# The Telecommunications and Data Acquisition Progress Report 42-107

## July–September 1991

E. C. Posner
Editor

November 15, 1991

National Aeronautics and
Space Administration

**Jet Propulsion Laboratory**
California Institute of Technology
Pasadena, California

The research described in this publication was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

# Preface

This quarterly publication provides archival reports on developments in programs managed by JPL's Office of Telecommunications and Data Acquisition (TDA). In space communications, radio navigation, radio science, and ground-based radio and radar astronomy, it reports on activities of the Deep Space Network (DSN) in planning, in supporting research and technology, in implementation, and in operations. Also included is standards activity at JPL for space data and information systems and reimbursable DSN work performed for other space agencies through NASA. The preceding work is all performed for NASA's Office of Space Operations (OSO). The TDA Office also peforms work funded by two other NASA program offices through and with the cooperation of the Office of Space Operations. These are the Orbital Debris Radar Program (with the Office of Space Station) and 21st Century Communication Studies (with the Office of Aeronautics and Exploration Technology).

In the search for extraterrestrial intelligence (SETI), *The TDA Progress Report* reports on implementation and operations for searching the microwave spectrum. In solar system radar, it reports on the uses of the Goldstone Solar System Radar for scientific exploration of the planets, their rings and satellites, asteroids, and comets. In radio astronomy, the areas of support include spectroscopy, very long baseline interferometry, and astrometry. These three programs are performed for NASA's Office of Space Science and Applications (OSSA), with the Office of Space Operations funding DSN operational support.

Finally, tasks funded under the JPL Director's Discretionary Fund and the Caltech President's Fund that involve the TDA Office are included.

This and each succeeding issue of *The TDA Progress Report* will present material in some, but not necessarily all, of the following categories:

OSO Tasks:
    DSN Advanced Systems
        Tracking and Ground-Based Navigation
        Communications, Spacecraft–Ground
        Station Control and System Technology
        Network Data Processing and Productivity
    DSN Systems Implementation
        Capabilities for Existing Projects
        Capabilities for New Projects
        New Initiatives
        Network Upgrade and Sustaining
    DSN Operations
        Network Operations and Operations Support
        Mission Interface and Support
        TDA Program Management and Analysis
    Ground Communications Implementation and Operations
    Data and Information Systems
    Flight–Ground Advanced Engineering
    Long-Range Program Planning

OSO Cooperative Tasks:
    Orbital Debris Radar Program
    21st Century Communication Studies

# Contents

## OSO TASKS
## DSN Advanced Systems
### TRACKING AND GROUND-BASED NAVIGATION

### COMMUNICATIONS, SPACECRAFT–GROUND

### STATION CONTROL AND SYSTEM TECHNOLOGY

## DSN Systems Implementation
### CAPABILITIES FOR NEW PROJECTS

### NETWORK UPGRADE AND SUSTAINING

# DSN Operations

## TDA PROGRAM MANAGEMENT AND ANALYSIS

# OSO Cooperative Tasks

## TWENTY-FIRST CENTURY COMMUNICATION STUDIES

# OSSA Tasks

## SEARCH FOR EXTRATERRESTRIAL INTELLIGENCE

November 15, 1991

# A Receiver Design for the Superconducting Cavity-Maser Oscillator

R. T. Wang and G. J. Dick
Communications Systems Research Section

*A new frequency standard has been demonstrated with the aid of a double phase-locked loop (PLL) receiver. A superconducting cavity-maser oscillator (SCMO) and a hydrogen maser are combined to show the medium-term performance of the hydrogen maser together with improved short-term performance made possible by the SCMO. The receiver, which generates a 100-MHz signal with reduced noise, is phase-locked to (and may be used in place of) the 100-MHz hydrogen-maser output. The maser signal, 2.69xxx-GHz SCMO output, and a 100-MHz quartz-crystal oscillator are optimally combined by the receiver. A measured two-source fractional frequency stability of $2 \times 10^{-14}$ was obtained for a measuring time of $\tau = 1$ sec, and $1 \times 10^{-15}$ at $\tau = 1{,}000$ sec. The 1-sec value is approximately 10 times lower than that for hydrogen masers, while the 1,000-sec value is identical to hydrogen-maser performance. The design is based on phase-noise models for the hydrogen maser, the SCMO, and quartz-crystal oscillators for offset frequencies down to $1 \times 10^{-6}$ Hz.*

## I. SCMO Operation

The superconducting cavity-maser oscillator (SCMO) is an all-cryogenic, helium-cooled oscillator with superior stability at short measuring times [1–4]. It differs from other superconducting cavity-stabilized oscillator (SCSO) designs [5–7] in its use of a very rigid ($Q \approx 10^9$) sapphire-filled stabilizing cavity and in its all-cryogenic design, with excitation provided by an ultralow-noise cryogenic ruby maser.

The three-cavity oscillator, which consists of a ruby maser, coupling cavity, and a high-$Q$ lead-on-sapphire cavity, has been discussed previously [2]. Oscillation at a frequency of 2.69xxx GHz results from maser amplification in the ruby material, where a population inversion is generated by a 13.1-GHz pump signal. A match between ruby and resonator frequencies is effected by application to the ruby of an approximately 500-gauss bias field. Output power of the oscillator is $> 10^{-9}$ W, more than 1,000 times larger than that of the hydrogen maser, which makes possible $\sqrt{1000} \approx 30$ times higher stability at short measuring times.

The SCMO was recently modified to allow its frequency to be actively adjusted to match that of the hydrogen maser. A coil has been installed on the ruby housing to modify the magnetic bias field by application of a DC current. The coil produces a tuning sensitivity of $7 \times 10^{-12}$ per mA, with a range of approximately $10^{-10}$ without significant heating. This range, although only 1/1000 that of a quartz-crystal voltage-controlled oscillator (VCO), is sufficient to accommodate the typical SCMO drift of $4 \times 10^{-13}$/day in long-term operation.

## II. Loop Design and Oscillator Noise Models

A double phase-locked loop (PLL) was designed to optimally combine SCMO and hydrogen-maser stabilities. Figure 1 shows a block diagram of that design. Here, one loop locks the phase of a quartz-crystal VCO to the SCMO, while the second loop locks the SCMO/VCO combination to a hydrogen maser through the tuning coil in the SCMO. Design goals are to preserve SCMO short-term stability through the first PLL and optimize the second PLL so that long-term performance of the hydrogen maser is preserved without significantly degrading SCMO performance at $\tau = 1$ sec measuring time. The VCO–SCMO loop characteristics, with a bandwidth of approximately 1,000 Hz, do not significantly impact stability performance when $\tau \geq 1$ sec. However, the SCMO–hydrogen-maser loop does, and so deserves more attention.

While most measurements of the stability of frequency standards are expressed as an Allan deviation $\sigma_y(\tau)$ in the time domain, the results do not directly apply to loop design. This is because the stability (for any measuring time $\tau$) of a source that combines several standards will depend on performance of the contributing standards at *every* $\tau$.

However, the performance of any standard can also be expressed in the frequency domain as a spectral density of phase fluctuations $S_\phi(f)$, and the spectral densities for the various standards at any frequency $f$ simply add, with multiplicative constants determined by loop characteristics at that frequency. Roughly speaking, the unity-gain frequency for the loop will match the crossover frequency between the spectral densities for any two standards being combined.

Although stability measurement data are abundantly available for the DSN hydrogen masers, there existed no measurement of close-in phase noise ($f < 1$ Hz) when the authors undertook their receiver design. Thus, they created a phase-noise model for which the calculated Allan deviation $\sigma_y(\tau)$ matches in detail the results of stability measurements in JPL's frequency standards test facility (FSTF).

Details of this model appear in Table 1. Included for three separate noise components are: noise type, the Fourier frequency window for which the component is dominant, Allan deviation, and phase spectral density. The authors' calculation methodology for generating the Allan deviation corresponding to any phase-noise model is described in the Appendix.

Figure 2 shows a comparison of this model with the results of a very recent calculation of hydrogen-maser close-in phase noise using the same raw data from which Allan deviations were calculated. These data consist of measurements of the time for zero-crossings of a 1-Hz difference frequency using a 100-MHz + 1-Hz reference. The difference between the model and the measured data is within 3 dB for the region from $1 \times 10^{-4}$ to $1 \times 10^{-1}$.

Measurements of the spectral density of phase fluctuations were available for the other two frequency sources. The authors had previously made measurements of the phase noise for the SCMO/SCSO combination for offset frequencies down to 0.01 Hz, and in the absence of any better information, they ascribed half of the measured noise to the SCMO. Other tests of the SCMO with a hydrogen maser as reference indicate a somewhat lower value for offsets below 0.01 Hz. Information for the 100-MHz quartz-crystal oscillator was obtained from the manufacturer.

Figure 3 shows a plot of the phase noises, measured and inferred, for the three oscillators. A smaller contribution due to input noise for the operational amplifier in loop no. 2 is also shown. Crossover frequencies of approximately 0.04 Hz and approximately 1,000 Hz are readily identifiable.

## III. Loop-Parameter Optimization

As shown by the dotted line in Fig. 3, calculated performance for the combined source closely approximates the best of the three sources at every frequency. However, deviations from "ideal" performance are significant, as Fig. 4 shows in expanded form, for second-order loops with various loop bandwidths and damping factors. Parameter optimization requires a measure of "goodness" for loop action. It is often the case that performance is optimized for a time period which is much longer than that associated with action of the loop. In this case such an optimization measure can be, for example, rms deviation of the phase error of the following oscillator from the one being followed. In the case at hand, such a single high- or low-frequency measure is not useful because frequencies of concern lie both above and below the crossover frequency.

As shown in Fig. 4, performance deviates significantly from the ideal for frequencies above and below the crossover frequency itself. Some applications of the combined source will, for example, emphasize high- over low-frequency performance (or short-term stability for a combined source at the expense of long-term stability). The authors have chosen instead a balanced optimization criterion, which both equalizes and minimizes peak deviations on either side of the crossover.

Successive steps approaching this optimized condition are shown in Fig. 4. A first guess—matching the loop bandwidth to the crossover frequency (0.04 Hz) with a damping factor of $1/\sqrt{2}$—resulted in a large low-frequency peak, substantially worse than any shown in the figure. Equivalent high- and low-frequency performances can be observed in the examples on a diagonal from upper right to lower left. Optimized performance, as shown at the lower left, results primarily from an increased damping factor.

The authors' calculations show continuing small improvement (< 0.2 dB) as the damping factor is further increased. This indicates that on the basis of this criterion alone, an infinite damping factor, corresponding to a first-order loop, would be preferable. However, the larger frequency drift rate of the SCMO, as compared with the hydrogen-maser, necessitates the use of a second-order loop to prevent a slow buildup of phase error, which corresponds to a frequency offset, between the sources. The value of 3 for the damping factor is chosen to give good noise performance and drift immunity combined with relatively rapid loop response.

While spectral densities give a complete representation of performance of the standard, performance in the time domain is a more familiar and widely used measure of performance. Figure 5 shows a calculated Allan deviation for the composite spectral density for the combined sources. It is surprising that stability near the crossover point here is a little better than that of either source, while it closely follows the SCMO and the hydrogen maser at short and longer times, respectively. The slight improvement at the crossover occurs because the Allan deviation for either component standard depends significantly on frequencies both above and below the crossover frequency, and the composite has the advantages of both at the crossover itself.

The calculated combined source shows a stability of $1 \times 10^{-14}$ at 1 sec and $1 \times 10^{-15}$ from 1,000 sec onward. The Allan deviation of the combined source follows very closely the best source in the time window of 1–10,000 sec.

## IV. Measurements

To confirm the stability of the combined source, a second identical system will be needed. Presently available are hydrogen-masers to test performance at longer measuring times and the SCSO for short-term characterization [5]. Figure 6 shows measurements that demonstrate the short-term stability of the combined source at a test frequency of 100 MHz using the SCSO as a reference. During this test, the poor long-term stability of the SCSO reference substantially degraded the measured values for times $\tau > 30$ sec. However, results at shorter times are as much as 10 times lower than for two hydrogen masers in a similar test. The slight increase for $\tau < 3$ sec has previously been identified as characteristic of the SCSO reference [6], which indicates performance for the authors' combined source of $\sigma_y(\tau) < 1 \times 10^{-14}$ at $\tau = 1$ sec.

Figure 7 shows the Allan deviation of fractional frequency fluctuations for the SCMO locked to a hydrogen maser, using a second hydrogen maser as reference. A special feature is the stability of $6 \times 10^{-14}$ at 1 sec, a value that is one maser's stability. The transition from "one-maser" noise below $\tau = 30$ sec and "two-maser" noise for longer times causes the slight kink shown in the figure.

## V. Conclusions

The authors have designed, built, and demonstrated a receiver that enables the SCMO to operate in conjunction with a hydrogen maser to form a new standard, one which combines their complementary performance over the range of measuring times from 1 to 10,000 sec. Calculated performance for the combined source shows spectral performance $S_\phi(f)$ within 1.4 dB of the best of the two sources over the frequency range $10^{-6}$ Hz $< f < 10^4$ Hz. Allen deviation is calculated to be within 7 percent of the best of the two sources for measuring times of 1 sec $\leq \tau \leq 10^4$ sec, and somewhat surprisingly, slightly better than either standard at their crossover. Performance for the combined standard was demonstrated in separate experiments using different ultrastable frequency sources to be at least as good as $2 \times 10^{-14}$ at 1-sec measuring times and $1 \times 10^{-15}$ at 1,000 sec. Significant aspects of this experiment are a new time window for scientific experiments and a unique demonstration that combines two different types of ultrastable microwave oscillators.

The combined SCMO/hydrogen-maser source underwent a field test at Goldstone Deep Space Communications Complex for 72 days (May–July 1991). The performance was the same as reported above. This exercise was to prepare the SCMO for a gravitational wave detection experiment in connection with Galileo in 1992.

# Acknowledgments

Special thanks to the following members of the Communications Research Section: W. A. Diener, for his assistance in ultrastable measurements, R. E. Taylor, for easy access to the capabilities of the Test Facility and assistance with the SCMO field tests, Dr. C. A. Greenhall, for software enhancements meeting the authors' special requirements, T. K. Tucker, for his assistance and support with hydrogen-maser measurements, and Dr. L. Maleki, for his many helpful suggestions.

Table 1. Contributions to the total noise of the DSN hydrogen maser used for loop design

| Noise type | Fourier frequency window | Allan deviation | Phase spectral density |
|---|---|---|---|
| Flicker frequency | $1 \times 10^{-6}$ to $4 \times 10^{-4}$ | $1 \times 10^{-15}$ | $7.5 \times 10^{-15}/f^3$ |
| White frequency | $4 \times 10^{-4}$ to $4.5 \times 10^{-2}$ | $3 \times 10^{-14}/\tau^{1/2}$ | $1.8 \times 10^{-11}/f^2$ |
| Flicker phase | $4.5 \times 10^{-2}$ to $1 \times 10^{-1}$ | $2 \times 10^{-13}/\tau$ | $4 \times 10^{-10}/f$ |

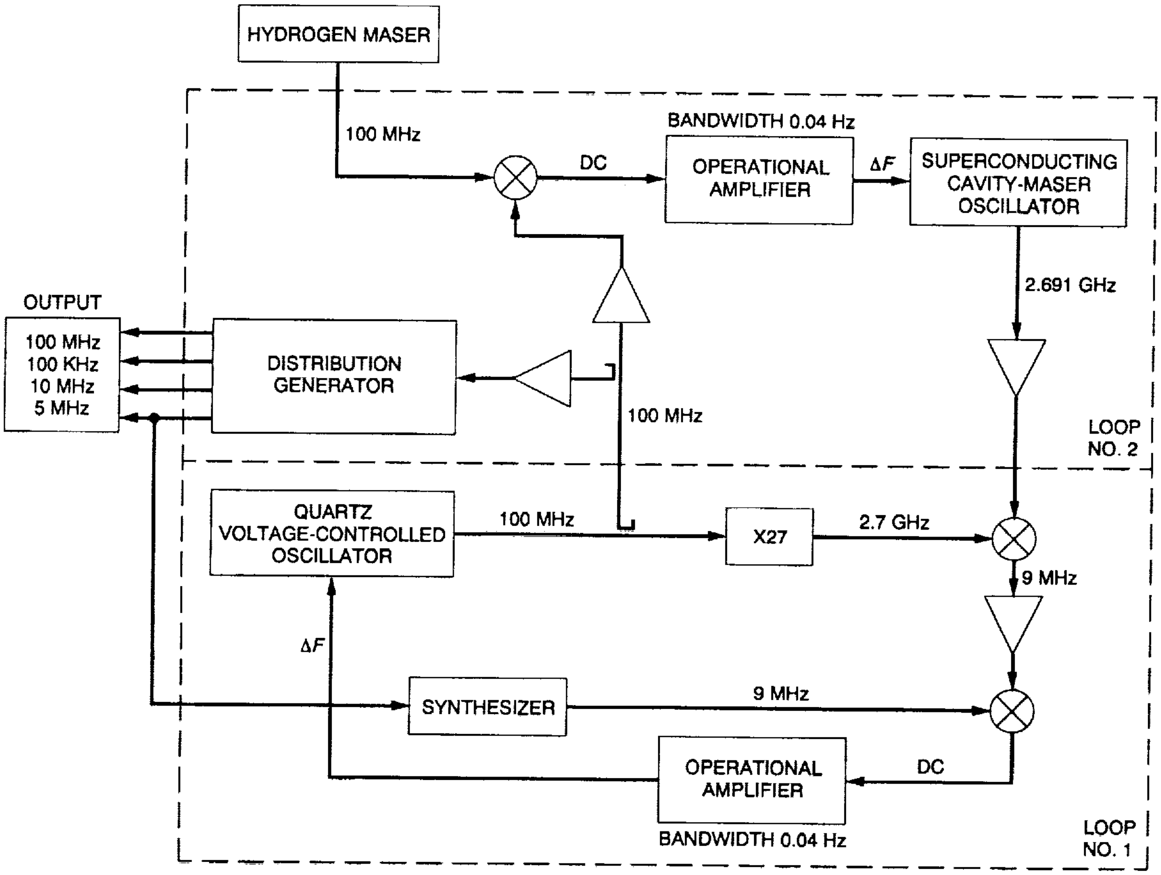


Fig. 1. Schematic of SCMO double-loop receiver at 100 MHz. The two loops are VCO/SCMO and SCMO/hydrogen maser. The required input is the 100-MHz hydrogen-maser signal, and the output has four different frequencies: 100, 10, 5, and 0.1 MHz.

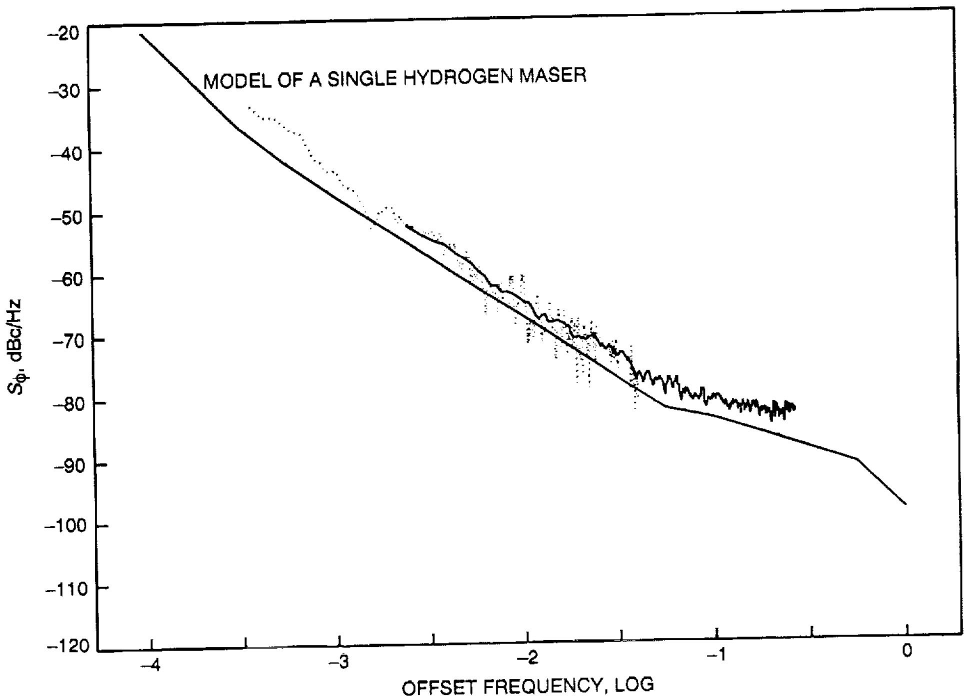


Fig. 2. A close-in phase-noise plot. The solid line shows the model of a single hydrogen maser; two sets of measured pair-data were shown. The difference is within 3 dB for the region from $1 \times 10^{-4}$ to $1 \times 10^{-1}$ Hz.

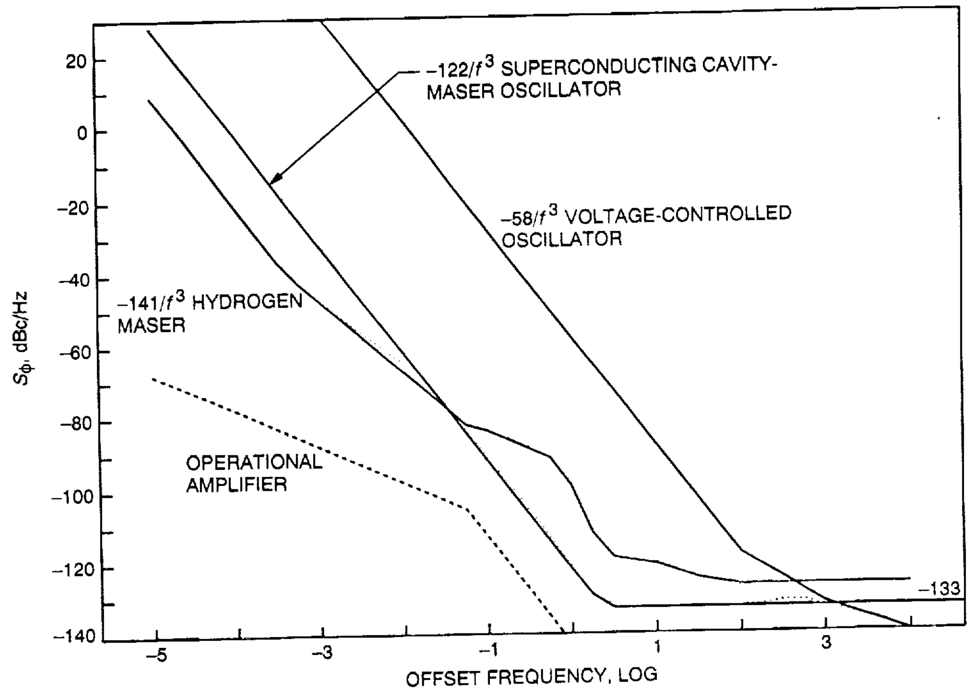


Fig. 3. Phase-noise plot of the three oscillators at 100 MHz. Calculated receiver performance, shown by the dotted line, closely follows the lowest noise of the three oscillators in the Fourier frequency window of $1 \times 10^{-5}$ to $10^{4}$ Hz. Noise due to the operational amplifier in loop no. 2 is shown to be much lower than oscillator noise contributions.

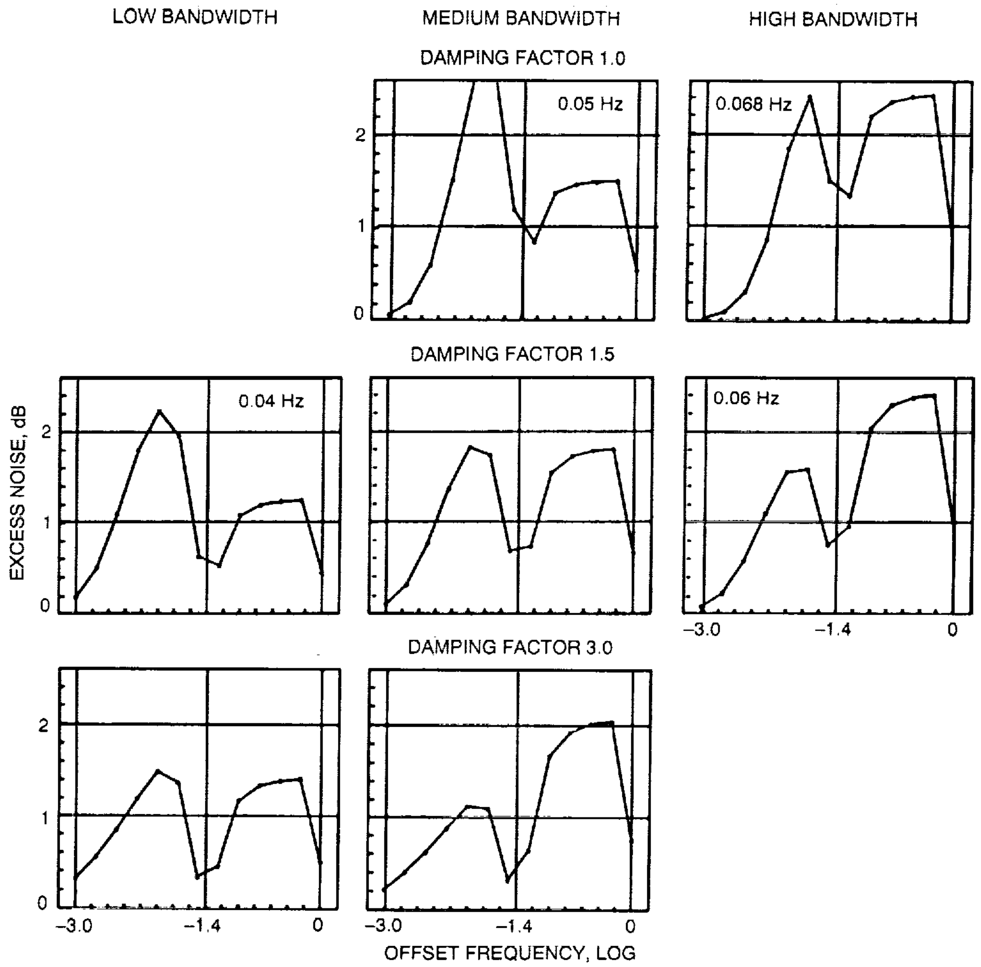


**Fig. 4. Optimization of the loop parameters is accomplished by minimizing calculated excess phase noise at the receiver output compared with the best of the constituent oscillators at any given offset frequency. Bandwidths for the various graphs increase from left to right and damping factor increases from top to bottom. Horizontal and vertical scales are identical for each subgraph. Best overall performance, as shown at the lower left, corresponds to a bandwidth of 0.04 Hz (a value equal to the SCMO–hydrogen-maser crossover frequency) and a relatively large damping factor.**

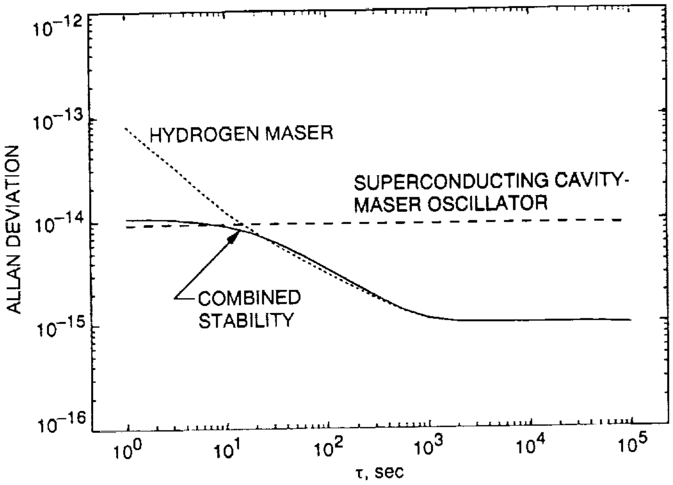


Fig. 5. Allan deviation derived from combined phase noise data (Fig. 3). The expected performance of the combined signal is 1 × $10^{-14}$ at 1 sec and 1 × $10^{-15}$ at 1,000 sec and beyond.

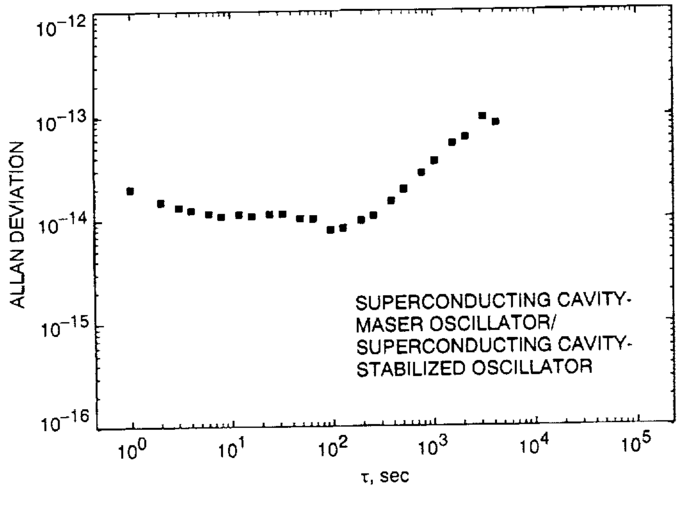


Fig. 6. Two sample Allan deviations of the SCMO tested with SCSO at 100 MHz. The measured frequency stability is 2 × $10^{-14}$ at 1 sec. Structure below $\tau$ = 100 sec is primarily due to the SCSO reference.

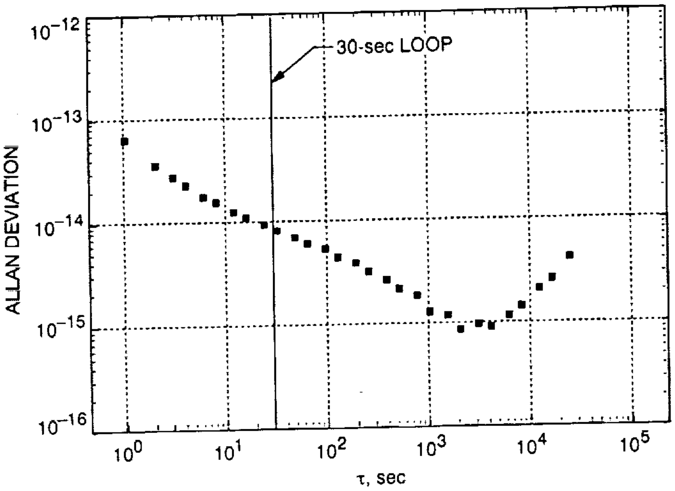


Fig. 7. Two sample Allan deviations of the SCMO locked to one hydrogen maser; the second maser is used as a reference. The vertical line shows the 30-sec loop time constant. The transition from "one-maser" noise below $\tau$ = 30 sec and "two-maser" noise for longer times causes the slight kink shown in the figure.

# Appendix

## Calculation of the Allan Deviation

The Allan deviation of variation of fractional frequency $\sigma_y^2(\tau)$ for any radio frequency (RF) signal can be derived from its spectral density of phase noise $S_\phi(f)$ by using [8]

$$\sigma_y^2(\tau) = \int_0^\infty C(\tau) \times S_\phi(f) \times \sin^4(\pi\tau f)df$$

where

$$C(\tau) = \frac{2}{\pi^2\tau^2\nu_o^2}$$

and $\nu_0$ is the RF frequency. In order to use the numerical integration routines in a popular mathematics program [9] to evaluate this relatively straightforward integral, it is necessary to "condition" the problem in two ways.

First, because significant contributions to the integral are due to frequency components that vary by many orders of magnitude, a change in variable is required, which results in an effectively logarithmic frequency scale. Second, calculation at high frequencies, as compared with $1/\tau$, is hampered by rapid variation in the $\sin^4$ term, which should be replaced by its average value (3/8) as the frequency is raised. It is important for the function that does the replacement to show *very complete* elimination of the average term as the argument $\pi\tau f$ goes to zero so as not to dominate the $(\pi\tau f)^4$ dependence of the $\sin^4$ term.

The change of variable required is straightforward, replacing the frequency $f$ with $e^g$ and the differential $df$ with $e^g dg$. Lower and upper limits to the integration were typically $-6\ln(10)$ and $4\ln(10)$, which corresponds to an integration over frequencies from $10^{-6}$ Hz to $10^4$ Hz.

Elimination of rapid variation is accomplished by replacing

$$\sin^4(\pi\tau f)$$

by

$$Dk\left[\frac{\tau f}{n}\right] \times \left[\sin^4(\pi\tau f) - \frac{3}{8}\right] + \frac{3}{8}$$

where $n$ represents the number of cycles of variation in the $\sin^4$ term that are to be explicitly integrated, and $Dk$ is a decay function defined by

$$Dk(x) = e^{-x\ln(2)}$$

for $x \geq 1$, and

$$Dk(x) = 1 - e^{-\left(\frac{\ln(2)}{x}\right)}$$

for $x \leq 1$. Features of this decay function are extremely rapid approach to 1 for small $x$, rapid approach to 0 for large $x$, and continuous first derivative at the crossover point $x = 1$. The authors found $n = 3$ cycles sufficient to reduce fractional errors to less than $10^{-4}$.

Taken together, these substitutions allow a difficult integral to be accurately evaluated by means of "canned" integration routines.

# References

[1] S. Thakoor, D. M. Strayer, G. J. Dick, and J. E. Mercereau, "A Lead-on-Sapphire Superconducting Cavity of Superior Quality," *J. Appl. Phys.*, vol. 59, pp. 854–858, February 1, 1986.

[2] G. J. Dick and D. M. Strayer, "Development of the Superconducting Cavity Maser as a Stable Frequency Source," *Proceedings of the 38th Annual Frequency Control Symposium*, pp. 435–446, May 29–June 1, 1984.

[3] R. T. Wang, G. J. Dick, and D. M. Strayer, "Operational Parameters for the Superconducting Cavity Maser," *Proceedings of the 20th Annual Precise Time and Time Interval (PTTI) Planning and Applications Meeting*, Vienna, Virginia, pp. 345–354, November 29, 1988.

[4] R. T. Wang and G. J. Dick, "Improved Performance of the Superconducting Cavity Maser at Short Measuring Time," *Proceedings of the Annual Frequency Control Symposium*, Baltimore, Maryland, vol. 44, pp. 89–93, May 23, 1990.

[5] S. R. Stein and J. P. Turneaure,"The Development of The Superconducting Cavity Stabilized Oscillator," *Proceedings of the Annual Frequency Control Symposium*, vol. 27, pp. 414–420, June 12–14, 1973.

[6] S. R. Stein, "Space Application of Superconductivity: Resonators for High Stability Oscillators and Other Applications," *Cryogenics*, vol. 22, pp. 363–371, July 1980.

[7] A. J. Giles, S. K. Jones, D. G. Blair, and M. J. Buckingham, "A High Stability Microwave Oscillator based on a Sapphire Loaded Superconducting Cavity," *Proceedings of the Annual Frequency Control Symposium*, Denver, Colorado, vol. 43, pp. 89-93, May 31-June 2, 1989.

[8] J. Rutman and J. Uebersfeld, "A Model for Flicker Frequency Noise of Oscillators," *Proceedings of the IEEE*, vol. 60, no. 2, pp. 233–235, February 1972.

[9] MathCad 2.0 by MathSoft, Inc., Cambridge, Massachusetts.

S2-32
53205

TDA Progress Report 42-107 November 15, 1991

P. 9

N92-14239

# Suppressed Carrier Full-Spectrum Combining

D. H. Rogstad
Tracking Systems and Applications Section

*A technique to accomplish full-spectrum arraying where all the telemetry power is put into the subcarrier sidebands (suppressed carrier) is described. The matched filter needed in each antenna prior to cross-correlation for deriving the coherence delay and phase offsets is an open-loop version of the telemetry phase-lock loop provided in the Advanced Digital Receiver. In analogy with a Costas-loop telemetry receiver, a "squaring loss" is derived, and a signal-to-noise ratio for the cross-correlation loop phase is presented.*

## I. Introduction

Normally, as a spacecraft travels farther from Earth and the telemetry signal-to-noise ratio (SNR) gets poorer, two system parameter trade-offs come into play. First, the telemetry modulation index is usually increased so more transmitter power is moved from the carrier to the telemetry signal, thereby improving telemetry SNR. This, of course, may result in a carrier signal that is significantly harder to acquire and track. The limit for this trade-off is full modulation where *no* carrier power is present. In this case, the carrier signal frequency must be acquired and tracked using a less-than-optimal Costas phase-lock-loop technique. The capability to Costas-loop track is not presently available in the DSN, but is planned as part of the new Block V receivers.

The second trade-off that comes into play is the rate at which telemetry data are transmitted back to Earth; this rate can be reduced, resulting in an improved SNR per telemetry bit. This has the unfortunate consequence of also reducing the total amount of data that can be returned during the critical encounter phase of a mission.

A technique often applied within the DSN to overcome these constraints (short of building larger antennas, even lower noise receivers, or employing more advanced data encoding/decoding methods) is antenna arraying to enhance the SNR of the received telemetry [1]. Perhaps the most recent major applications of antenna arraying occurred during the Voyager 2 encounter with Uranus and the Voyager 2 encounter with Neptune. In each case, arraying provided adequate telemetry SNR employing data rates higher than would have been possible otherwise.

At least four different techniques, which depend on the details of the observing circumstance, have been used for arraying or combining the signals from several antennas. These include symbol-stream combining, baseband arraying, carrier arraying (or more correctly, carrier aiding), and full spectrum combining. In the various discussions of these techniques, the terms arraying and combining are usually used interchangeably.

The first technique, symbol-stream combining, works well when each ground-based antenna/receiver is capable of locking on the spacecraft telemetry stream and demodulating it down to the soft symbol stream (the raw telemetry data before the decision of whether a given bit is 1 or 0 is made). Each stream must then be delayed to line up symbols, followed by weighting and combining. One major advantage of this technique is that the data stream from each antenna arrives at only the telemetry rate (typically on the order of 100 kbaud), facilitating its transmission to the combining location, either in real time or through a recording medium. The International Cometary Explorer (ICE) mission successfully used symbol-stream combining in a primary mode to gain approximately 2 dB in SNR [2]. A short time later, this combining technique was used on Voyager 2 to gain slightly less than 3 dB improvement [3]. More recently, symbol-stream combining was used as a backup for baseband arraying during the Voyager 2 encounter with Neptune [4].

The second technique, baseband arraying, works when the telemetry SNR is high enough to permit locking on and tracking of the carrier signal at each antenna, but too low to reliably maintain subcarrier lock and symbol sync. In this case, the baseband signal obtained after carrier demodulation is delayed, weighted, and added to a corresponding baseband signal from the other antenna. Then subcarrier demodulation and symbol sync are accomplished on the combined stream. The baseband bandwidth is on the order of 3 MHz and therefore somewhat more difficult to transmit or record for combining. During the Voyager 2 encounter with Uranus, this technique was used effectively with the Parkes Radio Astronomy antenna in Australia [5].

The third technique, carrier aiding, is useful when one antenna in the array can acquire and maintain carrier lock, but the other antennas cannot, either because they are smaller or have higher noise receivers. If the antennas to be arrayed are in proximity, the lock signal from the more sensitive antenna can be sent to the other antennas in real time and used to aid their locking on the telemetry carrier. The resulting baseband signals are then treated as they are in baseband arraying. This technique has been used quite recently to successfully array a 70-m antenna and a 34-m antenna at Goldstone while looking at the Pioneer 11 spacecraft.[1]

Finally, full-spectrum combining of an open-loop spacecraft telemetry signal can be used when the carrier is too weak to track, or when it is not convenient to track, at any single antenna. One form of this technique was used to coherently add the signal from all the antennas of the Very Large Array (VLA) during the Voyager 2 Neptune encounter [6]. Following is a description of a significantly improved version of full-spectrum combining, with comments on how it differs from the original.

By way of introduction, note that the telemetry signal transmitted from a spacecraft can be represented by the equation [7]

$$S[t] = \sqrt{2P}\{\cos\Delta \sin[\Omega_c t] + d[t] \sin \Delta \Sigma s_k[t]\}$$

where each term in the summation (odd $k$ from 1 to $\infty$) is

$$s_k[t] = (2/\pi)(1/k)\{g_{uk} \sin[(\Omega_c + k\Omega_{sc})t] - g_{lk} \sin[(\Omega_c - k\Omega_{sc})t]\}$$

corresponding to the subcarrier sidebands. The amplitude corresponds to a signal total power of $P$. Figure 1 is a graph of this signal.

In these expressions,

$t$ = time

$\Omega_c$ = the transmitter carrier frequency

$\Omega_{sc}$ = the telemetry subcarrier frequency

$\Delta$ = the modulation index of the subcarrier

$k$ = the subcarrier harmonic number

$d[t]$ = the telemetry modulation

$g_{uk}$ = the upper sideband gain at frequency $k$

$g_{lk}$ = the lower sideband gain at frequency $-k$

$P$ = the signal total average power

For simplicity, assume the modulation index is 90 deg, so that only the suppressed carrier case is treated (this will likely be true when full-spectrum combining is used to put all possible power into the telemetry). Also, assume the gain factors are all unity so the summation over $k$ can be treated as a square wave (the effects of this on the end result are minor). The signal then becomes

$$S[t] = \sqrt{2P}d[t] \cos[\Omega_c t]\mathbf{S}[\Omega_{sc} t] \qquad (1)$$

where the outlined **S** represents a square wave with zero crossings coincident with a sine wave of the same argument.

[1] T. Peng, "Carrier Array Demonstration at Goldstone with Pioneer 11," JPL Interoffice Memorandum 3393-90-62 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 11, 1990.

The spacecraft signal in Eq. (1) is received by several antennas, located at various delays, $\tau$, from the spacecraft. These delays include geometric as well as nondispersive media and instrument components (as well as a large component corresponding to the distance from the spacecraft to the center of the Earth and common to all antennas, which will be ignored because it differences out in the final results). In addition, each antenna's receiver downconverts the signal to a carrier intermediate frequency, $\Omega_{if} = (\Omega_c - \Omega_{lo})$, using a local oscillator frequency, $\Omega_{lo}$. The combined effect on the signal received at the $m$th antenna is

$$S_m[t] = \sqrt{2P}d[t - \tau_m] \cos[\Omega_{if}t - \Omega_c\tau_m + \theta_{om}]$$

$$\times \mathbf{S}[\Omega_{sc}(t - \tau_m)] + N_m \qquad (2)$$

where $\theta_{om}$ is an unknown, slowly varying (its change in a time equal to the difference in the delays between the various antennas is negligible) residual phase due to various dispersive media and instrument effects. $N_m$ is the noise of variance $\sigma_{om}$ in the bandpass containing the signal.

The goal of the combining technique is to coherently add this signal from several antennas to obtain an improved SNR for telemetry extraction. To achieve this goal, the received signals must be delayed and phase shifted to correct for the relative effects between the various antennas that are represented in Eq. (2) before they can be added. In the following analysis, assume the signal (plus noise) modeled in Eq. (2) is digitally sampled at the Nyquist rate and processed with accuracy sufficient to ignore quantization and round-off effects.

## II. Delay and Phase Shift

Following the steps outlined in Fig. 2, begin by delaying the signal received at the $m$th antenna by $\hat{\tau}'_m$, based on a best a priori model, $\tau'_m$, of the antenna/spacecraft system (the caret on this delay signifies a quantized version of the model delay; the prime indicates a model delay). From this,

$$S_m[t] = \sqrt{2P}d[t - \Delta_m] \cos[\Omega_{if}(t + \hat{\tau}'_m) - \Omega_c\tau_m + \theta_{om}]$$

$$\times \mathbf{S}[\Omega_{sc}(t - \Delta_m)] \qquad (3)$$

where the quantity $\Delta_m = \tau_m - \hat{\tau}'_m$ is the residual between the actual delay and the value used in data processing.

The delay operation is followed by a phase shift that involves first the generation of a quadrature version of Eq. (3) using a hybrid and then a complex multiplication by an expression of the form

$$E_{pm}[t] = \exp[-j\theta_{pm}] \qquad (4)$$

The final signal becomes

$$Sm[t] = \sqrt{2P}d[t - \Delta_m] \exp[j\{\Omega_{if}(t + \hat{\tau}'_m) - \Omega_c\tau_m + \theta_m\}]$$

$$\times \mathbf{S}[\Omega_{sc}(t - \Delta_m)] \qquad (5)$$

where $\theta_m = \theta_{om} - \theta_{pm}$ is the resulting phase offset of the signal. Note that the signal is now a complex quantity, having both in-phase and quadrature-phase components.

By adjusting $\hat{\tau}'_m$ and $\theta_{pm}$, it is possible to bring the signals from each antenna into coherence for combining. An error signal will be generated by comparing the signal in Eq. (5) between pairs of antennas and then using this signal to control the delay and phase adjustment of the telemetry streams from these antennas relative to each other.

## III. Matched Filter

The first step toward obtaining an error signal is to perform a matched filtering operation on the signal. When implementing full spectrum arraying with the VLA during the Voyager Neptune encounter, this "matched" filtering consisted simply of narrowing the VLA receiver bandpass from its nominal 50 MHz down to about 8 MHz with baseband filters. While hardly optimal, such a procedure is acceptable if the SNR at each antenna is reasonable, as was the case for Voyager. However, in general one needs a more optimal approach. A true matched filtering (providing a better power SNR by as much as a factor of 8 MHz/43 kHz, or ~ 180 in this case) is properly accomplished through the implementation of two procedures: (1) carrier demodulation followed by (2) subcarrier demodulation.

### A. Carrier Demodulation

Carrier demodulation is accomplished by coherently detecting the signal at the carrier intermediate frequency with a locally generated signal of the form

$$E_{cm}[t] = \exp[-j\{\Omega'_{if}(t + \hat{\tau}'_m) - \Omega'_c\tau'_m + \theta_{cm}\}] \qquad (6)$$

where the primed quantities are the best guesses for their unprimed counterparts, and $\theta_{cm}$ is an unknown arbitrary phase. The result of this process is a signal with real and imaginary parts representing the in-phase ($I$) and quadrature-phase ($Q$) components of the demodulated carrier. After low-pass filtering,

$$< I_{cm} > \; = (1/2)\sqrt{2P}d[t - \Delta_m] \; \cos\, [\Phi_{cm}]\mathbf{S}[\Omega_{sc}(t - \Delta_m)] \tag{7}$$

$$< Q_{cm} > \; = (1/2)\sqrt{2P}d[t - \Delta_m] \; \sin\, [\Phi_{cm}]\mathbf{S}[\Omega_{sc}(t - \Delta_m)] \tag{8}$$

where

$$\Phi_{cm} = (\Omega_c - \Omega_c')t - \Omega_c\Delta_m + \Omega_c'\Delta_m' + \theta_m - \theta_{cm} \tag{9}$$

is the residual carrier demodulation phase and, with good modeling, is only a slowly varying function of time. In Eq. (9), the quantity $\Delta_m' = \tau_m' - \hat{\tau}_m'$ is the model residual delay.

### B. Subcarrier Demodulation

Subcarrier demodulation is accomplished by coherently detecting the signals in Eqs. (7) and (8) at the subcarrier frequency using

$$E_{sm}[t] = j\mathbb{E}[-j\{\Omega_{sc}'(t - \Delta_m') + \theta_{sm}\}] \tag{10}$$

where $\Omega_{sc}'$ is the best guess of the subcarrier-oscillator frequency, $\theta_{sm}$ is an unknown arbitrary phase, and the outlined $\mathbb{E}$ represents a complex square wave with components $\mathbf{C} + j\mathbf{S}$. Again, the results of this process are signals with real and imaginary parts representing the in-phase and quadrature-phase components of this subcarrier demodulation. After filtering, there are four components:

$$< I_{cm}I_{sm} > = (1/2)\sqrt{2P}d[t - \Delta_m]$$
$$\times \cos\, [\Phi_{cm}]\mathbf{C}^+[\Phi_{sm}] + N_{IIm} \tag{11}$$

$$< Q_{cm}I_{sm} > = (1/2)\sqrt{2P}d[t - \Delta_m]$$
$$\times \sin\, [\Phi_{cm}]\mathbf{C}^+[\Phi_{sm}] + N_{QIm} \tag{12}$$

$$< I_{cm}Q_{sm} > = (1/2)\sqrt{2P}d[t - \Delta_m]$$
$$\times \cos\, [\Phi_{cm}]\mathbf{S}^+[\Phi_{sm}] + N_{IQm} \tag{13}$$

$$< Q_{cm}Q_{sm} > = (1/2)\sqrt{2P}d[t - \Delta_m]$$
$$\times \sin\, [\Phi_{cm}]\mathbf{S}^+[\Phi_{sm}] + N_{QQm} \tag{14}$$

where

$$\Phi_{sm} = (\Omega_{sc} - \Omega_{sc}')t - \Omega_{sc}\Delta_m + \Omega_{sc}'\Delta_m' - \theta_{sm} \tag{15}$$

is the residual (slowly varying) subcarrier demodulation phase. Specifically, the real component of Eq. (10) when mixed with Eqs. (7) and (8) and then filtered gives Eqs. (11) and (12); the imaginary component of Eq. (10) when mixed with Eqs. (7) and (8) and then filtered gives Eqs. (13) and (14). The outlined $\mathbf{C}^+$ and $\mathbf{S}^+$ quantities represent the convolution of appropriate square waves (sine or cosine) as a function of their offset in phase; these waves have been lowpass filtered. They are, in fact, triangle waves as a function of this phase argument, with peak amplitude 1. The four independent noises, $N_{IIm}$, $N_{QIm}$, $N_{IQm}$, and $N_{QQm}$, have been included explicitly at this point, each having a mean square variance of

$$\sigma_m^2 = \; (1/4)\sigma_m^2/n_a \tag{16}$$

In this equation, recall that $\sigma_m$ is the noise deviation per data point (noise coherence interval) as represented by $N_m$ in Eq. (2), while the 1/4 factor is due to averaging over the IF cycles and $n_a$ is the number of data points over which the lowpass filters average. It is desirable that the value of $n_a$ correspond to a length of time equivalent to a telemetry bit interval (the largest it can be when the telemetry, represented by the function $d(t)$, is unknown). The goal of the delay and phase modeling is to minimize the changes in residual delay and phase so this integration can be as long as possible before the cross-correlation is performed. Note that to achieve an $n_a$ corresponding to a full telemetry bit interval, the symbol-sync on each of the data streams must be obtained independently. This is possible in the face of low SNR, since one can integrate over many telemetry bits to acquire sync.

## IV. Cross-Correlation

The final step in the detection process requires the "squaring" of the signal to eliminate the telemetry data,

$d[t]$, which has an amplitude of $\pm 1$ but an unknown code. This is accomplished by cross-multiplying, or correlating, the signals from pairs of antennas. Unfortunately, this is a noisy process because of the multiplying of noise by noise, and noise by signal that comes about when generating the desired signal-by-signal product. If the $SNR_i$ of the incoming data is low ($<1$), then the $SNR_o$ of the outgoing signal will be a function of the square of $SNR_i$, which deteriorates rapidly with decreasing $SNR_i$. The loss of SNR due to this effect in a Costas loop is usually called "squaring loss." Here it more appropriately should be called "cross-correlation loss."

By taking some of the possible products (traceable from their names) of Eqs. (11) through (14) pair-wise for two antennas (the $m$th and the $n$th), for the signal-by-signal part,

$$< II_m QI_n > = (P/2)d^+ \; [|\Delta_m - \Delta_n|] \cos[\Phi_{cm}] \times \sin[\Phi_{cn}] \; \mathbf{C}^+ [\Phi_{sm}] \; \mathbf{C}^+ [\Phi_{sn}] \qquad (17)$$

$$< QI_m II_n > = (P/2)d^+ \; [|\Delta_m - \Delta_n|] \sin[\Phi_{cm}] \times \cos[\Phi_{cn}] \; \mathbf{C}^+ [\Phi_{sm}] \; \mathbf{C}^+ [\Phi_{sn}] \qquad (18)$$

$$< IQ_m QQ_n > = (P/2)d^+ \; [|\Delta_m - \Delta_n|] \cos[\Phi_{cm}] \times \sin[\Phi_{cn}] \; \mathbf{S}^+ [\Phi_{sm}] \; \mathbf{S}^+ [\Phi_{sn}] \qquad (19)$$

$$< QQ_m IQ_n > = (P/2)d^+ \; [|\Delta_m - \Delta_n|] \sin[\Phi_{cm}] \times \cos[\Phi_{cn}] \; \mathbf{S}^+ [\Phi_{sm}] \; \mathbf{S}^+ [\Phi_{sn}] \qquad (20)$$

where

$$d^+ [|\Delta_m - \Delta_n|] \; = < d[t - \Delta_m] d[t - \Delta_n] > \qquad (21)$$

is the time average of a convolution of the telemetry data as a function of the delay residuals. This function has a maximum value of 1 at the origin, and drops linearly to zero at an absolute delay difference of one telemetry bit length, assuming $d[t]$ takes on values of $\pm 1$, and successive data bits are uncorrelated.

If Eq. (17) is subtracted from Eq. (18), and Eq. (19) is subtracted from Eq. (20),

$$Q_{\mathbf{CC}} = (P/2)d^+ [|\Delta_m - \Delta_n|] \sin[\Phi_{cm} - \Phi_{cn}] \times \mathbf{C}^+ [\Phi_{sm}] \; \mathbf{C}^+ [\Phi_{sn}]$$

$$Q_{\mathbf{SS}} = (P/2)d^+ [|\Delta_m - \Delta_n|] \sin[\Phi_{cm} - \Phi_{cn}] \times \mathbf{S}^+ [\Phi_{sm}] \; \mathbf{S}^+ [\Phi_{sn}]$$

When these two expressions are added,

$$Q_{\hat{\mathbf{C}}} = (P/2)d^+ [|\Delta_m - \Delta_n|] \; \sin[\Phi_{cm} - \Phi_{cn}] \times \hat{\mathbf{C}}^+ \; [\Phi_{sm} - \Phi_{sn}] \qquad (22)$$

In this expression, $\hat{\mathbf{C}}^+[\Phi_{sm} - \Phi_{sn}]$ is the time average of the sum of $\mathbf{C}^+[\Phi_{sm}]\mathbf{C}^+[\Phi_{sn}]$ and $\mathbf{S}^+[\Phi_{sm}]\mathbf{S}^+[\Phi_{sn}]$ and is simply a function of the difference of $\Phi_{sm}$ and $\Phi_{sn}$. In fact, this function is a slightly smoothed version of the triangle function represented by $\mathbf{C}^+[\Phi]$ (hence, the caret on $\mathbf{C}^+$) with a maximum value of 2/3 at $\Phi = 0$.

By taking all other possible products of Eqs. (17) through (20) for two antennas and combining these in a fashion analogous to that used to derive Eq. (22),

$$I_{\hat{\mathbf{C}}} = (P/2)d^+ [|\Delta_m - \Delta_n|] \cos[\Phi_{cm} - \Phi_{cn}] \times \hat{\mathbf{C}}^+ \; [\Phi_{sm} - \Phi_{sn}] \qquad (23)$$

$$Q_{\hat{\mathbf{S}}} = (P/2)d^+ [|\Delta_m - \Delta_n|] \sin[\Phi_{cm} - \Phi_{cn}] \times \hat{\mathbf{S}}^+ \; [\Phi_{sm} - \Phi_{sn}] \qquad (24)$$

$$I_{\hat{\mathbf{S}}} = (P/2)d^+ [|\Delta_m - \Delta_n|] \cos[\Phi_{cm} - \Phi_{cn}] \times \hat{\mathbf{S}}^+ \; [\Phi_{sm} - \Phi_{sn}] \qquad (25)$$

The total cross-correlation noise on each of the components given in Eqs. (22) through (25) is due to the sum of the products of the various signals and noises explicitly shown in Eqs. (11) through (14). It is straightforward, but tedious, to show that these cross-correlation noise components are all independent of each other, have zero mean value, and all have the same maximum mean square variance of

$$\sigma_x^2 = 4\sigma^2(P/6 + \sigma^2)/n_t \qquad (26)$$

In Eq. (26), $P$ is the power in the original telemetry signal, $\sigma^2$ is the mean square noise variance per telemetry bit for the two uncorrelated signals making up this cross product from Eq. (16),

$$\sigma^2 = (\sigma_m^2 + \sigma_n^2)/2 \qquad (27)$$

and $n_t$ is the number of telemetry bits over which the cross-correlation sum is averaged.

## V. Residual Delay and Phase Estimation

Following the practice used in very long baseline interferometry (VLBI) cross-correlation, each of these signals can be determined for several delay lags (i.e., for several values of $|\Delta_m - \Delta_n|$ around the best a priori estimate). Then, based on the signal amplitudes in each lag, the correct delay difference can be determined and used to adjust the delay of one of the data streams relative to the other to maintain data stream *delay* coherency. In actual fact, when the telemetry bits are many microseconds in length, the delay difference can be estimated a priori to a small fraction of this bit length, eliminating the need for multilag correlation.

The phase differences in Eqs. (22) through (25) can be written explicitly as

$$\Phi_{cm} - \Phi_{cn} = \Omega_c'(\Delta_m' - \Delta_n') - \Omega_c(\Delta_m - \Delta_n)$$

$$+ \theta_{om} - \theta_{on} - \theta_{pm}$$

$$+ \theta_{pn} - \theta_{cm} + \theta_{cn} \qquad (28)$$

and

$$\Phi_{sm} - \Phi_{sn} = \Omega_{sc}'(\Delta_m' - \Delta_n') - \Omega_{sc}(\Delta_m - \Delta_n)$$

$$- \theta_{sm} + \theta_{sn} \qquad (29)$$

where $\Delta_k = \tau_k - \hat{\tau}_k'$ and $\Delta_k' = \tau_k' - \hat{\tau}_k'$ are the residual delay and the model residual delay for the $k$th antenna.

Assuming (1) the frequencies can be specified to within a few hertz, (2) the rate-of-change of the delay differentials is no greater than a few picoseconds per hour (quite easily obtained, based on VLBI experience), and (3) the instrument phases are indeed slowy varying (a few tens of millihertz), it is possible to estimate these phase differences and feed them back to the appropriate phase mixers in one of the antenna data streams to maintain data stream *phase* coherency. Note that for the carrier phase in Eq. (28), this correction can be accomplished in any one of three phase mixers, while for the subcarrier phase in Eq. (29), it must be performed in the subcarrier demodulation mixer.

In actual practice, the subcarrier phase residual in Eq. (29) can be kept small. This is because the delay residuals and the antenna clock offsets can usually be maintained at less than 25 nsec, which, for a subcarrier frequency on the order of 1 MHz, will result in no more than 10 deg of phase error. Therefore, the only residual that need be estimated is that for the carrier phase.

## VI. Signal-to-Noise Ratio

Estimating the noise in measuring the cross-correlation phase, as represented in Eq. (28), is completely analogous to determining the same quantity for a Costas phase-lock loop when tracking telemetry. As noted above, the Costas-loop algorithm involves the multiplying of quadrature signal components from a single antenna, while cross-correlation involves multiplying analogous signal components, but crosswise from a pair of antennas.

Using the quantities given in Eqs. (11) and (12) from the same (e.g., $m$th) antenna, the Costas-loop equivalent of Eq. (22) can be derived:

$$Q_{\mathbf{C}} = (P/4)d^{+}[0] \sin[2\Phi_{cm}] \, \mathbf{C}^{+} \, [\Phi_{sm}] \qquad (30)$$

In an analogy with Eq. (26), the mean square variance for this Costas component is

$$\sigma_c^2 = \sigma_m^2(P/2 + \sigma_m^2)/n_t \qquad (31)$$

From these two equations, the signal-to-noise ratio for measuring the carrier phase can be derived:

$$SNR_c = \rho/\{1 + 1/(2R_d)\} = \rho S_{Lc} \qquad (32)$$

where $\rho = 2P \; n_a \; n_t/\sigma_o^2$ is the equivalent SNR in the final loop bandwidth, $R_d = P \; n_a/\sigma_o^2$ is the data (symbol) SNR, and $S_{Lc}$ is the squaring loss as a function of $R_d$ due to multiplication of signal by noise, and noise by noise, in the Costas-loop process:

$$S_{Lc}[R_d] = 1/\{1 + 1/(2R_d)\} \qquad (33)$$

Following an analogous procedure using Eqs. (22) and (26), the SNR for measuring the cross-correlation phase can be derived:

$$SNR_x = \rho\ (1/3)[1/\{1 + 1/(2R_d/3)\}] = \rho S_{Lx} \qquad (34)$$

The squaring loss here is the same as that for the Costas loop, except for scaling:

$$S_{Lx}[R_d] = \ (1/3)S_{Lc}[(1/3)R_d] \qquad (35)$$

The larger loss in the cross-correlation case is due to the fact that the function $\hat{\mathbf{C}}^+$ in Eq. (22) has a peak value of 2/3, while the function $\mathbf{C}^+$ in Eq. (30) has a peak value of 1. The only way to overcome this loss is by making $n_t$, the number of symbols over which integration takes place, larger. This is analogous to making the Costas loop bandwidth narrower.

Figure 3 presents plots of these two squaring losses as a function of $R_d$. It is evident that to obtain an SNR for the cross-correlation phase equal to that for the Costas-loop phase, integration must be at least eight times longer when $R_d$ is less than 1.

Because of instability in the frequency of the spacecraft transmitter, increasing the integration time (or equivalently, narrowing the loop bandwidth) of the Costas-loop case beyond about 1 sec is not possible without serious SNR loss. But for the cross-correlation case, this instability largely "common-modes out" from one antenna to the other, and therefore integration periods up to 100 sec are easily obtained with only slight SNR degradation. The limit on this integration time is set by the differential tropspheric effects between the two antennas.

## VII. Conclusions

A technique has been outlined for performing full-spectrum combining of telemetry signals. It is possible to implement a device to accomplish this type of arraying by using the DSN Advanced Receiver (ARX). ARX components would serve as the matched filters while new hardware would have to be added to provide the incoming delay and phase rotation. The real practicality for the use of full-spectrum combining in the DSN ultimately rests on the difficulty and expense of implementing these matched filters. It is expected that continued advancement in VLSI technology will easily provide inexpensive solutions in the near future.

An alternative approach for a demonstration of the basic ideas would be to record telemetry data with the DSN wide channel bandwidth (WCB) VLBI system and use the Block II VLBI correlator to accomplish the combining. Only a slight modification in hardware together with some software upgrades would be needed to test the technique. The one-bit sampling of this system, however, would limit its value as a combiner in any real application.

# References

[1] J. W. Layland, A. M. Ruskin, D. A. Bathker, R. C. Rydgig, D. W. Brown, B. D. Madsen, R. C. Clauss, G. S. Levy, S. J. Kerridge, M. J. Klein, C. E. Kohlhase, J. I. Molinder, R. D. Shaffer, and M. R. Traxler, "Interagency Array Study Report," *TDA Progress Report 42-74*, vol. April–June 1983, Jet Propulsion Laboratory, Pasadena, California, pp. 117–148, August 15, 1983.

[2] W. J. Hurd, F. Pollara, M. D. Russell, and B. Siev, "Intercontinental Antenna Arraying by Symbol Stream Combining at ICE Giacobini-Zinner Encounter," *TDA Progress Report 42-84*, vol. October–December 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 220–228, February 15, 1986.

[3] W. J. Hurd, J. Rabkin, M. D. Russell, B. Siev, H. W. Cooper, T. O. Anderson, and P. U. Winter, "Antenna Arraying of Voyager Telemetry Signals by Symbol Stream Combining," *TDA Progress Report 42-86*, vol. April–June 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 131–142, August 15, 1986.

[4] D. W. Brown, W. D. Brundage, J. S. Ulvestad, S. S. Kent, and K. P. Bartos, "Interagency Telemetry Arraying for Voyager Neptune Encounter," *TDA Progress Report 42-102*, vol. April–June 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 91–118, August 15, 1990.

[5] D. W. Brown, H. W. Cooper, J. W. Armstrong, and S. S. Kent, "Parkes–CDSCC Telemetry Array: Equipment Design," *TDA Progress Report 42-85*, vol. January–March 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 85–110, May 15, 1986.

[6] J. S. Ulvestad, "Phasing the Antennas of the Very Large Array for Reception of Telemetry From Voyager 2 at Neptune Encounter," *TDA Progress Report 42-94*, vol. April–June 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 257–273, August 15, 1988.

[7] J. Yuen, ed., *Deep Space Telecommunications System Engineering*, JPL Publication 82-76, Jet Propulsion Laboratory, Pasadena, California, p. 191, July 1982.

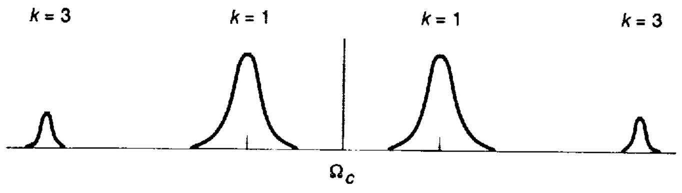


Fig. 1. Spacecraft telemetry signal.

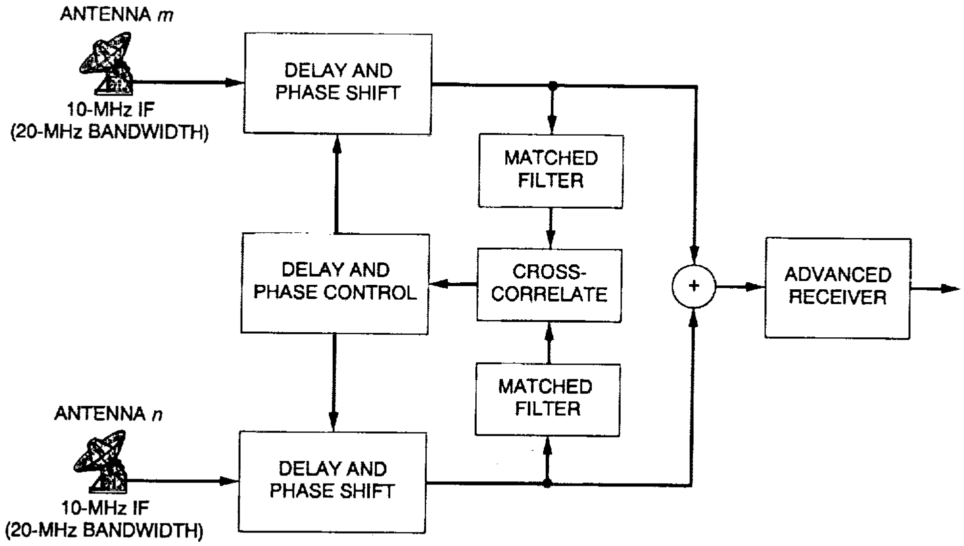


Fig. 2. Delay and phase shift.

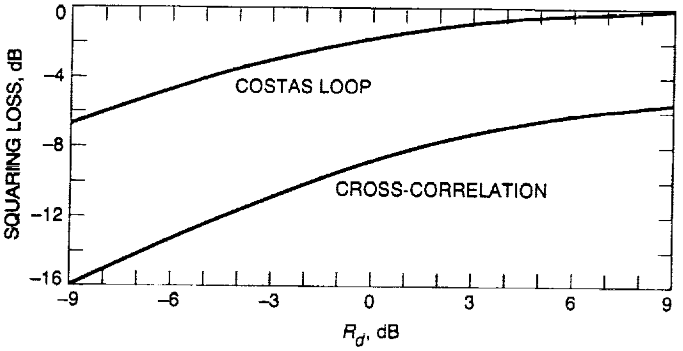


Fig. 3. Squaring loss.

S3-61
53206

p-6

# A Comparison of the Fractal and JPEG Algorithms

K.-M. Cheung and M. Shahshahani
Communications Systems Research Section

*A proprietary fractal image-compression algorithm and the Joint Photographic Experts Group (JPEG) industry standard algorithm for image compression are compared. In every case, the JPEG algorithm was superior to the fractal method at a given compression ratio according to a root-mean-square criterion and a peak signal-to-noise criterion.*

## I. Introduction

Fractal image compression has attracted much publicity in recent years. It has been suggested that one can achieve compression ratios of the order of thousands to one by the application of fractal algorithms. Some researchers successfully generated certain images, with very small databases, by using fractal algorithms. These images generally consisted of natural objects, and the memory requirement for, e.g., a realistic looking tree, was about 120 bytes. However, the applicability of these methods to general image compression and the achievement of the phenomenal compression ratios of thousands to one have been viewed with general skepticism. In order to make a comparative study of the fractal versus Joint Photographic Experts Group (JPEG) standard algorithms, the authors sent ten images to a vendor, Iterated Systems Inc., Norcross, Georgia. These images were compressed by their proprietary fractal algorithms, and reconstructed using their decompression package. The compression ratios were from about five to one to twenty to one. The mean square errors and peak signal-to-noise ratio (SNR) were compared to the corresponding ones for the JPEG algorithms at the same compression ratios. The latter approach proved to be superior in every case according to these criteria.

The theoretical foundation and the practical implementation of the fractal method for image generation is described in Section II. The relevant portion of the JPEG algorithm is briefly described in Section III, and finally in Section IV the results of the comparative study are given.

## II. Fractals

There are various ways of defining fractals. The framework proposed by J. Hutchinson [1] has been the most successful approach for the study of fractals. To describe this method let $\mathcal{S} = \{S_1, \cdots, S_n\}$ be a finite set of affine transformations of $\mathbf{R}^q$. This means that if $x \in \mathbf{R}^q$, then the effect of $S_i = (A_i, v_i)$ on $x$ is given by

$$S_i(x) = A_i(x) + v_i$$

where $A_i$ is a linear transformation of $\mathbf{R}^q$ and $v_i \in \mathbf{R}^q$. It will be assumed that $A_i$'s are nonsingular and $S_i$'s are contracting, i.e., $\|S_i(x) - S_i(y)\| < \|x - y\|$ for all $x, y \in \mathbf{R}^q$. Then the affine transformation $S_{i_1} \cdots S_{i_k}$ is also contracting and, therefore, has a unique fixed point that will be denoted by $F_{i_1,\cdots,i_k}$. The fractal set $\mathcal{F}(\mathcal{S})$ associated to $\mathcal{S}$ is, by definition,

$$\mathcal{F}(\mathcal{S}) = \text{closure}(\{F_{i_1,\cdots,i_k} \mid \text{all } 0 \le i_1,\cdots,i_k \le n$$

$$\text{and } k = 1,2,3,\cdots\})$$

$\mathcal{F}(\mathcal{S})$ is a compact subset of $\mathbf{R}^q$. The *self-similarity* property of fractals is expressed by the fundamental equation

$$\mathcal{F}(\mathcal{S}) = \bigcup_{i=1}^{n} S_i[\mathcal{F}(\mathcal{S})]$$

In fact,

**Theorem 1.** From [1], $\mathcal{F}(\mathcal{S})$ is the unique compact set $K \subset \mathbf{R}^q$ with the property

$$K = \bigcup_{i=1}^{n} S_i(K) \tag{1}$$

Notice that each $S_i(K)$ is a replica of $K$, so that Eq. (1) does indeed express the self-similarity property of fractals. This characterization of fractals is also important in practical applications.

**Example 1.** Let $K$ be a convex polygon in $\mathbf{R}^2$ with vertices $v_1,\cdots,v_n$, and let $S_1,\cdots,S_n$ be the affine transformations $S_i(x) = v_i + \alpha_i(x - v_i)$, where $0 < \alpha_i < 1$. If the $\alpha_i$'s are not too small, then $\bigcup S_i(K) = K$ and, consequently, $\mathcal{F}(\mathcal{S}) = K$ by Theorem 1 of fractal sets. Thus, every convex polygon can be realized as an $\mathcal{F}(\mathcal{S})$ for some $\mathcal{S}$. On the other hand, it is not hard to see that the boundary $\partial K$ of the convex polygon $K$ *cannot* be realized as a fractal set.

To generate a fractal image, one starts with a set $\mathcal{S} = \{S_1,\cdots,S_n\}$ of affine transformations. For every product $S_{i_l}\cdots S_{i_2}S_{i_1} = S = (A,v)$ of length $l \le k$, for some pre-assigned value $k$, one computes its unique fixed point $F_{i_l},\cdots,F_{i_1} = F = (I-A)^{-1}(v)$. Note that since $S$ is contracting, $I - A$ is invertible. The coordinates $F_1$ and $F_2$ of $F$ are multiplied by a normalizing factor $N$, and then quantized to the nearest integer to obtain a point $p$ with integral coordinates $(p_1,p_2) \in \mathbf{Z}^2$. A point $q = (q_1,q_2)$ on the screen is white (black) as it is (or is not) of form $(p_1,p_2)$, as described above.

The fractal set $\mathcal{F}(\mathcal{S})$, thus constructed, corresponds to a black-and-white image. In order to introduce grey levels into $\mathcal{F}(\mathcal{S})$, one would like to use the density of the points $\{F_{i_1},\cdots,F_{i_k}\}$ as a measure of the brightness of the pixels. To do so, it is convenient to regard the procedure for the generation of a fractal set as a random process. In fact, let $p_1,\cdots,p_n$ be positive real numbers such that $\Sigma p_i = 1$, and $x \in \mathbf{R}^q$. Consider the random process $\mathcal{X}$ where $x \to S_i(x)$ with probability $p_i$. This process has a unique stationary distribution $\mu$. The stationarity condition is expressed by the equation

$$\mu = \Sigma\, p_i S_{i*}(\mu) \tag{2}$$

where $S_{i*}(\mu)$ is the transform of the measure $\mu$ under the affine transformation $S_i$. Note that Eq. (2) is a more precise version of the fundamental self-similarity property expressed by Theorem 1. The measure $\mu$ is the mathematical representation of a grey-scale image on the screen. The support of the measure $\mu$ is the fractal set $\mathcal{F}(\mathcal{S})$, and is independent of the choice of positive numbers $p_i$. Furthermore, if $\mathcal{E}_k$ denotes the discrete probability measure naturally assigned to the set $\{F_{i_1},\cdots,F_{i_l} \mid l \le k$ and all $i_1,\cdots,i_l\}$, then

$$\mathcal{E}_k \to \mu \quad \text{weakly}$$

if all $p_i = 1/n$.

By introducing an alternative method for generating the fractal set $\mathcal{F}(\mathcal{S})$, one can take advantage of the probabilities $\{p_i\}$ in actual image generation. Consider a realization of the random process $\mathcal{X}$. This can be interpreted as follows: Using a random-number generator, one generates a sequence of integers $\{i_1, i_2, \cdots\}$, where $1 \le i_k \le n$ and the integer $j$ is chosen with probability $p_j$. A realization of the process $\mathcal{X}$ is then the sequence of points $S_{i_1}(z), S_{i_2}S_{i_1}(z), \cdots$. Let $\mathcal{M}_k$ be the discrete probability measure assigned to the set $\{S_{i_1}(z), S_{i_2}S_{i_1}(z), \cdots, S_{i_k}\cdots S_{i_2}S_{i_1}(z)\}$. Then, one can show by standard arguments that

**Theorem 2.** With probability 1, $\mathcal{M}_k \to \mu$ weakly.

After possibly multiplying by a factor $N$ and quantizing, one can regard a point $S_{i_l}\cdots S_{i_2}S_{i_1}(z)$ in a realization of the process $\mathcal{X}$ as the point $q_{i_l,\cdots,i_1} = q = (q_1,q_2) \in \mathbf{Z}^2$. One can make a histogram of the number of times the points of the integer lattice $\mathbf{Z}^2$ are generated in this fashion. Grey levels are accordingly assigned to the points so that the points generated more often have higher intensity (are whiter) than those generated fewer times. The additional input consisting of the positive numbers $\{p_1,\cdots,p_n\}$

allows one to have better control over the density of the points, i.e., the grey levels.

In order for any procedure to have practical applications, it must be stable relative to the variation of the parameters. In the case of fractals, one would like the fundamental property of self-similarity, as expressed by Theorem 1, to have the necessary stability. This means that if the transformations $S_i$ are such that $\bigcup_{i=1}^{n} S_i(K)$ is approximately identical with $K$, then the fractal set $\mathcal{F}(\mathcal{S})$ is also approximately identical with $K$. It is not difficult to establish, in a quantitative sense, the validity of this stability, which yields a slight generalization of Theorem 1. It is this mild generalization of Hutchinson's theorem that has been widely publicized by M. Barnsley as the *Collage Theorem* [2].

Just as in Example 1, where Theorem 1 was used to generate convex polygons as fractal sets, other natural-looking objects were generated by appealing to the approximate version and the experimental insight into the effects of the variation of the parameters on $\mu$. In actual applications, it is also convenient to decompose the matrix $A_i$ as a product

$$A_i = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b\alpha & 0 \\ 0 & b'\alpha \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

where $|bb'| = 1$ and $0 < \alpha < 1$. The advantage of using this decomposition is that the effects of the parameters can be more readily understood. The matrix

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is a rotation through angle $\theta$. The matrix

$$\begin{pmatrix} b\alpha & 0 \\ 0 & b'\alpha \end{pmatrix}$$

represents scaling by factors $b\alpha$ and $b'\alpha$, and

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$$

can be regarded as a *twist*. The numbers $b$, $b'$, and $\alpha$ should be chosen such that $|b\alpha|$, $|b'\alpha| < 1$. By specializing the parameters $b$, $b'$, and $a$ to 1, 1, and 0, respectively, one can already generate a number of very complex and natural-looking images. The interested reader should view these points as hints or general guidelines for experimentation with fractal image generation.

Using the fractal method, a number of images were generated by one of the authors and others (see, e.g., [2] and [3]). Fractal methods have been used recently at JPL for computer simulation of certain images. Since these images were generated by storing only the parameters of a few affine transformations, one would like to believe that the fractal method can be used for efficient data, or more specifically, image compression. The systematization of the fractal method so that it becomes applicable to general image compression has been attempted by a number of researchers. One approach is to try to find a set $\mathcal{S} = \{S_1, \cdots, S_n\}$ of affine transformations and positive numbers $\{p_i\}$ such that the corresponding stationary distribution is a good approximation to a given image. While in theory this is possible, the number of the affine transformations may be so large that the result will have no practical value. Furthermore, since most objects do not exhibit, even remotely, the self-similarity property that is an essential feature of fractals, this direct approach is probably a futile one. By segmentation of an image, one will have better control over the choice of the affine transformations. However, the startling compression ratios of thousands to one will not be achievable. The most successful attempt of the application of fractals to image compression has been by Iterated Systems Inc. While their methodology is a well-guarded trade secret, the authors have tested the results of their fractal compression scheme against the JPEG baseline algorithm. This will be discussed in more detail in Section IV.

## III. JPEG Algorithms

With the advent of multi-media services offered by the 64-Kbit/sec Integrated Services Digital Networks (ISDN), there is a strong urge to define a standard for applications as diverse as photo-videotex, desktop publishing, graphic arts, color facsimile, photojournalism, medical systems, and many others. The JPEG was formed under the joint auspices of the International Standards Organization (ISO) and the Comité Consultatif International de Téléphone et Télécommunication (CCITT) at the end of 1986 for the purpose of developing an international standard for the compression and decompression of continuous-tone, still-frame, monochrome, and color images.

The JPEG-proposed algorithm has three major components. The first is a baseline system that provides a simple and efficient algorithm that is adequate for most image-coding applications. The second is a set of extended system features that allows the baseline system to satisfy a broader range of applications. Among these optional features are 12-bit/pixel input, progressive sequential and hierarchical build-up, and arithmetic coding. The third is an independent, differential, pulse-code modulation (DPCM) scheme for applications that require lossless compression. The following is a brief description of the JPEG baseline system only. The reader is referred to the JPEG proposal [4] for a complete description of the components and the algorithms.

The JPEG baseline system is a transform-based algorithm consisting of three stages. The first stage is a discrete cosine transform (DCT). The output of the DCT is then quantized and in the final stage the quantized output is encoded by variable length codes. The original image is partitioned into 8 × 8 pixel blocks and each block is independently transformed by the DCT. The transform coefficients are then quantized using a user-defined quantization template that is fixed for all blocks. Each component of the quantization template is an 8-bit integer and is passed to the receiver as part of the header information that is required for every image. Up to four different quantization templates can be specified; for example, different quantization templates may be used for the different components of a color image. The JPEG baseline system supplies two default quantization templates: one for the luminance component (the $Y$-component) and the other for the two chrominance components (the $I$ and $Q$ components). The top-left coefficient in the two-dimensional DCT block [i.e., the (0, 0) coefficient] is the DC coefficient and is proportional to the average brightness of the spatial block. The remaining coefficients are called the AC coefficients. After quantization, the DC coefficient is encoded with a lossless DPCM scheme using the quantized DC coefficient from the previous block as a one-dimensional predictor. For the baseline system, up to two separate Huffman tables for encoding the resulting differential signal can be specified in the header information. A default Huffman table for DC encoding is given in the JPEG proposal. The encoding of the quantized AC coefficients uses a combination of runlength and Huffman coding techniques. There are many zeros in the quantized AC coefficients, especially in the high frequencies. The AC coefficients that are close (respectively, far) in location to (0, 0) are the low (respectively, high) frequencies. Typically, high frequencies have low energies. The two-dimensional block of quantized coefficients is transformed into a one-dimensional vector using a zigzag reordering so that the coefficients are arranged in approximately decreasing order of their average energy. This creates a combination of nonzero values at the beginning of the vector and long runs of zeros thereafter. To encode the AC coefficients, each nonzero coefficient is first described by a composite 8-bit value, denoted by $I$, of the form (in binary notation)

$$I = NNNNSSSS$$

The four least significant bits, $SSSS$, define a category for the coefficient amplitude. The values in category $k$ are in the range $(2^{k-1}, 2^k - 1)$ or $(-2^k + 1, -2^{k-1})$, where $1 \leq k \leq 10$. Given a category, it is necessary to send an additional $k$ bits to completely specify the sign and magnitude of a coefficient within that category. The four most significant bits in the composite value, i.e., $NNNN$, give the position of the current coefficient relative to the previous nonzero coefficients. The runlengths specified by $NNNN$ can range from 0 to 15, and a separate symbol $I = 11110000 = 240$ is defined to represent a runlength of 16 zeros. In addition, a special symbol, $I = 0$, is used to code the end of a block (EOB), which signals that all the remaining coefficients in the block are zero. Therefore, the total symbol set contains 162 members (10 categories × 16 runlength values + 2 additional symbols). The output symbols for each block are then Huffman coded and are followed by the additional bits required to specify the sign and exact magnitude of the coefficient in each of the categories. Up to two separate Huffman tables for the AC coefficients can be specified in the baseline system. A default runlength/Huffman table for AC encoding is given in the JPEG proposal.

## IV. The Comparison

For the comparison of the fractal and the JPEG algorithms, a set of ten 320 × 200 8-bit grey-scale images in Sun raster format was sent to Iterated Systems Inc. Some of these images are often used for image-compression experiments, and others are planetary images. The complete list is (1) an air scene, (2) baboons, (3) a couple in a room, (4) Lena (portrait of a young woman often used in image compression tests), (5) peppers, (6) light effects pattern, (7) the moon, (8) Miranda, and (9) and (10) two images of the planet Saturn. Each of the images was compressed to eleven files of sizes ranging from 3 to 13.5 Kbytes. The reconstruction was done by the decompression software *P.OEM*$^{\text{TM}}$ *Developers' Kit—Grayscale Still* developed by Iterated Systems Inc. for this purpose and the format conversion developed by one of the authors. The reconstructed images were then compared to the original to obtain the mean-square-error (MSE) and the peak-SNR values versus the number of bits per pixel (i.e., compression ratio). Peak SNR is defined as $10\log_{10}(255)^2/MSE$, and is measured in dB's. Similar curves were obtained for the JPEG baseline algorithm as implemented at JPL [5]. The results are given in Figs. 1 and 2 for images (5) and (8). It is clear from the figures that the JPEG algorithm is superior to the fractal algorithm by as much as 6 dB at low compression ratios and 2 dB at high compression ratios. The curves presented in this article contain the range of compression ratios obtained by Iterated Systems Inc. for the ten images referred to earlier. It should be pointed out that no conclusion regarding the effectiveness of the fractal method at higher compression ratios is warranted at this time. Qualitatively, at high compression ratios of about 16:1 to 20:1, the fractal scheme exhibits *solar flare*, while the JPEG baseline algorithm suffers from the *tiling effect.* No post-processing was done with the application of the JPEG algorithm. The authors are unaware of any post-processing in the fractal algorithm. The tiling effect can be somewhat alleviated by the application of certain filters which involve weighted averages of nearby neighbors. The authors know of no method for dealing with the solar flare effect.

## V. Conclusion

A comparative study of the JPEG baseline algorithm and the fractal method was conducted by the authors. In terms of the mean square error and peak SNR, the JPEG algorithm was superior in every case.

## References

[1] J. Hutchinson, "Fractals and Self-Similarity," *Indiana University Journal of Mathematics*, vol. 30, no. 5, pp. 713–747, 1981.

[2] M. Barnsley, *Fractals Everywhere*, New York: Academic Press, 1988.

[3] P. Diaconis and M. Shahshahani, "Products of Random Matrices and Computer Image Generation," *Contemporary Mathematics*, vol. 50, pp. 163–173, American Mathematical Society, Providence, Rhode Island, 1986.

[4] W. B. Pennebaker to ISO Group X3L2.8, *JPEG Draft Technical Specification Revision 8*, IBM, Yorktown Heights, New York, August 17, 1990.

[5] F. Pollara and S. Arnold, "Emerging Standards for Still Image Compression: A Software Implementation and Simulation Study," *TDA Progress Report 42-104*, vol. October–December 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 98–101, February 15, 1991.

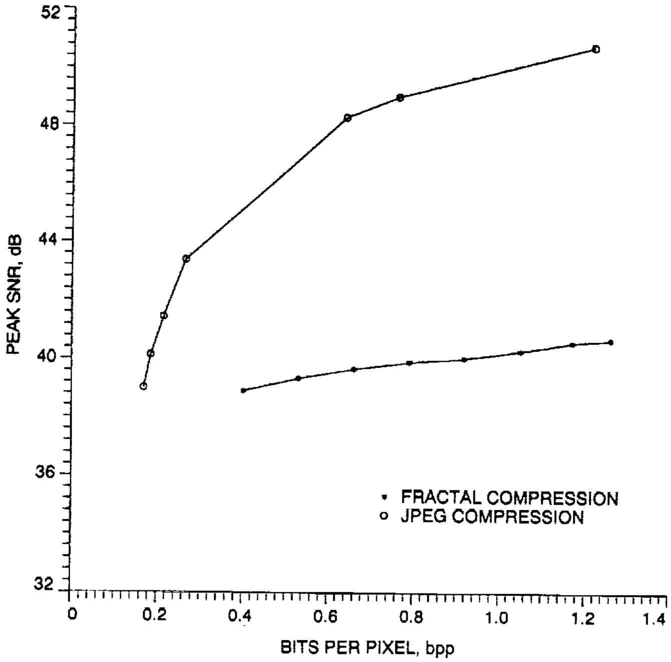


**Fig. 1. Fractal performance on peppers.**

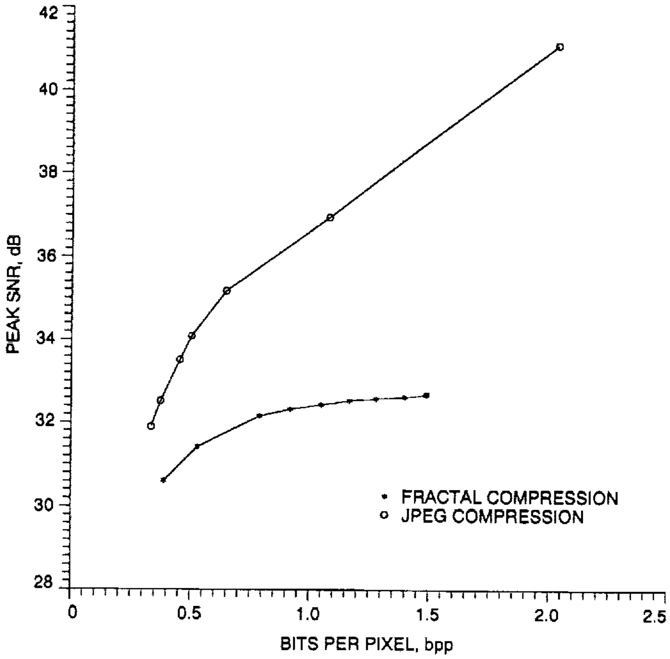


**Fig. 2. Fractal performance on Miranda.**

November 15, 1991

# A Minimalist Approach to Receiver Architecture[1]

O. Collins
Johns Hopkins University
Baltimore, Maryland

*This article describes new signal processing techniques for DSN radios and presents a proposed receiver architecture, as well as experimental results on this new receiver's analog front end. The receiver's design employs direct downconversion rather than high speed digitization, and it is just as suitable for use as a space-based probe relay receiver as it is for installation at a ground antenna. The advantages of having an inexpensive, shoe-box-size receiver, which could be carried around to antennas of opportunity, used for spacecraft testing or installed in the base of every antenna in a large array are the force behind this project.*

## I. Introduction

This article reports on research in progress into the design of a smaller, cheaper, and more reliable receiver for the Deep Space Network. Here, a receiver is defined as everything in the chain between the output of the first noise-floor-setting amplifier stage (the maser in the DSN) and the input to the telemetry detection and decoding equipment. The hardest link to forge in this chain is the receiver's analog front end, where the rf signal is downconverted and filtered before it is digitized. Therefore, this article concentrates on the basic signal processing scheme and analog hardware of the receiver, but not on the complex mixer to process the digitized outputs from the receiver's analog front end. The design presented exactly corrects amplitude and phase mismatch at one frequency, the carrier, or a calibration frequency; digital equalization can then eliminate phase and amplitude irregularities over the whole passband.

The proposed receiver architecture is a departure from the conventional approach to high-performance digital receiver design. Most designs push the point at which the signal is digitized closer and closer to the rf carrier frequency, grabbing each new advance in the speed of analog-to-digital conversion and digital signal-processing equipment, and exchange analog mixers and filters for their digital counterparts. Well-founded faith in the ability of digital electronics to process signals without degradation steers people in this direction. All such designs will, however, be limited by their sample-and-hold circuits. In fact, analog-to-digital converters are never really the limitation, since if sufficiently fast sample-and-hold circuits are available, they can simply be paralleled in order to capture higher bandwidth signals. For example, today's fastest digital oscilloscopes use four sample-and-hold amplifiers, which are fired in cascade. A sample and hold is very similar to a mixer, yet harder to build. Both operations can be implemented with a switch that opens and closes repetitively. However, the duty cycle of the switch in the mixer is 50 percent, whereas that in the sample and hold has to be made as short as possible.

[1] Work performed under a contract between JPL and Johns Hopkins University through the TDA Office.

The design outlined below replaces a high-speed sample and hold with a pair of mixers in a direct downconversion configuration. Preliminary results show that 80 dB of spurious-free dynamic range, far more than is necessary to cope with all except the very purest sources of interference, is achievable.

## II. Receivers

The task of a receiver is in effect to apply a matched filter to each symbol of the transmitted data stream. If the communications link is operating without coding, then the output of this filter will be a 1 or a 0 if the transmitter alphabet has only two letters and will represent one bit of information transmitted. For a larger transmitter alphabet, the output of the matched filter will be the index of the signal that was most likely sent. When the link uses coding, the receiver may benefit from not discarding so much of the analog information available about the signal. For example, for a binary signal set, the output of the matched filter should be a quantized version of the probability that a 1 was more likely sent than a 0, or something directly related to it. Binary phase-shift keying (BPSK) is one of the simplest forms of modulation used and is usually the choice for a communications system in which bandwidth is of little consequence, such as the link from a deep-space probe to the Earth. The signal structure and matched filter for BPSK in the presence of Gaussian noise appear in Fig. 1.

Each symbol is correlated with the transmitter's carrier. The result of this correlation is directly related to the probability that the symbol sent was a 1, e.g., if the voltage is positive then a 1 was more likely sent than a 0.

The block diagram of the matched filter in Fig. 1 is appealingly simple, and this article presents some experimental investigations into how closely a real receiver can approach Fig. 1's simplicity. Inasmuch as a balanced mixer is an analog multiplier, one might be tempted to try implementing the block diagram directly. The flaw with this scheme is the requirement that the mixer and subsequent very high-gain amplifier perform well at DC, where $1/f$ noise causes difficulties. That is, many data patterns will place significant energy very close to the carrier. This dilemma can be solved by proper coding; however, the remaining problem of carrier tracking, i.e., producing a replica of the transmitter's carrier at the receiver, is extremely hard to solve, since the problems of shielding and component nonidealities are close to intractable.

The superheterodyne receiver shown in Fig. 2 is the traditional solution. This receiver isolates a piece of the radio frequency spectrum, which contains the desired signal and converts it down in frequency while also amplifying it. The staged downconversion and amplification eliminate the hazard of feedback and raise the signal level sufficiently so that, by the time its frequency gets close to DC, $1/f$ amplitude noise in the mixer diodes and the following amplifiers is no longer important. Unfortunately, each stage of conversion is not only expensive, but also degrades the signal because the multiplications are imperfect. The sharpness of the available filters determines how many stages are necessary to drop from the rf carrier frequency to baseband. Thus, either higher $Q$ filters or a higher digitizing rate are necessary to simplify the analog electronics in Fig. 2. The narrower an analog filter is, the harder it is to make it linear phase, i.e., to make it behave like a uniform delay for all frequencies in its passband. If this distortion were constant, it would not pose a problem since no information is lost, and, in theory, a digital equalizer could be used to correct the phase distortions of the analog filter. Unfortunately, high-$Q$ analog filters change with temperature and are not uniform over their passbands. Thus, even if the filter were in a thermostatically controlled oven, the equalizer would have to change every time the receiver was tuned to a different frequency.

## III. New Approaches

This section presents a receiver design in which only low-$Q$, low-pass filters are needed, and most of the amplification takes place at a low enough frequency so that oscillation caused by feedback is not a problem. Only an extremely small amount of the incoming signal energy will appear close to DC. Even this loss could be eliminated by proper shaping of the transmitted signal. (For example, the pinned state convolutional codes described in [1], which the Galileo Big Viterbi Decoder is already able to decode, have this spectrum shaping capability.) The basic block diagram appears in Fig. 3.

The incoming signal is split in two and each half is mixed with one of two carriers which have the same frequency, but are as close to 90 deg out of phase as possible. The outputs from these two mixers are amplified, low-pass filtered, and digitized. The only filters necessary are the anti-aliasing filters for the analog-to-digital converters, and as each mixer sees only half the incoming power, dynamic range requirements are eased. If $\beta = 0$, then the two low-pass filtered signals are merely the two components of the Hadamard transform of the input signal and can be processed as shown in Fig. 4 to recover the spectrum one filter width to either side of the local oscillator.

This processing scheme is simply that of the image rejection mixer which has been around since the 1950s [2,3] and has been well studied. Researchers have constructed whole receivers based on this idea [4,5]. The most recent and successful design is in [6].[2] These designs, however, were all limited to fairly low frequencies (<70 MHz) and to applications where keeping costs low was more important than achieving the ultimate in performance. The crucial difficulty was that of making accurate quadrature power splitters—i.e., keeping $\beta$, the offset from perfect quadrature, very close to zero. These designs did not recognize the capability of a change in low-frequency signal processing to compensate for imperfections in the quadrature power splitter. When the quadrature power splitter is imperfect, all of the information in the original signal is still present, but the axes of the coordinate system in which the signal is viewed are no longer the usual orthogonal $X$ and $Y$.

Figure 5 shows a simple geometric way of viewing the signal in its skewed reference frame. Both $X$ and $P$ are the outputs of the two receiver channels. For clarity, Fig. 5 shows the complement of $\beta$, called $a$. Here 74 deg is greater than the worst skew that might occur in practice. Elementary trigonometry demonstrates that $Y$, the output of a channel in perfect quadrature, can be obtained easily as a linear combination of the two measured quantities $X$ and $Y$.

These straightforward results prove that proper baseband processing can eliminate the effects of small phase errors in the in-phase and quadrature local oscillators. Amplitude mismatches are of course even easier to deal with, as a simple rescaling is all that is required. Figure 6 shows the exceptionally simple baseband processing scheme necessary to eliminate both amplitude and phase mismatches, leaving a perfect in-phase and quadrature pair ready to feed into a complex mixer. Figure 3 envisions that all this processing will be done digitally; however, there is no reason why equivalent analog techniques would not work.

Unfortunately, the gain and phase mismatches will probably not remain constant over a reasonable operating temperature range and certainly not remain constant as the receiver is tuned. Thus, the constants in Fig. 6 really experience small variations with time scales on the order of a minute, and the remaining challenge is figuring out what these mismatches are in real time. One possibility is to switch periodically the input of the receiver over to a calibration oscillator. This oscillator would have no stringent stability requirements since only the relative phases and amplitudes of the two channels are important. A more elegant solution, however, is to synchronously detect a continuously modulated tone. The amplitude of the tone will be low enough so that it can be subtracted out digitally without degrading the information-bearing signal, and the synchronous detector will ignore interference. Thus, calibration and reception can proceed simultaneously. Exactly the same technique can be used to calibrate the delay through the whole receiver system.

The two anti-aliasing filters ahead of the analog-to-digital converters will, of course, not be absolutely the same. These manufacturing imperfections will introduce, between the channels, a phase shift which varies over the filters' passband. Digital equalization can, however, compensate for these imperfections. The reason is that, just as with channel-to-channel phase mismatch, the phase shift does not cause any loss of information. Although digital equalization could, in principle, be used in any receiver, the direct downconversion architecture makes it completely practical, since the filters are low $Q$ and thus stable with time and temperature. Also, the equalizer does not have to change as the receiver is tuned. One way to adjust this equalizer is to sweep the calibration tone.

## IV. Hardware

Elementary trigonometry demonstrates the soundness of the correction shown in Fig. 5. However, these new ideas concern radio frequency circuits where poorly modeled effects can be very important. Thus, it is not obvious that machinery that implements the different block diagrams can in fact be built. Some sort of experimental verification is necessary. All sorts of mundane but thorny engineering complexities presented themselves in the construction of a prototype. The proposed design requires the low-noise, low-distortion amplification of signals ranging from audio to a few megahertz in the presence of rf local oscillator feed-through from the mixer, which is more than 100 dB stronger. Similarly, the diode quads used in the mixers must perform reasonably close to DC. The extremely rapid plunge in frequency has, however, some advantages that ease requirements on the mixers and amplifiers. Certain distortion products that plague multi-stage superheterodyne designs, especially those which use a stage of upconversion to improve image rejection, are completely absent in this direct downconversion approach. Most of the amplification takes place at sufficiently low frequencies so that substantial amounts of feedback can be used to reduce distortion. A great deal of effort has gone into the design of such amplifiers for transcontinental coaxial cables [7,8].

[2] Amplification of this reference by a personal communication from Polly Estabrook in the Communications Systems Research Section, Jet Propulsion Laboratory, Pasadena, California is acknowledged.

Construction of one of the two channels shown in Fig. 3 began in late 1990 in order to evaluate the following: the susceptibility of the mixers to second-order intermodulation products caused by radio frequency interference, the phase and amplitude stability of the mixers and filters, and the intricacies of the detailed design of the impedance-matching network and low-noise amplifier following the mixer. The design used a diode quad driven by a 1.8-volt local oscillator as the mixer. It was tested at both 100 and 500 MHz. These frequencies were chosen because of test equipment limitations, rather than because of any intrinsic circuit limitations. Without improvement, the device should work well above 1 GHz, but will still require an initial stage of downconversion in order to operate at X-band (8.5 GHz). Going directly from X-band to baseband would have its fascinations, but would also require that the receiver sit right on the back of the maser, as cabling losses are too great to transmit the signal any distance. If the low-noise amplifier has a temperature of 20 K, then 40 dB of amplification is necessary before the signal enters the mixer in order to override the attenuation of the mixer and the noise of the subsequent low-frequency amplifiers. With more effort, the noise figure at the mixer input could be reduced by 13 dB; the amount of amplification necessary would then drop by a similar amount. Such an improvement would be useful for a spacecraft command or probe relay receiver, but would not be of much use to the ground-based DSN.

The construction of the device presented many sticky problems in analog electrical engineering, though none was sufficiently interesting to report on. Figure 7 shows the disconcerting first view of the baseband spectrum. Power line pickup by the impedance-matching network following the mixer caused the spikes. Many iterations later, the same spectrum analyzer produced Fig. 8, which shows 80 dB of clear two-tone dynamic range. Two pure tones were fed into the mixer through a power combiner. The harmonics were produced by the test sources themselves, not by the mixer or amplifier.

Figure 9 demonstrates the design's freedom from second-order intermodulation. The mass of tones on the right is a 90-percent AM-modulated carrier. The single peak on the left comes from a second signal generator, which was set 70 dB down in order to check for overload in the spectrum analyzer. Figure 10 shows the first of the dual-channel tests and demonstrates the amplitude tracking of the pair of low-pass filters. Their tracking relative to one another, the important criterion when digital equalization is used, is even better. The difference varied by less than 0.005 dB, the limit of measurement, while the temperature of the two filters cycled over a 10-deg C range. When a similar plot for phase is available, the performance of the complete receiver can be predicted with confidence.

## V. Conclusion

This article has presented a new way of digitally dealing with the imperfections of the analog hardware in a communications receiver, and has presented sufficient measurements to afford reasonable confidence that the overall architecture has no disastrous flaws. The design is unique in its incorporation of self calibration. The correction for unknown cable transmission delays between the receiver and antenna can also be handled by the same techniques used to measure the phase shift through the pair of low-pass filters. Cutting the number of analog filters in the design to the bare minimum and placing these at a point where the signal frequency does not change with receiver tuning gives the receiver great stability. All of the first-order (i.e., not time varying) imperfections in the two analog filters can be equalized out digitally, leaving only gradual second-order drifts to contend with. The great intrinsic stability of this receiver architecture makes it also suitable for use as a radio science receiver.

# References

[1] O. Collins, "Coding Beyond the Computational Cutoff Rate," Ph.D. dissertation, California Institute of Technology, Pasadena, California, 1989.

[2] D. E. Norgaard, "The Phase-Shift Method of Single Sideband Signal Generation," *Proceedings of the IRE*, vol. 44, pp. 1718–1735, December 1956.

[3] D. K. Weaver, "A Third Method of Generation and Detection of Single Sideband Signals," *Proceedings of the IRE*, vol. 44, pp. 1703–1705, December 1956.

[4] D. Treleaven and D. Wadsworth, "FSK Receiver Uses Direct Conversion," *R. F. Design*, San Diego: Cardiff Publishing Company, pp. 30–38, July 1986.

[5] J. K. Goatcher, M. W. Neale, and I. A. W. Vance, "Noise Considerations in an Integrated Circuit VHF Radio Receiver," *Proceedings of Conference on Radio Receivers and Associated Systems*, IERE no. 50, pp. 49–59, 1981.

[6] P. Estabrook, "The Design of a Mobile Radio Receiver Using a Direct Conversion Architecture," Ph.D. dissertation, Stanford University, Stanford, California, 1989.

[7] R. F. Molander "Amplification and Regulation—the L-4 Line Repeaters," *Bell Labs Record*, vol. 45, pp. 225–229, July/August 1967.

[8] "The L-5 Coaxial Cable System," *Bell System Technical Journal*, vol. 53, pp. 1897–2267, December 1974.

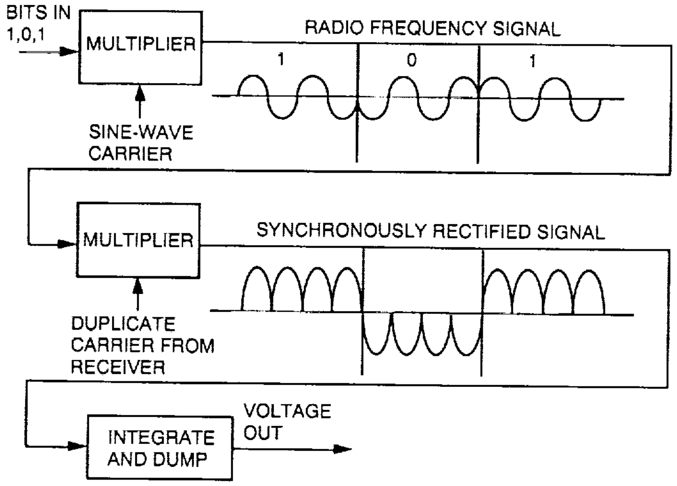


**Fig. 1. BPSK detection.**

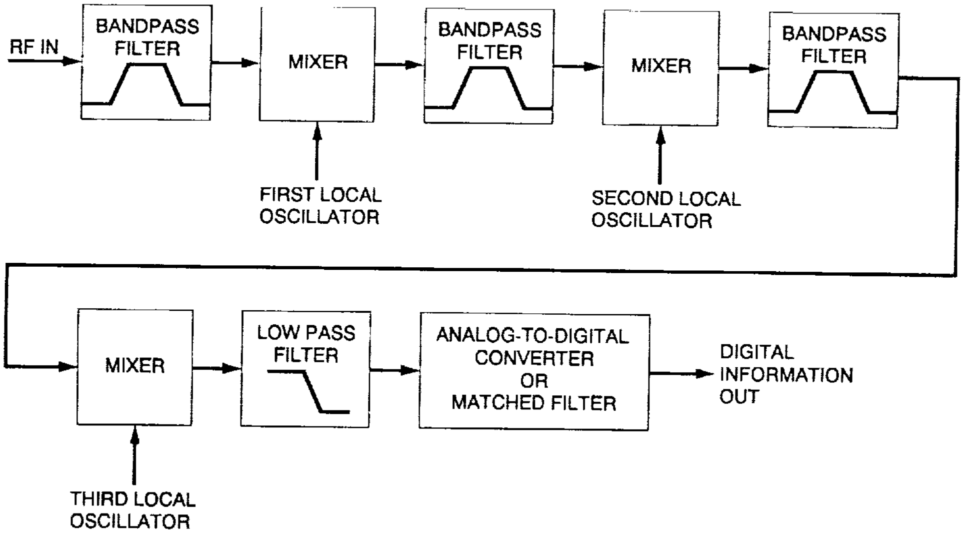


**Fig. 2. Staged downconversion.**

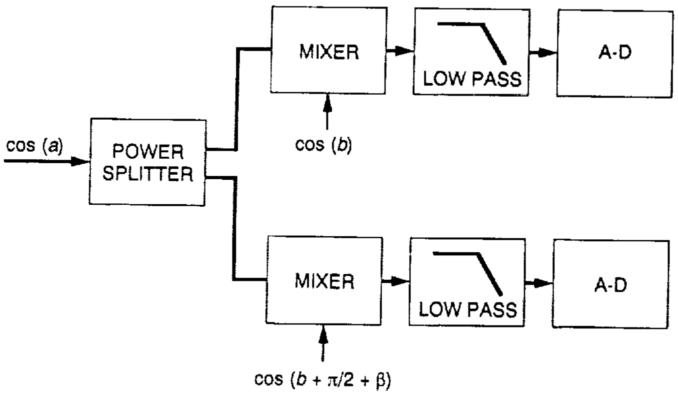


**Fig. 3. Direct downconversion architecture.**

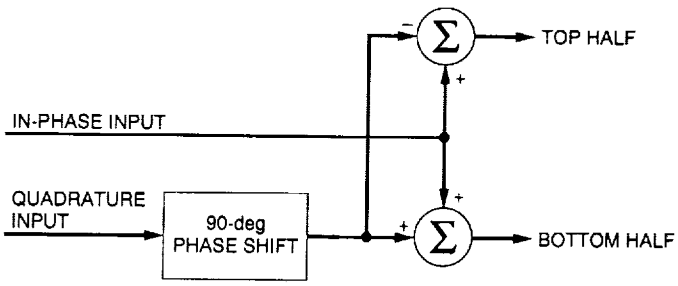


**Fig. 4. Two image rejection mixers in one.**

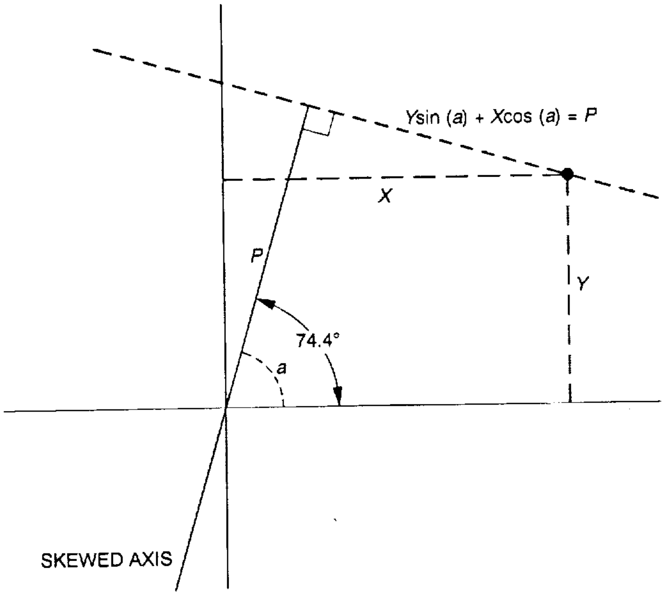


Fig. 5. Signal geometry.

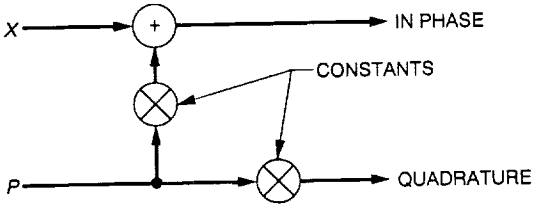


Fig. 6. Phase corrector.

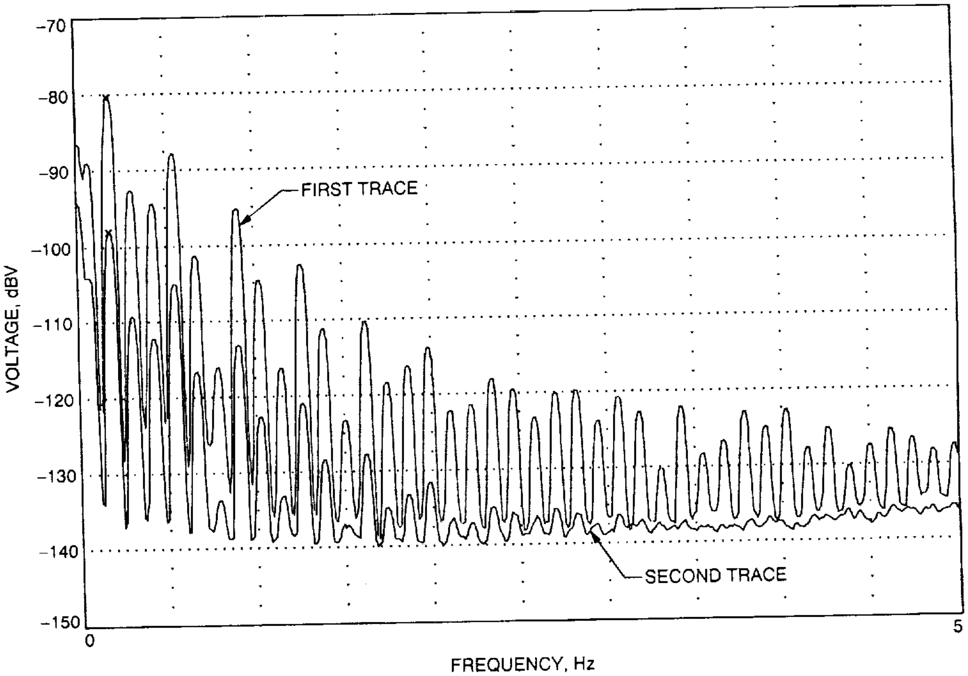


Fig. 7. Second-order distortion.

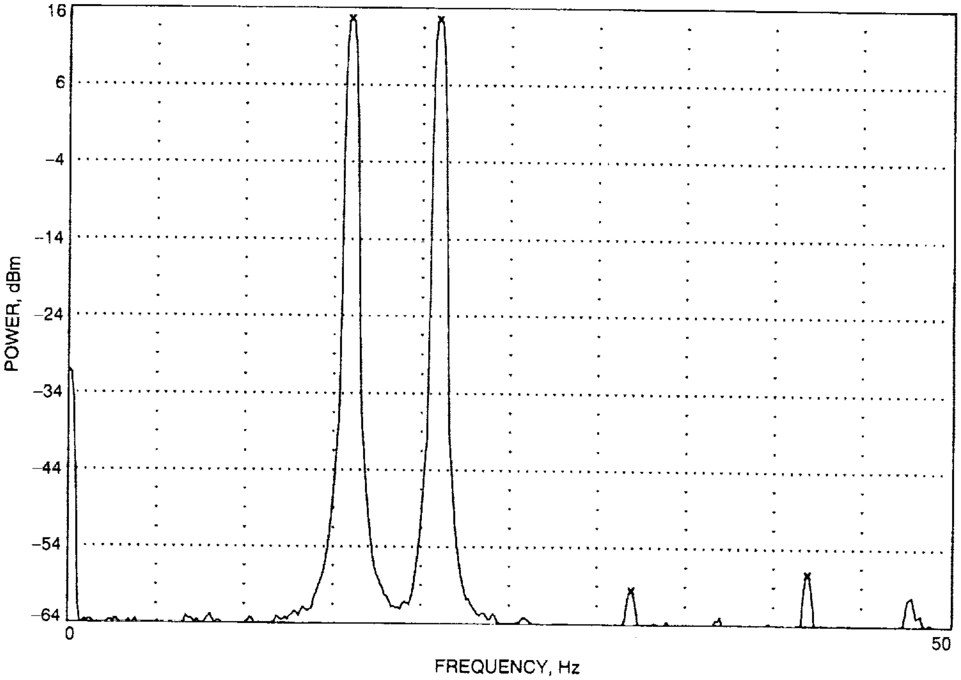


Fig. 8. Two-tone intermodulation distortion.

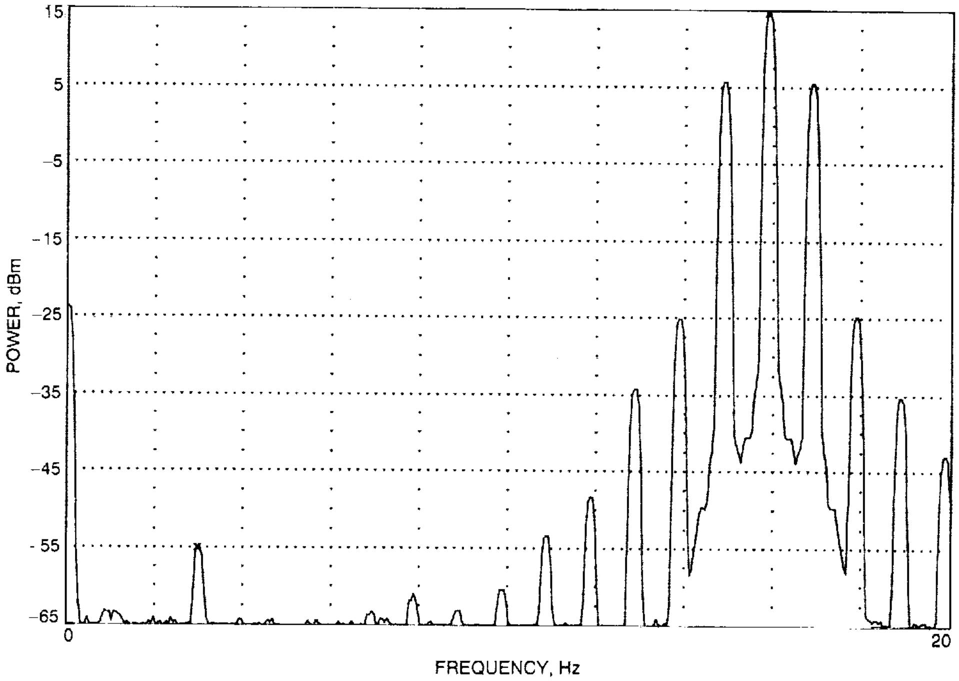


Fig. 9. Filter magnitude match.

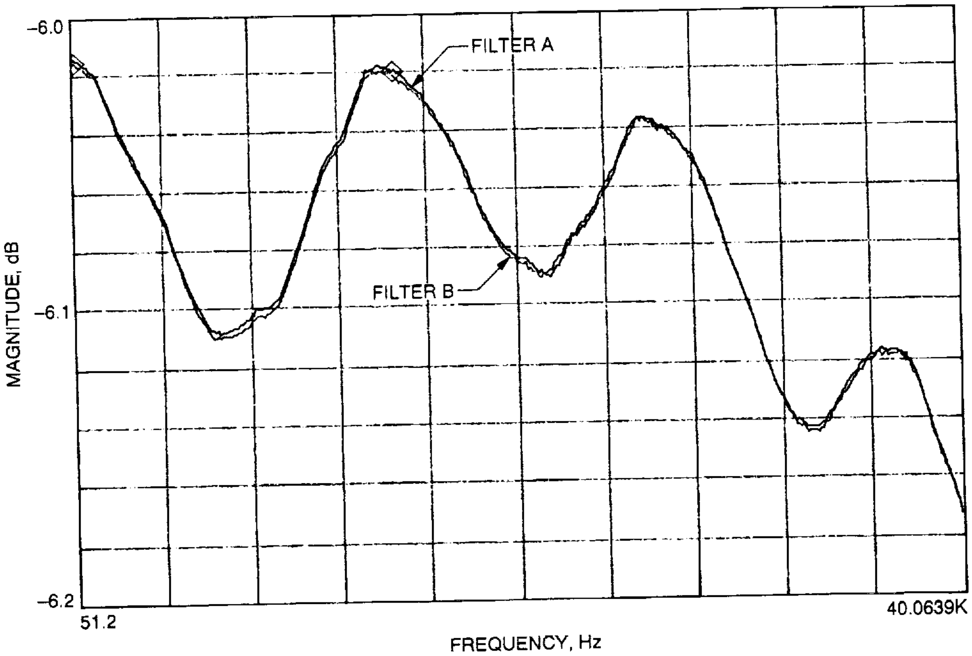


Fig. 10. Mixer output.

S5-61

53208

P28

TDA Progress Report 42-107

November 15, 1991

# Determinate-State Convolutional Codes[1]

O. Collins and M. Hizlan
Johns Hopkins University
Baltimore, Maryland

*A determinate-state convolutional code is formed from a conventional convolutional code by pruning away some of the possible state transitions in the decoding trellis. The type of staged power transfer used in determinate-state convolutional codes proves to be an extremely efficient way of enhancing the performance of a concatenated coding system. This article analyzes the decoder complexity and free distances of these new codes and provides extensive simulation results of their performance at the low signal-to-noise ratios where a real communications system would operate. The article concludes with concise, practical examples.*

## I. Introduction

In a determinate-state convolutional code, some of the possible branches of the trellis are pruned away, usually by employing an outer algebraic code, and the remaining paths in the convolutional code's trellis gain power from these "pinned" state transitions. The beauty of this technique is that it allows a concatenated coding system's performance to approach more closely the power of the concatenated code viewed as a single entity, while the decoding complexity remains comparable to the traditional sequential approach. This article will concentrate on convolutional rate $1/N$ inner codes, but trellis codes, block codes suitable for soft decoding, or more complex state machine-based codes can all use the same technique. The examples in this article, with one exception, will use algebraic outer codes for the conventional reason, i.e., the existence of decoding algorithms whose computational effort is only a polynomial in the error-correcting capability of the code. A short illustration in the next section defines the basic concept of determinate-state convolutional codes; the article then goes on to explore the decoding complexity of these new codes and their free distances. The article also presents design techniques and practical illustrations. Simulation is the only way to explore codes at the very low signal-to-noise ratios at which the inner codes in high- performance concatenated systems operate. Hence, extensive figures appear at the end of the article; these data can be used to design coding systems beyond what the article presents. The Appendix explains how the simulations were performed.

## II. The Elements of Power Transfer

Figure 1 shows what happens to the trellis of a constraint length-3 convolutional code when one of the bits going into the encoder is known to be zero.

[1] Work supported in part by a contract between Johns Hopkins University and JPL through the TDA Office, and by a grant from the National Science Foundation.

This known bit could come from the successful decoding of an outer code or from something much simpler, such as a synchronization pattern. The example of an interleaved synchronization pattern provides a quick introduction to the idea of determinate-state convolutional codes and is an extreme case of a whole continuum of different degrees of certainty which the decoder may have about the information bits. As an interesting aside, most of the simpler varieties of such outside information, such as the different letter frequencies in ASCII encoded text or the statistics of pixel differences in a binary image, are easy to incorporate in the decoding procedure.

Suppose, for instance, that every eighth bit of data going in a (7,1/2) convolutional encoder is part of a synchronization word and that the code is operating at an $E_b/N_0$ of 1.2 dB. Figure A-1 shows that the bit error rate out of the decoder will be 2.7 percent. After having acquired synchronization, one can incorporate the knowledge of every eighth bit into the decoding procedure. Figure A-5 gives the bit error rates of the remaining bits when one out of every eight is known and shows that the error rate of the remaining 7/8 of the bits drops after incorporating this information to a little under 0.92 percent. However, the power used in sending these known bits must be accounted for. The proper adjustment to the bottom axis of Fig. A-5, to allow direct power comparison with Fig. A-1, is 10 log (7/8), which is 0.58 dB. Looking back to Fig. A-1 at the 1.8-dB point, there is an error rate of 0.89 percent. Thus, almost all of the power from the known bits has transferred to those that remain, i.e., $1.8 - 1.2 = 0.6$. The overall effect is almost the same as sending a signal back to the encoder, stopping the transmission from framing information.

The successful decoding of an outer code can provide the same certainty as knowledge of a synchronization pattern. If every eighth bit in the data stream going into the (7,1/2) code discussed above is a symbol in an outer binary code, and the parameters of this code are such that it always decodes with extremely high probability when the error rate is 2.7 percent, then a second soft decoding operation gives those bits which are not doubly protected just the same error rate they had when every eighth bit was part of a synchronization pattern. The information bits of the outer code have, in effect, been sent for free. A moderately complex code will require about twice the minimum possible redundancy of $H(0.027) = 0.18$ at the 2.7-percent error rate. This simple example incorporates the essence of the article.

It is the concept of power transfer that makes determinate-state convolutional codes work. The idea of reinitializing (essentially an extreme form of state pinning) a constraint length-7 decoder by using the bytes from an 8-bit interleaved Reed-Solomon (RS) code word has been tested before in [1-3] with varying degrees of success. Performance gains were either not any larger than other easier decoder improvements, such as increasing the truncation depth, or required many decoding trials of the Reed-Solomon code, with different positions erased. Adjusting the rates of the different code words in a block of interleaved Reed-Solomon codes will significantly improve the performance of a system using decoder reinitialization, but, as this article goes on to show, when the pinned symbol size is greater than the constraint length, power transfer is always inefficient. The article proceeds with an analysis of the properties of both determinate-state convolutional codes and their decoders.

## III. Decoding Complexity and Implementation

Figure 2 demonstrates what happens as known bits percolate through an encoder shift register. The checkered circles represent zeros or ones that are known in advance and the box follows the shift register window of a constraint length-3 convolutional encoder. No additional argument is required to show that the total number of encoder states is cut in half for every known bit in the shift register.

The situation is, however, slightly more complex for the decoder since it performs one computational step for every possible state transition of the encoder rather than for every state, and thus the real measure of computational cost is the number of branches in the trellis. A return to the pruned trellis diagram in Fig. 1 confirms that the correct measure of complexity for a determinate decoding operation is just the average of $2^{K'}$, where $K'$ is the effective constraint length of the pruned code, i.e., the number of unknown bits in the shift register. In fact, considering the trellis, it is possible to collapse the row of states immediately preceding a known bit so that if a known bit comes along once every $K+1$ bits, the peak decoder memory can be halved, and if a known bit comes along once in every $K$ bits, the state diagram could look just like that of a $K-1$ code.

The decoder for a convolutional code in which some of the bits are known is straightforward to construct since it is identical to the decoder for a time-varying code of lower constraint length. This result derives from the capacity to represent the encoder for a determinate-state convolutional code as a conventional convolutional encoder with time-varying generator polynomials whose constraint length is the effective constraint length of the determinate code, i.e., the number of unknown bits in the shift register. The output from each of the encoder's generator polynomials is either inverted or not depending on the known bits in the shift register. Figure 3 shows a rate $1/2, K = 12$ encoder with four known bits.

The two constants in Fig. 3 can be computed using Eq. (1)

$$(\vec{g}_i \cdot \vec{k}) \oplus r_i = r_i' \qquad (1)$$

where $\vec{k}$ is the vector of information bits that are known to be one, $\vec{g}_i$ is the $i$th generator polynomial, and $r_i$ is the bit which will be sent over the channel as the $i$th symbol. The clocking of the shift register is stopped as known bits enter.

This unusual view of the encoder can be used to construct an efficient decoder by using Eq. (1) to modify the signs of the received symbols, i.e., $r_i$ is the bit that indicates whether the received symbol is a zero or a one and $r_i'$ is the sign bit fed to the decoder. Since the two exclusive-or operations will cancel each other out, this preprocessing step will eliminate any requirement that determinate bits must be fed into the decoder itself. Only the need to recirculate the accumulated state metrics, i.e., to update them without exchanging them in order to handle the pauses in the Fig. 3 encoder, distinguishes the determinate decoder from a conventional Viterbi decoder of reduced $K$.

Equipping a conventional decoder to perform a maximum likelihood estimate when some bits are known is also straightforward. The only necessary additions are two wires that make it possible to force all the states to choose either the zero branch or the one branch, thus overriding the normal selection of the lower of the two incoming metrics. The DSN constraint length-15 decoder, described in [4], incorporates precisely this feature. The forcing lines also simplify the testing of the decoder since they allow error bursts to be inserted artificially. This decoder was demonstrated in early 1991.

## IV. Determinate Code Properties

Another way of mechanizing the decoding process where subsets of the data stream are known is to give infinite weight to those branches that are not part of a possible path through the trellis. Since the shortest path between two points in a graph will remain the shortest when any line not on this path is lengthened,

$$S_i \mid \{S_\alpha, S_\beta, S_\gamma, \cdots\} = S_i \qquad (2)$$

where $S_i$ is the value of the $i$th symbol in the maximum likelihood estimate of the decoded data stream, and $S_i \mid \{S_\alpha, S_\beta, S_\gamma, \cdots\}$ is the estimate repeated with the knowledge that the bits in positions $\alpha$, $\beta$, $\gamma$, etc. are correct. Thus, side information, e.g., known bits, is beneficial only if it requires a change in the estimate of the transmitted data stream. Similarly, if a sequence of $k$ known symbols is sufficient to specify the encoder state, then

$$S_i \mid \{S_{i-k-1}, S_{i-k}, S_{i-k+1}, \cdots, S_{i-1}\} = S_i \mid \{Q\} \qquad (3)$$

$$S_i \mid \{S_{i+1}, S_{i+2}, S_{i+3}, \cdots, S_{i+k}\} = S_i \mid \{Q'\} \qquad (4)$$

where $\{Q\}$ is any subset of $S_1$ through $S_{i-1}$ that includes $S_{i-k-1}$ through $S_{i-1}$, and $\{Q'\}$ is any subset of $S_j$, where $j > i$, that includes $S_{i+1}$ through $S_{i+k}$. Thus, if at any point in the decoding process enough side information is available to specify the encoder state, then the future decoding operations are no longer coupled to the past through the encoder memory. If the length of the sequence of known bits is shorter than that required to specify the encoder state fully, then the coupling will be decreased, but not eliminated, so re-decoding can relieve interleaving requirements. Equations (3) and (4) also show that there is intrinsic inefficiency in using 8-bit symbols with a $K = 7$ code, since pinning more than $K$ bits in a row produces no additional power transfer.

The free distance of a determinate-state convolutional code can be obtained by making the same slight modifications to a decoder that are used to find the free distance of an ordinary convolutional code. While being fed all zero symbols, the decoder is forced out of the all-zero state into the state with all zeros and a single one. The total amount of distance that accumulates on the shortest path which brings the decoder back to the all-zero state is the free distance. Now, however, the decoder has to work with a pruned trellis, and so not all of the possible departure times from the all-zero state are equivalent; potentially, they may all have to be explored to find the free distance. When the information bit sequence that produces the minimum-weight burst is short, however, the problem of free-distance determination is trivial. If the run of unknown bits is longer than a sequence of information bits that produces a minimum-weight burst, the free distance of the root convolutional code cannot change as the result of state pinning. In particular, if a single isolated one produces a minimum-weight burst, as it does for the NASA (7,1/2) code, then the free distance cannot increase no matter how many known bits there are. Even if a single bit remains unfixed, the minimum burst producing one will slide into that position.

## V. First Example

The first example will replace a set of identical interleaved Reed-Solomon codes with a collection of different codes, which together have the same redundancy as the original set, but allow an overall gain of 0.5 dB over the old coding arrangement at its design error rate of $10^{-6}$ bit error rate (BER). The new system will actually have a lower error rate than 1 in a million. The inner code is the NASA (15,1/4) code used on the Galileo spacecraft. This is the (15,1/4) code whose performance is graphed and tabulated at the end of this article. The outer code used on the spacecraft to send compressed images is a (255,223) 8-bit Reed-Solomon code. The new outer codes will also be 8-bit Reed-Solomon codes, but with different amounts of redundancy. This example is chosen because an easily understood, very routine approach to code construction yields substantial improvements. For symbol errors within one Reed-Solomon code word to be almost independent of one another, they must be interleaved to a depth of 8, and to gain half a decibel, the redundancy will simply be shifted among the eight code words of an interleaving block, while keeping the average number of parity check symbols per code word fixed at 32. Since the redundancy of the outer code will remain constant, the energy gain will come from a lowering of the operating point on the inner code. The operating point of the inner code in the conventional system is 0.33 dB; in the proposed system, it is −0.2 dB.

In order to upper-bound the decoded bit error rate reliably, assume that if any Reed-Solomon code fails to decode, then all of its bits as well as the bits of any other undecoded Reed-Solomon code words in the frame experience a 50-percent error rate. Any concerns over frame-to-frame error propagation can be eliminated by inserting a sequence of 15 known bits after each frame; the frames are long enough so that the power penalty is negligible. Although this bound will be quite far above the actual error rate of the code, the difference measured in information bit energy will be small; i.e., little additional effort is required to lower the error rate of a coding system from one in a million to one in a billion. Figure A-10 shows how the 8-bit symbol error rate of the (15,1/4) code drops as additional side information becomes available. Table 1 shows the symbol error rates for each decoding operation and the contribution of the different Reed-Solomon code words to the average redundancy.

The decoder complexity of the new code design will be slightly less than three times the decoder complexity of the root constraint length-15 code; the last decoding step has complexity equal to that of a constraint length-7 code, and the second and third steps have complexity slightly below that of the first.

## VI. Second Example

The second example, which has its source in the original Voyager communications system, addresses those situations where the outer code must be very simple, either because the encoder needs to be small, e.g., taking up only a small fraction of a chip, or because the short decoding delays characteristic of convolutional codes must be preserved; e.g., a compressed voice circuit operating at 2.5 Kbps should not have a coding delay of more than 250 bits. This example also shows how a communications system can be improved without changing the encoder. Voyager used a (7,1/2) convolutional code to send uncompressed images back to Earth at a bit error rate of $5 \times 10^{-3}$. Decoding errors, which very seldom affected more than two adjacent 8-bit pixels, were corrected either by low-pass filtering of the image or manually. A fraction of the data, however, came from other instruments that demanded no more than one error in a million bits, and these data were encoded first by a Golay code and then combined with the imaging data and sent to the convolutional encoder. For this example, one out of every eight bits will be covered with a Golay code, which is slightly larger than the fraction on Voyager. The bursts produced by a $K = 7$ code are short enough so that the bits of a Golay code word will experience independent errors when they are spaced 8 bits apart in the data stream.

A re-decoding operation using the information from a successful decoding of the (24,12) Golay code will cut the error rate of the (7,1/2) convolutional code by a factor of 3. The probability of five or more errors in a Golay code word is so small that incorrect decodings have a negligible effect on the error rate of the bits following a second decoding. The errors produced by an incorrect Golay decoding can be upper-bounded by assuming that an error burst covers all those bits intermixed with the Golay code word, plus the bits for several constraint lengths on either side. The convolutional code's constraint length would have to be increased to 9 in order to achieve the same improvement. If a similar interleaved Golay code was used on the (15,1/4) convolutional code, the error rate would drop by a factor of 5. The second decoding operation increases the decoding complexity by a little more than a quarter since one determinate bit is always in the shift register, and two are in the shift register 7/8 of the time. Increasing the convolutional code's constraint length in order to achieve similar gains would raise the decoding complexity by a factor of 32.

In this example, the energy per information bit measured at an error rate of $10^{-3}$ does not decrease; instead, the second decoding step allows the error rate of 1/8 of the bits to be reduced to $10^{-6}$ without any increase in $E_b/N_0$ and with only an extremely small increase in encoder complexity. Variations on this scheme might be very useful for certain types of data compression systems that transmit both the parameters of a predictor and the differences between the predictor and the source; these differences can tolerate a much larger error rate.

## VII. Third Example

The third example will give a constraint length-11 convolutional code the same performance as the constraint length-15 code put aboard Galileo. The outer code will remain an 8-bit Reed-Solomon code, and the average number of redundant symbols per 255 symbol code word will stay fixed at 32. For a constraint length-11 code, an interleaving depth of four is sufficient to achieve symbol independence, and the re-decoding operation will take place after one code word of an interleaving block has decoded and every fourth 8-bit symbol is thus known. Only a single determinate decoding operation is necessary to increase the performance of a constraint length-11 code to a little better than that of the constraint length-15 code.

The operating point of the $K = 11$ code is 0.3 dB, giving an 8-bit symbol error rate of 4.92 percent, and 68 parity check symbols will bring the probability of Reed-Solomon decoder failure below $7 \times 10^{-8}$. After the pinned decoding operation, the average error rate of the remaining symbols will drop to 1.25 percent. Twenty redundant symbols on each of the three remaining Reed-Solomon code words will reduce the bit error rate to $10^{-6}$. Of course, if the first Reed-Solomon code fails to decode, then a determinate decoding operation will not be possible. The three remaining, much less powerful, code words will be left to cope with a symbol error rate, which will usually be worse than the average symbol error rate before a determinate decoding operation. Thus, when the first code fails, there is very little point in trying to decode any of the other three, and the decoder will just pass the bits from the convolutional encoder out without further processing. Simulation results show that the average bit error rate over a failed block is 3.2 percent. The effect of the primary Reed-Solomon code failure on the error rate of the bits in the other three code words is thus an increase of less than 0.22 percent. Reference [5] shows that Reed-Solomon decoder errors are so infrequent they can be ignored. The superior error rate of the bits in the primary code word will bring the aggregate system error performance below 1 in a million.

The average number of redundant symbols per code word is thus $1/4 \times 68 + 3/4 \times 20 = 32$, the same as the NASA standard outer code. The operating point of the inner convolutional code on Galileo is 0.33 dB, so the state pinned system is not only 1/8 as complex as the conventional, but performs slightly better. A second soft decoding operation could even improve the performance of the constraint length-11 code well beyond that of the Galileo code.

## VIII. Summary

This article has examined what happens to a convolutional code when some of the state transitions are predetermined, usually by the successful decoding of an outer code. This type of technique has to justify itself by being more efficient than other methods, such as increasing the convolutional code's constraint length or passing erasure information to the outer code. The examples were therefore constructed to demonstrate improved performance lucidly, without the ambiguities that varying the overall code rate would introduce. The examples were not constructed to present an optimal system. Such a coding system would require codes between BCH and RS codes, i.e., with a symbol size of 4 or 5 bits and 100 or more symbols in a code word. The hybridized Reed-Solomon codes described in [4] can fulfill these requirements.

The results in this article show that in a concatenated coding system, a determinate decoding operation is a more efficient way to expend computational resources than an increase of the constraint length by 1. A second and third determinate decoding operation can often be justified. Other types of coupling between the inner and outer codes do not approach the gains possible with determinate decoding. Even perfect erasure declaration, which would cut the redundancy in the outer code by half, would not achieve the performance gains demonstrated.

Determinate decoding is an especially useful technique where the encoder must be extraordinarily simple or where all of the information bits do not require equal error protection. Many types of compressors produce such outputs. The bits coming out of a determinate decoding operation may have a low enough error rate so that no outer code is needed to cover them. Furthermore, techniques presented in this article offer an especially efficient way of using surplus speed in a Viterbi decoder or of trading decoding speed for performance; e.g., the same machine might be used for deep-space probes which demand the highest possible coding gains and for near-Earth satellites which demand high data rates.

Table 1. Symbol error rates

| Symbols known | Error rate, percent | R/S redundancy | | | | Average |
|---|---|---|---|---|---|---|
| None | 10.4 | 104 | × | 1/8 | = | 13 |
| Every eighth | 6.25 | 68 | × | 1/8 | = | 8.5 |
| Every fourth | 1.13 | 26 | × | 1/4 | = | 6.5 |
| Every second | 0.106 | 8 | × | 1/2 | = | 4 |
| Total | | | | | | 32 |

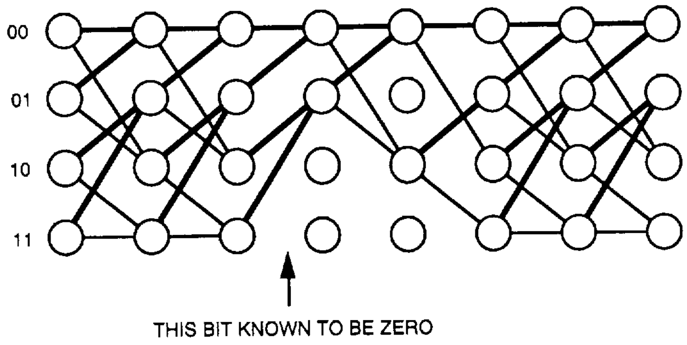


Fig. 1. Constraint length-3 convolutional code trellis.

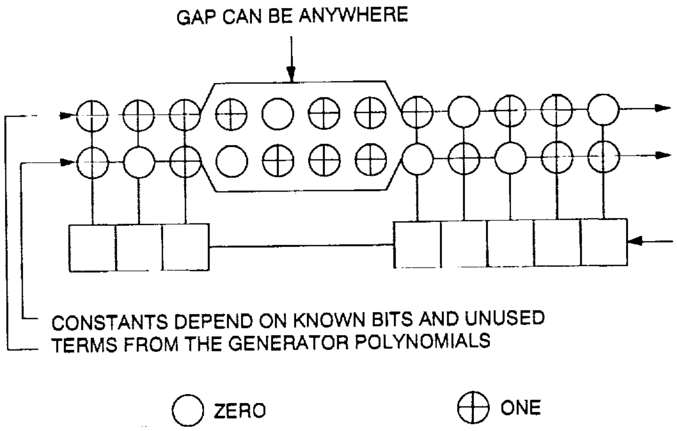

GAP CAN BE ANYWHERE
CONSTANTS DEPEND ON KNOWN BITS AND UNUSED
TERMS FROM THE GENERATOR POLYNOMIALS
ZERO
ONE

Fig. 3. A rate 1/2, *K* = 12 encoder with four known bits.

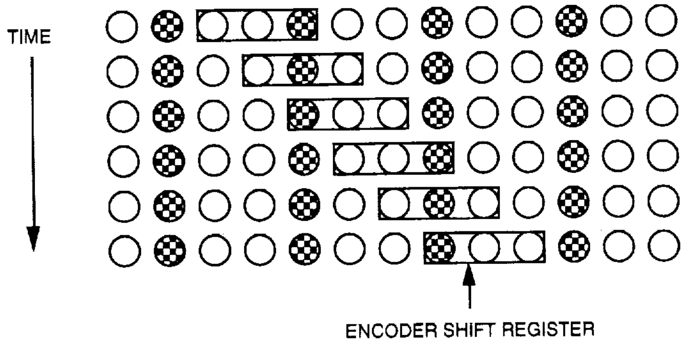


Fig. 2. Encoder shift register.

# Appendix

The simulations in this article were performed by encoding sequences of pseudo-randomly generated information bits, adding white Gaussian noise for a specified signal-to-noise ratio, quantizing the resulting noisy symbols into 8-bit symbols, and decoding the result. The use of pseudo-random information bits rather than all zeros or all ones eliminated the possibility that a programming error would produce a higher than normal performance.

Random number generators for both the information bits and the noise were adapted from *Numerical Recipes in C* [6]. The pseudo-random bit generator employs the algorithm irbit2 [6] with a primitive polynomial modulo 2 of order 32 with nonzero coefficients (32, 7, 5, 3, 2, 1, 0). Hence, the resulting sequence of pseudo-random bits does not repeat for more than 4,000,000,000 bits. The white Gaussian noise generator uses gasdev [6] with the routine ran1 [6]. Again, the resulting sequence has a virtually infinite period for simulation purposes.

The noisy encoded symbols are quantized before decoding into 8-bit symbols based on the desired signal-to-noise ratio, so that the expected quantizer saturation rate is 0.1 percent. The decoder employs the Viterbi algorithm modified to do trace-backs rather than entire path updates as in [4]. This saves a tremendous amount of computer time and makes simulation of codes with constraint lengths as high as 15 feasible on available workstations. Decoding is done on blocks of 170 bits after a 170-bit trace back through the trellis. For a detailed description of the algorithm, as well as its hardware implementation, see [4]. The decoding algorithm used is a direct translation of the one given there. Free distances and related computation can be performed by employing this decoding algorithm in conjunction with a forced starting state. Such free distances will then always include the effects of state pinning.

Each code has been simulated for a set of signal-to-noise ratios likely to produce the most relevant range of error rates. The number of bits decoded for each signal-to-noise ratio run is 2,000,000 for constraint length-15 codes, 3,000,000 for constraint length-13 codes, and 4,000,000 for all other codes. In each case, a signal-to-noise ratio run consists of two runs with half the total number of bits and different seeds for random number generators. The 95-percent confidence interval for the simulation results, based on applying counting statistics to the error bursts, is not larger than 0.11 dB in the worst case, and is typically much smaller.

A subset of the simulation runs have been presented as figures. Figures A-1 to A-4, showing symbol error rate, are used to design coding systems with Reed-Solomon or other similar outer codes. Figures A-5 to A-7 showing bit error rates, are useful for comparing the effectiveness of power transfer for different patterns of forced bits, i.e., for exploring the differences between having the determinate bits spread out and having them occur in bunches. Figures A-8 to A-10 show how the effective signal-to-noise ratio of the inner code changes as different numbers of symbols are forced; thus, they allow an easy evaluation of the benefits of multiple re-decoding operations. The data from all of the simulation runs appear in Tables A-1 through A-12.

The (7,1/2) code was used on Voyager, and the (15,1/4) is now flying on Galileo. The (11,1/4) code was made up at the computer keyboard by one of the authors. Tables are available from the authors containing a title bar with the code name, free distance of the code, the generating polynomial of the code, and the known symbol frequency (with the resulting power added in decibels). Bit error rates (BER) and symbol error rates (SER) are presented for 1-bit, 4-bit and 8-bit symbols as a function of signal-to-noise ratio (SNR) in each table. SER @ *n* stands for symbol error rate for the *n*th symbol after the known symbol.

# References

[1] L. N. Lee, "Concatenated Coding Systems Employing a Unit Memory Convolutional Code and a Byte Oriented Decoding Algorithm," *IEEE Trans. Communications*, vol. COM-25, pp. 1064–1074, October 1977.

[2] G. W. Zeoli, "Coupled Decoding of Block-Convolutional Concatenated Codes," *IEEE Trans. Communications*, vol. COM-21, pp. 219–226, March 1973.

[3] E. Paaske, "Improved Decoding for a Concatenated Coding System Recommended by CCSDS," *IEEE Trans. Communications*, vol. COM-38, pp. 1138–1144, August 1990.

[4] O. Collins, "Coding Beyond the Computational Cutoff Rate," Ph.D. thesis, California Institute of Technology, Pasadena, California, 1989.

[5] R. J. McEliece and L. Swanson, "On the Decoder Error Probability for Reed-Solomon Codes," *IEEE Transactions on Information Theory*, vol. IT-32, pp. 701–703, September 1986.

[6] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Wettering, *Numerical Recipes in C—The Art of Scientific Computing*," Cambridge: Cambridge University Press, 1988.

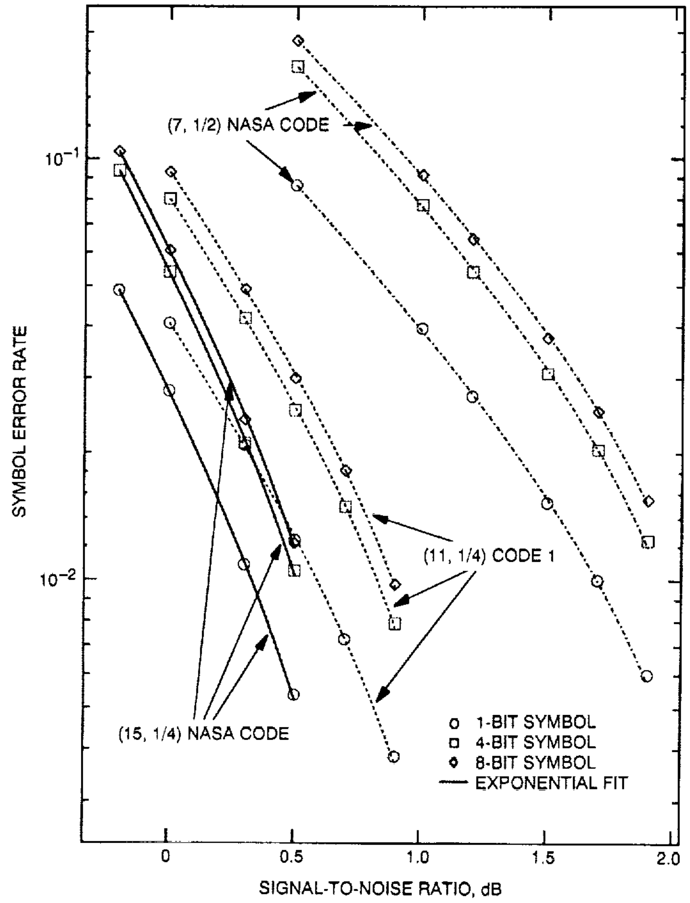


**Fig. A-1. Symbol (and bit) error rate with no symbol known.**

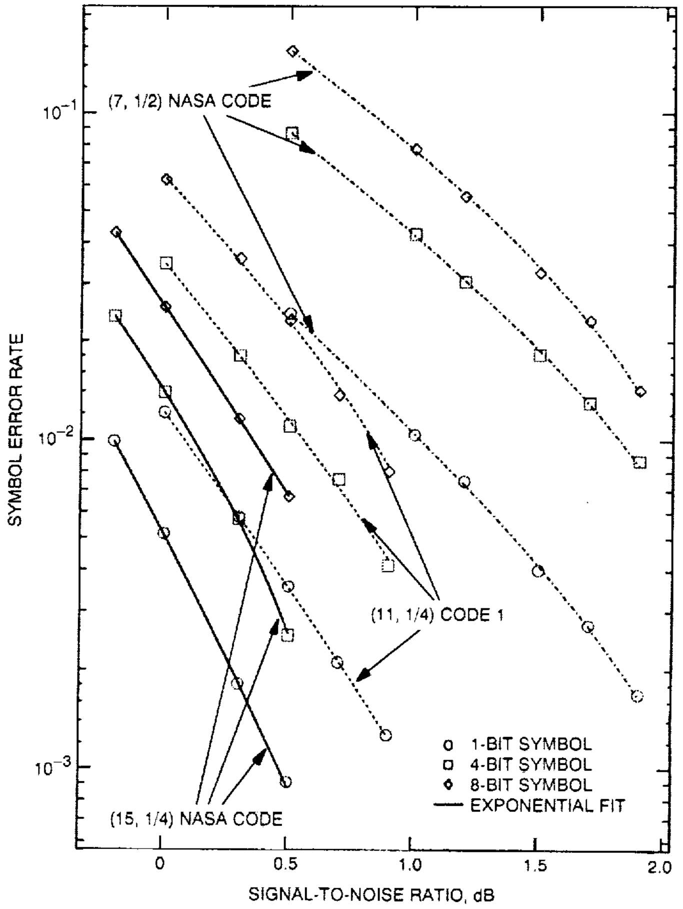


**Fig. A-2. Symbol error rate with every eighth symbol known.**

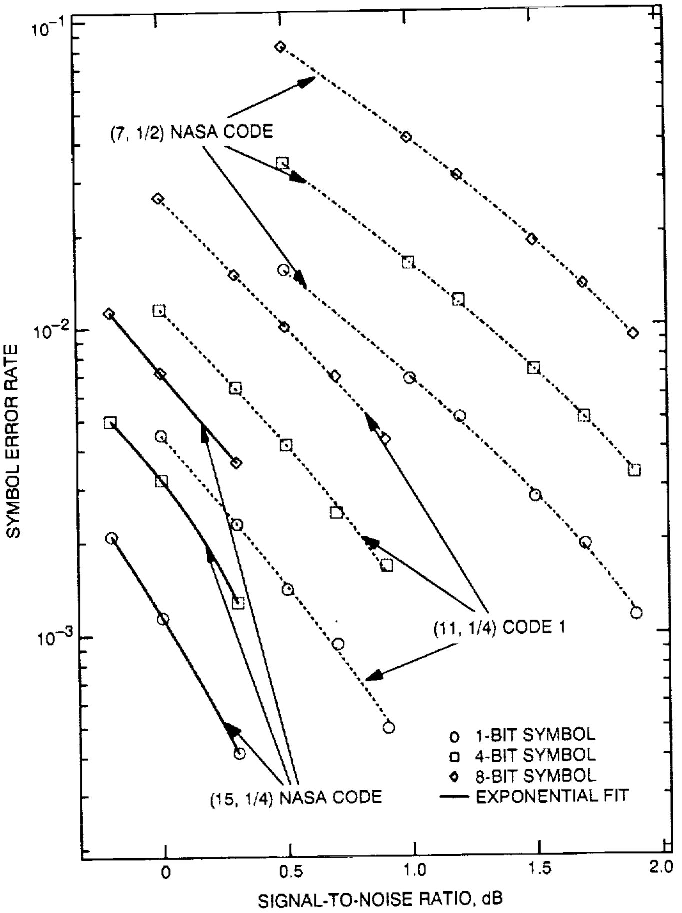


**Fig. A-3. Symbol error rate with every fourth symbol known.**

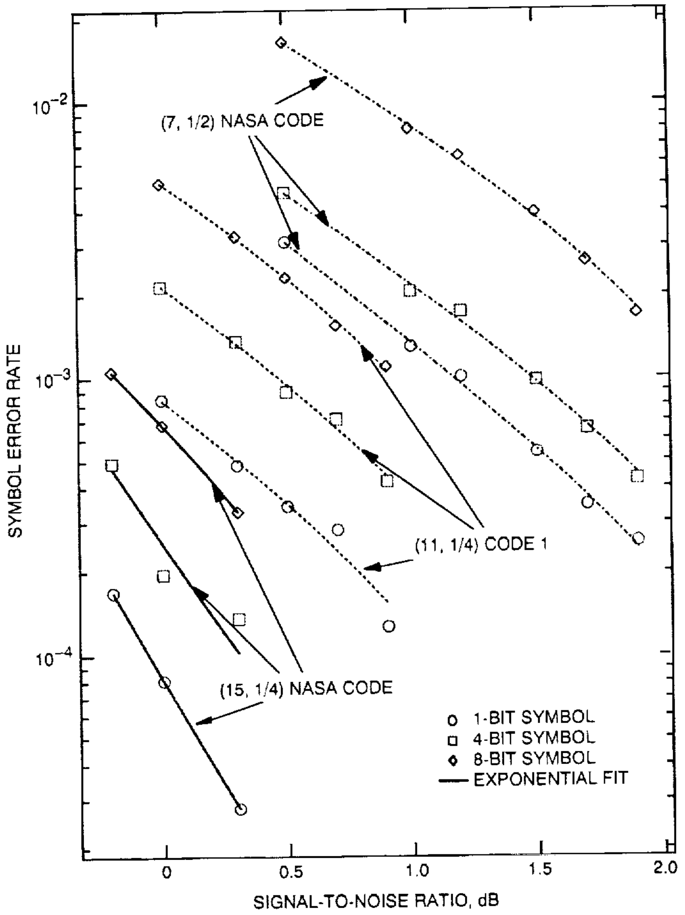


**Fig. A-4. Symbol error rate with every other symbol known.**

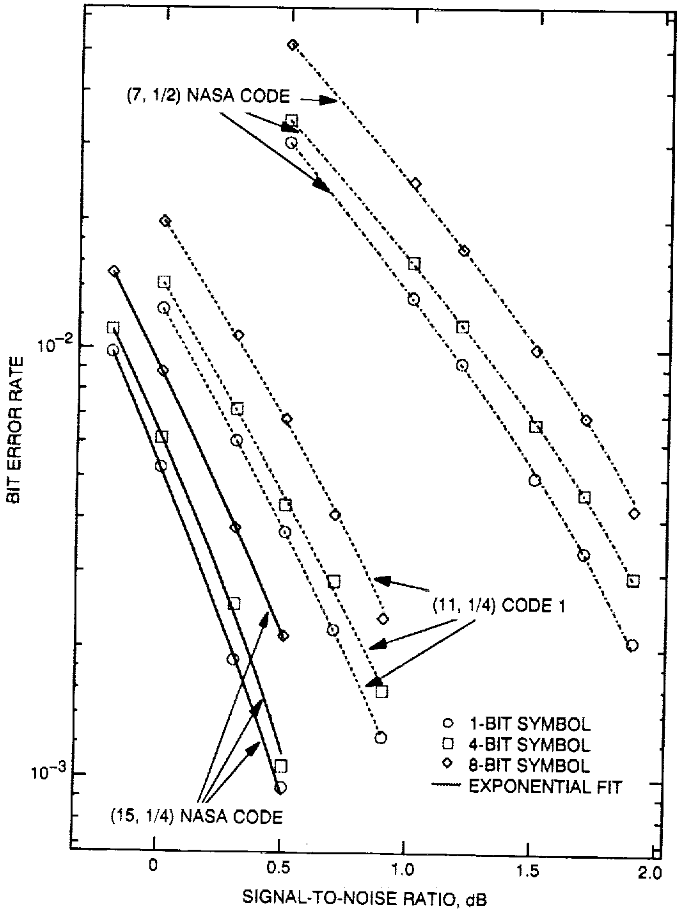


**Fig. A-5. Bit error rate with every eighth symbol known.**

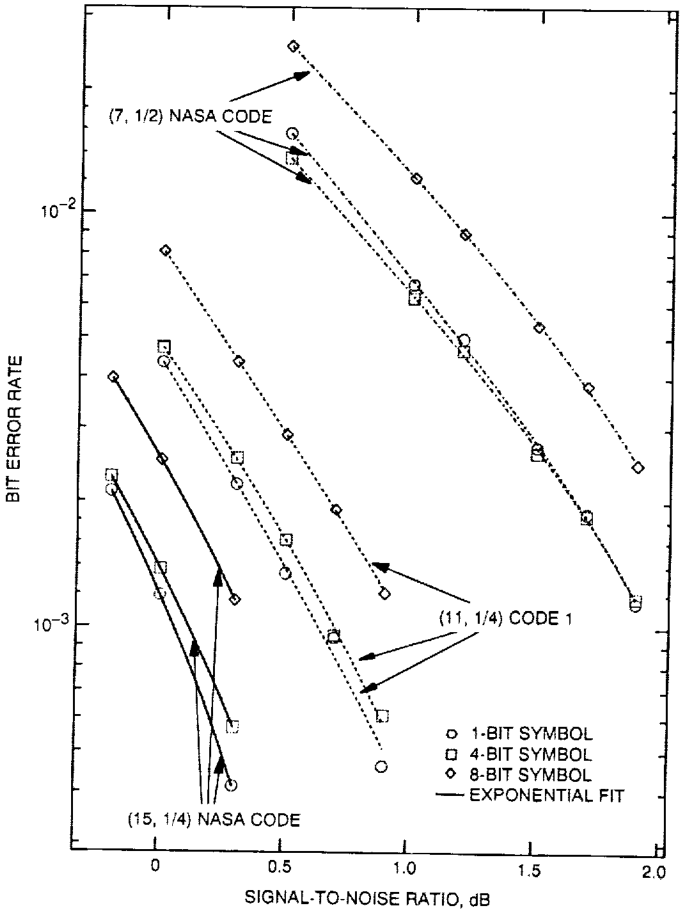


**Fig. A-6. Bit error rate with every fourth symbol known.**

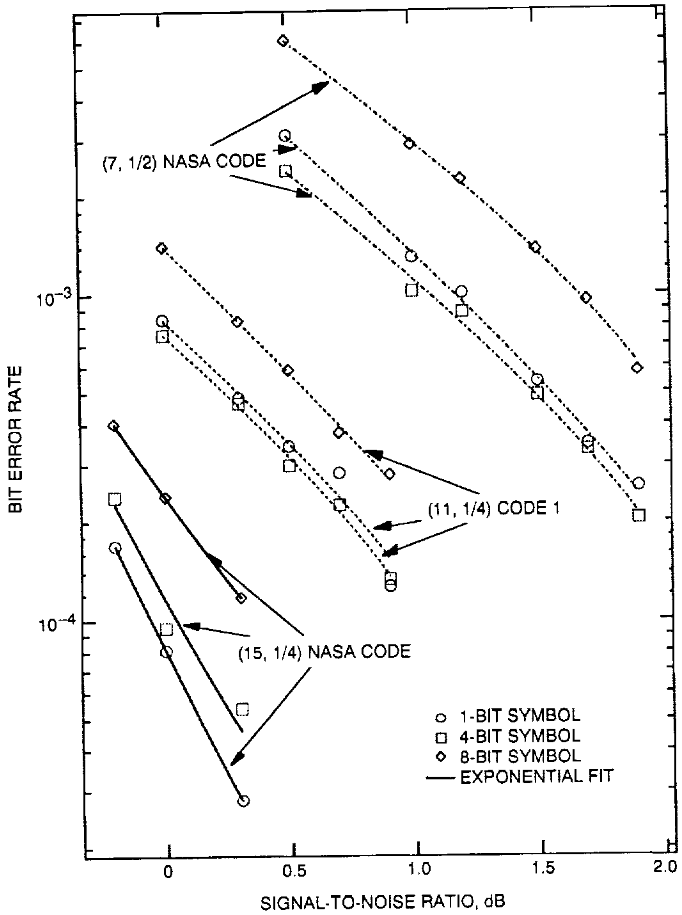


Fig. A-7. Bit error rate with every other symbol known.

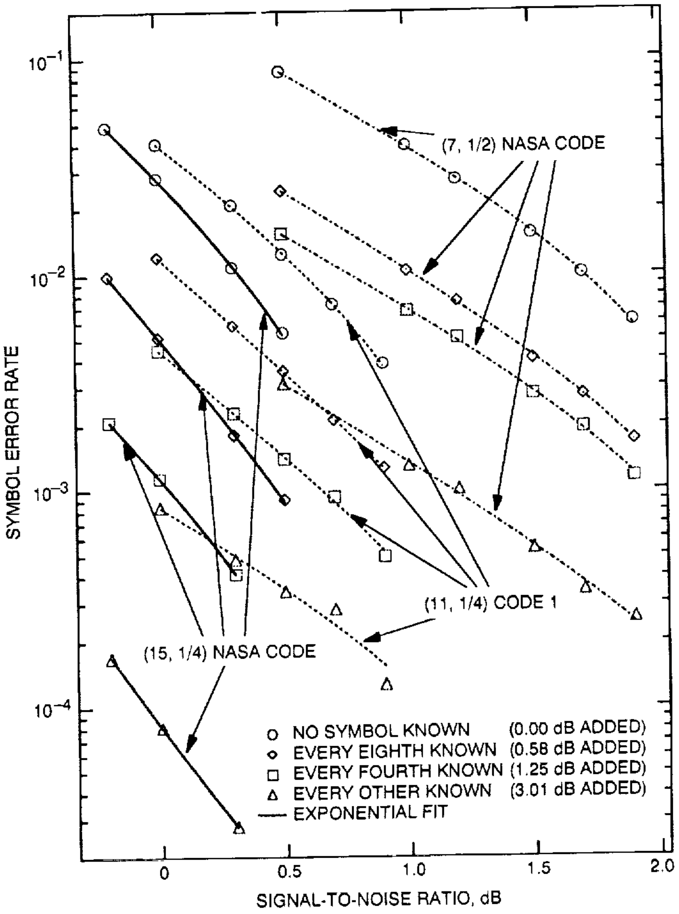


Fig. A-8. One-bit symbol error rates with different numbers of known bits.

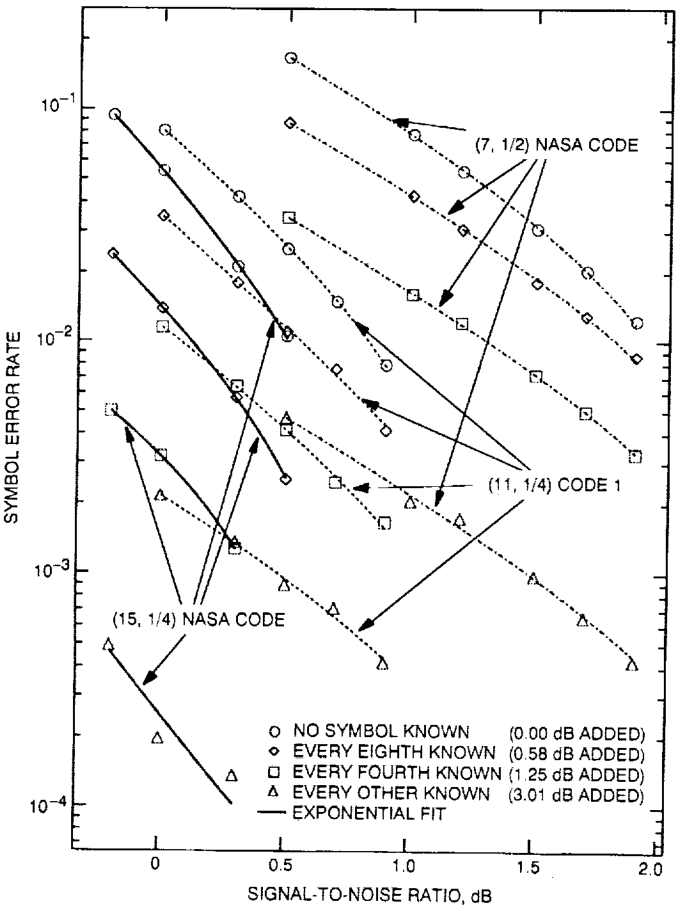


**Fig. A-9. Four-bit symbol error rates with different numbers of known bits.**

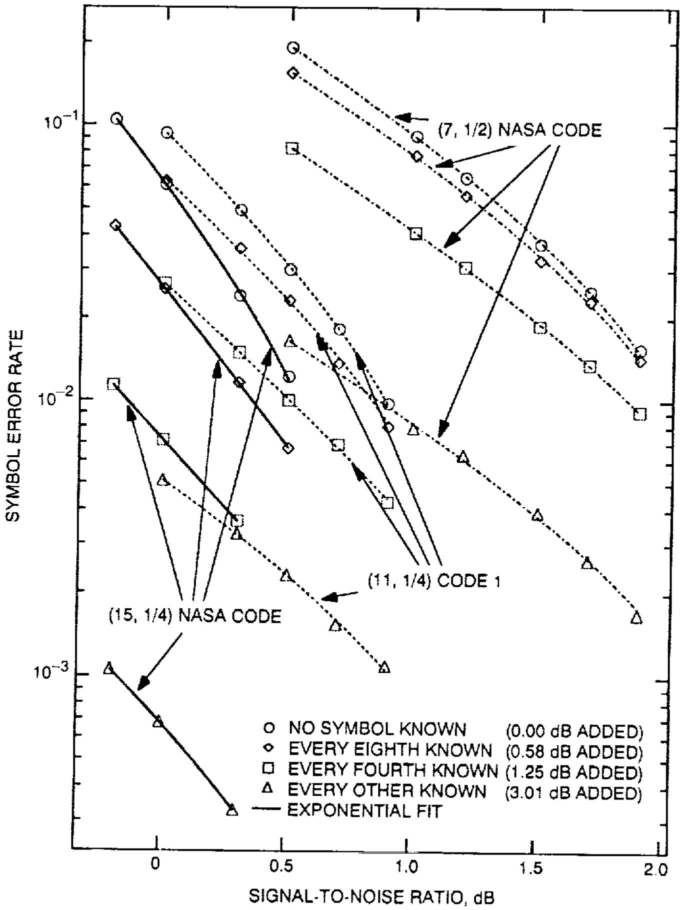


**Fig. A-10. Eight-bit symbol error rates with different numbers of known bits.**

Table A-1. No symbol known to the decoder, (7, 1/2) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | | No symbol known<br>to the decoder.<br>(dB added = 0.00) | | | |
|---|---|---|---|---|---|---|---|
| (7, 1/2)<br>NASA<br>Code | 10 | 1111001<br>1011011 | | | | | |
| *SYMBOL SIZE* | SNR (dB) | 0.5 | 1.0 | 1.2 | 1.5 | 1.7 | 1.9 |
| *1-Bit* | SER | 8.65e-02 | 3.95e-02 | 2.72e-02 | 1.53e-02 | 1.00e-02 | 6.00e-03 |
| *4-Bit* | SER | 1.66e-01 | 7.77e-02 | 5.41e-02 | 3.08e-02 | 2.04e-02 | 1.24e-02 |
| *8-Bit* | SER | 1.91e-01 | 9.15e-02 | 6.46e-02 | 3.75e-02 | 2.51e-02 | 1.55e-02 |

Table A-2. Every eighth symbol known to the decoder, (7, 1/2) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | | Every eighth<br>symbol known<br>to the decoder.<br>(dB added = 0.58) | | | |
|---|---|---|---|---|---|---|---|
| (7, 1/2)<br>NASA<br>Code | 10 | 1111001<br>1011011 | | | | | |
| *SYMBOL SIZE* | SNR (dB) | 0.5 | 1.0 | 1.2 | 1.5 | 1.7 | 1.9 |
| *1-Bit* | BER | 3.03e-02 | 1.32e-02 | 9.26e-03 | 5.01e-03 | 3.36e-03 | 2.09e-03 |
| | SER @ 1 | 3.00e-02 | 1.31e-02 | 9.30e-03 | 4.90e-03 | 3.33e-03 | 2.15e-03 |
| | SER @ 2 | 3.20e-02 | 1.41e-02 | 9.75e-03 | 5.35e-03 | 3.59e-03 | 2.22e-03 |
| | SER @ 3 | 3.17e-02 | 1.39e-02 | 9.72e-03 | 5.29e-03 | 3.57e-03 | 2.18e-03 |
| | SER @ 4 | 2.43e-02 | 1.04e-02 | 7.49e-03 | 4.02e-03 | 2.73e-03 | 1.69e-03 |
| | SER @ 5 | 3.01e-02 | 1.31e-02 | 9.16e-03 | 4.99e-03 | 3.27e-03 | 2.10e-03 |
| | SER @ 6 | 3.23e-02 | 1.42e-02 | 9.84e-03 | 5.38e-03 | 3.62e-03 | 2.18e-03 |
| | SER @ 7 | 3.15e-02 | 1.36e-02 | 9.56e-03 | 5.12e-03 | 3.42e-03 | 2.08e-03 |
| *4-Bit* | BER | 3.42e-02 | 1.59e-02 | 1.14e-02 | 6.66e-03 | 4.59e-03 | 2.95e-03 |
| | SER @ 1 | 4.68e-02 | 2.15e-02 | 1.55e-02 | 8.64e-03 | 5.62e-03 | 3.83e-03 |
| | SER @ 2 | 7.00e-02 | 3.35e-02 | 2.44e-02 | 1.43e-02 | 9.98e-03 | 6.42e-03 |
| | SER @ 3 | 8.31e-02 | 4.03e-02 | 2.94e-02 | 1.74e-02 | 1.22e-02 | 7.86e-03 |
| | SER @ 4 | 8.66e-02 | 4.26e-02 | 3.06e-02 | 1.82e-02 | 1.30e-02 | 8.66e-03 |
| | SER @ 5 | 8.35e-02 | 4.03e-02 | 2.86e-02 | 1.74e-02 | 1.20e-02 | 7.95e-03 |
| | SER @ 6 | 7.11e-02 | 3.38e-02 | 2.38e-02 | 1.47e-02 | 9.70e-03 | 6.35e-03 |
| | SER @ 7 | 4.79e-02 | 2.13e-02 | 1.53e-02 | 8.94e-03 | 5.76e-03 | 3.52e-03 |
| *8-Bit* | BER | 5.14e-02 | 2.45e-02 | 1.72e-02 | 1.00e-02 | 6.93e-03 | 4.22e-03 |
| | SER @ 1 | 7.91e-02 | 3.62e-02 | 2.75e-02 | 1.64e-02 | 1.12e-02 | 6.78e-03 |
| | SER @ 2 | 1.27e-01 | 6.10e-02 | 4.52e-02 | 2.76e-02 | 1.87e-02 | 1.14e-02 |
| | SER @ 3 | 1.48e-01 | 7.36e-02 | 5.24e-02 | 3.27e-02 | 2.18e-02 | 1.32e-02 |
| | SER @ 4 | 1.55e-01 | 7.76e-02 | 5.57e-02 | 3.27e-02 | 2.32e-02 | 1.42e-02 |
| | SER @ 5 | 1.46e-01 | 7.43e-02 | 5.26e-02 | 3.09e-02 | 2.28e-02 | 1.43e-02 |
| | SER @ 6 | 1.25e-01 | 6.31e-02 | 4.41e-02 | 2.62e-02 | 1.91e-02 | 1.28e-02 |
| | SER @ 7 | 7.91e-02 | 3.84e-02 | 2.61e-02 | 1.58e-02 | 1.13e-02 | 6.91e-03 |

Table A-3. Every fourth symbol known to the decoder, (7, 1/2) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | | Every fourth symbol known to the decoder. (dB added = 1.25) | | | |
|---|---|---|---|---|---|---|---|
| (7, 1/2) NASA Code | 10 | 1111001<br>1011011 | | | | | |
| *SYMBOL SIZE* | SNR (dB) | 0.5 | 1.0 | 1.2 | 1.5 | 1.7 | 1.9 |
| *1-Bit* | BER | 1.56e-02 | 6.74e-03 | 4.99e-03 | 2.71e-03 | 1.88e-03 | 1.15e-03 |
| | SER @ 1 | 1.48e-02 | 6.31e-03 | 4.68e-03 | 2.54e-03 | 1.70e-03 | 1.11e-03 |
| | SER @ 2 | 1.54e-02 | 6.77e-03 | 5.05e-03 | 2.76e-03 | 1.93e-03 | 1.14e-03 |
| | SER @ 3 | 1.66e-02 | 7.14e-03 | 5.23e-03 | 2.83e-03 | 2.01e-03 | 1.20e-03 |
| *4-Bit* | BER | 1.36e-02 | 6.25e-03 | 4.67e-03 | 2.62e-03 | 1.86e-03 | 1.18e-03 |
| | SER @ 1 | 2.60e-02 | 1.20e-02 | 8.93e-03 | 5.09e-03 | 3.60e-03 | 2.40e-03 |
| | SER @ 2 | 3.42e-02 | 1.61e-02 | 1.21e-02 | 7.18e-03 | 5.00e-03 | 3.29e-03 |
| | SER @ 3 | 2.62e-02 | 1.22e-02 | 9.22e-03 | 5.30e-03 | 3.53e-03 | 2.26e-03 |
| *8-Bit* | BER | 2.53e-02 | 1.22e-02 | 8.96e-03 | 5.37e-03 | 3.85e-03 | 2.47e-03 |
| | SER @ 1 | 5.94e-02 | 2.94e-02 | 2.18e-02 | 1.32e-02 | 9.74e-03 | 6.25e-03 |
| | SER @ 2 | 8.23e-02 | 4.10e-02 | 3.08e-02 | 1.88e-02 | 1.36e-02 | 9.18e-03 |
| | SER @ 3 | 6.02e-02 | 2.92e-02 | 2.15e-02 | 1.36e-02 | 9.11e-03 | 6.13e-03 |

Table A-4. Every other symbol known to the decoder, (7, 1/2) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | | Every other symbol known to the decoder. (dB added = 3.01) | | | |
|---|---|---|---|---|---|---|---|
| (7, 1/2) NASA Code | 10 | 1111001<br>1011011 | | | | | |
| *SYMBOL SIZE* | SNR (dB) | 0.5 | 1.0 | 1.2 | 1.5 | 1.7 | 1.9 |
| *1-Bit* | BER | 3.11e-03 | 1.30e-03 | 1.01e-03 | 5.37e-04 | 3.46e-04 | 2.55e-04 |
| | SER @ 1 | 3.11e-03 | 1.30e-03 | 1.01e-03 | 5.37e-04 | 3.45e-04 | 2.55e-04 |
| *4-Bit* | BER | 2.41e-03 | 1.02e-03 | 8.78e-04 | 4.87e-04 | 3.32e-04 | 2.04e-04 |
| | SER @ 1 | 4.68e-03 | 2.06e-03 | 1.74e-03 | 9.78e-04 | 6.52e-04 | 4.26e-04 |
| *8-Bit* | BER | 6.05e-03 | 2.90e-03 | 2.25e-03 | 1.37e-03 | 9.52e-04 | 5.77e-04 |
| | SER @ 1 | 1.65e-02 | 7.95e-03 | 6.36e-03 | 3.94e-03 | 2.64e-03 | 1.70e-03 |

Table A-5. No symbol known to the decoder, (11, 1/4) first code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | No symbol known to the decoder. (dB added = 0.00) |
|---|---|---|---|
| (11, 1/4) First Code | 23 | 10001011011<br>11101010001<br>10110101001<br>11000100101 | |

| *SYMBOL SIZE* | SNR (dB) | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|---|
| *1-Bit* | SER | 4.07e-02 | 2.09e-02 | 1.24e-02 | 7.25e-03 | 3.84e-03 |
| *4-Bit* | SER | 8.03e-02 | 4.17e-02 | 2.52e-02 | 1.49e-02 | 7.92e-03 |
| *8-Bit* | SER | 9.30e-02 | 4.92e-02 | 3.00e-02 | 1.82e-02 | 9.77e-03 |

Table A-6. Every eighth symbol known to the decoder, (11, 1/4) first code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | Every eighth symbol known to the decoder. (dB added = 0.58) |
|---|---|---|---|
| (11, 1/4) First Code | 23 | 10001011011<br>11101010001<br>10110101001<br>11000100101 | |

| *SYMBOL SIZE* | SNR (dB) | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|---|
| *1-Bit* | BER | 1.24e-02 | 6.05e-03 | 3.72e-03 | 2.21e-03 | 1.24e-03 |
| | SER @ 1 | 1.18e-02 | 5.89e-03 | 3.68e-03 | 2.08e-03 | 1.20e-03 |
| | SER @ 2 | 1.26e-02 | 6.08e-03 | 3.77e-03 | 2.21e-03 | 1.30e-03 |
| | SER @ 3 | 1.27e-02 | 6.11e-03 | 3.85e-03 | 2.26e-03 | 1.25e-03 |
| | SER @ 4 | 1.21e-02 | 5.80e-03 | 3.57e-03 | 2.11e-03 | 1.27e-03 |
| | SER @ 5 | 1.24e-02 | 6.24e-03 | 3.69e-03 | 2.33e-03 | 1.24e-03 |
| | SER @ 6 | 1.27e-02 | 6.21e-03 | 3.75e-03 | 2.27e-03 | 1.23e-03 |
| | SER @ 7 | 1.23e-02 | 6.05e-03 | 3.69e-03 | 2.22e-03 | 1.20e-03 |
| *4-Bit* | BER | 1.42e-02 | 7.18e-03 | 4.29e-03 | 2.87e-03 | 1.59e-03 |
| | SER @ 1 | 2.29e-02 | 1.15e-02 | 6.42e-03 | 4.47e-03 | 2.26e-03 |
| | SER @ 2 | 2.93e-02 | 1.54e-02 | 9.21e-03 | 6.26e-03 | 3.38e-03 |
| | SER @ 3 | 3.30e-02 | 1.72e-02 | 1.08e-02 | 7.38e-03 | 3.98e-03 |
| | SER @ 4 | 3.46e-02 | 1.80e-02 | 1.10e-02 | 7.58e-03 | 4.14e-03 |
| | SER @ 5 | 3.34e-02 | 1.71e-02 | 1.05e-02 | 7.50e-03 | 4.10e-03 |
| | SER @ 6 | 2.93e-02 | 1.50e-02 | 9.04e-03 | 6.20e-03 | 3.61e-03 |
| | SER @ 7 | 2.30e-02 | 1.14e-02 | 6.78e-03 | 4.37e-03 | 2.51e-03 |
| *8-Bit* | BER | 1.98e-02 | 1.07e-02 | 6.82e-03 | 4.10e-03 | 2.36e-03 |
| | SER @ 1 | 3.04e-02 | 1.68e-02 | 1.10e-02 | 7.04e-03 | 4.19e-03 |
| | SER @ 2 | 4.95e-02 | 2.78e-02 | 1.72e-02 | 1.10e-02 | 6.94e-03 |
| | SER @ 3 | 6.00e-02 | 3.38e-02 | 2.16e-02 | 1.32e-02 | 8.13e-03 |
| | SER @ 4 | 6.25e-02 | 3.58e-02 | 2.31e-02 | 1.37e-02 | 8.03e-03 |
| | SER @ 5 | 5.96e-02 | 3.36e-02 | 2.19e-02 | 1.28e-02 | 7.39e-03 |
| | SER @ 6 | 4.95e-02 | 2.72e-02 | 1.78e-02 | 1.15e-02 | 6.02e-03 |
| | SER @ 7 | 3.08e-02 | 1.72e-02 | 1.10e-02 | 7.18e-03 | 4.02e-03 |

Table A-7. Every fourth symbol known to the decoder, (11, 1/4) first code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | |
|---|---|---|---|
| (11, 1/4) First Code | 23 | 10001011011<br>11101010001<br>10110101001<br>11000100101 | Every fourth symbol known to the decoder. (dB added = 1.25) |

| *SYMBOL SIZE* | SNR (dB) | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|---|
| *1-Bit* | BER | 4.34e-03 | 2.20e-03 | 1.34e-03 | 9.42e-04 | 4.63e-04 |
| | SER @ 1 | 4.17e-03 | 2.15e-03 | 1.27e-03 | 9.81e-04 | 4.39e-04 |
| | SER @ 2 | 4.47e-03 | 2.28e-03 | 1.40e-03 | 9.31e-04 | 4.94e-04 |
| | SER @ 3 | 4.36e-03 | 2.16e-03 | 1.35e-03 | 9.15e-04 | 4.56e-04 |
| *4-Bit* | BER | 4.71e-03 | 2.54e-03 | 1.61e-03 | 9.55e-04 | 6.12e-04 |
| | SER @ 1 | 9.96e-03 | 5.41e-03 | 3.40e-03 | 2.09e-03 | 1.34e-03 |
| | SER @ 2 | 1.15e-02 | 6.40e-03 | 4.13e-03 | 2.47e-03 | 1.66e-03 |
| | SER @ 3 | 9.86e-03 | 5.37e-03 | 3.54e-03 | 2.10e-03 | 1.33e-03 |
| *8-Bit* | BER | 8.05e-03 | 4.35e-03 | 2.90e-03 | 1.91e-03 | 1.21e-03 |
| | SER @ 1 | 2.00e-02 | 1.12e-02 | 7.83e-03 | 5.21e-03 | 3.16e-03 |
| | SER @ 2 | 2.66e-02 | 1.49e-02 | 1.00e-02 | 6.90e-03 | 4.28e-03 |
| | SER @ 3 | 2.01e-02 | 1.13e-02 | 7.52e-03 | 5.32e-03 | 3.33e-03 |

Table A-8. Every other symbol known to the decoder, (11, 1/4) first code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | |
|---|---|---|---|
| (11, 1/4) First Code | 23 | 10001011011<br>11101010001<br>10110101001<br>11000100101 | Every other symbol known to the decoder. (dB added = 3.01) |

| *SYMBOL SIZE* | SNR (dB) | 0.0 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|---|
| *1-Bit* | BER | 8.40e-04 | 4.83e-04 | 3.42e-04 | 2.82e-04 | 1.27e-04 |
| | SER @ 1 | 8.40e-04 | 4.83e-04 | 3.42e-04 | 2.82e-04 | 1.27e-04 |
| *4-Bit* | BER | 7.54e-04 | 4.64e-04 | 2.98e-04 | 2.25e-04 | 1.33e-04 |
| | SER @ 1 | 2.16e-03 | 1.36e-03 | 8.92e-04 | 7.10e-04 | 4.20e-04 |
| *8-Bit* | BER | 1.41e-03 | 8.30e-04 | 5.86e-04 | 3.76e-04 | 2.79e-04 |
| | SER @ 1 | 5.10e-03 | 3.26e-03 | 2.32e-03 | 1.55e-03 | 1.10e-03 |

Table A-9. No symbol known to the decoder, (15, 1/4) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | | No symbol known to the decoder. (dB added = 0.00) | |
|---|---|---|---|---|---|
| (15, 1/4) NASA Code | 35 | 100010110011001<br>100111010100101<br>111011011110011<br>101110101000111 | | | |
| *SYMBOL SIZE* | SNR (dB) | -0.2 | 0.0 | 0.3 | 0.5 |
| *1-Bit* | SER | 4.88e-02 | 2.80e-02 | 1.09e-02 | 5.37e-03 |
| *4-Bit* | SER | 9.36e-02 | 5.41e-02 | 2.11e-02 | 1.05e-02 |
| *8-Bit* | SER | 1.04e-01 | 6.07e-02 | 2.40e-02 | 1.22e-02 |

Table A-10. Every eighth symbol known to the decoder, (15, 1/4) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | | Every eighth symbol known to the decoder. (dB added = 0.58) | |
|---|---|---|---|---|---|
| (15, 1/4) NASA Code | 35 | 100010110011001<br>100111010100101<br>111011011110011<br>101110101000111 | | | |
| *SYMBOL SIZE* | SNR (dB) | -0.2 | 0.0 | 0.3 | 0.5 |
| *1-Bit* | BER | 9.77e-03 | 5.24e-03 | 1.87e-03 | 9.48e-04 |
| | SER @ 1 | 9.46e-03 | 5.20e-03 | 2.06e-03 | 9.96e-04 |
| | SER @ 2 | 9.60e-03 | 5.32e-03 | 1.78e-03 | 9.52e-04 |
| | SER @ 3 | 9.95e-03 | 5.34e-03 | 1.93e-03 | 9.76e-04 |
| | SER @ 4 | 9.89e-03 | 5.14e-03 | 1.81e-03 | 9.08e-04 |
| | SER @ 5 | 9.86e-03 | 5.34e-03 | 1.84e-03 | 9.80e-04 |
| | SER @ 6 | 9.88e-03 | 5.23e-03 | 1.84e-03 | 9.60e-04 |
| | SER @ 7 | 9.78e-03 | 5.10e-03 | 1.80e-03 | 8.64e-04 |
| *4-Bit* | BER | 1.11e-02 | 6.12e-03 | 2.53e-03 | 1.06e-03 |
| | SER @ 1 | 1.88e-02 | 1.00e-02 | 4.05e-03 | 1.50e-03 |
| | SER @ 2 | 2.18e-02 | 1.26e-02 | 5.10e-03 | 2.05e-03 |
| | SER @ 3 | 2.40e-02 | 1.32e-02 | 5.57e-03 | 2.61e-03 |
| | SER @ 4 | 2.38e-02 | 1.39e-02 | 5.71e-03 | 2.54e-03 |
| | SER @ 5 | 2.34e-02 | 1.33e-02 | 6.00e-03 | 2.50e-03 |
| | SER @ 6 | 2.15e-02 | 1.22e-02 | 5.30e-03 | 2.21e-03 |
| | SER @ 7 | 1.80e-02 | 9.68e-03 | 4.18e-03 | 1.71e-03 |
| *8-Bit* | BER | 1.50e-02 | 8.80e-03 | 3.80e-03 | 2.13e-03 |
| | SER @ 1 | 2.29e-02 | 1.38e-02 | 4.61e-03 | 3.33e-03 |
| | SER @ 2 | 3.40e-02 | 2.04e-02 | 8.38e-03 | 5.06e-03 |
| | SER @ 3 | 4.05e-02 | 2.46e-02 | 1.10e-02 | 6.59e-03 |
| | SER @ 4 | 4.29e-02 | 2.54e-02 | 1.16e-02 | 6.69e-03 |
| | SER @ 5 | 4.12e-02 | 2.47e-02 | 1.15e-02 | 6.27e-03 |
| | SER @ 6 | 3.47e-02 | 1.98e-02 | 9.82e-03 | 4.83e-03 |
| | SER @ 7 | 2.27e-02 | 1.27e-02 | 6.18e-03 | 3.10e-03 |

Table A-11. Every fourth symbol known to the decoder, (15, 1/4) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | Every fourth symbol known to the decoder. (dB added = 1.25) |
|---|---|---|---|
| (15, 1/4) NASA Code | 35 | 100010110011001<br>100111010100101<br>111011011110011<br>101110101000111 | |

| *SYMBOL SIZE* | SNR (dB) | -0.2 | 0.0 | 0.3 | |
|---|---|---|---|---|---|
| *1-Bit* | BER | 2.11e-03 | 1.19e-03 | 4.13e-04 | |
| | SER @ 1 | 2.14e-03 | 1.24e-03 | 3.92e-04 | |
| | SER @ 2 | 2.09e-03 | 1.14e-03 | 4.12e-04 | |
| | SER @ 3 | 2.10e-03 | 1.18e-03 | 4.34e-04 | |
| *4-Bit* | BER | 2.28e-03 | 1.37e-03 | 5.70e-04 | |
| | SER @ 1 | 4.48e-03 | 2.66e-03 | 1.14e-03 | |
| | SER @ 2 | 4.98e-03 | 3.19e-03 | 1.27e-03 | |
| | SER @ 3 | 4.38e-03 | 2.62e-03 | 1.15e-03 | |
| *8-Bit* | BER | 3.97e-03 | 2.50e-03 | 1.16e-03 | |
| | SER @ 1 | 9.23e-03 | 5.71e-03 | 2.58e-03 | |
| | SER @ 2 | 1.13e-02 | 7.14e-03 | 3.63e-03 | |
| | SER @ 3 | 8.86e-03 | 5.62e-03 | 2.74e-03 | |

Table A-12. Every other symbol known to the decoder, (15, 1/4) NASA code

| CODE | FREE DISTANCE | GENERATING POLYNOMIAL | Every other symbol known to the decoder. (dB added = 3.01) |
|---|---|---|---|
| (15, 1/4) NASA Code | 35 | 100010110011001<br>100111010100101<br>111011011110011<br>101110101000111 | |

| *SYMBOL SIZE* | SNR (dB) | -0.2 | 0.0 | 0.3 | |
|---|---|---|---|---|---|
| *1-Bit* | BER | 1.69e-04 | 8.10e-05 | 2.80e-05 | |
| | SER @ 1 | 1.69e-04 | 8.10e-05 | 2.80e-05 | |
| *4-Bit* | BER | 2.39e-04 | 9.50e-05 | 5.40e-05 | |
| | SER @ 1 | 4.92e-04 | 1.96e-04 | 1.36e-04 | |
| *8-Bit* | BER | 4.03e-04 | 2.39e-04 | 1.18e-04 | |
| | SER @ 1 | 1.06e-03 | 6.80e-04 | 3.28e-04 | |

36-61
53209

November 15, 1991

P, 16

# Some Partial-Unit-Memory Convolutional Codes

K. Abdel-Ghaffar
University of California, Davis

R. J. McEliece[1]

G. Solomon[2]

*This article presents the results of a study of a class of error-correcting codes called partial-unit-memory convolutional codes, or PUM codes for short. This class of codes, though not entirely new, has until now remained relatively unexplored. This article shows that it is possible to use the well-developed theory of block codes to construct a large family of promising PUM codes. Indeed, at the end of the article the performances of several specific PUM codes are compared with that of the Voyager standard (2, 1, 6) convolutional code. It was found that these codes can outperform the Voyager code with little or no increase in decoder complexity. This suggests that there may very well be PUM codes that can be used for deep-space telemetry that offer both increased performance and decreased implementational complexity over current coding systems.*

## I. Introduction

This article gives a general construction for, and several interesting examples of, partial-unit-memory (PUM) convolutional codes. First, some definitions and notation are established.

A convolutional code $\mathcal{C}$ of length $n$ and dimension $k$ over a field $F$ is defined by an encoder

$$G(D) = G_0 + G_1 D + \cdots + G_M D^M \qquad (1)$$

where $G_0, G_1, \ldots, G_M$ are $k \times n$ matrices with entries from $F$.[3] The ratio $R = k/n$ is called the rate of the code, and $M$ is the code's memory. If $u(D) = u_0 + u_1 D + u_2 D^2 + \cdots$ is the input to the encoder (where the $u_i$'s are elements of $F$, and $D$ is an indeterminate), then $x(D) = u(D)G(D)$ is the output, which is also called a codeword. The encoder is said to be noncatastrophic if no infinite-weight input produces a finite-weight output. The free distance, $d_{\text{free}}$, of a convolutional code $\mathcal{C}$ is defined to be the minimum weight of any nonzero codeword $x(D) \in \mathcal{C}$. If the encoder $G(D)$ is noncatastrophic, then $d_{\text{free}}$ is the minimum weight of all codewords $u(D)G(D)$ generated by inputs $u(D)$ of finite weight. All other things being equal, it is generally desirable to have the quantity

[1] Consultant to the Communications Systems Research Section from the California Insitute of Technology.

[2] Independent consultant to the Communications Systems Research Section.

[3] In this article, it is always assumed that $F = GF(2)$, but most or all of the results generalize easily to other finite fields.

$Q = Rd_{\text{free}}$ as large as possible, since $Q$ is the asymptotic coding gain of the code, which is a good measure of the communications improvement afforded by the use of the code.

A convolutional code $\mathcal{C}$ has state complexity $m$ if the sum of the maximum degrees of the rows of $G(D)$ is $m$. This terminology reflects the fact that a physical encoder for $\mathcal{C}$ based on $G(D)$ has $2^m$ states. It is desirable to have $m$ as small as possible, since the computational complexity of the Viterbi decoding algorithm for $\mathcal{C}$ is proportional to $2^m$.

The notation "$[n, k, m, d]$ code" is introduced to describe a convolutional code of length $n$, dimension $k$, state complexity $m$, and free distance $d$. The notation "$(n, k, m)$" code is sometimes used to describe the same code without explicitly referring to its free distance. For example, in this notation, an $[n, k, d]$ block code, which can be viewed as a convolutional code with $M = 0$, is both an $[n, k, 0, d]$ and an $(n, k, 0)$ convolutional code. For a given $n$, $k$, and $m$, a code for which $d_{\text{free}}$ is as large as possible is said to be an optimal (more properly, distance optimal) code.

A convolutional code with $M = 0$ is just a block code. A convolutional code with $M = 1$ is called a unit-memory convolutional code. Unit-memory codes seem to form a class that lies halfway between block and convolutional codes. They were first studied seriously by Lee [5], who found a number of interesting examples of unit-memory convolutional codes. Thommesen and Justesen [10] have obtained bounds on the performance of unit-memory codes, and Justesen, Paaske, and Ballan [3] have constructed a class of unit-memory codes which they call quasi-cyclic codes. This article studies another subclass of unit-memory codes called partial-unit-memory codes.

For a unit-memory convolutional code, the state complexity is just the number of nonzero rows of $G_1$. If some of the rows of $G_1$ are zero, i.e., if $m < k$, then it is said that the code is a partial-unit-memory (PUM) convolutional code. Partial-unit-memory codes were introduced by Lauer [4], who constructed several optimal PUM codes. Some general constructions and further examples of PUM codes (under the name finite-state codes) were given in [7] and [8]. This article should be viewed as a continuation of these earlier studies, in which, among other things, it is shown that many of these earlier results follow from the authors' methods. (For example, Lauer's equidistant PUM codes appear in Example 2 in Section III of this article, and the general construction of Theorem 5 in [7] appears as Corollary 4 in Section II.)

Here is a summary of this article. Section II gives a general construction for PUM codes based on the existence of certain block codes. Roughly speaking, the main result is that if there exist two distinct $[n, k, d_0]$ block codes with a common $[n, k^*, d^*]$ subcode, then there exists a noncatastrophic $[n, k, k - k^*, d]$ PUM code with $d \geq \min(d^*, 2d_0)$. (As a point of comparison, Theorem 5 in [7] shows that if there exists a single $[n, k, d_0]$ block code with an $[n, k^*, d^*]$ subcode, then there exists a noncatastrophic $[n, k - 1, k - k^* - 1, d]$ PUM code with $d \geq \min(d^*, 2d_0)$.) Since there is a huge existing catalog of block codes, what this means is that it is possible to construct a very large number of interesting PUM codes. This is illustrated in Section III with several examples called Hamming, Reed-Muller, and Golay PUM codes. While these examples are possibly interesting and potentially important, the authors believe that they have only scratched the surface, and hope that future authors, using these techniques, or ones of their own devising, will unearth many more examples.

## II. Main Results

**Theorem 1.** Suppose that $\mathcal{C}_0$ is an $[n, k, d_0]$ linear block code, $\mathcal{C}_1$ is an $[n, k, d_1]$ linear block code, and $\mathcal{C}_0 \neq \mathcal{C}_1$. Suppose further that $\mathcal{C}_0$ and $\mathcal{C}_1$ contain a common subcode $\mathcal{C}^*$, which is an $[n, k^*, d^*]$ code. Then there exists a noncatastrophic $[n, k, m, d]$ PUM convolutional code, with $m = k - k^*$ and $d \geq \min(d^*{}_{\text{,}} d_0 + d_1)$.

**Proof:** Begin by choosing $k \times n$ generator matrices $G_0$ and $G_1$ for $\mathcal{C}_0$ and $\mathcal{C}_1$ of the form

$$G_0 = \begin{bmatrix} K^* \\ K_0 \end{bmatrix} \qquad G_1 = \begin{bmatrix} K^* \\ K_1 \end{bmatrix} \tag{2}$$

where $K^*$ is a $k^* \times n$ generator matrix for $\mathcal{C}^*$. Note that both $K_0$ and $K_1$ are $m \times n$ matrices; for future reference, let $\mathcal{C}_0^*$ and $\mathcal{C}_1^*$ be the corresponding codes, i.e.,

$$\mathcal{C}_0^* = \langle K_0 \rangle \tag{3}$$

$$\mathcal{C}_1^* = \langle K_1 \rangle \tag{4}$$

Next, define the matrix $G_1^0$ as

$$G_1^0 = \begin{bmatrix} \mathrm{O} \\ K_1 \end{bmatrix} \tag{5}$$

where O is a $k^* \times n$ matrix of 0's. Then define a $k \times n$ polynomial matrix $G(D)$ as follows:

$$G(D) = G_0 + G_1^0 D \tag{6}$$

Plainly, $G(D)$ is the generator matrix for an $(n, k, m)$ PUM code. The proof will be complete when the following two things are shown: (1) $d_{\text{free}} \geq \min(d^*, d_0 + d_1)$, and (2) $G(D)$ is noncatastrophic. Begin with the assertion about $d_{\text{free}}$.

Assume that $u(D) = u_0 + u_1 D + u_2 D^2 + \cdots$ is a finite nonzero input sequence, where $u_0 \neq 0$ and each $u_i$ is a $k$-dimensional row vector. Then the corresponding output sequence is $x(D) = u(D)G(D) = x_0 + x_1 D + x_2 D^2 + \cdots$, where

$$\left.\begin{aligned} x_0 &= u_0 G_0 \\ x_i &= u_i G_0 + u_{i-1} G_1^0, \qquad \text{for } i \geq 1 \end{aligned}\right\} \tag{7}$$

If $u = (\mu_1, \mu_2, \ldots, \mu_k)$ is a $k$-dimensional vector, $u^L$ (the left part of $u$) and $u^R$ (the right part of $u$) are defined as follows:

$$\left.\begin{aligned} u^L &= (\mu_1, \ldots, \mu_{k^*}) \\ u^R &= (\mu_{k^*+1}, \ldots, \mu_k) \end{aligned}\right\} \tag{8}$$

Then [see Eq. (2)], for any vector $u$, one has

$$uG_0 = u^L K^* + u^R K_0, \qquad uG_1^0 = u^R K_1 \tag{9}$$

After combining Eq. (8) with Eq. (6), one has

$$\left.\begin{aligned} x_0 &= u_0^L K^* + u_0^R K_0 \\ x_i &= u_i^R K_0 + [u_i^L, u_{i-1}^R] G_1, \qquad \text{for } i \geq 1 \end{aligned}\right\} \tag{10}$$

Now, either $[u_i^L, u_{i-1}^R] = 0$ for all $i \geq 1$, or not. It will be shown that weight $(x(D)) \geq d^*$ in the first case, and that weight $(x(D)) \geq d_0 + d_1$ in the second case.

If $[u_i^L, u_{i-1}^R] = 0$ for all $i \geq 1$, then by Eq. (6) and Eq. (9), one has

$$x(D) = u_0 G_0 = u_0^L K^* \tag{11}$$

which means that $x(D)$ is a nonzero vector in the rowspace of $K^*$, i.e., a nonzero codeword in $\mathcal{C}^*$, and so weight $(x(D)) \geq d^*$ in this case.

If, on the other hand, $[u_i^L, u_{i-1}^R] \neq 0$ for some $i \geq 1$, let $M$ denote the largest such index. Then, $u_M^R = 0$, and Eq. (9) implies

$$x(D) = u_0 G_0 + \cdots + [u_M^L, u_{M-1}^R] G_1 D^M \tag{12}$$

But $u_0 G_0$ is a nonzero word from $\mathcal{C}_0$, and so has weight $\geq d_0$, and $[u_M^L, u_{M-1}^R] G_1$ is a nonzero word from $\mathcal{C}_1$, and so has weight $\geq d_1$; thus, weight $(x(D)) \geq d_0 + d_1$ in this case, which proves the assertion about $d_{\text{free}}$.

It remains to be shown that $G(D)$ can be chosen noncatastrophically. Lemma 1, which follows, tells, in principle, whether a given $G(D)$ is catastrophic or not. Lemma 2 then tells that it is always possible to choose the matrices $G_0$ and $G_1$ so that $G(D)$ is noncatastrophic.

**Lemma 1.** Let the linear transformation $T : \mathcal{C}_0 \rightarrow \mathcal{C}_1$ be defined by $uG_0 \rightarrow uG_1$. Then $G(D)$ is noncatastrophic if and only if every subspace of $\mathcal{C}_0$ fixed by $T$ is a subspace of $\mathcal{C}^*$.

**Proof:** Denote the rows of $K^*$ by $(x_1, x_1, \ldots, x_{k^*})$, the rows of $K_0$ by $(y_1, y_1, \ldots, y_m)$, and the rows of $K_1$ by $(z_1, z_2, \ldots, z_m)$. Then $T$ is completely characterized by the $k$ values

$$Tx_i = x_i, \qquad \text{for } i = 1, 2, \ldots, k^* \tag{13}$$

$$Ty_i = z_i, \qquad \text{for } i = 1, 2, \ldots, m \tag{14}$$

Note that Eq. (13) says that $T$ not only fixes $\mathcal{C}^*$, it fixes $\mathcal{C}^*$ pointwise.

It is first assumed that every $T$-fixed subspace of $\mathcal{C}_0$ is a subspace of $\mathcal{C}^*$, and then shown that $G(D)$ is noncatastrophic. Let $u(D)$ be a nonzero input such that the corresponding output $x(D)$ is finite, i.e., $x_i = 0$ for $i > i_0$. If one defines

$$a_i = u_i^L K^* \in \mathcal{C}^* \tag{15}$$

$$b_i = u_i^R K_0 \in \mathcal{C}_0^* \tag{16}$$

one has, by Eq. (9),

$$x_i = a_i + b_i + Tb_{i-1}, \qquad \text{for } i \geq 1 \tag{17}$$

Thus, since it is assumed that $x_i = 0$ for $i > i_0$, it follows that (recall that the codes are binary)

$$Tb_{i-1} = b_i + a_i, \qquad \text{for } i > i_0 \tag{18}$$

so that $\langle b_{i_0}, b_{i_0+1}, \dots \rangle + \mathcal{C}^*$ is a $T$-fixed subspace of $\mathcal{C}_0$. However, it is assumed that all $T$-fixed subspaces of $\mathcal{C}_0$ are subspaces of $\mathcal{C}^*$, and so $b_i \in \mathcal{C}^*$ for all $i \geq i_0$. But since the rows of $K_0$ are linearly independent of the rows of $K^*$, this means that $b_i = 0$, and so $u_i^R = 0$, for all $i \geq i_0$. But then, by Eq. (17), $a_i = 0$, and so $u_i^L = 0$, for all $i > i_0$. Thus, the input $u(D)$ is necessarily finite, and so $G(D)$ is noncatastrophic.

Conversely, suppose that $B$ is a nonzero $T$-fixed subspace of $\mathcal{C}_0$ that properly contains $\mathcal{C}^*$. Choose $a_0 \in \mathcal{C}^*$ and $b_0 \in B - \mathcal{C}_0^*$ arbitrarily. Now, since $B$ is $T$-fixed, $Tb_0$ is also in $B$, and so it can be decomposed uniquely into the sum of an element of $\mathcal{C}^*$, which is called $a_1$, and an element of $\mathcal{C}_0^*$, which is called $b_1$. Note that since $b_0 \neq 0$, then $b_1 \neq 0$ also, for otherwise $T$ would map the $k^* + 1$ dimensional space $\mathcal{C}^* + \langle b_0 \rangle$ into the $k^*$-dimensional space $\mathcal{C}^*$. This process is continued inductively by constructing an infinite sequence of pairs $(a_i, b_i)$, with $a_i \in \mathcal{C}^*$, $b_i \in \mathcal{C}_0^*$, $b_i \neq 0$, such that

$$Tb_i = a_{i+1} + b_{i+1}, \qquad \text{for } i \geq 0 \tag{19}$$

Now for each $i \geq 1$, define the vector $u_i$ as follows:

$$u_i^L K^* = a_i \tag{20}$$

$$u_i^R K_0 = b_i \tag{21}$$

Since $b_i \neq 0$, then by Eq. (20), $u_i \neq 0$, and so the sequence $(u_i)$ is an infinite sequence of nonzero elements. It will now be shown that, if $(u_i)$ is the input, then the corresponding output is finite. One has

$$\begin{aligned} Tb_i &= T(u_i^R K_0) \\ &= T([0, u_i^R]G_0) \\ &= [0, u_i^R]G_1, \qquad \text{by definition of } T \\ &= u_i^R K_1 \end{aligned}$$

and so by Eq. (9), for $i \geq 1$,

$$\begin{aligned} x_i &= u_i^L K^* + u_i^R K_0 + u_{i-1}^R K_1, \qquad &\text{by Eq. (9)} \\ &= a_i + b_i + Tb_{i-1}, \qquad &\text{by Eqs. (19) and (20)} \\ &= 0, \qquad &\text{by Eq. (18)} \end{aligned}$$

Thus, the infinite input sequence $(u_i)$ produces a finite output sequence $(x_i)$, and so $G(D)$ is catastrophic, as was asserted. □

**Corollary 1.** If $\mathcal{C}_0 \cap \mathcal{C}_1 = \mathcal{C}^*$, then any generator matrix of the form Eq. (6) is noncatastrophic.

**Proof:** If $B$ is a $T$-fixed subspace of $T$, then since $T : \mathcal{C}_0 \to \mathcal{C}_1$ and $B = T(B)$, $B$ is a subspace of both $\mathcal{C}_0$ and $\mathcal{C}_1$. Thus, $B \subseteq \mathcal{C}_0 \cap \mathcal{C}_1 = \mathcal{C}^*$, and so by Lemma 1, $G(D)$ is noncatastrophic. □

Lemma 1 allows one to tell, in principle, whether or not a given $G(D)$ is catastrophic. Corollary 1 assures that if $\mathcal{C}_0 \cap \mathcal{C}_1 = \mathcal{C}^*$, then nothing can go wrong. However, if $\mathcal{C}_0 \cap \mathcal{C}_1 \supset \mathcal{C}^*$, more work is necessary to find a noncatastrophic generator matrix. Lemma 2, which follows, gives an explicit construction for a noncatastrophic $G(D)$ in the general case.

**Lemma 2.** Suppose that $\mathcal{C}_0$ and $\mathcal{C}_1$ are subspaces of $V_n(F)$, the $n$-dimensional vector space over $F$, with $\mathcal{C}_0 \neq \mathcal{C}_1$ but $\dim(\mathcal{C}_0) = \dim(\mathcal{C}_1) = k$, and that $\mathcal{C}^*$ is a subspace of both $\mathcal{C}_0$ and $\mathcal{C}_1$, with $\dim(\mathcal{C}^*) = k^*$. Then, if $(u_1, u_2, \dots, u_{k^*})$ is a basis for $\mathcal{C}^*$, there exist bases for $\mathcal{C}_0$ and $\mathcal{C}_1$ of the form

$$\left.\begin{aligned} \langle \mathcal{C}_0 \rangle &= (u_1, \dots, u_{k^*}, \ \alpha_1, \dots, \alpha_m) \\ \langle \mathcal{C}_1 \rangle &= (u_1, \dots, u_{k^*}, \ \beta_1, \dots, \beta_m) \end{aligned}\right\} \tag{22}$$

such that the linear transformation $T : \mathcal{C}_0 \to \mathcal{C}_1$ defined by

$$\left.\begin{aligned} Tu_i &= u_i, \qquad &\text{for } i = 1, \dots, k^* \\ T\alpha_i &= \beta_i, \qquad &\text{for } i = 1, \dots, m \end{aligned}\right\} \tag{23}$$

fixes no subspace of $\mathcal{C}_0$ larger than $\mathcal{C}^*$.

**Proof:** Begin by constructing two descending sequences of subspaces $(\mathcal{A}_i)$ and $(\mathcal{B}_i)$:

$$\mathcal{C}_0 = \mathcal{A}_0 \supset \mathcal{A}_1 \supset \cdots \supset \mathcal{A}_{N+1} = \mathcal{C}^*$$

$$\mathcal{C}_1 = \mathcal{B}_0 \supset \mathcal{B}_1 \supset \cdots \supset \mathcal{B}_{N+1} = \mathcal{C}^*$$

such that

$$\dim(\mathcal{A}_i) = \dim(\mathcal{B}_i), \qquad i = 0, 1, \ldots, N+1$$

$$\mathcal{A}_{i+1} = \mathcal{A}_i \cap \mathcal{B}_i, \qquad i = 0, 1, \ldots, N$$

Figure 1 illustrates the construction of Lemma 2. This construction can be done inductively as follows. Assume that $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_i$ and $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_i$ have already been constructed. (For $i = 0$, this simply requires setting $\mathcal{A}_0 = \mathcal{C}_0$ and $\mathcal{B}_0 = \mathcal{C}_1$.) Let $\mathcal{A}_{i+1} = \mathcal{A}_i \cap \mathcal{B}_i$. If $\mathcal{A}_{i+1} = \mathcal{C}^*$, define $\mathcal{B}_{i+1} = \mathcal{C}^*$, $N = i$, and stop. Otherwise, one has $\mathcal{C}^* \subset \mathcal{A}_{i+1} \subset \mathcal{B}_i$, and so by Lemma A2 in Appendix A, there exists a subspace $\mathcal{B}_{i+1} \neq \mathcal{A}_{i+1}$ such that $\mathcal{C}^* \subset \mathcal{B}_{i+1} \subset \mathcal{B}_i$, with $\dim \mathcal{B}_{i+1} = \dim \mathcal{A}_{i+1}$.

Now define integers $k_0, \ldots, k_N$ by $k_i = \dim(\mathcal{A}_i) - k^*$, so that

$$k_N < k_{N-1} < \cdots < k_0 = m$$

Next, using Lemma A1 in Appendix A, choose bases $\langle u_1, \ldots, u_{k^*}, \ \alpha_1, \ldots, \alpha_m\rangle$ for $\mathcal{C}_0$ and $\langle u_1, \ldots, u_{k^*}, \ \beta_1, \ldots, \beta_m\rangle$ for $\mathcal{C}_1$ such that

$$\left.\begin{aligned} \langle u_1, \ldots, u_{k^*}, \ \alpha_1, \ldots, \alpha_{k_i}\rangle &= \mathcal{A}_i \\ \langle u_1, \ldots, u_{k^*}, \ \beta_1, \ldots, \beta_{k_i}\rangle &= \mathcal{B}_i \end{aligned}\right\} \qquad (24)$$

for $i = 1, 2, \ldots, N$. Now define the transformation $T$ as in Eq. (22). Plainly, $T$ fixes $\mathcal{C}^*$ pointwise, and also, from Eq. (23),

$$T : \mathcal{A}_i \rightarrow \mathcal{B}_i, \qquad \text{for } i = 0, 1, \ldots, N \qquad (25)$$

If now $\mathcal{D}$ is a subspace of $\mathcal{C}_0$ fixed by $T$, then $\mathcal{D} \subseteq \mathcal{C}_0 = \mathcal{A}_0$, and $\mathcal{D} = T(\mathcal{D}) \subseteq T(\mathcal{C}_0) = \mathcal{C}_1 = \mathcal{B}_0$, so that $\mathcal{D} \subseteq \mathcal{A}_0 \cap \mathcal{B}_0 = \mathcal{A}_1$. Also, $\mathcal{D} = T(\mathcal{D}) \subseteq T(\mathcal{A}_1) = \mathcal{B}_1$, using Eq. (24), so that $\mathcal{D} \subseteq \mathcal{A}_1 \cap \mathcal{B}_1 = \mathcal{A}_2$. Continuing inductively, one finds that in fact $\mathcal{D} \subseteq \mathcal{A}_3, \ldots, \mathcal{D} \subseteq \mathcal{A}_{N+1} = \mathcal{C}^*$. Thus, a linear transformation $T$ is constructed such that any $T$-fixed subspace of $\mathcal{C}_0$ is a subspace of $\mathcal{C}^*$. ❐

Lemma 2 tells how to construct noncatastrophic generator matrices $G_0$ and $G_1$: Just let the rows of $G_0$ be the vectors $(u_1, \ldots, u_{k^*}, \ \alpha_1, \ldots, \alpha_m)$, and let the rows of $G_1$ be the vectors $(u_1, \ldots, u_{k^*}, \ \beta_1, \ldots, \beta_m)$ in Eq. (21). Then, the mapping $T : \mathcal{C}_0 \rightarrow \mathcal{C}_1$ defined by $uG_0 \rightarrow uG_1$ is the same as the mapping described in Lemma 2, and so the resulting $G(D)$ is noncatastrophic. This completes the proof of Theorem 1. ❐

**Corollary 2.** Suppose that $\mathcal{C}_0$ is an $[n, k, d_0]$ linear block code, and that $\mathcal{C}^*$ is an $[n, k^*, d^*]$ code, which is a subcode of $\mathcal{C}_0$. If the automorphism group of $\mathcal{C}^*$ contains a permutation that does not fix $\mathcal{C}_0$, then there exists an $[n, k, m, d]$ PUM convolutional code, with $m = k - k^*$ and $d \geq \min(d^*, 2d_0)$.

**Proof:** Let $\pi$ be an automorphism of $\mathcal{C}^*$ that does not fix $\mathcal{C}_0$, and let $\mathcal{C}_1 = \mathcal{C}_0^{\pi}$. Then, $\mathcal{C}_1$ is an $[n, k, d_0]$ code not equal to $\mathcal{C}_0$. Now apply Theorem 1. ❐

**Corollary 3.** If $\mathcal{C}_0$ is an $[n, k, d_0]$ linear block code that contains the all-ones vector, and if $k \neq 1, n-1, n$, then there exists an $(n, k, k-1)$ PUM code with $d_{\text{free}} \geq 2d_0$.

**Proof:** Here, Corollary 2 is applied, with $\mathcal{C}^*$ being the $[n, 1, n]$ code consisting of the two vectors $[00\cdots]$ and $[11\cdots1]$. Clearly $\mathcal{C}^*$ is fixed by all permutations of $\{1, 2, \ldots, n\}$. Furthermore, the only binary linear codes that are fixed by all permutations of $\{1, 2, \ldots, n\}$ have dimensions 0, 1, $n-1$, or $n$, so there must be an automorphism of $\mathcal{C}^*$ that doesn't fix $\mathcal{C}_0$. Thus, by Corollary 2, there exists an $(n, k, k-1)$ PUM code with $d_{\text{free}} \geq \min(2d_0, n)$. However, since $k \geq 2$, the mimimum distance $d_0$ of $\mathcal{C}_0$ must be $< n$; and since $\mathcal{C}_0$ contains the all-ones vector, there must be a word of weight $n - d_0$. Hence, $n - d_0 \leq d_0$, and so $d_0 \leq n/2$. Hence, $\min(2d_0, n) = 2d_0$, so that in fact $d_{\text{free}} \geq 2d_0$. ❐

**Corollary 4.** (Same as Theorem 5 in [7]). Suppose that $\mathcal{C}_0$ is an $[n, k, d_0]$ linear block code, $\mathcal{C}^*$ is an $[n, k^*, d^*]$ code that is a subcode of $\mathcal{C}_0$, and $k - k^*2$. Then, for every integer $i$ in the range $1 \leq i \leq k - k^* - 1$, there is a noncatastrophic $[n, k-i, k-i-k^*, d]$ PUM code with $\delta \geq \min(d^*, 2d_0)$.

**Proof:** Let $\mathcal{C}_0'$ be any $(n, k-i)$ subcode of $\mathcal{C}_0$ that contains $\mathcal{C}^*$. (The conditions on $i$ guarantee that $\dim \mathcal{C}^* z \leq \dim \mathcal{C}_0' \leq \dim \mathcal{C}_0$, so this is possible.) By Lemma A2, there exists a subcode $\mathcal{C}_1'$ not equal to $\mathcal{C}_0'$ but having the same dimension, and also lying between $\mathcal{C}_0$ and $\mathcal{C}^*$. Thus, $\mathcal{C}_0'$ and $\mathcal{C}_1'$ are both $[n, k-i, d']$ block codes, with $d' \geq d_0$. By applying Theorem 1 to the codes $\mathcal{C}_0'$, $\mathcal{C}_1'$, and $\mathcal{C}^*$, one obtains a noncatastrophic $[n, k-i, k-i-k^*, d]$ PUM code with $\delta \geq \min(d^*, 2d') \geq \min(d^*, 2d_0)$. ❐

## III. Examples

In this section, four examples of PUM codes are presented, that were constructed with the help of the results in Section II. Example 1 describes a Hamming $[8,4,3,8]$ PUM code, which was originally discovered by Lauer [4]. Example 2 gives a generalization of Example 1 to a class of Reed-Muller $[\nu 2^\mu, \mu+1, \mu, \nu 2^\mu]$ PUM codes, one code for each pair of positive integers $(\mu, \nu)$ except $(1,1)$ and $(2,1)$. (The code of Example 1 corresponds to the pair $(3,1)$.) The codes in Example 2 were also found, using different methods, by Lauer. Finally, in Examples 3 and 4, two new PUM Golay codes, with parameters $[24,12,7,12]$ and $[24,12,10,16]$, are presented.

**Example 1 (a Hamming PUM Code).** Let $\mathcal{C}_0'$ be the $[7,4,3]$ binary cyclic code with generator polynomial $g_0(x) = 1 + x + x^3$, and let $\mathcal{C}_1'$ be the $[7,4,3]$ binary cyclic code with generator polynomial $g_1(x) = 1 + x^2 + x^3$. Take as a generator matrix for $\mathcal{C}_0'$ the $4 \times 7$ binary matrix $G_0'$, whose rows are $g_1(x)g_0(x)$, $g_0(x)$, $xg_0(x)$, and $x^2 g_0(x)$, and for $\mathcal{C}_1'$ the $4 \times 7$ binary matrix $G_1'$, whose rows are $g_0(x)g_1(x)$, $g_1(x)$, $xg_1(x)$, and $x^2 g_1(x)$, i.e.,

$$G_0' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

$$G_1' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Now, if each code is extended to length 8 by appending an overall parity-check, one obtains codes $\mathcal{C}_0$ and $\mathcal{C}_1$, both of which are binary $[8,4,4]$ codes with generator matrices

$$G_0 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

$$G_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Since $\mathcal{C}_0' \cap \mathcal{C}_1'$ is the binary $[7,1,7]$ repetition code, it follows that $\mathcal{C}_0 \cap \mathcal{C}_1$ is the binary $[8,1,8]$ repetition code. Thus, in Theorem 1, $\mathcal{C}^*$ can be taken to be the $[8,1,8]$ repetition code with the $1 \times 8$ generator matrix

$$K^* = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

It follows from Corollary 1 that the matrix

$$G(D) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1+D & 1+D & 1 & D & 1+D & 0 & 0 & 0 \\ 1+D & 0 & 1+D & 1 & D & 1+D & 0 & 0 \\ 1+D & 0 & 0 & 1+D & 1 & D & 1+D & 0 \end{pmatrix}$$

generates a noncatastrophic $[8,4,3,8]$ PUM code. Furthermore, from [4] (Formula (3) with $L = 0$) or [7] (Corollary 1 to Theorem 1, with $L = 1$), any $(8,4,3)$ convolutional code must have $d_{\text{free}} \leq 8$, so this code is optimal.[4] □

**Example 2 (Some Reed-Muller PUM Codes).** Let $\mu$ and $\nu \geq 1$ be positive integers. Let $A_\mu$ be the $[2^\mu, \mu+1, 2^{\mu-1}]$ first-order Reed-Muller code, and let $B_\mu$ be the $[2^\mu, 1, 2^\mu]$ zeroth-order Reed-Muller code (a repetition code), which is a subcode of $A_\mu$.[5] Now let $\mathcal{C}_0(\mu,\nu)$ be the $[\nu 2^\mu, \mu+1, \nu 2^{\mu-1}]$ code obtained by repeating $A_\mu$ $\nu$ times, and let $\mathcal{C}^*(\mu,\nu)$ be the $[\nu 2^\mu, 1, \nu 2^\mu]$ code obtained by repeating $B_\mu$ $\nu$ times. Then, according to Corollary 3, unless $(\mu,\nu) = (1,1)$ or $(2,1)$, there exists a noncatastrophic $[\nu 2^\mu, \mu+1, \mu, \nu 2^\mu]$ PUM code. These codes are all optimal by the above-cited bounds in [4] or [7]. (This family of codes was originally constructed by Lauer [4], using a different approach. He called them equidistant PUM

[4] This code first appeared in the literature in [4], Table 1. It is apparently used by the Soviets in their *Regatta* space communication system.

[5] MacWilliams and Sloane [6], Chapter 13, is a good reference for Reed-Muller codes.

codes. A similar family of $[2^\mu, \mu, \mu - 1, 2^\mu]$ codes was constructed in [7], Example 4.) □

**Example 3 (the [24, 12, 7, 12] Golay PUM Code).** It is well known that there exists a [24, 12, 8] binary linear code; viz., the famous Golay code. It turns out that there are two isomorphic copies of the Golay code that contain a common [24, 5, 12] subcode, so that by Theorem 1, there exists a noncatastrophic [24, 12, 7, 12] PUM convolutional code. In this example, the construction of this code is detailed.

Define, for $A, B, C, D, E, F \in GF(8)$, the following two functions:

$$f_{A,B,C}(x, y) = \mathrm{Tr}(Axy) + \mathrm{Tr}\big((B + Cy)x^6\big) \qquad (26)$$

$$g_{D,E,F}(x, y) = \mathrm{Tr}\big((Dy + E)x\big) + \mathrm{Tr}\big((Fy)x^6\big) \qquad (27)$$

where $\mathrm{Tr}(x) = x + x^2 + x^4$ is the trace mapping from $GF(8)$ to $GF(2)$. Then, if $\beta$ is a fixed nonzero element of trace 0 in $GF(8)$, the following set of $2^{12}$ length-24 vectors is a [24, 12, 8] Golay code, which is called $A_0$:

$$A_0 = \big[f_{A,B,C}(x, \beta) + \epsilon_0 | f_{A,B,C}(x, \beta^2) + \epsilon_1 | \times f_{A,B,C}(x, \beta^4) + \epsilon_2\big]_{x \in GF(8)} \qquad (28)$$

In Eq. (28), the parameters $A$, $B$, and $C$ assume all values in $GF(8)$, and the parameters $\epsilon_0$, $\epsilon_1$, and $\epsilon_2$ assume all values in $GF(2)$. The proof that $A_0$ is indeed a [24, 12, 8] code appears in Appendix B as Lemma B5.

Similarly, the following set of $2^{12}$ length-24 vectors is another [24, 12, 8] Golay code, which is called $B_0$:

$$B_0 = \big[g_{D,E,F}(x, \beta) + \delta_0 | g_{D,E,F}(x, \beta^2) + \delta_1 | g_{D,E,F}(x, \beta^4) + \delta_2\big]_{x \in GF(8)} \qquad (29)$$

In Eq. (29), the parameters $D$, $E$, and $F$ assume all values in $GF(8)$, and the parameters $\delta_0$, $\delta_1$, and $\delta_2$ assume all values in $GF(2)$. The proof that $B_0$ is indeed a [24, 12, 8] code appears in Appendix B as Lemma B6.

In Appendix B (Lemma B10), it is shown that $A_0 \cap B_0$ is a [24, 9, 8] code consisting of the following set of vectors:

$$A_1 = [f_{A,0,C}(x, \beta) + \epsilon_0 | f_{A,0,C}(x, \beta^2) + \epsilon_1 | \times f_{A,0,C}(x, \beta^4) + \epsilon_2]_{x \in GF(8)} \qquad (30)$$

The code $A_1$, in turn, contains a [24, 5, 12] subcode consisting of the following set of vectors:

$$A_3 = \big[f_{0,0,C}(x, \beta) + \epsilon_0 | f_{0,0,C}(x, \beta^2) + \epsilon_1 | \times f_{0,0,C}(x, \beta^4) + \epsilon_0 + \epsilon_1\big]_{x \in GF(8)} \qquad (31)$$

(see Lemma B8 in Appendix B). Finally, $A_3$ contains a [24, 2, 16] subcode $A_4$:

$$A_4 = [\epsilon_0 | \epsilon_1 | \epsilon_0 + \epsilon_1]_{x \in GF(8)} \qquad (32)$$

(see Lemma B9). It follows from Theorem 1 that there exists both a [24, 12, 7, 12] code and a [24, 12, 10, 16] code, and by the bounds in [4] and [7], they are optimal.[6] To actually construct noncatastrophic generator matrices for these two codes, however, more work is necessary. Here are the needed intermediate subspaces (see Fig. 2):

$$B_1 = \big[g_{0,E,F}(x, \beta) + \delta_0 | g_{0,E,F}(x, \beta^2) + \delta_1 | \times g_{0,E,F}(x, \beta^4) + \delta_2\big]_{x \in GF(8)} \qquad (33)$$

$$A_2 = \big[f_{0,0,C}(x, \beta) + \epsilon_0 | f_{0,0,C}(x, \beta^2) + \epsilon_1 | \times f_{0,0,C}(x, \beta^4) + \epsilon_2\big]_{x \in GF(8)} \qquad (34)$$

$$B_2 = \big[g_{0,e,0}(x, \beta) + \delta_0 | g_{0,e,0}(x, \beta^2) + \delta_1 | \times g_{0,e,0}(x, \beta^4) + \delta_2\big]_{x \in GF(8)} \qquad (35)$$

In Eq. (35), $e$ assumes only the two values 0 and 1.

[6] In [8], [24, 5, 12] and [24, 2, 16] subcodes of a [24, 12, 8] Golay code were found, which led, via Corollary 4, to the construction of both [24, 11, 6, 12] and [24, 11, 9, 16] PUM codes.

In order to show that the subspaces in Fig. 2 behave as depicted, the following must be proved:

$$A_0 \cap B_0 = A_1 \tag{36}$$

$$A_1 \cap B_1 = A_2 \tag{37}$$

$$A_2 \cap B_2 = A_3 \tag{38}$$

These relationships are proved in Appendix B in Lemmas B10, B11, and B12.

It thus follows that for the [24, 12, 7, 12] code, a noncatastrophic choice for $G_0$ and $G_1$ is as follows:

$$G_0 = \begin{bmatrix} f_{001} \\ f_{00\beta} \\ f_{00\beta^2} \\ \mathbf{011} \\ \mathbf{101} \\ \mathbf{100} \\ f_{100} \\ f_{\beta 00} \\ f_{\beta^2 00} \\ f_{010} \\ f_{0\beta 0} \\ f_{0\beta^2 0} \end{bmatrix} \qquad G_1 = \begin{bmatrix} f_{001} \\ f_{00\beta} \\ f_{00\beta^2} \\ \mathbf{011} \\ \mathbf{101} \\ g_{010} \\ \mathbf{100} \\ g_{0\beta 0} \\ g_{0\beta^2 0} \\ g_{100} \\ g_{\beta 00} \\ g_{\beta^2 00} \end{bmatrix}$$

Here, $f_{A,B,C}$ denotes the length-24 vector obtained by taking $\epsilon_0 = \epsilon_1 = \epsilon_2 = 0$ in Eq. (28), and $g_{A,B,C}$ denotes the length-24 vector obtained by taking $\delta_0 = \delta_1 = \delta_2 = 0$ in Eq. (29). The first five rows of $G_0$ and $G_1$ are identical, and they generate the [24, 5, 12] subcode referred to above. In binary, these two matrices are as follows:

$$G_0 = \begin{bmatrix} 01110100 & 00111010 & 01001110 \\ 00111010 & 10011100 & 10100110 \\ 10011100 & 01001110 & 11010010 \\ 00000000 & 11111111 & 11111111 \\ 11111111 & 00000000 & 11111111 \\ 11111111 & 00000000 & 00000000 \\ 00101110 & 01011100 & 01110010 \\ 01011100 & 10111000 & 11100100 \\ 10111000 & 01110010 & 11001010 \\ 11101000 & 11101000 & 11101000 \\ 01110100 & 01110100 & 01110100 \\ 00111010 & 00111010 & 00111010 \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 01110100 & 00111010 & 01001110 \\ 00111010 & 10011100 & 10100110 \\ 10011100 & 01001110 & 11010010 \\ 00000000 & 11111111 & 11111111 \\ 11111111 & 00000000 & 11111111 \\ 10010110 & 10010110 & 10010110 \\ 11111111 & 00000000 & 00000000 \\ 00101110 & 00101110 & 00101110 \\ 01011100 & 01011100 & 01011100 \\ 00101110 & 01011100 & 01110010 \\ 01011100 & 10111000 & 11100100 \\ 10111000 & 01110010 & 11001010 \end{bmatrix}$$

❒

**Example 4 (the [24, 12, 10, 16] Golay PUM Code).** The [24, 5, 12] binary linear code of Example 3 contains a [24, 2, 16] subcode, so that by Theorem 1, there exists a noncatastrophic [24, 12, 10, 16] PUM convolutional code. In this example, the construction of this code is detailed. The subspaces $A_0$, $B_0$, $A_1$, $B_1$, $A_2$, and $A_4$ defined in Example 3 are used. Additionally, subspaces $B_2'$, $A_3'$, and $B_3'$ are defined as follows:

$$B_2' = \big[g_{0,E,0}(x,\beta) + \delta_0 | g_{0,E,0}(x,\beta^2) + \delta_1 | g_{0,E,0}(x,\beta^4) + \delta_2\big]_{x \in GF(8)} \tag{39}$$

$$A_3' = [\epsilon_0 | \epsilon_1 | \epsilon_2]_{x \in GF(8)} \tag{40}$$

$$B_3' = \big[g_{0,e,0}(x,\beta) + \delta_0 | g_{0,e,0}(x,\beta^2) + \delta_1 | g_{0,e,0}(x,\beta^4) + \delta_0 + \delta_1\big]_{x \in GF(8)} \tag{41}$$

In Eq. (41), $e$ assumes only the two values 0 and 1.

The proof that the subspaces behave as depicted in Fig. 3, i.e., that $A_2 \cap B_2' = A_3'$ and $A_3' \cap B_3' = A_4$, is given in Appendix B, Lemmas B13 and B14.

Now one can see that a noncatastrophic choice for $G_0$ and $G_1$ for this code is as follows:

$$G_0 = \begin{bmatrix} \mathbf{011} \\ \mathbf{101} \\ \mathbf{100} \\ f_{001} \\ f_{00\beta} \\ f_{00\beta^2} \\ f_{100} \\ f_{\beta 00} \\ f_{\beta^2 00} \\ f_{010} \\ f_{0\beta 0} \\ f_{0\beta^2 0} \end{bmatrix} \qquad G_1 = \begin{bmatrix} \mathbf{011} \\ \mathbf{101} \\ g_{010} \\ \mathbf{100} \\ g_{0\beta 0} \\ g_{0\beta^2 0} \\ f_{001} \\ f_{00\beta} \\ f_{00\beta^2} \\ f_{100} \\ f_{\beta 00} \\ f_{\beta^2 00} \end{bmatrix}$$

In binary, these matrices are

$$G_0 = \begin{bmatrix} 00000000 & 11111111 & 11111111 \\ 11111111 & 00000000 & 11111111 \\ 11111111 & 00000000 & 00000000 \\ 01110100 & 00111010 & 01001110 \\ 00111010 & 10011100 & 10100110 \\ 10011100 & 01001110 & 11010010 \\ 00101110 & 01011100 & 01110010 \\ 01011100 & 10111000 & 11100100 \\ 10111000 & 01110010 & 11001010 \\ 11101000 & 11101000 & 11101000 \\ 01110100 & 01110100 & 01110100 \\ 00111010 & 00111010 & 00111010 \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 00000000 & 11111111 & 11111111 \\ 11111111 & 00000000 & 11111111 \\ 10010110 & 10010110 & 10010110 \\ 11111111 & 00000000 & 00000000 \\ 00101110 & 00101110 & 00101110 \\ 01011100 & 01011100 & 01011100 \\ 01110100 & 00111010 & 01001110 \\ 00111010 & 10011100 & 10100110 \\ 10011100 & 01001110 & 11010010 \\ 00101110 & 01011100 & 01110010 \\ 01011100 & 10111000 & 11100100 \\ 10111000 & 01110010 & 11001010 \end{bmatrix}$$

□

The codes in Examples 1, 3, and 4 are quite interesting as combinatorial objects, but they have potential for applications. To illustrate, Fig. 4 shows a plot of the performance of these three codes and the NASA standard [2, 1, 6, 10] (non-PUM) code on an additive white Gaussian channel. Figure 4 shows that the low-complexity Hamming [8, 4, 3, 8] code is only a bit weaker than the NASA code, while the two Golay codes are both a bit stronger. Since the state complexity of the [24, 12, 7, 12] Golay code is only 1 greater than that of the NASA code, it may be that there is a relatively low-complexity decoding algorithm for this code, whose performance will significantly exceed that of the NASA code. In any case, these performance curves certainly justify a serious study of efficient decoding algorithms for these and other PUM codes.

## Acknowledgment

The authors thank Ivan Onyszchuk of the Communications Systems Research Section for providing the curves used in Fig. 4.

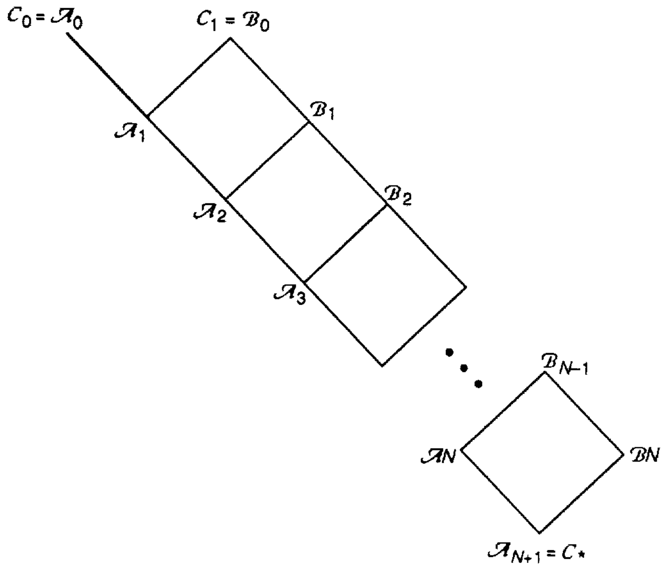


Fig. 1. The construction of Lemma 2.

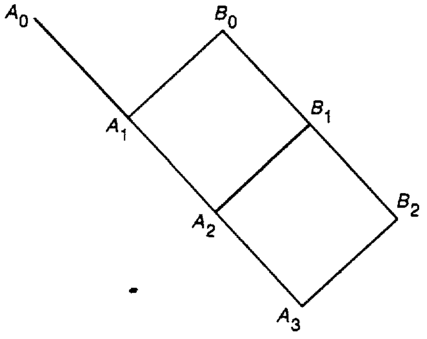


Fig. 2. The subspaces needed for the construction of a noncatastrophic encoder for the [24, 12, 7, 12] PUM code.

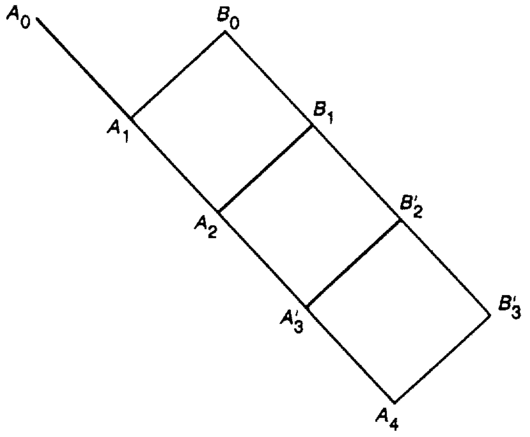


Fig. 3. The subspaces needed for the construction of a noncatastrophic encoder for the [24, 12, 10, 16] code.

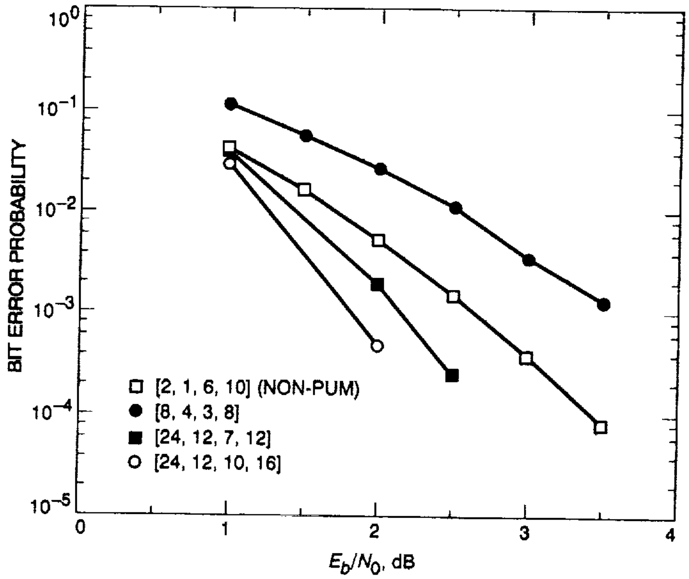


Fig. 4. Performance curves for three PUM codes, compared with the NASA standard [2, 1, 6, 10] code, on an additive white Gaussian channel.

# Appendix A
## Two Results From Linear Algebra

In this Appendix, two simple results from linear algebra are provided that are needed in the proof of Lemma 2.

**Lemma A1.** If $\langle 0\rangle = \mathcal{A}_0 \subsetneq \mathcal{A}_1 \subsetneq \cdots \subsetneq \mathcal{A}_m = \mathcal{V}$ is an ascending chain of subspaces of an $n$-dimensional vector space $\mathcal{V}$, with $\dim \mathcal{A}_i = k_i$, then there exists a basis $\langle \alpha_1, \ldots, \alpha_n\rangle$ for $\mathcal{V}$ such that

$$\langle \alpha_1, \ldots, \alpha_{k_i}\rangle = \mathcal{A}_i, \qquad \text{for } i = 1, \ldots, m \tag{A-1}$$

**Proof:** One proceeds recursively, as follows. Choose a basis $\langle \alpha_1, \ldots, \alpha_{k_1}\rangle$ for $\mathcal{A}_1$. If $m = 1$, one is done. Otherwise, by using a standard result in linear algebra [2, Lemma 4.2.5], the basis $\langle \alpha_1, \ldots, \alpha_{k_1}\rangle$ for $\mathcal{A}_1$ can be extended to a basis $\langle \alpha_1, \ldots, \alpha_{k_1}, \ldots, \alpha_{k_2}\rangle$ of $\mathcal{A}_2$, etc. ❒

**Lemma A2.** Suppose that $\mathcal{V}$ is an $n$-dimensional vector space over $F$, and $\mathcal{S}$ and $\mathcal{T}$ are subspaces of $\mathcal{V}$ with $\mathcal{S} \subset \mathcal{T} \subset \mathcal{V}$. Then there exists a subspace $\mathcal{T}' \neq \mathcal{T}$ such that $\dim \mathcal{T}' = \dim \mathcal{T}$ and $\mathcal{S} \subset \mathcal{T}' \subset \mathcal{V}$.

**Proof:** Suppose that $\dim \mathcal{S} = k$ and $\dim \mathcal{T} = k + j$, where $j > 0$. By Lemma A1, it is possible to find a basis for $\mathcal{V}$ of the form

$$\langle \mathcal{V}\rangle = \langle \alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_j, \gamma_1, \ldots, \gamma_h\rangle$$

where $k + j + h = n$, and

$$\langle \alpha_1, \ldots, \alpha_k\rangle = \mathcal{S}$$

$$\langle \alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_j\rangle = \mathcal{T}$$

If $h \geq j$, define $\mathcal{T}'$ as follows:

$$\mathcal{T}' = \langle \alpha_1, \ldots, \alpha_k, \gamma_1, \ldots, \gamma_j\rangle$$

If, on the other hand, $h < j$, define $\mathcal{T}'$ as follows:

$$\mathcal{T}' = \langle \alpha_1, \ldots, \alpha_k, \gamma_1, \ldots, \gamma_h, \beta_1, \ldots, \beta_{j-h}\rangle$$

❒

# Appendix B

## Proofs Needed in Examples 3 and 4

In this Appendix, the assertions made in Section III about the subspaces $A_0$, $A_1$, $A_2$, $A_3$, $A_3'$, $A_4$, $B_0$, $B_1$, $B_2$, $B_2'$, and $B_3'$ of $V_{24}(2)$ are proved.

**Lemma B1.** Let $A$ and $B$ be elements of $GF(8)$ such that $\mathrm{Tr}(Ax + Bx^6) = 0$ for all $x \in GF(8)$. Then $A = B = 0$.

**Proof:** Since $\mathrm{Tr}(y) = y + y^2 + y^4$ and $y^8 = y$ for all $y \in GF(8)$, it follows that

$$\mathrm{Tr}(Ax + Bx^6) = Ax + A^2x^2 + A^4x^4 + Bx^6 + B^2x^5 + B^4x^3 \tag{B-1}$$

for all $x \in GF(8)$. Thus, the equation $\mathrm{Tr}(Ax + Bx^6) = 0$ is a polynomial equation of sixth degree with 8 roots in $GF(8)$, and hence, the coefficients of the polynomial must be zero, i.e., $A = B = 0$, as asserted. □

**Lemma B2.** Let $f_{A,B,C}(x, y)$ be defined as in Eq. (26), and suppose there exists a nonzero element $y^*$ in $GF(8)$ such that

$$f_{A,B,C}(x, y^*) = 0, \qquad \text{for all } x \in GF(8) \tag{B-2}$$

Then, $A = 0$ and $B = Cy^*$. Further, if Eq. (B-2) holds and if $f_{A,B,C}(x, y)$ is not identically zero, then for any $y \neq y^*$, the number of solutions $x \in GF(8)$ to $f_{A,B,C}(x, y) = 0$ is exactly four.

**Proof:** If Eq. (B-2) holds, then by Eq. (26), one has

$$\mathrm{Tr}\big((Ay^*)x + (B + Cy^*)x^6\big) = 0, \qquad \text{for all } x \in GF(8) \tag{B-3}$$

Then, since $y^* \neq 0$, Lemma B1 implies that $A = 0$ and $B + Cy^* = 0$, i.e., $B = Cy^*$. This proves the first statement of the Lemma. To prove the second statement, assume that Eq. (B-2) holds and $f_{A,B,C}(x, y)$ is not identically zero. Then, since it is already known that $A = 0$ and $B = Cy^*$, it must be true that $C \neq 0$, so that the equation $f_{A,B,C}(x, y) = 0$ becomes

$$\mathrm{Tr}\big((Cy^* + Cy)x^6\big) = 0 \tag{B-4}$$

Since $C \neq 0$ and $y \neq y^*$, it follows that $Cy^* + Cy \neq 0$, so that Eq. (B-4) has the form $\mathrm{Tr}(Dx^6) = 0$, with $D \neq 0$. But since for $z \in GF(8)$, $\mathrm{Tr}(z) = 0$ has exactly four solutions, viz., $z = 0$, $\beta$, $\beta^2$, $\beta^4$, it follows that Eq. (B-4) has exactly four solutions. □

**Lemma B3.** Let $g_{D,E,F}(x, y)$ be defined as in Eq. (27), and suppose there exists a nonzero element $y^*$ in $GF(8)$ such that

$$g_{D,E,F}(x, y^*) = 0, \qquad \text{for all } x \in GF(8) \tag{B-5}$$

Then, $F = 0$ and $E = Dy^*$. Further, if Eq. (B-5) holds and if $g_{D,E,F}(x, y)$ is not identically zero, then for any $y \neq y^*$, the number of solutions $x \in GF(8)$ to $g_{D,E,F}(x, y) = 0$ is exactly four.

**Proof:** The proof of Lemma B3 is similar to the proof of Lemma B2 and is omitted. □

**Lemma B4.** Let $A$, $B$, $C$, $D$, $E$, and $F$ be elements of $GF(8)$ such that

$$f_{A,B,C}(x, y) = g_{D,E,F}(x, y) \tag{B-6}$$

for all $x \in GF(8)$, for two distinct values of $y$, say, $y = y_1$ and $y = y_2$. Then, $A = D$, $C = F$, and $B = E = 0$.

**Proof:** In view of the definitions in Eq. (26) and Eq. (27) of $f$ and $g$, the given conditions are equivalent to

$$\begin{aligned} &\mathrm{Tr}\big((Ay + Dy + E)x \\ &\quad + (B + Cy + Fy)x^6\big) = 0, \qquad \text{for all } x \in GF(8) \end{aligned} \tag{B-7}$$

for $y = y_1, y_2$. Thus, according to Lemma B1, $Ay + Dy + E = 0$ and $B + Cy + Fy = 0$ for $y = y_1, y_2$. The two equations $Ay_i + Dy_i + E = 0$ imply that $A = D$ and $E = 0$, and the two equations $Cy_i + Fy_i + B = 0$ imply that $C = F$ and $B = 0$. □

**Lemma B5.** The code $A_0$ defined in Eq. (28) is a $[24, 12, 8]$ code.

**Proof:** The mapping from 6-tuples $[A, B, C, \epsilon_0, \epsilon_1, \epsilon_2]$ to codewords in $A_0$ is linear. The kernel of this mapping is the set of 6-tuples such that the corresponding codeword is 0, i.e.,

$$f_{A,B,C}(x, \beta^{2^i}) + \epsilon_i = 0, \qquad \text{for } i = 0, 1, 2 \tag{B-8}$$

for all $x \in GF(8)$. By substituting $x = 0$ into these equations, one finds that $\epsilon_i = 0$ for $i = 0, 1, 2$, so that in fact

$$f_{A,B,C}(x, \beta^{2^i}) = 0, \qquad \text{for } i = 0, 1, 2 \tag{B-9}$$

for all $x \in GF(8)$. It then follows from Lemma B2 that $A = B = C = 0$. Thus, the kernel of the mapping contains only the 6-tuple $[0, 0, 0, 0, 0, 0]$, and so the mapping is one-to-one. But since the set of 6-tuples is 12-dimensional, it follows that the code is also 12-dimensional. Thus, $A_0$ is a $(24, 12)$ code. It remains to prove that its minimum distance is 8.

To show that the minimum distance is 8, first consider the $(24, 9)$ code $A_0'$ defined by Eq. (28) with $\epsilon_0 = \epsilon_1 = \epsilon_2 = 0$. Each word in $A_0'$ has three 8-bit segments, viz., the 8 bits corresponding to the function $f_{A,B,C}(x, y)$ for $y = \beta$, $\beta^2$, and $\beta^4$. Since in each segment the bit corresponding to $x = 0$ is 0, each segment may in fact be viewed as a 7-bit codeword with components indexed by the consecutive powers of a primitive root of $GF(8)$. Thus, for the "$A, B, C$" codeword in $A_0'$, the $y$-segment's $i$th component is given by $f_{A,B,C}(\beta^i, y)$, where

$$\begin{aligned} f_{A,B,C}(x, y) &= \mathrm{Tr}\big(Ayx + (B + Cy)x^6\big) \\ &= (Ay)x + (A^2y^2)x^2 + (B^4 + C^4y^4)x^3 \\ &\quad + (A^4y^4)x^4 + (B^2 + C^2y^2)x^5 + (B + Cy)x^6 \end{aligned} \tag{B-10}$$

It follows that each 7-bit segment is a codeword in the $(7, 6)$ binary cyclic code with generator polynomial $g(x) = x - 1$. In particular, each segment has even weight. The value of the weights modulo 4 can be computed by a theorem of McEliece-Solomon [9, Theorem 1] or [1, Theorem 16.33], which says that if an even-weight binary vector $a = (a_0, \ldots, a_{n-1})$ is described by its Mattson-Solomon (MS) polynomial (discrete Fourier transform) $A(x) = A_0 + \cdots + A_{n-1}x^{n-1}$, i.e., if

$$a_i = \sum_{j=0}^{n-1} A_j \beta^{-ij} \tag{B-11}$$

where $\beta$ is a primitive $n$th root of unity, and if $\Gamma_2(a) = \sum_{j=0}^{(n-1)/2} A_j A_{n-j}$, then $w(a) \equiv 2\Gamma_2(a) \pmod 4$. The MS polynomials for the 7-bit segments are given by Eq. (B-10), and so the value of $\Gamma_2$ for the $y$-segment is

$$\begin{aligned} \Gamma_2(y) &= (Ay)(B + Cy) + (A^2y^2)(B^2 + C^2y^2) \\ &\quad + (B^4 + C^4y^4)(A^4y^4) \\ &= (AB + A^4C^4)y + (A^2B^2 + AC)y^2 \\ &\quad + (A^4B^4 + A^2C^2)y^4 \\ &= \mathrm{Tr}\big((AB + A^4C^4)y\big) \end{aligned} \tag{B-12}$$

If the three segments are combined into one 21-bit word, the overall weight is still even, and the overall weight mod 4 is determined by the sum of the $\Gamma_2$s, viz.,

$$\begin{aligned} \Gamma_2 &= \Gamma_2(\beta) + \Gamma_2(\beta^2) + \Gamma_2(\beta^4) \\ &= \mathrm{Tr}\big((AB + A^4C^4)(\beta + \beta^2 + \beta^4)\big) \\ &= \mathrm{Tr}(0) = 0 \end{aligned} \tag{B-13}$$

Thus, each 7-bit segment has weight 0, 2, 4, or 6, and the overall weight is divisible by four. Furthermore, if one of the segments has weight zero, then by Lemma B2 either the other two segments are both zero, or else the other two segments have weight 4. It follows that the weights in the $(24, 9)$ code are 0, 8, 12, and 16. Now the original code $A_0$ is obtained from $A_0'$ by complementing some or all of the segments, i.e., by replacing a segment of weight $w$ with one of weight $8 - w$. Thus, in $A_0$, the segments have weight 0, 2, 4, 6, or 8. But since $8 - w \equiv w \pmod 4$, the weights in $A_0$ must also be divisible by four, and so in $A_0$ the only weights that can occur are 0, 4, 8, 12, 16, 20, and 24. The weight 4 can only occur as $0 + 0 + 4$ and $0 + 2 + 2$. Both of these cases can be eliminated by observing that since a zero-weight segment can only occur in an uncomplemented segment, and Lemma B2 says that if a codeword in $A_0$ has a zero-weight segment, then either the other two segments both have weight 8, or both have weight 4. Weight 20 is ruled out by observing that the complement of a word of weight 20 is a word of weight 4. □

Lemma B5 says that the code $A_0$ is a $[24, 12, 8]$ binary linear code. According to MacWilliams and Sloane [6, Section 20.6], such a code must be equivalent to the Golay code. The next Lemma indicates that the code $B_0$ defined in Eq. (29) is also equivalent to the Golay code.

**Lemma B6.** The code $B_0$ defined in Eq. (29) is a $[24, 12, 8]$ code.

**Proof:** The proof is virtually the same as the proof of Lemma B5. The key difference is that in place of Eq. (B-10), one has for the code $B_0$

$$\begin{aligned} g_{D,E,F}(x, y) &= \mathrm{Tr}\big((Dy + E)x + (Fy)x^6\big) \\ &= (Dy + E)x + (D^2y^2 + E^2)x^2 \\ &\quad + (F^4y^4)x^3 + (D^4y^4 + E^4)x^4 \\ &\quad + (F^2y^2)x^5 + (Fy)x^6 \end{aligned} \qquad \text{(B-14)}$$

so that in place of Eq. (B-12) one has

$$\begin{aligned} \Gamma_2(y) &= (Dy + E)(Fy) + (D^2y^2 + E^2)(F^2y^2) \\ &\quad + (F^4y^4)(D^4y^4 + E^4) \\ &= (EF + D^4F^4)y + (E^2F^2 + DF)y^2 \\ &\quad + (E^4F^4 + D^2F^2)y^4 \\ &= \mathrm{Tr}\big((EF + D^4F^4)y\big) \end{aligned} \qquad \text{(B-15)}$$

and in place of Eq. (B-13) one has

$$\begin{aligned} \Gamma_2 &= \Gamma_2(\beta) + \Gamma_2(\beta^2) + \Gamma_2(\beta^4) \\ &= \mathrm{Tr}\big((EF + D^4F^4)(\beta + \beta^2 + \beta^4)\big) \\ &= \mathrm{Tr}(0) = 0 \end{aligned} \qquad \text{(B-16)}$$

Further details are omitted here. □

**Lemma B7.** The code $A_1$ defined in Eq. (30) is a $[24, 9, 8]$ code.

**Proof:** Just as in the proof of Lemma B5, the mapping from 6-tuples $[A, 0, C, \epsilon_0, \epsilon_1, \epsilon_2]$ to codewords in $A_1$ is a linear, one-to-one mapping, which implies that $A_1$ is a $(24, 9)$ code. Since $A_1$ is a subcode of $A_0$, its minimum distance must be $\geq 8$. There are, however, many codewords in $A_1$ of weight 8, e.g., that obtained by taking $A = C = 0$, $\epsilon_0 = 1$, and $\epsilon_1 = \epsilon_2 = 0$. □

**Lemma B8.** The code $A_3$ defined in Eq. (31) is a $[24, 5, 12]$ code.

**Proof:** Using the formula Eq. (26) for $f_{A,B,C}(x, y)$, one finds that any codeword in $A_3$ can be represented as follows:

$$\begin{aligned} &\big[\mathrm{Tr}(C\beta x^6) + \epsilon_0 \,|\, \mathrm{Tr}(C\beta^2 x^6) + \epsilon_1 \,|\, \mathrm{Tr}(C\beta^4 x^6) \\ &\quad + (\epsilon_0 + \epsilon_1)\big]_{x \in GF(8)} \end{aligned} \qquad \text{(B-17)}$$

Two cases are considered: $C = 0$ and $C \neq 0$. If $C = 0$, then the codeword in Eq. (B-17) becomes $[\epsilon_0 | \epsilon_1 | \epsilon_0 + \epsilon_1]$, which is either identically zero or has weight 16. If $C \neq 0$, then since there are exactly 4 elements in $GF(8)$ with trace 0, the codeword in Eq. (B-17) has weight 12. Thus, the only weights that occur in $A_3$ are 0, 12, and 16, and so $A_3$ is a $[24, 5, 12]$ code, as asserted. □

**Lemma B9.** The code $A_4$ defined in Eq. (32) is a $[24, 2, 16]$ code.

**Proof:** According to the definition in Eq. (32), each codeword in $A_4$ has three 8-bit segments. Either all three segments are identically zero, or else one segment is zero and the other two have weight 8. Thus, in $A_4$ the only weights that occur are 0 and 16, so that $A_4$ is a $[24, 2, 16]$ code, as asserted. □

**Lemma B10.** $A_0 \cap B_0 = A_1$.

**Proof:** Note that from Eq. (26) and Eq. (27), $f_{A,0,C}(x, y) = g_{A,0,C}(x, y)$. Thus, the code $A_0 \cap B_0$ contains the code $A_1$ as defined in Eq. (30). To prove the opposite inclusion, note that by the definitions Eq. (28) and Eq. (29) of $A_0$ and $B_0$, any word in the intersection will produce an equation of the form $f_{A,B,C}(x, \beta^{2^i}) + \epsilon_i = g_{D,E,F}(x, \beta^{2^i}) + \delta_i$ for all $x \in GF(8)$, for $i = 0, 1, 2$. By substituting $x = 0$ on both sides of this equation, one gets $\epsilon_i = \delta_i$, so that in fact, $f_{A,B,C}(x, \beta^{2^i}) = g_{D,E,F}(x, \beta^{2^i})$ for all $x \in GF(8)$, for $i = 0, 1, 2$. By Lemma B4, $A = D$, $C = F$, $E = 0$, and $B = 0$, so that a word in the intersection must be of the form Eq. (30), i.e., it must lie in $A_1$. □

**Lemma B11.** $A_1 \cap B_1 = A_2$.

**Proof:** Given the definitions in Eq. (30) and Eq. (33) of $A_1$ and $B_1$, any word in the intersection $A_1 \cap B_1$ will produce an equation of the form $f_{A,0,C}(x, \beta^{2^i}) + \epsilon_i = g_{0,E,F}(x, \beta^{2^i}) + \delta_i$, for all $x \in GF(8)$ and $i = 0, 1, 2$. By substituting $x = 0$ on both sides of these equations, one gets $\epsilon_i = \delta_i$, so that in fact one has $f_{A,0,C}(x, \beta^{2^i}) = g_{0,E,F}(x, \beta^{2^i})$. By Lemma B4, this implies $A = E = 0$ and $C = F$. Thus, the intersection $A_1 \cap B_1$ is exactly the same as $A_2$, as defined in Eq. (34). □

**Lemma B12.** $A_2 \cap B_2 = A_3$.

**Proof:** Given Lemma B4 and the definitions Eq. (34) and Eq. (35) of $A_2$ and $B_2$, this result is immediate. □

**Lemma B13.** $A_2 \cap B_2' = A_3'$.

**Proof:** If a word in $A_2$, as defined in Eq. (34), is the same as a word in $B_2'$, as defined in Eq. (39), then by setting $x = 0$, one finds that $\epsilon_0 = \delta_0$, $\epsilon_1 = \delta_1$, and $\epsilon_2 = \delta_2$. Thus, also $f_{0,0,C}(x, y) = g_{0,E,0}(x, y)$ for all $x \in GF(8)$ and $y = \beta, \beta^2, \beta^4$. It then follows from Lemma B4 that $C = E = 0$, and so a word in the intersection $A_2 \cap B_2'$ must be of the form described in Eq. (40). □

**Lemma B14.** $A_3' \cap B_3' = A_4$.

**Proof:** If a word in $A_3'$, as defined in Eq. (40), is the same as a word in $B_3'$, as defined in Eq. (41), then by setting $x = 0$, one finds that $\epsilon_0 = \delta_0$, $\epsilon_1 = \delta_1$, and $\epsilon_2 = \delta_0 + \delta_1$. Thus, also $f_{0,0,0}(x, y) = g_{0,e,0}(x, y)$ for all $x \in GF(8)$ and $y = \beta, \beta^2, \beta^4$. It then follows from Lemma B4 that $e = 0$, and so a word in the intersection $A_3' \cap B_3'$ must be of the form described in Eq. (32). □

# References

[1] E. R Berlekamp, *Algebraic Coding Theory*, rev. ed., Laguna Hills, California: Aegean Park Press, 1984.

[2] I. N. Herstein, *Topics in Algebra*, 2nd ed., New York: Wiley and Sons, 1975.

[3] J. Justesen, E. Paaske, and M. Ballan, "Quasi-Cyclic Unit Memory Convolutional Codes," *IEEE Trans. Inform. Theory*, vol. IT-36, no. 3, pp. 540–547, May 1990.

[4] G. S. Lauer, "Some Optimal Partial-Unit-Memory Codes," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 2, pp. 240–243, March 1979.

[5] L.-N. Lee, "Short Unit Memory Byte-Oriented Convolutional Codes Having Maximal Free Distance," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 3, pp. 349–352, May 1976.

[6] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, Amsterdam: North-Holland, 1977.

[7] F. Pollara, R. McEliece, and K. Abdel-Ghaffar, "Finite-State Codes," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 5, pp. 1083–1088, September 1988.

[8] F. Pollara, K.-M. Cheung, and R. J. McEliece, "Further Results on Finite-State Codes," *TDA Progress Report 42-92*, vol. October–December 1987, pp. 56–62, February 15, 1988.

[9] G. Solomon and R. McEliece, "Weights of Cyclic Codes," *J. Combinatorial Theory*, vol. 1, no. 4, pp. 459–475, December 1966.

[10] C. Thommesen and J. Justesen, "Bounds on Distances and Error Exponents of Unit Memory Codes," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 637–649, September 1983.

S7-32

63210
P. 8

# A Portable Ku-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation

T. Y. Otoshi, S. R. Stewart, and M. M. Franco
Ground Antennas and Facilities Engineering Section

*This article presents the design of a Ku-band test package for the new beam-waveguide (BWG) antenna at 11.7–12.2 GHz. Results of linear polarization measurements with the test package on the ground are also presented. This report is the fifth in a series of articles concerned with test-package design and performance.*

## I. Introduction

A new 34-m beam-waveguide (BWG) antenna has been built at Deep Space Station 13 (DSS 13) in the Goldstone Deep Space Communications Complex near Barstow, California. This antenna is designed to be efficient at X-, Ku-, and Ka-bands, and it is the first NASA tracking antenna to use a BWG design. The antenna's focal points for the center-pass mode are shown in Fig. 1.

The unique methodology used to test the new BWG antenna included making measurements at the Cassegrain focal point and then at the final focal point in the pedestal room. Measurements made at the Cassegrain focal point F1 give information concerning panel settings, subreflector defocus, and far-field antenna patterns. Measurements made at the pedestal room focal point F3 provide amplitude and phase information on the entire BWG system, which includes the main reflector, subreflector, and six BWG mirrors. Degradations caused by the BWG mirror systems are determined by comparing the measured parameters at the two focal points. Previous articles [1–4] describe the successful employment of X- and Ka-band test packages for the new DSS 13 BWG antenna at 8.45 and 32 GHz.

This is the fifth article in a series of reports on the design and performance of X-, Ka-, and Ku-band test packages that were developed specifically for testing the BWG antenna. The Ku-band test package is used primarily for making holographic measurements [5]. In this article, only a description of the Ku-band test package and results from a linear polarization test will be given. In another report, it was shown that the Ku-band test package was used successfully at the BWG antenna Cassegrain focal point F1 to obtain holographic panel-setting information.[1] The Ku-band test package will be employed in the near future for making holographic measurements at F3 and also for making BWG frequency-stability measurements, as described in a companion article [6].

[1] B. L. Seidel and D. J. Rochblatt, Chapter 4, *DSS-13 Beam-Waveguide Antenna Project, Phase 1 Final Report,* JPL D-8451 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 15, 1991.

## II. Ku-Band Test-Package Design

Figure 2 shows the system block diagram of the Ku-band test package. Depicted are such Cassegrain front-end microwave components as the horn, cosine taper, circular waveguide rotary joint, the circular waveguide to waveguide-rectangular (WR) 75 transition, and the low-noise amplifier (LNA). The microwave system is designed to operate over a Ku-band frequency range of 11.7–12.2 GHz. For holographic measurements, an Eikontech Holographic Receiver system was used to downconvert the Ku-band frequency to an intermediate frequency (IF) of approximately 482 MHz.

The 22-dBi horn assembly has an aperture diameter of 5.004 in. and tapers linearly over a length of 18.414 in. to a diameter of 0.968 in. A 4-in.-long cosine taper provides a gradual transition from the horn output diameter of 0.968 in. to the circular rotary joint diameter of 0.879 in. The Ku-band circular waveguide rotary joint is a scale model of the DSN version used in the X-band test package [1] and other Cassegrain cone assemblies. To enable linear polarization to be varied to the desired rotation angle, the LNA is supported by a pair of orthogonal brackets with an axle-hole design that permits the amplifier (and follow-up coaxial cable) to be manually rotated 360 deg. A clamp is used to prevent further rotation of the assembly once the optimum linear polarization-output orientation is found for the incoming signal.

Testing the antenna at F1 and F3 required the test package to be convertible from 29- to 22-dBi horn configurations. The conversion is accomplished by using horn extensions of the same linear taper going from an aperture diameter of ~13.46 to 5.004 in., over a length of 38.58 in. Figure 3 shows the Ku-band test package in the 29-dBi horn configuration installed at the Cassegrain focal point F1. Figure 4 shows the test package in the 22-dBi horn configuration ready for installation on the mounting table at focal point F3. The height of the 29-dBi horn configuration, as measured from the horn aperture to the base of the frame assembly, is about 11 ft, while the height of the 22-dBi horn test-package configuration is about 39 in. shorter.

## III. Test Results

Figure 5 is a photograph of the Ku-band test package undergoing linear polarization tests at JPL. A block diagram of the test setup appears in Fig. 6. The transmit horn was pyramidal (Scientific-Atlanta model 12-7.0), with aperture dimensions of 5.04 × 3.73 in. and WR 112 at its input. The separation distance of 44.2 in. corresponded to 1.72 $D^2/\lambda$, where $\lambda$ is the free-space wavelength for the test frequency of 11.95 GHz, and $D$ corresponds to the 5.04-in. dimension of the pyramidal horn. Linear polarization tests were performed by rotating the lower half of the test-package circular rotary joint and measuring received signal levels with the equipment shown in Fig. 6.

The results of the polarization tests are given in Fig. 7. Good agreement was obtained between theoretical and experimental data. The measured depths of the nulls were ~−47 dB down from the peak when the two horns were cross-polarized with respect to each other. The test results verified that the Ku-band test-package system would receive linearly polarized signals at desired rotation angles and reject cross-polarized signals.

The Y-factor noise-temperature measurements made with aperture-ambient and liquid-nitrogen loads showed that the Ku-band operating noise temperature was 172 K, as defined at the Ku-band horn aperture. Most of this operating noise temperature was due to the 160 K LNA in the system (see Fig. 2).

## IV. Concluding Remarks

The Ku-band test package was tested in an on-the-ground configuration and was found to perform according to the design goals. Other reports showed that the test package was stable and performed well when used at the BWG Cassegrain focal point F1 for holographic measurements at 46.5- , 37.0- , and 12.7-deg. elevation angles.[2]

[2] Ibid.

## Acknowledgment

The authors gratefully acknowledge H. Reilly who provided the Ku-band rotary joint for the test package.

## References

[1] T. Y. Otoshi, S. R. Stewart, and M. M. Franco, "A Portable X-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation—Part I: Design and Ground Tests," *TDA Progress Report 42-103*, vol. July–September 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 135–150, November 15, 1990.

[2] T. Y. Otoshi, S. R. Stewart, and M. M. Franco, "A Portable X-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation—Part II: Tests on the Antenna," *TDA Progress Report 42-105*, vol. January–March 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 54–68, May 15, 1991.

[3] T. Y. Otoshi, S. R. Stewart, and M. M. Franco, "A Portable Ka-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation—Part I: Design and Ground Tests," *TDA Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 249–265, August 15, 1991.

[4] T. Y. Otoshi, S. R. Stewart, and M. M. Franco, "A Portable Ka-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation—Part II: Tests on the Antenna," *TDA Progress Report 42-106*, vol. April–June 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 266–282, August 15, 1991.

[5] D. J. Rochblatt and Y. Rahmat-Sammi, "Effects of Measurement Errors on Microwave Antenna Holography," *IEEE Transactions on Antennas and Propagation*, vol. 39, no. 7, pp. 933–942, July 1991.

[6] T. Y. Otoshi, "A Proposed Far-field Method for Frequency Stability Measurements on the DSS 13 Beam-Waveguide Antenna," *TDA Progress Report 42-107*, vol. July–September 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 81–87, November 15, 1991.

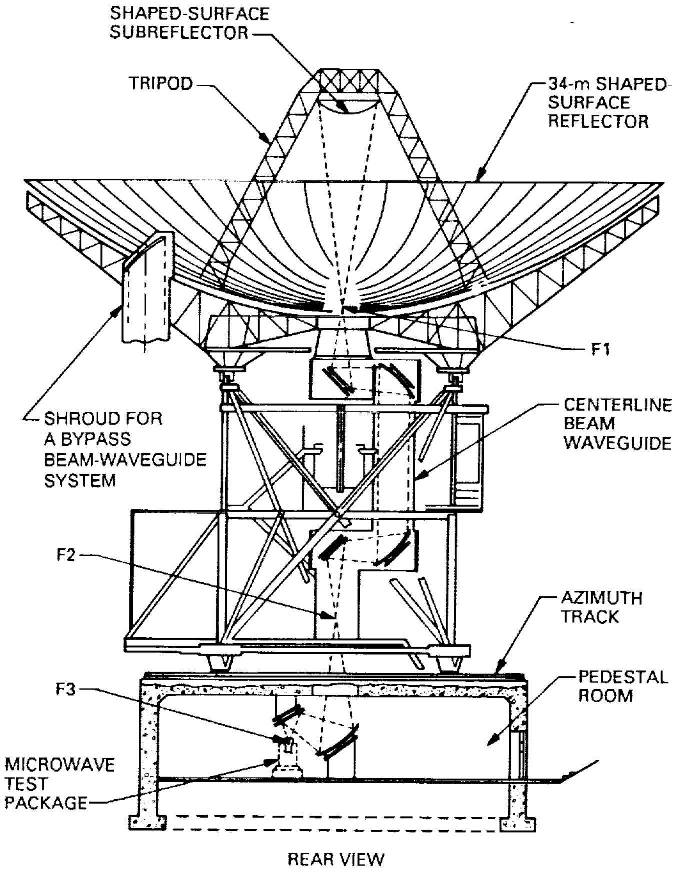

SHAPED-SURFACE SUBREFLECTOR
TRIPOD
34-m SHAPED-SURFACE REFLECTOR
F1
SHROUD FOR A BYPASS BEAM-WAVEGUIDE SYSTEM
CENTERLINE BEAM WAVEGUIDE
F2
AZIMUTH TRACK
PEDESTAL ROOM
F3
MICROWAVE TEST PACKAGE
REAR VIEW

**Fig. 1. The BWG antenna in the center-pass mode, showing focal points F1, F2, and F3 for testing with the portable test package.**

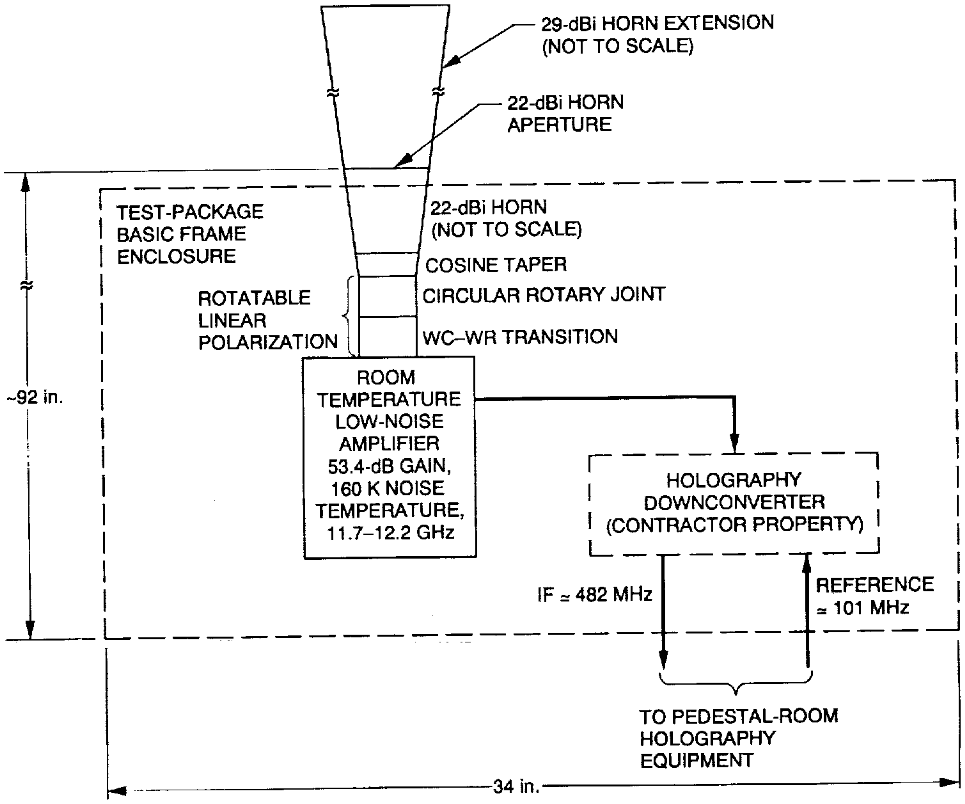

29-dBi HORN EXTENSION (NOT TO SCALE)
22-dBi HORN APERTURE
TEST-PACKAGE BASIC FRAME ENCLOSURE
22-dBi HORN (NOT TO SCALE)
COSINE TAPER
ROTATABLE LINEAR POLARIZATION
CIRCULAR ROTARY JOINT
WC–WR TRANSITION
ROOM TEMPERATURE LOW-NOISE AMPLIFIER 53.4-dB GAIN, 160 K NOISE TEMPERATURE, 11.7–12.2 GHz
~92 in.
HOLOGRAPHY DOWNCONVERTER (CONTRACTOR PROPERTY)
IF ≃ 482 MHz
REFERENCE ≃ 101 MHz
TO PEDESTAL-ROOM HOLOGRAPHY EQUIPMENT
34 in.

**Fig. 2. A block diagram of the Ku-band test-package system.**

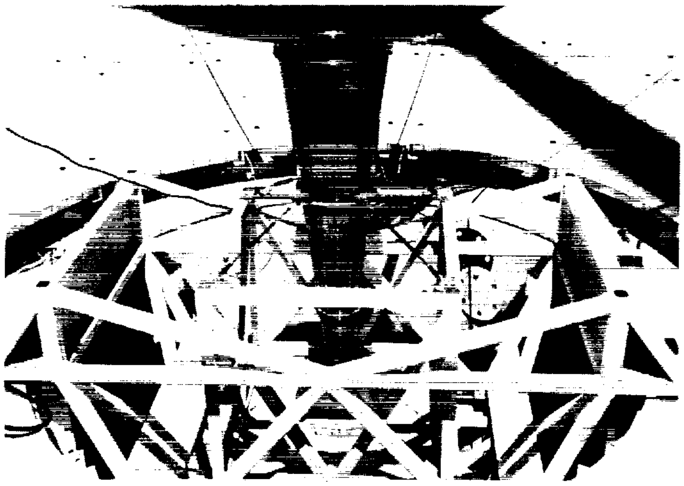


Fig. 3. The Ku-band test package mounted at F1 in the 29-dBi horn configuration.

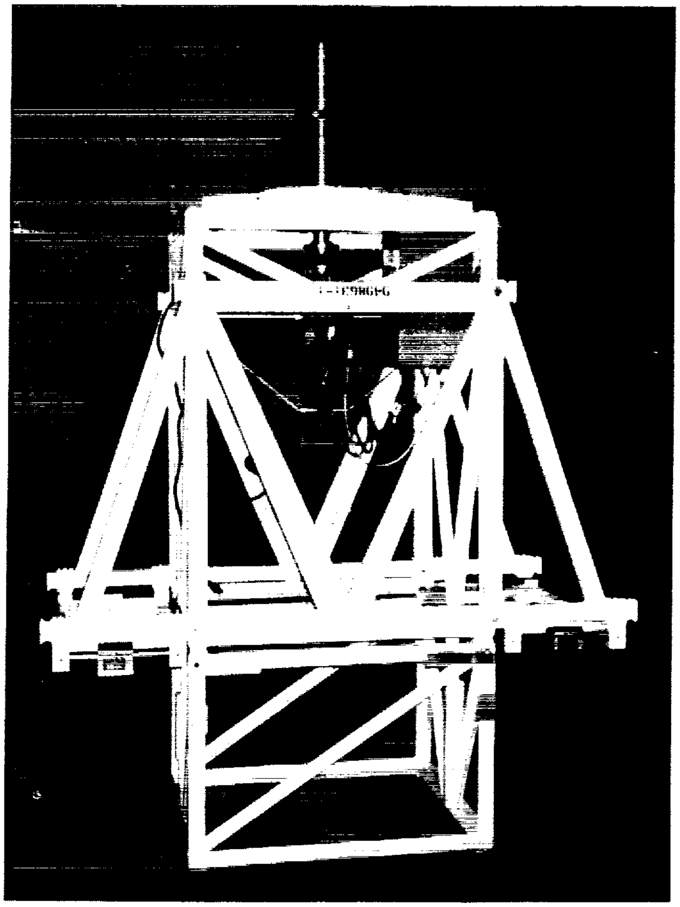


Fig. 4. The Ku-band test package assembly in the 22-dBi horn configuration prior to installation on the mounting table at F3.

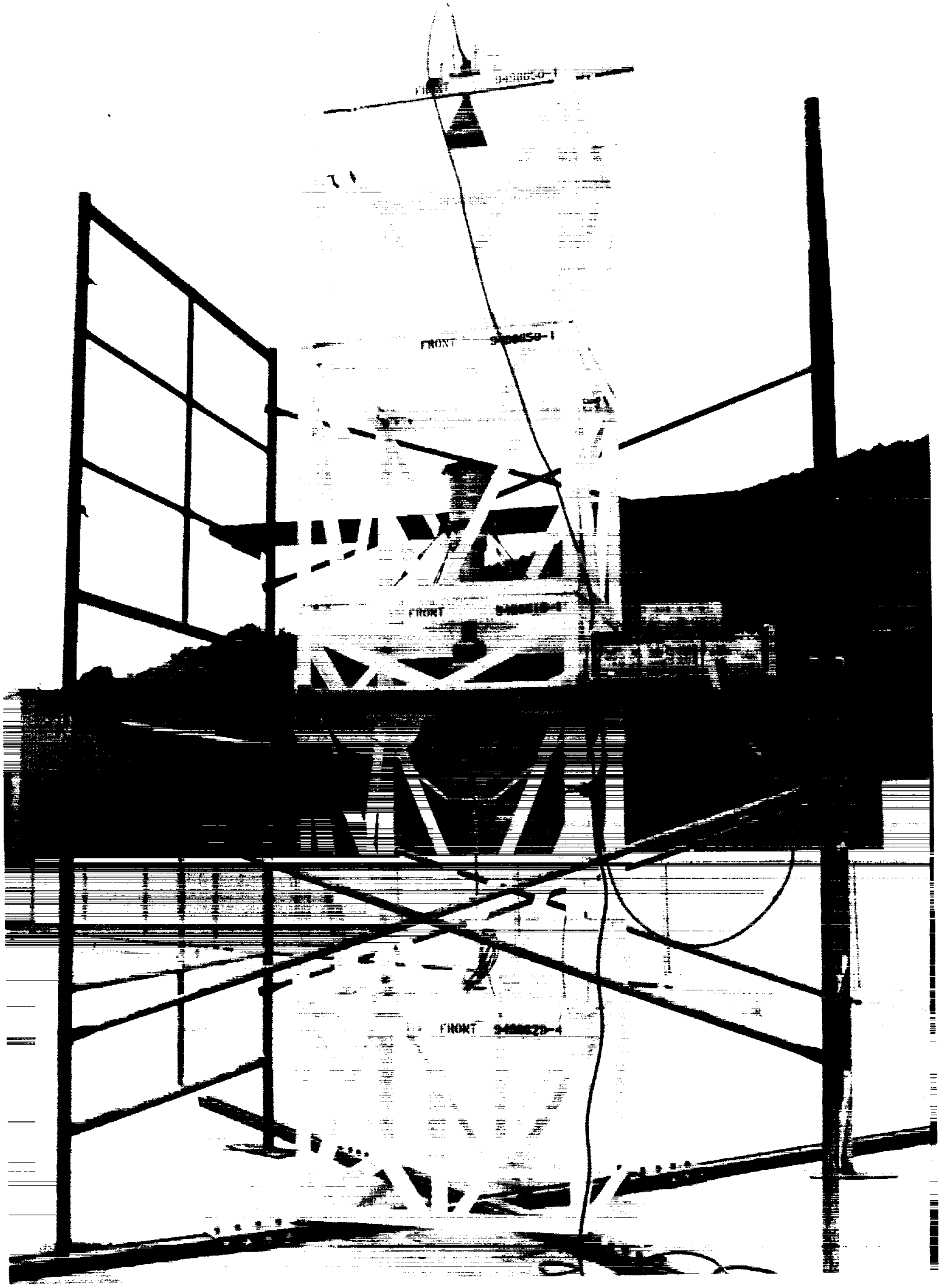


Fig. 5. The test setup for linear polarization tests on the Ku-band test package at JPL.

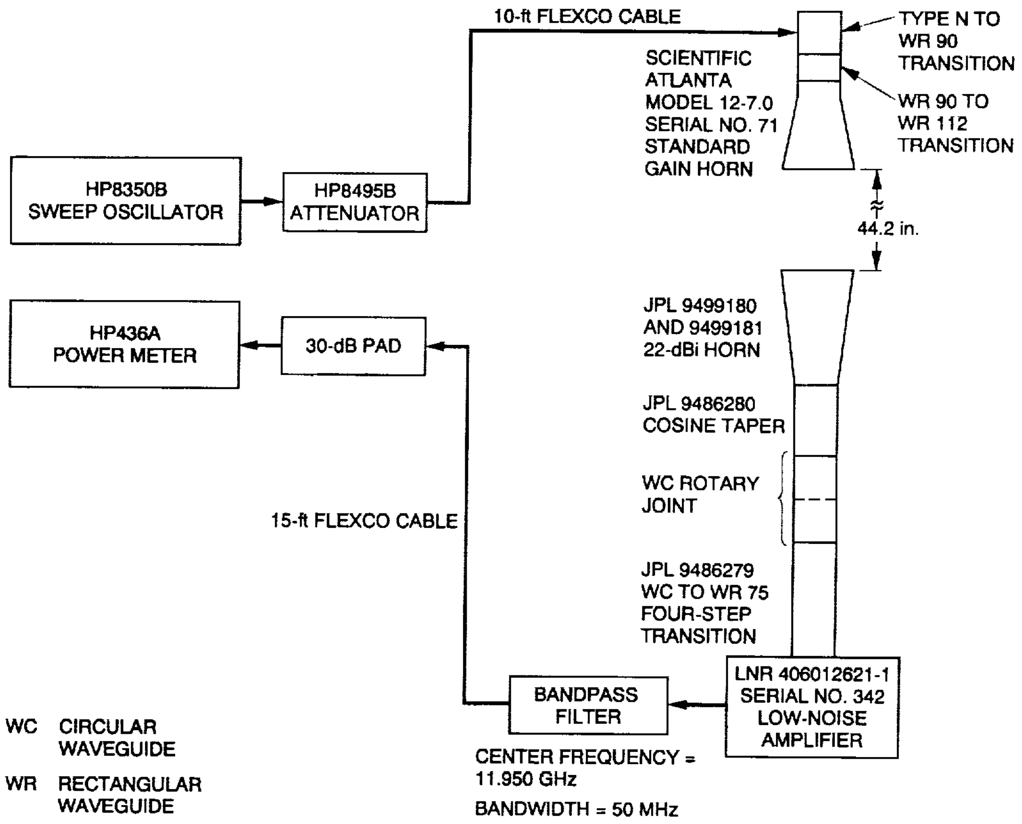

10-ft FLEXCO CABLE
TYPE N TO WR 90 TRANSITION
SCIENTIFIC ATLANTA MODEL 12-7.0 SERIAL NO. 71 STANDARD GAIN HORN
WR 90 TO WR 112 TRANSITION
HP8350B SWEEP OSCILLATOR
HP8495B ATTENUATOR
44.2 in.
HP436A POWER METER
30-dB PAD
JPL 9499180 AND 9499181 22-dBi HORN
JPL 9486280 COSINE TAPER
WC ROTARY JOINT
15-ft FLEXCO CABLE
JPL 9486279 WC TO WR 75 FOUR-STEP TRANSITION
BANDPASS FILTER
LNR 406012621-1 SERIAL NO. 342 LOW-NOISE AMPLIFIER
WC CIRCULAR WAVEGUIDE
WR RECTANGULAR WAVEGUIDE
CENTER FREQUENCY = 11.950 GHz
BANDWIDTH = 50 MHz

Fig. 6. A block diagram of the linear polarization test setup at 11.950 GHz.

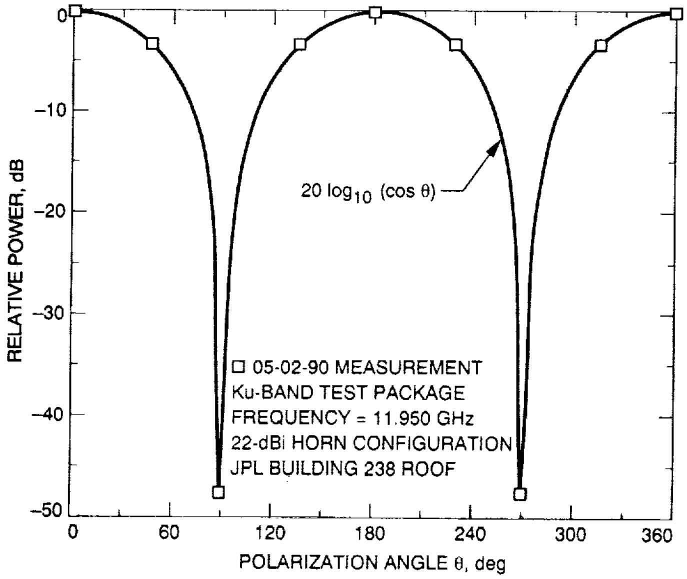


Fig. 7. Theoretical and experimental polarization loss versus polarization angle.

S8-32
532 //
P- 9

# A Proposed Far-Field Method for Frequency-Stability Measurements on the DSS 13 Beam-Waveguide Antenna

T. Y. Otoshi
Ground Antennas and Facilities Engineering Section

*A method is presented for measuring the frequency stability of the new beam-waveguide (BWG) antenna at DSS 13. This method is relatively inexpensive and primarily utilizes equipment that is already available. Another desirable feature of the method is that a far-field signal will be used for the measurement. In concert with the goal of employing new technology developments, a fiber-optic system will be used at 12 GHz to carry a reference antenna signal to the BWG antenna Ku-band test-package location in the pedestal room.*

## I. Introduction

This article presents a method for measuring the frequency stability of the antenna optics portion of the new DSS 13 beam-waveguide (BWG) system [1]. The antenna optics portion, hereafter referred to as the antenna, is defined as the portion of the antenna system that includes the main reflector, subreflector, the six mirrors, and the feedhorn assembly located at the final focal point in the pedestal room. The ultimate goal is to develop an accurate far-field method for measuring the frequency stability, group-delay stability, and phase-delay stability of the antenna at X- and Ka-band.

Frequency stability [2] is closely related to group-delay and phase-delay stability in that all three data types are derived from measurements or calculations of small phase perturbations. In the case of frequency stability, measured phase deviations over specified time intervals are converted to fractional frequency deviations of the carrier frequency. In the case of group-delay stability, changes of phase deviations from linear phase characteristics are measured over a specified spanned bandwidth as functions of time. Similarly, in the case of phase-delay stability, the measured changes of phase as functions of time are converted to equivalent pathlength changes for the carrier microwave frequency. Scientists who perform gravity-wave experiments are interested in measurements of frequency stability, while those who perform ranging and very-long-baseline interferometry (VLBI) experiments are interested in group- and phase-delay stability, respectively.

Previously, measurements of the frequency stability of Deep Space Network (DSN) tracking systems did not include the antenna. The methods employed in the past to measure frequency, group-delay, and phase-delay stability of an antenna are discussed in the Appendix. The disadvantages of those methods are also discussed.

The new far-field method discussed here takes advantage of recent advances made in fiber-optic technology. Most of the equipment required for making frequency-stability measurements with this new method is readily available.

## II. Methodology

Figure 1 shows a block diagram of the proposed experimental setup, which utilizes far-field transponder and beacon signals available from geostationary satellites at elevation angles between 12 and 47 deg. The method will utilize the already existing Ku-band microwave feedhorn and test package [3] to receive the satellite signals at various focal points of the BWG system at DSS 13. In addition, a phase-detector Allan-deviation measurement instrument [4] and a 10-ft reference antenna, which are already available, will be utilized for this project.

The only new development required is a Ku-band fiber-optic cable system (including modulators, isolators, and amplifiers) that is capable of transmitting the reference antenna signal (12 GHz) over the 50-m distance to the pedestal room. Tests by G. F. Lutes and R. T. Logan [5] of fiber-optic systems at 8.4 GHz demonstrated frequency stability of about $7 \times 10^{-16}$ for sampling times of 10 sec, which is approximately two orders of magnitude better than the frequency stability of typical hydrogen masers.

One advantage of the proposed method is that it does not require reference signals that are coherent with the station clock. Another advantage is that through the use of a far-field signal propagating through both the test antenna and a reference antenna, phase changes and atmospheric and ionospheric variations tend to cancel out when differenced in the output. Available equipment and state-of-the art technologies are utilized. Furthermore, the experimental setup utilizes fiber-optic technology, which seems to be the way of the future, and is in concert with other DSN–JPL goals.

The initial system will enable measurements of the stability of the BWG system at fixed elevation angles as functions of time and weather conditions (wind, ambient temperatures). Even though the measurements will be made at Ku-band, the information can be extrapolated to other frequencies, provided that the variations are related mostly to mechanical structural deformation. This is just the first step toward getting an accurate measurement of frequency stability of the antenna optics portion of a large antenna.

A variation and possible improvement of the proposed method is to move the reference antenna to the front of the subreflector. Then the possibility exists for using a spacecraft signal or radio sources for phase- and group-delay calibrations at various elevation and azimuth angles.

Early in 1995, X- and Ku-band far-field signals from Earth-orbiting spacecraft will become available for a Space VLBI Project.[1] These same signals can be used to calibrate the stability of the BWG antenna at various elevation angles with the configuration described above. For this Space VLBI Project, a Soviet spacecraft named Radioastron will be launched and will transmit X- and Ku-band downlink frequencies of 8.473 GHz and 15.06 GHz, respectively. A Japanese spacecraft named VSOP will be launched and will transmit a Ku-band downlink frequency of 14.2 GHz. The worldwide VLBI experiments to be performed with the Soviet and Japanese spacecraft represent an international group effort whose participants include the United States, Japan, the Soviet Union, Western Europe, Canada, and Australia.

## III. Concluding Remarks

There is a strong need to develop a far-field frequency-stability measurement technique for measuring antenna effects that does not tie up an additional DSN tracking antenna (fully implemented with phase calibrators, low-noise amplifier, VLBI processors, etc.), as is necessary for the VLBI or the connected-element interferometry (CEI) techniques. The processing from the system proposed in this article is simpler, less expensive, and enables useful data to be obtained in a shorter time frame.

It is suggested that the proposed work be the pioneering effort to uncover major problems in a short time frame, while the VLBI and CEI efforts are carried out in parallel for the long-term solution. Technology developed from this work can be transferred to such other projects as holography and the gravity-wave experiments. An opportunity exists to demonstrate that fiber-optic cables can be used to carry microwave frequencies (X- or Ku-bands) over long pathlengths with only negligible degradation of the signal (both phase and amplitude). The spin-offs from such a joint effort will potentially benefit other projects in the future.

[1] *Orbiting VLBI Subnet C/D Review*, JPL D-8361 (internal document), Jet Propulsion Laboratory, Pasadena, California, April 3, 1991, and *Design Requirements: DSN Orbiting VLBI Subnet*, JPL DM515606A (internal document), Jet Propulsion Laboratory, Pasadena, California, May 16, 1991.

## Acknowledgment

The use of fiber optics for the experimental setup was suggested by G. F. Lutes of the Frequency and Timing Group in the Communications Systems Research Section.

C-2

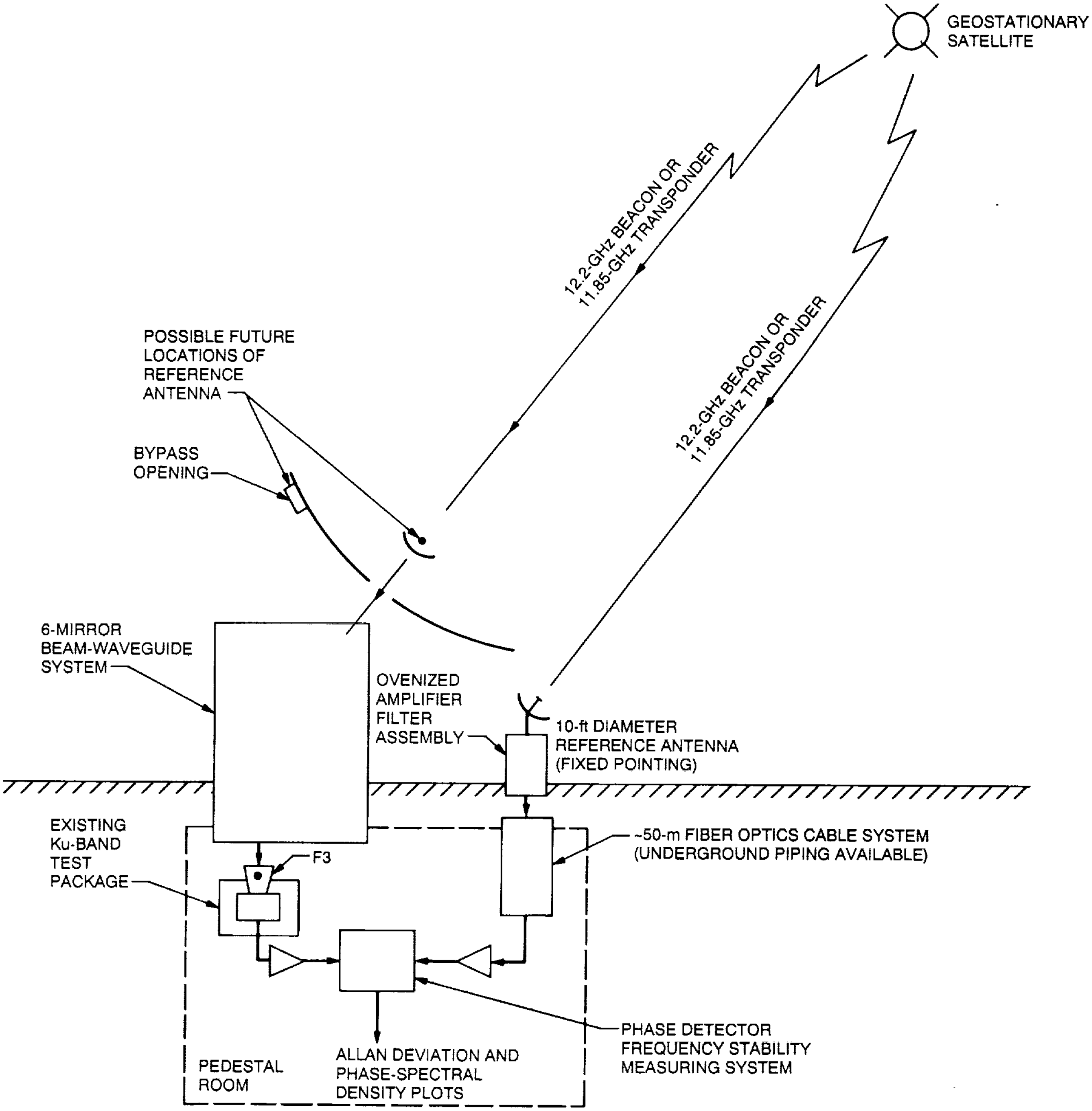

GEOSTATIONARY SATELLITE
12.2-GHz BEACON OR 11.85-GHz TRANSPONDER
12.2-GHz BEACON OR 11.85-GHz TRANSPONDER
POSSIBLE FUTURE LOCATIONS OF REFERENCE ANTENNA
BYPASS OPENING
6-MIRROR BEAM-WAVEGUIDE SYSTEM
OVENIZED AMPLIFIER FILTER ASSEMBLY
10-ft DIAMETER REFERENCE ANTENNA (FIXED POINTING)
EXISTING Ku-BAND TEST PACKAGE
F3
~50-m FIBER OPTICS CABLE SYSTEM (UNDERGROUND PIPING AVAILABLE)
PHASE DETECTOR FREQUENCY STABILITY MEASURING SYSTEM
PEDESTAL ROOM
ALLAN DEVIATION AND PHASE-SPECTRAL DENSITY PLOTS

**Fig. 1. Proposed method for measurement of frequency stability of beam-waveguide portion of DSS 13 BWG antenna.**

# Appendix

# Previous Methods Employed To Measure Frequency, Group-Delay, and Phase-Delay Stability of Antenna Systems

## I. Directional Coupler Below Feed Horn

The frequency stability of a transmit–receive system wäs successfully measured [6,7], but the measurements were restricted to the portion of the system below a direction coupler located just below the feedhorn. The coupler method does not test the portion of the system that includes the feedhorn, subreflector, and main reflector.

## II. Dish-Mounted Probes

In the past, a ranging system and zero-delay dish-mounted probes were used to measure multipath-caused errors on the group delay of a large antenna [8,9]. Similar dish-mounted probe tests were also performed using VLBI methods and network analyzer techniques.[1,2]

Dish-mounted probes are highly susceptible to multipath errors for both ranging and VLBI. Although time-domain techniques [10] have proved to be successful for separating multipath errors, the dish-mounted probe technique can only yield knowledge of antenna RF stability at a single point on the main reflector surface and not the entire surface.

## III. Collimation-Tower Technique

A VLBI method was used to measure the group- and phase-delay stability of the DSS 13 26-m antenna by using a VLBI noise source at the DSS 13 collimation tower and a 6-ft reference antenna mounted at the top edge of the main reflector [11]. A real-time VLBI correlator used to process the data showed good repeatability for a short duration. The disadvantage of the collimation-tower technique is that the collimation tower is not in the far-field and, in fact, is usually at only about 1/10 the required far-field distance. Furthermore, because the collimation tower source is usually located at an elevation angle of ~ 5 deg, this method is very susceptible to errors caused by multipath signals bouncing from the ground into the reference and test antennas. In addition, the multipath signals received by the test and reference antennas do not remain constant over long periods of time.

## IV. Far-Field Methods

The Viking and Voyager spacecraft were used as far-field illuminators to determine the group-delay changes due to subreflector defocusing [12,13].[3] The DSN ranging system was used, and group-delay measurements using the ranging system have a precision of ~ 0.5 nsec, which is inadequate for making long-term group-delay stability measurements. Charged particle changes on the uplink and downlink signals are difficult to take into account.

A VLBI technique was used to investigate the effects of antenna structural deformation on measured VLBI group delays of the DSS 14 64-m antenna when using the DSS 13 26-m antenna as the reference antenna [14]. The precision of these measurements is estimated to be about ±1 cm for a spanned bandwidth of 40 MHz after removing the effects of unwanted multipath reflections between the subreflector and the Mod-3 cone-support platform.

Resch recently proposed using VLBI and CEI techniques (needing phase calibrators and real-time correlators and another antenna, such as the DSS 14 70-m antenna) to test the stability of the DSS 13 BWG antenna.[4] The main disadvantage of this method is that it is difficult to separate out the frequency instabilities of the antenna under test when the stability of the reference antenna (DSS 14) is unknown. Phase calibrators have presented problems in the past for VLBI measurements and may require more development to improve their reliability.

[1] T. Y. Otoshi, "RTOP 61, Microwave Phase Calibration Work Unit Accomplishments FY 80," *DSN Advanced Systems Review* (internal document), Jet Propulsion Laboratory, Pasadena, California, June 1980.

[2] T. Y. Otoshi, unpublished network analyzer group-delay data on dish-mounted horn at DSS 13, April 1977.

[3] In [13], subreflector experiments were performed by T. Otoshi.

[4] G. Resch, "Radiometric Testing of the DSS 13 Beam Waveguide Antenna," IOM 335.0-90-38 (internal document), Jet Propulsion Laboratory, Pasadena, California, September 10, 1990.

# References

[1] T. Y. Otoshi, S. R. Stewart, and M. M. Franco, "A Portable X-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation—Part I: Design and Ground Tests," *TDA Progress Report 42-103*, vol. July–September 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 135–150, November 15, 1990.

[2] C. G. Greenhall, "Frequency Stability Review," *TDA Progress Report 42-88*, vol. October–December 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 200–212, February 15, 1987.

[3] T. Y. Otoshi, S. R. Stewart, and M. M. Franco, "A Portable Ku-Band Front-End Test Package for Beam-Waveguide Antenna Performance Evaluation," *TDA Progress Report 42-107*, vol. July–September 1991, Jet Propulsion Laboratory, Pasadena, California, pp.73–80, November 15, 1991.

[4] B. L. Conroy and D. Le, "Measurement of Allan Variance and Phase Noise at Fractions of a Millihertz," *Rev. Sci Instruments*, vol. 61, no. 6, pp. 1720–1723, June 1990.

[5] G. F. Lutes and R. T. Logan, "Status of Frequency and Timing Reference Signal Transmission by Fiber Optics," *Proceedings of the 45th Annual IEEE Symposium on Frequency Control*, to be published 1991.

[6] T. Y. Otoshi and M. M. Franco, "DSS 13 Frequency Stability Tests Performed During May 1985 Through March 1986," *TDA Progress Report 42-86*, vol. April–June 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 1–15, August 15, 1986.

[7] T. Y. Otoshi and M. M. Franco, "DSS 13 Frequency Stability Tests," *TDA Progress Report 42-89*, vol. January–March 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 1–20, May 15, 1987.

[8] C. T. Stelzried, T. Y. Otoshi, and P. D. Batelaan, "S/X Band Experiment: Zero Delay Device Antenna Location," *DSN Progress Report 42-20*, vol. January and February 1974, Jet Propulsion Laboratory, Pasadena, California, pp. 64–68, April 15, 1974.

[9] T. Y. Otoshi, "S-Band Zero–Delay Device Multipath Tests on the 64-m Antenna at DSS 43, DSS 63, and DSS 14," *DSN Progress Report 42-29*, vol. July and August 1975, Jet Propulsion Laboratory, Pasadena, California, pp. 20–32, October 15, 1975.

[10] T. Y. Otoshi, "An FM/CW Method for the Measurement of Time Delays of Large Cassegrain Antennas," *The Telecommunications and Data Acquisition Progress Report 42-66*, vol. September and October 1981, Jet Propulsion Laboratory, Pasadena, California, pp. 49–59, December 15, 1981.

[11] T. Y. Otoshi, L. E. Young, and W. V. T. Rusch, "VLBI Collimation Tower Technique for Time-Delay Studies of a Large Ground Station Communications Antenna," *IEEE Trans. on Antennas and Propagation*, vol. AP-33, no. 5, pp. 549–556, May 1985.

[12] T. Y. Otoshi and D. L. Brunn, "Multipath Tests on 64-m Antennas Using the Viking Orbiter-1 and -2 Spacecraft as Far-Field Illuminators," *DSN Progress Report 42-31*, vol. November and December 1975, Jet Propulsion Laboratory, Pasadena, California, pp. 41–49, February 15, 1976.

[13] D. W. Green, "Validation of Roundtrip Charged-Particle Calibration Derived from S- and X-Band Doppler Via DRVID Measurements," *DSN Progress Report 42-55*, vol. November and December 1979, Jet Propulsion Laboratory, Pasadena, California, pp. 30–40, February 15, 1980.

[14] T. Y. Otoshi and L. E. Young, "An Experimental Investigation of the Changes of VLBI Time Delays Due to Antenna Structural Deformations," *TDA Progress Report 42-68*, vol. January and February 1981, Jet Propulsion Laboratory, Pasadena, California, pp. 8–16, April 15, 1982.

-7-32

53212 P 8 ✓

should be Deep Space Station (DSS)

abstract correction needed

# The L-/C-Band Feed Design for the DSS 14 70-Meter Antenna (Phobos Mission)

P. H. Stanton and H. F. Reilly, Jr.
Ground Antennas and Facilities Engineering Section

*A dual-frequency (1.668 and 5.01 GHz) feed was designed for the DSS 14 70-m antenna to support the Soviet Phobos Mission. This antenna system was capable of supporting telemetry, two-way Doppler, and very long baseline interferometry (VLBI). VLBI and two-way Doppler information on the Phobos spacecraft was acquired with this antenna in 1989.*

## I. Introduction

Two Soviet spacecraft were launched in 1988 to observe Phobos and Mars. In cooperation with this Soviet project, the United States agreed to provide tracking and telecommunications from the DSS 14 70-m antenna at Goldstone. A dual-frequency (1.668 and 5.01 GHz) L-/C-band feedhorn was designed and installed on the antenna to support this project. This antenna system was capable of handling telemetry and two-way Doppler and was part of a very long baseline interferometry (VLBI) system. VLBI and two-way Doppler information on the Phobos 2 spacecraft was acquired with the antenna early in 1989. Unfortunately, both the Phobos 1 and the Phobos 2 spacecraft ceased functioning properly before the mission was completed.

To handle a simultaneous downlink frequency of 1.67 GHz and an uplink frequency of 5.01 GHz, a dual-frequency feed was designed which was composed of an L-band dual-mode horn [1] enclosing a thin, coaxially mounted, C-band surface-wave antenna [2]. A photograph of this dual-frequency feed is shown in Fig. 1.

## II. Dual-Frequency Feed Design

A dual-frequency feed, composed of a horn enclosing a coaxially mounted surface-wave antenna, is described in [3]. This type of feed has the ability to efficiently illuminate a Cassegrain antenna at two frequencies simultaneously. The design of this feed is facilitated by the relatively low radio frequency (RF) interaction obtainable between the horn and the surface-wave antenna. A frequency spread of 3:1 for the Phobos mission's feed allows nearly independent adjustment of the beamwidths and phase centers. Beyond the general requirements of simultaneous L-/C-band operation and low impact on the existing 70-m system, Table 1 lists the primary RF requirements of the 70-m antenna with this dual-frequency feed. A drawing of the L-/C-band feedhorn is presented in Fig. 2.

## A. Surface-Wave C-Band Feed

In the design of a C-band feed for the 70-m antenna at Goldstone, the following needs were addressed:

(1) A common L-/C-band phase center for simultaneous operation.

(2) A suitable RF radiation pattern; e.g., beamwidth (10 dB) ~ 26 deg, approximate Gaussian shape, and low cross-polarized radiation.

(3) Low interference with the L-band horn's radiation pattern.

(4) Ability to accommodate 15 kW continuous RF power.

(5) Circular polarization.

A surface-wave antenna that could be mounted coaxially inside the L-band horn was selected to satisfy the above needs. This C-band feed was composed of two main parts: an artificial dielectric rod, in this case a disc-on-rod [2] that supports a surface wave, and a launching horn that excites a surface wave on the disc-on-rod. The major challenge of this design was to obtain a high surface-wave ($HE_{11}$-mode) launching efficiency. Higher launching efficiency allows greater isolation from the L-band horn and greater control over the C-band RF radiation pattern. Improved launching efficiency was accomplished in two ways. First, the $TE_{11}$-mode smooth-wall launching horn used in [2] was discarded and an $HE_{11}$-mode corrugated horn was used for a better match with the $HE_{11}$-mode surface-wave antenna. Second, the surface-impedance taper at the input of this antenna was optimized for high surface-wave launching efficiency.

Over most of the length of the disc-on-rod, the surface-impedance profile, controlled by varying the disc diameter along a constant diameter rod, was designed to balance two conflicting needs: first, to couple the surface wave tightly enough to minimize its interaction with the surrounding L-band horn and, second, to couple the surface wave as loosely as practical to minimize the resistive loss. Over the length of this disc-on-rod, after the input launching taper, the disc diameter was gradually made smaller in a series of straight sections and tapers. Finally, the surface wave was coupled loosely enough to achieve the needed effective aperture at the radiating end of the disc-on-rod. The length of the disc-on-rod was adjusted to collocate its phase center with that of the surrounding L-band horn (approximately 20 in. inside the aperture). Details of the disc-on-rod and launching horn are shown in Figs. 3 and 4.

The 15 kW of continuous C-band power transmitted on the surface-wave antenna presented three problems. First, the removal of the disc-on-rod's resistive heat loss of approximately 600 W was accomplished with an internal water flow (0.6 gallons per minute) through the entire length of the rod. Second, the disc-on-rod was primarily supported (in a high-power field) by aramid string tension members at regular intervals along its length. This dielectric string was small in diameter (0.025 in.) and nearly transparent to the L-/C-band propagation. Third, in order to protect the L-band low-noise amplifier from the transmitted C-band power, isolation was provided by tight coupling of the surface wave near the base of the L-band horn and the addition of a bandpass filter in the L-band waveguide.

Two additional components were designed as part of the C-band feed: first, a coaxial waveguide, $TE_{11}$-mode, quarterwave-plate polarizer, and second, a rectangular-waveguide-to-coaxial-waveguide transition in which an inductive post was mounted to impedance-match the C-band feed to the transmitter.

## B. L-Band Feed

The L-band feed was made up of a dual-mode "Potter" horn [4], a four-arm waveguide network to bypass the C-band input, a suppressor for unwanted modes in the circular waveguide, and a circular-to-rectangular-waveguide transition. The L-band feedhorn was a sheet metal dual-mode antenna with a $TE_{11}/TM_{11}$-mode converter step (see Fig. 2). The L-band received signal was extracted by a symmetric, four-arm network just forward of the C-band launcher, and the signal was recombined just behind the C-band input. This L-band network not only made room for the C-band input, but also served as a circular polarizer by the fact that one opposing pair of arms was one quarter of a guide wavelength longer than the other pair.

The L-band horn was impedance-matched over a 5 percent bandwidth primarily by adjusting the axial position of the C-band launch horn relative to that of the L-band horn and placing inductive posts in the four waveguides of the bypass network near its junction with the L-band horn.

The L-band mode suppressor consisted of diagonally opposed probes inserted into a circular waveguide orthogonal to the desired $TE_{11}$ mode's E-field and phased to couple higher order propagating modes ($TM_{01}$ and $TE_{21}$) to a resistive load. Without this suppressor, higher order mode resonances occurred within the feed at L-band.

## III. Results

The dual-frequency feed was put through a series of RF tests[1] including pattern, voltage standing wave ratio (VSWR), L-/C-band isolation, high power, and noise temperature. Typical amplitude and phase RF radiation patterns for this L-/C-band feed are shown in Figs. 5 and 6. The VSWR was less than 1.14:1 at L-band and less than 1.08:1 at C-band. The isolation provided by this feed geometry between the C-band input and the L-band output was greater than 38 dB. The high-power tests included a 15 kW run for 4 hr continuously with no anomalies. The L-band feed noise temperature was 8.15 K.

The dual-frequency feed was mounted on the DSS 14 70-m antenna and system tests were performed. Details of the 70-m system test with this L-/C-band feed in place are given in [5]. Table 2 summarizes the test results. The system was in compliance with the requirements listed in Table 1 except that the L-band system temperature was 35.3 K instead of the required 35 K or less.

[1] For a detailed account of these tests, see M. Gatti, *C-/L-Band Feed Test Summary*, JPL Interoffice Memorandum 3334-88-052 (internal document), Radio Frequency and Microwave Subsystems Section, October 18, 1988.

## References

[1] J. Withington, "DSN 64-Meter Antenna L-Band (1668-MHz) Microwave System Performance Overview," *TDA Progress Report 42-94*, vol. April–June 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 294–300, August 14, 1988.

[2] S. A. Brunstein and R. F. Thomas, "Characteristics of a Cigar Antenna," *JPL Quarterly Technical Review*, vol. 1, no. 2, Jet Propulsion Laboratory, Pasadena, California, pp. 87–95, July 1971.

[3] S. Narasimhan and M. S. Sheshadri, "Propagation and Radiation Characteristics of Dielectric Loaded Corrugated Dual-Frequency Circular Waveguide Horn Feeds," *IEEE Trans. Antenna Propagat.*, vol. AP-27, no. 6, pp. 858–860, November 1979.

[4] P. D. Potter, "A New Horn Antenna with Suppressed Sidelobes and Equal Beamwidths," *Microwave J.*, vol. VI, no. 6, pp. 71–78, June 1963.

[5] S. Gatti, A. J. Freiley, and D. Girdner, "RF Performance Measurement of the DSS-14 70-Meter Antenna at C-Band/L-Band," *TDA Progress Report 42-96*, vol. October–December 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 117–125, February 15, 1989.

Table 1. RF requirements

| Parameter | C-band | L-band |
|---|---|---|
| Bandwidth, MHz | 5010 ±10 | 1668 ± 40 |
| Gain, dBi | — | $\geq 57$ |
| Effective isotropic radiated power, dBm | $\geq +138$ | — |
| Noise temperature, K | — | $< 35$ |
| Polarization | Right circular | Left circular |
| Ellipticity, dB | $< 2$ | $< 2$ |
| Power, kW | $\geq 15$ | — |

Table 2. Measured RF performance

| Parameter | C-band | L-band |
|---|---|---|
| Bandwidth, MHz | 5010 ±10 | 1668 ± 40 |
| Gain, dBi | — | 59.5 |
| Effective isotropic radiated power, dBm | +138.3 | — |
| Noise temperature, K | — | 35.3 |
| Polarization | Right circular | Left circular |
| Ellipticity, dB | $< 2$ | $< 2$ |
| Power, kW | $\geq 15$ | — |

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

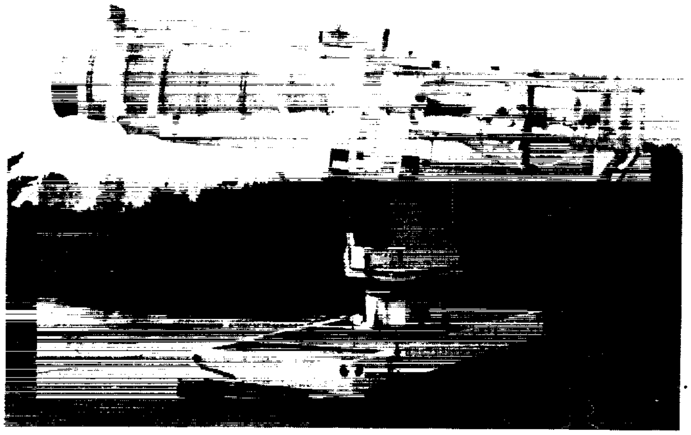


Fig. 1. L-/C-band feed.

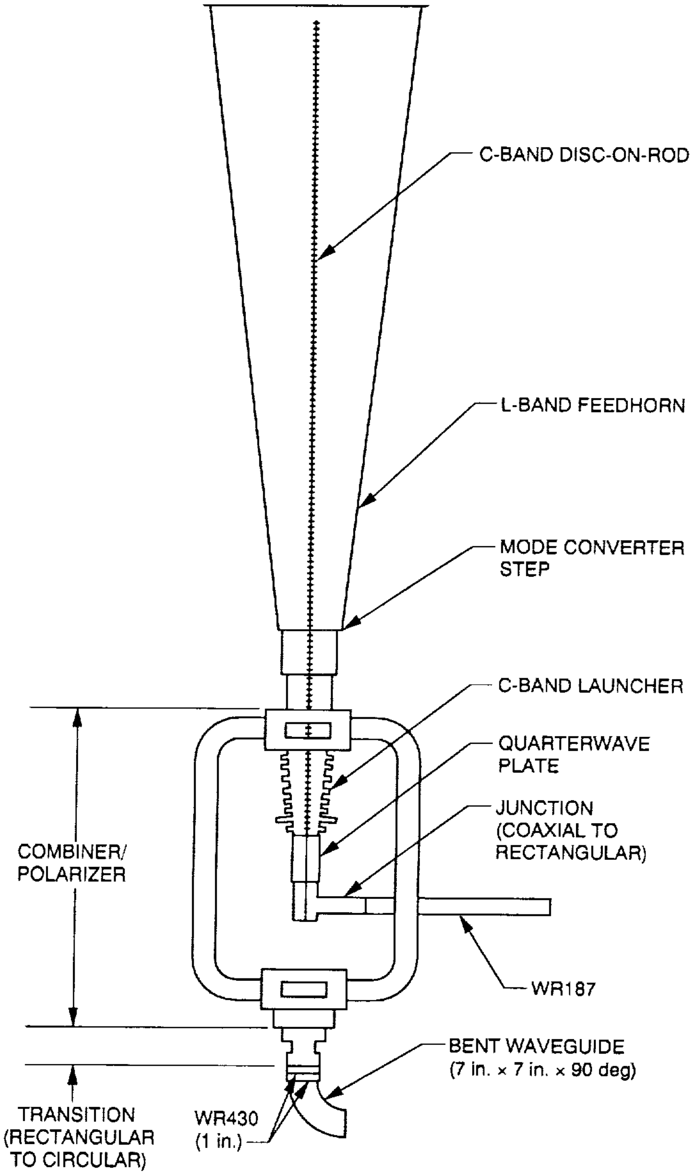


Fig. 2. L-/C-band dual-frequency feed system.

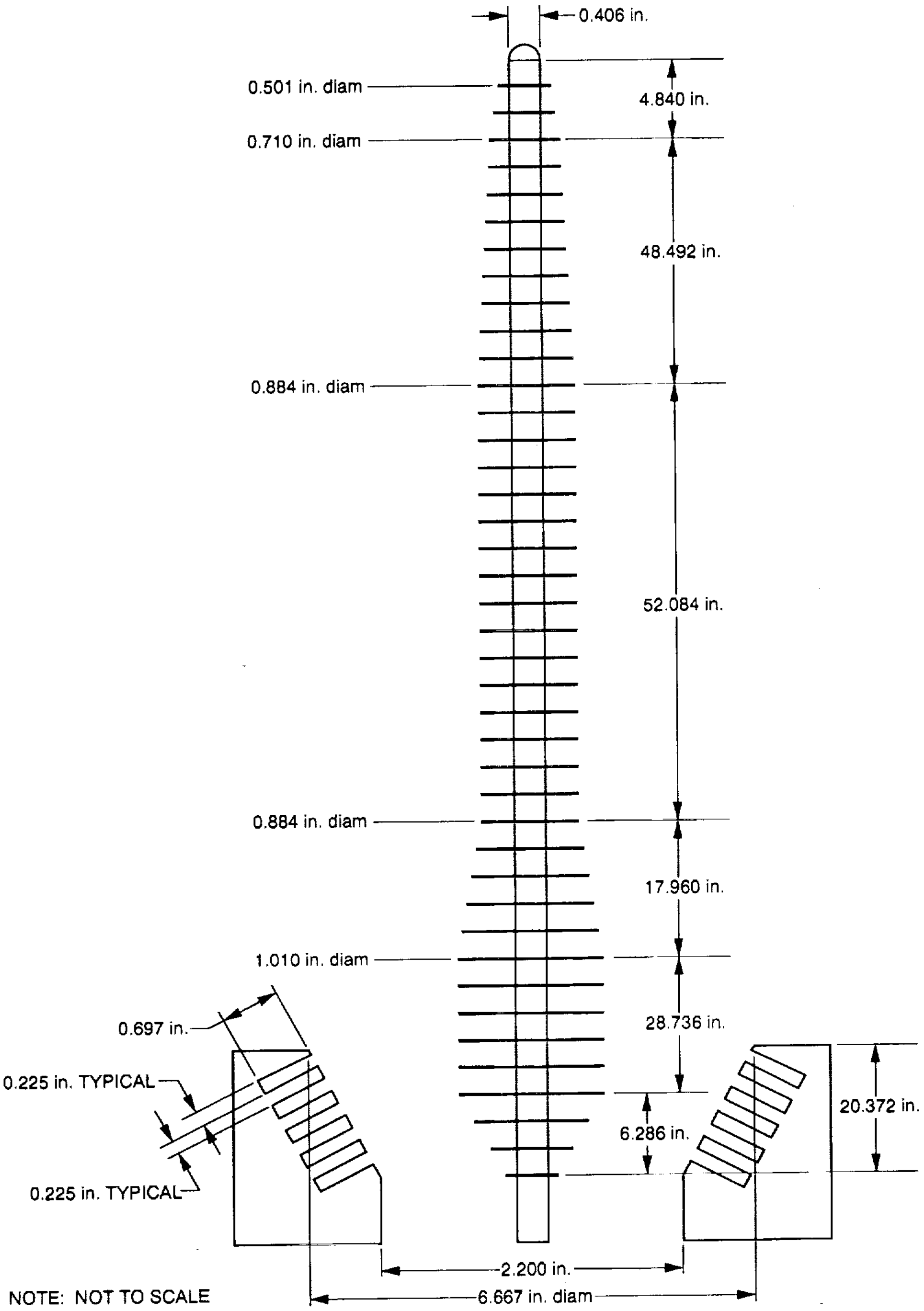


**Fig. 3. C-band disc-on-rod and launching horn.**

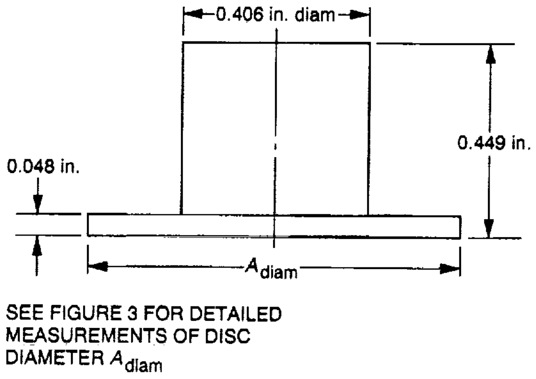

0.406 in. diam
0.449 in.
0.048 in.
$A_{diam}$
SEE FIGURE 3 FOR DETAILED MEASUREMENTS OF DISC DIAMETER $A_{diam}$

**Fig. 4. Element detail of disc-on-rod.**

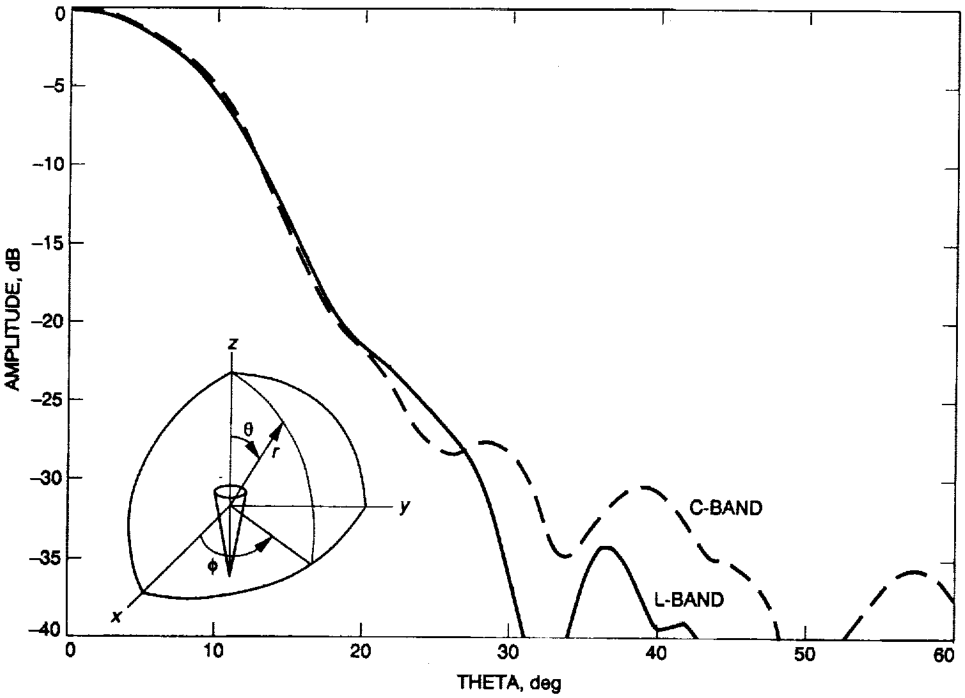


**Fig. 5. Amplitude of L-/C-band feed pattern.**

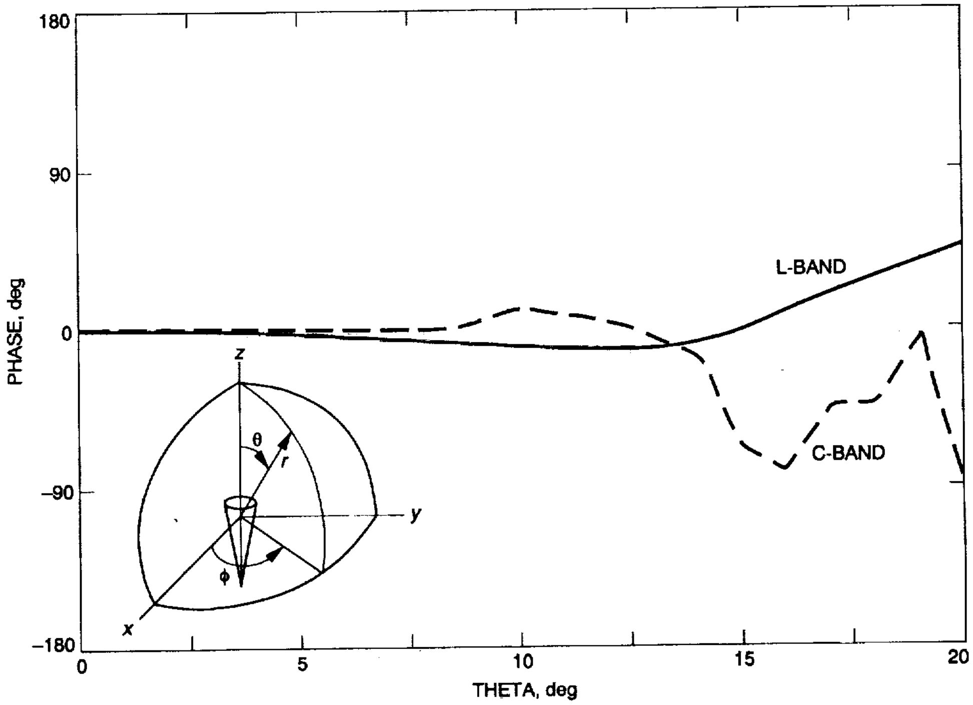


**Fig. 6. Phase of L-/C-band feed pattern.**

# Mark IVA Antenna Control System Data Handling Architecture Study

H. C. Briggs and D. B. Eldred
Guidance and Control Section

*A high-level review was conducted to provide an analysis of the existing architecture used to handle data and implement control algorithms for NASA's Deep Space Network (DSN) antennas and to make system-level recommendations for improving this architecture so that the DSN antennas can support the ever-tightening requirements of the next decade and beyond. It was found that the existing system is seriously overloaded, with processor utilization approaching 100 percent. A number of factors contribute to this overloading, including dated hardware, inefficient software, and a message-passing strategy that depends on serial connections between machines. At the same time, the system has shortcomings and idiosyncrasies that require extensive human intervention. A custom operating system kernel and an obscure programming language exacerbate the problems and should be modernized. A new architecture is presented that addresses these and other issues. Key features of the new architecture include a simplified message passing hierarchy that utilizes a high-speed local area network, redesign of particular processing function algorithms, consolidation of functions, and implementation of the architecture in modern hardware and software using mainstream computer languages and operating systems. The system would also allow incremental hardware improvements as better and faster hardware for such systems becomes available, and costs could potentially be low enough that redundancy would be provided economically. Such a system could support DSN requirements for the foreseeable future, though thorough consideration must be given to hard computational requirements, porting existing software functionality to the new system, and issues of fault tolerance and recovery.*

## I. Introduction

The objective of this study was to provide an independent assessment of the capabilities of the current antenna control system data handling assemblies and to make recommendations for future subsystem development to be used as a guide during planned upgrades in the 1992 to 1995 time frame. The study focused on the data handling architecture of the antenna control system as it exists in the field and on its ability to handle the task of providing operator monitoring and control of the system. The system is built upon an elaborate message processing and transmission protocol that clearly shows the impact of ad-hoc system evolution in the form of late and lost messages. The study addresses the handling of the monitor and control data messages and the suitability of the design to sustain adequate message handling services.

The specific algorithms of the antenna control system have been reviewed, and a number were identified as having the potential to significantly impact the ability of the system to maintain timely responses. These algorithms include predict processing, correction for systematic errors, automatic boresighting correction, display generation, and collection of monitor data. Where alternate algorithms that require fewer system resources can be identified, these have been incorporated into the assessment and recommendations. Furthermore, although few criticisms have been levied against the lowest level servo-loop algorithms, it would be prudent to allocate increased system resources to them in order to support the evolution that should be possible in a healthy system with margin for growth.

The study was not constrained by implementation considerations associated with transition to a new system, system outages, and planned maintenance periods. These considerations are, however, real limitations on any upgrade plan, and, although they have not been addressed specifically, these issues have been reflected implicitly in the proposed data handling architecture in that the number of proposed changes to the system is minimized.

## II. Overview of the Current Architecture

Each antenna at a given site includes its own antenna control system. The Control Monitor Console (CMC) operator builds a link consisting of a Link Monitor Console (LMC), the Antenna Pointing Assembly (APA), and an antenna to enable communications with a spacecraft or for radio astronomy. Each APA is time-shared among up to three links. Because of the critical role of the APA in the link, a second APA is held in reserve as a backup, and can be automatically switched into action in case of failure. The control system architecture is defined as the sum of the subsystem functions, their assignment to hardware elements, and the resulting topology (Fig. 1). The antenna physical interfaces are not shown, but they consist of angle encoders, gimbal motor rate commands, and safety interconnect logic connected to the Antenna Servo Controller (ASC), Antenna Control and Monitor (ACM), Master Equatorial Controller (MEC), and Subreflector Controller (SRC).

The principal functions of the antenna control system consist of antenna pointing control, antenna performance monitoring, and maintenance support and fault isolation.[1] Partitioning these functions into subunits begins with the APA, which receives and processes directives addressed to the control system. The major APA functions include[2]

(1) Processing predicts. The APA receives predicts from the LMC specifying antenna attitude and corresponding time at coarsely spaced intervals. The predicts are interpolated in time and coordinate transformed.

(2) Performing automatic boresight correction (Conical Scan [CONSCAN]). Receiver automatic gain control (AGC) data are collected over a 30-sec period and processed using a fast Fourier transform algorithm.

(3) Monitoring status for performance and safety and providing operator displays.

The Antenna Control Subassembly (ACS), one of which is dedicated to each antenna, is assigned to an APA by the CMC for a given link. The ACS must be capable of operating successfully even during a switch from one APA to an alternative or a change in the controlling CMC. The major functions of the ACS include[3]

(1) Pointing the antenna as directed by the APA. The predicts are refined, more coordinate transformations are applied, and systematic error corrections are added.

(2) Participating in automatic boresight pointing by providing pointing angles to the APA and executing scan-drive pointing commands.

(3) Monitoring and configuring antenna equipment.

(4) Processing systematic error corrections. The ACS incorporates site-dependent systematic error corrections into the predicts. These corrections are used to offset biases introduced by gravity sag, atmospheric refraction, encoder bias, and axis misalignments. The systematic error correction tables are prepared off-line by the antenna engineer and stored with other configuration data in the LMC.

The ASC, MEC, and SRC control servo loops are similar to one another when viewed at the functional level.

[1] *DSN System Functional Requirements Document 820-20 Rev. A, General Requirements and Policies Through 1988*, JPL D-5081, vol. 1, change 3 (internal document), Jet Propulsion Laboratory, Pasadena, California, pp. 3–5, June 15, 1980.

[2] Op. cit., p. 3C-3.

[3] Op. cit., p. 3C-8.

On those antennas with an MEC, the ACS sends pointing commands to the MEC and the ASC is slaved to follow, with error signals originating from an autocollimator. The ACS provides corrected pointing predicts to the MEC and to the ASC as backup information, but corrects each predict using systematic error tables for the individual controller. In the event of a failure or loss of lock in the MEC, the ASC drops back to executing pointing commands from the ACS. At a high level, the servo-loop controllers perform the following functions:

(1) Executing the antenna pointing commands via a servo control loop.

(2) Monitoring and reporting the status of equipment.

## III. Analysis of the Current Architecture

Prior to any explanation of the problems that prompted the current study, it should be clearly understood that the system is currently fulfilling its intended role. Taken as a whole, the Mark IVA is a complex system that has served admirably for some 6 years. Improvements and fixes have been applied over this time in a layered fashion, and expectations have changed because of advances in technology and improved understanding of the practical requirements based upon experience. Nevertheless, there are anomalies on record, functions that are inoperable, and a pervasive sense among the operators that the control system gradually grinds to a halt during heavy loading after a few hours of use.

Well-defined symptoms of problems have been observed. Operators have experienced slow update rates for critical display information, lost messages, and floods of warnings that ultimately originate from a single event. This combination of problems has two effects. First, the congestion in the communications channels is aggravated, possibly causing interference with other functions. Second, the operator's ability to respond correctly and in a timely fashion to an event is impaired, because of the time delays associated with critical messages and the distraction caused by a flood of messages. Currently, each subsystem generates its own set of messages for the operator when problems are identified. Although problems are rated on an urgency scale from one to five, causality is difficult to trace. Thus, messages tend to proliferate through the system. One result of this is that the APA processor duty cycle has been observed at close to 100 percent.[4] Though an upgrade to a new, higher capacity processor for the APA has already been approved, such an approach treats the symptom of the problem rather than the root cause, which lies in the system connectivity and functional allocations.

Many of the anomalies are related to new requirements or unmet expectations. For example, the antenna and subreflector positions are not logged, and the displays generated by the system are not as useful as they might be. The changes required to correct these anomalies can be adequately negotiated and implemented in the future, just as the many other changes to the system have been performed over the past years. Only a few of these reported anomalies reflect the system inadequacies that are the subject of this study. The system change process does draw attention, however, to a principal shortcoming of the system reflected in the lack of reserve Central Processing Unit (CPU) capacity to easily execute software repair and improvement. Furthermore, some algorithms and practices (e.g., CONSCAN) consume resources inefficiently, although they do not appear to have a major impact on the system performance. On the whole, the functional design of the system is adequate, given the original requirements and the use of the system as it has evolved.

Within the scope of this study there are several identifiable features of this architecture that impact the performance of the system. These features are candidates for change in a modified architecture:

(1) The message passing system needs improvement. A significant portion of the observed behavior of the system is tied to message passing response, since any message or command originating at the LMC must pass through several subsystems prior to reaching its addressee. Most of these intermediaries perform some processing of the message, either translating or expanding its contents. This multi-level hierarchical message passing system results in significantly delayed information when any subsystem is highly loaded.

(2) The control subsystems were implemented in computer hardware that, while current for the time, is significantly outdated by current standards. The hierarchical partitioning of the system may in fact be a derivative of the limitations of the then-current hardware. An alternative architecture built upon more capable hardware would place major functions entirely within one subsystem.

(3) The communication channels slow the performance of the system, given the basis in message passing and the hierarchical topology of the system.

[4] R. D. Rasmussen and N. T. Brady, *Mark IVA Antenna Control System Data Handling Study Final Report* (internal document), Jet Propulsion Laboratory, Pasadena, California, February 9, 1990.

(4) The existing hardware topology constrains the potential functional topology options. The physical serial communications lines severely restrict the possibility of reallocating functions or repartitioning message paths.

(5) Logging and archiving of monitor statistics are not performed in a systematic manner. Typically, the operators disable the monitor data logging function because it slows down the system.

As a result of the existing multilevel architecture, detailed knowledge of the antenna subsystem is embedded in several subsystem elements. The operator is forced to know detailed, low-level start-up, operation, and fault isolation sequences. Hardware-specific errors are reported by every system. Operability and maintainability of the system would be substantially improved if each antenna subsystem received higher level commands and reported summary error conditions in a condensed, predigested format. Ideally, the operator might, for example, command initiation of a track for a specific spacecraft, and the antenna control subsystem would configure, point, and monitor automatically, based upon preconfigured tables and supervisory functions. While the antenna control proceeds automatically, a monitor function associated with the antenna would collect exceptions, filter and process messages, and provide critical advisories to the operator.

## IV. Specific Recommendations

### A. Predict Processing

Predict processing is a core function of the system and provides the directives for pointing the antenna control system. The system's communication and computation burdens would be significantly reduced if a single interpolation algorithm were used by the servo controller. Further reduction in computation could be achieved by using a single coordinate system for all predicts and delaying the conversion to the antenna-peculiar coordinate system until the predicts reach the servo controller.

Implementing these changes would reduce computations, memory and disk storage requirements, and communication channel message volumes. The algorithms appear to be well within the capability of current hardware. This amounts to the replacement of extensive tabular data with an on-demand generator function.

### B. Application of Systematic Error Corrections

The predicted pointing angles are corrected in the ACS for such systematic errors as gravity sag and atmospheric refraction. Application of these corrections is not computationally intensive, since it involves combinations of table entries and simple interpolations. The only change recommended is to provide automatic selection of the proper tables without operator intervention, given the operational mode.

### C. Generation of Systematic Error Correction Functions

Determination of the proper systematic corrections is severely hampered with the current method of data monitoring. Logging should be reinstated, and the systematic error modeling and creation of the correction tables should continue to be done off-line. The antenna engineer should be supported with investigative access to the antenna, perhaps under the maintenance mode, to assist in identifying error mechanisms and parameters. The Program Office should continue to invest in the development of on-line error-tracking and system calibration capabilities. Future enhancements might include the incorporation of real-time calibration and error tracking.

### D. Generation of Operator Displays

Data for inclusion in displays are generated at all levels of the control system, and most of the displays themselves are generated in the APA. In general, displays should not be generated by any element of the antenna control system that must also execute a real-time control function such as CONSCAN, servo loops, or exception monitoring. The operator console or a new unit in the antenna control system should be used to build the displays directly from a new database without recurring inquiries acted upon by the servo-loop systems. This implies an attendant change from generation of data on demand to a policy of maintaining a database of status information. Monitor and display data should be passed in a compact, binary mode using agreed-upon data structures.

### E. Single Point Antenna Monitor and Control

The monitor and control functions contained in the current functional architecture of the antenna control system should be collected and utilized as a single point of representation for the antenna. The intent is to partition low-level knowledge of the configuration and error handling to functions within the antenna, leaving the LMC operator with a consolidated, high-level interface. This function, named herein the "antenna monitor and control function," is not significantly different from the original design role of the ACM subsystem. However, the ACM as presently implemented is too close to the mechanical interfaces of the antenna hardware and, except for the watchdog activity, executes little monitoring and no control.

Several implementation issues support the reinstatement of an effective monitor and control function besides raising the level of the operator interface. An interface between the control room Local Area Network (LAN) and the antenna network is required to minimize undesired message traffic, buffer signals for the potentially long distance to the antenna, and allow for media differences between the two network domains. Thus, a computer system that hosts the monitor and control function that is located in the control room can serve as the internetwork gateway and electronic interface.

## V. Proposed Architecture

### A. Functional Architecture

To alleviate the observed performance problems of the current system, modifications to the architecture could incorporate one or more of the following (the top-level system diagram that incorporates these modifications is shown in Fig. 2):

(1) A simplified message-passing hierarchy where the path from any sender to any receiver includes at most one, and preferably no, intermediate subsystems or logical tasks. The intention is to reduce message passing delays; when accompanied by consolidation of tasks within more powerful servo controllers, this can be readily accomplished.

(2) Implementation of a network local to the antenna-control system that eliminates the point-to-point low-speed communication channels. This would establish each subsystem as a peer on the network and provide the physical mechanism to support direct addressing of all messages. Most reasonable network implementations will also provide reliable communications at substantially higher speeds than the present RS-232 protocol. Careful consideration must be given to the distance requirements, which, for some antennas, might reach 10 km from the control room to the antenna. Again, most reasonable network implementations can meet this requirement, although careful design is required, and slightly higher costs can be expected for the long-distance segments of the network.

(3) Redesign of certain functions to reduce the computation, storage, and message volume associated with the algorithms. Specifically, predict processing might employ a single coordinate system and a single interpolation algorithm.

(4) Consolidation of antenna monitor and control functions and implementation in a system that resides at the interface between the control room network and the antenna network.

(5) Utilization of new, modern 32-bit commercial computer hardware with a real-time operating system kernel and network-based communications. This would provide significant flexibility in the design of the new architecture task structure and provide a basis for potential incorporation of fault tolerance through semiautomatic reconfiguration. The principal motivation for this change is to provide increased processing power and high-speed, standardized communications at relatively low cost.

The system functions would be retained and allocated to the logical units of the modified architecture. The principal differences are

(1) CONSCAN, or automatic boresighting, is executed entirely within the servo controller.

(2) Systematic errors are corrected at the servo controller.

(3) The monitor and control function oversees operation of the antenna subsystem and maintains a database that is used to generate displays for the LMC.

### B. Principal Features of the Modified Architecture

The topology of the modified architecture represents a simpler hierarchy by replacing the APA and ACS with a single unit that is logically a part of the antenna control system and provides a network interface and high-level monitor and control functions. Elimination of the functions currently allocated to the APA and ACS is not proposed. All existing functions should be retained, although a few should be reformulated and hosted in new subsystems.

The antenna control subsystems are depicted as peers on a local network. This network might be implemented in several ways, for example, in Ethernet between separate computers or in shared memory among multiple processors on one common backplane. Such a network should support a virtual communications system defined in software, as might be done with Unix sockets or Transmission Control Protocol/Internet Protocol (TCP/IP) channels, and would provide reliable, high-speed communications at very low cost.

The architecture retains the MEC and SRC units, principally as logical functions. It might be entirely possible to host these less intensive functions as tasks on the same CPU with the ASC. Hardware such as an MC68040 on the VMEbus or an i80486 on the Multibus II provide sufficient capacity to execute the current functions and servo loops. Economical systems with either Ethernet or backplane networks are readily available.

An additional subsystem, the Antenna Monitor and Control (AMC), has been included in the architecture to provide a high-level interface to configure and control the antenna and to monitor the status of the antenna system. The AMC provides local access for maintenance, monitoring, or calibration, and provides data collection, analysis, and monitoring for identification of systematic errors, equipment failures, or anomalies in support of reconfiguration, and potentially even intelligent real-time oversight of operational status. These functions are probably quite similar to the original intent of the ACM, although the ACM as it has evolved is too intimately associated with the physical interfaces to the antenna to also host higher level functions and network connection needs.

The AMC takes on the display generation functions currently resident within the APA. Routine monitor data are collected and stored in the database within the AMC and utilized to generate displays as required by the operator. All high-level operator commands are expanded and sent directly to the subsystem that hosts the targeted task. For example, predicts and interpolation parameters are messaged directly to the pointing task resident on the ASC where the servo loop performs the interpolation and correction when required.

## VI. Implementation Options

Beyond the recommendation to rebuild the system in modern commercial software and hardware, several implementation and configuration options are available. The commercial market is well developed, with excellent product selection and price competition. A good development environment based upon Computer Aided Software Engineering (CASE) tools for real-time systems should be utilized and the control software rewritten based upon a real-time kernel. Where possible, the existing HAL-S and PL/M code should be converted to a current language that can be supported by the CASE tools and available programmers.

In the hardware arena, configuration options for the computers that host the antenna control system functions range from replicating the current configuration, which is one CPU per chassis per subsystem, to one chassis with a few CPUs hosting all functions. Within this implementation framework, several development options exist that can contribute to improved reliability and the incorporation of fault tolerance. The necessity of such improvements depends entirely on the yet-to-be-developed detailed reliability requirements.

## VII. Summary and Recommendations

This study focused on the architecture of the antenna control system and its ability to handle the task of providing operator monitoring and control of the system as it has evolved and exists in the field. Recommendations for modifications to the existing architecture have been proposed and further investigation and design in key areas is required prior to final planning for the system upgrade. This section briefly summarizes these areas.

Detailed assembly-level requirements must be derived for such functions as error reporting, operator control of lower level devices, and status presentations. The system documentation consistently neglects the development of coherent propagation of requirements to the lower levels of the servo controllers.

With derived requirements in hand, the modified architecture design can be continued. The immediate need is for a budgetary cost estimate for the upgrade. This can be obtained once CPU loads, communication channel capacities, and memory sizes are estimated from the functional descriptions and requirements. Since the system's overall costs are not very sensitive to the exact number of Single Board Computers (SBCs), attention should be directed to firming up the architecture configuration.

Considerable attention is required in the area of specifying the reliability requirements for the system to the level of detail that permits selection of a particular method. The current central requirement based upon a guaranteed not-to-exceed downtime is insufficient. Other derived requirements in the form of probability of failure, fail-safe or fail-tolerant design paradigm, or worst case failure are required to enable a design choice among such reliability enhancements as voting, hot spares, and fault-tolerant subsystems.

Development of an on-line error tracking and system calibration capability needs to be done first so that requirements can be integrated into the planned upgrade. Most probably, a development period of several years will be required and installation in the antenna systems might occur after the upgrade period. With an early start on the development, capacity might be installed to accept later deliveries, or at least the system can incorporate the necessary "hooks" and access ports for eventual installation.

There are significant constraints associated with maintaining system services during the modification period that were excluded from the current study. Several means to achieve a phased, testable installation exist, and the ramifications and details would be worked out in the development plan.

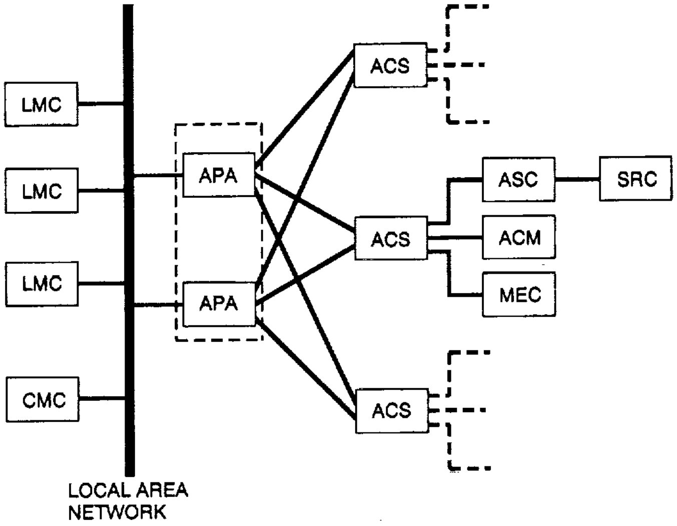


Fig. 1. Current system architecture.

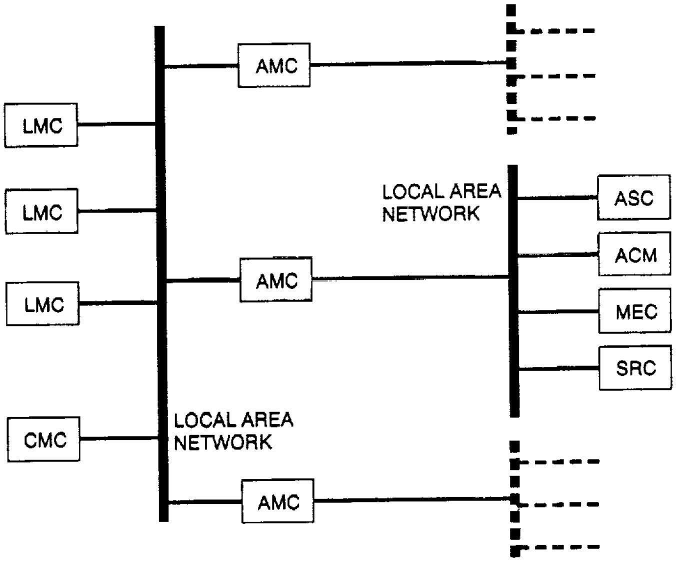


Fig. 2. Modified system architecture.

511-32

53214 p-9

# Expected Antenna Utilization and Overload

E. C. Posner
Office of Telecommunications and Data Acquisition

*This article develops the trade-offs between the number of antennas at a DSN Deep-Space Communications Complex and the fraction of continuous coverage provided to a set of hypothetical spacecraft, assuming random placement of the spacecraft passes during the day. The trade-offs are fairly robust with respect to the randomness assumption. A sample result is that a three-antenna complex provides an average of 82.6 percent utilization of facilities and coverage of nine spacecraft that each have 8-hour passes, whereas perfect phasing of the passes would yield 100 percent utilization and coverage. One key point is that sometimes fewer than three spacecraft are visible, so an antenna is idle, while at other times, there aren't enough antennas, and some spacecraft do without service. This point of view may be useful in helping to size the network or to develop a normalization for a figure of merit of DSN coverage.*

## I. Introduction

In deciding how much antenna (and signal processing) capability is appropriate for the DSN, user requirements need to be known. One also needs to know how the spacecraft to be supported are distributed in the sky, or rather in the duration of their pass over a single DSN Deep-Space Communications Complex. This article assumes a random independent pass distribution. This might hold over many years, but in a particular year or even decade there will be bunching—for example, the outer planets at southern declination during the 1980s. Or, there might be several orbiters around the same planet, e.g., Mars, but not in the same antenna beam. Thus, actual near-term facilities' decisions need actual view period knowledge. In the long term, though, randomness may be a good assumption.

The differences between the model presented here and the actual DSN are as follows. First, this article considers one Complex in isolation, instead of all three Complexes together. This would give a correct picture if the view periods at each Complex were the same, and all spacecraft required continuous coverage. The view periods are not the same, of course, so to find total coverage, one would have to sum the coverage at each Complex. Second, no distinction is made here between types of facilities at a Complex, e.g., between 34-m high efficiency (HEF) and 70 m. This leads to a harder problem than the one considered here. However, the methods of this article are relevant for the 70 m alone for spacecraft that can only be supported by the 70 m. A third, and the most significant, difference between the model and reality is that the model assumes that all spacecraft require continuous coverage. Some may and some may not require coverage, and passes can be moved around to meet requirements or critical spacecraft events can be scheduled to occur when a facility is in view. To find the fraction of requirements that can be supported in the realistic situation is a harder problem to state and solve than the one considered here. Nevertheless, this article is useful because it still gives an estimate for the expected time when there is nothing to track at all, an important parameter to know. Finally, a fourth difference is that actual view periods are not random-independent, but occur according to mechanical laws. As argued above, for the long term, randomness may be as good an assumption as can be made. The assumption that all pass lengths are equal is not essential in what follows, but it does simplify the formulas; it is not counted as a difference between the simple model and reality.

Thus, it is useful to consider the problem for random passes. At the very least, this can be used as a calibrator of the actual utilization, sort of a zero point for a figure of merit on the adequacy of DSN facilities. This article develops the random-pass model and applies it to cases involving realistic numbers of spacecraft and antennas. The methods used provide expected values quite easily. To determine probabilities such as "what is the probability that three antennas can service 90 percent of five spacecraft with random 12-hour passes?" is harder and will not be attempted here. Note also that continuous coverage requirements are assumed since this is a model that can be solved and often occurs.

Section II develops the model, Section III solves it for expected values, and Section IV works out some cases of relevance to the size of typical DSN deep-space mission sets. Section V, the concluding section, compares the theoretical results with some actual spacecraft view periods. The results on percent coverage for the actual view periods are not too far off the theoretical expected values.

## II. The Model

Think of the 24-hour day as the circumference of a circle, and the passes as connected intervals of arc. The circumference length is assumed equal to 1, and the length of a pass is $\beta$ (where $0 < \beta < 1$), which is the same for all spacecraft in this model, although this assumption is not essential to solve the model. The circumference of a circle is used as if each previous and succeeding day had the same spacecraft visibilities. Although the visibility periods or pass times precess somewhat from one day to the next, this is a minor effect.

As in Fig. 1, there are $n$ spacecraft corresponding to $n$ intervals of arc $X_i(\omega)$, where $0 \leq \omega < 1$ is the phase around the circle and the random variable $X_i$ is 0 if the spacecraft is not visible and 1 if it is visible. It is noted here that the expected value method does not require these intervals of arc to be connected, even though they are connected for deep-space spacecraft.

The $n$ spacecraft are also viewed as $n$ independent random variables $\{X_i\}$ defined on the circle. The pass lengths are all assumed to be equal to $\beta$, so that the expected value of the $0-1$ random variable $X_i$ is $\beta$. The reason is that $\beta$ is the probability that the pass contains a particular point on the circle, since the $\beta$ of the circle is covered by spacecraft $i$. The method used in this article would work only if the *expected* pass length were $\beta$; it is not actually required that the pass lengths be $\beta$ with certainty, nor even that a "pass" be connected. What *is* important in this model is that the spacecraft is to be supported every time it is visible—for example, no data dumps. With a minor modification of this method, the lengths of the passes can be spacecraft dependent, as was said above. Another assumption is that all antennas are assumed equivalent here—for example, no distinction is made between the 34-m and 70-m antennas.

In Fig. 1, intervals of time during the day (intervals of arc on the circle) are shown where no spacecraft is trackable because none is visible. There are regions where a certain number of spacecraft are visible: only one spacecraft, exactly two, exactly three, and exactly four (the maximum as assumed in the figure). All antennas (say there are three) would be idle during times when there is no spacecraft to track, two of the three would be idle during times with only one spacecraft visible, one would be idle during times with only two spacecraft visible, and all three would be busy and all visible spacecraft being tracked during times when exactly three spacecraft are visible. When four spacecraft are visible, all three antennas would be busy, but one spacecraft would not be tracked. In this case, some requirements are not being met. The rest of the article finds the expected fractions of time for all these conditions of visibility and antenna utilization.

## III. Analytic Expressions

Recall that $X_i(\omega) = 0$ or 1 depending upon whether spacecraft $i$ (one of $n$ spacecraft) is not visible or is visible at "time" $\omega$ on the circle of circumference 1. The expected value

$$E(X_i) = \beta \tag{1}$$

for all $i$, where $\beta$ is the relative pass length. Thus, $\beta = 1/2$ corresponds to 12-hour passes. The expectation is $\beta$, because all rotations of the pass around the circle are equally likely to be selected by the random pass generator.

The random fraction $U_0(\omega)$ of the circle, or day, where none of the $n$ spacecraft is visible is

$$U_0(\omega) = \prod_{i=1}^{n} [1 - X_i(\omega)] \tag{2}$$

This is because $U_0(\omega) = 1$ precisely for these $\omega$ points where all $X_i(\omega) = 0$, i.e., no spacecraft is visible. The expected fraction $\alpha_0$ of a day where no spacecraft is visible, by the independence assumed for the random variables $X_i(\omega)$, is

$$\alpha_0 = E[U_0(\omega)] = E\prod_{i=1}^{n}[1 - X_i(\omega)]$$

$$= \prod_{i=1}^{n} E[1 - X_i(\omega)] = (1-\beta)^n$$

so

$$\alpha_0 = (1-\beta)^n \tag{3}$$

More generally, let $U_k(\omega)$ be the random fraction of a day during which exactly $k$ spacecraft are visible, where $0 \leq k \leq n$. It is given by the following, not very useful, expression:

$$U_k(\omega) = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_k}\prod_{l=1}^{k} X_{i_l}(\omega) \times \prod_{j \neq \text{any } i_l} [1 - X_j(\omega)] \tag{4}$$

This merely makes $U_k(\omega) = 1$ precisely when exactly $k$ spacecraft (such as $i_1, \cdots, i_k$) are visible to the tracking station. Let $\alpha_k$ be the expected time during which exactly $k$ spacecraft are visible. On taking the expectation, Eq. (4) becomes the more useful result

$$\alpha_k = E[U_k(\omega)] = \sum_{i_1}\cdots\sum_{i_k}\beta^k(1-\beta)^{n-k}$$

which becomes

$$\alpha_k = \binom{n}{k}\beta^k(1-\beta)^{n-k}, \qquad \text{for } 0 \leq k \leq n \tag{5}$$

Now suppose there are $r$ interchangeable antennas at the Complex. Here $1 \leq r \leq n$ is of interest, since all spacecraft are supported all the time when $r = n$. What are the expected total requirements $E_r$ that are not met? Good units to use for $E_r$ are spacecraft days. However, the metric $E_r$ does not consider exactly *which* of the spacecraft are not being supported. The expression for $E_r$ is

$$E_r = \sum_{k=r+1}^{n} (k-r)\alpha_k, \qquad \text{for } 1 \leq r < n \tag{6}$$

Here $k - r$ is the number of visible spacecraft not being tracked. Using Eq. (5), this becomes

$$E_r = \sum_{k=r+1}^{n} (k-r)\binom{n}{k}\beta^k(1-\beta)^{n-k}, \text{ for } 1 \leq r \leq n \tag{7}$$

This sum can also start at $k = r$.

As a check, when $r = 0$ (no antennas), $E_r = n\beta$ (the total number of spacecraft days visible). Equation (7) gives

$$E_0 = \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!}\beta^k(1-\beta)^{n-k}$$

$$= \sum_{k=1}^{n}\frac{n(n-1)!}{(k-1)![n-1-(k-1)]!}\beta^k(1-\beta)^{[n-1-(k-1)]}$$

$$= n\beta\sum_{l=0}^{n-1}\binom{n-1}{l}\beta^l(1-\beta)^{n-1-l}$$

$$= n\beta[\beta + (1-\beta)]^{n-1}$$

$$= n\beta$$

where the sum above is evaluated by the binomial theorem. Therefore, this checks.

The expected fraction of requirements not supported, $F_r$, is merely

$$F_r = \frac{E_r}{n\beta} \tag{8}$$

The expected number of antennas that are idle when $r$ antennas are installed at a Complex, $I_r$, is easily found, too

$$I_r = \sum_{i=0}^{r-1}(r-i)\alpha_i \tag{9}$$

Here $r-i$ is the number of antennas that are idle when $i$ spacecraft are visible, where $0 \leq i \leq r$. Using Eq. (5), this becomes

$$I_r = \sum_{i=0}^{r-1}(r-i)\binom{n}{i}\beta^i(1-\beta)^{n-i} \tag{10}$$

The above sum can, of course, be run up to $r$.

As a check, the average number of antennas in use, $r - I_r$, plus the average number $E_r$ of spacecraft visible but not being tracked, must equal $n\beta$, the number of spacecraft days

$$\sum_{i=0}^{r} i\alpha_i + r\sum_{i=r+1}^{n}\alpha_i + \sum_{i=r+1}^{n}(i-r)\alpha_i = \sum_{i=0}^{n} i\binom{n}{i}\beta^i(1-\beta)^{n-i}$$

The latter sum is the expected number of heads $i$ in $n$ independent coin flips when the probability of a head is $\beta$. This expectation is of course $n\beta$, which checks. This identity is useful in computation, so it is stated below:

$$r - I_r + E_r = n\beta \tag{11}$$

## IV. Performance Metrics

Let, for example, $r = n-1$, i.e., there is one less antenna than spacecraft. Equation (7) becomes

$$E_{n-1} = \beta^n$$

If $n = 4$ spacecraft and $r = 3$ antennas, while $\beta = 1/2$ (12-hour passes), it follows that $E_3 = (1/2)^4 = 1/16$. The average fraction of requirements not supported is $E_r/n\beta = E_3/(4\times1/2) = E_3/2 = 1/32$, or 3-percent unsupported spacecraft on average. To state this another way, to achieve this 97-percent support, an $I_3$ must be tolerated, from Eq. (10), of

$$I_3 = \sum_{i=0}^{2}(3-i)\binom{4}{i}\cdot\frac{1}{16} = \frac{3}{16} + \frac{2\cdot 4}{16} + \frac{1\cdot 6}{16} = \frac{17}{16}$$

On average, 1-1/16 antennas (out of three) must stand idle to provide 97-percent coverage of four spacecraft, each of which has 12-hour passes. This is true even noting that on average only two of the four spacecraft are visible and yet there are three antennas. The facilities' utilization is only $[3-(17/16)]/3 = 31/48 = 64.6$ percent. On average, one has to tolerate a 64.6-percent facilities' utilization in tracking time to provide 97-percent support to four spacecraft with 12-hour passes per day, each of which must be fully supported.

Table 1 presents and Fig. 2 graphs for $\beta = 1/2$ (12-hour passes) the fraction of tracking requirements not met $E_r/n\beta$ and the antenna-idle fraction $I_r/r$ as a function of $r$ for $n = 3$ to 7 spacecraft, $r$ going from 1 to 7. This shows what price has to be paid in apparently idle facilities (ignoring maintenance, upgrades, and radio/radar astronomy, etc.) in order to meet a given fraction of continuous coverage requirements. The idle capacity curves are almost linear. It is clear that there does need to be some apparent surplus capacity in the Network to achieve good coverage.

## V. Comparison With History

An experiment was performed using view period data for 1990 and $\beta = 1/3$ (8-hour passes). Specifically, a random day of year was picked (Greenwich Mean Time, GMT, day 191, July 10, 1990), and the visibility of $n = 9$ particular deep-space spacecraft (actually, eight spacecraft and one planet) over Goldstone was found. The eight deep-space spacecraft were Pioneers 10 and 11, Voyagers 1 and 2, Magellan, Galileo, the International Comet Explorer (ICE), and Giotto; the planet was Saturn. All passes would have been longer than 8 hours, but the 8-hour interval (perhaps involving the days before or after) centered at maximum elevation was used because the case being considered was $\beta = 1/3$. Pioneer 12 (Venus Orbiter) was not included because it orbits Venus, while Magellan was nearly in Venus orbit (insertion on August 10, 1990). This situation was not "random." In fact, the view periods for Giotto and Saturn were accidentally virtually identical.

Also, there was a much larger time with no spacecraft visible than had been expected. All indications are that these targets bunched.

The expected fraction of time $\alpha_k$ of the 24 hours during which $k$ spacecraft would be visible according to Eq. (5) is presented for $0 \leq k \leq 9$ in Table 2, together with the actual fraction $\hat{\alpha}_k$ as determined from the DSN view-period database evaluated for July 10, 1990, by the TDA Mission Support and DSN Operations Office. The raw-data view-period midpoints are presented in Table 3. The results are not far off from the expected values. This is especially noteworthy considering two facts. First, if there were a time when all nine spacecraft were visible, then there would be at least 8 hours when no spacecraft at all was visible: $\hat{\alpha}_9 > 0 \Rightarrow \hat{\alpha}_0 > 1/3$. Second, the view periods as shown above were quite bunched, as observed above. The three antennas actually covered 76.9 percent of the requirements on that day, compared with the expected 82.6 percent. Thus, the fraction obtained is quite robust with respect to the randomness assumption. This fraction is calculated from Eqs. (6) and (8), which give the expected fraction supported by $r = 3$ antennas with nine spacecraft, $\beta = 1/3$, as $1 - F_r = 1 - E_r/(n\beta) = 1 - E_3/3 = 1 - 0.174 = 0.826$.

## VI. Summary

The article has presented a model for the number of antennas needed to meet various fractions of full-coverage requirements for various numbers of spacecraft with view periods of random phase during the day. The trade-off between idle antennas and fractional tracking requirements met was clearly shown. More requirements met translates into more idle facilities. The model can be used to help calibrate the adequacy of facilities' plans in the longer term when mission sets are not so certain. It can also provide a zero calibration for the fraction of idle facilities in the existing Network. It seems robust with respect to the assumption of independent view periods.

## Acknowledgment

The author thanks Gene Winn of the TDA Mission Support and DSN Operations Office for obtaining the view period data used in Section V with the assistance of the Bendix Field Engineering Corporation.

Table 1. Expected requirements not met and expected idle fraction versus expected number of antennas, 12-hour passes

| No. of spacecraft, $n$ | No. of antennas, $r$ | $E_r/n\beta$ | $I_r/r$ |
|---|---|---|---|
| 3 | 1 | 0.4167 | 0.1250 |
| 3 | 2 | 0.0833 | 0.3125 |
| 3 | 3 | 0 | 0.5000 |
| 4 | 1 | 0.5313 | 0.0625 |
| 4 | 2 | 0.1875 | 0.1875 |
| 4 | 3 | 0.0313 | 0.3490 |
| 4 | 4 | 0 | 0.5000 |
| 5 | 1 | 0.6125 | 0.0313 |
| 5 | 2 | 0.2875 | 0.1094 |
| 5 | 3 | 0.0875 | 0.2396 |
| 5 | 4 | 0.0125 | 0.3828 |
| 5 | 5 | 0 | 0.5000 |
| 6 | 1 | 0.6719 | 0.0156 |
| 6 | 2 | 0.3750 | 0.0625 |
| 6 | 3 | 0.1563 | 0.1563 |
| 6 | 4 | 0.0417 | 0.2813 |
| 6 | 5 | 0.0052 | 0.4031 |
| 6 | 6 | 0 | 0.5000 |
| 7 | 1 | 0.7165 | 0.0078 |
| 7 | 2 | 0.4487 | 0.0352 |
| 7 | 3 | 0.2277 | 0.0990 |
| 7 | 4 | 0.0848 | 0.1992 |
| 7 | 5 | 0.0201 | 0.3141 |
| 7 | 6 | 0.0022 | 0.4180 |
| 7 | 7 | 0 | 0.5000 |

Table 2. The fraction of time that zero through nine spacecraft are visible, 8-hour passes ($\beta$ = 1/3)

| Expected from random model | | Observed on one particular day | |
|---|---|---|---|
| No. of spacecraft, $k$ | Fraction $\alpha_k$ of time $k$ visible | No. of spacecraft, $k$ | Fraction of $\hat{\alpha}_k$ of time $k$ visible |
| 0 | 0.0260 | 0 | 0.1382 |
| 1 | 0.1171 | 1 | 0.0681 |
| 2 | 0.2341 | 2 | 0.1431 |
| 3 | 0.2731 | 3 | 0.1542 |
| 4 | 0.2048 | 4 | 0.2993 |
| 5 | 0.1024 | 5 | 0.1972 |
| 6 | 0.0341 | 6 | 0 |
| 7 | 0.0073 | 7 | 0 |
| 8 | 0.0009 | 8 | 0 |
| 9 | 5 x $10^{-5}$ | 9 | 0 |
| 6–9 | 0.0424 | 6–9 | 0 |

**Table 3. Nine view periods, July 10, 1990, Goldstone (8-hour passes centered on maximum elevation, DSS 14)**

Pass durations

| Spacecraft or planet | Start of pass, GMT | Midpoint, GMT | End of pass, GMT |
|---|---|---|---|
| Voyager 1 | 00:56 | 04:56 | 08:56 |
| Pioneer 11 | 02:14 | 06:14 | 10:14 |
| Voyager 2 | 03:29 | 07:29 | 11:29 |
| Giotto | 04:11 | 08:11 | 12:11 |
| Saturn (planet) | 04:12 | 08:12 | 12:12 |
| Galileo | 10:36 | 14:36 | 18:36 |
| ICE | 12:18 | 16:18 | 20:18 |
| Pioneer 10 | 13:23 | 17:23 | 21:23 |
| Magellan | 13:37 | 17:37 | 21:37 |

Visibility durations and spacecraft tracked

| Time period, GMT | Duration, hr | No. of spacecraft visible | No. of spacecraft tracked | No. of spacecraft not tracked |
|---|---|---|---|---|
| 21:37–00:56 | 3:19 | 0 | 0 | 0 |
| 00:56–02:14 | 1:18 | 1 | 1 | 0 |
| 02:14–03:29 | 1:15 | 2 | 2 | 0 |
| 03:24–04:11 | 0:42 | 3 | 3 | 0 |
| 04:11–04:12 | 0:01 | 4 | 3 | 1 |
| 04:12–08:56 | 4:44 | 5 | 3 | 2 |
| 08:56–10:14 | 1:18 | 4 | 3 | 1 |
| 10:14–10:36 | 0:22 | 3 | 3 | 0 |
| 10:36–11:29 | 0:53 | 4 | 3 | 1 |
| 11:29–12:11 | 0:42 | 3 | 3 | 0 |
| 12:11–12:12 | 0:01 | 2 | 2 | 0 |
| 12:12–12:18 | 0:06 | 1 | 1 | 0 |
| 12:18–13:23 | 1:05 | 2 | 2 | 0 |
| 13:23–13:37 | 0:14 | 3 | 3 | 0 |
| 13:37–18:36 | 4:59 | 4 | 3 | 1 |
| 18:36–20:18 | 1:42 | 3 | 3 | 0 |
| 20:18–21:23 | 1:05 | 2 | 2 | 0 |
| 21:23–21:37 | 0:14 | 1 | 1 | 0 |

Recapitulation[a]

| $k$ | Total minutes | Fraction $\hat{\alpha}_k$[b] |
|---|---|---|
| 0 | 199 | 0.1382 |
| 1 | 98 | 0.0681 |
| 2 | 206 | 0.1431 |
| 3 | 222 | 0.1542 |
| 4 | 431 | 0.2993 |
| 5 | 284 | 0.1972 |
| 6–9 | 0 | 0 |

[a] Average number of spacecraft not tracked = 111/160 = 0.694. Fraction of requirements not met = 111/(3 x 160) = 37/160 = 23.1 percent. Fraction of requirements met = 76.9 percent.

[b] Actual fraction with $k$ spacecraft, $\alpha_k$.

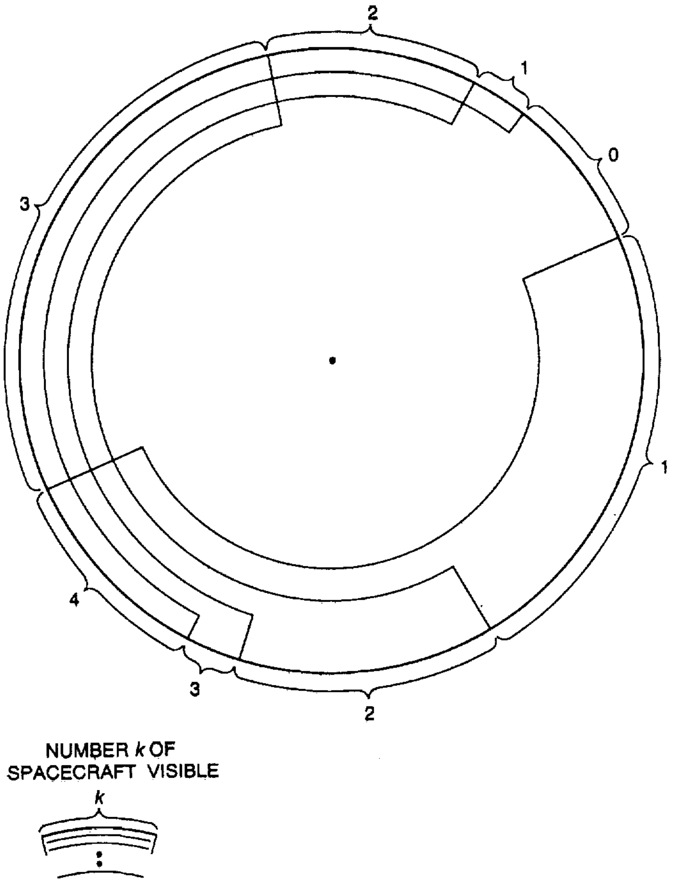


**Fig. 1. Four spacecraft tracked around the clock, 12-hour passes.**

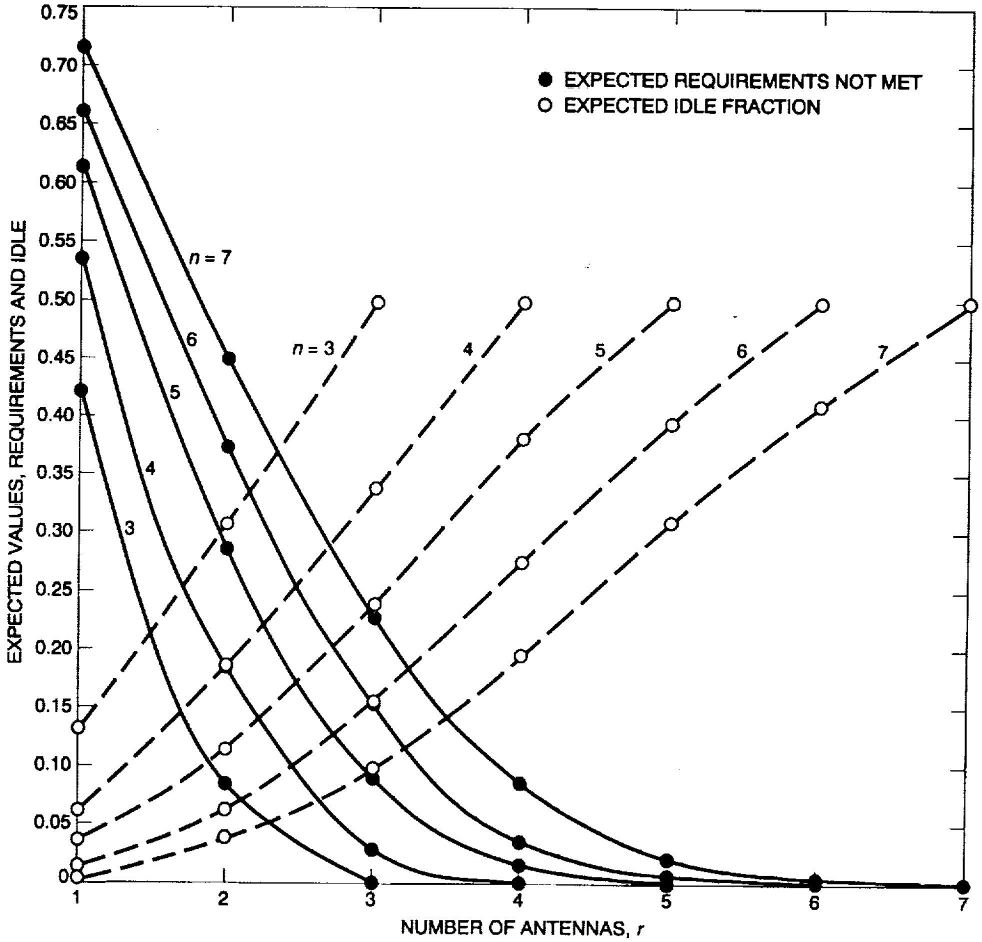


Fig. 2. Requirements not met and idle capacity versus number of antennas.

5/2-13

53215

P- 19

TDA Progress Report 42-107

November 15, 1991

N92-14249

# Trajectory and Navigation System Design for Robotic and Piloted Missions to Mars

S. W. Thurman
Navigation Systems Section

S. E. Matousek
Mission Design Section

*Future Mars exploration missions, both robotic and piloted, may utilize Earth-to-Mars transfer trajectories that are significantly different from one another, depending upon the type of mission being flown and the time period during which the flight takes place. The use of new or emerging technologies for future missions to Mars, such as aerobraking and nuclear rocket propulsion, may yield navigation requirements that are much more stringent than those of past robotic missions, and are very difficult to meet for some trajectories. This article explores the interdependencies between the properties of direct Earth-to-Mars trajectories and the Mars approach navigation accuracy that can be achieved using different radio metric data types, such as ranging measurements between an approaching spacecraft and Mars-orbiting relay satellites, or Earth-based measurements such as coherent Doppler and very long baseline interferometry. The trajectory characteristics affecting navigation performance are identified, and the variations in accuracy that might be experienced over the range of different Mars approach trajectories are discussed. The results predict that three-sigma periapsis altitude navigation uncertainties of 2 to 10 km can be achieved when a Mars-orbiting satellite is used as a navigation aid.*

## I. Introduction

The exploration of Mars to date has been accomplished by unmanned spacecraft using low-energy ballistic transfer trajectories to reach their destinations. NASA's ambitious plans for future Mars exploration call for a variety of robotic and piloted spacecraft to investigate the Red Planet from orbit and on its surface. These missions may employ new technologies, such as nuclear rocket propulsion, which make it possible to send large payloads to Mars along high-energy trajectories that are inaccessible to current chemically propelled launch-vehicle/upper-stage combinations. Another concept receiving serious consideration is aerobraking, in which a spacecraft executes a controlled passage through the Martian atmosphere to decelerate into a closed orbit or to initiate a descent to the surface of the planet. Aerobraking can also be employed by a spacecraft already orbiting Mars to modify its orbit.

The successful use of aerobraking for orbit insertion (called aerocapture) or direct entry and landing may require approach navigation accuracies that are much more stringent than those typically needed to support a propulsive orbit insertion, depending upon the target orbit (or landing point) and the characteristics of the aerobrake vehicle itself. For example, previous studies of navigation requirements for Mars aerocapture have found that aerocapture vehicles of moderate maneuver capability (maximum lift-to-drag ratios of 0.5 to 0.7) must be delivered to within 5 to 20 km in altitude and 30 to 50 km along the flight path (downtrack) at the nominal atmospheric entry point, which typically occurs just prior to closest approach [1,2]. This is in contrast to an altitude delivery requirement at closest approach of about 300 km for the Mars Observer mission, which will perform a propulsive orbit insertion.[1] Spacecraft using high-thrust nuclear propulsion (nuclear-thermal rocket engines, in which a solid or gaseous core reactor is used to heat a working fluid such as hydrogen) will probably also require greater delivery accuracies than Mars-Observer–class missions, as they may possess Mars approach velocities of up to 10 km/sec [3]; in contrast, this figure will be about 2.5 km/sec for Mars Observer.

Several studies have analyzed the performance of different pre-aerocapture approach navigation schemes at Mars [1–2,4–6]. These studies, which have addressed a relatively small subset of possible Mars approach trajectories, investigated radio and optical data that provide direct measurements of the Mars-relative trajectory of an approaching spacecraft: spacecraft onboard optical imaging of Martian moons, ranging measurements between a Mars-orbiting communications relay satellite and the approaching spacecraft, and Earth-based dual-spacecraft radio interferometry, again using a Mars relay satellite in conjunction with the approaching spacecraft. This article describes a preliminary assessment of the impact of different approach trajectories, arising from different types of direct Earth–Mars transfer trajectories, on the performance of the radio navigation schemes involving a Mars relay satellite listed above, focusing on the performance needed to support the use of aerobraking. The guidance accuracy that can be achieved by modern robotic spacecraft is also investigated briefly to provide some indication of the relative importance of guidance errors versus orbit determination errors in Mars approach navigation system design.

## II. Direct Transfer Trajectories

There are many different trajectories that can be used to reach Mars from the Earth. Current launch vehicle capabilities limit the available trajectories to those with reasonable launch energies for spacecraft of modest (less than about 6000 kg) mass. Avoidance of excessive transit time generally limits the range of possible trajectories to those that take less than one full revolution around the Sun. After these initial constraints are taken into account, direct transfer trajectories known as type-1 (transfer angle between 0 and 180 deg) or type-2 (transfer angle between 180 and 360 deg) are left to consider. Transfers that involve a flyby of Venus after launch from Earth are also possible, but these trajectories are beyond the scope of this study [7].

### A. Trajectory Characteristics

For a given launch opportunity, either a type-1 or a type-2 trajectory can be selected. Type-1 trajectories generally have shorter transit times than type-2 trajectories; however, in most cases type-1 trajectories also require a higher launch energy than type-2 trajectories for a given launch opportunity. Within the general categories of type-1 or -2 lie other trajectory options. One obvious choice is to optimize for a minimum launch energy (generally defined in terms of the parameter called $C_3$, with units of km$^2$/sec$^2$, or hyperbolic excess launch velocity, $V_\infty$, which is equal to $\sqrt{C_3}$). Another alternative is to minimize the arrival velocity at Mars. Unfortunately, trajectories optimized for minimum launch energy have greater arrival velocities at Mars than trajectories optimized for minimum Mars arrival velocity. Conversely, trajectories optimized for minimum arrival velocity at Mars possess larger launch energies at Earth than trajectories optimized for minimum launch energy.

Figure 1 shows four possible trajectories for the 1998 launch opportunity. Each of these trajectories corresponds to a type-1 or type-2 transfer, further subdivided into a minimum launch energy case and a minimum arrival velocity case. While trajectories for different launch opportunities would, by necessity, differ somewhat from those shown in Fig. 1, the basic appearance of the different trajectory types relative to each other would not change significantly. Figure 2 depicts the Mars arrival geometry corresponding to the type-1 minimum launch energy trajectory, assuming

[1] P. B. Esposito, *Mars Observer Navigation Plan*, JPL D-3820, Rev. C (internal document), Jet Propulsion Laboratory, Pasadena, California, June 5, 1990.

a polar circular target orbit with an altitude of 700 km. Figure 3 depicts the Mars arrival geometry for the type-2 minimum launch energy trajectory, assuming that the target orbit is the same as that in Fig. 2. In both Figs. 2 and 3, the frames showing the view from "above" the ecliptic plane represent how the trajectory would look when viewed from the ecliptic north pole looking south at the ecliptic plane, the mean plane of the Earth's orbit. Also in Figs. 2 and 3, note that the angle between the Earth–Mars radial line and the type-1 (Fig. 2) incoming trajectory is significantly different from the type-2 (Fig. 3) trajectory.

### B. Trajectory Design Issues

There are many factors that can influence the selection of a particular Earth–Mars trajectory. However, it is possible to single out a few major constraints that affect trajectory selection and, consequently, Mars approach navigation performance. The first, and most obvious, constraint on a trajectory is that it must deliver a spacecraft to Mars. The energy imparted from the launch vehicle system (launch vehicle, upper stages, and any additional boost stages) to the spacecraft at injection must match the spacecraft velocity to the velocity required at a particular point in space to follow a given transfer trajectory. There are many different options for meeting this constraint [8, 9]. For robotic missions, the trajectory design process generally consists of evaluating trade-offs between minimizing the launch energy, and hence the injection velocity, and the Mars arrival velocity, subject to criteria derived from the mission objectives. For piloted missions, this process is further complicated by the additional constraint that a return leg is also needed, therefore the launch energy from Mars and the arrival velocity at Earth for a return trip must also be considered along with the corresponding parameters for the Earth–Mars trajectory (Soldner [5] describes the round-trip trajectory design problem for piloted missions).

Table 1 summarizes the range of launch energies and Mars arrival velocities for minimum launch energy and minimum arrival energy type-1 and type-2 trajectories, obtained from an analysis of Mars launch opportunities between 1995 and 2020. Nuclear-rocket-propelled spacecraft may be able to utilize fast "sprint"- trajectories, which are type-1-class trajectories with larger launch energies and arrival velocities than the optimized trajectories given in Table 1.

The range of launch azimuths available from a particular launch site is another constraint that must be considered. The launch-azimuth constraints for the Kennedy Space Center (KSC) are shown in Fig. 4. Range safety considerations call for a launch trajectory over water for the early part of the flight (it should be noted that not all of the allowable azimuths shown in Fig. 4 are necessarily available because of islands in certain areas of the allowable envelope). From KSC, the available range of launch azimuths effectively restricts the injection asymptote declination (the inclination of the injection velocity vector relative to the Earth's equatorial plane) to the range from about −53 to +53 deg. This restriction can, in turn, make it very difficult to achieve the injection velocity vector required to utilize some direct Earth–Mars transfers for certain launch opportunities. For example, the type-1 minimum launch energy trajectory for the 2001 launch opportunity requires an injection asymptote declination of 54 deg, which is unreachable from KSC with current U. S. launch vehicles because of launch azimuth constraints. Even if the required injection asymptote declination can be reached, trajectories with large declination m̃agnitudes generally require greater launch energies from a near-equatorial launch site such as KSC, effectively reducing the available payload mass for the mission. Other mission design considerations such as the length of the daily launch window, desired Mars arrival geometry, and target orbit influence the transfer trajectory design process as well [9].

## III. Navigation Accuracy Analysis

The objective of approach-phase navigation is to deliver a spacecraft to a chosen aim point at a desired time. The navigation system for this task may consist of many different physical elements, located both on the spacecraft and on the Earth, but it must perform two primary functions, regardless of the means employed: orbit determination, which is the process of determining the current and predicted future flight path of a spacecraft, and guidance (maneuver analysis/design), which is the process of planning and executing trajectory correction maneuvers (TCMs) that will remove known deviations of the spacecraft from the intended flight path and will satisfy other mission constraints. The overall navigation, or delivery, accuracy achieved by the end-to-end navigation system depends on the accuracy to which both the orbit determination and guidance functions are performed.

### A. Maneuver Analysis

Guidance for interplanetary spacecraft is normally carried out using propulsive maneuvers of short duration for flight path control (the exception being spacecraft employing low-thrust nuclear or solar-electric propulsion systems that may operate continuously for extended periods) to achieve a desired close flyby of a target body or to decelerate into a closed orbit upon arrival. In this section, approximate estimates of the navigation uncertainty due to guidance errors are developed using propulsion-system performance data representative of modern robotic spacecraft. In subsequent sections, approximate orbit-determination accuracy estimates are developed and used along with the guidance-error estimates to compute statistics for the overall navigation altitude error at the vacuum periapsis point (closest approach).

Maneuver calculations are most often performed using an asymptotic, or "$B$-plane," coordinate system, defined in Fig. 5. The origin of this system is the center of mass of the target planet. The $B$-plane coordinates describing the trajectory are defined in terms of the orthogonal unit vectors $\hat{S}$, $\hat{T}$, and $\hat{R}$. $\hat{S}$ is parallel to the incoming asymptote of the approach hyperbola, while $\hat{T}$ usually lies in either the ecliptic plane or the equatorial plane of the target body; $\hat{R}$ completes the triad. The aim point is defined by $\vec{B}$, known as the "miss" vector, and the desired arrival time, which is expressed in terms of the linearized time of flight (LTOF), is defined as the time before closest approach, if it is assumed that the miss vector has zero magnitude. Both maneuver-execution errors and orbit-determination errors are normally characterized by a three-sigma $B$-plane dispersion ellipse, shown in Fig. 5, and the three-sigma uncertainty in linearized time of flight. In Fig. 5, SMAA is the semimajor axis of the dispersion ellipse, while SMIA is the semiminor axis of the dispersion ellipse.

During a mission, the miss vector and linearized time of flight corresponding to a spacecraft's actual trajectory are estimated repeatedly during the orbit-determination process and compared with their desired values. If the current estimated aim point is sufficiently removed from the desired aim point, then a TCM must be performed at some point to remove this deviation. The placement and design of TCMs must take into account a great many considerations; these have been described in much greater detail than can be given here by Hintz and Chadwick [10, 11].

For roughly the final 10 to 14 days before encounter, a spacecraft approaching Mars will have a nearly constant velocity with respect to the planet, directed along the Mars–spacecraft radial line, until it is within 12 to 24 hr of periapsis (closest approach) [8]. During this period, small changes in the $B$-plane coordinates resulting from a small, instantaneous spacecraft velocity change (an excellent approximation for most TCMs) vary roughly linearly with time. This relationship can be expressed as

$$\Delta\vec{B} = \mathbf{K}\Delta\vec{v} \tag{1}$$

where

$$\Delta\vec{B} = [\Delta B{\cdot}T,\ \Delta B{\cdot}R, \Delta LTOF]^T$$

$$\Delta\vec{v} = [\Delta v_T, \Delta v_R, \Delta v_S]^T$$

$$\mathbf{K} \approx \begin{bmatrix} t & 0 & 0 \\ 0 & t & 0 \\ 0 & 0 & t/V_\infty \end{bmatrix}$$

and

$\Delta B{\cdot}T, \Delta B{\cdot}R$ = changes in $T$ and $R$ components of $\vec{B}$, respectively

$\Delta LTOF$ = change in linearized time of flight

$\Delta v_T, \Delta v_R, \Delta v_S$ = $T, R$, and $S$ velocity increments

$t$ = time to go before closest approach

$V_\infty$ = hyperbolic approach velocity

The approximation given for the $\mathbf{K}$ matrix in Eq. (1) effectively assumes that the target planet has no mass. It has been shown that for a small planet such as Mars, Eq. (1) is also a fairly good approximation (±20 percent) until roughly the final 12 to 24 hours of the approach phase [12].

In a typical robotic mission, a TCM to remove execution errors from earlier maneuvers will be scheduled at about 10 days prior to encounter. This point is near enough to encounter to effect fairly small changes in the aim point, but far enough out so that there is sufficient time to redetermine the orbit, and design and execute a final TCM at 1 to 2 days out, if necessary. The error covariance matrix for the $B$-plane coordinates prior to each maneuver is just the sum of the orbit-determination error covariance and the guidance-error covariance at the maneuver epoch, assuming that the orbit-determination errors and guidance errors are independent:

$$\mathbf{\Lambda}_{\Delta\vec{B}} = \mathbf{\Lambda}_{OD} + \mathbf{\Lambda}_{G} \tag{2}$$

where

$\mathbf{\Lambda}_{\Delta\vec{B}}$ = $B$-plane coordinate error covariance matrix

$\mathbf{\Lambda}_{OD}$ = $B$-plane coordinate orbit determination error covariance

$\Lambda_G$ = $B$-plane coordinate guidance error covariance

The guidance-error covariance reflects the $B$-plane coordinate uncertainty obtained upon completion of the previous maneuver. Equation (1) can be inverted to compute a maneuver to correct for a known aim-point error, $\Delta\vec{B}$:

$$\Delta\vec{v} = \mathbf{K}^{-1}\Delta\hat{\vec{B}} \tag{3}$$

where

$\Delta\hat{\vec{B}}$ = orbit-determination estimate of $\Delta\vec{B}$ at maneuver epoch

Since the maneuver computed using Eq. (3) must by necessity be based on an estimate of $\Delta\vec{B}$, it becomes apparent that the accuracy of the maneuver will be limited by orbit-determination accuracy. Hence, it is desirable that maneuvers be executed only when the orbit-determination uncertainty at the maneuver epoch is small relative to the size of the guidance errors to be removed from the trajectory.

After a maneuver, the $B$-plane coordinate error covariance, assuming that the orbit-determination errors and maneuver-execution errors are independent, is

$$\Lambda_{\Delta\vec{B}} = \Lambda_{OD} + \mathbf{K}\Lambda_E\mathbf{K}^T \tag{4}$$

where

$\Lambda_E$ = maneuver-execution error covariance

The $B$-plane coordinate covariance in Eq. (4) becomes the guidance-error covariance in Eq. (2) for the next maneuver. When a maneuver $\Delta\vec{v}$ is computed, the errors in the orbit-determination estimate of the trajectory at that time result in an erroneous computation; hence, orbit-determination errors are effectively translated into guidance errors as each successive maneuver is performed. Maneuver-execution errors, caused by imperfect execution of the planned maneuver, are typically broken into fixed errors and proportional errors, both in $\Delta\vec{v}$ magnitude and direction. Representative three-sigma values for large robotic spacecraft such as Galileo and Cassini are about 1.0 mm/sec fixed magnitude and direction, 5.0 percent proportional magnitude, and 10.0 mrad/axis proportional direction.

## B. Guidance-Error Calculations

Using Eqs. (2), (3), and (4), approximate guidance (maneuver-execution error) dispersion ellipses can be calculated for TCMs performed near Mars. The results given below were computed for TCMs assumed to be located at 10 days (hereafter referred to as TCM1) and 1 day (hereafter referred to as TCM2) prior to encounter, respectively.

The $B$-plane coordinate error covariance prior to TCM1 must be specified to determine the expected magnitude of this maneuver. The guidance-error covariance at this point was assumed to be a spherical-error ellipsoid, with a radius equal to 150 km, which is the semimajor axis (one-sigma) of the predicted $B$-plane dispersion ellipse for Mars Observer before its final TCM, 10 days prior to orbit insertion.[2] It should be noted here that the linearized time-of-flight uncertainty actually represents the position uncertainty in the $\hat{S}$ direction divided by the hyperbolic approach velocity, $V_\infty$; therefore, a spherical position uncertainty ellipse is easily converted into an appropriate $B$-plane coordinate covariance. The orbit-determination error covariance was also assumed to be spherical, with a radius (one-sigma) of 10 km. This figure is representative of anticipated Earth-based radio-only tracking performance about 10 years from now, and is based on the study performed by Konopliv and Wood [4].

To compute the magnitude statistics of TCM1, Eq. (3) must first be used to compute the $\Delta\vec{v}$ covariance at 10 days out, using the assumed $B$-plane coordinate covariance. From the $\Delta\vec{v}$ covariance, the expected value and standard deviation of the maneuver magnitude can be obtained from the Monte Carlo simulation data describing maneuver magnitude statistics presented by Bollman and Chadwick [13]. The expected magnitude of TCM1 was found to be 28 cm/sec, with a standard deviation of 12 cm/sec; the three-sigma magnitude is then 64 cm/sec. The largest three-sigma execution error component, using the three-sigma maneuver execution statistics given above, was found to be 3 cm/sec. This value was assumed to apply to all three components of TCM1, resulting in a post-TCM1, three-sigma, $B$-plane guidance dispersion ellipse that is circular, with a radius of about 26 km. The post-TCM1 guidance and orbit-determination $B$-plane dispersions (three-sigma) and the root-sum-square (RSS) navigation (orbit-determination errors plus guidance errors) three-sigma dispersion ellipse are shown in Fig. 6(a). The corresponding post-TCM1 total LTOF uncertainty (three-sigma) is equal to 40 km/$V_\infty$.

[2] Op. cit.

To compute the execution-error statistics for TCM2, the same process described above for TCM1 was repeated, with one modification. It was assumed that the orbit-determination $B$-plane covariance prior to TCM2 was small relative to the guidance-error covariance, which is just the post-TCM1 error covariance. It will be shown in the next section that this is a good assumption when a Mars-orbiting relay satellite is available as a navigation aid. The expected magnitude and standard deviation of TCM2 were then a function only of the post-TCM1 $B$-plane covariance. The three-sigma TCM2 magnitude was found to be 55 cm/sec. The largest maneuver-execution error component was again assumed to apply to all three spatial components of the maneuver, to construct a conservative estimate of the execution error dispersions. The three-sigma $B$-plane guidance dispersion ellipse for TCM2 is shown in Fig. 6(b). The corresponding three-sigma LTOF uncertainty was about 2.4 km/$V_\infty$. The calculation of the orbit-determination $B$-plane dispersions needed to compute statistics for the total post-TCM2 navigation error uncertainty is the subject of the next section.

## C. Orbit-Determination Analysis

To effectively support the final TCM (TCM2) before encounter, the errors in the trajectory solution used to compute the maneuver must be small relative to the guidance errors to be corrected, as discussed above. The orbit-determination errors at the time of TCM2 must also be small enough that the navigation errors at encounter, which include TCM2 maneuver-execution errors as well as orbit-determination errors, will not exceed the allowable requirements. Thus, after the first approach phase TCM (TCM1) is performed at 10 days out as assumed in this analysis, the approach trajectory must be redetermined accurately within 9 days, to support TCM2.

The Mars approach orbit-determination accuracy that can be achieved with conventional Earth-based radio metric data is fundamentally limited by errors in knowledge of the geocentric position and velocity of Mars itself, until the motion of the approaching spacecraft becomes dominated by the Martian gravity field. This does not occur until the last few hours or days before closest approach, depending on the approach trajectory characteristics. A spacecraft already orbiting Mars, since it is closely tied to the planet gravitationally, can be used as a radio navigation aid for an approaching spacecraft in two different ways. Ranging measurements between the two spacecraft have been shown to be potentially capable of determining the Mars-relative position and velocity of the approaching spacecraft to within a few kilometers and centimeters/second, respectively, although there exist tracking geometries that may yield significantly degraded performance [1,4,7]. Simultaneous tracking of a Mars orbiter or lander and an approaching spacecraft using Earth-based delta very long baseline interferometry (ΔVLBI), when used in conjunction with conventional Doppler and ranging data, has also been shown to be capable of similar accuracies [5]. In this section, the orbit-determination accuracy that can be obtained from both of these techniques is illustrated using approximate calculations of $B$-plane dispersions for short data arcs acquired near Mars.

**1. Spacecraft–Spacecraft Ranging.** The tracking geometry for spacecraft–spacecraft ranging measurements is depicted in Fig. 7. It has been shown previously [6] that the range observable, $\rho$, can be written simply as

$$\rho = r\left[1 - (2/r) r_s \cos\ \delta\ \cos\left(\alpha_s - \alpha\right) + r_s^2/r^2\right]^{1/2} \quad (5)$$

where

$r$ = distance from approach spacecraft to center of Mars

$\delta$ = spacecraft declination relative to satellite orbit plane

$\alpha$ = spacecraft right ascension in satellite orbit plane

$r_s$ = relay-satellite orbital radius

$\alpha_s$ = relay-satellite true anomaly

It can be seen from Eq. (5) that the range observable will enable a complete determination of the time histories of the approach spacecraft spherical coordinates relative to the satellite orbit plane. The ephemeris (position and velocity) of the relay satellite generally must also be estimated along with the approach spacecraft trajectory. To investigate the orbit-determination performance of spacecraft–spacecraft ranging, statistics associated with a weighted least-squares estimate of the $B$-plane coordinates describing the approach trajectory can be readily computed from the partial derivatives of Eq. (5) with respect to the approach trajectory and the relay-satellite orbit, and the error covariance assumed for the ranging data.

To compute the $B$-plane statistics, each ranging measurement, designated $z$, is assumed to consist of the actual range value and a zero-mean additive noise, $\nu$:

$$z = \rho + \nu \quad (6)$$

Small changes in a series of ranging measurements, $\Delta\vec{z}$, from range values computed using an a priori estimate of the approach trajectory, are related to small changes in the vector of estimated parameters, $\Delta\vec{x}$, from their a priori values through a linearized matrix equation:

$$\Delta\vec{z} = \mathbf{A}\Delta\vec{x} + \vec{v} \tag{7}$$

where

$$\mathbf{A} = \begin{bmatrix} \partial z_1/\partial\,\vec{x} \\ \partial z_2/\partial\,\vec{x} \\ \cdot \\ \cdot \\ \partial z_n/\partial\,\vec{x} \end{bmatrix}$$

For spacecraft–spacecraft ranging covariance analysis, the estimated parameters were the epoch $B$-plane parameters, the magnitude and orientation of the asymptotic approach velocity vector, the position and velocity of the relay satellite at epoch, and a range measurement bias and bias rate, for a total of 14 parameters, all of which are constants. The error covariance for a weighted least-squares estimate of $\vec{x}$, designated $\mathbf{\Lambda_x}$, is

$$\mathbf{\Lambda_x} = \left[\tilde{\mathbf{\Lambda}}_{\mathbf{x}}^{-1} + \mathbf{A}^T\mathbf{\Lambda}_\nu^{-1}\mathbf{A}\right]^{-1} \tag{8}$$

In Eq. (8), $\tilde{\mathbf{\Lambda}}_{\mathbf{x}}$ is the a priori error covariance for the estimated parameters, and $\mathbf{\Lambda}_\nu$ is the error covariance for the noise-induced range measurement errors.

The assumptions used in the "baseline" spacecraft–spacecraft ranging scenario are given in Table 2. As in the guidance error computations, the approach spacecraft was assumed to nominally move at a constant velocity relative to Mars, since this is a good approximation for the trajectory until very near encounter. The approach velocity, $V_\infty$, was chosen to be a midrange value, given the arrival velocity ranges from Table 1. The declination of the incoming velocity vector, given in Table 2 to be 20 deg, is defined with respect to the satellite orbit plane. The $\hat{T}$ axis (see Fig. 5), is taken here to lie in the Martian equatorial plane; therefore, by setting $\theta$ equal to zero, the miss vector lies in the Martian equatorial plane as well. In Table 2, the parameter $h_p$ is the periapsis altitude for the actual hyperbolic flight path, whose incoming asymptote is coincident with the constant velocity trajectory used for the analysis. This value of $h_p$ is representative of aerobraking approach trajectories used in previous studies [1,2].

$B$-plane dispersion ellipses calculated using three different values of range acquisition distance, the distance from Mars at which ranging data are first acquired, and two different values of the approach trajectory declination are shown in Fig. 8. In all cases shown in Fig. 8, the data cutoff point was assumed to be 24 hr prior to closest approach, and it was further assumed that the relay satellite was always in view of the approach spacecraft, so that ranging data were acquired continuously. Since data are taken up until the time of TCM2, it is implicitly assumed that the orbit-determination and TCM2 computations are performed onboard the approach spacecraft. The three-sigma LTOF uncertainty in all cases was less than 0.03 sec (equivalent to 120 m). These cases represent the performance that might be obtained with two-way ranging data. The range acquisition distance that can be achieved, which is seen in Fig. 8 to have a significant impact on orbit-determination performance, will depend upon the antenna sizes and transmitter power available on the two spacecraft, and the link frequency as well. The largest acquisition distance used, 2 million km, is reached about 5.8 days before encounter, while the minimum distance, 1 million km, is reached only 2.9 days from encounter.

In this analysis, the range measurement accuracy was assumed to vary linearly with the range between the two spacecraft (see Table 2); this behavior was found to be representative of a power-limited spacecraft–spacecraft ranging system in an earlier investigation [1]. It should be remembered that the relay-satellite ephemeris was estimated along with the trajectory of the approach spacecraft. The relay-satellite a priori position and velocity uncertainties given in Table 2 are representative of the level of accuracy that can typically be achieved using Earth-based Doppler tracking data. Since it takes time to estimate the relay-satellite ephemeris from the ranging data, along with the other estimated parameters, changes in the a priori relay-satellite covariance will affect the accuracy achieved for the approach spacecraft, especially in cases when the range acquisition distance is small and the data arc is thereby short in length.

In Fig. 8, note than the $B$-plane dispersion ellipses for the small declination cases, Fig. 8(b), are much larger than those for the corresponding cases in which baseline declination from Table 2 was used, Fig. 8(a). When the magnitude of the declination angle ($\delta$) is small, range measurements are relatively insensitive to small changes in declination, which in Fig. 8 corresponds roughly to the $\hat{R}$ direction. This behavior can be illustrated by taking the partial derivative of Eq. (5) with respect to $\delta$, which to first order is the sensitivity of range to a small change in $\delta$. If it is assumed that $r_s/r \ll 1$ (a good assumption until the last one or two days before closest approach), this partial derivative is approximately

$$\partial\rho/\partial\delta \sim r_s \sin \delta \cos(\alpha_s - \alpha) \tag{9}$$

From Eq. (9), it is apparent that when $\delta$ is small, it will be difficult to accurately determine the declination angle (and hence its rate of change as well) from ranging data.

**2. Earth-Based Doppler and Dual-Spacecraft Interferometry.** Earth-based VLBI tracking of a Mars orbiter or lander and a spacecraft approaching Mars provides a direct measure of the Mars-relative approach trajectory, without requiring any communication between the two spacecraft. A detailed description of the dual-spacecraft VLBI measurement technique and the error sources affecting this data type is given by Edwards, Folkner, Border, and Wood [5]. Two-way (coherent) Doppler tracking of the approach spacecraft can to some degree sense the Mars-relative spacecraft trajectory, but only when the spacecraft is within the gravitational influence of Mars, which does not occur until the last few hours or days prior to closest approach. Since Doppler data sense the spacecraft motion along the Earth–spacecraft line of sight, and VLBI data sense primarily the motion perpendicular to the line of sight, these two data types provide complementary information when used together.

The information content of Doppler data acquired during the planetary approach phase has been described by Bollman [14]. A dual-spacecraft VLBI observation, illustrated in Fig. 9, consists of the time delay of radio signals observed by two stations; the radio signals from one spacecraft are differenced with the time delay from the other spacecraft as observed by the same two stations. As mentioned earlier, the trajectory of a Mars orbiter or lander with respect to Mars can be accurately determined from Earth-based tracking data, since it is gravitationally (or physically in the case of a lander) tied to Mars. Assuming the position of one of the two spacecraft is well known with respect to Mars, the dual-spacecraft VLBI observable, $\Delta\tau$, is approximately

$$\Delta\tau \approx \frac{\vec{r}_B}{c} \cdot (\vec{r} - \vec{r}_p) \tag{10}$$

where

$\vec{r}_B$ = baseline vector between the two participating stations

$\vec{r}$ = unit vector pointing toward approach spacecraft

$\vec{r}_p$ = unit vector pointing toward Mars

$c$ = speed of light

During roughly the final two weeks before encounter, Eq. (10) becomes very nearly a function of the Mars-relative spacecraft position and the Earth baseline only:

$$\Delta\tau \approx \left(\frac{1}{c}\right) \frac{\vec{r}_B}{r} \cdot \vec{R}_{s/p} \tag{11}$$

where

$\vec{R}_{s/p}$ = approach spacecraft position with respect to Mars

$r$ = approach spacecraft distance from Earth

From Eq. (11), it can be seen that the precision of the dual-spacecraft VLBI observable is directly proportional to the length of the baseline and inversely proportional to the Earth–spacecraft distance.

The assumptions used for calculating Doppler/dual-spacecraft VLBI orbit determination performance are given in Table 3. The trajectory parameters used ($V_\infty, \theta, h_p$) were the same as those for the spacecraft–spacecraft ranging cases (see Table 2). The estimated parameters were the $B$-plane coordinates and the arrival velocity vector components, a total of six in all. Calculations were performed for viewing geometries corresponding to two different Mars approach trajectories, representing type-1 and type-2 minimum launch energy transfers for the 1998 launch opportunity, respectively. The encounter geometries for these two cases are those shown in Figs. 2 and 3. The Mars relay-satellite used for acquiring dual-spacecraft VLBI data was assumed to have an ephemeris uncertainty of 2.0 km (one-sigma, each component), which was treated as a random error affecting the data. In Table 3, the dual-spacecraft VLBI measurement uncertainties are given in units of distance (cm) instead of units of time, since the observable, Eq. (10), can be viewed as a measure of distance simply by removing the factor $1/c$. The measurement accuracy assumed for the dual-spacecraft VLBI data was that given by Edwards for observations made at X-band (8.4-GHz) frequencies [5]. Since Earth-based data would likely be processed on Earth, the data cutoff point was assumed to be 36 hr from encounter, allowing 12 hr for the ground processing needed for orbit determination and computation of the TCM2 maneuver at 24 hr from encounter.

$B$-plane and LTOF dispersions for three different dual-spacecraft VLBI data sets are shown in Fig. 10. The performance in the "baseline" cases, which include dual-spacecraft VLBI data acquired from two baselines formed by the DSN complexes near Goldstone, California; Madrid, Spain; and Canberra, Australia, is seen to be significantly better than that obtained when only one of the two DSN baselines is used. These results raise the question of whether spacecraft near Mars can be viewed from both the DSN Goldstone–Canberra and Goldstone–Madrid baselines for all possible Mars encounter dates. Figure 11 illustrates the overlap regions in which different portions of the celestial sphere can be viewed simultaneously from different pairs of DSN complexes. In Fig. 11, spacecraft declination is referred to the Earth's equatorial plane. Mars encounter declinations range from about −25.5 deg to +25.5 deg; for low-declination encounters, it can be seen from Fig. 11 that the Goldstone–Madrid baseline may not be able to view Mars and its vicinity. In fact, minimum elevation restrictions limit the lowest declination angle that can be effectively observed simultaneously by Goldstone and Madrid to about −20 deg.

### D. Total Periapsis Altitude Navigation Error

This section presents the three-sigma periapsis altitude uncertainties that could be obtained using the hypothetical guidance and orbit-determination scenarios developed in the previous sections. The statistics of the altitude error at periapsis can be readily calculated from the total navigation $B$-plane error covariance, consisting of orbit-determination and guidance-error statistics, at completion of the final trajectory correction maneuver. As stated in the Introduction, the periapsis altitude error that can be tolerated by aerobrake vehicles possessing moderate (0.5 to 0.7) lift-to-drag ratios is between 5 and 20 km, depending upon the target orbit; this requirement is much more stringent than the periapsis downtrack error requirement (30 to 50 km) for these vehicles, and will therefore be the focus of the remaining discussion.

The magnitude of the miss vector, $|\vec{B}|$, is related to the periapsis radius, $r_p$, through the following formula from two-body orbital mechanics:

$$|\vec{B}| = r_p\sqrt{1 + (2\mu)/(r_p V_\infty^2)} \qquad (12)$$

In Eq. (12), $\mu$ is the gravitational parameter of the target body. To first order, small errors in $r_p$ due to errors in $|\vec{B}|$ can be expressed through the partial derivative of Eq. (12), yielding

$$\Delta r_p \approx \frac{|\vec{B}|}{r_p + (\mu/V_\infty^2)} \Delta |\vec{B}| \qquad (13)$$

From Fig. 6(b), the three-sigma uncertainty in $\vec{B}$ due to maneuver execution errors in the final TCM is about 2.4 km. For a nominal periapsis altitude of 20 km ($r_p$ = 3417 km), this results in a three-sigma altitude uncertainty ranging from 1.95 km for an arrival velocity of 3.0 km/sec to 2.39 km for an arrival velocity of 10.0 km/sec. This guidance component of the altitude error represents the lower bound for the total navigation error. In looking at the orbit determination $B$-plane dispersions in Figs. 8 and 10, it can be seen that in most cases the guidance errors are small relative to the orbit-determination errors.

The previous section showed that the orbit-determination performance of spacecraft–spacecraft ranging varies with the declination of the approach trajectory with respect to the relay-satellite orbit plane and the maximum distance over which ranging data can be acquired. To investigate the sensitivity of the total altitude navigation error at periapsis to changes in $V_\infty$ using spacecraft–spacecraft ranging for orbit determination, three-sigma altitude uncertainties were calculated over a range of $V_\infty$ values for two different values of acquisition range. The error modeling assumptions used were those given in Table 2. The results are shown in Fig. 12; the minimum value of altitude uncertainty is about 2 km, which is primarily due to the guidance error component of the total navigation error. The calculations were repeated for circular orbits of different altitudes, ranging from 17,000 km (24.6-hr period, shown in Fig. 12) down to 5000 km (6.2-hr period). The results for the lower altitude orbits were not significantly different from those given in Fig. 12, and were therefore not shown, although this may not be the case for elliptic orbits [4]. In general, the data in Fig. 12 indicate that relatively large acquisition ranges may be needed to meet aerocapture approach navigation requirements for higher energy approach trajectories.

The navigation performance obtained when Earth-based Doppler and dual-spacecraft VLBI data are used for orbit determination may also vary with the Mars arrival velocity. The variation in periapsis altitude uncertainty with $V_\infty$ for this case is shown in Fig. 13, for both the 1998 type-1 and type-2 trajectory geometries used previously (the error modeling assumptions used were those given in Table 3). The curve for the type-2 trajectory in Fig. 13 ends at $V_\infty$ = 6 km/sec since this was found to be roughly the upper bound for type-2 trajectories (see Table 1). The curve for the type-1 trajectory extends to $V_\infty$ = 10 km/sec since the high-energy trajectories that might be followed by spacecraft utilizing nuclear rocket propulsion would be type-1-class trajectories. The behavior seen in Fig. 13 indicates that the impact of large values in this case is much less severe than that for spacecraft–spacecraft ranging. This is due to the fact that for the Earth-based data set, the length of the data arc is 8.5 days, regardless of the value of $V_\infty$, whereas in the spacecraft–spacecraft ranging cases, the acquisition range constraint effectively reduces the length of the ranging data arc as $V_\infty$ increases. Overall, though, the results in both Figs. 12 and 13 suggest that 2- to 10-km-altitude delivery accuracies can be achieved over a wide range of arrival velocities and viewing geometries using conventional impulsive guidance methods coupled with either spacecraft–spacecraft ranging or Earth-based dual-spacecraft VLBI.

The final sensitivity analysis investigated the impact of the Doppler data accuracy on the navigation performance that uses Earth-based Doppler and dual-spacecraft VLBI data. Figure 14 shows the variation in periapsis altitude navigation uncertainty with the Doppler data weight (accuracy) for the Doppler/dual-spacecraft VLBI baseline scenarios described in Table 3. The value of $V_\infty$ used for all calculations shown in Fig. 14 was 4.0 km/sec. The Doppler accuracy used in the original baseline scenarios, 1.0 mm/sec, is representative of the performance of the current DSN Doppler system at S-band (2.3 GHz). At X-band (8.4 GHz), DSN Doppler accuracy is about 0.1 mm/sec, except for Sun–Earth–spacecraft angles of less than roughly 10 deg. In Fig. 14, the guidance error causes the altitude uncertainty curve to be essentially flat for Doppler weights of 0.1 mm/sec or better, while at the other extreme, once the Doppler weight reaches about 5.0 mm/sec, the altitude uncertainty curve becomes flat once again, indicating that the Doppler data are no longer affecting the altitude estimate. However, it appears that increasing the Doppler data accuracy from 1.0 mm/sec to 0.1 mm/sec may yield a significant improvement in performance, although it must be noted here that systematic error sources known to affect Doppler data, but not explicitly modeled in this analysis, may cause this improvement to be much less than that shown in Fig. 14 for the ideal case.

## IV. Conclusions

Before stating any specific conclusions, it must be emphasized that the results of this analysis are products of the assumptions and error models used. Although the assumptions made for such parameters as maneuver-execution error statistics and data accuracies were, intentionally, conservative, the error models used to predict orbit-determination performance were relatively simple and did not include all error sources that may be present in actuality, but only those considered most significant. Previous experience with the kinds of approximations and assumptions used in this study suggest that the navigation-error statistics derived from these scenarios could be in error by as much as 20 percent, compared with results obtained with more complete error models.

Radio metric data types using a Mars-orbiting spacecraft as a navigation aid were found to be capable of delivering three-sigma periapsis altitude navigation errors of 2 to 10 km over a fairly wide range of Mars arrival velocities and viewing geometries. This level of performance equals or nearly meets that needed to support aerobraking for Mars orbit insertion by aerobrake vehicles possessing moderate lift-to-drag ratios. In most cases, the guidance-error contribution to the total navigation-error uncertainty was small relative to the orbit-determination errors. For spacecraft–spacecraft ranging data acquired from a Mars relay satellite, the orbit-determination performance was found to be sensitive to changes in the Mars arrival velocity, the declination of the approach trajectory with respect to the satellite orbit plane, and the maximum distance over which ranging data can be acquired.

The orbit-determination accuracy obtained from Earth-based Doppler/dual-spacecraft VLBI data sets was comparable to that obtained from spacecraft–spacecraft ranging data when two DSN baselines are used for obtaining dual-spacecraft VLBI data, but was much poorer when only one baseline was used. In addition, it was found that Doppler/dual-spacecraft VLBI performance was much less sensitive to changes in the Mars arrival velocity than that of spacecraft–spacecraft ranging data. Because of visibility restrictions for the DSN Goldstone–Madrid baseline, it may not be possible to obtain dual-spacecraft VLBI data from both of the currently available DSN baselines (Goldstone–Madrid and Goldstone–Canberra) for Mars encounter declinations (relative to the Earth's equator) less than about −20 deg.

# References

[1] K. M. Spratlin, ed., *1989 Lunar/Mars Initiative Guidance, Navigation and Control Final Report*, CSDL-P-2838, The Charles Stark Draper Laboratory, Inc., Cambridge, Massachusetts, February 1990.

[2] S. W. Shepperd, D. P. Fuhry, and T. J. Brand, "Onboard Preaerocapture Navigation Performance at Mars," paper AAS 91-119, AAS/AIAA Spaceflight Mechanics Meeting, Houston, Texas, February 11–13, 1991.

[3] J. K. Soldner, "Round-Trip Mars Trajectories: New Variations on Classic Mission Profiles," paper AIAA-90-2932, AIAA/AAS Astrodynamics Conference, Portland, Oregon, August 20–22, 1990.

[4] A. K. Konopliv and L. J. Wood, "High-Accuracy Mars Approach Navigation with Radio Metric and Optical Data," paper AIAA-90-2907, AIAA/AAS Astrodynamics Conference, Portland, Oregon, August 20–22, 1990.

[5] C. D. Edwards, W. M. Folkner, J. S. Border, and L. J. Wood, "Spacecraft–Spacecraft Interferometry for Planetary Approach Navigation," paper AAS 91-181, AAS/AIAA Spaceflight Mechanics Meeting, Houston, Texas, February 11–13, 1991.

[6] S. W. Thurman and J. A. Estefan, "Mars Approach Navigation Using Doppler and Range Measurements to Surface Beacons and Orbiting Spacecraft," paper AAS 91-118, AAS/AIAA Spaceflight Mechanics Meeting, Houston, Texas, February 11–13, 1991.

[7] A. C. Young, J. A. Mulqueen, and J. E. Skinner, *Mars Exploration, Venus Swingby and Conjunction Class Mission Modes, Time Period 2000 to 2045*, NASA Technical Memorandum 86477, George C. Marshall Space Flight Center, Huntsville, Alabama, August 31, 1984.

[8] V. A. Lee and S. W. Wilson, "A Survey of Ballistic Mars-Mission Profiles," *J. Spacecraft Rockets*, vol. 4, no. 2, pp. 129–142, February 1967.

[9] A. B. Sergeyevsky, G. C. Snyder, and R. A. Cunniff, *Interplanetary Mission Design Handbook*, vol. 1, part 2, JPL Publication 82-43, Jet Propulsion Laboratory, Pasadena, California, September 15, 1983.

[10] G. R. Hintz, "An Interplanetary Targeting and Orbit Insertion Maneuver Design Technique," *J. Guidance Control*, vol. 5, no. 2, pp. 210–217, March–April 1982.

[11] G. R. Hintz and C. Chadwick, "Design and Analysis Techniques for Trajectory Correction Maneuvers," paper AIAA-84-2014, AIAA/AAS Astrodynamics Conference, Seattle, Washington, August 20–22, 1984.

[12] W. E. Bollman and M. G. Wilson, "Planetary Trajectory Correction Maneuver Dynamics on Approach Hyperbolic Trajectories," paper AIAA-86-2117, AIAA/AAS Astrodynamics Conference, Williamsburg, Virginia, August 18–21, 1986.

[13] W. E. Bollman and C. Chadwick, "Statistics of $\Delta V$ Magnitude for a Trajectory Correction Maneuver Containing Deterministic and Random Components," paper AIAA-82-1429, AIAA/AAS Astrodynamics Conference, San Diego, California, August 9–11, 1982.

[14] W. E. Bollman, "An Approximate Solution to the Analytical Partials of the Spacecraft's Geocentric Range-Rate During the Pre-Encounter Phase of a Planetary Mission," *JPL Space Programs Summary 37-52*, vol. 2, pp. 34–37, May–June 1968.

**Table 1. Launch energy and Mars arrival velocity ranges (optimized Earth–Mars direct transfers, 1995–2020)**

| Trajectory type | Launch energy, km$^2$/sec$^2$ | | Arrival velocity, km/sec | |
|---|---|---|---|---|
| | Avg. | Range | Avg. | Range |
| Minimum launch energy (type-1) | 12.2 | 8.0–19.0 | 4.1 | 2.7–6.0 |
| Minimum launch energy (type-2) | 10.9 | 8.0–17.0 | 3.6 | 2.5–6.0 |
| Minimum arrival velocity (type-1) | 20.0 | 8.0–31.0 | 3.6 | 2.3–4.0 |
| Minimum arrival velocity (type-2) | 16.6 | 9.0–31.0 | 2.9 | 2.4–4.0 |

**Table 2. Spacecraft–spacecraft ranging baseline scenario**

Approach spacecraft trajectory:
$V_\infty = 4.0$ km/sec, $\delta = 20.0$ deg, $\theta = 0.0$ deg, $h_p = 20.0$ km

Relay-satellite orbit:
Period = 24.62 hr (Mars-synchronous), altitude = 17,030.6 km

Ranging measurement accuracy, sample rate:
$\sigma_\rho = \rho/22,000$ (m), sample rate = 6 points/hr

A priori approach spacecraft uncertainties (one-sigma):
$\Delta\vec{B}\cdot\hat{T}, \Delta\vec{B}\cdot\hat{T} = 15.0$ km, $\Delta LTOF = 3.57$ sec, $\Delta V_\infty = 2.0$ cm/sec (each component)

A priori relay-satellite uncertainties (one-sigma):
Position = 2.0 km, velocity = 1.0 cm/sec (each component)

A priori ranging system uncertainties (one-sigma):
Range bias = 33.3 nsec (10.0 m), bias drift = $1.0 \times 10^{-11}$ sec/sec (3.0 mm/sec)

**Table 3. Earth-based Doppler/dual-spacecraft VLBI baseline scenarios**

| Type-1 minimum launch energy | Type-2 minimum launch energy |
|---|---|
| $r = 1.5 \times 10^8$ km | $r = 2.8 \times 10^8$ km |
| $\delta = -16.6$ deg[a] | $\delta = -14.4$ deg[a] |
| $\sigma_{VLBI} = 8.0$ cm | $\sigma_{VLBI} = 5.0$ cm |
| $\sigma_{Dop} = 1.0$ mm/sec | $\sigma_{Dop} = 1.0$ mm/sec |

A priori approach spacecraft uncertainties (one-sigma):

$\Delta\vec{B}\cdot\hat{T}, \Delta\vec{B}\cdot\hat{T} = 15.0$ km, $\Delta LTOF = 3.75$ sec, $\Delta V_\infty = 2.0$ cm/sec (each component)

Doppler data schedule:

Continuous data (sample rate = 1 point/min) from E −10 days to E −1.5 days

Dual-spacecraft VLBI data schedule:

1 point/day each from DSN Goldstone–Madrid and Goldstone–Canberra
baselines from E −10 days to E −1.5 days

(9 points/baseline, 18 points total)

[a] With respect to Earth's equatorial plane.

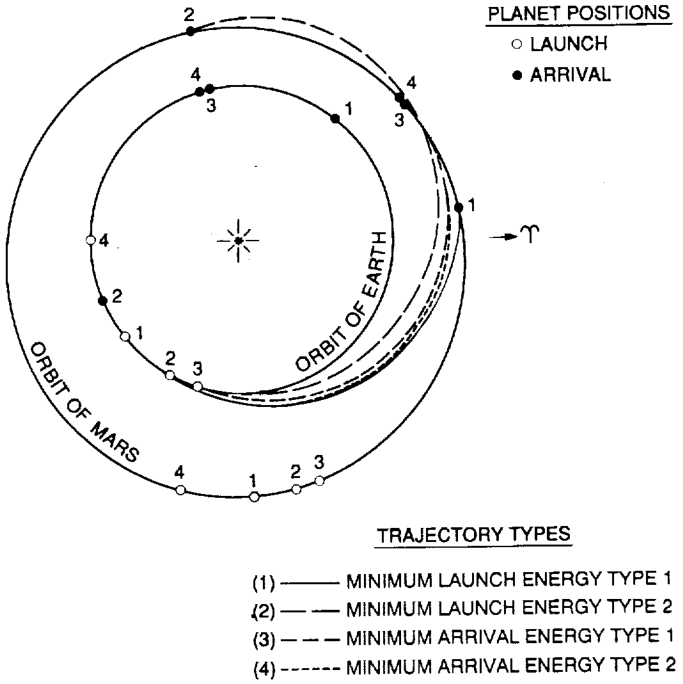

PLANET POSITIONS
○ LAUNCH
● ARRIVAL
ORBIT OF EARTH
ORBIT OF MARS
♈
1
2
3
4
TRAJECTORY TYPES
(1) ——— MINIMUM LAUNCH ENERGY TYPE 1
(2) — — MINIMUM LAUNCH ENERGY TYPE 2
(3) – – – MINIMUM ARRIVAL ENERGY TYPE 1
(4) ------ MINIMUM ARRIVAL ENERGY TYPE 2

**Fig. 1. Direct Earth–Mars transfer trajectories, 1998 launch opportunity.**

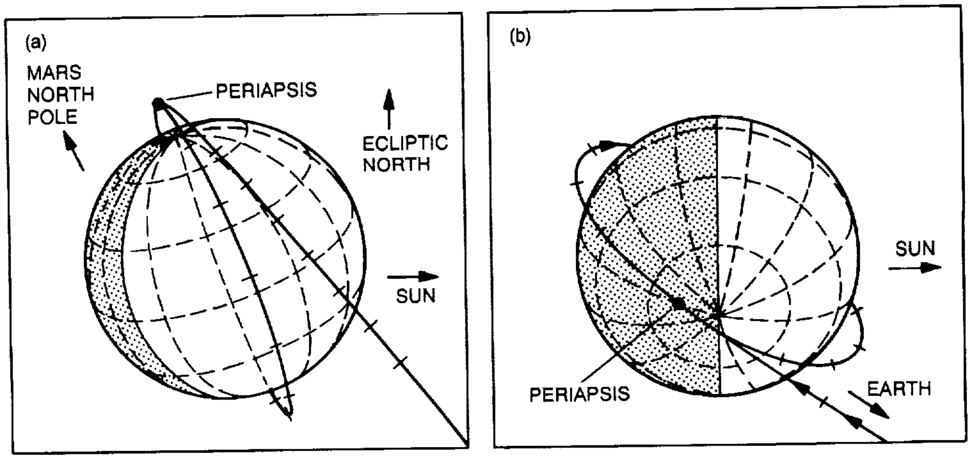


**Fig. 2. Mars arrival geometry for type-1 transfer (1998 opportunity, minimum launch energy): (a) view from Earth, and (b) view from above ecliptic plane.**

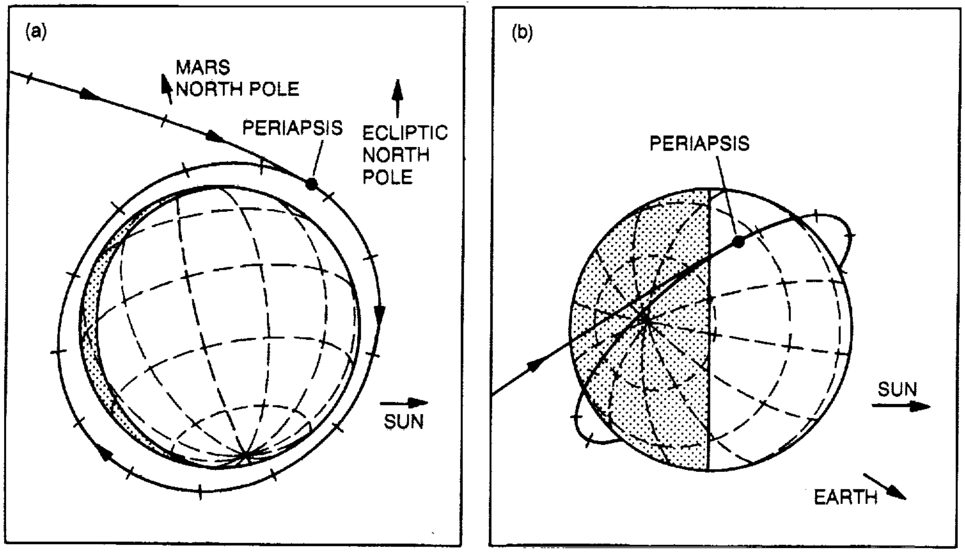


**Fig. 3. Mars arrival geometry for type-2 transfer (1998 opportunity, minimum launch energy): (a) view from Earth, and (b) view from above ecliptic plane.**

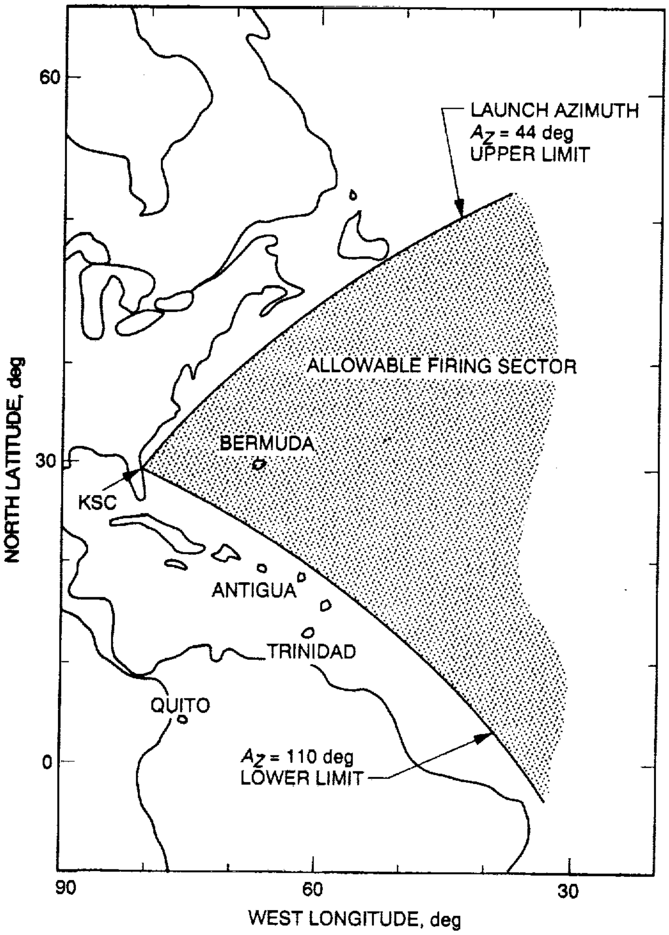


**Fig. 4. Launch azimuth constraints for Kennedy Space Center.**

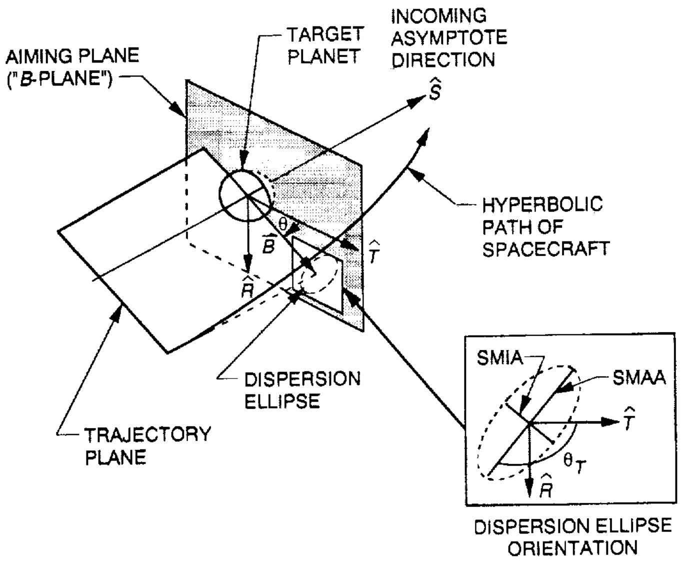


**Fig. 5. Asymptotic coordinate system definition.**

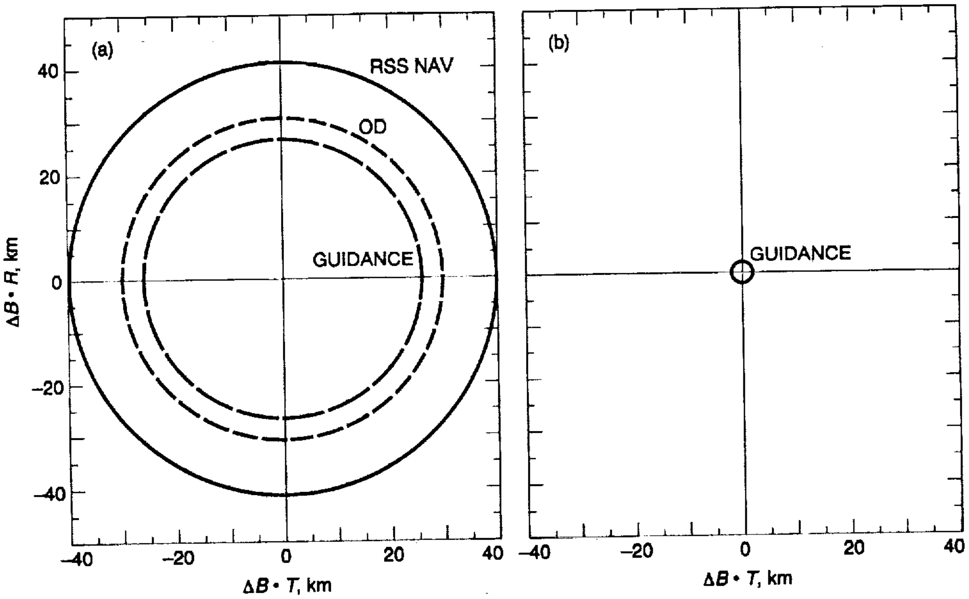


**Fig. 6. Post-TCM1 and -TCM2 three-sigma *B*-plane dispersions: (a) post-TCM1, and (b) post-TCM2.**

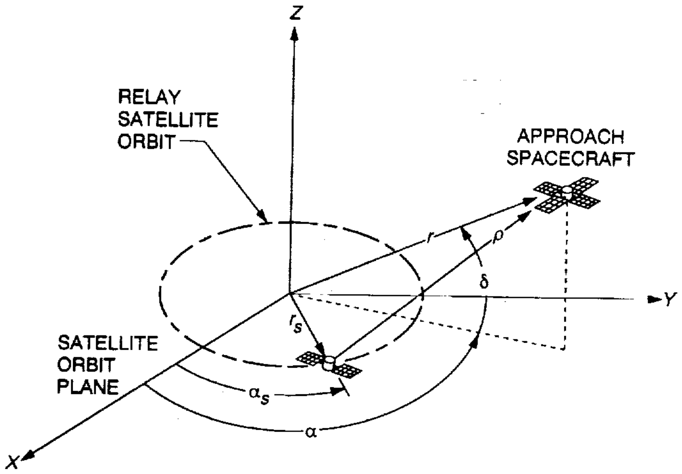


**Fig. 7. Spacecraft–spacecraft tracking geometry.**

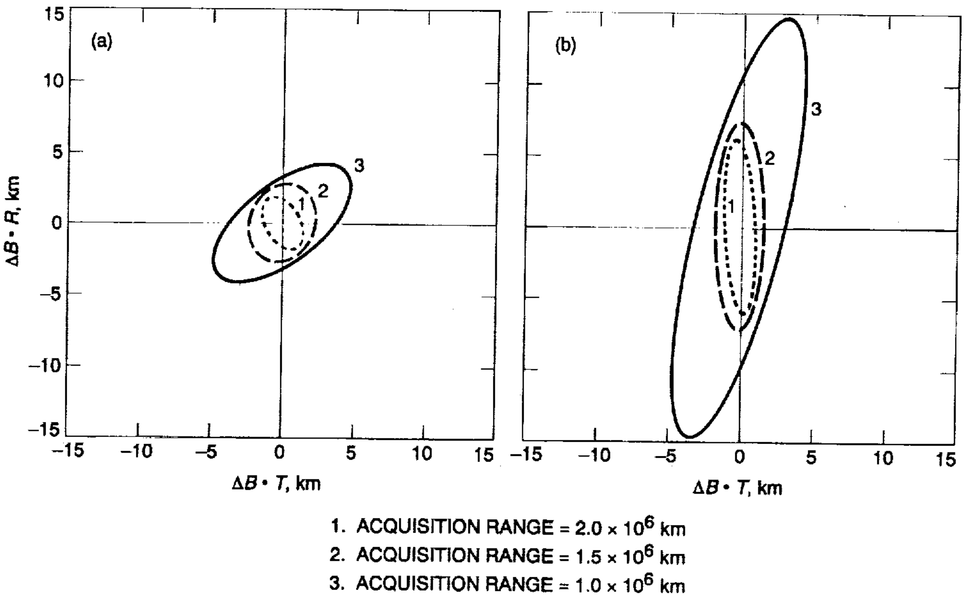


**Fig. 8. Three-sigma *B*-plane dispersions for spacecraft–spacecraft ranging cases (24-hr data cutoff): (a) $\delta$ = 20 deg, and (b) $\delta$= 5 deg.**

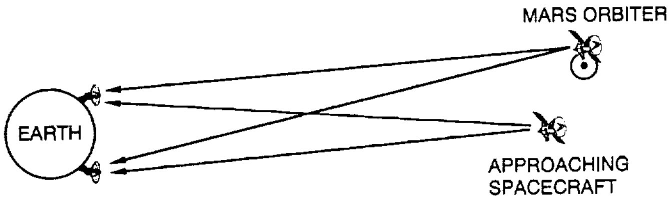
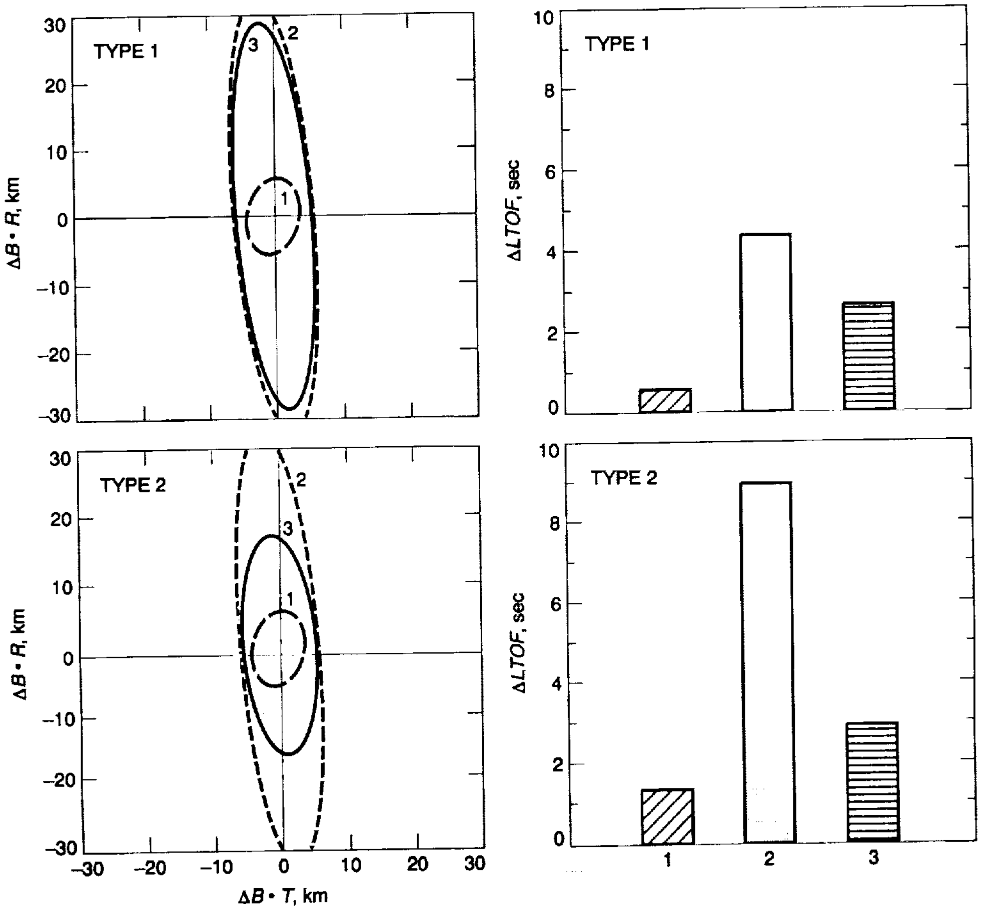


Fig. 9. Dual-spacecraft VLBI observation.



1. VLBI DATA FROM TWO BASELINES
2. VLBI DATA FROM GOLDSTONE-CANBERRA ONLY
3. VLBI DATA FROM GOLDSTONE-MADRID ONLY

Fig. 10. Three-sigma *B*-plane dispersions and LTOF uncertainty for Earth-based Doppler/ dual-spacecraft VLBI cases (36-hr data cutoff).

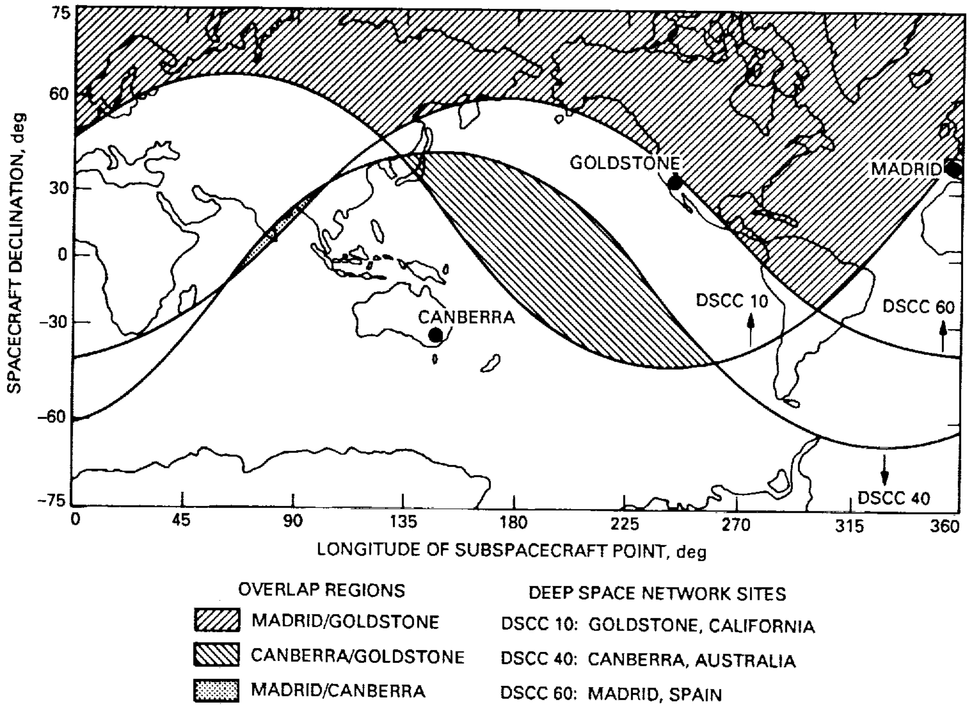


**Fig. 11. Deep Space Network intercontinental baseline visibility regions.**

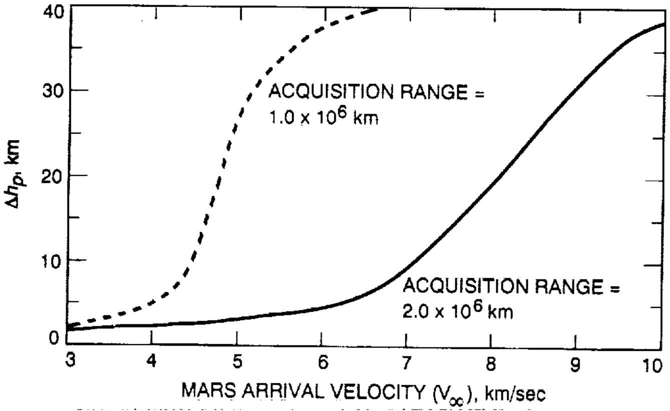


**Fig. 12. Three-sigma altitude uncertainty versus Mars arrival velocity (spacecraft–spacecraft ranging).**

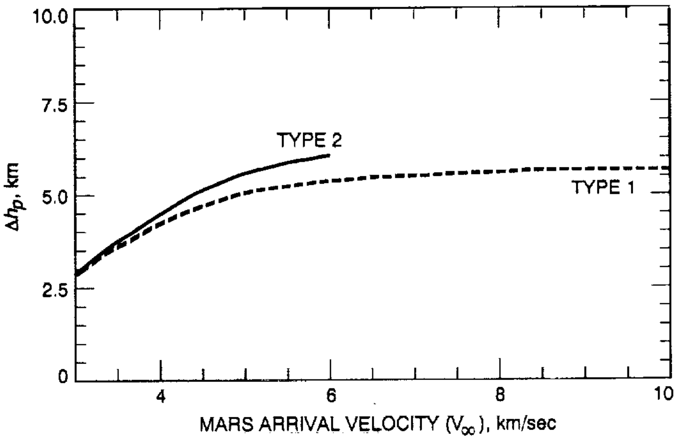


**Fig. 13. Three-sigma altitude uncertainty versus Mars arrival velocity (DSN Doppler/dual-spacecraft VLBI).**

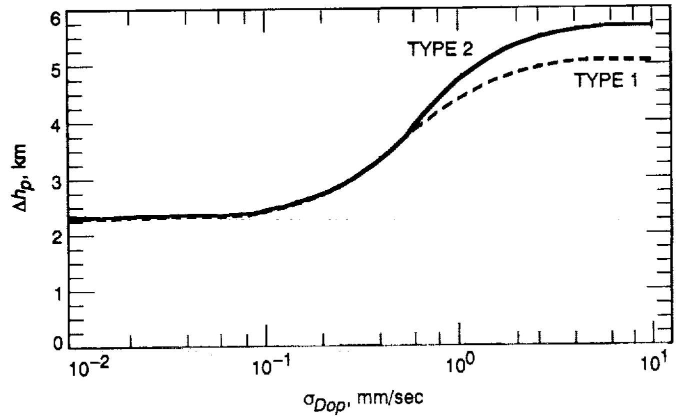


**Fig. 14. Three-sigma altitude uncertainty versus Doppler weight (DSN Doppler/dual-spacecraft VLBI).**

# A 640-MHz 32-Megachannel Real-Time Polyphase-FFT Spectrum Analyzer

G. A. Zimmerman, M. F. Garyantes, M. J. Grimm
Communications Systems Research Section

B. Charny
Spacecraft Telecommunications Equipment Section

*A polyphase-fast-Fourier-transform (FFT) spectrum analyzer being designed for NASA's Search for Extraterrestrial Intelligence (SETI) Sky Survey at the Jet Propulsion Laboratory is described. By replacing the time-domain multiplicative window preprocessing with polyphase filter processing, much of the processing loss of windowed FFTs can be eliminated. Polyphase coefficient memory costs are minimized by effective use of run-length compression. Finite word length effects are analyzed, producing a balanced system with 8-bit inputs, 16-bit fixed-point polyphase arithmetic, and 24-bit fixed-point FFT arithmetic. Fixed-point renormalization midway through the computation is seen to be naturally accommodated by the matrix FFT algorithm proposed. Simulation results validate the finite word length arithmetic analysis and the renormalization technique.*

## I. Introduction

A $2^{25}$ (33,554,432) channel, 640-MHz-wide polyphase-fast-Fourier-transform (FFT) spectrum analyzer is being designed at the Jet Propulsion Laboratory for the Search for Extraterrestrial Intelligence (SETI) Sky Survey. This spectrum analyzer will be used to separate two 320-MHz-wide polarizations into channels approximately 20 Hz wide for input to SETI signal detection algorithms. Construction of a prototype windowed-FFT spectrum analyzer [1] with 40 MHz of bandwidth and $2^{21}$ (2,097,152) channels has recently been completed. The new spectrum analyzer design, similar to the prototype machine in many respects, is functionally divided into eight identical 80-MHz, 4-megachannel, real-input polyphase-FFT filter banks, each implemented as a pipelined special-purpose hardware signal processor. The spectrum analyzer functions consist of polyphase preprocessing, a 4-megapoint matrix-algorithm FFT and trigonometric recombination ("real-adjust") to compute the positive half of an 8-megapoint real FFT.

Other than the increases in bandwidth and number of channels, the main architectural difference from the prototype spectrum analyzer system is that the new design is a polyphase-FFT spectrum analyzer rather than a windowed-FFT spectrum analyzer. The advantages of polyphase-FFT spectrum analysis, as well as a review of the supporting theory, are presented in [2] and are only touched on here. Similarly, features that the spectrum analyzer has in common with the prototype system can be found in [1] and are not described in detail here.

The remainder of this article will be divided into the following four main sections: a general description of the spectrum analyzer, a description of the polyphase-FFT filter bank implementation, a discussion of finite word length effects and the fixed-point arithmetic implementation, and results from system simulation.

## II. General Description

The 320-MHz dual-polarization system is divided into four 80-MHz subbands per polarization. Each of the 80-MHz subbands, or "slices," is identical to the others, and each can be operated independently as an 80-MHz spectrum analyzer. A functional block diagram of the 320-MHz dual-polarization system is shown in Fig. 1.

Like the prototype system, each 80-MHz slice of the spectrum analyzer is a pipelined architecture, allowing all stages of the polyphase-FFT algorithm to execute concurrently. As in the prototype system, a stage "bypass" capability and stimulus and response buffers provide built-in testability. Each 80-MHz slice of the system will have an 8-bit analog/digital (A/D) converter as its input, and 24-bit fixed-point arithmetic will be used for the FFT portion. The fixed-point arithmetic will be implemented in an application-specific integrated circuit (ASIC), jointly developed with the Telecommunications and Data Acquisition (TDA) Advanced Systems very large scale integration (VLSI) program. This saves both memory and arithmetic relative to a floating-point implementation. In addition, because all of the FFT is completely performed in 24-bit fixed-point arithmetic, the transforms before and after the matrix transposition, or "corner turn," can use identical boards, saving a unique board design. In contrast, the prototype performs the column FFTs in 16-bit fixed-point arithmetic and the row FFTs in 32-bit floating-point arithmetic.

Architectural improvements in the FFT portion of the new spectrum analyzer include replacing two spectrum-length (4-complex-megapoint) double buffers with single buffers and the removal of a third spectrum-length double buffer.[1] This improvement significantly reduces the amount of memory in the spectrum analyzer, reducing both size and cost.

The user-loadable window function in the prototype system has been replaced by polyphase filter preprocessing. The polyphase preprocessor is capable of operating as a window function, because a windowed discrete Fourier transform (DFT) is a degenerate case of a polyphase DFT filter bank. For more information on polyphase DFT filter banks, see [2].

As in the prototype system, the FFT is implemented using a matrix-style DFT pipe. By decomposing the transform into shorter row and column transforms, the long-delay memories and coefficient ("twiddle factor") storage required for the 4-megapoint FFT are concentrated in a single matrix transposition and complex multiplication stage, making the actual FFT arithmetic boards much simpler. The spectrum analyzer performs a 4-megapoint ($2^{22}$) DFT as 4096 point column DFTs followed by 1024 point row DFTs with multiplication of the matrix entries by complex rotation factors between the row and column transforms. The row and column DFTs are each performed as pipelined radix-4 FFTs.

While the prototype system was implemented in wire-wrap technology, the greater bandwidth and the highly repetitive nature of the new system require implementation using multilayer printed circuit boards. Because the new system takes advantage of improvements in technology, while it is functionally quite similar to the prototype, the detailed hardware designs are all new. A functional block diagram is shown in Fig. 2.

## III. Polyphase FFT Structure

Because the spectrum analyzer channels are considered independently by the signal detection algorithms, the worst case processing loss [3], that is, the maximum attenuation of a continuous-wave (CW) signal in a channel's passband (in dB) plus the ratio of a channel's equivalent noise bandwidth to the channel spacing (in dB), is crucial to the system's sensitivity to weak signals. This is rarely less than 3 dB for windowed FFTs [3]. This is a result of the fact that for a windowed FFT, the time aperture over which the signal is considered (in seconds) is exactly equal to the reciprocal of the FFT channel spacing (in hertz).

[1] R. Brown, *Using Single Buffers and Data Reorganization to Implement a Multi-Megasample FFT* (internal document), Jet Propulsion Laboratory, Pasadena, California.

Consider an FFT with channel spacing $N_t$ times finer than the desired spacing. This FFT operates on a time aperture $N_t$ times longer than the FFT that would provide the desired channel spacing. Call the shorter FFT time aperture $N_1$ samples and the longer FFT $N_1 N_t$ samples. Now, apply a time-domain multiplicative window with the desired low-pass-filter transfer function to the long FFT, compute the FFT, and discard all but every $N_t$th FFT channel, i.e., retain only those bins whose center frequencies correspond to a center frequency in the shorter length $N_1$ transform. Now shift the $N_1 N_t$ sample input vector by $N_1$ samples and repeat the procedure. This process is equivalent to a polyphase-FFT filter bank, implementing a bank of identically shaped finite impulse response (FIR) band-pass filters, each centered on the bin center frequencies of the shorter, $N_1$-long FFT.

In summary, the polyphase-FFT filter bank operates on a time aperture larger than its resolution, allowing bin shapes to encompass any transfer function that can be implemented as an $N_t N_1$ tap FIR filter. With as few as 8 taps per polyphase branch, it is possible to reduce the worst case processing loss to less than 1 dB. In fact, by including the polyphase-filter preprocessing step prior to computing the FFT, as much as 2.6 to 2.9 dB can be gained in worst-case processing loss over Hanning or Blackman windowed FFTs. In large FFT systems such as this one, the polyphase preprocessing has a small computational cost when compared to the FFT. However, substantial gains over windowed-FFT techniques are made even for systems with 2- or 4-tap polyphase branches, making polyphase preprocessing more than appropriate for smaller spectrum analyzer systems. For more discussion of polyphase DFT filter banks, see [2].

The main cost incurred in the polyphase preprocessing is in memory. The polyphase preprocessing requires that $N_t N_1$ input points must be stored at any one time, whereas in the windowed FFT system only $N_1$ points of storage were required. This storage requirement is mitigated somewhat by the fact that the points to be stored are of the input word length that is typically much shorter than the FFT arithmetic word length. In addition, the polyphase coefficients may be quantized to only slightly longer than the input word length without significant loss. In our implementation, the input word length is 8 bits, and 12 bits is sufficient for the polyphase coefficients. Due to the fact that our desired bin transfer functions are based on ideal band-pass filters, the prototype low-pass polyphase transfer function does not deviate significantly from an ideal low-pass filter. Correspondingly, the polyphase coefficients do not deviate significantly from the Fourier transform of an ideal low-pass filter, the Sinc function. As a result, the maximum rate of change of coefficients is limited, and a minimum run length can be found for quantized coefficients. For 12-bit coefficient quantization, the minimum run length is slightly more than 2048 coefficients, allowing for effective use of run-length compression to minimize the coefficient memory cost.

The minimum run length is greater than the row (second) FFT length, guaranteeing that there will be at most one transition per row. As a result, the rows may be separately run-length compressed, simply by identifying two coefficient values and the location of the transition. This is advantageous in a pipelined signal processor, where the order of data points into the board may be scrambled both within columns and by columns, because it allows run-length coding to be efficiently implemented as shown in Fig. 3.

Hardware simplifications result by setting the real and imaginary polyphase coefficients equal. Since the FFT is to be performed with the real-adjust algorithm, the real samples represent even-time indices while the imaginary samples represent the odd-time indices. Constraining real and imaginary polyphase coefficients to be identical means that the impulse response of the prototype low-pass FIR filter is constrained to change values only on even samples. Such a constraint will alter the transfer function of the resulting filter. One can use a matrix formulation of the DFT ($N_t N_1$ by 2 real-only points) to examine the transfer function of the resulting filter. Since the impulse response changes value only on even samples, the initial two-column DFTs ($N_t N_1$) are identical, and, in fact, both are the desired low-pass prototype transfer functions centered on the zero-frequency bin. The relevant portions of the entire transfer function are computed by taking the two-point DFTs of the "twiddle-in-the-middle" complex rotation factors $\exp\left(-j2\pi[(\text{row no.} \times \text{ col. no.})N_1]\right)$ in the passband and near-transition-band regions of the low-pass initial DFT results. The resulting transfer function is mainly low pass, with a replicant passband, attenuated by 135 dB, centered at the Nyquist frequency ($\pi$). For a system with a 41-dB signal-to-noise ratio (SNR) CW input, this results in a spur 28 dB below the noise floor in the output, requiring about 360,000 accumulations or about 5.2 hours to reach a 0-dB bin output SNR. It is also important to note that this spur is at the limit of the 24-bit representation, since the power of the smallest 24-bit number ($2^{23} + 0j$) is 141.5 dB down from the largest power value ($-1 - j$), and, as such, is at about the same level as output arithmetic and quantization noise. A simulation was performed with the real and imaginary coefficients constrained to be identical, and the results validated the above analysis.

## IV. Finite Word Length Effects

With 8-bit input quantization, 16-bit fixed-point polyphase arithmetic and 24-bit fixed-point FFT arithmetic provide balanced system performance, with the input quantization (A/D) noise dominating. The spectrum analyzer has been designed with 8-bit input quantization based on considerations of A/D converter technology and the maximum expected RFI levels.

Noise due to finite word lengths can be divided between the A/D converter and the spectrum analyzer arithmetic. A balanced system would have roughly equal contributions from each source. The purpose of the SETI system is to detect small signals in the presence of interference and noise. Arithmetic and quantization noise will degrade the system performance by effectively increasing the input noise level, decreasing the SNR. The resulting noise power out will be:

$$\begin{aligned}\sigma^2_{out} &= \sigma^2_{in} + \sigma^2_{NQ} + \sigma^2_{Arith} \\ &= \sigma^2_{in}\left(1 + \frac{\sigma^2_{NQ}}{\sigma^2_{in}}\left(1 + \frac{\sigma^2_{Arith}}{\sigma^2_{NQ}}\right)\right)\end{aligned}$$

where $\sigma^2_{in}$ and $\sigma^2_{out}$ are the input and output noise power, and $\sigma^2_{NQ}$ and $\sigma^2_{Arith}$ are the quantization and arithmetic noise, respectively. It is apparent that the sensitivity loss is

$$\frac{\sigma^2_{out}}{\sigma^2_{in}} = 1 + \frac{\sigma^2_{NQ}}{\sigma^2_{in}}\left(1 + \frac{\sigma^2_{Arith}}{\sigma^2_{NQ}}\right)$$

To put the input noise and the two digital noise factors into the same units, a translation from rms noise (volts) to quantization levels must be used; however, given a fixed input quantizer word length, it is apparent that the loss will be controlled by the ratio of the arithmetic to quantization noise.

Using the techniques in [4], the noise due to fixed-point multiplication round offs present in an output can be computed. This computation assumes that all round-off noise sources are white, mutually uncorrelated, zero mean, and uniformly distributed with a maximum value of 1/2 of one least significant bit. A radix-4 decimation-in-frequency FFT implementation was assumed. Each radix-4 FFT stage scales the data by 1/4 to ensure no overflows occur. Blind overflow-protection scaling of the FFT stages would result in few significant bits to represent uncorrupted, noise-only or weak signal outputs, so the 24-bit words must be automatically renormalized at some point in the computation. This renormalization occurs at the corner-turn memory. Making the usual assumption of white, zero-mean, uniformly distributed input quantizer noise, noise due to the input quantizer was computed. The resulting ratio of the arithmetic noise to input quantization noise is given in Table 1. These results were used to calculate losses for the various input SNRs examined in the next section.

Note that renormalizing the data amplifies all sources of noise prior to the renormalization. When the renormalization shift is small, the majority of the arithmetic noise is due to the final stages of arithmetic, after the renormalization. Larger renormalization shifts reduce the relative effect of the fixed-point arithmetic. Note that the amount of renormalization shift depends on the maximum peak in the spectrum. In most cases, strong interference will be wideband relative to a spectrum analyzer channel (19 Hz), allowing a significant (>1) renormalization shift to be used.

In a pipelined processor implementation, renormalization requires a buffer capable of holding the entire vector. The matrix FFT algorithm was chosen for the FFT implementation based on experience with the SETI prototype [1]. The initial reason for this choice was because the algorithm concentrates pipeline delay memories and coefficient memory in a central matrix transposition and vector multiplication stage. In fixed-point FFTs, another advantage becomes apparent: the matrix transposition buffer provides a natural location for the vector renormalization. Renormalization can therefore be performed at the corner-turn memory, prior to the twiddle multiplication stage, without the cost of an additional buffer.

## V. Simulation Results

The resulting system was simulated on a SUN-4 computer workstation with the appropriate word lengths, validating the analytical results and the renormalization approach. The simulated system had the following features:

(1) A 12-tap-per-branch polyphase filter with 12-bit coefficients.

(2) 16-bit fixed-point multiply-accumulate arithmetic in the polyphase filter computation.

(3) 24-bit fixed-point arithmetic in the FFT, twiddle multiply, and real-adjust sections.

(4) 32-bit IEEE floating-point power calculation.

(5) A radix-4 4096 point by radix-4 1024 point matrix style 4-megapoint complex FFT with real-adjust and automatic renormalization between the row and column transforms.

(6) 8, 10, or 16 bits of quantization at the input (A/D conversion).

The simulations were performed to test the performance in the presence of a strong signal (75 percent full scale on the input quantizer), measuring both output SNR and two-tone dynamic range. The simulations demonstrated 90-dB, close-in, two-tone dynamic range, detecting a weak signal 90 dB down from a strong signal 10 FFT channels away.

## A. Finite Word Length SNR Degradation

The measured SNR out was compared with the SNR that would be output if the input quantization and arithmetic were infinitely precise. The ideal, infinite precision SNR is given by:

$$SNR_{ideal\ out}\ (\text{dB}) = SNR_{in}\ (\text{dB}) + Gain_{FFT}\ (\text{dB}) + Gain_{PP}\ (\text{dB}) - EQNB_{PP}\ (\text{dB})$$

where $Gain_{FFT}$ is the gain due to the number of channels in the FFT ($2^{22}$ channels = 66.23 dB), $Gain_{PP}$ is the gain of the polyphase-DFT filter transfer function, dependent on the signal's frequency within the resolved FFT channel (−0.268 dB at 0.4 bins offset), and $EQNB_{PP}$ is the equivalent noise bandwidth of the polyphase-DFT filter transfer function relative to an ideal bandpass filter (0.346 dB in this 12-tap case). For our test cases, the strong signal frequency was held constant at 10 percent of the Nyquist frequency, resulting in an offset of 0.4 bins from the center frequency of the target bin.

Losses were predicted using the conventional round-off and quantization noise models, assuming white noise from each source. The results of the simulations are given in Table 2.

At lower SNRs there is excellent agreement with theory. For 8-bit input quantization, the A/D converter noise and spurs dominate the arithmetic contribution, as predicted; they can be observed from the lower SNR 8-bit measurements and a comparison of the 41-dB and 47-dB SNR 8-bit predictions and measurements with the 10- and 16-bit predictions and measurements. It may be noted that at high SNRs, the predictions deviate from the measurements. While the measured losses for 8-bit inputs exceed the predictions, those for 10- and 16-bit inputs are less than the predicted loss. At high input SNRs, the assumption of white quantization and computational noise sources is violated. This is particularly true for the 8-bit input quantization noise with greater than 40 dB input SNR [5]. Since the 8-bit case behaves differently than either the 10- or 16-bit cases at these SNR levels, the increased loss for 8-bit input at 41- and 47-dB input SNR is attributed to high SNR input quantizer effects. It is important to notice that at input SNRs above 40-dB (output SNRs above 106 dB), quantization spurs from an 8-bit input quantizer become noticeable in the output, defining the upper limit of the spectrum analyzer's useful range in the presence of strong CWs.

Identifying the cause of decreased loss at high SNRs requires a closer examination of the sources of quantization noise. As the SNR increases, the loss due to the 24-bit fixed-point arithmetic grows. Unlike the 8-bit input quantization cases, the loss with 16-bit input quantization is almost entirely due to the arithmetic, and with 10-bit input, a large portion (78 percent) is due to the arithmetic. Due to overflow-preventing attenuation in the FFT arithmetic, which scales the data by 1/4 each stage, the lion's share of the FFT noise contribution comes from the final stages of FFT and real-adjust arithmetic. With high SNR CW inputs, beyond the 40-dB design range of the instrument, a significant fraction of the output noise power underflows the 24-bit fixed-point precision and is mapped to zero. This is especially apparent with 60-dB input SNR, where 12.5 percent (one-eighth) of the output values are complex zeros, and an additional 45 percent have either a zero real or imaginary component. Since arithmetic noise is computed from multiplication round offs, and multiplication by zero is an exact operation, the result is a decreased number of noise sources when high SNR tones are input. As a result, the arithmetic noise can be reduced by as much as 57.4 percent, or 3.71 dB in the 60-dB input case. This is sufficient to account for the discrepancies observed.

## B. Two-Tone Dynamic Range

Measurements of the system's close-in two-tone dynamic range were performed. The two-tone dynamic range is defined as the maximum ratio of strong to weak signal levels at which a weak signal can be detected. Detection was performed without accumulating, providing a lower bound on the actual two-tone dynamic range of the instrument. Two-tone tests confirmed that a tone 90 dB down from the strong signal and 10 bins away is detectable, without accumulation, as the maximum spectral peak outside of ±4 bins from the strong signal.

The minimum detectable small signal level, allowing accumulation, and hence the two-tone dynamic range are determined by the maximum spectral level outside of the immediate spectral neighborhood of the strong signal. The maximum spur-to-noise ratio defines the minimum detectable SNR. The minimum detectable signal must be only slightly greater than the maximum spur. Given the very large number of noise samples in the spectrum, the ratio of the maximum to the average of the sample set would be tightly constrained (11.6 dB to 12.4 dB, 10 percent to 90 percent probability) if the output were white Gaussian noise only and contained no spurs. Some measured values and the probabilities that they are due to white Gaussian noise alone are given in Table 3.

Measurements made with greater than 8-bit inputs indicate that the maximum spur levels observed were due to the input quantization. These measurements confirm that a system with 8-bit inputs would have a two-tone dynamic range of 41 dB + $Gain_{FFT}$ + $\min(Gain_{PP})$ − $EQNB_{PP}$ − 13.0 dB = 93.6 dB. Examples of detection of a weak signal 80 dB down from a strong signal 10 channels away are shown in Fig. 4.

## VI. Conclusions

This article has described the latest results in a continuing effort to build more sensitive and broader bandwidth multichannel spectrum analyzers [1,6]. The thrusts of the current effort, described in this article, have been in the areas of signal processing, machine architecture, and technology utilization. The introduction of the polyphase-FFT provides signal processing gain superior to windowed FFTs, nearly independent of frequency location within a bin. Improvements in machine architecture have allowed the use of fixed-point arithmetic and have saved cost in both memory and arithmetic parts. Through the effective utilization of advances in technology, the bandwidth and number of channels of a single processor unit have each been doubled. The signal processor will be capable of operation at the limits of its 8-bit A/D converter, allowing up to 41-dB input SNR with less than 1-dB loss in sensitivity and exhibiting greater than 90-dB two-tone dynamic range.

# References

[1] M. P. Quirk, H. C. Wilck, M. F. Garyantes, and M. J. Grimm, "A Wideband, High-Resolution Spectrum Analyzer," *TDA Progress Report 42-93*, vol. January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 188–198, May 15, 1988.

[2] G. A. Zimmerman and S. Gulkis, "Polyphase–Discrete Fourier Transform Spectrum Analysis for the Search for Extraterrestrial Intelligence Sky Survey," *TDA Progress Report 42-107*, vol. July–September 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 141–154, November 15, 1991.

[3] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, vol. 66, no. 1, 1978.

[4] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Chapter 9, Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

[5] M. J. Flanagan, "The Behavior of Quantization Spectra as a Function of Signal-to-Noise Ratio," *TDA Progress Report 42-107*, vol. July–September 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 155–168, November 15, 1991.

[6] G. A. Morris and H. C. Wilck, "JPL $2^{20}$ Channel 300 MHz Bandwidth Digital Spectrum Analyzer," presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, Oklahoma, 1978.

**Table 1. Arithmetic noise/input quantization noise**

| Renormalization shift | Input quantization | |
|---|---|---|
| | 8 bits, dB | 10 bits, dB |
| 0 | −0.64 | 11.40 |
| 1 | −6.66 | 5.39 |
| 2 | −12.66 | −0.62 |

**Table 2. Simulation results**

| Input quantization, bits | $\mathrm{SNR}_{in}$, dB | Ideal $\mathrm{SNR}_{out}$, dB | Predicted loss, dB | Measured loss, dB |
|---|---|---|---|---|
| 8 | 30 | 95.61 | 0.05 | 0.02 |
| 8 | 35 | 100.61 | 0.15 | 0.18 |
| 8 | 38 | 103.61 | 0.29 | 0.41 |
| 8 | 41 | 106.61 | 0.56 | 0.84 |
| 8 | 47 | 112.61 | 1.91 | 2.53 |
| 10 | 47 | 112.61 | 0.52 | 0.22 |
| 16 | 41 | 106.61 | 0.11 | −0.065 |
| 16 | 60 | 125.61 | 4.70 | 1.415 |

**Table 3. Maximum noise spur to average > 3500 bins from strong signal**

| Input, bits | $\mathrm{SNR}_{in}$, dB | Strong signal/ weak signal | Max spur/ noise average, dB | Probability (max./avg. > $x$ dB) (noise alone), percent |
|---|---|---|---|---|
| 8 | 30 | Single tone | 11.60 | `90 |
| 8 | 35 | Single tone | 11.71 | 80 |
| 8 | 38 | Single tone | 11.54 | >90 |
| 8 | 41 | Single tone | 12.80 | 2 |
| 10 | 47 | Single tone | 11.49 | >90 |
| 8 | 41 | 80 dB | 12.87 | 2 |
| 8 | 41 | 90 dB | 13.00 | 0.9 |
| 16 | 41 | 80 dB | 11.61 | 90 |
| 16 | 41 | 90 dB | 11.49 | >90 |

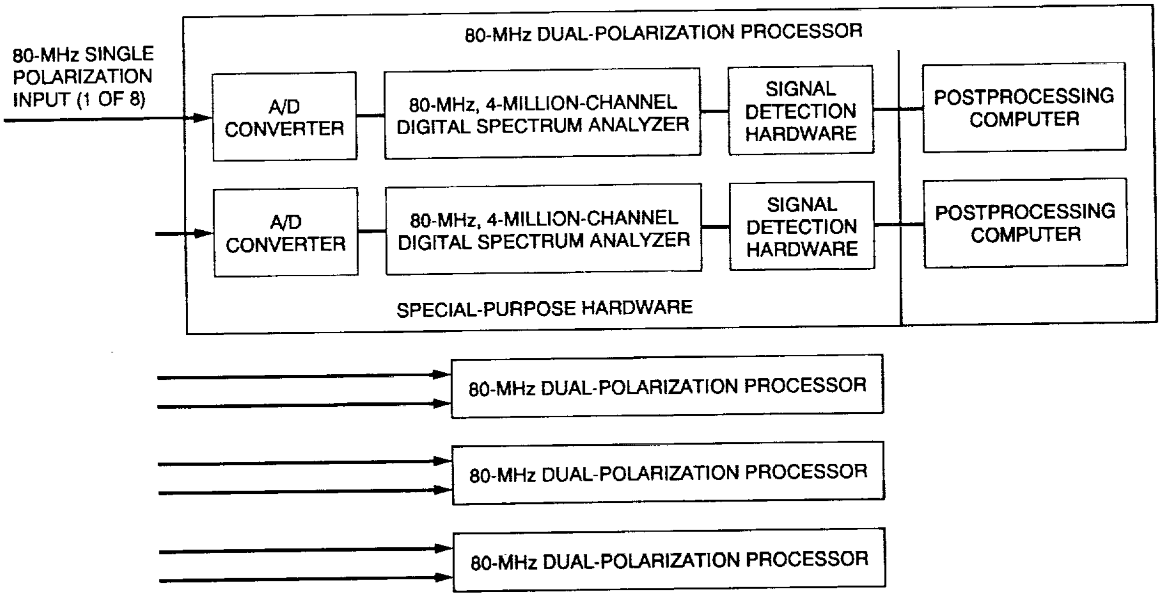

80-MHz SINGLE POLARIZATION INPUT (1 OF 8)
80-MHz DUAL-POLARIZATION PROCESSOR
A/D CONVERTER
80-MHz, 4-MILLION-CHANNEL DIGITAL SPECTRUM ANALYZER
SIGNAL DETECTION HARDWARE
POSTPROCESSING COMPUTER
A/D CONVERTER
80-MHz, 4-MILLION-CHANNEL DIGITAL SPECTRUM ANALYZER
SIGNAL DETECTION HARDWARE
POSTPROCESSING COMPUTER
SPECIAL-PURPOSE HARDWARE
80-MHz DUAL-POLARIZATION PROCESSOR
80-MHz DUAL-POLARIZATION PROCESSOR
80-MHz DUAL-POLARIZATION PROCESSOR

**Fig. 1. SETI-MOP sky survey operational signal processor (320-MHz dual polarization).**

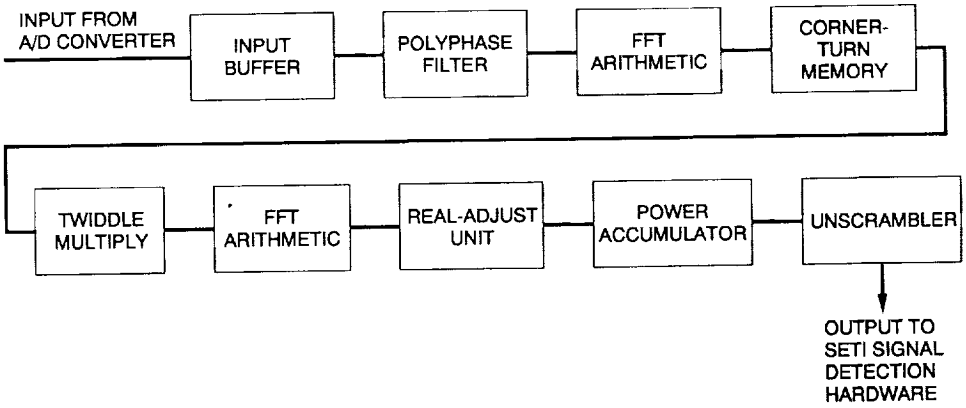


**Fig. 2. An 80-MHz, 4-million-channel digital spectrum analyzer.**

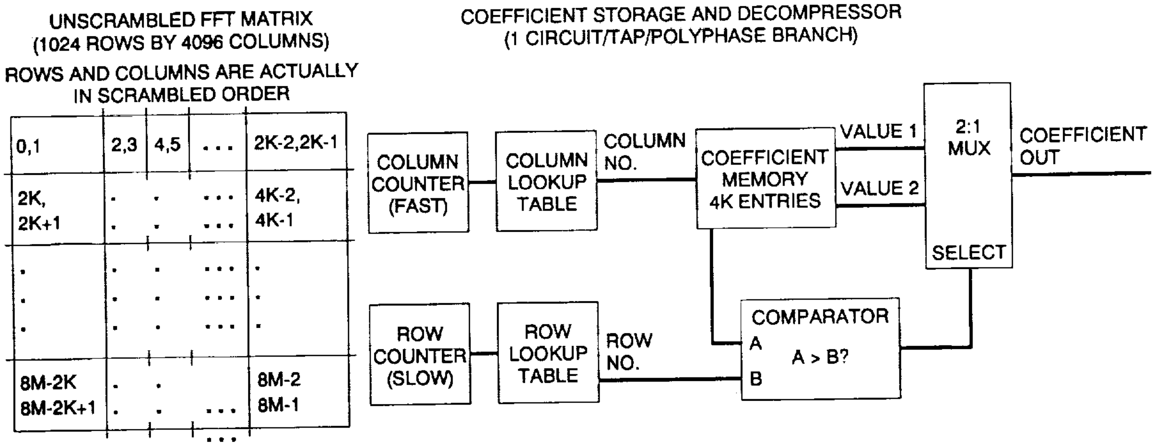


**Fig. 3. Compression of polyphase filter coefficients.**

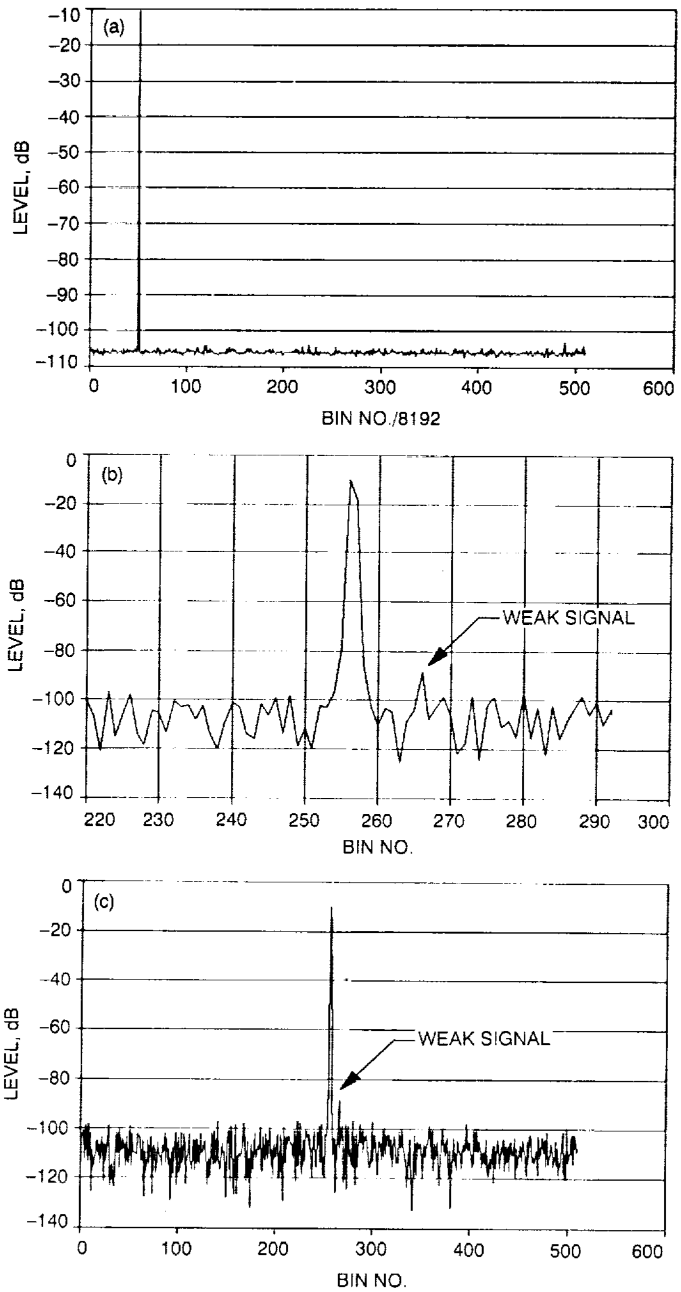


**Fig. 4. Two-tone dynamic range test, 80-dB case: (a) full 4-megapoint spectrum with 8K compression by maximum; (b) signal region uncompressed; and (c) expanded signal region uncompressed.**

5/4- 55
55217
P- 14

# Polyphase–Discrete Fourier Transform Spectrum Analysis for the Search for Extraterrestrial Intelligence Sky Survey

G. A. Zimmerman
Communications Systems Research Section

S. Gulkis
Space Physics and Astrophysics Section

*The sensitivity of a matched filter-detection system to a finite-duration continuous wave (CW) tone is compared with the sensitivities of a windowed discrete Fourier transform (DFT) system and an ideal bandpass filter-bank system. These comparisons are made in the context of the NASA Search for Extraterrestrial Intelligence (SETI) microwave observing project (MOP) sky survey. A review of the theory of polyphase-DFT filter banks and its relationship to the well-known windowed-DFT process is presented. The polyphase-DFT system approximates the ideal bandpass filter bank by using as few as eight filter taps per polyphase branch. An improvement in sensitivity of ~3 dB over a windowed-DFT system can be obtained by using the polyphase-DFT approach. Sidelobe rejection of the polyphase-DFT system is vastly superior to the windowed-DFT system, thereby improving its performance in the presence of radio frequency interference (RFI).*

## I. Introduction

The Search for Extraterrestrial Intelligence (SETI) sky survey will use a spectrum-analysis system with a 640-MHz total bandwidth and 33,554,432 ($2^{25}$) channels divided equally between two 320-MHz-wide polarizations [1]. The purpose of the system is to detect finite-duration continuous wave (CW) tones. Cost and technology constrain the number of channels in the system. To evaluate the resulting sensitivity loss, a study was undertaken of the relative sensitivities of theoretically optimal detection and systems with a fixed number of channels.

Spectrum analysis by means of a windowed discrete Fourier transform (DFT), usually implemented as a fast Fourier transform (FFT), is used in a wide variety of fields [2]. For the SETI sky survey, as in many applications, a windowed DFT has been proposed to channelize the input bandwidth [3]. The windowed DFT isolates interference in the frequency domain and allows CW signal-detection algorithms to operate on the narrowband output channels of the DFT, which improves the signal-to-noise ratio (SNR) at which signal detection must take place. However, the superior sidelobe rejection of windowed DFT's does not come without a performance penalty. This penalty is well-known and is tabulated in [2] for various window functions. For most common window functions, the worst-case processing loss is between 3 to 4 dB. For the SETI sky survey, this loss would add to that incurred by fixing the number of channels. Recovering this processing loss would be equivalent to halving the system temperature or doubling the antenna's collecting area.

A multirate digital signal-processing technique, known as a polyphase-DFT filter bank, is proposed to recover most of the windowed-DFT processing loss while providing superior interference isolation. The theory of polyphase-DFT filter banks has been established in the multirate digital-signal-processing community for some years now. However, unlike windowed-DFT techniques, the filter bank theory is not, in general, well-known. It is the intent of this article to propagate the theory of polyphase-DFT filter banks as a technique related to the use of windows with DFT's, and it is suggested that applications using windowed DFT's might benefit by switching to polyphase-DFT filter banks.

The first part of this article compares the sensitivity of the theoretically optimal matched filter-detection system with the sensitivities of a Hanning windowed-DFT system and an ideal bandpass filter-bank system. The second part reviews the theory of polyphase-DFT filter banks and their relation to windowed DFT's. In the third part, the performance of simple polyphase-DFT filter preprocessing is examined.

## II. Comparative Sensitivities

The best possible sensitivity that can be expected with the SETI sky survey will be achieved in the absence of radio frequency interference (RFI) with the use of a matched filter detector [4]. It is therefore of interest to calculate the matched filter-detection system sensitivity and use that as a standard to judge how well the SETI sky survey compares with the optimum. The flux sensitivity that can be achieved with a matched filter system is given in this section. This sensitivity is also compared with two systems with the same fixed number of channels: a Hanning windowed-DFT system of 30-Hz resolved equivalent noise bandwidth and an ideal bandpass filter bank of 20-Hz resolved bandwidth.

The SETI sky survey will sweep at a constant angular rate on the sky with a DSN 34-m high-efficiency (HEF) antenna. As the antenna transits a source on the sky, the source will be convolved by the antenna's beam shape. For this analysis, the authors assume that the antenna pattern sweeping past a CW point source can be approximated by a square-wave pulse of width equal to the half-power beamwidth (HPBW) of the antenna. The SNR for a matched filter system is then given by

$$SNR = \frac{A\phi\tau}{kT}$$

where $A$ is the effective area of the antenna, $\phi$ is the received flux, $\tau$ is the approximate time to move the telescope one HPBW, and $kT$ is the spectral noise density of the receiver and background. Most of these are system parameters. The authors assume a 34-m diameter antenna with a 60-percent effective collecting area and a system temperature of 25 K. The telescope is assumed to be driven at a constant rate of 0.2 deg per second across the sky. The constant scan rate gives a pulse duration, and hence a resolved matched filter bandwidth, which varies with frequency. The pulse duration time is approximately $\tau = 3/\nu$ sec, when $\nu$ is expressed in GHz.

To determine the minimum detectable flux density, one must first determine the minimum detectable SNR. Two factors determine the SNR required for a detection by the SETI sky survey. The first is a requirement limiting to 0.1 the probability of missing a signal. This parameter is the same for both the matched filter system and any system with a fixed number of channels. The second parameter required to determine the detectable SNR is the probability of false alarm ($P_{FA}$). Since a detection implies that a sky location and frequency should be reobserved, the amount of available look-back time is the main constraint on this parameter. The SETI sky survey requirement on $P_{FA}$ is formulated as 3,000 hits per 320-MHz survey over the entire sky. Since the antenna beamwidth varies with frequency, the number of resolvable points on the sky increases as the square of the frequency, decreasing the required $P_{FA}$ inversely in the same manner. Since the matched filter system has a resolved bandwidth that varies with frequency, the number of matched filter channels in a 320-MHz survey varies inversely with frequency. As a result, the required $P_{FA}$ for the matched filter system decreases approximately inversely with frequency, while the required $P_{FA}$ for a system with a fixed number of channels per bandwidth decreases as the inverse square of the frequency. There is, therefore, considerably more variation in the SNR required for detection for a system with a fixed number of channels than for a matched filter system.

Having defined the SNR's required for detection, as well as the effective area, the system temperature, and the pulse duration, the flux sensitivity $\phi$ can be computed. The matched filter system sensitivity can be straightforwardly calculated, as can the sensitivity for an ideal bandpass filter system, given a fixed number of channels. A windowed DFT system, however, has additional losses that must be considered if the signal-detection process is to be performed independently on individual DFT-channel outputs. A windowed DFT has two sources of loss relative to an ideal bandpass filter bank with identical channel spacing. These are shown in Fig. 1. The first is an expansion of the equivalent noise bandwidth of a DFT channel relative to the channel spacing. This loss is independent of the position of a signal within the DFT channel, and is approximately 1.76 dB for Hanning windowed DFT's. The second loss, commonly called "scallop loss," is due to the transfer function of a DFT bin. The scallop loss is dependent on the position of a signal within the DFT channel and varies from no loss to 1.42 dB, worst case, in the Hanning windowed DFT.

The sensitivities of the three systems can now be computed and are shown in Fig. 2. Losses of the ideal bandpass filter-bank system and the worst-case Hanning windowed DFT system are shown in Fig. 3. A number of conclusions can be drawn from these calculations. First, a sky survey with a fixed number of channels ranges from 1.9 to 7.5 dB less sensitive than the theoretical optimum. Second, the windowed-DFT approach is approximately 3 dB less sensitive than an ideal bandpass filter bank with the same channel spacing. Recovering the loss due to the windowed DFT would reduce the maximum loss of the sky survey relative to the optimum to approximately 4.5 dB. The authors therefore present an approximation to an ideal bandpass filter bank.

## III. Polyphase-DFT Filter Banks

Consider a bank of $K$ filters based on the same lowpass prototype impulse response $h(n)$, each centered at a different center frequency, $\omega_k = 2\pi k/K, k = 0, 1, \ldots,\ K-1$. Since the filters share the same prototype impulse response, but are just shifted in frequency by $\omega_k$, the impulse response of each is simply $h_k(n) = h(n)e^{j\omega_k n}$, and all have identically shaped passbands (centered at different frequencies, of course). Let

$$y_k(n) = \sum_{p=1}^{KN_t} x(n-p)h_k(p), \quad h_k(n) = h(n)e^{j2\pi kn/K}$$

where $KN_t$ is the length of the finite-impulse response filter $h(n)$. Note that if the $h(n)$ prototype filter is an ideal lowpass filter, with a passband $-\pi/K < \omega < \pi/K$, then this arrangement covers the spectrum with no losses and no overlapping bins. Furthermore, if the output is decimated $K$:1, as in Fig. 4, then there is no oversampling.

Let $p = \ell k - i$; $\quad i = 0, 1, \ldots,\ K-1$, and consider the $K$:1 decimated outputs of the filters so that $n = mK$. Then

$$y_k(mK) = \sum_{i-0}^{K-1} \sum_{\ell=1}^{N_t} x\left[(m-\ell)K + i\right] h_k(\ell K - i)$$

Keep in mind that only one out of every $K$ samples from each filter $h_k$ is selected. It is apparent that the output recognizes only a given input sample multiplied by every $K$th filter coefficient. The subset of the filter coefficients that multiplies a given input data point depends on the phase of the input data point relative to the $K$:1 decimation. To represent the double summation form above, the authors introduce the polyphase filter structure for the decimated filter $h_k$ shown in Fig. 5. The $i$th polyphase branch of the decimated filter $h_k$ is $\overline{p}_{i,k}$, defined as: $\overline{p}_{i,k}(m) = h_k(mK - i)i = 0, 1, \ldots, K-1$, and the branch input signals are $x_i(m) = x(mK + i)$.

Now, since $h_k(n) = h(n)e^{j(2\pi k/Kn)}$:

$$\begin{aligned} \overline{p}_{i,k}(m) &= h_k(mK-i) = h(mK-i)e^{j(2\pi k/K)}e^{-j\omega_k i} \\ &= h_k(mK-i)e^{j\omega_k i} = \overline{p}_{i,0}(m)e^{-j\omega_k i}; \\ &\quad i, k = 0, 1, \ldots,\ k-1 \end{aligned}$$

Note that all the filters are now defined in terms of outputs of the prototype filter's polyphase implementation. Hence:

$$y_k(mK) = \sum_{i=0}^{K-1} e^{-j(2\pi ki/K)} \sum_{l=1}^{N_t} x_i(m-\ell)\overline{p}_{i,0}(\ell)$$

For ease of notation, define $\overline{p}_i(m) = \overline{p}_{i,0}(m)$, and call the outputs of these prototype filter polyphase branches $z_i(m)$, where $i$ is the polyphase branch number. The resulting structure for filter $h_k$ (centered at $\omega_k$) is shown in Fig. 6.

The output is clearly: $y_k(m) = \sum_{i=0}^{K-1} z_i(m)e^{-j\omega_k i}$, which is bin $k$ of the DFT of the $z_i$'s for a fixed $m$ over the branch index $i$. Since the $z_i$'s are the same for all $K$ filters $h_k, k = 0, 1, \ldots, K-1$, the entire bank can be synthesized by computing the $z_i$'s and taking the DFT, as shown in Fig. 7.

In summary, by designing one lowpass filter of length $N = K(N_t)$ taps and dividing it up into $K$ polyphase branches of $N_t$ taps each, and by applying a $K$-point DFT to their outputs as shown above, one can synthesize a filter bank of equally spaced (on the DFT-bin center frequencies), identically shaped filters.

This concludes the derivation of the filter-bank structure; a more detailed explanation can be found in [5]. In Chapter 3, sections 3.3.2 and 3.3.3 of [5], Crochiere and Rabiner derive polyphase structures in general, and in Chapter 7, sections 7.2.1, 7.2.2, and 7.2.3, they describe uniform DFT filter banks for the "critically sampled" (number of filter bands, $K$ = decimation factor) case described here. An alternative derivation of the polyphase-DFT filter bank can be constructed by using an inverse DFT and a counterclockwise commutation of the input data points. An advantage of the forward-DFT approach is that the first polyphase branch to receive a data point each commutator cycle also provides the first data point to the DFT. In the inverse DFT approach, the first data point of the sequence on which the inverse DFT is to be performed is provided by the last polyphase branch to receive its input. This data ordering can be important in the design of real-time signal processors.

The additional processing necessary to produce a polyphase-DFT filter bank from a DFT is a separable, preprocessing step, similar to multiplication by a time-domain window function in the windowed DFT. In fact, the windowed DFT is a special, degenerate case of a polyphase-DFT filter bank. Consider the case where $N_t = 1$, then the prototype filter is of the same length as the DFT. The structure then becomes, in fact, an ordinary windowed DFT, with the window function being the set of filter coefficients.

$$\left[z_i(m) = h(i)\ \chi(mK + i)\right]$$

Since the polyphase structure is just a front-end process to a DFT, a polyphase system with programmable filter coefficients can easily be changed into a simple windowed DFT. This is accomplished by setting all the coefficients to zero except for one tap per polyphase branch, e.g., all except tap 1 of each branch. Furthermore, any finite impulse response (FIR) polyphase-DFT filter bank can conceptually be converted to an equivalent windowed DFT as follows:

(1) Take the samples of the impulse response of the prototype lowpass filter as the window coefficients.

(2) Perform an $N_t K$-point windowed DFT by using consecutive time samples.

(3) Take only the outputs of bins whose indexes are integer multiples of $N_t$, i.e., bins $N_t, 2N_t, 3N_t, \ldots$, and discard the other bin outputs.

(4) Continue performing steps 2 and 3, and shift the start of the input sequence to the windowed DFT by $K$ samples each time so that the first DFT is on samples $x(0)$ to $x\left[N_t(K-1)\right]$, the second DFT is on samples $x(K)$ to $x\left[(N_t+1)(K-1)\right]$, and the $i$th DFT is on samples $x\left[(i-1)K\right]$ to $x\left\{\left[N_t + (i-1)\right](K-1)\right\}$.

The reader will notice that while this method produces the same output as the polyphase-DFT filter bank, it requires, in general, a large amount of computation to do so, and is not a recommended approach.

## IV. Performance of Polyphase-DFT Filter Banks

Examples of polyphase-DFT filter-bank performance appear in Table 1. Note that the total number of inputs (the time aperture) affecting each output is $KN_t$ samples; however, for a fixed resolution in absolute frequency (hertz), the prototype lowpass filter $h(n)$ must approximate a truncated sync function whose major weighting (main lobe, between the $-1$ and $+1$ nulls) is of a *fixed time duration* proportional to the reciprocal of the desired lowpass bandwidth. Hence, the impulse response of this longer time aperture looks and acts like an ideal lowpass filter with a truncated time aperture. The advantage of this longer time aperture is that it allows the filter-transfer function to have a flatter passband and sharper transition, which provides both increased sensitivity to desired signals as well as increased interference immunity. An example of an $N_t = 8$-tap polyphase-filter transfer function versus a Hanning windowed transfer function is shown in Fig. 8.

Some quick filter designs, with Parzen (Riesz) windowed lowpass filters, produced the following worst-case processing losses, as defined by Harris [2]:

Worst-Case Loss (dB)

$$= 10\log\left[\frac{\text{Equivalent Noise Bandwidth}}{\text{Input Bandwidth/Number of Channels}}\right]$$

$$+ \text{Minimum Gain } (\pm 0.5 \text{ bins offset})$$

As usual with digital FIR filter design, optimization can be performed to trade off sidelobe levels for worst-case losses. Transfer functions for these filter designs are shown in Figs. 9–16. Table 2 shows sample responses for these Parzen smoothed lowpass filters.

Computationally the $N_t$-tap polyphase structure requires $N_t$ multiplications and $N_t - 1$ additions per real data-point input. In comparison, a $K$-point radix-2 fast Fourier transform (FFT) requires $1.5(\log_2 K)$ real additions and $\log_2 K$ real multiplications, for a total of $2.5(\log_2 K)$ operations per real data-point input. A 12-tap filter is, therefore, slightly less computationally intensive than a 1,024-point radix-2 FFT. For large FFT systems, such as the SETI sky survey system, the polyphase-DFT filter bank offers a method of significantly improving the system with a relatively small computational increase. However, since the polyphase preprocessing stage must store $N_t$ vectors, each of which is the length of the FFT to be performed, the memory requirements for a large polyphase-DFT filter bank, such as SETI's, can be significant. The required data storage for a $K$-channel filter bank is $N_t K$ times the input word length. For a 4-Mchannel filter bank with 8-bit complex inputs, such as SETI's, this results in $8N_t$ Mbytes of memory just for data. In addition, filter coefficients must be stored, but these can often take advantage of symmetries and simple data-compression techniques.

## V. Conclusions

Like the application of a window to the sequence prior to the DFT in windowed DFT's, the additional processing involved in polyphase-DFT filter banks is a preprocessor to the conventional DFT. In fact, the conventional windowed DFT is the simplest case of a polyphase-DFT filter bank. Moreover, the polyphase-DFT filter bank can operate on a time aperture larger than the conventional DFT. As a result, it is possible to easily synthesize DFT-bin (window) transfer functions with significantly less frequency-scallop loss, less noise-bandwidth expansion, and faster sidelobe falloff than is possible with conventional windowed-DFT techniques. Furthermore, the computational cost is generally minimal as compared with that of the DFT. For example, the computational cost of a 64-point radix-2 FFT produces a transfer function with a worst-case processing loss of 1 dB, as defined by Harris in his analysis of windows [2]. The best windows in [2] have a worst-case loss of 3 dB. By recovering much of the processing loss inherent in windowed-DFT spectrum analysis as well as providing superior isolation of narrowband interference, polyphase-DFT spectrum analysis provides a relatively inexpensive means for the SETI sky survey to increase its sensitivity by at least 2 dB, which provides a sensitivity that ranges between 2.9 to 5.5 dB from optimal.

# References

[1] G. A. Zimmerman, B. Charny, M. F. Garyantes, and M. J. Grimm, "A 640-MHz 32-Megachannel Real-Time Polyphase-FFT Spectrum Analyzer," *TDA Progress Report 42-107*, vol. October–December 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 132–140, November 15, 1991.

[2] F. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, January 1978.

[3] M. P. Quirk, H. C. Wilck, M. F. Garyantes, and M. J. Grimm, "A Wideband, High-Resolution Spectrum Analyzer," *TDA Progress Report 42-93*, vol. January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 188–198, May 15, 1988.

[4] B. Oliver, "Parametric Relations in a Full Sky Search," *NASA SP 419*, NASA, Washington, D.C., p. 129, 1977.

[5] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall Signal Processing Series, Englewood Cliffs, New Jersey: Prentice-Hall, 1983.

**Table 1. Examples of polyphase-DFT filter-bank performance**

| Taps per branch, $N_t$ | Worst-case processing loss, dB | Interference rejection: offset bins to >70-dB attenuation |
|---|---|---|
| 2 | 2.503 | 7 |
| 4 | 1.513 | 4 |
| 6 | 1.236 | 3 |
| 8 | 1.048 | 3 |
| 12 | 0.653 | 2 |
| 16 | 0.543 | 2 |

**Table 2. Sample responses for Parzen smoothed lowpass filters**

| Taps per branch, $N_t$ | Min. gain at 0.5 bin, dB | Equivalent noise bandwidth, bins/Hz | Min. gain, dB | Worst-case loss, dB |
|---|---|---|---|---|
| 2 | −0.492 | 1.589/30.302 | −0.492 | 2.503 |
| 4 | −0.244 | 1.335/25.460 | −0.258 | 1.513 |
| 6 | −0.128 | 1.239/23.638 | −0.304 | 1.236 |
| 8 | −0.077 | 1.184/22.581 | −0.315 | 1.048 |
| 12 | −0.295 | 1.083/20.649 | −0.308 | 0.653 |
| 16 | −0.252 | 1.058/20.181 | −0.298 | 0.543 |
| 1[a] | −3.908 | 1.016/19.372 | | 3.976 |
| 1[b] | −1.423 | 1.516/28.908 | | 3.229 |

[a] Rectangular window.
[b] Hanning window.

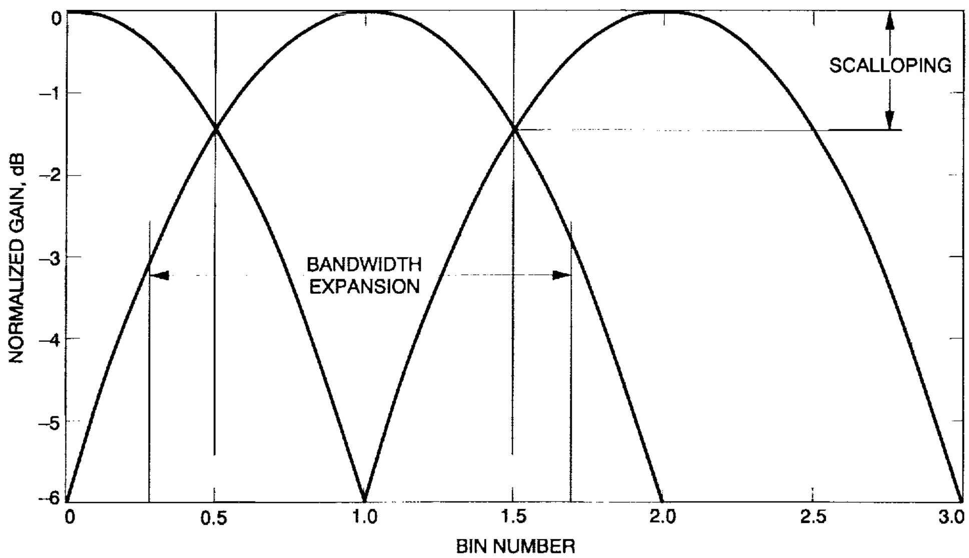


**Fig. 1. Losses in Hanning windowed discrete Fourier transforms (DFTs).**

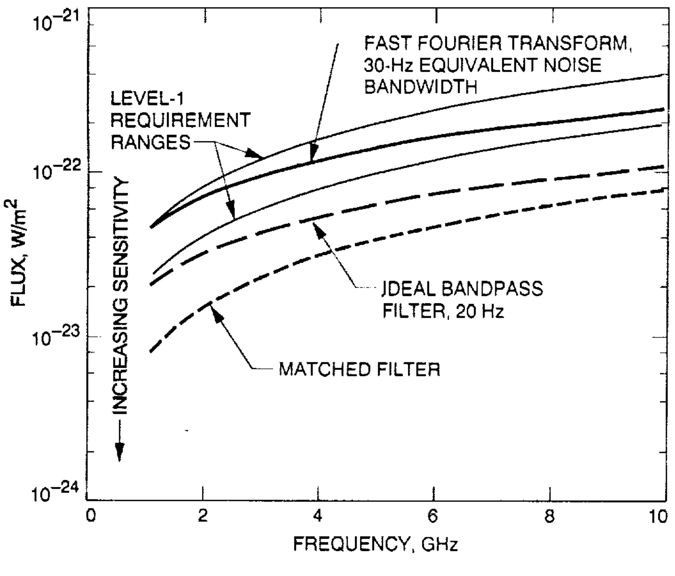


**Fig. 2. A comparison of system sensitivities.**

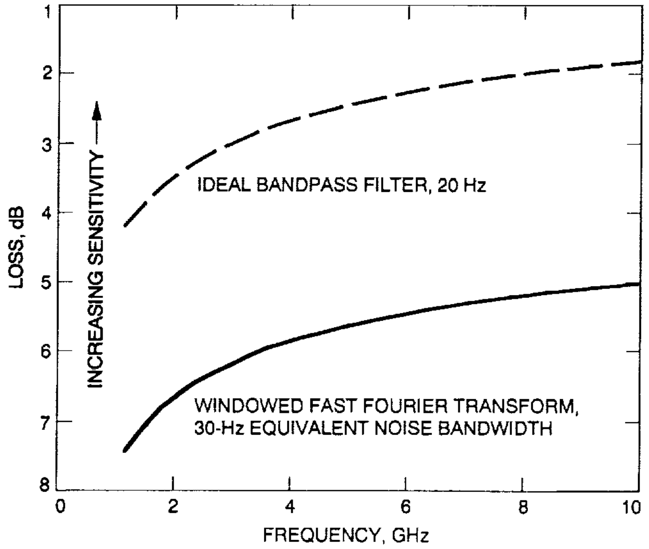


**Fig. 3. A comparison of losses relative to the theoretical optimum.**

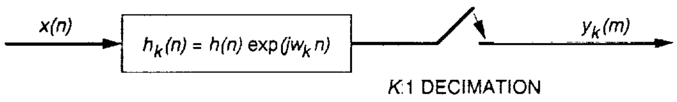


**Fig. 4. A *K*:1 decimated bandpass filter from a lowpass prototype.**

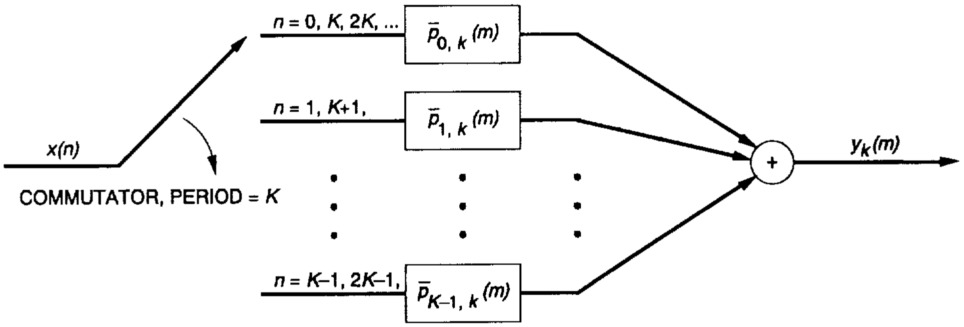


Fig. 5. General polyphase structure for the $K$:1 decimated filter $h_k$.

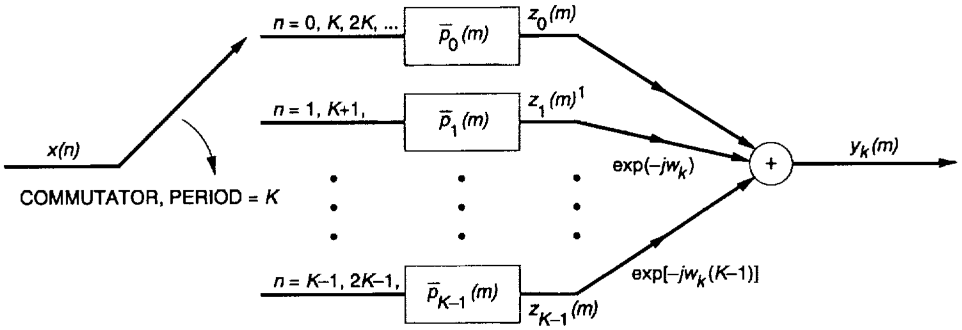


Fig. 6. Polyphase structure for the $k$th $K$:1 decimated filter $h_k$, with center frequency $\omega_k = 2\pi k/K$.

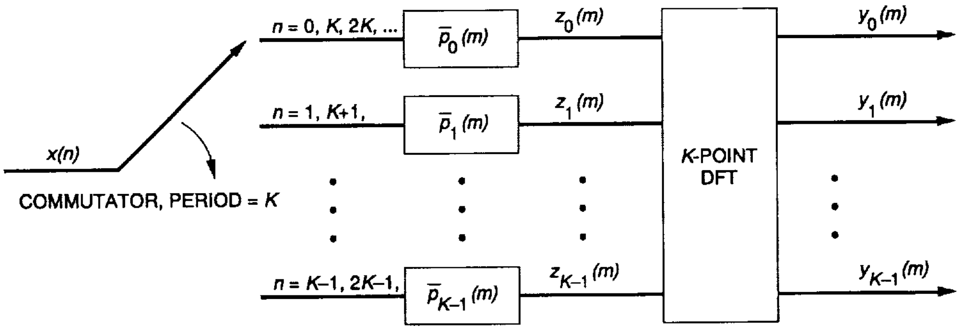


Fig. 7. Discrete Fourier transform implementation for a bank of identically spaced, identically shaped filters.

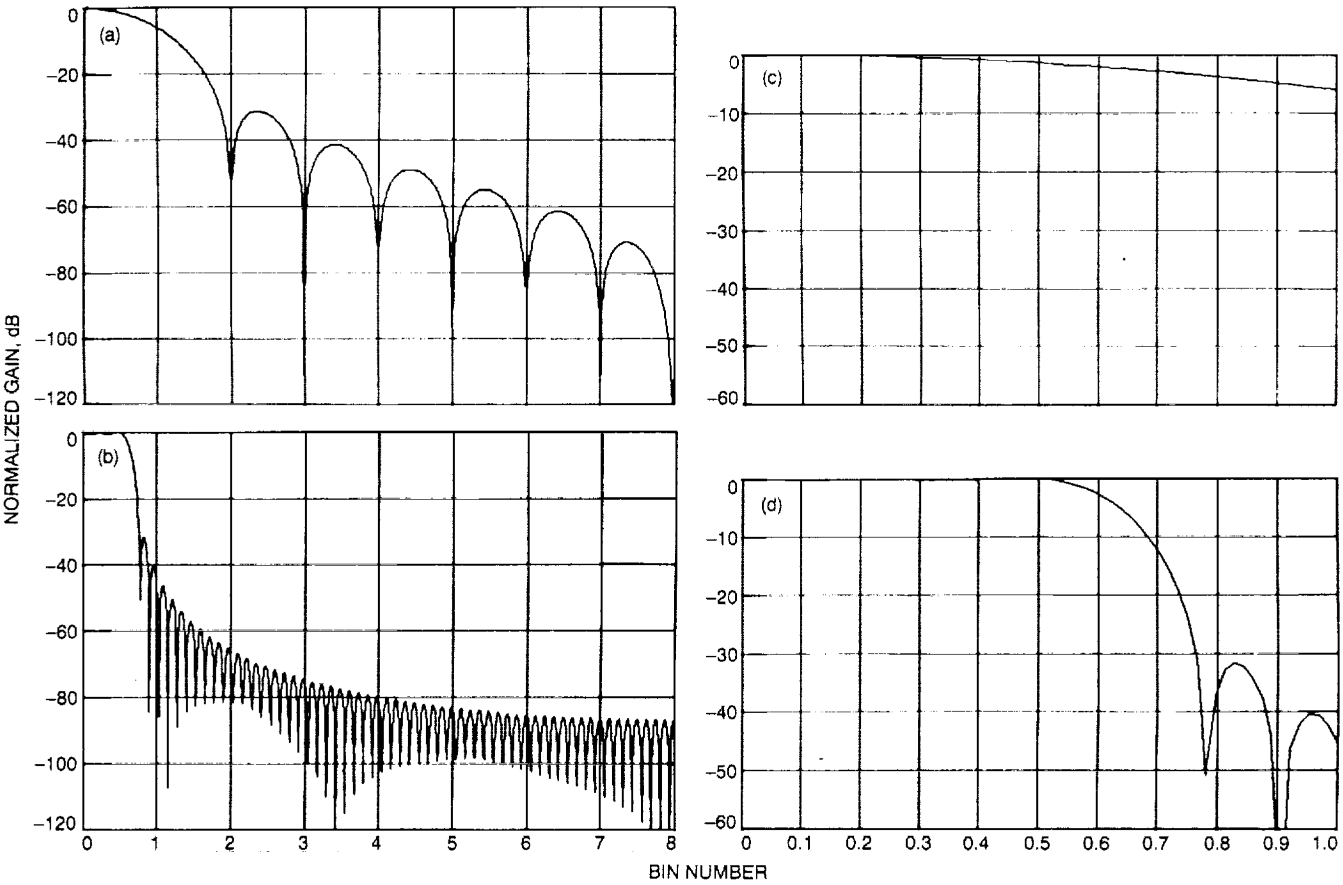


**Fig. 8. A comparison of a Hanning windowed discrete Fourier transform with an 8-tap polyphase discrete Fourier transform: (a) Hanning window; (b) 0.65-bin lowpass filter, Parzen smoothed; (c) Hanning window, close up; and (d) 0.65-bin lowpass filter, Parzen smoothed, close up.**

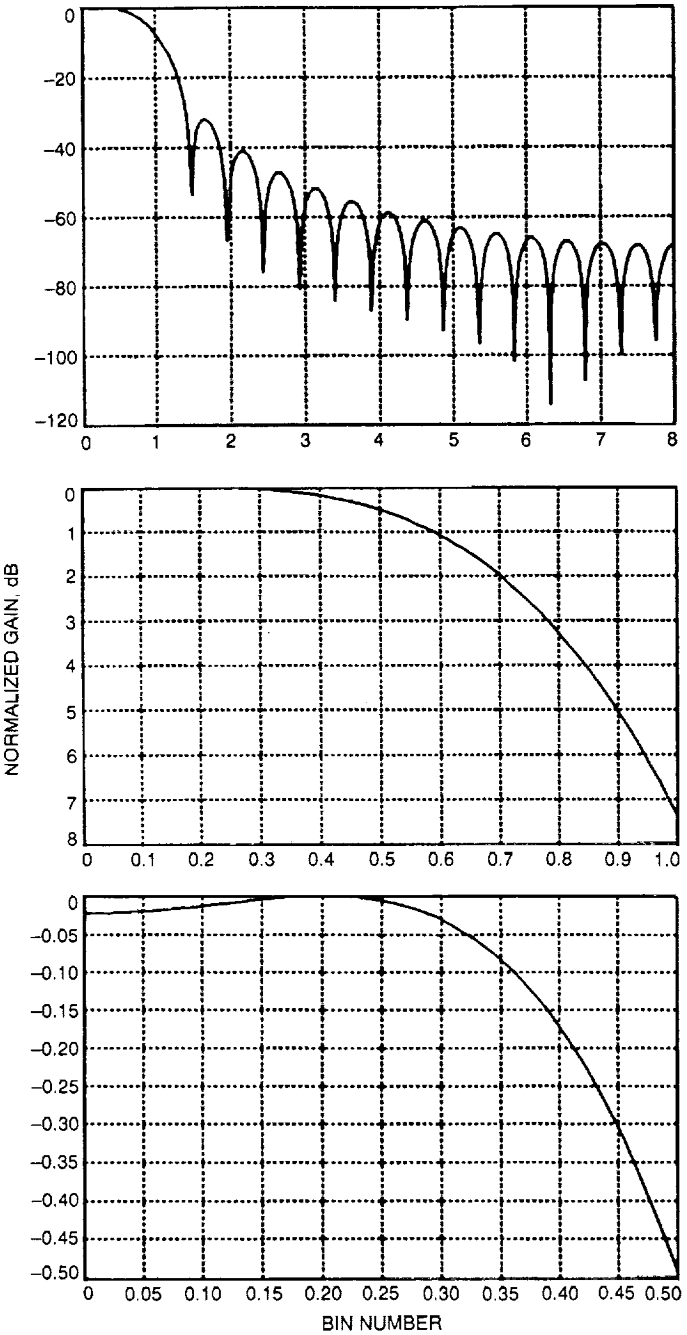


**Fig. 9. Sample response with two close-up views of a 2-tap, 0.95-bin Parzen smoothed lowpass filter.**

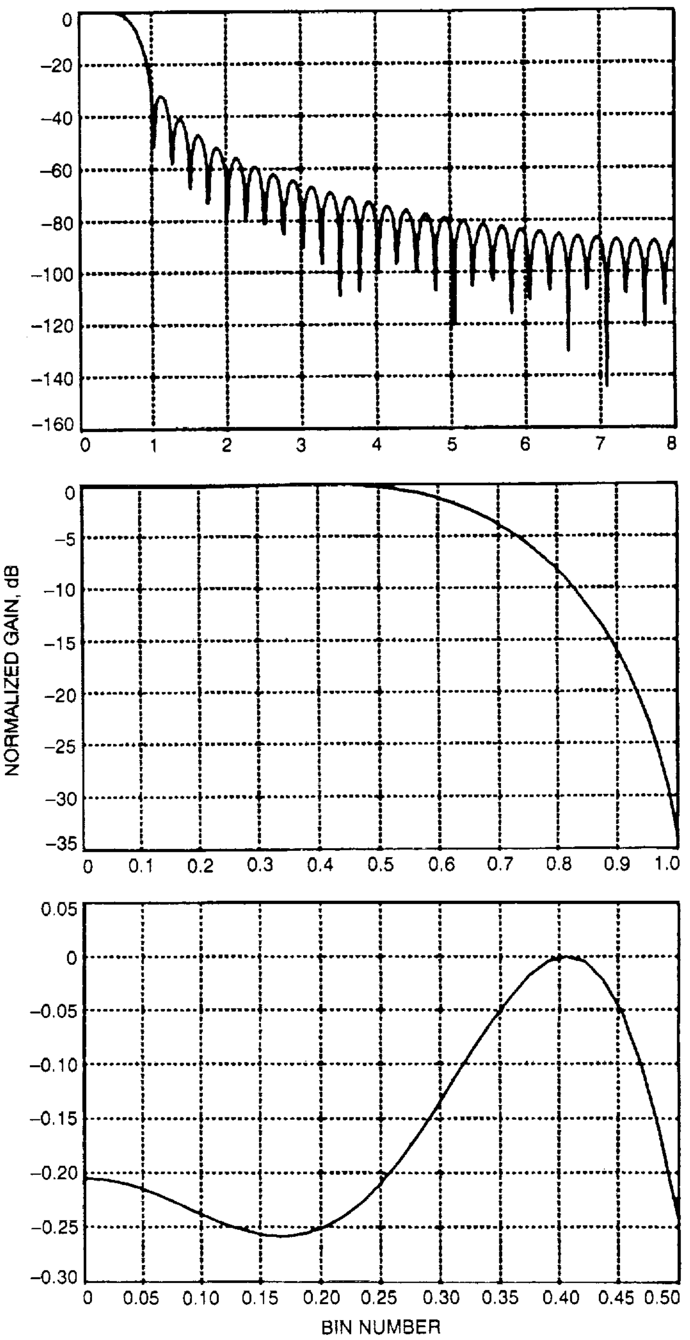


**Fig. 10. Sample response with two close-up views of a 4-tap, 0.76-bin Parzen smoothed lowpass filter.**

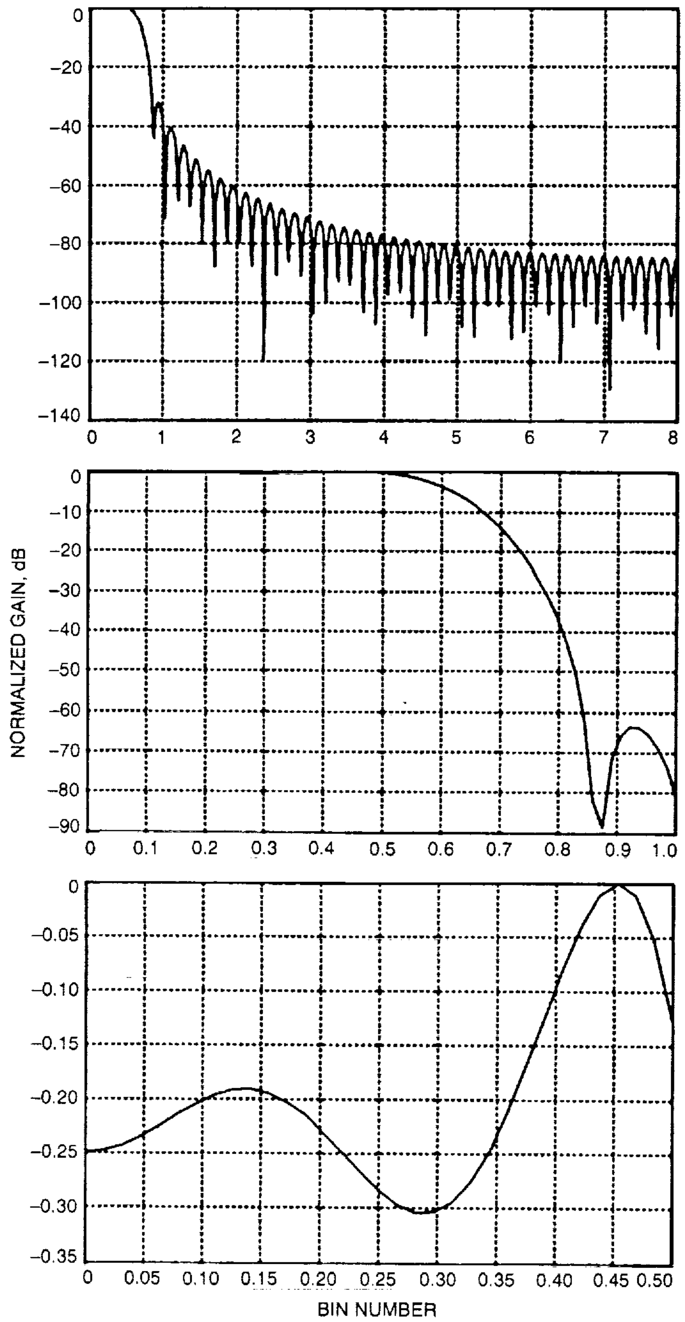


**Fig. 11. Sample response with two close-up views of a 6-tap, 0.69-bin Parzen smoothed lowpass filter.**

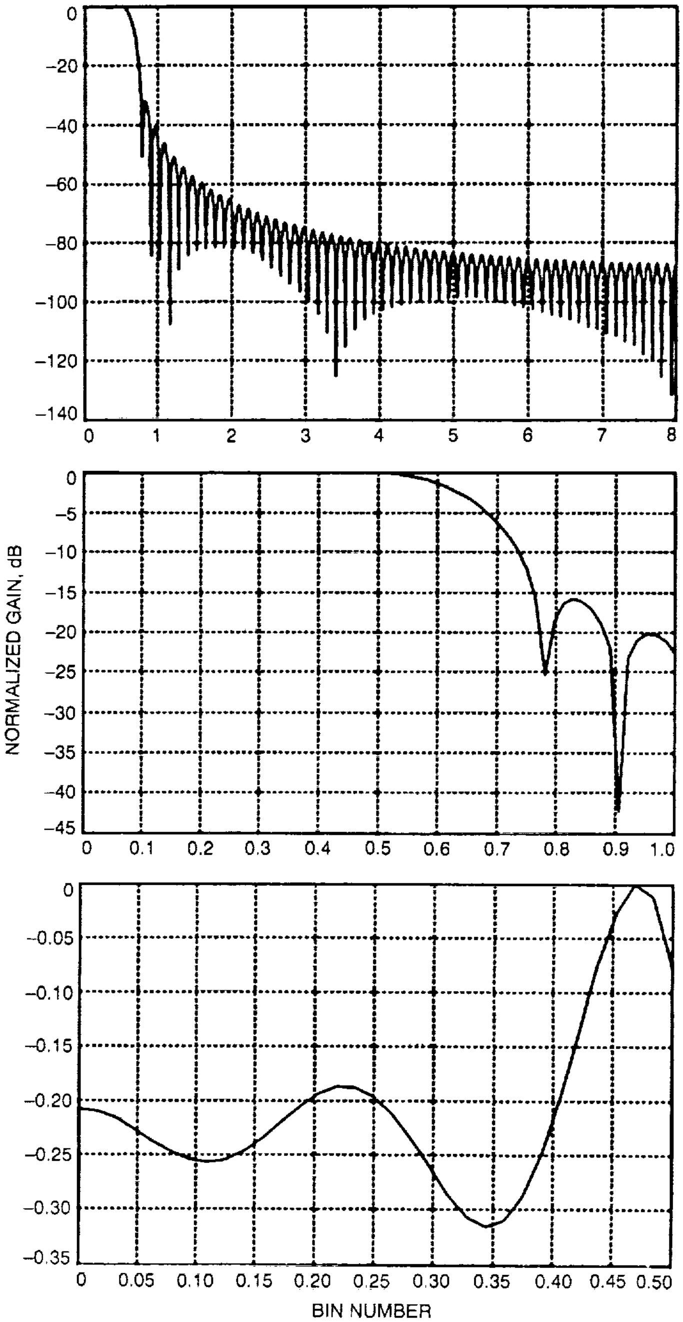


**Fig. 12. Sample response with two close-up views of an 8-tap, 0.65-bin Parzen smoothed lowpass filter.**

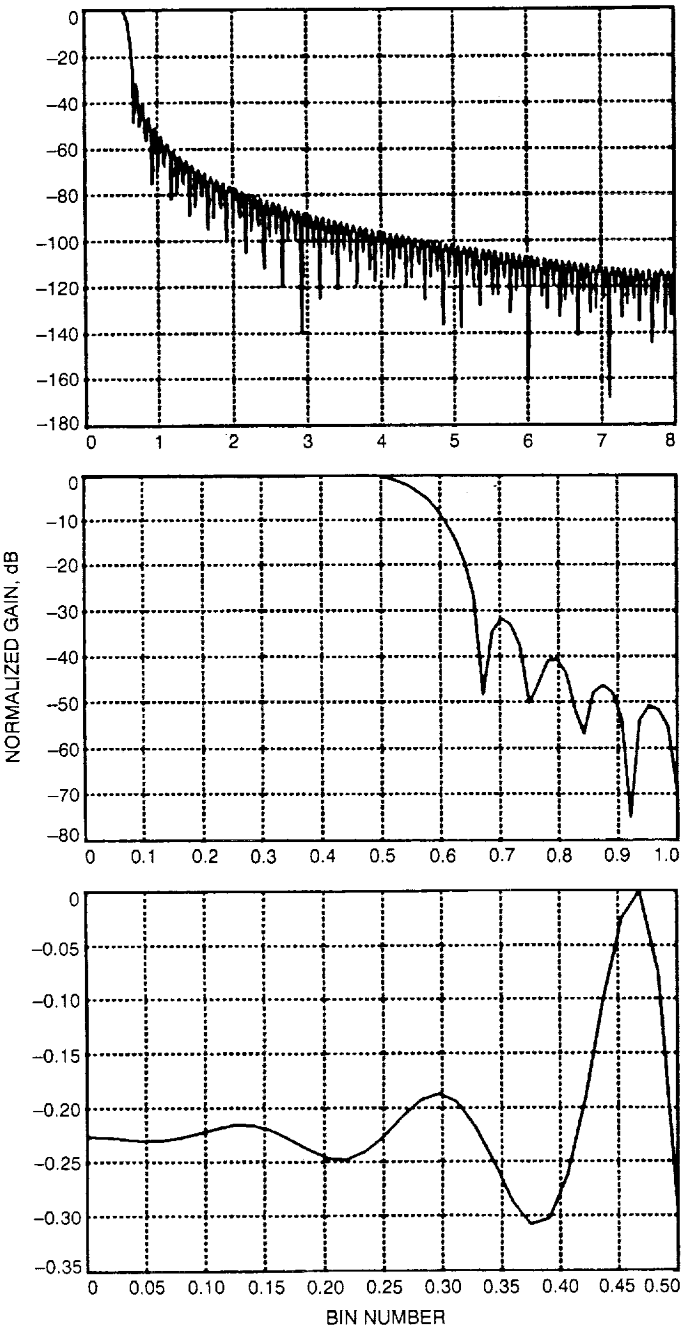


**Fig. 13. Sample response with two close-up views of a 12-tap, 0.585-bin Parzen smoothed lowpass filter.**

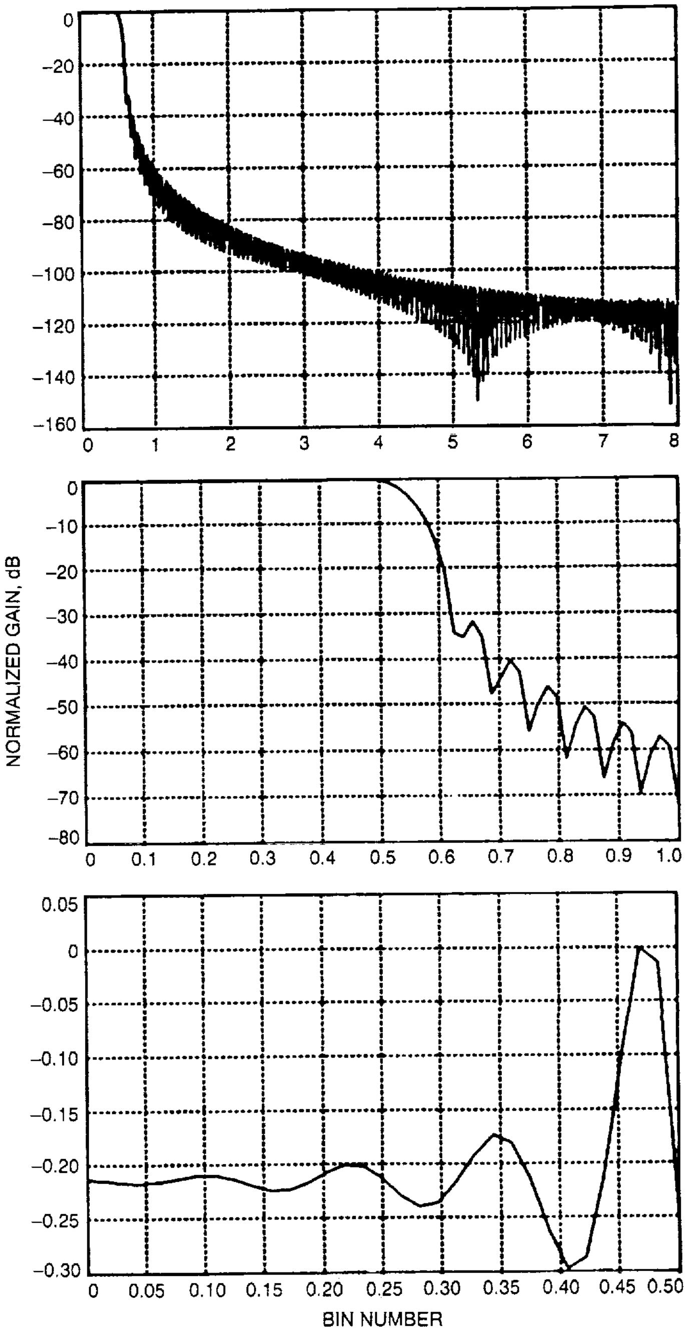


**Fig. 14. Sample response with two close-up views of a 16-tap, 0.565-bin Parzen smoothed lowpass filter.**

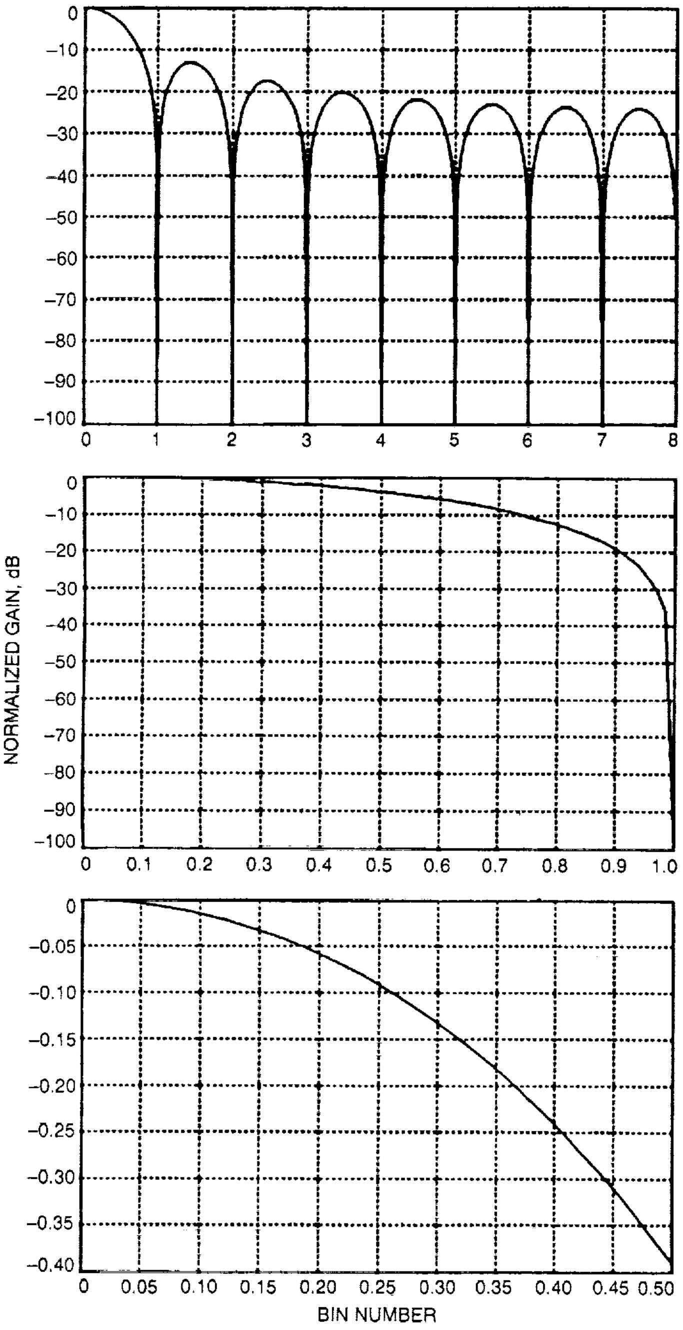


**Fig. 15. Sample response with two close-up views of a rectangular window.**

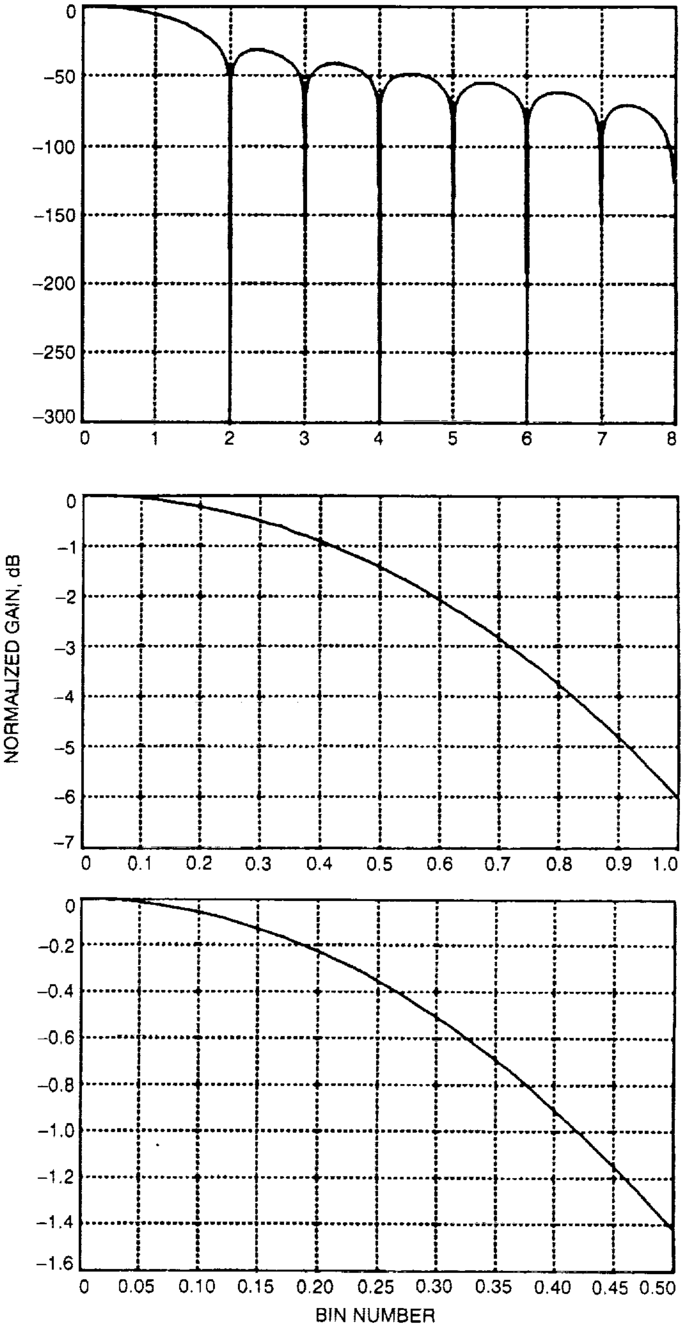


**Fig. 16. Sample response with two close-up views of a Hanning window.**

5/5-32

53218

P-14

# The Behavior of Quantization Spectra as a Function of Signal-to-Noise Ratio

M. J. Flanagan[1]
Communications Systems Research Section

*An expression for the spectrum of quantization error in a discrete-time system whose input is a sinusoid plus white Gaussian noise is derived. This quantization spectrum consists of two components: a white-noise floor and spurious harmonics. The dithering effect of the input Gaussian noise on both components of the spectrum is considered. Quantitative results in a discrete Fourier transform (DFT) example show the behavior of spurious harmonics as a function of the signal-to-noise ratio (SNR). These results have strong implications for digital reception and signal analysis systems. At low SNRs, spurious harmonics decay exponentially on a log-log scale, and the resulting spectrum is white. As the SNR increases, the spurious harmonics figure prominently in the output spectrum. A useful expression is given that roughly bounds the magnitude of a spurious harmonic as a function of the SNR.*

## I. Introduction

This work was inspired by consideration of a 2-million channel spectrum analyzer built by the Digital Projects Group of the Communications Systems Research Section [1]. This spectrum analyzer is a prototype of a larger system that will be used in the sky-survey portion of the Search for Extraterrestrial Intelligence (SETI) project. After computer simulations were performed, 8-bit input quantization was observed to pose the greatest limitation to the dynamic range of the spectrum analyzer. This is because quantization is a nonlinear process that generates spurious harmonics in the spectrum of the quantizer output.

Previous work by Bennett [2] considered the spectra of quantized signals when the system input has "energy uniformly distributed throughout a definite frequency band and with the phases of the components randomly distributed." Hurd [3] developed an expression for the correlation function of a quantized sine wave plus Gaussian noise and examined the case where the input noise spectrum is rectangular narrow-band and the signal-to-noise ratio (SNR) is small. Quantization error spectra are most commonly assumed to be white [4]. This article derives an

[1] The author is also a graduate student in Electrical Engineering at the California Institute of Technology.

expression for the spectrum of a quantized sine wave plus white Gaussian noise. An SNR transition region where the spectrum goes from being filled with spurious harmonics to white is presented. This transition is due to the dithering effect of the input Gaussian noise. A rule of thumb is given bounding the size of a spurious harmonic as a function of the SNR. Implications for digital reception and signal analysis systems are considered.

## II. Power Spectrum of the Quantization Error

Consider the quantizer system in Fig. 1 with input $x$ and output $y$. One can write:

$$y \; = \; Q[x] \; = \; x \; - \; e \tag{1}$$

where $Q[\,]$ is the quantization operator and $e$ is the quantization error. When $Q[\,]$ is a uniform mid-tread symmetric quantizer with a staircase input–output relation as in Fig. 2, $e$ can be expressed as a sawtooth function of $x$ as in Fig. 3. Assuming an infinite quantizer (or equivalently, no quantizer saturation), one can write a Fourier series expansion for $e(x)$ as in [5]:

$$e(x) \; = \; -\frac{\Delta}{j2\pi} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^k}{k} \exp\left(\frac{j2\pi kx}{\Delta}\right) \tag{2}$$

Now consider the system in Fig. 4 where the input is $A\sin(\omega_0 t + \phi) + z(t)$ and $z(t)$ is zero mean, Gaussian noise with variance $\sigma^2$. The continuous-time signal $e(t)$ can be written as

$$e(t) =$$

$$-\frac{\Delta}{j2\pi} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^k}{k} \exp\left(\frac{j2\pi kx}{\Delta}\left(A\sin\left(\omega_0 t + \phi\right) + z\left(t\right)\right)\right) \tag{3}$$

The autocorrelation function of $e(t)$ is defined as

$$R_e(t, t+\tau) \; = \; E\{e(t)e(t+\tau)\} \tag{4}$$

Returning to Fig. 4, one sees that $x[n] = e(nT)$ is a discrete-time random process. The autocorrelation function of $x[n]$ can be expressed as

$$\begin{aligned} R_x[n, n+k] &= E\{x[n]x[n+k]\} \\ &= E\{e(nT)e(nT+kT)\} = R_e(nT, nT+kT) \end{aligned} \tag{5}$$

If the phase $\phi$ of the input sinusoid is a random variable uniformly distributed between 0 and $2\pi$, $\overline{R}_x[k]$ can be computed as

$$\overline{R}_x[k] \; = \; \frac{1}{2\pi}\int_0^{2\pi} R_x[n, n+k]d\phi \tag{6}$$

Leaving the details to Appendix A, one obtains

$$\overline{R}_x[k] \; = \; \begin{cases} \dfrac{\Delta^2}{12} + \dfrac{\Delta^2}{\pi^2}\displaystyle\sum_{l=1}^{\infty}\frac{(-1)^l}{l^2}\exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right) J_0\left(\frac{2\pi Al}{\Delta}\right); & k = 0 \\ \displaystyle\sum_{\substack{n-\infty \\ n \ odd}}^{\infty} S_n \exp(jn\omega_0 Tk); & k \neq 0 \end{cases} \tag{7}$$

where

$$S_n \; = \; \frac{\Delta^2}{\pi^2}\left(\sum_{l=1}^{\infty}\frac{(-1)^l}{l}\exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right) J_n\left(\frac{2\pi Al}{\Delta}\right)\right)^2 \tag{8}$$

and $J_n(x)$ is the $n$th-order Bessel function. The power spectrum of the discrete-time random process $x[n]$ is

$$S(\omega) \; = \; \sum_{k=-\infty}^{\infty} \overline{R}_x[k]e^{j\omega Tk} \tag{9}$$

Finally, the power spectrum can be expressed as

$$S(\omega) = N_Q + \sum_{\substack{n-\infty \\ n\ odd}}^{\infty} 2\pi S_n \delta((\omega T - n\omega_0 T) \bmod 2\pi) \quad (10)$$

where

$$N_Q = \frac{\Delta^2}{12} + \left( \frac{\Delta^2}{\pi^2} \sum_{l=1}^{\infty} \frac{(-1)^l}{l^2} \exp\left( - \frac{2\pi^2\sigma^2}{\Delta^2} l^2 \right) \right.$$

$$\left. \times J_0\left( \frac{2\pi A l}{\Delta} \right) \right) - \sum_{\substack{n-\infty \\ n\ odd}}^{\infty} S_n$$

and $S_n$ is defined in Eq. (8) (details of this derivation are left to Appendix A). The $N_Q$ term in Eq. (10) represents a frequency-independent quantization noise floor (see Appendix A). The infinite sum in Eq. (10) specifies the phase-averaged magnitude and location of all spurious harmonics in the frequency domain. Depending on the value of $\omega_0 T$, spurious harmonics can be spread throughout the frequency domain or lie concentrated at only a few frequencies. This complicates digital reception of weak signals in the presence of stronger interferers. While a strong interferer may be easily identified and filtered out, the spurious harmonics generated by a strong interferer would require more complex filtering techniques. A spectrum (taken from a discrete Fourier transform) with spurs is shown in Fig. 5.

When $\sigma > \Delta$,

$$S(\omega) \;=\; \frac{\Delta^2}{12} + O\left( \exp\left( \frac{-2\pi^2\sigma^2}{\Delta^2} \right) \right) + < \text{spur term} >$$

where the <spur term> is a delta function with weight zero or weight $O(\exp(-4\pi^2\sigma^2/\Delta^2))$, depending on whether or not a spur was located at that frequency. These results are consistent with those in [5].

With the exception of the $\Delta^2/12$ term, all the components in the power spectrum in Eq. (10) are dependent on the ratio $\sigma/\Delta$. When this ratio is much greater than 1, the resulting spectrum is essentially $\Delta^2/12$ and white. In this manner, the input Gaussian noise has a dithering effect on the spectrum. For other values of $\sigma/\Delta$, it is not immediately obvious how the spectrum will appear. For this reason, an example involving a discrete Fourier transform (DFT) is presented in the next section.

## III. DFT Example

This example provides a quantitative analysis of the manner in which spurious harmonics are dithered due to additive white Gaussian noise. In particular, an SNR region where the magnitude of a spurious harmonic decays exponentially on a log-log scale is presented. Consider the DFT of the signal $x[n]$ in Fig. 4. One can write

$$E\{|X[k]|^2\} = \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{r=0}^{N-1} E\{x[n]x[r]\} e^{\frac{-j2\pi k(n-r)}{N}} \quad (11)$$

For the purpose of exposition, the $\omega_0 T$ product is chosen to equal $\pi/8$, and the value of $k$ is $3N/16$. Thus, a case is analyzed where the infinite number of spurious harmonics are aliased into only a few frequency bins. In this example, the phase is not treated as a random variable as was done in the previous section. Treating the phase as a constant allows a more general phase-dependent solution to be obtained.

By observing bin $3N/16$, one is examining the spectral sample that contains the spurious harmonic specified by an arrow in Fig. 5. The spectrum in Fig. 5 was generated using the following parameters: $A = 0.5$, $\phi = 0.8147576$, $N = 1024$, $\Delta = 1/127$, and SNR $= 10\log_{10}(A^2/2\sigma^2) =$ 50 dB.

Leaving the details to Appendix B, one can evaluate Eq. (11):

$$Y \;=\; E\{|X[3N/16]|^2\} = N_{Q1} + N_{Q2} + N_{Q3} + S_Q \quad (12)$$

where

$$N_{Q1} \;=\; \frac{\Delta^2}{12N} \quad (13)$$

and

$$\left.\begin{aligned}
N_{Q2} &= \frac{\Delta^2}{16\pi^2 N}\sum_{m=0}^{7}\sum_{l=1}^{\infty}\frac{(-1)^l \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right)}{l^2}\left[\cos\left(\frac{2\pi A\Theta_m l}{\Delta}\right)+\cos\left(\frac{2\pi A\Xi_m l}{\Delta}\right)\right]\\
N_{Q3} &= -\frac{\Delta^2}{16\pi^2 N}\sum_{m=0}^{7}\left[\left(\sum_{l=1}^{\infty}\frac{(-1)^l \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right)}{l}\sin\left(\frac{2\pi A\Theta_m l}{\Delta}\right)\right)^2\right.\\
&\quad\left.+\left(\sum_{l=1}^{\infty}\frac{(-1)^l \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right)}{l}\sin\left(\frac{2\pi A\Xi_m l}{\Delta}\right)\right)^2\right]\\
S_Q &= \frac{\Delta^2}{\pi^2}\left|\sum_{l=1}^{\infty}\frac{(-1)^l \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right)}{l}F\left(16\pi,3,16,\frac{2\pi A l}{\Delta}\right)\right|^2
\end{aligned}\right\}\qquad (14)$$

with $F(\ )$, $\Theta_m$, and $\Xi_m$ as defined in Appendix C.

The quantities in Eq. (12) are easily computed. In fact, for the SNRs of interest, only a small number of terms is required to adequately represent the infinite sums over $l$ (see Appendix D for more details). The spectral sample $Y$ consists of two types of components: a white-noise-floor term and a spurious-harmonic term. As seen in Appendix B, the values of $N_{Q_1}$, $N_{Q_2}$, and $N_{Q_3}$ are independent of bin number (frequency), and thus represent a white-noise floor. As a check, these quantities obey the $1/N$ (where $N$ is the number of points in the DFT) processing rule for white noise in a DFT. The value $S_Q$ represents the spurious harmonic component of the spectral sample. This term is independent of the DFT size $N$ and is responsible for the spectral sample becoming large when the SNR is high. In effect, this limits the dynamic range of the digital spectrum analyzer.

Figures 6(a) and (b) show how the value of the spectral sample changes as the SNR varies. The units are decibels relative to the carrier (dBc). The following parameters are held constant: $A = 0.5$, $\phi = 0.8147576$, $N = 1024$, and $\Delta = 1/127$. The quantizer step size $\Delta$ was chosen to simulate an 8-bit input quantizer ($\Delta \approx 2^{-B+1}$ where $B = 8$). An experimental curve (simulation) is presented with the theoretical curve to show the excellent agreement. For each SNR used in the experiments, 10,000 spectra were accumulated and rescaled.

The horizontal line in Fig. 6(a) indicates the value of $N_{Q_1}$. This would be the value of the spectral sample under traditional quantization error assumptions [4]. Below 40 dB SNR (for the 8-bit input quantizer), the value of the spectral sample $Y$ reduces to essentially $N_{Q_1}$. Above 70 dB SNR, the value of the spectral sample does not change by more than a few decibels. The transition region in Fig. 6(a) coincides with $\sigma/\Delta$ approaching and exceeding unity.

Figure 7 plots $S_Q$, the spurious harmonic portion of the spectral sample, as a function of SNR. The units are decibels relative to the carrier (dBc) with the same conditions as above. When $\sigma/\Delta$ approaches unity, this value decays exponentially on the log-log scale. From Eq. (14), one can estimate the behavior of $S_Q$ when $\sigma/\Delta$ exceeds unity to obtain a rule of thumb. Assume that the first term in the infinite sum over $l$ in Eq. (14) dominates, and bound $F(\ )$ by its maximum value (from Appendix C, $|F(\ )| \leq 1$). This provides a rough bound on the maximum value of $S_Q$. In particular, compute $S_{QdBc} = 10\log_{10}(2S_Q/A^2)$. Recall that $\mathrm{SNR} = 10\log 10(A^2/2\sigma^2)$. After manipulating,

$$S_{QdBc} = 10\log_{10}\left(\frac{2\Delta^2}{\pi^2 A^2}\right) - \frac{20\pi^2}{ln 10}\frac{A^2}{\Delta^2}10^{\frac{-SNR}{10}} \qquad (15)$$

This expression is easily inverted to express the SNR as a function of $S_{QdBc}$. As seen in Fig. 7, the rule of thumb in Eq. (15) nicely describes the behavior of $S_Q$ even when $\sigma/\Delta$ is less than one. Asymptotically, Eq. (15) levels off at high SNRs to a constant value, which is consistent with empirical observations. This expression should prove useful to system designers concerned with dynamic-range limitations imposed by input quantization.

## IV. Conclusions

Spurious harmonics pose complex filtering problems for digital reception systems. These harmonics also limit the dynamic range of digital spectrum analyzers. Expressions have been obtained describing the spectrum of quantization error when the input is a noisy sinusoid. An example involving a DFT has provided quantitative information about the behavior of spurious harmonics in the frequency domain as a function of the SNR. The input Gaussian noise dithers the output spectrum when the ratio $\sigma/\Delta$ exceeds 1, where $\sigma^2$ is the noise variance, and $\Delta$ is the step size in the input quantizer. A useful rule of thumb has been derived that roughly bounds the magnitude of a spurious harmonic as a function of the SNR.

## Acknowledgment

Simulations in this report were done in part on the JPL-CRAY. The author thanks the Supercomputer User Group for their assistance.

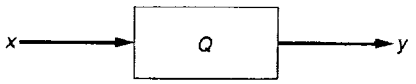


Fig. 1. Quantizer system diagram.

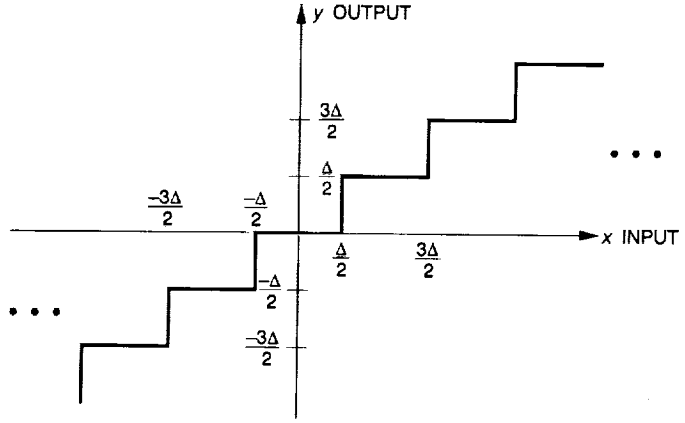


Fig. 2. Quantizer input–output relation.

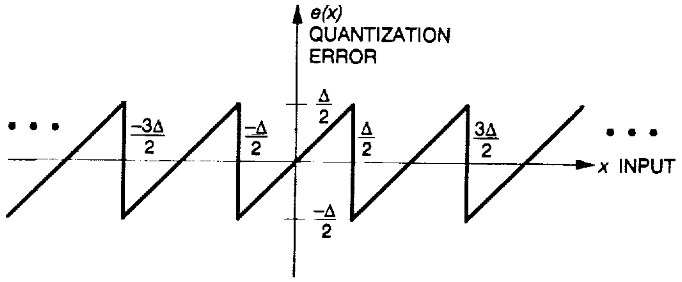


Fig. 3. Quantization error input–output relation.

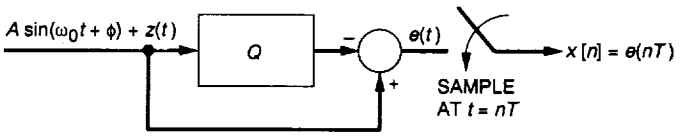


Fig. 4. Generation of discrete-time quantization error.

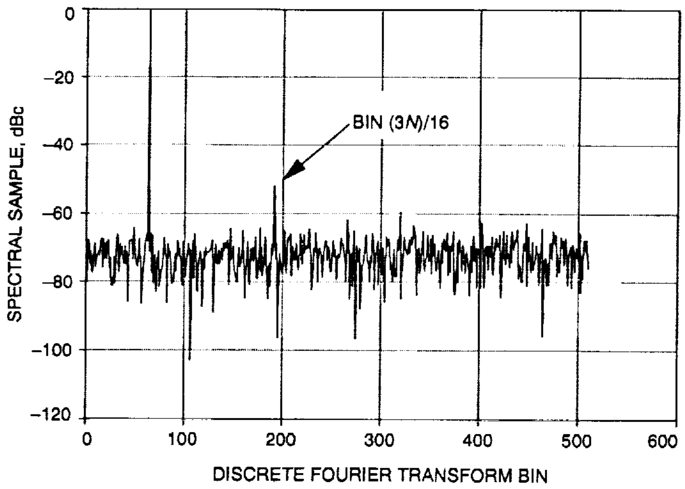


Fig. 5. DFT spectrum, 50-dB SNR input.

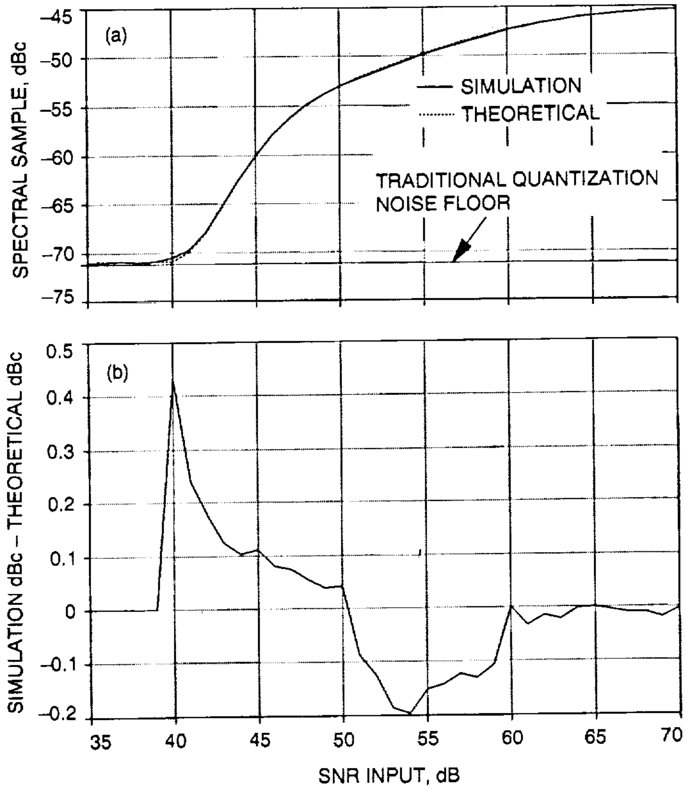


**Fig. 6. Spectral sample versus (a) quantizer input SNR, and (b) quantizer input SNR, dBc difference.**

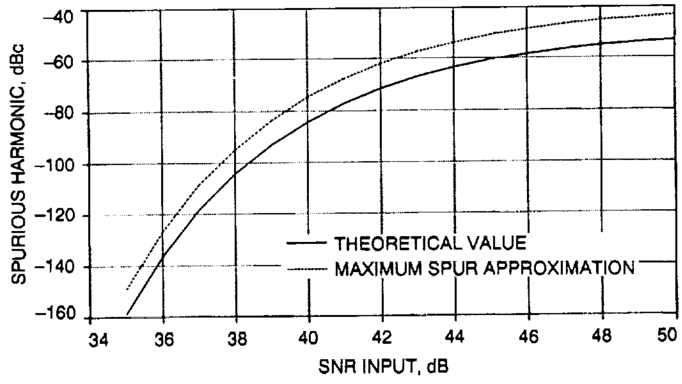


**Fig. 7. Spurious harmonics versus quantizer input SNR.**

# Appendix A

## Derivation of the Power Spectrum

From Eq. (4), first consider $\tau = 0$:

$$R_e(t,t) = -\frac{\Delta^2}{4\pi^2} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^{l+k}}{lk}$$

$$\times \exp\left(\frac{j2\pi A}{\Delta}(l+k)\sin(\omega_0 t + \phi)\right)$$

$$\times E\left\{\exp\left(\frac{j2\pi(l+k)}{\Delta}z(t)\right)\right\}$$

Using the characteristic function of a zero-mean Gaussian random variable $z$ with variance $\sigma^2$, it is known that $E\{\exp(j\alpha z)\} = \exp(\alpha^2/2\sigma^2)$ [6]. Therefore,

$$R_e(t,t) = -\frac{\Delta^2}{4\pi^2} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^{l+k}}{lk}$$

$$\times \exp\left(\frac{j2\pi A}{\Delta}(l+k)\sin(\omega_0 t + \phi)\right)$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}(l+k)^2\right)$$

Changing indices so that $l + k = m$ and translating the condition $k \neq 0$ into $m \neq l$, one obtains

$$R_e(t,t) =$$

$$\frac{\Delta^2}{4\pi^2} \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^m}{l(l-m)}$$

$$\times \exp\left(\frac{j2\pi Am}{\Delta}\sin(\omega_0 t + \phi)\right)\exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}m^2\right)$$

Separating the case when $m = 0$ and rearranging the order of summation,

$$R_e(t,t) = \frac{\Delta^2}{4\pi^2} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{1}{l^2} + \frac{\Delta^2}{4\pi^2} \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} (-1)^m$$

$$\times \exp\left(\frac{j2\pi Am}{\Delta}\sin(\omega_0 t + \phi)\right) \sum_{\substack{l=-\infty \\ l\neq 0 \\ l\neq m}}^{\infty} \frac{1}{l(l-m)}$$

Noting that the sum involving $1/l^2$ evaluates to $\pi^2/3$ [7] and (after partial fraction manipulation) that the sum involving $1/l(l-m)$ evaluates to $2/m^2$, one can write

$$R_e(t,t) = \frac{\Delta^2}{12}$$

$$+ \frac{\Delta^2}{2\pi^2} \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} \frac{(-1)^m \exp\left(\frac{j2\pi Am}{\Delta}\sin(w_0 t + \phi)\right)}{m^2}$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}m^2\right) \tag{A-1}$$

Now consider $\tau \neq 0$:

$$R_e(t, t+\tau) =$$

$$-\frac{\Delta^2}{4\pi^2} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^{l+k}}{lk}$$

$$\times \exp\left(\frac{j2\pi A}{\Delta}[k\sin(\omega_0 t + \phi) + l\sin(\omega_0 \tau + \phi)]\right)$$

$$\times E\left\{\exp\left(\frac{j2\pi}{\Delta}(kz(t) + lz(t+\tau))\right)\right\}$$

Since the input noise is white, $z(t)$ and $z(t+\tau)$ are independent random variables ($\tau \neq 0$), and the expectation of the product becomes the product of the expectations. When one uses the same characteristic function method detailed above,

$$R_e(t,t+\tau) = -\frac{\Delta^2}{4\pi^2} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^{l+k}}{lk}$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}(l^2+m^2)\right)$$

$$\times \exp\left(\frac{j2\pi A}{\Delta}[k\sin(\omega_0 t+\phi)\right.$$

$$\left. + l\sin(\omega_0 t + \omega_0 \tau + \phi)]\right) \qquad \text{(A-2)}$$

As indicated in Eq. (5), to obtain the discrete-time autocorrelation function $R_x[n, n+k]$, replace $t$ with $nT$ and $\tau$ with $kT$ in Eqs. (A-1) and (A-2). The evaluation of Eq. (6) involves changing the order of integration and summation until only terms involving the phase $\phi$ are inside the integral. It is useful at this time to use the Jacobi-Anger formula [7]:

$$e^{jx\sin\phi} = \sum_{p=-\infty}^{\infty} J_p(x)e^{jp\phi}$$

One can now evaluate

$$\overline{R}_x[0] = \frac{\Delta^2}{12} + \frac{\Delta^2}{2\pi^2} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^l}{l^2}$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right) \sum_{p=-\infty}^{\infty} J_p\left(\frac{2\pi A l}{\Delta}\right) e^{jp\omega_0 T n}$$

$$\times \frac{1}{2\pi}\int_0^{2\pi} e^{jp\phi} d\phi$$

The integral above will be 1 when $p = 0$ and 0 otherwise. Noting that the resulting expression inside the above sum is an even function of $l$, one obtains the first half of Eq. (7). In evaluating $\overline{R}_x[k]$ when $k \neq 0$, one again employs the Jacobi-Anger formula and interchanges the order of summation and integration so that the appropriate terms from Eq. (A-2) are inside the integral. Using the orthogonality of exponentials and noting that $J_p(x) = (-1)^p J_{-p}(x)$, one obtains

$$\overline{R}_x[k] = -\sum_{p=-\infty}^{\infty} (-1)^p \exp(-jp\omega_0 Tk)\frac{\Delta^2}{4\pi^2}$$

$$\times \left( \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \frac{(-1)^l J_p\left(\frac{2\pi A l}{\Delta}\right)}{l} \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}l^2\right) \right)^2$$

Consider the sum over $l$. When $p$ is even, this sum will be zero since $J_p(x) = (-1)^p J_p(-x)$. When $p$ is odd, the sum over $l$ is an even function of $l$, and one can reduce the double-sided infinite sum to a single-sided infinite sum. By changing the index from $p$ to $-n$, one obtains the second half of Eq. (7).

In evaluating Eq. (9), one can write

$$S(\omega) = \overline{R}_x[0] + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \overline{R}_x[k]e^{j\omega Tk}$$

Note that the first term, $\overline{R}_x[0]$, is independent of the frequency, $\omega$. Using the notation in Eq. (7),

$$S(\omega) = \overline{R}_x[0]$$

$$+ \left( \sum_{k=-\infty}^{\infty} \sum_{\substack{n-\infty \\ n\ odd}}^{\infty} S_n \exp(jn\omega_0 Tk)\exp(j\omega Tk) \right)$$

$$- \sum_{\substack{n-\infty \\ n\ odd}}^{\infty} S_n$$

Note that the third term above is independent of frequency. Writing out the expression for $\overline{R}_x[0]$, one arrives at Eq. (10) after noting

$$\sum_{k=-\infty}^{\infty} e^{jxk} = 2\pi \sum_{k=-\infty}^{\infty} \delta(x - 2\pi k) = 2\pi\delta(x \bmod 2\pi) \qquad \text{(A-3)}$$

where $\delta$ ( ) is the Dirac delta function.

# Appendix B

## Derivation of the DFT Problem

From Eq. (11), one can write

$$Y = E\{|X[k]|^2\}$$

$$= \frac{1}{N^2}\sum_{n=0}^{N-1} E\{x[n]x[n]\}$$

$$+ \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{\substack{r=0\\ r\neq n}}^{N-1} E\{x[n]x[r]\}\exp\left(\frac{-j2\pi k(n-r)}{N}\right) \tag{B-1}$$

Consider the first term above. Note that it is independent of the bin number $k$ (i.e., independent of frequency). Recalling the results of Eqs. (5) and (A-1), rearranging the order of summation, and using the Jacobi-Anger formula leaves one with

$$\frac{1}{N^2}\sum_{n=0}^{N-1} R_e(nT, nT) = \frac{\Delta^2}{12N}$$

$$+ \frac{\Delta^2}{2\pi^2 N}\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}\frac{(-1)^m}{m^2}$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}m^2\right)\sum_{p=-\infty}^{\infty} J_p\left(\frac{2\pi Am}{\Delta}\right)e^{jp\phi}$$

$$\times \frac{1}{N^2}\sum_{n=0}^{N-1} e^{\frac{j\pi n}{8}}$$

Recall that $\omega_0 T = \pi/8$ and $k = 3N/16$ in this example. The sum over $n$ is nonzero only when $p = 0 \bmod 16$. So,

$$\frac{1}{N^2}\sum_{n=0}^{N-1} R_e(nT, nT) = N_{Q_1} + N_{Q_2}$$

where $N_{Q_1} = \Delta^2/12N$, and

$$N_{Q_2} = \frac{\Delta^2}{2\pi^2 N}\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}\frac{(-1)^m}{m^2}$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}m^2\right)\sum_{p=-\infty}^{\infty} J_{16p}\left(\frac{2\pi Am}{\Delta}\right)e^{j16\phi p}$$

The sum over $p$ above is evaluated in Appendix C. Inserting the result of Appendix C, interchanging the order of summations, and using Euler's rule yields the second part of Eq. (13).

Now return to the second part of Eq. (B-1):

$$W = \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{\substack{r=0\\ r\neq n}}^{N-1} E\{x[n]x[r]\}\exp\left(\frac{-j2\pi k(n-r)}{N}\right)$$

$$= \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{\substack{r=0\\ r\neq n}}^{N-1} R_e(nT, rT - nT)\exp\left(\frac{-j2\pi k(n-r)}{N}\right)$$

where $R_e(\ )$ is defined in Eq. (A-2). Evaluating this further,

$$W =$$

$$-\frac{\Delta^2}{4\pi^2}\sum_{\substack{l=-\infty\\ l\neq 0}}^{\infty}\sum_{\substack{m=-\infty\\ m\neq 0}}^{\infty}\frac{(-1)^{l+m}}{lm}\exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}(l^2+m^2)\right)$$

$$\times\left[\frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{r=0}^{N-1}\exp\left(\frac{j2\pi A}{\Delta}[l\sin(\omega_0 Tn + \phi)\right.\right.$$

$$\left.+ m\sin(\omega_0 Tr + \phi)]\right)\exp\left(\frac{j2\pi k(n-r)}{N}\right)$$

$$\left.-\frac{1}{N^2}\sum_{n=0}^{N-1}\exp\left(j\frac{2\pi A(l+m)}{\Delta}\sin(\omega_0 Tn + \phi)\right)\right]$$

Note that the second term inside the big brackets is independent of the bin number $k$. Further separating the terms inside the big brackets, one can write

$$W = S_Q + N_{Q_3} \qquad \text{(B-2)}$$

Consider the $N_{Q_3}$ term (this is the term independent of $k$). Employing the Jacobi-Anger formula and evaluating the sum over $n$ as before,

$$N_{Q_3} =$$

$$\frac{\Delta^2}{4\pi^2 N} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} \frac{(-1)^{(l+m)} \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}(l^2+m^2)\right)}{lm}$$

$$\times \sum_{k=-\infty}^{\infty} J_{16k}\left(\frac{2\pi A}{\Delta}(l+m)\right) e^{j16\phi k}$$

The results of Appendix C can be used to evaluate the infinite sum over $k$. Again, after manipulating the definition of $F(\ )$ in Appendix C as detailed above, one finally obtains the third part of Eq. (13).

Finally, consider $S_Q$ from Eq. (B-2):

$$S_Q = -\frac{\Delta^2}{4\pi^2} \sum_{\substack{l=-\infty \\ l\neq 0}}^{\infty} \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} \frac{(-1)^{l+m}}{lm}$$

$$\times \exp\left(-\frac{2\pi^2\sigma^2}{\Delta^2}(l^2+m^2)\right)$$

$$\times \left[\frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{r=0}^{N-1} \exp\left(\frac{j2\pi A}{\Delta}[l\sin(\omega_0 Tn+\phi)\right.\right.$$

$$\left.\left. + m\sin(\omega_0 Tr+\phi)]\right) \exp\left(\frac{-j2\pi k(n-r)}{N}\right)\right]$$

Call the sum in the big brackets $V$, and employ the Jacobi-Anger formula. Using the values of $\omega_0 T$ and $k$ from Section III, and assuming that $N$ is a power of 2 ($N \geq 16$), the value of $V$ reduces to

$$V = \left(\sum_{p=-\infty}^{\infty} J_{3+16p}\left(\frac{2\pi Al}{\Delta}\right) e^{j16\phi p}\right)$$

$$\times \left(\sum_{q=-\infty}^{\infty} J_{-3+16q}\left(\frac{2\pi Am}{\Delta}\right) e^{j16\phi q}\right)$$

Since $J_n(x) = (-1)^n J_n(-x)$ and $\pm 3 + 16p$ is odd, one can change the double-sided infinite sums to single-sided infinite sums. By changing the index on $q$ to $-q$, recalling that $J_{-n}(x) = (-1)^n J_n(x)$, and using the formula in Appendix C, one gets Eq. (14).

# Appendix C
## Evaluating an Infinite Bessel Sum

Define

$$F(x,y,z,t) = \sum_{k=-\infty}^{\infty} e^{jxk} J_{y+zk}(t)$$

where $x$ and $y$ are integers and $z \neq 0$. Using the integral definition of the Bessel function of integer order [7],

$$F(x,y,z,t) =$$

$$\sum_{k=-\infty}^{\infty} e^{jxk} \frac{j^{(-y+zk)}}{\pi} \int_{0}^{\pi} \cos(y\theta + zk\theta) e^{jt\cos(\theta)} d\theta$$

After changing the order of integration and summation, one can apply Euler's rule to express cos( ) in terms of complex exponentials. Next, equate the infinite sums of complex exponentials to infinite sums of Dirac delta functions as in Eq. (A-3). Only a finite number of delta functions will remain inside the limits of integration. Note that the remainder of this analysis assumes (for simplicity) that no delta functions lie on the limits of integration. For the DFT example in this article, this condition is satisfied. Using the sifting and scaling properties of delta functions, one finally obtains

$$F(x,y,z,t) =$$

$$\frac{j^{-y}}{z} \sum_{m=0}^{\lfloor z/2 \rfloor - 1} \left[\exp\left(j\left(t\Theta_m + y\theta_m\right)\right) + \exp\left(j\left(t\Xi_m - y\xi_m\right)\right)\right]$$

where

$$\Theta_m = \frac{\pi}{2} - \frac{x}{z} + \frac{2\pi}{z}\left(\left\lfloor \frac{x}{2\pi} - \frac{z}{4} \right\rfloor + m\right)$$

$$\xi_m = \frac{\pi}{2} + \frac{x}{z} - \frac{2\pi}{z}\left(\left\lfloor \frac{x}{2\pi} - \frac{z}{4} \right\rfloor - m\right)$$

and $\Theta_m = \cos(\theta_m)$ and $\Xi_m = \cos(\xi_m)$.

# Appendix D

## Truncating an Infinite Sum

This appendix considers bounds on the error introduced by truncating the following infinite sum:

$$x = \sum_{l=1}^{\infty} \frac{e^{-\alpha^2 l^2}}{l^n} F(l)$$

$$= \sum_{l=1}^{L} \frac{e^{-\alpha^2 l^2}}{l^n} F(l) + \sum_{l=L+1}^{\infty} \frac{e^{-\alpha^2 l^2}}{l^n} F(l)$$

$$= \hat{x} + error$$

where $n \geq 1$ and $|F(l)| \leq 1$. The error can be bounded as follows:

$$|error| \leq \sum_{l=L+1}^{\infty} \frac{e^{-\alpha^2 l^2}}{l^n} |F(l)| \leq \sum_{l=L+1}^{\infty} \exp(-\alpha^2 l^2)$$

Let $\alpha^2 = 1/2\sigma^2$. Then,

$$|error| \leq \sqrt{2\pi\sigma^2} \sum_{l=L+1}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}}$$

The above sum can be visualized as the integral from $x = L$ to $x = \infty$ of a discontinuous step-like function whose value is constant over the interval between two adjacent integers. The value over an interval is equal to the Gaussian density evaluated at the right-most portion of the interval. A little thought will show that this integral is strictly less than the integral of a Gaussian distribution from $x = L$ to $x = \infty$. Therefore,

$$|error| < \frac{\sqrt{\pi}}{2\alpha} \operatorname{erfc}(\alpha L)$$

where erfc( ) is the complementary function defined in [7].

# References

[1] M. P. Quirk, M. F. Garyantes, H. C. Wilck, and M. J. Grimm, "A Wide-Band High-Resolution Spectrum Analyzer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-36, no. 12, December 1988.

[2] W. R. Bennett, "Spectra of Quantized Signals," *Bell Systems Technical Journal*, vol. 27, pp. 446–472, July 1948.

[3] W. J. Hurd, "Correlation Function of Quantized Sine Wave Plus Gaussian Noise," *IEEE Transactions on Information Theory*, vol. IT-13, no. 1, pp. 65–68, January 1967.

[4] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Chapter 3, Englewood Cliffs, New Jersey: Prentice-Hall, 1989.

[5] L. E. Brennan and I. S. Reed, "Quantization Noise in Digital Moving Target Indication Systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-2, no. 6, pp. 655–658, November 1966.

[6] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, pp. 115–116, 1984.

[7] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Washinton, D. C.: U. S. Department of Commerce, National Bureau of Standards, 1972.