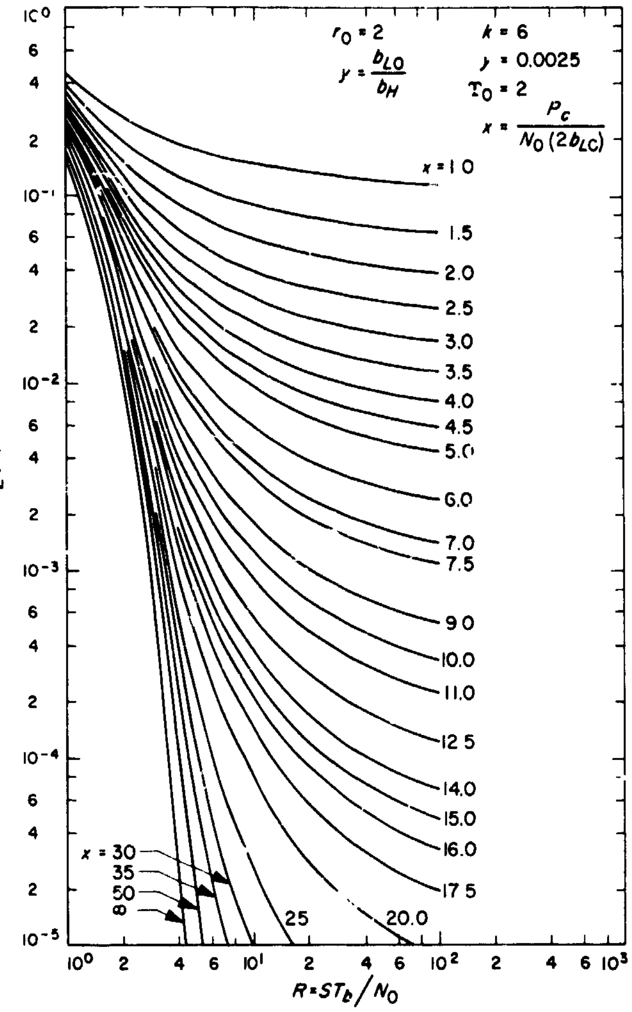


Fig. 9. Word-error probability vs signal-to-noise ratio R for various values of the signal-to-noise ratio x (k = 6, one-way)

In the case of two-way lock, system performance for $k = 6$ bit orthogonal codes is illustrated in Fig. 10 for $\rho_1 = 20$ and various values of x_2. The same carrier tracking loop design point is used for this case as was used for the one-way lock case. Notice that in this sequence of figures as the signal-to-noise ratio in the ground receiver's design point loop bandwidth, x, increases without limit, the deleterious effects of the up-link noise introduce an irreducible error probability. This irreducible error depends upon the amount of carrier phase jitter introduced by the vehicle's carrier tracking loop. This irreducible

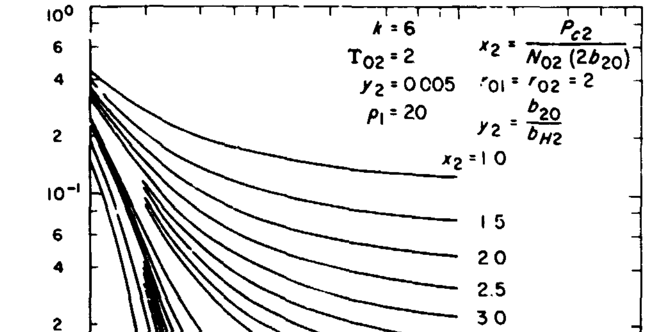


Fig. 10. Word-error probability vs signal-to-noise ratio R_2 for various values of the signal-to-noise ratio x_2 ($k = 6$, $\rho_1 = 20$, two-way)

error probability can be made arbitrarily small by increasing the up-link transmitter power. In fact, it is easy to show that the irreducible error probability, say $P_{ir}(k)$, is given by

$$P_{ir}(k) = 2\int_{\pi/2}^{\pi} p_n(\phi)\,d\phi$$

$$P_{ir}(k) = \lim_{R\to\infty} P_E(k) = 2\int_{\pi/2}^{\pi} p_n(\phi)\,d\phi$$

which is the probability that the phase error exceeds $\pi/2$, i.e., Prob [$|\phi| > \pi/2$]. This says that P_{ir} is independent of the code, and it depends only upon the design of the carrier tracking loops, the available power in the carrier components, and the channel noise. Thus, for given channel conditions and fixed loop parameters, large transmitter output power capability is certainly desirable.

For $k \geqq 5$, the performance of a block-coded digital communication system using biorthogonal codes is essentially the same as one that uses orthogonal codes (Ref. 10). Hence for $k \geqq 5$, the results presented can be applied to the design of systems whose code dictionaries are biorthogonal.

References

1. Lindsey, W. C., "Optimal Design of One-Way and Two-Way Coherent Communication Links," *IEEE Trans. Commun. Technol.*, Vol. COM-14, pp. 418–431, Aug. 1966.
2. Lindsey, W. C., "Determination of Modulation Indexes and Design of Two-Channel Coherent Communication Systems," *IEEE Trans. Commun. Technol.*, Vol. COM-14, pp. 229–237, Apr. 1967.
3. Lindsey, W. C., "Design of Block-Coded Communication Systems," *IEEE Trans. Commun. Technol.*, Vol. COM-15, No. 4, pp. 525–534, Aug. 1967.
4. Lindsey, W. C., *Performance of Phase Coherent Receivers Preceded by Bandpass Limiters*. Technical Report 32-1162, Jet Propulsion Laboratory, Pasadena, Calif., Sept. 15, 1967. Also to be published in *IEEE Trans. on Commun. Technol.*, 1968.
5. Stiffler, J. J., "Synchronization Methods for Block Codes," *IRE Trans. Inform. Theory*, Vol. IT-8, pp. S 25–S 34, Sept. 1962.
6. Tausworthe, R. C., *Theory and Practical Design of Phase-Locked Receivers*. Technical Report 32-819. Jet Propulsion Laboratory, Pasadena, Calif., Feb. 15, 1966.
7. Lindsey, W. C., and Charles, F. J., *A Model Distribution For The Phase Error in Second-Order Phase-Locked Loops*. Technical Report 32-1017. Jet Propulsion Laboratory, Pasadena, Calif., Oct. 31, 1966.
8. Charles, F. J., and Lindsey, W. C., "Some Analytical and Experimental Phase-Locked Loop Results For Low Signal-to-Noise Ratios," *Proc. IEEE*, Vol. 54, pp 1152–1166, Sept. 1966.
9. Lindsey, W. C., and Weber, L. C., "[illegible] the Theory of Automatic Phase Control," in *Stochastic Optimization and Control*. John Wiley and Sons, Inc., New York, 1968.
10. Golomb, S., *Digital Communications With Space Applications*. Prentice Hall, Inc., Englewood Cliffs, N. J., 1964.

F. Communications Systems Development: A Digital Demonstration of Sequential Decoding and Comparison With Block-Coded Systems, *P. Stanek*

1. Introduction

Sequential decoding of tree-coded data is theoretically a highly efficient scheme on a wide variety of channels. Specifically, both high information rates (bits per symbol) and high data rates (bits per second) may be achieved with low-error probabilities and modest equipment investment. Basic information on sequential decoding is contained in Ref. 1; the interrelations of the physical features of a theoretical communications system for the gaussian noise case are shown in Ref. 2. To determine the feasibility of sequential decoding, using a general-purpose digital computer in the role of decoder, and to discover realistic operating parameters for such a scheme, an extensive simulation was conducted using these theoretical techniques for the discrete memoryless case.

This article describes this simulation and compares sequential decoding with other schemes that might be applied to the same communications system. One such scheme is the maximum likelihood decoding of orthogonal and biorthogonal block codes at corresponding information rates and on a simulated channel model derived from a discrete time version of the gaussian channel within 5-dB of capacity. For this case, it is shown that sequential decoding exhibits an undetected bit-error probability at least several orders of magnitude less than that of these optimum block codes.

This advantage is partly offset in a real-time decoding system by the appearance of erasures in the output data at a rate entirely dependent on the decoder's speed compared to the data rate. It will be seen that such erasures may be recovered simply by increasing the decoder's speed, and that for a constant erasure rate, a speed increase of ten times allows a signal energy-to-noise ratio decrease of nominally 1 dB. Moreover, even if a reasonable and nearly optimum erasure strategy is adopted for block decoding, the undetected bit-error probability is a function of the block erasure probability and cannot be reduced to that observed for sequential decoding unless an erasure rate of 50% or more is allowed. The erasures from block decoding cannot be recovered without a corresponding increase in undetected bit-error probability.

2. Comparison of Optimum Systems

For an arbitrary, binary-input discrete memoryless channel with inputs ±1, outputs y_k, $1 \leqq k \leqq K$, and transition probabilities $p[y_k | \pm 1]$, a block code with t code words $X^1, \cdots, X^t$ of block length n can be found for which an optimum decoding procedure allowing erasures produces errors and erasures which satisfy

$$A2^{-nE^*(R)} \leqq p\,[\text{error}] \leqq A2^{-nE_*(R)} \qquad (1)$$

and

$$A^{-1}2^{\,nE^*(R)} \leqq p\,[\text{erasure}] \leqq A^{-1}2^{-nE_*(R)} \qquad (2)$$

An optimum decoder minimizes expected error probability given equally likely input probabilities for each code word. The probabilities in these inequalities are given per code word, and such a code has an information rate R, defined by $R = (\log t)/n$. The constant A does not depend on the code, and the exponent functions are defined by

$$\begin{aligned} E_*(R) &= \max_{0<\rho\leqq 1} \{E_0(\rho) - \rho R\}, \qquad R \geqq R_{crit} \\ &= E_0(1) - R, \qquad R \leqq R_{crit} \end{aligned} \qquad (3)$$

$$E^*(R) = \max_{0<\rho} \{E_0(\rho) - \rho R\} \qquad (4)$$

where

$$E_0(\rho) = -\log \sum_{k=1}^{K} \left[\frac{1}{2} p\,[y_k | 1]^{1/(1+\rho)} + \frac{1}{2} p\,[y_k | -1]^{1/(1+\rho)}\right]^{1+\rho} \qquad (5)$$

The number $E_0(1)$ is also called R_{comp} and the upper and lower bounds satisfy $E_*(R) \leqq E^*(R)$ with equality whenever $R \geqq R_{crit}$. That rate for which $E_*(R) = E^*(R) = 0$ is channel capacity. Further important rates are defined by $R_\rho = E_0(\rho)/\rho$. These exponents and their interpretation are thoroughly explained in Refs. 3 and 4. The lower bounds are "sphere-packing" and cannot be transgressed, while the upper bounds are obtained by random-coding arguments and merely assert the existence of codes with the prescribed error behavior without exhibiting any. The probabilities, p [error] and p [erasure], are code word error and erasure probabilities after symbol and word synchronization is achieved.

For the same channel, a tree code of constraint length ν bits and rate $R = 1/B$, for integer B, can be mechanized with ν shift registers and B adders (Ref. 2). Viterbi (Ref. 5) shows that, if an optimum decoding procedure is used, the error probability per bit in a sufficiently long tree code is bounded as

$$p\,[\text{error}] \geqq 2^{-\nu B E_0(\rho)} \qquad (6)$$

where

$$R = R_\rho \qquad (7)$$

Moreover, Yudkin (Ref. 6) has shown that, if the Fano algorithm (Ref. 2) is used for sequential decoding, the undetected bit-error probability is exponentially upper-bounded with the same exponent. Hence, if the information rate is R_{comp}, Fano's algorithm will produce fewer than $N 2^{-\nu}$ bit errors in decoding a tree of N bits (or NB channel symbols). For a fixed information rate and a fixed channel, Eqs. (1) and (2) show that performance of a block-coded system can be improved only by increasing block length, a situation which quickly leads to unacceptable impracticalities. For a tree-coded system at the same rate on the same channel, performance is improved, according to Eq. (6), by increasing constraint length. There is little difficulty in using long constraint lengths in practical decoders for tree-coded systems at high information rates. As an example, doubling the constraint length would increase the decoding program used in this simulation study by an average of six machine instructions executed per branch; whereas for fixed constraint length, lowering the information rate requires a significantly larger memory, but no increase in executed instructions for decoding.

As ρ increases, the rates R_ρ decrease. The exponent function $E^*(R)$ can be obtained as the convex hull of the family of straight lines of slope $-\rho$ which intercept the R-axis at R_ρ. At capacity, $\rho = 0$; at R_{comp}, $\rho = 1$. Geometrically, it can be seen that optimum tree codes enjoy a significant advantage over optimum block codes of the same rate, on the basis of error probability per bit. For example, if channel capacity is ½, an optimum block code of information rate R_{comp} and $n = 50$, an unbelievably enormous number for a practical system, has a bit-error rate greater than 2^{-25}, while an optimum tree code of the same rate with constraint length 35 and tree size 1024 bits, not at all unreasonable parameters, has a bit-error rate less than 2^{-25}.

3. The "Optimum" Decoders

In the event that code words are equally likely, a maximum likelihood decoder can be shown to be error-minimizing. For given code $X^1, \cdots, X^t$, each X^i of length n, a maximum likelihood decoder will produce the code word X when Y is received if

$$p\,[Y|X] > p\,[Y|X'] \text{ for all } X' \neq X \tag{8}$$

Instead of such a decoder, it will be convenient to consider one which, for fixed σ, produces X in case

$$p\,[Y|X] > p[Y|X']\,2^{\sigma} \text{ for all } X' \neq X, \sigma \geqq 0 \tag{9}$$

and which has no output if no X satisfies Eq. (9). This latter event is called *erasure* and the constant A of Eqs. (1) and (2) becomes then $2^{-\sigma\rho/(1+\rho)}$, where ρ is a solution of Eq. (7), and clearly depends on the decoding strategy. By using Eq. (9) instead of Eq. (8), extra reliability is gained by assuring that X is more likely than its nearest competitor by at least some fixed amount, predictable in advance.

Because of the various probability assumptions, Eq. (9) can be rewritten

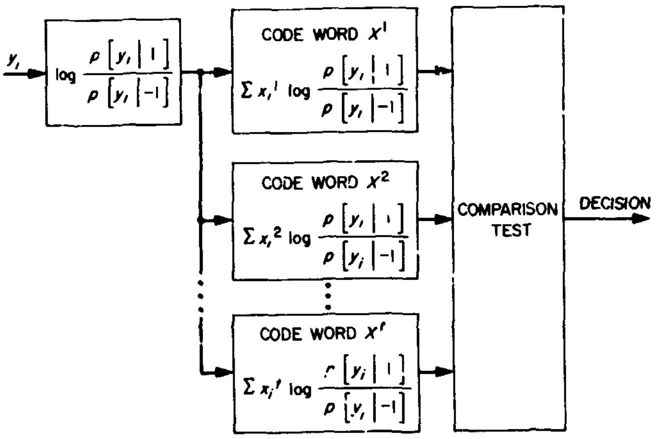
$$\sum_{i=1}^{n} \log \frac{p\,[y_i|x_i]}{p\,[y_i|x'_i]} > \sigma \text{ for all } X' \neq X, \sigma \geqq 0 \tag{10}$$

where y_i is the ith coordinate of Y, as are x_i and x'_i of X and X', respectively. Note that if $x_i = x'_i$, the ith summand is 0. Since the code words are written in the alphabet $\{\pm 1\}$, the decoding rule becomes

$$\sum_{x_i \neq x'_i} x_i \log \frac{p\,[y_i|1]}{p\,[y_i|-1]} > \sigma \text{ for all } X' \neq X, \sigma \geqq 0 \tag{11}$$

and a mechanization is visualized as in Fig. 11. The erasure parameter σ and the strategy in using it will be taken up later.

Two things are of importance here. First, according to Eqs. (1) and (2), the only way to improve error performance in the optimum case is to increase block size, reduce information rate, or increase the erasure threshold. The erasures produced by the decoder of Fig. 11 cannot be recovered by other changes within that system. Secondly, the box in that figure labeled "comparison test" contains a number of operations which are exponential in block size and such a decoder, even as a special-purpose device, will limit data rate. Hence, the optimum decoder becomes less than optimum in the ordinary meaning of that word.



y_i
$\log \frac{p[y_i|1]}{p[y_i|-1]}$
CODE WORD X^1
$\Sigma x_i^1 \log \frac{p[y_i|1]}{p[y_i|-1]}$
CODE WORD X^2
$\Sigma x_i^2 \log \frac{p[y_i|1]}{p[y_i|-1]}$
CODE WORD X^t
$\Sigma x_i^t \log \frac{p[y_i|1]}{p[y_i|-1]}$
COMPARISON TEST
DECISION

Fig. 11. Maximum likelihood decoder block diagram

To the extent, then, that practical considerations limit block size, information rate, and so on, a coding–decoding scheme which produces error and erasure statistics satisfying external requirements ought to be the design goal.

4. Theoretical Parameters of Sequential Decoding

The principal problem in designing a tree-coded sequentially-decoded communications system relates to the variable decoding time per bit (or block of bits) and the relative persistence of long tree searches. Berlekamp and Jacobs (Ref. 7) have analyzed this search problem for any sequential decoding algorithm and estimate the distribution function of the random variable C, which is the average number of branches examined in decoding a block, as

$$p\,[C > L] \geq DL^{-\rho} \tag{12}$$

where D is a constant depending on the decoder, and ρ is found again by solving Eq. (7). In deriving an expression such as Ineq. (12), it is typically assumed that the decoder has made no previous mistakes, that searching continues as long as required, and that no errors are committed. One branch examination is called a computation, and, since a computation time depends on the decoder's speed and the efficiency of its programming, the data rate must be chosen in light of this constraint.

To operate effectively, a sequential decoder must have a speed advantage in computation time over data rate; i.e., it must be able to search several branch paths to modest depth during a single bit time. To accommodate longer searches, an effective countermeasure to the variable decoding time is a temporary storage buffer for incoming channel symbols. This would allow considerable searching of likely paths to a significant depth, while the speed advantage would allow "catching-up" after a difficult portion of the correct path. This problem is discussed in Ref. 8. If the decoder lags behind incoming symbols further than buffer size while searching, subsequent incoming symbols are lost and decoding cannot proceed. By comparing Eq. (6) with Ineq. (12), it will be seen that this event, viz., buffer overflow, is far more likely to occur than an undetected bit error. Since buffer overflow terminates decoding and since such an event will eventually occur, some method of restarting decoding after overflow must be devised.

If the data stream is divided into blocks of fixed size and a known sequence inserted between blocks, then buffer overflow will terminate decoding only within the block in which overflow occurs and the decoder can be restarted at the beginning of the next block. The output of such a scheme would consist of decoded bits and occasional blocks of erasures. It is important to note that these erasures resulted from the inability of the decoder to search enough paths in time. Hence, if the incoming data were recorded for decoding later, much of the erased information could be recovered, nearly error free, by an off-line decoder which would be given as much search time as needed for decoding. No such comparable procedure is available for block decoding.

Constraint length for tree codes is defined as the smallest number ν such that two branch paths, anywhere in the tree, which have a segment of ν consecutive branches, anywhere on each path, corresponding to identical segments of ν information bits will encode subsequent bits identically. Therefore, a sequential decoding algorithm such as the Fano algorithm will commit an error if it accepts a wrong path as most likely for at least one constraint length. If this is the case, the decoder will produce, most likely, ν bits in error. By varying ν, two extreme situations become apparent. For very large ν, essentially no errors are made and the decoder performs as predicted by Ineq. (12). For small ν, errors are made, but the decoder accepts paths more freely and so decodes faster. This latter possibility can be used to advantage in the real-time scheme combined with off-line decoding of erasures, as outlined above. Again, the design criteria is a reasonable output data rate maintained at low-error probabilities.

5. A Channel Model

The discrete memoryless channel chosen for this simulation is the quantized version of the binary input, continuous output, additive normal noise channel. According to Ref. 1, for unquantized outputs such a channel can be modeled as a radio channel with inputs a function of time $s(t)$ given by

$$s(t) = \begin{cases} (2E_S)^{1/2}\cos\left(\omega_0 t + \dfrac{\pi}{2}\right) \text{for input symbol } 1 \\ (2E_S)^{1/2}\cos\left(\omega_0 t - \dfrac{\pi}{2}\right) \text{for input symbol } -1 \end{cases} \tag{13}$$

for $0 \leq t < \tau$, where τ is the symbol time and E_S is the received energy per symbol. On the additive normal noise channel with noise level N_0, the output can be taken to be

$$y = \pm\left(\frac{2E_S}{N_0}\right)^{1/2} + n \tag{14}$$

where n is normally distributed of zero mean and unit variance. For the quantized version, $K = 8$ is suggested in Ref. 2 (and is used in this simulation), and transition probabilities $p\,[y_k\,|\,1]$, $p\,[y_k\,|\,-1]$ for $1 \leq k \leq 8$ are given by

$$\left.\begin{aligned}
p\,[y_1\,|\,1] &= \int_{-\infty}^{-1.5} (2\pi)^{-1/2} \exp\left[-\frac{(x + 2E_S/N_0)^2}{2}\right] dx \\
p\,[y_k\,|\,1] &= \int_{-1.5+0.5(k-2)}^{-1.5+0.5(k-1)} (2\pi)^{-1/2} \exp\left[-\frac{(x + 2E_S/N_0)^2}{2}\right] dx, \qquad 2 \leq k \leq 7 \\
p\,[y_8\,|\,1] &= \int_{1.5}^{\infty} (2\pi)^{-1/2} \exp\left[-\frac{(x + 2E_S/N_0)^2}{2}\right] dx \\
p\,[y_k\,|\,-1] &= p\,[y_{9-k}\,|\,1], \qquad 1 \leq k \leq 8
\end{aligned}\right\} \tag{15}$$

The basic parameter of the continuous output version is the symbol energy-to-noise ratio E_S/N_0. The transition probabilities have been tabulated, and a listing for the range -5 to -1 appears in Table 2. For information rate R, the bit energy-to-noise ratio is given by

$$\frac{E_B}{N_0} = \frac{E_S}{RN_0} \tag{16}$$

and the parameter ρ is shown as a function of E_B/N_0 for fixed rates ½, ⅜, and ⅓ in Fig. 12, for transition probabilities given in Eq. (15).

Any digital communications scheme applied to the continuous model performs a quantization of some kind at some point. The point of view adopted here is that a quantization is performed on the channel symbols, and

Table 2. Transition probabilities for various signal-to-noise ratios per symbol

E_S/N_0, dB	$p\,[y_1\,\|\,1]$	$p\,[y_2\,\|\,1]$	$p\,[y_3\,\|\,1]$	$p\,[y_4\,\|\,1]$	$p\,[y_5\,\|\,1]$	$p\,[y_6\,\|\,1]$	$p\,[y_7\,\|\,1]$	$p\,[y_8\,\|\,1]$
−5.0	0.240	0.178	0.197	0.171	0.116	0.061	0.025	0.011
−4.8	0.246	0.180	0.197	0.169	0.113	0.060	0.025	0.010
−4.6	0.252	0.181	0.197	0.167	0.111	0.058	0.024	0.010
−4.4	0.259	0.183	0.196	0.165	0.109	0.056	0.023	0.009
−4.2	0.265	0.184	0.196	0.163	0.107	0.054	0.022	0.009
−4.0	0.272	0.185	0.195	0.161	0.104	0.053	0.021	0.008
−3.8	0.279	0.187	0.195	0.159	0.102	0.051	0.020	0.008
−3.6	0.286	0.188	0.194	0.157	0.099	0.049	0.019	0.007
−3.4	0.293	0.189	0.193	0.155	0.097	0.047	0.018	0.007
−3.2	0.301	0.190	0.192	0.152	0.094	0.046	0.017	0.007
−3.0	0.309	0.192	0.191	0.150	0.092	0.044	0.016	0.006
−2.8	0.317	0.193	0.190	0.147	0.089	0.042	0.016	0.006
−2.6	0.326	0.194	0.189	0.144	0.086	0.041	0.015	0.005
−2.4	0.335	0.194	0.188	0.142	0.084	0.039	0.014	0.005
−2.2	0.344	0.195	0.186	0.139	0.081	0.037	0.013	0.005
−2.0	0.353	0 196	0.184	0.136	0.078	0.035	0.013	0.004
−1.8	0.363	0.196	0.183	0.133	0.076	0.034	0.012	0.004
−1.6	0.373	0.197	0.181	0.130	0.073	0.032	0.011	0.004
−1.4	0.383	0.197	0.178	0.126	0.070	0.030	0.010	0.003
−1.2	0.394	0.197	0.176	0.123	0.067	0.029	0.010	0.003
−1.0	0.405	0.197	0.174	0.120	0.065	0.027	0.009	0.003

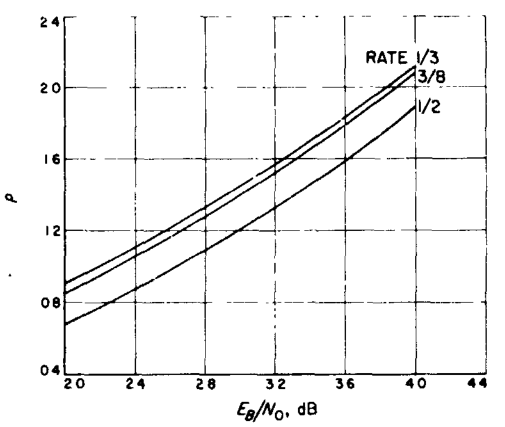


Fig. 12. Theoretical ρ parameter for 3-bit quantization

the decoder looks at a discrete memoryless channel. This practice minimizes the amount of special-purpose equipment between the antenna and decoder. While results and comparisons in this article are given on the basis of a theoretical E_B/N_0, they have been derived empirically from the discrete memoryless channel with the given channel transition probabilities. Further applications of this work to other communications schemes which are described in terms of energy-to-noise ratios are valid only to the extent that they represent a discrete channel with probabilities matching those listed in Table 2.

6. Simulation Results

In addition to the channel model, the important system parameters chosen for the study are:

Channel model	8-level quantized additive normal noise
Information rate	⅓ information bits per channel symbol
Constraint length	24 information bits
Block size	2048 information bits
Buffer size	512 information bits
Coding	systematic convolutional tree code

These parameters were chosen as a compromise between theoretical virtues of sequential decoding and conditions imposed by the digital computer (in this case, a 24-bit octal machine) and are somewhat variable, except for information rate. Vanous constraint lengths in multiples of 12 and various block sizes in multiples of 512 could be used. Just as a starting point, with these parameters undetected bit-error probability for transition probabilities, for which $R_1 = ⅓$, is 10^{-4}, theoretically. (From Fig. 12, $E_B/N_0 = 2.2$ dB in this case.)

Figure 13 is a block diagram of the receiving system envisioned, and it should be noted that energy-to-noise ratios are measured at the input to the decoder.

There are three timing problems to solve with this system, viz., symbol synchronizing, block synchronizing, and output data formatting. The symbol problem is solved by the receiver which generates a timing pulse to interrupt the computer, causing it to process the next received symbol from the converter.

In order to provide a single-channel capability, a block synchronization technique was added to the decoding program. Each tree path ends in a known sequence of 24 bits, i.e., one constraint length. At the end of a tree, this ensures that the last few bits before this sequence will be decoded properly. For a systematic code, then, of the last 72 symbols, every third one is known. In addition to these, the initial 8 information bits of the next block were also fixed in advance, and consequently 48 channel symbols are known *a priori* at the end and beginning of consecutive blocks. Synchronizing is achieved by searching for this pattern in the incoming symbol stream by requiring good correlation on any multiple of the 48 channel symbols. Once block synchronization is declared, decoding begins at the correct place. Following an overflow. the decoder moves ahead to the start of the next block according to its previous reference. If overflows occur in consecutive blocks, for example in four or five consecutive blocks, it becomes reasonable to declare a system failure and the program returns to the block synchronizing mode.

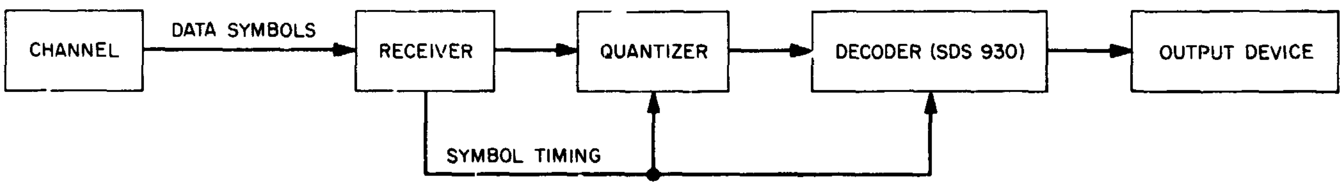


Fig. 13. Receiving system block diagram

Block synchronizing could also be derived from the decoder alone by the following argument. If the decoder begins anywhere in the symbol stream with an incorrect block reference, it will not find the correct path because there isn't one, and so will overflow and not decode. Said contrapositively, decoding implies synchronizing. The converse is not true since the decoder will sometimes overflow even with correct block reference. The situation is different in the case of optimum block coding in which the decoder is in reality a block synchronizer and synchronizing implies decoding.

Figure 14 shows theoretical undetected bit-error probabilities for constraint lengths 24, 36, and 48, and observed error probabilities for length 24. No errors were observed for 36 and 48. This part of the experiment observed bit errors in the decoder output when the decoder is allowed as long as required to decode. Undetected errors, that is, errors produced by the decoder to the output device, were observed at 2 energy levels, E_B/N_0 of 1.2 and 2.2. At 1.2 dB, the bit-error rate was nearly 0.5, and at 2.2 dB it was less than 10^{-4}. From 2.2 dB, the experiment continued in increments of 0.2 dB through $E_B/N_0 = 4.0$, and no further bit errors were observed. The sample size at each energy level was 2 million information bits, or 6 million channel symbols.

As can be seen from Fig. 12 and Inequality (12), the performance of a sequential decoder is very sensitive to small changes in energy-to-noise ratio. Part of the difficulty in this experiment was finding a range of energy levels and data rates over which anything of statistical interest could be observed in a reasonable length of time. The next part of the experiment was an attempt to verify Inequality (12) and show $p[C > L] \simeq L^{-\rho}$. Plotted on log–log paper, this distribution function should be a straight line of slope $-\rho$. For each of the six cases $E_B/N_0 = 4.6$, 4.0, 3.4, 2.7, 2.2, and 1.2, two million information bits were processed and the decoding time observed. Time during which the computer performed input and output functions was not recorded. The results are presented in Fig. 15 and show straight-line behavior in all cases except the last in which the decoder was in error in almost half of the decoded bits. The measured slopes show close agreement with the theoretical values of Fig. 12, except for E_B/N_0 of 4.6, in which case the decoder is about twice as fast as theoretically expected. Note that decoding time decreases as undetected errors increase, as predicted.

It is apparent that, in a real-time situation, buffer overflow is much more likely than undetected error. The final phase of the experiment was designed to determine the overflow probability. Quantized channel symbols were transmitted to the decoder in a serial stream at a fixed data rate. Tree synchronization was achieved by searching for the special interblock symbols. A buffer overflow caused the computer to record that event and wait for the start of the next block. In addition to decoding, the

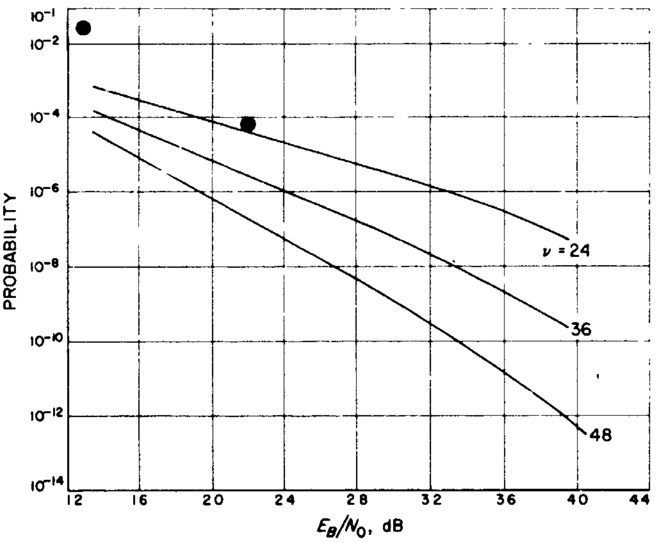


Fig. 14. Theoretical error probability for constraint lengths (ν) 24, 36, and 48 and observed error probabilities for constraint length 24

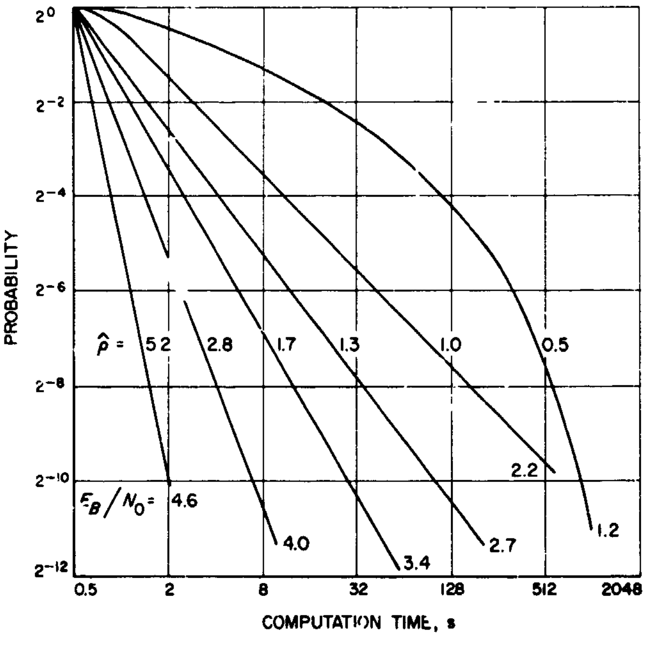


Fig. 15. Observed distribution of computation time for constraint length 24

computer performed input and output functions. Two output devices were tested, the line printer and the magnetic tape recorder. The line printer appears worse largely because the computer spends extra time formatting output bits for printing, whereas in the magnetic tape case, the output was merely recorded and read later with another program.

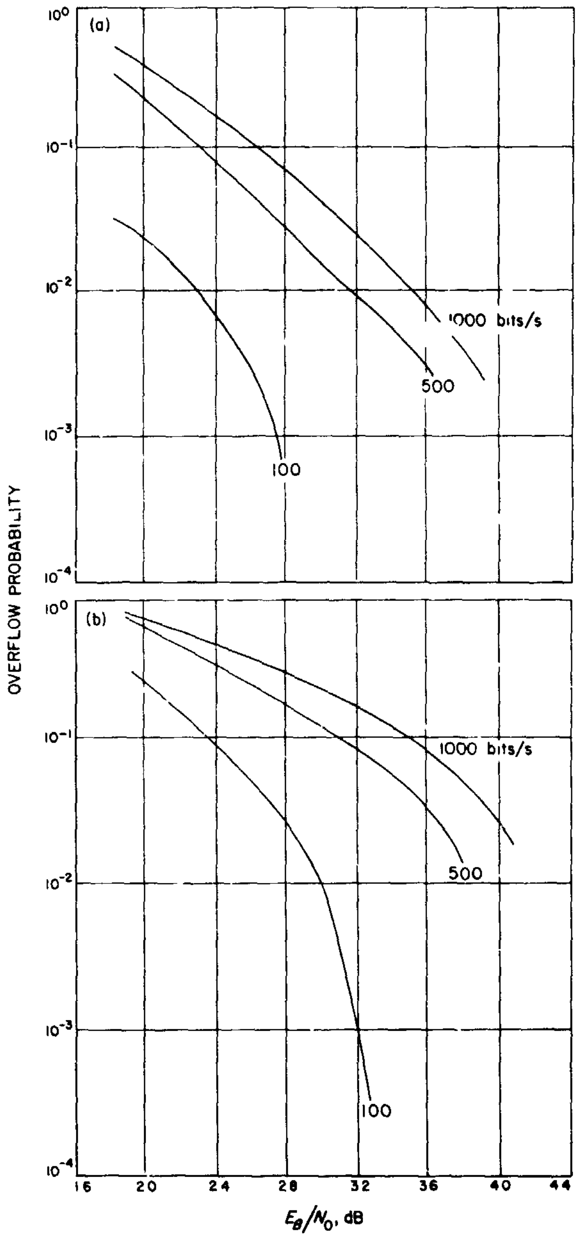


Fig. 16. Observed erasure probability for: (a) magnetic tape as output device, and (b) line printer as output device

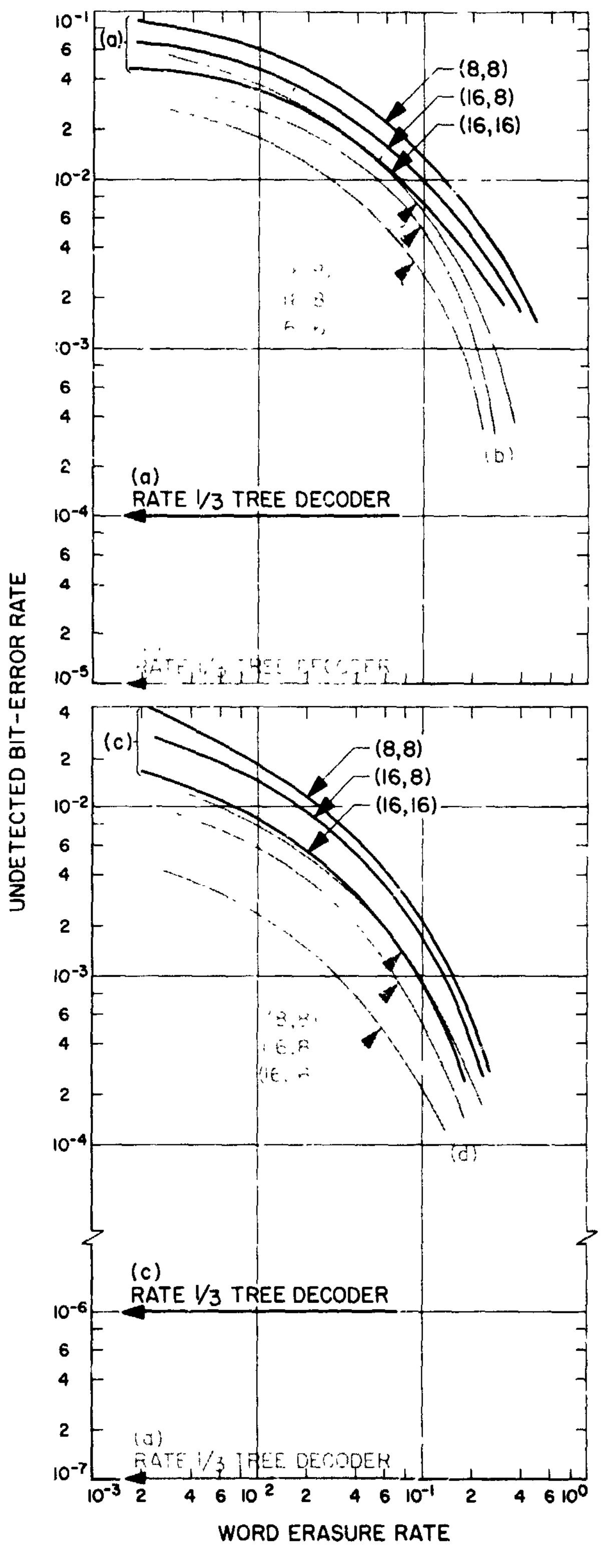


Fig. 17. Comparison of erasure rate vs undetected bit-error rate for E_B/N_0 of: (a) 2.2, (b) 2.8, (c) 3.4, and (d) 4.0 dB

Figures 16a and 16b show the results of the real-time simulation for magnetic-tape and line-printer output, respectively. The experiment was conducted over the dynamic range of 1.8 to 4.0 dB in steps of 0.2 dB at three data rates, 100, 500, and 1000 bits/s on input (300, 1500, 3000 symbols/s) and overflow probabilities recorded. For each case, 2 million bits were processed, and the graphs are not continued beyond the point where no overflows were recorded. No undetected errors were observed.

7. Comparison With Orthogonal Block Codes

The erasure versus error performance of orthogonal and biorthogonal codes, using the optimum decoder of *Subsection 3* in place of the sequential decoding algorithm in the decoder of Fig. 13, can be calculated and then compared with the simulation results for tree codes. It's easy to verify that, for the additive normal noise channel, the optimum block decoder of Fig. 13 is actually a correlation decoder. Figure 17 shows the relation between word erasure probability and undetected bit-error probability for fixed E_B/N_0 of 2.2, 2.8, 3.4, and 4.0 dB. In each case, the horizontal line represents the performance of the sequential decoder. A fixed-speed advantage for the decoder over data rate results in a fixed-overflow probability, while increasing that speed advantage decreases overflow probability and maintains the same undetected bit-error probability.

For the block codes and likelihood decoding with erasures. the undetected bit-error rate is a function of the word erasure rate. In the range of 2.2- to 4.0-dB bit energy-to-noise ratio, this error rate is inferior by many orders of magnitude to the performance of the tree decoder unless an unrealistically high erasure rate, above 0.5, can be tolerated. This erasure probability can be improved by either accepting, for a fixed code, a higher bit-error rate, or by operating at a lower information rate. [The orthogonal (8,3) and biorthogonal (16,3) have rates ⅜ and ½, respectively.] This latter possibility is limited by Eq. (16) and the performance of the symbol synchronizer. For example, a (16,4) orthogonal code has $E_S/N_0 = -3.82$ for $E_B/N_0 = 2.2$ dB. The erasures resulting from likelihood decoding of block codes can never be recovered with other changes in the system, whereas erasures from sequential decoding, if recorded, could be decoded offline with the error rate indicated on the graph.

For high information rates, low bit energy-to-noise ratio, and a general-purpose digital computer, sequential decoding provides superior performance to optimum decoding of orthogonal block codes.

References

1. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communications Engineering*, Chap. 6. John Wiley & Sons, Inc., New York, 1965.
2. Jacobs, I. M., "Sequential Decoding for Efficient Communication from Deep Space," *IEEE Trans. Commun. Technol.*, COM-15, No. 4, pp. 492–501, Aug. 1967.
3. Gallager, R. G., "A Simple Derivation of the Coding Theorem," *IEEE Trans. Inform. Theory*, IT-11, pp. 3–18, Jan. 1965.
4. Shannon, C. E., Gallager, R. G., and Berlekamp, E., "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels," *Inform. Contr.*, Vol. 10, pp. 65–103, Jan. 1967.
5. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Inform. Theory*, IT-13, pp. 260–269, Apr. 1967.
6. Yudkin, H. L., *Channel State Testing in Information Decoding*, Ph. D. thesis Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., Sept. 1964.
7. Berlekamp, E., and Jacobs, I. M., "A Lower Bound to the Distribution of Computation for Sequential Decoding," *IEEE Trans. Inform. Theory*, IT-13, pp. 167–174, Apr. 1967.
8. Savage, J. E., "Sequential Decoding: The Computation Problem," *Bell Syst. Tech. J.*, Vol. 45, pp. 149–176, Jan. 1966.

G. Communications Systems Development: The Optimum Cross-Correlation Function for a First-Order Tracking Loop Under Unit Power Constraint, *J. W. Layland*

1. Introduction

In SPS 37-41, Vol. IV, pp. 270–272, and SPS 37-43, Vol. IV, pp. 321–323, Stiffler proved that the unconstrained optimum cross-correlation function for a first-order tracking loop is a square wave and developed a minimum mean-square-error approximation to this cross-correlation function under the additional constraint that both the received and local reference signals have unit power. Subsequent work, reported in SPS 37-50, Vol. III, pp. 284–287, determined the optimum unit power local reference signal for use when the received signal is a square wave. This article describes a more precise result obtained for the optimum cross-correlation function when both the received and local reference signals have unit power but are otherwise unconstrained.

2. Problem Formulation

The probability density function of the phase error in a first-order loop due to additive white gaussian noise has been shown to be (Ref. 1)

$$p(\phi) = C \exp\{-\alpha \int^{\phi} \rho_{rA}(\eta)\, d\eta\} \qquad (1)$$

where C is a normalizing constant, $\rho_{rA}(\eta)$ denotes the normalized cross-correlation function between $r(t)$, the received signal, and $A(t)$, the local reference signal, and $\alpha = 4/N_0K(PA/Pr)^{1/2}$ denotes loop signal-to-noise ratio (SNR). The functions $r(t)$ and $A(t)$ will be determined such that they minimize

$$E_k = \int_{\pi}^{\pi} |\phi|^k \exp\left\{-\alpha \int^{\phi} \rho_{rA}(\eta)\, a\eta\right\} d\phi \qquad (2)$$

for some integer k. If $k = 0$, this maximizes $P(0)$. For other k, this minimizes the un-normalized kth absolute central moment of the distribution. Normalized moments could be substituted at the expense of increased computational difficulty. The initial results presented do not depend on k; final numerical results are obtained for $k = 2$ only.

3. General Results

Since $A(t)$ and $r(t)$ are periodic with period 2π, they can be represented in the form

$$\left.\begin{aligned} r(t) &= \sum_{n=-\infty}^{\infty} r_n e^{j(nt+\phi_n)} \\ A(t) &= \sum_{n=-\infty}^{\infty} a_n e^{j(nt+\psi_n)} \end{aligned}\right\} \qquad (3)$$

where r_n, a_n are real, $r_n = r_{-n}$, $a_n = a_{\ n}$, $\phi_n = -\phi_{\ n}$, and $\psi_n = -\psi_n$. If Eq. (3) is inserted in Eq. (2), the lower limit of the integration over η is set to -2π, and the unit power constraint is appended, the optimization criterion becomes

$$E_k = \int_{-\pi}^{\pi} |\phi|^k \exp\left\{-\alpha \sum_n \frac{j}{n} a_n r_n (1 - e^{jn\phi}) e^{j(\psi_n - \phi_n)}\right\} d\phi + \lambda_1 \left\{\sum r_n^2 - 1\right\} + \lambda_2 \left\{\sum a_n^2 - 1\right\} \qquad (4)$$

For convenience, denote $\psi_n - \phi_n = \delta_n$. It is clear from Eq. (4) that E_k depends upon δ_n, rather than ψ_n or ϕ_n individually. To determine the optimum δ_n, compute

$$\frac{\partial E_k}{\partial \delta_n} = \int_{-\pi}^{\pi} |\phi|^k \exp\left\{-\alpha \sum a_n r_n \frac{j}{n} (1 - e^{jn\phi}) e^{j\delta_n}\right\} \cdot a_n r_n \left[\left(-\frac{j\alpha}{n}\right)(1 - e^{jn\phi})(je^{j\delta_n}) + \left(\frac{j\alpha}{n}\right)(1 - e^{-jn\phi})(-je^{-j\delta_n})\right] d\phi \qquad (5)$$

$$\frac{\partial E_k}{\partial \delta_n} = \int_{\pi}^{\pi} |\phi|^k \exp\{-\alpha \cdots\} \cdot \frac{2\alpha\, a_n r_n}{n} \{\cos\delta_n (1 - \cos n\phi) + \sin\delta_n \sin n\phi\}\, d\phi \qquad (6)$$

Note first that if $P(\phi)$ is symmetric about $\phi = 0$, $E\{|\phi|^k \sin n\phi\} = 0$; hence, $\partial E/\partial\delta_n = 0$ if $\cos\delta_n = 0$, i.e., if $\delta_n = \pm\pi/2$. Furthermore, if $\delta_n = \pm\pi/2$ for all n, then $P(\phi)$ will be symmetric about $\phi = 0$. Hence $\delta_n = \pm\pi/2$ for all n is a sufficient condition for an extremum of E_k.

The determination of the optimum coefficients a_n, r_n can be simplified by the following argument: Whatever the optimum values of $\{a_n\}$ and $\{r_n\}$ are, there will be some fixed amount of power in the nth component of both signals. Call this $k_n^2 = a_n^2 + r_n^2$ and assume that, by some means, k_n has been determined without finding a_n or r_n. Finding these then reduces to finding the $\{a_n\}$ which minimizes

$$E_k = \int_{\pi}^{\pi} |\phi|^k \exp\left\{-\alpha \sum_n \frac{j}{n} a_n (k_n^2 - a_n^2)^{1/2} e^{j\delta_n} (1 - e^{jn\phi})\right\} d\phi \qquad (7)$$

To do this compute

$$\begin{aligned} \frac{\partial E_k}{\partial a_n} &= \int_{-\pi}^{\pi} |\phi|^k \exp\{-\alpha \cdots\} \frac{r_n^2 - a_n^2}{r_n} \cdot \frac{\alpha}{n} \{2\sin\delta_n - 2\sin(\delta_n + n\phi)\}\, d\phi \\ &= \operatorname{sign}\{n\} \operatorname{sign}\{\delta_n\}\, C_n \frac{r_n^2 - a_n^2}{r_n} \end{aligned} \qquad (8)$$

where C_n is a positive constant, and sign $\{x\}$ denotes the algebraic sign of x.

If sign $\{n\}$ sign $\{\delta_n\}$ is positive, $\partial E_k/\partial a_n > 0$ for $a_n^2 < r_n^2$ and $\partial E_k/\partial a_n^2 < 0$ for $a_n^2 > r_n^2$, which implies that the minimum E_k occurs for either $a_n^2 = 0$ or $a_n^2 = k_n^2$. If, however, $a_n^2 = k_n^2$, then $r_n^2 = 0$, and if either a_n^2 or r_n^2 is zero, then the contribution to the correlation function from the nth component is zero. Therefore, if sign $\{n\}$ sign $\{\delta_n\}$ is positive, $k_n^2 = 0$.

If sign $\{n\}$ sign $\{\delta_n\}$ is negative, then $\partial E_k/\partial a_n^2 < 0$ for $a_n^2 < r_n^2$ and $\partial E_k/\partial a_n > 0$ for $a_n^2 > r_n^2$. Therefore, $a_n^2 = r_n^2$.

Use of these two results reduces E_k to the form

$$E_k = \int_{-\pi}^{\pi} |\phi|^k \exp\left\{-\alpha \sum_{n>0} 2a_n^2 \left(\frac{1-\cos n\phi}{n}\right)\right\} d\phi + \lambda \left\{\sum_{n>0} a_n^2 - \frac{1}{2}\right\} \tag{9}$$

The optimum $\{a_n\}$ are the solutions to the equations

$$\frac{\partial E_k}{\partial a_n} = a_n \left[\int_{-\pi}^{\pi} |\phi|^k \left(\frac{1-\cos n\phi}{n}\right) \exp\left\{-\alpha \sum_{m>1} 2\, a_m^2 \left(\frac{1-\cos m\phi}{m}\right)\right\} d\phi - \lambda'\right] \tag{10}$$

for all n. It is a relatively easy matter to show that the functions $A(t)$ and $r(t)$ are band-limited. For large α, $\cos\phi$ is approximately $1 - \phi^2/2$ for all ϕ for which $p(\phi)$ is not essentially zero: so for $n = 1$,

$$\lambda' \approx \frac{1}{2 \cdot p(0)} E\{|\phi|^{k+2}\} \text{ if } \alpha \text{ is large} \tag{11}$$

For any n,

$$\frac{1-\cos n\phi}{n} \leq \frac{2}{n}$$

and hence for any α,

$$\lambda' \leq \frac{2}{n \cdot p(0)} E\{|\phi|^k\} \text{ unless } a_n = 0 \tag{12}$$

Combining these two requirements:

$$\frac{1}{2} \frac{E\{|\phi|^{k+2}\}}{E\{|\phi|^k\}} \leq \frac{2}{n} \tag{13}$$

But

$$\frac{E\{|\phi|^{k+2}\}}{E\{|\phi|^k\}} = C_1 \alpha^{-C_2} \tag{14}$$

where C_1 depends upon k, and $p(\phi)$, and C_2 depends only on $p(\phi)$ and is in the range $1 \leq C_2 \leq 2$. Therefore,

$$n \leq \frac{4}{C_1} \alpha^{C_2} \text{ unless } a_n = 0 \tag{15}$$

For very small α, the exponential term in Eq. (10) can be expanded in a series and terms of higher than first order in α ignored. Solution of the resultant set of equations shows that $a_1 = 1/(2)^{1/2}$, $a_n = 0$ for $n \neq 1$ is the only solution allowed. The band limit thus extends, as expected, to small α. It may be noted that the solution to Eq. (10) is not unique, since $a_1 = 1/(2)^{1/2}$, $a_n = 0$ for $n \neq 1$ is a solution for any α. However, the solution to Eq. (10) with the maximum possible number of non-zero components should be unique, and should also represent the true minimum, since the resultant $\rho_{rA}(\eta)$ will have the steepest slope in the vicinity of 0 of any of the possible solutions.

4. Numerical Results

Equation (10) has been subjected to an iterative numerical solution for $k = 2$ and for various values of loop SNR α. A typical resultant power spectrum and the associated cross-correlation function are shown in Fig. 18. With the exception that the even harmonics are slightly suppressed, this is very similar to the main lobe of a $(\sin x/x)^2$ spectrum, the spectrum of a pseudo noise (PN) sequence of length (approximately) $\alpha/2$. The variance of the resultant phase error in a loop employing optimum signals is plotted as a function of α in Fig. 19. Also shown, for comparison, is the phase error variance obtained using the first lobe of a $(\sin x/x)^2$ spectrum with $\alpha/2$ components. Since the minimum is very broad, very little loss is suffered by use of this simply generated signal. The bottom line in this figure corresponds to the phase-error variance which would result from use of Stiffler's non-realizable optimum cross-correlation function. The 3-dB difference in performance appears to be due solely to the imposition of the realizability constraint of unit power.

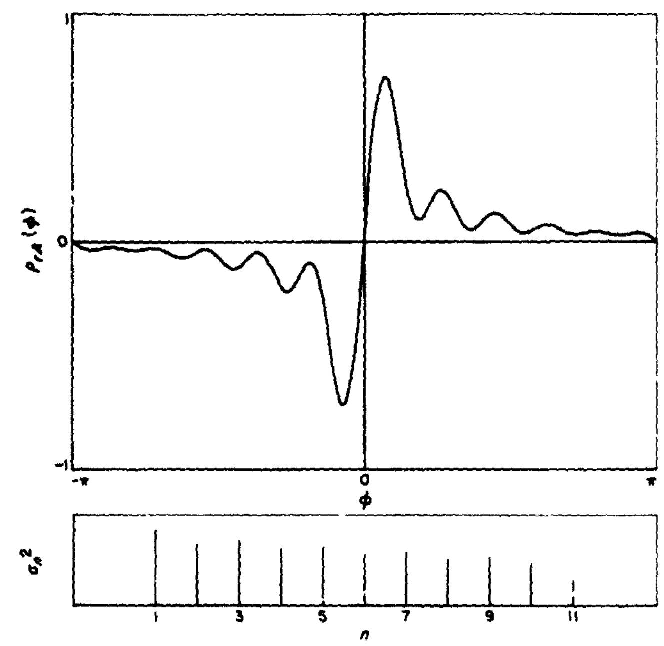


Fig. 18. Typical optimum power spectrum and cross-correlation function ($\alpha = 32$, 11 non-zero components)

5. Comparison to a PN Range Tracking Loop

Since one of the main uses of a very high SNR tracking loop is in range measurement, it is of interest to compare the performance attainable by using the optimum waveforms for such a loop with that obtained when using the binary PN sequences which are typically used. A PN waveform with p digits has a series expansion given by

$$\mathrm{PN}(t) = -\frac{1}{p} + \frac{2}{\pi}(p+1)^{1/2} \sum_{n=1}^{\infty} \frac{\sin\left(\frac{n\pi}{p}\right)}{n}$$

$$\times \cos\left(\frac{2n\pi}{p} t + \phi_n\right) \qquad (16)$$

The main lobe of this power spectrum has already been shown to be an effective approximation to the optimum $r(t)$. The local reference signal for a PN wave is usually constructed as

$$\mathrm{PNR}(t) = \frac{\mathrm{PN}(t + 1/2)}{(2)^{1/2}} - \frac{\mathrm{PN}(t - 1/2)}{(2)^{1/2}}$$

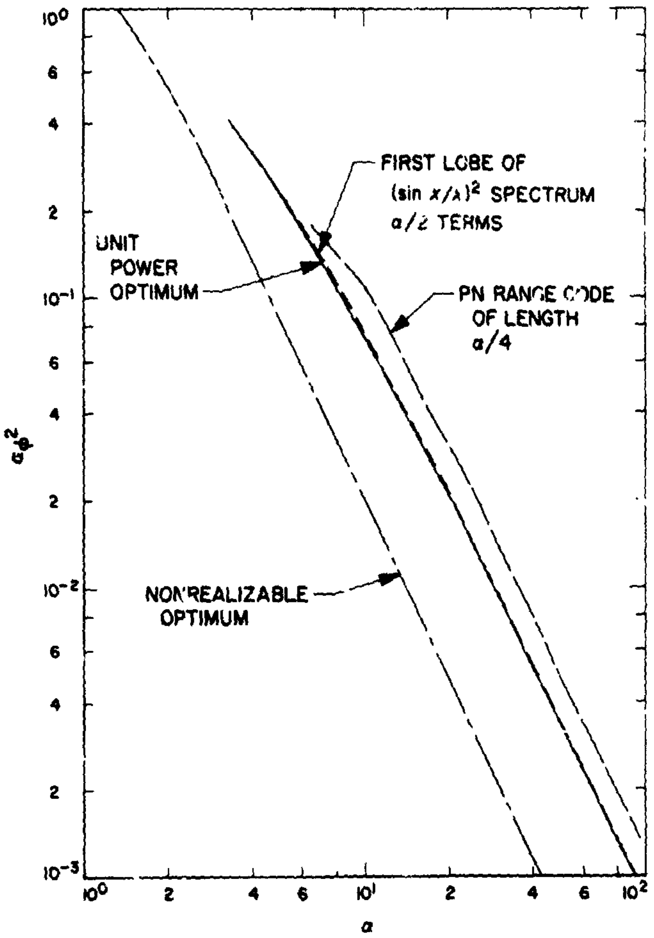


Fig. 19. Loop phase-error variance σ_ϕ^2 vs loop SNR α for various cross-correlation functions

This signal possesses the expansion

$$\mathrm{PNR}(t) = \frac{2}{\pi}[2(p+1)]^{1/2} \sum_{n=1}^{\infty} \frac{\sin^2\left(\frac{n\pi}{p}\right)}{n}$$

$$\times \cos\left(\frac{2n\pi}{p} t + d_n + \frac{\pi}{2}\right) \qquad (17)$$

The phase relationship of PNP (t) to PN (t) is the same as that of the optimal reference for components in the range $2k \leq n/p \leq 2k + 1$ and phase-reversed for $2k + 1 \leq n/p \leq 2k + 2$, all k. It would appear that tracking performance could be improved by filtering to remove all frequency components above $n = p$. In addition, a factor of $\sin(n\pi/p)$ modifies each term of PNR (t). The effect of this is shown in Fig. 20, which shows a comparison between the phase variance which results in a first order tracking loop using the usual PN system and using

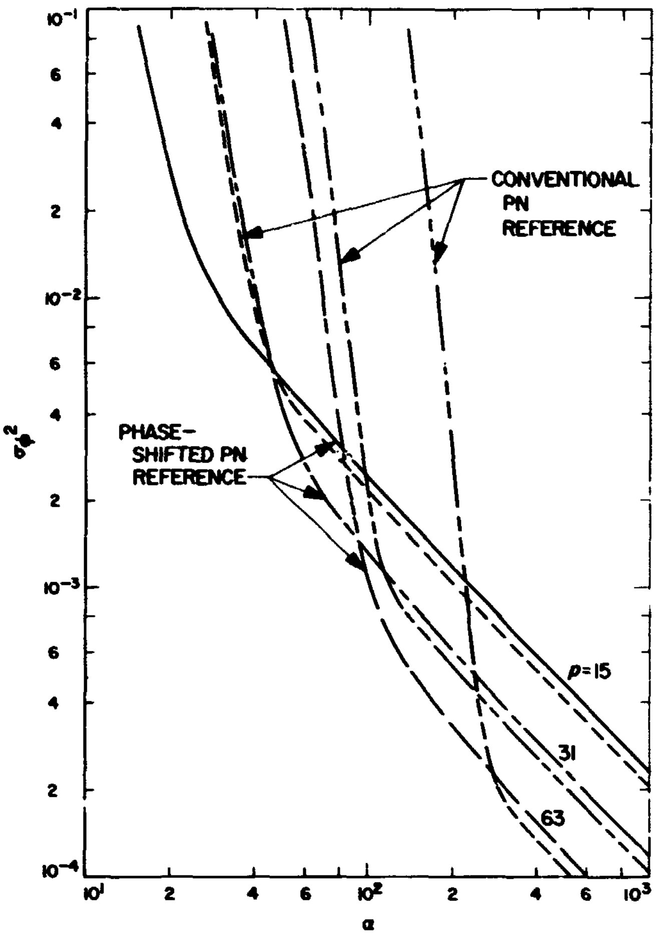


Fig. 20. Comparison of phase-error variance σ_ϕ^2 for conventional PN reference and phase-shifted PN reference, for fixed code length p as a function of α

a phase-shifted PN for local reference signal, considering frequency components $n \leq p$ only. The curves show that the usual PN system is poorer than the modified one for $\alpha < 4p$ and indicate that the best usual PN system to use has code length $p \approx \alpha/4$. The upper line of Fig. 19 corresponds to the phase variance in a usual PN system with code-length $\alpha/4$. An improvement of approximately 1 dB in effective loop SNR can be obtained by use of a phase-shifted PN reference signal as opposed to the reference signal usually implemented.

Reference

1. Viterbi, A. J., "Phase Locked Loop Dynamics In the Presence of Noise by Fokker-Planck Techniques," *IEEE Proc.*, pp. 1737–1753, Dec. 1963.

H. Information Processing: Disjoint Cycles From the de Bruijn Graph, *H. Fredricksen*

1. The de Bruijn Diagram

a. Description. An n-bit shift register (Fig. 21) is a set of n storage registers with logic which defines their contents at any point in time. The contents of the ith storage register at time t is equal to the contents of the $(i-1)$st storage register at time $t-1$, for $2 \leq i \leq n$. A feedback function $f(x_1, x_2, \cdots, x_n)$ determines the contents of the first register x_1 at time t from the contents of the n registers $x_1, x_2, \cdots, x_n$ at time $t-1$. The contents of the register at time t, regarded as a binary number or a binary vector, is called the *state* of the register. At the end of each time interval, determined by an external clock, there is a transition from one state to the next.

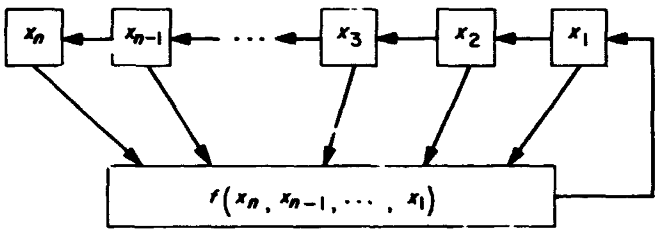


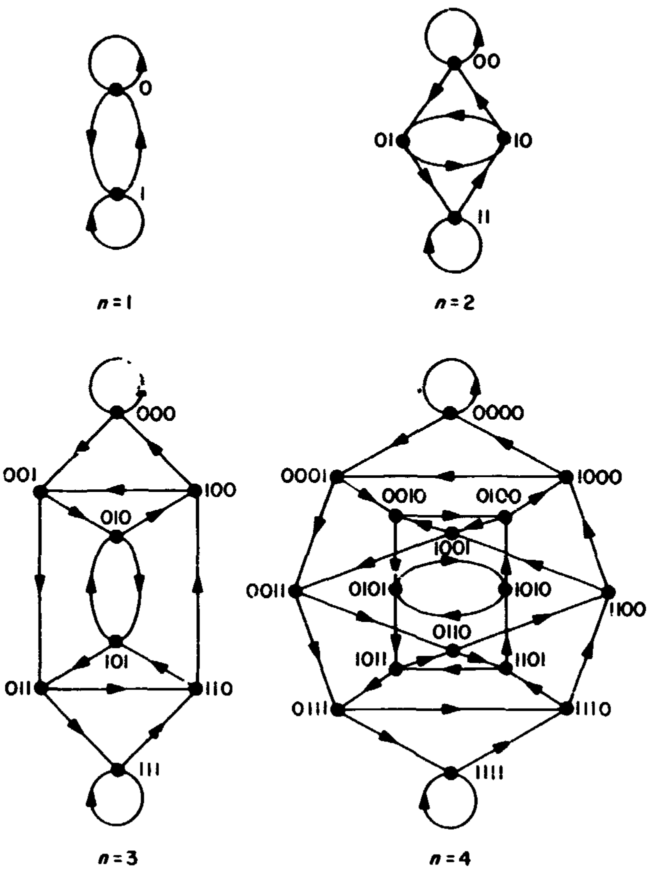
Fig. 21. General shift register

Since there are 2^n vectors defined by a register of length n, there are 2^n states for the shift register. The diagram of all possible state transitions is called the *de Bruijn diagram*. The de Bruijn diagrams for $n = 1,2,3,4,5$ are shown in Figs. 22 and 23. Each node in a diagram has two possible successors and two possible predecessors. These diagrams contain all possible transition patterns for their shift registers.

Transition patterns in the de Bruijn diagram are determined by the feedback function of the shift register. For the function to be well defined, we require that each state have only one successor. Then the feedback function chooses exactly one path for the exit from each state of the diagram. If we change the feedback function so that a state maps into the other state possible, we say we have chosen the alternate successor for the state.

b. Cycles of the de Bruijn diagram. Let f be the feedback function which defines the state transitions. If a succession of k state transitions leads from state s_i back to s_i, i.e.,

$$s_j = f(s_i), \qquad s_k = f(s_j) = f^2(s_i), \cdots, \qquad s_i = f^k(s_i)$$



n = 1 *n* = 2

n = 3 *n* = 4

Fig. 22. de Bruijn graphs for *n* = 1,2,3,4

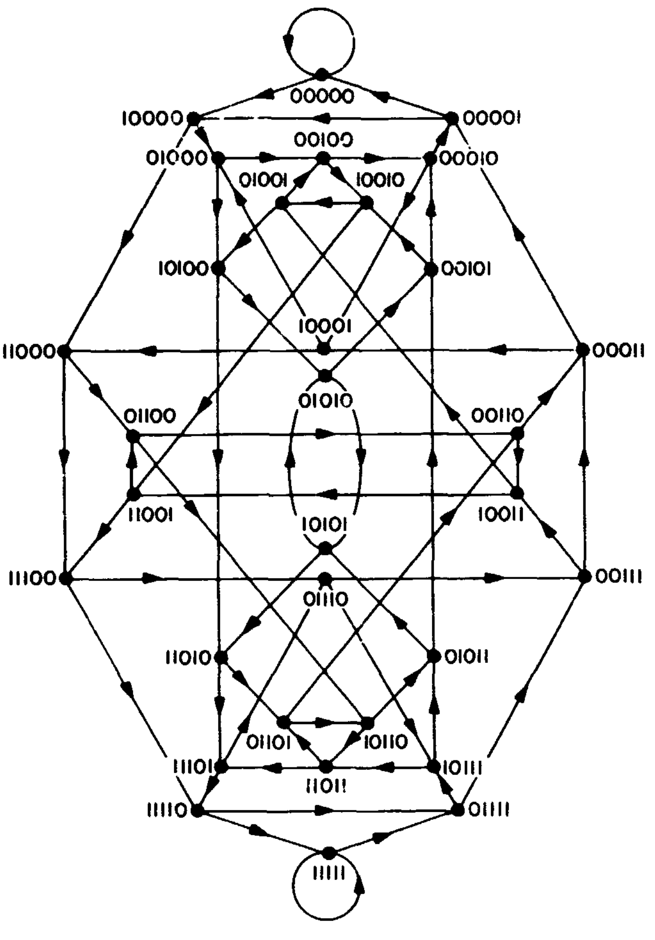


Fig. 23. de Bruijn graph for *n* = 5

we say the states $s_i, f(s_i), f^2(s_i), \cdots, f^{k-1}(s_i)$ form a *cycle* of length k in the diagram. The cycle can be described as the k-tuple of *zeros* and *ones* which are the feedback values of the states on the cycle. Equivalently, the cycle could be described in decimal notation by the decimal equivalent of the binary representation of the states which make up the cycle. It will often be convenient to use each of these representations in what follows.

We would also like to restrict the truth tables so that each state has a unique predecessor as well as a unique successor. This will decompose the de Bruijn diagram in such a way that every state will be on a unique cycle. Golomb (Ref. 1, p. 115) gives a condition that insures that a feedback function yield pure cycles. We state that condition here.

Theorem 1. The feedback function for a shift register yields pure cycles if the last variable enters linearly into the feedback function.

Then the feedback function $f(x_1, x_2, \cdots, x_n)$ can be represented as $g(x_1, x_2, \cdots, x_{n-1}) + x_n$. When the feedback function is so representable, the half of the truth table where $x_n = 1$ is the complement of the half of the truth table where $x_n = 0$. The truth table of the function g contains all the information about the shift register. When we speak of the truth table of the shift register in what follows, we shall mean the truth table of the function g. The context will make it clear when we wish to discuss an arbitrary feedback function f.

c. Particular cycle decompositions. The question as to whether a cycle can be found which contains all the states of the diagram has been answered in the affirmative by de Bruijn (Ref. 2). The number of de Bruijn cycles is also given in Ref. 2. He shows there are $2^{2^{n-1}-n}$ de Bruijn cycles in the graph. Other authors (Refs. 1 and 3) give alternate proofs. In *Subsection* 2, we discuss the distribution of the de Bruijn cycles by their weight.

The existence of cycles of all lengths from length 1 to length 2^n from a register of length n is shown in Golomb (Ref. 1, p. 192).

Other feedback functions yield special cycle decompositions which are of interest. Two simple functions shall be discussed in *Subsection* 2 and in a future article. They are the pure-cycling register and the complementing-cycling register. The pure-cycling register is given by the feedback function $f(x_1, x_2, \cdots, x_n) = x_n$, $(g \equiv 0)$. The complementing-cycling register is given by the feedback function

$$f(x_1, x_2, \cdots, x_n) = x_n + 1, \qquad (g \equiv 1)$$

Golomb (Ref. 1) shows the number of cycles determined by the pure-cycling register is given by

$$Z(n) = \frac{1}{n} \sum_{d/n} \phi(d)\, 2^{n/d}$$

where the summation is over all divisors d of n and ϕ is the Euler ϕ function. He shows $Z(n)$ is even for all $n > 2$. The number of cycles determined by the complementing-cycling register is

$$Z^*(n) = \frac{1}{2} Z(n) - \frac{1}{2n} \sum_{2d/n} \phi(2d)\, 2^{n/2d}$$

Here the summation is over only the even divisors of n.

Golomb makes the conjecture (Ref. 1, p. 174) that the maximum number of cycles into which the de Bruijn graph can be decomposed is equal to $Z(n)$. This conjecture will be discussed in a future article.

2. Distribution of Truth Tables by Number of Cycles and Weight

a. Boundary of the table. Consider the set of all feedback functions on the shift register. We determine the various cycle decompositions from the de Bruijn graph. For a register of length n, there are $2^{2^{n-1}}$ truth tables.

For $n = 3, 4, 5$, we group the truth tables according to their weight and to the number of cycles they generate. We only need consider the half of the truth table where $x_n = 0$ since the half where $x_n = 1$ is just its complement. For $n = 3$, there are only two free variables x_1 and x_2. There are four possible value pairs which these two variables can take on.

Corresponding to each value pair we have a two-fold choice of 0 or 1 for the feedback function at that position. This gives us a total of 16 different truth tables. The weight of the truth tables ranges between 0 and 4 and the distribution of truth tables, by their weight and by the number of cycles they produce, is given in Table 3.

Table 3. Cycle decomposition table for $n = 3$

		Number of cycles			
		1	2	3	4
Weight of truth table	0	C	0	0	1
	1	0	0	4	0
	2	0	5	0	1
	3	2	0	2	0
	4	0	1	0	0

F_0, the pure-cycling register, is the register of weight 0. The four cycles generated by F_0 are (0), (1), (001), and (011). No other feedback truth table yields more cycles than F_0. There is one truth table which ties F_0 for the maximum. This is F_6, where the subscript is the decimal representation of the truth table given by the values that the variables take on. For F_6 the variable pair x_2, x_1 take on the values $f(0,0) = 0$, $f(0,1) = 1$, $f(1,0) = 1$, $f(1,1) = 0$ and the subscript is given by

$$s = f(0,0)\,2^3 + f(0,1)\,2^2 + f(1,0)\,2 + f(1,1)$$

F_6 yields the cycle structure (0), (1), (01), (0011).

A change in a single position in the truth table will cause a change of one in the number of cycles from one truth table to the next, either increasing or decreasing by one the number of cycles (Ref. 1). For $n = 3$, there are four possible single changes which could be made in the truth table. If we start from F_0, the four changes are all $0 \rightarrow 1$ changes resulting in the four truth tables F_1, F_2, F_4, F_8 all of weight 1. In every case, we find the number of cycles decreases when we make this change. This is the result of two cycles joining and forming a single cycle. From F_2 or F_4, an additional change results in F_6, which splits the cycle which has been formed to form two new cycles.

For $n = 4, 5$, the cycle decomposition tables are given in Tables 4 and 5. The truth tables F_{18}, F_{72}, F_{90} all yield $Z(4) = 6$ cycles.

The behavior in the decomposition tables shown is typical of the general behavior of these tables. Figure 24 is a picture of the typical decomposition table. We state

Table 4. Cycle decomposition table for $n = 4$

Weight of truth table	Number of cycles					
	1	2	3	4	5	6
0	0	0	0	0	0	1
1	0	0	0	0	8	0
2	0	0	0	26	0	2
3	0	0	44	0	12	0
4	0	37	0	32	0	1
5	12	0	40	0	4	0
6	0	22	0	6	0	0
7	4	0	4	0	0	0
8	0	1	0	0	0	0

Table 5. Cycle decomposition table for $n = 5$

Weight of truth table	Number of cycles							
	1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	16	0
2	0	0	0	0	0	114	0	6
3	0	0	0	0	476	0	84	0
4	0	0	0	1253	0	552	0	15
5	0	0	2036	0	2132	0	200	0
6	0	1736	0	5050	0	1197	0	25
7	576	0	6488	0	4098	0	278	0
8	0	4056	0	7326	0	1467	0	21
9	960	0	6684	0	3572	0	224	0
10	0	2892	0	4338	0	767	0	11
11	448	0	2652	0	1210	0	58	0
12	0	736	0	962	0	121	0	1
13	64	0	368	0	124	0	4	0
14	0	52	0	62	0	6	0	0
15	0	0	12	0	4	0	0	0
16	0	0	0	1	0	0	0	0

certain theorems here which relate to the character of the general decomposition table.

Theorem 2. The line $[k, Z(n) - k]$ defines the (achieved) upper border of the decomposition table.

Theorem 3. The line $[2^{n-1} - k, Z^*(n) - k]$ is the lower-left boundary of the table.

There are three other border lines on the decomposition table which are of some interest. To state the location of the right-hand border of the table is to answer the conjecture on the maximum number of cycles from a register. In what follows, we shall assume that the conjecture is correct. We show below that the right-hand border is at least as long as twice the number k of $[\alpha, r(\alpha)]$ pairs on a cycle which have their respective alternate successors

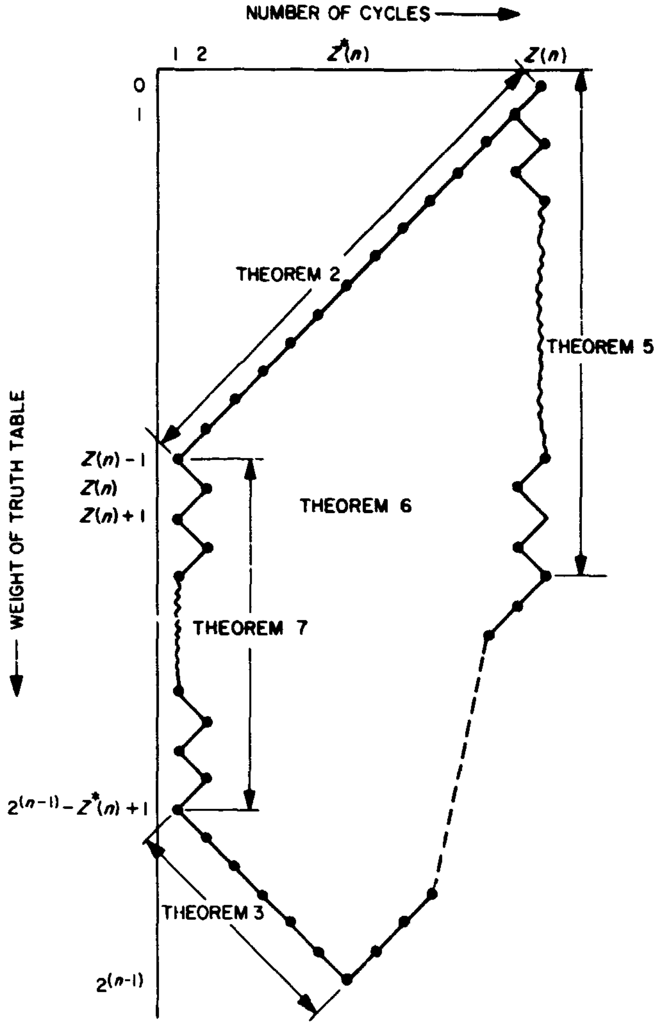


Fig. 24. General cycle decomposition table

on one (other) cycle. (The reverse function r is defined below.) No examples are known of a truth table of weight greater than $2k$ and $Z(n)$ cycles. We also show that there are at least 2^k examples of truth tables with $Z(n)$ cycles.

In the cycle decomposition tables for $n = 3, 4, 5$, we note that there is more than one truth table having $Z(n)$ cycles. This is true for all $n \geq 3$. Consider the two cycles from the pure-cycling register $(000 \cdots 01)$ and $(00 \cdots 011)$ each of length n. We could change the successor of

$$\overbrace{00 \cdots 01}^{n} \text{ and of } \overbrace{010 \cdots 0}^{n}$$

and have a truth table of weight 2 which had $Z(n)$ cycles. The two new cycles would be $(0 \cdots 011)$ of length $n + 1$ and $(0 \cdots 01)$ of length $n - 1$.

The n-tuples $00 \cdots 01$ and $010 \cdots 0$ can be related to one another. We can say $010 \cdots 0$ is the *reverse* of $0 \cdots 01$, where the *reverse function* r is defined by

$$r(\alpha) = r(a_n, a_{n-1}, \cdots, a_1) = (a_n, a_1, a_2, \cdots, a_{n-1})$$

The reverse function is an order 2 function.

Consider the half of the truth table where $a_n = 0$. We can separate the positions, by the reverse operation, into pairs. Some positions will be self-reverse and will not pair with any other position.

Suppose we can find a pair $\alpha, r(\alpha)$ on the same pure-cycle for which the pair $\alpha^*, r(\alpha)^*$, the respective alternate successors, is on another pure-cycle. Then we can change the successor of α and $r(\alpha)$ and not alter the number of cycles. Suppose there exist k such pairs. We can show the following:

Theorem 4. There exist $\geqq 2^k$ truth tables yielding $Z(n)$ cycles.

Theorem 5. There exist truth tables yielding $Z(n)$ cycles having weight $\geqq 2k$.

For $n = 7$, there are 15 $[\alpha, r(\alpha)]$ pairs. So we have at least 2^{15} truth tables which have $Z(n)$ cycles. This is from a truth table set of 2^{64} truth tables. There are examples of truth tables of weight 30 producing $Z(n)$ cycles.

Because of the unsettled nature of the $Z(n)$ conjecture, the lower right-hand border is also unsettled. The left-hand border consists of the so-called de Bruijn sequences. All of the 2^n nodes of the graph are on one cycle. We discuss this border below.

Assume the border has been established. We make the following statement about the interior of the table.

Theorem 6. There are no (weight, number of cycles) pairs interior to the table, for which truth tables are possible, that do not occur. (The weight and number of cycles must be of the same parity for $n > 2$.)

In Theorems 2 and 3, we exhibited an upper and lower boundary. This left-hand end of the boundary in each case came at the place where all 2^n nodes were on one cycle. In the first case, the weight of the truth table was $Z(n) - 1$, and in the second case, the weight was $2^{n-1} - Z^*(n) + 1$. We showed these values were the lower and upper limits, respectively, for the weight of a truth table having exactly one cycle associated with it. We can show that every odd weight between these limits has a truth table of that weight associated with a de Bruijn cycle or sequence.

Theorem 7. There exists a de Bruijn cycle for every odd weight between $Z(n) - 1$ and $2^{n-1} - Z^*(n) + 1$.

b. de Bruijn cycles.

Distribution by weight. We showed in *Paragraph a*, above, that de Bruijn cycles exist for every length register. We also showed that there is a minimum and a maximum weight for a truth table which produces a de Bruijn cycle. Finally, we showed that every value between the minimum and the maximum had a truth table of that weight which defined a de Bruijn cycle. The number of all such cycles has been given by de Bruijn to be $2^{2^{n-1}-n}$ (Ref. 2). For $n = 3, 4, 5$, the number of de Bruijn cycles is 2,16,2048. These are the number of de Bruijn cycles we show in *Subsection 2-a*. We also classify the de Bruijn cycles by the weight of their truth table.

A well-known graph-theoretic theorem can be applied to find the number of de Bruijn cycles of maximum and minimum weight. For the cycles of maximum weight, we form the decomposition of the space of 2^n nodes by the complementing cycling register. Label the cycles formed as $A_1, A_2, \cdots, A_{Z^*(n)}$. We form a labeled graph containing $Z^*(n)$ nodes in the following way:

Connect A_i to A_j if A_i contains a vector whose alternate successor is on $A_j, j \neq i$. Label the arc from A_i to A_j with the number of such vectors on A_i. We form a matrix B with entries b_{ij} = label on the arc from A_i to A_j. Also form the diagonal matrix $C = (c_{ij})$ whose entries are given by c_{ii} equal to the sum of the labels on arcs entering A_i. The theorem states that the number of rooted trees of the graph is equal to the determinant of the minor of d_{11}, where A_1 is a root of the graph and $D = (d_{ij}) = C - B$. The number of rooted trees is equal to the number of de Bruijn cycles of maximum weight. Applying the theorem, we find the number of de Bruijn cycles of maximum weight for $n = 3, 4, 5, 6, 7$ are $2, 4, 64, 2^{14}, 3 \times 2^{26}$. The first four values were checked on the computer by exhaustive search and the truth tables listed. For $n = 7$, the time required to check all de Bruijn cycles of weight 57 is prohibitive.

The matrix B is symmetric. This can be seen by noting for $a \epsilon A$ which has its alternate successor $a^* \epsilon B$, there is $\bar{a}$, the $2^{n-1} - 1$ complement of a on B with its alternate successor on A. That is, if $a = a_1, a_2, \cdots, a_n$, then by the complementing cycling register,

$$A = (a_1, a_2, \cdots, a_n, \bar{a}_1, \cdots, \bar{a}_n)$$

If a has its alternate successor, $a^* = a_2, a_3, \cdots, a_n, a_1$ on B, then $B = (a_2, \cdots, a_n, a_1, \bar{a}_2, \cdots, \bar{a}_n, \bar{a}_1)$. The complement of a, $\bar{a} = a_1, \bar{a}_2, \cdots, \bar{a}_n$ is on B with its alternate successor $(\bar{a})^* = \bar{a}_2, \cdots, \bar{a}_n, a_1$ on A.

For the de Bruijn cycles of minimum weight, we employ the pure-cycle decomposition. We form the graph in the same way as above. An equivalent determinant is taken to find the number of de Bruijn cycles of minimum weight.

We form a table with the number of de Bruijn cycles of minimum and maximum weight for the first few values of n.

n	Number of cycles of maximum weight = $[C_{max}(n)]$	Number of cycles of minimum weight = $[C_{min}(n)]$
1	1	1
2	1	1
3	2	2
4	4	12
5	2^6	$2^6 \cdot 3^2$
6	2^{14}	$2^{14} \cdot 3^4 \cdot 5^2$
7	$2^{26} \cdot 3$	$2^{28} \cdot 3^5 \cdot 5^3 \cdot 13$

For $n = 1, 2, 3$, the minimum weight equals the maximum weight. For $n = 4$, there are no other de Bruijn cycles. It is interesting to note that

$$C_{max}(n) \mid C_{min}(n); \text{ also, } C_{max}(n) \mid C_{max}(n+1) \text{ and } C_{min}(n) \mid C_{min}(n+1)$$

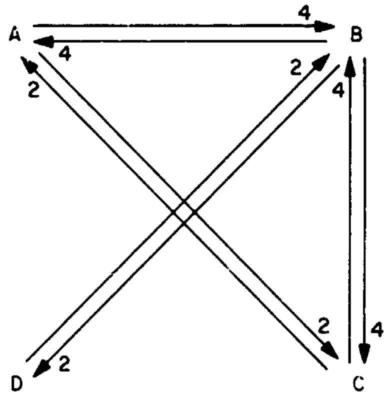
Also, from Table 5 we see that $C_{max}(5)$ divides the number of de Bruijn cycles of any weight.

We give examples to illustrate the method for the case $n = 5$. For the de Bruijn cycles of maximum weight, we form the cycle decomposition of the complementing cycling register on five variables. With maximum weight of de Bruijn cycles, $n = 5$, an example of the graph is:

Cycle	Set of vectors
A	0, 1, 3, 7, 15, 31, 30, 28, 24, 16
B	2, 5, 11, 23, 14, 29, 26, 20, 8, 17
C	4, 9, 19, 6, 13, 27, 22, 12, 25, 18
D	10, 21

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 6 & -4 & -2 & 0 \\ B & -4 & 10 & -4 & -2 \\ C & -2 & -4 & 6 & 0 \\ D & 0 & -2 & 0 & 2 \end{array}$$

Matrix C–B



where D is a root of the graph. We evaluate the determinant of the minor of the D,D position

$$\begin{vmatrix} 6 & -4 & -2 \\ -4 & 10 & -4 \\ -2 & -4 & 6 \end{vmatrix} = 64$$

For the number of de Bruijn cycles of minimum weight, we form the cycle decomposition under the pure-cycling register for $n = 5$. The graph of the cycle connections is given in Fig. 25. With minimum weight of de Bruijn

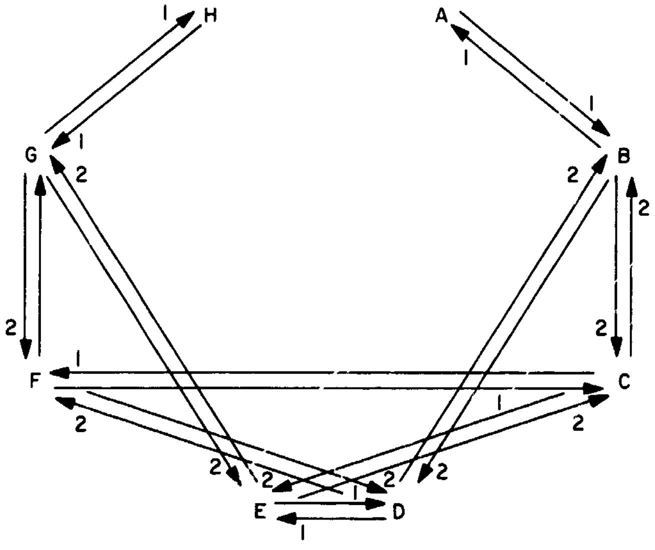


Fig. 25. Graph of cycle connections for pure-cycle register

cycles, $n = 5$, an example of the graph is:

Cycle	Set of vectors
A	0
B	1, 2, 4, 8, 16
C	3, 6, 12, 24, 17
D	5, 10, 20, 9, 18
E	7, 14, 28, 25, 19
F	11, 22, 13, 26, 21
G	15, 30, 29, 27, 23
H	31

$$\begin{array}{c|rrrrrrrr} & A & B & C & D & E & F & G & H \\ \hline A & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ B & -1 & 5 & -2 & -2 & 0 & 0 & 0 & 0 \\ C & 0 & -2 & 5 & 0 & -2 & -1 & 0 & 0 \\ D & 0 & -2 & 0 & 5 & -1 & -2 & 0 & 0 \\ E & 0 & 0 & -2 & -1 & 5 & 0 & -2 & 0 \\ F & 0 & 0 & -1 & -2 & 0 & 5 & -2 & 0 \\ G & 0 & 0 & 0 & 0 & -2 & -2 & 5 & -1 \\ H & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{array}$$

Matrix C–B

A is a root of the graph (Fig. 25). We evaluate the determinant of the minor of the A, A position

$$\begin{vmatrix} 5 & -2 & -2 & 0 & 0 & 0 & 0 \\ -2 & 5 & 0 & -2 & -1 & 0 & 0 \\ -2 & 0 & 5 & -1 & -2 & 0 & 0 \\ 0 & -2 & -1 & 5 & 0 & -2 & 0 \\ 0 & -1 & -2 & 0 & 5 & -2 & 0 \\ 0 & 0 & 0 & -2 & -2 & 5 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{vmatrix} = 576 = 2^6 \cdot 3^2$$

The lexicographically least de Bruijn cycle. An example is given in Ford (Ref. 4) to show that de Bruijn cycles exist for all orders. We start with a register of length n filled with *zeros*. Take for the vector $a_n, a_{n-1}, \cdots, a_1$ its odd successor $a_{n-1}, a_{n-2}, \cdots, a_1, 1$ if possible. If the odd successor has been used, i.e.,

$$\bar{a}_n, a_{n-1}, \cdots, a_1 \rightarrow a_{n-1}, \cdots, a_1, 1$$

we use the even successor. If we follow this construction, we have a de Bruijn cycle of order n. If we list the de Bruijn cycles in lexicographic order, with 1 preceding 0, the cycle thus formed is the lexicographically least de Bruijn cycle.

Theorem 8. The truth table of the lexicographically least de Bruijn cycle has weight $Z(n) - 1$.

Corollary. The truth table of lexicographically greatest de Bruijn cycle L is weight $Z(n) - 1$.

References

1. Golomb, S. W., *Shift Register Sequences.* Holden-Day, Inc., San Francisco, Calif., 1967.
2. Van Aardenne-Ehrenfest, T., and de Bruijn, N. G., "Circuits and Trees in Oriented Linear Graphs," *Simon Stevin,* Vol. 28, p. 203, 1951.
3. Hall, M., Jr., *Combinatorial Theory.* Blaisdell Publishing Company, Waltham, Mass., 1967.
4. Ford, L. R., Jr., *A Cyclic Arrangement of M-tuples,* Report P-1071. Rand Corporation, Santa Monica, Calif., Apr. 23, 1957.

I. Information Processing: Estimating the Correlation Between Two Normal Distributions When Only the Means are Known, *I. Eisenberger*

1. Introduction

Let x and y denote two jointly normal random variables distributed $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, with correlation ρ, and let $\{x_i, y_i\}$ be a set of n independent pairs of sample values. In SPS 37-50, Vol. III, pp. 287–289, a linear unbiased estimator of ρ is given by

$$\tilde{\rho} = \frac{\left(\frac{\pi}{2}\right)^{1/2}}{2n} \left[\frac{1}{\sigma_1} \sum_{i=1}^{n} (x_i - \mu_1) \operatorname{sgn}(y_i - \mu_2) + \frac{1}{\sigma_2} \sum_{i=1}^{n} (y_i - \mu_2) \operatorname{sgn}(x_i - \mu_i) \right]$$

The estimator $\tilde{\rho}$ has two disadvantages:

(1) The moments of x and y must be known; a somewhat unrealistic assumption.

(2) Although the efficiency of $\tilde{\rho}$ relative to the maximum likelihood estimator is quite high when ρ is near zero, it is quite poor for ρ close to ±1. For example, for $\rho = 0$, eff $(\tilde{\rho}) = 0.778$, while for $\rho = 0.8$, eff $(\tilde{\rho}) = 0.098$.

In practical situations, however, it often occurs that, although the variances are unknown, nevertheless the means are known. Under these conditions, and assuming without loss of generality that $\mu_1 = \mu_2 = 0$, we propose in this article the asymptotically unbiased estimator of ρ given by

$$\hat{\rho} = \frac{1}{2}\left[\frac{\sum_{i=1}^{n} x_i \operatorname{sgn} y_i}{\sum_{i=1}^{n} |x_i|} + \frac{\sum_{i=1}^{n} y_i \operatorname{sgn} x_i}{\sum_{i=1}^{n} |y_i|}\right] \tag{1}$$

For $\rho = \pm 1$, $y_i = \pm(\sigma_2/\sigma_1)x_i$, so that x_i sgn $y_i = \pm|x_i|$ and y_i sgn $x_i = \pm|y_i|$. Thus, $\hat{\rho}$ has the property that as $\rho \to \pm 1$, $\operatorname{var}(\hat{\rho}) \to 0$. It will also be shown that although the asymptotic efficiency of $\hat{\rho}$ decreases as ρ increases (decreases) from zero, it does so at a much slower rate than does the efficiency of $\tilde{\rho}$. For example, for $\rho = 0$, $\operatorname{eff}(\hat{\rho}) = 0.778$ and for $\rho = 0.8$, $\operatorname{eff}(\hat{\rho}) = 0.504$.

2. The Asymptotic Variance of $\hat{\rho}$ and Its Efficiency

It is not difficult to show the following:

$$E(|x|) = \sigma_1\alpha, \qquad \operatorname{var}(|x|) = \sigma_1^2(1 - \alpha^2)$$

$$E(|y|) = \sigma_2\alpha, \qquad \operatorname{var}(|y|) = \sigma_2^2(1 - \alpha^2)$$

$$E(x|0 < x < \infty) = \sigma_1\alpha, \qquad E(x^2|0 < x < \infty) = \sigma_1^2$$

where

$$\alpha^2 = \frac{2}{\pi}$$

It is also well known that

$$E(x|y) = \rho\frac{\sigma_1}{\sigma_2}y, \qquad E(x^2|y) = \sigma_1^2(1 - \rho^2) + \rho^2\frac{\sigma_1^2}{\sigma_2^2}y^2$$

To derive the mean and variance of x sgn y, we consider the conditional random variable $x|0 < y < \infty$. One has

$$E(x|0 < y < \infty) = E[E(x|y|0 < y < \infty)] = E\left[\rho\frac{\sigma_1}{\sigma_2}y|0 < y < \infty\right] = \rho\frac{\sigma_1}{\sigma_2}\cdot\sigma_2\alpha = \rho\sigma_1\alpha$$

$$E(x^2|0 < y < \infty) = E[E(x^2|y|0 < y < \infty)] = E\left[\sigma_1^2(1 - \rho^2) + \rho^2\frac{\sigma_1^2}{\sigma_2^2}y^2|0 < y < \infty\right] = \sigma_1^2(1 - \rho^2) + \frac{\rho^2\sigma_1^2}{\sigma_2^2}\cdot\sigma_2^2 = \sigma_1^2$$

Thus,

$$\operatorname{var}(x|0 < y < \infty) = \sigma_1^2 - \rho^2\sigma_1^2\alpha^2 = \sigma_1^2(1 - \rho^2\alpha^2)$$

Similarly,

$$E(x| -\infty < y < 0) = -\rho\sigma_1\alpha$$

$$\operatorname{var}(x| -\infty < y < 0) = \sigma_1^2(1 - \rho^2\alpha^2)$$

It now becomes obvious that one has

$$E(x \operatorname{sgn} y) = \rho\sigma_1\alpha, \qquad \operatorname{var}(x \operatorname{sgn} y) = \sigma_1^2(1 - \sigma^2\alpha^2)$$

$$E(y \operatorname{sgn} x) = \rho\sigma_2\alpha, \qquad \operatorname{var}(y \operatorname{sgn} x) = \sigma_2^2(1 - \rho^2\alpha^2)$$

Now let

$$\sum_{i=1}^{n} x_i \operatorname{sgn} y_i = u_1, \qquad \sum_{i=1}^{n} y_i \operatorname{sgn} x_i = u_2$$

$$\sum_{i=1}^{n} |x_i| = v_1, \qquad \sum_{i=1}^{n} |y_i| = v_2$$

Eq. (1) can then be written as

$$\hat{\rho} = \frac{1}{2}\left[\frac{u_1}{v_1} + \frac{u_2}{v_2}\right]$$

As an approximation to the variance of $\hat{\rho}$, we will take the asymptotic variance. Thus, one has

$$\operatorname{var}(\hat{\rho}) = \frac{1}{4}\left[\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial\hat{\rho}}{\partial u_i}\frac{\partial\hat{\rho}}{\partial u_j}\operatorname{cov}(u_i,u_j) + \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial\hat{\rho}}{\partial v_i}\frac{\partial\hat{\rho}}{\partial v_j}\operatorname{cov}(v_i,v_j) + 2\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial\hat{\rho}}{\partial u_i}\frac{\partial\hat{\rho}}{\partial v_j}\operatorname{cov}(u_i,v_j)\right] \tag{2}$$

where each partial derivative is to be evaluated at $u_i = E(u_i)$ and $v_i = E(v_i)$, for $i = 1, 2$. Evaluation of the partial derivations gives

$$\frac{\partial\hat{\rho}}{\partial u_1} = \frac{1}{n\sigma_1\alpha}, \qquad \frac{\partial\hat{\rho}}{\partial u_2} = \frac{1}{n\sigma_2\alpha}$$

$$\frac{\partial\hat{\rho}}{\partial v_1} = \frac{-\rho}{n\sigma_1\alpha}, \qquad \frac{\partial\hat{\rho}}{\partial v_2} = \frac{-\rho}{n\sigma_1\alpha}$$

One also has,

$$\operatorname{var}(u_i) = n\sigma_i^2(1-\rho^2\alpha^2), \operatorname{var}(v_i) = n\sigma_i^2(1-\alpha^2), i = 1,2$$

$$\operatorname{cov}(u_1,v_1) = n\operatorname{cov}(x\operatorname{sgn} y, |x|)$$

$$\operatorname{cov}(u_2,v_2) = n\operatorname{cov}(y\operatorname{sgn} x, |y|)$$

$$\operatorname{cov}(u_1,v_2) = n\operatorname{cov}(x\operatorname{sgn} y, |y|)$$

$$\operatorname{cov}(u_2,v_1) = n\operatorname{cov}(y\operatorname{sgn} x, |x|)$$

$$\operatorname{cov}(u_1,u_2) = n\operatorname{cov}(x\operatorname{sgn} y, y\operatorname{sgn} x)$$

$$\operatorname{cov}(v_1,v_2) = n\operatorname{cov}(|x|,|y|)$$

We will illustrate a method of computing the above covariances by deriving cov $(x \operatorname{sgn} y, |x|)$ in some de*a*il. Noting that

$$x\operatorname{sgn} y\cdot|x| = \begin{cases} x^2 \text{ if } x, y \gtrless 0 \\ -x^2 \text{ if } x \gtrless 0, y \lessgtr 0 \end{cases}$$

one has

$$E(x \operatorname{sgn} y \cdot |x|) = \frac{2}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \left\{ \int_0^\infty \int_0^\infty x^2 \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma\ \sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right] dx\,dy \right.$$

$$\left. - \int_\infty^0 \int_0^\infty x^2 \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right] dx\,dy \right\} \qquad (3)$$

$$= \frac{4}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \int_0^\infty \int_0^\infty x^2 \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right] dx\,dy - \sigma_1^2 \qquad (4)$$

since the *sum* of the two integrals in Eq. (3) equals σ_1^2.

By means of the transformation $x = ty$, Eq. (4) becomes

$$E(x \operatorname{sgn} y \cdot |x|) = \frac{4}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \int_0^\infty \int_0^\infty t^2 y^3 \exp\left\{ -y^2 \left[\frac{1}{2(1-\rho^2)} \left(\frac{t^2}{\sigma_1^2} - \frac{2\rho t}{\sigma_1\sigma_2} + \frac{1}{\sigma_2^2} \right) \right] \right\} dy\,dt - \sigma_1^2$$

Integrating first with respect to y and then with respect to t results in

$$E(x \operatorname{sgn} y \cdot |x|) = \alpha^2\sigma_1^2 \left[\rho(1-\rho^2)^{1/2} + \frac{1}{\alpha^2} + \sin^{-1}\rho \right] - \sigma_1^2$$

$$\operatorname{cov}(x \operatorname{sgn} y, |x|) = \alpha^2\sigma_1^2 \left[\rho(1-\rho^2)^{1/2} + \frac{1}{\alpha^2} + \sin^{-1}\rho \right] - \sigma_1^2 - \sigma_1^2\rho\alpha^2 = \sigma_1^2\alpha^2 [\rho(1-\rho^2)^{1/2} + \sin^{-1}\rho - \rho]$$

In a similar manner, one obtains the following:

$$\operatorname{cov}(y \operatorname{sgn} x, |y|) = \sigma_2^2\alpha^2 [\rho(1-\rho^2)^{1/2} + \sin^{-1}\rho - \rho]$$

$$\operatorname{cov}(x \operatorname{sgn} y, |y|) = \operatorname{cov}(y \operatorname{sgn} x, |x|) = \rho\sigma_1\sigma_2(1-\rho^2)$$

$$\operatorname{cov}(x \operatorname{sgn} y, y \operatorname{sgn} x) = \alpha^2\sigma_1\sigma_2 [(1-\rho^2)^{1/2} + \rho\sin^{-1}\rho - \rho^2]$$

$$\operatorname{cov}(|x|, |y|) = \alpha^2\sigma_1\sigma_2 [(1-\rho^2)^{1/2} + \rho\sin^{-1}\rho - 1]$$

Substituting the above expressions in Eq. (2) and simplifying finally results in

$$\operatorname{var}(\hat{\rho}) = \frac{(1-\rho^2)}{2n} \left[\frac{\pi}{2} + (1-\rho^2)^{1/2} - \rho\sin^{-1}\rho \right]$$

The maximum-likelihood estimator of ρ when the variances are unknown and the means are both zero is given by

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 \right)^{1/2}}$$

The asymptotic variance of r is

$$\operatorname{var}(r) = \frac{(1-\rho^2)^2}{n}$$

Defining the efficiency of $\hat{\rho}$ as

$$\operatorname{eff}(\hat{\rho}) = \frac{\operatorname{var}(r)}{\operatorname{var}(\hat{\rho})}$$

Table 6 gives the variance of $\hat{\rho}$ and its efficiency for values of ρ between 0 and 0.9.

Table 6. Variance and efficiency of $\hat{\rho}$

ρ	n var $(\hat{\rho})$	eff $(\hat{\rho})$
0.0	1.2854	0.778
0.1	1.2651	0.775
0.2	1.2050	0.765
0.3	1.1072	0.748
0.4	0.9755	0.723
0.5	0 8156	0.690
0.6	0.6351	0.645
0.7	0.4443	0.585
0.8	0.2872	0.504
0.9	0.09989	0.380

3. Estimating ρ

Two sets of samples, $\{x_i\}$ and $\{y'_i\}$, each containing 200 sample values, were drawn from a table of random numbers in which the entries are independent and distributed $N(0, 1)$. The transformation

$$y_i = 0.8\,x_i + 0.6\,y'_i$$

was then performed. Consequently, each x_i and y_i can be assumed to be distributed $N(0, 1)$ with a correlation of 0.8. Then $\hat{\rho}$ and r were calculated and found to be

$$\hat{\rho} = 0.8221$$

$$r = 0.8235$$

Two new sets of 200 values each were drawn from the same table of random numbers and paired at random, so that one can assume that $\rho = 0$. The results for this case were

$$\hat{\rho} = 0.0828$$

$$r = 0.0559$$

J. Information Processing: The Distribution of the Ratio of Two Jointly Normal Random Variables, *I. Eisenberger*

1. Introduction

Let x and y be random variables, distributed $N(m_1, \sigma_1^2)$ and $N(m_2, \sigma_2^2)$, respectively. The distribution of the ratio $t = x/y$ is derived in Ref. 1, under the assumption that x and y are independent. In that report, the hypothesis that $\sigma_2 = \sigma_1$ was tested, using quantiles, against the alternative hypotheses that $\sigma_2 = \theta\sigma_1$, when σ_1 was unknown. The test statistics that were used in order to eliminate dependence on σ_1 were ratios of the sums of two sets of quantiles and, in order to specify a critical region, it was necessary to determine the distribution of t. In this note, we derive the distribution of t under the assumption that x and y are jointly normal with correlation ρ. It will be shown that the density function, $g(t)$, is given by

$$g(t) = \frac{\sigma_1\sigma_2(1-\rho^2)^{1/2}\exp\left\{-\frac{1}{2}\left[\frac{m_1^2\sigma_2^2 - 2\rho\, m_1 m_2\,\sigma_1\sigma_2 + m_2^2\sigma_1^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}\right]\right\}}{\pi(\sigma_2^2t^2 - 2\rho\,\sigma_1\sigma_2 t + \sigma_1^2)} + \frac{\sigma_2 t(m_1\sigma_2 - \rho\, m_2\sigma_1) + \sigma_1(m_2\sigma_1 - \rho\, m_1\sigma_2)}{(2\pi)^{1/2}(\sigma_2^2t^2 - 2\rho\,\sigma_1\sigma_2 t + \sigma_1^2)^{3/2}}$$

$$\times \exp\left\{-\frac{1}{2}\left[\frac{(m_2t - m_1)^2}{\sigma_2^2t^2 - 2\rho\,\sigma_1\sigma_2 t + \sigma_1^2}\right]\right\}\left\{-1 + 2F\left[\frac{\sigma_2 t(m_1\sigma_2 - \rho m_2\sigma_1) + \sigma_1(m_2\sigma_1 - \rho m_1\sigma_2)}{\sigma_1\sigma_2(1-\rho^2)^{1/2}(\sigma_2^2t^2 - 2\rho\,\sigma_1\sigma_2 t + \sigma_1^2)^{1/2}}\right]\right\} \qquad (1)$$

where

$$F(x) = \frac{1}{(2\pi)^{1/2}}\int_{-\infty}^{x} e^{-1/2\,t^2}\,dt$$

The distribution of the random variable t may also be a useful approximation when one is considering the distribution of the ratio of the sums of two sets of sample values

$$R = \frac{\sum_{i=1}^{n_1} u_i}{\sum_{i=1}^{n_2} w_i}$$

where n_1 and n_2 are large. The central limit theorem, when applicable, assures us that the numerator and denominator of R are approximately normal and hence the distribution of R can be approximated by the distribution of t. It should be observed, however, that, whereas no moment of t of positive order is finite, the moments of R may exist.

2. Discussion

The joint density function of x and y is given by

$$g_1(x,y) = K\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - \frac{2\rho(x-m_1)(x-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}$$

where

$$K = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}}$$

Putting $x = ty$, one sees that, since the Jacobian of the transformation is $|y|$, the joint density of t and y is given by

$$g_2(t,y) = K|y|\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(ty-m_1)^2}{\sigma_1^2} - \frac{2\rho(ty-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\} \qquad \begin{matrix} -\infty < y < \infty \\ -\infty < t < \infty \end{matrix}$$

$$= K|y|\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{y^2(\sigma_2^2t^2 - 2\rho\sigma_1\sigma_2 t + \sigma_1^2)}{\sigma_1^2\sigma_2^2} - \frac{2y[\sigma_2 t(m_1\sigma_2 - \rho m_2\sigma_1) + \sigma_1(m_2\sigma_1 - \rho m_1\sigma_2)]}{\sigma_1^2\sigma_2^2}\right.\right.$$

$$\left.\left. + \frac{m_1^2\sigma_2^2 - 2\rho\, m_1 m_2\,\sigma_1\sigma_2 + m_2^2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right]\right\}$$

$$= K|y|\exp\left[-\frac{A}{2}y^2 + By + C\right] \tag{2}$$

where

$$A = \frac{\sigma_2^2t^2 - \rho\,\sigma_1\sigma_2 t + \sigma_1^2}{(1-\rho^2)\,\sigma_1^2\sigma_2^2}$$

$$B = \frac{\sigma_2 t(m_1\sigma_2 - \rho\, m_2\sigma_1) + \sigma_1(m_2\sigma_1 - \rho\, m_1\sigma_2)}{(1-\rho^2)\,\sigma_1^2\sigma_2^2}$$

$$C = \frac{-(m_1^2\sigma_2^2 - 2\rho\, m_1m_2\,\sigma_1\sigma_2 + m_2^2\sigma_1^2)}{2(1-\rho^2)\,\sigma_1^2\sigma_2^2}$$

By completing the square in y, Eq. (2) becomes

$$K|y|\exp\left[-\frac{A}{2}\left(y - \frac{B}{A}\right)^2 + C + \frac{B^2}{2A}\right] \tag{3}$$

The density function of t can now be obtained by integrating out y in Eq. (3). Accordingly,

$$g(t) = K\exp\left(C + \frac{B^2}{2A}\right)\left\{\int_0^\infty y\exp\left[-\frac{A}{2}\left(y-\frac{B}{A}\right)^2\right]dy - \int_{-\infty}^0 y\exp\left[-\frac{A}{2}\left(y-\frac{B}{A}\right)^2\right]dy\right\} \tag{4}$$

By use of the transformation $Z = (A)^{1/2}(y - B/A)$, Eq. (4) becomes

$$g(t) = \frac{K\exp\left(C + \frac{B^2}{2A}\right)}{(A)^{1/2}}\left[\int_{-B/(A)^{1/2}}^{\infty}\left(\frac{Z}{(A)^{1/2}} + \frac{B}{A}\right)e^{-1/2Z^2}\,dZ - \int_{-\infty}^{-B/(A)^{1/2}}\left(\frac{Z}{(A)^{1/2}} + \frac{B}{A}\right)e^{-1/2Z^2}\,dZ\right]$$

$$= \frac{\exp(C)}{\pi\sigma_1\sigma_2 A(1-\rho^2)^{1/2}} + \frac{B\exp\left(C + \frac{B^2}{2A}\right)}{\sigma_1\sigma_2 A^{3/2}[2\pi(1-\rho^2)]^{1/2}}\left[-1 + 2F\left(\frac{B}{(A)^{1/2}}\right)\right]$$

which, after simplification, becomes Eq. (1).

If $m_1 = m_2 = 0$, $g(t)$ takes on the relatively simple form

$$g(t) = \frac{\sigma_1\sigma_2(1-\rho^2)^{1/2}}{\pi(\sigma_2^2 t^2 - 2\rho\,\sigma_1\sigma_2 t + \sigma_1^2)} \qquad (5)$$

The transformation $v = \sigma_2 t$ converts $g(t)$ in Eq. (5) to the density function of a Cauchy distribution of the form

$$h(v) = \frac{\lambda}{\pi[\lambda^2 + (v-\mu)^2]}$$

where, in this case,

$$\lambda = \sigma_1(1-\rho^2)^{1/2}$$

$$\mu = \rho\sigma_1$$

Reference

1. Eisenberger, I., *Tests of Hypotheses and Estimation of the Correlation Coefficient using Quantiles I*, Technical Report 32-718. Jet Propulsion Laboratory, Pasadena, Calif., June 1, 1965.

K. Astrometrics: Pulsar Observations, *R. M. Goldstein*

1. Introduction

Two of the recently discovered (Ref. 1) pulsating radio sources, or pulsars, have been observed at the Jet Propulsion Laboratory's Goldstone Deep Space Communication Complex (Mars deep space station). The signals from these pulsars are known (Refs. 1 and 2) to have extremely regular repetition periods, although the amplitude within a pulse and from pulse to pulse varies erratically. The radio frequency of each pulse has been observed (Ref. 3) to decrease with time, following the dispersion relationship of electromagnetic propagation through a medium containing free electrons. Presumably, the signals near the source contain a wide band of frequencies. Since the group velocity for waves in such a medium is less for the lower frequencies, the received signals have the form of a sliding tone, or whistle, with the higher frequencies arriving before the lower.

2. Magnetic Field Measurement

The familiar equation for index of refraction (Ref. 3) is

$$n^2 = 1 - \frac{\frac{Ne^2}{m\epsilon_0}}{\omega^2 \pm \frac{e}{m}B\omega} \qquad (1)$$

where N is the electron density, ω is 2π times the frequency, e is the electron charge, m is the electron mass, ϵ_0 is the permittivity of space, and B is the component of any magnetic field along the line of sight. The ± sign depends on the relation of the direction of the circularly polarized waves to the direction of the magnetic field.

This fact gives us the possibility of measuring directly the interstellar magnetic field, averaged along the line of sight. By observing the change of the time of arrival of the pulses with the antenna switched from left- to right-handed circular polarization, a measure of the field is obtained. Although this method is not as sensitive as the utilization of Faraday rotation (Ref. 4), it does not require a polarized source.

3. The Data

The data collected is in the form of spectrograms of the signals. A bandwidth of 3 MHz, centered at 84 MHz, and with a resolution of 50 kHz was investigated. Time was divided into 15.4-ms slices, and an independent spectrogram was taken for each slice. Because of the periodic nature of the pulsars, the signal-to-noise ratio can be enhanced greatly by averaging together corresponding sets of spectra from many pulses.

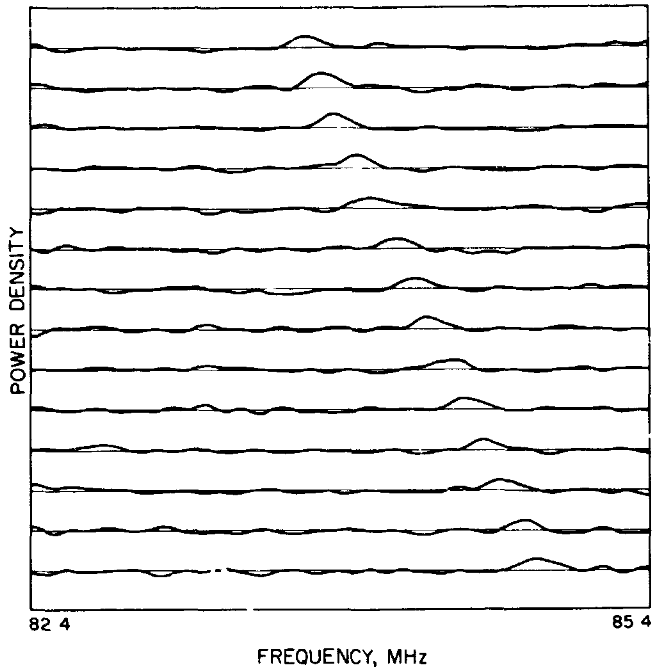


Fig. 26. Set of spectra of CP 1919 taken in successive 15.4-ms time intervals and averaged over 1350 pulses

A sample set of spectra from CP 1919 is given in Fig. 26. It shows the time–frequency history of the signals, averaged over 1350 pulses. Signals from CP 1919 are seen to enter the spectrograms from the high-frequency side and move rapidly through them towards the low. It follows from Eq. (1) that, if the observed effect is indeed caused by dispersion, the relationship between f and t is

$$f = \frac{k}{(t - t_0)^{1/2}} \tag{2}$$

The data from each set of spectrograms was processed to determine the constants k and t_0 by the method of least squares. The central frequency of the pulse in each spectrum was obtained by convolving the data with the expected pulse shape—a maximum-likelihood procedure if the shape is perfectly known.

The results of the least square fit is given in Fig. 27. As can be seen, there is a close fit to the theoretical curve. Note that, at 83 MHz, the pulse has been delayed (dispersed) by almost 7½ s. Table 7 summarizes the values of k obtained, along with the corresponding frequency sweep rates and integrated electron densities.

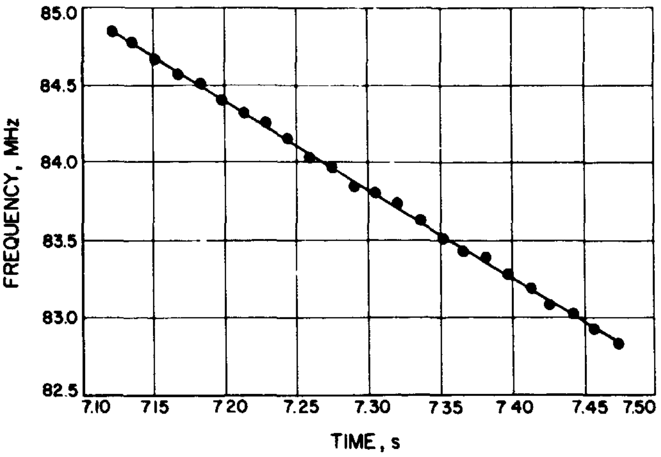


Fig. 27. Least square fit of the function $f = k/(t - t_0)^{1/2}$ to the data from CP 1919

From Eq. (1), it follows that the change in time of arrival, ΔT, that occurs when the mode of circular polarization is switched is

$$\Delta T = \frac{4(t - t_0)eB}{\omega m}$$

We found that the measured ΔT was not statistically significant for either source. However, an upper limit for the integrated magnetic field can be set. From the standard deviation of the time-of-arrival estimates (0.0006 s), that of the magnetic field measurement is found to be

$$\pm 0.62 \times 10^{-3}\,\mathrm{G}$$

The time-of-arrival measurements have also allowed us to determine the repetition period of the pulses to surprising accuracy. A very small difference of timing between the pulses and the signal sampling equipment produces a cumulative drift of the time-frequency trajectory of the pulses. Our measurements, corrected for the earth's orbital velocity and rotation, are given in Table 7. For CP 1919, the period matches very closely to that published in Refs. 5 and 6, in distinction to that of Ref. 1.

Table 7. Values of k, sweep rates, and integrated electron densities

Pulsar	Frequency, MHz	df/dt at 84 MHz, MHz/s	$\int N\,dl$, pc/cm^3	Period, s
CP 1919	$226/(t)^{1/2}$	5.81	12.4	1.3373008 ± 3
CP 0834	$231\,(t)^{1/2}$	5.56	12.9	1.2737620 ± 3

Using the best-fit relation (Eq. 2), we displaced each spectrum of a set to a common frequency origin and then averaged them. The results are given in Figs. 28a and b. These figures, then, show the spectral characteristics of the average pulse.

The frequency structure of these two sources is quite similar to the time structure already reported (Ref. 7). They both have a basic triangular shape and sudden onset and termination. There was no evidence of any power outside of the main pulse.

The peak power density, average power, and average bandwidth of these pulsars are given in Table 8.

References

1. Hewish, A., et al, *Nature*, Vol. 217, p. 709, 1968.
2. Davies, J. G., et al, *Nature*, Vol. 217, p. 910, 1968.
3. Stratton, J. A., *Electromagnetic Theory*, p. 329. McGraw-Hill Book Co., New York, 1941.
4. Smith, F. G., *Nature*, Vol. 218, p. 325, 1968.
5. Radhakrishnan, V., et al, *Nature*, Vol. 218, p. 229, 1968.
6. Moffet, A. T., and Ekers, R. D., *Nature*, Vol. 218, p. 227, 1968.
7. Lyne, A. G., and Rickett, B. J., *Nature*, Vol. 218, p. 326, 1968.

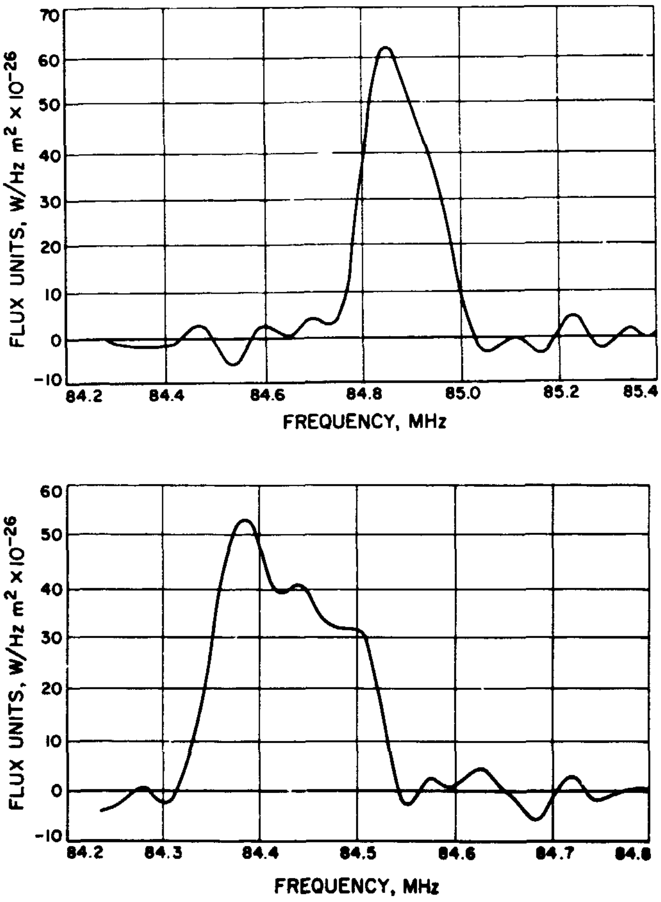


Fig. 28. Instantaneous spectrum of the average pulse of: (a) CP 1919, and (b) CP 0834

Table 8. Peak power density, average power, and average bandwidth

Pulsar	Peak power density, ω/Hz/m^2 × 10^{-26}	Average power, ω/m^2 × 10^{-21}	Average bandwidth, kHz
CP 1919	63	49	77
CP 0834	53	36	69

L. Astrometrics: Optimum Range Gates, *A. Garsia,*[3] *E. Rodemich, and H. Rumsey, Jr.*

1. Introduction

Let $\mathcal{F}_\delta$ denote the family of functions $\Lambda(x)$ satisfying the following conditions:

$\Lambda(x)$ is positive definite and continuous on the real axis (1a)

$$\Lambda(x) = 0 \;\forall\; |x| \geqq \delta \tag{1b}$$

$$\Lambda(0) = 1 \tag{1c}$$

Our problem is to calculate

$$C_\delta = \max_{\Lambda \in \mathcal{F}_\delta} \int_{-\delta}^{\delta} |\Lambda(x)|^2 \, dx \tag{2}$$

This question has arisen in trying to maximize the average power of the received signal in JPL's planetary radar system. The main conclusion is that the present system is nearly optimum from the analytic standpoint and certainly the best from the standpoint of equipment simplicity. In radar mapping, $\Lambda(x)$ depends on the hardware used. Since

$$\int_{-\delta}^{\delta} |\Lambda(x)|^2 \, dx \tag{3}$$

is proportional to the average power received, per unit power sent, any $\Lambda(x)$ which maximizes this integral would correspond to a best possible hardware. The present version uses $\Lambda(x)$, a triangle function obtained as the correlation function of a maximum-length shift-register sequence.

[3] Consultant, Mathematics Department, University of California, San Diego, California.

Let $\mathcal{C}_\delta$ be the family of functions $\Lambda(x)$ satisfying

$$\Lambda(x) \quad \text{is positive definite and periodic of period } 2\pi \tag{4a}$$

$$\Lambda(x) = 0 \text{ for } \delta \leqq |x| \leqq \pi \tag{4b}$$

$$\Lambda(0) = 1 \tag{4c}$$

A companion to the above problem is that of finding

$$D_\delta = \max_{\Lambda \in \mathcal{C}_\delta} \int_{-\delta}^{\delta} |\Lambda(x)|^2 \, dx$$

We shall see that the two problems are related and, indeed, when $\delta < \pi$,

$$C_\delta < D_\delta$$

Furthermore, it can be shown that

$$\lim_{\delta \to 0} \frac{C_\delta}{D_\delta} = 1$$

It can be seen that both these problems are special cases of a general question which can be formulated on a large class of abelian groups. The first arises when the group in question is the real line, and the second arises when the group is the circle.

We shall not go deeper here into these matters, but we shall be guided by these considerations and refer to the first problem as the "line" case and the second as the "circle" case. When the integers or the Nth roots of unity (for a fixed N) are taken as the basic group, we obtain two problems which are closely related to the line and circle cases, respectively.

The case of the "integers" can be stated as follows. We define $\mathcal{L}_N$ as the family of sequences $\{a_n\}$ such that

$$\{a_n\} \quad \text{is positive definite} \tag{5a}$$

$$a_n = 0 \ \forall \ |n| > N \tag{5b}$$

$$a_0 = 1 \tag{5c}$$

We then seek

$$I_N = \max_{\{a_n\} \in \mathcal{L}_N} \sum_{\nu=-N}^{N} a_\nu^2$$

It can be shown that $\forall N \geqq 1$

$$C_1 \leqq \frac{1}{N+1} I_N \tag{6}$$

and indeed

$$\lim_{N \to \infty} \frac{1}{N+1} I_N = C_1 \tag{7}$$

Inequality (6) can be used to get some very sharp upper bounds for C_1 and, therefore, since (as it can be easily shown) $C_\delta = \delta C_1$, also upper bounds for $C_\delta \forall \delta > 0$.

In this article, we shall establish, among other things, that for the line problem there is a unique maximizing function and that this maximizing function can be calculated to any degree of accuracy by a successive approximation method which is suitable to use with a computer. We have not succeeded in finding an explicit formula for the maximizing function, although such a function can be shown to have some rather remarkable properties. In fact, we shall see that the extremal function for the line case is also the solution of several other maximum problems.

The circle case is open. At this moment, we are not in possession even of an existence proof let alone uniqueness for the maximizing function. As we shall see, the circle case leads to some very interesting, and as far as we know, unsolved problems for the circle group.

2. The Factorization

A wide variety of functions of $\mathcal{G}_\delta$ and $\mathcal{C}_\delta$ can be obtained as follows. For the line case, we start with a real or complex valued function $\beta(x)$ which satisfies

$$\beta(x) \quad \text{is defined in } (-\infty, +\infty) \text{ and is square integrable} \tag{8a}$$

$$\beta(x) = 0 \ \forall \ |x| \geqq \frac{\delta}{2} \tag{8b}$$

$$\int_{-\delta/2}^{\delta/2} |\beta(x)|^2 \, dx = 1 \tag{8c}$$

then set

$$\Lambda(x) = \int_{-\infty}^{+\infty} \beta(x+t)\overline{\beta(t)} \, dt \tag{9}$$

It is easily seen that such $\Lambda(x) \epsilon \mathcal{P}_\delta$. Indeed, Properties (1b) and (1c) are obvious and (1a) follows from the identity

$$\sum_{ij} \Lambda(x_i - x_j)\,\xi_i \bar{\xi}_j = \int_{-\infty}^{+\infty} \left| \sum_i \beta(x_i + t)\,\xi_i \right|^2 dt$$

Similarly, in the circle case, we start with a function $\alpha(x)$ satisfying

$$\alpha(x) \quad \text{is periodic of period } 2\pi \text{ and square integrable} \tag{10a}$$

$$\alpha(x) = 0 \text{ for } \frac{\delta}{2} \leqq |x| \leqq \pi \tag{10b}$$

$$\int_{-\delta/2}^{\delta/2} |\alpha(x)|^2\,dx = 1 \tag{10c}$$

We then set

$$\Gamma(x) = \int_{-\pi}^{\pi} \alpha(x+t)\,\overline{\alpha(t)}\,dt$$

Again, it is easily shown that $\Gamma(x) \epsilon \mathcal{C}_\delta$ for any such choice of $\alpha(x)$.

It is compelling at this point to ask whether or not such representations are always possible for functions of $\mathcal{P}_\delta$ and $\mathcal{C}_\delta$. It is clear that this would introduce a considerable simplification on our maximum problem, since the conditions on $\beta(x)$ and $\alpha(x)$ are very simple and easy to handle.

However, the remarkable fact which distinguishes the line case from the circle case is that this factorization holds in the former but not in the latter case.

The factorization result can be stated as follows:

Theorem 1. Given a function $\Lambda(x)$ *which is positive definite on the real axis and vanishes for* $|x| \geqq \delta$, *then there is a function* $\beta(x)$ *which is square integrable and vanishes for* $|x| > \delta/2$ *such that*

$$\Lambda(x) = \int_{-\infty}^{+\infty} \beta(x+t)\,\overline{\beta(t)}\,dt \tag{11}$$

This result can be stated and proved as a theorem on entire functions of exponential type (Ref. 1, pp. 124–126). In this form, it is also stated without proof in a paper of Krein (Ref. 2). However, there is no need to use such sophisticated tools. We give a proof in this article using a method which we discovered quite independently of the above mentioned works and which yields at the same time an interesting viewpoint. We also indicate briefly why the corresponding factorization is not, in general, possible in the periodic case. From these considerations, it follows that (when $\delta \leqq \pi$)

$$C_\delta \leqq D_\delta$$

Examples may also be given which show that strict inequality holds.

Some sort of substitute for the factorization can be established in the periodic case. It reads as follows: every function of C_δ can be written in the form

$$\Gamma(x) = \int_{-\pi}^{\pi} \alpha(x+t)\,\bar{\Lambda}(t)\,dt \tag{12}$$

where $\Lambda(t) \epsilon \mathcal{P}_{\delta_1}$, $\alpha(x)$ is periodic and vanishes for $\delta_2 \leqq |x| \leqq \pi$, and $\delta_1 + \delta_2 \leqq \delta$. Furthermore, the Fourier transform

$$\tilde{\alpha}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{inx}\,\alpha(x)\,dx$$

has all its zeros on the real axis and is non-negative at every integer where the Fourier transform

$$\tilde{\Lambda}(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{inx}\,\Lambda(x)\,dx$$

is different from zero.

These conditions are necessary and sufficient for the function $\Gamma(x)$, given by Eq. (12), to belong to the class $\mathcal{C}_\delta$; unfortunately, they are not easy to work with.

3. Symmetrization

We have seen that every function $\Lambda(x) \epsilon \mathcal{P}_\delta$ can be written in the form

$$\Lambda(x) = \int_{-\infty}^{+\infty} \beta(x+t)\,\overline{\beta(t)}\,dt \tag{13}$$

where $\beta(x)$ satisfies the conditions

$$\beta(x) = 0 \;\forall\; |x| \geqq \frac{\delta}{2} \tag{14a}$$

$$\int_{-\infty}^{+\infty} |\beta(x)|^2\,dx = 1 \tag{14b}$$

This result not only does simplity considerably our maximization problem but can also be used to narrow down our search for the maximal function.

Indeed, we show here β may, without loss, be restricted to be non-negative and symmetrically decreasing. More precisely, we show:

Theorem 2. Let $\mathcal{F}'_\delta$ be the subclass of $\mathcal{F}_\delta$ of functions $\Lambda(x)$ which admit a factorization of the form Eq. (13) with a $\beta(x)$ satisfying in addition to the conditions in Eq. (14) also

$$\beta(x) \geqq 0 \tag{15a}$$

$$\beta(x) = \beta(-x) \tag{15b}$$

$$\beta(x) \geqq \beta(y) \text{ when } 0 \leqq x \leqq y \tag{15c}$$

then

$$\max_{\Lambda \in \mathcal{F}_\delta} \int_{-\infty}^{+\infty} |\Lambda(x)|^2 dx = \max_{\Lambda \in \mathcal{F}'_\delta} \int_{-\infty}^{+\infty} |\Lambda(x)|^2 dx$$

4. Existence

Symmetrization yields a very quick path to existence of the maximizing function. Indeed, let

$$\Lambda_n(x) = \int_{-\infty}^{+\infty} \beta_n(x+t)\,\beta_n(t)\,dt \tag{16}$$

be a sequence of functions in the class $\mathcal{F}'_\delta$ (i.e., each β_n satisfies the Conditions (14) and (15) such that

$$\lim_{n\to\infty} \int_{-\infty}^{+\infty} |\Lambda_n(x)|^2 dx = C_\delta \tag{17}$$

Since each β_n is non-increasing for $x \geqq 0$,

$$x\beta_n^2(x) \leqq \int_0^x \beta^2(t)\,dt \leqq 1$$

so, by symmetry,

$$\beta_n(x) \leqq \frac{1}{(|x|)^{1/2}} \;\forall\; x \neq 0 \tag{18}$$

By a well known argument, we can produce a function $\beta(x)$ on $(-\infty, +\infty)$ that is symmetric and non-decreasing for $x \geqq 0$ and a subsequence $\{\beta_{n_k}(x)\}$ such that

$$\lim_{k\to\infty} \beta_{n_k}(x) = \beta(x) \tag{19}$$

at all points of continuity of $\beta(x)$.

Note that, by Eq. (19) and Fatou's lemma, $\beta(x)$ will also satisfy

$$\beta(x) = 0 \;\forall\; |x| \geqq \frac{\delta}{2} \tag{20a}$$

$$\int_{-\infty}^{+\infty} \beta^2(x)\,dx \leqq 1 \tag{20b}$$

Furthermore, from Ineq. (18), we get

$$\beta_{n_k}(x+t)\,\beta_{n_k}(t) \leqq \frac{1}{(|x+t|\,|t|)^{1/2}}$$

So at least, for $x \neq 0$, from Lebesgue's dominated convergence theorem, we get

$$\lim_{k\to\infty} \int_{-\infty}^{+\infty} \beta_{n_k}(x+t)\,\beta_{n_k}(t)\,dt = \int_{-\infty}^{+\infty} \beta(x+t)\,\beta(t)\,dt$$

In other words,

$$\lim_{k\to\infty} \Lambda_{n_k}(x) = \Lambda(x)$$

when

$$\Lambda(x) = \int_{-\infty}^{+\infty} \beta(x+t)\,\beta(t)\,dt$$

We have $|\Lambda_n(x)| \leqq 1 \;\forall\; n$, so again by dominated convergence

$$C_\delta = \lim_{k\to\infty} \int_{-\infty}^{+\infty} [\Lambda_{n_k}(x)]^2 dx = \int_{-\delta}^{\delta} [\Lambda(x)]^2 dx$$

However, this result, combined with Eq. (20) and the definition of C_δ, implies that equality must actually hold in Eq. (20b). Thus, $\Lambda(x)$ must belong to $\mathcal{F}'_\delta$ and indeed must be a maximizing function.

The above argument is the one by which existence of the maximizing function was first obtained. Later on, we found another path to existence which we shall present in the *Subsection* 5 since it follows a rather interesting and fruitful line of reasoning.

5. The Integral Equation, Another Path to Existence

Before proceeding with our second proof of existence, we should observe that the considerations at the end of

Subsection 4 yield a result which will be rather crucial for our later considerations, namely:

Theorem 3. If

$$\Lambda(x) = \int_{-\infty}^{+\infty} \beta(x+t)\,\beta(t)\,dt, \qquad \beta(x) = 0 \text{ for } |x| \geqq \frac{\delta}{2}$$

is a maximizing function in $\mathcal{G}_\delta$, then $\beta(x)$ satisfies the integral equation

$$C_\delta \beta(x) = \int_{-\delta/2}^{\delta/2} \Lambda(x-t)\,\beta(t)\,dt \qquad \forall\ |x| \leqq \frac{\delta}{2} \tag{21}$$

In *Subsection 6,* we shall show that this function is unique.

6. Uniqueness

It will be convenient at this point to work with $\delta = 1$. This involves no loss since we can always reduce ourselves to this case by a change of scale. We shall use here and in the discussion that follows $\mathcal{G}$ and C to mean $\mathcal{G}_1$ and C_1.

Our point of departure to show uniqueness of the maximizing function will be Eq. (21). However, we shall write it in the form[4]

$$C\beta(x) = \beta \times \beta \times \beta(x)\,\chi(x)\ \forall\ x \tag{22}$$

where χ indicates convolution and

$$\chi(x) = \begin{cases} 1 \text{ for } |x| \leqq \frac{1}{2} \\ 0 \text{ for } |x| > \frac{1}{2} \end{cases}$$

Let us say that a function $\beta(x)$ on $(-\infty, +\infty)$ is "admissible" if and only if it satisfies the conditions

$$\beta(x) = \beta(-x) \geqq 0\ \forall\ x \tag{23a}$$

$$\int_{-\infty}^{+\infty} \beta^2(x)\,dx = 1 \tag{23b}$$

$$\beta(x) \geqq \beta(y)\ \forall\ 0 \leqq x \leqq y \tag{23c}$$

$$\beta(x) = 0\ \forall\ |x| \geqq \frac{1}{2} \tag{23d}$$

Our uniqueness result can then be stated as follows:

Theorem 4. There is at most one admissible function $\beta(x)$ which satisfies the equation

$$\lambda\beta(x) = \beta \times \beta \times \beta(x)\,\chi(x) \tag{24}$$

for some $\lambda \geqq 2/3$.

From this result, it follows immediately that there is only one function $\Lambda(x)$ in each class $\mathcal{G}'_\delta$ (see *Subsection 3* for definition of $\mathcal{G}'_\delta$) which maximizes the integral

$$I(\Lambda) = \int_{-\infty}^{+\infty} \Lambda^2(x)\,dx \tag{25}$$

Uniqueness in $\mathcal{G}_\delta$ can be established by showing that every $\Lambda \in \mathcal{G}_\delta$ for which $I(\Lambda) = C_\delta$ is necessarily in $\mathcal{G}'_\delta$. This given, it is easy to show that the equation

$$C_\delta \beta(x) = \beta \times \tilde{\beta} \times \beta(x)\,\chi(x), \qquad \tilde{\beta}(x) = \beta(-x)$$

has a unique solution $\beta(x)$ satisfying the conditions

$$\int_{-\infty}^{+\infty} \beta^2(x)\,dx = 1 \tag{26a}$$

$$\beta(x) = 0\ \forall\ |x| \geqq \frac{\delta}{2} \tag{26b}$$

7. The Successive Approximation Method

In this subsection, we shall denote our extremal function by β_0. We know that this function satisfies the equation

$$\lambda_0 \beta_0(x) = \beta_0 \times \beta_0 \times \beta_0(x)\,\chi(x),$$

$$\chi(x) = \begin{cases} 1 \text{ for } |x| \leqq \frac{1}{2} \\ 0 \text{ for } |x| > \frac{1}{2} \end{cases} \tag{27}$$

where λ_0 is the extremal constant.[5]

This given, we can try to obtain β_0 by a successive approximation method of the form

$$\lambda_n \beta_{n+1}(x) = \beta_n \times \beta_n \times \beta_n(x)\,\chi(x) \tag{28}$$

[4] Recall that $\beta(x)$, for the maximizing function, has been shown to be symmetric. Thus, $\Lambda(x) = \beta \times \beta(x)$.

[5] This is the constant we denoted by C_1 in the Introduction.

where λ_n is determined each time by the condition that

$$\| \beta_{n+1} \| = \left[\int_{-\frac{1}{2}}^{\frac{1}{2}} \beta_{n+1}^2 (x)\, dx \right]^{\frac{1}{2}} = 1$$

This is indeed the method we shall use, and we shall show that the iterates in Eq. (28) do converge geometrically to β_0 when the initial function β_1 is taken to be sufficiently close to β_0 itself, in particular when $\beta_1 = \chi$.

This is an obvious method to use; however, the estimates needed to complete it are not so obvious and are rather delicate.

Let us introduce for each $n \geqq 1$ the function

$$\Delta_n (x) = \beta_n (x) - \theta_n \beta_0 (x) \tag{29}$$

where

$$\theta_n = (\beta_n, \beta_0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \beta_n (x)\, \beta_0 (x)\, dx \tag{30}$$

Since β_0 is normalized, we see that for each n

$$(\Delta_n, \beta_0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \Delta_n (x)\, \beta_0 (x)\, dx = 0$$

Our goal is to show that under suitable circumstances we have

$$\| \Delta_{n+1} \| \leqq \rho \| \Delta_n \| \ \forall \ n \tag{31}$$

for some constant $0 < \rho < 1$. This, of course, implies that

$$\| \beta_n - \beta_0 \| \rightarrow 0$$

geometrically as $n \rightarrow \infty$.

For each admissible β set

$$F (\beta) = \beta \times \beta \times \beta (x)\, \chi (x) \tag{32}$$

Inequality (31) can then be written in the form

$$\| F (\beta_n) - (F (\beta_n), \beta_0)\, \beta_0 \| \leqq \rho \| \Delta_n \| \, \| F (\beta_n) \| \tag{33}$$

A simple geometric argument shows that for any $\alpha > 0$ and any F

$$\frac{\| F - (F, \beta_0)\, \beta_0 \|}{\| F \|} \leqq \frac{\| F - \alpha\beta_0 \|}{\alpha} \tag{34}$$

Indeed, Ineq. (34) (when it is not trivial) simply says that the sine of the angle between the directions of β_0 and F is always less than the sine of the angle between β_0 and the tangent to the circle through F with center at the point $\alpha\beta_0$.

This given, we can assure Ineq. (33) if we can find an α for which

$$\| F (\beta_n) - \alpha\beta_0 \| \leqq \alpha\rho \| \Delta_n \| \tag{35}$$

To this end, using the Relation (29) into (32), in view of (27), we get

$$F (\beta_n) = \{ \theta_n^3 \lambda_0 \beta_0 + 3\theta_n^2 \beta_0 \times \beta_0 \times \Delta_n + 3\theta_n \beta_0 \times \Delta_n \times \Delta_n + \Delta_n \times \Delta_n \times \Delta_n \}\, \chi$$

This suggests taking $\alpha = \theta_n^3 \lambda_0$ in Ineq. (35). With this choice of α, we get

$$\| F (\beta_n) - \alpha\beta_0 \| \leqq 3\theta_n^2 \| \chi\beta_0 \times \beta_0 \times \Delta_n \| + 3\theta_n \| \chi\beta_0 \times \Delta_n \times \Delta_n \| + \| \chi\Delta_n \times \Delta_n \times \Delta_n \| \tag{36}$$

We shall need to estimate the three terms on the right-hand side of this inequality as accurately as possible.

For the first term, we use the eigenfunction expansion

$$\beta_0 \times \beta_0 (x - y) = \lambda_0 \beta_0 (x)\, \beta_0 (y) + \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu (x)\, \phi_\nu (y) \tag{37}$$

and obtain, using the orthogonality of Δ_n and β_0,

$$\| \beta_0 \times \beta_0 \times \Delta_n \|^2 \leqq \lambda_e^2 \| \Delta_n \|^2 \tag{38}$$

where λ_e denotes the next largest eigenvalue corresponding to an even eigenfunction of the kernel $\beta_0 \times \beta_0 (x - y)$ in $[-1/2, 1/2] \times [-1/2, 1/2]$.

For the second term, we observe that by Schwarz's inequality

$$\int_{-1/2}^{1/2} \left[\int_{-1/2}^{1/2} \beta_0 \times \Delta_n (x - t) \Delta_n (t) \, dt \right]^2 dx \leqq \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} [\beta_0 \times \Delta_n (x - t)]^2 \, dt \, dx \, \| \Delta_n \|^2 = \beta_0 \times \beta_0 \times \Delta_n \times \Delta_n (0) \quad \| \Delta_n \|^2$$

Using again the eigenfunction expansion Eq. (37), we then get

$$\| \beta_0 \times \Delta_n \times \Delta_n \| \leqq (\lambda_e)^{1/2} \| \Delta_n \|^2 \tag{39}$$

The last term is easiest to estimate. We get

$$\| \Delta_n \times \Delta_n \times \Delta_n \| \leqq \lambda_0 \| \Delta_n \|^3 \tag{40}$$

We see that for the method to be accomplished we need

$$3\theta_n^2 \lambda_e \| \Delta_n \| + 3\theta_n (\lambda_e)^{1/2} \| \Delta_n \|^2 + \lambda_0 \| \Delta_n \|^3 \leqq \theta_n^3 \lambda_0 \rho \| \Delta_n \|$$

Simplifying and noticing that

$$\| \Delta_n \|^2 = 1 - \theta_n^2$$

we get

$$\frac{3}{\theta_n} \frac{\lambda_e}{\lambda_0} + \frac{3}{\theta_n^2} \left(\frac{\lambda_e}{\lambda_0} \right)^{1/2} \frac{1}{(\lambda_0)^{1/2}} (1 - \theta_n^2)^{1/2} + \frac{1}{\theta_n^3} (1 - \theta_n^2) \leqq \rho \tag{41}$$

In *Subsection 6*, we showed that $\lambda_e / \lambda_0 \leqq 1/8$ and we proved $\lambda_0 \geqq 2/3$. We see then that Ineq. (41) will be satisfied if

$$\frac{3}{8} \frac{1}{\theta_n} + \frac{3}{\theta_n^2} \left(\frac{1}{8} \right)^{1/2} \left(\frac{3}{2} \right)^{1/2} (1 - \theta_n^2)^{1/2} + \frac{1}{\theta_n^3} (1 - \theta_n^2) \leqq \rho \tag{42}$$

Note now, if we do establish this relation with a $\rho < 1$, then we shall have

$$1 - \theta_{n+1}^2 = \| \Delta_{n+1} \|^2 \leqq \rho^2 \| \Delta_n \|^2 < (1 - \theta_n^2)$$

In other words, the θ_n's increase.

However, since the function

$$g(\theta) = \frac{3}{8} \frac{1}{\theta} + \frac{3}{\theta^2} \left(\frac{1}{8} \right)^{1/2} \left(\frac{3}{2} \right)^{1/2} (1 - \theta^2)^{1/2} + \frac{1}{\theta^3} (1 - \theta^2)$$

decreases as θ increases, we see that in order for Ineq. (42) to hold for all n, all we need is to assure that it holds for $n = 1$.

From the form of $g(\theta)$, it is easy to see that if β_1 is sufficiently close to β_0, we shall have

$$g(\theta_1) < 1$$

To verify this relation for

$$\beta_1(x) = \chi(x) = \begin{cases} 1 \text{ for } |x| \leq \frac{1}{2} \\ 0 \text{ for } |x| > \frac{1}{2} \end{cases}$$

we need a careful estimate of θ for such a function. This can be achieved by means of the following inequality, the proof of which is immediate.

Lemma 1. Let $\phi(x)$ be measurable in $[\alpha, \beta]$ and let

$$0 < a \leq \phi(x) \leq b \;\; \forall \;\; x \in [\alpha, \beta]$$

$$\frac{1}{\beta - \alpha} \int_\alpha^\beta \phi^2(x)\, dx \geq 1$$

then

$$\frac{1}{\beta - \alpha} \int_\alpha^\beta \phi(x)\, dx \geq \frac{1 + ab}{a + b}$$

For our extremal function β_0, the hypotheses of the lemma are satisfied with $a = (2)^{1/2}/2$ and $b = 1/(\lambda_0)^{1/2}$. Indeed, $\beta_0(x)$ is symmetric around zero and does not increase away from zero. Furthermore, we know

$$\beta_0\left(\frac{1}{2}\right) = \beta_0\left(-\frac{1}{2}\right) = \frac{(2)^{1/2}}{2}$$

and from the eigenfunction expansion Eq. (37), we get

$$1 = \int_{-1/2}^{1/2} \beta_0^2(x)\, dx = \beta_0 \times \beta_0(0) \geq \lambda_0 \beta_0^2(0)$$

Simple arithmetic then gives

$$\theta_1 = (\beta_0, \chi) = \int_{-1/2}^{1/2} \beta_0(x)\, dx \geq 0.965$$

Substituting this value of θ_1 in $g(\theta)$, we then get

$$g(\theta_1) \leq 0.831$$

Thus, the convergence of the iterates in Eq. (28) for this choice of β_1 is established.

Remark 1. The method establishes more than the convergence of β_n to β_0. We can write

$$\|\beta_{n+1} - \beta_0\|^2 = \frac{2}{1 + \theta_{n+1}} \|\Delta_{n+1}\|^2$$

If we have

$$\|\Delta_{n+1}\|^2 \leq \rho \|\Delta_n\|$$

then, since this implies $\theta_{n+1} \geq \theta_n$, we get

$$\|\beta_{n+1} - \beta_0\|^2 \leq \rho^2 \frac{1 + \theta_n}{1 + \theta_{n+1}} \|\beta_n - \beta_0\|^2 \leq \rho^2 \|\beta_n - \beta_0\|^2$$

The triangle inequality then gives

$$\|\beta_n - \beta_0\| \leq \frac{1}{1 - \rho} \|\beta_{n+1} - \beta_n\| \;\; \forall \;\; n$$

In other words, we can tell how close we are to β_0 at any step of the iteration by seeing how close is β_{n+1} to β_n.

It is not difficult to see that the estimates presented in this subsection yield the following results:

Theorem 5. Let β be any admissible function and let $\theta = (\beta_0, \beta)$ if

$$g(\theta) = \frac{3}{8}\frac{1}{\rho} + \frac{3}{\theta^2}\left(\frac{1}{8}\right)^{1/2}\left(\frac{3}{2}\right)^{1/2}(1 - \theta^2)^{1/2} + \frac{1}{\theta^3}(1 - \theta^2) < 1$$

then

$$\|\beta - \beta_0\| \leq \frac{1}{1 - g(\theta)} \left\| \frac{F(\beta)}{\|F(\beta)\|_2} - \beta \right\|$$

Theorem 6. Let $F(x)$ be a (possibly nonlinear) operator on some hilbert space $\mathscr{H}$. Let β_0 be an eigenfunction of $F(x)$, i.e., let

$$F(\beta_0) = \lambda_0 \beta_0$$

Assume that for some $\delta > 0$

$$\rho = \sup_{\left\{\begin{array}{c}\|x - (x, \beta_0)\beta_0\| \leq \delta \\ x \in \mathscr{H}, \|x\| = 1\end{array}\right\}} \left(\inf_\alpha \frac{\|F(x) - \alpha\beta_0\|}{\alpha \|x - (x, \beta_0)\beta_0\|} \right) < 1$$

Then the sequence of iterates

$$\beta_{n+1} = \frac{F(\beta_n)}{\|F(\beta_n)\|}$$

does converge towards β_0, and indeed

$$\|\beta_n - \beta_0\| \leqq \rho^{n-1} \|\beta_1 - \beta_0\| \leqq \frac{\rho^{n-1}}{1-\rho} \|\beta_2 - \beta_1\|$$

provided

$$\|\beta_1 - (\beta_1\beta_0)\beta_0\| \leqq \delta$$

Using theorem 5, we were able to calculate our extremal constant to 24 decimal figures. The result ot this calculation gives

$$C_1 \cong 0.686981293033114600949413$$

The value attainable with the present range-gated radar is $C_1 = 2/3$, obtained with the triangular correlation function of a maximal-length shaft-register sequence. Thus, the present system is near optimum.

8. Some Final Remarks

It can be shown that our extremal function $\beta_0(x)$ is in $(-1/2, 1/2)$ the restriction of an entire function. This follows from the integral equation by successive differentiation and a judicious estimation of the resulting terms. Although straightforward, this calculation is quite intricate and is omitted.

It would be interesting if $\beta_0(x)$ could be expressed in terms of familiar functions or if C_1 itself turns out to be related to some of the classical constants.

It is interesting to note that our final result in *Subsection 7* can be put in the form

$$\|\beta_n - \beta_0\| \leqq \frac{\rho^{n-1}}{1-\rho} \|\beta_2 - \beta_1\| \qquad (43)$$

where $\beta_n (n = 1, \cdots)$ is the outcome of the nth iteration of the successive approximation method, β_0 is the function we want to calculate, and ρ is a constant we can explicitly estimate. Thus, we can calculate our unknown function and constant with any degree of accuracy.

However, our proof of Ineq. (43) is non-constructive. The same holds for our existence proof for β_0. This says that it is quite possible to obtain explicit estimates (thereby estimates that can be used in the applications) by entirely non-constructive arguments.

References

1. Boas, R. P., *Entire Functions*. Academic Press, New York, 1954.
2. Krein, M., *Comptes Rendus (Doklady) de l' Acad. des Sci. de l' U.R.S.S.*, Vol. XXVI, No. 1, pp. 17–22, 1940.

M. Data Compression Techniques: Product Entropy of Gaussian Distributions, *E. C. Posner, E. R. Rodemich, and H. Rumsey, Jr.*

1. Introduction

This article is a study of the product epsilon entropy of mean-continuous gaussian processes. That is, a given mean-continuous gaussian process on the unit interval is expanded into its Karhünen expansion. Along the kth eigenfunction axis, a partition by intervals of length ϵ_k is made, and the entropy of the resulting discrete distribution is noted. The infimum of the sum over k of these entropies subject to the constraint that $\Sigma\,\epsilon_k^2 \leqq \epsilon^2$ is the *product epsilon entropy* of the process. It is shown that the best partition to take along each eigenfunction axis is the one in which 0 is the midpoint of an interval in the partition. Furthermore, the product epsilon entropy is finite if and only if $\Sigma\,\lambda_k \log 1/\lambda_k$ is finite, where λ_k is the kth eigenvalue of the process. When the above series is finite, the values of ϵ_k which achieve the product entropy are found. Asymptotic expressions for the product epsilon entropy are derived in some special cases. The problem arises in the theory of data compression.

The work is motivated by the problem of data compression, the efficient representation of data for the purpose of information transmission. We shall consider the case in which the data to be represented consists of a sample function from a mean continuous gaussian process, X, on the unit interval. Our basic problem is how to transmit (over a noiseless channel) information as to which sample function of X occurred. We assume that the recipient of the transmitted data has full knowledge of the statistics of the process. In particular, he knows the Karhünen expansion (Ref. 1) of the process; namely

$$X(t) = \sum_{k=1}^{\infty} \lambda_k y_k \phi_k(t) \qquad (1)$$

where the y_k are mutually independent-unit normal random variables (they determine which sample function of the processes occurred); the $\phi_k(t)$ are the (orthonormal) eigenfunctions of the process; they are known *a priori*, as are the λ_k, which are the eigenvalues of the process. We note that the series in Eq. (1) converges with probability 1. If

$$R(s,t) = E[X(s)X(t)] \tag{2}$$

is the covariance function of the process, then

$$R(s,t) = \Sigma \lambda_k \phi_k(s) \phi_k(t) \tag{3}$$

the convergence being uniform on the unit square. Furthermore, $R(s,t)$ is jointly continuous in s and t. The functions ϕ_k are continuous and satisfy the integral equation

$$\lambda_k \phi_k(s) = \int_0^1 R(s,t)\phi_k(t)\,dt \tag{4}$$

where the λ_k are non-negative and are the eigenvalues of this integral equation. It follows that

$$\Sigma \lambda_k = \int_0^1 R(s,s)\,ds < \infty \tag{5}$$

In the special case when all but a finite number of the λ_k are zero, the process X is just a finite dimensional gaussian distribution. The interesting cases, from the point of product entropy, turn out to be the one-dimensional processes and the infinite-dimensional processes.

In the data compression problem, we wish to represent the sample functions of the known process X. By Eq. (1) we can fully describe a sample function $X(t)$ by specifying the values of the y_k which occur in Eq. (1). We shall call y_k the projection of the process along the kth coordinate axis.

Our final assumption concerning the nature of our problem is the requirement that the information which is transmitted must be adequate to locate the sample function in some set of (L_2) diameter at most ϵ.

The data compression procedure we propose is as follows: Observe $X(t)$ and compute its projections, y_k, along the coordinate axes. Quantize the kth coordinate axis into intervals of diameter at most ϵ_k. For each k, transmit the index of the interval which actually occurred. If the ϵ_k satisfy

$$\Sigma \epsilon_k^2 \leq \epsilon^2 \tag{6}$$

then, with probability 1, when the intervals of uncertainty are known, the original sample function is known to within a set which is a hyper-rectangle of diameter at most ϵ.

Our main concern in this article is to study the entropy of the above procedure. We observe that this entropy does not depend on the eigenfunctions, ϕ_k, of the process, but only on the eigenvalues, λ_k. This is because any two mean-continuous gaussian processes with the same λ_k possess measure-preserving isometries between the Hilbert spaces generated by their ϕ_k. It follows that assumptions about stationarity, band-limiting, etc., are relevant only insofar as they help estimate the λ_k.

A definition of epsilon entropy for mean-continuous stochastic processes is found in Ref. 2. The entropy defined in Ref. 2 is upper-bounded by the product epsilon entropy considered here; for it uses partitions by arbitrary measurable sets of diameter at most ϵ, instead of hyper-rectangles of diameter at most ϵ. It can be shown[6] that the epsilon entropy of a mean-continuous gaussian process on the unit interval is always finite. It turns out, however, that product entropy is finite if and only if $\Sigma \lambda_k \log 1/\lambda_k$ converges. A further discussion of data compression in a general setting is in preparation.[7]

Subsection 2 treats the one-dimensional case. We show that the best ϵ-partition (the ϵ-partition with least entropy) is that partition by intervals of length ϵ which contains the interval $(-\epsilon/2, \epsilon/2)$. We treat the cases of large and small ϵ separately. Techniques of analytic function theory are necessary.

In *Subsection 3*, we show that the product epsilon entropy, $J_\epsilon(X)$, of a mean-continuous gaussian process on the unit interval is finite if and only if $\Sigma \lambda_k \log 1/\lambda_k$ is finite. In case $J_\epsilon(X)$ is finite, we give a product partition whose entropy equals $J_\epsilon(X)$.

Subsection 4 gives an asymptotic form for $J_\epsilon(X)$ when the eigenvalues satisfy a relation of the form $\lambda_k \sim Bk^{-p}$. In particular, for the Weiner process, $J_\epsilon(X) \sim C/\epsilon^2$ as $\epsilon \to 0$, where C is a constant between 6 and 7.

Subsection 5 considers a general lower bound $L_\epsilon(X)$ for $J_\epsilon(X)$. We show that if

$$\sum_{k=n}^{\infty} \lambda_k = 0(n\lambda_n)$$

[6] Posner, E. C., Rodemich, E. R., and Rumsey, H., Jr., *Epsilon Entropy of Gaussian Processes* (to be published).

[7] Posner, E. C., and Rodemich, E. R., *Epsilon Entropy and Data Compression* (to be published).

then the ratio J_ϵ/L_ϵ remains bounded as ϵ tends to 0; and if

$$\sum_{k=n}^{\infty} \lambda_k = o\,(n\lambda_n)$$

then $J_\epsilon \sim L_\epsilon$ as $\epsilon \to 0$.

This last result implies that, when

$$\sum_{n}^{\infty} \lambda_k = o\,(n\lambda_n)$$

product ϵ-entropy is asymptotically as good as ϵ-entropy for small ϵ. As an application of our techniques, we show that for a stationary band-limited gaussian process on the unit interval, with well-behaved spectrum,

$$J_\epsilon(X) \sim \frac{\log^2 \dfrac{1}{\epsilon}}{2 \log \log \dfrac{1}{\epsilon}}$$

Subsection 6 presents an application of theorem 5 to band-limited processes.

2. The One-Dimensional Normal Distribution

In this subsection, we consider a normal random variable of mean 0 on the line. We show that the ϵ-partition of the line with least entropy is the "centered partition consisting of non-overlapping intervals of length ϵ, and containing the interval $(-\epsilon/2, \epsilon/2)$.

We need a series of six lemmas to prove this result, which is theorem 1. The first lemma shows that we need only consider portions consisting of non-overlapping intervals of length ϵ. Lemmas 2–3 show that the centered partition is best (has smallest entropy) if $\epsilon \geqq 3$. Lemmas 4–6 are devoted to showing that the centered partition is best when $\epsilon \leqq \pi$.

We begin by defining the entropy of a countable partition U of the real line under a probability measure: Let the probabilities of the sets of U be denoted by p_i. Then the entropy $H(U)$ of the partition U is the (Shannon) entropy of the discrete probability distribution $\{p_i\}$, that is

$$H(U) = \sum_i p_i \log \frac{1}{p_i} \tag{7}$$

The term "epsilon entropy" in the following lemma refers to the definition of Ref. 2: the epsilon entropy of a separable metric space with a probability distribution on the Borel sets is the infimum of the entropies of all partitions of the space by measurable sets of diameters at most ϵ.

For conciseness, the statement of the lemma neglects the behavior of the partition on sets of probability zero. More precisely, the sets of positive probability in an optimal partition can be intervals of length ϵ with sets of probability zero omitted.

Lemma 1. Let X be the real line with a probability distribution μ on the Borel sets of X such that μ has a density $p(x)$ which achieves its maximum value at 0, is monotonic on $(0, \infty)$, and even $[p(-x) = p(x)]$. Then the ϵ-entropy of X is attained only by a partition which consists of consecutive intervals of length ϵ (or one which agrees with such a partition on the interval supporting μ if this interval is finite).

The hypothesis of unimcdality of the description is essential for the conclusion of lemma 1. The distribution need not be symmetric, however. This assumption was used to simplify the treatment of a partition in which the interval containing zero has length less than ϵ. In the problem at hand, lemma 1 implies that for gaussian distributions, the epsilon entropy is attained only for a partition by consecutive intervals of length ϵ. We are thus led to the following definition:

Definition. Let X be the real line with the probability distribution of a normal random variable with mean zero and variance 1. Let $h(\epsilon, \alpha)$ be the entropy of the partition of X by intervals of length ϵ centered at the points $\epsilon(k - \alpha), k = 0, \pm 1, \pm 2, \cdots$: $h(\epsilon, 0)$ is denoted by $h(\epsilon)$, the entropy of the centered ϵ partition of X.

Lemmas 3 and 6 below show that for any $\epsilon > 0$ we have $h(\epsilon, \alpha) \geqq h(\epsilon)$, with equality only if α is an integer. We first define two functions and state some of their properties. Let $P(\epsilon, z)$ be the probability of the interval of length ϵ centered at ϵz, so that

$$P(\epsilon, z) = \int_{(z-1/2)\epsilon}^{(z+1/2)\epsilon} \phi(y)\, dy = \int_{(z-1/2)\epsilon}^{(z+1/2)\epsilon} \exp\left(-\frac{y^2}{2}\right) \frac{dy}{(2\pi)^{1/2}} \tag{8}$$

where ϕ is the normal density function. Since

$$P(\epsilon, z) \sim \frac{1}{\left(z - \frac{1}{2}\right)\epsilon} \phi\left[\left(z - \frac{1}{2}\right)\epsilon\right]$$

for large z, all the series which we encounter will converge absolutely; we need make no further mention of convergence.

Define

$$F(z) = F(\epsilon, z) = \log \frac{P\left(\epsilon, z - \frac{1}{2}\right)}{P\left(\epsilon, z + \frac{1}{2}\right)}$$

$$= \log \frac{\int_{z-1}^{z} \exp\left(-\frac{\epsilon^2 y^2}{2}\right) dy}{\int_{z}^{z+1} \exp\left(-\frac{\epsilon^2 y^2}{2}\right) dy}$$

The following lemma lists some of the properties of P and F.

Lemma 2. The following seven properties hold for the functions P and F:

$$P(\epsilon, -z) = P(\epsilon, z)$$

$$F(-z) = -F(z)$$

$$\epsilon^2 \left(z - \frac{1}{2}\right) \leq \frac{\partial}{\partial z} \log \frac{1}{P(\epsilon, z)}, \qquad \frac{\partial}{\partial z} \log \frac{1}{P(\epsilon, z)} < \epsilon^2 z \qquad \text{for} \qquad z > 0$$

$$0 < F'(z) \text{ for } z, \epsilon > 0, \qquad F'(z) \leq \frac{3}{2}\epsilon^2 \qquad \text{for} \qquad z > \frac{1}{2}$$

$$\epsilon^2 \left(z - \frac{1}{2}\right) < F(\epsilon, z) < \epsilon^2 z \qquad \text{for} \qquad z > 0$$

$$F(\epsilon, z) > 4\int_0^{\epsilon z} \phi(y)\, dy - 4\int_{\epsilon - \epsilon z}^{\epsilon} \phi(y)\, dy \qquad \text{for} \qquad 0 < z < \frac{1}{2}$$

$$F(\epsilon, z) \text{ is increasing in } \epsilon \text{ for fixed } z > 0$$

The next lemma proves theorem 1 for large ϵ. However, the difficult case is the case of small ϵ.

Lemma 3. If $\epsilon \geq 3$, $h(\epsilon, \alpha)$ assumes its minimum value only when α is an integer.

To complete the proof of theorem 1, we shall have to study the function $h(\epsilon,\alpha)$ very carefully. This is because for small ϵ

$$\frac{\partial}{\partial\alpha}h(\epsilon,\alpha) = 0\left[\exp\left(-\frac{2\pi^2}{\epsilon^2}\right)\right]$$

so that h is very flat as $\epsilon \to 0$.

The rapid convergence of the series for $h(\epsilon,\alpha)$ ensures that it is C^∞. From the periodicity of the function in α, it follows that it is the sum of a convergent Fourier series:

$$h(\epsilon,\alpha) = \frac{1}{2}C_0(\epsilon) + \sum_{n=1}^{\infty} C_n(\epsilon)\cos(2n\pi\alpha) \tag{9}$$

where

$$C_n(\epsilon) = 2\int_{-\frac{1}{2}}^{\frac{1}{2}} h(\epsilon,\alpha)\cos(2n\pi\alpha)\,d\alpha = 2\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} P(\epsilon,k-\alpha)\log\frac{1}{P(\epsilon,k-\alpha)}\cos(2n\pi\alpha)\,d\alpha$$

We interchange the order of integration and summation here; after the substitution $k - \alpha = x$, we have

$$C_n(\epsilon) = 2\int_{-\infty}^{\infty} P(\epsilon,x)\log\frac{1}{P(\epsilon,x)}\cos(2n\pi x)\,dx \tag{10}$$

To get useful inequalities for these coefficients, we need to investigate the properties of $P(\epsilon,z)$ as an entire function of the complex variable z.

Define

$$\left.\begin{aligned} Q(\epsilon,z) &= \frac{(2\pi)^{\frac{1}{2}}}{\epsilon}\exp\left(\frac{z^2}{2\epsilon^2}\right)P\left(\epsilon,\frac{z}{\epsilon^2}\right) \\ \text{so that}\qquad Q(\epsilon,z) &= \int_{-\frac{1}{2}}^{\frac{1}{2}}\exp\left(-zy-\frac{\epsilon^2y^2}{2}\right)dy \end{aligned}\right\} \tag{11}$$

which shows that $Q(\epsilon,z)$ is an even entire function of z of exponential type. Hence, it can be expressed in terms of the canonical product of its zeros $\pm\zeta_1, \pm\zeta_2, \cdots$ as

$$Q(\epsilon,z) = Q(\epsilon,0)\prod_{k=1}^{\infty}\left(1-\frac{z^2}{\zeta_k^2}\right) \tag{12}$$

Thus, information about the zeros of $Q(\epsilon,z)$ would be quite useful, and the next lemma furnishes the required information.

Lemma 4. The zeros $\{\pm\zeta_k\}$ of $Q(\epsilon,z)$ are all distinct and are on the imaginary axis for $0 < \epsilon < \epsilon_0 = 4.309 \cdots$. Furthermore, under the appropriate indexing, we have

$$2\pi k < \frac{\zeta_k}{i} < 2\pi(k+1), \qquad k = 1,2,\cdots \tag{13}$$

Next, lemma 4 will be applied to get estimates for the Fourier coefficients $C_n(\epsilon)$ of $h(\epsilon,\alpha)$. This is the content of lemma 5.

Lemma 5. If $0 < \epsilon < \epsilon_0$,

$$C_1(\epsilon) \leqq -2\exp\left(-\frac{2\pi^2}{\epsilon^2}\right)[1 - P(\epsilon,0)] \tag{14}$$

and, for $n \geqq 2$,

$$|C_n(\epsilon)| \leqq \exp\left(-\frac{2\pi^2n^2}{\epsilon^2}\right)[2 + 4P(\epsilon,0)] + \frac{2\epsilon}{n(2\pi)^{\frac{1}{2}}}\sum_{k=1}^{n-1}\exp\left[-\frac{2\pi^2(2nk-k^2)}{\epsilon^2}\right] \tag{15}$$

Lemma 6. For $0 < \epsilon \leqq \pi$, $h(\epsilon, \alpha) > h(\epsilon)$ when α is not an integer.

Theorem 1. The ϵ-entropy $H_\epsilon(X)$ of the real line X under a one-dimensional gaussian distribution with mean 0, variance σ^2 is $h(\epsilon/\sigma)$. The only ϵ-partition of the line with this entropy is the partition into consecutive intervals of length ϵ with one interval centered at zero.

Proof. We can assume $\sigma = 1$, since the general case follows by a change of scale. By lemma 1, the only ϵ-partitions whose entropy can be the ϵ-entropy of the space are those which subdivide the line into intervals of length ϵ. We run through all these partitions by taking the partition into ϵ-intervals with one interval centered at $-\epsilon\alpha$, $0 \leqq \alpha < 1$. These partitions have entropies $h(\epsilon, \alpha)$, so that

$$H_\epsilon(X) = \inf_{0 \leqq \alpha < 1} h(\epsilon, \alpha)$$

By lemmas 3 and 6, for any positive ϵ, this infimum is assumed only at $\alpha = 0$, which proves theorem 1.

The final lemma of this subsection lists some properties of the function $h(\epsilon)$. These properties are interesting in themselves, and they are also needed at various points throughout the remainder of this article.

Lemma 7. For $0 < \epsilon < \infty$, $h'(\epsilon) < 0$ and $[h'(\epsilon)/\epsilon]' > 0$. The function $h'(\epsilon)/\epsilon$ varies monotonically from $-\infty$ to 0 for ϵ on $(0, \infty)$. Also, the following asymptotic formulas hold:

as $\epsilon \to 0$,

$$\left.\begin{aligned} h(\epsilon) &\sim \log\frac{1}{\epsilon} \\ h'(\epsilon) &\sim -\frac{1}{\epsilon} \end{aligned}\right\} \qquad (16)$$

as $\epsilon \to \infty$,

$$\left.\begin{aligned} h(\epsilon) &\sim \frac{\epsilon}{2(2\pi)^{1/2}} \exp\left(-\frac{\epsilon^2}{8}\right) \\ h'(\epsilon) &\sim -\frac{\epsilon^2}{8(2\pi)^{1/2}} \exp\left(-\frac{\epsilon^2}{8}\right) \end{aligned}\right\} \qquad (17)$$

Now that we have gotten "preliminaries" about the one-dimensional gaussian distribution out of the way, we can begin to study the case of arbitrary mean-continuous gaussian processes on the unit interval.

3. The Product Epsilon Entropy Function $J_\epsilon(X)$

In this subsection, we define the product ϵ-entropy, $J_\epsilon(X)$, of a mean-continuous gaussian process X on the unit interval. The main results are contained in theorem 2. We find a necessary and sufficient condition for $J_\epsilon(X)$ to be finite. In the case when $J_\epsilon(X)$ is finite, we show how to construct a product ϵ-partition with entropy equal to $J_\epsilon(X)$.

In order to define the product ϵ-entropy function $J_\epsilon(X)$, we first consider the class π'_ϵ of all *product ϵ-partitions* of $L_2(0,1)$. A product ϵ-partition of $L_2(0,1)$ is the cartesian product of ϵ_k-partitions of the kth coordinate axis in the Karhünen expansion of the process, where $\Sigma\, \epsilon_k^2 \leqq \epsilon^2$. Thus product ϵ-partitions consist of hyper-cubes of diameter at most ϵ. Next define π_ϵ to be the subclass of π'_ϵ of partitions in which a countable collection of the sets have a union with probability 1. A product partition in π_ϵ includes a denumerable partition of a subset of X of probability 1. By the *entropy* of the product partition we mean the entropy of this denumerable partition.

The product epsilon is defined as

$$J_\epsilon(X) = \infty \text{ if } \pi_\epsilon \text{ is empty}$$

$$J_\epsilon(X) = \inf_{U \in \pi_\epsilon} H(U) \text{ if } \pi_\epsilon \text{ is not empty}$$

The entropy $H(U)$ is defined as in Eq. (7) over the sets of U of positive probability.

It turns out that π_ϵ is empty if the series Eq. (19) diverges, and otherwise $J_\epsilon(X)$ is finite.

Our first lemma shows how to compute tne entropy of a product partition in terms of the entropies of its one-dimensional partitions.

Lemma 8. Let the probability space X be the product of a sequence of probability spaces $X_1, X_2, \cdots$, with product measure. If U_k is a partition of X_k, $k = 1, 2, \cdots$, and U the product partition of X, then

$$H(U) = \sum_{k=1}^{\infty} H(U_k)$$

This is to be interpreted to mean that if the union of countably many sets of U does not have probability 1, then $H(U)$ is infinite.

The next two lemmas taken together show that, for a mean-continuous gaussian process on the unit interval, either π_ϵ is empty for all $\epsilon > 0$, or else π_ϵ contains a partition of finite entropy for all $\epsilon > 0$.

Lemma 9. Let $X(t)$ be a mean-continuous gaussian process on the unit interval. Let U be a product partition of $L_2(0,1)$ obtained as the product of partitions U_k of the coordinate axes by intervals of lengths ϵ_k. Then the following three conditions are equivalent:

(1) The union of countably many sets of U has probability 1

(2) U contains a set of positive probability

(3) With probability 1, all but a finite number of components of an element of $L_2(0,1)$ lie in the unique interval containing zero in the partition of that coordinate.

If the partitions U_k are centered, these conditions are also equivalent to

$$\sum_{k=1}^{\infty} \frac{\lambda_k^{1/2}}{\epsilon_k} \phi\left(\frac{\epsilon_k}{2\lambda_k^{1/2}}\right) < \infty$$

where ϕ is the unit normal density function, and $\{\lambda_k\}$ are the eigenvalues of the process.

Lemma 10. For $k = 1, 2, \cdots$, let U_k be a given ϵ_k'-partition of the kth coordinate axis. Let $\Sigma \epsilon_k'^2$ converge, and let a countable subpartition of the product partition, $U = \pi_k U_k$, cover a set of probability 1 in $L_2(0,1)$. Then for every $\epsilon > 0$ there exist ϵ_k-partitions, V_k, of the kth coordinate axis such that

$$\epsilon^2 = \sum \epsilon_k^2$$

and

$$\sum H(V_k) < \infty$$

Lemma 11. For a mean-continuous gaussian process on $(0,1)$ with eigenvalues $\lambda_n = \sigma_n^2, n = 1, 2, \cdots$, the product ϵ-entropy is given by

$$J_\epsilon(X) = \inf_{\Sigma \epsilon_k^2 = \epsilon^2} \sum_{k=1}^{\infty} h\left(\frac{\epsilon_k}{\sigma_k}\right) \qquad (18)$$

Proof. With each product ϵ-partition of X, we can associate a sequence $\{\epsilon_k\}$ such that the partition of the kth component space X_k is an ϵ_k-partition, and $\Sigma \epsilon_k^2 = \epsilon^2$. For given $\{\epsilon_k\}$, the minimum possible entropy of the partition of X_k is $h(\epsilon_k/\sigma_k)$, by theorem 1. Hence, Eq. (18) follows from lemma 8. Lemma 11 is proved.

Equation (18) reduces the problem of finding an optimal product ϵ-partition to the problem of selecting an optimal set, $\{\epsilon_k\}$, of quantizations for the coordinate axes. The next theorem solves this problem and gives a necessary and sufficient condition for $J_\epsilon(X)$ to be finite.

Theorem 2. The product ϵ-entropy $J_\epsilon(X)$ of a mean-continuous gaussian process on $(0,1)$ with eigenvalues $\{\lambda_k\}$ is finite if and only if

$$\sum \lambda_k \log \frac{1}{\lambda_k} < \infty \qquad (19)$$

If this condition is satisfied, the equations

$$h'(\delta_k) = -A\lambda_k\delta_k, \qquad k = 1, 2, \cdots \qquad (20)$$

have a unique solution $\{\delta_k\}$ with A such that

$$\Sigma \lambda_k \delta_k^2 = \epsilon^2 \qquad (21)$$

Then

$$J_\epsilon(X) = \sum_{k=1}^{\infty} h(\delta_k) \qquad (22)$$

On the other hand, if Eq. (19) is violated, Eqs. (20) and (21) have no solution. The condition Eq. (19) is also the condition that there be a countable subpartition of some product epsilon partition covering a set of probability 1.

Proof. Set $\sigma_k = \lambda_k^{1/2}$. We want to minimize

$$J(\epsilon_1, \epsilon_2, \cdots) = \sum h\left(\frac{\epsilon_k}{\sigma_k}\right)$$

subject to condition $\Sigma \epsilon_k^2 = \epsilon^2$. Equation (20) is the condition for a minimum, by the method of Lagrange multipliers, if $\delta_k = \epsilon_k/\sigma_k$. To avoid justifying the use of this method in an infinite-dimensional space, we will consider finite dimensional subspaces of X.

First we show that Eqs. (20) and (21) have a (unique) solution for any $\epsilon > 0$, if any, only if Eq. (19) is satisfied. According to lemma 7, for any $A > 0$ there is a unique solution $\{\delta_k\}$ of Eq. (20); each δ_k is a monotonic decreasing function of A, and

$$\lim_{A \to 0+} \delta_k = \infty, \qquad \lim_{A \to \infty} \delta_k = 0$$

For a given value of A, $A\lambda_k \to 0$ as $k \to \infty$. Hence for k sufficiently large, δ_k is so large that we can conclude from lemma 7 that

$$h'(\delta_k) = -C_k \delta_k^2 \exp\left(-\frac{1}{8}\delta_k^2\right)$$

where

$$\frac{1}{16(2\pi)^{1/2}} < C_k < 1$$

Then we have

$$C_k \delta_k \exp\left(-\frac{1}{8}\delta_k^2\right) = A\lambda_k$$

which implies

$$\delta_k^2 \sim 8 \log \frac{1}{\lambda_k} \tag{23}$$

We see that the series Eq. (21) is finite if and onl, if Eq. (19) is satisfied. If Eq. (19) holds, then the monotone dependence of δ_k on A shows that the series in Eq. (21) is a strictly decreasing function of A, taking all positive values as A ranges over all positive values. Therefore, Eqs. (20) and (21) have a unique solution.

Notice also that the existence of a solution of Eqs. (20) and (21) implies that $J_\epsilon(x)$ is finite, for if we put $\epsilon_k = \sigma_k \delta_k$, then

$$\Sigma \epsilon_k^2 = \epsilon^2$$

and

$$J_\epsilon(X) \leqq \sum h\left(\frac{\epsilon_k}{\sigma_k}\right) = \Sigma h(\delta_k)$$

This series converges, for by lemma 7,

$$h(\delta_k) \sim -\delta_k h'(\delta_k) = A\delta_k^2 \lambda_k$$

Now let $X^{(n)}$ be the product of the first n coordinate spaces. By lemma 11,

$$J_\epsilon(X^{(n)}) = \inf_{\sum_1^n \epsilon_k^2 = \epsilon^2} \sum_{k=1}^{n} h\left(\frac{\epsilon_k}{\sigma_k}\right)$$

This sum is a continuou function over the positive 2^n-tant of the sphere $\Sigma \epsilon_k^2 = \epsilon^2$, approaching infinity at the boundaries. Hence the infimum is assumed at some point, and we have there

$$\frac{h'\left(\dfrac{\epsilon_k}{\sigma_k}\right)}{\sigma_k} = -A^{(n)} \epsilon_k, \qquad k = 1, \cdots, n$$

where $A^{(n)}$ is a positive constant. Let $\epsilon_k = \delta_k^{(n)} \sigma_n$ be a solution of this system of equations, which lies on the sphere. Then

$$\frac{h'(\delta_k^{(n)})}{\delta_k^{(n)}} = -A^{(n)}\lambda_k, \qquad k = 1, \cdots, n \tag{24}$$

and

$$\sum_{k=1}^{n} \lambda_k \delta_k^{(n)^2} = \epsilon^2 \tag{25}$$

For any value of $A^{(n)}$, the solutions of Eq. (24) are uniq by lemma 7. Furthermore, as $A^{(n)}$ varies from 0 to ∞, each $\delta_k^{(n)}$ varies monotonically from ∞ to 0. Thus, there is a unique value of $A^{(n)}$ at which Eq. (25) is satisfied. We have

$$J_\epsilon(X) \geqq J_\epsilon(X^{(n)}) = \sum_{k=1}^{n} h(\delta_k^{(n)}) \tag{26}$$

This can be done for any n. In particular, for the numbers $A^{(n+1)}$ and

$$\{\delta_k^{(n+1)}\},$$

$$\delta_1^{(n+1)}, \cdots, \delta_n^{(n+1)}$$

are solutions of Eq. (24) with $A^{(n)}$ replaced by $A^{(n+1)}$, and

$$\sum_{k=1}^{n} \lambda_k \delta_k^{(n+1)^2} \leqq \epsilon^2$$

It follows that $A^{(n+1)} \geqq A^{(n)}$. Define

$$\bar{A} = \lim_{n \to \infty} A^{(n)}$$

$\bar{A}$ is either a positive real number or ∞.

First suppose $\bar{A} = \infty$. Then as $n \to \infty$, $A^{(n)}\lambda_1 \to \infty$ and $\delta_1^{(n)} \to 0$. From Eq. (26),

$$J_\epsilon(X) \geqq h(\delta_1^{(n)}) \to \infty$$

so $J_\epsilon(X) = \infty$. It follows from above that in this case Eq. (19) is violated.

Now let $\bar{A}$ be finite, and let $\{\bar{\delta}_k\}$ be the solution of Eq. (20) when $A = \bar{A}$. Since $A^{(n)} \leqq \bar{A}$,

$$\sum_{k=1}^{n} \lambda_k \bar{\delta}_k^2 \leqq \sum_{k=1}^{n} \lambda_k \delta_k^{(n)^2} = \epsilon^2$$

hence

$$\sum_{k=1}^{\infty} \lambda_k \bar{\delta}_k^2 \leqq \epsilon^2$$

This shows that there is a value A^* of A for which the solution of Eq. (20) satisfies Eq. (21), and $A^* \leqq \bar{A}$. Denoting this solution by $\{\delta_k^*\}$, we have

$$\sum_{k=1}^{n} \lambda_k \delta_k^{*2} \leqq \epsilon^2$$

hence $A^* \geqq A^{(n)}$, for all n. It follows that $\bar{A} = A^*$.

For each k, we have $\delta_k^{(n)} \rightarrow \bar{\delta}_k$ as $n \rightarrow \infty$. From Eq. (26), if $m \leqq n$,

$$J_\epsilon(X) \geqq \sum_{k=1}^{m} h(\delta_k^{(n)})$$

Letting $n \rightarrow \infty$, then $m \rightarrow \infty$, we obtain

$$J_\epsilon(X) \geqq \sum_{k=1}^{\infty} h(\bar{\delta}_k)$$

On the other hand, we have seen above that this series is the entropy of an ϵ-product partition of X. Therefore, equality holds, and Eq. (22) is true. The last assertion of the theorem follows from lemmas 9 and 10. This completes the proof of theorem 2.

Corollary. $J_\epsilon(X)$ is a continuous function of ϵ.

Proof. This is a consequence of the formulas of theorem 2. The asymptotic formula (23) is uniform over any interval $0 < A_1 \leqq A \leqq A_2 < \infty$. Thus the series in Eqs. (21) and (22) are uniformly convergent. It follows that these series are continuous functions of A. Since ϵ, given by Eq. (21), is a strictly decreasing function of A, A, and $J_\epsilon(X)$ are continuous functions of ϵ This proves the corollary.

We remark that when the λ_k are written in non-increasing order, condition (19) is equivalent to

$$\Sigma \lambda_k \log k < \infty$$

Also note that Eq. (19) is the entropy of the distribution $\{\lambda_k\}$, provided the λ_k are normalized so that $\Sigma \lambda_k = 1$. The occurrence of the entropy of the eigenvalues in this way appears to be fortuitous.

4. Some Special Processes

In this subsection, we shall consider a class of gaussian processes whose product ϵ-entropies can be estimated for small ϵ by theorem 2. We begin with some general remarks on product ϵ-entropy.

Let X be a finite-dimensional mean-continuous gaussian process on $(0, 1)$. That is, X has only a finite number of non-zero eigenvalues, $\lambda_1, \cdots, \lambda_n$, say. It is a consequence of theorem 2 and lemma 7 that

$$J_\epsilon(X) \sim n \log \frac{1}{\epsilon}$$

as $\epsilon \rightarrow 0$. For this reason the interesting processes to now consider, from the point of view of product ϵ-entropy, are the infinite-dimensional ones.

The first thing we observe about an infinite-dimensional process X is that, as $\epsilon \rightarrow 0$, its product ϵ-entropy must increase faster than any positive multiple of $\log 1/\epsilon$. To verify this, let $X^{(n)}$ be the finite-dimensional process obtained from X by setting $\lambda_k = 0$ for $k > n$. Then as $\epsilon \rightarrow 0$

$$J_\epsilon(X) \geqq J_\epsilon(X^{(n)}) \sim n \log \frac{1}{\epsilon}$$

Since n was arbitrary, this proves our assertion.

In the final *Subsection 5*, we shall develop some techniques which are more generally applicable than theorem 2. For the present, however, we shall consider mean-continuous gaussian processes on $(0, 1)$ whose eigenvalues satisfy a relation of the form

$$\lambda_k \sim Bk^{-p} \text{ as } k \rightarrow \infty$$

where $B > 0$ and $p > 1$ are constants. Special cases of these processes arise as solutions of the stochastic differential equation

$$\frac{d^nX}{dt^n} + a_{n-1}\frac{d^{n-1}X}{dt^{n-1}} + \cdots + aX = b_m\frac{d^mN}{dt^m} + \cdots + bN$$

where $N(t)$ is white gaussian noise of spectral density $1/2$ and the a's and b's are constants with $b_m \neq 0$ and

$n > m$. For these processes, $R(s, t) = E[X(s)X(t)]$ can be found as well as the λ_k. However, for our purposes it is enough to know that

$$\lambda_k \sim Bk^{-p}$$

where $B > 0$ and $p = 2(n - m)$. This is true for stationary processes by Ref. 3 and apparently is also true for nonstationary processes. The most important special case is the Weiner process, for which

$$dX/dt = N, R(s,t) = \min(s,t)$$

and

$$\lambda_k = \frac{1}{\pi^2\left(k - \frac{1}{2}\right)^2}, \qquad k = 1, 2, \cdots$$

The main result of this subsection is the following theorem which gives an asymptotic formula for $J_\epsilon(X)$ as $\epsilon \to 0$.

Theorem 3. Let X be a mean-continuous gaussian process on the unit interval with eigenvalues $\{\lambda_n\}$ such that

$$\lambda_n \sim Bn^{-p}$$

$B > 0$, $p > 1$. Then, as $\epsilon \to 0$,

$$J_\epsilon(X) \sim \epsilon^{-2/(p-1)}\left(\frac{2B}{p-1}\right)^{1/(p-1)} \times \left\{\int_0^\infty \left[-\frac{h'(x)}{x}\right]^{1-1/p} x\,dx\right\}^{p/(p-1)} \qquad (27)$$

Corollary. For the Weiner process on $(0, 1)$

$$J_\epsilon(X) \sim \frac{C}{\epsilon^2}$$

as $\epsilon \to 0$, with

$$C = \frac{2}{\pi^2}\left\{\int_0^\infty [-xh'(x)]^{1/2}dx\right\}^2 = 6.711 \cdots$$

Proof. We apply theorem 3 with $B = 1/\pi^2, p = 2$, and evaluate the integral numerically to prove this corollary.

The ϵ-entropy $H_\epsilon(X)$ of the Weiner process has been considered,[6] where $H_\epsilon(X)$ is the infimum of the entropies of all countable partitions or sets of probability 1 in $L_2[0,1]$ by measurable sets of diameters at most ϵ. Thus, $H_\epsilon(X) \leq J_\epsilon(X)$. However, it has been shown[r] that for the Weiner process

$$\frac{17}{32\epsilon^2} < H_\epsilon(X) < \frac{1}{\epsilon^2}$$

(the notation $U < V$ means $\limsup U/V \leq 1$). Thus, for the Weiner process,

$$\liminf_{\epsilon \to 0} \frac{J_\epsilon(X)}{H_\epsilon(X)} \geq 6.711 \cdots$$

This means that, for small ϵ, the product ϵ-partition on the average requires at least 6.7 times as many bits to transmit the outcome of the process as does the optimal ϵ-partition.

5. The Order of Magnitude of $J_\epsilon(X)$

In this final subsection, a useful lower bound $L_\epsilon(X)$ for $J_\epsilon(X)$ is considered. Conditions on the eigenvalues λ_k are given, which guarantee that $J_\epsilon(X) = 0[L_\epsilon(X)]$, or even $J_\epsilon(X) \sim L_\epsilon(X)$. Since $L_\epsilon(X)$ is a lower bound for the epsilon entropy $H_\epsilon(X)$, these results imply that $H_\epsilon(X)$ is of the same order as, or even asymptotically equal to, $J_\epsilon(X)$, so that not much is lost by the restriction to product partitions in these cases. Finally, these results are applied to a stationary band-limited gaussian process on the unit interval to obtain a simple asymptotic expression for $J_\epsilon(X)$ in that case.

The lower bound $L_\epsilon(X)$ derived[6] for the ϵ-entropy $H_\epsilon(X)$ of a gaussian process X is as follows: Assume $\epsilon^2 < \Sigma\,\lambda_k$. Define the number $b = b(\epsilon)$ by

$$\epsilon^2 = \Sigma\frac{\lambda_k}{1 + b\lambda_k} \qquad (28)$$

Then

$$L_\epsilon(X) = \frac{1}{2}\Sigma\log(1 + b\lambda_k) \qquad (29)$$

Since

$$L_\epsilon(X) \leq H_\epsilon(X) \text{ and } H_\epsilon(X) \leq J_\epsilon(X), L_\epsilon(X)$$

also provides a lower bound for $J_\epsilon(X)$.

The next lemma gives a lower bound for $L_\epsilon(X)$, which is actually the bound we shall be using.

Lemma 12. Let X be a mean-continuous gaussian process on $[0, 1]$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$. Define $\lambda(x), x \geq 1$, as the function such that

$$\lambda(n) = \lambda_n, \qquad n = 1, 2, \cdots,$$

and $x\lambda(x)$ is linear on each interval $(n, n+1)$. For $\epsilon^2 < \lambda_1$, define the function $y = y(\epsilon)$ to be the smallest root on $(1, \infty)$ of the equation

$$y\lambda(y) = \epsilon^2$$

Then

$$L_\epsilon(X) \geq \int_\epsilon^{\lambda_1^{1/2}} [y(t) - 1] \frac{dt}{t} + 0(1) \tag{30}$$

as $\epsilon \to 0$.

The next lemma estimates the number $A = A(\epsilon)$ given by Eqs. (20) and (21) in terms of the function $y(\epsilon)$ of the preceding lemma. To make these estimates, certain restrictions must be put on the eigenvalues λ_k; these restrictions imply that the influence of the eigenvalues far out is not too large.

Lemma 13. Let $A = A(\epsilon)$ be the number in the solution of Eqs. (20) and (21). If the gaussian process X has an infinite number of positive eigenvalues, and

$$\sum_{k=n}^{\infty} \lambda_k = 0(n\lambda_n)$$

when the eigenvalues are arranged in non-increasing order, then Eqs. (20) and (21) have a solution, and $A\epsilon^2 = 0[y(\epsilon)]$ as $\epsilon \to 0$. If the stronger condition

$$\sum_{k=n}^{\infty} \lambda_k = o(n\lambda_n) \tag{31}$$

holds, then $A\epsilon^2 \sim y(\epsilon)$.

The main result of this subsection is the following theorem.

Theorem 4. Let X be a mean-continuous gaussian process on the unit interval with infinitely many non-zero eigenvalues $\{\lambda_n\}$ arranged in non-increasing order. If

$$\sum_{k=n}^{\infty} \lambda_k = 0(n\lambda_n)$$

then, as $\epsilon \to 0$, we have

$$J_\epsilon(X) = 0[L_\epsilon(X)]$$

If the stronger condition

$$\sum_{k=n}^{\infty} \lambda_k = o(n\lambda_n) \tag{32}$$

holds, we have

$$J_\epsilon(X) \sim L_\epsilon(X)$$

An important consequence of theorem 4 is the next result, which has been proved within theorem 4.

Theorem 5. Let X be a mean-continuous gaussian process on the unit interval with infinitely non-zero eigenvalues $\{\lambda_n\}$ arranged in non-increasing order. If

$$\sum_{k=n}^{\infty} \lambda_k = 0(n\lambda_n)$$

then

$$J_\epsilon(X) = 0\left(\int_\epsilon^{\lambda_1^{1/2}} y(t) \frac{dt}{t}\right)$$

If the stronger condition

$$\sum_{k=n}^{\infty} \lambda_k = o(n\lambda_n)$$

holds, then

$$J_\epsilon(X) \sim \int_\epsilon^{\lambda_1^{1/2}} y(t) \frac{dt}{t}$$

Note that theorem 5 applies in the case of theorem 3, but gives less precise information.

Since

$$J_\epsilon(X) \geq H_\epsilon(X) \geq L_\epsilon(X)$$

Theorem 4 can be thought of as a condition for

$$J_\epsilon(X) = 0[H_\epsilon(X)]$$

or

$$J_\epsilon(X) \sim H_\epsilon(X)$$

In the former case, X can be transmitted by product partitions with a number of bits not worse than the optimal system by more than a constant multiple. For processes with the stronger property (Eq. 32), the product partition system is asymptotically as good as the best possible system as $\epsilon \to 0$. It can, moreover, be shown that $J_\epsilon(X)$ can be finite and yet not $0[H_\epsilon(X)]$.

6. Application of Theorem 5 to Band-Limited Processes

Let X be a mean-continuous stationary gaussian process on the real line whose covariance function

$$\rho(\tau) = R(s, s+\tau)$$

has Fourier transform $dS(f)$ with support in some finite interval. Suppose $dS(f) = a(f)\,df$ with $a(f)$ continuous. Then when X is restricted to the unit interval, it is known [Ref. 4, lemma 2] that

$$\lambda_n \sim \frac{1}{n}(Cn)^{-2n}$$

for some constant C. It is seen that

$$y(\epsilon) \sim \frac{\log\frac{1}{\epsilon}}{\log\log\frac{1}{\epsilon}}$$

Theorem 5 then implies

$$J_\epsilon(X) \sim \int_\epsilon^{\lambda_1^{1/2}} \left(\frac{\log\frac{1}{t}}{\log\log\frac{1}{t}}\right)\left(\frac{dt}{t}\right)$$

so that

$$J_\epsilon(X) \sim \frac{1}{2}\frac{\left(\log\frac{1}{\epsilon}\right)^2}{\log\log\frac{1}{\epsilon}} \tag{33}$$

Equation (33) shows that band-limited processes are not much more random than finite-dimensional distributions, since $J_\epsilon(X)$ does not increase much more rapidly than a constant times $\log 1/\epsilon$. This is to be expected, since the sample functions are analytic with probability 1.

References

1. Loéve, M., *Probability Theory—Foundations, Random Sequences,* Sec. 34.5. D. van Nostrand Co. Inc., New York, 1955.
2. Posner, E. C., Rodemich, E. R., and Rumsey, H., Jr., "Epsilon Entropy of Stochastic Processes," *Ann. Math. Statist.*, Vol. 38, pp. 1000–1020, 1967.
3. Widom, H., "Asymptotic Behavior of Eigenvalues of Certain Integral Operators," *Arch. Ration. Mech.*, Vol. 17, pp. 215–229, 1964.
4. Widom, H., "Asymptotic Behavior of Eigenvalues of Certain Integral Equations," *Trans. Amer. Math. Soc.*, Vol. 109, pp. 278–295, 1963.

N. Data Compression Techniques: Estimators of the Parameters of an Extreme-Value Distribution Using Quantiles, *I. Eisenberger*

1. Introduction

The statistical theory of extreme values for large samples has been applied by Posner (Ref. 1) to the problem of estimation of low probability of error in threshold communications receivers. The extreme value distribution function that he considers is of the form

$$G(x) = \exp\left\{-\exp\left[-\frac{1}{\beta}(x-\alpha)\right]\right\},$$

$$-\infty < x < \infty\,, \qquad \beta > 0$$

where α, the mode of the distribution, and β, a scale parameter, are unknown and hence must be estimated. Posner, after making a change of parameters, derives the maximum-likelihood equations, the solutions of which give the maximum-likelihood estimators of his parameters. He then suggests a novel method for obtaining good first approximations.

The purpose of this article is to provide optimum or near-optimum asymptotically unbiased estimators of α and β using k quantiles when the sample size is large and both α and β are unknown. First we estimated α for $k = 1, 2, 3, \cdots, 10$, assuming β unknown. Then we estimated β for $k = 2, 3, 4, \cdots, 10$, assuming α unknown. Finally, since the orders of the k quantiles which give optimum or near-optimum estimators of α are not those which give optimum or near-optimum estimators of β, we derived estimators of α and β using the same k quantiles, for $k = 2, 4, 6, 8$, and 10. The orders of the k quantiles are taken to be those which minimize

$$\mathrm{var}(\hat{\alpha}) + C\,\mathrm{var}(\hat{\beta})\,, \qquad C = 1, 2$$

These estimators are designated as suboptimum. The efficiencies of the quantile estimators relative to the maximum-likelihood estimators were also determined.

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 pth quantile or the quantile of order p of the distribution, denoted by ζ_p^*, is defined as the root of the equation $H(\zeta_p^*) = p$; that is

$$p = \int_{-\infty}^{\zeta_p^*} dH(x) = \int_{-\infty}^{\zeta_p^*} h(x)\,dx$$

The corresponding sample quantile z_p is defined as follows: If the sample values are arranged in non-decreasing order of magnitude

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

then $x_{(i)}$ is called the ith order statistic and

$$z_p = x_{([np]+1)}$$

where $[np]$ is the greatest integer $\leq np$.

If $h(x)$ is differentiable in some neighborhood of each quantile considered, it has been shown (Ref. 2) that the joint distribution of any number of quantiles is asymptotically normal as $n \to \infty$ and that, asymptotically,

$$E(z_p) = \zeta_p^*$$

$$\mathrm{var}(z_p) = \frac{p(1-p)}{nh^2(\zeta_p^*)}$$

$$\rho_{12} = \left[\frac{p_1(1-p_2)}{p_2(1-p_1)}\right]^{1/2}$$

where ρ_{12} is the correlation between z_{p_1} and z_{p_2}, $p_1 < p_2$. We will denote by $F(x)$ and $f(x) = F'(x)$ the distribution tunction and density function, respectively, of the standardized extreme-value distribution; that is

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \exp(-e^{-x})$$

where

$$f(x) = \exp[-x - \exp(-x)]$$

Thus, denoting by ζ_p the pth quantile of the standardized distribution, one has

$$p = \int_{-\infty}^{\zeta_p^*} g(x)\,dx = \int_{-\infty}^{\zeta_p^* - \alpha/\beta} f(x)\,dx = \int_{-\infty}^{\zeta_p} f(x)\,dx$$

Hence, one sees that, asymptotica

$$E(z_p) = \zeta_p^* = \beta\zeta_p + \alpha$$

and, since

$$g(\zeta_p^*) = \frac{1}{\beta} f(\zeta_p)$$

$$\zeta_p = -\ln(-\ln p)$$

$$f(\zeta_p) = \exp\{\ln(-\ln p) - \exp[\ln(-\ln p)]\}$$

$$= -p \ln p$$

one also has

$$\mathrm{var}(z_p) = \frac{\beta^2 p(1-p)}{nf^2(\zeta_p)} = \frac{\beta^2 (1-p)}{np(\ln p)^2}$$

Since n is assumed to be large, the statistical analysis to be given will be based on the asymptotic distribution of the sample quantiles.

3. Unbiased Estimators of α Using Quantiles

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the quantile estimators of α and β, respectively, and let $\tilde{\alpha}$ and $\tilde{\beta}$ denote the corresponding maximum-likelihood estimators. We then define the efficiency of $\hat{\alpha}$ and $\hat{\beta}$ as

$$\mathrm{eff}(\hat{\alpha}) = \frac{\mathrm{var}(\tilde{\alpha})}{\mathrm{var}(\hat{\alpha})}$$

$$\mathrm{eff}(\hat{\beta}) = \frac{\mathrm{var}(\tilde{\beta})}{\mathrm{var}(\hat{\beta})}$$

Using large-sample theory, a long and involved calculation, which will be omitted, gives the asymptotic results

$$\mathrm{var}(\tilde{\alpha}) = \frac{1.10867\,\beta^2}{n}$$

$$\mathrm{var}(\tilde{\beta}) = \frac{0.60793\,\beta^2}{n}$$

The only linear unbiased estimator of α using one sample quantile, when β is unknown, is given by

$$\hat{\alpha} = z$$

where z is of order $p = e^{-1} = 0.3679$. For, since

$$\zeta = -\ln(-\ln p)$$

one has

$$E(\hat{\alpha}) = E(z) = E(\beta\zeta + \alpha) = E[-\beta \ln(\ln e) + \alpha] = \alpha$$

The variance and efficiency of $\hat{\alpha}$ are given by

$$\operatorname{var}(\hat{\alpha}) = \operatorname{var}(z) = \frac{(1-p)\beta^2}{np(\ln p)^2} = \frac{e-1}{n} \approx \frac{1.7182\beta^2}{n}$$

$$\operatorname{eff}(\hat{\alpha}) = 0.6452$$

The best linear unbiased estimator of α using $k > 1$ quantiles is of the form

$$\hat{\alpha} = \sum_{i=1}^{k} a_i z_i$$

Since

$$E(\hat{\alpha}) = \sum_{i=1}^{k} a_i (\beta\zeta_i + \alpha) = \beta \sum_{i=1}^{k} a_i\zeta_i + \alpha \sum_{i=1}^{m} a_i$$

for $\hat{\alpha}$ to be unbiased when β is unknown, the restrictions

$$\left.\begin{aligned} \sum_{i=1}^{k} a_i &= 1 \\ \sum_{i=1}^{k} a_i\zeta_i &= 0 \end{aligned}\right\} \tag{1}$$

must be placed on the coefficients a_i. Moreover, for maximum efficiency, the values of the a_i and the orders of the quantiles should be chosen so as to minimize $\operatorname{var}(\hat{\alpha})$, subject to the above restrictions. For fixed values of p_i, $i = 1, 2, \cdots, k$, the first part of the optimization procedure can be carried out using two Lagrange multipliers. Since

$$\operatorname{var}(\hat{\alpha}) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j \sigma_{ij}$$

where σ_{ij} is the covariance between z_i and z_j, form the function

$$R_1(a_1, \cdots, a_k) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i a_j \sigma_{ij} + \lambda_1 \sum_{i=1}^{k} a_i\zeta_i + \lambda_2 \sum_{i=1}^{k} a_i$$

Differentiating $R_1(a_1, \cdots, a_k)$ with respect to a_i, $i = 1, 2, \cdots, k$, results in

$$\frac{\partial R_1}{\partial a_i} = 2a_i\sigma_i^2 + 2 \sum_{\substack{j=1 \\ j \neq i}}^{k} a_j\sigma_{ij} + \lambda_1\zeta_i + \lambda_2, \qquad i = 1, 2, \cdots, k$$

Setting the k partial derivatives equal to zero and adding Eqs. (1) provides a system of $k + 2$ linear equations, the simultaneous solutions of which give the values of the a_i that minimize $\operatorname{var}(\zeta)$, in terms of the moments of the sample quantiles. By varying the orders of the k quantiles, one can then determine the optimum p_i and a_i for maximizing $\operatorname{eff}(\hat{\alpha})$. This procedure was carried out for $k = 2$, 3 and 4.

For $k > 4$, in order to simplify the calculations, a modification of the above method for determining the optimum a_i for fixed values of the p_i was adopted. If one assumes that the z_i are independent, one has

$$\left.\begin{aligned} R_2(a_1, \cdots, a_k) &= \sum_{i=1}^{k} a_i^2\sigma_i^2 + \lambda_1 \sum_{i=1}^{k} a_i\zeta_i + \lambda_2 \sum_{i=1}^{k} a_i \\ \frac{\partial R_2}{\partial a_i} &= 2a_i\sigma_i^2 + \lambda_1\zeta_i + \lambda_2 = 0, \qquad i = 1, 2, \cdots, k \\ a_i &= A_i\lambda_1 + B_i\lambda_2 \end{aligned}\right\} \tag{2}$$

where

$$A_i = \frac{-\zeta_i}{2\sigma_i^2}, \qquad B_i = -\frac{1}{2\sigma_i^2}$$

From Eqs. (1), one then has

$$\lambda_1 \sum_{i=1}^{k} A_i + \lambda_2 \sum_{i=1}^{k} B_i = 1$$

$$\lambda_1 \sum_{i=1}^{k} A_i \zeta_i + \lambda_2 \sum_{i=1}^{k} B_i \zeta_i = 0$$

Solving the above equations for λ_1 and λ_2 and then substituting these values in Eq. (2) results in

$$a_i = \frac{A_i \sum_{j=1}^{k} B_j \zeta_j - B_i \sum_{j=1}^{k} A_j \zeta_j}{\sum_{j=1}^{k} \sum_{m=1}^{k} A_j B_m \zeta_m - \sum_{j=1}^{k} \sum_{m=1}^{k} B_j A_m \zeta_m},$$

$$i = 1, 2, \cdots, k$$

This procedure was carried out for $k = 5, 6, \cdots, 10$, resulting in near-optimum estimators of α. Table 9 lists the optimum and near-optimum estimator of α for $k = 10$ and its efficiency. The high efficiencies ($> 95\%$) achieved for $k > 5$ indicate that the efficiency lost by adopting the simplified method of determining the a_i was not excessive.

4. Unbiased Estimators of β Using Quantiles

The best linear unbiased estimator of β when α is unknown is of the form

$$\hat{\beta} = \sum_{i=1}^{k} b_i z_i$$

Since

$$E(\hat{\beta}) = \beta \sum_{i=1}^{k} b_i \zeta_i + \alpha \sum_{i=1}^{k} b_i$$

one must impose the restrictions

$$\left.\begin{aligned} \sum_{i=1}^{k} b_i \zeta_i &= 1 \\ \sum_{i=1}^{k} b_i &= 0 \end{aligned}\right\} \qquad (3)$$

Thus, one sees immediately that β cannot be estimated using a single quantile and when two quantiles are used $b_1 = -b_2$. The procedure for determining the optimum b_i is similar to that for determining the optimum a_i when α is being estimated. Form

$$R_3(b_1, \cdots, b_k) = \sum_{i=1}^{k} \sum_{j=1}^{k} b_i b_j \sigma_{ij}$$

$$+ \lambda_1 \sum_{i=1}^{k} b_i \zeta_i + \lambda_2 \sum_{i=1}^{k} b_i$$

Set $\partial R_3 / \partial b_i = 0$, $i = 1, 2, \cdots, k$, add Eqs. (3) and solve for the b_i. Then by varying the p_i, the optimum b_i and p_i will be determined. This was done for $k = 2$ and 3. For $k > 3$, two procedures were used to determine near-optimum estimators. For odd values of k, the simplified method used to estimate α was adopted, resulting in

$$b_i = \frac{B_i \sum_{j=1}^{k} A_j - A_i \sum_{j=1}^{k} B_j}{\sum_{j=1}^{k} \sum_{m=1}^{k} A_j B_m \zeta_m - \sum_{j=1}^{k} \sum_{m=1}^{k} B_j A_m \zeta_m},$$

$$i = 1, 2, \cdots, k$$

For even values of k, the estimator was formed given by

$$\hat{\beta} = \sum_{j=1}^{k/2} b_j \left(\frac{z_{k-j+1} - z_j}{\zeta_{k-j+1} - \zeta_j} \right)$$

It is readily seen that

$$E(\hat{\beta}) = \beta \sum_{j=1}^{k/2} b_j$$

so that the only restriction required is

$$\sum_{i=1}^{k/2} b_i = 1 \qquad (4)$$

Let $W_j = z_{k-j+1} - z_j$. If we assume that the W_i are independent, then one has, using one Lagrange multiplier

$$R_4(b_1, \cdots b_{k/2}) = \sum_{j=1}^{k/2} \frac{b_j^2 \sigma_j^2}{(\zeta_{k-j+1} - \zeta_j)^2} + \lambda \sum_{j=1}^{k/2} b_j$$

$$\frac{\partial R_4}{\partial b_j} = \frac{2 b_j \sigma_j^2}{(\zeta_{k-j+1} - \zeta_j)^2} + \lambda = 0$$

$$b_j = \lambda D_j$$

where

$$\sigma_j^2 = \operatorname{var}(W_j)$$

$$D_j = \frac{-(\zeta_{k-j+1} - \zeta_j)^2}{2\sigma_j^2}$$

Using Eq. (4), one obtains

$$\sum_{j=1}^{k/2} b_j = \lambda \sum_{j=1}^{k/2} D_j = 1$$

$$\lambda = \frac{1}{\sum_{j=1}^{k/2} D_j}$$

and, finally

$$b_j = \frac{D_j}{\sum_{j=1}^{k/2} D_j}$$

Table 10 lists the optimum and near-optimum estimators of β and its efficiency for $k = 10$. Efficiencies in excess of 90% were found for $k > 6$.

5. Suboptimum Estimators of α and β Using the Same Quantiles

One can see from Tables 9 and 10 that the optimum and near-optimum quantiles for estimating α are not optimum or near-optimum for estimating β. For k-quantile estimators of α and β, one can, of course, select the $2k$ optimum or near-optimum quantiles and estimate both parameters independently. However, suppose, for example, one wishes to achieve maximum data compression of space telemetry by using the same k quantiles to estimate the two parameters. Which quantiles should be used? Using the optimum quantiles for estimating one parameter, in order to estimate the other, results in a substantial loss of efficiency. For instance, for $k = 8$, if one uses to estimate α the near-optimum quantiles for estimating β, eff $(\hat{\alpha})$ drops from 0.9725 to 0.8263, while estimating β with the near-optimum quantiles for estimating α results in eff $(\hat{\beta}) =$ 0.4807 instead of 0.9317.

What is required then is a method, based on a reasonable criterion, for determining suboptimum quantiles to be used to estimate both α and β. The method we propose here is as follows: Determine the orders of the quantiles which minimize $\operatorname{var}(\hat{\alpha}) + C \operatorname{var}(\hat{\beta})$ and form unbiased estimators of α and β using the quantiles thus specified. This was done, for $C = 1$ and 2, for $k = 2, 4, 6, 8,$ and 10. The estimators for $C = 1$ are given in Table 11, and the estimators for $C = 2$ are given in Table 12. A comparison of Tables 11 and 12 with Tables 9 and 10 showed that if one uses $2k$ suboptimum quantiles to estimate α and β simultaneously, the efficiencies of both estimators are greater than the efficiencies of the corresponding optimum or near-optimum k quantile estimators.

6. Estimating Functions of α and β Using Quantiles

The mean μ and the standard deviation of the distribution with distribution function $G(x)$ are given by

$$\mu = C\beta + \alpha$$

$$\sigma = \frac{\pi\beta}{6^{1/2}}$$

where $C = 0.5772$ denotes Euler's constant. Quantile estimators of μ and σ, and their variances, are given by

$$\hat{\mu} = C\hat{\beta} + \hat{\alpha}$$

$$\operatorname{var}(\hat{\mu}) = C^2 \operatorname{var}(\hat{\beta}) + \operatorname{var}(\hat{\alpha}) + 2C \operatorname{cov}(\hat{\alpha}, \hat{\beta})$$

$$\hat{\sigma} = \frac{\pi\hat{\beta}}{6^{1/2}}$$

$$\operatorname{var}(\hat{\sigma}) = \frac{\pi^2}{6 \operatorname{var}(\hat{\beta})}$$

Table 9. Optimum and near-optimum estimators of α and their efficiencies when β is unknown ($k = 10$)

k	Estimators $\hat{\alpha}$	eff $(\hat{\alpha})$
1	$z(0.3679)$	0.6452
2	$0.5570\, z(0.1797) + 0.4430\, z(0.6023)$	0.8156
3	$0.3514\, z(0.1041) + 0\ 4089\, z(0.3705) + 0.2397\, z(0.7365)$	0.8863
4	$0.2423\, z(0.0676) + 0.3306\, z(0.2474) + 0.2838\, z(0.5193) + 0.1433\, z(0.8187)$	0.9226
5	$0.1691\, z(0.0466) + 0.2729\, z(0.1735) + 0.2763\, z(0.3837) + 0.1976\, z(0.6412) + 0.0841\, z(0.8763)$	0.9436
6	$0.1277\, z(0.0342) + 0.2220\, z(0.1294) + 0.2489\, z(0.2924) + 0.2124\, z(0.5051) + 0.1361\, z(0.7305) + 0.0529\, z(0.9131)$	0.9569
7	$0.0970\, z(0.0262) + 0.1779\, z(0.0969) + 0.2174\, z(0.2231) + 0.2091\ (0.3959) + 0.1637\, z(0.5954)$ $+ 0.0984\, z(0.7903) + 0.0365\, z(0.9352)$	0.9660
8	$0.0771\, z(0.0208) + 0.1462\, z(0.0757) + 0.1894\, z(0.1767) + 0.1961\, z(0.3188) + 0.1713\, z(0.4916) + 0.1250\, z(0.6746)$ $+ 0.0700\, z(0.8413) + 0.0249\, z(0.9525)$	0.9725
9	$0.0637\, z(0.0169) + 0.1247\, z(0.0622) + 0.1669\, z(0.1961) + 0.1806\, z(0.2669) + 0.1680\, z(0.4164) + 0.1362\, z(0.5805)$ $+ 0.0934\, z(0.7427) + 0.0496\, z(0.8793) + 0.0169\, z(0.9650)$	0.9771
10	$0.0547\, z(0.0146) + 0.1080\, z(0.0529) + 0.1480\, z(0.1239) + 0.1654\, z(0.2274) + 0.1610\, z(0.3573) + 0.1395\, z(0.5041)$ $+ 0.1069\, z(0.6561) + 0.0695\, z(0.7976) + 0.0351\, z(0.9088) + 0.0119\, z(0.9733)$	0.9806

Table 10. Optimum and near-optimum estimators of β and their efficiencies when α is unknown ($k = 10$)

k	Estimators $\hat{\beta}$	eff $(\hat{\beta})$
2	$0.3345\, [z(0.8326) - z(0.0262)]$	0.6635
3	$0.3440\, z(0.8159) - 0.2289\, z(0.0413) - 0.1151\, z(0.00624)$	0.7152
4	$0.1139\, [z(0.9290) - z(0.00701)] + 0.2360\, [z(0.7193) - z(0.0504)]$	0.8304
5	$0.1167\, z(0.9268) + 0.2336\, z(0.7100) - 0.1356\, z(0.0681) - 0.1448\, z(0.0227) - 0.0700\, z(0.00328)$	0.8509
6	$0.0510\, [z(0.9644) - z(0.00273)] + 0.1294\, [z(0.8496) - z(0.0185)] + 0.1817\, [z(0.6457) - z(0.0715)]$	0.8979
7	$0.0517\, z(0.9649) + 0.1350\, z(0.8428) + 0.1720\, z(0.6430) - 0.1019\, z(0.0867) - 0.1239\, z(0.0397)$ $- 0.0960\, z(0.0114) - 0.0369\, z(0.00159)$	0.9079
8	$0.0264\, [z(0.9798) - z(0.00129)] + 0.0743\, [z(0.9120) - z(0.00827)] + 0.1225\, [z(0.7838) - z(0.0309)]$ $+ 0.1464\, [z(0.5995) - z(0.0881)]$	0.9317
9	$0.0264\, z(0.9809) + 0.0795\, z(0.9089) + 0.1258\, z(0.7738) + 0.1329\, z(0.5976) - 0.0807\, z(0.1014) - 0.1037\, z(0.0549)$ $- 0.0980\, z(0.0220) - 0.0614\, z(0.00589) - 0.0208\, z(0.000831)$	0.9366
10	$0.0159\, [z(0.9866) - z(0.000775)] + 0.0462\, [z(0.9428) - z(0.00448)] + 0.0824\, [z(0.8561) - z(0.0159)]$ $+ 0.1113\, [z(0.7279) - z(0.0437)] + 0.1218\, [z(0.5601) - z(0.1036)]$	0.9509

Table 11. Sub-optimum estimators of α and β for $c = 1$

k	Estimators	eff
2	$\hat{\alpha} = 0.5671\, z(0.0865) + 0.4329\, z(0.7338)$	0.7334
	$\hat{\beta} = 0.4836\, [z(0.7338) - z(0.0865)]$	0.5661
4	$\hat{\alpha} = 0.1067\, z(0.0172) + 0.4025\, z(0.1388) + 0.3825\, z(0.5548) + 0.1083\, z(0.8783)$	0.8930
	$\hat{\beta} = 0.1879\, [z(0.8783) - z(0.0172)] + 0.2919\, [z(0.5548) - z(0.1388)]$	0.7632
6	$\hat{\alpha} = 0.0512\, z(0\ 00674) + 0.1661\, z(0.0463) + 0.2836\, z(0.1884) + 0.2769\, z(0.4496) + 0.1663\, z(0.7442)$ $+ 0.0559\, z(0.9331)$	0.9434
	$\hat{\beta} = 0.0930\, [z(0.9331) - z(0.00674)] + 0.1939\, [z(0.7442) - z(0.0463)]$ $+ 0.2009\, [z(0.4496) - z(0.1884)]$	0.8479
8	$\hat{\alpha} = 0.0243\, z(0.00343) + 0.0808\, z(0.0204) + 0.1628\, z(0.0769) + 0.2252\, z(0.2169) + 0.2232\, z(0.3986)$ $+ 0.1659\, z(0.6409) + 0.0891\, z(0.8412) + 0.0287\, z(0.9596)$	0.9647
	$\hat{\beta} = 0.0517\, [z(0.9596) - z(0.00343)] + 0.1216\, [z(0.8412) - z(0.0204)] + 0.1662\, [z(0.6409) - z(0.0769)]$ $+ 0.1497\, [z(0.3986) - z(0.2169)]$	0.8955
10	$\hat{\alpha} = 0.0118\, z(0.00162) + 0.0424\, z(0.00962) + 0.0942\, z(0.0356) + 0.1533\, z(0.1010) + 0.1909\, z(0.2322)$ $+ 0.1894\, z(0.3746) + 0.1534\, z(0.5822) + 0.1001\, z(0.7681) + 0.0495\, z(0.9031) + 0.0150\, z(0.9760)$	0.9743
	$\hat{\beta} = 0.0286\, [z(0.9760) - z(0.00162)] + 0.0752\, [z(0.9031) - z(0.00962)] + 0.1202\, [z(0.7681) - z(0.0356)]$ $+ 0.1393\, [z(0.5822) - z(0.1010)] + 0.1200\, [z(0.3746) - z(0.2322)]$	0.9259

Table 12. Sub-optimum estimators of α and β for $c = 2$

k	Estimators	eff
2	$\hat{\alpha} = 0.5592\, z(0.0606) + 0.4408\, z(0.7569)$	0.6863
	$\hat{\beta} = 0.4374\, [z(0.7569) - z(0.0606)]$	0.6077
4	$\hat{\alpha} = 0.0918\, z(0.0136) + 0.4170\, z(0.1117) + 0.3971\, z(0.5918) + 0.0941\, z(0.8929)$	0.8678
	$\hat{\beta} = 0.1649\, [z(0.8929) - z(0.0136)] + 0.2799\, [z(0.5918) - z(0.1117)]$	0.7882
6	$\hat{\alpha} = 0.0478\, z(0.00524) + 0.1609\, z(0.0368) + 0.2953\, z(0.1632) + 0.2851\, z(0.4792)$ $+ 0.1597\, z(0.7711) + 0.0512\, z(0.9421)$	0.9307
	$\hat{\beta} = 0.0811\, [z(0.9421) - z(0.00524)] + 0.1794\, [z(0.7711) - z(0.0368)]$ $+ 0.2003\, [z(0.4792) - z(0.1632)]$	0.8615
8	$\hat{\alpha} = 0.0215\, z(0.00251) + 0.0766\, z(0.0161) + 0.1617\, z(0.0633) + 0.2372\, z(0.1982) + 0.2334\, z(0.4176)$ $+ 0.1617\, z(0.6758) + 0.0828\, z(0.8618) + 0.0251\, z(0.9662)$	0.9576
	$\hat{\beta} = 0.0439\, [z(0.9662) - z(0.00251)] + 0.1106\, [z(0.8618) - z(0.0161)] + 0.1602\, [z(0.6758) - z(0.0633)]$ $+ 0.1512\, [z(0.4176) - z(0.1982)]$	0.9046
10	$\hat{\alpha} = 0.0108\, z(0.00129) + 0.0399\, z(0.00790) + 0.0911\, z(0.0299) + 0.1533\, z(0.0875) + 0.1992\, z(0.2217)$ $+ 0.1971\, z(0.3824) + 0.1529\, z(0.6077) + 0.0963\, z(0\ 7886) + 0.0460\, z(0.9141) + 0.0134\, z(0.9795)$	0.9700
	$\hat{\beta} = 0.0250\, [z(0.9795) - z(0.00129)] + 0.0688\, [z(0.9141) - z(0.00790)] + 0.1143\, [z(0.7886) - z(0.0299)]$ $+ 0.1379\, [z(0.6077) - z(0.0875)] + 0.1208\, [z(0.3824) - z(0.2217)]$	0.9312

A percentage point x_p of the distribution is defined by

$$p = \exp\left\{- \exp\left[- \frac{1}{\beta}(x_p - \alpha)\right]\right\}$$

Then one has

$$\frac{1}{\beta}(x_p - \alpha) = -\ln(-\ln p)$$

$$x_p = -\beta \ln(-\ln p) + \alpha$$

A quantile estimator of x_p and its variance are given by

$$\hat{x}_p = -\hat{\beta} \ln(-\ln p) + \hat{\alpha}$$

$$\operatorname{var}(\hat{x}_p) = [\ln(-\ln p)]^2 \operatorname{var}(\hat{\beta}) + \operatorname{var}(\hat{\alpha}) - 2 \ln(-\ln p) \operatorname{cov}(\hat{\alpha}, \hat{\beta})$$

One might wish to estimate the probability that x will not exceed some threshold value x_0. Thus,

$$\operatorname{pr}(x < x_0) = S = \exp\left\{-\exp\left[-\frac{1}{\beta}(x - \alpha)\right]\right\}$$

and a quantile estimator of S is given by

$$\hat{S} = \exp\left\{-\exp\left[-\frac{1}{\hat{\beta}}(x - \hat{\alpha})\right]\right\}$$

The approximate variance of $\hat{S}$ is given by

$$\operatorname{var}(\hat{S}) \cong \frac{(S \ln S)^2}{\beta^2} \{\operatorname{var}(\hat{\alpha}) + [\ln(-\ln S)]^2 \operatorname{var}(\hat{\beta}) - 2 [\ln(-\ln S)] \operatorname{cov}(\hat{\alpha}, \hat{\beta})\}$$

7. Estimating α and β From Real Data Using Quantiles

In order to obtain a sample quantile z_p of order p from a sample of size n drawn from a population with distribution function $G(x)$, a table of random digits can be used. A set of n k-digit numbers is drawn from the table and the sample quantile of order p, say v_p, is determined from this sample. Then the desired sample quantile z_p of $G(x)$ is obtained by solving for z_p in the equation

$$(v_p + 0.5)\, 10^{-k} = G(z_p)$$

This procedure was adopted in order to obtain sample quantiles necessary for estimating $\alpha = 0$ and $\beta = 1$. Two sets of sample values, sample A and sample B, each of size 500, were drawn from a table of random digits (Ref. 3). For each sample, the suboptimum quantiles were determined for both $C = 1$ and $C = 2$, and used to estimate α and β. The results are as follows ($\hat{\alpha}_k$ and $\hat{\beta}_k$ will denote the estimates of α and β using k suboptimum quantiles):

From sample A, with $C = 1$

$\hat{\alpha}_2 = 0.0006$	$\hat{\beta}_2 = 1.0059$
$\hat{\alpha}_4 = 0.0576$	$\hat{\beta}_4 = 0.9640$
$\hat{\alpha}_6 = 0.0436$	$\hat{\beta}_6 = 0.9625$
$\hat{\alpha}_8 = 0.0387$	$\hat{\beta}_8 = 0.9943$
$\hat{\alpha}_{10} = 0.0291$	$\hat{\beta}_{10} = 1.0044$

From sample A, with $C = 2$

$\hat{\alpha}_2 = -0.0044$	$\hat{\beta}_2 = 1.0018$
$\hat{\alpha}_4 = 0.0400$	$\hat{\beta}_4 = 0.9664$
$\hat{\alpha}_6 = 0.0333$	$\hat{\beta}_6 = 0.9795$
$\hat{\alpha}_8 = 0.0396$	$\hat{\beta}_8 = 0.9798$
$\hat{\alpha}_{10} = 0.0257$	$\hat{\beta}_{10} = 0.9875$

From sample B, with $C = 1$

$\hat{\alpha}_2 = -0.0339$	$\hat{\beta}_2 = 0.9621$
$\hat{\alpha}_4 = -0.0354$	$\hat{\beta}_4 = 1.0589$
$\hat{\alpha}_6 = -0.0173$	$\hat{\beta}_6 = 1.0443$
$\hat{\alpha}_8 = -0.0256$	$\hat{\beta}_8 = 1.0309$
$\hat{\alpha}_{10} = -0.0167$	$\hat{\beta}_{10} = 1.0444$

From sample B, with $C = 2$

$\hat{\alpha}_2 = -0.0468$	$\hat{\beta}_2 = 0.9949$
$\hat{\alpha}_4 = -0.0282$	$\hat{\beta}_4 = 1.0334$
$\hat{\alpha}_6 = -0.0286$	$\hat{\beta}_6 = 1.0500$
$\hat{\alpha}_8 = -0.0252$	$\hat{\beta}_8 = 1.0375$
$\hat{\alpha}_{10} = -0.0092$	$\hat{\beta}_{10} = 1.0387$

References

1. Posner, E. C., "The Application of Extreme Value Theory to Error-Free Communication," *Technometrics*, Vol. 7, No. 4, pp. 517–529, Nov. 1965.
2. Cramer, H., *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N. J., 1946.
3. The Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates*. The Free Press, Glencoe, Ill., 1955.

O. Data Compression Techniques: Mass Spectrogram Data Compression by the Slope Threshold Method, *L. Kleinrock*

A complete description of the slope threshold method of data compression is given in SPS 37-49, Vol. III, pp. 325–328.

The data used for this experiment was randomly generated by choosing 33 integers from the set

$$\{1, 2, 3, \cdots, 100\}$$

with replacement, and using these as the time location of 33 peaks for the mass spectrogram. At each of these points, the non-negative amplitude of the peak was chosen from a geometric distribution whose mean was 10. Each peak was then converted into a triangular pulse with height equal to the chosen peak value and with a base of width equal to 2 time units, centered on the original time location. The sum of these triangles resulted in the generated data shown in Fig. 29.

In Fig. 30, we show the results of compressing the mass spectrogram using the slope threshold method for the various values of a shown and with $b = 0$ (see SPS 37-49, Vol. III, pp. 325–328, for details of the compression algorithm). In Fig. 31, we show the result of compression by periodic sampling. Table 13 lists the parameter values and the rms error, as well as the Posner norm ϵ_1 (the rms error is merely ϵ_0), for the two sampling methods. This

Table 13. Experimental results

Figure	a	b	Period	ϵ_0	ϵ_1	Gross compression ratio
29	1	0	—	0	0	1.56
30a	2	0	—	0.1	0.17	1.67
30b	3	0	—	0.28	0.47	1.79
30c	4	0	—	0.52	0.88	2.0
30d	5	0	—	0.75	1.19	2.28
30e	6	0	—	1.05	1.71	2.44
30f	7	0	—	1.27	2.1	2.7
30g	8	0	—	1.79	2.62	2.94
30h	11	0	—	3.15	4.56	4.16
30i	15	0	—	3.6	5.3	5.0
30j	20	0	—	4.55	6.33	5.26
30k	25	0	—	6.35	8.04	7.7
30l	30	0	—	22.4	23.5	16.7
31a	—	—	1	0	0	1.0
31b	—	—	2	5.23	9.07	2.0
31c	—	—	3	5.5	9.67	3.0
31d	—	—	4	6.26	11.14	4.0
31e	—	—	5	6.1	10.3	5.0
31f	—	—	6	6.8	11.55	6.0
31g	—	—	7	7.8	11.7	7.0

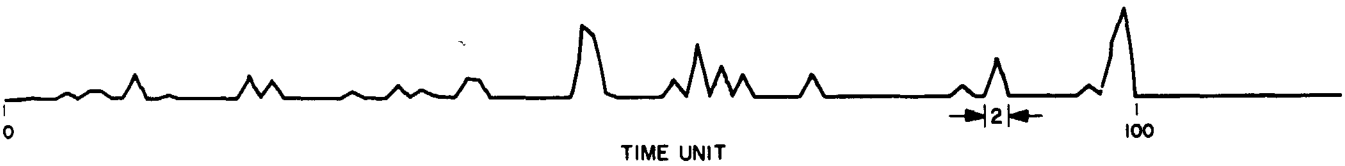


Fig. 29. Randomly generated mass spectrogram

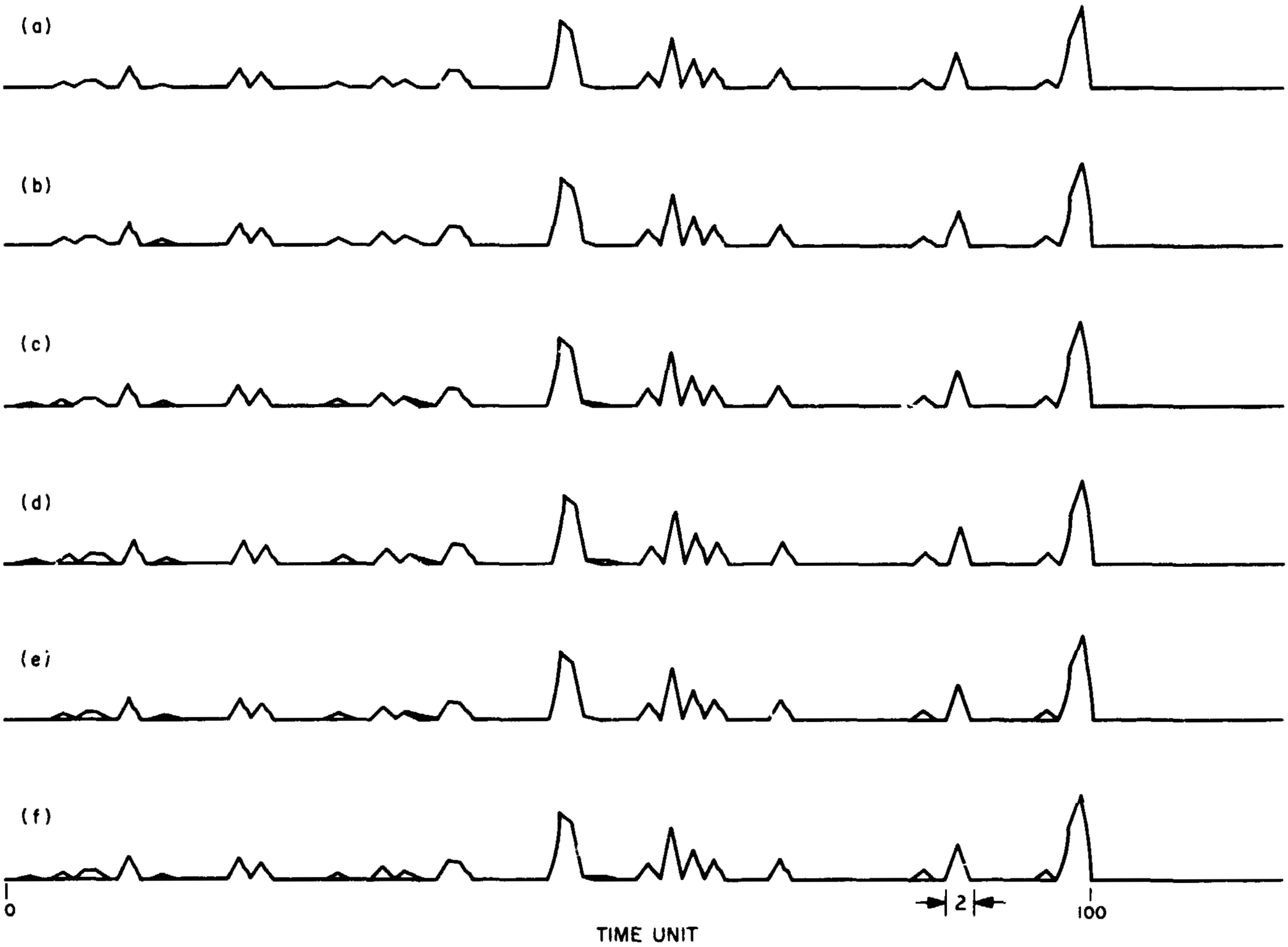


Fig. 30. Slope threshold sampling: (a) a = 2, (b) a = 3, (c) a = 4, (d) a = 5, (e) a = 6, (f) a = 7, (g) a = 8, (h) a = 11, (i) a = 15, (j) a = 20, (k) a = 25, (l) a = 30

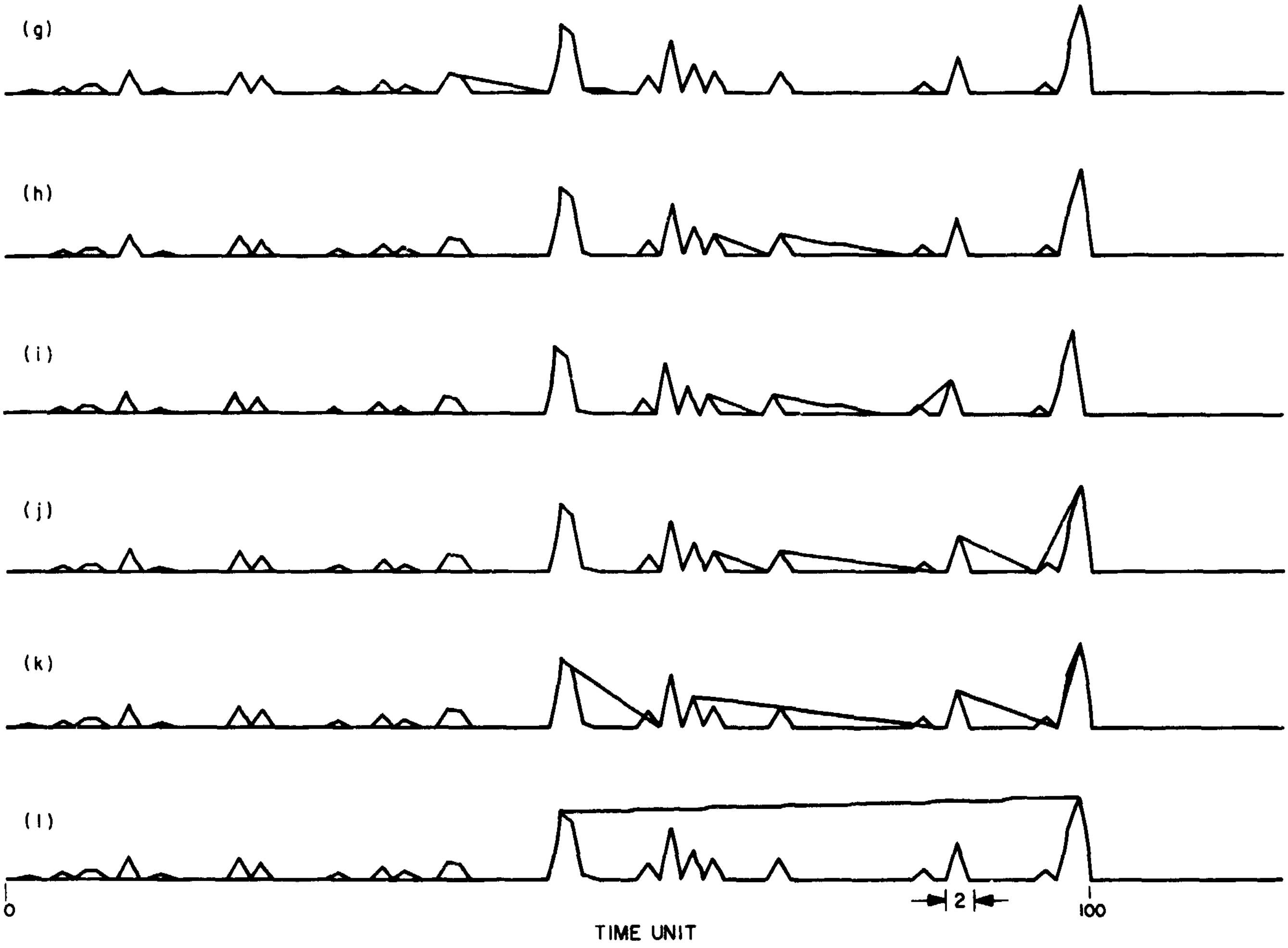


Fig. 30 (contd)

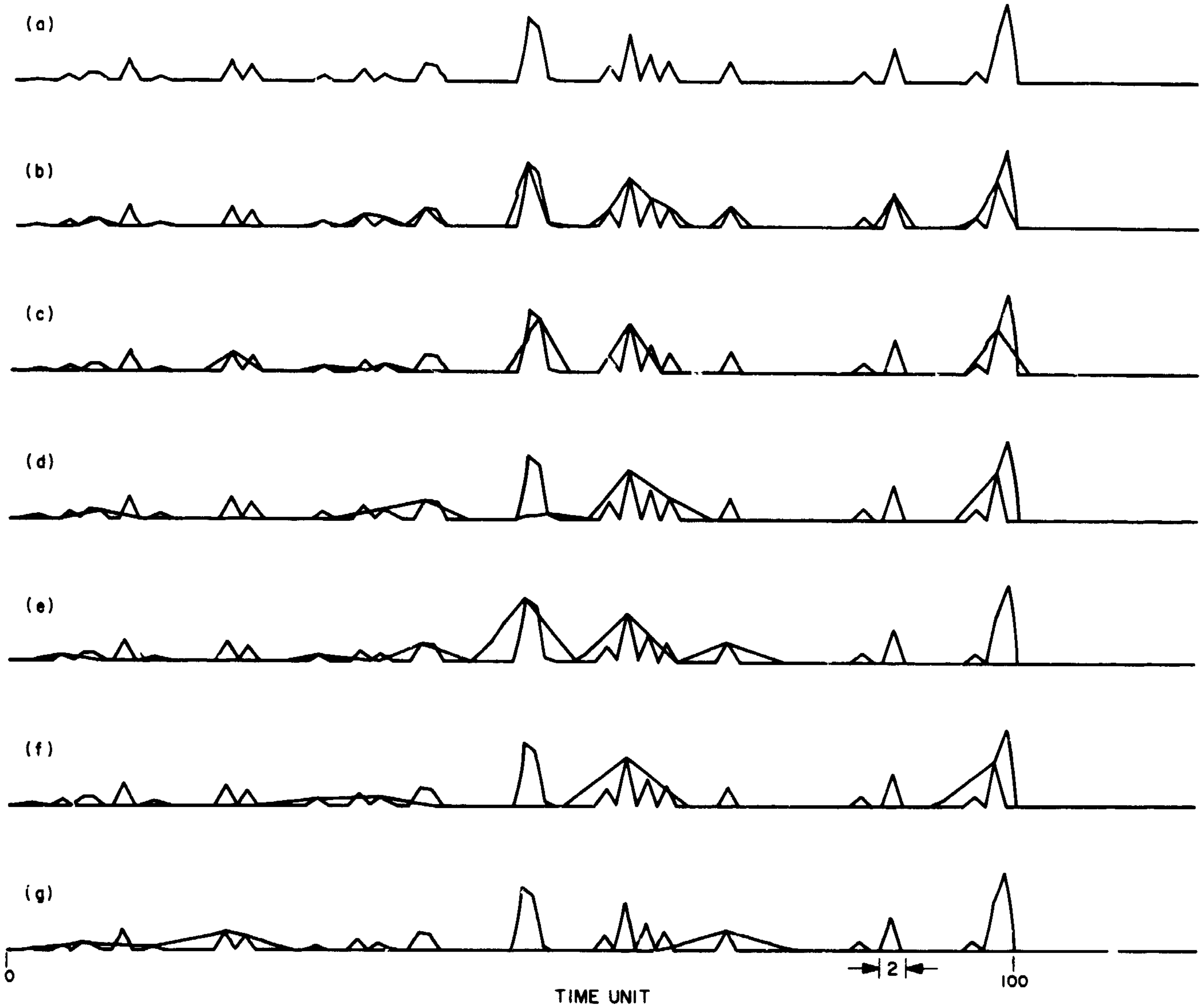


Fig. 31. Periodic sampling: (a) period = 1, (b) period = 2, (c) period = 3, (d) period = 4, (e) period = 5, (f) period = 6, (g) period = 7

norm, suggested by E. C. Posner, is defined as

$$\epsilon_\alpha = \left[\frac{1}{N}\sum_{n=1}^{N}(f_n - \hat{f}_n)^2 + \frac{\alpha}{N-1}\sum_{n=2}^{N}(\Delta f_n - \Delta\hat{f}_n)^2\right]^{1/2}$$

Figures 30 and 31 show the reconstructed function, f_n, superimposed on the original data, f_n. For example, in Fig. 30f, we see that f_n has missed a number of peaks. It is interesting to observe the behavior of the Posner norm, which is designed to measure the mean-squared amplitude error plus α times the mean-squared slope error. Figure 30b and Table 13 show that ϵ_1 is 70% larger than ϵ_0, indicating that the slope error is almost as significant as the amplitude error. The extreme case shown in Fig. 30l shows that the slope error is insignificant compared to the amplitude error; in Fig. 30k, the slope error is almost the same as in Fig. 30l, but the amplitude error is much reduced. We conclude that the use of the Posner norm here is more significant as the mean-squared amplitude error decreases.

In Fig. 32, we plot the Posner norm (for $\alpha = 0, 1$) as a function of gross compression ratio. We observe for moderate compression ratios (less than 3) that the slope threshold method of sampling is far superior to periodic

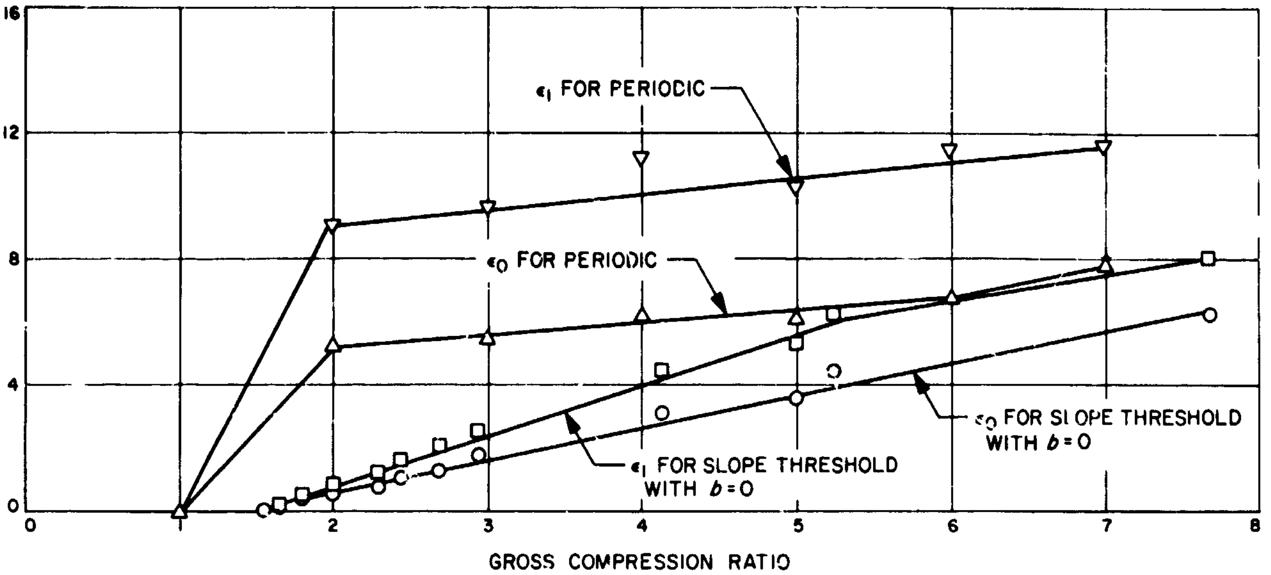


Fig. 32. Comparison of periodic sampling and slope threshold sampling, using Posner norm ($\alpha = 0, 1$)

sampling. In this range, we observe only slight distortion of the peaks. However, if one were to transmit only the peaks themselves (of which there are less than 33), one obtains a gross compression ratio of approximately 3 at small cost. We therefore conclude that the slope threshold method of data compression for mass spectrogram is not practical.

P. Data Compression Techniques: Estimating the Correlation Between Two Normal Populations Using Quantiles of Conditional Distributions,

I. Eisenberger

1. Introduction

The problem of estimating the parameters of a univariate normal distribution using quantiles when the sample size is large is considered in Ref. 1, where estimators of the mean and standard deviation are given using up to twenty quantiles. If a set of pairs of sample values taken from a bivariate normal distribution is given, one must also estimate the correlation in order to completely describe the distribution. The problem of estimating the correlation coefficient ρ using quantiles is considered in Refs. 2, 3, and 4, where asymptotically unbiased estimators of ρ are constructed using up to eight sample quantiles. However, before constructing the estimators, it was necessary to perform a linear transformation on the sample pairs in order to obtain a new set of *independent* pairs. Since, from the viewpoint of data compression of space telemetry, this procedure is not entirely satisfactory due to the equipment complexity, it was felt that a new approach to the problem of estimating ρ was desirable.

It is reasonable to conjecture that if one considers the quantiles of the conditional distribution of, say, y, it might be possible to construct satisfactory estimators of ρ without a transformation of variables. As a result of the ensuing investigation, quantile estimators of ρ will be given when the quantiles are taken from the conditional distribution of y given that x lies in specified intervals, for a large sample size. These estimators are very nearly unbiased, with good efficiencies relative to the maximum-likelihood estimator when ρ is not too large.

2. Review of Quantiles

To define a quantile, consider a sample of n independent values, $x_1, x_2, \cdots, x_n$, taken from a distribution of a continuous type with distribution function $G(x)$ and density function $g(x)$. The quantile of order p of the distribution or population, denoted by ζ_p, is defined as the root of the equation $G(\zeta_p) = p$: that is,

$$p = \int_{-\infty}^{\zeta_p} dG(x) = \int_{-\infty}^{\zeta_p} g(x)\,dx$$

The corresponding *sample* quantile Z_p is defined as follows: If the sample values are arranged in nondecreasing order of magnitude

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

then $x_{(i)}$ is called the ith order statistic and

$$Z_p = x_{([np]+1)}$$

where $[np]$ is the greater integer $\leqq np$.

If $g(x)$ is differentiable in some neighborhood of each quantile value considered, it has been shown (Ref. 5) that the joint distribution of any number of quantiles is asymptotically normal as $n \to \infty$ and that, asymptotically,

$$E(Z_p) = \zeta_p$$

$$\operatorname{var}(Z_p) = \frac{p(1-p)}{ng^2(\zeta_p)}$$

$$\rho_{12} = \left[\frac{p_1(1-p_2)}{p_2(1-p_1)}\right]^{1/2}$$

where ρ_{12} is the correlation between Z_{p_1} and Z_{p_2}, $p_1 < p_2$. Since n is assumed to be large, the statistical analysis to be given will be based on the asymptotic distribution of the sample quantiles. We will denote by Z_i the quantile of order p_i, and it should always be assumed that $p_i < p_j$ when $i < j$.

3. The Distribution and Moments of $v = y \mid a < x < b$

Given a set of n independent pairs of sample values, $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, taken from two jointly normal standard distributions with distribution functions $F(x)$ and $F(y)$, density functions $f(x)$ and $f(y)$, and joint density $h(x, y)$, we derive the distribution of y given that x lies in the interval $a < x < b$, that is, we consider the random variable $v = y \mid a < x < b$. Denoting by $G(v)$ and $g(v)$ the distribution function and density function of v, respectively, one has

$$G(v) = \operatorname{pr}(V < v) = \frac{\int_{-\infty}^{v}\int_{a}^{b} h(x, t)\,dx\,dt}{\int_{a}^{b} f(x)\,dx} \tag{1}$$

Differentiating Eq. (1) with respect to v results in

$$g(v) = \frac{\partial G(v)}{\partial v} \frac{\int_a^b h(x, v)\,dx}{F(b) - F(a)}$$

$$= \frac{\dfrac{1}{2\pi(1-p^2)^{1/2}} \exp\left(-\dfrac{1}{2}v^2\right)\displaystyle\int_a^b \exp\left\{-\dfrac{1}{2}\left[\dfrac{x-\rho v}{(1-\rho^2)^{1/2}}\right]^2\right\}dx}{F(b) - F(a)}$$

$$= \frac{\dfrac{1}{(2\pi)^{1/2}} \exp\left(-\dfrac{1}{2}v^2\right)\displaystyle\int_{a-\rho v/(1-\rho^2)^{1/2}}^{b-\rho v/(1-\rho^2)^{1/2}} \exp\left(-\dfrac{1}{2}Z^2\right)dZ}{F(b) - F(a)}$$

$$= f(v)\left[\frac{F\left(\dfrac{b-\rho v}{[1-\rho^2]^{1/2}}\right) - F\left(\dfrac{a-\rho v}{[1-\rho^2]^{1/2}}\right)}{F(b) - F(a)}\right] \tag{2}$$

We will derive the mean and variance of v from the moments of the truncated variable $x \mid a < x < b$. It is shown in SPS 37-38, Vol. IV, pp. 252–258, that the mean μ_x and variance σ_x^2 of this truncated variable are given by

$$\mu_x = \frac{f(a) - f(b)}{F(b) - F(a)}$$

$$\sigma_x^2 = 1 + \frac{af(a) - bf(b)}{F(b) - F(a)} - \left[\frac{f(b) - f(a)}{F(b) - F(a)}\right]^2$$

It is also well known that the mean and variance of the conditional distribution of $y|x$ are given by

$$E(y|x) = \rho x$$

$$\mathrm{var}(y|x) = 1 - \rho^2$$

Now,

$$E(v) = E(y|a < x < b) = E[E(y|x|a < x < b)]$$

$$= E(\rho x|a < x < b)$$

$$= \rho\mu_x$$

Similarly,

$$E(v^2) = E(y^2|a < x < b) = E[E(y^2|x|a < x < b)]$$

$$= E(1 - \rho^2 + \rho^2 x^2|a < x < b)$$

$$= 1 - \rho^2 + \rho^2(\sigma_x^2 + \mu_x^2)$$

Thus, one has

$$\mu_v = E(v) = \rho\mu_x$$

$$\sigma_v^2 = \mathrm{var}(v) = 1 + \rho^2(\sigma_x^2 - 1)$$

4. Estimators of ρ Using Quantiles

We divide the x-axis into the six intervals I_k: $a_k < x < b_k$, $k = 1, 2, \cdots, 6$, where

$$a_1 = -\infty$$

$$a_2 = -a_6$$

$$a_3 = -a_5$$

$$a_4 = 0$$

$$b_k = a_{k+1} \text{ for } k = 1, \cdots, 5$$

$$b_6 = \infty$$

This partitions the x-axis into three pairs of symmetric regions. For each region, we will estimate ρ using two pairs of optimum symmetric quantiles taken from the set of y values such that the corresponding x values fall into the given region. Denoting by $\hat{\rho}_k$ the estimator of ρ from the v_i of I_k, we then form the estimator

$$\bar{\rho}_y = \sum_{i=1}^{3} C_i(\hat{\rho}_i + \hat{\rho}_{7-i})$$

determining the C_i so as to minimize $\mathrm{var}(\bar{\rho})$ under the condition that

$$2\sum_{i=1}^{3} C_i = 1$$

Thus, let Z_i be a sample quantile of order p_i taken from a set $\{v_k\}$, for $i = 1, 2, 3, 4$ such that $p_4 = 1 - p_1$ and $p_3 = 1 - p_2$. Then

$$E(Z_i) = \zeta_i = \mu_v + \sigma_v\zeta_i^* = \rho\mu_x + \sigma_v\zeta_i^*$$

where ζ_i^* is the population quantile of order p_i of the standardized distribution of v. Although strictly speaking, the sample size of each of the sets $\{v_i\}$ is a random quantity, we will take as the variance of Z_i the approximation

$$\mathrm{var}(Z_i) = \frac{p_i(1 - p_i)}{m_i g^2(\zeta_i)}$$

where

$$m_i = n\,\mathrm{pr}(a_i < x < b_i)$$

This means that we are taking as the sample size the expected number of x's falling in the interval $a < x < b$.

Forming the estimator

$$\hat{\rho}_i = \frac{0.1918(Z_1 + Z_4) + 0.3082(Z_2 + Z_3)}{\mu_x} \qquad (3)$$

where

$$p_1 = 0.1068, \qquad p_3 = 0.6488$$

$$p_2 = 0.3512, \qquad p_4 = 0.8932$$

one has

$$E(\hat{\rho}_i) = \frac{0.1918[2\rho\mu_x + \sigma_v(\zeta_1^* + \zeta_4^*)] + 0.3082[2\rho\mu_x + \sigma_v(\zeta_v^* + \zeta_3^*)]}{\mu_x}$$

$$= \rho + \sigma_v\left[\frac{0.1918(\zeta_1^* + \zeta_4^*) + 0.3082(\zeta_2^* + \zeta_3^*)}{\mu_x}\right] \qquad (4)$$

The orders of the quantiles and the values of the coefficients of $\hat{\rho}_i$ were chosen for two reasons. First, when $\rho = 0, g(v) = f(v)$ and the numerator of Eq. (3) becomes the best unbiased estimator of the mean of a normal distribution using four quantiles, and hence has the smallest variance. Secondly, it was found after repeated trials that if the estimator $\bar{\rho}_y$ using one set of quantiles had a smaller variance than the estimator had using another set, when $\rho = 0$, then the same result held when $\rho \neq 0$.

The variance of $\bar{\rho}_y$ depends upon the choice of a_5 and a_6, that is, on how we partition the positive x-axis into three regions. It was determined that if one chooses $a_5 = 0.8$ and $a_6 = 1.5$, the resulting estimator $\bar{\rho}_y$ will be very nearly optimum. However, the optimum choice of the C_i for given values of a_5 and a_6 depends upon the value of ρ. If one determines the C_i from

$$C_i = \frac{\dfrac{1}{\operatorname{var}(\hat{\rho}_i)}}{\displaystyle\sum_{k=1}^{6} \frac{1}{\operatorname{var}(\hat{\rho}_k)}}$$

$\operatorname{var}(\bar{\rho}_y)$ will be minimized for a given value of ρ, but since ρ is not known in advance, one set of the C_i must be chosen for all possible values of ρ. It was found that by using in $\bar{\rho}_y$ the optimum values of the C_i for $\rho = 0.5$, very little loss in efficiency resulted for ρ between 0 and ± 9. Thus, the estimator $\bar{\rho}_y$ that we propose is given by

$$\bar{\rho}_y = 0.2632\,(\hat{\rho}_1 + \hat{\rho}_6) + 0.1920\,(\hat{\rho}_2 + \hat{\rho}_6) + 0.0448\,(\hat{\rho}_3 + \hat{\rho}_4)$$

and, because $\hat{\rho}_i$ is independent of $\hat{\rho}_j$ for $i \neq j$, the variance of $\bar{\rho}_y$ is given by

$$\operatorname{var}(\bar{\rho}_y) = 2\,[(0.2632)^2 \operatorname{var}(\hat{\rho}_1) + (0.1920)^2 \operatorname{var}(\hat{\rho}_2) + (0.0448)^2 \operatorname{var}(\hat{\rho}_3)]$$

The value of the bias term of $\hat{\rho}_i$, the second term of the right-hand side of Eq. (4), depends upon the degree of symmetry of $g(v)$. If $g(v)$ were symmetric, then $\zeta_1^* = -\zeta_4^*$, $\zeta_2^* = -\zeta_3^*$, and $E(\hat{\rho}) = \rho$. Fortunately, $g(v)$ is sufficiently symmetric, for ρ between 0 and 0.9, that the bias term is negligible. This is shown in Table 14, which lists the mean and variance of $\hat{\rho}_k$, $k = 1, \cdots, 6$, and $\bar{\rho}_y$, for ρ between 0 and 0.9.

Table 14. Mean and variance of $\hat{\rho}_k$ and $\bar{\rho}_y$ [$\bar{\rho}_y = 0.2632\,(\hat{\rho}_1 + \hat{\rho}_6) + 0.1920\,(\hat{\rho}_2 + \hat{\rho}_5) + 0.0448\,(\hat{\rho}_3 + \hat{\rho}_4)$]

ρ	$E(\hat{\rho}_1)$ $E(\hat{\rho}_6)$	$n \operatorname{var}(\hat{\rho}_1)$ $n \operatorname{var}(\hat{\rho}_6)$	$E(\hat{\rho}_2)$ $E(\hat{\rho}_5)$	$n \operatorname{var}(\hat{\rho}_2)$ $n \operatorname{var}(\hat{\rho}_5)$	$E(\hat{\rho}_3)$ $E(\hat{\rho}_4)$	$n \operatorname{var}(\hat{\rho}_3)$ $n \operatorname{var}(\hat{\rho}_4)$	$E(\rho_y)$	$\operatorname{var}(\bar{\rho}_y)$
0	0	4.328	0	6.145	0	26.232	0	1.158
0.1	0.0999	4.289	0.1000	6.087	0.1001	25.984	0.1000	1.147
0.2	0.1999	4.179	0.2001	5.910	0.2000	25.239	0.2000	1.116
0.3	0.2998	3.995	0.3001	5.615	0.3000	23.995	0.2999	1.064
0.4	0.3997	3.737	0.4001	5.202	0.3999	22.254	0.3999	0.991
0.5	0.4995	3.405	0.5001	4.669	0.5000	20.017	0.4998	0.896
0.6	0.5991	2.998	0.6001	4.020	0.5999	17.281	0.5996	0.781
0.7	0.6985	2.516	0.7000	3.252	0.6998	14.050	0.6992	0.645
0.8	0.7973	1.958	0.8000	2.367	0.7997	10.323	0.7986	0.487
0.9	0.8947	1.321	0.8998	1.364	0.8993	6.106	0.8971	0.308

It is of interest to determine the efficiency of $\bar{\rho}_y$ relative to several commonly used estimators involving all the sample values, such as:

(1) The maximum-likelihood estimator ρ^*, the solution of the equation

$$(\rho^*)^3 - c\,(\rho^*)^2 + (a + b - 1)\,\rho^* - c = 0$$

where

$$a = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$c = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

(2) The sample correlation coefficient r, given by

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left\{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]\right\}^{1/2}}$$

(3) The easily computed estimator $\tilde{\rho}$, given by

$$\tilde{\rho} = \frac{1}{n}\sum x_i y_i$$

The asymptotic variances of the above estimators are given by

$$\mathrm{var}\,(\rho^*) = \frac{(1-\rho^2)^2}{n(1+\rho^2)}$$

$$\mathrm{var}\,(r) = \frac{(1-\rho^2)^2}{n}$$

$$\mathrm{var}\,(\tilde{\rho}) = \frac{1+\rho^2}{n}$$

Defining the efficiency of $\bar{\rho}_y$ relative to any other estimator $\hat{\rho}$ as

$$\mathrm{eff}\,(\bar{\rho}_y) = \frac{\mathrm{var}\,(\hat{\rho})}{\mathrm{var}\,(\bar{\rho}_y)}$$

Table 15 gives the efficiency of $\bar{\rho}_y$ relative to the above three estimators.

By applying to the x-values the method described above for obtaining $\bar{\rho}_y$, one also obtains $\bar{\rho}_x$ with identical statistical properties. One can then form the final estimator $\bar{\rho}$ given by

$$\bar{\rho} = \frac{1}{2}(\bar{\rho}_x + \bar{\rho}_y)$$

Table 15. Efficiency of $\bar{\rho}_y$ relative to ρ^*, r, and $\tilde{\rho}$

ρ	eff $(\bar{\rho}_y)$		
	Relative to ρ^*	Relative to r	Relative to $\tilde{\rho}$
0	0.864	0.864	0.864
0.1	0.846	0.854	0.880
0.2	0.794	0.826	0.932
0.3	0.714	0.778	1.025
0.4	0.614	0.712	1.171
0.5	0.502	0.628	1.394
0.6	0.386	0.524	1.741
0.7	0.271	0.403	2.311
0.8	0.162	0.266	3.366
0.9	0.065	0.117	5.875

In order to compute var $(\bar{\rho})$, one must determine the correlation between a quantile Z_p of order p taken from $y\,|\,a < x < b$ and a quantile Z'_q of order q taken from $x\,|\,c < y < d$. If $E(Z_p) = \eta$ and $E(Z'_q) = \zeta$, then E. Rodemich has shown that the asymptotic correlation ρ_{pq} between Z_p and Z'_q is given by

$$\rho_{pq} = \frac{N}{[pq(1-p)(1-q)\,\mathrm{pr}\,(a<x<b)\,\mathrm{pr}\,(c<y<d)]^{1/2}}$$

where

$$N = \mathrm{pr}(a<x<\zeta, c<y<\eta) - p\,\mathrm{pr}\,(a<x<b, c<y<\eta)$$
$$- q\,\mathrm{pr}\,(a<x<\zeta, c<y<d)$$
$$+ pq\,\mathrm{pr}(a<x<b, c<y<d)$$

when $a < \zeta < b$ and $c < \eta < d$. If $\zeta < a$, the terms in N which contain the condition $z < x < \zeta$ become zero, and, similarly, for $\eta < c$. If $\zeta > b$, the condition $a < x < \zeta$ should be written $z < x < b$, and if $\eta > d$, $c < y < \eta$ should be written $c < y < d$.

The extensive computations necessary to compute var $(\bar{\rho})$ for even the simplest case, $\rho = 0$, were carried out for this case, resulting in

$$\mathrm{var}\,(\bar{\rho}\,|\,\rho = 0) = \frac{1}{4}[2\,\mathrm{var}\,(\rho_y) + 2\,\mathrm{cov}\,(\bar{\rho}_x, \bar{\rho}_y)]$$
$$= \frac{1}{2}(1.158 + 1.068) = 1.113$$

Estimators $\bar{\rho}'_y$ and $\bar{\rho}'_x$ with the same statistical properties as $\bar{\rho}_y$ and $\bar{\rho}_i$ can be obtained by a somewhat simpler procedure. To the set of y_i's such that $x_i \in I_k$ add the set of y_i's, *with their signs changed*, such that $x_i \in I_{7-k}$, for $k = 4, 5, 6$. Then form the quantile estimators $\tilde{\hat{\rho}}_k$ and $\tilde{\hat{\rho}}_y$, given by

$$\tilde{\hat{\rho}}_k = 0.1918\,(Z_1 + Z_4) + 0.3082\,(Z_2 + Z_3)\,,$$
$$k = 4, 5, 6$$

$$\tilde{\hat{\rho}}_y = 0.5264\,\hat{\rho}_6 + 0.3840\,\hat{\rho}_5 + 0.0896\,\hat{\rho}_4$$

Then one has

$$\mathrm{var}\,(\tilde{\hat{\rho}}_k) = \frac{1}{2}\,\mathrm{var}\,(\hat{\rho}_k)\,, \qquad k = 4, 5, 6$$

$$E\,(\bar{\rho}'_y) = E\,(\bar{\rho}_y)$$

$$\mathrm{var}\,(\bar{\rho}'_y) = \mathrm{var}\,(\bar{\rho}_y)$$

The estimator $\bar{\rho}'_x$ is obtained in a similar fashion.

5. Estimating ρ

Two sets of samples $\{x_i\}$ and $\{y'_i\}$, each containing 600 sample values, were drawn from a table of random numbers in which the entries are distributed $N\,(0, 1)$. The transformation

$$y_i = 0.6\,x_i + 0.8\,y'_i$$

was then performed. Consequently, each x_i and y_i can be assumed to be distributed $N\,(0, 1)$ with a correlation of $\rho = 0.6$. Using the method involving $x_i \operatorname{sgn} y_i$ and $y_i \operatorname{sgn} x_i$ to estimate ρ using quantiles resulted in the following:

$$\bar{\rho}_x = 0.5601\,, \qquad \bar{\rho}_y = 0.5909\,, \qquad \bar{\rho} = 0.5755$$

The following estimates were also obtained:

$$\rho^* = 0.5771\,, \qquad \tilde{\rho} = 0.5404\,, \qquad r = 0.5644$$

Two new sets of 600 values each were then drawn from the same table of random numbers and paired at random, so that one can assume that $\rho = 0$. The results were:

$$\bar{\rho}_x = -0.0070\,, \qquad \bar{\rho}_y = -0.0060\,, \qquad \bar{\rho} = -0.0065$$

$$\rho^* = 0.0027\,, \qquad \tilde{\rho} = -0.0041\,, \qquad r = -0.0102$$

References

1. 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 published as Technical Report 32-510, Jet Propulsion Laboratory, Pasadena, Calif., Oct. 1, 1963.
2. Eisenberger, I., *Tests of Hypotheses and Estimation of the Correlation Coefficient using Quantiles I*, Technical Report 32-718. Jet Propulsion Laboratory, Pasadena, Calif., June 1, 1965.
3. Eisenberger, I., *Tests of Hypotheses and Estimation of the Correlation Coefficient using Quantiles II*, Technical Report 32-755. Jet Propulsion Laboratory, Pasadena, Calif., Sept. 15, 1965.
4. Eisenberger, I., *Tests of Hypotheses and Estimation of the Correlation Coefficient using Six and Eight Quantiles*, Technical Report 32-1163. Jet Propulsion Laboratory, Pasadena, Calif., Jan. 1, 1968.
5. Cramer, H., *Mathematical Methods of Statistics*. Princeton University Press, Princeton, N. J., 1946.

XXI. Communications Elements Research

TELECOMMUNICATIONS DIVISION

A. RF Techniques: Switching Frequency Determination for the Nodding Subdish System, *T. Sato, W. V. T. Rusch, C. T. Stelzried, S. D. Slobin, O. B. Parham*

The Nodding Subdish System (NSS), used in the October 1967 lunar eclipse measurements (SPS 37-50, Vol. III, pp. 290–295) takes the place of the microwave switch used in conventional Dicke radiometers. The advantages of the NSS are the minimization of loss and the reduction of atmospheric scintillation effects.

Because the NSS is a mechanical device, the number of switching cycles per second is limited as excessive speed leads to rapid wear and possible self-destruction. An original switching frequency of 1.16 Hz was chosen to ensure NSS longevity. During the final system checks prior to the eclipse observation, the noise output of the radiometer was larger than anticipated. A series of radiometer noise output measurements were made at various switching rates selected to be non-harmonically related to 60 Hz. The general trend of these data suggested a noise decrease with increased switching frequency.

The radiometer was reconfigured into a conventional Dicke radiometer, a ferrite switch replacing the NSS, to allow switching frequencies up to 37 Hz. The results of this experiment are shown in Fig. 1. After measuring both balanced and unbalanced cases, a 2.7-Hz switching frequency was selected as a good compromise between sufficiently reduced noise and a reasonable NSS life expectancy.

The noise power spectrum of the radiometer output was measured using the non-real time digital spectrum analyzer shown in Fig. 2. The radiometer was switched between a high-temperature and ambient load to produce a known output at the 2.7-Hz switching frequency. The output spectrum, given in Fig. 3, shows that the noise power spectral density at 2.7 Hz corresponds to 35°K.

These data show that low switching rates rapidly compromise radiometer performance, and that a detailed knowledge of the radiometer components' noise characteristics must be known to select an optimum switching frequency. Further study is required in this general area.

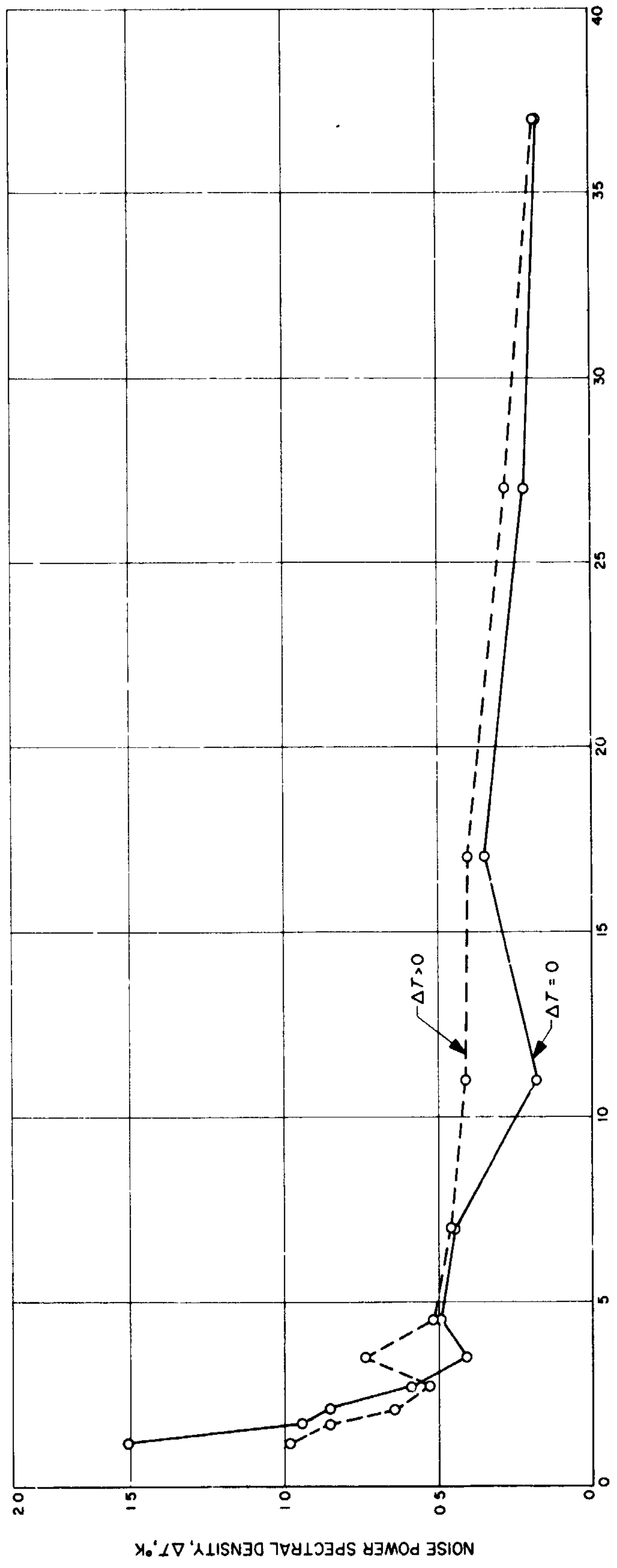


Fig. 1. Experimental results of radiometer reconfiguration

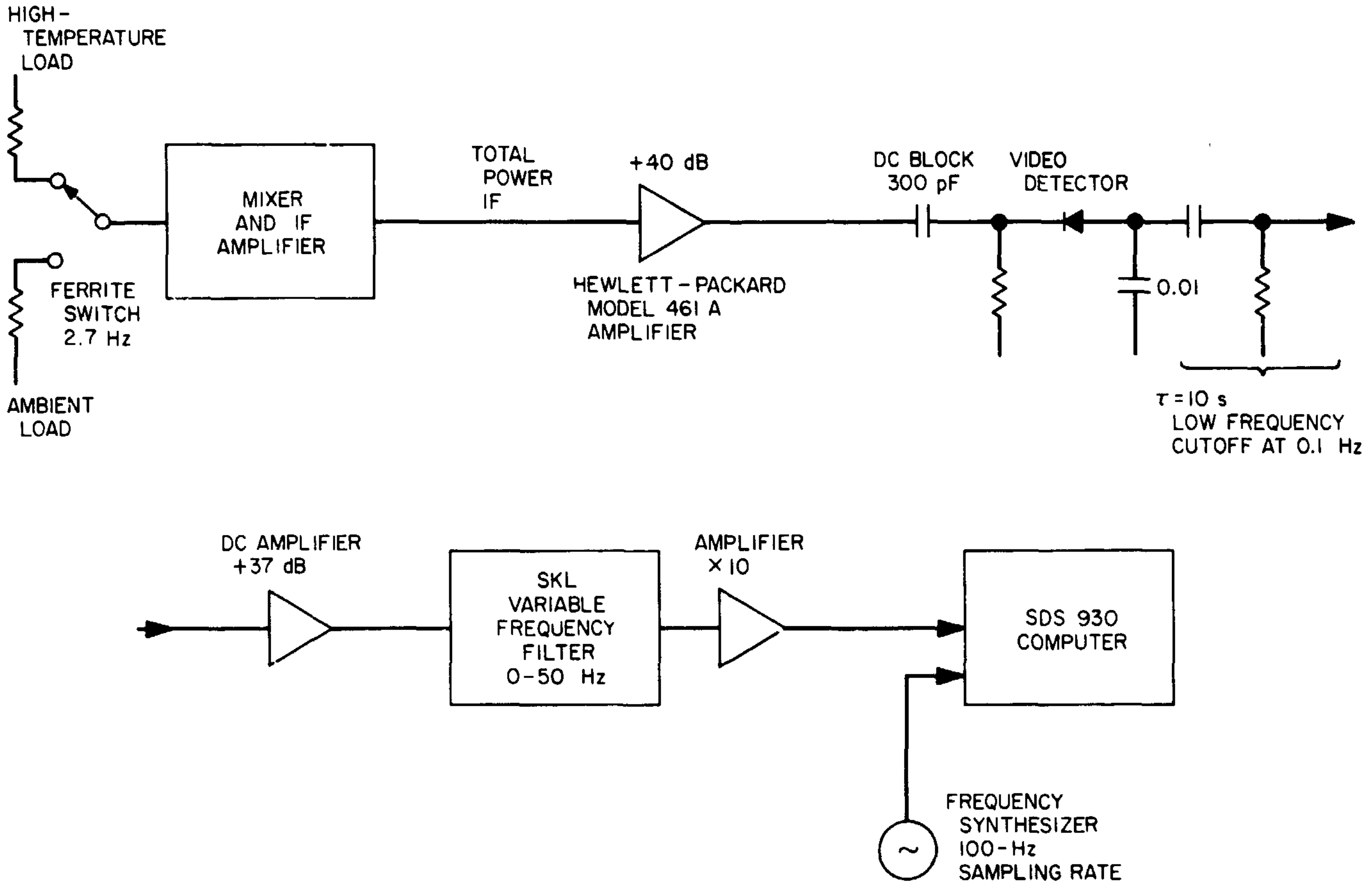


Fig. 2. Instrumentation for radiometer noise and gain-change measurements

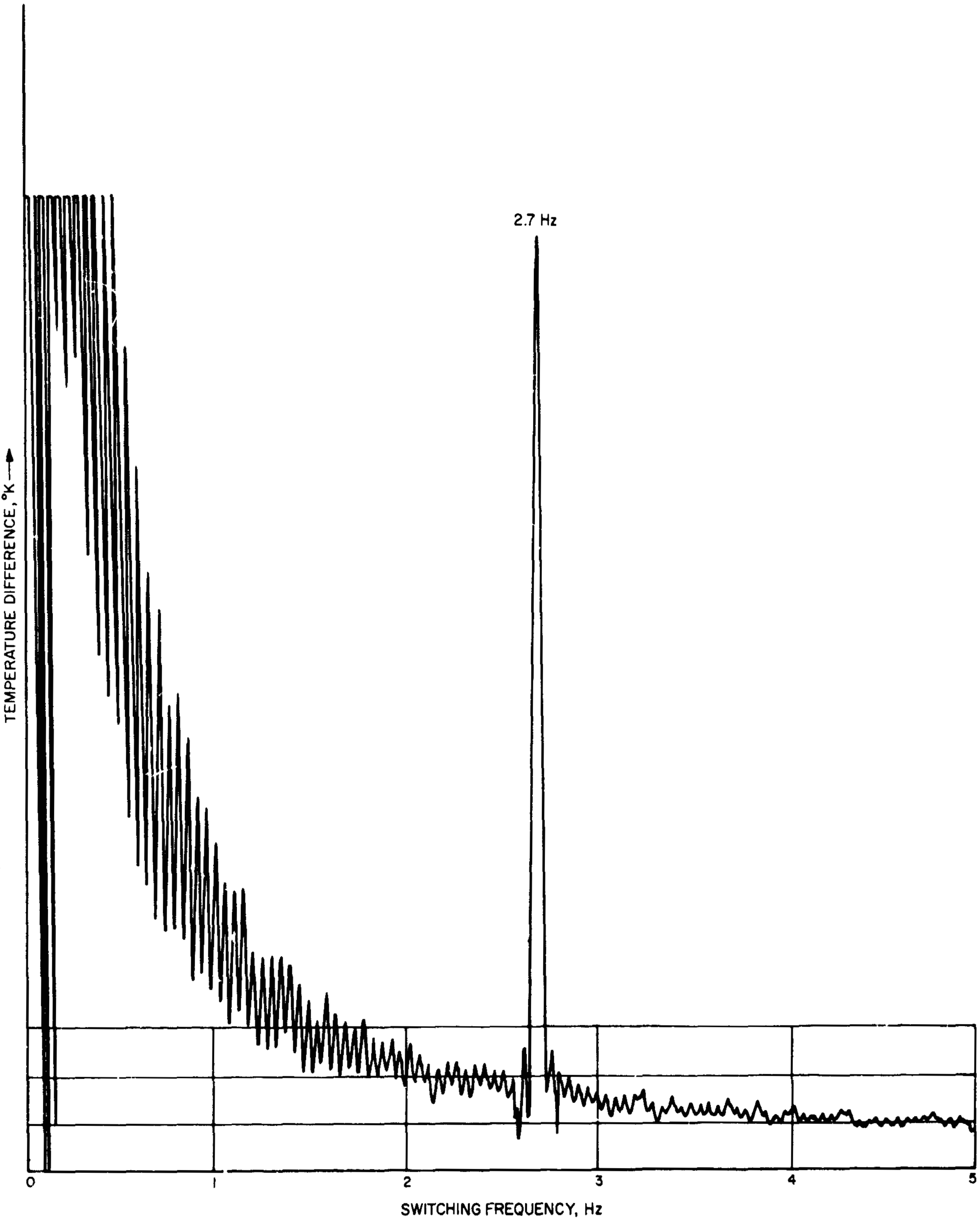


Fig. 3. Radiometer noise spectrum

B. Precision Calibration Techniques: Microwave Thermal Noise Standards, *C. Stelzried*

1. Introduction

Calibrated microwave thermal-noise standards (Ref. 1) are used for microwave radiometry, antenna temperature calibrations, loss measurements (SPS 37-41, Vol. III, p. 83), low-noise amplifier performance evaluation and low-level continuous-wave signal-level calibrations (Ref. 2). A typical thermal-noise standard consists of a matched resistive element thermally isolated by a uniform transmission line. The transmission line is usually fabricated from copper-plated stainless steel and has distributed temperatures and transmission loss factors. Although thermal-noise standards have been constructed without the use of transmission lines by pointing an antenna beam directly at bulk termination material (Ref. 3), the calibration cf these standards is complicated by the antenna characteristics (side lobes, etc.). The present discussion is limited to the use of a transmission line with matched termination.

Microwave thermal-noise standards are usually designated hot, ambient, or cold, depending upon whether the resistive element is above, at, or below ambient temperature. The construction and calibration techniques used in hot or cold loads are similar. The primary difference is the method used to obtain temperature equilibrium of the resistive element. Hot loads normally use electrical heaters or boiling liquids with a high boiling point (e.g., water), and cold loads normally use refrigeration or boiling liquids with a low boiling point (e.g., liquid helium). Ambient loads are the easiest to fabricate and calibrate, requiring only a matched termination with a suitable thermal heat sink and thermometer.

2. Theory

Nyquist's theorem (Ref. 4), including the zero-point energy (Ref. 5), states that the available termination noise power P is given by

$$P = \frac{1}{2} hfB + \frac{hfB}{\exp(hf/kT) - 1} \qquad (1)$$

where

T = termination temperature, °K

k = Boltzmann's constant, 1.38054×10^{-23} J-°K^{-1}

h = Planck's constant, 6.6256×10^{-34} J-s

B = bandwidth, Hz

f = frequency, Hz

Assuming $hf/kT \ll 1$,

$$P = kTB \qquad (2)$$

Consider a thermal-noise standard, as shown in Fig. 4, consisting of a termination at temperature T and a transmission line with distributed temperatures and propagation constants. The problem is to determine the noise power or noise temperature at the output reference point. Signify the propagating noise power, transmission line thermal temperature, and propagation constant at x by P_x, T_x, and $2\alpha_x$.

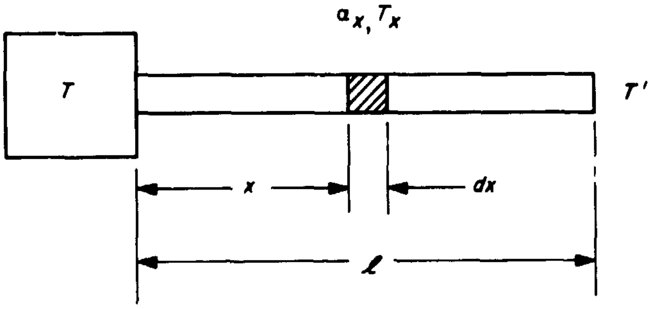


Fig. 4. Thermal noise sta.ıdard with loss and temperatur? of the transmission line as a function of position

The propagating noise power can be separated into two parts: (1) from the termination, attenuated by the transmission line, and (2) from the noise contribution of the lossy transmission line. The noise power at the reference output due to the termination is given by P/L (the termination noise power divided by the total line loss). Total line loss L is given by

$$L = \exp(2\alpha l) = \exp\left(\int_0^l 2\alpha_x dx\right) \qquad (3)$$

The noise power generated by a transmission line element of length dx is

$$kBT_x(1 - \exp(2\alpha_x dx)) \simeq kBT_x(2\alpha_x dx) \qquad (4)$$

The contribution at the reference output is given by dividing by the transmission line loss from x to the output reference

$$\frac{kBT_x 2\alpha_x dx}{\exp\left(\int_x^l 2\alpha_x dx\right)} = \left(\frac{2kB}{L}\right) T_x \alpha_x L_x dx \qquad (5)$$

where

$$L_x = \exp\left(\int_0^x 2\alpha_x dx\right)$$

is the loss from the source to x. The total noise power at the output reference is found by integrating and adding the contribution from the termination

$$P' = \frac{2kB}{L}\int_0^l \alpha_x L_x T_x dx + \frac{P}{L} \qquad (6)$$

Dividing by kB gives the noise temperature (Ref. 6)

$$T' = T'' + \frac{T}{L} \qquad (7)$$

where

$$T'' = \frac{2}{L}\int_0^l \alpha_x L_x T_x dx$$

is the contribution from the transmission line. If α_x and T_x are treated as constants α and T_P, then $L_x = \exp(2\alpha x)$ and

$$T' = \left(1 - \frac{1}{L}\right)T_P + \frac{T}{L} \qquad (8)$$

A useful expansion for small losses is given by

$$\frac{1}{L} = 1 - \mathcal{L} + \frac{1}{2}\mathcal{L}^2 + \cdots \qquad (9)$$

where

$$\mathcal{L} = 2\alpha l = \frac{L, \text{dB}}{10\log_{10}e} \simeq 0.23026\,L, \text{dB}$$

Then

$$T' = T + (T_P - T)\left(\mathcal{L} - \frac{1}{2}\mathcal{L}^2 + \cdots\right) \qquad (10)$$

Other solutions are presented in Table 1 for various combinations of transmission line temperature and propagation constant distributions.

3. Calibration Errors

The most critical measurement in the calibration of the noise temperature of a thermal-noise standard is usually the transmission line loss. For example, if the loss and temperature distributions are constant, the error in T' due to loss measurement errors is [assuming a small loss and differentiating Eq. (10)]

$$\Delta T' \simeq 0.23026\,(T_P - T)\,\Delta L, \text{dB} \qquad (11)$$

To determine T' to an accuracy of better than 0.1°K for a liquid helium cooled termination requires better than a 0.002-dB measurement accuracy.

The contribution of an ambient temperature transmission line with a 0°K termination is approximately (0.23026 $T_P\Delta L$, dB), or 6.7°K/0.1 dB. As seen from Eq. (11), the transmission line loss has no net effect with an ambient termination (assuming the transmission line and termination are at the same temperature T_P).

For precision measurements, it is necessary to account for the pressure inside the dewar with cryogenically cooled terminations. In this case, replace the termination temperature T with

$$T_c + C\Delta P \qquad (12)$$

where

T_c = cryogenic liquid boiling temperature at standard pressure °K (approximately 77.36°K for liquid nitrogen and 4.216°K for liquid helium)

C = cryogenic liquid pressure constant, °K/torr (approximately 0.010987°K/torr for liquid nitrogen and 0.001352°K/torr for liquid helium)

ΔP = barometric pressure greater than standard, (760 torr)

In cryogenically cooled terminations, it is necessary to maintain the termination material in temperature equilibrium with the boiling liquid (unless the termination material temperature is determined by means other than the boiling temperature of the liquid). This can be accomplished by submerging the termination material in the liquid, or by providing a very low thermal heat path to the liquid relative to the thermal heat path to the outside environment.

Table 1. Tabulated solutions of the theoretical noise temperature at the output of a transmission line with various temperature and loss distributions[a]

Transmission-line parameters	Solved parameter	Exact solution	Approximate solution (assuming $\mathcal{L} \ll 1$)
Constant temperature distribution T_P and constant propagation loss distribution α	T''	$\left(1 - \frac{1}{L}\right) T_P$	$T_P\left(\mathcal{L} - \frac{1}{2}\mathcal{L}^2 + \cdots\right)$
	T'	$\left(1 - \frac{1}{L}\right) T_P + \frac{T}{L}$	$T + (T_P - T)\left(\mathcal{L} - \frac{1}{2}\mathcal{L}^2 + \cdots\right)$
	$(T_P - T')$	$\frac{T_P - T}{L}$	$(T_P - T)\left(1 - \mathcal{L} + \frac{1}{2}\mathcal{L}^2 + \cdots\right)$
Constant temperature distribution T_P and linear propagation loss distribution from α_1 to α_2 $\alpha = (\alpha_1 + \alpha_2)/2$	T''	$\left(1 - \frac{1}{L}\right) T_P$	$T_P\left(\mathcal{L} - \frac{1}{2}\mathcal{L}^2 + \cdots\right)$
	T'	$\left(1 - \frac{1}{L}\right) T_P + \frac{T}{L}$	$T + (T_P - T)\left(\mathcal{L} - \frac{1}{2}\mathcal{L}^2 + \cdots\right)$
	$(T_P - T')$	$\frac{1_P - T}{L}$	$(T_P - T)\left(1 - \mathcal{L} + \frac{1}{2}\mathcal{L}^2 + \cdots\right)$
Linear temperature distribution from T to T_P and constant propagation loss distribution α	T''	$\left(1 - \frac{1 - \frac{1}{L}}{\mathcal{L}}\right) T_P - \left(\frac{1}{L} - \frac{1 - \frac{1}{L}}{\mathcal{L}}\right) T$	$\frac{1}{2} T_P \left(\mathcal{L} - \frac{1}{3}\mathcal{L}^2 + \cdots\right)$
	T'	$\left(1 - \frac{1 - \frac{1}{L}}{\mathcal{L}}\right) T_P + \left(\frac{1 - \frac{1}{L}}{\mathcal{L}}\right) T$	$T + \frac{1}{2}(T_P - T)\left(\mathcal{L} - \frac{1}{3}\mathcal{L}^2 + \cdots\right)$
	$(T_P - T')$	$\frac{(T_P - T)\left(1 - \frac{1}{L}\right)}{\mathcal{L}}$	$(T_P - T)\left(1 - \frac{1}{2}\mathcal{L} + \frac{1}{6}\mathcal{L}^2 + \cdots\right)$
Linear temperature distribution from T_1 to T_2 and constant propagation loss distribution α	T'	$\left(1 - \frac{1 - \frac{1}{L}}{\mathcal{L}}\right) T_2 - \left(\frac{1}{L} - \frac{1 - \frac{1}{L}}{\mathcal{L}}\right) T_1 + \frac{T}{L}$	$T + (\mathbf{T} - T)\mathcal{L} - \frac{1}{6}(T_2 + 2T_1 - 3T)\mathcal{L}^2 + \cdots$
Linear temperature distribution from T_1 to T_2 and linear propagation loss distribution from α_1 to α_2 $\alpha = (\alpha_1 + \alpha_2)/2$	T''		$\mathbf{T}\mathcal{L} - \frac{1}{6}(T_2 + 2T_1)\mathcal{L}^2 +$
	T'	$T_2 - \frac{T_1}{L} - \frac{2(T_2 - T_1)}{\sqrt{2(\mathcal{L}_2 - \mathcal{L}_1)}}\left[D\left(\frac{\mathcal{L}_2}{\sqrt{2(\mathcal{L}_2 - \mathcal{L}_1)}}\right) - \frac{1}{L}D\left(\frac{\mathcal{L}_1}{\sqrt{2(\mathcal{L}_2 - \mathcal{L}_1)}}\right)\right] + \frac{T}{L}$ where $D(x)$ is Dawson's integral of x	$T + (\mathbf{T} - T)\mathcal{L} - \frac{1}{6}(T_2 + 2T_1 - 3T)\mathcal{L}^2 + \cdots$

[a] $\mathcal{L}_i = 2\alpha_i l = L_i$, dB/10 $\log_{10} e$
$\mathcal{L} = (\mathcal{L}_1 + \mathcal{L}_2)/2$
$\mathbf{T} = (T_1 + T_2)/2$
$T = \exp(2\alpha l)$

The magnitude of the error made by assuming $hf/kT \ll 1$ in Eq. (2) can be estimated by considering the higher order terms. In this case (valid for $hf/kT < 4\pi^2$), we have, from Eq. (1)

$$P = \frac{1}{2} hfB + kTB \left[1 - \frac{hf}{2kT} + \frac{1}{12} \frac{hf^2}{kT} + \cdots \right] \tag{13}$$

The correction term $(hf/kT)^2/12$ contributes less than 2% error at an operating frequency of 10 GHz for T greater than 1°K. Some authors have expressed doubt concerning the inclusion of the zero-point energy term (Ref. 7). It should be noted that Eq. (13) reduces to the same correction term if the zero-point energy is neglected when the calibrated terminations are used to perform temperature-difference calibrations in an actual radiometer.

Other sources of error include the inaccuracies in the temperature and loss distribution calibrations, nonhomogeneous transmission line effects, and microwave mismatches.

References

1. Stelzried, C. T., "A Liquid Helium-Cooled Coaxial Termination," *Proc. IRE*, Vol. 49, No. 7, p. 1224, July 1961.
2. Stelzried, C. T., and Reid, M. S., "Precision Power Measurements of Spacecraft CW Signal Level with Microwave Noise Standards," *IEEE Trans. Inst. Meas.*, Vol. IM-15, No. 4, p. 318, Dec. 1966.
3. Singer, A., Ulrich, R. R., and Naess, E., *Thermal Calibrators in Millimeter-Wave Radiometry*, TM-67-2. Harry Diamond Laboratories, Washington, D.C., Mar. 1967.
4. Nyquist, H., "Thermal Agitation of Electrical Charge in Conductors," *Phys. Rev.*, Vol. 32, p. 110, July 1928.
5. Siegman, A. E., "Zero-Point Energy as the Source of Amplifier Noise," *Proc. IRE*, p. 633, Mar. 1961.
6. *IRE Standards on Electron Tubes: Definitions of Terms, 1962* (62 IRE 7.S2), *Proc. IEEE*, p. 434, Mar. 1963.
7. MacDonald, D. K. C., *Noise and Fluctuations: An Introduction*, p. 37. John Wiley & Sons, New York, 1962.

C. RF Breakdown Studies: RF Breakdown in Coaxial Transmission Lines, *R. Woo*

1. Introduction

A scheme for presenting breakdown data was discussed in SPS 37-45, Vol. IV, pp. 323–330 and SPS 37-46, Vol. IV, pp. 259–263. A series of breakdown experiments have been conducted for the 50-Ω coaxial transmission line configuration in frequency range of 4–800 MHz. These measurements yielded breakdown data for fd values of 20–600 MHz–cm.

2. Results

The breakdown data obtained are shown in Fig. 5. Two experimental setups were used: (1) 10–150 MHz lumped-circuit test set (Ref. 1), and (2) 150–800 MHz transmission line test set (Ref. 2). The data are plotted in terms of similarity parameters and, as can be seen, the scaling correspondence between data obtained from both test sets is remarkably good (within reproducibility of the data). It must be pointed out that the transmission line test-set frequency in one case is as high as seven times that of the lumped-circuit test set. There is a spread in the results for $fd = 100$ MHz–cm (Fig. 5b) at the lower values of pd. This is not surprising since, as will be discussed below, this corresponds to a region of several transitions, and breakdown conditions are somewhat dependent on surface conditions.

3. Discussion

The data of Fig. 5 can be combined with that obtained previously to form the composite breakdown plot shown in Fig. 6. S. C. Brown and A. D. MacDonald (Ref. 3) showed that breakdown data can be represented by a three-dimensional surface using similarity parameters. Figure 6 defines this three-dimensional surface with breakdown power as the vertical axis and fd and $p\lambda$ as the horizontal axes (see Fig. 7). The similarity parameters of Fig. 6 are, however, more useful to a design engineer than those used by Brown and MacDonald. For a given coaxial line operating at a particular frequency, the engineer computes the corresponding fd, and, by referring to Fig. 6, he has the breakdown behavior as a function of pressure. In addition, he has information on the effects of changing either frequency or line size.

The fd vs $p\lambda$ plane shown in Fig. 8 is very useful in understanding the breakdown processes involved. Although the various limits are indicated in the form of lines, it should be pointed out that these are meant to indicate transition rather than abrupt change. The mean free path limit serves to separate ionization breakdown from multipacting breakdown. The term "ionization breakdown" encompasses all breakdown processes where the dominant electron production mechanism is ionization by electron collision. This type of breakdown occurs when $p\lambda$ is greater than the mean free-path limit because, under these conditions, the electron mean free

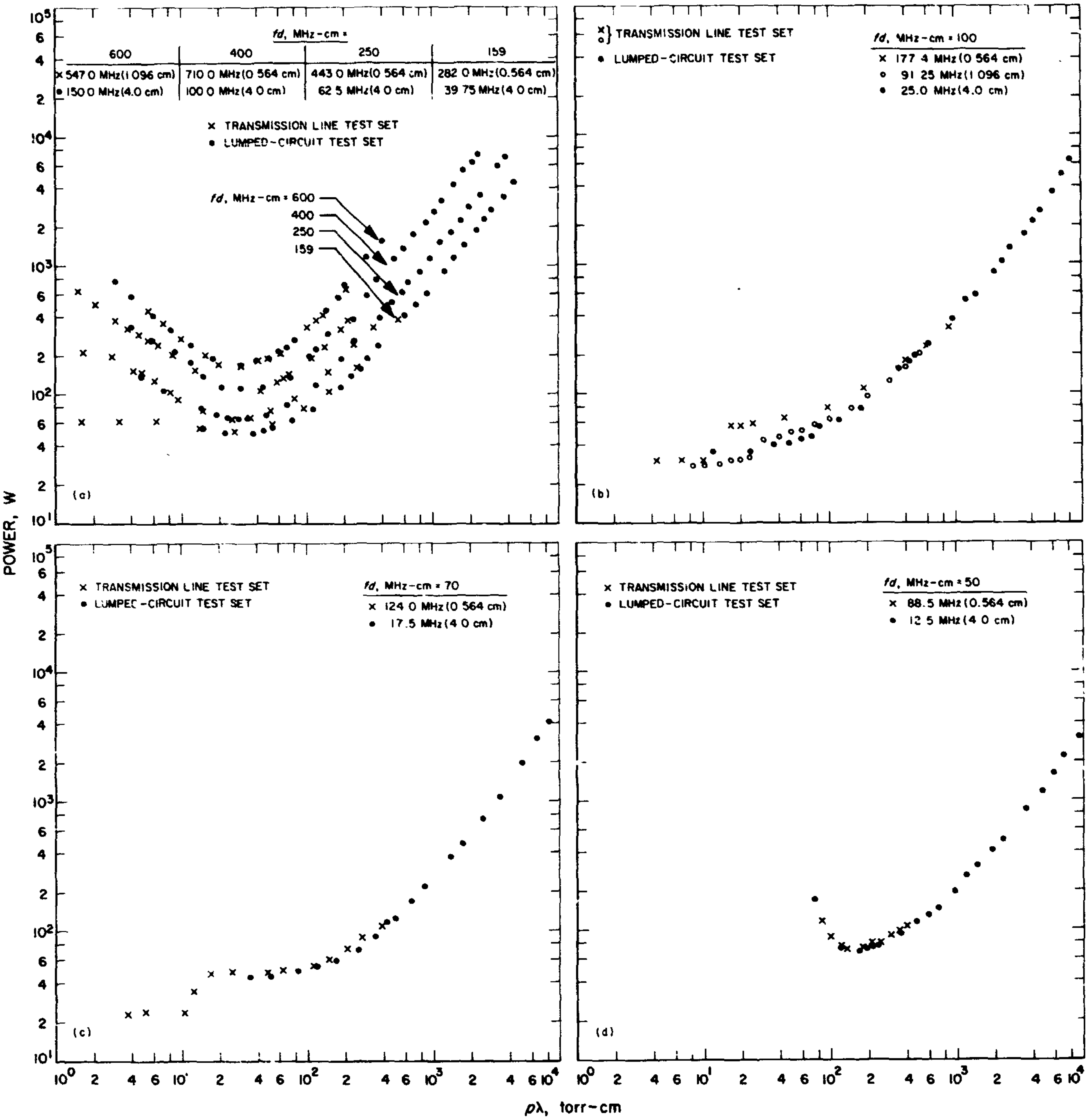


Fig. 5. RF breakdown data plotted in terms of similarity parameters

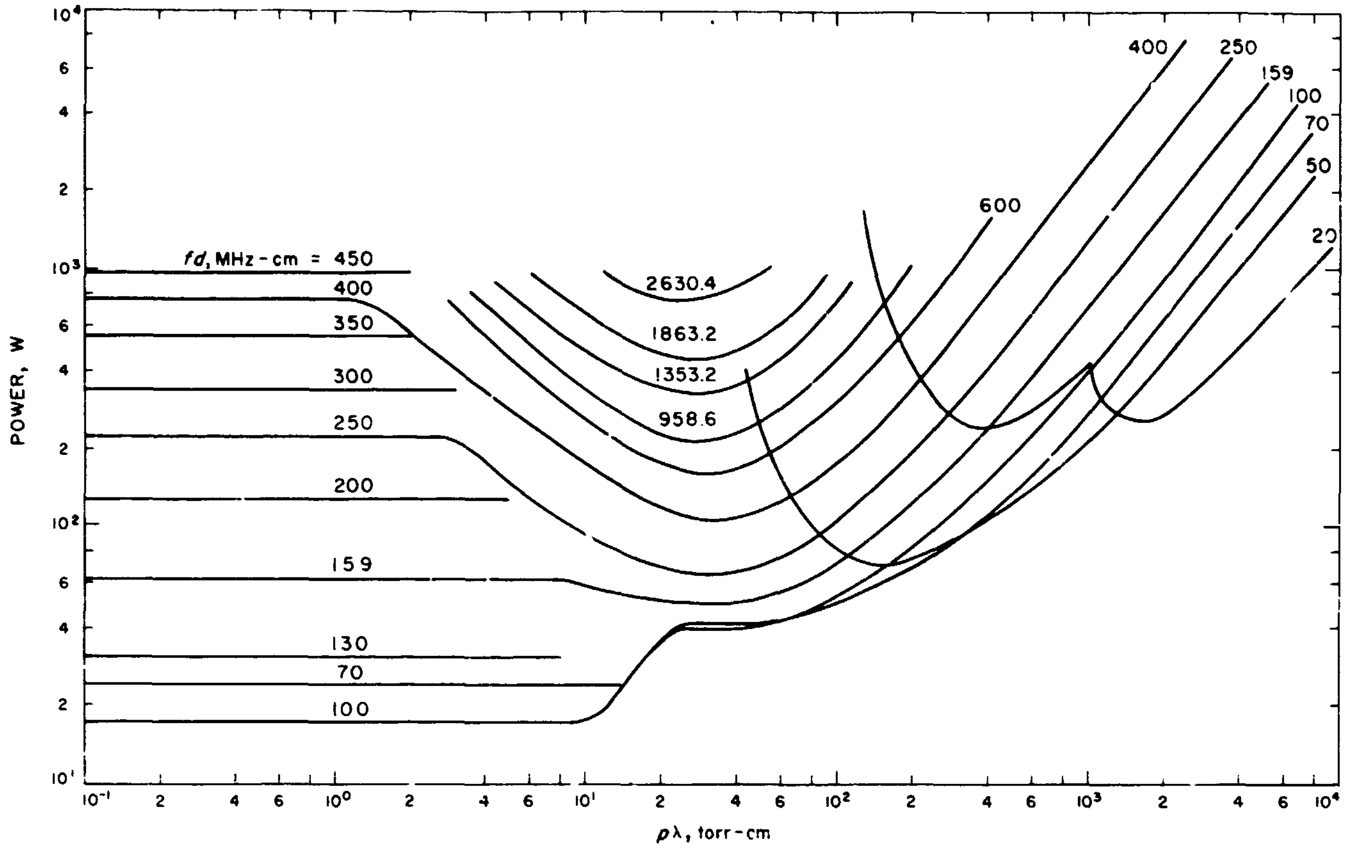


Fig. 6. Unified plot for RF breakdown in 50-Ω coaxial transmission line

path is shorter than the gap distance. When discussing ionization breakdown, it is convenient to think of it in terms of the two ranges of fd presented below.

a. $fd > 100$ MHz–cm. Under these conditions, frequency is sufficiently high and the gap distance sufficiently large that the electrons are not swept out of the discharge region by the field as in the case of dc breakdown. Instead, the electrons are concentrated in the center of the discharge region and slowly diffuse away towards the electrodes. The speeds are so low that the electrons produce, essentially, no secondary effects at the electrode surfaces. Breakdown of this type is termed diffusion-controlled or microwave breakdown (Ref. 4). This, in many ways, is the simplest high-frequency breakdown since only two main processes are involved; electrons are produced through ionization by electron collision and are removed by diffusion to the walls. In certain gases, electrons are also effectively lost by attachment to gas molecules.

The minimum of the diffusion-controlled curves occurs at approximately $p\lambda = 30$ torr–cm. The $p\lambda = 30$ torr–cm line is called the collision frequency transition. At the collision frequency transition, the applied frequency and the electron–molecule collision frequency are approximately equal, and energy transfer to the electrons from the field is at a maximum. If pressure is increased, the electron–molecule collision frequency increases, the energy gained by electrons from the field per mean free path decreases, and the breakdown level correspondingly increases. In a perfect vacuum, the electrons oscillate with their velocity 90 deg out of phase with the RF field, and no energy is gained by the electrons from the field. The electrons gain energy from the field only by undergoing collisions with the gas molecules. A decrease in pressure from the collision frequency transition corresponds to an increase in loss of energy transfer from the field to the electrons. Breakdown power, therefore, rises with decreasing pressure.

b. $fd < 100$ MHz–cm. When the applied frequency is sufficiently low or the gap distance sufficiently short, the amplitude of oscillation of the electron cloud approaches the gap distance and the electrodes enter the breakdown picture. This situation occurs when fd is less than

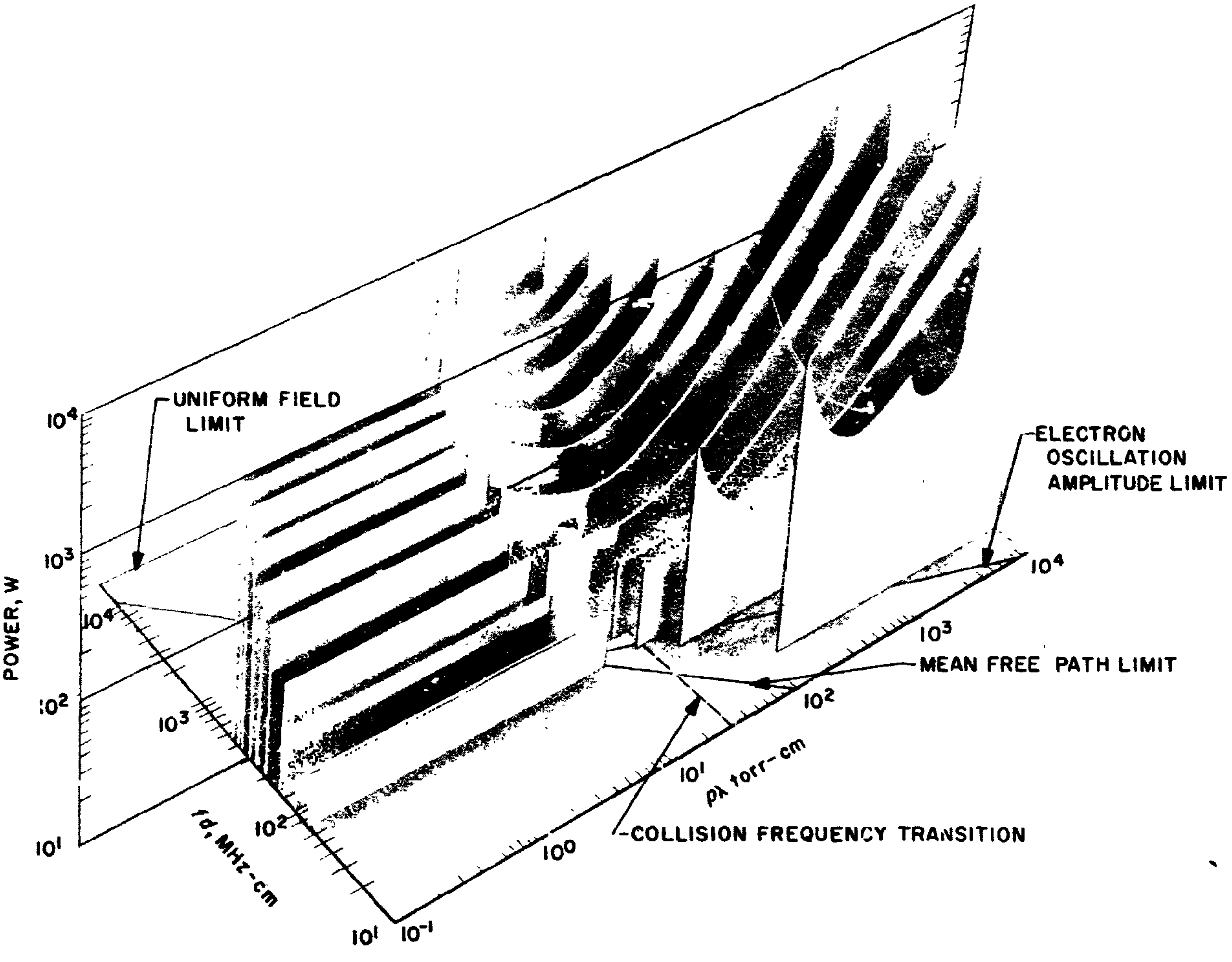


Fig. 7. Three-dimensional surface representing RF breakdown in 50-Ω coaxial transmission line

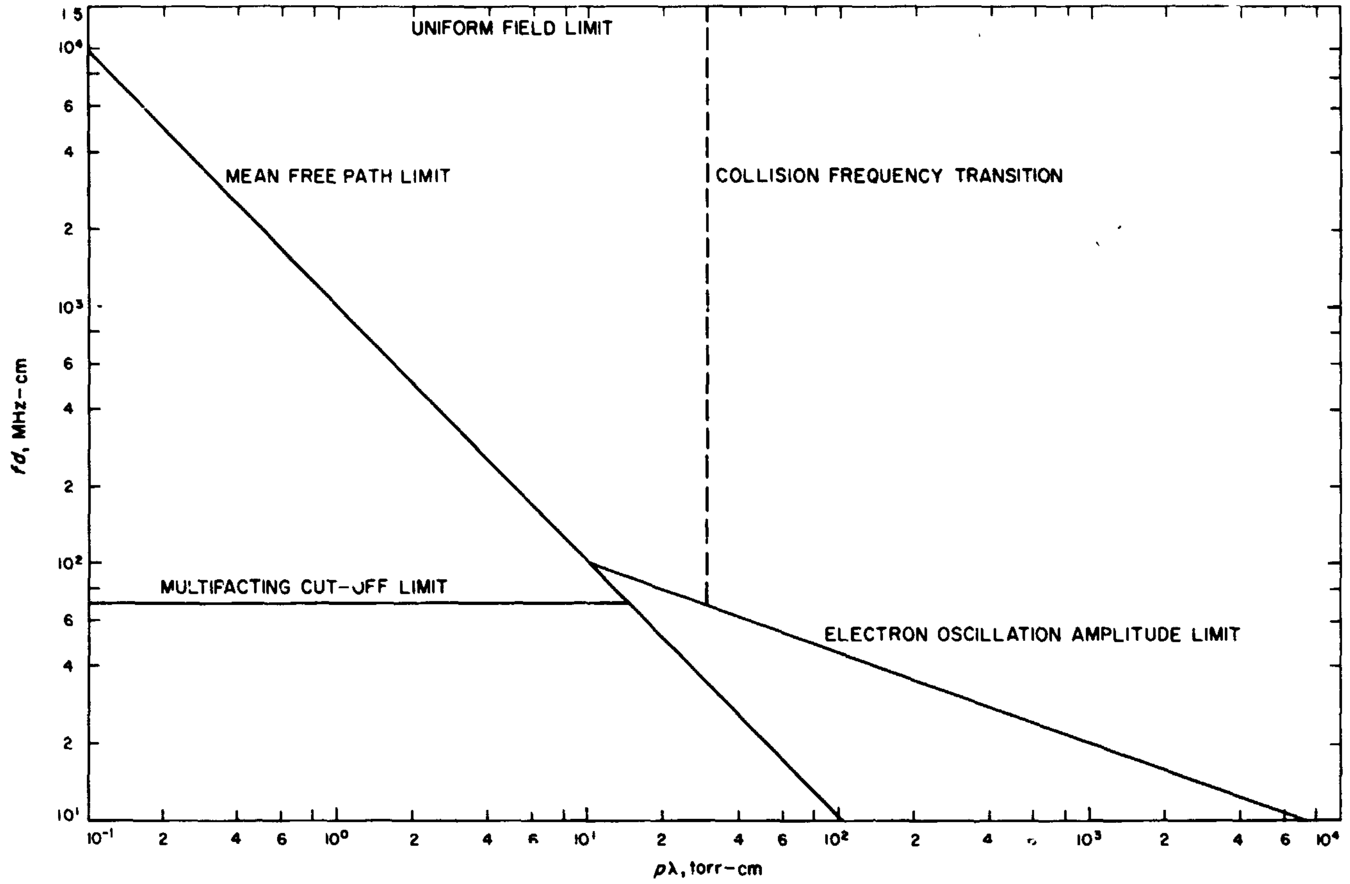


Fig. 8. $p\lambda$–fd plane showing limits of breakdown processes

100 MHz–cm. Under such conditions, the loss of electrons is governed by mobility. Brown (Ref. 5) has termed this type of breakdown mobility-controlled breakdown. It must be emphasized that the transition from diffusion-controlled to mobility-controlled breakdown is gradual and occurs at approximately 100 MHz–cm. The oscillation amplitude limit corresponds to the condition for which the amplitude of oscillation of the electron cloud is equal to the gap distance. At this limit, electrons are lost to the electrodes and the power required for breakdown rises rapidly. This behavior is illustrated in the data for $fd = 50$ and 20 MHz–cm in Fig. 6. In the case of $fd = 20$ MHz–cm, another minimum is observed if pressure is further decreased. This additional minimum appears when fd $\lesssim$ 20 MHz–cm. This region has been studied extensively by Gill and von Engel (Ref. 6) who attribute the additional minimum to the ions. At this additional minimum, the amplitude of oscillation of the ion cloud is equal to the gap distance, and the ions impinging on the electrodes release secondary electrons. Electrons are, therefore, produced by ion bombardment of the electrodes.

When $p\lambda$ is less than the mean free path limit, the electron mean free path is longer than the gap distance and secondary electron emission is the electron production mechanism. Under these conditions, secondary electron resonance or multipacting breakdown occurs. Although multipacting has been adequately covered elsewhere (Refs. 1, 2, 5, and 7), the following are points worth mentioning in connection with the multipacting data of Fig. 6:

(1) The multipacting data of Fig. 6 corresponds to the lower breakdown boundary. The upper boundary, above which multipacting will not occur, is not shown in Fig. 6.

(2) For fd less than the multipacting cut-off limit of $fd \sim 70$ MHz–cm, multipacting will not occur.

(3) Multipacting is independent of pressure.

(4) Multipacting breakdown power levels are very sensitive to surface and outgassing conditions. In general, this is not the case for ionization breakdown.

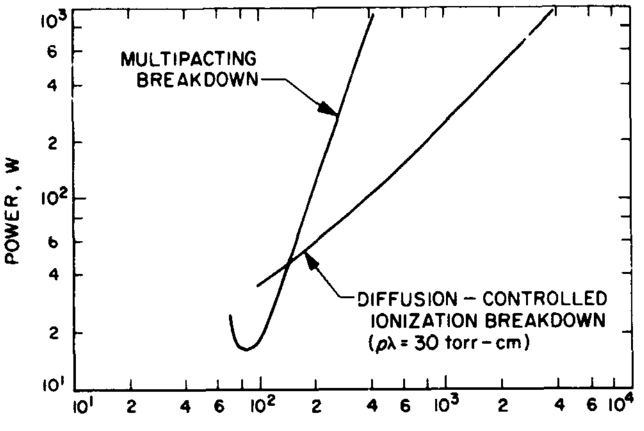


Fig. 9. Power handling capability in terms of *fd*

(5) As fd is increased, breakdown power levels rise more rapidly for multipacting than for ionization breakdown (see Fig. 9). Therefore, for a fixed power level, ionization breakdown will cover a wider range of experimental variables than multipacting.

From the above discussion, the reason is clear for the spread in the data of Fig. 5b at the low values of $p\lambda$. A transitional region between diffusion-controlled and mobility-controlled breakdown is represented by $fd =$ 100 MHz–cm. At approximately $p\lambda = 10$ torr–cm, there is also a transition between multipacting and ionization breakdown. The $fd = 100$ MHz–cm corresponds to the minimum energy boundary in the case of multipacting, and breakdown data is especially sensitive to surface conditions.

Engineers are, in general, interested in the minimum power-handling capability of a given component. The breakdown power levels along the collision frequency transition are shown in Fig. 9 as a function of fd, thus giving tne minima of the diffusion-controlled breakdown curves. The multipacting breakdown data are also included for comparison. As can be seen, for $fd > 145$ MHz–cm, the ionization breakdown level is lower than the multipacting breakdown level, while the reverse is true for $fd < 145$ MHz–cm.

4. Concluding Remarks

Breakdown data obtained for the 50-Ω coaxial transmission line are summarized in Fig. 6, which is concise and compact and should prove to be valuable to the design engineer. When using Fig. 6, the design engineer should be aware of the various breakdown processes involved and, consequently, the accuracy to be expected from these curves. It must be remembered that the data in Fig. 6 were obtained through carefully controlled experimental conditions. When testing a component for breakdown, the engineer must assure himself that he is measuring the pressure level in the area where breakdown occurs. The materials of the component should have a low outgassing rate and be relatively clean. The breakdown procedures should be similar to the ones used in obtaining the data of Fig. 6. Figure 2 gives the breakdown power levels for air. Ionization breakdown is dependent on the type of gas while multipacting is not. The power levels of Fig. 6 correspond to a perfectly matched transmission line. If mismatches exist in the line, the breakdown power level must be correspondingly derated.

Finally, the scheme of data presentation of Fig. 6 can be used for configurations other than the 50-Ω coaxial transmission line. Similar curves can also be obtained for gases other than air.

References

1. Woo, R., "Multipacting Discharges Between Coaxial Electrodes," *J. Appl. Phys.*, Vol. 39, pp. 1528–1533, 1968.
2. Woo, R., "Multipacting Breakdown in Coaxial Transmission Lines," *Proc. IEEE* (Letters), Vol. 56, pp. 776–777, 1968.
3. MacDonald, A. D., and Brown, S. C., "Limits for the Diffusion Theory of High Frequency Gas Discharge Breakdown," *Phys. Rev.*, Vol. 76, pp. 1629–1633, 1949.
4. MacDonald, A. D., *Microwave Breakdown in Gases*. John Wiley & Sons, Inc., New York, 1966.
5. Brown, S. C., *Handbuch der Physik*, Vol. 22, pp. 531–575. Edited by S. Flugge. Springer-Verlag, Berlin, 1956.
6. Gill, E. W. B., and von Engel, A., "Starting Potentials of Electrodeless Discharges," *Proc. Roy. Soc. London*, Ser. A197, pp. 107–124, 1949.
7. Woo, R., and Ishimaru, A., "A Similarity Principle for Multipacting Discharges," *J. Appl. Phys.*, Vol. 38, pp. 5240–5244, 1967.

D. Spacecraft Antenna Research: 400-MHz Coaxial Cavity Radiator, Part II, *K. Woo*

1. Introduction

The power handling capability of the 400-MHz coaxial cavity radiator (SPS 37-48, Vol. III, pp. 238–240) at very low pressures has been determined. The ionization breakdown of the antenna occurs at as low as 76 W in air and 62 W in 100% CO_2. The multipacting breakdown was not observed up to an input power level of 100 W (operating limit of the feeding hybrid).

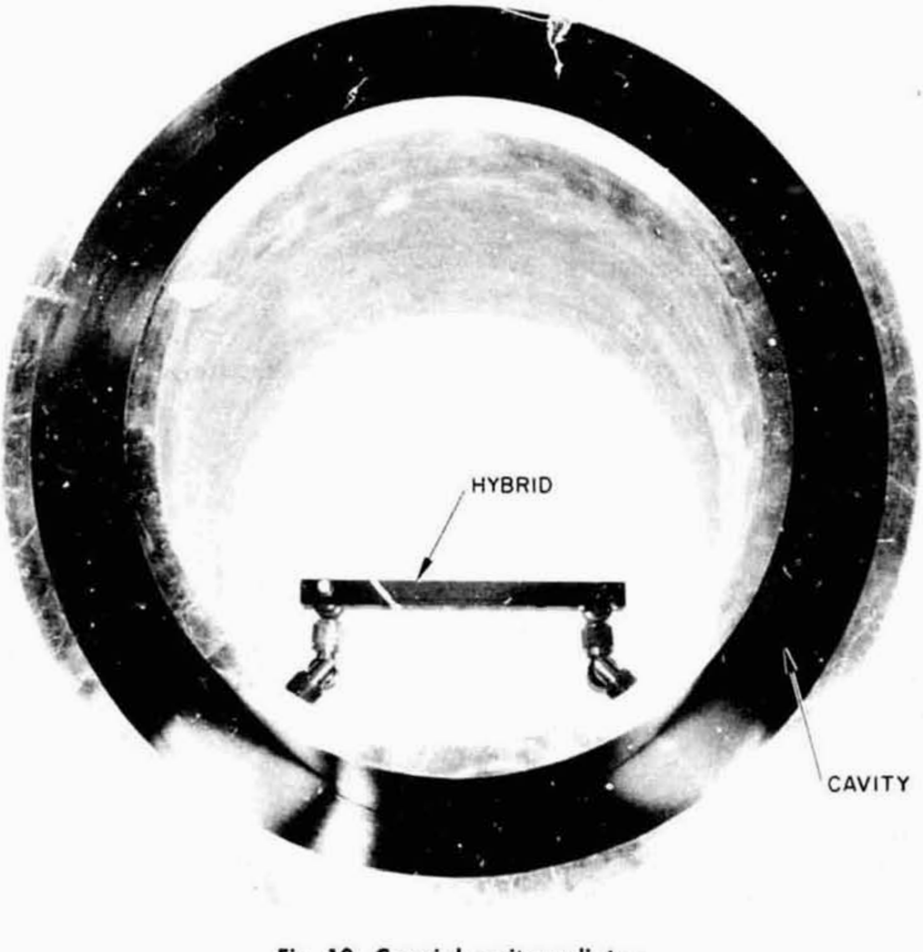


Fig. 10. Coaxial cavity radiator

2. Antenna Design

The design of the antenna is shown in Figs. 10 and 11. The coaxial cavity is excited by two orthogonal probes. The input feeds of the probes are connected to the two output terminals (having a 90-deg phase difference) of a 3-dB hybrid fed by the incoming line. For the purpose of preventing breakdown in the input feeds of the probes, and between the cavity walls and the probes, teflon insulators are used to fill up each input feed (between outer and center conductors) and they extend out into the cavity to wrap completely around each probe (see Fig. 11). With this arrangement, the voltage standing-wave ratio looking into each input feed with the other terminated is 1.25. When energized, the antenna radiates circularly polarized waves.

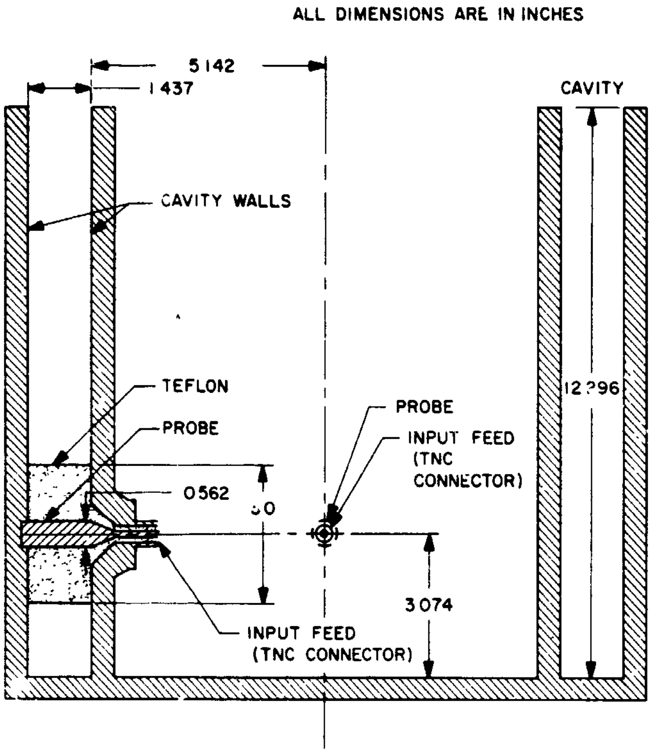


Fig. 11. Cavity and feed configuration

3. Test Results

The power handling capability of the antenna was determined at the JPL Voltage Breakdown Facility. The antenna was tested in the vacuum chamber first with air, and then with 100% CO_2. The ionization breakdown power level of the antenna is shown in Fig. 12 as a function of pressure near and at where the power-handling capability of the antenna is least. The ionization breakdown of the antenna occurs at as low as 76 W (at 0.28 torr) in air, and 62 W (at 0.25 torr) in 100% CO_2. In both cases, the breakdown took place at the aperture of the antenna. The multipacting breakdown (tested at 10^{-5} torr) was not observed up to an input power level of 100 W (operating limit of the feeding hybrid).

To increase the antenna power-handling capability, the following modifications are being implemented:

(1) The aperture of the existing antenna is being flared.

(2) A new cavity having a wider slot width is being fabricated.

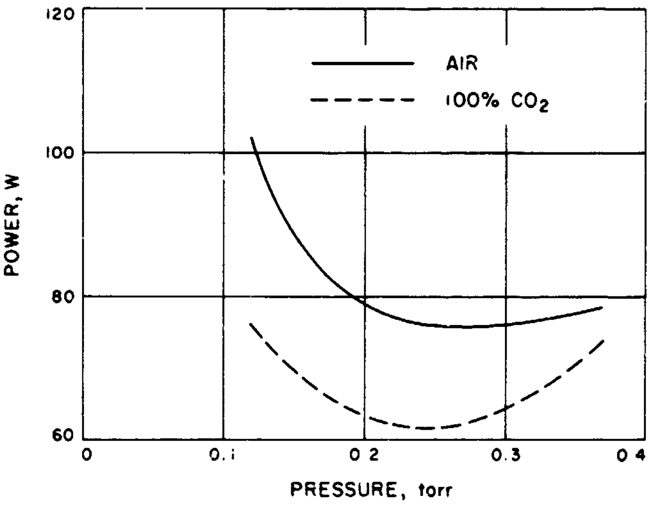


Fig. 12. Ionization breakdown characteristics

XXII. Spacecraft Telemetry and Command

TELECOMMUNICATIONS DIVISION

A. Multiple-Mission Telemetry System: Bit-Sync Lock Detector Evaluation, *N. Burow and A. Vaisnys*

The multiple-mission telemetry system (MMTS) bit tracking and detection functions are accomplished by means of a mission-dependent program in the TCP computer. In the original demonstration and *Mariner* Mars 1969 versions of this program, an estimate of the ratio of energy per bit to noise spectral density (ST/N_0) is used as an in-lock indicator. The threshold value of ST/N_0 is entered via typewriter and is a function of the expected ST/N_0.

A preliminary analysis of ST/N_0 estimation in the bit-sync loop was presented by Dr. J. Layland in SPS 37-48, Vol. III, pp. 209–212. Additional analytical work, and suggested ST/N_0 thresholds, are presented in *Chapter XX-G* of this volume. This article describes the approach used in evaluating the ST/N_0 estimator as a lock detector.

The overall test configuration is shown in Fig. 1. The MMTS demonstration bit-sync program was modified to output ST/N_0 samples on magnetic tape in groups of 1000. Measurements were made for input ST/N_0 of 0,

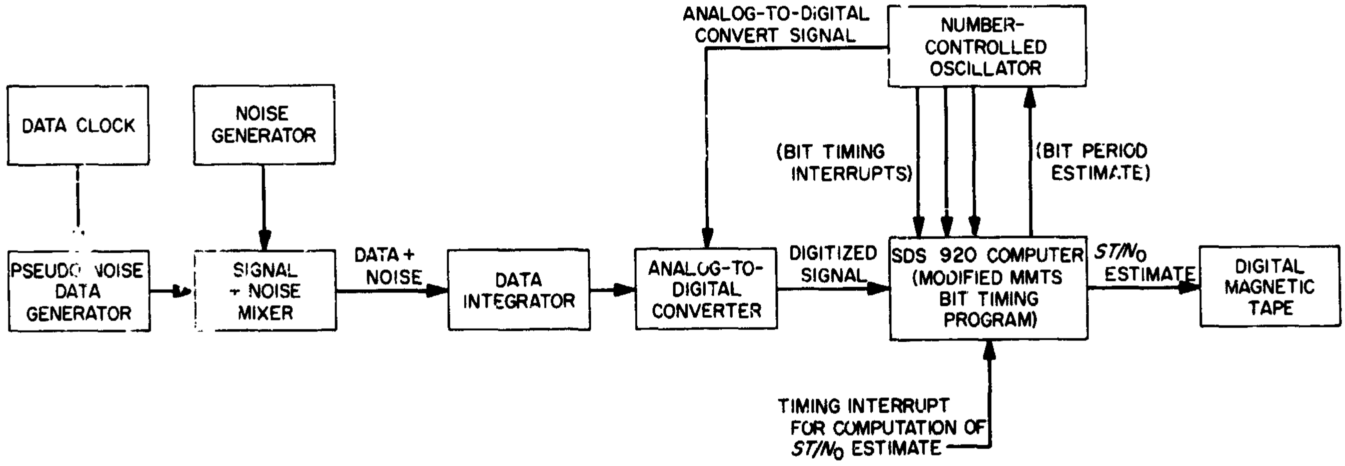
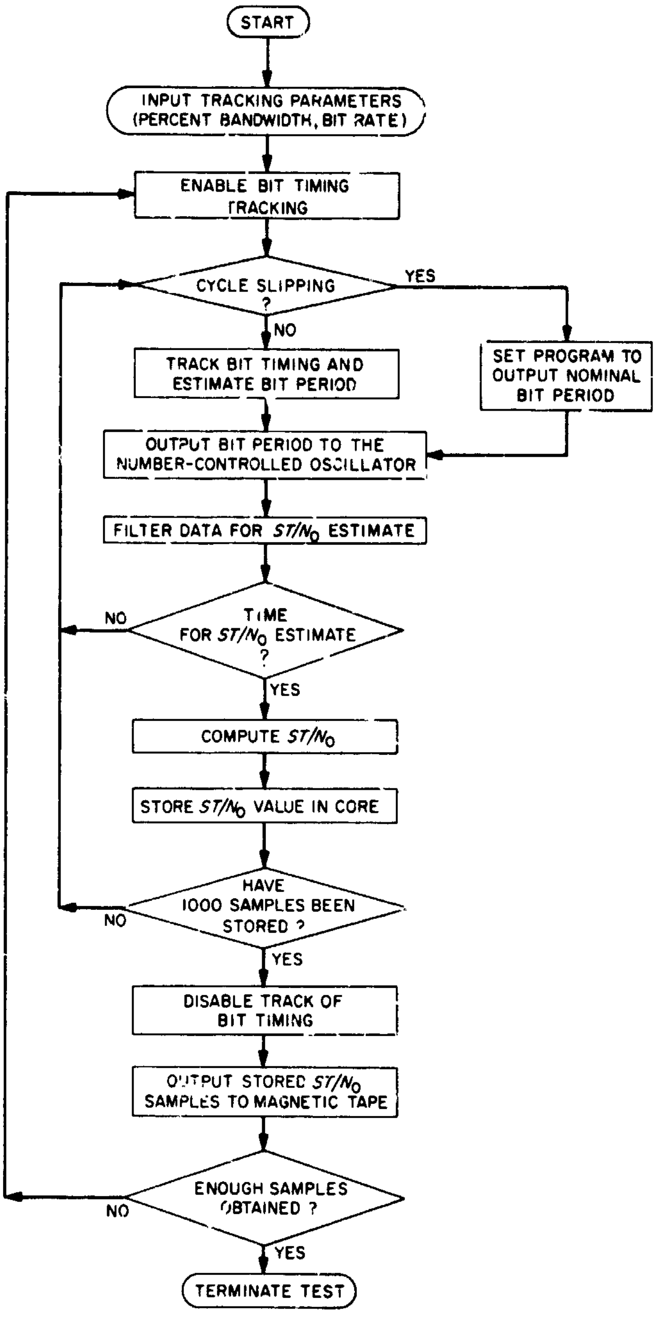

ANALOG-TO-DIGITAL CONVERT SIGNAL
NUMBER-CONTROLLED OSCILLATOR
DATA CLOCK
NOISE GENERATOR
(BIT TIMING INTERRUPTS)
(BIT PERIOD ESTIMATE)
PSEUDO NOISE DATA GENERATOR
SIGNAL + NOISE MIXER
DATA + NOISE
DATA INTEGRATOR
ANALOG-TO-DIGITAL CONVERTER
DIGITIZED SIGNAL
SDS 920 COMPUTER (MODIFIED MMTS BIT TIMING PROGRAM)
ST/N0 ESTIMATE
DIGITAL MAGNETIC TAPE
TIMING INTERRUPT FOR COMPUTATION OF ST/N0 ESTIMATE

Fig. 1. ST/N_0 lock detector evaluation test configuration



START
INPUT TRACKING PARAMETERS (PERCENT BANDWIDTH, BIT RATE)
ENABLE BIT TIMING TRACKING
CYCLE SLIPPING ?
YES
NO
TRACK BIT TIMING AND ESTIMATE BIT PERIOD
SET PROGRAM TO OUTPUT NOMINAL BIT PERIOD
OUTPUT BIT PERIOD TO THE NUMBER-CONTROLLED OSCILLATOR
FILTER DATA FOR ST/N_0 ESTIMATE
NO
TIME FOR ST/N_0 ESTIMATE ?
YES
COMPUTE ST/N_0
STORE ST/N_0 VALUE IN CORE
NO
HAVE 1000 SAMPLES BEEN STORED ?
YES
DISABLE TRACK OF BIT TIMING
OUTPUT STORED ST/N_0 SAMPLES TO MAGNETIC TAPE
NO
ENOUGH SAMPLES OBTAINED ?
YES
TERMINATE TEST

Fig. 2. Flow diagram of modified bit timing program

2.5, 5.2, 7.5, and 10 dB, both with the bit-sync loop locked and cycle slipping. The data samples were filtered in a bandwidth equivalent to 0.3% of the bit rate for input values of $ST/N_0 \geq 7.5$ dB, and 0.1% of the bit rate for input values of $ST/N_0 < 7.5$ dB. Each test contains a minimum of 10,000 independent samples of ST/N_0 estimate. This required that samples be taken at least [1/(bandwidth × bit rate)] seconds apart. For convenience, a bit rate of 250 bits/s was arbitrarily selected, yielding sample rates of 0.75 samples per second for the 0.3% bandwidth and 0.25 samples per second for the 0.1% bandwidth. For the frequency offset or cycle slipping measurements, an offset of 3.6% of the bit rate was used. Figure 2 is an abbreviated flow diagram showing the operation of the modified bit timing program.

The data tapes were processed using a data-analysis computer program, and a histogram of the ST/N_0 estimate was plotted for each value of input ST/N_0. The complete set of plots is included in *Chapter XX-G* of this volume. Figure 3 is a summary of the results showing the spread of each ST/N_0 estimate probability density together with the proposed lock thresholds.

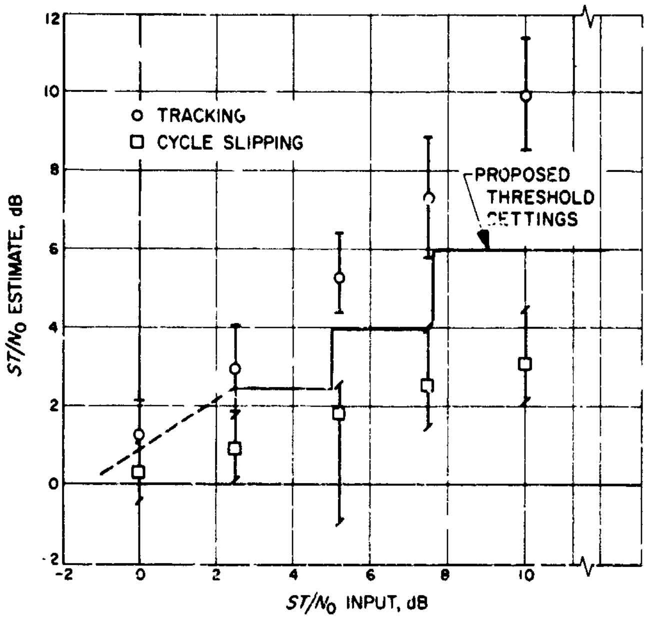


Fig. 3. ST/N_0 estimate distributions

B. Relay Telemetry Modulation System Development, *C. Carl*

The overall objective of this development effort is the design and test of telemetry modulation systems for relay-link applications, such as between a planetary entry capsule and a nearby orbiting or flyby spacecraft.

The breadboard evaluation of a proposed relay link for a *Mariner* 1971-type mission is continuing as previously described in SPS 37-50, Vol. III, pp. 326–331. That article described the test results of an audio equivalent RF transmitter–receiver followed by a bit synchronizer.

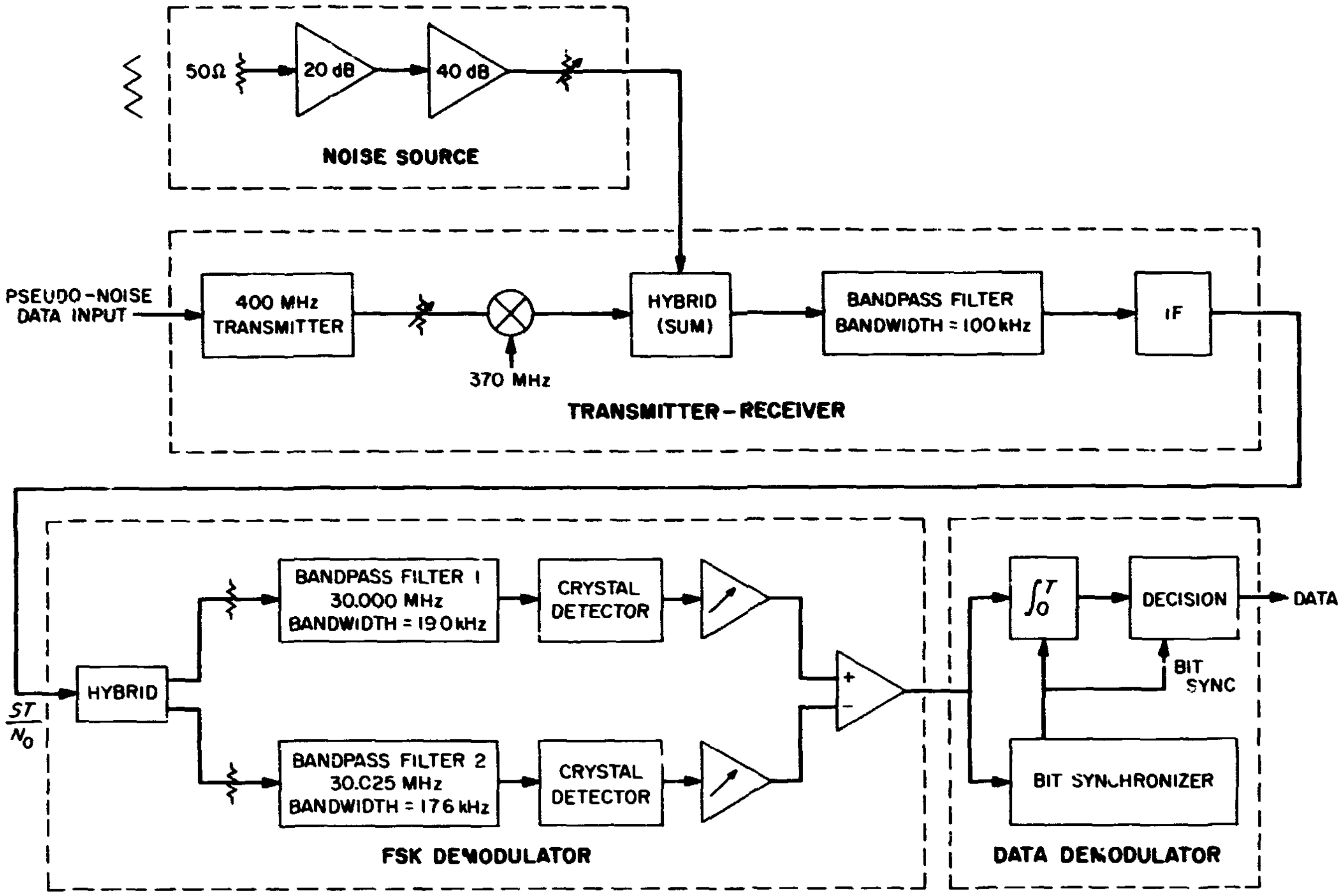


Fig. 4. Experimental relay link

The audio equivalent transmitter–receiver pair has been replaced by a breadboard 400 MHz FSK transmitter and receiver[1] for the purposes of running complete link compatibility tests. The configuration is as shown in Fig. 4. Random data modulates the transmitter; the down-converted transmitter output, at 30 MHz, is mixed with broadband noise, to establish a controlled signal-energy to noise-density ratio (ST/N_0) at the receiver IF. After IF amplification, the signal is FSK-demodulated by the conventional topology consisting of crystal filters, square-law detectors, channel-balance amplifiers, and subtractor. Finally, the bit synchronizer and data detector recover data and bit-sync timing from the noisy FSK-demodulated data stream.

The noise bandwidths of bandpass filters 1 and 2 were averaged and that value (18.3 kHz) used for determining ST/N_0 and N, the IF bandwidth to bit-rate ratio ($N = 36.6$). The bit synchronizer uses the absolute-value phase detector topology and 60-Hz loop bandwidth ($2B_L$) as described in the referenced SPS.

The first bit-error and acquisition-time tests have been completed and are shown in Figs. 5 and 6, respectively. The theoretical performance curve of Fig. 5 is extracted from Boyd.[2] The hardline bit-sync data is in excellent agreement with theory. Using bit sync derived from the bit synchronizer, a 0.3–0.4 dB loss is observed; this value of sync loss was also observed in the audio-equivalent receiver tests. The 0.9 probability of acquisition time for frequency offsets (Δf) of 2.0 and 4.0 Hz. The values are also consistent with those obtained with the audio receiver.

Extensive bit-error and acquisition time tests are scheduled to determine the performance of this RF relay link as a function of RF limiting, square law versus linear envelope detectors, and channel unbalance.

[1] The RF equipment has been developed under NASA Code 186-68-04-08, Relay-Link RF Systems.

[2] Boyd, D. W., *Performance of FSK Systems with Large Uncertainty in the Carrier Frequency*, Apr. 3, 1967 (JPL internal document).

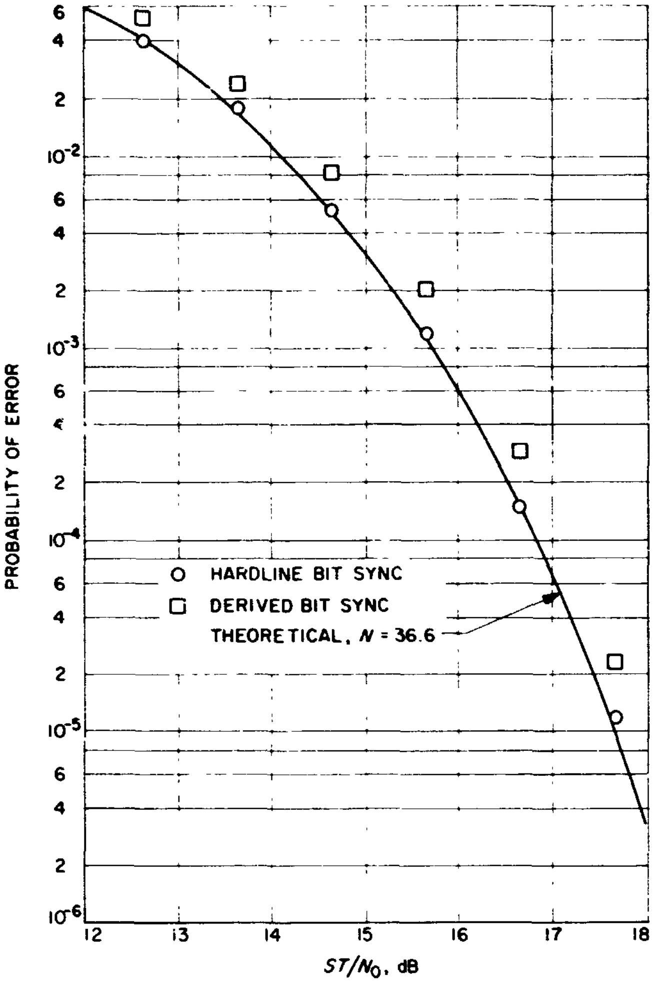


Fig. 5. FSK bit-error test

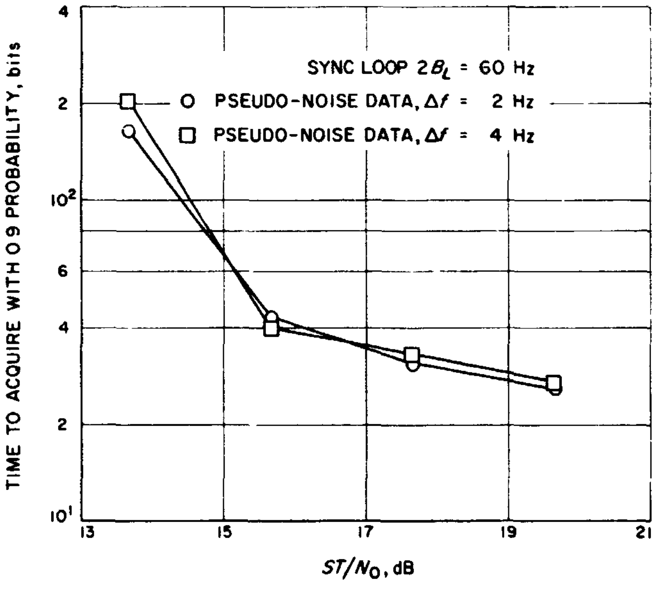


Fig. 6. FSK bit-sync acquisition time test

XXIII. Spacecraft Radio

TELECOMMUNICATIONS DIVISION

A. *Lunar Orbiter V* Side-Looking Radar Experiment, *R. L. Horttor*

1. Introduction

For some time the Laboratory has been developing surface imaging or mapping radar systems applicable to lunar and planetary missions. Present spacecraft ordinarily have telecommunication elements that are very similar to the elements used in such radar systems. Preliminary investigation has shown that the S-band ranging transponder with a high-gain antenna could serve as a side-looking radar. To demonstrate this idea, an experiment was performed on January 24, 1968, using the S-band ranging transponder and high-gain antenna of the *Lunar Orbiter V* spacecraft in flight. A description of the experiment and the derivations of the mapping equations are presented in this article.

2. Experiment

The equipment used in this experiment is different from that of the usual side-looking radar, because the radar transmitter and receiver are widely separated. As far as is known, a bistatic side-looking radar experiment has never been performed before. With reference to Fig. 1, the actual radar signal is transmitted from the spacecraft, reflected from the lunar surface, and received at the Mars DSS. In order to keep time and frequency references, the ranging modulation is actually transmitted from the Mars DSS to the spacecraft, routed through the transponder, and retransmitted on a different carrier frequency.

As described, the experiment communication link contains three time-varying delay times. Proper tracking of the round trip delay and doppler is the crux of the data-processing problem.

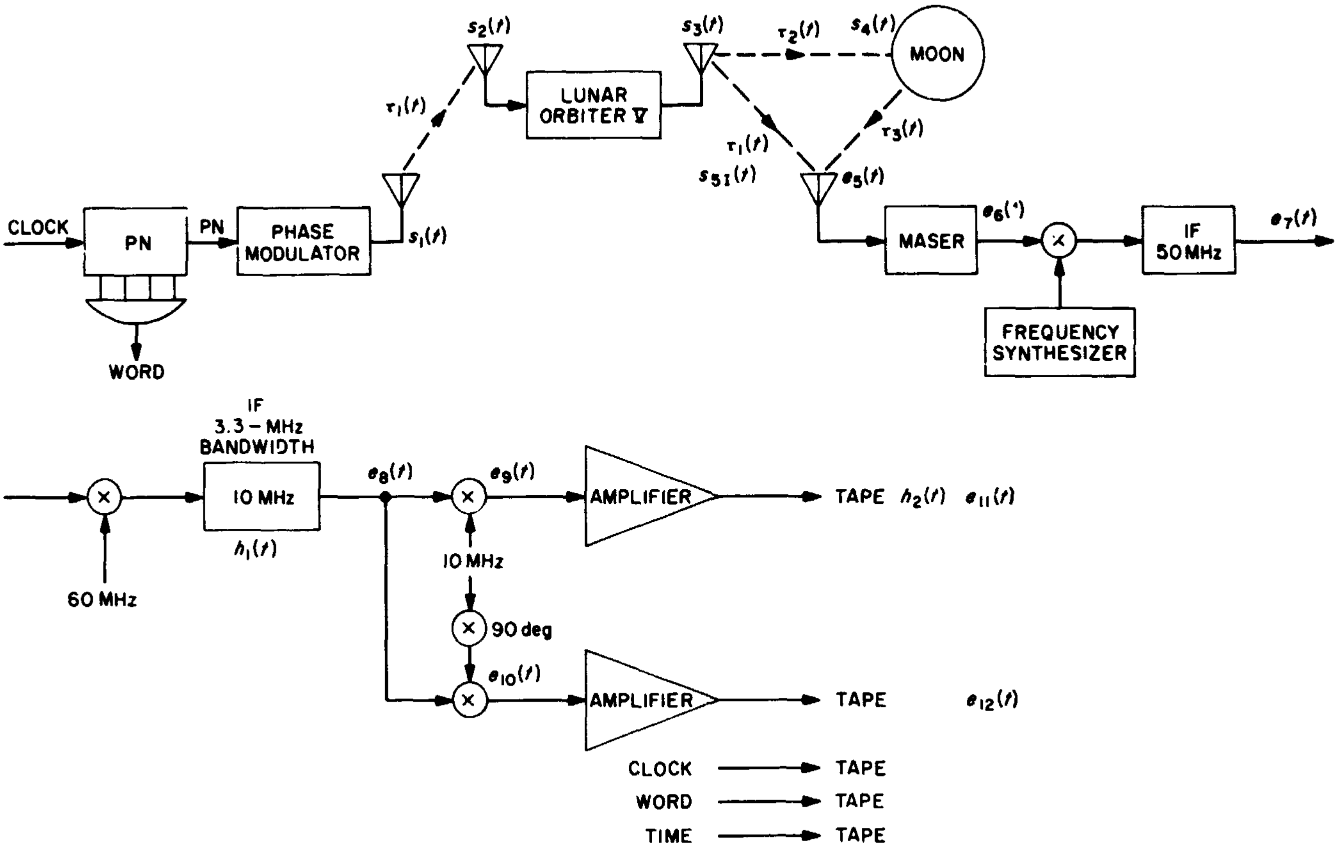


Fig. 1. Block diagram of bistatic side-looking radar experiment and signal flow model

The received signal is recorded in phase quadrature at baseband. Range resolution is achieved using a pseudonoise (PN) code biphase-modulated on the carrier. The received signal is multiplied by an identical locally generated PN sequence. The portion of the received signal whose modulation is synchronized to the local code can be separated from the rest by a low-pass filter. This signal corresponds to a narrow strip at constant range from the spacecraft, as shown in Fig. 2. Each point within that strip passes through the lines of constant doppler caused by the motion of the spacecraft. A filter which matches that motion-induced phase behavior can resolve individual point scatterers. This is the basic principle of the side-looking radar.

The surface resolution is determined by the radar beam incidence angle, the PN code bit length, and the bandwidths of the spacecraft transponder, the DSN receiver IF amplifier, and the tape recorder. The bit rate chosen for the experiment was a ⅓-MHz clock rate, allowing a 3-μs bit time. This is 1.0 km in slant range, corresponding to about 1.4 km on the surface.

The code length was long enough to keep the spacecraft direct signal and the surface signal unambiguous. Also, the longer the code, the better the suppression of the direct signal. However, searching for the surface reflected signal gets more difficult as the code is lengthened. The code chosen was 1023 bits long.

Signal strength calculations were based on resolving a 1-km square on the surface. Predictions showed a 3-dB signal-to-noise ratio. Such a noisy picture should reveal a recognizable shape, such as a large crater. Analysis of the data is not yet complete.

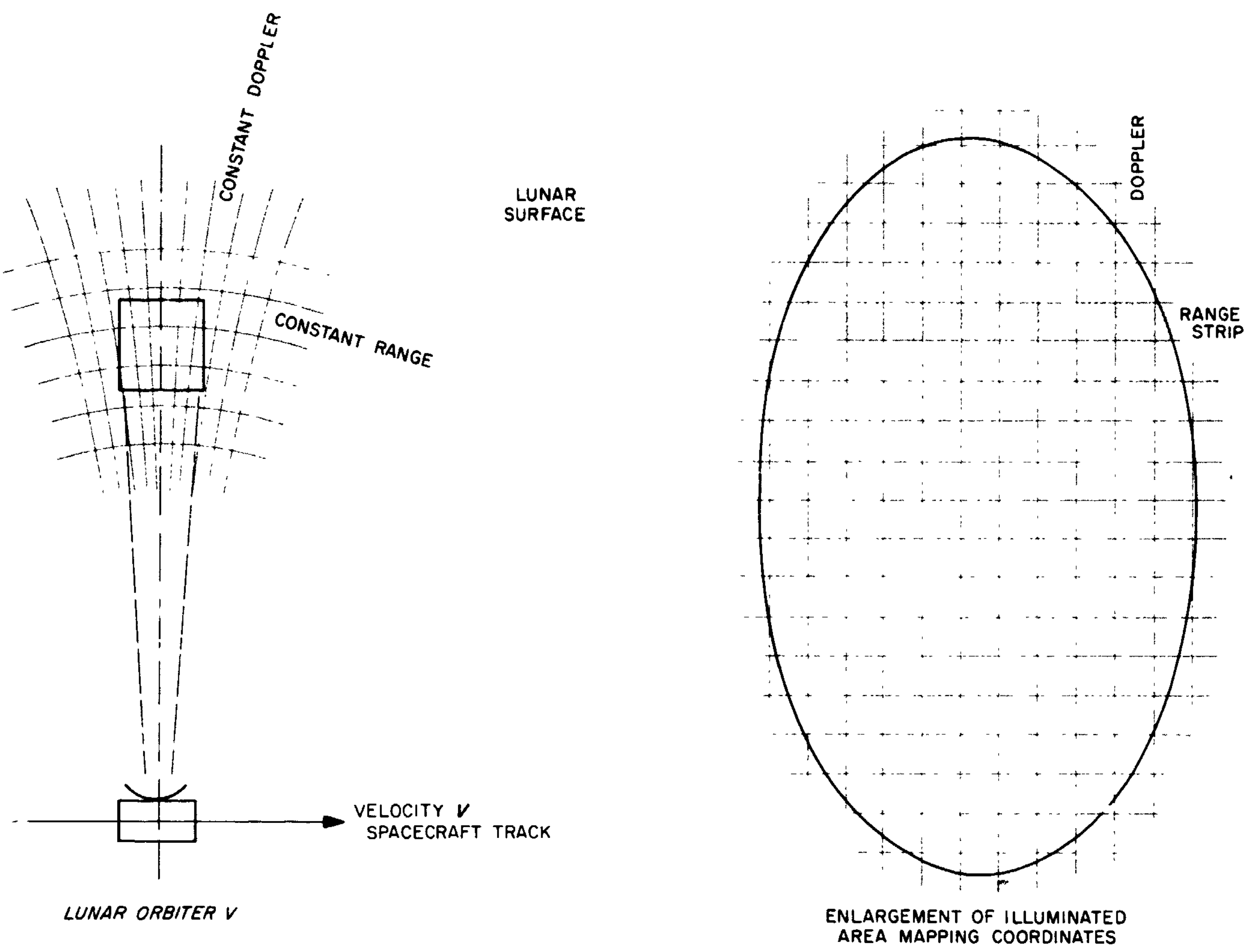


Fig. 2. Lunar Orbiter V side-looking radar mapping coordinates

3. Analysis

This section presents the signal flow from the earth station to the spacecraft to the lunar surface and back to the station. Further operations are performed, culminating in the system response to a point reflector on the lunar surface. A map is the superposition of many responses from the surface features.

a. Communication link round trip time delay. Each link in the process is characterized by the time delay $\tau_i(t)$. The delay is a function of time because the velocity of light is finite, and the three elements in the system are moving with respect to one another. Figure 1 has already identified these links. Only the time delay of the spacecraft-Moon link, $\tau_2(t)$, is of real interest, because it is the particular nature of its variation which allows resolution along the direction of travel.

The modulation of the transmitted signal is a PN code. One length of the code is denoted by $x(t)$ and is T sec long. The modulating signal is then

$$X(t) = \sum_{n=-\infty}^{\infty} x(t-nT), \qquad x(t) = \pm 1 \tag{1}$$

Assuming a modulation index β, the transmitted signal is

$$s_1(t) = A_1 2^{1/2} \cos(\omega_1 t + \beta X(t)) \tag{2}$$

The signal at the spacecraft is attenuated and delayed.

$$s_2(t) = A_2 2^{1/2} \cos(\omega_1 [t - \tau_1(t)] + \beta X [t - \tau_1(t)]) \tag{3}$$

At the spacecraft, the transponder retransmits the modulation on a different carrier frequency.

$$s_3(t) = A_3 2^{1/2} \cos\left(\omega_3 \left[t - \frac{\omega_1}{\omega_3}\tau_1(t)\right] + \beta X [t - \tau_1(t)]\right) \tag{4}$$

On the lunar surface, the signal has additional delay $\tau_2(t)$ and a time-varying amplitude factor caused by motion through the antenna illumination pattern.

$$s_4(t) = A_4(t - \tau_2(t)) 2^{1/2} \cos\left(\omega_3 \left[t - \frac{\omega_2}{\omega_3}\tau_1(t - \tau_2(t)) - \tau_2(t)\right] + \beta X [t - \tau_1(t - \tau_2(t)) - \tau_2(t)]\right) \tag{5}$$

At the earth station the signal has additional time delay $\tau_3(t)$, an attenuated power factor $A_5(°)$, and white gaussian lunar background noise. The phase factor θ_1 is arbitrary but fixed for each scatterer, while θ_2 is uniformly random.

$$\begin{aligned} e_5(t) = {} & A_5 [t - \tau_2(t - \tau_3(t)) - \tau_3(t)] 2^{1/2} \cos\left[\omega_3 \left(t - \frac{\omega_1}{\omega_3}\tau_1 [t - \tau_2(t - \tau_3(t)) - \tau_3(t)] - \tau_2(t - \tau_3(t)) - \tau_3(t)\right)\right. \\ & \left. + \beta X \left(t - \tau_1 [t - \tau_2(t - \tau_3(t)) - \tau_3(t)] - \tau_2(t - \tau_3(t)) - \tau_3(t)\right) + \theta_1\right] \\ & + n_1(t) \cos(\omega_3 t + \theta_2) + n_2(t) \sin(\omega_3 t + \theta_2) \end{aligned} \tag{6}$$

Let the total time delay be denoted by $\tau(t)$.

$$\begin{aligned} \tau(t) &= \tau_1(t - \tau_a(t)) + \tau_a(t) \\ \tau_\phi(t) &= \frac{\omega_1}{\omega_3}\tau_1(t - \tau_a(t)) + \tau_a(t) \\ \tau_a(t) &= \tau_2(t - \tau_3(t)) + \tau_3(t) \end{aligned} \tag{7}$$

The functions will later be calculated by the ephemeris and trajectory data and will contain the desired phase behavior of the spacecraft-to-surface link. The received signal is, therefore,

$$e_5(t) = A_5(t - \tau_a(t)) 2^{1/2} \cos[\omega_3(t - \tau_\phi(t)) + \beta X(t - \tau(t)) + \theta_1] + n_1(t) \cos(\omega_3 t + \theta_2) + n_2(t) \sin(\omega_3 t + \theta_3) \tag{8}$$

In addition, there is the direct component from the spacecraft omni-antenna. Denote it by $s_{5I}(t)$, the interference component.

$$s_{5I}(t) = A_{5I} 2^{1/2} \cos\left(\omega_3 \left[t - \frac{\omega_1}{\omega_3}\tau_1(t - \tau_1(t)) - \tau_1(t)\right] + \beta X [t - \tau_1(t - \tau_1(t)) - \tau_1(t)]\right) \tag{9}$$

Let the interference signal time delay be $\tau_I(t)$. Then

$$s_{5I}(t) = A_{5I}\, 2^{1/2} \cos\left[\omega_3 (t - \tau_{I\phi}(t)) + \beta X(t - \tau_I(t))\right] \tag{10}$$

where

$$\tau_I(t) = \tau_1(t - \tau_1(t)) + \tau_I(t)$$

$$\tau_{I\phi}(t) = \frac{\omega_1}{\omega_3} \tau_1(t - \tau_1(t)) + \tau_1(t) \tag{11}$$

This component could be coherently tracked by the receiver, if desired.

b. Receiver signal flow. At the receiver, signal is amplified by the maser front end and mixed with the local oscillator frequency ω_{LO}. The receiver is not tracking, hence ω_{LO} is constant.

$$e_6(t) = (s_5(t) + S_{5I}(t) + n_5(t))\, 2^{1/2} \cos \omega_{LO} t$$

The frequency difference $(\omega_3 - \omega_{LO})$ is 50 MHz and is denoted by ω_{50}. Double frequency terms are dropped.

$$\begin{aligned} e_6(t) &= A_5(t - \tau_a(t)) \cos\left[\omega_{50} t - \omega_3 \tau_\phi(t) + \beta X(t - \tau(t)) + \theta_1\right] + A_{5I} \cos\left[\omega_{50} t - \omega_3 \tau_{I\phi}(t) + \beta X(t - \tau_I(t))\right] \\ &\quad + n_1(t) \frac{1}{2^{1/2}} \cos(\omega_{50} t + \theta_2) - n_2(t) \frac{1}{2^{1/2}} \sin(\omega_{50} t + \theta_2) \end{aligned} \tag{12}$$

Mixing with a 60-MHz reference produces signal on a 10-MHz IF. Make substitutions for the phase terms $\omega_3 \tau_\phi(t)$ and $\omega_3 \tau_{I\phi}(t)$.

$$\phi(t) = \omega_3 \tau_\phi(t) \tag{13a}$$

$$\phi_I(t) = \omega_3 \tau_{I\phi}(t) \tag{13b}$$

$$e_7(t) = e_6(t)\, 2 \cos \omega_{60} t$$

$$\begin{aligned} e_7(t) &= A_5(t - \tau_a(t)) \cos\left[\omega_{10} t + \phi(t) - \beta X(t - \tau(t)) - \theta_1\right] + A_{5I} \cos\left[\omega_{10} t + \phi_I(t) - \beta X(t - \tau_I(t))\right] \\ &\quad + n_1(t) \frac{1}{2^{1/2}} \cos(\omega_{10} t - \theta_2) - n_2(t) \frac{1}{2^{1/2}} \sin(\omega_{10} t - \theta_2) \end{aligned} \tag{14}$$

After mixing to 10 MHz, the signal is passed through the 10-MHz IF filter, with impulse response $h_1(t)$. The filter bandwidth is 3.3 MHz. The convolution integral uses the dummy variable ρ_1.

$$\begin{aligned} e_8(t) = \int_0^\infty d\rho_1\, h_1(\rho_1) \Big\{ & A_5(t - \tau_a(t - \rho_1) - \rho_1) \cos\left[\omega_{10}(t - \rho_1) + \phi(t - \rho_1) - \beta X(t - \tau(t - \rho_1) - \rho_1) - \theta_1\right] \\ & + A_{5I} \cos\left[\omega_{10}(t - \rho_1) + \phi_I(t - \rho_1) - \beta X(t - \tau_I(t - \rho_1) - \rho_1)\right] \\ & \times n_1(t - \rho_1) \frac{1}{2^{1/2}} \cos(\omega_{10}(t - \rho_1) - \theta_2) - n_2(t - \rho_1) \frac{1}{2^{1/2}} \sin(\omega_{10}(t - \rho_1) - \theta_2) \Big\} \end{aligned} \tag{15}$$

Remove the IF frequency and record the result on magnetic tape. Since the signal is now at baseband, quadrature components must be kept.

$$e_9(t) = e_8(t)\, 2 \sin \omega_{10} t$$

$$e_{10}(t) = e_8(t)\, 2 \cos \omega_{10} t$$

Upon substitution these signals become

$$e_9(t) = \int_0^\infty d\rho_1 \, h_1(\rho_1) \, 2 \sin \omega_{10} t \Big\{ A_5 (t - \tau_a(t - \rho_1) - \rho_1) \cos [\omega_{10}(t - \rho_1) + \phi(t - \rho_1) - \beta X(t - \tau(t - \rho_1) - \rho_1) - \theta_1]$$

$$+ A_{5I} \cos [\omega_{10}(t - \rho_1) + \phi_I(t - \rho_1) - \beta X(t - \tau(t - \rho_1) - \rho_1)]$$

$$+ n_1(t - \rho_1) \frac{1}{2^{1/2}} \cos (\omega_{10}(t - \rho_1) - \theta_2) - n_2(t - \rho_1) \frac{1}{2^{1/2}} \sin (\omega_{10}(t - \rho_1) - \theta_2) \Big\} \tag{16a}$$

$$e_{10}(t) = \int_0^\infty d\rho_1 \, h_1(\rho_1) \, 2 \cos \omega_{10} t \Big\{ A_5 (t - \tau_a(t - \rho_1) - \rho_1) \cos [\omega_{10}(t - \rho_1) + \phi(t - \rho_1) - \beta X(t - \tau(t - \rho_1) - \rho_1) - \theta_1]$$

$$+ A_5 \cos [\omega_{10}(t - \rho_1) + \phi_I(t - \rho_1) - \beta X(t - \tau_I(t - \rho_1) - \rho_1)]$$

$$+ n_1(t - \rho_1) \frac{1}{2^{1/2}} \cos (\omega_{10}(t - \rho_1) - \theta_2) - n_2(t - \rho_2) \frac{1}{2^{1/2}} \sin (\omega_{10}(t - \rho_1) - \theta_2) \Big\} \tag{16b}$$

Expanding the sinusoidal products and discarding the double-frequency terms gives

$$e_9(t) = \int_0^\infty d\rho_1 \, h_1(\rho_1) \Big\{ A_5 (t - \tau_a(t - \rho_1) - \rho_1) \sin [\omega_{10} \rho_1 - \phi(t - \rho_1) + \beta X(t - \tau(t - \rho_1) - \rho_1) + \theta_1]$$

$$+ A_{5I} \sin [\omega_{10} \rho_1 - \phi_I(t - \rho_1) + \beta X(t - \tau_I(t - \rho_1) - \rho_1)]$$

$$+ n_1(t - \rho_1) \frac{1}{2^{1/2}} \sin (\omega_{10} \rho_1 + \theta_2) - n_2(t - \rho_1) \frac{1}{2^{1/2}} \cos (\omega_{10} \rho_1 + \theta_2) \Big\} \tag{16c}$$

$$e_{10}(t) = \int_0^\infty d\rho_1 \, h_1(\rho_1) \Big\{ A_5 (t - \tau_a(t - \rho_1) - \rho_1) \cos [\omega_{10} \rho_1 - \phi(t - \rho_1) + \beta X(t - \tau(t - \rho_1) - \rho_1) + \theta_1]$$

$$+ A_{5I} \cos [\omega_{10} \rho_1 - \phi_I(t - \rho_1) + \beta X(t - \tau_I(t - \rho_1) - \rho_1)]$$

$$+ n_1(t - \rho_1) \frac{1}{2^{1/2}} \cos (\omega_{10} \rho_1 + \theta_2) + n_2(t - \rho_1) \frac{1}{2^{1/2}} \sin (\omega_{10} \rho_1 + \theta_2) \Big\} \tag{16d}$$

But $h_1(t)$ is a bandpass filter function

$$h_1(t) = h_{IF}(t) \, 2 \cos (\omega_{10} t + \theta_{IF}(t)) \tag{17}$$

where $h_{IF}(t)$ and $\theta_{IF}(t)$ are the amplitude and phase functions, respectively.

$$e_9(t) = \int_0 d\rho_1 \, h_{IF}(\rho_1) \, 2 \cos (\omega_{10} \rho_1 + \theta_{IF}(\rho_1))$$

$$\times \Big\{ A_5 (t - \tau_a(t - \rho_1) - \rho_1) \sin [\omega_{10} \rho_1 - \phi(t - \rho_1) + \beta X(t - \tau(t - \rho_1) - \rho_1) + \theta_1]$$

$$+ A_{5I} \sin [\omega_{10} \rho_1 - \phi_I(t - \rho_1) + \beta X(t - \tau_I(t - \rho_1) - \rho_1)]$$

$$\times n_1(t - \rho_1) \frac{1}{2^{1/2}} \sin (\omega_{10} \rho_1 + \theta_2) - n_2(t - \rho_1) \frac{1}{2^{1/2}} \cos (\omega_{10} \rho_1 + \theta_2) \Big\} \tag{18a}$$

$$
\begin{aligned}
e_{10}(t) = \int_0^\infty & d\rho_1\, h_{IF}(\rho_1) \cos(\omega_{10}\rho_1 + \theta_{IF}(\rho_1)) \\
& \times \Big\{ A_5(t - \tau_a(t-\rho_1) - \rho_1) \cos[\omega_{10}\rho_1 - \phi(t-\rho_1) + \beta X(t - \tau(t-\rho_1) - \rho_1) + \theta_1] \\
& + A_{5I} \cos[\omega_{10}\rho_1 - \phi_I(t-\rho_1) + \beta X(t - \tau_I(t-\rho_1) - \rho_1)] \\
& \times n_1(t-\rho_1) \frac{1}{2^{1/2}} \cos(\omega_{10}\rho_1 + \theta_2) + n_2(t-\rho_1) \frac{1}{2^{1/2}} \sin(\omega_{10}\rho_1 + \theta_2) \Big\}
\end{aligned}
\tag{18b}
$$

Expanding the sine and cosine products and discarding double-frequency terms gives

$$
\begin{aligned}
e_9(t) = \int_0^\infty d\rho_1\, h_{IF}(\rho_1) & \Big\{ A_5(t - \tau_a(t-\rho_1) - \rho_1) \sin[-\phi(t-\rho_1) + BX(t - \tau(t-\rho_1) - \rho_1) - \theta_{IF}(\rho_1) + \theta_1] \\
& + A_{5I} \sin[-\phi_I(t-\rho_1) + \beta X(t - \tau_I(t-\rho_1) - \rho_1) - \theta_{IF}(\rho_1)] \\
& + n_1(t-\rho_1) \frac{1}{2^{1/2}} \sin(\theta_2 - \theta_{IF}(\rho_1)) - n_2(t-\rho_1) \frac{1}{2^{1/2}} \cos(\theta_2 - \theta_{IF}(\rho_1)) \Big\}
\end{aligned}
\tag{19a}
$$

$$
\begin{aligned}
e_{10}(t) = \int_0^\infty d\rho_1\, h_{IF}(\rho_1) & \Big\{ A_5(t - \tau_a(t-\rho_1) - \rho_1) \cos[-\phi(t-\rho_1) + \beta X(t - \tau(t-\rho_1) - \rho_1) - \theta_{IF}(\rho_1) + \theta_1] \\
& + A_{5I} \cos[-\phi_I(t-\rho_1) + \beta X(t - \tau_I(t-\rho_1) - \rho_1) - \theta_{IF}(\rho_1)] \\
& + n_1(t-\rho_1) \frac{1}{2^{1/2}} \cos(\theta_2 - \theta_{IF}(\rho_1)) + n_2(t-\rho_1) \frac{1}{2^{1/2}} \sin(\theta_2 - \theta_{IF}(\rho_1)) \Big\}
\end{aligned}
\tag{19b}
$$

Let $h_2(t)$ be the filter associated with the tape recorder response.

$$
\begin{aligned}
e_{11}(t) = \int_0^\infty & d\rho_1 \int_0^\infty d\rho_2\, h_2(\rho_2)\, h_{IF}(\rho_1) \\
& \times \Big\{ A_5(t - \tau_a(t-\rho_1-\rho_2) - \rho_1 - \rho_2) \sin[-\phi(t-\rho_1-\rho_2) + \beta X(t - \tau(t-\rho_1-\rho_2) - \rho_1 - \rho_2) - \theta_{IF}(\rho_1) + \theta_1] \\
& + A_{5I} \sin[-\phi_I(t-\rho_1-\rho_2) + \beta X(t - \tau(t-\rho_1-\rho_2) - \rho_1 - \rho_2) - \theta_{IF}(\rho_1)] \\
& + n_1(t-\rho_1-\rho_2) \frac{1}{2^{1/2}} \sin(\theta_2 - \theta_{IF}(\rho_1)) - n_2(t-\rho_1-\rho_2) \frac{1}{2^{1/2}} \cos(\theta_2 - \theta_{IF}(\rho_1)) \Big\}
\end{aligned}
\tag{20a}
$$

$$
\begin{aligned}
e_{12}(t) = \int_0^\infty & d\rho_1 \int_0^\infty d\rho_2\, h_2(\rho_2)\, h_{IF}(\rho_1) \\
& \times \Big\{ A_5(t - \tau_a(t-\rho_1-\rho_2) - \rho_1 - \rho_2) \cos[-\phi(t-\rho_1-\rho_2) + \beta X(t - \tau(t-\rho_1-\rho_2) - \rho_1 - \rho_2) - \theta_{IF}(\rho_1) + \theta_1] \\
& + A_{5I} \cos[-\phi(t-\rho_1-\rho_2) + \beta X(t - \tau(t-\rho_1-\rho_2) - \rho_1 - \rho_2) - \theta_{IF}(\rho_1)] \\
& + n_1(t-\rho_1-\rho_2) \frac{1}{2^{1/2}} \cos(\theta_2 - \theta_{IF}(\rho_1)) + n_2(t-\rho_1-\rho_2) \frac{1}{2^{1/2}} \sin(\theta_2 - \theta_{IF}(\rho_1)) \Big\}
\end{aligned}
\tag{20b}
$$

*c. **Range code demodulation.*** Range gating or demodulation is performed by correlating the received signal with time shifted locally generated versions of the PN code modulation.

$$
\begin{aligned}
e_{13}(t) &= e_{11}(t)\, X(t - T_0) \\
e_{14}(t) &= e_{12}(t)\, X(t - T_0)
\end{aligned}
\tag{21}
$$

Both signals are passed through a low-pass filter for which $X(t)$ is rapidly varying and $\phi(t)$ is slowly varying.

$$
\begin{aligned}
e_{15}(t) &= \int_0^\infty d\rho_3\, h_3(\rho_3)\, e_{13}(t-\rho_3) \\
&= \int_0^\infty d\rho_3\, h_3(\rho_3)\, e_{11}(t-\rho_3)\, X(t-T_0-\rho_3) \\
e_{16}(t) &= \int_0^\infty d\rho_3\, h_3(\rho_3)\, e_{14}(t-\rho_3) \\
&= \int_0^\infty d\rho_3\, h_3(\rho_3)\, e_{12}(t-\rho_3)\, X(t-T_0-\rho_3)
\end{aligned}
\tag{22}
$$

Substituting the expression gives

$$
\begin{aligned}
e_{15}(t) = {} & \int_0^\infty d\rho_3 \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\rho_3)\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\, X(t-T_0-\rho_3) \\
& \times \Big\{ A_5\left(t - \tau_a(\rho_1-\rho_2-\rho_3) - \rho_1-\rho_2-\rho_3\right) \sin\left[-\phi(t-\rho_1-\rho_2-\rho_3)\right. \\
& + \beta X\left(t-\tau(t-\rho_1-\rho_2-\rho_3)-\rho_1-\rho_2-\rho_3\right) - \theta_{\mathrm{IF}}(\rho_1) + \theta_1 \big] \\
& + A_{5I} \sin\left[-\phi_I(t-\rho_1-\rho_2-\rho_3) + \beta X\left(t-\tau_I(t-\rho_1-\rho_2-\rho_3)-\rho_1-\rho_2-\rho_3\right) - \theta_{\mathrm{IF}}(\rho_1)\right] \\
& + n_1(t-\rho_1-\rho_2-\rho_3)\frac{1}{2^{1/2}} \sin\left(\theta_2-\theta_{\mathrm{IF}}(\rho_1)\right) - n_2(t-\rho_1-\rho_2-\rho_3)\frac{1}{2^{1/2}}\cos\left(\theta_2-\theta_{\mathrm{IF}}(\rho_1)\right)\Big\}
\end{aligned}
\tag{23a}
$$

$$
\begin{aligned}
e_{16}(t) = {} & \int_0^\infty d\rho_3 \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\rho_3)\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\, X(t-T_0-\rho_3) \\
& \times \Big\{ A_5\left(t - \tau_a(t-\rho_1-\rho_2-\rho_3) - \rho_1-\rho_2-\rho_3\right) \cos\left[-\phi(t-\rho_1-\rho_2-\rho_3)\right. \\
& + \beta X\left(t-\tau(t-\rho_1-\rho_2-\rho_3)-\rho_1-\rho_2-\rho_3\right) - \theta_{\mathrm{IF}}(\rho_1) + \theta_1 \big] \\
& + A_{5I} \cos\left[-\phi_I(t-\rho_1-\rho_2-\rho_3) + \beta X\left(t-\tau_I(t-\rho_1-\rho_2-\rho_3)-\rho_1-\rho_2-\rho_3\right) - \theta_{\mathrm{IF}}(\rho_1)\right] \\
& + n_1(t-\rho_1-\rho_2-\rho_3)\frac{1}{2^{1/2}} \cos\left(\theta_2-\theta_{\mathrm{IF}}(\rho_1)\right) + n_2(t-\rho_1-\rho_2-\rho_3)\frac{1}{2^{1/2}}\sin\left(\theta_2-\theta_{\mathrm{IF}}(\rho_1)\right)\Big\}
\end{aligned}
\tag{23b}
$$

Expand the sine and cosine products to separate the $\beta X(t)$ terms.

$$\sin(-\phi(t) + \beta X(t)) = -\sin\phi(t)\cos\beta X(t) + \cos\phi(t)\sin\beta X(t)$$

$$\cos(-\phi(t) + \beta X(t)) = \cos\phi(t)\cos\beta X(t) + \sin\phi(t)\sin\beta X(t)$$

But $X(t)$ is ± 1 only, so it may be removed from the arguments.

$$\sin(-\phi(t) + \beta X(t)) = -\cos\beta\sin(\phi(t)) + X(t)\sin\beta\cos(\phi(t))$$

$$\cos(-\phi(t) + \beta X(t)) = \cos\beta\cos(\phi(t)) + X(t)\sin\beta\sin(\phi(t))$$

Substituting into F.q. (23 a, b) gives

$$\begin{aligned} e_{15}(t) = \int_0^\infty d\rho_3 \int_\infty^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\rho_3)\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1) \Bigl\{ & A_5 (t - \tau_a (t - \rho_1 - \rho_2 - \rho_3) - \rho_1 - \rho_2 - \rho_3) \\ & \times \Bigl[X(t - T_0 - \rho_3)\, X(t - \tau(t - \rho_1 - \rho_2 - \rho_3) - \rho_1 - \rho_2 - \rho_3) \sin\beta \cos(\phi(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \\ & - X(t - T_0 - \rho_3) \cos\beta \sin(\phi(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \Bigr] \\ & + A_{5I} \Bigl[X(t - T_0 - \rho_3)\, X(t - \tau_I(t - \rho_1 - \rho_2 - \rho_3) - \rho_1 - \rho_2 - \rho_3) \sin\beta \cos(\phi_I(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1)) \\ & - X(t - T_0 - \rho_3) \cos\beta \sin(\phi_I(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1)) \Bigr] \\ & + X(t - T_0 - \rho_3) \Bigl[n_1(t - \rho_1 - \rho_2 - \rho_3) \frac{1}{2^{1/2}} \sin(\theta_2 - \theta_{\mathrm{IF}}(\rho_1)) - n_2(t - \rho_1 - \rho_2 - \rho_3) \frac{1}{2^{1/2}} \cos(\theta_2 - \theta_{\mathrm{IF}}(\rho_1)) \Bigr] \Bigr\} \end{aligned} \tag{24a}$$

$$\begin{aligned} e_{16}(t) = \int_0^\infty d\rho_3 \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\rho_3)\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1) \Bigl\{ & A_5 (t - \tau_a (t - \rho_1 - \rho_2 - \rho_3) - \rho_1 - \rho_2 - \rho_3) \\ & \times \Bigl[X(t - T_0 - \rho_3) \cos\beta \cos(\phi(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \\ & + X(t - T_0 - \rho_3)\, X(t - \tau(t - \rho_1 - \rho_2 - \rho_3) - \rho_1 - \rho_2 - \rho_3) \sin\beta \sin(\phi(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \Bigr] \\ & + A_{5I} \Bigl[X(t - T_0 - \rho_3) \cos\beta \cos(\phi_I(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1)) \\ & + X(t - T_0 - \rho_3)\, X(t - \tau_I(t - \rho_1 - \rho_2 - \rho_3) - \rho_1 - \rho_2 - \rho_3) \sin\beta \sin(\phi(t - \rho_1 - \rho_2 - \rho_3) + \theta_{\mathrm{IF}}(\rho_1)) \Bigr] \\ & + X(t - T_0 - \rho_3) \Bigl[n_1(t - \rho_1 - \rho_2 - \rho_3) \frac{1}{2^{1/2}} \cos(\theta_2 - \theta_{\mathrm{IF}}(\rho_1)) + n_2(t - \rho_1 - \rho_2 - \rho_3) \frac{1}{2^{1/2}} \sin(\theta_2 - \theta_{\mathrm{IF}}(\rho_1)) \Bigr] \Bigr\} \end{aligned} \tag{24b}$$

It has been stated previously that $h_3(t)$ changes rapidly compared to $X(t)$, but very slowly compared to $\tau_I(t)$, $\tau(t)$, $\phi(t)$ and $\phi_I(t)$. Hence, the integration over ρ_3 affects only $X(t)$ terms. The other factors may be removed from the ρ_3 integral. This statement implies the following:

$$\int_0^\infty d\rho_3\, h_3(\rho_3)\, X(t - T - \rho_3) \cong \bar{X}$$

$$\int_0^\infty d\rho_3\, h_3(\rho_3)\, X(t - T_0 - \rho_3)\, X(t - \tau(t - \rho_3 - \rho_2 - \rho_1) - \rho_1 - \rho_2 - \rho_3) \cong R_x(T_0 - \tau(t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \tag{25}$$

where $\overline{X}$ is the mean value of $X(t)$ and $R_x(\tau)$ is the autocorrelation function of $X(t)$. Equations (24a, b) are now

$$
\begin{aligned}
e_{15}(t) = \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1) \Big\{ & A_5 (t - \tau_a (t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \\
& \times \Big[\sin \beta R_x (T_0 - \tau (t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \cos (\phi (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \\
& - \overline{X} \cos \beta \sin (\phi (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \Big] \\
& + A_{5I} \Big[\sin \beta R_x (T_0 - \tau_I (t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \cos (\phi_I (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1)) \\
& - \overline{X} \cos \beta \sin (\phi (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1)) \Big] \Big\} + n_{15}(t)
\end{aligned}
\tag{26a}
$$

$$
\begin{aligned}
e_{16}(t) = \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1) \Big\{ & A_5 (t - \tau_a (t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \Big[\overline{X} \cos \beta \cos (\phi (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \\
& + \sin \beta R_x (t - \tau (t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \sin (\phi (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \Big] \\
& + A_{5I} \Big[\overline{X} \cos \beta \cos (\phi_I (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1)) \\
& + \sin \beta R_x (t - \tau_I (t - \rho_1 - \rho_2) - \rho_1 - \rho_2) \sin (\phi_I (t - \rho_1 - \rho_2) + \theta_{\mathrm{IF}}(\rho_1) - \theta_1) \Big] \Big\} + n_{16}(t)
\end{aligned}
\tag{26b}
$$

The noise processes are quasi-gaussian. Multiplication by $X(t)$ makes the process values at the PN code transition points undefined. But averaging in $h_3(t)$ applies the central limit theorem. Hence, the term "quasi-gaussian." Assume they are gaussian. The autocorrelation functions are equal.

$$
\begin{aligned}
R_{n_{15}}(\tau) = R_{n_{16}}(\tau) = & \int_0^\infty d\rho_6 \int_0^\infty d\rho_5 \int_0^\infty d\rho_4 \int_0^\infty d\rho_3 \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\rho_6)\, h_3(\rho_3)\, h_2(\rho_5)\, h_2(\rho_2) \\
& \times h_{\mathrm{IF}}(\rho_4)\, h_{\mathrm{IF}}(\rho_1)\, X(t - T_0 - \rho_3)\, X(t + \tau - T_0 - \rho_6) \\
& \times E \Big\{ n_1 (t - \rho_1 - \rho_2 - \rho_3)\, n_1 (t + \tau - \rho_4 - \rho_5 - \rho_6) \frac{1}{2} \sin^2 (\theta_2 - \theta_{\mathrm{IF}}(\rho_1)) \\
& + n_2 (t - \rho_1 - \rho_2 - \rho_3)\, n_2 (t + \tau - \rho_4 - \rho_5 - \rho_6) \frac{1}{2} \cos^2 (\theta_2 - \theta_{\mathrm{IF}}(\rho_1)) \Big\}
\end{aligned}
\tag{27}
$$

The cross terms of $n_1(t)$ and $n_2(t)$ have already been eliminated, since the two are independent. But the two noise processes are white with correlation $(\Phi/2)\,\delta(\tau)$.

$$
\begin{aligned}
R_{n_{15}}(\tau) = R_{n_{16}}(\tau) = & \int_0^\infty d\rho_6 \int_0^\infty d\rho_5 \int_0^\infty d\rho_4 \int_0^\infty d\rho_3 \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\rho_6)\, h_3(\rho_3)\, h_2(\rho_5)\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_4)\, h_{\mathrm{IF}}(\rho_1) \\
& \times X(t - T_0 - \rho_3)\, X(t + \tau - T_0 - \rho_6) \frac{\Phi}{4} \delta (\tau + \rho_1 - \rho_4 + \rho_2 - \rho_5 + \rho_3 - \rho_6)
\end{aligned}
\tag{28}
$$

Integrating over ρ_6 gives

$$R_{n_{15}}(\tau) = \frac{\Phi}{4}\int_0^\infty d\rho_5 \int_0^\infty d\rho_4 \int_0^\infty d\rho_3 \int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_3(\tau + \rho_1 - \rho_4 + \rho_2 - \rho_5 + \rho_3)\, h_3(\rho_3)\, h_2(\rho_5)\, h_2(\rho_2)$$

$$\times\, h_{\mathrm{IF}}(\rho_4)\, h_{\mathrm{IF}}(\rho_1)\, X(t - T_0 - \rho_3)\, X(t - T_0 + \rho_4 - \rho_1 + \rho_5 - \rho_2 - \rho_3) \tag{29}$$

For purposes of calculating $R_{n_{15}}(\tau)$ it is fair to assume that $h_2(t)$ and $h_{\mathrm{IF}}(t)$ do not affect $X(t)$, since the filters are wideband compared to $h_3(t)$. Treating them as unit impulses then gives

$$R_{n_{15}}(\tau) = \frac{\Phi}{4}\int_0^\infty d\rho_3\, h_3(\tau + \rho_3)\, h_3(\rho_3)\, X^2(t - T_0 - \rho_3) \tag{30}$$

But $X(t)$ is ±1, so $X^2(t)$ is constant.

$$R_{n_{15}}(\tau) = R_{n_{16}}(\tau) = \frac{\Phi}{4}\int_0^\infty d\rho_3\, h_3(\tau + \rho_3)\, h_3(\rho_3) \qquad -\infty < \tau < \infty \tag{31}$$

Equations (26a, b) may be simplified somewhat, since the phase and time delay functions are essentially unaffected by filters $h_2(t)$ and $h_{\mathrm{IF}}(t)$. Parts of the integrands may be moved outside the integrals.

$$e_{15}(t) = A_5(t - \tau_a(t))\Bigg\{\cos(\phi(t) - \theta_1)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\Big[\sin\beta R_x(T_0 - \tau(t) - \rho_1 - \rho_2)\cos\theta_{\mathrm{IF}}(\rho_1)$$

$$- \overline{X}\cos\beta\sin\phi_{\mathrm{IF}}(\rho_1)\Big] - \sin(\phi(t) - \theta_1)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)$$

$$\times\Big[\sin\beta R_x(T_0 - \tau(t) - \rho_1 - \rho_2)\sin\theta_{\mathrm{IF}}(\rho_1) + \overline{X}\cos\beta\cos\theta_{\mathrm{IF}}(\rho_1)\Big]\Bigg\}$$

$$+ A_{5I}\Bigg\{\cos(\phi_I(t))\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\Big[\sin\beta R_x(T_0 - \tau_I(t) - \rho_1 - \rho_2)\cos\theta_{\mathrm{IF}}(\rho_1)$$

$$- \overline{X}\cos\beta\sin\theta_{\mathrm{IF}}(\rho_1)\Big] - \sin\phi_I(t)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\, h_{\mathrm{IF}}(\rho_1)$$

$$\times\Big[\sin\beta R_x(T - \tau_I(t) - \rho_1 - \rho_2)\sin\theta_{\mathrm{IF}}(\rho_1) + \overline{X}\cos\beta\cos\theta_{\mathrm{IF}}(\rho_1)\Big]\Bigg\} + n_{15}(t) \tag{32a}$$

Similarly for $e_{16}(t)$

$$e_{16}(t) = A_5(t - \tau_a(t))\Bigg\{\cos(\phi(t) - \theta_1)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\Big[\sin\beta R_x(T_0 - \tau(t) - \rho_1 - \rho_2)\sin\theta_{\mathrm{IF}}(\rho_1)$$

$$+ \overline{X}\cos\beta\cos\theta_{\mathrm{IF}}(\rho_1)\Big] + \sin(\phi(t) - \theta_1)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)$$

$$\times\Big[\sin\beta R_x(T_0 - \tau(t) - \rho_1 - \rho_2)\cos\theta_{\mathrm{IF}}(\rho_1) - \overline{X}\cos\beta\sin\theta_{\mathrm{IF}}(\rho_1)\Big]\Bigg\}$$

$$+ A_{5I}\Bigg\{\cos\phi_I(t)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)\Big[\sin\beta R_x(T - \tau_I(t) - \rho_1 - \rho_2)\sin\phi_{\mathrm{IF}}(\rho_1) + \overline{X}\cos\beta\cos\theta_{\mathrm{IF}}(\rho_1)\Big]$$

$$+ \sin\phi_I(t)\int_0^\infty d\rho_2 \int_0^\infty d\rho_1\, h_2(\rho_2)\, h_{\mathrm{IF}}(\rho_1)$$

$$\times\Big[\sin\beta R_x(T_0 - \tau(t) - \rho_1 - \rho_2)\cos\theta_{\mathrm{IF}}(\rho_1) - \overline{X}\cos\beta\sin\theta_{\mathrm{IF}}(\rho_1)\Big]\Bigg\} + n_{16}(t) \tag{32b}$$

The expressions in Eqs. (32a, b) are not so formidable as they may seem. They contain both the modulated and carrier components of the reflected and direct signals. Observe that the carrier component may be suppressed by adjusting the modulation index β or the average value $\overline{X}$ of the PN code. This is more easily observed by assuming $h_{IF}(t)$ is very broad band compared to the code autocorrelation. In practice, most of the resolution degradation comes from the tape recorder $h_2(t)$. These are part of the assumptions used earlier in calculating $n_1(t)$ and $n_2(t)$.

$$h_{IF}(t) = \mu_0(t) \qquad \phi_{IF}(t) \triangleq 0 \tag{33}$$

Equations (32a, b) become

$$\begin{aligned} e_{15}(t) = A_5(t-\tau_a(t))&\left\{\cos(\phi(t)-\theta_1)\sin\beta\int_0^\infty d\rho_2\, h_2(\rho_2)\, R_x(T_0-\tau(t)-\rho_2)\right. \\ &\left.-\overline{X}\cos\beta\sin(\phi(t)-\theta_1)\right\} + A_{5I}\left\{\cos\phi_I(t)\sin\beta\int_0^\infty d\rho_2\, h_2(\rho_2)\, R_x(T_0-\tau_I(t)-\rho_2)\right. \\ &\left.-\overline{X}\cos\beta\sin(\phi_I(t))\right\} + n_{15}(t) \end{aligned} \tag{34a}$$

$$\begin{aligned} e_{16}(t) = A_5(t-\tau_a(t))&\left\{\sin(\phi(t)-\theta_1)\sin\beta\int_0^\infty d\rho_2\, h_2(\rho_2)\, R_x(T_0-\tau(t)-\rho_2)\right. \\ &\left.+\overline{X}\cos\beta\cos(\phi(t)-\theta_1)\right\} + A_{5I}\left\{\overline{X}\cos\beta\cos\phi_I(t)\right. \\ &\left.+\sin\phi_I(t)\sin\beta\int_0^\infty d\rho_2\, h_2(\rho_2)\, R_x(T_0-\tau_I(t)-\rho_2)\right\} + n_{16}(t) \end{aligned} \tag{34b}$$

Equations (34a, b) are the operating equations for analysis defining system mapping capability and the effects of interfering terms.

It is now instructive to demonstrate the two-dimensional ıature of Eqs. (34a, b) by showing how an array of point scatterers appear at $e_{15}(t)$ and $e_{16}(t)$. Assume that the surface is an array of point reflectors whose complex reflection coefficients are characterized by α_{ij} and θ_{ij}, the amplitude and phase angle, respectively. Furthermore, each resolution strip parallel to the vehicle track is associated with a time delay $\tau_j(t)$ and a phase variation $\phi_j(t)$. Within a strip, points are separated by their time occurrence t_i. Time delay $\tau_j(t)$ does not vary along a strip limited by antenna beam width. Signals $e_{15}(t)$ and $e_{16}(t)$ are then expressed by a double summation over (i, j).

$$\begin{aligned} e_{15}(t) = \sum_j &\left\{\sum_i \alpha_{ij}\, A_5(t-\tau_{aj}(t)-t_i)\left[\int_0^\infty d\rho_2\, h_2(\rho_2)\, R_x(T_0-\tau_j(t)-\rho_2)\right.\right. \\ &\left.\left.\times\sin\beta\cos(\phi_j(t-t_i)-\theta_{ij})-\overline{X}\cos\beta\sin(\phi_j(t-t_i)-\theta_{ij})\right]\right\} \\ &+ A_{5I}\left[\cos\phi_I(t)\sin\beta\int_0^\infty d\rho_2\, h_2(\rho_2)\, R_x(T_0-\tau_I(t)-\rho_2)\right. \\ &\left.-\overline{X}\cos\beta\sin\phi_I(t)\right] + n_{15}(t) \end{aligned} \tag{35a}$$

$$
\begin{aligned}
e_{16}(t) = \sum_j \Bigg\{ \sum_i \alpha_{ij} A_5 (t - \tau_{aj}(t) - t_i) \Bigg[\int_0^\infty d\rho_2 \, h_2(\rho_2) \, R_x (T_0 - \tau_j(t) - \rho_2) \\
\times \sin\beta \sin(\phi_j(t - t_j) - \theta_{ij}) + \overline{X} \cos\beta \cos(\phi_j(t - t_i) - \theta_{ij}) \Bigg] \Bigg\} \\
+ A_{5l} \Bigg[\sin\phi_l(t) \sin\beta \int_0^\infty d\rho_2 \, h_2(\rho_2) \, R_x (T_0 - \tau_j(t) - \rho_2) \\
+ \overline{X} \cos\beta \cos\phi_l(t) \Bigg] + n_{16}(t)
\end{aligned}
\tag{35b}
$$

Because of the peaked nature of $R_x(\tau)$, as shown in Fig. 3, the dominant term in Eq. (35a, b) is the one for which $T_0 \cong \tau_j(t)$. For this condition, the terms may be separated into signal from the *j*th surface strip plus interference. Thus $e_{15}(t)$ and $e_{16}(t)$ become

$$
\begin{aligned}
e_{15}(t) = \sum_i \alpha_{ij} A_5 (t - \tau_{aj}(t) - t_i) \Bigg[\int_0^\infty d\rho_2 \, h_2(\rho_2) \, R_x (T_0 - \tau_j(t) - \rho_2) \\
\times \sin\beta \cos(\theta_j(t - t_i) - \theta_{ij}) - \overline{X} \cos\beta \sin(\phi_j(t - t_i) - \theta_{ij}) \Bigg] \\
- \overline{X} \sum_{l \neq j} \sum_i \alpha_{il} A_5 (t - \tau_{al}(t) - t_i) \sin(\phi_l(t - t_i) - \beta - \theta_{il}) \\
- \overline{X} A_{5l} \sin(\phi_l(t) - \beta) + n_{15}(t)
\end{aligned}
\tag{36a}
$$

$$
\begin{aligned}
e_{16}(t) = \sum_i \alpha_{ij} A_5 (t - \tau_{aj}(t) - t_i) \Bigg[\int_0^\infty d\rho_2 \, h_2(\rho_2) \, R_x (T_0 - \tau_j(t) - \rho_2) \\
\times \sin\beta \sin(\phi_j(t - t_i) - \theta_{ij}) + \overline{X} \cos\beta \cos(\phi_j(t - t_i) - \theta_{ij}) \Bigg] \\
+ \overline{X} \sum_{l \neq j} \sum_i \alpha_{il} A_5 (t - \tau_l(t) - t_i) \cos(\phi_l(t) - \beta - \theta_{il}) \\
+ \overline{X} A_{5l} \cos(\phi_l(t) - \beta) + n_{16}(t)
\end{aligned}
\tag{36b}
$$

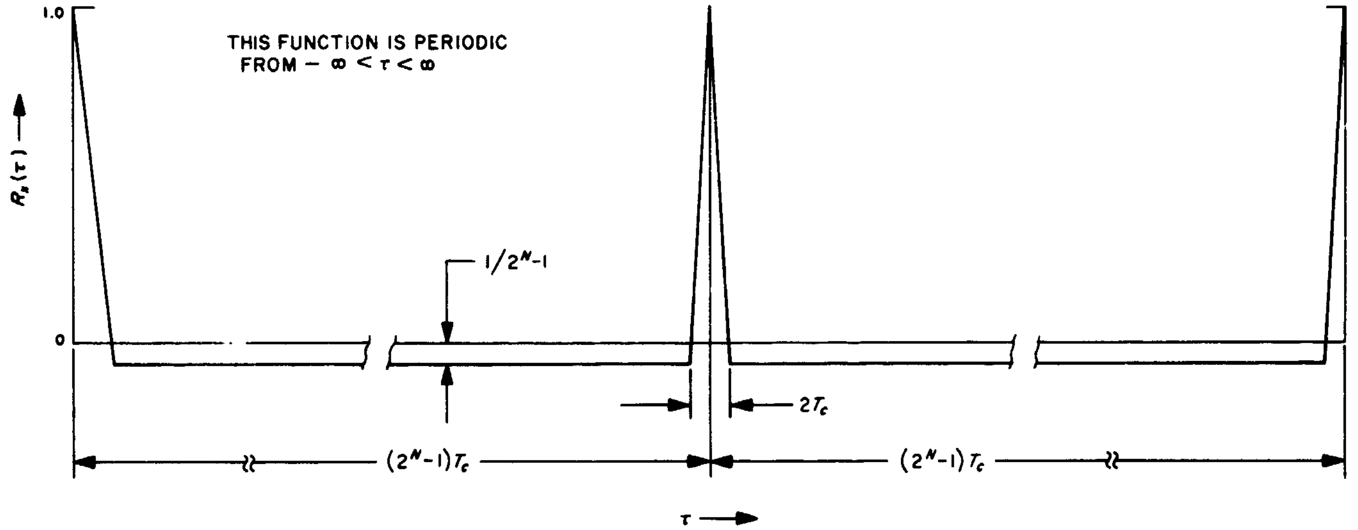


Fig. 3. Autocorrelation of range code X(t)

A glance at Fig. 3 indicates that $R_x(\tau)$ may be expressed as

$$\begin{aligned} R_x(\tau) &= [R_x(\tau) - \overline{X}] + \overline{X} \\ &= R_{0x}(\tau) + \overline{X} \end{aligned}$$

where $R_{0x}(\tau)$ is nonzero only in the region $\tau \cong 0,\ n(2^N - 1)T_c$. Equations (36a, b) become

$$\begin{aligned} e_{15}(t) = {} & \sin\beta \sum_i \alpha_{ij} A_5(t - \tau_{aj}(t) - t_i) \int_0^\infty d\rho_2\, h_2(\rho_2)\, R_{0x}(T_0 - \tau_j(t) - \rho_2) \cos(\phi_j(t - t_i) - \theta_{ij}) \\ & - \overline{X} \sum_j \sum_i \alpha_{ij} A_5(t - \tau_{aj}(t) - t_i) \sin(\phi_j(t - t_i) - \theta_{ij} - \beta) \\ & - \overline{X} A_{5I} \sin(\phi_I(t) - \beta) + n_{15}(t) \end{aligned} \tag{37a}$$

$$\begin{aligned} e_{16}(t) = {} & \sin\beta \sum_i \alpha_{ij} A_5(t - \tau_{aj}(t) - t_i) \int_0^\infty d\rho_2\, h_2(\rho_2)\, R_{0x}(T_0 - \tau_j(t) - \rho_2) \sin(\phi_j(t - t_i) - \theta_{ij}) \\ & + \overline{X} \sum_j \sum_i \alpha_{ij} A_5(t - \tau_{aj}(t) - t_i) \cos(\phi_j(t - t_i) - \beta - \theta_{ij}) \\ & + \overline{X} A_{5I} \cos(\phi_I(t) - \beta) + n_{16}(t) \end{aligned} \tag{37b}$$

Note how the return signal consists of a portion limited by the modulation to the strip and a portion from the whole area contributed by the carrier component caused by the code average value.

It is important to make $\overline{X}$ as small as possible. This is done by making the PN code as long as possible. For a code of length $(2^N - 1)$,

$$R_x(nT_c) = 1 \qquad n = 0, \pm(2^N - 1), \pm 2(2^N - 1), \cdots$$

$$T_c = \text{bit period}$$

$$R_x(\tau) = -\frac{1}{2^N - 1}$$

$$\overline{X} = -\frac{1}{2^N - 1}$$

The map of the surface is reproduced in the following way. Signals $e_{15}(t)$ and $e_{16}(t)$ are combined in a single sideband mixer and passed through a filter matched to $A_5(t)$ and $\phi_j(t)$. Depending on the detailed nature of $\phi_j(t)$, the scatterers for each j-strip are resolved. The process is repeated for each j. The results are mapped on a (T_0, t) plane. A strip parallel to the vehicle track will be reproduced along the contour $T_0 = \tau_j(t)$. Because of the relative motion between station, spacecraft, and moon, $\tau_j(t)$ is a function of time, and hence, a strip maps into a curved strip, in general. However, T_0 is the time reference of the locally generated PN code. If provision is made for tracking the variable portion of $\tau_j(t)$ by making

$$T_0 = \tau(t) + \tau$$

$$\tau_j(t) = \tau(t) + \tau_j$$

then

$$T_0 - \tau_j(t) = \tau - \tau_j$$

and the output map is fixed, because the mapping coordinate would be fixed to (τ, t). The analysis of phase processing of $\phi_j(t)$ and the corresponding need for phase tracking will appear as convenience allows.

B. Power Spectral Densities for Binary Frequency-Shift-Keyed Waveforms, *D. W. Boyd*

1. Introduction

In designing a communications receiver, it is important to know the power spectral density of the received waveform. This quantity defines the distribution of average signal power versus frequency and is useful primarily for locating the frequency bands of most interest. In this article we shall specialize certain general results from Ref. 1 for the power spectra of binary frequency-shift-keyed waveforms.

2. Basic Assumptions and Definitions

Following Ref. 1, we assume that the transmitted waveform is given by

$$\left.\begin{aligned} u(t) &= u_1(t), \qquad nT \leq t < (n+1)T \\ &= u_2(t), \qquad n = 0, 1, 2, \cdots \end{aligned}\right\} \tag{1}$$

where

$$\left.\begin{aligned} u_1(t) &= A\cos(2\pi f_1 t + \theta_n) \\ u_2(t) &= A\cos(2\pi f_2 t + \phi_n) \end{aligned}\right\} \tag{2}$$

We shall distinguish between two different cases for the $u_k(t), k = 1, 2$.

In both cases, the choice of the $u_k(t)$ is made independently and with equal probability for each interval of length T. In the first case, discontinuous phase frequency-shift-keying (FSK), the values of θ_n and ϕ_n are unconstrained from interval to interval. This corresponds to the case of switching between two independent oscillators. In the second case, continuous phase FSK, the initial values of the phase at $t = 0$ are $\phi_0 = \theta_0 = \phi$, and the succeeding values ϕ_n, θ_n are chosen so as to make the phase of $u(t)$ continuous at the transition points. This corresponds to the case of shifting the frequency of a single oscillator.

For each case, we shall specify the power spectra to be:

(1) One-sided, that is, specified completely in terms of positive frequencies.

(2) Approximations obtained by neglecting terms of the order of $1/(f + f_k)^2$ compared to terms which vary like $1/(f - f_k)^2$. The contributions of the neglected terms become appreciable only when the f_k are smaller than the signaling frequency, $f_s = 1/T$. For situations in which we shall be interested, this will never occur.

3. Discontinuous Phase FSK

For discontinuous phase FSK, we shall consider two subcases:

$$(f_2 - f_1) \neq mf_s \qquad (m \text{ an integer})$$
$$(f_2 + f_1) \neq mf_s$$
$$(f_2 - f_1) \neq \frac{(m+1)}{2} f_s$$
$$(f_2 + f_1) \neq \frac{(m+1)}{2} f_s$$

and

$$f_1 \text{ and } f_2 \text{ arbitrary}$$

In the first case from Eq. (76) of Ref. 1, we have the power spectrum given by

$$w_u(f) = (A^2/8\delta)(f - f_1) + (A^2/8\delta)(f - f_2) + \frac{\sin^2\left[\pi\left(\frac{f - f_1}{f_s}\right)\right]}{8\left[\pi\left(\frac{f - f_1}{f_s}\right)\right]^2} + \frac{\sin^2\left[\pi\left(\frac{f - f_2}{f_s}\right)\right]}{8\left[\pi\left(\frac{f - f_2}{f_s}\right)\right]^2} \tag{3}$$

We see from Eq. (3) that the spectrum consists of impulses at f_1 and f_2 with the familiar $(\sin^2 x)/x^2$ form centered about these impulses.

For the second case of discontinuous phase FSK, we have to use Eq. (15) of Ref. 1. This equation is much more complicated and involves the values of ϕ_n and θ_n explicitly. However, if we assume that ϕ_n and θ_n are independent random variables with uniform distributions over the interval $[0, 2\pi]$, we can average the power spectrum given in Eq. (15) of Ref. 1 over ϕ_n and θ_n. Doing this, we obtain exactly the same expression as given in Eq. (3). Thus, for our purposes, Eq. (3) completely specifies the power density spectrum for discontinuous phase FSK. If for any reason ϕ_n and θ_n take on particular values, this statement will no longer be true, and we will have to go back to Eq. (15) of Ref. 1.

4. Continuous Phase FSK

For continuous phase FSK, we shall also corsider two subcases:

$$f_2 - f_1 \neq mf_s \qquad (m \text{ an integer})$$

$$f_2 + f_1 \neq mf_s$$

$$f_2 - f_1 \neq \frac{(m+1)}{2} f_s$$

$$f_2 + f_1 \neq \frac{(m+1)}{2} f_s$$

and

$$f_2 - f_1 \approx mf_s \qquad (m \text{ an integer})$$

$$f_2 + f_1 \neq mf_s$$

$$f_2 - f_1 \neq \frac{(m+1)}{2} f_s$$

$$f_2 + f_1 \neq \frac{(m+1)}{2} f_s$$

Other cases are treated in Ref. 1, but these two should include most practical situations of interest. In the first case from Eq. (48) of Ref. 1, we have

$$w_u(f) = \frac{2A^2 \sin^2\left[\left(\frac{f-f_1}{f_s}\right)\pi\right] \sin^2\left[\left(\frac{f-f_2}{f_s}\right)\pi\right] \left[\frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right)} - \frac{1}{2\pi\left(\frac{f-f_2}{f_s}\right)}\right]^2}{f_s\left\{1 - 2\cos\left[\frac{2\pi f - \frac{1}{2}(2\pi f_2 + 2\pi f_1)}{f_s}\right]\cos\left[\left(\frac{f_2-f_1}{f_s}\right)\pi\right] + \cos^2\left[\left(\frac{f_2-f_1}{f_s}\right)\pi\right]\right\}} \tag{4}$$

Although the behavior of $w_u(f)$ is not as transparent as it was before, it appears that the spectrum has peaks in the vicinities of f_1 and f_2.

For the second case, in which $f_2 - f_1 = mf_s$, we use Eqs. (52), (53), and (54) of Ref. 1 to obtain

$$w_u(f) = w_v(f) + (A^2/8)\left[\delta(f-f_1) + \delta(f-f_2)\right] \tag{5}$$

where

$$w_v(f) = (A^2/2f_s)\sin^2\left[\pi\left(\frac{f-f_1}{f_s}\right) - \frac{m\pi}{2}\right]\left[\frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right)} - \frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right) - 2\pi m}\right]^2 \qquad \text{for } m \text{ even} \tag{6}$$

$$w_v(f) = (A^2/2f_s)\cos^2\left[\pi\left(\frac{f-f_1}{f_s}\right) - \frac{m\pi}{2}\right]\left[\frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right)} - \frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right) - 2\pi m}\right]^2 \qquad \text{for } m \text{ odd} \tag{7}$$

Since m is an integer, the above expressions can be simplified:

$$w_v(f) = (A^2/2f_s)\sin^2\left[\pi\left(\frac{f-f_1}{f_s}\right)\right]\left[\frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right)} - \frac{1}{2\pi\left(\frac{f-f_1}{f_s}\right) - 2\pi m}\right]^2 \qquad \text{for all } m \tag{8}$$

Here we see once again that impulses at f_1 and f_2 are combined with a continuous spectrum.

5. General Observations

One of the most important characteristics to consider is the behavior of the spectra as a function of frequency. By studying Eqs. (3), (4), and (8), we conclude:

(1) The spectra for discontinuous phase FSK fall off as $1/f^2$ for large f.

(2) The spectra for continuous phase FSK fall off as $1/f^4$ for large f.

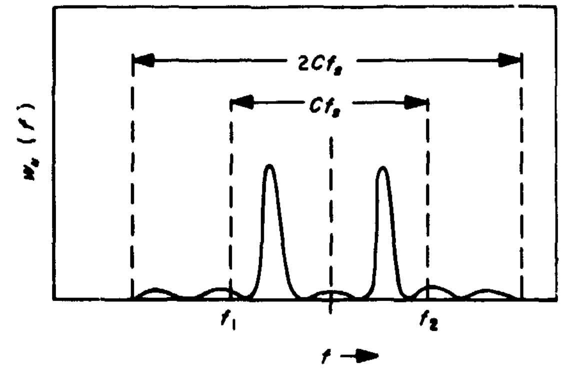
This observation is important in designing practical systems. For example, in systems with small relative difference frequency

$$C = \frac{f_2 - f_1}{f_s} \tag{9}$$

we would, ideally, like to use a continuous phase oscillator to minimize overlap from each of the two frequencies. Practically, it will be a question of how much the frequency of a real oscillator can be pulled.

Another way to compare the behavior of the spectra versus frequency is to calculate the percentage of total power in an arbitrary bandwidth. A convenient bandwidth to consider is $2Cf_s$, for which we calculate $P_n \equiv$ percentage of total power within a bandwidth $2Cf_s$ centered about $(f_1 + f_2)/2$. Figure 4a shows the bandwidth defined above, and Fig. 4b gives representative values of P_n for various C. Included for comparison are values for a phase-shift-keyed (PSK) spectrum centered about $(f_1 + f_2)/2$. As can be seen from the figure, continuous phase FSK is by far the most efficient in terms of having the most power in the smallest bandwidth.

(a)



(b)

C	P_n, PSK	P_n, FSK DISCONTINUOUS PHASE	P_n, FSK CONTINUOUS PHASE
1	0.903	0.926	0.984
2	0.95	0.967	0.995
3	0.966	0.977	0.996
4	0.975		
5	0.98		

Fig. 4. (a) Bandwidth $2Cf_s$ (b) Percentage of total power P_n in bandwidth Cf_s

Another interesting characteristic is the shape of the spectra as a function of C. It is clear from Eq. (3) that the spectrum for discontinuous phase FSK is just the properly separated sum of the impulses and the $(\sin^2 x)/x^2$ terms. Thus for large C, when the overlap between the two terms is negligible, the shape of the spectrum in the region of f_1 and f_2 is a constant independent of C. The same sort of behavior for the continuous phase spectra can be deduced from Eqs. (4), (5), and (8). Substituting Eq. (9) into Eq. (4) and simplifying, we obtain.

$$\frac{w_u(f)f_s}{A^2} = \frac{\sin^2\left[\left(\frac{f-f_1}{f_s}\right)\pi\right]\sin^2\left[\left(\frac{f-f_1}{f_s}\right)\pi - C\pi\right]}{2\left\{1 - 2\cos\left[2\pi\left(\frac{f-f_1}{f_s}\right) - \pi C\right]\cos\pi C + \cos^2\pi C\right\}}\left\{\frac{\pi C}{\pi\left(\frac{f-f_1}{f_s}\right)\left[\pi\left(\frac{f-f_1}{f_s}\right) - \pi C\right]}\right\}^2 \tag{10}$$

It is convenient to consider the spectrum as a function of the normalized variable

$$x = \frac{(f - f_1)}{f_s} \tag{11}$$

so that we have

$$\frac{w_u(f)f_s}{f_s} = \frac{\sin^2(\pi x)\sin^2(\pi x - \pi C)}{2\{1 - 2\cos(2\pi x - \pi C)\cos(\pi C) + \cos^2(\pi C)\}}\left[\frac{\pi C}{\pi x(\pi x - \pi C)}\right]^2 \tag{12}$$

a function which is symmetrical about $C/2$. If we let

$$x = C + \Delta, \qquad |\Delta| < 1 \tag{13}$$

and expand all the trigonometric identities, we obtain

$$\frac{w_u(f)\,f_s}{A^2} = \frac{[\sin \pi C \cos \pi\Delta + \cos \pi C \sin \pi\Delta]^2 \sin^2(\pi\Delta)}{2\{1 - 2\cos(2\pi\Delta)\cos^2(\pi C) + 2\sin(2\pi\Delta)\sin(\pi C)\cos(\pi C) + \cos^2(\pi C)\}} \left[\frac{\pi C}{\pi(C+\Delta)\pi\Delta}\right]^2 \tag{14}$$

Now if $C >> 1$, the C over $(C + \Delta)$ in the last bracket will cancel, and we will have an expression which depends only on Δ and periodic functions of C with period 1. What this means practically is that in the region of interest around f_1 and f_2 ($x = 0$ and $x = C$) the spectrum for e.g., $C = 15.2$ is approximately the same as the spectrum for $C = 16.2$, with a diffeient separation. Some of the characteristic shapes will be identified in *Subsection 6*. Using similar reasoning, we also arrive at the same conclusion for Eq. (8).

The same sort of arguments also show that the following properties of the spectra *in the region of* f_1 and f_2 ($x = 0$ and $x = C$) hold:

(1) The spectrum for continuous phase FSK with $f_2 - f_1 = mf_s$ is approximately equal to that for discontinuous phase FSK for large C.

(2) For large C, there is a rotational symmetry about $x = 0$ and $x = C_0$ for the continuous phase spectra for $C = C_0 + \beta$ and $C = C_0 - \beta$, where $0 < \beta < 1$.

The practical implication of the second statement is that we can determine the shape of the spectrum for $C = 15.2$ by looking at the spectrum for $C = 14.8$ and rotating that portion of it in the vicinity of $x = 15$ (or $x = 0$) about the point $x = 15$ (or $x = 0$). Further explanations of this symmetry property of the examples are given in *Subsection 6*.

To summarize, we list the symmetry properties which we have outlined:

(1) Behavior versus frequency: continuous phase FSK falls off as $1/f^4$ and discontinuous phase FSK falls off as $1/f^2$.

(2) Shape of spectra in region of f_1 and f_2 ($x = 0$ and $x = C$) as function of C for large C:

(a) Discontinuous phase, shape is same for all C.

(b) Continuous phase, shape for C is approximately the same as for $C + 1$.

(c) Spectrum for continuous phase with

$$f_2 - f_1 = mf_s$$

is approximately equal to spectrum for discontinuous phase.

(d) Rotational symmetry about $x = 0$ and $x = C_0$ for $C = C_0 + \beta$ and $C = C_0 - \beta$.

The first property is the most fundamental; the others are pointed out to give a better insight to the behavior of the spectra.

6. Plots

Figures 5a to 5k show plots of the spectra for various values of C. In each case we have plotted only the continuous portion of the spectrum as a function of the normalized variable $x = (f - f_1)/f_s$. The plots shown are symmetrical about the point $x = C/2$; $x = 0$ corresponds to f_1, and $x = C$ corresponds to f_2. For purposes of comparison, we have included plots of the PSK spectrum centered on $x = C/2$. In each figure the numbered curves correspond to the following functions:

Curve 0—PSK

$$\frac{w_u(f)\,f_s}{A^2} = \frac{\sin^2(\pi x - \pi C/2)}{2(\pi x - \pi C/2)^2} \tag{15}$$

Curve 1—Discontinuous phase FSK

$$\frac{w_u(f)\,f_s}{A^2} = \frac{1}{8}\left[\frac{\sin^2(\pi x)}{(\pi x)^2} + \frac{\sin^2(\pi x - \pi C)}{(\pi x - \pi C)^2}\right] \tag{16}$$

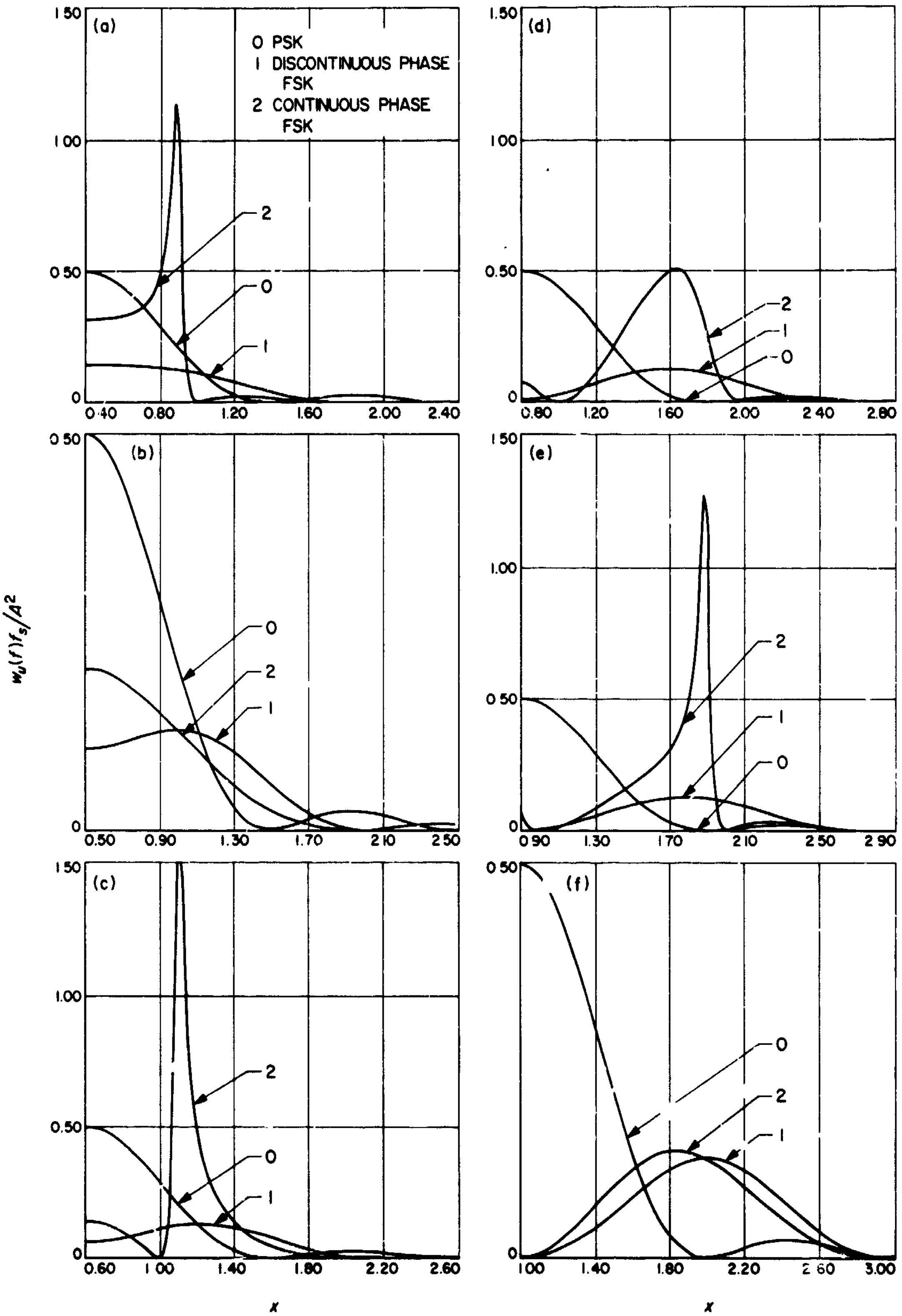


Fig. 5. Power spectra (a) for C = 0.8, (b) for C = 1.0, (c) for C = 1.2, (d) for C = 1.6, (e) for C = 1.8, (f) for C = 2

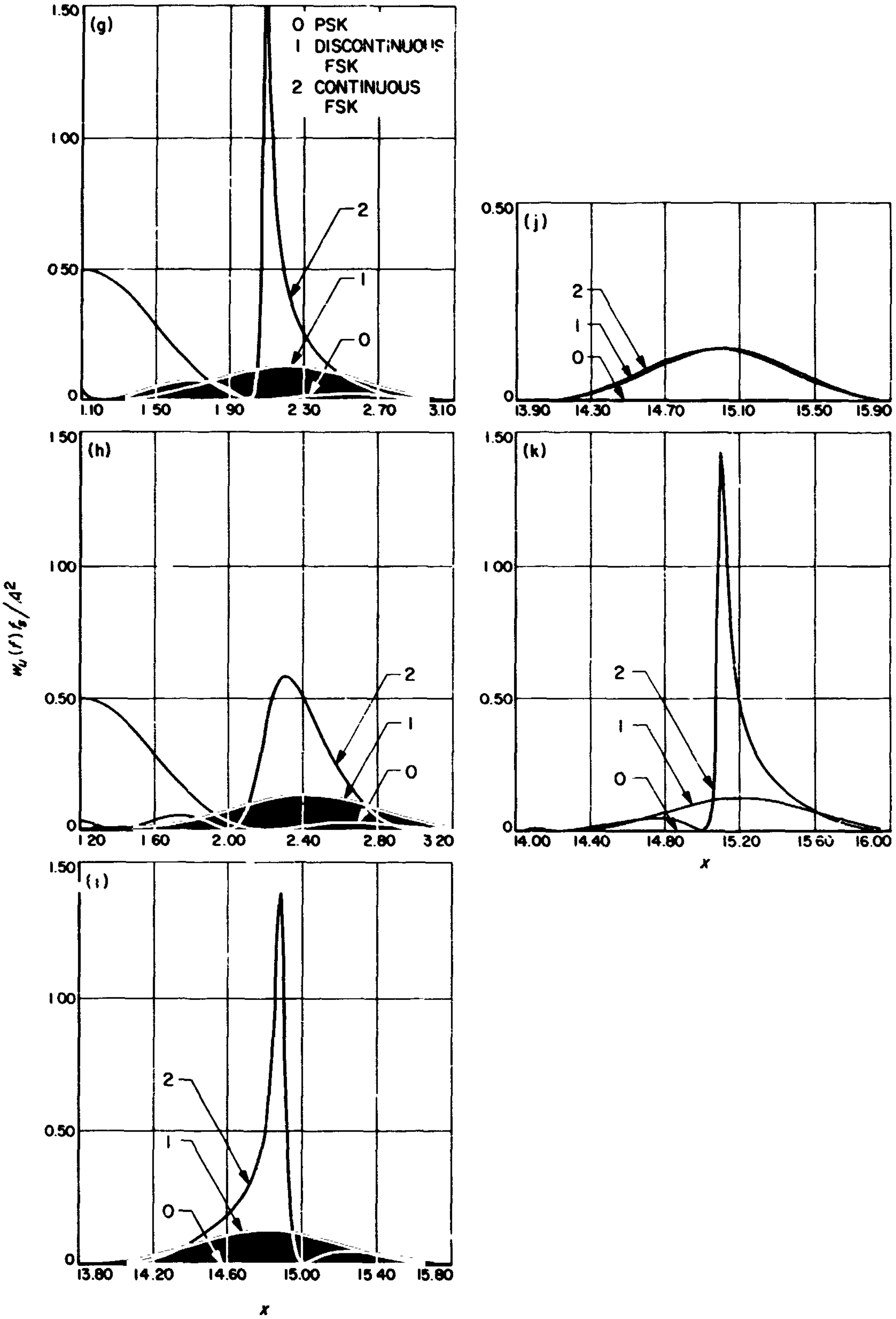


Fig. 5 (contd). Power spectra (g) for C = 2.2, (h) for C = 2.4, (i) for C = 14.8, (j) for C = 15, (k) for C = 15.2

Curve 2—Continuous phase FSK

$$\frac{w_u(f)f_s}{A^2} = \frac{\sin^2(\pi x)}{2}\left[\frac{1}{2\pi x} - \frac{1}{2\pi x - 2\pi m}\right]^2 \quad (17)$$

$$\text{for } C = m = \text{integer}$$

$$\frac{w_u(f)f_s}{A^2} = \frac{\sin^2(\pi x)\sin^2(\pi x - \pi C)}{2\{1 - 2\cos(2\pi x - \pi C)\cos\pi C + \cos^2\pi C\}} \times \left[\frac{\pi C}{\pi x(\pi x - \pi C)}\right]^2 \quad (18)$$

$$\text{for } C \neq \text{integer}$$

The total transmitted power is the same in each case. The y-axis is the value of $w_u(f)f_s/A^2$ and x-axis is the value of x. Since we have only plotted the continuous portions of the spectra, impulses must be added to the curves corresponding to Eqs. (16) and (17).

Figures 5d to 5g in particular illustrate the effect on the continuous phase spectrum of changing C. For C less than 2 but greater than 1.5 we have a peak inside the point $x = 2$. As C approaches 2, the peak becomes more pronounced and moves closer to the point $x = 2$, until we obtain an impulse for $C = 2$. For C greater than 2, but less than 2.5, we observe a similar behavior, except that the peak is outside the point $x = 2$. A similar behavior can be expected as C varies through any integer value.

The general properties discussed in *Subsection* 5 should be evident from the plots, particularly Figs. 5i and 5j. The PSK spectrum has decayed to a negligible level in these figures, and the symmetry relations discussed are clear.

Another point of interest is that the spectral peaks for continuous phase FSK become less pronounced as C varies away from integer values. For example, compare Fig. 5d with Fig. 5f. To obtain sharp spectral peaks, $f_2 - f_1$ must be approximately an integer value. This property may be important in system design.

Reference

1. Bennett, W. R., and Rice, S. O., "Spectral Density and Autocorrelation Functions Associated With Binary Frequency-Shift Keying," *Bell System Technical Journal*, pp. 2355–2385, Sept. 1963.

XXIV. Future Projects

ADVANCED STUDIES

A. Science Utility of Automated Roving Vehicles,

R. G. Brereton

1. Introduction

The geology of the earth has been synthesized from a prodigious amount of data that was contributed from many observations and scientific disciplines and acquired over several decades. There is every reason to suppose that knowledge about lunar geology (i.e., knowledge of the structure and processes of the lunar interior, the composition, structure, and processes of the lunar surface, and the history of the moon) will be unfolded in the same way. Although working hypotheses have matured through experience and the terrestrial sphere is available as an accessible geological example, the true picture of lunar geology can only be formed from much new data that will have to be acquired from a wide range of surface location, structures, and physiographic provinces. The very nature of the lunar exploration task suggests that a surface mobility system will be required to acquire the needed data. Several types of designs for this mobility system have already been proposed.

Previous studies have indicated that a separable rover, delivered as payload by a *Surveyor* and hence limited to a total mass of 100 to 200 lb, could be useful in local surveys; however, it is recognized that such small vehicles, with a payload capability of about 20 lb, would have only marginal utility for most roving missions. At the same time, a reasonable upper limit on size for an automated rover would seem to be that of the local scientific survey module, whose mass is more than 1000 lb and whose size is compatible with a *Saturn* V launch. Between these lower and upper size–mass limits, there is probably a feasible vehicle design that can perform the required roving vehicle mission, while still being small and light enough to be delivered as an integral package by *Centaur*, alternately as payload aboard a single launch, manned *Apollo* mission, or as a separable payload item aboard an unmanned soft-landed vehicle intermediate in size between *Centaur* and *Saturn* V.

In the past, discussions and designs for roving vehicle systems have been constrained by specifying the size, weight, power, etc., of the roving vehicle to fit it into a particular launch vehicle, or the vehicle design has been constrained by a specified program or operating mode. These constraints, however justified, have tended to limit considerations regarding full scientific utility of automated roving vehicles.

The scientific instruments carried on a lunar roving vehicle will vary with the function of the programmed scientific task, although one basic vehicle design will probably suffice for all tasks provided reasonable range and mobility requirements are satisfied. This basic design must incorporate an imaging system and a navigation system that can perform both the vehicle guidance and navigation function, and also support the task or science function. It is expected that the vehicle will travel slowly over the surface (at the rate of a few kilometers per hour at most) and will be long-lived. Lifetime of the vehicle will, of course, be a function of the particular science task that the rover is programmed for; but, in general, the science requirements call for a vehicle lifetime measured in months to perhaps years. This would suggest that the prime source of power be nuclear, solar, or a combination of these. Telecommunications are not critical for operation anywhere on the front face of the moon; however, for backside operation, an orbiter relay link would be required.

There are four basic science tasks or separate missions that an automated roving vehicle can be useful for in a program of lunar exploration. Vehicles utilized in this way can be expected to provide significant new data about the moon that may not be available through other cost-comparable techniques. Each of these tasks has its place in the overall lunar exploration program; any plan that defines the most feasible and economical lunar exploration program must consider a mix of these roving vehicle tasks with other lunar missions, both manned and unmanned.

2. Imaging System

The automated roving vehicle will require an imaging system for purposes of navigation and guidance and for terrain assessment. The system should have stereometric, polarimetric, and colorimetric capabilities and possibly telescopic lens combinations to allow a close look at features with minimum amount of vehicle travel and shuffling. The specific objectives of the imaging system on the rover are:

(1) Provide near-real-time images that can be used to guide the roving vehicle.

(2) Acquire dimensionally stable images from which topographic maps can be made by photogrammetric methods.

(3) Provide reconnaissance-eye-type geological information in color.

(4) Provide near-field information on surface structure and texture, with the capability to detect particle sizes down to at least 0.5 mm.

A variety of sensors and camera systems could perhaps be adapted to the roving vehicle mission; however, the selected system should meet the following requirements that are believed to be essential to the objectives of the roving vehicle mission.

(1) A stereographic baseline of the camera system of preferably 3 ft but no less than 1 ft. The baseline may be vertical or horizontal.

(2) A measurement of the local vertical to 0.5 deg at each position of the roving vehicle from which an image is obtained.

3. Science Tasks

a. Sample acquisition. The *Apollo* sample return experiment is recognized as one of the most important in the entire lunar program, since it affords the opportunity for elaborate earth-based investigation of the isotopic composition, chemistry, mineralogy, and physical state of the lunar surface material. To extend this experiment beyond the *Apollo* landing locations appears highly desirable. Therefore, a possible mission for a small rover is the collection of samples along an extended (up to hundreds of kilometers) traverse, followed by the delivery of the samples to a collection point where they would be returned to earth, presumably by an *Apollo* spacecraft. The traverse could be either from one *Apollo* landing point to another or from an unmanned vehicle landing point to an *Apollo* site. It is assumed that any special packaging requirement for samples to be returned to earth could be accomplished by the astronaut at rendezvous. It has been suggested (Ref. 1) that the automated rover be capable of traverses up to 500 km with at least 100 stations for observation and sample collection, and be capable of carrying 25 kg of samples collected and individually packaged in 100- to 250-g containers. It would appear that the real limitation for this mission is not the weight of samples that can be conveniently carried by the rover to the rendezvous point or transported by *Apollo* back to earth, but rather the time required to acquire meaningful samples. It does not seem that random sampling along a profile is the most desirable technique; however, it may turn out to be the most practical one. Samples should be acquired from select locations (outcrops, etc.); this will require considerable observer effort and time, and much stop and go maneuvering for the automated roving vehicle.

The minimum scientific instrumentation for this type of mission would include an imaging and navigation device plus techniques for acquiring lunar samples. To perform the latter task, two separate sampling modes are desirable—one for hard rock material, and another for sifting the particulate material that appears to form much of the lunar surface. Nash (Ref. 2) has given an excellent discussion of the strategy, principles, and instrument requirements for sampling planetary surfaces.

The imaging device would perform several functions. During traverse, it would observe the general lay-of-the-land, its structure, stratification, and topographic form, and color changes and rock textures down to at least 0.5 mm; therefore, it would indicate interesting areas for sampling. The device would also be used to locate sample stations with respect to identifiable lunar surface features to within 100 m on base maps or orbiter photographs.

It would be desirable to equip this type of rover with a device for elemental chemical analysis that could be used in a reconnaissance mode, and, in conjunction with the imaging device, for selecting meaningful samples. A number of lightweight instruments using techniques such as alpha scattering, neutron activation, and nondispersive X-ray emission spectroscopy seems to be suitable for this operation.

Table 1 presents some information about a typical science payload for an automated rover designed for the sample acquisition task.

Table 1. Science instruments for sample acquisition rover

Imaging system **(8 lb; 2 W)**
This instrument would provide images for guidance and positioning of the rover and for sample selection. Stereo, color, and resolution to at least 1 mm is desirable.
Elemental analysis **(8 lb; 4 W, during operation)**
The ins'rument (nondispensive X-ray emission spectroscopy) is formed from a radioactive excitation source, a gas fill-d proportional counter for detecting a signal, an amplifier and deployment mechanism. In operation, the instrument excitation source and sensor must be deployed to the lunar surface.
Particulate sampler **(5 lb; 2 W)**
The suggested instrument is a so-called rigid helical conveyor with drill tip. It would be capable of sampling the typical lunar soil to a depth of perhaps 5 in. It size-sorts particles so as to diminish the content of those over 500 μm and reject those over 1000 μm. Device would have two functions—acquire samples and distribute them to sample containers.
Hard rock drill **(10 lb; 25 W)**
This is a rotary impact drill capable of sampling rock material as hard as dense basalt. The instrument has a depth capability of about 1 ft. Device would have two functions—acquire samples from hard rock and distribute them to sample containers.
Sample container **(50 lb, full; 2 W)**
Desire about 100 sample containers for 0.25- to 0.5-lb samples.

b. In situ analysis. One of the most obvious and perhaps more important roles of the automated roving vehicle in the lunar program will be geological reconnaissance, or the ability to extend the local measurements of *Surveyor* or *Apollo* into the surrounding area. Only a very small area of the moon is expected to be explored by manned missions of the near future; therefore, a properly instrumented automated rover capable of probing the environs of the moon out from *Apollo* sites should have an important mission in a lunar exploration program. Although an automated rover, however instrumented, can never be expected to replace the on-site geologist, a properly instrumented rover can be expected to provide: (1) survey type data on the geochemistry of the moon to include information about the kind, origin, and distribution of lunar rocks and minerals; and (2) reconnaissance imagery bearing on lunar physiography, surface structures and stratigraphy. These data will contribute to the understanding of the moon and indicate areas of high interest for planning future missions. This type of rover mission can serve a useful scientific purpose in both regional and local studies.

It should be realized that a chemical basis alone is incapable of classifying the many diverse products of rock-forming processes. Thus, chemical elemental analysis experiments will not distinguish crystalline rock from volcanic glass or ash with the same chemical composition, nor a physical mixture of local debris from a crystalline rock. The accepted schemes of rock classification are based on texture (the size, shape, and geometrical relation of grains in a rock) and the identification of the minerals in the rock. From these parameters, information regarding the nature, geologic history, and origin of the rock may be defined.

On the basis of the above, the scientific instrumentation for this type of rover mission should include: (1) an imaging device, (2) an array of geochemical instruments, and (3) a sample acquisition preparation device. The imaging device would perform the same function as on the sample acquisition task.

The sample acquisition preparation device would be the same as described for the previous task. Samples of lunar surface material would be obtained by the particulate sampler or hard rock drill, and this material would be distributed to the geochemical instruments.

As a minimum set, the array of geochemical instruments must include methods for elemental analysis, phase analysis, and study of rock textures. The suggested instruments here are an X-ray spectrometer (Ref. 3) for elemental analysis, an X-ray diffractometer (Ref. 4) for mineral phase determination, and a petrographic microscope (Ref. 5) that could observe crushed rock samples in transmitted light. These instruments were previously considered for both *Surveyor* and rover missions. Table 2 presents a typical science payload for an automated rover designed for the *in situ* analysis task.

In addition to the above instruments, it may be desirable to include a gas chromatograph in the payload for this vehicle. The chromatograph would provide an analysis of the volatile constituents in lunar surface material.

Table 2. Science instruments for *in situ* analysis rover

Imaging system (8 lb; 2 W)
This instrument would provide images for guidance and positioning of the rover and also for geological eyeball type information. Stereo, color, and system resolution to at least 0.5 mm is desirable.
X-ray diffractometer (15 lb; 4 W)
The X-ray diffractometer will be used to conduct mineralogical analyses of lunar surface material acquired at a number of fixed points on a roving vehicle traverse. The primary objective of this instrument is to identify the types and relative abundance of the various crystalline phases expected to be present in a lunar sample. The instrument will provide diffraction data of sufficient quality to identify any of the major rock-forming and accessory minerals.
X-ray spectrometer (15 lb; 4 W)
The X-ray spectrometer will be used to conduct an elemental analysis of lunar surface material acquired at a number of fixed points on a roving vehicle traverse. The mode of analysis can detect elements from sodium through uranium; however, only those elements from sodium through nickel are expected to be present in sufficient quantity to allow detection.
Petrographic microscope (15 lb; 4 W)
The petrographic microscope would provide textural and optical information on rocks and particulate material from the lunar surface.
Particulate sampler (5 lb; 2 W)
The suggested instrument is a so-called rigid helical conveyor with drill tip. It would be capable of sampling the typical lunar soil to a depth of perhaps 5 in. It size-sorts particles so as to diminish the content of those over 500 μm and reject those over 1000 μm. Device would have two functions—acquire samples and distribute them to the geochemical instruments above.
Hard rock drill (10 lb; ~5 W)
This is a rotary impact drill capable of sampling rock material as hard as dense basalt. The instrument has a depth capability of about 1 ft. Device would have two functions—acquire samples from hard rock and distribute them to the geochemical instruments above.

c. Traverse geophysics. Traverse geophysics has a very special place in the lunar exploration program. It can provide data toward the solution of problems that can be solved only by the combined techniques of surface mobility and geophysical instrumentation. Traverse geophysics using automated roving vehicles is not a panacea for all the problems of lunar exploration; however, it is a powerful tool for providing data on the subsurface of the moon and when these data are correlated with lunar geology and multiple working hypotheses, they can provide an informative picture of the possible structure and processes of the lunar crust. The choice of scientific instruments for traverse task is quite large, because geophysical techniques and instrumentation have become more diversified through the effect of space age technology and the revolutionary growth of science that has taken place since 1940. For example, 10 years ago, a magnetic survey was usually accomplished with a field balance magnetometer,[1] which measured only one component of the earth's magnetic field. Thus, a magnetic survey to measure the magnitude of the geomagnetic field vector required two separate survey operations with two separate magnetometers (one survey and instrument to measure the horizontal component and another to measure the vertical component). Today, this same operation can be carried out with one small and completely

[1] Designed by A. Schmidt; manufactured by Askania Werke, Berlin.

portable instrument called a proton procession magnetometer at a fraction of the time and at perhaps greater accuracy. Space age technology has similarly affected seismic, electrical, radioactive, and gravity instruments and their application.

Although new technology has affected geophysical instrument design and its application, particularly in the sense that it makes the roving vehicle traverse geophysics mission feasible, classical geophysical experiments in magnetism, gravity, and seismic prospecting appear to be most practical for the early traverse missions, as their data are more understood and interpretable in terms of terrestrial analogs. An imaging system and laser ranging experiment should also be a part of the minimum science package for the traverse geophysics task.

The imaging system as previously described would be suitable for the traverse geophysics task. This instrument would serve as the eyes of the rover for navigation, guidance, and positioning and, in addition, support the geophysical experiments by providing eyeball type geological information at each measurement site.

It has been suggested that the present absence of a strong internal magnetic field for the moon may reduce the effectiveness of standard magnetic surveying techniques for understanding deep structural features; however, the absence of this field may now enhance the detection of remnant magnetism that could have considerable cosmogonic significance. Also, because the difference in measured susceptibility between acid and basic rocks, between nickel–iron meteorites and silicate rocks, and even between chondrites and silicate rocks is large, it is probable that they have become polarized by external fields, relic lunar field, or flowage. Therefore, magnetic survey techniques may prove to be a valuable tool for mapping contacts, providing criteria for distinguishing impact and volcanic features, and, in general, providing new data on the structure and processes of the lunar surface.

The traverse operation will require a three-component orthogonal magnetometer of the flux-gate, proton procession, or optical pumping type. The last two types are favored because they provide absolute magnitude data. It is desirable that the magnetometer operate continuously; i.e., operate both during station stops and while the rover is in traverse An accuracy of ±5 γ is desirable. This suggests that the magnetometer sensor must be compensated for both perm and induced magnetic interference from the roving vehicle, or else removed from its vicinity during measurements. A base control for monitoring external fluctuations and changes in the lunar magnetic field is required. This could be provided by *Apollo* lunar surface experiments package (ALSEP) science or an emplaced science station (ESS) package containing a magnetometer. The base control station should be located in the survey or traverse area, but a separation up to 500 km could be tolerated.

The surface gravity of the moon is only one-sixth that of the earth; therefore, gravity anomalies on the moon resulting from a density contract in lunar material will comprise a larger part of the total-field measurement than similar measurements on earth. Lunar gravity anomalies may be caused by local near-surface density contracts in rock units, as between the regolith and basement rocks, or perhaps by regional isostatic phenomena where the moon's crust is out of isostatic equilibrium because of ancient frozen tidal effects or crustal overloading by ejecta from large meteor impacts. The Carpathians and Apennines are possible examples of crustal overloading from the Imbrium impact event. Surface gravity data from profiles across virtually all lunar structures and contacts are desirable as these data may be critical to an understanding of the origin and evolution of these features and even the moon itself.

The best gravity instrument for the traverse task is probably a conventional spring–mass gravimeter in contrast to a torsion balance or pendulum. The state-of-the-art in design of these instruments is highly advanced. Terrestrial gravimeters are required to detect changes in gravity of the order of 1 part/10^7 The lunar instrument, because of the lower gravity on the moon and the resulting higher ratio between anomaly and total gravity, should be calibrated to detect changes of the order of 1 part/10^6 over a dynamic change of 500 mgals. Considerable care in instrument design to control long term temperature and mechanical drift will be required, since it is unlikely that the vehicle on a traverse mission can be returned to a previous station to measure instrument drift.

It will be necessary to make both free-air and Bouguer corrections to the gravity observations. These corrections, if not observed, could mask regional trends and even local anomalies. An integrating tiltmeter, supplemented with data from base map and imaging system, can provide the information for this correction. A terrain correction may be needed locally; a tidal correction to account for differential alignment between the sun, earth, and moon will also be required.

The active seismic experiment, as defined here, can be considered a ?hallow exploration technique for probing the uppermost ' ilometer of the moon, and designed to measure the depth of the lunar regolith and its elastic and phyical properties, seismic wave velocities, and rock contacts, and, in general, to provide subsurface data on the moon's structure and stratification. Both refraction and reflection techniques can be useful, for one phenomenon rarely occurs without the other. The geological picture that is emerging for the lunar surface suggests a low density regolith overlying a denser unchurned basement. The study of this may present an ideal problem for the seismograph.

The experiment consists of a seismometer which can be deployed to the lunar surface under the rcver, an auger for shot-hole preparation, and approximately 100 charges weighing 0.25 lb each for providing a seismic energy source. The charges would be activated separately by radio command from the rover. In terrestrial seismic prospecting, where mobility and backtracking are not problems, the normal prccedure is to use one shot point in conjunction with six or so detector–seismometers. Energy arrival times for each detector from the one source are plotted as a function of time and distance to portray refraction, reflection, and subsurface structure. On the moon, where backtracking and locating emplaced detector–seismometers cculd be difficult, a suggested procedure is to emplace a series of, say, five charges on a rover traverse and then, from a selected position with the seismometer on, each charge would be remotely detonated in turn. A few seconds recovery time must be allowed between detonations to avoid record wipe-out of subsequent events by the first. A shot-point spread of about 50 ft would be used at first to carefully define the average velocity in the regolith, or for precision probing for a suspected contact or structure; however, operating procedures would evolve through experience. Once the velocity function has been determined for the regolith or a particulai marker bed, its depth at the other locations and during the traverse could possibly be determined from one shot point and the reflection record. Time anomalies between the regolith and basement rocks could be very high, but between the ejecta blanket from one event and underlying ash, or say earlier ejecta, the time anomalies might be small. Therefore, timing accuracy should be controlled to at least 0.005 s and shot-point distance measured to 2%.

Instrument design for the seismometer and radio-controlled charge are not problems; however, a design for the auger and a technique for charge emplacement need further study.

The laser ranging experiment on the traverse mission would define a series of fixed points over the lunar surface. This type of information has a direct application to problems in selenodesy, lunar mapping, and celestial mechanics, and, in addition, the experiment will provide information to assist in the free-air and Bouguer corrections for the gravity experiment. The definition of a geodetic arc for various radii of the moon will define the size and shape of the moon and locate frozen tidal effects. The equipment for the laser experiment would consist of a transmitter at an appropriate terrestrial observatory and a reflecting antenna and its points and orienting mechanism on the rover.

Table 3 presents a typical science payload for an automated rover designed for the traverse geophysics task.

d. Special tasks. In addition to the tasks that have already been discussed, there are several others that can be logically programmed and considered feasible for an automated roving vehicle.

Instrument deployment. There are areas on the moon that are too rough, or inaccessible, for landing vehicles. It may be desirable to place scientific instrumentation there as part of an ALSEP or ESS net.

Deep seismic. A rover in conjunction with a fixed seismic station, as could be provided by ALSEP or ESS, presents a technique for long profiling and deep seismic probing. The rover, equipped with an auger for shot-hole preparation and a large number of charges, would seismic-profile out from the fixed seismic station to the limit of reception and then make additional profiles in and out from the station as dictated by the structure under investigation or by technology. The individual charges would be detonated by radio command from rover. The size of the charge will be a function of the seismic noise on the moon, the elastic properties of the regolith and deeper rock layers, and also the depth of the shot point.

Exploration. The rover as considered here would be instrumented to explore for strategic materials in the lunar crust. Such an operation would, of course, be carried out in close cooperation with other types of lunar missions. The particular exploration task will determine the best scientific payload for each mission. Water search is an example of this type of mission, where the rover would be sent into a promising area selected from physiographic and sound geological principles. This type of operation may well be very demanding of the rover

Table 3. Science instruments for traverse geophysics rover

Imaging system **(8 lb; 2 W)**
This instrument would provide images for guidance, navigation, and positioning of rover and, in addition, it would support the geophysical experiments by providing eyeball type geological information at each measurement site. Stereo, color, and system resolution to at least 0.5 mm is desirable.
Magnetometer **(6 lb; 1 W)**
This instrument on the rover would provide a profile of the magnetic intensity over various types of lunar structures and features. Instrument should have an inherent accuracy of ±5 γ. External control by ALSEP or other station magnetometer needed.
Gravimeter **(15 lb; 2 W)**
This experiment will provide information on density differences and the isostatic equilibrium of the moon. Will be operated in a profile mode at fixed stations. Station separation will be determined from structural considerations and operating experience.
Seismometer **(40 lb; 1 W)**
This experiment will determine the depth of the lunar regolith and its elastic and physical properties, seismic wave velocities, and rock contacts, and, in general, provide subsurface data on the moon's structure and stratification. At least 100 separate seismic events are programmed in 0.25-lb charges. Timing accuracy to 0.005 s and shot-point distance measurement to 2% are required.
Auger **(10 lb; 5 W)**
Will prepare a 12-in.-deep, 3-in.-diam hole for the seismic charges.
Laser and radio ranging experiment **(10 lb; 2 W)**
The objective of this experiment is to determine the location and altitude of a roving vehicle gravity site by both a laser ranging instrument and radio ranging technique.

mobility system, since rough surface operation on volcanic, rille, certain types of ejecta blanket, and other terrains will be required.

Site certification. Later *Apollo* missions, landing in potentially hazardous regions, may call for advanced certification of sites, or survey information may be needed on a potential site for a lunar base or observatory. A properly instrumented rover could provide these data extending over a much greater area than is possible with fixed-point vehicles, such as *Surveyor,* with higher resolution viewing than can be reasonably accomplished by orbiters, and with the main advantage of many tactile measurements over the entire area of the landing site.

Science missions. Included in this category would be tasks in which rover moved a baseline for ranging or geodetic measurements, provided a platform for scientific study of lunar degassing at points of observed emanating gases, conducted variable baseline electromagnetic studies, or provided an automated technique for shallow drilling and down-hole *in situ* geophysics including heat flow studies. The list of possibilities is obviously quite large.

4. Conclusions

The automated lunar roving vehicle presents a versatile technique with wide scientific utility for a program of lunar exploration. Some of the scientific tasks that must be done on the moon are unique to the rover capability; i.e., unique in the sense that a rover presents not only the most practical way to do the tasks, but perhaps presents the only way they will be likely to get done in the foreseeable future.

With the exception of the instrument deployment task, where the rover may be required to deliver an ALSEP or an ESS package, or perhaps in the deep seismic probing task where heavy seismic-source charges could be desirable, the total weight of the science subsystem is less than 100 lb.

References

1. *Lunar Science and Exploration,* 1967 Summer Study, NASA SP-157, p. 228. National Aeronautics and Space Administration, Washington, 1967.
2. Nash, D. B., *Sampling of Planetary Surface Solids for Unmanned In Situ Geological and Biological Analysis: Strategy, Principles, and Instrument,* Technical Report 32-1225. Jet Propulsion Laboratory, Pasadena, Calif., Nov. 15, 1967.
3. Metzger, A. E., *An X-Ray Spectrograph for Lunar Surface Analysis,* Technical Report 32-669. Jet Propulsion Laboratory, Pasadena, Calif., Oct. 16, 1964.
4. Gupta, K. D., et al. (1966), A Combined Focusing X-Ray Diffractometer and Nondispersive X-Ray Spectrometer for Lunar and Planetary Analysis, Advanced in X-Ray Analysis Vol. 9, Plenum Press.
5. Loomis, A. A., *A Lunar and Planetary Petrography Experiment,* Technical Report 32-785. Jet Propulsion Laboratory, Pasadena, Calif., July 15, 1965.