The Telecommunications and Data Acquisition Progress Report 42-99

July–September 1989

E. C. Posner
Editor

November 15, 1989

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Telecommunications and Data Acquisition Progress Report 42-99

July–September 1989

E. C. Posner
Editor

November 15, 1989

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.

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) and its associated Ground Communications Facility (GCF) in planning, in supporting research and technology, in implementation, and in operations. Also included is TDA-funded activity at JPL on 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 performs 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 Exploration).

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 support by the Office of Space Operations for the station support time.

Finally, tasks funded under the JPL Director's Discretionary Fund and the Caltech President's Fund which 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
- Communications Implementation and Operations
- Data and Information Systems
- Flight-Ground Advanced Engineering

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

OSSA Tasks:

Search for Extraterrestrial Intelligence
Goldstone Solar System Radar
Radio Astronomy

Discretionary Funded Tasks

Contents

OSO TASKS
DSN Advanced Systems
TRACKING AND GROUND-BASED NAVIGATION

Determination of Earth Orientation Using the Global Positioning System 1
A. P. Freedman
NASA Code 310-10-61-87-02

Simple Analytic Potentials for Linear Ion Traps 12
G. R. Janik, J. D. Prestage, and L. Maleki
NASA Code 310-10-62-15-00

Microwave Oscillator With Reduced Phase Noise by Negative Feedback Incorporating Microwave Signals With Suppressed Carrier 20
G. J. Dick and J. Saunders
NASA Code 310-10-62-34-00

Effect of Laser Frequency Noise on Fiber-Optic Frequency Reference Distribution 34
R. T. Logan, Jr., G. F. Lutes, and L. Maleki
NASA Code 310-10-62-16-00

Thermal Coefficient of Delay for Various Coaxial and Fiber-Optic Cables 43
G. Lutes and W. Diener
NASA Code BG 310-10-62-16-00

COMMUNICATIONS, SPACECRAFT-GROUND

Performance of the All-Digital Data-Transition Tracking Loop in the Advanced Receiver 60
U. Cheng and S. Hinedi
NASA Code 310-30-70-04-02

Costas Loop Lock Detection in the Advanced Receiver 72
A. Mileant and S. Hinedi
NASA Code 310-30-70-04-02

Photon Statistical Limitations for Daytime Optical Tracking 90
W. M. Folkner and M. H. Finger
NASA Code 310-20-67-89-02

STATION CONTROL AND SYSTEM TECHNOLOGY

Memory Management in Traceback Viterbi Decoders 98
O. Collins and F. Pollara
NASA Code 310-30-71-83-02

Some Easily Analyzable Convolutional Codes 105
R. McEliece, S. Dolinar, F. Pollara, and H. Van Tilborg
NASA Code 310-30-71-83-02

Quantization Effects in Viterbi Decoding Rate 1/*n* Convolutional Codes 115
I. M. Onyszchuk, K.-M. Cheung, and O. Collins
NASA Code 310-30-72-88-01

Big Viterbi Decoder (BVD) Results for (7,1/2) Convolutional Code 122
J. Statman, J. Rabkin, and B. Siev
NASA Code 310-30-72-88-01

Fast Transform Decoding of Nonsystematic Reed-Solomon Codes 130
T. K. Truong, K.-M. Cheung, I. S. Reed, and A. Shiozaki
NASA Code 310-30-70-87-02

Application of Adaptive Least-Squares Algorithm to Multi-Element Array Signal Reconstruction 141
R. Kumar
NASA Code 310-30-70-89-01

FLIGHT-GROUND ADVANCED ENGINEERING

The Effects of Sinusoidal Interference on the Second-Order Carrier Tracking Loop Preceded by a Bandpass Limiter in the Block IV Receiver 161
C. J. Ruggier
NASA Code 315-20-50-00-05

DSN Systems Implementation

NETWORK SUSTAINING

Disturbance Torque Rejection Properties of the NASA/JPL 70-Meter Antenna Axis Servos 170
R. E. Hill
NASA Code 314-30-42-01-44

CAPABILITIES FOR EXISTING PROJECTS

Parkes Radio Science System Design and Testing for Voyager Neptune Encounter 189
T. A. Rebold and J. F. Weese
NASA Code 314-40-41-81-11

CAPABILITIES FOR NEW PROJECTS

32-GHz Performance of the DSS-14 70-Meter Antenna: 1989 Configuration 206
M. S. Gatti, M. J. Klein, and T. B. H. Kuiper
NASA Code 310-20-64-89-00

Errata 220

Determination of Earth Orientation Using the Global Positioning System

A. P. Freedman
Tracking Systems and Applications Section

Modern spacecraft tracking and navigation require highly accurate Earth-orientation parameters. For near-real-time applications, errors in these quantities and their extrapolated values are a significant error source. A globally distributed network of high-precision receivers observing the full Global Positioning System (GPS) configuration of 18 or more satellites may be an efficient and economical method for the rapid determination of short-term variations in Earth orientation.

A covariance analysis utilizing the JPL Orbit Analysis and Simulation Software (OASIS) has been performed to evaluate the errors associated with GPS measurements of Earth orientation. These GPS measurements appear to be highly competitive with those from other techniques and can potentially yield frequent and reliable centimeter-level Earth-orientation information while simultaneously allowing the oversubscribed Deep Space Network (DSN) antennas to be used more for direct project support.

I. Introduction

Knowledge of the Earth's orientation in space is critical to the operation of NASA's Deep Space Network (DSN). Unless the orientation is closely monitored, the variable rotation of the Earth can lead to errors in spacecraft navigation. In near-real-time, high-precision spacecraft-tracking applications, the need for up-to-date Earth-orientation information is particularly crucial. The Magellan mission to Venus, for example, requires that Earth-rotation errors be kept under 30 cm;[1] by the mid-1990s, missions are envisioned that would require continuously available Earth-orientation knowledge accurate to 3 cm.[2]

The Navstar Global Positioning System (GPS) is a network of orbiting radio transmitters designed for navigation purposes that is revolutionizing terrestrial distance determinations. The completion of the full satellite network by the early 1990s promises to extend recent improvements in regional geodetic measurements with GPS to a global scale [1, 2]. The capability of GPS to pinpoint receiver locations at the centimeter or subcentimeter level in a ter-

[1] T. F. Runge, "UTPM calibration accuracy for Magellan," JPL IOM 335.5-87.81 (internal document), Jet Propulsion Laboratory, Pasadena, California, April 30, 1987.

[2] R. Treuhaft and L. Wood, "Revisions in the differential VLBI error budget and applications for navigation in future missions," JPL IOM 335.4-601 (internal document), Jet Propulsion Laboratory, Pasadena, California, December 31, 1986.

restrial reference frame that is precisely tied to an inertial frame suggests that GPS may be effective in monitoring Earth-orientation changes.

Earth orientation consists of three parts: the angle of rotation of the Earth about its rotation axis relative to a mean rotation angle (UT1–UTC), the position of the current axis of rotation of the Earth with respect to a reference axis tied to the crust and mantle of the solid Earth (polar motion), and the orientation of the rotation axis in inertial space (precession and nutation). The first two of these parameters, UT1 and polar motion, collectively known as UTPM (Fig. 1), vary as a result of angular momentum exchange between the solid parts of the Earth and its atmosphere, oceans, and fluid core. These Earth-orientation components can vary rapidly and unpredictably. Nutations and precession are primarily products of Earth's interactions with other celestial bodies and are largely periodic; they will not be dealt with in this article.

Earth orientation[3] is currently being monitored by a number of precise geodetic techniques: Very Long Baseline Interferometry (VLBI), Satellite Laser Ranging (SLR), and Lunar Laser Ranging (LLR). These techniques can currently achieve measurement accuracies of up to 2–3 cm over time scales as short as one hour [3]. The turnaround time necessary to collect and process these raw data can be quite long, however. VLBI from IRIS (International Radio Interferometric Surveying) provides daily UT1–UTC and five-day UTPM data, while SLR from CSR (Center for Space Research, University of Texas, Austin) provides three-day polar-motion data. The results, however, are not usually available until a week or more after the epoch at which the measurements are valid. This delay is not acceptable for many DSN navigation needs. VLBI data obtained by JPL using the DSN antennas are known as TEMPO (Time and Earth Motion Precision Observations) and can be processed rapidly when necessary; demands for radio-telescope time limit the frequency of observations to weekly, however. Prompt reduction of data can be critical to navigation since Earth orientation is continuously changing. Earth rotation, in particular, is highly variable, and changes of 25 cm per day have been known to occur. Thus no system presently active is likely to meet the long-term DSN needs of regular, high-precision, daily monitoring of Earth orientation with data reduction times of less than one day.

VLBI, SLR, and LLR are, in addition, labor intensive and require significant investments in equipment and personnel. In VLBI, large radio telescopes are required—telescopes whose valuable observation time is in great demand at sites such as the DSN. The laser techniques use dedicated stations to obtain data but are subject to the vagaries of local weather conditions; thus they are not a reliable source for regular daily measurements.

If GPS technology could produce Earth-orientation measurements of a quality comparable to that produced by VLBI, SLR, and LLR, it would free up significant amounts of time on the DSN and other VLBI networks currently being used to monitor Earth orientation. It would, moreover, allow measurements to be made more frequently than is now practical with VLBI. In addition, the use of radio-frequency energy would enable GPS systems to be much less sensitive to weather than optical systems. GPS would not replace these VLBI, SLR, and LLR techniques (due to systematic difficulties described below); rather, GPS systems would be employed in a synergistic combination with these present-day techniques to enhance overall performance.

Frequent high-precision Earth-orientation data is also of value for scientific studies. Little is known about exchanges of angular momentum between the solid Earth and the atmosphere, or about the excitation of polar motions, at periods of a week or less. Continuous GPS monitoring would help to extend our knowledge to higher frequencies, with benefit to geophysics, meteorology, and astronomy. Weather forecasting, in particular, may benefit from independent estimates of daily atmospheric angular momentum as provided by geodetic Earth-orientation measurements.

A GPS receiver and data-processing system is scheduled to be in place at each of the DSN sites within a year to enable highly accurate, near-real-time ionospheric calibration in support of deep-space missions that transmit at a single frequency.[4] This system is designed to have

[3] Earth orientation is measured in a variety of units. Polar motion is essentially an angular displacement, while rotational variations can be expressed either as angular or temporal displacements. Both can also be expressed as a distance corresponding to the angular displacement measured at one Earth radius. Thus 1 cm at the Earth's surface is equivalent to an angular distance of approximately 0.3 marcsec or 1.6 nrad. The time it takes for the Earth to rotate through this angle and move a point at the equator to the east by 1 cm is approximately 0.02 msec of time. Thus 30 cm corresponds to 0.65 msec of Earth rotation or 9.6 marcsec (47 nrad) of polar motion.

[4] C. J. Vegos, "DSCC Media Calibration Subsystem (DMD), Functional Design Review (Level D)," JPL 834-30, vol. 1 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 1, 1987.

a data turnaround capability of about 12 hours. These receivers are also expected to be part of an international, global GPS tracking network for the TOPEX/POSEIDON mission [4]. By 1992, therefore, a GPS system should be in place to support the continuous monitoring of UTPM.

This article documents a covariance analysis evaluating the potential of GPS for measuring Earth orientation. The assumed satellite constellation is that originally proposed by the U.S. Department of Defense as the operational configuration. It consists of 18 satellites with 12-hour orbits and lying in six orbit planes equidistantly placed in longitude, three satellites per orbit plane. This constellation enables at least five satellites to be seen and tracked most of the time from anywhere on the Earth's surface [5]. Although the Air Force has subsequently modified this constellation to include up to 24 satellites, these changes should not significantly alter the conclusions of this study. The network of ground receivers is assumed to consist of six sites regularly spaced around the globe (three of these coincident with the DSN sites).

These sites are assumed to be equipped with receivers that yield two distinct observables: "carrier phase," based on measurements of the radio frequency (RF) carrier that is transmitted by the GPS satellites, and "pseudorange," based on a precise modulation of the transmitted signal. Both observables are indicators of the distance between a satellite and a receiver. Pseudorange consists of the light travel time between the two points, plus any clock offsets of the receiver and transmitter. It is often corrupted by multipath effects and is therefore the noisier data type. Carrier phase monitors the relative position change between the satellite and ground station. It is a cleaner data set, but the absolute distance is made ambiguous by a constant bias (equal to an integer number of wavelengths) for each continuously measured satellite arc.

Contemporary receiver capabilities are better for carrier phase than for pseudorange, but the quality of pseudorange data is rapidly improving. The GPS receiver being installed at the DSN sites can now achieve, under optimum conditions, pseudorange noise levels as low as 5 cm (averaged over 30 minutes), as well as carrier phase noise levels well below 0.5 cm [6]. By about 1992, when the full GPS constellation is active, such high-precision pseudorange and carrier phase data should be routinely available from numerous sites around the globe.

II. Covariance Analysis

This study utilized the Orbit Analysis and Simulation Software (OASIS) program, developed at JPL for the covariance and simulation analysis of Earth-orbiting satellites [7]. It consists of a number of independent modules: PV integrates the satellite orbits and computes the variational partial derivatives for satellite-related parameters; REGRES-PMOD generates simulated observations and their measurement partials; OAFILTER does the actual covariance analysis, i.e., using specified uncertainties of the data and the models, the program estimates the uncertainties in desired parameters; and UDIGEST generates the desired output. Parameters can be either estimated (adjusted) or "considered"; "considered" parameters are treated as systematic error sources [8]. Adjusted parameters can be modeled either as constants or as stochastic variables.

Parameters that can be adjusted include

(1) satellite epoch states, i.e., their initial positions and velocities

(2) various satellite force-model parameters such as solar-radiation pressure and Y-bias

(3) satellite and station clock offsets

(4) station locations

(5) wet-zenith tropospheric path delay for each station

(6) Earth-orientation parameters

(7) gravitational harmonic coefficients and the value of GM (the gravitational constant, G, multiplied by the mass of the Earth, M)

(8) geocenter offset

(9) carrier phase biases

To determine Earth orientation with GPS, one needs to know the precise position and orientation of a set of points on the Earth's surface with respect to an inertial reference frame as a function of time. The GPS satellites provide an orbital reference frame that is not truly inertial but is slowly varying with time. Uncertainties in the GPS orbits can be reduced through estimation of parameters (1), (2), and (7) described above. The distances between satellites and receivers can be determined from the data after removing the effects of parameters (3), (4), (5), and (9). Determining the orientation of the satellite constellation in inertial space may be achieved by fixing the locations of a few ground receivers. These sites are known as fiducial sites and are tied by local ground surveys to nearby, colocated VLBI antennas whose relative positions in inertial space are known precisely [9]. The origin of this VLBI frame may not be coincident with the Earth's center

of mass as determined by the satellites; this geocenter offset (parameter 8) can also be estimated. Thus the satellite framework can be constrained in inertial space, and the movements of the solid Earth within that framework, such as Earth orientation, can be observed.

Appropriate a priori uncertainties for all estimated parameters are needed to strengthen the solution. Deciding which parameters to estimate or consider, what a priori values to use, and which data to include depends on the physical problem of interest.

The following questions were addressed in this study:

(a) How many GPS measurements are necessary to generate Earth-orientation values with a precision comparable to present techniques? In other words, how long need the observation periods be in order to produce useful Earth-orientation data?

(b) Are both pseudorange and carrier phase data types needed, and what maximum data noise is permissible for each type to allow adequate resolution of Earth-orientation parameters?

(c) How important are the effects of solar-radiation pressure, station location errors, tropospheric uncertainties, and geocenter errors on Earth-orientation estimation? Do these parameters need to be estimated along with satellite states and Earth-orientation parameters, or can they be considered?

The parameter-estimation strategy is listed in Table 1. Earth-orientation parameters and their rates of change are all estimated with a priori uncertainties at least as large as the uncertainties expected in VLBI-provided UTPM prior to a GPS measurement, and at least a factor of 10 larger than the final, desired uncertainties. Satellite states and solar-radiation parameters are also estimated, with a priori sigmas comparable to the known uncertainties of the broadcast satellite ephemerides and of solar-radiation effects. Both DSN and non-DSN station locations are estimated, with the three DSN sites comprising the fiducial network constrained more tightly than the non-DSN sites. A priori sigmas for the station locations correspond to the present-day uncertainties of the VLBI baselines. The wet-zenith troposphere delay is estimated with an a priori sigma that is appropriate for a dry climate if surface meteorology data are available, and is more than adequate for wetter regions if water vapor radiometers are used.

The geocenter, carrier phase biases, and clock errors were all estimated with large, effectively unconstrained a priori values. One station clock was fixed as a reference clock. All parameters were estimated as constants over the observing period, except for the clock errors, which were modeled as white noise [10].

The assumed data noise levels are appropriate for a DSN GPS receiver operating either in a poor, noisy propagation environment (20-cm pseudorange, 1-cm carrier phase) or in a reasonably good environment (5 cm, 0.5 cm, respectively). Note that these data noise levels assume individual measurements averaged over a 5-minute interval, or "batch."

Details of the software capabilities and modeling strategies for most parameters are described in [10]. Six matrix rotations[5] are applied to station-location vectors to convert them from an Earth-fixed reference frame (1903.0 CIO frame) to a geocentric inertial system (J2000). Three of these rotations correspond to UT1 and the two components of polar motion. For this study, Earth-orientation rates are also needed. The UTPM components are thus modeled as $\theta(t) = a + b(t - t_0)$, where $\theta(t)$ is a UTPM parameter residual, a is the estimated value of the constant component at some epoch t_0, and b is the estimate of the rate component.

III. Results and Discussion

Presented below are the results of covariance analyses for three distinct models summarized in Table 2. In model A, only pseudorange is employed, and these data have a high noise level. This model thus represents a "worst case" scenario for determining Earth orientation. Model B is identical to A, except that higher quality pseudorange data are assumed. Model C represents a near-optimal situation with regard to data quality: Both high-quality pseudorange (5 cm) and carrier phase (0.5 cm) are employed jointly. Note that this "best-case" model represents the expected quality of the data. In all of these models, station locations, wet-zenith tropospheric delays, solar pressure, clock offsets, geocenter offset, and carrier phase biases are assumed to be estimated along with Earth-orientation parameters and satellite epoch states.

The Earth-rotation parameter, UT1–UTC, is not expected to be directly measurable by GPS. An error in the satellite-node longitudes cannot unambiguously be sepa-

[5] W. I. Bertiger, "Non-force models module," *OASIS Mathematical Description, V. 1.0*, JPL D-3139 (internal document), Jet Propulsion Laboratory, Pasadena, California, April 1986.

rated from uncertainty in UT1. This is a problem common to all satellite geodesy, including SLR. If, however, GPS can determine the change over time of UT1–UTC, then this change can be combined with an initial value from VLBI measurements to yield the full UT1–UTC as a function of time.

Figure 2 illustrates the ability of GPS to monitor the rate of change of UT1–UTC, also known as length-of-day (LOD). With eight hours of observation, none of the models has achieved the measurement precision of present-day techniques. This present-day capability, available from VLBI or SLR after a processing delay of a week or longer, is indicated on Figs. 2, 4, and 5 by an arrow. With 16 hours of observation, however, model C, with its high-quality pseudorange and carrier phase data, can resolve LOD to better than 5 cm. By 24 hours, both models B and C show error levels comparable to or better than current uncertainties. The best scenario, model C, predicts subcentimeter LOD accuracy with 24 hours of GPS tracking.

This powerful ability to measure the rate of change of UT1–UTC enables the estimation of UT1–UTC with the help of VLBI. Because GPS and VLBI receivers are colocated at DSN sites, the VLBI and GPS reference frames should be precisely defined with respect to each other, yielding a high-quality GPS tie to inertial space. By integrating LOD over time, the total change in UT1–UTC can be estimated and added to an initial value determined from VLBI. If UT1–UTC can be measured accurately to 2 cm at an initial epoch, for example, daily estimates of LOD with GPS will add less than one centimeter per day to this uncertainty. To remain within 30 centimeters of the true UT1–UTC, VLBI measurements of UT1–UTC may only be needed monthly. This is illustrated in Fig. 3. Combining and smoothing all the data from various sources will yield better UT1 estimates, but these will only be available much later, after additional data are obtained. Thus GPS and VLBI are complementary techniques that can be combined synergistically to yield improved UT1 estimates.

Figure 4 illustrates the improvement with observing time of the Y component of polar motion (PMY). The X component (PMX) behaves in a similar manner and is not illustrated here. By 16 hours, the "best case" scenario of model C shows measurement accuracy comparable to present techniques, while by 24 hours model B (high-quality pseudorange) also yields an acceptable predicted error. The error estimates for all three models seem to be converging at the few-centimeter level. This is a limitation controlled mainly by the a priori station location uncertainties (constrained at 3–5 centimeters for each of three components).

Figure 5 shows the estimate of the error in the rate of change of PMY as a function of observing time; the rate of change of PMX error exhibits similar temporal variations. The behavior illustrated here is similar to that of the LOD component (Fig. 2). All three models show rapid improvement with time; by 24 hours, models B and C again achieve high-quality rate measurements. In the best case (model C), the error is down at the few-mm/day level.

It appears that GPS determination of Earth orientation is feasible for LOD, polar motion, and polar-motion rates. With less than 24 hours of tracking, these components will be as well-determined as measurements by current techniques. Recall, however, that current techniques only provide these values many days after the measurements are taken, whereas GPS is expected to deliver the results to the DSN within 12 hours. The best estimates come, as expected, with the smallest data noise. Combining high-quality pseudorange and carrier phase data reduces the uncertainties in Earth-orientation parameters most rapidly, achieving present-day measurement precision within approximately 16 hours. High-quality pseudorange data alone (model B) is the next-best option, measuring UTPM to an acceptable level within 24 hours.

Do Earth-orientation data continue to improve if observations are extended beyond 24 hours? And can UTPM be estimated every day with consistently high accuracy? To answer these questions, parameters that are not expected to change with time, such as station locations and GPS orbit epoch states, should be modeled as constant over the entire observing period. Earth-orientation parameters, which do change with time and for which one hopes to observe the most rapid variations possible, need to be periodically "reset," i.e., their uncertainties inflated at regular intervals during the observing session while subsequent data attempt to constrain their then-current values.

Figure 6 illustrates this situation. Data are taken for 48 hours, with all parameters except Earth orientation and clock errors modeled as constant over this time period. Just after 24 hours, the uncertainties of the Earth-orientation parameters are reset at their a priori values (while clock errors continue to be modeled as white noise). Since additional data need only constrain Earth orientation, these parameters improve much more rapidly than in the first 24 hours. For comparison, Earth-orientation parameters estimated as constant over the entire 48 hours are also shown. This figure is produced with model B, using high-quality pseudorange only; performance is significantly enhanced if carrier phase data are included.

At 24 hours, PMY error is estimated to be 3 cm. Twelve hours after resetting (i.e., at 36 hours) the error is already down to 6 cm (versus > 10 cm in the first 12 hours); by 48 hours, the error is again at 3 cm. This compares with an error of 2 cm if PMY were estimated as constant over the full 48 hours. Although the 48-hour estimate after resetting is no better than the 24-hour estimate, the error drops much more rapidly in the second 24 hours than during the first. Thus polar-motion measurements every 12 to 16 hours may be possible with this technique, i.e., after other parameters have been suitably constrained during the first 24 hours.

UT1–UTC and PMY rates show even more pronounced improvement. LOD at 24 hours shows an error of about 2 cm; continuing to 48 hours brings this down to < 0.5 cm, while resetting at 24 hours yields 3 cm at 36 hours and 1 cm at 48 hours. The polar-motion rate improves from 5 cm/day at 24 hours to < 0.5 cm/day after 48 hours, and to < 1.5 cm/day if reset after 24 hours.

All these values are comparable to or better than present-day measurement capabilities whose time resolutions are 24 hours or more. Thus long arcs of GPS data show promise for frequent high-quality Earth-orientation measurements. Future studies will model Earth orientation stochastically, treating each UTPM parameter as a random walk whose standard deviation is allowed to grow in a manner consistent with the empirical behavior of UTPM.

To see whether station locations, solar-radiation pressure, geocenter location, and wet-zenith troposphere delay are all truly necessary to be adjusted along with satellite orbits and Earth orientation, a number of covariance runs were performed in which these parameters were considered. One of these is shown in Fig. 7. In this model, 20-cm pseudorange and 1-cm carrier phase were assumed. Solar pressure and geocenter position were adjusted along with satellite epoch states and Earth orientation, while station locations and wet-zenith troposphere delays were considered. All the Earth-orientation parameters and their rates show a high sensitivity to the considered parameters, as the formal error (labelled "Data") is dwarfed by the uncertainties due to considered effects. The station location errors, considered at 5 cm for each component, dominate the total error ("Total RSS sigma"), while the wet troposphere delay, considered at 1 cm, has a smaller but significant effect.

These results suggest that, with 24 hours of data, station locations and wet-zenith troposphere delays do need to be adjusted in the estimation in order to generate reliable UTPM estimates (as has, in fact, been done). Additional consider runs (not shown) demonstrate that solar pressure is a significant error source even at 12 hours and needs to be adjusted, whereas the geocenter location does not seem to have a significant effect on Earth orientation at an uncertainty level of 10 cm. Since consider errors scale linearly with the uncertainty in the considered parameter, if station locations are known to better than 1 cm in each component, their effects will not be significant at 24 hours. Similarly, it is not necessary to adjust the wet-zenith troposphere delay if it is known to better than 1 cm.

IV. Conclusions

This covariance analysis demonstrates that

(1) High-quality length-of-day and polar-motion data can be obtained with GPS in under 24 hours, with a precision comparable to or greater than present-day techniques.

(2) UT1–UTC may be reliably determined if periodic reference frame ties are performed, and initial values of UT1–UTC measured, with VLBI using colocated receivers.

(3) The best combination of data types is high-quality pseudorange plus carrier phase, although high-quality pseudorange (data noise < 5 cm) alone performs well.

(4) It is necessary at present to adjust solar-radiation pressure and Y-bias, station locations, and wet-zenith troposphere delay in addition to satellite states and Earth-orientation parameters. However, if the a priori uncertainties in the station positions or troposphere can be reduced, they need not be included in the estimation process.

In conclusion, it appears that GPS will be a useful addition to the collection of techniques currently employed to measure Earth orientation, and it may provide a reliable, economical method of monitoring in near-real-time high-frequency variations in UTPM. For the DSN, GPS methods may considerably reduce the demand for antenna time needed to measure Earth-orientation parameters while simultaneously enhancing the parameters' accuracy.

References

[1] C. L. Thornton, J. L. Fanselow, and N. A. Renzetti, "GPS-based geodetic measurement systems," *Space Geodesy and Geodynamics*, A. J. Anderson and A. Cazenave, eds., London: Academic Press, pp. 197–218, 1986.

[2] I. I. Mueller and S. Zerbini, eds., *Lecture Notes in Earth Sciences, Vol. 22: The Interdisciplinary Role of Space Geodesy*, (proceedings of an international workshop held in Erice, Sicily, July, 1988), Berlin: Springer-Verlag, 1989.

[3] W. E. Carter and D. S. Robertson, "Project Polaris and IRIS: Monitoring polar motion and UT1 by very long baseline interferometry," *Space Geodesy and Geodynamics*, A. J. Anderson and A. Cazenave, eds., London: Academic Press, pp. 269-279, 1986.

[4] R. E. Neilan and W. G. Melbourne, "GPS global tracking system in support of missions to the planet Earth," *CSTG Bulletin No. 11, New Satellite Missions for Solid Earth Studies—Status and Preparations*, C. Reigber (ed.), Munich, Federal Republic of Germany: Deutsches Geodätisches Forschungsinstitut (DGFI), for The Commission on International Coordination of Space Techniques for Geodesy and Geodynamics (CSTG), pp. 129–140, June 1989.

[5] B. W. Parkinson and S. W. Gilbert, "NAVSTAR: Global Positioning System—ten years later," *Proc. IEEE*, vol. 71, no. 10, pp. 1177–1186, October 1983.

[6] R. E. Neilan, T. H. Dixon, T. K. Meehan, W. G. Melbourne, J. A. Scheid, J. N. Kellogg, and J. L. Stowell, "Operational aspects of CASA UNO 1988—The first large-scale international GPS geodetic network," *IEEE Trans. Instr. Meas.*, vol. 38, no. 2, pp. 648–651, April 1989.

[7] S. C. Wu and C. L. Thornton, "OASIS—A new GPS covariance and simulation analysis software system," *Proceedings First International Symposium on Precise Positioning with GPS–1985*, C. Goad (ed.), Rockville, Maryland, vol. 1, pp. 337–345, April 15–19, 1985.

[8] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation*, Orlando, Fla.: Academic Press, 1977.

[9] J. M. Davidson et al., *The Spring 1985 high precision baseline test of the JPL GPS-based geodetic system: A final report*, JPL Publication 87-35, Jet Propulsion Laboratory, Pasadena, California, November 15, 1987.

[10] S. M. Lichten and J. S. Border, "Strategies for high-precision Global Positioning System orbit determination," *J. Geophys. Res.*, vol. 92, no. B12, pp. 12751–12762, November 10, 1987.

Table 1. Parameter-estimation strategy (a priori sigmas)

Earth-orientation parameters	
UT1–UTC rate	10^{-7} ($\sim$ 10 msec/day)
PMX, PMY	80 nrad ($\sim$ 16 marcsec)
PMX, PMY rates	10^{-11} rad/sec ($\sim$ 200 marcsec/day)
Satellite parameters (18-satellite constellation)	
X, Y, Z positions	10 m (each component)
X, Y, Z velocities	1 mm/sec (each component)
Solar-radiation pressure (X, Z)*	50 percent
Y-Bias*	10^{-12} km/sec^2 (100 percent)
Station parameters (6 stations with global distribution)	
DSN station locations*	3 cm (each component)
Non-DSN station locations*	5 cm (each component) (both 5 cm, if considered)
Wet-zenith troposphere delay*	10 cm (1 cm, if considered)
Other parameters	
Geocenter*	100 m (each component) (10 cm, if considered)
Carrier phase biases	10 km
Satellite and station clocks (modeled as white noise)	1 km (except one station)
Data noise	
Pseudorange	20 cm, 5 cm (5-min batches)
Carrier phase	1 cm, 0.5 cm (5-min batches)

* These parameters were considered in the covariance analyses discussed in the text, but are estimated in the models shown in Table 2.

Table 2. Models

Model	Data
A ("worst case model")	20-cm pseudorange only
B	5-cm pseudorange only
C ("best case model")	5-cm pseudorange 0.5-cm carrier phase

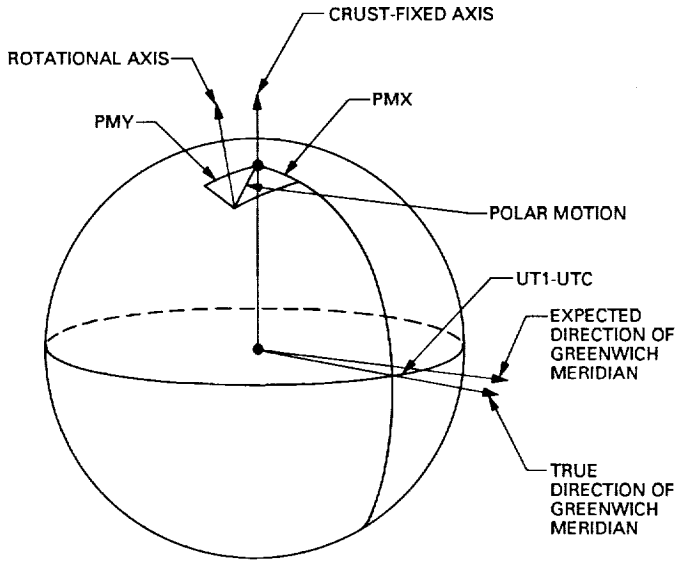


Fig. 1. Schematic illustration of the components of Earth orientation: polar motion (PMX, PMY) and UT1–UTC.

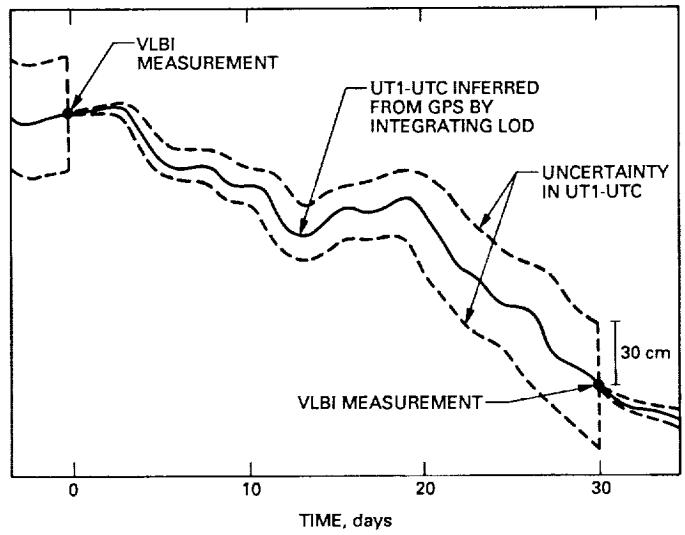


Fig. 3. Illustration of how precise periodic VLBI measurements of UT1–UTC can be combined with daily GPS LOD measurements to constrain the uncertainty in UT1–UTC.

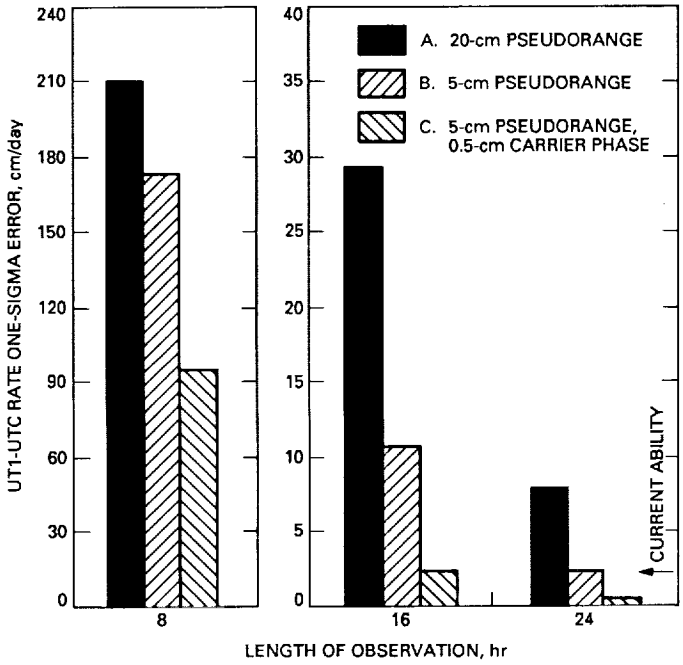


Fig. 2. Predicted uncertainty in the rate of change of UT1–UTC as a function of observation time. The arrow at right indicates present-day measurement capability in this and subsequent figures. Length-of-day uncertainty can also be found from this figure, as LOD = (UT1–UTC rate) × 1 day. The units of LOD are cm.

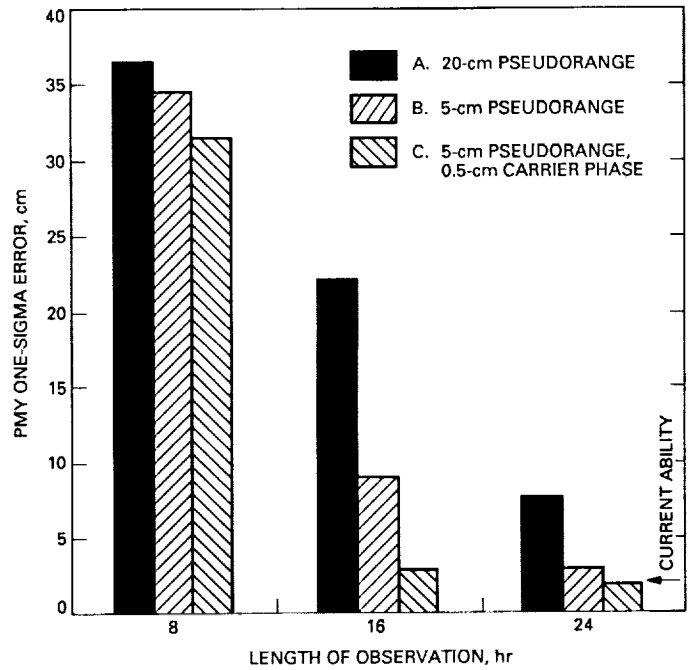


Fig. 4. Predicted uncertainty in Y polar motion (PMY) as a function of observing time. X polar motion (PMX) behaves in a similar manner.

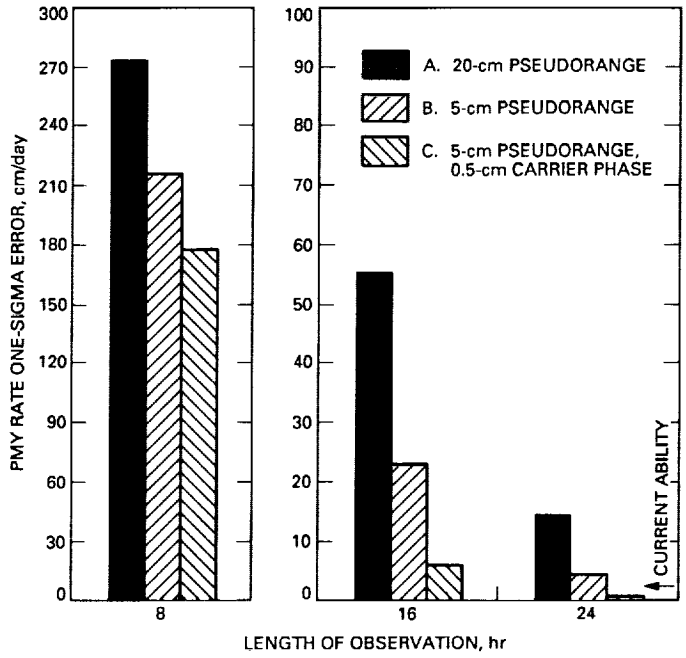


Fig. 5. Predicted uncertainty in the rate of change of Y polar motion as a function of observing time. The PMX rate behaves similarly.

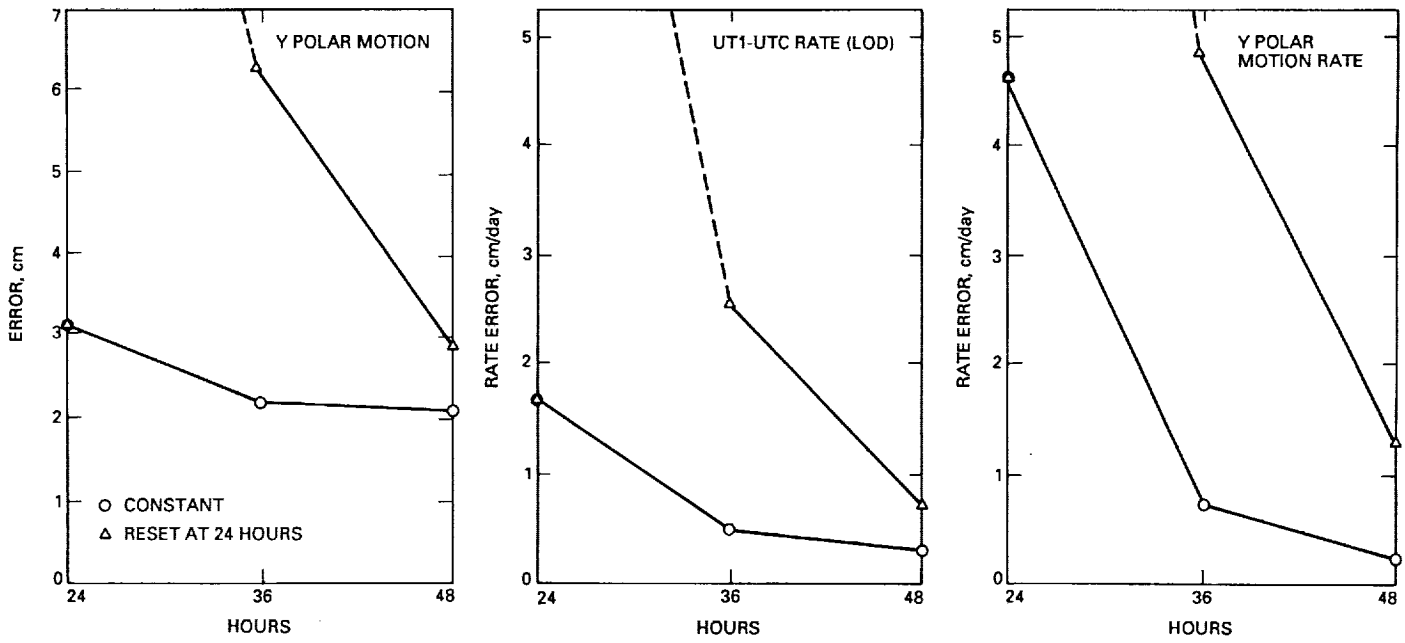


Fig. 6. Predicted improvement in estimates of UTPM after two days. Two cases are shown in which Earth-orientation parameters (o) are modeled as constant over 48 hours and in which these parameters (△) are reset to large values after 24 hours. Model B is used, employing only high-quality pseudorange data.

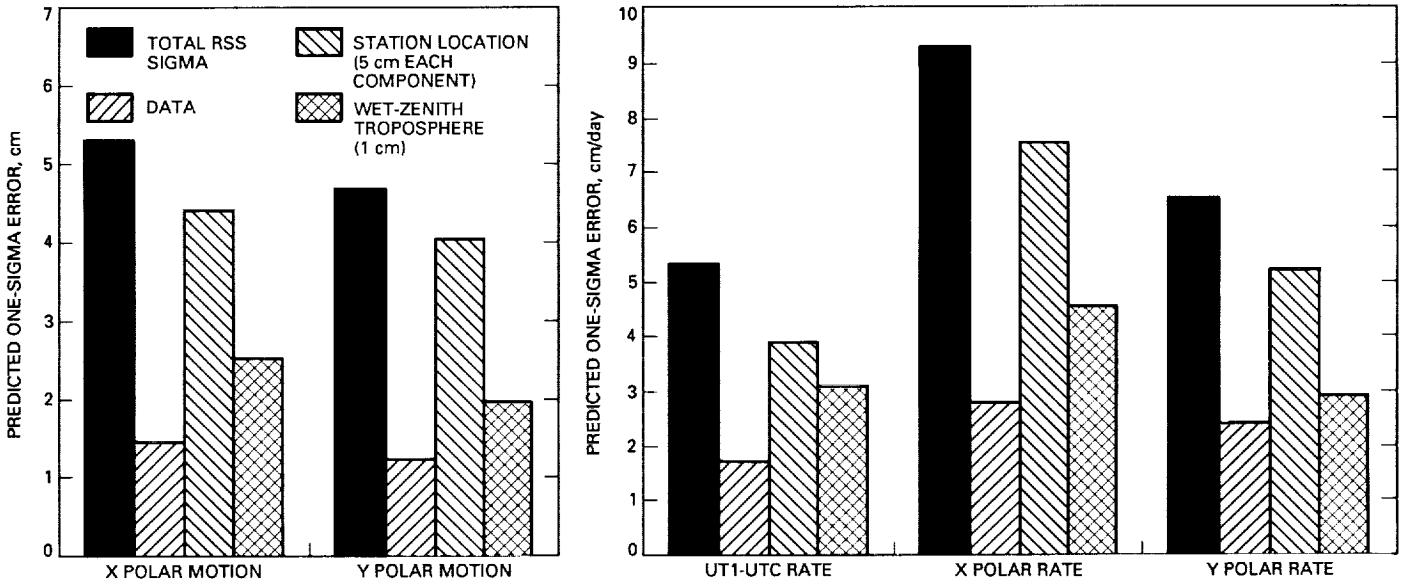


Fig. 7. Consider errors at 24 hours. Station locations and wet-zenith troposphere delay were considered with uncertainties of 5 cm per component and 1 cm, respectively. In this covariance analysis, satellite epoch states, Earth orientation, solar pressure, clocks, carrier phase biases, and the geocenter were all adjusted. The assumed data noise was 20 cm pseudorange and 1 cm carrier phase. The "Total RSS sigma" denotes the root sum square of the adjusted error estimate ("Data") and the two considered errors.

TDA Progress Report 42-99

November 15, 1989

Simple Analytic Potentials for Linear Ion Traps

G. R. Janik, J. D. Prestage, and L. Maleki
Communications Systems Research Section

A simple analytical model has been developed for the electric and ponderomotive (trapping) potentials in linear ion traps. This model was used to calculate the required voltage drive to a mercury trap, and the result compares well with experiments. The model gives a detailed picture of the geometric shape of the trapping potential and allows an accurate calculation of the well depth. The simplicity of the model allowed an investigation of related, more exotic trap designs which may have advantages in light-collection efficiency.

I. Introduction

Radio frequency (RF) quadrupole ion traps have great importance to the development of new atomic frequency standards and high-precision measurements. The three-dimensional quadrupole trap has mainly been used. The ideal geometry for the electrodes of this trap is hyperboloids of revolution, which produce a pure quadrupole field. Practical traps are built with a different, more open geometry because of the need to collect light efficiently from the ions. Ion fluorescence is used to determine its quantum state, and the collection efficiency is an important factor in the signal-to-noise ratio. Deviations in the electric field and trapping potential from a pure quadrupole are usually not considered in detail due to the difficulty in calculating them.

In order to increase the number of trapped ions without degrading the frequency stability, the Time and Frequency Systems Research Group recently introduced into frequency-standard research a linear trap based on the quadrupole mass spectrometer [1]. A side benefit of this trap geometry is the ability to calculate the electric and trapping potentials to good accuracy with a simple analytic model. In this article, the model is developed and used to predict an ion resonance frequency, which has been experimentally measured. The article then shows how the model can also be used to investigate similar but more exotic geometries which may be advantageous in some applications.

The linear trap consists of four parallel cylindrical rods arranged with their centers on the corners of a square. An RF voltage is applied to the rods so that nearest neighbors have opposite polarity. This creates an alternating two-dimensional quadrupole electric field between the rods. The field confines ions along the center axis of the trap by the ponderomotive force, just as in a quadrupole mass spectrometer. Two endcap electrodes with a DC bias voltage applied confine the ions axially. This article considers only the center axial region, far enough from the endcaps so that the field is essentially two-dimensional. The DC fields from the endcaps decay exponentially along the axis,

so that this restriction applies to the vast majority of the trapping volume.

The electric fields due to four cylindrical rods have been studied for years by designers of mass spectrometers and particle accelerators. In order to achieve the closest approximation to a true quadrupole field, it has been found that the ratio of the rod radius to the distance between the rod center and the trap axis reaches its optimum value at 0.5342. This value was first determined by measurements on a quadrupole accelerator magnet [2] and later reproduced with numerical calculations [3].

A radius-spacing ratio of 0.5342 produces a geometry that is too closed to allow efficient light collection from ions. A pure quadrupole field is not really necessary if ion confinement is the only goal; hence, in this case the ratio can be reduced. The trap discussed here has a ratio of 0.25, and even smaller ratios might be desirable. The smallness of the ratio leads directly to a simple approximation, namely, the fields produced by infinitely thin rods. A conducting rod with an applied voltage has some induced charge, and as the rod diameter is reduced, all the charge coalesces into a line. This idea can be used to transform a two-dimensional boundary-value problem into a much simpler calculation of the potential produced by fixed sources. All the calculations that follow are based on fields generated by an array of uniform parallel line charges.

II. Model for a Four-Rod Trap

Since the model field is two-dimensional, the method of complex variables can be used. This is not necessary, but makes the calculation of the trapping potentials a little easier. Adopting the notation of Landau and Lifshitz [4], a positive line charge at $z = z_0$ produces a complex potential $w = -\log(z - z_0)$, whose real part is the ordinary electric potential. Dimensionless quantities are now used to calculate the geometric form of the potentials. Scale factors are introduced later to calculate real trap parameters. The model for the quadrupole trap consists of two negative line charges at $z = \pm i$ and two positive line charges at $z = \pm 1$. This produces the complex potential

$$w = \ln\left(\frac{z^2+1}{z^2-1}\right)$$

The ponderomotive trapping potential is proportional to the square of the electric field F [5]. To calculate this, note that $-F_x + iF_y = dw/dz$, so that

$$|F|^2 = \left|\frac{dw}{dz}\right|^2 = \frac{16|z|^2}{|z^4-1|^2}$$

Changing to polar coordinates, it is found that

$$|F|^2 = \frac{16r^2}{r^8 - 2r^4\cos 4\phi + 1} \equiv \Gamma G(r,\phi) \qquad (1)$$

where $\Gamma = 16$. One can expand G about the origin to obtain

$$G(r,\phi) \approx r^2\left[1 + 2r^4\cos 4\phi + r^8(1 + 2\cos 8\phi) + \cdots\right]$$

where the leading term r^2 is an isotropic harmonic potential, and the higher-order terms have at least fourfold symmetry. It turns out that any two-dimensional field configuration that vanishes at a point produces an isotropic harmonic trapping potential in lowest order about that point. In the trap configurations analyzed later, the function G is always defined to have a leading term r^2, and Γ is used for the numerical factor.

The trapping potential model function $G(x,y)$ is shown in a three-dimensional plot in Fig. 1, and shows a center well rising up to singularities at the rods. Halfway between the rods are saddle points beyond which the potential drops again. The height of the saddle points sets the maximum energy an ion can have and still stay trapped, i.e., the well depth. The well depth can be calculated by setting $\phi = \pi/4$ in Eq. (1) and finding the maximum value of $G(r,\pi/4)$. The maximum occurs at $r_s = 3^{-1/4} = 0.760$, and has a value of $G_s = 9/16\sqrt{3} = 0.3248$, where the subscript s is the saddle point.

Now that the electric and trapping potentials produced by four line charges have been calculated, a real trap may be modeled. The electric equipotentials of the model are plotted as contours in Fig. 2. Close to the charges, they have a nearly circular shape. The equipotentials of a real trap are exactly circular at the electrode surfaces. The deviation of the model's equipotential from circularity is calculated by finding its horizontal and vertical "diameters" and taking the difference. The electric potential V is the real part of w, and is given by

$$e^{2V} = \frac{x^4 + y^4 + 2(x^2 - y^2 + x^2y^2) + 1}{x^4 + y^4 + 2(y^2 - x^2 + x^2y^2) + 1} \equiv B^2 \qquad (2)$$

The equipotential surrounding the charge at $z = 1$ is to be examined here, so first the intersection of the equipotential with the x-axis is found by using

$$e^V = \frac{|1 + x^2|}{|1 - x^2|} = B \qquad (3)$$

For a fixed value of V, Eq. (3) has two positive and two negative solutions. If the smallest positive solution is denoted x_0, and the other positive solution is denoted x_1, then Eq. (3) produces the following relationships:

$$x_1 = \frac{1}{x_0} = \sqrt{\frac{B+1}{B-1}} \tag{4a}$$

$$d \equiv x_1 - x_0 = \frac{1-x_0^2}{x_0} = \frac{2}{\sqrt{B^2-1}} \tag{4b}$$

$$c \equiv \frac{x_0+x_1}{2} = \frac{1+x_0^2}{x_0} = \frac{B}{\sqrt{B^2-1}} \tag{4c}$$

which are valid for $B > 1$. Notice that there is a symmetry between B and $1/B$ corresponding to the symmetry between V and $-V$. The case of $B < 1$ applies to the region near the charges at $z = \pm i$.

Point x_0 corresponds to the inner surface of a rod, x_1 to the outer surface of a rod, d to the horizontal diameter of the rod, and c to the center of the rod. The "rod" defined here is a hypothetical electrode shaped so that four of them will produce a field distribution equivalent to the four line charges. To calculate the vertical diameter, one can use the intercepts $\pm y_0$ on the line $x = c$. Using Eq. (2) results in

$$y_0^2 = \frac{2B\left(\sqrt{(B^2+2)} - B\right) - 1}{B^2-1} \approx \frac{1}{B^2}$$

The approximation is very good for $B \geq 2$, and improves as B increases. The fractional difference of the rod diameters, $(d - 2y_0)/d$, scales as $1/(2B^2)$, and is consistent with the idea that the equipotentials become more circular as the line charge is approached. A curious feature of the model is evident in Eq. (4c). The center of the rod c depends on B and does not coincide with the line charge at $x = 1$. Seen another way, the line charge sits at the geometric mean of the inner and outer surfaces of the rod (x_0 and x_1) and only approaches the arithmetic mean as B becomes large.

So far, it has been assumed that the boundary is infinitely far away. This is allowed since the potential V in Eq. (2) tends toward zero for large r. The fact that the total charge is zero assures this. Denison [3] has numerically calculated the effect of a circular grounded boundary at $r = 1.65$ for an optimized quadrupole with a radius-spacing ratio of 0.5342. He found that the boundary changed the optimum ratio by about 1 percent. Although a more open trap may be more sensitive to a boundary, practical traps will have their boundaries at greater distances than 1.65. Therefore, boundaries will be ignored in this treatment.

In order to use the model to calculate the parameters of a trap, the "squashed" rods of the model are associated with the round electrodes of the trap, and the linear dimensions are scaled appropriately. Dimensions of the real trap are given in capital letters; model dimensions are in lower case. The distance of the line charges from the trap axis is A, and they will not be exactly at the center of the electrodes. The inner and outer x-axis intercepts of the trap rods are called X_0 and X_1, respectively. Therefore, the dimensionless coordinates are given by

$$x_0 = \frac{X_0}{A} \tag{5a}$$

$$x_1 = \frac{X_1}{A} \tag{5b}$$

$$A^2 = X_0 X_1 \tag{5c}$$

The mercury ion linear trap has dimensions $X_0 = 7.62$ mm and $X_1 = 12.7$ mm, implying $A = 9.84$ mm (compared to 10.2 mm for the true center), $x_0 = 0.774$, $x_1 = 1.29$, and $B = 3.99$. The horizontal and vertical diameters of the equivalent model trap rod differ by 3 percent, providing a good approximation to a circle.

In order to test the predictive power of the model, the applied voltage needed for the mercury ion trap is calculcated to produce a natural ion resonance frequency ω of $2\pi \times 48.5$ kHz. For small displacements, the leading r^2 term in the trapping potential leads to harmonic motion. The ions are detected by amplifying the current they induce on the trap rods at the natural frequency, and the amplifiers are tuned to 48.5 kHz. The driving voltage is applied at a frequency Ω of $2\pi \times 500$ kHz in a balanced mode, so that the peak voltage on two opposing rods is V_0 and the voltage on the other two rods is $-V_0$ with respect to the vacuum system. The balanced drive keeps the trap axis at zero potential, which allows the application of only DC bias to the endcaps. The amplitude of the applied RF voltage is ramped until a resonance signal appears, indicating that the natural frequency of the trap matches the frequency to which the amplifiers are tuned.

Calculating the required applied voltage uses the fact that on the inner surface of the rod x_0, the electric potential V equals the applied voltage V_0, so that

$$\frac{V}{V_0} = \frac{\ln|(1+x^2)/(1-x^2)|}{\ln|(1+x_0^2)/(1-x_0^2)|} = \frac{\ln|(1+x^2)/(1-x^2)|}{\ln B}$$

and the electric field is

$$|E|^2 = \frac{\Gamma V_0^2 G(x,y)}{(A \ln B)^2}$$

The ponderomotive trapping potential Ψ is given by

$$\Psi = \frac{e^2 E^2}{4m\Omega^2} = \frac{\Gamma e^2 V_0^2 G(x,y)}{4m\Omega^2 (A \ln B)^2} \tag{6}$$

where m is the mass of the ion [5]. Using the harmonic approximation for G leads to an expression for the natural resonance frequency

$$\omega = \frac{\sqrt{\Gamma}\, eV_0}{\sqrt{2}\, m\Omega^2 A^2 \ln B} \tag{7}$$

For mercury isotope 199 and the trap parameters given above, Eq. (7) predicts a peak drive voltage of 93.7 V. The experimental value is 100 ± 5 V, giving agreement almost within experimental error. The well depth can also be calculated using Eqs. (6) and (7) as

$$\Psi_s = \frac{G_s m\omega^2 A^2}{2} \tag{8}$$

yielding a value of 3.00 eV for the parameters here.

III. Model for a Two-Rod Trap

The simplicity of this model makes it useful for analyzing other two-dimensional trap geometries. Any system that produces a point with a vanishing electric field can potentially trap ions. The simplest system consists of two rods driven by the same voltage with respect to a distant boundary. This configuration has a zero field point midway between the two rods. A trap of this type was demonstrated with oil droplets by Straubel [6] very early on, and may be of use in frequency standards due to its wider viewing angle. The model for this trap is two equal negative line charges at $z = \pm i$. The complex potential is

$$w = \ln(z^2 + 1)$$

and the geometric form of the ponderomotive potential is

$$G(r,\phi) = \frac{r^2}{r^4 + 2r^2 \cos 2\phi + 1}$$

with $\Gamma = 4$. Contours of the electric equipotentials and a plot of the ponderomotive potential are shown in Figs. 3 and 4. The two saddle points are on the x-axis at $r_s = 1$, and have the value $G_s = 1/4$. The electric potential can be written as

$$e^{2V} = x^4 + y^4 + 2(x^2 - y^2 + x^2 y^2) + 1 \equiv B^2 \tag{9}$$

One can calculate the intercepts of an equipotential on the y-axis around the charge at $z = i$ and obtain relations analogous to those of the four-rod trap:

$$y_0 = \sqrt{1 - B} \tag{10a}$$

$$y_1 = \sqrt{1 + B} \tag{10b}$$

for $B < 1$. Combining Eqs. (10a) and (10b) to calculate the scaling law for a real trap results in $2A^2 = Y_0^2 + Y_1^2$, where Y_0 and Y_1 are the inner and outer surfaces of the rod. For a trap with rods the same size and spacing as two opposing rods of the mercury ion trap, $A = 10.5$ mm. This time the line charge is more distant than the rod center (10.2 mm). A calculation of the diameter difference gives a deviation from roundness of 3 percent.

In order to calculate the applied voltage, the surrounding boundary must be taken into account, because the bars must be driven with respect to something. According to Eq. (9) and Fig. 3, the equipotentials for large r become roughly circular with a value of $2 \ln r$, and it is assumed that the boundary follows one of these contours. The potential on the trap axis is zero and the potential on the rod is $\ln B$. The applied voltage will be proportional to $2 \ln r - \ln B$, where $r = R/A$ and R is the radius of the surrounding boundary. The trapping potential can then be written as

$$\Psi = \frac{\Gamma e^2 V_0^2 G}{4m\Omega^2 A^2 (2 \ln r - \ln B)^2}$$

The dimensions of the vacuum system correspond to a value of 2.5 for r. In order to duplicate the natural resonance of 48.5 kHz with the two-rod trap, a peak drive voltage of 364 V is necessary, an increase by a factor of 3.9 over the four-rod trap. The well depth is still governed by Eq. (8), and depends on the boundary only through the voltage necessary to maintain ω. The well depth then is 2.63 eV. The price paid for an increased viewing angle is a larger drive voltage and a slightly smaller well depth.

An added complication with this trap is that the electric potential at the trap axis is now oscillating with respect to the vacuum system at the drive frequency. This makes an added AC bias to the endcaps necessary to keep them

at a constant potential above the trap axis. Straubel did not use endcaps, but relied upon fringing fields from the ends of his rods for axial confinement.

IV. Models for a Three-Rod Trap

Although the viewing angle is larger in the two-rod trap, the trapping point is still between the two rods. It may be desirable for some applications to trap well outside the electrode structure. This becomes possible with a three-rod trap. A simple example of a three-rod trap model consists of a positive line charge at the origin flanked by two equal negative line charges on the real axis at $z = \pm 1$. The resulting complex and ponderomotive potentials are:

$$w = \ln\left(\frac{z^2 - 1}{z}\right)$$

and

$$|F|^2 = \frac{r^4 + 2r^2\cos 2\phi + 1}{r^2(r^4 - 2r^2\cos 2\phi + 1)} \qquad (11)$$

The electric field vanishes at $z = \pm i$, and these are the trapping points. The electric and trapping potentials are shown in Figs. 5 and 6.

If Eq. (11) is expanded about the trapping point $z = i$, then $\Gamma = 4$. The expression for G is too complicated to be very useful. The lowest saddle point can be found from Eq. (11) at $r_s = \sqrt{2+\sqrt{5}} = 2.058$, and $\phi = \pi/2$ where $G_s = 0.0217$. This gives a well depth of 200 meV for the parameters used here. The question of the applied voltage is slightly more complicated than in the previous two cases. Since the total charge is not zero, the boundary must again be included in the calculation. There are now three voltages involved: the central rod V_1, the outer rods V_2, and the boundary V_3. The ratios between these voltages can be calculated from the expression for the potential

$$e^{2V} = \frac{x^4 + y^4 + 2(y^2 - x^2 + x^2y^2) + 1}{x^2 + y^2} \equiv B^2$$

and the various intercepts. For the center rod ($B_1 > 1$), the relation between the B value and the dimensionless diameter is

$$d_1 = B_1\left(\sqrt{1 + \frac{4}{B_1^2}} - 1\right) \approx \frac{2}{B_1}$$

and for the outer rods ($B_2 < 1$)

$$d_2 = B_2$$

and

$$c_2 = \sqrt{1 + \frac{B_2^2}{4}}$$

Once again, the intercepts of the outer rod obey the relations of Eq. (5). The boundary is again assumed to conform to a long-range equipotential at dimensionless radius r with value $\ln r$. All the potentials can be offset by $\ln r$ to keep the boundary at ground and obtain the following form for the applied voltages:

$$\frac{V_1}{V_2} = \frac{\ln(B_1/r)}{\ln(B_2/r)}$$

Using a value of 2.5 for r and the same rod diameters and spacing as before obtains peak applied voltages of 49.4 V and -214 V for V_1 and V_2. A suitably tapped transformer could be used to provide these drive voltages. Alternatively, the ratio of the inner and outer rod diameters could be adjusted to force $B_1B_2 = r^2$ so that $V_1 = -V_2$. Even with this simplification, an additional AC voltage is necessary to bias the endcaps as in the two-rod case.

The influence of the boundary can be removed by choosing the total charge to be zero. This can be accomplished by doubling the center charge. Unfortunately, there is no trapping point with this configuration while the charges are in a line. Moving the center charge down along the imaginary axis to $z = -ib$ creates a trapping point at $z = i/b$. The case of $b = 1$ has been analyzed, and the results are shown here. The electric and trapping potentials are plotted in Figs. 7 and 8. The scaling factor Γ is 1/4, and the saddle point is on the y-axis at $y_s = 1.839$ with $G_s = 0.0728$. The electric potential is

$$e^{2V} = \frac{x^4 + y^4 + 2(y^2 - x^2 + x^2y^2) + 1}{(x^2 + y^2 + 2y + 1)^2}$$

and the x intercepts of the outer rods still follow the relationships of Eqs. (4) and (5), with B replaced by $1/B_1$ (for $B_1 < 1$). The y intercepts of the lower rod obey the relationship below (for $B_2 > 1$):

$$y_{0,1} = \frac{-B_2 \pm \sqrt{2B_2 - 1}}{B_2 - 1}$$

and Eq. (5). Once again, the voltages applied to the rods will not be equal and opposite unless the diameters are adjusted so that $B_1 = 1/B_2$. The endcaps need only DC bias because the potential is zero at the trapping point.

The small value of Γ, which is 1/4, makes the drive voltage for the outer rods eight times that of the four-rod trap for the same ion resonance frequency. This fact, plus the small well depth, may limit the usefulness of this trap. In both three-rod traps, however, the low saddle point lies only on one side of the trapping region. The potential barrier on the opposite side is much higher. These traps may be useful for trapping macroscopic particles, where gravity plays an important role. The trap rods could be oriented horizontally so that the low saddle point lies above the trapping region.

V. Conclusions

This article has presented a simple model for linear ion traps that permits the accurate calculation of trapping parameters and gives a detailed picture of the potentials. The model was used to analyze some new trap geometries, and their advantages and disadvantages were discussed. Each trapping geometry is characterized by the two parameters Γ and G_s, which determine the dependence of its natural resonance frequency and well depth on the applied RF voltage and trap dimensions.

Acknowledgments

The authors thank C. Greenhall and G. J. Dick for helpful discussions.

References

[1] J. D. Prestage, G. J. Dick, and L. Maleki, "New Ion Trap for Frequency Standard Applications," *J. Appl. Phys.*, vol. 66, pp. 1013–1017, 1989.

[2] I. E. Dayton, F. C. Shoemaker, and R. F. Mozley, "The Measurement of Two-Dimensional Fields, Part II: Study of a Quadrupole Magnet," *Rev. Sci. Instr.*, vol. 25, pp. 485–489, 1954.

[3] D. R. Denison, "Operating Parameters of a Quadrupole in a Grounded Cylindrical Housing," *J. Vac. Sci. Tech.*, vol. 8, pp. 266–269, 1970.

[4] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Oxford, UK: Pergamon Press, pp. 13–14, 1960.

[5] H. G. Dehmelt, "Radiofrequency Spectroscopy of Stored Ions, I: Storage," *Adv. At. Mol. Phys.*, vol. 3, pp. 53–154, 1967.

[6] H. Straubel, "Kurze Originalmitteilungen," *Naturwissenschaften*, vol. 42, pp. 506–507, 1955.

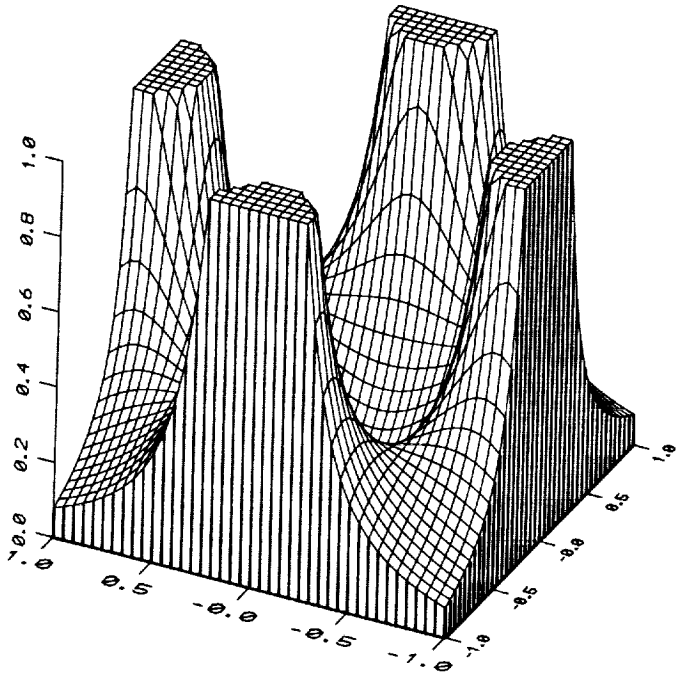


Fig. 1. The ponderomotive potential function G for a four-rod quadrupole trap is plotted in three dimensions. The singularities are truncated at the value 1.0 for clarity.

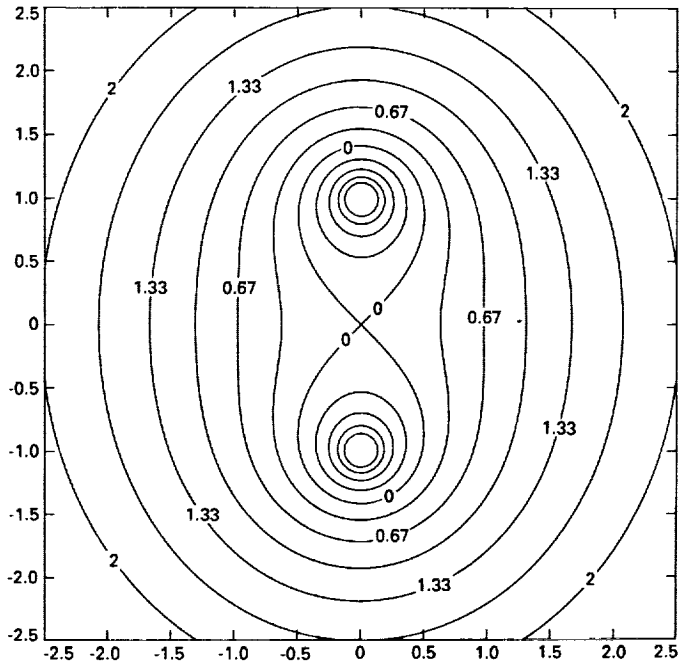


Fig. 3. The electric potential for a two-rod trap is shown in this contour plot. The range of potentials shown is −1 to 2 at intervals of 1/3.

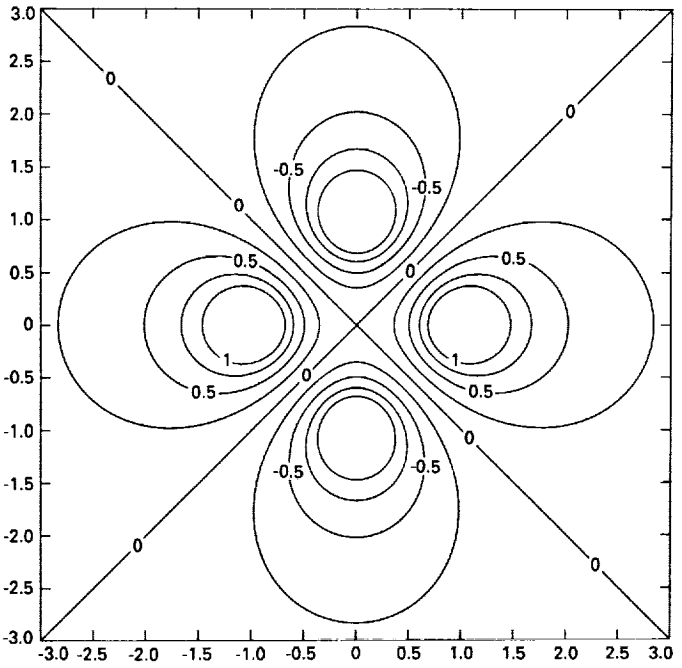


Fig. 2. The electric potential for a four-rod trap is shown in this contour plot. The range of potentials shown is −1 to 1 at intervals of 0.25.

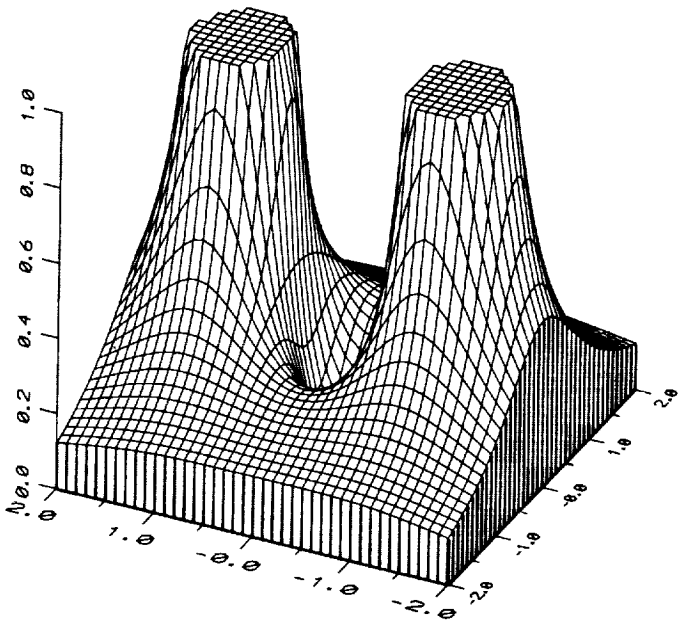


Fig. 4. The ponderomotive potential function G for a two-rod quadrupole trap is plotted in three dimensions. The singularities are truncated at the value 1.0 for clarity.

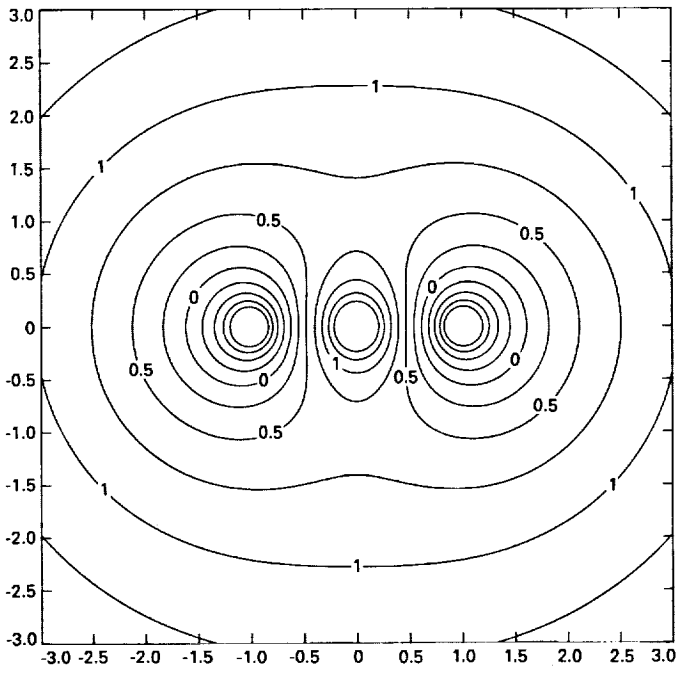


Fig. 5. The electric potential for a three-rod trap is shown in this contour plot. The range of potentials shown is −1 to 1.5 at intervals of 0.25.

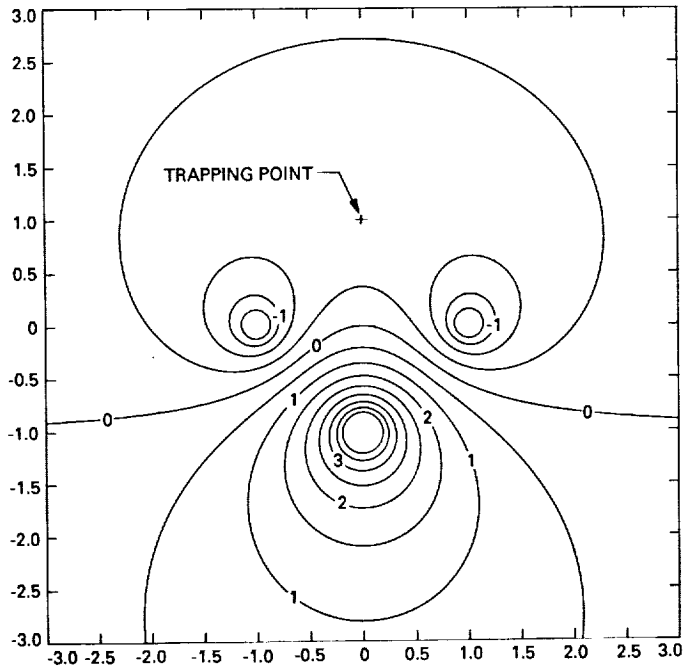


Fig. 7. The electric potential for a three-rod trap with zero net line charge is shown in this contour plot. The range of potentials shown is −2 to 4 at intervals of 0.5.

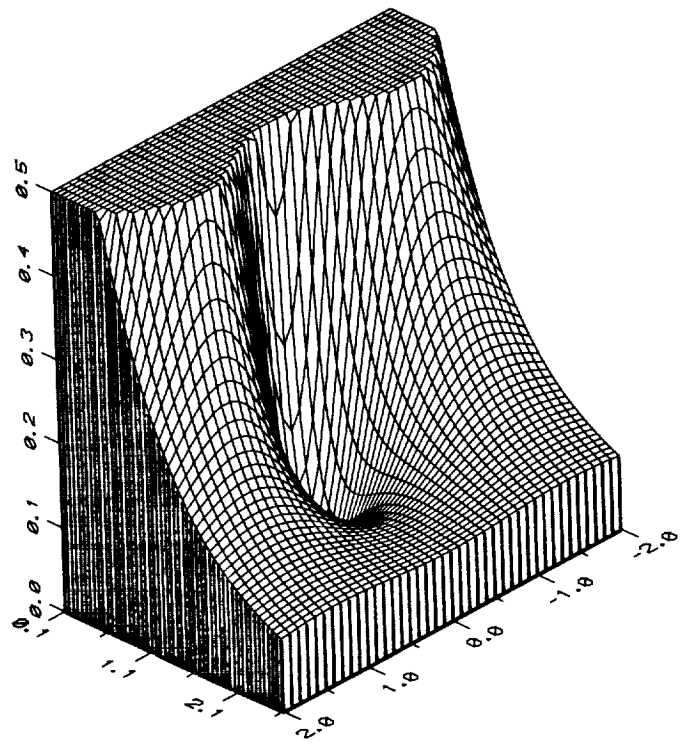


Fig. 6. The ponderomotive potential function $|F^2|$ for a three-rod quadrupole trap is plotted in three dimensions. The singularities are truncated at the value 0.5 to emphasize the potential well at $z = i$.

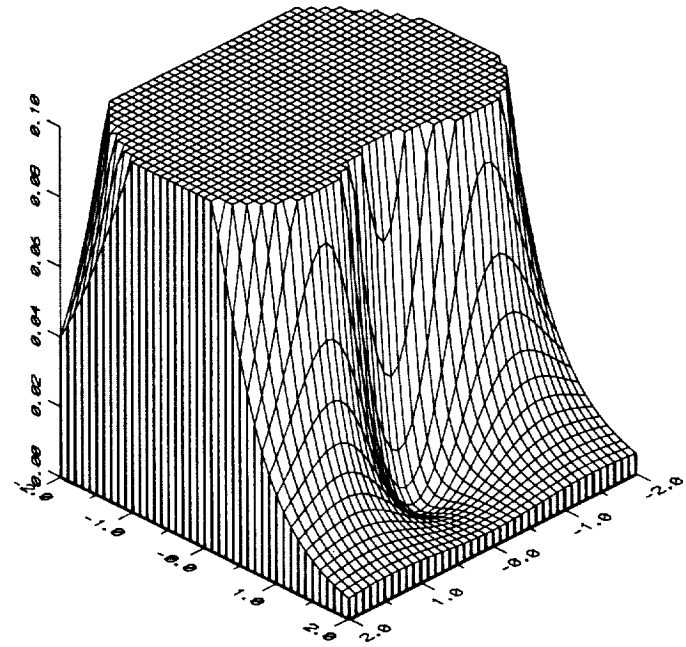


Fig. 8. The ponderomotive potential function $|F^2/4|$ for a three-rod quadrupole trap with zero net line charge is plotted in three dimensions. The singularities are truncated at the value 0.1 to emphasize the potential well at $z = i$.

Microwave Oscillator With Reduced Phase Noise by Negative Feedback Incorporating Microwave Signals With Suppressed Carrier

G. J. Dick and J. Saunders
Communications Systems Research Section

This article develops and analyzes oscillator configurations which reduce the effect of $1/f$ noise sources for both direct feedback and stabilized local oscillator (STALO) circuits. By appropriate use of carrier suppression, a small signal is generated which suffers no loss of loop phase information or signal-to-noise ratio. This small signal can be amplified without degradation by multiplicative amplifier noise, and can be detected without saturation of the detector. Together with recent advances in microwave resonator Qs, these circuit improvements will make possible lower phase noise than can be presently achieved without the use of cryogenic devices.

I. Introduction

Phase fluctuations in microwave oscillators show a characteristic $1/f^3$ spectral density for frequencies very close to the carrier. The spectral density $S_\phi(f)$ can be expressed in terms of rad^2 per Hz bandwidth at an offset frequency f from the carrier. At larger frequency offsets ($f > 10$ kHz), fluctuations decrease more slowly, typically approaching a more or less constant value for frequencies of 100 kHz and above. This article is concerned with reduction of the fluctuations very close to the carrier by the use of circuit techniques not previously applied to microwave oscillators.

In order to achieve a microwave signal with the lowest possible noise at all offset frequencies, a variety of technologies is presently used [1, 2, 3]. Typically, high stability for small offsets ($f < 100$ Hz) is obtained by locking the microwave oscillator to a harmonic of a low-frequency (5 MHz) bulk acoustic wave (BAW) quartz-crystal oscillator. For offset frequencies in the range of 100 Hz $< f <$ 10 kHz, further stabilization may be provided by a surface acoustic wave (SAW) oscillator operating at $\approx$ 500 MHz. The microwave oscillator itself provides the best possible stability only for relatively large offset frequencies ($f \geq$ 10 kHz).

While all the sources listed above show $1/f^3$ type phase fluctuations for small values of offset frequency f, the microwave oscillator itself shows by far the highest noise. This is due to two contributing factors: the low Q available for microwave resonators and the large $1/f$

phase noise in available microwave devices. (The action of the oscillation loop converts this $1/f$ phase noise into $1/f$ frequency noise, which is mathematically equivalent to $1/f^3$ phase noise [1, 16].) Thus, while BAW and SAW quartz crystals show Qs of 10^6 and 10^5 respectively, microwave cavities and dielectric resonators are limited to Qs in the range 1–3 $\times 10^4$. In like manner, active devices which are available at the lower operating frequencies of the quartz devices show $1/f$ phase noise of -140 dB per Hz or lower at an offset frequency of $f = 1$ Hz, while the best 8–12 GHz (X-band) amplifiers have noise of -120 dB per Hz [13, 14].

Recently, a new type of microwave resonator has been demonstrated with Qs 10 to 1000 times larger than previously available [4–11]. This sapphire whispering-gallery-mode resonator allows the intrinsic Q of the sapphire itself to be utilized by isolating the fields to the sapphire element and away from lossy metallic container walls. These resonators have shown Qs of 2×10^5 at room temperature and 3×10^7 at 77 K. Tests of an 8-GHz oscillator stabilized by such a sapphire resonator (a stabilized local oscillator, or STALO) show $1/f^3$ noise 10 dB lower than previously reported for any 8–12 GHz (X-band) source and only 22 dB higher than the best quartz-crystal stabilized oscillator [9, 12]. If the device noise in this oscillator could be reduced by 20 dB or more, the need for quartz stabilization would be eliminated, and a new oscillator capability would become available.

The following sections describe two substantially different implementations of an idea in which negative feedback in the oscillator is generated by means of a suppressed-carrier microwave signal. In one implementation, the signal is fed to a semiconducting phase detector to enhance its sensitivity while avoiding saturation. In this STALO configuration, phase-detector noise is effectively reduced by the enhanced sensitivity. In the second implementation, direct RF feedback of a signal with suppressed carrier induces both oscillation and negative phase feedback. In this case, the degree of improvement in $1/f$ phase noise over that which characterizes the amplifier itself is equal to the degree of negative feedback which can be achieved without oscillation in unwanted modes.

A comparison of the two implementations shows the second (direct RF feedback) to be somewhat trickier in concept (involving both negative-phase feedback and positive-amplitude feedback in the same loop), and simpler in realization. Conditions of stability require the use of a filter with a performance that is expected to limit the usable loop gain to the 20–30 dB range. The STALO-type implementation, while more elaborate, is somewhat less tricky, and should allow larger gains to be used. Both should show crystal-oscillator-type performance in a room-temperature 10-GHz (X-band) oscillator using a whispering-gallery-mode sapphire resonator with an intrinsic Q of 2×10^5, and allow dramatic improvements in the state of the art with cooled resonators and higher Qs.

II. Background

Figures 1 and 2 show conventional microwave self-excited oscillator and STALO configurations, together with an identification φf the in-oscillator and oscillator output noise spectral densities. In the self-excited oscillator shown in Fig. 1, the oscillation condition requires that the phase shift around the complete feedback loop comprising the amplifier, resonator, and interconnections be a multiple of 2π. With this condition satisfied, any phase fluctuation in the microwave amplifier must be accompanied by an opposite shift of equal magnitude in the resonator. For slow phase fluctuations ($f \ll \nu/Q$), the characteristic phase slope of the resonator $\delta\phi/\delta\nu = 2Q/\nu$ implies a corresponding slow fluctuation in the frequency of the oscillator. Here f represents the fluctuation frequency, ν the microwave frequency, Q the quality factor of the resonator, and ϕ the phase of the microwave signal. In this way, a power spectral density of phase fluctuations for the amplifier $S_\phi(f)\,|_{amp}$ results in oscillator output frequency noise

$$S_y(f)\,|_{out} = (2Q)^{-2} S_\phi(f)\,|_{amp}$$

or the mathematically equivalent output phase fluctuations

$$S_\phi(f)\,|_{out} = (2Q)^{-2} \left(\frac{\nu}{f}\right)^2 S_\phi(f)\,|_{amp} \qquad (1)$$

where $y \equiv \delta\nu/\nu$ is the fractional frequency deviation.

Figure 2 shows the schematic diagram for a STALO in which the frequency variations of a noisy microwave source are cancelled by a feedback loop that detects the consequent phase shifts across a high-Q resonator to generate a frequency-correction voltage. The phase from the reference loop is adjusted to produce the proper sign of the correction voltage and attain maximum sensitivity by operating in the high slope region of the mixer output versus reference phase relation. In the limit of large loop gain, stable equilibrium requires that the phases at the two input ports of the mixer be in quadrature (mixer output = zero). A significant advantage of the STALO is that the properties of the feedback loop are particularly easy to control, since the signal is mixed down to baseband (near zero frequency). This allows the use of active filters with narrow

bandwidths and sophisticated response shapes which are not possible at microwave frequencies. A second advantage is that the $1/f$ noise for X-band mixers (-135 dB/f per Hz at 10 GHz) [15] is better than that which is available from the best amplifiers (-110 to -120 dB/f) [13, 14]. The analysis from the self-excited oscillator can be adapted to the STALO by noting that the only difference is that the phase detection and correction has been moved from the resonator in Fig. 1 to the mixer-amplifier-oscillator combination in Fig. 2. The phase noise of the RF amplifier of Fig. 1 is replaced by that of a mixer. Consequently, the output phase fluctuations are described by

$$S_\phi(f)\,|_{out} = (2Q)^{-2}\left(\frac{\nu}{f}\right)^2 S_\phi(f)\,|_{mix}$$

As a consequence, the performance of a direct-feedback oscillator with an amplifier having $1/f$ noise of -120 dB per Hz at 1 Hz ($S_\phi(f)\,|_{amp} = 10^{-12}/f$ rad^2 per Hz) is

$$S_\phi(f)\,|_{out} = 10^{-12}(2Q)^{-2}\left(\frac{\nu^2}{f^3}\right)$$

while a STALO using a mixer with $S_\phi(f)\,|_{mix} = 10^{-13.5}/f$ rad^2 per Hz will be 15 dB quieter. These device noise levels represent the quietest components presently available, giving a clear advantage to the STALO configuration.

III. Oscillator Configurations for Reduced Noise

Three new oscillator configurations for low phase noise are detailed in this section. The first two are STALO configurations that reduce the effect of mixer noise by increasing its sensitivity by the use of a suppressed-carrier signal. Of these two, the first achieves increased sensitivity by operating the high-Q resonator at higher power than would otherwise be possible. In the second, a similar effect is achieved by use of a low-level RF amplifier. The third oscillator configuration uses direct RF feedback of a sort that simultaneously achieves positive-amplitude feedback and negative-phase feedback. In this way excess gain of the amplifier is used to reduce its effective phase noise.

The STALO shown in Fig. 3 forms the basis for the new designs. It differs in implementation from Fig. 2 in that the signal from the cavity to the mixer is not taken from a second coupling port but is instead taken from the signal reflected from the input port. A circulator separates this signal from the forward-driving signal. At critical coupling and on resonance, the returned signal is identically zero. However, it is the superposition of two equal signals, one of which emanates from the cavity, and a second, reflected signal which is derived from the driving signal with a constant phase shift. This reflected signal does not significantly affect the operation of the mixer at resonance since it is in quadrature with the signal at the other mixer port. Thus the two STALOs will have approximately identical performance. Figure 4 shows the returned signal for small errors in local oscillator frequency. While the amplitude goes through zero on resonance, a phase reversal takes place in which the in-phase signal on one side becomes out of phase on the other, allowing a linear dependence of mixer output voltage on the frequency error, as required for effective feedback. Instead of viewing the mixer as a phase detector, it is seen as projecting the component of the signal at the r input onto the phase of that at the l input.

A. STALO Design for Enhanced Phase Detector Sensitivity

In this configuration, enhanced sensitivity in the phase detector is achieved by means of relatively high power in the high-Q resonator. It has the advantages of simplicity and absence of any amplifier to introduce added phase noise if carrier suppression is incomplete. A disadvantage is that power limitations in the microwave source or high-Q resonator restrict the available improvement factor.

As previously discussed and as shown in Fig. 5(a), suppression of the carrier at the r port of the mixer in Fig. 3 has only incidental consequence regarding mixer sensitivity, since the suppressed part of the signal is in quadrature with the reference signal at the l port. (The part of the signal due to frequency variations, $\pm\delta\nu$, is in phase with the reference and so is detected in any case.) However, things are not quite identical to the conventional STALO shown in Fig. 2. Suppression of the carrier at the r port allows the power to the high-Q resonator to be increased without saturating the mixer. This increased power results in an enhanced sensitivity of the mixer output voltage to frequency variations $\pm\delta\nu$.

Figure 6 describes such a circumstance. Besides the increased power levels in oscillator and resonator, the only difference from Fig. 3 is an appropriately weaker coupling to the mixer's l port. Mixers typically saturate at signal levels on the order of 20 milliwatts, while frequency sources and resonators can operate at power levels up to 1 watt or even higher. The resultant increase in sensitivity of up to 17 dB reduces the consequence of mixer noise by the same factor. For a mixer with flicker noise of -135 dB per Hz at 1 Hz offset, the effective noise could be reduced to a value of -152 dB per Hz.

B. STALO Design Using RF Amplification

Figure 7 shows a further modification of the STALO which can result in improved performance. Here the small, nominally zero signal returned from the resonator is amplified before it enters the mixer. Two effects of this addition are easy to understand. The loop gain will be increased by the added gain, an effect which may be compensated for in the design of the baseband amplifier. Secondly, the gain of the amplifier will increase the sensitivity of the mixer output to phase error in the resonator without significantly affecting mixer noise. Thus the effective mixer phase noise is reduced by the amount of amplifier gain. This can be a very substantial improvement.

The third effect of this modification is a little more complicated. Since amplifiers are somewhat more noisy than mixers (−120 dB versus −135 dB at 1 Hz offset as previously discussed), a crucial point is the proper analysis of the contribution of amplifier noise. The kind of noise under discussion is not additive noise, which would be independent of any large signal also present, but instead is multiplicative noise, which transforms a large signal by slightly modifying its amplitude and phase. (Additive noise in good amplifiers is insignificant except at offset frequencies $f > \approx 10$ kHz where the $1/f$ multiplicative noise is relatively small.) Figure 5(a) shows in phasor form the cavity signals with and without carrier suppression corresponding to the oscillators shown in Figs. 3 and 2, respectively. The added signals due to small frequency variation in the local oscillator are also shown. These added signals are detected by the mixer in order to allow feedback circuitry to cancel the frequency variations. The effect of multiplicative phase noise in the amplifier for the two cases is shown in Fig. 5(b). It is clear from Fig. 5(b) that this noise source generates signals which are indistinguishable from those caused by actual frequency variations, and which are due only to the presence of the coherent carrier. Thus a reduction in amplifier noise is effected which is proportional to the degree of carrier suppression of the microwave signal at its input.

Oscillator phase noise is thus determined by a combination of mixer noise (reduced by amplifier gain) and amplifier noise (reduced by the degree of carrier suppression). For example, if mixer noise of -135 dB/f per Hz were reduced by 25 dB of amplifier gain to -160 dB/f per Hz, and amplifier noise of -120 dB/f per Hz were reduced by 40 dB of carrier suppression to the same value, the combined noise of -157 dB/f per Hz would determine oscillator performance. For a loaded Q of 10,000 (intrinsic $Q = 20{,}000$), this would allow an oscillator phase noise of $S_\phi(f)|_{osc} = -43$ dB/f^3 per Hz, a value superior to any room-temperature microwave oscillator to date. For room temperature and thermoelectrically cooled sapphire resonators with Qs of 10^5 and 10^6, performance would be superior to that of any available source at $S_\phi(f)|_{osc} = -63$ dB/f^3 per Hz and -83 dB/f^3 per Hz, respectively.

C. Oscillator With RF Feedback

1. Noise reduction. Figure 1 shows a conventional microwave oscillator excited by direct RF feedback. As discussed earlier, any slow phase shift in the amplifier is converted by the feedback process into a frequency shift of the oscillator output as required by the condition of constant phase shift around the loop ($\phi_{loop} = 2n\pi$), with the conversion constant depending on the resonator Q. The total signal returned from the input port to the resonator is the superposition of two parts, a reflected constant signal equal in magnitude to the input signal, and an emitted part proportional to the instantaneous RF amplitude in the resonator. In the previous section, the critically coupled case was discussed, where on resonance the net returned signal was identically zero. If instead the cavity is slightly over-coupled, so that the signal emitted from the resonator is larger than the reflected signal, the net signal returned from the resonator will not be zero on resonance but will have a small, constant value. Figure 8(a) shows the configuration for an oscillator in which the small returned signal is amplified and returned to the cavity to induce oscillation. RF feedback of a phase and magnitude that allows oscillation on resonance will also induce negative phase feedback that reduces the effect of amplifier phase fluctuations on the oscillator frequency.

Figure 9(a) shows these various signals in phasor form for the case $Q \gg 1$, with weak magnetic coupling to the resonator achieved by means of an iris at the end of a waveguide or by a shorted coaxial line. Shown are the forward signal voltage amplitude $\vec{f}$, reflected signal $\vec{r}$ (directly reflected by the coupling port), emitted signal $\vec{e}$, and net returned signal $\vec{n}$, as shown for the condition of resonance (test frequency = resonator frequency.) Signals are measured at the effective plane of the weak coupling port. For frequencies significantly outside the bandwidth of the resonator, the net returned signal is approximately equal to $\vec{r}$; thus, out-of-bandwidth oscillation is prevented by the phase reversal between $\vec{n}$ and $\vec{r}$. In order to achieve oscillation at resonance, the gain and phase shift around the loop must regenerate $\vec{f}$ from $\vec{n}$. For the slightly overcoupled case shown, this corresponds to a net phase shift $2\vec{n}\pi$, and gain to overcome the signal amplitude reduction $|\ \vec{f}/\vec{r}\ |$. While small losses due to transmission through various circuit elements must also be made up by the gain element, they can be ignored for this analysis.

Intrinsic and external Qs, Q_i and Q_e, describe the effect of resonator and coupling losses and combine to form the loaded Q_l

$$Q_l^{-1} = Q_i^{-1} + Q_e^{-1}$$

which defines the operational bandwidth of the resonator. Using standard circuit analysis, the amplitude of the resonator response to the forward signal can be written:

$$\frac{|\vec{e}|}{|\vec{f}|} = \frac{2q}{q+1} \times \frac{1}{\sqrt{1+\left(2Q_l\delta\nu/\nu_o{}^2\right)}} \tag{2}$$

where $q \equiv Q_i/Q_e$ is a loading factor, ν_o is the resonance frequency, and $\delta\nu$ is the frequency deviation from resonance. The phase of the resonator response is similarly given by

$$\tan(\phi) = \frac{2Q_l\delta\nu}{\nu_o} \tag{3}$$

for a slope at resonance of

$$\frac{\delta\phi}{\delta\nu} = \frac{2Q_l}{\nu_o} \tag{4}$$

Now calculate the amplifier gain required for oscillation at resonance ($\delta\nu \approx 0$). From the nature of the coupling,

$$\vec{r} = -\vec{f} \tag{5}$$

and from the equation above

$$\vec{e} = \vec{f} \times \frac{2q}{q+1}$$

These can be combined to give

$$\vec{n} = \vec{r} + \vec{e} = \vec{f} \times \frac{q-1}{q+1}$$

requiring a gain of

$$G = \frac{|\vec{f}|}{|\vec{n}|} = \frac{q+1}{q-1} \tag{6}$$

for oscillation. The over-coupled condition depicted in Fig. 9 corresponds to $q > 1$. If $q \equiv 1 + \delta q$, the gain requirement can be rewritten as

$$G = \frac{2}{\delta q} + 1 \tag{7}$$

For the oscillator shown in Fig. 1, a small phase shift θ in the amplifier gives a corresponding phase shift $-\theta$ across the resonator, with a resultant frequency shift given by the phase slope in Eq. (4). The configuration shown in Fig. 8(a) gives a reduced resonator phase shift and thus reduced frequency shift. This reduction is now calculated. Figure 9(b) shows the self-consistent phasor diagram for the oscillator of Fig. 8(a), with a slight amplifier phase shift θ. The instantaneous frequency is determined by the smaller angle ϕ, together with the resonator phase slope $2Q_l/\nu_o$; thus, the ratio ϕ/θ describes the phase noise reduction of Fig. 7 compared to Fig. 1. The diagram is oriented so that the direction of e is constant. Graphically solving $\vec{n} = \vec{e} + \vec{r}$ shows that the effect of a rotation θ between $\vec{n}$ and $\vec{f}$ due to the amplifier gives an angle ϕ between $\vec{f}$ and $\vec{e}$. The angle ϕ in turn determines the frequency shift via Eq. (3). A straightforward evaluation for the geometry shown gives

$$\theta = \phi + \sin^{-1}\left(\frac{|\vec{r}|}{|\vec{n}|} \times \sin(\phi)\right)$$

which can be approximated in the limit of small angles to give

$$\phi/\theta \approx \frac{1}{1 + \dfrac{|\vec{r}|}{|\vec{n}|}} \tag{8}$$

This approximation will hold to high degree of accuracy because of the very small value of the phase fluctuations involved. Using Eqs. (6) and (5), Eq. (8) can be rewritten in terms of the amplifier gain as

$$\phi/\theta \approx \frac{1}{1+G}$$

The factor ϕ/θ describes the reduction in phase variation across the cavity compared to that across the amplifier. It also describes the improvement in performance due to the circuit in Fig. 8(a). Combining this result with Eq. (1), the phase-noise performance of the oscillator is

$$S_\phi(f)\,|_{out} = (\phi/\theta)^2(2Q_l)^{-2}\left(\frac{\nu}{f}\right)^2 S_\phi(f)\,|_{amp}$$

For example, for $G = 10$ (20 dB of amplification), ϕ/θ is 1/11 (22 dB of noise reduction). If the amplifier has noise

of -120 dB/f per Hz (or $S_\phi(f)\,|_{amp} = 10^{-12}/f$ rad^2 per Hz), oscillator performance will be

$$S_\phi(f)\,|_{out} = 0.8 \times 10^{-14}(2Q_l)^{-2}\left(\frac{\nu^2}{f^3}\right)$$

which for a frequency of 10 GHz and Q_l of 10^5 gives

$$S_\phi(f)\,|_{out} = 2 \times 10^{-5}/f^3$$

or -47 dB per Hz at $f = 1$ Hz offset frequency. For this same case, Eq. (7) shows that the degree of over-coupling required is

$$\delta q = \frac{2}{G-1} = \frac{2}{9}$$

or

$$\frac{Q_i}{Q_e} = 1 + \delta q = \frac{11}{9}$$

and

$$Q_l = \frac{1}{Q_e^{-1} + Q_i^{-1}} = Q_i \times \frac{1}{11/9+1} = 0.45 \times Q_i$$

2. Loop stability. The oscillation condition for the direct RF feedback configuration is such that the directly reflected signal has a net phase shift of 180 deg with respect to the signal emitted from the resonator. Thus the loop will not oscillate at frequencies outside the passband of the stabilizing resonator (where very little signal is absorbed or emitted by the resonator), unless the phase is shifted by other elements in the circuit. Unfortunately, the path length of the circuit alone is sufficient to add such a shift if the frequency is slightly varied. Thus, for a path length of 10λ, a 180-deg phase shift will occur with a 5-percent frequency change. Even worse, at these ± 5-percent frequencies, the loop gain will be nearly G, a value much greater than the loop gain at the desired mode (approximately unity). These oscillations do not involve any high-Q resonance, and can only be prevented by the introduction of a filter whose function is to reduce the gain (as the frequency is varied away from ν_o) to a value less than unity by the time the phase shift due to all sources reaches 180 deg. Because each stage of a filter introduces a phase shift of ≈ 90 deg by the time substantial attenuation is achieved, a single-stage filter must be used. The phase margin for the rest of the circuit is thus reduced to ≈ 90 deg by the filter's presence. Because a single-stage filter attenuates relatively slowly as the frequency is varied, a narrow bandwidth is required. For the example here, with a path length of 10λ and for $G = 10$ (20 dB), filter attenuation of 20 dB must be achieved at a frequency offset of 2.5 percent, where the circuit length alone introduces a 90-deg shift. This requires a filter bandwidth of ≈ 0.25 percent, a value that can be achieved with low loss using conventional techniques.

The expressions for amplitude response and phase shift for a single-stage filter are identical to those already presented in Eqs. (2) and (3). The shift due to a transmission line is given by

$$\delta\phi = 2\pi\left(\frac{p}{\lambda}\right)\left(\frac{\delta\nu}{\nu_o}\right)$$

where p is the effective path length. Figure 8(b) shows an oscillator with the added filter; Figs. 10 and 11 show the amplitude and phase response for a single-stage filter with $Q = 3000$, and the added phase due to an effective path length of 1 m. For Fig. 11, an attenuation of 35 dB is generated before the total phase shift reaches 180 deg, thus allowing $\leq 35\ dB$ of loop gain for the reduction of amplifier phase noise. It can be seen from the figures that the attenuation of this filter is nearly 35 dB by the time the total phase variation due to filter and transmission path reaches 180 deg. While 3000 is a relatively high Q for a single-mode filter, it can be attained with low loss. It is clear that usable gain and consequent usable noise reduction of 20 to 30 dB can be achieved with this technique.

IV. Conclusions

Typical low-noise 10-GHz (X-band) oscillators use a single transistor or other active semiconducting device for excitation; while a more elaborate STALO configuration is used for the lowest possible phase noise. With the development of sapphire whispering-gallery-mode resonators with Qs above 10^5 to 10^7 at 10 GHz, the possibilities have been considerably enhanced. Whereas lower frequency SAW or BAW quartz-crystal oscillators had far lower noise than their higher frequency counterparts, they are rivalled by an X-band oscillator using the sapphire resonator [9]. Together with the improved oscillator circuits developed here, such a resonator may make possible close-in phase noise lower than that of any noncryogenic frequency source. Furthermore, cooling by means of thermoelectric coolers or liquid nitrogen may make practical frequency sources with greatly reduced phase noise.

New design configurations for STALOs and direct RF oscillators allow reduced phase noise in comparison to conventional configurations. By appropriate use of carrier sup-

pression, a small signal is generated which suffers no loss of loop-phase information or signal-to-noise ratio. This small signal can be amplified without degradation by multiplicative amplifier noise, and can be detected without saturation of the detector.

Figure 12 shows phase-noise calculations for an improved sapphire whispering-gallery-mode oscillator and STALO for configurations as shown in Figs. 8 and 7. Use of a cryogenic sapphire resonator allows a further improvement of 20 to 43 dB. Quality factors are assumed to be $Q_i = 2 \times 10^5$ at room temperature, $Q_i = 2 \times 10^6$ at 170 K, and $Q_i = 3 \times 10^7$ at 77 K. Noise plots for various conventional 10-GHz frequency sources are also shown. The multiplied 5-MHz crystal oscillator presently represents the best performance available at X-band.

References

[1] W. P. Robins, *Phase Noise in Signal Sources*, IEE Telecommunications Series 9, London: Peter Peregrinus Ltd., 1984.

[2] G. K. Montress, T. E. Parker, M. J. Loboda, and J. A. Greer, "Extremely Low Phase Noise SAW Resonators and Oscillator: Design and Performance," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. UFFC-35, no. 6, pp. 657–667, 1988.

[3] R. G. Rogers, "Theory and Design of Low Phase Noise Microwave Oscillators," *Proc. 42nd Ann. Symposium on Frequency Control*, pp. 301–303, 1988.

[4] D. G. Blair and S. K. Jones, "High-Q Sapphire Loaded Superconducting Cavities and Application to Ultrastable Clocks," *IEEE Trans. Magnetics*, vol. MAG-21, pp. 142–145, March 1985.

[5] A. Giles, S. Jones, and D. Blair, "A High Stability Microwave Oscillator Based on a Sapphire Loaded Superconducting Cavity," *Proc. 43rd Ann. Symposium on Frequency Control*, pp. 89–93, 1989.

[6] V. B. Braginsky, V. P. Mitrofanov, and V. I. Panov, *Systems With Small Dissipation*, Chicago: University of Chicago Press, pp. 85–89, 1985.

[7] V. I. Panov and P. R. Stankov, "Frequency Stabilization of Oscillators With High-Q Leucosapphire Dielectric Resonators," *Radiotekhnika i Electronika*, vol. 31, no. 213, 1986 (in Russian).

[8] G. J. Dick and D. M. Strayer, "Measurements and Analysis of Cryogenic Sapphire Dielectric Resonators and DRO's," *Proc. 41st Ann. Symposium on Frequency Control*, pp. 487–491, 1987.

[9] J. Dick and J. Saunders, "Measurement and Analysis of a Microwave Oscillator Stabilized by a Sapphire Dielectric Ring Resonator for Ultra-Low Noise," *Proc. 43rd Ann. Symposium on Frequency Control*, pp. 107–114, 1989.

[10] X. H. Jiao, P. Guillon, and L. A. Bermudez, "Resonant Frequencies of Whispering-Gallery Dielectric Resonator Modes," *IEE Proceedings*, vol. 134, pt. H, pp. 497–501, 1987.

[11] M. Gastine, L. Courtois, and J. L. Dormann, "Electromagnetic Resonances of Free Dielectric Spheres," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-15, pp. 694–700, December 1967.

[12] M. M. Driscoll, "Low Noise Signal Generation Using Bulk- and Surface-Acoustic-Wave Resonators," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. UFFC-35, no. 3, pp. 426–434, 1988.

[13] T. E. Parker, "Characteristics and Sources of Phase Noise in Stable Oscillators," *Proc. 41st Ann. Symposium on Frequency Control*, pp. 99–110, 1987.

[14] C. P. Lusher and W. N. Hardy, "Effects of Gain Compression, Bias Conditions, and Temperature on the Flicker Phase Noise of an 8.5 GHz GaAs MESFET Amplifier," *IEEE Trans. on Microwave Theory and Techniques*, vol. 37, pp. 643–646, April 1989.

[15] F. L. Walls, A. J. D. Clements, C. M. Felton, M. A. Lombardi, and M. D. Vanek, "Extending the Range and Accuracy of Phase Noise Measurements," *Proc. 42nd Ann. Symposium on Frequency Control*, pp. 432–441, 1988.

[16] J. A. Barnes, R. Chi, L. S. Cutler, D. J. Healey, D. B. Leeson, T. A. McGunigal, J. A. Mullen, Jr., W. L. Smith, R. L. Sydnor, R. F. C. Vessot, and G. M. R. Winkler, "Characterization of Frequency Stability," *IEEE Trans. Instr. and Meas.*, vol. IM-20, pp. 105–120, May 1971.

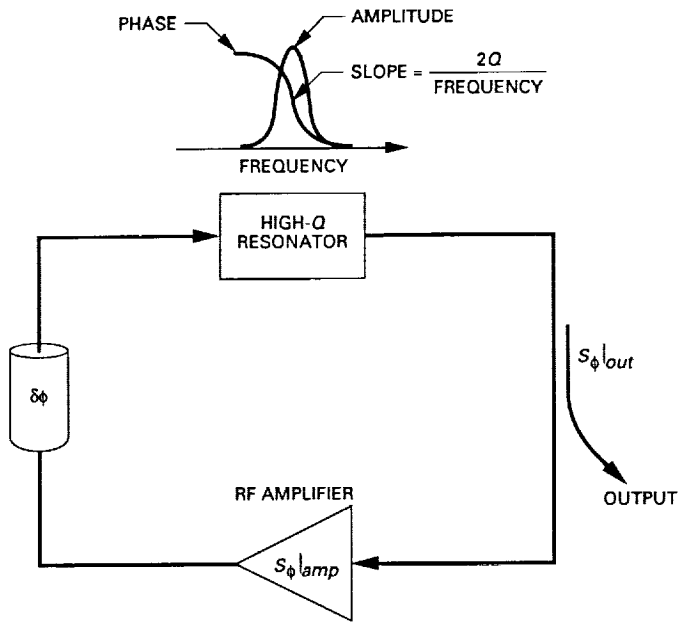


Fig. 1. Block diagram of simple oscillator with direct RF feedback. Output phase noise is derived from amplifier noise together with phase slope of resonator; phase is adjusted to give 2 $n\pi$ radians around loop at the center of the sapphire resonator passband.

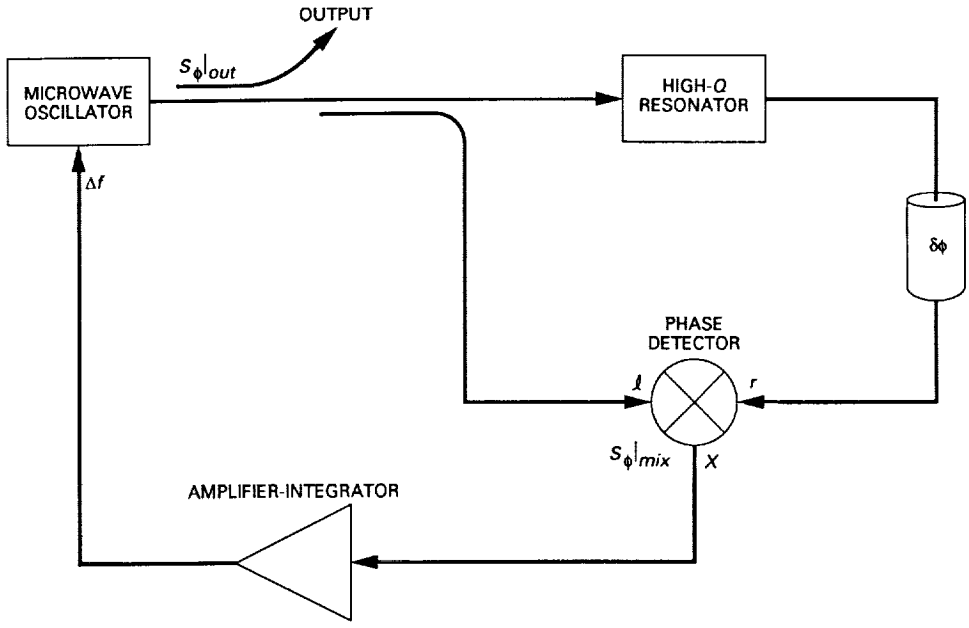


Fig. 2. Block diagram of stabilized local oscillator (STALO) with double-balanced mixer type phase detector. Mixer noise plays the same role as amplifier noise in Fig. 1; phase is adjusted to give *l* and *r* signals in quadrature.

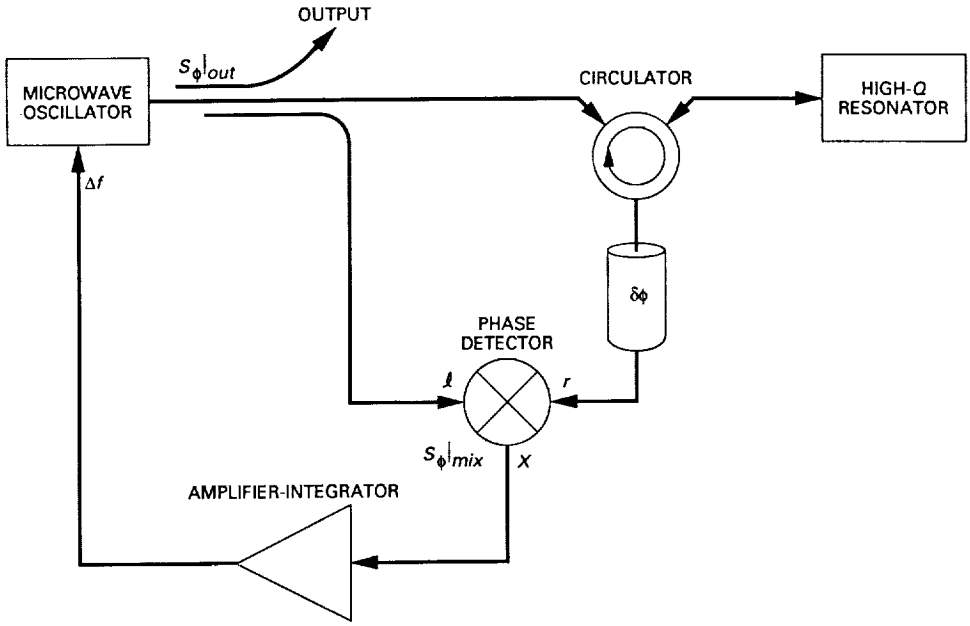


Fig. 3. STALO with different configuration but functionally identical to that of Fig. 2. Signal returned from resonator is superposition of the resonator signal and the constant reflected signal.

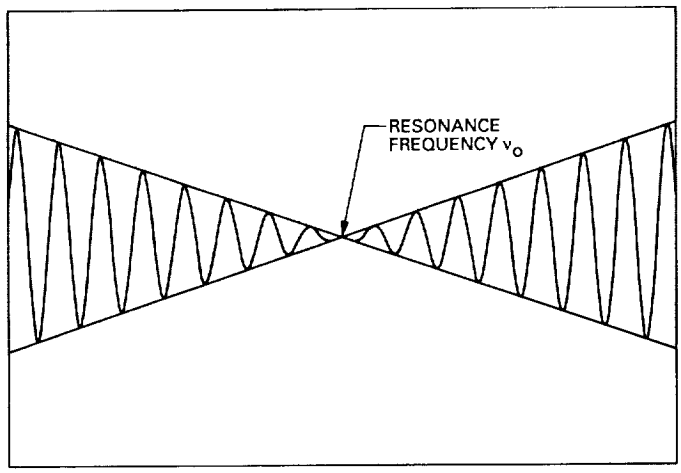


Fig. 4. RF envelope of returned signal for critical coupling as frequency ν is varied. Phase inversion at center allows linear dependence in mixer output with frequency for arbitrarily small frequency errors.

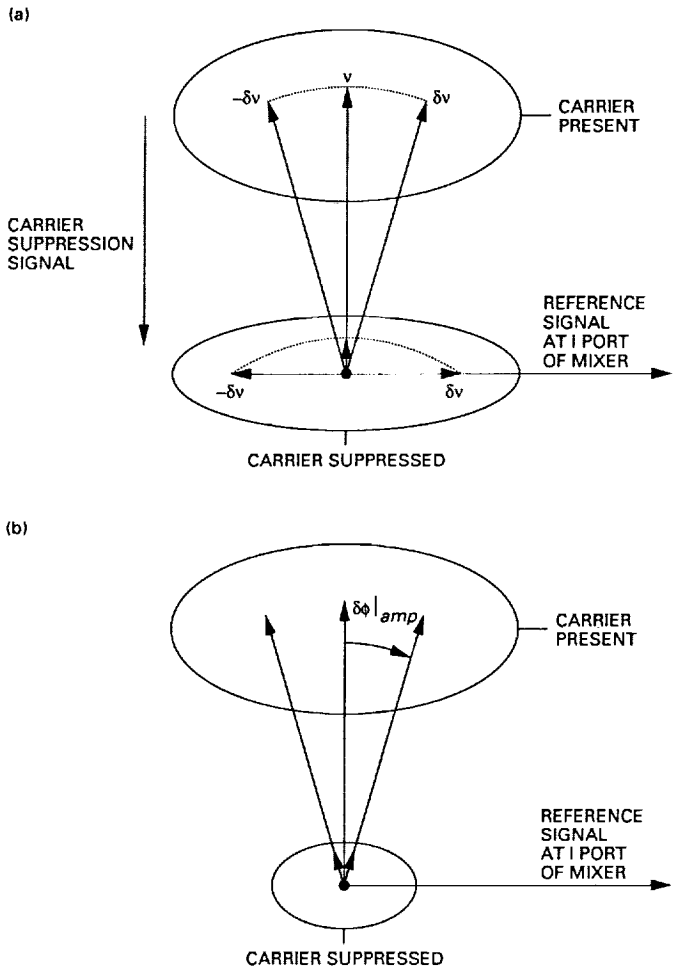


Fig. 5. Phasor diagram showing the effect, with and without carrier suppression, of (a) frequency error and (b) amplifier phase noise on the RF resonator signal. Also shown are the constant carrier suppression signal and mixer reference signal. Note that the effect of frequency error on the component of the RF resonator signal in phase with the reference is unchanged by 20 dB carrier suppression, while the effect of amplifier phase noise is reduced by 20 dB.

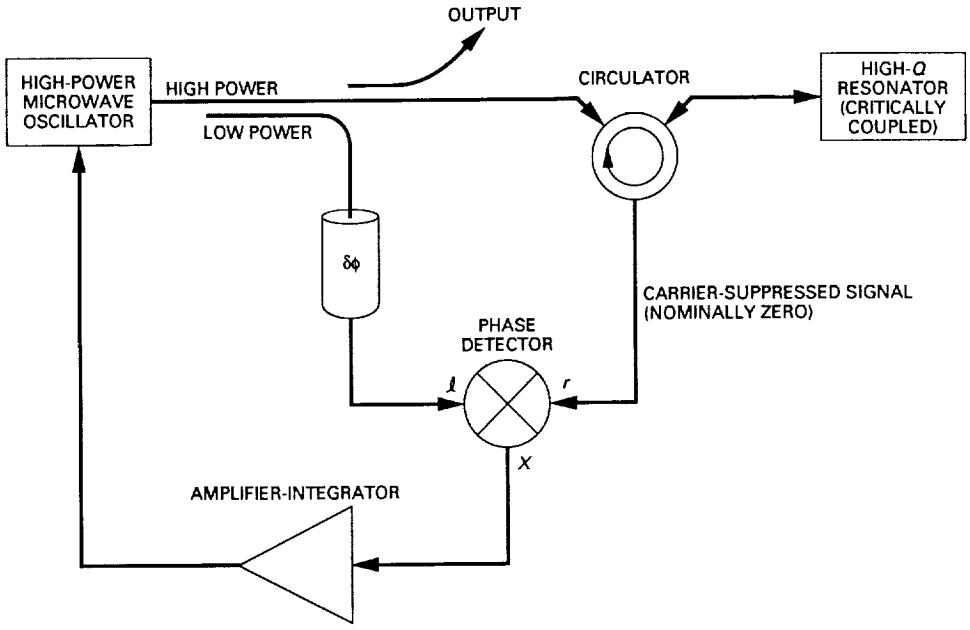


Fig. 6. Configuration to allow increased power into the high-Q resonator without saturating the mixer; the resultant increase in sensitivity of up to 17 dB reduces the consequence of mixer noise by the same factor.

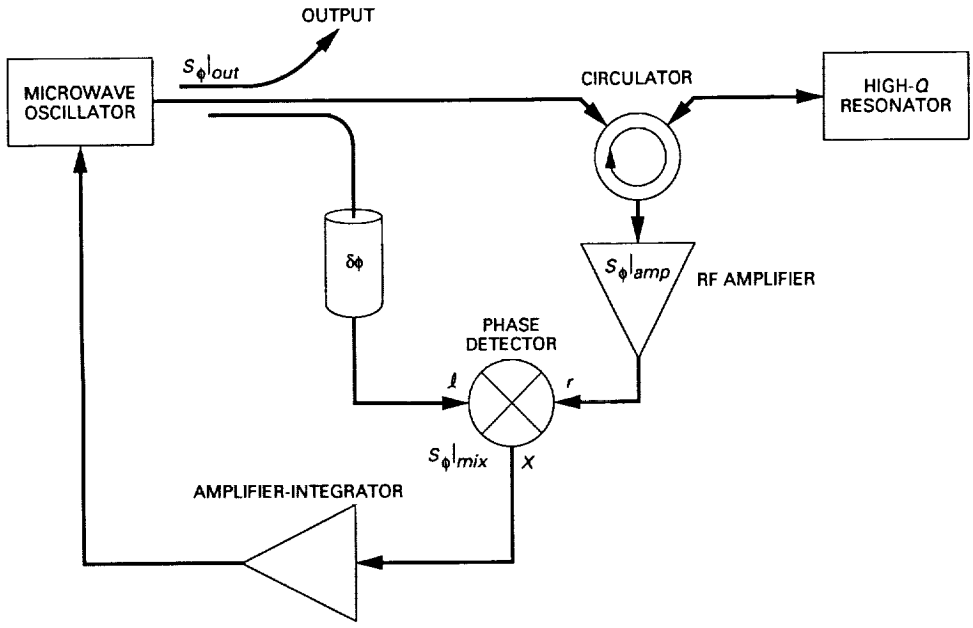
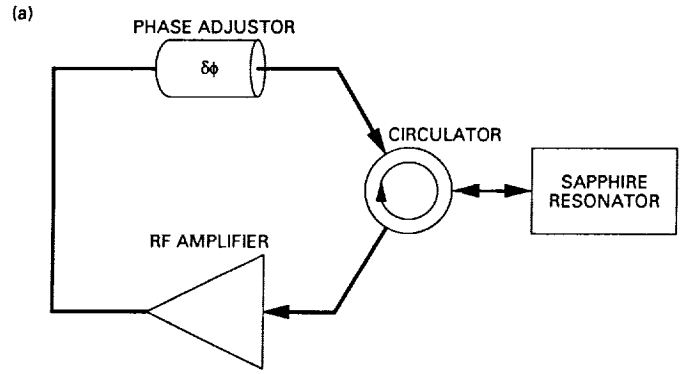


Fig. 7. STALO configuration for reduced phase noise. Critical coupling to resonator and operation very near ν_o allow insensitivity to phase noise of amplifier; amplifier gain allows reduction of phase detector noise.



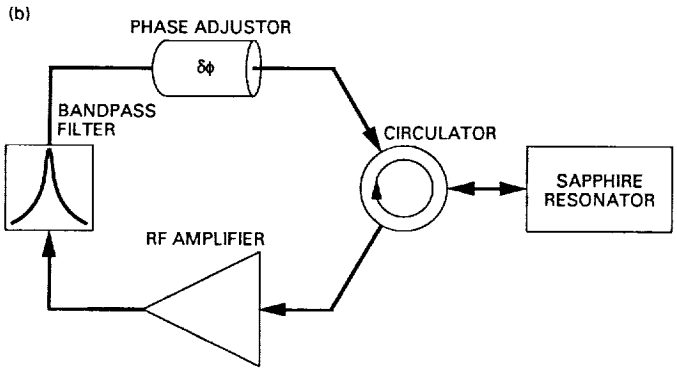


Fig. 8. Configuration of reflection oscillator: (a) with direct RF feedback, (b) added filter prevents spurious oscillation.

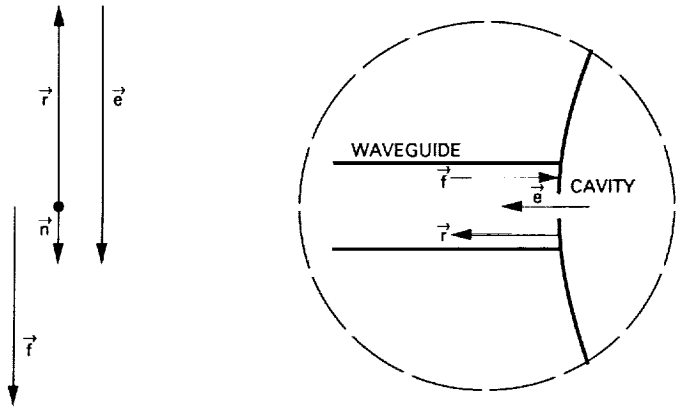


Fig. 9(a). Phasor diagram showing forward, reflected, emitted, and net signal amplitudes $\vec{f}$, $\vec{r}$, $\vec{e}$, and $\vec{n}$, for a slightly over-coupled resonator at resonance.

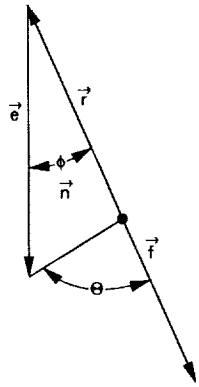


Fig. 9(b). Self-consistent phasor diagram showing the effect of amplifier phase shift θ on operation of the oscillator shown in Fig. 7. Feedback gain is the same as in Fig. 9(a).

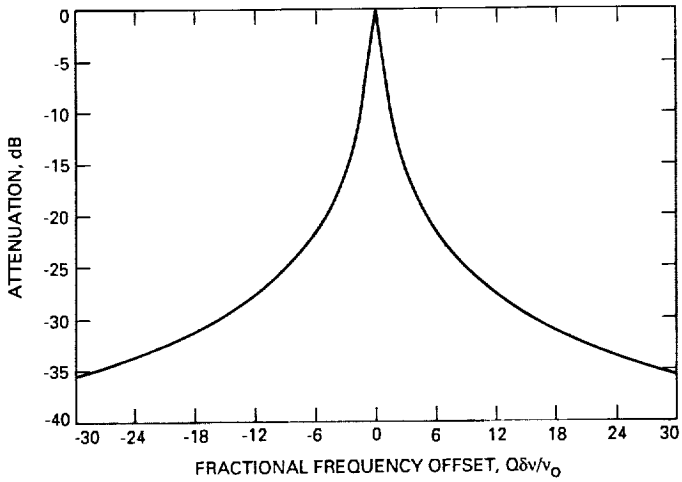


Fig. 10. Amplitude response of the single-stage filter in Fig. 8(b). More stages cannot be used because the added phase shift would allow oscillation within the filter bandwidth.

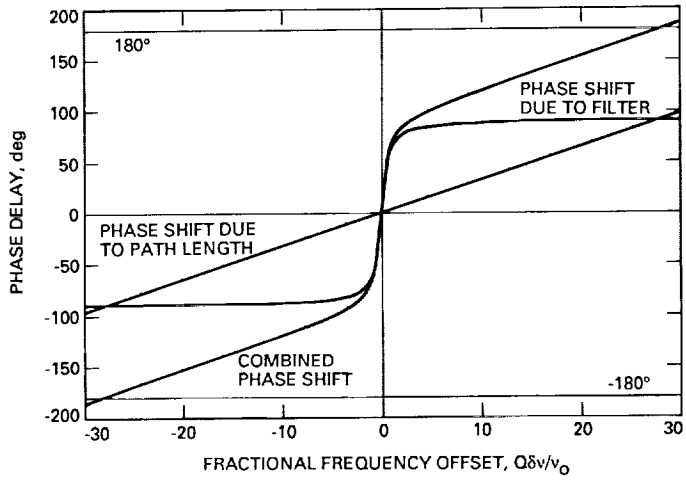


Fig. 11. Phase response of the single-stage filter together with phase response due to path lengths, for path length of 1 m, Q = 3000, and ν_o = 8 GHz. Path-length induced phase shift would be greater for lower *Q* or a longer path length.

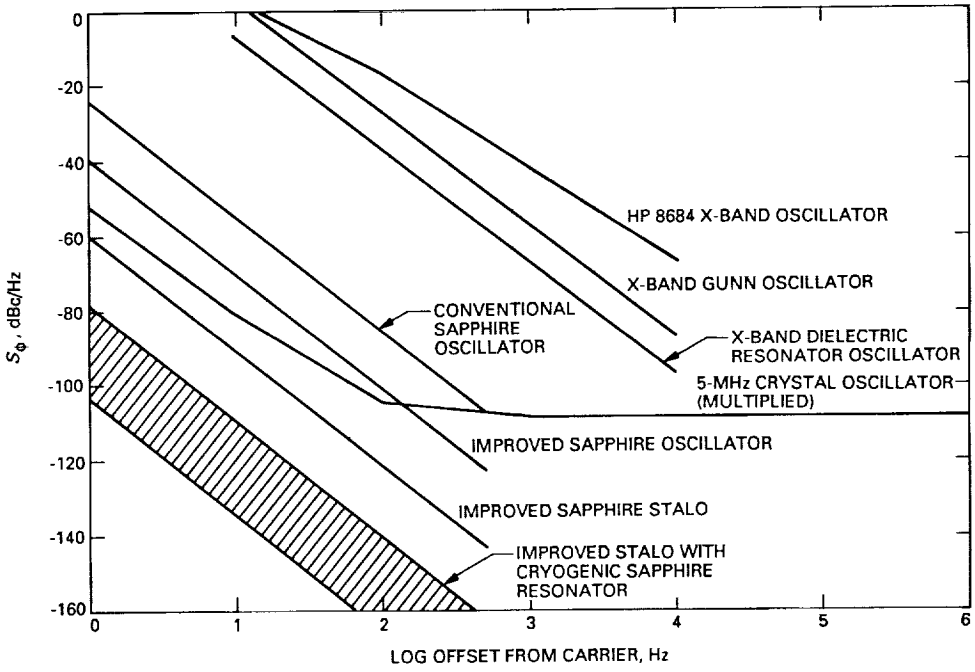


Fig. 12. Phase-noise calculations for improved sapphire whispering-gallery-mode oscillator and STALO shown in Figs. 8 and 7. Use of a cryogenic (170 K to 77 K) sapphire resonator allows further improvement by 20 to 43 dB.

TDA Progress Report 42-99

November 15, 1989

Effect of Laser Frequency Noise on Fiber-Optic Frequency Reference Distribution

R. T. Logan, Jr., G. F. Lutes, and L. Maleki
Communications Systems Research Section

This article presents an analysis of the effect of the linewidth of a single-longitudinal-mode laser on the frequency stability of a frequency reference transmitted over single-mode optical fiber. The interaction of the random laser frequency deviations with the dispersion of the optical fiber is considered to determine theoretically the effect on the Allan deviation (square root of the Allan variance) of the transmitted frequency reference. It is shown that the magnitude of this effect may determine the limit of the ultimate stability possible for frequency reference transmission on optical fiber, but is not a serious limitation to present system performance.

I. Introduction

Ultrastable fiber optic transmission of hydrogen maser reference signals is presently operational at the Goldstone facility of the NASA/JPL Deep Space Network [1]. This capability supports radio science experiments such as Connected Element Interferometry (CEI) by enabling phase-coherent arraying of widely separated antennas in real time. Also, distribution of a centralized maser reference throughout the entire complex eliminates the need for a hydrogen maser frequency standard at each Deep Space Station, with substantial cost savings and increased reliability.

A reference signal produced by a hydrogen maser frequency standard is presently distributed over distances up to 29 km, with differential fractional frequency stability $\frac{\delta f}{f} \approx 10^{-15}$ for 1000-second averaging times. Although the present fiber optic distribution capability is as stable as the hydrogen maser frequency standard, the ideal distribution system should be an order of magnitude more stable than the distributed signal. With the promise of trapped-ion frequency standards [2] and superconducting cavity masers [3] that will both provide more stable frequency references, fiber optic link stability of 10^{-18} at 1000 seconds will be required for stable distribution.

The demanding requirements that a frequency reference distribution system must meet necessitate the examination of all sources of instability at levels far beyond the needs of typical analog and digital fiber optic communication systems. Presently, laser source amplitude noise and thermal variations of the optical fiber have been identified as the limiting factors to distribution system performance. Improved lasers with lower intensity noise and single-longitudinal-mode operation will be employed in the near future. A thorough examination of fiber optic sys-

tems and components has indicated that laser frequency deviations may limit system performance as lower amplitude noise lasers become available. However, a quantitative analysis of the effect of laser frequency noise on a narrow-band frequency distribution system has not previously been performed.

The present analysis theoretically estimates the effects of laser frequency fluctuations on the amplitude and phase stability of a frequency reference transmitted on single-mode optical fiber. An expression for the phase noise spectral density of the modulation signal due to the frequency noise spectral density of the laser is derived and then used to calculate the expected Allan deviation of the transmitted reference signal. The laser-induced phase noise is shown to depend on the modulation signal frequency, fiber length, and fiber dispersion, as well as the magnitude of the laser frequency fluctuations. It is shown that the differential frequency stability of a single-mode fiber optic link is fundamentally limited by laser frequency noise. Thus, as laser intensity noise is reduced, the laser frequency noise will limit transmission stability.

II. Fiber-Laser Interaction

The dispersion of optical fiber causes various optical frequencies to travel with different velocities. Optical carrier frequency deviations couple with the dispersion of the fiber to produce random phase deviations in the envelope of a modulation signal, thereby degrading its phase stability. Optical fiber acts as a frequency discriminator to translate random frequency deviations of the laser into random phase deviations of the RF modulation envelope. Although every effort is made to operate the laser at the minimum dispersion point of the fiber, the slope of the fiber index of refraction versus wavelength is typically not zero. As the laser frequency deviates, the signal experiences changes in the fiber index of refraction that cause phase shifts of the modulation envelope.

Under bias current modulation, a semiconductor laser exhibits changes in wavelength, or chirp, that are synchronous with the bias modulation due to the change in refractive index of the laser gain medium [4]. Lasers also exhibit random frequency fluctuations due to the quantum phenomenon of electron-hole recombination in the gain media, with an attendant change of refractive index [5]. Temperature excursions of the laser diode also affect the index of refraction and the lasing wavelength, but with time constants from minutes to hours.

In digital transmission systems, laser chirp is the predominant limit on transmission distance [6, 7]. These wide-band systems are sensitive to phase deviations of the modulation envelope at all frequencies. However, in such systems, spontaneous emission noise is ignored since chirp is the overwhelming effect [7].

In contrast to digital or wide-band analog transmission systems, frequency distribution systems employ a narrow-band loop filter at the output of the fiber receiver. Therefore, high-frequency deviations of the modulation envelope phase are averaged out, leaving only the laser noise within the loop bandwidth. An analysis of the effect of close-to-carrier laser frequency noise on a long-distance frequency reference transmission system has not been published (to our knowledge), so the effects of laser frequency noise have not been known. Also, at the levels of signal to noise ratio (SNR) and frequency stability of the frequency distribution systems under consideration, it has been unclear what role laser frequency noise plays in determining the ultimate system stability attainable. The present analysis provides an estimate of the role of intrinsic laser frequency noise in a narrow-band frequency distribution system to determine the level at which system performance might become limited.

Since high-frequency laser chirp can be neglected in a narrow-band system, the present analysis considers only the intrinsic laser frequency noise within the bandwidth of the output filter. As such, the analysis applies to any type of laser system, although semiconductor lasers are typically used. Externally modulated Nd:YAG lasers at 1318 nm may be an attractive alternative to semiconductor lasers for long-haul analog signal transmission. The present analysis applies equally well to these types of lasers by substitution of the appropriate parameters.

III. Analysis

Intrinsic laser frequency noise has its origins in the discrete random photons spontaneously emitted into the lasing mode that cause random frequency changes of the laser wavelength [8–10]. The high-frequency character of this noise is well-known. It is basically flat within the modulation bandwidth, peaking at the relaxation oscillation resonance of the laser diode cavity, usually in the tens-of-gigahertz region [8, 9]. Within tens of kHz of the carrier, the frequency noise exhibits a $1/f$ character [10]. Close-to-carrier measurements of laser noise are limited to within about 10 kIIz, due to the physical difficulty of fabricating frequency discriminators with sufficient resolution at optical frequencies. It is this low-frequency FM noise that is of interest for the analysis of narrow-band frequency distribution systems.

We desire an expression for the phase noise spectral density of the modulation signal as a function of the spectral density of laser frequency fluctuations. The resultant phase noise density may then be used to calculate the expected Allan deviation of the reference signal, provided the character of the laser frequency noise is known.

Consider a single-longitudinal-mode laser coupled to a single-mode fiber. The laser output is amplitude modulated at RF frequency Ω. The phase delay for the modulation signal envelope along the fiber is given by

$$\phi = \frac{2\pi n L \Omega}{c} \quad \text{(rad)} \tag{1}$$

where n is the fiber index of refraction, L is the fiber length, and c is the speed of light in a vacuum. It is assumed that the laser operates in a single longitudinal mode at $\lambda = 1.3$ μm. Now, consider the effect of a perturbation, such as a change in ambient temperature, on the refractive index of the fiber. This causes a phase change

$$d\phi = \frac{dn\, 2\pi L \Omega}{c} \quad \text{(rad)} \tag{2}$$

Multiplying the numerator and denominator on the right-hand side by $d\lambda$ gives

$$d\phi = \frac{2\pi L \Omega\, d\lambda}{c}\left(\frac{dn}{d\lambda}\right) \quad \text{(rad)} \tag{3}$$

By writing $d\lambda$ in terms of the laser frequency ν, the phase deviations may be expressed in terms of the laser frequency deviations, which have the same (random) time dependence. Thus

$$d\phi(t) = d\nu(t)\frac{2\pi L \Omega \lambda^2}{c^2}\left(-\frac{dn}{d\lambda}\right) \quad \text{(rad/sec)} \tag{4}$$

For the analysis of frequency stability, it is more convenient to look at the last expression in the frequency domain by Fourier transforming as follows:

$$S_\phi(f) = S_\nu(f)\left[\frac{-2\pi L \Omega \lambda^2}{c^2}\frac{dn}{d\lambda}\right]^2 \quad \text{(rad}^2\text{/Hz)} \tag{5}$$

In this expression, $S_\phi(f)$ is the spectral density of the phase fluctuations at an offset frequency f from the RF signal; the fluctuations are induced by the spectrum of random frequency deviations, $S_\nu(f)$, of the laser.

The variation of the effective fiber index of refraction with wavelength $\frac{dn}{d\lambda}$ depends on the waveguide parameters and material composition of the fiber. The measured result for typical single-mode fiber at 1300 nm is [11]

$$\frac{dn}{d\lambda} = 270.1 \quad \text{m}^{-1} \tag{6}$$

Inserting this value into Eq. (5) and substituting the appropriate constants produces the simple relation

$$S_\phi(f) = S_\nu(f)\, L^2 \Omega^2 (1.02 \times 10^{-51}) \quad \text{(rad}^2\text{/Hz)} \tag{7}$$

The above quantity is the mean-square phase-noise spectral density at an offset f from the modulation signal, Ω. This is the spectrum which would be observed if a perfect oscillator (i.e., an oscillator with no phase noise) modulated the laser and if the output of the photodetector were compared to a second perfect oscillator, as in a phase noise measurement system.

IV. Numerical Estimates

The FM noise spectrum of distributed feedback-type (DFB) single-mode lasers typically used in fiber optic distribution systems exhibits a power-independent $1/f$ character at low frequencies (below about 1 MHz). In the modulation band, the FM noise is white and inversely proportional to optical power [10]. The physical mechanism responsible for the $1/f$ behavior is thought to be the trapping of carriers due to impurities and interfacial boundaries, but it is not fully understood. The white portion of the spectrum is due to spontaneous-emission events and is adequately modeled by theory [5, 8].

The frequency noise spectrum of typical DFB lasers has been measured experimentally [10]. The above-mentioned physical mechanisms may be modeled as

$$S_\nu(f) = \frac{C}{P} + \frac{K}{f} \tag{8}$$

where P is the average output power of the laser and f is the frequency offset from the optical carrier. C and K are empirically determined constants. For the Fujitsu DFB laser diodes measured [10], $C = 1.5 \times 10^4$ (Hz · W), and $K = 5.8 \times 10^{11}$ (Hz^2).

The frequency noise spectrum of a typical DFB laser is illustrated in Fig. 1. As laser power is increased, the white portion of the noise spectrum decreases proportional to P^{-1}. A numerical estimate of the additive RF phase noise

requires that only the $1/f$ portion of $S_\nu(f)$ be inserted into Eq. (7), which gives

$$S_\phi(f) = \frac{5.8 \times 10^{11}}{f} \Omega^2 L^2 (1.05 \times 10^{-51}) \ (\text{rad}^2/\text{Hz})$$

$$= L^2 \Omega^2 \frac{5.9 \times 10^{-40}}{f} \ (\text{rad}^2/\text{Hz}) \qquad (9)$$

The $1/f$ laser frequency noise is converted to $1/f$, or "flicker," phase noise in the fiber optic distribution system. This level of $1/f$ phase noise depends on the inherent quantum fluctuations of the laser frequency and represents the ultimate phase noise floor of the system.

For flicker phase noise, the Allan deviation (square root of the Allan variance) may be calculated from the following relation [12]:

$$\sigma_y(\tau) = \sqrt{\frac{3}{(2\pi)^2 \tau^2} \frac{S_\phi(f) f}{\Omega^2} \ln(8.88 f_h \tau)} \qquad (10)$$

where f_h is the frequency cutoff of the phase noise. In this case, f_h is one-half the bandwidth of the output filter.

Substituting Eq. (9) into the last expression, the modulation frequency cancels, and the expression for the Allan deviation reduces to

$$\sigma_y(\tau) = \sqrt{\frac{3}{(2\pi)^2 \tau^2} L^2 \, 6 \times 10^{-40} \ln(8.88 f_h \tau)} \qquad (11)$$

The laser-induced flicker phase noise thus sets the minimum bias level of the $1/\tau$ section of the Allan deviation plot. The fact that the last expression does not depend on the RF modulation frequency, Ω, is significant. This implies that moving to higher or lower modulation frequencies for reference signal distribution will not alter the laser frequency noise "floor" of the Allan deviation.

For purposes of comparison, we consider an actual frequency distribution link in use at the NASA/JPL Goldstone Deep Space Communications Complex. The longest frequency distribution run is 29 km. Assuming that the output filter bandwidth, f_h, is 10 Hz, the Allan deviation, calculated from Eq. (11), is

$$\sigma_y(\tau) \simeq \frac{4.1 \times 10^{-16}}{\tau} \qquad (12)$$

The relation between the laser-frequency noise-limited Allan deviation of the 29-km link and the Allan deviation of a typical hydrogen maser is plotted in Fig. 2. This represents the ultimate frequency stability attainable with such a link, provided all other noise sources are negligible.

V. Present State of the Art

The ultimate link stability plotted in Fig. 2 will only be attained if all other noise sources are negligible. In reality, other noise sources do contribute to the link stability. This is illustrated in Fig. 3, where an actual measurement of the 29-km Goldstone link is plotted in addition to the maser and the ultimate-stability-link curves of Fig. 2.

In present-day systems, the Allan deviation $1/\tau$ intercept is set by the SNR of the fiber link, which is determined by the laser intensity noise. The SNR of a typical high-performance analog link is 120 dB/Hz. Figure 4 illustrates the 1-sec Allan deviation as a function of fiber length. It is immediately apparent from Fig. 4 that the laser frequency noise does not limit frequency distribution system performance at this time, since the laser SNR dominates. As lower amplitude-noise lasers become available, the laser frequency noise floor of the Allan deviation may begin to limit frequency reference distribution system performance.

Figure 5 depicts fiber link Allan deviation at 1 second as a function of link SNR. The laser relative intensity noise (RIN) sets the SNR of the fiber link for short distances [13]. Also shown is the Allan deviation floor due to laser frequency noise for a 29-km link. This plot shows clearly that laser frequency noise limits the frequency stability "floor" of the fiber link to $4 \times 10^{-16}/\tau$ for SNR $\geq$ 145 dB/Hz. Systems with as high as 150 dB/Hz SNR may be achievable with externally modulated high-power semiconductor-diode-pumped Nd:YAG solid-state lasers, or through the use of squeezed light generated directly from semiconductor lasers. As these system improvements are realized, the low-frequency $1/f$ FM noise of the laser may begin to limit system performance.

A final observation: Since the fiber optic transmission system converts laser frequency noise to RF phase noise, it may be the case that a stabilized fiber optic link comprises a very accurate system for measuring the frequency deviations of lasers close to the optical carrier. This approach is under consideration for future research.

VI. Conclusion

At present, the noise floor of fiber optic distribution systems is determined by the laser signal to noise ratio

(SNR) in the RF modulation band. However, lasers with lower relative intensity noise (RIN) or those which use squeezed light promise increases in link SNR, and passive and active temperature-stabilization schemes can improve link stability at long averaging times. As these improvements in components and systems are realized, the fundamental limit for frequency stability due to laser frequency noise may be reached.

The present analysis provides the contribution to the phase noise of a transmitted frequency reference due to single-mode laser frequency deviations. Through interaction with the dispersion of the fiber, the $1/f$ FM noise close to the optical carrier is converted to $1/f$ phase noise close to the RF reference signal. The additive $1/f$ laser-induced phase noise is a function of the fiber dispersion and length and determines the ultimate Allan deviation floor of the fiber optic distribution system in the 1- to 100-second region.

For the longest fiber optic frequency distribution link in the NASA/JPL Deep Space Network (29 km), using data for commercially available DFB lasers, the analysis indicates that the link Allan deviation is limited to $4 \times 10^{-16}/\tau$ (for averaging times between 1 second and 100 seconds). This stability limit will be reached at link SNR of 145 db/Hz, which is 25 dB better than the present system.

Further increases in SNR will not yield higher link stability unless laser frequency noise is decreased as well. The laser FM noise stability limit is two orders of magnitude higher stability than the best current frequency standard, which indicates that laser frequency noise will not limit fiber optic frequency distribution capability in the foreseeable future.

Acknowledgments

The authors thank G. J. Dick, R. Sydnor, C. Greenhall, and L. Primas for numerous helpful discussions.

References

[1] L. E. Primas, G. Lutes, Jr., and R. Sydnor, "Fiber Optic Frequency Transfer Link," *Proceedings of the 42nd Annual Symposium on Frequency Control*, Baltimore, Maryland, pp. 478–484, June 1–3, 1988.

[2] J. Prestage, G. Janik, G. Dick, and L. Maleki, "New Ion Trap for Frequency Standard Applications," *Proceedings of the 43rd Annual Symposium on Frequency Control*, Denver, Colorado, pp. 135–142, May 31–June 2, 1989.

[3] G. J. Dick, R. T. Wang, and D. M. Strayer, "Operating Parameters for the Superconducting Cavity Maser," *Proceedings of the 20th Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting*, Vienna, Virginia, pp. 345–354, November 29–December 1, 1988.

[4] L. Chinlon, G. Eisenstein et al., "Fine structure of frequency chirping and FM sideband generation in single-longitudinal-mode semiconductor lasers under 10-GHz direct intensity modulation," *Applied Physics Letters*, vol. 46, no. 1, pp. 12–14, January 1, 1985.

[5] K. Vahala and A. Yariv, "Semiclassical Theory of Noise in Semiconductor Lasers: Part I, Part II," *IEEE Journal of Quantum Electronics*, vol. QE-19, no. 6, pp. 1096-1109, June 1983.

[6] J. C. Cartledge and G. S. Burley, "The Effect of Laser Chirping on Lightwave System Performance," *IEEE Journal of Lightwave Technology*, vol. 7, no. 3, pp. 568–573, March 1989.

[7] D. C. Atlas, A. F. Elrefaie, M. B. Romeiser, and D. G. Daut, "Chromatic dispersion limitations due to semiconductor laser chirping in conventional and dispersion-shifted single-mode fiber systems," *Optics Letters*, vol. 13, no. 11, pp. 1035–1037, November 1988.

[8] B. Daino, P. Spano, M. Tamburrini, and S. Piazzolla, "Phase Noise and Spectral Lineshape in Semiconductor Lasers," *IEEE Journal of Quantum Electronics*, vol. QE-19, no. 3, pp. 266–270, March 1983.

[9] R. Schimpe and W. Harth, "Theory of FM Noise in Single-Mode Injection Lasers," *Electronics Letters*, vol. 19, no. 4, pp. 136–137, February 17, 1983.

[10] K. Kikuchi, "Effect of $1/f$-type FM noise on semiconductor laser linewidth residual in high-power limit," *IEEE Journal of Quantum Electronics*, vol. 25, no. 4, pp. 648–688, April 1989.

[11] Corning Telecommunications Products Data Sheet, "SMF-21 CPC3 Single-Mode Optical Fiber," Corning, New York, February 1987.

[12] "Characterization of Frequency Stability," NBS Technical Note 394, U.S. Department of Commerce, National Bureau of Standards, Washington, D.C., October 1970.

[13] K. Y. Lau, "Signal-to-Noise Calculation for Fiber Optics Links," *TDA Progress Report 42-58*, vol. May–June 1980, Jet Propulsion Laboratory, Pasadena, California, pp. 41–48, August 15, 1980.

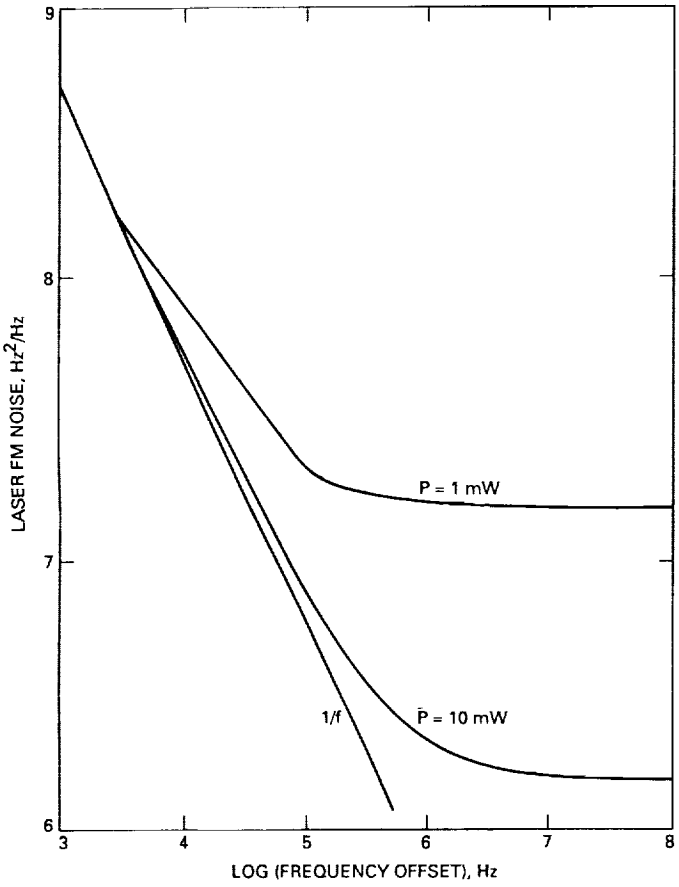


Fig. 1. Typical DFB laser frequency noise spectrum.

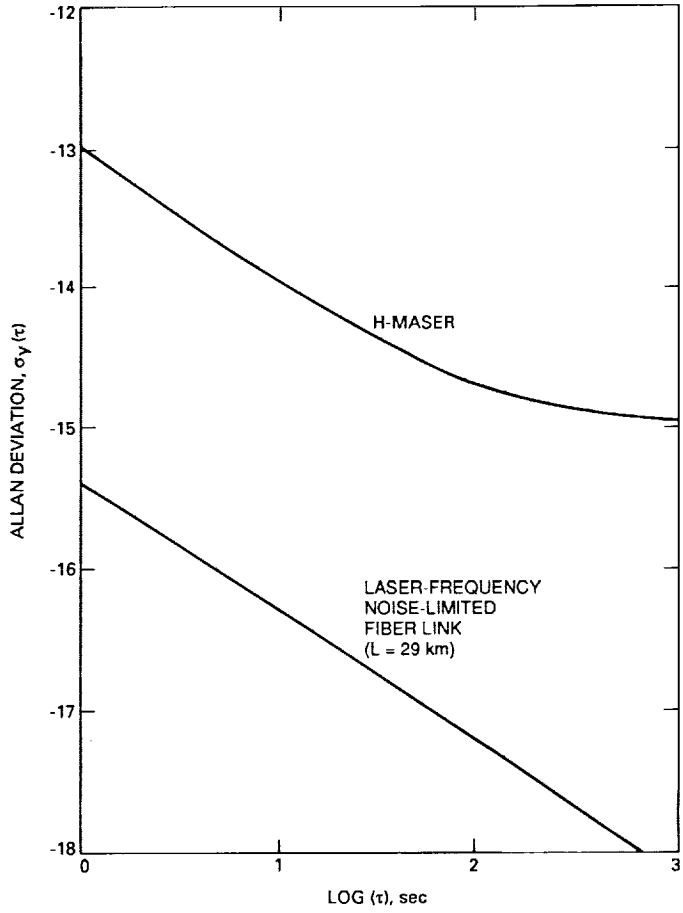


Fig. 2. Hydrogen maser and laser FM noise-limited fiber optic link frequency-stability comparison.

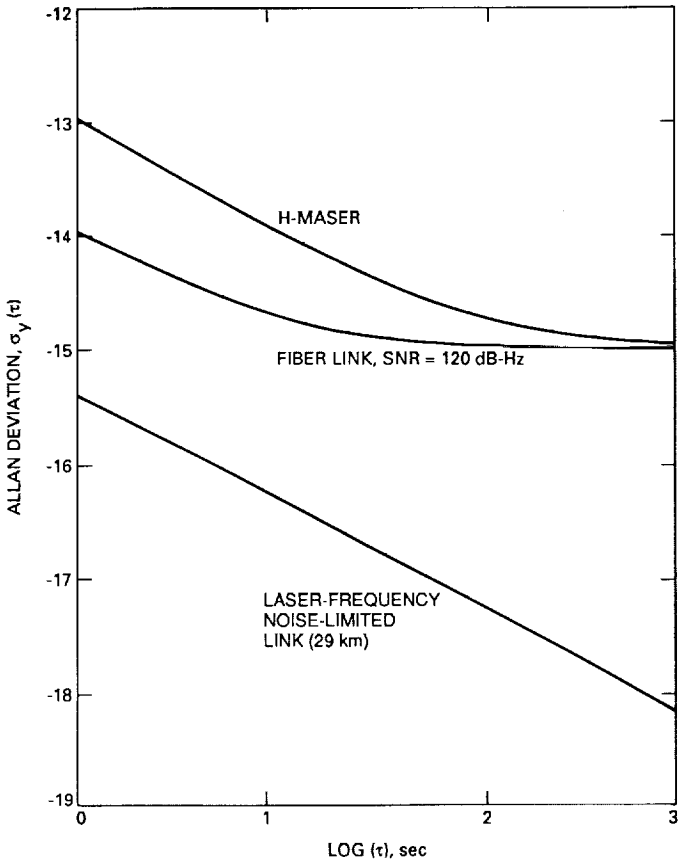


Fig. 3. Comparison of Allan deviation: H-maser, actual 29-km link, and theoretical FM noise-limited link.

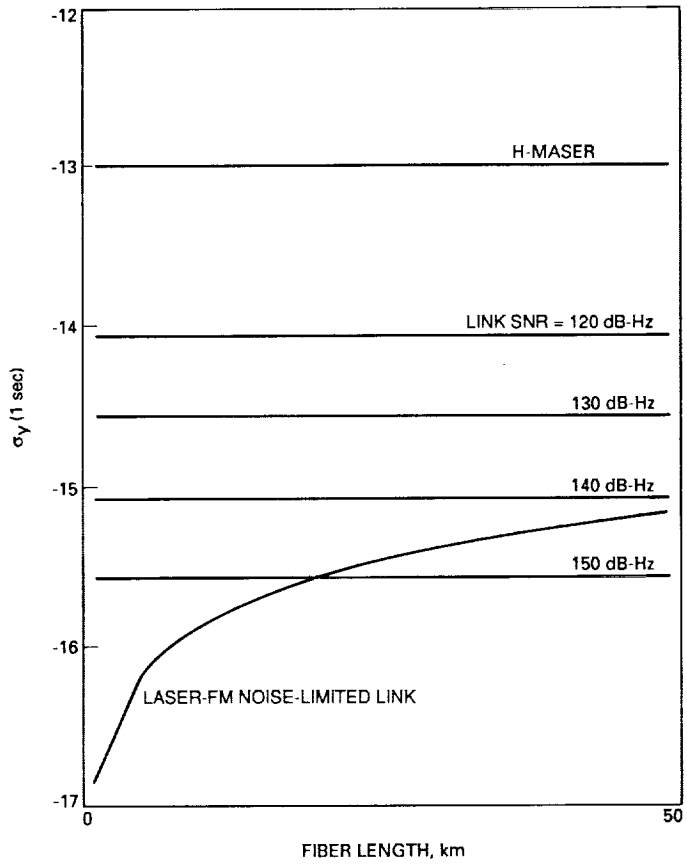


Fig. 4. Comparison of frequency stability at one second versus fiber length for maser, fiber optic links of various SNR, and noise-limited laser FM.

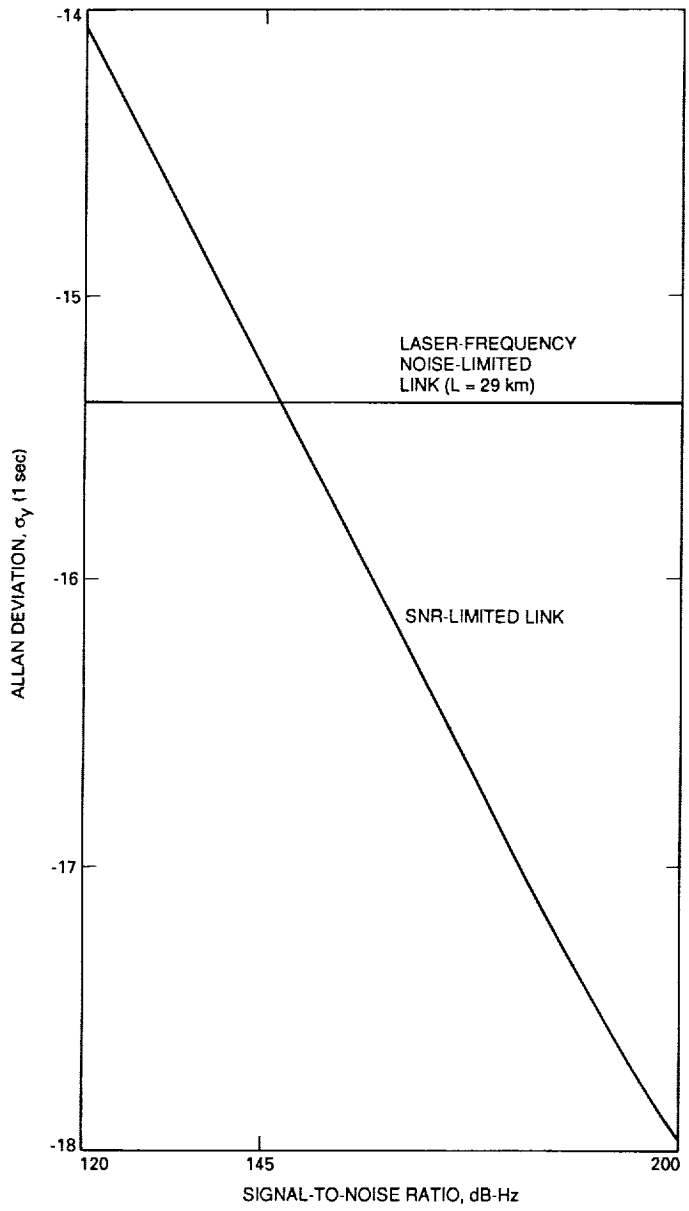


Fig. 5. Comparison of fiber link frequency stability at one second versus SNR with laser FM noise-limited 29-km link.

Thermal Coefficient of Delay for Various Coaxial and Fiber-Optic Cables

G. Lutes and W. Diener
Communications Systems Research Section

This article presents data on the thermal coefficient of delay for various coaxial and fiber-optic cables, as measured by the Frequency and Timing Systems Engineering Group and the Time and Frequency Systems Research Group. The measured pressure coefficient of delay is also given for the air-dielectric coaxial cables. The article includes a description of the measurement method and a description of each of the cables and its use at JPL and in the DSN. An improvement in frequency and phase stability by a factor of ten is possible with the use of fiber optics.

I. Introduction

Highly stable frequency and timing reference signals generated by atomic frequency standards enable the NASA/JPL Deep Space Network (DSN) to make precise measurements of relative time and position. These measurements are used to locate spacecraft and guide them to their destinations. They are also used to support radio science, radio and radar astronomy, very long baseline interferometry, and geodynamics.

Within a Deep Space Communications Complex (DSCC), high-stability distribution systems are used to distribute the frequency and timing reference signals derived from the frequency standard to the subsystems that use them. These distribution systems use coaxial or fiber-optic cables as the transmission medium. Delay changes in these cables are often the major contributor to the phase and frequency instability of the distributed reference signals.

Detailed data on delay stability of cables is generally not available from manufacturers. Since this information is needed to design frequency and timing distribution systems, the Time and Frequency Systems Research Group and the Frequency and Timing Systems Engineering Group have measured the thermal coefficient of delay (TCD) and the pressure coefficient of delay (PCD) for various coaxial and fiber-optic cables. This article presents the results of these measurements. A key finding is that an improvement in frequency and phase stability by a factor of ten is possible, compatible with anticipated requirements on the DSN for supporting gravitational-wave experiments and connected-element interferometry.

II. Background

Delay changes degrade the phase and frequency stability of a signal transmitted through a transmission line [1]. When the temperature of a transmission line changes, the

result is a corresponding delay change through the transmission line. A pressure change in an air-dielectric transmission line results in an additional delay change, since a dielectric constant change results from a pressure change.

The TCD is a measure of delay change in a signal path that results from a temperature change. Its value is often given in parts per million per deg Celsius (ppm/deg C), which is the change in delay divided by the total delay normalized to 1 million units. In equation form it is

$$\text{TCD} = \frac{\Delta t(10^6)}{t\Delta T} \tag{1}$$

where Δt is the change in delay through a signal path, t is the nominal delay through the signal path, and ΔT is the change in temperature. Similarly, the pressure coefficient of delay (PCD) is a measure of delay change in a signal path that results from a pressure change. Like TCD, the value of PCD is often given in ppm/psi.

In practice, phase measurements can be made with much higher resolution and accuracy than direct delay measurements. Therefore, to obtain the data presented in this article, phase changes were measured and converted to delay changes. Since phase delay has a linear relationship to time delay, Eq. (1) can be rewitten in terms of phase as

$$\text{TCD} = \frac{\Delta\Theta(10^6)}{\Theta\Delta T} \tag{2}$$

where $\Delta\Theta$ is the change in phase delay through a signal path, and Θ is the nominal phase delay through the signal path.

The TCD and PCD of various cables have been measured, and the data presented in this article will enable designers to identify a cable with suitable stability performance for a particular application.

III. Measurement Method

Figure 1 shows a block diagram of the measurement system. A reference signal is separated by an RF power splitter into two signals. One of these signals is passed through the cable under test, and the other signal is used as a reference. The signal at the output of the cable under test drives one port of an RF phase detector. The reference signal from the RF power splitter passes through a manual RF phase shifter and drives the other port of the RF phase detector. A lowpass filter on the output of the RF phase detector eliminates the RF signals, leaving only the DC component, which is measured with a DC voltmeter. A temperature-controlled test chamber contains the cable under test.

The phase detector's sensitivity is measured in volts per deg phase, and must be calibrated before measurements can be made. For most phase detectors, the output voltage vs. phase curve, commonly called the S-curve, is sinusoidal if the proper input signal levels are applied. When this is the case, it is only necessary to measure the peak voltage out of the phase detector in order to calibrate it. The peak voltage is obtained by adjusting the manual phase shifter for 0 deg or 180 deg phase difference between the signals at the phase detector input ports. In terms of the peak voltage, the slope of the S-curve, in volts per deg phase, near zero volts (90 deg) is

$$K_\Theta \approx \frac{\pi V_p}{180} \tag{3}$$

where V_p is the peak voltage out of the RF phase detector. If the phase detector curve is not sinusoidal, it is necessary to determine the slope of the curve by other means.

The measurements are made with the phase difference between the signals applied to the phase detector set near 90 deg. This is where the output of the phase detector is near zero volts. For these measurements, the phase detector output voltage vs. phase change is assumed to be linear in this region. However, the total phase-delay change in the cable under test should be no more than ± 30 deg over the test temperature range. This keeps errors due to the nonlinearity of the S-curve to less than 5 percent. A lower test frequency permits testing very long cables without exceeding a 30-deg phase-delay change over the test temperature range.

To make the measurement, the temperature of the test chamber is set to a nominal value, usually 25 deg C. The manual phase shifter is adjusted to obtain a 90-deg phase difference between the signals at the ports of the RF phase detector. For this phase difference, the output voltage from the RF phase detector is near zero. Once this zero-volt reference is established, the temperature in the test chamber is changed in steps. For each temperature change in the test chamber, the voltage out of the phase detector changes, indicating a phase (delay) change. The value of the phase-delay change is

$$\Delta\Theta = \frac{\Delta E}{K_\Theta} \tag{4}$$

where ΔE is the change in voltage due to a temperature change.

When the voltage change stabilizes, its new value is recorded and the temperature in the test chamber is stepped to a new value. This can take from 20 min to 1 hr per step. The temperature is normally changed in steps of 5 deg C. This usually results in a large enough phase change to be accurately measured. Yet the phase change is not so large that it results in significant error due to the nonlinearity of the phase-detector curve.

From Eqs. (2), (3), and (4), the TCD in terms of the known and measured parameters is

$$\mathrm{TCD} = \frac{\Delta E(10^6)(3 \times 10^8)\alpha}{2fl\pi V_p \Delta T} \tag{5}$$

where α is the propagation constant for the cable, f is the measurement frequency used, l is the physical length of the cable, and ΔT is the change in temperature of the transmission line.

A plot of the phase detector output vs. time and temperature for a 39.6-m length of RG-223/U cable is shown in Fig. 2. Each step in phase was the result of a temperature step of 5 deg C in the test chamber. The total temperature range is from 35 deg C at the bottom of the plot to 15 deg C at the top of the plot. The vertical scale shows 50 mV per major division (50 mV/div), and the horizontal scale shows time in 1 hr per major division (1 hr/div). The phase detector's peak output V_p was measured at 1.06 V. The propagation factor α for the cable is given by the manufacturer as 0.659. A 100-MHz test frequency was used. Using the example of the step between 30 deg C and 25 deg C, shown enclosed in dotted lines at the left side of Fig. 2, the change in voltage E is $1.2(5 \times 10^{-2}) = 6 \times 10^{-2}$. Evaluating Eq. (5) for this change in delay using the above information,

$$\mathrm{TCD} = \frac{(6 \times 10^{-2})(10^6)(3 \times 10^8)(0.659)}{2(10^8)(39.6)\pi(1.06)5}$$

$$= 90\mathrm{ppm/deg\ C\ at\ 27.5\ deg\ C}$$

which is the average temperature of this step.

IV. Results

The cables that have been measured in the Frequency Standards Laboratory (FSL) at JPL are loose-tube single-mode fiber-optic cable, low-TCD fiber-optic cable, and metal-based coaxial cables RG-223/U, SF-214, F242-VV-2400-AOB, F645-EIA-5160-AO, HCC-12-50J, 64-500, and 64-875. Each of these cables and its application at JPL and in the DSN is described in this section. The TCD for each cable is given and the PCD is given for those cables that are normally pressurized.

Loose-tube single-mode fiber-optic cable is used between Deep Space Stations (DSSs) at the Goldstone DSCC. These cables, supplied by several manufacturers, use Corning single-mode fiber and are very similar in design. Figure 3 depicts the general cable design. The performance characteristics are virtually identical for all of the manufacturers, and are given in Table 1 [2, 3, 4]. A graph of TCD with respect to temperature is shown in Fig. 4, which compares this type of cable to low-TCD optical fiber and the best coaxial cable (64-875).

A new optical fiber with a very low TCD has been tested in the FSL. This fiber is manufactured by Sumitomo Electric, and cables containing four fibers of this type are now being procured. They will be tested for distribution of several types of signals in the DSN, including frequency references, time references, local oscillator signals, and intermediate frequency (IF) signals. Table 2 lists the physical and performance characteristics of this new fiber [5]. Figure 5 shows a graph of its TCD with respect to temperature.

RG-223/U is a general-purpose coaxial cable in common use at JPL and in the DSN. It is manufactured by a number of companies, including Times Wire and Cable. Table 3 lists its important physical and performance characteristics, and Fig. 6 shows a graph of its TCD with respect to temperature [6].

SF-214 is also a general-purpose cable in common use at JPL and in the DSN. This cable is manufactured by Times Wire and Cable. It has lower loss than RG-223, and is used where this characteristic is important. It is also used in the antenna wrap-ups in the DSN where the cable must be flexed. Table 4 lists the important physical and performance characteristics of SF-214, and Fig. 7 shows a graph of its TCD with respect to temperature [6].

F242-VV-2400-AOB is a 3/8-in.-diameter coaxial cable with a corrugated outer conductor for flexibility. It is used in the FSL for test applications where the cable must have low TCD and be flexible to accommodate various test configurations. It is manufactured by Flexco Microwave. Table 5 lists its important physical and performance characteristics [7], and Fig. 8 shows a graph of its TCD with respect to temperature.

F645-EIA-5160-AO is a 1-in.-diameter coaxial cable with a corrugated outer conductor, manufactured by

Flexco Microwave. It was tested for possible use in the DSN. Table 6 lists its important physical and performance characteristics [7], and Fig. 9 shows a graph of its TCD with respect to temperature. The PCD of F645 cable with respect to air pressure is shown in Fig. 10.

HCC-12-50J is a 1/2-in.-diameter hardline cable used in the FSL where delay stability is critical but the cable is not flexed. It is manufactured by Cablewave Systems. Table 7 lists its important physical and performance characteristics [8] and Fig. 11 shows a graph of its TCD with respect to temperature.

Two pressurized hardline air-dielectric cables, 64-500 and 64-875, are used in the DSN where the lowest TCD and low attenuation are needed. The 64-875, a 7/8-in.-diameter cable, has better delay stability and lower attenuation than the 64-500 cable, which has a 1/2-in. diameter. However, the 64-875 cable is considerably more expensive and harder to work with. These two cables were manufactured by Prodelin, which was sold to Cablewave, who now manufactures them. Cablewave has informed the cable Cognizant Operations Engineer (COE) that they will stop manufacturing these cables in the near future. Table 8 lists the important physical and performance characteristics for 64-500 cable, and Figs. 12 and 13 show graphs of its TCD and PCD. Table 9 lists the important physical and performance characteristics of 64-875 [9], and Figs. 14 and 15 show graphs of its TCD and PCD.

V. Conclusion

Several fiber-optic cables and coaxial cables used at JPL and in the DSN have been measured to determine their TCD. The PCD of the air-dielectric cables was also measured. The plots of TCD and PCD given here are meant to guide the user in choosing a coaxial or fiber-optic cable for use in applications requiring high delay stability. Many of the technical parameters needed to make tradeoff decisions are given. The costs of these cables as given in the tables are to be taken as a guide only. For very large quantities, e.g., tens of kilometers, the costs are tied largely to material costs and market conditions, and the cost of short lengths of cable may be three to four times the large-quantity cost. In the normal operating range of temperatures from 15 to 35 deg C, the new low-TCD fiber-optic cable with superior delay stability permits signals to be transmitted with unprecedented frequency and phase stability. For instance, compared to the best coaxial cable at 25 deg C, the use of low-TCD fiber would improve the frequency and phase stability of a transmitted signal by more than ten times.

Acknowledgment

The authors wish to thank Al Kirk for developing the procedure used to measure the cable delay stability vs. temperature given in this article.

References

[1] G. Lutes, "High Stability Frequency and Timing Distribution Using Semiconductor Lasers and Fiber-Optic Links," *O-E Lase '89*, Los Angeles, California, pp. 263–271, January 17–20, 1989.

[2] "SMF-21 CPC3, Single-Mode Optical Fiber," specification sheet, Corning Glass Works, Corning, New York 14831, September 1987.

[3] "Singlemode Fiber Optic Cable," specification sheet no. OT-01/HP/7-85, Siecor Corporation, 489 Siecor Park, Hickory, North Carolina 28603-0489.

[4] A. Bergman, S. T. Eng, A. R. Johnston, and G. F. Lutes, "Temperature Dependence of Phase for a Single-Mode Fiber Cable," *Proc. Third International Conference on Integrated Optics and Optical Fiber Communications*, OSA-IEEE, San Francisco, California, p. 60, April 27–29, 1981.

[5] Data sheet no. 88-31, supplied by Sumitomo Electric Industries, Inc. to the Jet Propulsion Laboratory on June 1, 1988.

[6] "RF Transmission Line Catalog and Handbook," catalog no. TL-6, Times Wire and Cable, 358 Hall Avenue, Wallingford, Connecticut 06492, 1972.

[7] "Precision Flexible Cable Catalog," Flexco Microwave, Inc., PO Box 174, Karrville Road, Port Murray, New Jersey 07865, 1984.

[8] "Antenna and Transmission Line Systems Catalog," catalog no. 500A, Cablewave Systems, 60 Dodge Avenue, North Haven, Connecticut 06473, 1979.

[9] "General Catalog 1776," Prodelin, 1976. Information now available from Cablewave Systems, 60 Dodge Avenue, North Haven, Connecticut 06473.

Table 1. Physical and performance characteristics of loose-tube, single-mode, fiber-optic cable

Cutoff wavelength	1130 to 1270 nm
Core concentricity	$< 1\mu m$
Cladding diameter	$125 \pm 3\mu m$
Coating diameter	$250 \pm 15\mu m$
Core diameter	$8.7\mu m$
Spot size	$10\mu m$
Dispersion of 1285- to 1350-nm wavelength	< 3.5 psec/nm-km
Optical loss, maximum	0.5 dB/km
RF bandwidth	> 100 GHz-km
Number of fibers	18 and 24
Nominal weight	165 kg/km
Maximum diameter	15.1 mm
Temperature range	−40 to +70 deg C
Maximum tensile rating	
During installation	2700 N
Long-term after installation	600 N
Minimum bend radius	
During installation	300 mm
Free bend, installed	150 mm
Crush resistance, long-term installed	50 N/cm
Maximum vertical rise	175 m
Price	$0.25/fiber-meter to $0.35/fiber-meter

Table 2. Physical and performance characteristics of low-TCD, single-mode optical fiber cable

Cutoff wavelength	1260 nm
Core concentricity	$< 0.1\mu m$
Cladding diameter	$125.4\mu m$
Coating diameter	$815.0\mu m$
Spot size	$9.6\mu m$
Zero-dispersion wavelength	1305 nm
Zero-dispersion slope	< 0.083 psec/nm-km
Optical loss, maximum	0.32 dB/km
RF bandwidth	> 100 Ghz-km
Price	≈ $4/fiber-meter

Table 3. Physical and performance characteristics of RG-223/U coaxial cable

Inner conductor	Silver-covered copper; outside diameter 0.035 in.
Dielectric	Solid polyethylene; outside diameter 0.116 in.
Outer conductor	Two shielding braids, silver-covered copper
Jacket material	Black polyvinylchloride
Cable outside diameter	3/16 in.
Minimum bend radius	1.0 in.
Weight	0.034 lbs/ft
Impedance	50 ohms
Nominal capacitance	30.8 pf/ft
Maximum operating temperature range	−40 to +80 deg C
Maximum operating voltage	1900 volts RMS
Propagation constant	0.659
Nominal loss characteristics, dB/100 ft	
10 MHz	1.35
50 MHz	3.0
100 MHz	4.3
200 MHz	6.0
400 MHz	8.8
1 GHz	16.5
3 GHz	36.0
5 GHz	51.0
10 GHz	85.0
Price, < 300 meters	≈ $4.50/m

Table 4. Physical and performance characteristics of SF-214 coaxial cable

Inner conductor	Seven strands of 0.0296-in. silver-covered copper; outside diameter 0.089 in.
Dielectric	Solid polyethylene; outside diameter 0.285 in.
Outer conductor	Silver-covered copper, braided flat round composites
Jacket material	Black polyvinylchloride
Cable outside diameter	0.450 in.
Minimum bend radius	2.0 in.
Weight	0.144 lbs/ft
Impedance	50 ohms
Cutoff frequency	13.7 GHz
Nominal capacitance	30.8 pf/ft
Maximum operating temperature range	−55 to +80 deg C
Maximum operating voltage	5000 volts RMS
Propagation constant	0.659
Minimum recommended bend radius	2.0 in.
Nominal loss characteristics, dB/100 ft	
10 MHz	Not available
50 MHz	Not available
100 MHz	2.0
200 MHz	2.9
400 MHz	4.1
1 GHz	7.0
3 GHz	13.0
5 GHz	18.0
10 GHz	27.0
Price, < 300 meters	≈ $20/m

Table 5. Physical and performance characteristics of F242-VV-2400-AOB coaxial cable

Inner conductor	Silver-covered copper; outside diameter 0.081 in.
Dielectric	Air; spline polytetrafluoroethylene spacer; outside diameter 0.200 in.
Outer conductor	Soldered strip-wound (corrugated) silver-covered copper; outside diameter 0.330 in.
Jacket material	Fluorinated ethylene propylene
Cable outside diameter	3/8 in.
Minimum bend radius	1.0 in.
Weight	Not available
Impedance	50 ohms
Cutoff frequency	20 GHz
Nominal capacitance	Not available
Maximum operating temperature range	−55 to +200 deg C
Maximum operating voltage	5000 volts RMS
Propagation constant	0.80
Nominal loss characteristics, dB/100 ft	
10 MHz	Not available
50 MHz	Not available
100 MHz	2.5
200 MHz	4.0
400 MHz	6.0
1 GHz	8.5
3 GHz	15.0
5 GHz	19.5
10 GHz	37.0
Price, < 1 km	≈ $102/m
> 5 km	≈ $30/m

Table 6. Physical and performance characteristics of F645-EIA-5160-AO coaxial cable

Inner conductor	Stranded silver-covered copper
Dielectric	Air; spline polytetrafluoroethylene spacer
Outer conductor	Soldered strip-wound (corrugated) silver-covered copper; outside diameter 1.025 in.
Jacket material	Fluorinated ethylene propylene
Cable outside diameter	1.065 in.
Minimum bend radius	5.5 in.
Weight	Not available
Impedance	50 ohms
Cutoff frequency	6 GHz
Nominal capacitance	Not available
Maximum operating temperature range	−55 to +200 deg C
Maximum operating voltage	5000 volts RMS
Propagation constant	0.79
Nominal loss characteristics, dB/100 ft	
10 MHz	Not available
50 MHz	Not available
100 MHz	0.60
200 MHz	0.95
400 MHz	1.50
1 GHz	2.70
3 GHz	5.0
5 GHz	6.80
10 GHz	10.0
Price, < 1 km	≈ $148/m
> 5 km	≈ $40/m

Table 7. Physical and performance characteristics of HCC-12-50J coaxial cable

Inner conductor	Copper-clad aluminum; outside diameter 0.155 in.
Dielectric	Air; spiral polyethylene spacer; outside diameter 0.338 in.
Outer conductor	Corrugated copper; outside diameter 0.484 in.
Jacket material	Black polyethylene
Cable outside diameter	0.618 in.
Minimum bend radius	5 in.
Weight	0.16 lbs/ft
Impedance	50 ohms
Cutoff frequency	11.3 GHz
Nominal capacitance	Not available
Maximum operating temperature range	−55 to +80 deg C
Maximum operating voltage	Not available
Propagation constant	0.915
Nominal loss characteristics, dB/100 ft	
10 MHz	0.26
50 MHz	0.59
100 MHz	0.85
200 MHz	1.3
400 MHz	1.8
1 GHz	2.9
3 GHz	5.0
5 GHz	7.0
10 GHz	10.0
Price, > 5 km	≈ $12/m

Table 8. Physical and performance characteristics of 64-500 pressurized hardline air-dielectric cable

Inner conductor	Copper-clad aluminum; outside diameter 0.167 in.
Dielectric	Air; six polyethylene tubes; outside diameter 0.456 in.
Outer conductor	Aluminum; outside diameter 0.530 in.
Jacket material	Black polyethylene
Cable outside diameter	0.550 in.
Minimum bend radius	5 in.
Weight	0.22 lbs/ft
Impedance	50 ohms
Cutoff frequency	Not available
Nominal capacitance	Not available
Maximum operating temperature range	−55 to +80 deg C
Maximum operating voltage	3.4 volts RMS
Propagation constant	0.855
Nominal loss characteristics, dB/100 ft	
10 MHz 0.24	
50 MHz 0.53	
100 MHz 0.75	
200 MHz 1.1	
400 MHz 1.5	
1 GHz 2.4	
3 GHz 4.6	
5 GHz 6.4	
10 GHz 10.0	
Price, < 1 km	≈ $35/m
> 5 km	≈ $10/m

Table 9. Physical and performance characteristics of 64-875 pressurized hardline air-dielectric cable

Inner conductor	Copper-clad aluminum; outside diameter 0.311 in.
Dielectric	Air; six polyethylene tubes; outside diameter 0.837 in.
Outer conductor	Aluminum; outside diameter 0.953 in.
Jacket material	Black polyethylene
Cable outside diameter	1.023 in.
Minimum bend radius	10 in.
Weight	0.46 lbs/ft
Impedance	50 ohms
Cutoff frequency	Not available
Nominal capacitance	Not available
Maximum operating temperature range	−55 to +80 deg C
Maximum operating voltage	6.0 volts RMS
Propagation constant	0.855
Nominal loss characteristics, dB/100 ft	
10 MHz	0.13
50 MHz	0.30
100 MHz	0.43
200 MHz	0.6
400 MHz	0.86
1 GHz	1.4
3 GHz	2.7
5 GHz	4.0
10 GHz	Not available
Price, < 1 km	≈ $66/m
> 5 km	≈ $18/m

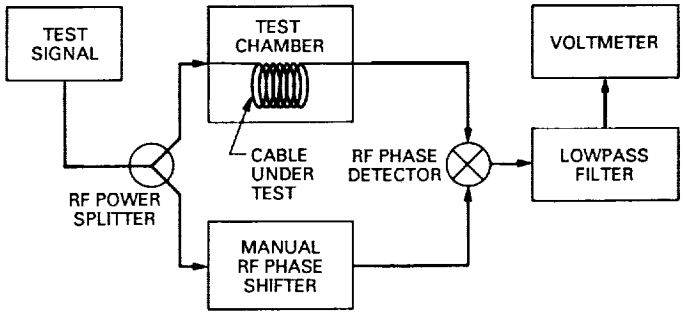


Fig. 1. Block diagram of the system used to measure cable thermal coefficient of delay (TCD) and pressure coefficient of delay (PCD).

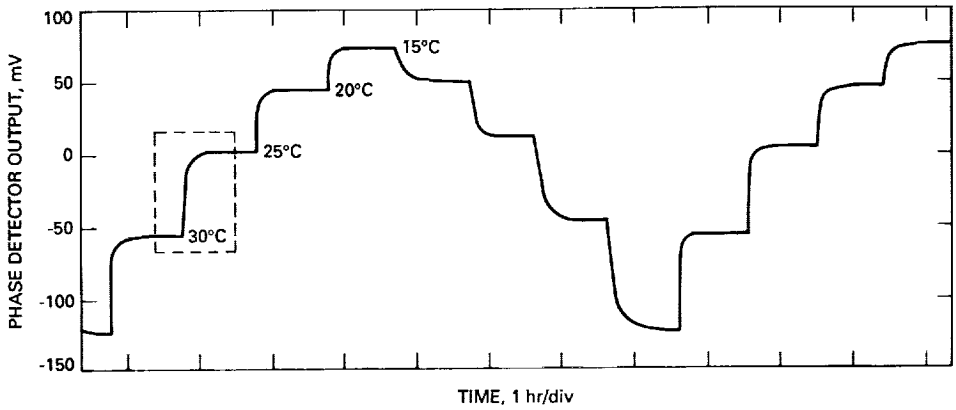


Fig. 2. Example of recorded data for RG-223/U coaxial cable.

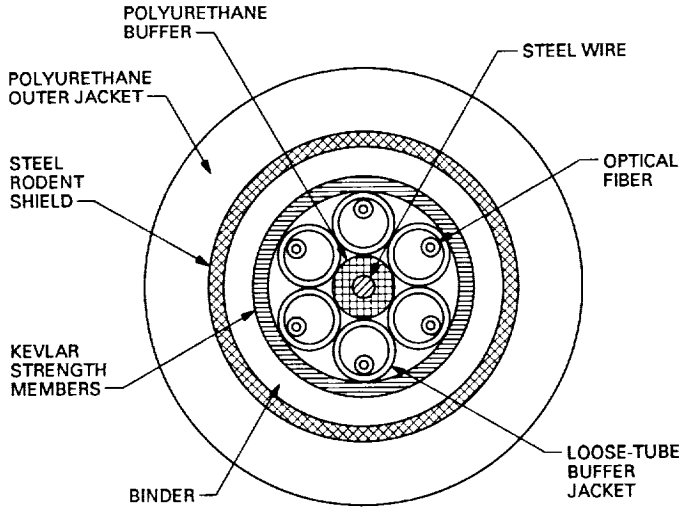


Fig. 3. Construction of a typical loose-tube fiber-optic cable.

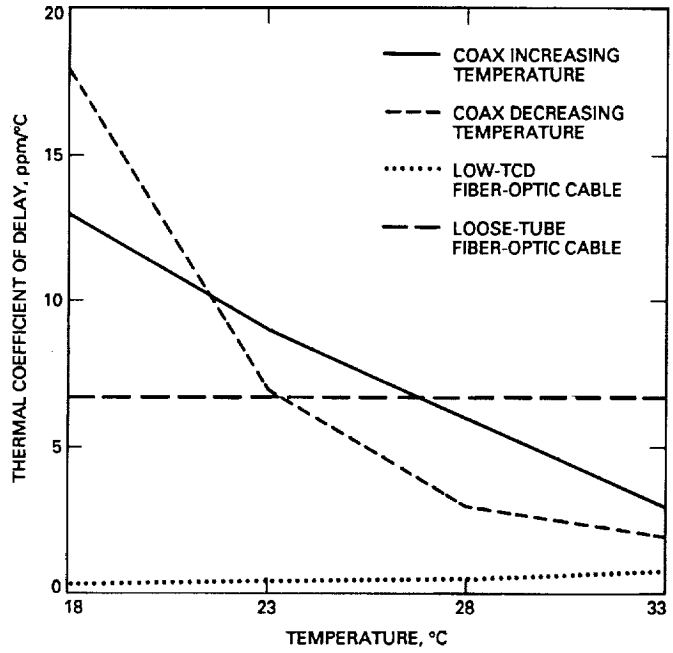


Fig. 4. A comparison of the measured TCD of loose-tube, single-mode, fiber-optic cable, the best coaxial cable (64-875), and low-TCD fiber-optic cable.

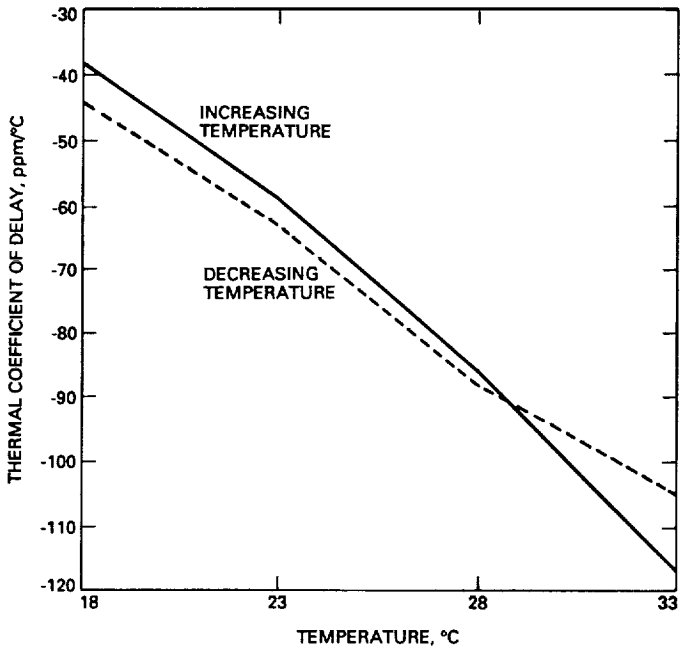


Fig. 6. Measured TCD of RG-223/U coaxial cable.

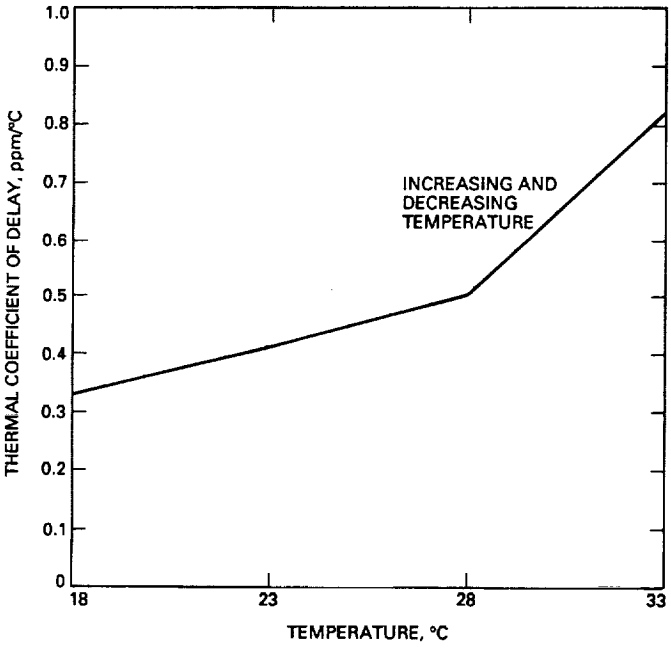


Fig. 5. Measured TCD of Sumitomo Electric optical fiber.

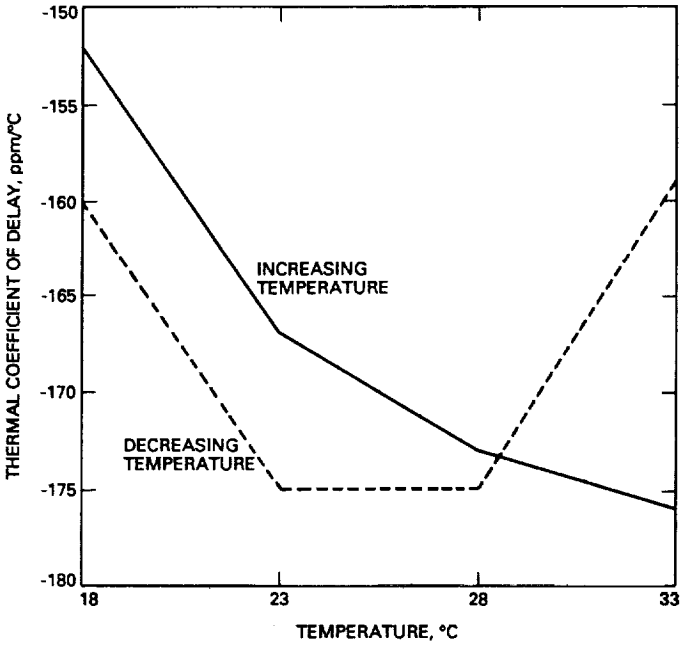


Fig. 7. Measured TCD of SF-214 coaxial cable.

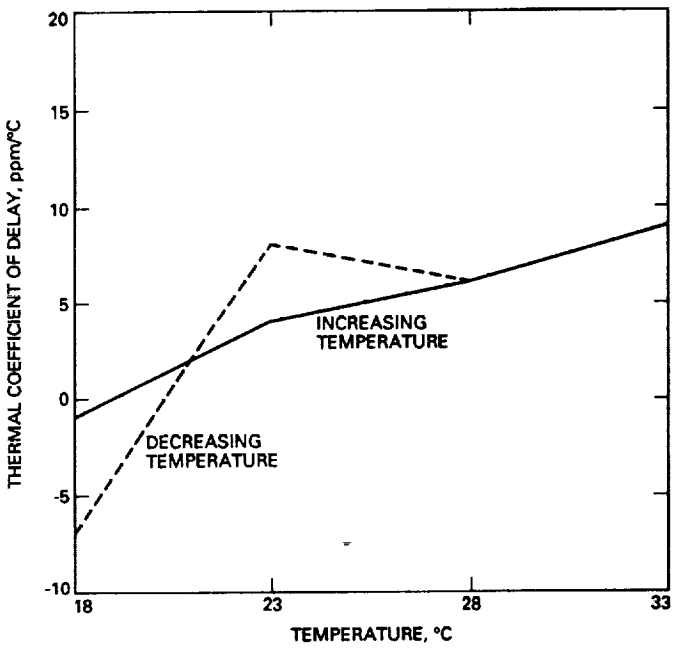


Fig. 8. Measured TCD of F242 coaxial cable.

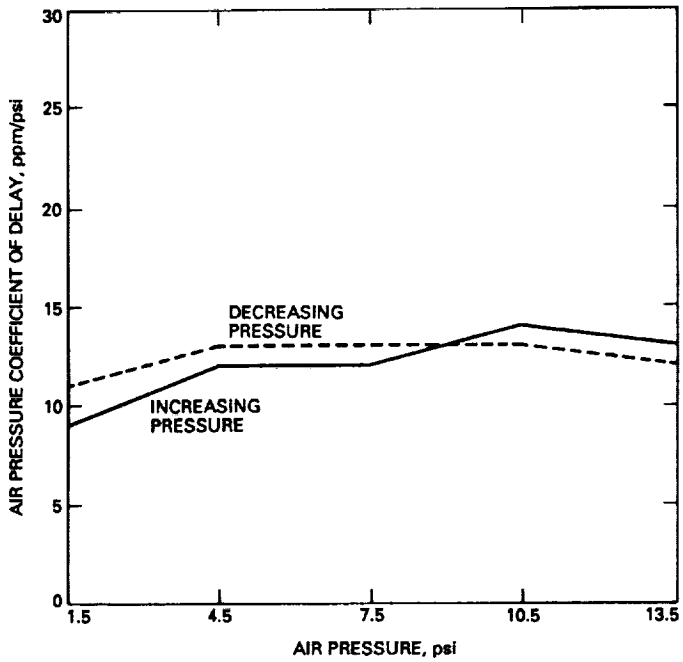


Fig. 10. Measured PCD of F645 coaxial cable.

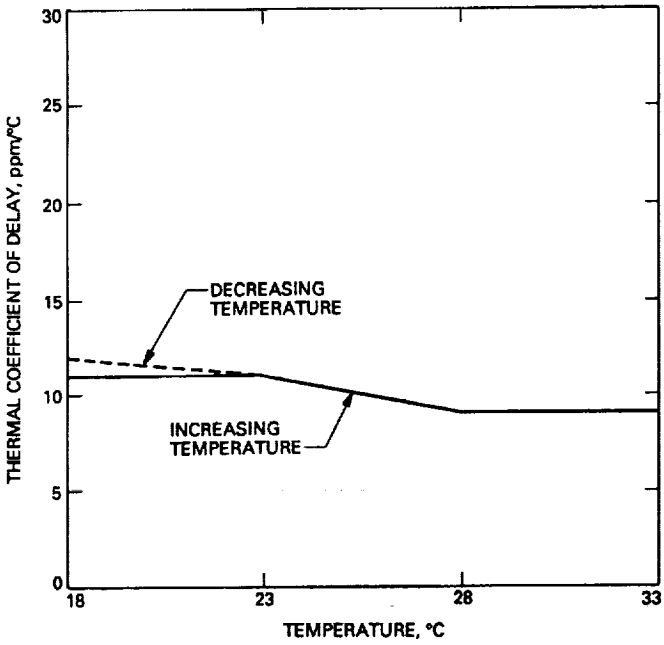


Fig. 9. Measured TCD of F645 coaxial cable.

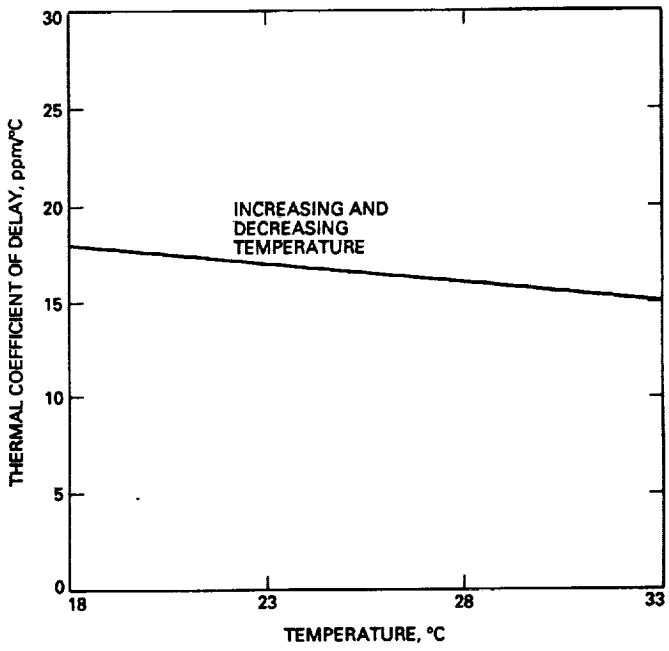


Fig. 11. Measured TCD of HCC-12-50J coaxial cable.

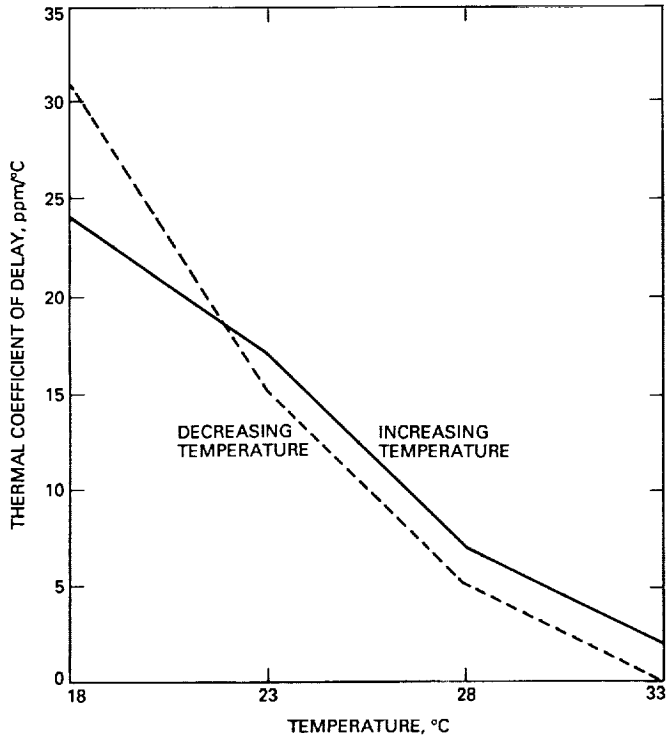


Fig. 12. Measured TCD of 64-500 coaxial cable.

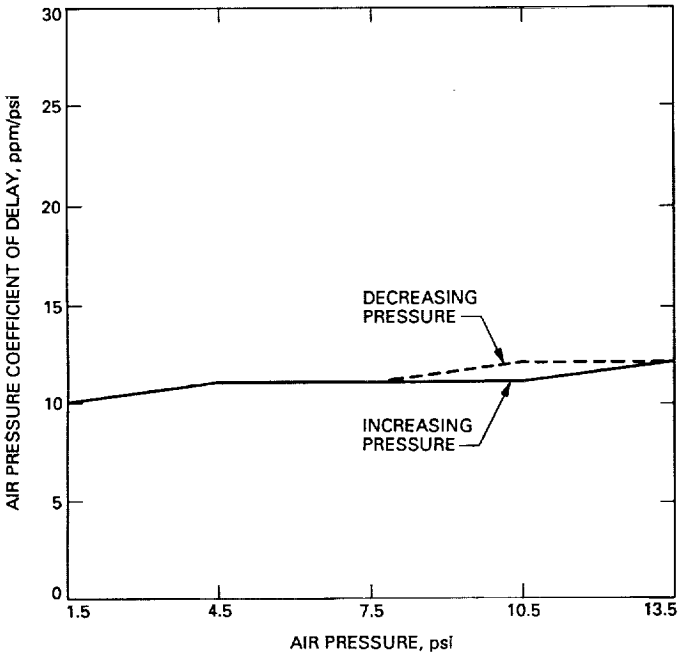


Fig. 13. Measured PCD of 64-500 coaxial cable.

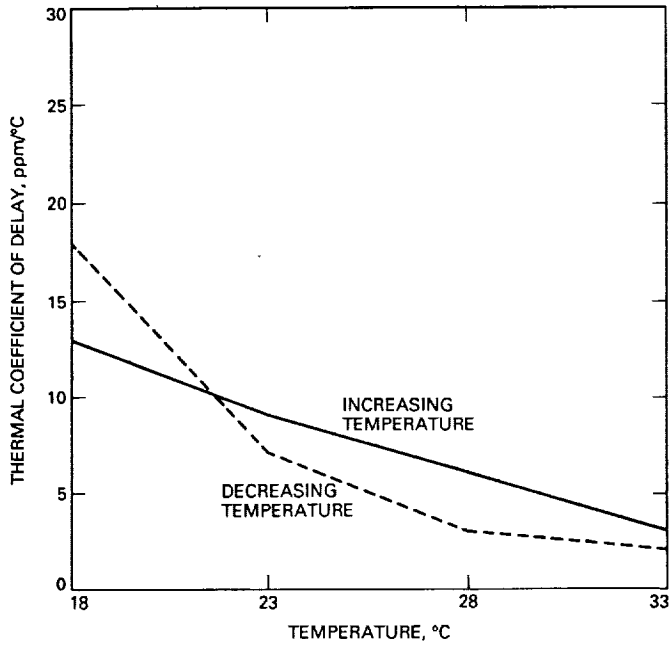


Fig. 14. Measured TCD of 64-875 coaxial cable.

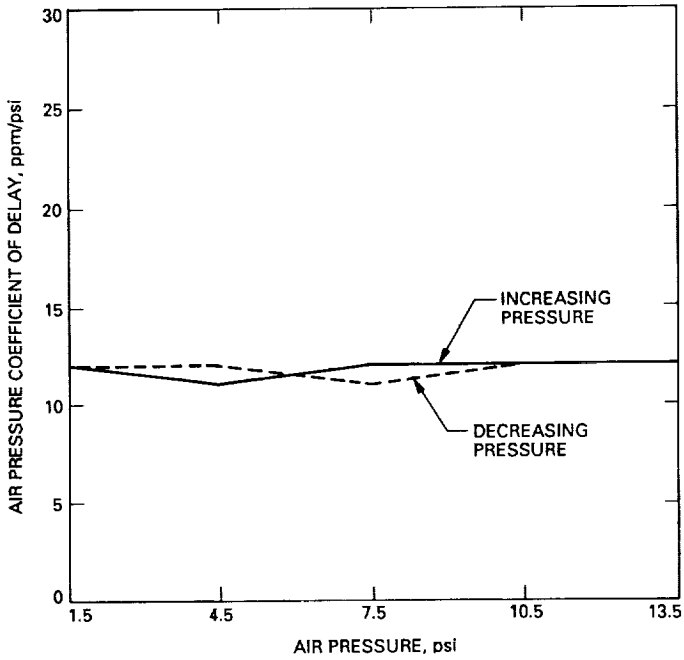


Fig. 15. Measured PCD of 64-875 coaxial cable.

N90-19440

TDA Progress Report 42-99

November 15, 1989

Performance of the All-Digital Data-Transition Tracking Loop in the Advanced Receiver

U. Cheng and S. Hinedi
Communications Systems Research Section

This article describes the performance of the all-digital data-transition tracking loop (DTTL) with coherent or noncoherent sampling. The effects of few samples per symbol and of noncommensurate sampling rates and symbol rates are addressed and analyzed. Their impacts on the loop phase-error variance and the mean time to lose lock (MTLL) are quantified through computer simulations. The analysis and preliminary simulations indicate that with three to four samples per symbol, the DTTL can track with negligible jitter because of the presence of Earth Doppler rate. Furthermore, the MTLL is also expected to be large enough to maintain lock over a Deep Space Network track.

I. Introduction

In modern digital communication systems, analog-to-digital conversion (ADC) is performed as far toward the front end as possible using available technology. Usually, the received signal is amplified and then downconverted to the appropriate frequency for digital conversion. Thereafter, various system functions are performed digitally, including carrier, subcarrier, and symbol synchronization, as well as signal detection and decoding. Depending on the application, either the baseband signals (inphase and quadrature) or the intermediate frequency (IF) signal can be sampled. Furthermore, the sampling clock can be free-running or controlled by the symbol-synchronization loop. In the latter case, the sampling clock can be adjusted to obtain an integer number of samples per cycle of the IF signal, or to obtain an integer number of samples per received symbol. All of these issues affect the final architecture and design of a receiver and influence the amount of cross-coupling among the various loops.

Since sampling is done up front, the various tracking loops need to be implemented digitally. The classical analog integrate-and-dump (I&D) filters, which are typically part of the loop arms (inphase and quadrature), must be replaced by digital accumulators. This article investigates the performance of the all-digital data-transition tracking loop (DTTL) with small noninteger numbers of samples per symbol. In the previous version of the Advanced Receiver (ARX I) [1], the sampling was performed synchronously with the symbol rate, and a large number of samples per symbol were available. In the current version of the Advanced Receiver (ARX II), the sampling is performed asynchronously and the sample clock is independent of the symbol rate. At the highest required data rate

of 6.6 Msymbols/sec and the processing rate of 20 MHz for the Block V receiver, only about three samples per symbol are obtained.

Some analytical results for the phase-error variance of the analog DTTL were first derived in [2], where the input was an analog signal and symbol and midphase detection were performed with analog I&D filters. Later, the analysis was reworked [3], taking into account variations of the equivalent noise spectrum with respect to normalized phase error.

The interest here is in the loop response and performance of all-digital DTTLs, where digital symbol detection and digital midphase accumulation are used. There are two sampling scenarios: one is to sample the signal instantaneously, and the other is to obtain the sample by I&D sampling of the signal. The instantaneous sampling technique can be used when the sampling rate is significantly higher than the symbol rate. The I&D sampling technique should be used when the number of samples per symbol is small. If the received symbol waveform is a perfect square wave, the samples by instantaneous sampling all have equal amplitude. The samples by I&D sampling also have equal amplitude, except for the first sample of every symbol which has a different polarity from the symbol immediately preceding it. These changes in symbol polarity are referred to as the transition boundaries. The first sample after each transition boundary has a smaller amplitude than other samples due to integration across the transition boundary. The all-digital DTTL can operate on either type of sample. It is worthwhile noting at this time that when the received signal is filtered and instantaneously sampled, the process can be modeled to the first degree by I&D sampling of an ideal waveform. Thus, I&D sampling can be thought of as a tool to model the filtering operations in the receiver.

The components in the all-digital DTTL affected by the type of sampling are the symbol detector and the midphase accumulator. For instantaneous sampling, the symbol detector accumulates all samples in the current symbol epoch. The midphase detector accumulates all samples in the current transition-detection window. For I&D sampling, the problem is slightly different. In this case, even when a sample is in the current symbol epoch, most of its energy may be from the previous symbol epoch. Therefore, a more sophisticated rule is needed to determine whether a particular sample should be used for the detection of a particular symbol.

A reasonable criterion is to include a particular sample in the current symbol if more than half of its energy is from the current symbol. This criterion leads to a simple rule for the operation of the symbol and midphase detectors. The rule is as follows: the first sample after each symbol boundary should belong to the previous symbol if the time offset between the sample and the symbol boundary is less than half of the sampling interval; otherwise, the sample belongs to the current symbol. A sample mark is one-half of a sampling interval ahead of its respective sample time for I&D sampling, and is the respective sample time for instantaneous sampling. Thus, the rule can be restated: the symbol detector accumulates all samples with their sample marks in the current symbol epoch, and the midphase detector accumulates all samples with their sample marks in the current transition-detection window. Therefore, the DTTL with I&D sampling is similar to that with instantaneous sampling if the concept of a sample mark is used.

To simplify the mathematical analysis, the effects of unequal I&D sample amplitude immediately following transition boundaries are ignored, and instead equal amplitude for all I&D samples is assumed.

To illustrate the differences between analog and all-digital DTTLs, the noiseless case is considered first. Note that if the input is an analog signal, the midphase integrator can produce a nonzero error voltage no matter how small the phase error is. Thus, a correction voltage can be generated at every symbol transition whenever a phase error exists. Therefore, the analog DTTL has infinite resolution for phase detection.

In contrast, the all-digital DTTL has only finite resolution for phase detection. This is illustrated in the following example. Suppose that there is an even number of samples per symbol. When a symbol transition occurs, the digital midphase accumulator can produce a nonzero voltage only if the phase error causes sample slipping (assuming samples of equal amplitude). As long as the phase error stays within a range of values that avoids sample slipping, the loop always generates a no-error signal. This range of undetectable phase errors accounts for the finite resolution of the all-digital DTTL. The more samples per symbol that are used, the higher the achievable resolution, and the closer the all-digital DTTL is to its analog counterpart. A key question is the impact of the all-digital DTTL's finite resolution on the phase-error variance for few samples per symbol (say, four or five samples).

Another issue in an all-digital implementation is the effect of a noninteger number of samples per symbol. If the sampling clock is driven by the symbol-synchronization loop, the number of samples per symbol can be made an exact even integer, which reduces the self-noise generated

in the midphase accumulator (as discussed later). Under that sampling scenario, the sampling clock is constantly adapting as the data rate changes due to Doppler or other effects. One disadvantage of that scheme is that no fixed time base is available in the system. On the other hand, if the sampling clock is free-running and is derived from a fixed frequency standard, the sampling period is fixed, although the symbol rate may change. This may result in a noninteger number of samples per symbol. A model is derived in this article to analyze the performance of the DTTL where the sampling rate and the data rate are noncommensurate. Other issues such as the mean time to lose lock (MTLL), probability of symbol error, probability of losing lock, and error variance are also investigated via simulations. In Section II, a general analysis of the loop is presented in handling several scenarios. A discussion and comparison with simulation results are given in Section III, and the conclusion is given in Section IV.

II. Analysis

The performance of the all-digital data-transition tracking loop with noncoherent sampling is analyzed here. The block diagram of the all-digital DTTL is delineated in Fig. 1. The input $r(i)$ to the DTTL can be obtained by instantaneous sampling or by I&D sampling. For the Advanced Receiver II, the number of samples per symbol becomes small at high data rates, and therefore the I&D sampling technique is used. In the subsequent derivation, equal-amplitude samples are assumed.

Noncoherent sampling means that the sampling clock runs independently of the estimated symbol phase, i.e., the sampling time interval and the sampling time do not change with the estimated symbol phase. This is not an issue if there are many samples per symbol. The problem becomes complex as the number of samples per symbol decreases. The proposed Advanced Receiver II has about three to four samples per symbol at high data rates (the goal is 6.6 Msymbols per sec). Noncoherent sampling results in a noninteger number of samples per symbol. All of these factors affect the performance of the DTTL by changing its S-curve and by introducing self-noise. Considered here are the probability of loss of lock, the MTLL, the degradation of the symbol detection, and the phase-error variance. An approximate theory is presented for a first-order DTTL. The approach is to derive the S-curve and then solve the Fokker-Planck equation to get the density function of the phase error. The phase-error variance and the degradation of the symbol detection can be evaluated from the phase-error density function.

To illustrate the phenomenon of self-noise, a simple example is shown in Fig. 2, where there are five samples per symbol. Assumed are no thermal noise and perfect tracking at a particular moment. The output of the symbol-transition detector is not zero because it sums three samples from the first symbol and two samples from the second symbol (Fig. 2a). Notice that this situation occurs for every symbol interval as long as the loop maintains perfect tracking. The nonzero output of the loop filter will gradually drag the loop away from the perfect tracking condition until the polarity of the output of the symbol-transition detector changes (Fig. 2b). The loop filter cannot eliminate this type of self-noise. The problem is more complex if the number of samples per symbol is not an integer. In order to describe this phenomenon, three useful parameters are introduced here. Let β denote the number of samples per symbol, which may not be an integer, and α denote the offset of the first sample mark in a received symbol from the symbol boundary. By convention, α is normalized and is measured as a percentage of the sampling interval. If β is an integer, α remains constant; if β is not an integer, α varies from symbol to symbol. Let the received symbol be numbered 0, 1, 2, ..., and let the value of α at the first symbol be denoted as $\alpha_0 \equiv \gamma$, which is referred to as the initial sampling offset. The values of α at the subsequent symbols, namely, α_1, α_2, ..., can be computed from β and γ. The number of sample marks in a transition-detection window and the number of sample marks in a symbol-detection window are functions of α. Thus the output of the symbol detector and that of the transition detector fluctuate from symbol to symbol as α_i. This subject will be further discussed later.

Another important observation about the DTTL with noncoherent sampling is that there is inherent phase error due to finite samples per symbol. To illustrate this phenomenon, consider the example shown in Fig. 3, where every symbol contains four samples. As long as the estimated phase lies between t_1 and t_2, the error signal is always zero (or nearly zero if the received symbol does not have a perfect square waveform or if there are unequal-strength samples from I&D sampling), and the DTTL remains in tracking. However, unresolved phase ambiguity still exists within the interval from t_1 to t_2. Mathematically, this phase ambiguity can be explained by a step-like S-curve. This phenomenon might have little effect on symbol detection performance if straight accumulation is used to detect the symbols. However, if weighted accumulation is used to detect the symbols, the phase ambiguity can introduce misweighting and thus degrade performance. The symbol-error probability can be obtained by simulation.

Before proceeding to the mathematical analysis, examine the all-digital DTTL block diagram again in Fig. 1. The error-signal accumulator between the loop filter $F(z)$ and the multiplier performs an averaging function so that

the subsequent loop filter can operate at a slower speed. The loop bandwidth is determined primarily by the loop filter $F(z)$. Thus the presence of the accumulator is simply for hardware convenience. In the following analysis, the DTTL is considered without the error-signal accumulator.

A. Mathematical Model

Assuming that the carrier and subcarrier (if any) have been removed in an ideal fashion, the received waveform is given by

$$r(t) = \sqrt{S}\sum_k a_k\, p(t - kT) + n(t)$$

where S is the data power, $n(t)$ is white Gaussian noise with two-sided power spectral density $N_0/2$ W/Hz, $a_k = \pm 1$ represents the polarity of the kth symbol, and $p(t)$ is the square-wave function having value 1 for $0 \le t < T$ and having value 0 elsewhere. With I&D sampling, the ith sample can be expressed as

$$r(i) = \sqrt{S}a_k + n(i) \tag{1}$$

where it is assumed that the ith sample is derived from the kth symbol, $n(i)$ is a zero-mean Gaussian random variable with variance $\sigma^2 = N_0/(2T_s)$, and T_s is the sampling interval. Note that equal sample strength is assumed in Eq. (1).

Let the phase error λ (in cycles) be defined as

$$\lambda = \frac{\theta - \hat{\theta}}{2\pi}$$

where θ is the actual received symbol phase and $\hat{\theta}$ is the estimated symbol phase. Note that λ should have a value between -0.5 and 0.5. The error signal is affected by the locations of samples within their respective received symbols. In order to quantify this effect, a set of twelve Δ functions is defined, six for the $\lambda \ge 0$ case and six for the $\lambda < 0$ case. They are the numbers of sample marks contained in their respective intervals defined in Fig. 4(a). The output of the inphase accumulator $x(k)$ and the output of the midphase accumulator $y(k)$ can be expressed in terms of the Δ functions. If $\lambda \ge 0$, then

$$x(k) = \sqrt{S}(\Delta_1 a_k + \Delta_2 a_{k+1}) + n_1(k) + n_2(k) + n_3(k)$$

$$x(k+1) = \sqrt{S}(\Delta_3 a_{k+1} + \Delta_4 a_{k+2}) + n_4(k) + n_5(k) + n_6(k)$$

and

$$y(k) = \sqrt{S}(\Delta_5 a_k + \Delta_6 a_{k+1}) + n_2(k) + n_3(k) + n_4(k) \tag{2}$$

where $n_j(k)$, $1 \le j \le 6$ are zero-mean Gaussian random variables with their respective variances $(\Delta_1 - \Delta_5)\sigma^2$, $\Delta_5\sigma^2$, $\Delta_2\sigma^2$, $(\Delta_6 - \Delta_2)\sigma^2$, $(\Delta_3 + \Delta_2 - \Delta_6)\sigma^2$, and $\Delta_4\sigma^2$.

The Δ functions are computed using the following equations:

$$\Delta_1 = \lfloor \beta - \alpha \rfloor - \lfloor \lambda\beta - \alpha \rfloor$$

$$\Delta_2 = \lfloor (1+\lambda)\beta - \alpha \rfloor - \lfloor \beta - \alpha \rfloor$$

$$\Delta_3 = \lfloor 2\beta - \alpha \rfloor - \lfloor (1+\lambda)\beta - \alpha \rfloor$$

$$\Delta_4 = \lfloor (2+\lambda)\beta - \alpha \rfloor - \lfloor 2\beta - \alpha \rfloor \tag{3}$$

$$\Delta_5 = \begin{cases} \lfloor \beta - \alpha \rfloor - \lfloor (1+\lambda-(W/2))\beta - \alpha \rfloor & \text{if } W/2 > 1+\lambda \\ 0 & \text{if } W/2 < 1+\lambda \end{cases}$$

$$\Delta_6 = \begin{cases} \lfloor (1+\lambda+(W/2))\beta - \alpha \rfloor - \lfloor \beta - \alpha \rfloor & \text{if } W/2 > 1+\lambda \\ \lfloor (1+\lambda+(W/2))\beta - \alpha \rfloor - \lfloor (1+\lambda-(W/2))\beta - \alpha \rfloor & \text{if } W/2 < 1+\lambda \end{cases}$$

if $\lambda > 0$, and

$$\Delta_1' = \lfloor (1+\lambda)\beta - \alpha \rfloor + 1$$

$$\Delta_2' = \begin{cases} \lfloor -\lambda\beta + \alpha \rfloor & \text{if } -\lambda\beta + \alpha \text{ is not an integer} \\ \lfloor -\lambda\beta + \alpha \rfloor + 1 & \text{if } -\lambda\beta + \alpha \text{ is an integer} \end{cases}$$

$$\Delta_3' = \lfloor (2+\lambda)\beta - \alpha \rfloor - \lfloor \beta - \alpha \rfloor$$

$$\Delta_4' = \lfloor \beta - \alpha \rfloor - \lfloor (1+\lambda)\beta - \alpha \rfloor \tag{4}$$

$$\Delta_5' = \begin{cases} \lfloor \beta - \alpha \rfloor - \lfloor (1+\lambda-(W/2))\beta - \alpha \rfloor & \text{if } W/2 > -\lambda \\ \lfloor (1+\lambda+(W/2))\beta - \alpha \rfloor - \lfloor (1+\lambda-(W/2))\beta - \alpha \rfloor & \text{if } W/2 < -\lambda \end{cases}$$

$$\Delta_6' = \begin{cases} \lfloor (1+\lambda+(W/2))\beta - \alpha \rfloor - \lfloor \beta - \alpha \rfloor & \text{if } W/2 > -\lambda \\ 0 & \text{if } W/2 < -\lambda \end{cases}$$

---

if $\lambda < 0$, where $\lfloor y \rfloor$ is the greatest integer strictly less than y.

In the above equations, W is the width of the transition-detection window. The derivations of the twelve Δ functions are similar. Two examples are given here, Δ_2 and Δ_3'. To derive Δ_2, the beginning of the kth received symbol is used as the reference point. The number of sample marks in the kth received symbol is $\lfloor \beta - \alpha \rfloor$. The number of sample marks from the beginning of the kth received symbol to the end of the kth estimated symbol is $\lfloor (1+\lambda)\beta - \alpha \rfloor$. Equation (3) follows by observing that the number of sample marks from the end of the kth received symbol to the end of the kth estimated symbol is Δ_2. To derive Δ_3', the beginning of the kth received symbol is also used as the reference point. The number of sample marks in the kth received symbol is $\lfloor \beta - \alpha \rfloor$. The number of sample marks from the beginning of the kth received symbol to the end of the $(k+1)$th estimated symbol is $\lfloor (2+\lambda)\beta - \alpha \rfloor$. Equation (4) follows by observing that the number of sample marks from the end of the kth received symbol to the end of the $(k+1)$th estimated symbol is Δ_3'.

The error signal $e(k)$ is given by

$$e(k) = z(k)y(k)$$

The conditional S-curve is defined by

$$g(\lambda|\alpha) = E_{n,s}\{e(k)|\lambda, \alpha\}$$

where $E_{n,s}$ represents the conditional expectation on λ with respect to the noise and the signal. Following similar mathematical manipulation as in [2],

$$\frac{4E\{e(k)|\lambda \geq 0,\alpha\}}{\beta\sqrt{S}} = \frac{\Delta_6}{\beta}\left\{\operatorname{erf}\left(\sqrt{(\Delta_3+\Delta_4)E_s/\beta}\right) + \operatorname{erf}\left((\Delta_3-\Delta_4)\sqrt{E_s/(\beta(\Delta_3+\Delta_4))}\right)\right\}$$

$$- \frac{\Delta_5+\Delta_6}{\beta}\operatorname{erf}\left(\sqrt{(\Delta_1+\Delta_2)E_s/\beta}\right)$$

$$- \frac{\Delta_5-\Delta_6}{\beta}\operatorname{erf}\left((\Delta_1-\Delta_2)\sqrt{E_s/(\beta(\Delta_1+\Delta_2))}\right)$$

$$+ \frac{\Delta_6-\Delta_2}{\sqrt{\pi\beta(\Delta_3+\Delta_4)E_s}}\left\{\exp\left(-\frac{\Delta_3+\Delta_4}{\beta}E_s\right) + \exp\left(-\frac{(\Delta_3-\Delta_4)^2}{(\Delta_3+\Delta_4)\beta}E_s\right)\right\}$$

$$- \frac{\Delta_2+\Delta_5}{\sqrt{\pi\beta(\Delta_1+\Delta_2)E_s}}\left\{\exp\left(-\frac{\Delta_1+\Delta_2}{\beta}E_s\right) + \exp\left(-\frac{(\Delta_1-\Delta_2)^2}{(\Delta_1+\Delta_2)\beta}E_s\right)\right\}$$

where E_s is the symbol energy-to-noise ratio, namely,

$$E_s = \frac{ST}{N_0}$$

Using the same approach yields a similar result for the $\lambda < 0$ case:

$$\frac{4E\{e(k)|\lambda < 0,\alpha\}}{\beta\sqrt{S}} = \frac{\Delta'_5+\Delta'_6}{\beta}\operatorname{erf}\left(\sqrt{(\Delta'_3+\Delta'_4)E_s\beta}\right)$$

$$+ \frac{\Delta'_5-\Delta'_6}{\beta}\operatorname{erf}\left((\Delta'_4-\Delta'_3)\sqrt{E_s/(\beta(\Delta'_3-\Delta'_4))}\right)$$

$$- \frac{\Delta'_5}{\beta}\left\{\operatorname{erf}\left(\sqrt{(\Delta'_1+\Delta'_2)E_s/\beta}\right) + \operatorname{erf}\left((\Delta'_1-\Delta'_2)\sqrt{E_s/(\beta(\Delta'_1+\Delta'_2))}\right)\right\}$$

$$+ \frac{\Delta'_4-\Delta'_6}{\sqrt{\pi\beta(\Delta'_3+\Delta'_4)E_s}}\left\{\exp\left(-\frac{\Delta'_3+\Delta'_4}{\beta}E_s\right) + \exp\left(-\frac{(\Delta'_3-\Delta'_4)^2}{(\Delta'_3+\Delta'_4)\beta}E_s\right)\right\}$$

$$- \frac{\Delta'_5+\Delta'_4}{\sqrt{\pi\beta(\Delta'_1+\Delta'_2)E_s}}\left\{\exp\left(-\frac{\Delta'_1+\Delta'_2}{\beta}E_s\right) + \exp\left(-\frac{(\Delta'_1-\Delta'_2)^2}{(\Delta'_1+\Delta'_2)\beta}E_s\right)\right\}$$

Observe that $g(\lambda|\alpha)$ is the (unconditional) S-curve if α is a constant. If α changes rapidly from symbol to symbol, the loop filter will smooth its effect on the error signal. Therefore, the S-curve is obtained by averaging the above equations over a certain set of values of α, which is determined by the initial sampling offset and the number of samples per symbol. This problem is addressed in the subsequent discussion.

If β is an integer, the value of α remains constant from symbol to symbol, namely, $\alpha_i = \gamma$ for all i. For this case, the S-curve can be determined for a given γ. When γ is 0.5, the S-curve is centered; otherwise, the S-curve is biased slightly to one side. The phase error is certainly biased if the S-curve is biased. If β is an odd integer, the error signal is not zero when the phase error is zero. This is a source of self-noise, as discussed before.

Next, consider the effect of a noninteger number of samples per symbol on the S-curve. To illustrate the concept, assume that $\beta = 4.1$. Suppose that the initial sampling offset γ is 0.7. Clearly, $\alpha_0 = \gamma = 0.7$, $\alpha_1 = 0.6$, $\alpha_2 = 0.5$, $\alpha_3 = 0.4$, $\alpha_4 = 0.3$, $\alpha_5 = 0.2$, $\alpha_6 = 0.1$, $\alpha_7 = 0$, $\alpha_8 = 0.9$, $\alpha_9 = 0.8$, $\alpha_{10} = 0.7$, and so on. Consider a system with a normalized symbol rate of 1 Hz and a one-sided loop filter bandwidth of 0.05 Hz. The error-signal fluctuation due to variation of α is averaged by the loop filter in the same way as the fluctuation due to thermal noise. Therefore, the S-curve is obtained by averaging the error signal with respect to noise and all possible values of α. For the example given here, the set of values for α is $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Figure 5 shows the conditional S-curve for the $\beta = 4.1$ case with $\alpha = 0.5$. Figure 6 shows the unconditional S-curve for the $\beta = 4.1$ case after averaging over values of α belonging to the set $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. There is a phase ambiguity area in the conditional S-curve in Fig. 5, i.e, zero error signal for nonzero phase error. The phase ambiguity is removed in the S-curve in Fig. 6 due to averaging over α. Notice the small bias of the S-curve due to the particular set of values for α.

Another example is the $\beta = 4.11$ case. The fractional part of β, i.e., 0.11, can be decomposed into the two components 0.1 and 0.01. Both components contribute to the values of α_i. For instance, if $\alpha_0 = \gamma = 0.7$, then $\alpha_1 = 0.59 = 0.7 - 0.1 - 0.01$, and $\alpha_2 = 0.48 = 0.7 - 0.2 - 0.02, \ldots$ and so on. The same loop bandwidth is assumed as before. The variation of α due to the component 0.1 changes quickly relative to the loop bandwidth, and the variation due to the component 0.01 changes slowly relative to the loop bandwidth. The S-curve is obtained approximately by averaging the error signal over all possible values of α due to the fast component. The initial sampling offset will drift slowly from time to time due to the slow component. The slow component does not affect the instantaneous S-curve, but it does affect the S-curve gradually by changing the initial sampling offset. Therefore, the slow component does contribute to the overall phase-error density function. Note that the conditional density function for the phase error can be obtained for any given initial sampling offset. The phase-error density function can then be derived by averaging the conditional density function over the initial sampling offset. The distribution function for the initial sampling offset can be assumed to be uniform between 0 and 1 or can be determined by simulation.

Partitioning the fractional part of β into slow and fast components can only be done approximately, and they are determined by the loop bandwidth. Let $\beta = n_\beta + f_{\beta,1} + f_{\beta,2}$, where n_β is the greatest integer less than or equal to β, $f_{\beta,1}$ is the fast component, and $f_{\beta,2}$ is the slow component. The choice of $f_{\beta,1}$ and $f_{\beta,2}$ is made solely by experience. In the following discussion, a criterion is provided for justifying the choice of $f_{\beta,1}$.

Let γ be the initial sampling offset. The set of values, denoted by $D_{\beta,\gamma}$ which $\{\alpha_i\}$ can take on, can be determined from γ and β using the following procedure: suppose that $f_{\beta,1}$ contains k digits after the decimal point (for instance, if $f_{\beta,1} = 0.15$, then $k = 2$). The basic incremental unit g_β is defined as

$$g_\beta = \frac{\mathrm{GCD}(10^k f_{\beta,1},\, 10^k)}{10^k}$$

where $\mathrm{GCD}(a, b)$ is the greatest common divisor between a and b. Then $D_{\beta,\gamma}$ is given by

$$D_{\beta,\gamma} = \left\{\gamma + m g_\beta | 0 \le m \le (1/g_\beta) - 1\right\}$$

For instance, if $\gamma = 0.1$ and $f_{\beta,1} = 0.15$, then $g_\beta = 0.05$, and $D_{\beta,0.1} = \{0.1 + 0.05m | 0 \le m \le 19\}$. In terms of $D_{\beta,\gamma}$, the S-curve is given by

$$g(\lambda|\gamma) = E_{\alpha \epsilon D_{\beta,\gamma}}\left\{g(\lambda|\alpha)\right\}$$

where the expectation is performed with respect to all values of α in $D_{\beta,\gamma}$, which are assumed equiprobable. Let the one-sided loop bandwidth be B_L and let the symbol rate be R. A valid choice for $f_{\beta,1}$ is to ensure that all values in $D_{\beta,\gamma}$ can occur in a $1/B_L$ time interval, i.e., to satisfy the following equation:

$$B_L \underline{\le} R g_\beta \qquad (5)$$

B. The Density Function and Variance of Phase Error

The steady-state Fokker-Planck Eq. (2) is given by

$$\frac{dP}{d\lambda}\Big\{A(\lambda|\gamma)P(\lambda|\gamma)\Big\} = \frac{1}{2}\frac{d^2}{d\lambda^2}\Big\{B(\lambda|\gamma)P(\lambda|\gamma)\Big\} \qquad (6)$$

In the above equation,

$$A(\lambda|\gamma) = -\frac{2B_L}{E_s N_0}\, g(\lambda|\gamma)$$

$$\beta(\lambda|\gamma) = \left(\frac{2B_L}{E_s N_0}\right)^2 S(0,\lambda|\gamma)$$

where $S(\omega,\lambda|\gamma)$ is the spectrum of the noise $n_\lambda(t)$ defined as

$$n_\lambda(t) = E_{n,s,\alpha\epsilon A_{\beta,\gamma}}\{e(k)e(k+m)|\lambda,\gamma\} - \Big(g(\lambda|\gamma)\Big)^2$$

The solution to Eq. (6) is of the form

$$P(\lambda|\gamma) = C\exp\left[\int_0^\lambda \frac{2A(y|\gamma) - \frac{dB(y|\gamma)}{dy}}{B(y|\gamma)}\, dy\right] \qquad (7)$$

In order to use the above equation, $S(0,\lambda|\gamma)$ must be found, which is a fairly complex task. In the subsequent derivation, it is assumed that $S(0,\lambda|\gamma) = WN_0T/4$. Thus Eq. (7) can be simplified to

$$P(\lambda|\gamma) = C\exp\left[-\frac{2E_s}{B_L RW}\int_0^\lambda g(y|\gamma);dy\right] \qquad (8)$$

The phase-error bias $E\{\lambda|\gamma\}$ is given by

$$E\{\lambda|\gamma\} = \int_{-1/2}^{1/2} \lambda P(\lambda|\gamma)\, d\lambda$$

The mean-square phase noise $\sigma^2(\lambda|\gamma)$ is given by

$$\sigma^2(\lambda|\gamma) = E\{\lambda^2|\gamma\} - (E\{\lambda|\gamma\})^2$$

The phase ambiguity phenomenon is a direct result of Eq. (8). Note that if $g(y|\gamma) = 0$ for $-\epsilon_1 < y < \epsilon_2$, the phase error is uniformly distributed between $-\epsilon_1$ and ϵ_2 when E_s approaches infinity.

III. Discussion and Numerical Results

Figure 7 shows the phase-error variance versus symbol signal-to-noise ratio (SNR) with an even integer number of samples per symbol. Notice that the phase-error variance approaches a limit as symbol SNR increases for the given number of samples per symbol. That limit of the phase-error variance is due to the phase ambiguity; thus it cannot be eliminated by increasing the symbol SNR. The phase ambiguity decreases as the number of samples per symbol increases.

Note that the phase ambiguity phenomenon may have an effect on the performance of weighted symbol detection. For illustration, Fig. 8 shows simulation results of the MTLL of the all-digital DTTL for various symbol SNRs and for 4 and 100 samples per symbol. The plot depicts normalized MTLL, which is the MTLL times the loop bandwidth. Notice that it usually takes a long time to simulate the MTLL performance; therefore, the loop operation was purposely simulated at a very low loop SNR (on the order of 3 to 9 dB) to guarantee loss of lock within a "practical" time period. It is clear that the four samples per symbol case ($\beta = 4$) loses lock more often than the $\beta = 100$ case. In the DSN, the symbol loop SNR is so high that the loop is expected to maintain lock over a whole track. It is still expected that the MTLL for the $\beta = 4$ case will be less than for the $\beta = 100$ case, but both of these will be large enough that lock is maintained over a whole track.

In a practical communication system, Doppler and Doppler rate are present due to the relative motion between transmitter and receiver. The effect of the Earth Doppler rate on a symbol rate of 6.6 Msymbols/sec is about 1 mHz/sec, which is enough to guarantee that the number of samples per symbol will not remain an exact integer for long. Consider a scenario designed for $\beta = 4$ samples per symbol, but due to Doppler rate, the actual number of samples per symbol is $\beta = 4.0000001$. In this scenario, the basic incremental unit is $g_\beta = 10^{-7}$. When the DTTL is operating with a 1-mHz-loop bandwidth, the time constant of the loop is about 1000 sec or 6.6 Gsymbols at a symbol rate of 6.6 Msymbols/sec. Since the loop is effectively averaging over all those symbols, the effect of the 10^{-7} basic incremental unit will be enough to smooth the composite S-curve as discussed earlier. This is because with a time constant of 1000 sec, Rg_β/B_L = 6.6 Gsymbols $\times 10^{-7} = 660 \gg 1$ (Eq. 5). This effectively smooths the S-curve so that the digital loop behaves like its equivalent analog counterpart. For a loop time constant of 1 sec (1-Hz loop bandwidth), $Rg_\beta/B_L = 0.66$. However, with a time constant of 0.2 sec (5-Hz loop bandwidth),

$Rg_\beta/B_L = 0.132$; therefore, self-noise might become considerable. But that case would still exhibit less self-noise than the exact four samples per symbol scenario. So overall, the Doppler rate helps in reducing the self-noise. Depending on the actual parameters, the self-noise degradation might become negligible. More simulations with Doppler rates are planned to verify this concept.

IV. Conclusion

The all-digital DTTL with coherent or noncoherent sampling is analyzed in this article. Two sampling schemes are considered, i.e., instantaneous sampling and I&D sampling. The theory presented here is valid for both sampling schemes. The effects of few samples per symbol and of noncommensurate sampling rates and symbol rates are addressed and analyzed. The phase ambiguity problem due to a small number of samples per symbol is illustrated, and it is shown that the phase ambiguity can be alleviated when there is a noninteger number of samples per symbol and the loop filter has appropriate bandwidth. A closed-form expression for the S-curve is derived for any number of samples per symbol. Finally, the interplay between the loop bandwidth and the number of samples per symbol in the reduction of self-noise is shown.

Acknowledgments

The authors would like to thank Joseph Statman, Dr. Ted Peng, Dr. William Hurd, Dr. Kurt Ware, and especially Dr. Brooks Thomas for bringing the "filling in" viewpoint to their attention.

References

[1] D. H. Brown and W. J. Hurd, "DSN Advanced Receiver: Breadboard Description and Test Results," *TDA Progress Report 42-89*, vol. January–March 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 48–66, May 15, 1987.

[2] W. C. Lindsey and T. O. Anderson, "Digital-Data Transition Tracking Loop," *International Telemetry Conference*, Los Angeles, California, pp. 259–271, October 8–10, 1968.

[3] M. K. Simon, "An Analysis of the Steady-State Phase Noise Performance of a Digital Data-Transition Tracking Loop," *Jet Propulsion Laboratory Space Programs Summary 37-55*, vol. III, Jet Propulsion Laboratory, Pasadena, California, pp. 54–62, January 31, 1969.

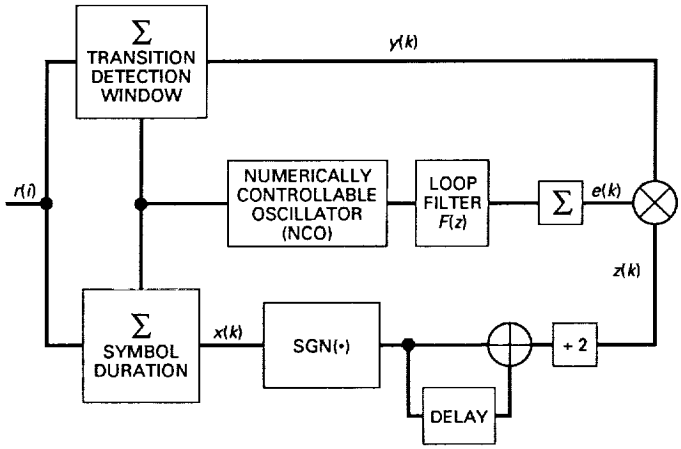


Fig. 1. Block diagram of the all-digital DTTL.

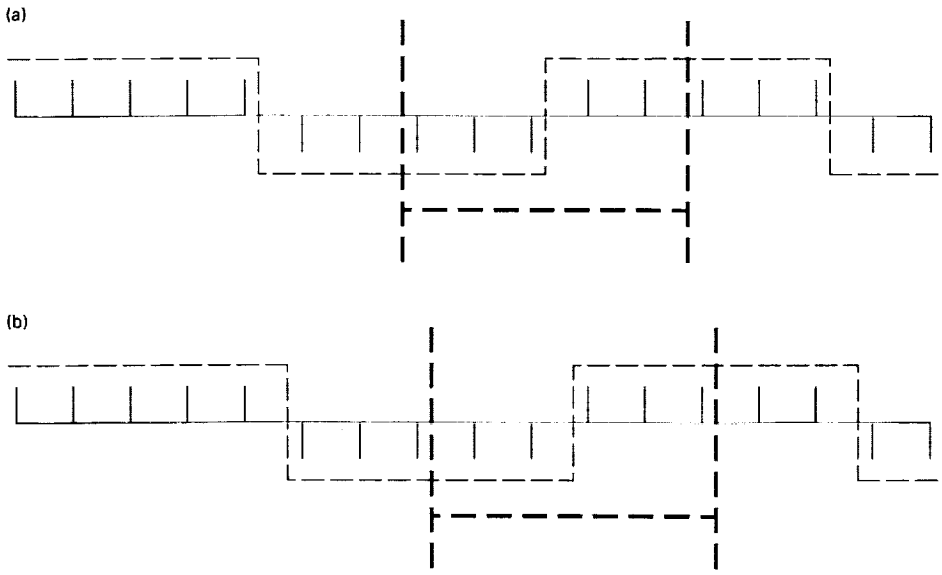


Fig. 2. Effect of odd number of samples per symbol: (a) local reference waveform is ahead of the received waveform, (b) local reference waveform is behind the received waveform.

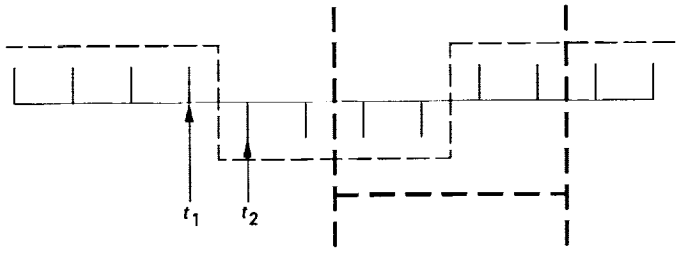


Fig. 3. The phase ambiguity phenomenon.

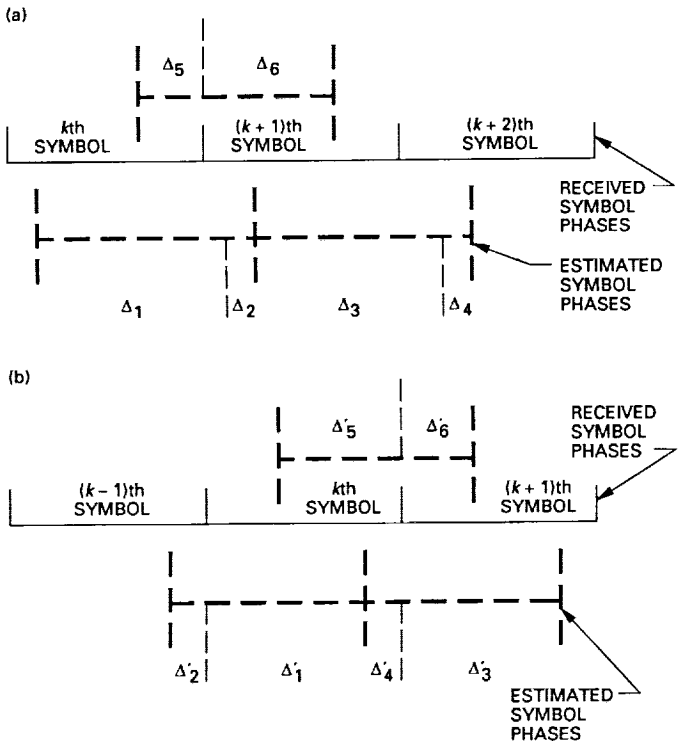


Fig. 4. The Δ functions: (a) $\lambda < 0$, (b) $\lambda \geq 0$.

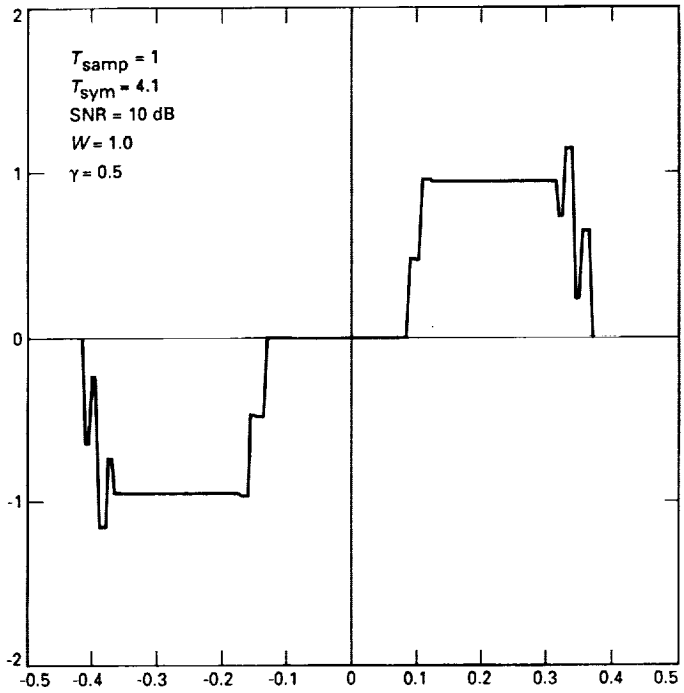


Fig. 5. Conditional S-curve for the $\beta = 4.1$ case.

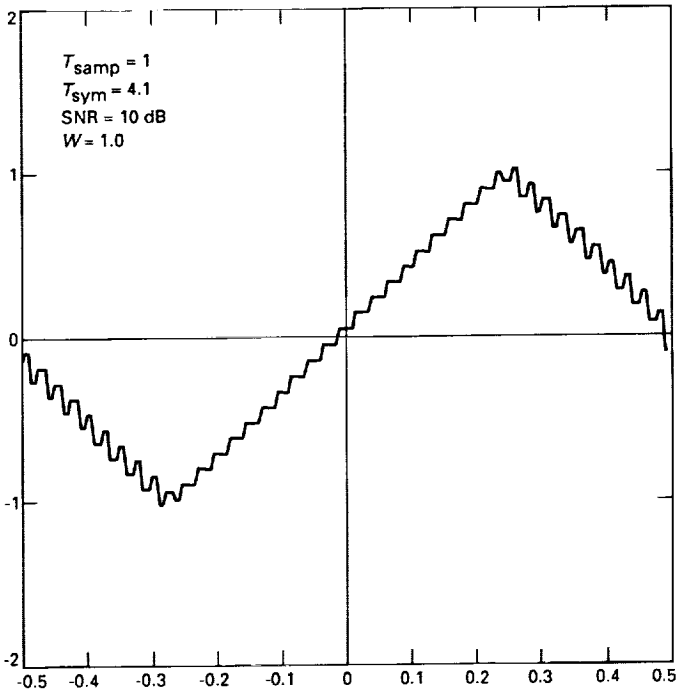


Fig. 6. S-curve for the $\beta = 4.1$ case by averaging over α belonging to the set $\{ 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 \}$.

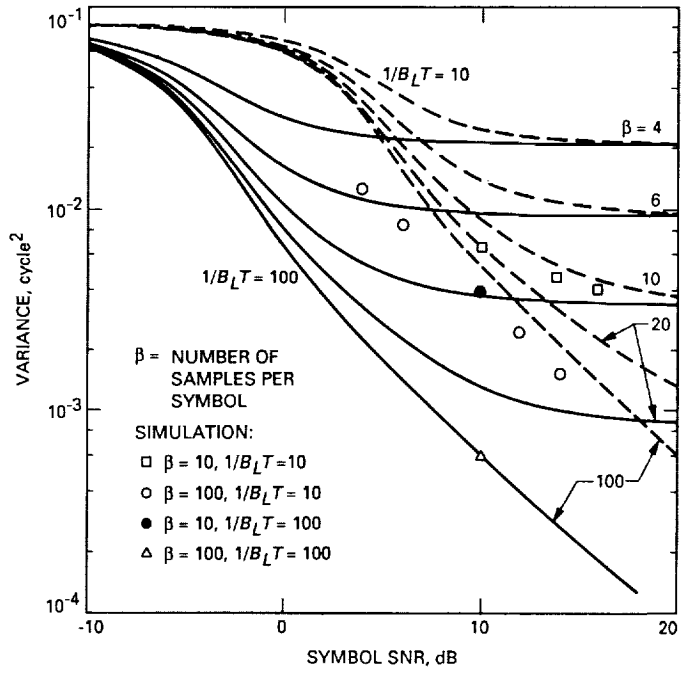


Fig. 7. Phase-error variance versus symbol SNR with even-integer number of samples per symbol.

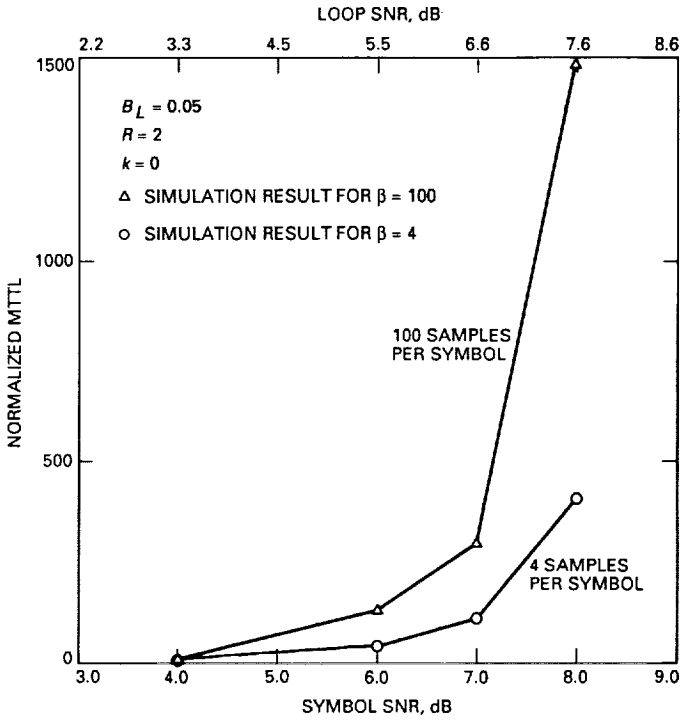


Fig. 8. Mean time to lose lock of all-digital DTTL versus symbol SNR for 4 and 100 samples per symbol.

TDA Progress Report 42-99

November 15, 1989

Costas Loop Lock Detection in the Advanced Receiver

A. Mileant
Telecommunications Systems Section

S. Hinedi
Communications Systems Research Section

The Advanced Receiver currently being developed uses a Costas digital loop to demodulate the subcarrier. Previous analyses of lock detector algorithms for Costas loops have ignored the effects of the inherent correlation between samples of the phase-error process. Accounting for this correlation is necessary to achieve the desired lock-detection probability for a given false-alarm rate. In this article, both analysis and simulations are used to quantify the effects of phase correlation on lock detection for the square-law and the absolute-value type detectors. Results are obtained which depict the lock-detection probability as a function of loop signal-to-noise ratio for a given false-alarm rate. The mathematical model and computer simulation show that the square-law detector experiences less degradation due to phase jitter than the absolute-value detector and that the degradation in detector signal-to-noise ratio is more pronounced for square-wave than for sine-wave signals.

I. Introduction

Costas loops are being used extensively in modern coherent communication systems to track both subcarriers and suppressed carriers. In many applications, residual carriers are being replaced by suppressed carriers as the latter dedicates the total transmitted power to both carrier tracking and symbol detection simultaneously. This has a clear advantage over residual carrier tracking, which requires a fraction of the total transmitted power delegated solely to that purpose, and hence, reduces the available power that can be used for symbol detection. The disadvantages of Costas loops are that they require symbol synchronization and suffer from an additional loss factor typically referred to as squaring loss, which is highly dependent on symbol energy-to-noise ratio. Squaring loss is the result of forming the product of the inphase and quadrature signals to wipe out the data modulation, in order to obtain a feedback error signal that is only a function of the instantaneous phase error [1]. Another disadvantage of suppressed carrier tracking is the issue of false lock, which occurs either as a result of accumulated delay [2] or during acquisition with frequency uncertainty greater than one-half the symbol rate [3, 4]. The latter can be detected through a false-lock indicator, as described in [3].

Lock detection is an important part of a tracking loop's operation and monitoring, as it provides an insight

into the tracking loop's behavior in real time. Lock detection basically serves as a binary indicator of whether the loop is tracking the received signal or not, and during loop start-up it also indicates whether or not the loop has acquired the phase of the signal. There are mainly two kinds of lock detectors for Costas loops, the $I^2 - Q^2$, or square-law detector, and the $| I | - | Q |$, or absolute-value detector [5]. Both have been analyzed in the past at high loop signal-to-noise ratio (SNR), which basically assumes zero phase-tracking error. At low loop SNR, the assumption of zero phase jitter becomes inadequate and results in system operating parameters that are different from their design counterparts. Thus, a new model is required which has to account for the phase jitter and the correlation between samples of the phase-error process. The latter is essential in an accurate performance prediction analysis, in order to operate the system at the desired lock-detection probability for a given false-alarm rate. The analysis of the detectors, including the phase correlation and assuming either a sinusoidal or a square wave signal, is presented in Section II. In the case of a square-wave, the analysis is general and includes any windowing operation as described in [6]. The discussion of the results and some simulated data are shown in Section III, followed by the conclusion in Section IV.

II. Lock-Detection Analysis

Suppressed carrier tracking for binary-phase shift keyed (BPSK) signals can be accomplished using a squaring loop or a Costas loop [7]. The squaring loop relies on wiping out the data by a squaring operation and tracking the resulting residual double-frequency component with a classical phase-locked loop. The squaring loop does not require symbol timing, but results in an additional noise term which becomes dominant at low SNR. On the other hand, the Costas loop implements a phase discriminator by forming the product of the inphase and quadrature signals. That too results in some degradation, commonly referred to as the squaring loss. Depending on the design, both loops can be implemented with identical performances for all practical purposes. The Costas loop has various derivatives, each approximating the maximum a posteriori (MAP) estimator at different SNRs [8]. For example, a hard-limiter can be included in the inphase arm to estimate the current symbol, and that results in less squaring loss at high SNRs.

This article is concerned with the lock detection for the all-digital IQ loop, which is also a derivative of the Costas loop with integrate-and-dump arm filters. All-digital refers to the fact that the input waveform to the loop is a sequence of samples and that the integrate-and-dump arm filters are digital accumulators. The IQ loop and square-law lock detector are depicted in Fig. 1 for the square-wave case, with the optional windowing operation on the quadrature channel. The analysis that follows will be applicable for both sinusoidal and square waves. The received waveform is digitized to produce the samples r_j, which are subsequently digitally mixed with the inphase and quadrature references. The outputs of the mixers, running at the sampling rate, are accumulated to detect the received symbols. It is assumed that there are L samples per symbol and that perfect symbol synchronization has been achieved. The accumulator outputs, now at the symbol rate, are multiplied together to wipe out the data and again accumulated to reduce the processing rate even further. The output, running at a new rate (referred to as the loop update rate), is the input to the digital loop filter which provides a frequency estimate to adjust the phase of the numerically controlled oscillator (NCO). The lock detector processes the arm accumulator outputs at the symbol rate, accumulates the result over M symbols, and provides a binary decision on the loop status.

In the I and Q branches of Fig. 1, the signals accumulated over L samples during the kth symbol interval are given by

$$x_{Ik} = d_k L\sqrt{P_D}\,\omega_k + n_{Ik} \qquad k = 1, \ldots, M$$

and (1)

$$x_{Qk} = d_k L\sqrt{P_D}\,v_k + n_{Qk} \qquad k = 1, \ldots, M$$

where

$$\omega_k = \left(1 - | u_k |\right), \quad u_k = \frac{2}{\pi}\phi_k, \quad | \phi_k | \leq \pi$$

$$v_k = \begin{cases} u_k & | \phi_k | \leq \pi W/2 \\ \mathrm{sgn}(\phi_k) W & \pi W/2 \leq | \phi_k | \leq \pi(1 - W/2) \\ 2\mathrm{sgn}(\phi_k) - u_k & \pi(1 - W/2) \leq | \phi_k | \leq \pi \end{cases} \tag{2}$$

for a square-wave subcarrier and

$$\omega_k = \cos\phi_k$$

$$v_k = \begin{cases} \sin\phi_k & | \phi_k | \leq \pi W/2 \\ \sin(\phi_k W) & \pi W/2 \leq | \phi_k | \leq \pi(1 - W/2) \\ \sin\phi_k & \pi(1 - W/2) \leq | \phi_k | \leq \pi \end{cases} \tag{3}$$

for a sine-wave subcarrier, where P_D is the average data power, d_k is the data value of the kth binary symbol (± 1 equally likely), ϕ_k is the subcarrier phase-estimation error (radians) at time k, W is the width of the window in the Q channel (i.e., the fraction of cycle of the reference signal which has nonzero value $W \leq 1$ [$W = 1$ means no window is used]), and n_{Ik}, n_{Qk} are zero-mean white Gaussian noise samples. From Eq. (1), the mean values and the variances of x_{Ik} and x_{Qk} conditioned on ϕ_k and d_k are given by

$$\mu_{Ik} = d_k L\sqrt{P_D}\,\omega_k \qquad \sigma_I^2 = L\sigma_n^2$$

and (4)

$$\mu_{Qk} = d_k L\sqrt{P_D}\,v_k \qquad \sigma_Q^2 = WL\sigma_n^2$$

where $\sigma_n^2 = N_0 B_n$ is the noise variance of a received sample, N_0 is the one-sided noise spectral density, and B_n is the Nyquist bandwidth. The equations that follow are applicable to both sine and square waves.

A. Square-Law Detector

The first algorithm considered detects the in-lock state by producing a signal that is proportional to the cosine of the phase error (in the case of a sinusoidal wave) and averaging it over several symbols before comparing it to a threshold τ. Referring to Fig. 1,

$$\sum_{k=1}^{M} y_k \gtrless \tau, \quad \text{where} \quad y_k \triangleq x_{Ik}^2 - x_{Qk}^2 \tag{5}$$

An estimation of the performance of this lock detector requires the first and second moments of x_{Ik}^2 and x_{Qk}^2 conditioned on ϕ_k, which are readily obtainable:

$$\overline{x_{Ik}^2} = \mu_{Ik}^2 + \sigma_I^2 \tag{6a}$$

$$\overline{x_{Qk}^2} = \mu_{Qk}^2 + \sigma_Q^2 \tag{6b}$$

$$\overline{x_{Ik}^4} = \mu_{Ik}^4 + 6\mu_{Ik}^2\sigma_I^2 + 3\sigma_I^4 \tag{7a}$$

$$\overline{x_{Qk}^4} = \mu_{Qk}^4 + 6\mu_{Qk}^2\sigma_Q^2 + 3\sigma_Q^4 \tag{7b}$$

where μ_{Ik}, μ_{Qk}, σ_I^2, and σ_Q^2 are given by Eq. (4). Using Eqs. (6) and (7), the variances of x_{Ik}^2 and x_{Qk}^2 become respectively

$$\text{var}(x_{Ik}^2) = 4\mu_{Ik}^2\sigma_I^2 + 2\sigma_I^4 \tag{8a}$$

and

$$\text{var}(x_{Qk}^2) = 4\mu_{Qk}^2\sigma_Q^2 + 2\sigma_Q^4 \tag{8b}$$

The mean value of y_k is obtained from Eqs. (4) and (6) in Eq. (5), namely

$$\mu_{y_k} = L^2 P_D(\omega_k^2 - v_k^2) + L\sigma_n^2(1 - W)$$

The variance of y_k, $\sigma_{y_k}^2$, will be the sum of variances of x_{Ik}^2 and x_{Qk}^2, and is obtained by using Eq. (4) in Eq. (8), i.e.,

$$\sigma_{y_k}^2 = 2L^2\sigma_n^4\left[\frac{2LP_D}{\sigma_n^2}(\omega_k^2 + Wv_k^2) + 1 + W^2\right]$$

The lock-detector signal z is the accumulation of My_k samples, which are highly correlated due to the phase-error samples. The mean value and variance of z, μ_z, and σ_z^2 are derived in the Appendix, where it is shown that

$$\mu_z = M\Big(L^2 P_D d + L\sigma_n^2(1 - W)\Big) \tag{9}$$

and

$$\sigma_z^2 = M^2L^4P_D^2(g - d^2) + 4ML^3P_D\sigma_n^2(f + hW)$$

$$+ 2ML^2\sigma_n^4(1 + W^2) \tag{10}$$

The parameters d, f, h, and g depend on the waveform type and are given by ($b \triangleq W\pi/2$)

$$d \simeq 1 - 2\left(\frac{2}{\pi}\right)^{1.5}\sigma_\phi$$

$$f = \left(1 - \frac{4}{\pi}\overline{|\phi|} + \frac{4}{\pi^2}\overline{\phi^2}\right)$$

$$h = \frac{4}{\pi^2}\overline{\phi^2}$$

$$g = 1 - \frac{8}{\pi}\overline{|\phi|} + \left(\frac{4}{\pi}\right)^2\frac{1}{M^2}\sum_{k=0}^{M-1} c(k)d(k)$$

where

$$\overline{|\phi|} \simeq \sqrt{\frac{2}{\pi}}\sigma_\phi$$

and

$$\overline{\phi^2} \simeq \sigma_\phi^2 \ \text{erf}\left(\frac{b}{\sqrt{2\sigma_\phi^2}}\right)$$

$$\sigma_\phi^2 = \left(\frac{\pi}{2}\right)^2 W\left(\frac{N_0 B_{sc}}{P_D}\right)\left(1+\frac{1}{2E_s/N_0}\right) \qquad (11)$$

for a square-wave subcarrier, and

$$d = \exp(-2\sigma_\phi^2)$$

$$f = 0.5\left(1+\exp(-2\sigma_\phi^2)\right)$$

$$h = 0.5\Big(1-\exp(-2\sigma_\phi^2)\Big)$$

$$g = \frac{1}{M^2}\sum_{k=0}^{M-1} c(k)d(k)$$

$$\sigma_\phi^2 = \left(\frac{N_0 B_{sc}}{P_D}\right) W\left(1+\frac{1}{2E_s/N_0}\right) \qquad (12)$$

for a sine-wave subcarrier. Note that Eqs. (11) and (12) specify the variance of the phase jitter, assuming the linear loop model. For example, at 15 dB of nominal loop SNR, the actual variance σ_ϕ^2 can be about 1 dB larger. Nominal loop SNR ρ is defined as $1/\sigma_\phi^2$, where σ_ϕ^2 is obtained from the linear model; B_{sc} is the one-sided noise bandwidth of the Costas loop, and E_s/N_0 is the symbol energy-to-noise ratio. The constants $c(k)$ and $d(k)$ are defined as follows:

$$c(k) = \begin{cases} M & \text{for } k = 0 \\ 2(M-k) & \text{for } k = 1,2,\ldots,M-1 \end{cases}$$

for both waveforms, but

$$d(k) \triangleq \int_{-b}^{b}\int_{-b}^{b} |\,\phi_1\,||\,\phi_2\,|\,p(\phi_1,\phi_2,\tau_k)d\phi_1 d\phi_2$$

for a square wave and

$$d(k) = \begin{cases} 0.5\Big(1+\exp(-2\sigma_\phi^2)\Big) & \text{for } k = 0 \\ \exp(-4\sigma_\phi^2)\cosh\Big(2\sigma_\phi^2 C(\tau_k)\Big) & k = 1,2,\ldots,M-1 \end{cases}$$

for a sine wave; $C(\tau_k)$ is the correlation function of the phase-error process in the tracking loop, which is assumed to be of the form given by Eq. (A-9) [9], and the second-order joint probability density function of the correlated phase process $p(\phi_1,\phi_2,\tau_k)$ is assumed as in Eq. (A-10). When the loop is in-lock and assuming high loop SNR, $\phi_k \to 0$ for all k. Hence $\omega_k^2 \to 1$ and $v_k^2 \to 0$, and the above mean value and variance of the detector's signal simplify to

$$\mu_z = M\Big(L^2 P_D + L\sigma_n^2(1-W)\Big)$$

$$\sigma_z^2 = 4ML^2\sigma_n^4\left[\frac{LP_D}{\sigma_n^2}+\frac{1+W^2}{2}\right]$$

which are true for both square-wave and sine-wave signals. Note that in the above equation the following relation holds:

$$\frac{LP_D}{\sigma_n^2} = \frac{2E_s}{N_0}$$

because $L = 2T_s B_n$ (Nyquist sampling), where T_s is the symbol time, $\sigma_n^2 = N_0 B_n$, and $T_s P_D = E_s$, the energy per symbol.

B. Absolute-Value Detector

For the absolute-value detector, the squaring operation in Fig. 1 is replaced by an absolute-value operation, and the algorithm defining the new lock detector becomes

$$\sum_{k=1}^{M} y_k \gtrless \tau, \quad \text{where} \quad y_k \triangleq |\,x_{Ik}\,| - |\,x_{Qk}\,| \qquad (13)$$

In order to estimate the performance of this lock detector, the first and second moments of $|\,x_{Ik}\,|$ and $|\,x_{Qk}\,|$ are needed, again assuming a white Gaussian noise process at

the phase-locked loop input. These moments, conditioned on ϕ_k, become

$$\overline{|x_{Ik}|} \triangleq r_{Ik}$$

$$= L\sqrt{P_D}\,\omega_k \operatorname{erf}\left(\sqrt{\frac{E_s}{N_0}\omega_k^2}\right) + \sqrt{\frac{2L}{\pi}}\,\sigma_n \exp\left(-\frac{E_s}{N_0}\omega_k^2\right)$$

$$\overline{|x_{Qk}|} \triangleq r_{Qk} \tag{14}$$

$$= L\sqrt{P_D}\,v_k \operatorname{erf}\left(\sqrt{\frac{E_s v_k^2}{N_0 W}}\right) + \sqrt{\frac{2LW}{\pi}}\,\sigma_n \exp\left(-\frac{E_s}{N_0}\frac{v_k^2}{W}\right)$$

The second moments of $|x_{Ik}|$ and $|x_{Qk}|$ are identical to the second moments of x_{Ik} and x_{Qk}, and are given by Eq. (6). The mean value of y_k follows from Eqs. (13) and (14):

$$\mu_{yk} \triangleq r_{Ik} - r_{Qk}$$

$$= L\sqrt{P_D}\left(\omega_k \operatorname{erf}\left(\sqrt{\frac{E_s}{N_0}\omega_k^2}\right) - v_k \operatorname{erf}\left(\sqrt{\frac{E_s v_k^2}{N_0 W}}\right)\right) + \sqrt{\frac{2L}{\pi}}\,\sigma_n\left(\exp\left(-\frac{E_s}{N_0}\omega_k^2\right) - \sqrt{W}\exp\left(-\frac{E_s}{N_0}\frac{v_k^2}{W}\right)\right) \tag{15}$$

and the variance of y_k is

$$\sigma_{yk}^2 = L^2 P_D\left(\omega_k^2 + v_k^2\right) + L\sigma_n^2(1+W) - \left(\overline{r_{Ik}}^2 + \overline{r_{Qk}}^2\right)$$

The lock detector's signal z is again obtained by adding My_k samples. The mean and variance of z are found by averaging the first two moments of z over the correlated phase process in the tracking loop. This is carried out in the Appendix and gives

$$\mu_z = M\mu_y$$

$$= M\left[L\sqrt{P_D}\left(\overline{\omega \operatorname{erf}\left(\sqrt{\frac{E_s}{N_0}\omega^2}\right)} - \overline{v \operatorname{erf}\left(\sqrt{\frac{E_s v^2}{N_0 W}}\right)}\right) + \sqrt{\frac{2L}{\pi}}\,\sigma_n\left(\overline{\exp\left(-\frac{E_s}{N_0}\omega^2\right)} - \sqrt{W}\,\overline{\exp\left(-\frac{E_s}{N_0}\frac{v^2}{W}\right)}\right)\right] \tag{16}$$

$$\sigma_z^2 = M\left(L^2 P_D\left(\overline{\omega^2} + \overline{v^2}\right) + L\sigma_n^2(1+W)\right) + \sum_{\substack{\text{all } i,j \\ i \neq j}}\sum\left(\overline{r_{Ii}r_{Ij}} + \overline{r_{Qi}r_{Qj}}\right) - M^2\left(\overline{r_I}^2 + \overline{r_Q}^2\right) \tag{17}$$

The bar over the product terms denotes expectation over the joint probability density function of ϕ_i, ϕ_j, assumed to be of the form given by Eq. (A-10), and r_{Ik}, r_{Qk} are defined by Eq. (14). Because no closed-form solutions for the above averaging operations are known, the averaging was done numerically. When the loop is in-lock and at high loop SNR, $\phi_k \to 0$ for all k. Hence, $\omega_k^2 \to 1$ and $v_k^2 \to 0$, and the above mean value and variance of the lock detector simplify to

$$\mu_z = M\left[L\sqrt{P_D}\operatorname{erf}\left(\sqrt{\frac{E_s}{N_0}}\right) + \sqrt{\frac{2L}{\pi}}\,\sigma_n\left(\exp\left(-\frac{E_s}{N_0}\right) - \sqrt{W}\right)\right] \tag{18}$$

$$\sigma_z^2 = M\left[L^2 P_D + L\sigma_n^2(1+W) - \left(L\sqrt{P_D}\operatorname{erf}\left(\sqrt{\frac{E_s}{N_0}}\right) + \sqrt{\frac{2L}{\pi}}\,\sigma_n \exp\left(-\frac{E_s}{N_0}\right)\right)^2 - \frac{2WL}{\pi}\sigma_n^2\right] \tag{19}$$

C. Probability of Detection and of False Indication

During subcarrier detection, each z sample is compared with a predefined threshold τ, and the lock detector decides that the loop is in-lock when $z > \tau$. It is possible that even when no signal is present, z will occasionally be larger than τ. In this case, the lock detector will mistakenly declare an in-lock condition. The probability of false indication is

$$P_f = \frac{1}{\sqrt{2\pi\sigma_{z0}^2}} \int_{\tau}^{\infty} \exp\left(-\frac{(z-\mu_{z0})^2}{2\sigma_{z0}^2}\right) dz$$

$$= \frac{1}{2}\operatorname{erfc}\left(\frac{\tau-\mu_{z0}}{\sqrt{2\sigma_{z0}^2}}\right) \tag{20}$$

where μ_{z0} and σ_{z0}^2 are the mean and variance of the lock-detector signal in the out-of-lock state, and $\operatorname{erfc}(x)$ is the complementary error function ($\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$, where $\operatorname{erf}(x)$ is the error function defined in the Appendix). For the square-law detector, μ_{z0} and σ_{z0}^2 are obtained from Eqs. (9) and (10) by making $P_D = 0$ (or, equivalently, assuming that $\overline{\omega_k}$ and $\overline{v_k}$ in Eq. 4 are zero), namely

$$\mu_{z0} = ML\sigma_n^2(1-W)$$

$$\sigma_{z0}^2 = 2ML^2\sigma_n^4(1+W^2)$$

whereas for the absolute-value detector, Eqs. (18) and (19) result in

$$\mu_{z0} = M\sqrt{\frac{2L}{\pi}}\,\sigma_n\left(1-\sqrt{W}\right)$$

$$\sigma_{z0}^2 = ML\sigma_n^2\left(1-\frac{2}{\pi}\right)(1+W) \tag{21}$$

Given a desired probability of false indication P_f, the threshold τ is obtained by solving Eq. (20) and setting it equal to

$$\tau = \sqrt{2\sigma_{z0}^2}\ \operatorname{erfc}^{-1}(2P_f) + \mu_{z0} \tag{22}$$

where $\operatorname{erfc}^{-1}(\cdot)$ is the inverse complementary error function. When the loop is in-lock, it can be argued via the central limit theorem that the random variable z is approximately Gaussian, with mean and variance obtained earlier for either the square-law or the absolute-value detector. For either detector, the probability of detection is

$$P_d = \frac{1}{\sqrt{2\pi\sigma_z^2}} \int_{\tau}^{\infty} \exp\left(-\frac{(z-\mu_z)^2}{2\sigma_z^2}\right) dz$$

$$= \frac{1}{2}\operatorname{erfc}\left(\frac{\tau-\mu_z}{\sqrt{2\sigma_z^2}}\right)$$

where μ_z and σ_z^2 are given by Eqs. (9) and (10) or by Eqs. (16) and (17). Defining the detector's SNR as

$$\mathrm{SNR}_z \triangleq \frac{\mu_z^2}{\sigma_z^2} \tag{23}$$

then, for $\mu_{z0} = 0$ ($W = 1$), the probability of detection in terms of SNR_z can be expressed as

$$P_d = \frac{1}{2}\operatorname{erfc}\left(\frac{\sigma_{z0}}{\sigma_z}\operatorname{erfc}^{-1}(2P_f) - \sqrt{\frac{\mathrm{SNR}_z}{2}}\right)$$

The above equation shows the dependence of the probability of detection on the detector's SNR. Phase jitter in the tracking loop degrades the detector's SNR by a factor D:

$$D = \mathrm{SNR}_z/\mathrm{SNR}_{z\,(ideal)} \tag{24}$$

where $\mathrm{SNR}_{z\,(ideal)}$ is the detector SNR, assuming infinite loop SNR, i.e., no phase jitter. $\mathrm{SNR}_{z\,(ideal)}$ is computed from Eq. (23) using the high-SNR expressions in Eqs. (18) and (19) for μ_z and σ_z^2. For a given W, M, loop SNR ρ ($\rho = 1/\sigma_\phi^2$ where σ_ϕ^2 is given by Eq. 11 or 12 repectively), and P_f, the detector's SNR must be increased approximately by the factor $1/D$ in order to achieve a desired probability of detection.

III. Discussion and Numerical Results

Computer simulation was performed in order to check the predictions of the analysis. Figure 2 depicts the probability of lock detection versus symbol energy-to-noise ratio E_s/N_0 for both sine-wave and square-wave signals, assuming ideal conditions, i.e., no phase jitter in the tracking loop. The square-law detector performs slightly better than the absolute-value detector for a given symbol SNR.

The degradation in detection probability for a finite-loop SNR is shown in Figs. 3(a) and 3(b) for the square-law and absolute-value detectors, respectively. The threshold

τ was set to achieve probabilities of false detection P_f of 10^{-1} and 10^{-4}, and detector SNR was set to achieve nominal probabilities of detection $\overline{P_d}$ of 0.99 and 0.90. Nominal probability refers to the case of no phase jitter in the loop. It is clear that sine waves produce less degradation than square waves, and this is true for both detection schemes.

The performance of both detectors is compared in Fig. 4 for square-wave signals only, since the difference in performance is almost negligible for sine-wave signals. The performance with respect to detector SNR is shown in Fig. 5 for a 15-dB loop SNR. The improvement in detection probability due to windowing is clear for both detectors and can result in several decibels. Finally, Fig. 6 depicts both theoretical and simulation points of detector SNR degradation D (defined in Eq. 24) versus loop SNR. The degradation in SNR is slightly larger for the absolute-value detector than for the square-law detector when tracking a square wave, but less when tracking a sine wave, and it can be as large as 3 dB depending on the operating parameters.

The results are summarized in Figs. 7 and 8. The detector SNR as a function of E_s/N_0 is shown in Fig. 7 for both infinite and 15-dB loop SNR, and for $W = 1.0$ and 0.25. When $W = 1$, a good rule of thumb is that the detector SNR varies linearly with E_s/N_0, with slope equal to 2/3 on a decibel scale. For different values of M, the curve will be scaled vertically in a linear fashion. Figure 8 depicts the detection probability for the square-law and absolute-value detectors respectively, as a function of E_s/N_0 for both infinite and 15-dB loop SNR. The conclusion from Fig. 8 is that when operating at low loop SNR (i.e., 15 dB), an extra 1.5-dB increase in E_s/N_0 or a comparable increase in M will achieve the detection probability which was designed for assuming infinite loop SNR.

For design purposes, Fig. 9 can be useful since it depicts both the detection probability and the required threshold as a function of M for both detectors. As a design example, suppose that the absolute-value detector is required to operate at $P_f = 10^{-4}$ and $P_d = 0.99$, and that the signal is a square wave with symbol rate $\tau_s = 80$ symbols per second, $E_s/N_0 = 0.0$ dB, and loop SNR = 15 dB with a quarter window ($W = 0.25$). Figure 9 indicates that at least 90 detector samples (y_k) are needed to achieve 0.99 probability of detection.

Setting $M = 100$ (integration time = 1.25 sec), τ is obtained using Eqs. (21) and (22), with $P_f = 10^{-4}$. Figure 9 predicts that τ should be set to 46, assuming that the outputs of the integrate-and-dump devices are scaled by the factor $1/\sqrt{2\sigma_n^2 L}$. Using Fig. 7 ($M = 30$), one can check that when the detector is in-lock, $\mathrm{SNR}_z \approx 13 + 10\log_{10}(100/30) = 16.3$ dB, where scaling was performed to extend the results of Fig. 7 for $M = 100$. This is confirmed in Fig. 5, which depicts $P_d = 0.99$ for $\mathrm{SNR}_z \approx 16$ dB.

IV. Conclusion

This article presents a mathematical model of the performance of two lock detectors for Costas loops: the square-law detector and the absolute-value detector. The model concentrates on the impact of phase jitter in the tracking loop on the performance of the lock detectors. Results of the analysis were verified by computer simulation and show that low loop SNRs result in a degradation in probability of lock detection, the amount of which is dependent on the scenario of interest. That decrease can be overcome by properly readjusting the design parameters. It was further shown that the square-law detector experiences less degradation due to phase jitter than the abolute-value detector, and that the degradation in detector signal-to-noise ratio is more pronounced for square-wave than for sine-wave signals.

Acknowledgments

The authors acknowledge Dr. Marvin Simon of Section 339 and Fred Krogh and William Snyder of Section 366 for their helpful discussions during the preparation of this article.

References

[1] J. Yuen, *Deep Space Telecommunications Systems Engineering*, New York: Plenum Press, 1983.

[2] W. C. Lindsey, *Synchronization Systems in Communication and Control*, New Jersey: Prentice-Hall, 1972.

[3] G. L. Hedin, J. K. Holmes, W. C. Lindsey, and K. T. Woo, "Theory of False Lock in Costas Loops," *IEEE Trans. on Comm.*, vol. COM-26, no. 1, pp. 1-12, January 1978.

[4] S. T. Kleinberg and H. Chang, "Sideband False-Lock Performance of Squaring, Fourth-Power, and Quadriphase Costas Loops for NRZ Data Signals," *IEEE Trans. on Comm.*, vol. COM-28, no. 8, pp. 1335-1342, August 1980.

[5] M. L. Olson, "False Lock Detection in Costas Demodulators," *IEEE Trans. on AES*, vol. AES-11, pp. 180-182, March 1975.

[6] W. Hurd and S. Aguirre, "A Method to Dramatically Improve Subcarrier Tracking," *IEEE Trans. on Comm.*, vol. COM-36, no. 2, pp. 238-243, February 1988.

[7] W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, New Jersey: Prentice-Hall, 1973.

[8] M. K. Simon, "On the Optimality of the MAP Estimation Loop for Carrier Phase Tracking BPSK and QPSK Signals," *IEEE Trans on Comm.*, vol. COM-27, no. 1, pp. 158-165, January 1979.

[9] J. K. Holmes, *Coherent Spread Spectrum Systems*, New York: John Wiley and Sons, 1982.

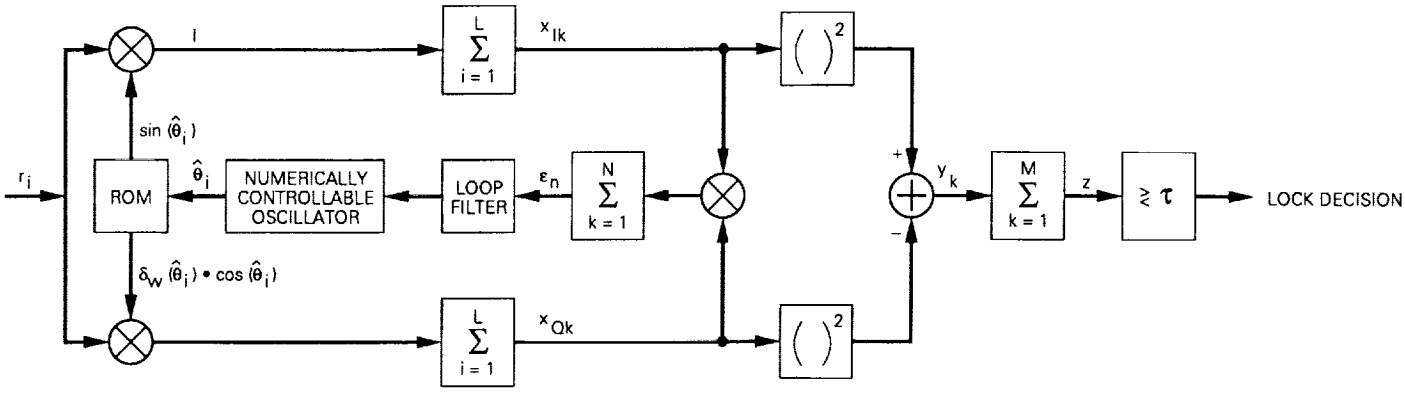


Fig. 1. Costas loop with the square-law detector.

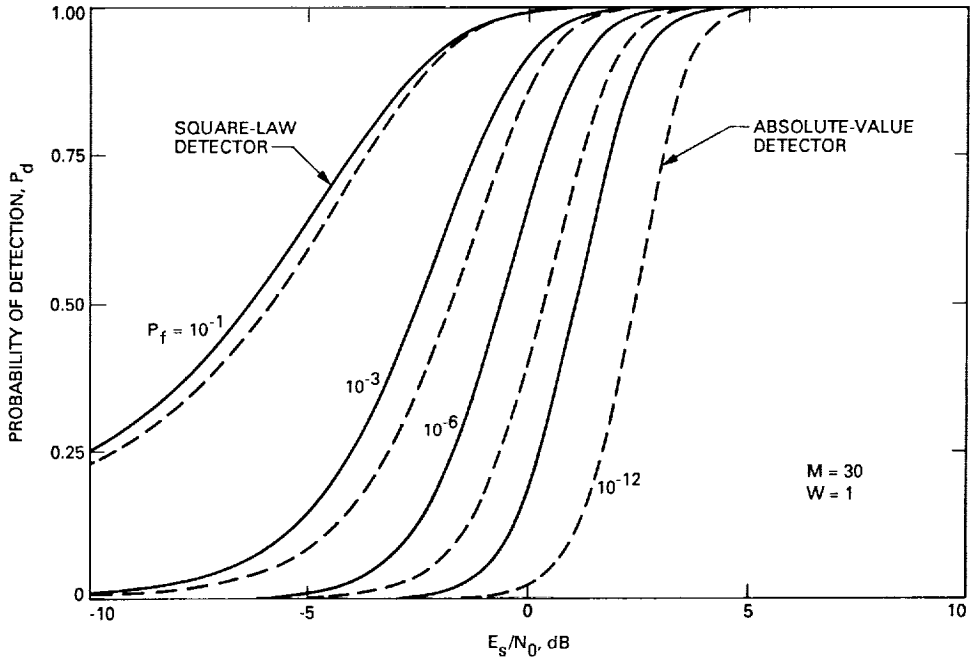


Fig. 2. Probability of detection versus E_s/N_0, infinite loop SNR, sine wave and square wave.

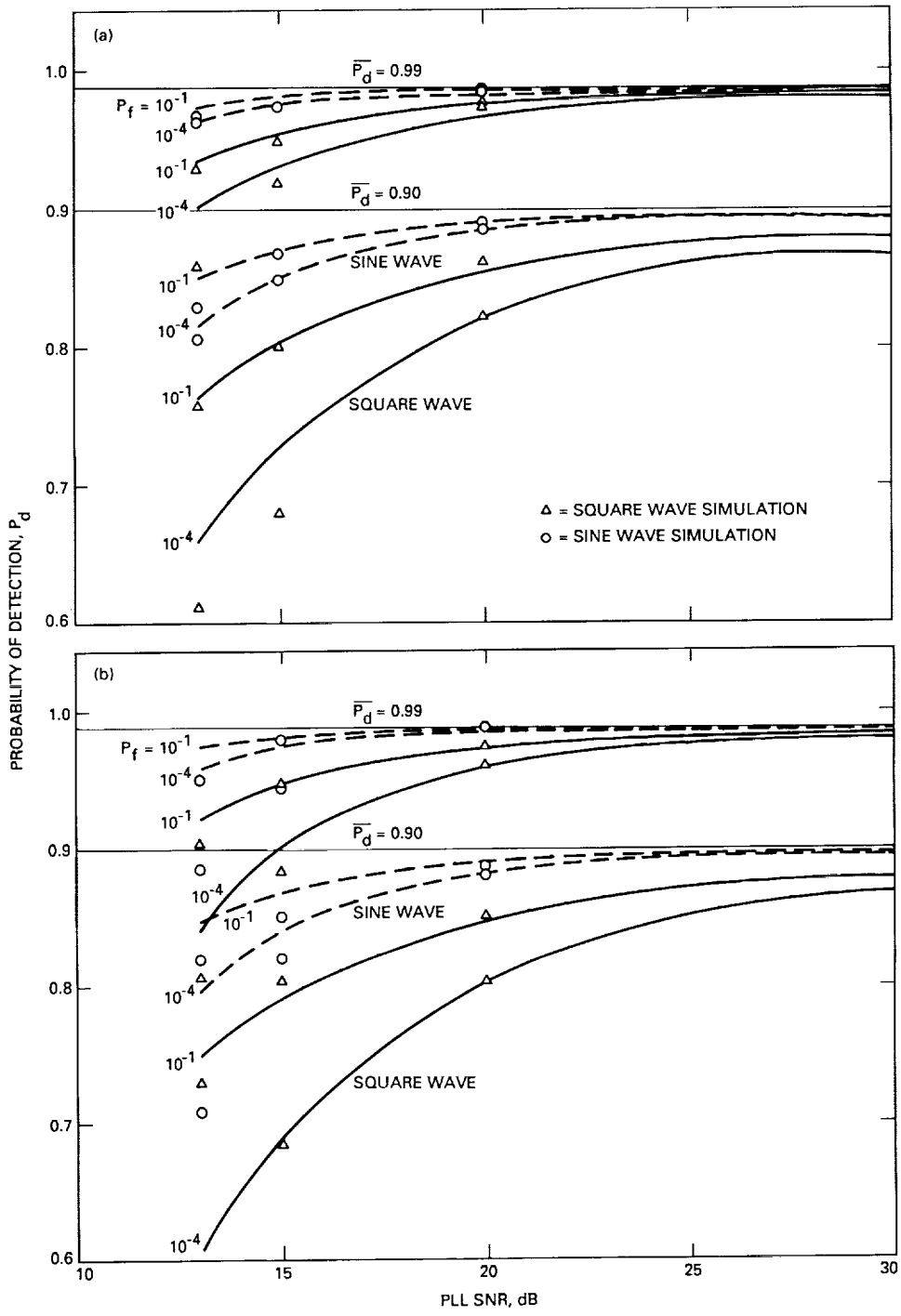


Fig. 3. Probability of detection versus loop SNR: (a) square-law detector, (b) absolute-value detector.

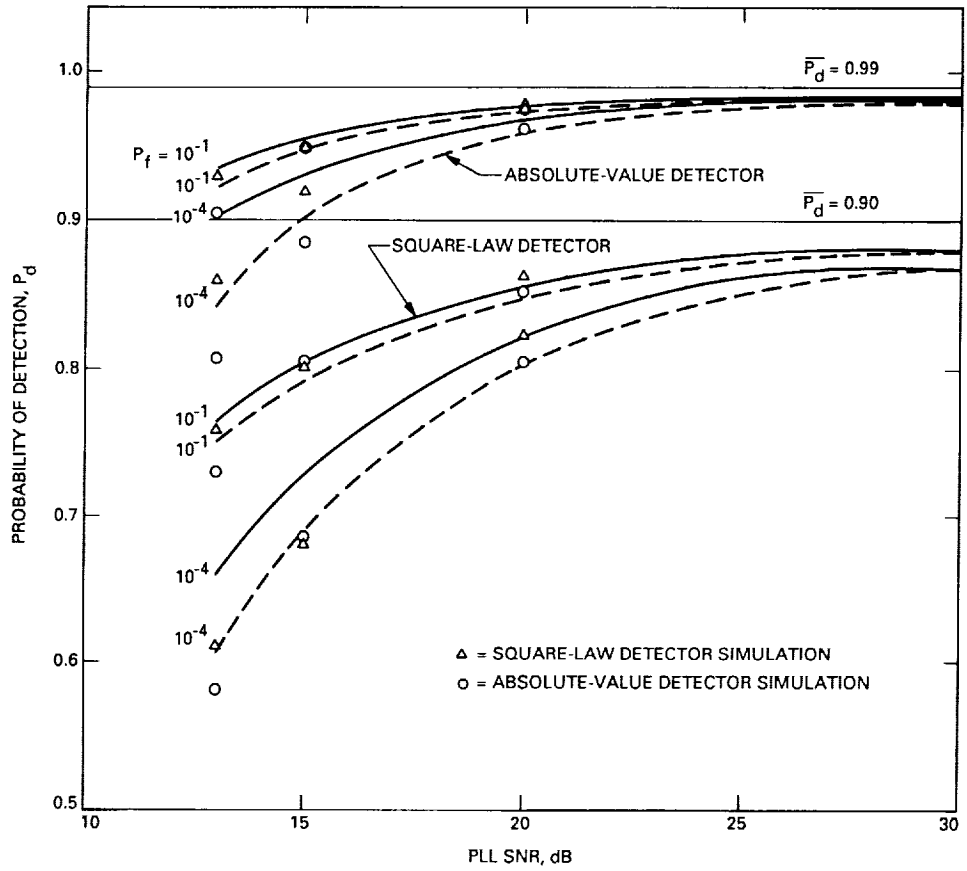


Fig. 4. Probability of detection versus loop SNR, square wave.

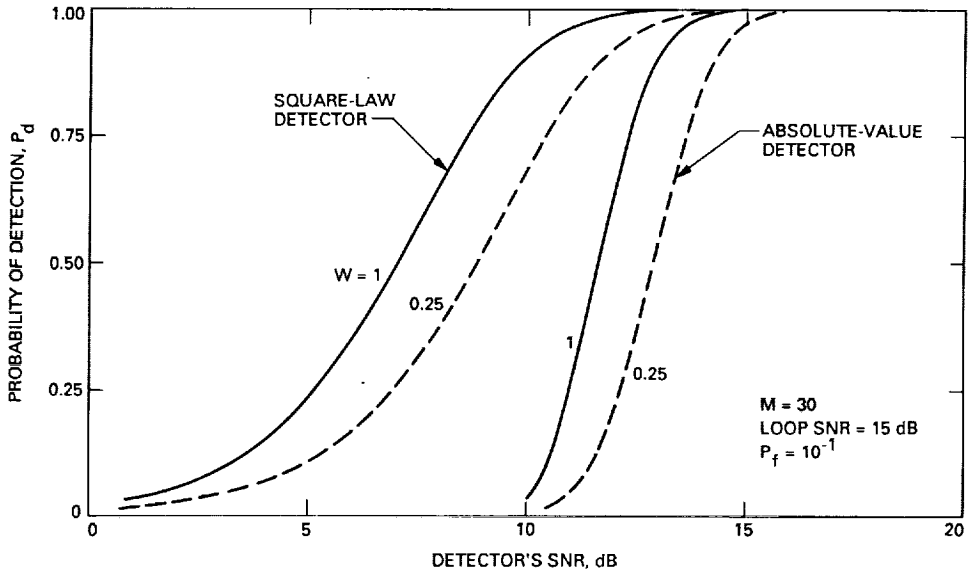


Fig. 5. Probability of detection versus detector's SNR.

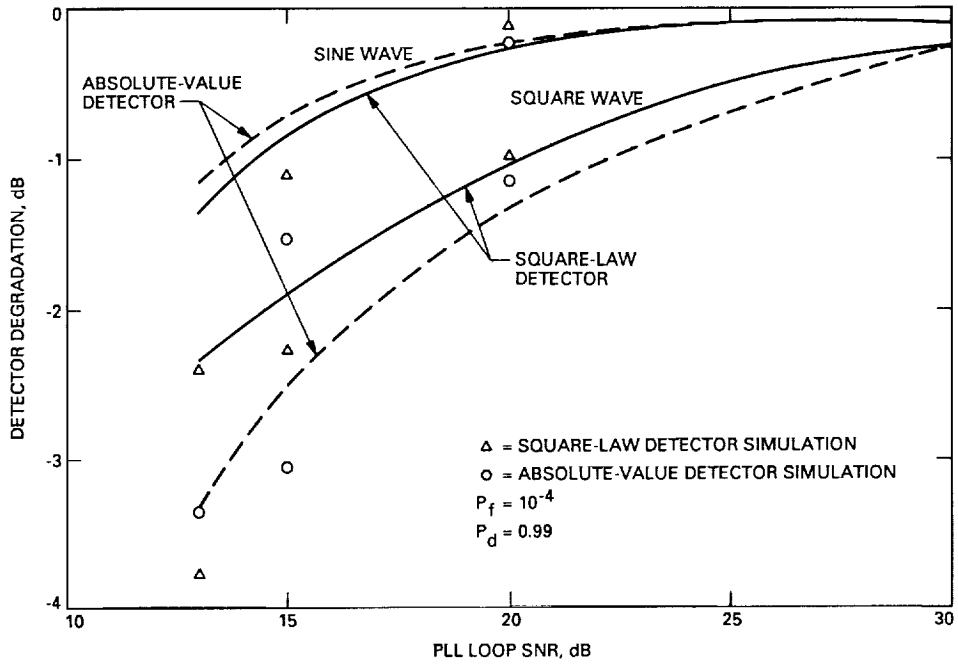


Fig. 6. Detector degradation versus phase-locked-loop SNR.

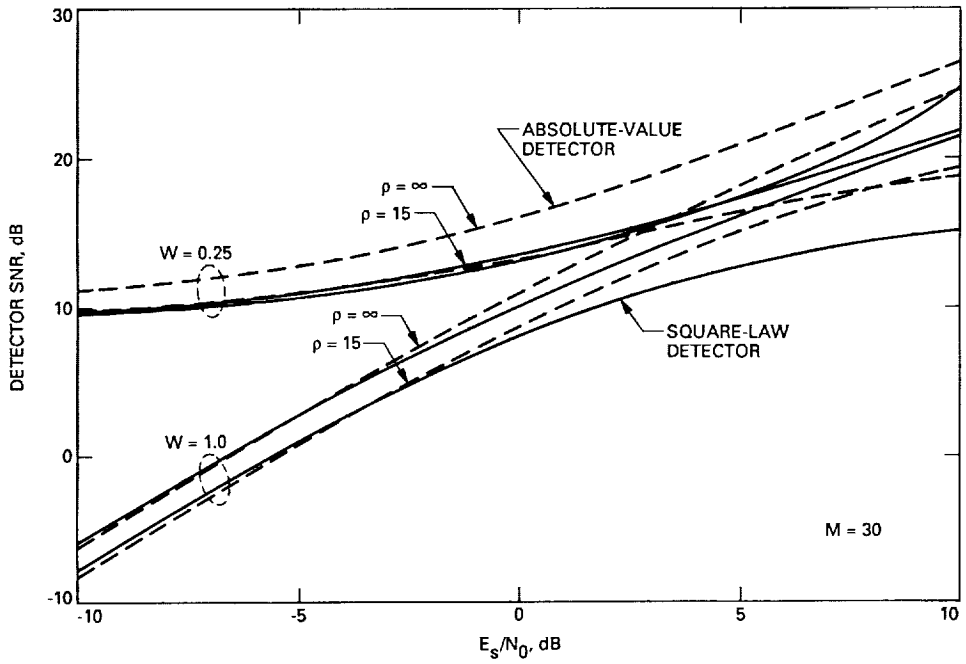


Fig. 7. Detector's SNR versus E_s/N_0.

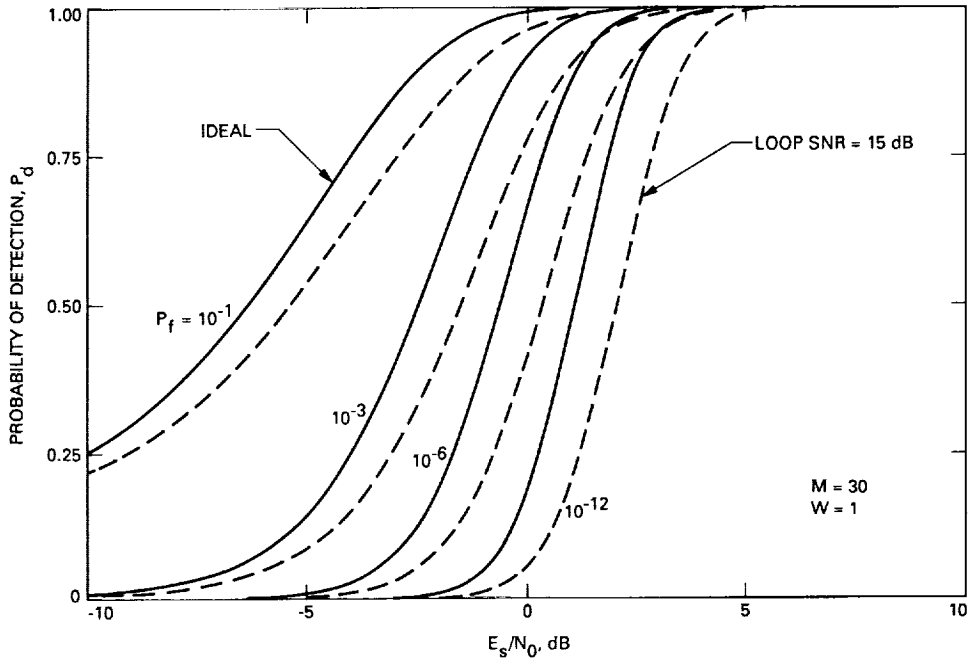


Fig. 8. Probability of detection versus E_s/N_0.

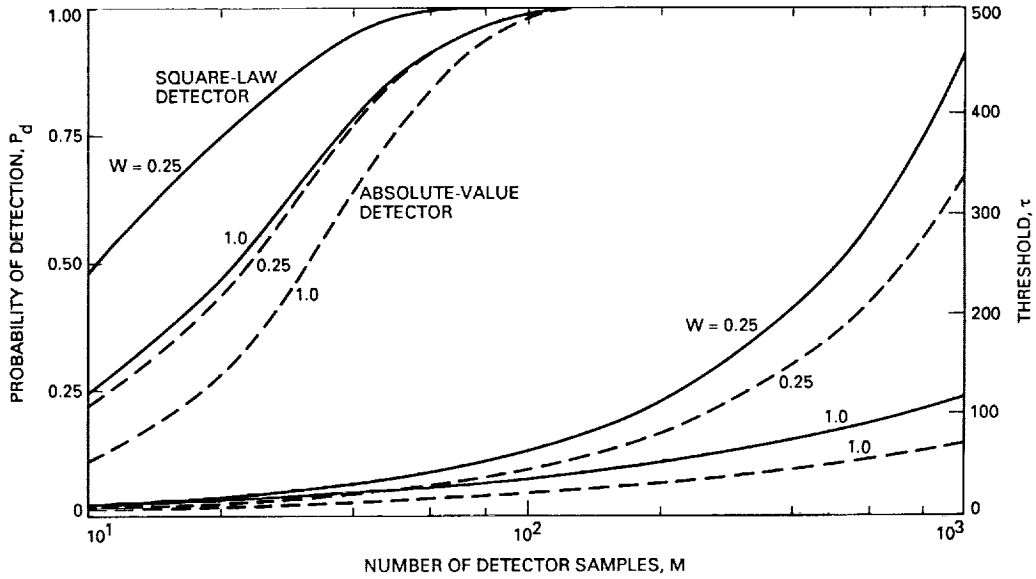


Fig. 9. Probability of detection and threshold τ versus number of detector samples *M*.

C-2

Appendix

Derivation of Detector's First Two Moments at Low Loop SNR

A. Square-Law Detector

Using Eq. (5) with Eq. (4), the expression for lock-detector signal is rewritten as follows

$$\begin{aligned} y_k &= L^2 P_D\left(\omega_k^2 - v_k^2\right) + \left(n_{Ik}^2 - n_{Qk}^2\right) \\ &\quad + 2d_k L\sqrt{P_D}\left(\omega_k n_{Ik} - v_k n_{Qk}\right) \\ &= a_k + b_k + c_k \\ z &= \sum_{k=1}^{M}\left(a_k + b_k + c_k\right) \end{aligned} \tag{A-1}$$

To assess the performance of the lock detector, the first two moments of z are needed. For a square-wave subcarrier, $\omega_k^2 - v_k^2 = 1 - \frac{4}{\pi} \mid \phi_k \mid$, and for a sine-wave subcarrier, $\omega_k^2 - v_k^2 = \cos 2\phi_k$. Assuming that ϕ is a zero-mean (no Doppler) Gaussian phase process, it can be shown that

$$\overline{\mid \phi_k \mid} = \sqrt{\frac{2}{\pi}}\sigma_\phi\left(1 - \exp\left(-\frac{b^2}{2\sigma_\phi^2}\right)\right) \tag{A-2}$$

and

$$\overline{\phi_k^2} = \sigma_\phi^2 \operatorname{erf}\left(\frac{b}{\sqrt{2\sigma_\phi^2}}\right) - \frac{2}{\pi}\sigma_\phi b \exp\left(-\frac{b^2}{2\sigma_\phi^2}\right) \tag{A-3}$$

where $b \triangleq \frac{W\pi}{2}$, σ_ϕ^2 is the variance of the phase-jitter process in the subcarrier loop, and erf (x) is the error function defined as

$$\operatorname{erf}(x) \triangleq \frac{2}{\sqrt{\pi}}\int_0^x \exp\left(-t^2\right)dt$$

At first glance, Eq. (A-3) seems to express the variance of the phase-error process as a nonlinear function of itself. This is not the case since $\overline{\phi_k^2}$ is the variance in the window of the loop, and σ_ϕ^2 is the variance integrated over the complete density. This subtle effect is due to the Gaussian assumption, which approximates a density over a finite interval by the Gaussian density which is over an infinite interval. Numerically, $\overline{\phi_k^2}$ and σ_ϕ^2 are very close for all practical values of loop SNR. Note that the moments of Eqs. (A-2) and (A-3) are independent of k. The expected value of z is

$$\mu_z = M\left(L^2 P_D d + L\sigma_n^2(1 - W)\right) \tag{A-4}$$

where d is the signal degradation factor due to phase jitter in the tracking loop, and $\overline{c_k} = 0$. Using Eq. (A-2) results in

$$d = 1 - 2\left(\frac{2}{\pi}\right)^{1.5}\sigma_\phi\left(1 - \exp\left(-\frac{b^2}{2\sigma_\phi^2}\right)\right)$$

for a square-wave subcarrier, and

$$d = \exp\left(-2\sigma_\phi^2\right)$$

for a sine-wave subcarrier. To compute the variance of z, Eq. (A-1) is used to get

$$\begin{aligned} z^2 &= A\sum_{i=1}^{M}\sum_{j=1}^{M} g_i g_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\left(n_{Ii}^2 - n_{Qi}^2\right)\left(n_{Ij}^2 - n_{Qj}^2\right) \\ &\quad + B\sum_{i=1}^{M}\left(\omega_k n_{Ii} - v_k n_{Qi}\right)^2 + C\sum_{i=1}^{M}\sum_{j=1}^{M} g_i\left(n_{Ij}^2 - n_{Qj}^2\right) \end{aligned}$$

where $A = L^4 P_D^2$, $B = 4L^2 P_D$, $C = 2L^2 P_D$, and $g_i = \left(\omega_i^2 - v_i^2\right) = \left(1 - \frac{4}{\pi} \mid \phi_i \mid\right)$ for a square-wave subcarrier. In the above equation, six terms were left out because their expected value is zero. Taking the expected value of z^2, first over the thermal noise, and second over the phase process in the tracking loop, the following is obtained:

$$\begin{aligned} \overline{z^2} &= L^4 P_D^2 M^2 g + 2ML^2\sigma_n^4(1 + W^2) \\ &\quad + M^2 L^2\sigma_n^4(1 - W)^2 \\ &\quad + 4ML^3 P_D\sigma_n^2(f + hW) \\ &\quad + 2M^2 L^3\sigma_n^2(1 - W)d \end{aligned} \tag{A-5}$$

where

$$g = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \overline{g_i g_j}$$

and f, h are given by

$$f = \overline{\omega_k^2} = \left(1 - \frac{4}{\pi} \overline{|\phi|} + \frac{4}{\pi^2} \overline{\phi^2}\right)$$
$$h = \overline{v_k^2} = \frac{4}{\pi^2} \overline{\phi^2} \tag{A-6}$$

for a square-wave subcarrier and

$$f = \overline{\omega_k^2} = \overline{\cos^2 \phi} = 0.5\left(1 + \exp(-2\sigma_\phi^2)\right)$$
$$h = \overline{\sin^2 \phi} = 0.5\left(1 - \exp(-2\sigma_\phi^2)\right) \tag{A-7}$$

for a sine-wave subcarrier. The variance of z can be found from the relation $\sigma_z^2 = \overline{z^2} - (\overline{z})^2$. Using Eqs. (A-4) and (A-5) results in

$$\sigma_z^2 = M^2 L^4 P_D^2 (g - d^2) + 4ML^3 P_D \sigma_n^2 (f + hW) + 2ML^2 \sigma_n^4 (1 + W^2) \tag{A-8}$$

Note that at high loop SNR, $g \to 1$, $d \to 1$, $f \to 1$, $h \to 0$, and Eqs. (A-4) and (A-8) reduce to

$$\mu_z = ML^2 P_D + ML\sigma_n^2 (1 - W)$$
$$\sigma_z^2 = 4ML^2 \sigma_n^4 \left[\frac{LP_D}{\sigma_n^2} + \frac{1 + W^2}{2}\right]$$

as they should. In order to evaluate g, one must know the correlation between samples of the phase-error process in the tracking loop. That was obtained by simulation and is shown in Fig. A-1. A good closed-form model is given by

$$R(\tau) = \sigma_\phi^2 C(\tau) \tag{A-9a}$$

where

$$C(\tau) = \left(1 - \frac{|B_L \tau|}{0.91}\right) \exp(-1.25 B_L \tau) \tag{A-9b}$$

and where σ_ϕ^2 is the closed-loop variance of the phase process and B_L is the one-sided loop bandwidth. In order to evaluate g, one needs to know the second-order joint density function of the phase-error process $p(\phi_i, \phi_j, \tau)$. As an approximation, it is assumed to be a two-dimensional Gaussian density specified by the means, variances, and correlation coefficient $R(\tau)$, which is obtained by simulation. Hence,

$$p(\phi_i, \phi_j, \tau) \approx \frac{1}{2\pi\sqrt{R^2(0) - R^2(\tau)}} \exp\left(-\frac{R(0)\phi_i^2 - 2R(\tau)\phi_i\phi_j + R(0)\phi_j^2}{2\left(R^2(0) - R^2(\tau)\right)}\right)$$
$$= \frac{1}{2\pi\sigma_\phi^2\sqrt{1 - C^2(\tau)}} \exp\left(-\frac{\phi_i^2 - 2C(\tau)\phi_i\phi_j + \phi_j^2}{2\sigma_\phi^2\left(1 - C^2(\tau)\right)}\right) \tag{A-10}$$

Here g can be evaluated from the following:

$$g = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \int_{-b}^{b} \int_{-b}^{b} g_i g_j p(\phi_i, \phi_j, \tau_{ij}) d\phi_i d\phi_j \tag{A-11}$$
$$\tau_{ij} = T_L(i - j) = t_i - t_j$$

Expanding Eq. (A-11) for the square-wave subcarrier results in

$$g = \left(1 - \frac{8}{\pi} \overline{|\phi|} + \left(\frac{4}{\pi}\right)^2 s\right)$$

where

$$s = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \int_{-b}^{b} \int_{-b}^{b} |\phi_1| |\phi_2| p(\phi_1, \phi_2, \tau_{ij}) d\phi_1 d\phi_2 \tag{A-12}$$

The correlation function in Eq. (A-10) is symmetric, i.e, $R(\tau_{ij}) = R(\tau_{ji})$, and depends only on the magnitude of the difference $k = |i - j|$. This property allows the double summation in Eq. (A-12) to be reduced to a single summation:

$$s = \frac{1}{M^2} \sum_{k=0}^{M-1} c(k) d(k)$$

where $c(0) = M$, $c(k) = 2(M-k)$ for $k = 1, 2, \ldots, M-1$, and

$$d(k) = \int_{-b}^{b} \int_{-b}^{b} \mid \phi_1 \mid\mid \phi_2 \mid p(\phi_1, \phi_2, \tau_k) d\phi_1 d\phi_2$$

Note that $\sum_{k=0}^{M-1} c(k) = M^2$ as it should. For $k = 0$, the above probability density function (pdf) reduces to a delta function times a zero-mean Gaussian pdf with variance σ_ϕ^2, so that

$$\overline{\mid \phi_1 \mid\mid \phi_2 \mid} = \sigma_\phi^2$$

Because

$$\overline{\mid \phi \mid} \approx \sqrt{\frac{2}{\pi}} \sigma_\phi$$

the first term of Eq. (A-8) can be rewritten as follows:

$$\sigma_{z1}^2 = M^2 L^4 P_D^2 \left(\frac{4}{\pi}\right)^2 \left(s - \frac{2}{\pi}\sigma_\phi^2\right)$$

It can be shown that the lower bound of the above equation equals

$$\sigma_{z1}^2 = M L^4 P_D^2 \left(\frac{4}{\pi}\right)^2 \left(1 - \frac{2}{\pi}\right) \sigma_\phi^2$$

Keeping all detector parameters constant, as the loop SNR decreases (i.e., σ_ϕ^2 increases), μ_z, σ_z^2, SNR_z, P_f, and P_d decrease.

For a sine-wave subcarrier,

$$g = \frac{1}{M^2} \sum_{k=0}^{M-1} c(k) d(k)$$

where $c(k)$ is the same as in the square-wave subcarrier case, and

$$d(k) = \begin{cases} 0.5\left(1 + \exp(-2\sigma_\phi^2)\right) & \text{for } k = 0 \\ \exp(-4\sigma_\phi^2) \cosh\left(2\sigma_\phi^2 C(\tau_k)\right) & k = 1, 2, \ldots, M-1 \end{cases}$$

B. Absolute-Value Detector

At low SNR in the tracking loop, the mean value of the detector's signal z is obtained by taking the expected value of y_k (Eq. 15) over the phase process in the tracking loop, and multiplying the result by M (the number of y_k samples):

$$\mu_z = M\left(\overline{r_{Ik}} - \overline{r_{Qk}}\right)$$

$$= M\left[L\sqrt{P_D}\left(\overline{\omega \operatorname{erf}\left(\sqrt{\frac{E_s}{N_0}}\omega^2\right)} - \overline{v \operatorname{erf}\left(\sqrt{\frac{E_s v^2}{N_0 W}}\right)}\right)\right.$$

$$\left. + \sqrt{\frac{2L}{\pi}}\sigma_n\left(\overline{\exp\left(-\frac{E_s}{N_0}\omega^2\right)} - \sqrt{W}\,\overline{\exp\left(-\frac{E_s}{N_0}\frac{v^2}{W}\right)}\right)\right]$$

where ω and v are defined by Eq. (2) or Eq. (3), and r_{Ik} and r_{Qk} are defined by Eq. (14). The variance of z is obtained from $\sigma_z^2 = \overline{z^2} - \overline{z}^2$, namely

$$\sigma_z^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} \overline{\mid x_{Ij} \mid\mid x_{Ij} \mid} + \sum_{i=1}^{M}\sum_{j=1}^{M} \overline{\mid x_{Qj} \mid\mid x_{Qj} \mid}$$

$$- 2\sum_{i=1}^{M}\sum_{j=1}^{M} \overline{\mid x_{Ij} \mid\mid x_{Qj} \mid} - \left(\sum_{i=1}^{M} \overline{\mid x_{Ij} \mid} - \overline{\mid x_{Qi} \mid}\right)^2$$

The following is now obtained:

$$\sigma_z^2 = M\left(L^2 P_D\left(\overline{\omega^2} + \overline{v^2}\right) + L\sigma_n^2(1+W)\right)$$

$$+ \sum_{\substack{\text{all } i,j \\ i \neq j}}\sum \left(\overline{r_{Ii} r_{Ij}} + \overline{r_{Qi} r_{Qj}}\right) - M^2\left(\overline{r_I}^2 + \overline{r_Q}^2\right)$$

where again r_I and r_Q are defined by Eq. (14). Unfortunately, closed-form solutions for most of the above equations are not obtainable and their evaluation has to be done numerically.

The double sum in the above equation equals

$$s = \sum_{k=1}^{M-1} c(k) d(k)$$

where $c(k) = 2(M-k)$ for $k = 1, 2, \ldots, M-1$, and

$$d(k) = \int_{-b}^{b} \int_{-b}^{b} \left[r_I(\phi_1) r_I(\phi_2) + r_Q(\phi_1) r_Q(\phi_2) \right] \times p(\phi_1, \phi_2, \tau_k) d\phi_1 d\phi_2$$

where the probability density function is again approximated by Eq. (A-10). An upper bound of σ_z^2 is

$$M\left(L^2 P_D + L\sigma_n^2(1+W) - \left(\overline{r_I}^2 + \overline{r_Q}^2\right)\right)$$

and a lower bound is

$$M\left(L^2 P_D(f+h) + L\sigma_n^2(1+W) - \left(\overline{r_I}^2 + \overline{r_Q}^2\right)\right)$$

where f and h are defined in Eqs. (A-6) and (A-7).

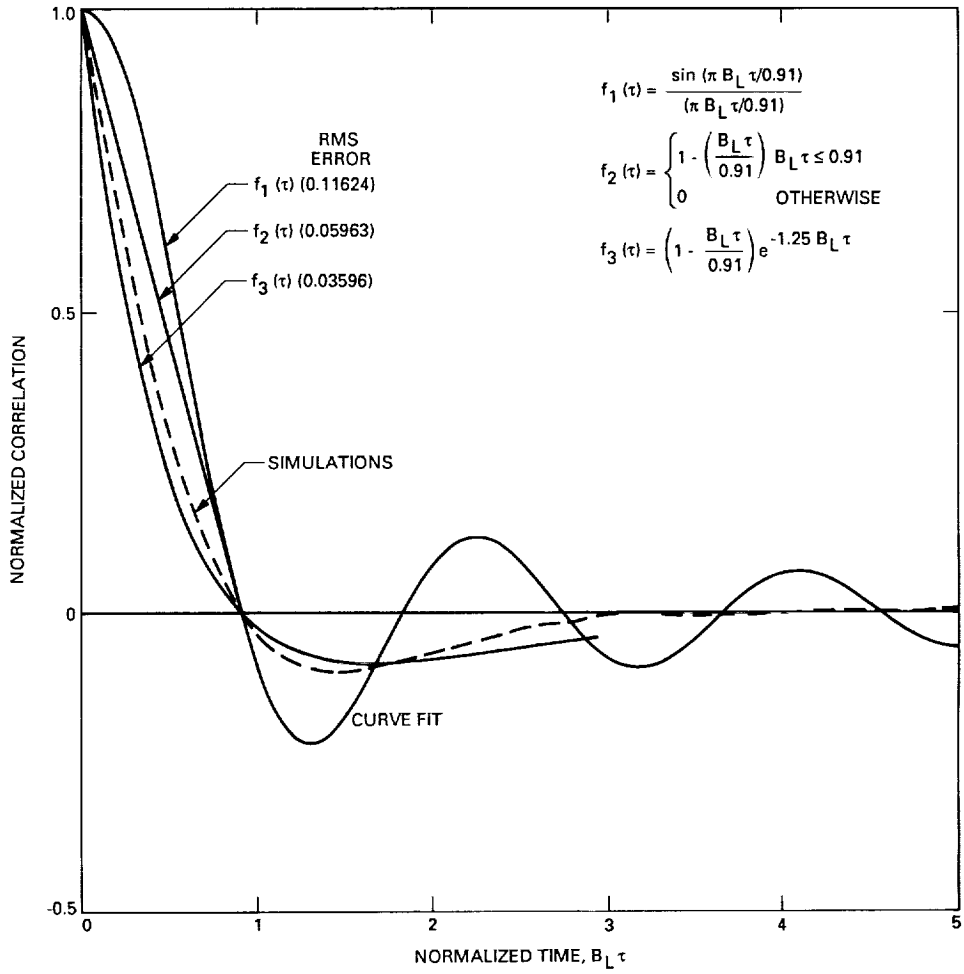


Fig. A-1. Correlation function of the phase process.

TDA Progress Report 42-99

November 15, 1989

48.

Photon Statistical Limitations for Daytime Optical Tracking

W. M. Folkner and M. H. Finger
Tracking Systems and Applications Section

Tracking of interplanetary spacecraft equipped with optical communications systems by using astrometric instruments is being investigated by JPL. Existing instruments are designed to work at night and, for bright sources, are limited by tropospheric errors. To provide full coverage of the solar system, astrometric tracking instruments must either be capable of daytime operation or be space-based. The integration times necessary for the ground-based daytime photon statistical errors to reach a given accuracy level (5 to 50 nanoradians) have been computed for an ideal astrometric instrument. The required photon statistical integration times are found to be shorter than the tropospheric integration times for the ideal detector. Since the astrometric accuracy need not be limited by photon statistics even under daytime conditions, it may be fruitful to investigate instruments for daytime optical tracking.

I. Introduction

Several observables for spacecraft navigation based on a laser telemetry system are being investigated. One such observable is the difference in direction between the spacecraft and a cataloged reference object in the same field of view of an astrometric telescope. The error for such a ground-based differential astrometric measurement includes the directional error of the reference source, the photon statistical error, and the error induced by index of refraction fluctuations in the troposphere. The best conditions for making differential astrometric measurements are at night, when the low background light levels lead to small statistical errors, and at high angles of elevation above the horizon, where the troposphere errors are smallest due to the short path length through the atmosphere.

Unfortunately, on any given day only a small portion of the solar system can be observed at high elevations during the night. In Section II, the daily available observation times for the planets are examined for the cases in which observations are restricted to nighttime and limited in elevation. All the planets suffer seasons in which no nighttime viewing is possible, corresponding to periods of low Sun-Earth-Probe (SEP) angles. While nighttime angular position measurements might augment other observable types during a fraction of a mission, critical parts of

a mission may take place at low SEP angles due to launch criteria. The restricted nighttime visibility implies that ground-based astrometric observations of laser-equipped spacecraft may have to cope with the daytime sky background.

In Section III, the photon statistical error of an astrometric measurement in the presence of a background is analyzed. In Section IV, some specific examples of the photon statistical error are presented. In Section V, a simple model of the tropospheric error is discussed. The relative size of these errors is affected by the angular separation of the reference source and the spacecraft. A larger separation is more likely to allow the use of brighter reference objects and result in a smaller photon statistical error. A small separation would require the utilization of fainter reference objects but would decrease the tropospheric error. The smallest usable field of view is set by the density of the catalog stars. It was expected that in the mid 1990s the Hipparchos mission would provide a catalog with an average of 2.5 stars per 1-degree by 1-degree field with initial accuracy of 10 nrad per component and proper motion uncertainty of 10 nrad/year [1]. The average brightness of reference sources in a 1-degree by 1-degree field is magnitude $m_v = 8$ [2]. Using magnitude $m_v = 8$ and 1 degree for the source-spacecraft separation, it is found that the integration times for daytime observations with 50 nrad accuracy are about 30 minutes.

For the error models used here, the tropospheric error for such measurements is larger than the photon statistical error. However, the daytime photon signal-to-noise ratio is very unfavorable. Existing astrometric instruments, designed for nighttime operation, are not able to work in this regime. Limitations imposed by real detectors or by systematic effects such as sky brightness variation may be much larger than the photon statistical error.

II. Nighttime Planetary Viewing Limitations

To demonstrate what fraction of a mission trajectory would be inaccessible to nighttime astrometric measurements, the number of minutes per day that each planet is visible from Goldstone in the night sky has been computed. Here "night" has been defined as the time that the sun is below -15 degrees elevation. This value for sun elevation was chosen to correspond roughly to the time of astronomical darkness. Since the astrometric error depends on elevation angle through the elevation dependence of the troposphere, two different minimum elevation cutoffs for the planets were included.

Figures 1–7 are plots of the number of minutes per day that Venus, Mars, Jupiter, and Saturn are above 10 degrees or 30 degrees elevation with the sun below -15 degrees for the time span of 1990 to 2000. Mercury is almost never visible in the dark sky and is not plotted. Venus is always at low elevation, and nighttime astrometric tracking is possible less than half of each year. The outer planets are unavailable for 25 percent or more of each year. These figures suggest that visibility limitations will severely affect the utility of astrometric tracking if it is limited to the nighttime hours.

III. Photon Statistics of an Astrometric Telescope

This section presents an estimate of the accuracy limitation placed by photon statistics on an astrometric telescope in the daytime. This category includes Ronchi-ruled telescopes and charge-coupled device (CCD) instruments. An ideal detector is able to record the position on the focal plane for each detected photon. Any real instrument will have larger photon statistical errors than this ideal instrument. CCD instruments are capable of dividing the image of a point source into many pixels and can approach the ideal detector scheme. However, limits on the size of CCD arrays limit the fields of view of such instruments. In a Ronchi telescope, a moving ruling is used to modulate the light incident on the detector with position information derived from the detected modulation [3]. For present designs, the field of view must be larger than the image of the star. This makes Ronchi telescopes more susceptible to background light problems than CCD instruments.

The ideal astrometric telescope tracks a source for a time T and records the plane-of-sky coordinates $(\xi, \eta) = \vec{\Omega}$ for each photon detected. The troposphere and instrument resolution cause the photons to be smeared in the plane of the sky. For simplicity, the photon spatial distribution about the true source direction is assumed to have the Gaussian form

$$I(\vec{\Omega} - \vec{\Omega}_s) = BT + \frac{ST}{2\pi\sigma^2} \exp\left[-\frac{(\vec{\Omega} - \vec{\Omega}_s)^2}{2\sigma^2}\right] \quad (1)$$

where B is the number of background photons per unit solid angle per unit time, $\vec{\Omega}_s$ is the true source direction, and S is the number of signal photons per unit time. This distribution is a reasonable approximation to a smeared point source in a uniform background and has been successfully used in fitting high-precision small-field images [4]. The width of the Gaussian is determined by the ap-

parent diameter of a point source. This apparent diameter as determined by the turbulence of the atmosphere is commonly called the seeing angle. The full width at half maximum of the Gaussian (2.35σ) is here set equal to the seeing angle.

Measurement of the source direction consists of recording the arrival of n photons and the plane-of-sky coordinates for each detected photon. The maximum-likelihood method can be used to estimate the source direction. The likelihood function $L(\vec{\Omega}_a)$ is the probability of recording n photons with the set of arrival directions $\{\vec{\Omega}_k\}$ given an assumed source direction $\vec{\Omega}_a$.

The photons are distributed in time according to Poisson statistics. The expected number of detected photons is given by

$$N = \int I(\vec{\Omega}) d^2\Omega \tag{2}$$

where the integral is taken over the field of view of the telescope. The probability of detecting n photons is then given in [5] as

$$P(n) = \frac{N^n e^{-N}}{n!} \tag{3}$$

The (normalized) probability of any one photon arriving with direction $\vec{\Omega}_k$ given that the source is at direction $\vec{\Omega}_a$ is

$$P(\vec{\Omega}_k|\vec{\Omega}_a) = \frac{I(\vec{\Omega}_k - \vec{\Omega}_a)}{N} \tag{4}$$

Assuming that the distribution for each photon-arrival direction is independent, the probability of finding n photons with the set of photon directions $\{\vec{\Omega}_k\}$ is

$$L(\vec{\Omega}_a) = \left(\frac{N^n e^{-N}}{n!}\right) \prod_{k=1}^{n} \frac{I(\vec{\Omega}_k - \vec{\Omega}_a)}{N} \tag{5}$$

The maximum-likelihood estimate of the source direction is the assumed source direction that maximizes the likelihood function. In the limit $N \gg 1$ the error associated with this estimate is given by the Konig-Kramer bound [6]

$$\frac{1}{\sigma_\xi^2} = \left\langle -\frac{\partial^2 \log L(\vec{\Omega}_a)}{\partial \xi_a^2} \right\rangle \Bigg|_{\vec{\Omega}_s} \tag{6}$$

where σ_ξ^2 is the variance of the source direction estimate in the ξ direction, and the expectation value is the average taken over all possible numbers of detected photons and sets of directions. Since the assumed photon distribution is symmetric, the error is the same for the η direction. With some work, the error expression of Eq. (6) can be applied to the likelihood function given in Eq. (5) to give the error in the source direction measurement as

$$\sigma_\xi^2 = \frac{\sigma^2}{ST} f\left(\frac{2\pi\sigma^2 B}{S}\right) \tag{7}$$

where the function $f(\alpha)$ is given by

$$f(\alpha) = \left[\frac{1}{2}\int_0^\infty \frac{x^3 e^{-x^2}}{\alpha + e^{-x^2/2}}\, dx\right]^{-1} \tag{8}$$

In the limit of no background ($B = 0$), $\alpha = 0$, and $f(0) = 1$; the directional error is then given by

$$\sigma_\xi = \frac{\sigma}{\sqrt{ST}} \tag{9}$$

Rewriting Eq. (9) gives the integration time to reach a specified accuracy σ_ξ as

$$T = \frac{\sigma^2}{\sigma_\xi^2 S} \tag{10}$$

In the limit of high background, the position error is given by

$$\sigma_\xi = \frac{\sigma}{\sqrt{ST}} \sqrt{\frac{8\pi\sigma^2 B}{S}} \tag{11}$$

Rewriting Eq. (11) gives

$$T = \left(\frac{\sigma^2}{\sigma_\xi^2 S}\right)\left(\frac{8\pi\sigma^2 B}{S}\right) \tag{12}$$

The effective signal-to-noise ratio for determining the source direction is $S/(8\pi\sigma^2 B)$ where the combination $8\pi\sigma^2 B$ is the effective background rate.

IV. Examples of Photon Statistical Integration Times

In this section, the integration times needed to reach a given accuracy for a reference star and for a spacecraft at 10 AU with a 2-W laser are given. The parameters assumed for the detector, the spacecraft, and the background

are summarized in Table 1. The derived photon rates and integration times are listed in Table 2.

The photon rate from a star of visual magnitude m_v is given by [2, 7] as approximately

$$S = 10^{-0.4m_v - 7.45} \frac{\mathrm{W}}{\mu\mathrm{m}\, m^2} \left(\frac{\lambda}{hc}\right) \Delta\lambda \frac{\pi}{4} d_r^2 \eta_a \eta_{ra} \eta_{ro} \eta_f \eta_d \tag{13}$$

where λ is the central wavelength, h is Planck's constant, c is the speed of light, $\Delta\lambda$ is the wavelength pass band, d_r is the telescope diameter, η_a is the atmosphere transmission factor, η_{ra} is the receiver obscuration factor, η_{ro} is the receiver optics efficiency, η_f is the narrow-band filter efficiency, and η_d is the detector quantum efficiency. Since refraction will cause image smearing if the wave band is too wide, a value $\Delta\lambda = 0.1$ μm, centered at $\lambda = 0.532$ μm, will be used below. For the receiver efficiencies given in Table 1 and for a star of magnitude $m_v = 8$ the signal rate is 6.8×10^5 photons/second.

The photon rate received from the spacecraft is given by [7] as

$$S = P_l \left(\frac{\lambda}{hc}\right) \left(\frac{\pi d_t}{\lambda}\right)^2 \left(\frac{\lambda}{4\pi r}\right)^2 \left(\frac{\pi d_r}{\lambda}\right)^2 \times \eta_{ta} \eta_{to} \eta_{tp} \eta_a \eta_{ra} \eta_{ro} \tag{14}$$

where P_l is the transmitted laser power, λ is the wavelength, d_t is the transmitter objective diameter, r is the spacecraft-receiver distance, η_{ta} is the transmitter obscuration factor, η_{to} is the transmitter optics efficiency, and η_{tp} is the transmitter pointing efficiency. The parameters used here are taken from an example by Kerr [8] and listed in Table 1. These factors combined with the nominal receiver used above yield a detected photon rate S of 2.6×10^4 photons/second for the spacecraft.

A critical parameter for the tracking telescope is the atmospheric seeing. At many sites, the daytime seeing is worse than the nighttime seeing by a factor of 5 to 10 [9, 10]. However, at some solar observatory sites, the seeing is 5 to 10 μrad ($\sim$1 to 2 arcseconds) both day and night. At the Sacramento Peak site, the daytime seeing angle is reported as being less than 10 μrad 80 percent of the time and better than 5 μrad 50 percent of the time during the day [11]. Since 10 μrad is similar to other reported daytime seeing values (see [11] and references therein), 10 μrad full width at half maximum will be taken as a nominal value. This corresponds to a value $\sigma = 4.3$ μrad for the photon distribution.

The background photon rate is given by

$$(8\pi\sigma^2)B = (8\pi\sigma^2)\, I \left(\frac{\lambda}{hc}\right) \Delta\lambda \frac{\pi}{4} d_r^2 \eta_{ro} \eta_{ra} \eta_f \eta_d \tag{15}$$

where I is the background spectral irradiance. The daytime background spectral irradiance depends on the weather, the sun elevation, and the Sun-Earth-Probe angle among other factors. The value for I of 100 W/(μm m^2 steradian) at 0.532 μm [12] will be used in the following example. Kerr gives values ranging from 30 W/(μm m^2steradian) at 90 degrees from the sun to 500 W/μm m^2 steradian at 10 degrees to the sun [8]. Using the receiver values of Table 1 and 2-arcsec seeing, the daytime background rate is 2.6×10^9 photons/second. Using a narrower filter will not improve the signal-to-noise ratio for the reference star since the signal photon rate depends on the bandwidth in the same way as the background; narrowing the filter bandwidth reduces the signal rate and increases the needed integration time. However, the laser signal is narrow band and narrowing the filter bandwidth improves the spacecraft signal-to-noise ratio.

Table 2 presents the integration times needed to reach 50-nrad or 5-nrad accuracy for the reference star and the spacecraft. Several different filter bandwidth options have been used. For the case of the 0.03-nm filter, the filter transmission efficiency is reduced to 0.4 [8]. The integration times for the photon statistics error to reach the 5-nrad level are not prohibitively long provided that the spacecraft and star are separately filtered. However, several factors may combine to lengthen the integration times. The photon rates for the reference star plus background are very high for photon counting devices. Detector dead time and noise will be significant effects. A Ronchi telescope integrates the background over a fixed field of view. Unless the field of view is limited to the size of the source image and precisely positioned, the photon statistics error will be much worse for a Ronchi telescope. In any case, the ruling modulation reduces the photon flux by a factor of 2, correspondingly increasing the integration time. If each coordinate is measured separately, the times will double again.

Effects other than counting problems are important. Since the integration time increases as the fourth power of the seeing, the site selection is critical. For the reference star, the signal-to-noise ratio may be worse than 10^{-3}. This may pose unrealistic requirements on the dynamic range and linearity of the detector. There is also the possibility that variations in the background intensity

over the period of integration could degrade the accuracy of the star position.

V. Tropospheric Integration Times

The photon statistics contribute an angular position error for the spacecraft and the reference object. The error in the angular difference between the spacecraft and reference object is also affected by variations in angle of arrival imposed by the troposphere. The integration time requirements imposed by the troposphere may be computed from a result by Lindegren [13]. Lindegren's result is derived from a frozen turbulence model of the atmosphere with a power law structure function. The turbulence causes angle of arrival variation as the frozen turbulence moves with the wind velocity. The variations in angle of arrival from two point sources are computed, given some assumptions about the atmosphere structure and averaging over wind direction. Lindegren's paper compares his model to several experimental results, including stellar position and solar diameter measurements, with reasonable agreement.

The expression for the error σ_θ in the difference angle θ between two sources is

$$\sigma_\theta = 6.3 \times 10^{-6} \theta^{\frac{1}{2}} T^{-\frac{1}{2}} \tag{16}$$

where σ_θ and θ are given in radians and the integration time T in seconds. Solving Eq. (16) for the integration time necessary for a given angular accuracy σ_θ gives

$$T = 4.0 \times 10^{-11} \theta^{\frac{1}{2}} \sigma_\theta^{-2} \tag{17}$$

This expression applies near zenith and with about 1-arcsec seeing. Results are expected to be worse for lower elevation angles and worse seeing.

For two objects 1 degree apart, this implies an integration time of 59 hours to reach 5-nrad accuracy and 35 minutes to reach 50-nrad accuracy. These times are long compared to those imposed by the photon statistics for a magnitude 8 star. There is a possible trade-off by narrowing the field of view and utilizing fainter reference stars. However, this implies a larger catalog effort. By including more reference objects in the field, there is the possibility of reducing the troposphere error.

VI. Discussion

The photon statistics do not rule out the operation of astrometric telescopes for daytime optical tracking of spacecraft. Given the simple assumptions used in Section III, the tropospheric error dominates the photon statistical error for a magnitude $m_v = 8$ star and 1-degree source-spacecraft separation. However, existing instruments are not capable of operating in the daytime because of the extremely poor signal-to-noise ratio and large photon fluxes. Finding a suitable detection scheme for daytime operation will be a challenge. Since the integration times required are many minutes, systematic effects and variations in the background level may wash out the position signal. It would be useful to try differential position measurement in the daytime at a solar observatory with a CCD camera to begin an investigation into systematic background effects.

References

[1] J. Kovalevsky, "Prospects for Space Stellar Astrometry," *Space Science Review*, vol. 39, pp. 1–63, 1984.

[2] B. L. Schumaker, "Apparent Brightness of Stars and Lasers," *TDA Progress Report 42-93*, vol. January–March 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 111–130, May 15, 1988.

[3] G. D. Gatewood, "The Multichannel Astrometric Photometer and Atmospheric Limitations in the Measurement of Relative Positions," *Astron. J.*, vol. 94, pp. 213–224, 1987.

[4] D. G. Dahn and C. C. Monet, "CCD Astrometry. I. Preliminary Results from the KPNO 4-m/CCD Parallax Program," *Astron. J.*, vol. 88, pp. 1489–1507, 1983.

[5] P. R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*, San Francisco: McGraw-Hill, 1969.

[6] M. Fisz, *Probability Theory and Mathematical Statistics*, New York: John Wiley and Sons, 1963.

[7] W. K. Marshall and B. D. Burk, "Received Optical Power Calculations Link Performance Analysis," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 32–40, November 15, 1986.

[8] E. L. Kerr, "Fraunhofer Filters to Reduce Solar Background for Optical Communications," *TDA Progress Report 42-87*, vol. July–September 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 48–55, November 15, 1986.

[9] E. S. Barker, "Site Testing with an Acoustic Sounder at McDonald Observatory," *Identification, Optimization, and Protection of Optical Telescope Sites*, R. L. Millis, O. G. Franz, H. D. Ables, and C. C. Dahn (eds.), Lowell Observatory, pp. 49–57, 1987.

[10] D. A. Erasmus, "Identification of Optimum Sites for Daytime and Nighttime Observations at Mauna Kea Observatory," *Identification, Optimization, and Protection of Optical Telescope Sites*, R. L. Millis, O. G. Franz, H. D. Ables, and C. C. Dahn (eds.), Lowell Observatory, pp. 86–93, 1987.

[11] P. N. Brandt, H. A. Mauter, and R. Smartt, "Day-time seeing statistics at Sacramento Peak Observatory," *Astron. Astrophys.*, vol. 188, pp. 163–168, 1987.

[12] W. K. Pratt, *Laser Communication Systems*, New York: John Wiley and Sons, 1969.

[13] L. Lindegren, "Atmospheric Limitations of Narrow-field Optical Astrometry," *Astron. Astrophys.*, vol. 89, pp. 41–47, 1980.

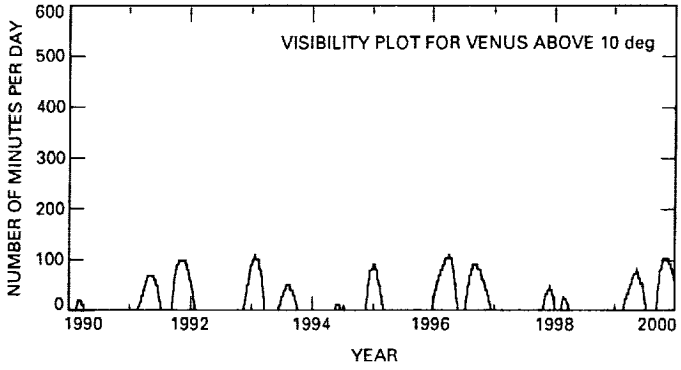


Fig. 1. Plot of the number of minutes per day (24 hours) for which Venus is visible above 10 degrees in the nighttime sky.

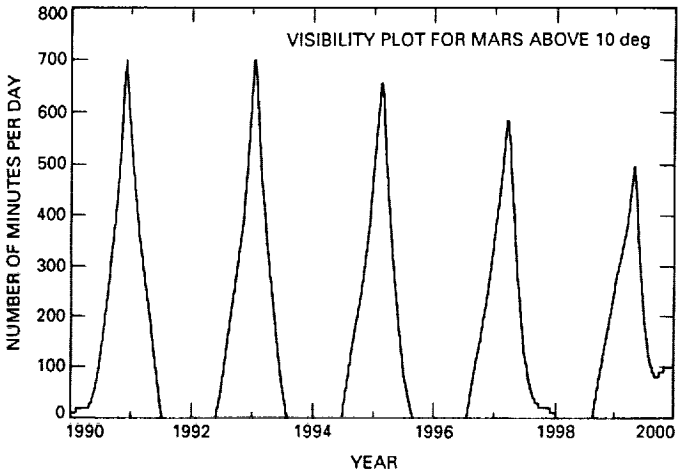


Fig. 2. Plot of the number of minutes per day (24 hours) for which Mars is visible above 10 degrees in the nighttime sky.

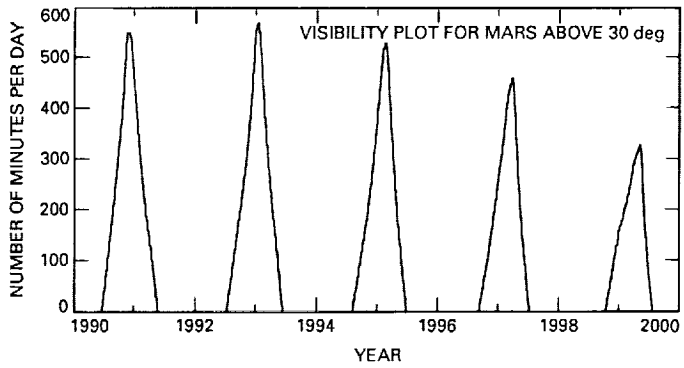


Fig. 3. Plot of the number of minutes per day (24 hours) for which Mars is visible above 30 degrees in the nighttime sky.

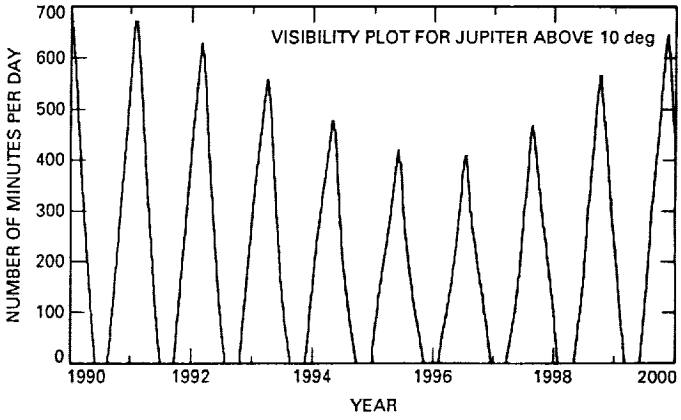


Fig. 4. Plot of the number of minutes per day (24 hours) for which Jupiter is visible above 10 degrees in the nighttime sky.

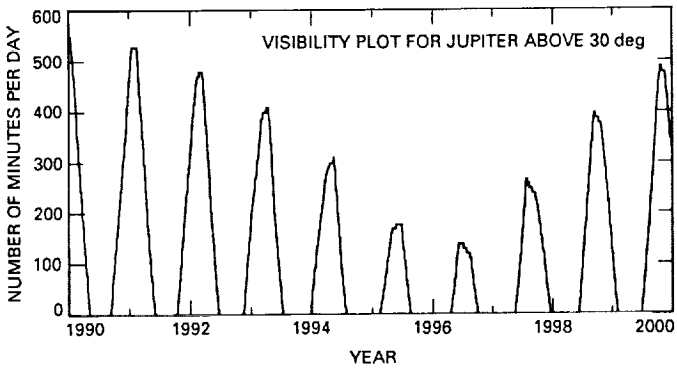


Fig. 5. Plot of the number of minutes per day (24 hours) for which Jupiter is visible above 30 degrees in the nighttime sky.

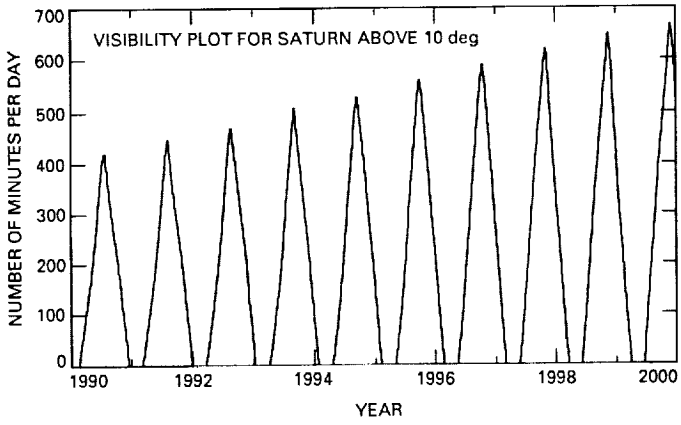


Fig. 6. Plot of the number of minutes per day (24 hours) for which Saturn is visible above 10 degrees in the nighttime sky.

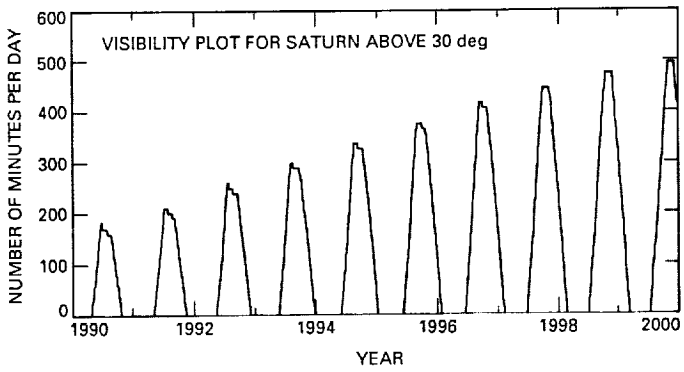


Fig. 7. Plot of the number of minutes per day (24 hours) for which Saturn is visible above 30 degrees in the nighttime sky.

TDA Progress Report 42-99 November 15, 1989

18.

Memory Management in Traceback Viterbi Decoders

O. Collins
Johns Hopkins University, Maryland

F. Pollara
Communications Systems Research Section

The new Viterbi decoder for long constraint length codes, under development for the DSN, stores path information according to an algorithm called "traceback." The details of a particular implementation of this algorithm, based on three memory buffers, are described. The penalties in increased storage requirement and longer decoding delay are offset by the reduced amount of data that needs to be exchanged between processors, in a parallel architecture decoder.

I. Introduction

A new, long constraint length Viterbi decoder [1] is under development for the DSN, and will be used to decode the constraint length $K = 15$ experimental code adopted by the Galileo mission. This article describes the traceback algorithm that is used in this decoder to store the most likely paths into each node of the decoder's trellis.

The *traceback* (TB) method is one of two known ways to store decisions made by the add-compare-select unit of a Viterbi decoder, and then provide decoded bits as output. The other method is the more traditional *register exchange* (RE) technique. The RE method is suitable for short constraint length decoders or for low-speed decoders due to the large amount of data that needs to be read, modified, and rewritten at each bit time.

The basic difference between these two methods is that the RE method stores the actual hypothesized information sequences (survivors), while the TB method stores the results of comparisons of paths converging into each node of the trellis. The maximum-likelihood path is then found by tracing back through the trellis a path, according to stored decisions. It is worth observing that the bits representing the results of these comparisons actually coincide with the information bits in convolutional codes where the state transitions are governed by a shift register. Therefore, the crucial difference between RE and TB methods lies in the organization of the memory used to store the survivors. The TB method is widely used in practice, but not as widely described in the literature. It is mentioned in [2], and described in [3] without any reference to required storage or resulting decoding delay.

This article deals with the details of one version of the traceback algorithm that is based on three pointers exploring three memory buffers. Details on the hardware implementation of this version of the algorithm may be found in [4]. If L branches are required as the minimum decoding depth ($L \approx 5K$ is a usually accepted rule, but $L \approx 10K$ is more realistic for low E_b/N_0 applications), enough storage

must be provided for $3L$ branches in order to perform the necessary buffering for this TB method. This penalty is not important in practice since inexpensive and slow off-chip memory can be used, while the RE method requires fast on-chip registers.

II. Memory Management for Traceback

The memory required for the traceback method is organized in three banks as shown in Fig. 1. Each bank is L bits long and 2^{K-1} bits high (the number of states), which gives a required storage of about 1 Mbyte, for $K = 15$ and $L = 170$ (not including additional storage for 2^{K-1} accumulated metrics). Any bit in this traceback memory can be accessed by an address consisting of the state j ($0 \leq j < 2^{K-1}$) and the bit memory pointer m ($0 \leq m < 3L$).

There are three basic operations going on in the memory banks every bit time:

(1) *Traceback*, which is a "read" operation and traces a path between states on the trellis by computing the next (backward) state address from the presently read memory content.

(2) *Decoding*, which is also a "read" operation and similar to traceback, except that it is performed on "older" data and it produces output information bits corresponding to the path being traced.

(3) *Writing* new data (decisions given by the add-compare-select unit), which moves forward on the trellis. These bits can be written in the locations just freed by the decoding operation.

Every L bit times a new traceback front is started in one memory bank from state 0, and a new decode front is started in a different memory bank at the state where the previous traceback ended. New data is written into the second memory bank as memory locations are freed by the decode operation. The third memory bank remains inactive. After a period of L bit times, the traceback/decode/write operations are switched to a different pair of memory banks, as described in detail below.

Among the three operations, the writing of new data is by far the most time consuming, since 2^{K-1} bits must be written for each information bit time. Read operations only access one bit per information bit time.

Given the amount of required memory, it is cheaper to use commercial RAM chips, rather than to design this memory into custom VLSI circuits.

III. The Traceback Algorithm in Detail

The evolution of memory operations needed for the traceback method is illustrated in Fig. 2, where the vertical axis represents the elapsed time and each box represents the status of memory at a given time. The variables' names are the same later used in the pseudocode description of the algorithm in Fig. 4.

At $time = 0$, a new traceback is started from state 0 ($state\_tb = 0$) at bit memory pointer $m = L - 1$ in the rightmost memory bank (bank = 0). The top row of Fig. 2 shows the memory bank number where traceback (tb) and decoding (dec) operate. This traceback will end in a certain $state\_tb$ at bit memory pointer $m = 0$.

Simultaneously, a decoding operation starts from state $state\_dec$ (this is initially an arbitrary value) at bit memory pointer $m = L-1$ and proceeds until $m = 0$, in the leftmost memory bank (bank 1). For each decoded bit, a full column of bits can be overwritten with new data. Notice that all three operations (traceback, decode, and write) evolve from right to left.

At the end of the first block of L bits, a new traceback starts from state 0 ($state\_tb = 0$) at bit memory pointer $m = 0$, moving from left to right in memory bank 1. At the same time, a decoding operation goes on in the middle bank (bank 2), starting at the state where the previous traceback ended ($state\_dec = state\_tb$) and at bit memory pointer $m = 0$. Also, new data is written in the bank doing decoding. All of these operations evolve from left to right.

After the end of the second block, a traceback and a decoding start at $m = L - 1$, both moving from right to left, in banks 2 and 0, respectively. New data is written in bank 0. After the third block, i.e., during the fourth block, traceback and decoding take place in banks 0 and 1 again, but all operations occur left-to-right, which is opposite to the direction for the first block. Notice that the read/write operations alternate between right-to-left and left-to-right sweeps, which implies that Fig. 2 has a repetition cycle of $3 \times 2 = 6$ blocks.

Only after three full blocks is decoding performed on data that has actually been written, rather than on initialized memory. Therefore the decoding delay is at least $3L$ bits. Since the decoded output is generated in reversed order (last bit first in each block), one must provide a buffer to reverse the output, bringing the decoding delay to $4L$ bits. Finally, the delay due to the encoder memory must be added, which yields a total decoding delay of $4L + K$ bits.

The flow diagram of memory operations is shown in Fig. 3, where the traceback and decoding operations are shown as sequential in time, even though they may happen simultaneously in a specialized hardware.

Figure 4 shows the pseudocode description of the TB algorithm for a convolutionally coded system using a (7,1/2) code and $L = 100$. The C-language version of this algorithm has been used to demonstrate this concept by software simulation and to verify the correctness of the algorithm. The memory is denoted by the three-dimensional array RAM[*state*][*m*][*bank*], M[*state*][.] stores the accumulated metrics, and *d*[.] represents the branch metrics.

IV. Advantages of Traceback for Parallel Processing

In a multiprocessor implementation, the Viterbi algorithm based on register exchange requires that the full survivor sequences be exchanged among processors, together with the accumulated metrics. It has been found [5] that the traceback method drastically reduces the communication bandwidth required between processors by eliminating the need for survivor exchanges. This reduction is achieved at the price of higher decoding delay.

Figure 5 shows a general parallel architecture, where the interconnection network is described in [6]. Since, among the three operations described in Section II, the write operation is the most demanding, it is performed concurrently in each processor and its local memory. Each processor operates sequentially on a certain number of states, and then exchanges the accumulated metrics through the interconnection network. The traceback/decoding operation may use a bus line to transfer information, consisting of traceback memory addresses sent to all memories and single bits coming from a particular memory, corresponding to the memory location referenced by a given address. The new address is computed from the old one and the data bit read from a memory. The address computation does not require parallelism, since it uses only one bit read per information bit. At the end of each block the last address found by the traceback unit is used to initialize the decoding unit. Further details on the hardware implementation may be found in [4].

References

[1] J. Statman, G. Zimmerman, F. Pollara, and O. Collins, "A Long Constraint Length VLSI Viterbi Decoder for the DSN," *TDA Progress Report 42-95*, Jet Propulsion Laboratory, Pasadena, California, pp. 134–142, November 15, 1988.

[2] G. C. Clark and J. B. Cain, *Error Correction Coding for Digital Communications*, Plenum Press, 1981.

[3] C. M. Rader, "Memory Management in a Viterbi Decoder," *IEEE Trans. on Communications*, vol. COM-29, no. 9, pp. 1399–1401, September 1981.

[4] O. Collins, *Coding Beyond the Computational Cutoff Rate*, Ph.D. Thesis, California Institute of Technology, May 1989.

[5] F. Pollara, "Concurrent Viterbi Algorithm with Traceback," *SPIE Proceedings*, vol. 696, p. 204–209, August 1986.

[6] O. Collins, F. Pollara, S. Dolinar, and J. Statman, "Wiring Viterbi Decoders (Splitting deBruijn Graphs)," *TDA Progress Report 42-96*, Jet Propulsion Laboratory, Pasadena, California, pp. 93–103, February 15, 1989.

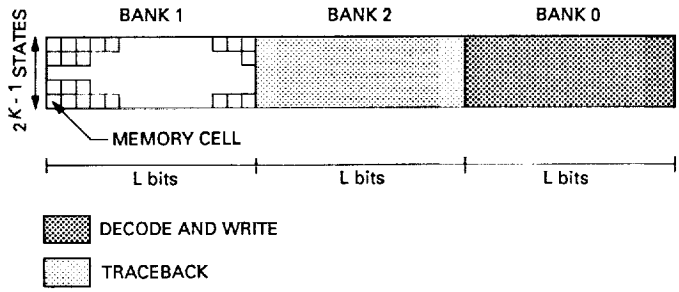


Fig. 1. Memory organization in three banks.

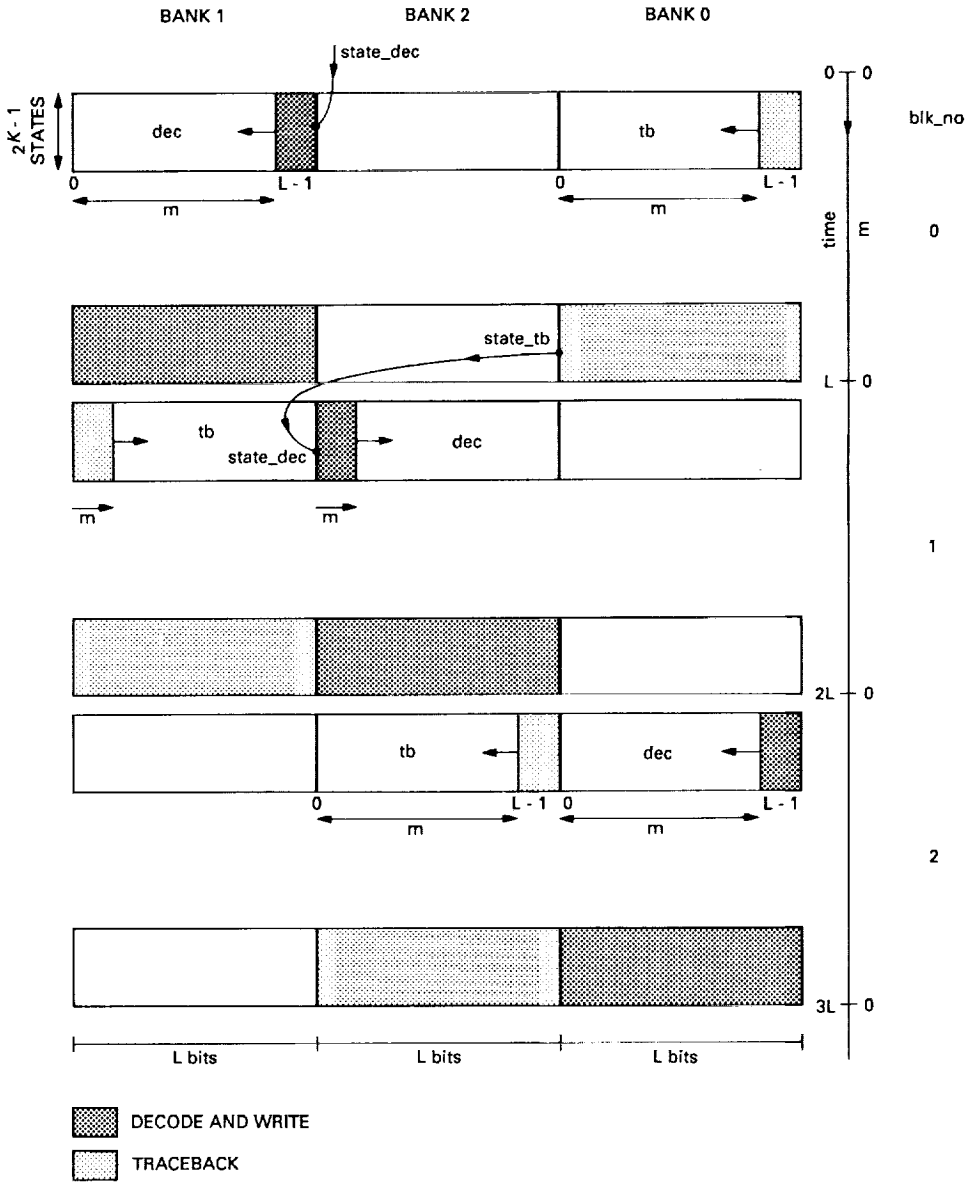


Fig. 2. Evolution of memory operations in the traceback method.

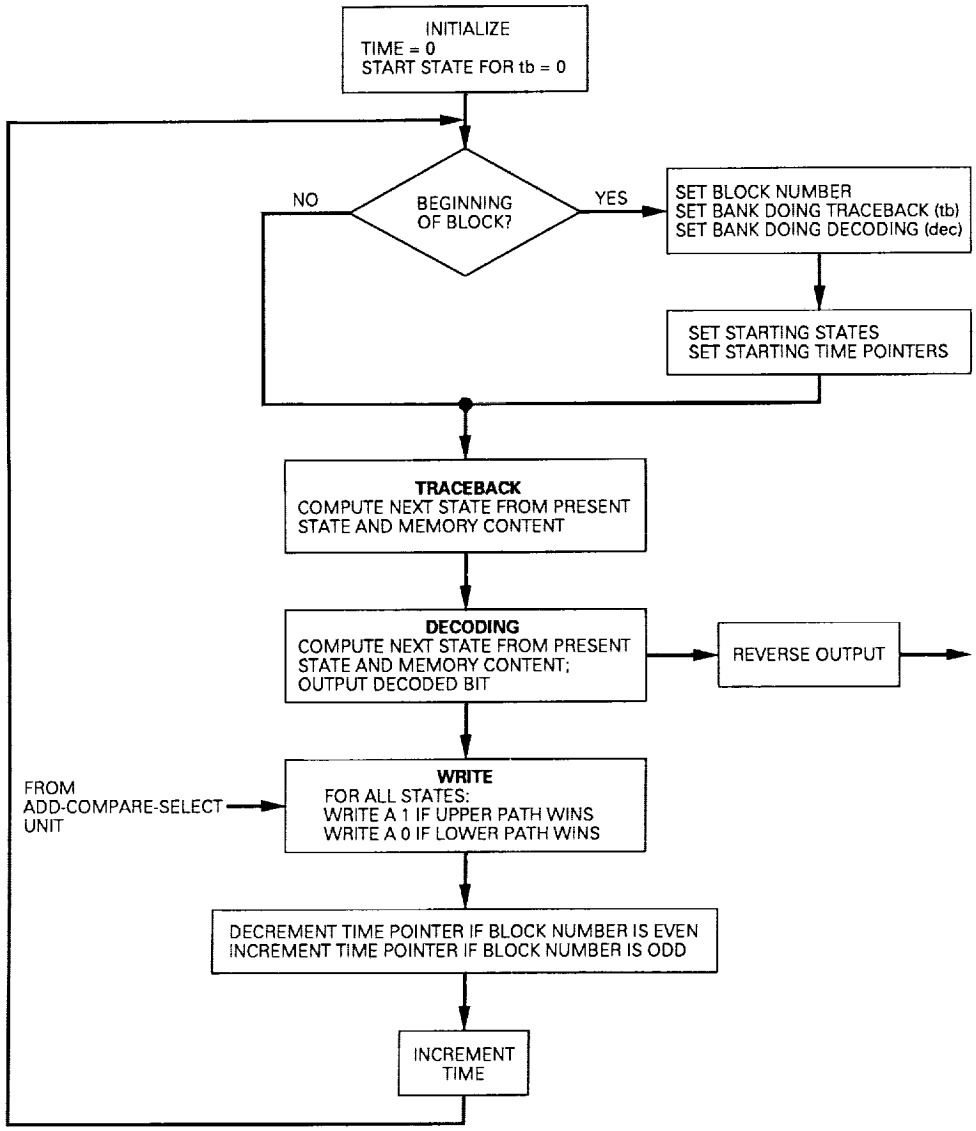


Fig. 3. Flow diagram of memory operations.

```
K=7                                          "constraint length"
NS=2^(K-1)                                   "number of states"
NS2=NS/2
L=100                                        "truncation length"
state_tb=0
time=0

WHILE time < max
{

m=time MOD L
bit_par=time MOD 2

IF(m==0)                                     "start a new traceback front"
{
blk_no=time/L
blk_par=blk_no MOD 2
tb=blk_no MOD 3                              "bank doing traceback = 0,1,2"
dec=(tb+1) MOD 3                             "bank doing decoding"
state_dec=state_tb                           "set starting state for decoding"
state_tb=0                                   "set starting state for new traceback"
FOR m FROM 0 TO L-1 {outr[m]=out[m]}         "buffer for order reversal"
}

IF(blk_par==0) {m1=L-m-1}                    "right to left"
ELSE           {m1=m}                        "left to right"

state_dec=(SHIFT RIGHT state_dec OF 1 BIT) OR (NS2*RAM[state_dec][t][dec])
                                             "decoding: address in bank=dec"

state_tb= (SHIFT RIGHT state_tb OF 1 BIT ) OR (NS2*RAM[state_tb ][t][tb ])
                                              "traceback: address in bank=tb"

out[m]= (SHIFT RIGHT state_dec OF K-2 BITS) AND 1        "output buffer"

FOR j FROM 0 TO NS-1                                     "add, compare, select"
        {
        i=SHIFT RIGHT j OF 1 BIT
        L0 = M[i]      [bit_par XOR 1] + d[K[j]]
        L1 = M[i OR NS2][bit_par XOR 1] + d[3-K[j]]
        IF(L1 < L0)       {                              "write one bit"
                          M[j][bit_par]=L1
                          RAM[j][m1][dec]=1
                          }
        ELSE              {
                          M[j][bit_par]=L0
                          RAM[j][m1][dec]=0
                          }
        }

PRINT outr[L-1-m]                            "print decoded bits in correct order"

time = time+1
}
```

Fig. 4. Pseudocode for traceback algorithm.

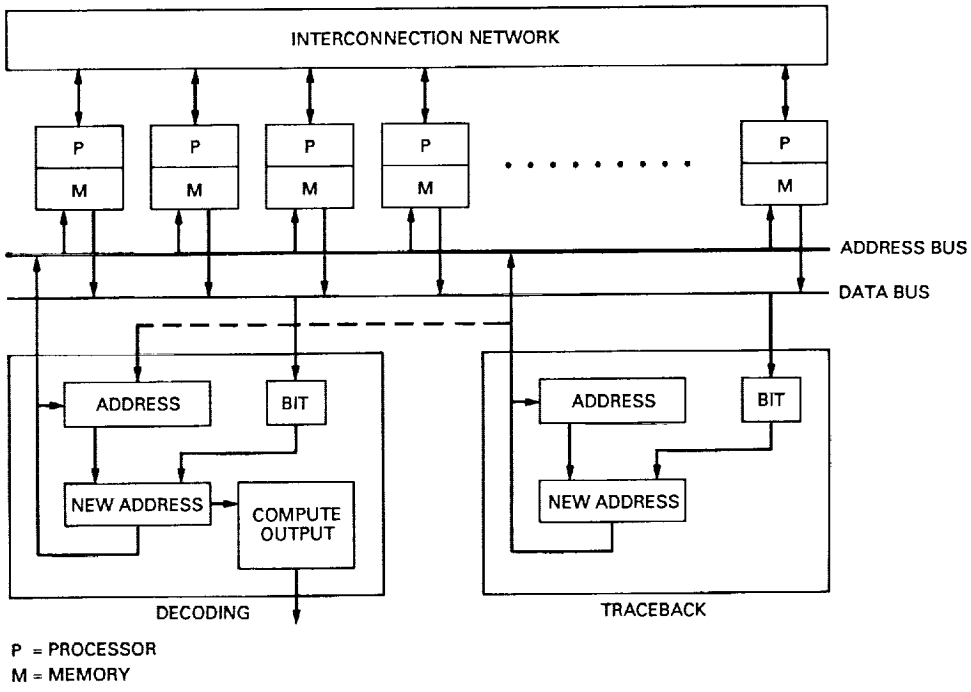


Fig. 5. Parallel traceback architecture.

Some Easily Analyzable Convolutional Codes

R. McEliece, S. Dolinar, and F. Pollara
Communications Systems Research Section

H. Van Tilborg
Eindhoven University, Mathematics Department, The Netherlands

Convolutional codes have played and will play a key role in the downlink telemetry systems on many NASA deep-space probes, including Voyager, Magellan, and Galileo. One of the chief difficulties associated with the use of convolutional codes, however, is the notorious difficulty of analyzing them. Given a convolutional code as specified, say, by its generator polynomials, it is no easy matter to say how well that code will perform on a given noisy channel. The usual first step in such an analysis is to compute the code's free distance; this can be done with an algorithm whose complexity is exponential in the code's constraint length. The second step is often to calculate the transfer function in one, two, or three variables, or at least a few terms in its power series expansion. This step is quite hard, and for many codes of relatively short constraint length, it can be intractable. However, we have discovered a large class of convolutional codes for which the free distance can be computed by inspection, and for which there is a closed-form expression for the three-variable transfer function. Although for large constraint lengths, these codes have relatively low rates, they are nevertheless interesting and potentially useful. Furthermore, the ideas developed here to analyze these specialized codes may well extend to a much larger class.

I. Introduction

In this article a class of binary $(n, 1)$, constraint length K, convolutional codes, called *zero-run length (ZRL) convolutional codes*, is defined and studied. These codes are interesting because they are easy to analyze. ZRL codes include as special cases orthogonal convolutional codes, the recent "superorthogonal codes" of Viterbi, and many others. None of the convolutional codes currently used in NASA missions belong to the ZRL class. For any ZRL code, it is possible to compute the free distance by inspection, and to write down the complete transfer function

$T(D,I,L)$, explicitly (see Theorem 7, below). Important variations of the transfer function, viz.

$$T_{\text{num}}(D) = T(D,1,1)$$

$$T_{\text{bit}}(D) = \frac{\partial T}{\partial I}(D,1,1)$$

$$T_{\text{len}}(D) = \frac{\partial T}{\partial L}(D,1,1)$$

are commonly used to overbound the probability of decoder error for these codes ([3], Section 9.3, or [4], Section 4.4). For arbitrary convolutional codes, these functions can be very complicated indeed (see [7]), but for any ZRL code these functions have simple, closed-form expressions (see Corollary 8).

II. Zero-Run Length Convolutional Codes

Any $(n,1)$, constraint length K convolutional code is characterized by a list of n generator polynomials $(g_1(x), \ldots, g_n(x))$, where $g_i(x) = g_{i,0} + g_{i,1}x + \cdots + g_{i,K-1}x^{K-1}$ is a polynomial of degree $K-1$ or less. The encoder for such a code consists of a shift register of length $K-1$, with one input and n outputs; the *state* of the encoder is defined to be the contents of the shift register. If $(s_1, \ldots, s_{K-1})$ is the current state, and s_0 is the current input, the next state is $(s_0, \ldots, s_{K-2})$ and the output, which we will call a *code segment*, is the n-tuple $(y_1, \ldots, y_n)$, where $y_i = \sum_{j=0}^{K-1} s_j g_{i,j}$.

1. Definition. An encoder state $s = (s_1 s_2 \cdots s_{K-1})$ is said to have *zero-run length* i, written "ZRL(s) = i" for short, if s contains exactly i leading zeros. For example, with $K = 5$, ZRL(0010) = 2, ZRL(0000) = 4, and ZRL(1001) = 0. In general, for an $(n,1)$, constraint length K, convolutional code, there will be 2^{K-1} states, but only K possible values for ZRL (0, 1, ..., $K-1$).

Note that if the encoder is in a state of zero-run length i, and the input is 0, the next state will have ZRL = $\min(i+1, K-1)$, whereas if the input is 1, the next state will have ZRL = 0. Thus the ZRL of the encoder's next state depends only on the current value of ZRL and the input. This fact is illustrated in Fig. 1, which shows the topology of states, organized according to the values of ZRL. In Fig. 1, the arrows marked with α's represent state transitions caused by 0 inputs, and the arrows marked with β's represent state transitions caused by 1 inputs. We will return to this state diagram in the proof of our main result, Theorem 7, below.

2. Definition. An $(n,1)$ convolutional code of constraint length K is said to be a *ZRL* code if the output weight depends only on the input and the ZRL of the state. The symbol u_i is used to denote the output weight if the encoder has ZRL = i and the input is 0, and the symbol w_i is used if ZRL = i and the input is 1. The u_i's and the w_i's are conveniently displayed in a $2 \times K$ matrix, called the *weight matrix* of the code:

$$W = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & \cdots & K-1 \end{array} \\ \begin{pmatrix} u_0 & u_1 & \cdots & u_{K-1} \\ w_0 & w_1 & \cdots & w_{K-1} \end{pmatrix} \end{array}$$

3. Example. Let $K = 3$. Then the $(4,1)$ convolutional code with generator polynomial list $(1, x, 1 + x^2, 1 + x + x^2)$ is a ZRL code. Since with $K = 3$ there is only one state with ZRL = 1, viz. 01, and only one state with ZRL = 2, viz. 00, in order to verify that this code is ZRL, one need only investigate the two states with ZRL = 0, i.e., 10 and 11. If the state is 10 and the input is 0, the output is (0101), whereas if the input is 1 the output is (1110). On the other hand, if the state is 11 and the input is 0, the output is (0110), and if the input is 1, the output is (1101). Thus, if the state has ZRL = 0, and the input is 0, the output weight is 2; and if the input is 1, the output weight is 3. Hence, the output weight indeed depends only on the state's ZRL, as required. The weight table for this code is as follows:

$$W = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 2 & 2 & 0 \\ 3 & 1 & 3 \end{pmatrix} \end{array}$$

4. Definition. The *profile* of an $(n,1)$, constraint length K ZRL convolutional code is the vector $(d_1, d_2, \ldots, d_K)$, where d_i is the Hamming weight of the output of the encoder, beginning in a state with ZRL = 0, with length i input sequence $0^{i-1}1$.

5. Lemma. In terms of the entries in the weight table, the profile of a ZRL convolutional code is

$$d_i = u_0 + u_1 + \cdots + u_{i-2} + w_{i-1}$$

for $i = 1, 2, \ldots, K$

Proof: If one starts in a state with ZRL = 0, and uses the input sequence 0^{i-1}, one passes through states

with ZRL $= 1, 2, \ldots, i-2$, causing outputs of weight $u_0, u_1, \ldots, u_{i-2}$, and arrives at a state with ZRL $= i-1$. The last input of 1 causes the encoder to move to a state with ZRL $= 0$ and to produce an output of weight w_{i-1}.

6. Example. Combining the weight table in Example 3 with Lemma 5, one finds that the profile of the code in Example 3 is $(3,3,7)$: $d_1 = w_0 = 3$; $d_2 = u_0 + w_1 = 2+1 = 3$; and $d_3 = u_0 + u_1 + w_2 = 2+2+3 = 7$.

III. Transfer Function for ZRL Codes

The following theorem is our main result. It gives the promised closed-form expression for the transfer function of a ZRL code in terms of its profile.

7. Theorem. For a ZRL convolutional code with profile $(d_1, \ldots, d_K)$, the three-variable transfer function is given by

$$T(D,I,L) = \frac{D^{d_K} I L^K}{1 - \sum_{i=1}^{K-1} D^{d_i} I L^i}$$

Proof: One begins by reviewing the definition of $T(D, I, L)$ for an arbitrary $(n,1)$, constraint length K, convolutional code. (See [3] or [4] for more details.)

Starting with the state diagram for the given code, which is the 2^{K-1} vertex deBruijn graph, each of the 2^K edges is labelled with a monomial in the three indeterminates D, I, and L, i.e., a term of the form $D^w I^\epsilon L$. The power w of D in the monomial represents the Hamming weight of the encoder output corresponding to the given state transition, and ϵ is either 0 or 1, according to whether the corresponding encoder input is zero or one. The resulting labelled, directed graph is called the "DIL state diagram" for the code.

In Fig. 2 is the DIL state diagram for a $K = 3$ ZRL code. For example, in Fig. 2 the edge from state 10 to 11 is labelled $D^{w_0} IL$. This is because the transition $10 \rightarrow 11$ is caused by an encoder input of 1, so that the exponent of I in the edge label is 1. State 10 has ZRL $= 0$, and by definition of a ZRL code, when the state has ZRL $= 0$ and the input is 1, the output weight is w_0; thus the exponent on D in the label is w_0. The other seven edge labels can be explained similarly.

A *path* of length m in the DIL state diagram is defined as a sequence of $m+1$ vertices such that each adjacent pair of vertices in the sequence is connected by a directed edge. For example, in Fig. 2, the vertex sequence $00 \rightarrow 10 \rightarrow 01 \rightarrow 00$ is a path of length 3. A path is completely specified by its initial vertex and the string of input bits corresponding to the vertex transitions, which we call the *input string* of the path. For example, the path $00 \rightarrow 10 \rightarrow 01 \rightarrow 00$ has initial vertex 00 and input string 100. The *weight* of a path is defined to be the product of the labels on its edges. For example, the path $00 \rightarrow 10 \rightarrow 01 \rightarrow 00$ in Fig. 2 has weight $D^{w_2+u_0+u_1} IL^3$.

The three-variable transfer function $T(D, I, L)$ is now defined to be *the sum of the weights of all paths from vertex* 0^{K-1} *back to vertex* 0^{K-1} *which have no intermediate returns to vertex* 0^{K-1}. Alternatively, $T(D, I, L)$ is the sum of the weights of all paths with initial vertex 0^{K-1} whose input string ends with 0^{K-1} but has no other substring equal to 0^{K-1}. (In [3, Section 9.3] these paths are called "fundamental paths.")

In principle, one can compute $T(D, I, L)$ for any convolutional code using the standard "transfer matrix method" described, for example, in [5, Sec. 4.7]. However, this method is essentially equivalent to inverting a $2^{K-1} \times 2^{K-1}$ matrix with three-variable monomial entries, and is not in general practical except for codes with extremely small constraint lengths [7]. However, for a ZRL code, one can simplify this calculation considerably, by first "collapsing" the state diagram by combining states with the same value of ZRL. In the collapsed state diagram, there will be K vertices, labelled $0, 1, \ldots, K-1$; vertex i will be connected by a directed edge to vertex j if there is any edge in the original (noncollapsed) DIL state diagram connecting a vertex with ZRL $= i$ to one with ZRL $= j$. The label on an edge in the reduced state diagram will be the same as the label on the corresponding edge in the original graph; the ZRL property implies that this rule is well defined.

The collapsing process is illustrated in Fig. 3, which shows the collapsed version of the graph in Fig. 2. Note, for example, that in Fig. 3 the edge from vertex 0 to vertex 1 is labelled $D^{u_0} L$. This is because in Fig. 2, both edges from a vertex with ZRL $= 0$ to a vertex with ZRL $= 1$, viz. $10 \rightarrow 01$ and $11 \rightarrow 01$, have the same label $D^{u_0} L$.

When the DIL state diagram for a constraint length K ZRL code is collapsed, the resulting state diagram will be identical to the state diagram in Fig. 1, where the labels α_i and β_i are given by

$$\alpha_i = D^{u_i} L$$

$$\beta_i = D^{w_i} IL$$

One can think of the collapsed state diagram of Fig. 1 as the state diagram of a finite-state machine, with input alphabet $\{0,1\}$ and output alphabet the set of monomials $D^w I^\epsilon L$. If this machine is in state i and its input is 0, its next state is $\min(i+1, K+1)$, and its output is $D^{u_i} L$; if it is in state i and its input is 1, its next state is 0 and its output is $D^{w_i} IL$. Note that, as for the original state diagram, any path in the collapsed state diagram is specified by its initial vertex and its input string. For example, the path $2 \to 0 \to 1 \to 2$ in the collapsed state diagram of Fig. 3 has initial vertex 2 and input string 100. Its weight is $D^{w_2+u_0+u_1} L^3 I$.

The important point is that the collapsed state diagram is equivalent to the original state diagram for purposes of computing the $T(D,I,L)$ transfer function for the ZRL code. This is because a path in the original DIL state diagram with an initial vertex with ZRL $= i$ and input string σ will have the same weight as a path in the collapsed state diagram with initial vertex i and the same input string σ. For example, the path in the state diagram of Fig. 2 with initial vertex 00 and input string 100 has weight $D^{w_2+u_0+u_1} L^3 I$, which is the same as the weight of the path in the collapsed state diagram of Fig. 3 with initial state 2 and input string 100.

It follows then that the $T(D,I,L)$ transfer function for a ZRL code is the sum of the weights of all paths in the collapsed state diagram of Fig. 1 from state $K-1$ back to state $K-1$, with no intermediate returns to state $K-1$. This transfer function is denoted by $T^*_{K-1,K-1}$. One way to compute $T^*_{K-1,K-1}$ is to remove the vertices $1, 2, \ldots, K-2$ from the state diagram, but to preserve the path label information by relabelling the remaining edges appropriately, as shown in Fig. 4. For example, in Fig. 4, the edge from vertex 0 to vertex $K-1$ is labelled $\alpha_0\alpha_1\cdots\alpha_{K-2}$; this is because in Fig. 1 there is exactly one path from vertex 0 to vertex $K-1$ that uses only the deleted vertices $\{1,2,\ldots,K-1\}$, viz. $012\cdots(K-1)$, and its weight is $\alpha_0\alpha_1\cdots\alpha_{K-2}$. Similarly, the loop at vertex 0 is relabelled to reflect the fact that there are $K-1$ paths from vertex 0 back to vertex 0 which use only the deleted vertices: $00, 010, 0120, \ldots, 012\cdots(K-2)0$, and the sum of the weights of these $K-1$ paths is $\beta_0 + \alpha_0\beta_1 + \cdots + \alpha_0\cdots\alpha_{K-3}\beta_{K-2}$, which is the label on the loop at vertex 0 in Fig. 4.

Once the state diagram has been reduced to only two states, the computation of the transfer function $T^*_{K-1,K-1}$ is straightforward. Any path from vertex $K-1$ back to vertex $K-1$ with no intermediate return to vertex $K-1$ in Fig. 4 must be of the form $(K-1)0\cdots0(K-1)$, and so the desired transfer function is equal to the weight of the path $(K-1)0(K-1)$ divided by 1 minus the weight of the loop at vertex 0, i.e.,

$$T^*_{K-1,K-1} = \frac{\alpha_0\alpha_1\cdots\alpha_{K-2}\beta_{K-1}}{1-\beta_0-\alpha_0\beta_1-\alpha_0\alpha_1\beta_2-\cdots-\alpha_0\cdots\alpha_{K-3}\beta_{K-2}}$$

If one substitutes the above values for α_i and β_i into this expression, and uses the definition of the profile, the expression for $T(D,I,L)$ in the statement of the theorem is obtained.

8. Corollary. For a ZRL convolutional code with profile $(d_1, \ldots, d_K)$, the free distance is d_K and

$$T_{\text{num}}(D) = \frac{D^{d_K}}{P(D)}$$

$$T_{\text{bit}}(D) = \frac{D^{d_K}}{P(D)^2}$$

$$T_{\text{len}}(D) = \frac{D^{d_K} Q(D)}{P(D)^2}$$

where the polynomials $P(D)$ and $Q(D)$ are defined by

$$P(D) = 1 - \sum_{i=1}^{K-1} D^{d_i}$$

$$Q(D) = K - \sum_{i=1}^{K-1} (K-i) D^{d_i}$$

Proof: This follows directly from Theorem 7 and the definitions of $T_{\text{num}}(D)$, $T_{\text{bit}}(D)$, and $T_{\text{len}}(D)$ given at the beginning of the article.

9. Example. Continuing Examples 3 and 6, the profile is $(3,3,7)$, and so $P(D) = 1-2D^3$, $Q(D) = 3-3D^3$. Thus, by Corollary 8, $d_{\text{free}} = 7$, and

$$\begin{aligned} T_{\text{num}}(D) &= \frac{D^7}{1-2D^3} \\ &= D^7 + 2D^{10} + 4D^{13} \\ &\quad + 8D^{16} + 16D^{19} + 32D^{22} + \cdots \end{aligned}$$

$$T_{\text{bit}}(D) = \frac{D^7}{(1-2D^3)^2}$$
$$= D^7 + 4D^{10} + 12D^{13} + 32D^{16} + 80D^{19} + 192D^{22} + \cdots$$

$$T_{\text{len}}(D) = \frac{D^7(3-3D^3)}{(1-2D^3)^2}$$
$$= 3D^7 + 9D^{10} + 24D^{13} + 60D^{16} + 144D^{19} + 336D^{22} + \cdots$$

IV. Superorthogonal and Ultraorthogonal Codes

Next, two important general classes of ZRL convolutional codes, the superorthogonal codes introduced by Viterbi [1] and the ultraorthogonal codes introduced here, are defined.

10. Definition. The *superorthogonal code* of constraint length K, denoted by S_K, is defined as follows: $S_1 = (1)$, and for $K \geq 2$, then S_K is a $(2^{K-2}, 1)$ code whose generator polynomials are all 2^{K-2} possible polynomials of the form $1 + g_1 x + \cdots + g_{K-2}x^{K-2} + x^{K-1}$.

11. Definition. The *ultraorthogonal code* of constraint length K, denoted by U_K, is defined as follows: $U_1 = (0)$, and for $K \geq 2$, then U_K is a $(2^{K-2}, 1)$ code whose generator polynomials are all 2^{K-2} possible polynomials of the form $g_1 x + \cdots + g_{K-2}x^{K-2} + x^{K-1}$.

12. Example. For $K = 3$ the code S_3 has generator polynomial list $(1 + x^2, 1 + x + x^2)$, and U_3 has generator polynomial list $(x^2, x + x^2)$.

13. Theorem. For all $K \geq 1$, the codes S_K and U_K are ZRL codes. The weight tables for the superorthogonal codes are as follows:

$$W(S_1) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{array}$$

$$W(S_2) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \quad 1 \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$W(S_3) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \quad 1 \quad 2 \\ \begin{pmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix} \end{array}$$

and, for $K \geq 3$

$$W(S_K) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \quad 1 \quad \cdots \quad K-3 \quad K-2 \quad K-1 \\ \begin{pmatrix} 2^{K-3} & 2^{K-3} & \cdots & 2^{K-3} & 2^{K-2} & 0 \\ 2^{K-3} & 2^{K-3} & \cdots & 2^{K-3} & 0 & 2^{K-2} \end{pmatrix} \end{array}$$

Similarly, the weight tables for the ultraorthogonal codes are as follows:

$$W(U_1) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{array}$$

$$W(U_2) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \quad 1 \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \end{array}$$

$$W(U_3) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \quad 1 \quad 2 \\ \begin{pmatrix} 1 & 2 & 0 \\ 1 & 2 & 0 \end{pmatrix} \end{array}$$

and, for $K \geq 3$

$$W(U_K) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} 0 \quad 1 \quad \cdots \quad K-3 \quad K-2 \quad K-1 \\ \begin{pmatrix} 2^{K-3} & 2^{K-3} & \cdots & 2^{K-3} & 2^{K-2} & 0 \\ 2^{K-3} & 2^{K-3} & \cdots & 2^{K-3} & 2^{K-2} & 0 \end{pmatrix} \end{array}$$

Proof: The key to the proof is the close relationship between the convolutional codes S_K and U_K and the first-order Reed-Muller (1RM) block codes, which are now described. The $(2^m, m+1)$ 1RM code can be defined by an $(m+1) \times 2^m$ generator matrix G_m which has as columns all possible binary $(m+1)$-tuples ending with 1. For example, with $m = 2$ the $(4, 3)$ 1RM code has generator matrix

$$G_2 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

It is known that all weights in the $(2^m, m+1)$ 1RM code are equal to 2^{m-1}, except for the all-zero word and the all-one word ([8], Chapter 13). If G_m^0 is defined to be the matrix obtained by adding a row of zeros at the top of G_m, and G_m^1 to be the matrix obtained by adding a row of ones at the top of G_m, then the columns of G_{K-2}^0 give the coefficients of the generator polynomials of U_K

and the columns of G^1_{K-2} give the generator polynomials of S_K. For example, again with $m = 2$,

$$G^0_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad G^1_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

It therefore follows that every (2^{K-2})-bit code segment in either of the codes S_K or U_K is a word in the $(2^{K-2}, K-1)$ 1RM code. In almost every case, this segment will have weight 2^{K-3}; the only other possibilities are weight 0 (the all-zero codeword) and weight 2^{K-2} (the all-one codeword). To analyze these exceptional cases, note that every linear combination of rows of either G^0_m or G^1_m is a word in the 1RM code. All such linear combinations will therefore have weight 2^{m-1}, with the following exceptions. In G^0_m, the empty linear combination, or the top row, give the all-zero codeword; and the bottom row, or the top row plus the bottom row, give all ones. In G^1_m, the empty linear combination or the top row plus the bottom row gives the all-zero codeword; and the top row or the bottom row gives all ones.

Therefore, in the ultraorthogonal code U_K, the code segment will be all zeros if and only if the state is 0^{K-1} and the input is 0, or the state is 0^{K-1} and the input is 1. Similarly, the code segment will be all ones if and only if the state is $0^{K-2}1$ and the input is zero, or the state is $0^{K-2}1$ and the input is 1. Thus, the output weight will be 2^{K-2} unless the state has ZRL $= K-1$ and the input is 0 or 1, in which case the output weight is 0, or if the state has ZRL $= K-2$ and the input is 0 or 1, in which case the output weight is 2^{K-1}. This is what the theorem states about the ultraorthogonal codes.

Similarly, in the superorthogonal code S_K, the code segment will be all zeros if and only if the state is 0^{K-1} and the input is 0, or the state is $0^{K-2}1$ and the input is 1. Similarly, the code segment will be all ones if and only if the state is 0^{K-1} and the input is 1, or the state is $0^{K-2}1$ and the input is 0. Thus, the output weight will be 2^{K-2} unless the state has ZRL $= K-1$ and the input is 0, or if the state has ZRL $= K-2$ and the input is 1, in which case the output weight is 0; or if the state has ZRL $= K-1$ and the input is 1, or if the state has ZRL $= K-2$, and the input is 0, in which case the output weight is 2^{K-1}. This is what the theorem states about the superorthogonal codes.

Theorem 13 provides many ZRL codes. The following definition and the discussion that follows will show how to use the superorthogonal and ultraorthogonal codes to build many other ZRL codes.

14. Definition. Given two convolutional codes, their *sum* is defined to be the convolutional code whose generator polynomial (g.p.) list is obtained by merging the g.p. lists for the original codes. Thus for example, the sum of the $(3,1)$ code with g.p. list $(1, 1+x, 1+x+x^2)$ and the $(2,1)$ code with g.p. list $(1+x^2, 1+x+x^2)$ is the $(5,1)$ code with g.p. list $(1, 1+x, 1+x^2, 1+x+x^2, 1+x+x^2)$. In general, the sum of an $(n_1, 1)$ convolutional code of constraint length K_1 and an $(n_2, 1)$ convolutional code of constraint length K_2 is an $(n_1+n_2, 1)$ convolutional code of constraint length $\max(K_1, K_2)$.

15. Lemma. If C_1 and C_2 are ZRL convolutional codes, with constraint lengths K_1 and K_2, respectively, with $K_1 \leq K_2$, then $C_1 + C_2$ is also ZRL, and the weight table for $C_1 + C_2$ is obtained from the weight tables W_1 and W_2 by first extending W_1 by repeating its last column $K_2 - K_1$ times, and then adding the two weight tables together.

Proof: If the two codes have the same constraint length, this is immediate. If, however, the two constraint lengths are different, and $K_1 < K_2$, C_1 can nevertheless be regarded as a convolutional code with constraint length K_2 in which the last $K_2 - K_1$ bits in the shift register are never used. States with ZRL values $K_1, K_1+1, \ldots, K_2-1$, will plainly behave just like the all-zeros state (with ZRL $= K_1 - 1$), and the extra $K_2 - K_1$ columns that appear in the weight matrix will be identical to the last column of the unextended weight matrix. The result now follows.

16. Example. The code of Example 3 is $S_1+U_2+S_3$, as may easily be verified. The corresponding weight tables are, by Theorem 13,

$$W(S_1) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$W(U_2) = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

$$W(S_3) = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

To obtain the weight matrix for $S_1 + U_2 + S_3$, first extend $W(S_1)$ and $W(U_2)$ to dimensions 2×3 by repeating the

respective last rows, and then adding the resulting matrices:

$$W = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 0 \\ 3 & 1 & 3 \end{pmatrix}$$

which is the same as was seen in Example 3.

17. Example. For any K, the code $\sum_{i=1}^{K}(S_i + U_i)$ is by Lemma 15 a ZRL code. In fact, this code has as generator polynomials all 2^K polynomials of degree $\leq K-1$; it is the orthogonal code of constraint length K.

18. Theorem. The profiles of the codes S_K are:

$$\begin{aligned} \text{profile}(S_1) &= (1) \\ \text{profile}(S_2) &= (0,2) \\ \text{profile}(S_3) &= (1,1,5) \\ \text{profile}(S_4) &= (2,4,4,12) \\ \text{profile}(S_5) &= (4,8,12,12,28) \\ &\vdots \\ \text{profile}(S_K) &= (2^{K-3}, 2\cdot 2^{K-3}, \ldots, \\ &\quad (K-2)2^{K-3}, (K-2)2^{K-3}, \\ &\quad (K+2)2^{K-3}) \end{aligned}$$

The profiles of the codes U_K are

$$\begin{aligned} \text{profile}(U_1) &= (0) \\ \text{profile}(U_2) &= (1,1) \\ \text{profile}(U_3) &= (1,3,3) \\ \text{profile}(U_4) &= (2,4,8,8) \\ \text{profile}(U_5) &= (4,8,12,20,20) \\ &\vdots \\ \text{profile}(U_K) &= (2^{K-3}, 2\cdot 2^{K-3}, \ldots, \\ &\quad (K-2)2^{K-3}, K2^{K-3}, K2^{K-3}) \end{aligned}$$

Proof: This follows by combining Theorem 13 and Lemma 5.

19. Example. By combining Theorems 7 and 18, one can obtain the transfer function for the superorthogonal codes. Indeed, if $z = D^{2^{K-3}}$, it follows from these theorems that for the superorthogonal code of constraint length K,

$$\begin{aligned} &T(D,I,L) \\ &= \frac{z^{K+2}IL^K}{1 - zIL(1 + zL + \cdots + z^{K-3}L^{K-3}) - z^{K-2}IL^{K-1}} \\ &= \frac{z^{K+2}IL^K(1-zL)}{1 - z(L+IL) - z^{K-2}IL^{K-1} + z^{K-1}(IL^{K-1} + IL^K)} \end{aligned}$$

an expression first found by Viterbi [1]. It follows then from Corollary 8 that $d_{\text{free}} = (K+2)2^{K-3}$ and

$$\begin{aligned} T_{\text{num}}(D) &= \frac{z^{K+2}(1-z)}{1 - 2z - z^{K-2} + 2z^{K-1}} \\ &= \frac{z^{K+2}(1-z)}{(1-2z)(1-z^{K-2})} \\ &= z^{K+2}\left\{ \frac{2^{K-3}}{(2^{K-2}-1)(1-2z)} \right. \\ &\quad \left. + \frac{(2^{K-3}-1) - z - 2z^2 - \cdots - 2^{K-4}z^{K-3}}{(2^{K-2}-1)(1-z^{K-2})} \right\} \end{aligned}$$

In the last expression, a two-term partial-fraction decomposition is seen (in braces) for the generating function $T_{\text{num}}(D)/z^{K+2}$. The coefficient of z^k in the expansion of the first term is

$$\frac{2^{K-3}}{2^{K-2}-1} \cdot 2^k$$

The coefficients of the expansion of the second term are periodic of period $K-2$, and each term is less than $1/2$ in absolute value. Since it is known that the coefficient of z^k in the combined expansion is an integer, it follows that this coefficient must be the integer closest to

$$\frac{2^{K-3}}{2^{K-2}-1} \cdot 2^k$$

Therefore, it has been proved that the coefficient of $D^{d_{\text{free}}+k2^{K-3}}$ in $T_{\text{num}}(D)$ for the superorthogonal code of constraint length K is

$$N_{d_{\text{free}}+k2^{K-3}} = \text{integer closest to } \frac{2^{K-3}}{2^{K-2}-1} \cdot 2^k$$

As a special case, it is found that the $(8,1)$, $K = 5$ superorthogonal code has $d_{\text{free}} = 28$, and the number of fundamental paths of weight $28 + 4k$ is the integer closest to $\frac{4}{7} \cdot 2^k$, i.e.,

$$T_{\text{num}}(D) = D^{28} + D^{32} + 2D^{32} + 5D^{36} + 9D^{40} + 18D^{44} + 37D^{48} + O(D^{52})$$

V. A Representation Theorem

If Theorem 13 is combined with Lemma 15, many ZRL codes can be constructed. It is surprising (and perhaps disappointing) that all such codes are constructed this way.

20. Theorem. An $(n,1)$ convolutional code C of constraint length K is ZRL if and only if it is the sum of copies of superorthogonal and ultraorthogonal codes:

$$C = \sum_{i=1}^{K} (m_i S_i + n_i U_i)$$

where m_i and n_i are integers denoting the multiplicities of S_i and U_i in the code C.

Proof: The proof of this theorem is lengthy and will be omitted.

The next lemma, when combined with Theorems 20 and 18, enables one to write down the transfer functions for any ZRL convolutional code.

21. Lemma. If C_1 and C_2 are ZRL convolutional codes, with constraint lengths K_1 and K_2 respectively, with $K_1 \leq K_2$, then the profile for the sum $C_1 + C_2$ is obtained from the profiles for C_1 and C_2 by first extending profile(C_1) to length K_2 by repeating its last entry $K_2 - K_1$ times, and then adding the two profiles together.

Proof: This follows by combining Lemma 15 with Lemma 5.

22. Example. The ZRL code in Example 3 is $C = S_1 + U_2 + S_3$, as was seen in Example 16. The corresponding profiles are, by Theorem 19,

$$\text{profile}(S_1) = (1)$$

$$\text{profile}(U_2) = (1,1)$$

$$\text{profile}(S_3) = (1,1,5)$$

To obtain C's profile, use Lemma 21. First extend the profiles of S_1 and U_2 to length 3 by repeating the last entries, and then add the resulting lists:

$$\text{profile}(C) = (1,1,1) + (1,1,1) + (1,1,5) = (3,3,7)$$

as was seen in Example 6. However, for the same values of n and K, one can get a larger d_{free} by considering the code $2S_3$ instead, since its profile is $2(1,1,5) = (2,2,10)$, so that $d_{\text{free}} = 10$. And in fact, for $n = 4$ and $K = 3$ this is the largest possible free distance, since the Plotkin bound for these parameters gives $d_{\text{free}} \leq 10$. In general, for $(n,1)$, $K = 3$ ZRL codes, the largest possible d_{free} is $\lfloor \frac{5n}{2} \rfloor$, achieved by $\lfloor \frac{n}{2} \rfloor S_3 + (n \bmod 2) S_2$, whereas the best possible d_{free} among all codes, ZRL or not, is $\lfloor \frac{8n}{3} \rfloor$, achieved by $\lfloor \frac{n+1}{3} \rfloor (1 + x^2) + \lfloor \frac{2n+1}{3} \rfloor (1 + x + x^2)$. The ratio of these two values approaches 16/15 as $n \to \infty$, and the smallest value of n for which these two values differ by as much as two is $n = 9$, where the best ZRL code $4S_3 + S_2$ has $d_{\text{free}} = 22$, but the code with g.p. list $(3(1 + x^2), 6(1 + x + x^2))$ has $d_{\text{free}} = 24$. However, even in this case the ZRL code may be competitive, since its T_{num} is

$$\frac{D^{22}}{1 - D^4 - D^6} = D^{22} + D^{26} + D^{28} + D^{30} + O(D^{32})$$

whereas the unrestricted code has

$$T_{\text{num}} = \frac{D^{24}(2 - D^6)}{1 - 3D^6 + D^{12}} = 2D^{24} + 5D^{30} + O(D^{36})$$

And indeed, an asymptotic analysis shows the rate of growth of the coefficients of $T_{\text{num}}(D)$ for the ZRL code to be $\approx (1.1577)^n$, whereas for the unrestricted code it is $\approx (1.1740)^n$. Thus, as discussed in [2], the ZRL code may perform better at low signal-to-noise ratios than the non-ZRL code.

VI. Summary

A class of convolutional codes, termed zero-run length (ZRL) convolutional codes, has been discovered for which the free distance can be computed by inspection, and for which there is a closed-form expression for the three-variable transfer function. This class of codes includes the superorthogonal codes introduced by Viterbi [1] and analogous "ultraorthogonal" codes introduced here. It has been found that, while ZRL codes are much more general than superorthogonal or ultraorthogonal codes, any ZRL code may be constructed as a combination ("sum") of superorthogonal and ultraorthogonal codes.

Although ZRL codes have very low rates for large constraint lengths, they are nevertheless interesting and potentially useful. Furthermore, many of the ideas developed here to analyze this class of specialized codes, such as the use of reduced state diagrams, might extend to other interesting code classes as well.

References

[1] A. J. Viterbi, "A New Class of Very Low Rate Convolutional Codes with Application to Spread Spectrum Multiple Access," preprint (April 1989).

[2] C.-C. Chao and R. J. McEliece, "On the Path Weight Enumerators of Convolutional Codes," Proc. 26th Ann. Allerton Conference, Univ. Illinois, pp. 1049–1058, October 1988.

[3] R. J. McEliece, *The Theory of Information and Coding*, Reading, Massachussetts: Addison Wesley, 1977.

[4] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, New York: McGraw-Hill, 1979.

[5] R. P. Stanley, *Enumerative Combinatorics, Vol. I*, Monterey, California: Wadsworth & Brooks Cole, 1986.

[6] R. J. McEliece, R. B. Ash, and C. Ash, *Introduction to Discrete Mathematics*, Boston: Random House, 1989.

[7] I. Onyszchuk, "Efficient methods for computing transfer functions for convolutional codes," in preparation.

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

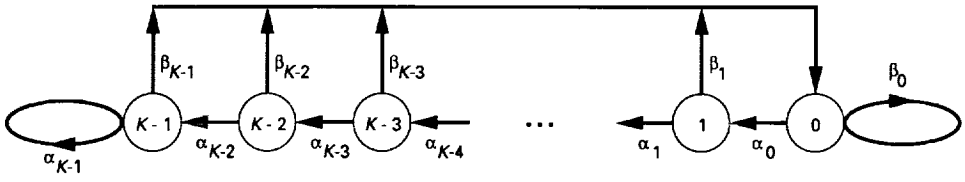


Fig. 1. Reduced state diagram for analyzing ZRL codes.

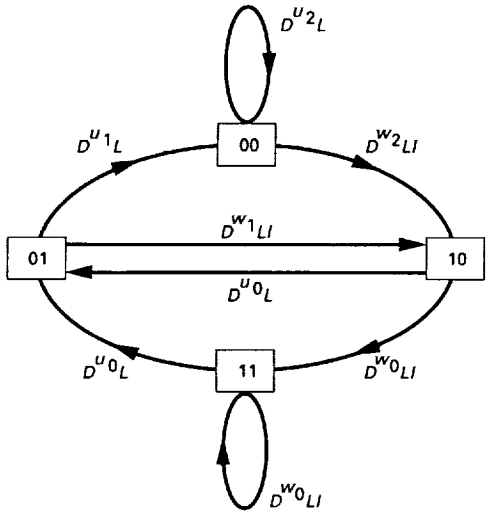


Fig. 2. The DIL state diagram for a K = 3 ZRL code.

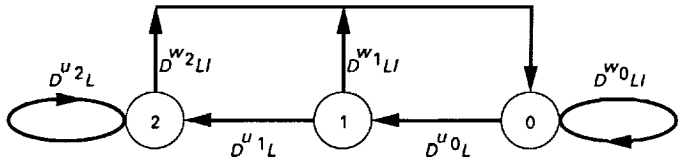


Fig. 3. The collapsed DIL state diagram for a K = 3 ZRL code (compare to Fig. 2).

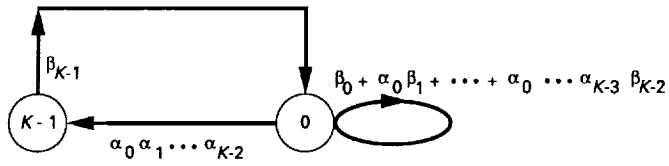


Fig. 4. The state diagram of Fig. 1, after the loop at state K−1 and the states 1, 2, . . ., K−2 have been eliminated.

78.

Quantization Effects in Viterbi Decoding Rate 1/n Convolutional Codes

I. M. Onyszchuk, K.-M. Cheung, and O. Collins
Communications Systems Research Section

A Viterbi decoder's performance loss due to quantizing data from the additive white Gaussian noise (AWGN) channel is studied. An optimal quantization scheme and branch metric calculation method are presented. The uniformly quantized channel capacity $C_u(q)$ is used to determine the smallest number of quantization bits q that does not cause a significant loss. The quantizer stepsize which maximizes $C_u(q)$ almost minimizes the decoder bit error rate (BER). However, a slightly larger stepsize is better, like the value that minimizes the Bhattacharyya bound. The range and renormalization of state metrics are analyzed, in particular for $K = 15$ decoders such as the Big Viterbi Decoder (BVD) for the Galileo mission. These results are required to design reduced hardware complexity Viterbi decoders with a negligible quantization loss.

I. Introduction

Theoretically, Viterbi decoding is a maximum-likelihood decoding algorithm for convolutional codes. In practice, the main performance loss results from quantizing input data with q bits. The decoder's hardware complexity and speed depend strongly upon q and the state metric register length ℓ. Therefore, these parameters must be chosen as the smallest values that do not cause a significant bit signal-to-noise ratio (E_b/N_0) loss. A constraint length $K = 15$ decoder performs double the computation of a $K = 14$ decoder, but requires about 0.1 dB less E_b/N_0 for a bit error rate (BER) of 0.005. Since part of the decoder's hardware complexity increases only linearly with q, even a 0.01-dB quantization loss is large. However, given that one must construct a fully parallel $K = 15$ (or $K = 7$) decoder, a slightly larger loss might be acceptable or required by hardware and speed constraints.

The uniformly quantized, additive white Gaussian noise (AWGN) channel capacity $C_u(q)$ is used to estimate the quantizer stepsize Δ and smallest q that result in a negligible loss. For each q, almost minimum BER occurs when Δ maximizes $C_u(q)$ or minimizes the Bhattacharyya bound γ. New methods are presented to minimize the state metric register length ℓ in bits. These estimates are verified by simulations of three codes : the constraint length $K = 7$, rate $R = 1/2$, NASA standard code; the new experimental $K = 15$, $R = 1/4$, Galileo code [1]; and the $K = 15$, $R = 1/6$, "2-dB" code [2].

These results are used to determine the best design parameters q, ℓ, and Δ for $K = 15$, rate 1/4 and rate 1/6 decoders. Using 6 quantization bits with 10-bit state metric registers would substantially reduce the Big Viterbi Decoder (BVD) [4] hardware complexity and allow the system clock frequency to decrease by a factor of 0.56 as compared to the current design, which has $q = 8$ and $\ell = 16$. Using $q = 5$ would cause an E_b/N_0 loss of 0.02 dB at a BER of 0.005 for the Galileo code or "2-dB" code, but there is no measurable E_b/N_0 loss when $q = 6$. Since the same losses occurred for 8-bit symbol error rates (SER), these results apply when an outer block code is concatenated with the convolutional code.

II. Branch Metrics

When an encoded 0 or 1 is mapped to +1 or −1, respectively, and then transmitted, the receiver's demodulator output is a conditionally Gaussian random variable y with mean $+m$ or $-m$ and the same variance $\sigma^2 = m^2/(2RE_b/N_0)$ as the zero-mean AWGN channel noise. (This holds for binary phase shift-keyed [BPSK] signaling with ideal coherent detection.)

For the AWGN channel, a Viterbi decoder finds the trellis path with minimum Euclidean distance (or equivalently, minimum negative inner product) to the received sequence. Thus, each trellis branch metric is the inner product of the length n branch label (with 0 and 1 replaced with +1 and −1) and the negative of a received vector $[y_1, y_2, \ldots, y_n]$. Hence, the decoder adds $-y_i$ or $+y_i$ (equivalently $(-y_i + |y_i|)/2$ or $(y_i + |y_i|)/2$ when σ is fixed, because incrementing or multiplying all branch metrics by a constant does not change the decoder's output) to the metrics of those branches with a +1 or −1 in position i. Therefore, the decoder may add $|y_i|$ to the metrics of branches having different signs in position i than that of y_i, and zero otherwise. This sign-magnitude method is used throughout this article because it halves the branch and state metric maximum ranges, as compared to using standard integer metrics [3,4]. For example, using this method in the Scarce-State $K = 7$, rate 1/2 decoder [5] would substantially decrease the chip circuitry.

III. Quantization

When zeros and ones are equally likely in the encoder input data,

$$\Pr(|y| = x) = \frac{1}{2}\Big[\Pr(y = x \mid +1) + \Pr(y = -x \mid +1)\Big] + \frac{1}{2}\Big[\Pr(y = x \mid -1) + \Pr(y = -x \mid -1)\Big]$$

$$= \Pr(y = x \mid +1) + \Pr(y = x \mid -1)$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[e^{-(x-m)^2/2\sigma^2} + e^{-(x+m)^2/2\sigma^2}\right]$$

In this article, $m = 0.84$ volts. The probability distribution function of $|y|$ (Fig. 1) suggests that more quantization levels are required for the $K = 15$ codes operating near 0 dB (high noise variance) than the NASA code at $E_b/N_0 = 2.25$ dB.

Let the random variable J be the quantized value of y and for $-2^{q-1} - 2 \le j \le 2^{q-1} - 2$ define

$$p_j = \Pr(J = j \mid +1) = \frac{1}{\sqrt{2\pi}\sigma}\int_{(j-0.5)\Delta}^{(j+0.5)\Delta} e^{-(y-m)^2/2\sigma^2}\,dy$$

For $j = \pm(2^{q-1} - 1)$, p_j is the above integral with limits $(j - 0.5)\Delta$ and $+\infty$, or $-\infty$ and $(j + 0.5)\Delta$.

Since $|J_1|, \ldots, |J_n|$ are summed to form branch metrics, the absolute error $|J_i - y_i|$ in quantizing y_i is also the contribution to the branch metric error incurred. A decoder using signed integers to represent J_i could conceptually use $0, \pm\Delta, \pm2\Delta, \ldots, \pm(2^{q-1} - 1)\Delta$ for any real number Δ, because multiplying all metrics by Δ has no effect. Therefore, the quantizer thresholds should be uniformly spaced Δ volts apart at $\pm\Delta/2, \pm3\Delta/2, \ldots, \pm(2^{q-1})\Delta/2$, because this minimizes the metric error defined above (and also any positive function of $J_i - y_i$). Thus, only uniform quantization schemes, characterized by q and Δ, are considered herein. (Several simulations of the NASA code using 3-bit integer branch metrics and nonuniform quantization schemes never produced lower BERs than using the best Δ).

For $q = 3$, J_i is normally one of 7 values from −3 to +3, so quantizer levels +4 and −4 are appended (Fig. 2) to decrease the BER near that for 8 levels and standard integer metrics. Thus, the maximum magnitude of J, $2^{q-1} - 1$, will be replaced by 4 instead of 3 for $q = 3$ throughout this article. In rate 1/2 decoders, a branch metric of 8 is decreased to 7 so that $q = 3$ bits still represent all possible values. Since $\Pr(|J_i| = j) = p_j + p_{-j}$, this event occurs with probability $(p_{+4} + p_{-4})^2$, which is only 0.11 for the NASA code at $E_b/N_0 = 2.25$ dB.

IV. Quantizer Stepsize

Ideally, the uniform quantizer stepsize Δ should minimize BER and SER over the decoder's operating range of channel noise levels. In practice, a Δ which almost minimizes the BER for the lowest expected E_b/N_0 will also nearly minimize both the BER and SER when E_b/N_0 increases by up to 1 dB. Simulations (described later) indicate that the Δ that maximizes channel capacity is near optimum.

Since the binary-input quantized AWGN channel is symmetric, capacity is achieved with equiprobable inputs:

$$C_u(q) = \sum_{j=-2^{q-1}+1}^{2^{q-1}-1} p(j \mid +1) \log_2 \left[\frac{2p(j \mid +1)}{p(j \mid +1) + p(j \mid -1)} \right]$$

$$= 1 - \sum_{j=-2^{q-1}+1}^{2^{q-1}-1} p_j \log_2 \left(1 + \frac{p_{-j}}{p_j} \right)$$

bits per channel use. Figure 3 shows how rapidly the maximum possible uniformly quantized AWGN channel capacity $C_u(q)$ approaches its limit for several noise variances; $C_u(3)$ is based upon the 9-level quantizer in Fig. 2 instead of using 8 or 7 levels. A $q = 4$ or $q = 5$ quantizer has 15 or 31 levels, respectively. The data points in Fig. 2 indicate $C_u(q)$ for integer values of q. The lines between data points are channel capacities when uniform quantizers have intermediate numbers of levels, such as 24.

The curves in Fig. 3 show that there is negligible capacity gain for $q > 6$, and in fact $C_u(5)/C_u(\infty) \geq 0.9975$ suggests that there will be a very small loss for $q = 5$. Figure 4 shows how the performance of the NASA code at $E_b/N_0 = 2.25$ dB varies with q and Δ. Observe that the minimum BER for $q = 5, 4$, and 3 increases roughly in proportion with the decrease in capacity. Also, for $q = 5$, there is a negligible loss and the BER increases extremely slowly for Δ greater than the optimum. Therefore, for $q \geq 4$, it is important to choose Δ larger instead of smaller than the best value. The labels C and γ in Fig. 4 indicate the stepsizes that respectively maximize $C_u(q)$ and minimize the Battacharyya bound parameter

$$\gamma = \sum_{j=-2^{q-1}+1}^{2^{q-1}-1} \sqrt{p_j p_{-j}}$$

which is a measure of the channel noise level: near 0 for high E_b/N_0 and approaching 1 for very noisy channels. The Δ that minimizes γ is the safest choice because it is slightly larger than the stepsize which minimizes BER. Also, minimizing γ yields the lowest BER for $q = 3$ with 9 quantizer levels (Fig. 4). Finally, the corresponding 8-bit SER curves are not shown because they have the same relative shape and spacing as the BER curves in Fig. 4. Many sets of software simulations were run for the NASA code and the Galileo code. The values of q were 3, 4, 5, or 6 and E_b/N_0 ranged from 0 dB to 3.5 dB.

In all simulations, the Δs which maximize $C_u(q)$ or minimize γ were, respectively, slightly smaller or larger than the Δ that minimized BER. For $q = 3$ or 4, the Δs which minimize the quantizer mean-square error or absolute error were too large.

The simulations in Fig. 5 for the $K = 15$ codes show that using $q = 5$ or 4 costs 0.02 dB or 0.05 dB at the BER of 0.005 required for images. These E_b/N_0 quantization losses are the same when the Viterbi decoder output becomes the input to an outer block decoder, because the 8-bit symbol error rate curves are spaced the same distance apart as the BER curves. In all simulations, the uniform spacing Δ was chosen to minimize γ.

V. State Metric Renormalization

For each received n-vector and encoder state, a Viterbi decoder finds the trellis path with least total branch metrics into the state. Since the state metrics are stored in ℓ-bit registers, occasionally they must all be decreased to avoid overflow. This renormalization can be accomplished by zeroing every register's most significant bit (msb), which is equivalent to subtracting $2^{\ell-1}$ from every metric if all registers have msb = 1. However, detecting when all 2^{K-1} metrics simultaneously have msb = 1 is impractical for a $K = 15$ decoder such as the BVD.

At each trellis level, let the random variable M be the difference between the maximum and minimum state metrics. If any state metric is $\geq 2^{\ell-1} + 2^{\ell-2}$, (its two most significant bits are 1) and $M < 2^{\ell-2}$, then all metrics are $\geq 2^{\ell-1}$, so every msb = 1. In the BVD, $\ell = 16$ was chosen to guarantee that $2^{\ell-2} > M$, and so a single state metric is monitored and renormalization occurs when the two most significant bits are 1. The following improved method should be used when ℓ is reduced so that $M \geq 2^{\ell-2}$. Let W be the maximum of the metrics of the all-zeros state, the all-ones state, and the state with a one input followed by $K - 2$ zeros. Since most state metrics differ from one of these three metrics by only a few $|J_i|$ contributions, W is close to the largest state metric

(Galileo code simulations verified this). Therefore, renormalization could occur when W exceeds a threshold such as $2^{\ell-1}+2^{\ell-2}+2^{\ell-3}$. If more metrics are monitored, the threshold can be set closer to $2^{\ell}-1$ because W will be closer to the largest state metric.

Definition. Let D be the maximum, over all nonzero states s, of the least-weight trellis path from the all-zeros state into state s.

Lemma. $M \leq D(2^{q-1}-1)$

Proof. Let b and w be the states with lowest and highest metrics. Since a convolutional code is linear, there exist two trellis paths from some state c, one into state b and the other into state w, whose branch labels differ in D or fewer positions. Since the maximum contribution to a branch metric by one J_i is $2^{q-1}-1$, the state metric of w is at most the state metric of b plus $D(2^{q-1}-1)$.

Corollary. In the absence of noise,

$$M = M_0 = D \cdot j_m$$

where $j_m = \lfloor 0.5 + m/\Delta \rfloor$ is the quantizer output when $+m$ volts is input.

For nonsystematic codes, D is near d_{free} and usually much less than $n(K-1)$, the maximum possible. The NASA code has $d_{\text{free}} = 10$, $D = 8$, and $n(K-1) = 12$. Since $D = 33$ for the Galileo code, $M_0 = 132$ for $q = 5$, $\Delta = 0.20$, and $m = 0.84$. Since $D = 50$ for the rate 1/6 "2-dB" code, $M_0 = 200$. Simulations for the Galileo and NASA codes show that M_0 is an upper bound on the mean of M when the channel is noisy and $2M_0$ is always greater than M.

As in the $q = 3$ case where levels $+4$ and -4 were adjoined, a rule for limiting branch metrics may be derived by computing their probabilities. Define

$$m(x) = p_0 + \sum_{j=1}^{2^{q-1}-1} (p_j + p_{-j})x^j$$

Then $\Pr(|J_i| = j) = \{m(x)\}_j$, the coefficient of x^j in $m(x)$. Since the largest possible branch metric is the sum of n independent values $|J_i|, \ldots, |J_n|$, it equals t with probability $\{[m(x)]^n\}_t$. Thus M could be reduced by limiting branch metrics.

Claim. $\Pr(M \geq t) \leq \sum_{i=t}^{D(2^{q-1}-1)} \left\{[m(x)]^D\right\}_i$

where the subscript i denotes the coefficient of x^i in the polynomial within the braces.

Proof. Let b and w be the states with lowest and highest metric. An upper bound on $\Pr(M=t)$ is obtained by considering the worst possible case: the survivor path for state w differs from that for state b in exactly D positions, and in these positions, the survivor path branch labels of w have a different sign than the received J_i. Then $\Pr(M = t)$ is the coefficient of x^t in $[m(x)]^D$.

To achieve a particular (very low) probability, t must be unrealistically large since the above bound is not very tight. This is fine, because t could be chosen as the least power of 2 such that $\Pr(M \geq t) \leq 10^{-5}$. Then setting $\ell = 1 + \log_2 t$ results in no loss of performance.

The current BVD design has $q = 8$ and $\ell = 16$ to accommodate $M \leq n(K-1)(2^{q-1}-1)$ and two extra bits for renormalization. This results in full maximum-likelihood decoder performance. However, using $q=6$, $\Delta = 0.14$, and $\ell = 10$ for the BVD operating at 0 to 1 dB E_b/N_0 would not increase the BER or SER detectably, but would reduce the decoder hardware. Furthermore, the system clock frequency and thus timing constraints would be reduced by a factor of 10/18.

When $\ell < 1 + \log_2[D(2^{q-1}-1)]$, then occasionally a state metric may overflow, whereupon it is immediately decreased by 2^{ℓ}, instead of $2^{\ell-1}$ at the next renormalization. Protecting against overflow is important because a state with a high metric might suddenly become one of the best states, causing the decoder to make wrong decisions. This can be avoided by setting state metrics that overflow equal to all ones $(2^{\ell}-1)$. Then states with very high metrics remain this way even after renormalization so they do not affect the decoder's output. An underflow is the event that occurs at renormalization when a state metric has msb $= 0$, in which case the metric is effectively increased by $2^{\ell-1}$. Rarely, underflows may occur because it is infeasible to continuously check all 2^{K-1} state metrics to find the least value. In conclusion, overflows can be prevented by extra hardware, but underflows will occasionally happen. In practice, always examining several state metrics gives a good approximation of the current metric size and range M. Hence, renormalization can take place so that overflows and underflows occur with very low probability.

Myth. When state metrics overflow or underflow, the decoder fails completely.

One million decoded bit simulations for $q = 5$ and 4 with short state metric registers having $\ell = 9$ and 8 bits,

respectively, yielded the same results as in Fig. 5, because the odd underflow or overflow that occurred did not significantly affect the output. This follows from the Viterbi decoder's robustness and tolerance of occasional state metric disruptions. Further shortening of the state metric registers to 8 and 7 bits resulted in a graceful BER increase, as though q was being decreased. This behavior is expected because the overall trellis path metric resolution is the decoder parameter, affected by input quantization, which influences decisions.

References

[1] J. Statman, G. Zimmerman, F. Pollara, and O. Collins, "A Long Constraint Length VLSI Viterbi Decoder for the DSN," *TDA Progress Report 42-95*, vol. July–September 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 134–142, November 15, 1988.

[2] J. H. Yuen and Q. D. Vo, "In Search of a 2-dB Coding Gain," *TDA Progress Report 42-83*, vol. July–September 1985, Jet Propulsion Laboratory, Pasadena, California, pp. 26–33, November 15, 1985.

[3] J. A. Heller and I. M. Jacobs, "Viterbi Decoding for Satellite and Space Communication," *IEEE Trans. Commun. Tech.*, vol. COM-19, pp. 835–848, October 1971.

[4] G. C. Clark and J.B. Cain, *Error-Correction Coding for Digital Communications*, New York: Plenum Press, 1981.

[5] T. Ishitani, K. Tansho, N. Miyahara, and S. Kato, "A Scarce-State-Transition Viterbi-Decoder VLSI for Bit Error Correction," *IEEE Journal of Solid-State Circuits*, vol. SC-22, pp. 575–581, August 1987.

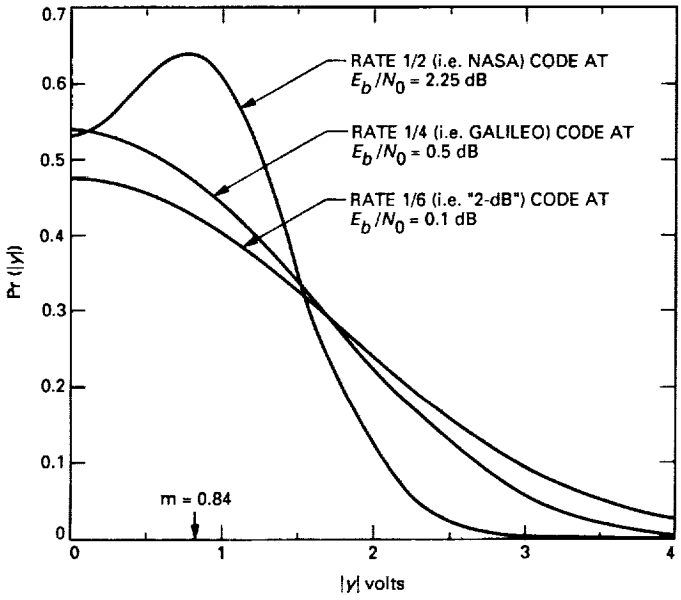


Fig. 1. Received signal magnitude distribution.

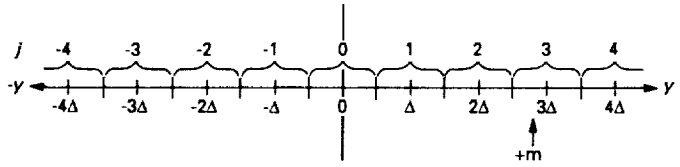


Fig. 2. Optimal quantization for 3-bit branch metrics.

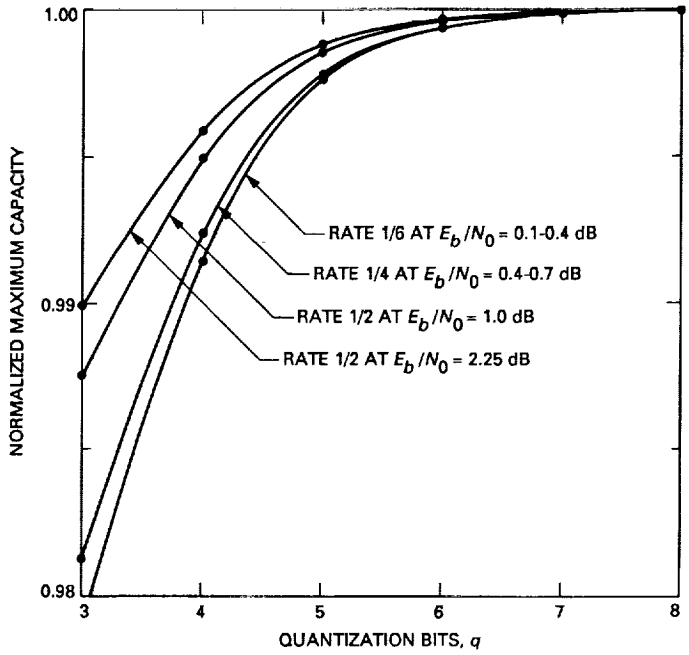


Fig. 3. Uniformly quantized AWGN channel capacities.

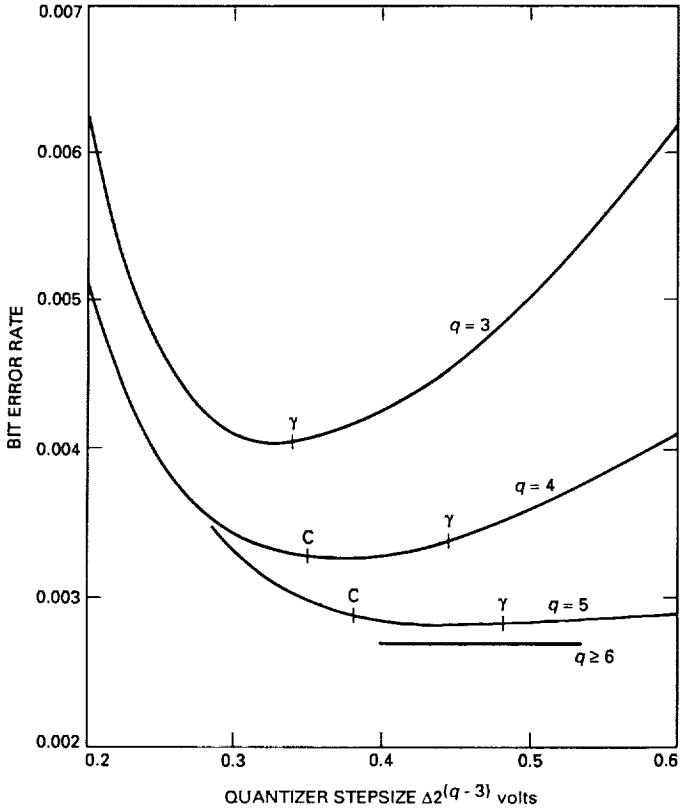


Fig. 4. Performance of the NASA code on the uniformly quantized AWGN channel.

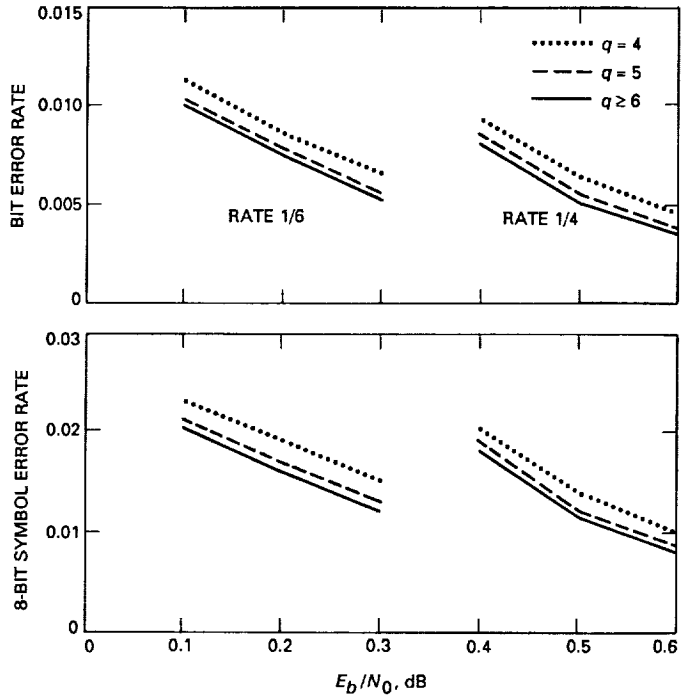


Fig. 5. $K = 15$ code simulations.

Big Viterbi Decoder (BVD) Results for (7,1/2) Convolutional Code

J. Statman, J. Rabkin, and B. Siev
Communications Systems Research Section

The Big Viterbi Decoder (BVD), capable of decoding convolutional codes with constraint lengths of up to 15, is under development for the DSN. As part of the development, a commercial single-chip (7,1/2) Viterbi decoder is used to enable early start of system integration. Tests of the integrated partial system (including simulator, input interfaces, output interfaces, and computer controls) were recently completed at the DSN Compatibility Test Area (CTA-21) at JPL. This article describes the system elements used for the demonstration and test results.

I. Introduction

The Big Viterbi Decoder (BVD) is under development for the Deep Space Network (DSN) [1]. It is intended to provide up to 1.8 dB improvement in link margin through the use of convolutional codes with larger constraint lengths. Specifically, the BVD is designed to operate with any convolutional code with constraint length of $K \leq 15$ and code rates $1/2, 1/3, \ldots, 1/6$. In contrast, the current equipment is designed for the standard DSN code, $K = 7$ and rate 1/2. The BVD prototype will be used in a May 1991 demonstration in conjunction with the Galileo mission. Following a successful demonstration, the decoder will be inserted into the DSN for use with Galileo and future missions.

A block diagram of the BVD, as initially planned, is shown in Fig. 1(a). The core of the decoding processing is performed in the Processor Assembly using 256 or 512 identical custom VLSI chips. All the other functions are performed in the Controller Assembly. These include transforming of input soft symbols, buffering data to the Processor Assembly, interfacing to output devices, and providing full self-test capability. Early on it was recognized that the Processor Assembly, especially the VLSI chips, is the time-critical element in the development schedule since no meaningful DSN-compatibility tests could be conducted without having a "decode" capability, which requires a full Processor Assembly. To overcome this bottleneck, a secondary "decode" path was introduced as shown in Fig. 1(b). The additional path uses a commercial QUALCOMM Q1401 (7,1/2) decoder chip and enables testing of many BVD functions well before the Processor Assembly is ready. A partial BVD, shown in Fig. 1(c), was completed and tested in the laboratory and in the DSN Compatibility Test Area (CTA-21) at JPL,

verifying decoder operation for (7,1/2) code and current DSN interfaces.

II. Functional Block Description

The system under test consists of a MULTIBUS I chassis with six boards: an Intel 80386/21 CPU and five custom digital boards. Figure 2 shows a more detailed functional block diagram of the five custom boards. These boards include functions required in DSN operation, as well as functions needed during development and testing. Some of the latter are especially critical during diagnostics and fault isolation, enabling failures to be traced to a specific board. The following is a description of these functions, by board.

A. Memory Board

This board includes 1 Mbyte of Electrically Erasable Programmable Read-Only Memory (EEPROM) and is used for object code and key parameter storage. It enables the BVD to accept program updates without a major interruption: the new code is downloaded from an IBM PC/AT computer into RAM and stored in the EEPROM. Upon reset, the program is read from the EEPROM into RAM and executed. This approach eliminates the operational problems associated with removing boards and replacing EPROMs (which must be erased under an ultraviolet light). The board also stores critical mission parameters to allow easy restart after a power glitch. If additional memory is required, multiple boards can be installed.

B. Encoder Board

The encoder (simulator) board provides a flexible source for encoded data test sequences for BVD self-test. In operation, a fairly comprehensive self-test with several million bits can be run at 1.1 Mbit/sec, requiring only a few seconds. The board includes an uncoded data source, encoder, and circuitry that gets calibrated noise samples from an external noise source. All functions are fully programmable. Simulated symbols are generated by passing bits from a programmable sequence generator through an encoder, and summing them with properly scaled noise. The encoder is implemented as a computer-loaded RAM, where each state of the encoder shift-register corresponds to an address of the RAM. This allows implementing of any convolutional encoder for $K \leq 15$ and $1/n$, $n \leq 6$ through a single computer loading of the RAM.

Other functions that the board performs are monitoring the mean and variance of simulated noise samples (received from an external source), generation of simulated bit and symbol clocks, and transfer of simulated bits to the comparator board.

C. SSA/BBA Interface Board

This board includes circuitry that generates system clocks, interfaces to the Symbol Synchronizer Assembly/ Baseband Assembly (SSA/BBA) for input symbols, reclocks input signals, and interfaces to the Time Code Translator (TCT) for time tagging. It also provides a simulated symbol RAM that allows testing of the SSA/BBA interface.

D. SNR Estimator Board

The circuitry on this board collects data for symbol signal-to-noise ratio (SNR) estimation. The approach is similar to that used in the Symbol Stream Combiner (SSC)[1] and requires computation of sum-of-squares and sum-of-absolute-values of the received symbols. Computing symbol SNR internal to the BVD provides an important diagnostic tool as well as a monitor of BVD health during real-time operation.

The board includes a scale RAM and an alternate symbol sign-flipper to allow for computer-controlled adjustment of these symbol attributes. The scale RAM allows one to scale the input symbols by a constant, which also enables testing for possible inversion in the SSA/BBA (the Galileo (15,1/4) code is nontransparent). The alternate symbol sign-flipper is required to compensate for the sign flipping used in encoders on board most JPL spacecraft. Another feature on this board is a coded data test RAM that allows the CPU to read and store symbols from either the SSA/BBA (in real time) or the simulated signal generated by the encoder board.

The board also includes provisions for future interfaces to the DSN through a First-In, First-Out (FIFO) buffer. This interface will be utilized when the new Telemetry Processor Assembly (TPA), currently under development, is sufficiently defined.

E. Comparator Board

The comparator board includes a (7,1/2) Viterbi decoder, a comparator that allows bit error rate (BER) data

[1] S. Dolinar, "A Lot of Things You Always Wanted to Know About Signal-to-Noise Estimation Methods (but Didn't Bother to Ask)," JPL Interoffice Memorandum 331-85.2-109 (internal document), Jet Propulsion Laboratory, Pasadena, California, January 22, 1986.

collection, symbol rate estimation circuitry, and various output interfaces.

The (7,1/2) Viterbi decoder section is centered on a QUALCOMM Q1401 commercial chip, and includes additional circuitry to provide for computer-controlled node synch. The Q1401 is designed for the Goddard (7,1/2) code, which uses the same polynomials as the DSN code but in a reversed order. A circuit that allows the BVD to operate with either polynomial order had been designed but had not been installed during the tests. It will be tested separately later.

The comparator function compares the decoded bit stream to a delayed version of the simulated unencoded bit stream. It includes a variable-delay buffer (used to align the two streams) and a set of counters that allows for collection of data for BER versus E_b/N_0 evaluation.

The output interface circuit receives decoded bits and status data from the (7,1/2) decoder chip or from the Processor Assembly, records these in a test RAM, and channels the data to external devices through a Frame Synchronization Subsystem (FSS) driver or a FIFO. The FIFO will be used for interface to the new TPA. An optional differential decoder is also part of the output circuit.

F. Other Boards

To complete the BVD, several more boards will be designed and manufactured. The Controller Assembly will have two more boards: the Processor Assembly Interface board and the Node Synch board. The Processor Assembly will have seventeen boards mounted in a complex custom backplane. The backplane is being designed and manufactured by Teradyne Connector Systems to JPL specifications, while the seventeen boards are being designed by JPL. Sixteen of these boards will be identical and will house custom VLSI chips, while the seventeenth will perform traceback and interface functions.

III. Laboratory and CTA-21 Test Results

Tests were conducted in the laboratory and in CTA-21 to validate BVD operation and compatibility with the DSN. It is important to note that as DSN interfaces are upgraded, similar tests will need to be conducted with the new interfaces. However, the interfaces validated here are the minimal set needed for the planned 1991 demonstration in DSS-14, namely symbol input from SSA/BBA and command/control through an RS232 to a terminal.

During the CTA-21 test, the BVD was connected as part of a telemetry string (Fig. 3). The test was conducted at high SNR and the BVD was connected to the station's SSA, replacing the current maximum-likelihood convolutional decoder (MCD). Successful decoding was demonstrated for several input sequence formats. The test used the Goddard code, i.e., reverse order for code polynomials, and will be repeated when the circuit is modified to handle both the DSN and Goddard codes. However, the key objective of verifying DSN compatibility was achieved.

IV. Conclusions

A partial BVD was constructed that has the capability to decode (7,1/2) convolutional codes. It was demonstrated in the laboratory and in the CTA-21 environment, validating DSN compatibility for the planned May 1991 demonstration.

Reference

[1] J. Statman, G. Zimmerman, F. Pollara, and O. Collins, "A Long Constraint Length VLSI Viterbi Decoder for the DSN," *TDA Progress Report 42-95*, vol. July–September 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 134–142, November 15, 1988.

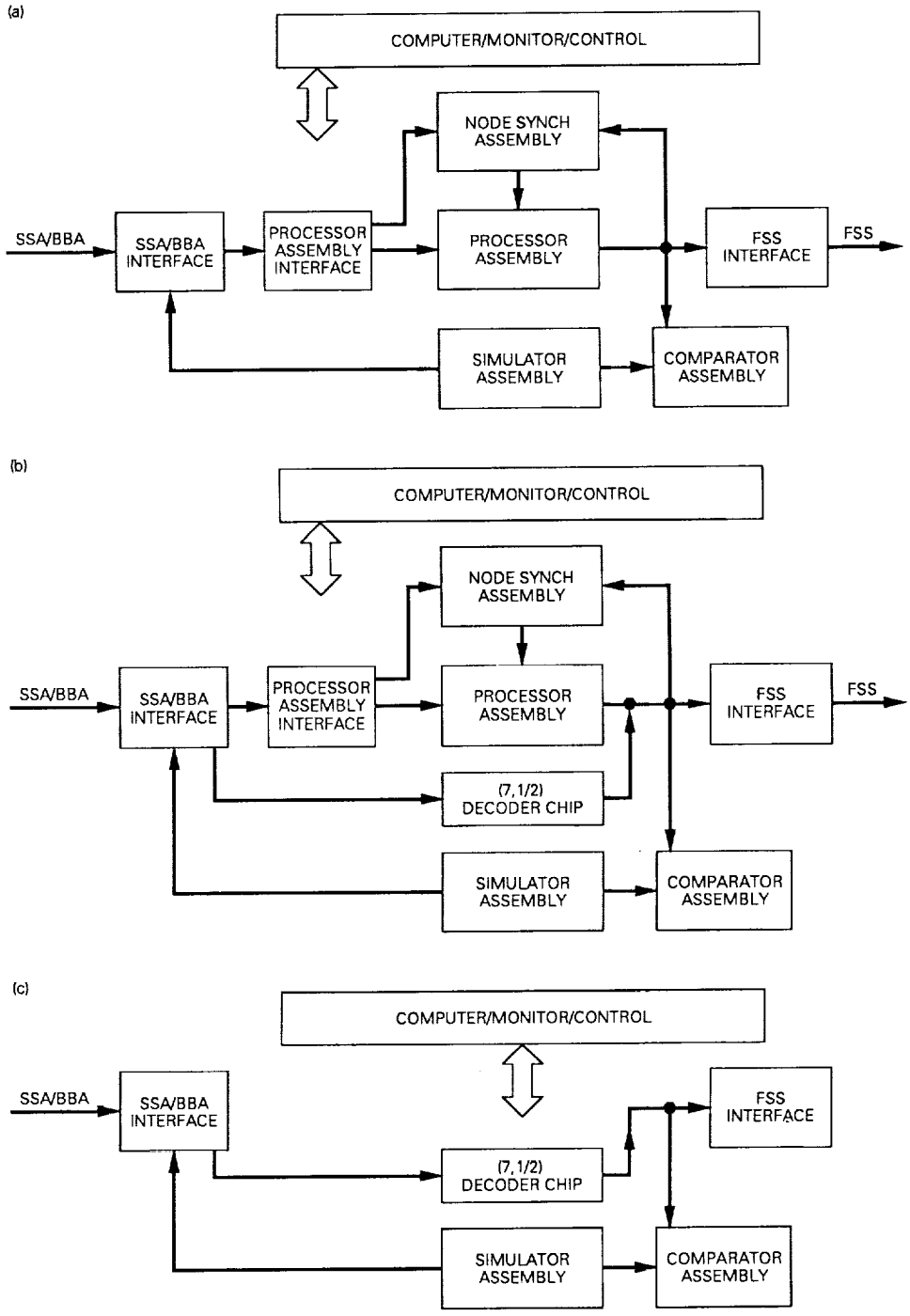
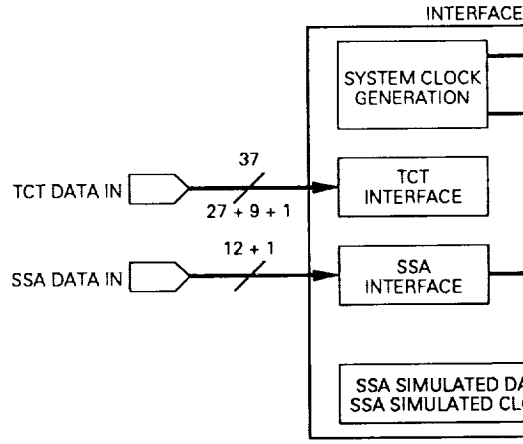
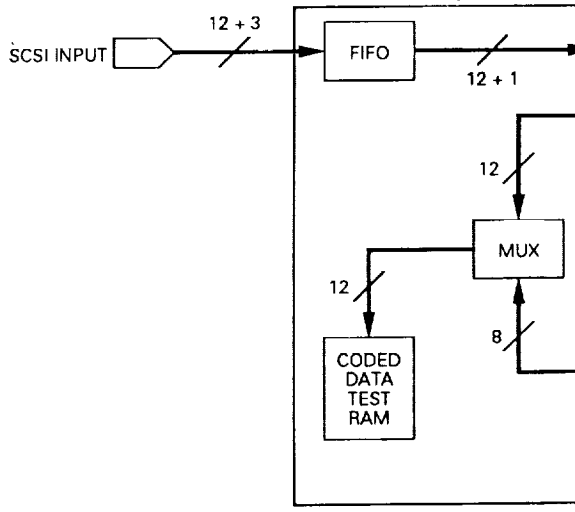
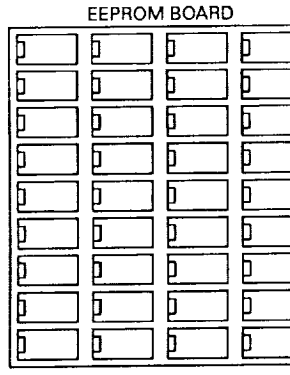


Fig. 1. Functional block diagram of Big Viterbi Decoder: (a) original, (b) modified to include (7,1/2) decoder chip, (c) during test.

FOLDOUT FRAME 1,







FOLDOUT FRAME 2

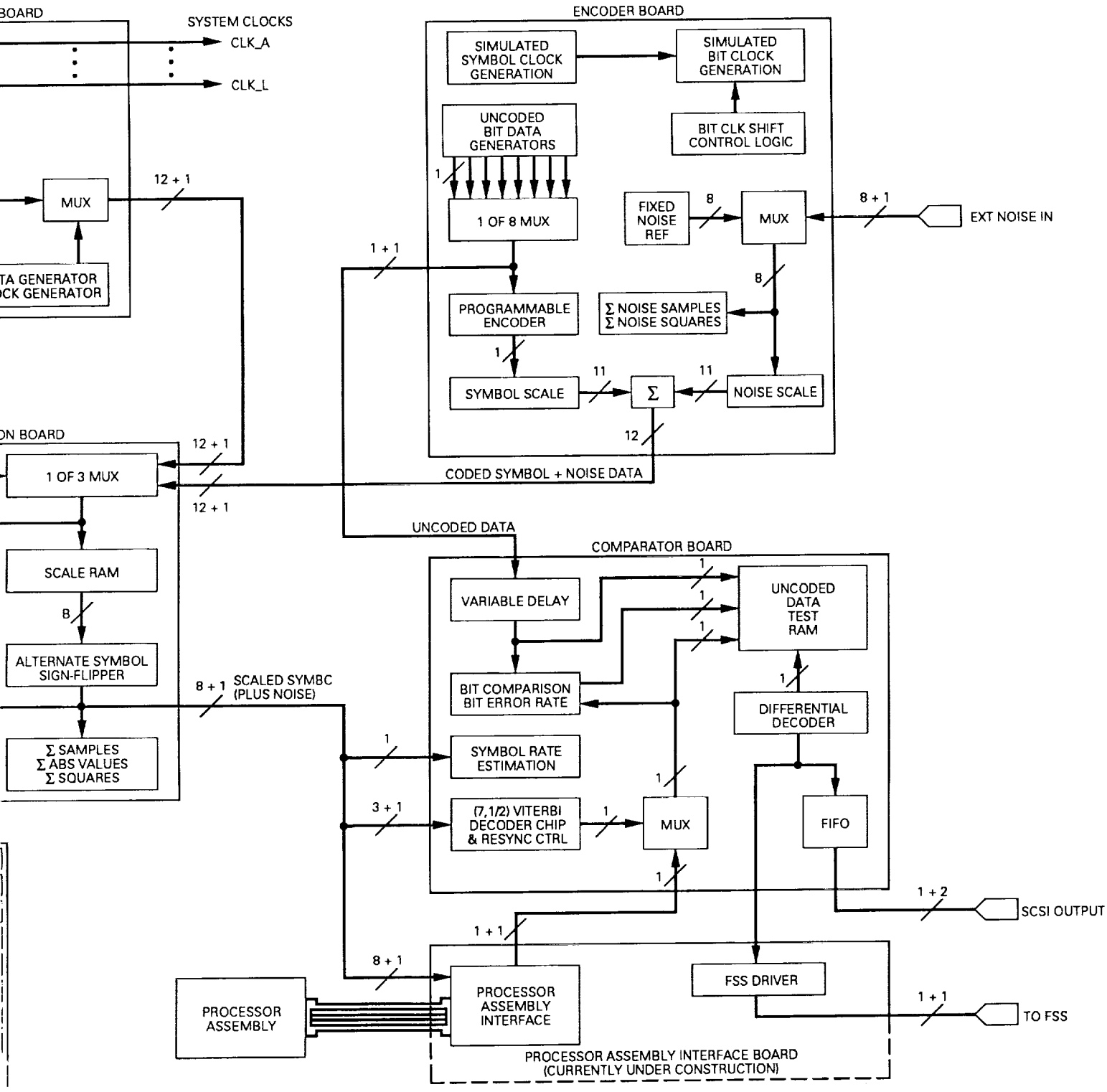


Fig. 2. Detailed block diagram of Big Viterbi Decoder.

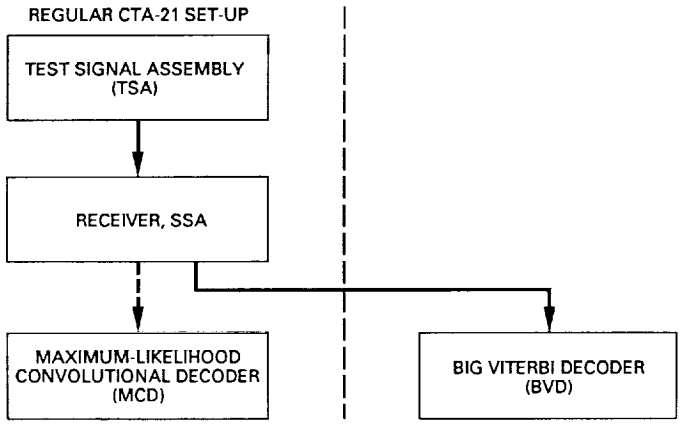


Fig. 3. CTA-21 test setup.

N90-19447

TDA Progress Report 42-99

November 15, 1989

Fast Transform Decoding of Nonsystematic Reed-Solomon Codes

T. K. Truong and K.-M. Cheung
Communications Systems Research Section

I. S. Reed
University of Southern California, Department of Electrical Engineering

A. Shiozaki
Osaka Electro-Communication University, Osaka, Japan

This article considers a Reed-Solomon (RS) code to be a special case of a redundant residue polynomial (RRP) code, and presents a fast transform decoding algorithm to correct both errors and erasures. This decoding scheme is an improvement of the decoding algorithm for the RRP code suggested by Shiozaki and Nishida [1], and can be realized readily on VLSI chips.

I. Introduction

Classes of redundant residue polynomial (RRP) codes were introduced first in [3,4]. These codes are constructed by use of the Chinese remainder theorem for polynomials over a finite field $GF(q)$. The codeword symbols of the RRP codes are expressed as polynomials over this field. The RRP codes can correct t error symbols with the aid of $2t$ redundant symbols.

Reed-Solomon (RS) codes constitute a subclass of RRP codes and are used in many sectors of today's industry. Some examples are the (255,223) 16-error-correcting RS code (NASA code) used in deep-space communications, the (31,15) 8-error-correcting RS code (JTIDS code) used in military communications, and the Cross Interleaving RS code (CIRC code) used in the compact-disc industry.

As Shiozaki [5] points out, by using the Chinese remainder theorem together with the Euclidean algorithm, an RRP code can be decoded without solving the error-locator polynomial and the error-evaluator polynomial. The decoder developed in [5] is a general frequency-domain implementation type depicted in the second block diagram in Fig. 9.2 of [2]. The advantage of the decoder in [5] over the decoder in [2] is that both the recursive extension and the inverse transform can be replaced by a single polynomial division. However, the method proposed by Shiozaki has the disadvantage that the reconstruction of the corrupted information polynomial $F'(x)$ from the received

symbols involves n polynomial multiplications in $GF(q)$, followed by the operation modulo $M(x)$, where n is the codeword length and $M(x)$ is a product of n polynomials. These operations severely lower the decoding speed.

This article considers RS codes to be a special case of the RRP codes and proposes to decode RS codes by the use of both the Fermat number transform [6,7] and the Euclidean algorithm. The Fermat number transform (FNT) eliminates polynomial multiplications and reduces the number of multiplications needed to reconstruct $F'(x)$ to $n \log_2 n$. The fast transform decoding scheme proposed in this article is faster than the decoding algorithm in [5].

II. Some Preliminaries on Finite Fields and the Fast Fermat Number Transform

Given that $GF(q)$ is a finite field, let $GF(q)[x]$ be the ring of polynomials over $GF(q)$.

Definition 1. The two polynomials $m_1(x)$ and $m_2(x)$ over $GF(q)$ are said to be relatively prime if and only if the greatest common multiple of $m_1(x)$ and $m_2(x)$ is a constant in $GF(q)$.

Definition 2. The two polynomials $m_1(x)$ and $m_2(x)$ over $GF(q)$ are said to be congruent modulo $m(x)$, i.e., $m_1(x) \equiv m_2(x) \bmod m(x)$ if and only if $m(x)$ divides $m_1(x) - m_2(x)$.

The Chinese remainder theorem is presented here for convenience; proof can be found in [11]. Let $M(x) = \prod_{i=1}^{r} m_i(x)$ be a product of pairwise relatively prime polynomials. Let $A_1(x), A_2(x), \ldots, A_r(x)$ be any r polynomials such that $\deg[A_i(x)] \leq \deg[m_i(x)], i = 1, 2, \ldots, r$. Finally, let $t_i(x)$ satisfy

$$\frac{M(x)}{m_1 x} t_i(x) \equiv 1 \bmod m_i(x) \quad \text{for } (i = 1, 2, \ldots, r)$$

There then exists one and only one polynomial $f(x)$ of $GF(q)[x]$ of degree satisfying $\deg[f(x)] \leq \deg[M(x)]$, which uniquely solves the system of congruences:

$$f(x) \equiv A_i(x) \bmod m_i(x)$$

The polynomial $f(x)$ is given by

$$f(x) \equiv \sum_{i=1}^{r} \frac{M(x)}{m_i(x)} t_i(x) A_i(x) \bmod M(x) \qquad (1)$$

Let $GF(q)$ be a finite field, let n be a number that divides $q - 1$, and let γ be a primitive nth root of unity. Define $(a_i)_{i=0}^{n-1}$ to be a sequence of n elements from $GF(q)$. A discrete Fourier transform of this sequence of length n is defined by

$$A_k \equiv \sum_{i=o}^{n-1} a_i \gamma^{ki} \bmod q \quad \text{for } (k = 0, 1, \ldots, n-1) \qquad (2a)$$

The inverse discrete Fourier transform of A_k is defined by

$$a_i \equiv n^{-1} \left[\sum_{k=0}^{n-1} A_k \gamma^{-ik} \right] \bmod q \quad \text{for } (i = 0, 1, \ldots, n-1) \qquad (2b)$$

A direct computation of the transform in Eq. (2a) or its inverse transform in Eq. (2b) requires $n(n-1)$ multiplications.

When q is a Fermat prime, the Fermat number transform (FNT) over $GF(q)$ can be used. A Fermat prime F_m is defined by

$$F_m = 2^{2^m} + 1 \quad \text{for } (m = 1, 2, 3, 4)$$

$$F_m = 2^{2^m} + 1 \quad \text{for } (m = 1, 2, 3, 4)$$

It is shown in [6,7] that integer 3 is a primitive $l = 2^{2^m}$th root of unity in $GF(F_m)$. Next, let n divide 2^{2^m}. Finally, suppose γ is a primitive nth root of unity in $GF(F_m)$ where

$$\gamma = 3^{l/n}$$

The purpose of an FNT of length n is to compute efficiently the transform sequence $(A_k)_{k=0}^{n-1}$ using Eq. (2a). On the other hand, the inverse Fermat number transform (IFNT) of length n reconstructs the sequence $(a_i)_{i=0}^{n-1}$ from the sequence $(A_k)_{k=0}^{n-1}$ via Eq. (2b). Since the order of γ is a power of 2 in $GF(F_m)$, the length of the sequence to be transformed is a power of 2. As a consequence, the very efficient FNT can then be used to yield a fast transform [6] which is analogous to the fast Fourier transform (FFT). In this case, the number of multiplications involved in evaluating such a transform sequence of length n is $n \log_2 n$ [8]. A new type of Fermat number multiplier is developed in [9]. More details about the FNT can be found in [6].

III. Nonsystematic RS Codes

First, a set of RRP codes is defined. As shown next, these codes are constructed using the Chinese remainder theorem for polynomials over a finite field $GF(q)$. Let $m_0(x), m_1(x), \ldots$, and $m_{n-1}(x)$ be n relatively prime polynomials, and

$$M(x) = \prod_{i=0}^{n-1} m_i(x)$$

Assume that the degree of each $m_i(x)$ is d, and that kd information symbols $\underline{u} = (u_0, u_1, \ldots, u_{kd-1})$ are represented by the information polynomial as

$$F(x) = u_0 + u_1 x + \cdots + u_{kd-1} x^{kd-1}$$

where $u_i \epsilon GF(q)$ and $k < n$. Then an RRP code is the residue representation of $F(x)$, that is,

$$\underline{\nu} = (A_0(x), A_1(x), \ldots, A_{n-1}(x))$$

where $A_i(x) \equiv F(x) \bmod m_i(x)$ and $\deg[A_i(x)] < d$. By the Chinese remainder theorem, $F(x)$ can be recaptured from $A_i(x)$. The vector corresponding to the polynomial $A_i(x)$ is named the ith symbol. A code vector $\underline{\nu}$ can correct error symbols less than or equal to t symbols if $n - k \geq 2t$ [3,4].

The following shows that RS codes are a subclass of RRP codes. In order to facilitate the fast encoding and decoding procedures, which make use of the fast FNT methods as described in Section II, the codeword length n is required to be a power of 2.

Let $m_0(x), m_1(x), \ldots, m_{n-1}(x)$ be n relatively prime polynomials given by

$$m_i(x) = x - \gamma^i \quad \text{for } (i = 0, 1, \ldots, n-1)$$

Also let the k information symbols

$$(u_0, u_1, \ldots, u_{k-1}), u_i \epsilon GF(q)$$

be denoted by the information polynomial

$$F(x) = u_0 + u_1 x + \cdots + u_{k-1} x^{k-1}$$

Then the equations

$$F(1) \equiv F(x) \bmod m_0(x)$$

$$F(\gamma) \equiv F(x) \bmod m_1(x), \ldots,$$

and

$$F(\gamma^{n-1}) \equiv F(x) \bmod m_{n-1}(x)$$

lead to a code vector $\underline{\nu}$ represented by

$$\underline{\nu} = (A_0, A_1, \ldots, A_{n-1}) = (F(1), F(\gamma), \ldots, F(\gamma^{n-1}))$$

The code vector $\underline{\nu}$ is a nonsystematic RS codeword. It is not difficult to see that $\underline{\nu} = (A_0, A_1, \ldots, A_{n-1})$ is just the FNT of the sequence $(u_0, u_1, \ldots, u_{k-1}, 0, \ldots, 0)$ and the k information symbols $(u_0, u_1, \ldots, u_{k-1})$, i.e., $F(x)$ can be recaptured by an IFNT on the code vector $\underline{\nu} = (A_0, A_1, \ldots, A_{n-1})$.

On the other hand, since an RS code is a special case of an RRP code, the information polynomial $F(x)$ can be recaptured also from $\underline{\nu} = (A_0, A_1, \ldots, A_{n-1})$ by the use of the Chinese remainder theorem. Let $t_i(x)$ denote a polynomial that satisfies

$$\frac{M(x)}{m_i(x)} t_i(x) \equiv 1 \bmod m_i(x) \quad \text{for } (i = 0, 1, \ldots, n-1) \quad (3)$$

where

$$M(x) = \prod_{i=0}^{n-1} M_i(x) = \prod_{i=0}^{n-1} (x - \gamma^i)$$

Then the information polynomial $F(x)$ can be reconstructed as

$$F(x) \equiv \left[\sum_{i=0}^{n-1} \frac{M(x)}{m_{i(x)}} t_i(x) A_i \right] \pmod{M(x)}$$

IV. Decoding RS Codes

As Shiozaki et al. [1,5] point out, by using the Chinese remainder theorem together with the Euclidean algorithm, the RRP codes, which include the RS codes, can be decoded without solving the error-locator polynomial and the error-evaluator polynomial. However, that method has the disadvantage that the reconstruction of the corrupted information polynomial $F'(x)$ from the received symbols involves n polynomial multiplications in $GF(q)$ followed by the operation modulo $M(x)$. These operations can significantly lower the decoding speed. A modified decoding scheme, which makes use of the fast transform technique

to bypass the tedious polynomial multiplications and modulo $M(x)$ operation, is given in the Appendix.

A. Decoding for Correcting Errors

The overall decoding of nonsystematic RS codes for correcting errors using the Euclidean algorithm is summarized in the following (see the Appendix for details):

1. Compute the IFNT of the received code word $\underline{v}' = (A'_0, A'_1, \ldots, A'_{n-1})$ from Eq. (A-1) in the Appendix to obtain

$$F'(x) = u'_0 + u'_1 x + \cdots + u'_{n-1} x^{n-1}$$

in Eq. (A-3). Next, calculate the degree of $F'(x)$. If $\deg[F'(x)] < k$, where k is the number of information symbols, then the information polynomial $F(x) = F'(x)$; otherwise, go to step 2.

2. To determine the error-locator polynomial $D(x)$ in Eq. (A-5) and $F'(x)D(x)$, apply the Euclidean algorithm to $M(x)$ defined in Eq. (3) and $F'(x)$. The initial values of the Euclidean algorithm are $p_1(x) = 0, p_0(x) = 1, r_{-1}(x) = M(x)$, and $r_0(x) = F'(x)$. The iterative procedure of the Euclidean algorithm terminates when $\deg[r_i(x)] < n - \lfloor (d-1)/2 \rfloor$ where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.

3. Compute $F(x)$ from Eq. (A-14).

A flowchart of a decoding algorithm to correct errors only is depicted in Fig. 1. An example of this decoding scheme is given in Example 1.

B. Decoding to Correct Errors and Erasures

Shiozaki [5] suggests a decoding scheme to correct both errors and erasures. This algorithm ignores the erasure locations and uses the Chinese remainder theorem and the Euclidean algorithm to decode the shortened RS codeword. However, the shortened codeword loses the FFT structure; thus, a fast transform decoding scheme cannot be used. In this section, an improved decoding scheme is suggested which uses the fast-transform techniques discussed in the previous sections to decode RS codewords with both errors and erasures.

Suppose an RS codeword is transmitted through a noisy channel. Let there be s erasure symbols and t error symbols in the codeword such that $2t + s \leq n - k$. Next, assume that the symbols at positions $k_1, k_2, \ldots, k_s$ are erasure symbols and that the symbols at positions $\ell_1, \ell_2, \ldots, \ell_t$ are error symbols. Finally, define

$$D_1(x) = \prod_{i=1}^{s} (x - \gamma^{k_i}) \qquad \text{(known)} \tag{4}$$

and

$$D_2(x) = \prod_{i=1}^{t} (x - \gamma^{\ell_i}) \qquad \text{(unknown)}$$

and

$$D(x) = D_1(x) D_2(x)$$

By an extension of the derivation of the key Eq. (A-9) given in the Appendix, the following key equation for both errors and erasures can be obtained:

$$-M(x)B(x) + F'(x)D_1(x)D_2(x) = F(x)D_1(x)D_2(x) \tag{5}$$

where $B(x)$ is as defined in Eq. (A-5) in the Appendix, $\deg[D_2(x)] \leq \lfloor (d-1-s)/2 \rfloor$, and $\deg[F(x)D_1(x)D_2(x)] \leq n - \lfloor (d-1-s)/2 \rfloor - 1$.

The Euclidean algorithm is an iterative procedure which can be used to find in Eq. (5) the greatest common divisor (GCD) of $M(x)$ and $F'(x)D_1(x)$ [10]. An important intermediate relationship among the polynomials of the Euclidean algorithm is given in the equation

$$F'(x)D_1(x)p_i(x) + M(x)s_i(x) = r_i(x) \tag{6a}$$

and

$$\deg[p_i(x)] + \deg[r_i(x)] < \deg[M(x)] \quad \text{for } -1 \leq i \leq m \tag{6b}$$

where i is the iterative index and $r_m(x)$ is the GCD of $F'(x)D_1(x)$ and $M(x)$. The algorithm involves four sequences of polynomials: $s_i(x)$, $p_i(x)$, $r_i(x)$, and $q_i(x)$. The initial conditions are set in accordance with the following rules:

1. For $\deg[F'(x)D_1(x)] \leq \deg[M(x)]$, set $s_{-1}(x) = 1$, $s_0(x) = 0$, $p_{-1}(x) = 0$, $p_0(x) = 1$, $r_{-1}(x) = M(x)$, and $r_0(x) = F'(x)D_1(x)$.

2. For $\deg[F'(x)D_1(x)] > \deg[M(x)]$, set $s_{-1}(x) = 0$, $s_0(x) = 1$, $p_1(x) = 1$, $p_0(x) = 0$, $r_{-1}(x) = F'(x)D_1(x)$, and $r_0(x) = M(x)$.

Since $2t + s \leq n - k$,

$$\deg[D_2(x)] + \deg[F(x)D_1(x)D_2(x)] = 2t + s + k - 1 < n = \deg[M(x)] \tag{7}$$

Therefore, let $2t + s \leq n - k$, $u = \lfloor (d-1-s)/2 \rfloor$, and $v = n - \lfloor (d-1-s)/2 \rfloor - 1$. By the proof of the theorem in the Appendix and Eqs. (5), (6a), (6b), and (7), there exists a unique index j in the Euclidean algorithm such that $D_2(x) = \lambda(x)p_j(x)$ and $F(x)D_1(x)D_2(x) = \lambda(x)r_j(x)$, where $\lambda(x)$ is some polynomial, $\deg[p_j(x)] \leq \lfloor (d-1-s)/2 \rfloor$, and $\deg[r_j(x)] < n - \lfloor (d-1-s)/2 \rfloor$. Thus, $F(x)$ can be reconstructed as follows:

$$F(x) = \frac{r_i(x)}{P_i(x)D_1(x)} \tag{8}$$

The overall decoding of nonsystematic RS codes for correcting errors and erasures using the Euclidean algorithm is summarized in the following steps:

1. Use step 1. from the description of decoding for correcting errors.

2. Compute the erasure-locator polynomial $D_1(x)$ from Eq. (4). Next, compare $\deg[F'(x)D_1(x)]$ with $\deg[M(x)]$. If $\deg[F'(x)D_1(x)] \leq \deg[M(x)]$, set $p_{-1}(x) = 0$, $p_0(x) = 1$, $r_{-1}(x) = M(x)$, and $r_0(x) = F'(x)D_1(x)$; otherwise, set $p_{-1}(x) = 1$, $p_0(x) = 0$, $r_{-1}(x) = F'(x)D_1(x)$, and $r_0(x) = M(x)$.

3. To determine the error-locator polynomial $D_2(x)$ and $F'(x)D(x)$, apply the Euclidean algorithm to $M(x)$ and $F'(x)D_1(x)$. The initial values of the Euclidean algorithm are defined in step 2; the iterative procedure of the algorithm terminates when $\deg[r_i(x)] < n - \lfloor (d-1-s)/2 \rfloor$.

4. Compute $F(x)$ from Eq. (8).

A flowchart of the decoding scheme for correcting both errors and erasures is given in Fig. 2. A depiction of this decoding scheme is presented in Example 2.

This simpler, faster transform-decoding scheme using the FNT for RS codes is particularly suitable for pipeline VLSI implementation. The transform-decoding scheme utilizes an efficient FNT to compute the corrupted information polynomial $F'(x)$ in a manner analogous to syndrome computation in the conventional decoding schemes. However, this new algorithm does not require the extra steps needed to solve the error-locator and error-evaluator polynomials.

C. Examples of the Two Decoding Methods

Example 1. Consider the Fermat prime $F_3 = 17$, and let $k = 4$. This is an (8,4) RS code over $GF(17)$, which is capable of correcting two errors or less. It is shown in [6] that $\gamma = 2$ is a primitive 8th root of unity. Also, for this case,

$$M(x) = \prod_{i=0}^{i=7}(x - \gamma^i) = x^8 - 1$$

Let the four information symbols be $\underline{u} = (2,3,1,4)$. Then $F(x) = 2 + 3x + x^2 + 4x^3$. An FNT on the sequence $(2,3,1,4,0,0,0,0)$ yields the codeword $\underline{\nu} = (10,10,14,13,13,2,5,0)$. Next, let the third and seventh symbols be erroneous. Thus, $\underline{e} = (0,0,5,0,0,0,15,0)$ and $\underline{\nu}' = (10,10,2,13,13,2,3,0)$, where $\underline{\nu}$, $\underline{e}$, and $\underline{\nu}'$ are as defined in the Appendix. After taking the IFNT on $\underline{\nu}'$, one obtains $F'(x) = 13 + 8x + 7x^2 + 16x^3 + 11x^4 + 5x^5 + 6x^6 + 12x^7$. The Euclidean algorithm stops after the second iteration to yield $r_2(x) = 10x^5 + 11x^4 + 9x^3 + 16x^2 + 16x + 5$ and $p_2(x) = 11x^2 + 11$. Then $F(x)$ is recaptured as

$$F(x) = \frac{r_2(x)}{p_2(x)} = 2 + 3x + x^2 + 4x^3$$

That is, $\underline{u} = (2,3,1,4)$.

Example 2. Consider the same codeword $\underline{\nu} = (10,10,14,13,13,2,5,0)$ given in Example 1. Let the first symbol be an error symbol, and the third and seventh symbols be erasure symbols. Thus, $\underline{e} = (1,0,5,0,0,0,15,0)$, and $\underline{\nu}' = (11,10,2,13,13,2,3,0)$. After the IFNT is taken of ν', one obtains $F'(x) = 11+6x+5x^2+14x^3+9x^4+3x^5+4x^6+10x^7$. Since the erasure symbols are at the third and seventh positions, $D_1(x) = (x-2^2)(x-2^6) = x^2+1$. Thus, $F'(x)D_1(x) = 10x^9+4x^8+13x^7+13x^6+14x^4+3x^3+16x^2+6x+11$. The Euclidean algorithm stops after the second iteration to yield $r_2(x) = x^6 + 12x^5 + 10x^4 + 16x^3 + 4x + 8$ and $p_2(x) = 13x + 4$. Then $F(x)$ is recaptured as

$$F(x) = \frac{r_2(x)}{p_2(x)D_1(x)} = 2 + 3x + x^2 + 4x^3$$

That is, $\underline{u} = (2,3,1,4)$.

V. Conclusions

In this article, a fast transform decoding scheme is introduced which is particularly suitable for VLSI implementation. This scheme first utilizes the highly efficient Fermat number transform to calculate the corrupted information polynomial $F'(x)$. It then uses the Euclidean algorithm to evaluate the information polynomial $F(x)$ directly, without going through the intermediate steps of solving the error-locator and error-evaluator polynomials. Thus, this fast-transform decoding scheme is faster and simpler than the decoding scheme in [1].

Acknowledgments

The authors wish to thank Dr. Les Deutsch, Dr. Robert J. McEliece, Dr. Ed Satorius, and Dr. Laif Swanson at the Jet Propulsion Laboratory for their mathematical suggestions pertaining to the research which led to this article.

References

[1] A. Shiozaki and F. Nishida, "A New Decoding Method of Redundant Residue Polynomial Codes," *Bull. Univ. of Osaka Prefecture*, series A, vol. 24, no. 1, pp. 101–112, 1975.

[2] R. Blahut, *Theory and Practice of Error Control Codes*, Reading, Massachusetts: Addison-Wesley, 1984.

[3] I. S. Reed and G. Solomon, "Polynomial Codes Over Certain Finite Fields," *SIAM*, vol. 8, no. 2, pp. 300–304, 1960.

[4] J. J. Stone, "Multiple-burst Error Correction with the Chinese Remainder Theorem," *SIAM*, vol. 11, no. 1, pp. 74–81, 1963.

[5] A. Shiozaki, "Decoding of Redundant Residue Polynomial Codes Using Euclid's Algorithm," *IEEE Trans. Information Theory*, vol. 34, no. 5, pp. 1351–1354, September 1988.

[6] R. C. Agarwal and C.S. Burrows, "Fast Convolution Using Fermat Number Transforms with Applications to Digital Filtering," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-22, no. 2, pp. 87–97, April 1974.

[7] I. S. Reed, T. K. Truong, and L. R. Welch, "The Fast Decoding of Reed-Solomon Codes Using Fermat Transforms," *IEEE Trans. Information Theory*, vol. IT-24, no. 4, pp. 497–499, July 1978.

[8] A. Oppenheim and R. Schafer, *Digital Signal Processing*, New York: Prentice-Hall, 1975.

[9] H. C. Shyu, T. K. Truong, I. S. Reed, I. S. Hsu, and J. J. Chang, "A New VLSI Complex Integer Multiplier Which Uses a Quadratic-Polynomial Residue System with Fermat Number," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no. 7, pp. 1076–1079, July 1987.

[10] R. McEliece, *The Theory of Information and Coding*, vol. 3 of *The Encyclopedia of Mathematics and Its Applications*, Reading, Massachusetts: Addison-Wesley, 1977.

[11] E. R. Berlekamp, *Algebraic Coding Theory*, New York: McGraw-Hill, 1968.

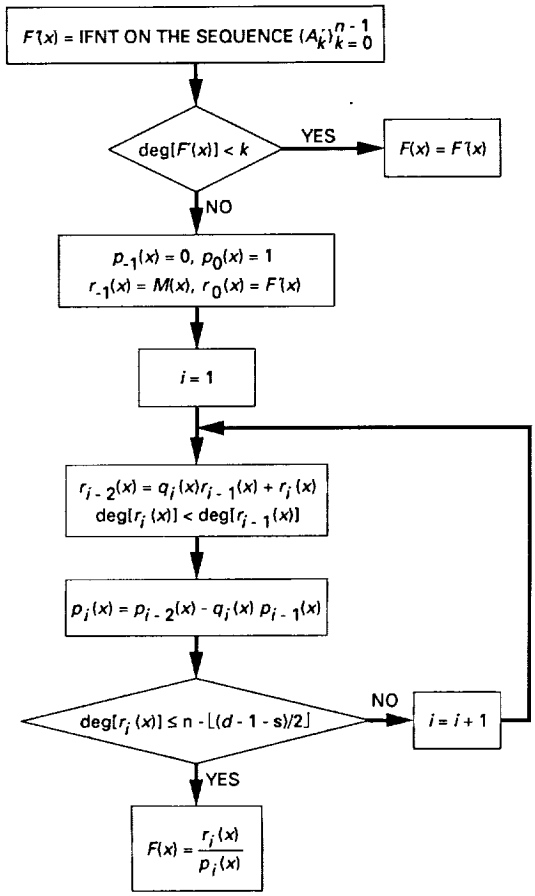


Fig. 1. Flowchart of decoding procedure to correct errors only.

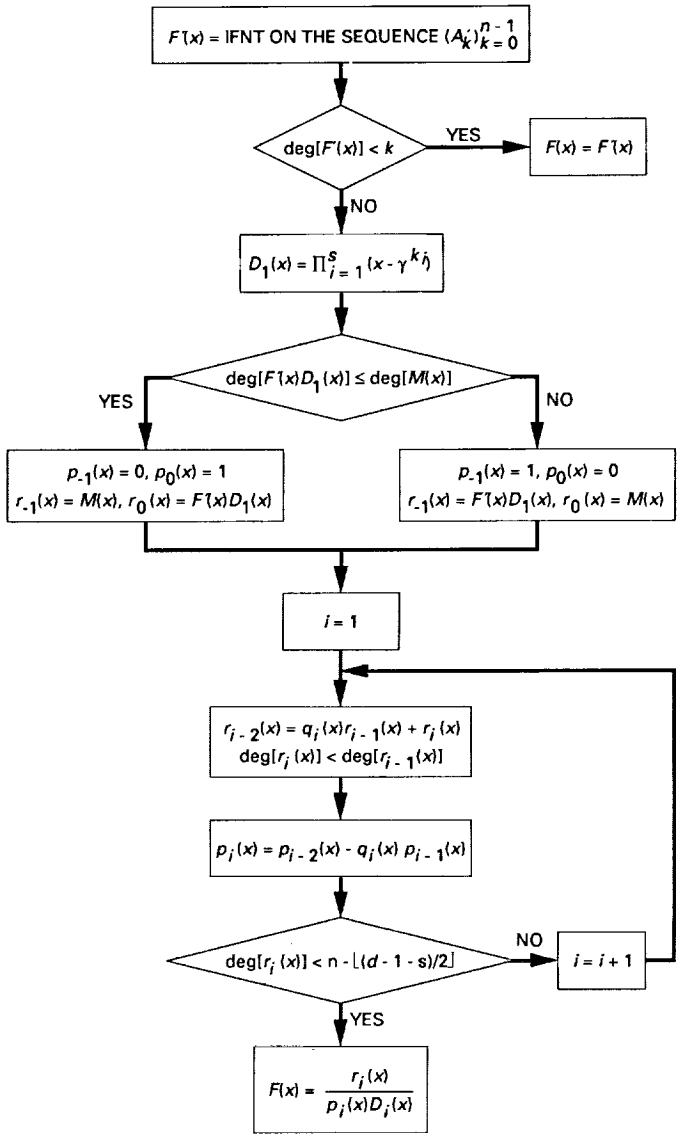


Fig. 2. Flowchart of decoding procedure to correct errors and erasures.

Appendix
Decoding RS Codes Using the Euclidean Algorithm

Suppose the codeword $\underline{\nu} = (A_0, A_1, \ldots, A_{n-1})$ is transmitted through a noisy channel. Assume that the symbols at positions $\ell_1, \ell_2, \ldots,$ and ℓ_t are in error. The received codeword $\underline{\nu}'$ is thus represented by

$$\underline{\nu}' = \underline{\nu} + \underline{e} = (A_0', A_1', \ldots, A_{n-1}' \tag{A-1}$$

where $\underline{e}$ is the error vector defined by

$$\underline{e} = (0, \ldots, e_{\ell_1}, 0, \ldots, e_{\ell_t}, \ldots, 0)V \tag{A-2}$$

Let $(u_0', u_1', \ldots, u_{n-1}')$ and $(w_0, w_1, \ldots, w_{n-1})$ be the inverse transforms of $\underline{\nu}'$ and $\underline{e}$ respectively. Also let $F'(x) = u_0' + u_1'x + \cdots + u_{n-1}'x^{n-1}$ be defined as the corrupted information polynomial, and $E(x) = w_0 + w_1x + \cdots + w_{n-1}x^{n-1}$ be defined as the error polynomial.

It is not difficult to see that the residue representations of $F'(x)$ and $E(x)$, modulo $m_i(x)$ are Eqs. (A-1) and (A-2) respectively. That is, $\underline{\nu}'$ and $\underline{e}$ can be written, respectively, as

$$\underline{\nu}' = \left(F'(1), F'(\gamma), \ldots, F'(\gamma^{n-1})\right)$$

and

$$\underline{e} = \left(E(1), E(\gamma), \ldots, E(\gamma^{n-1})\right)$$

From Section III, an RS codeword $\underline{\nu}$ is generated by an information polynomial $F(x)$ via the following:

$$\underline{\nu} = \left(F(1), F(\gamma), \ldots, F(\gamma^{n-1})\right)$$

Since $\underline{\nu}' = \underline{\nu} + \underline{e}$, one obtains $F'(\gamma^i) = F(\gamma^i) + E(\gamma^i)$ for $0 \leq i \leq n-1$. Thus, there are at least n values of x for which $F'(x)$ and $F(x) + E(x)$ are equal. It is obvious that $\deg[F(x)] < k, \deg[F'(x)] < n$, and $\deg[E(x)] < n$. Hence, by the fundamental theorem of algebra,

$$F'(x) = F(x) + E(x) \tag{A-3}$$

Since RS codes are a special case of RRP codes, it is shown in [5] that $F'(x)$ and $E(x)$ can also be reconstructed using the Chinese remainder theorem as follows:

$$F'(x) \equiv \left[\sum_{i=0}^{n-1} \frac{M(x)}{m_P(x)} t_i(x) A_i'\right] \bmod M(x)$$

and

$$E(x) \equiv \left[\sum_{i=1}^{t} \frac{M(x)}{m_{\ell_i}(x)} t_{\ell_i}(x) e_{\ell_i}\right] \bmod M(x) \tag{A-4}$$

Let

$$B(x) \equiv \left[\sum_{i=1}^{t} \frac{D(x)}{m_{\ell_i}(x)} t_{\ell_i}\right] \bmod D(x) \tag{A-5}$$

where

$$D(x) = \prod_{i=1}^{t} m_{\ell_i}(x)$$

is called the error-locator polynomial. The key equation of the decoding algorithm is derived from these relationships. First, let

$$A(x) = \frac{M(x)}{D(x)} \tag{A-6}$$

and

$$B'(x) = \sum_{i=1}^{t} \frac{D(x)}{m_{\ell_i}(x)} t_{\ell_i}(x) e_{\ell_i}$$

Then, by Eq. (A-4), one has

$$E(x) \equiv \left[\frac{M(x)}{D(x)} \sum_{i=1}^{t} \frac{D(x)}{m_{\ell_i}} t_{\ell_i}(x) e_{\ell_i}\right] \bmod M(x)$$

$$[A(x)B'(x)] \bmod A(x)D(x) \tag{A-7}$$

Equation (A-7) can now be re-expressed as:

$$E(x) = A(x)\left[\lambda(x)D(x) + B'(x)\right]$$

$$= A(x)\left[B'(x) \bmod D(x)\right] \tag{A-8}$$

where $\lambda(x)$ is some polynomial over the finite field. Using Eq. (A-5), a substitution of $A(x)$ and $B'(x)$ in Eq. (A-6) into Eq. (A-8) yields

$$E(x) = \frac{M(x)}{D(x)} B(x) \qquad \text{(A-9)}$$

The proof of Eq. (A-9) is similar to the proof given in [5]. A similar result of Eq. (A-9) is given by Blahut in theorem 9.1.1 of [2] using a spectral technique. The decoder in Fig. 9.2 of [2] applies the Euclidean algorithm to $M(x)$ and the $2t$ high-order coefficients of $E(x)$ to determine the error-locator polynomial $D(x)$. Then a recursive extension is used to compute the rest of the coefficients of $E(x)$ from the known $D(x)$. Finally, the inverse transform over $GF(2^n)$ of E_j is taken to recover the error pattern.

The next paragraph describes how the decoder developed in this article applies the Euclidean algorithm to the polynomials $M(x)$ and $F'(x)$ rather than to the usual $M(x)$ and the syndrome polynomial $S(x)$, i.e., the $2t$ high-order coefficients of $E(x)$. In other words, to determine polynomials $D(x)$ and $F(x)D(x)$, this new decoder applies the Euclidean algorithm to $M(x)$ and $F'(x)$. Thus, $F(x)$ can be reconstructed from $F(x) = F(x)D(x)/D(x)$. The advantage of this new decoder over the decoder developed in Fig. 9.3 of [2] is that both the recursive extension and the inverse transform can be replaced by a single polynomial division.

By combining Eqs. (A-3) and (A-9), the key equation is obtained as follows:

$$-M(x)B(x) + F'(x)D(x) = F(x)D(x) \qquad \text{(A-10)}$$

where $B(x)$ is defined as in Eq. (A-5).

The Euclidean algorithm is an iterative procedure to find the greatest common divisor (GCD) of $M(x)$ and $F'(x)$ [10]. An important intermediate relationship among the polynomials of the Euclidean algorithm is given in the following:

$$-M(x)s_i(x) + F'(x)p_i(x) = r_i(x) \qquad \text{(A-11)}$$

and

$$\deg[P_i(a)] + \deg[r_i(x)] < \deg[M(x)]$$

and

$$\text{for} -1 \le i \le m \qquad \text{(A-12)}$$

where i is the iterative index, and $r_m(x)$ is the GCD of $F'(x)$ and $M(x)$. The algorithm involves four sequences of polynomials: $s_i(x)$, $p_i(x)$, $r_i(x)$, and $q_i(x)$. The initial conditions are: $s_{-1}(x) = 1$, $s_0(x) = 0$, $p_{-1}(x) = 0$, $p_0(x) = 1$, $r_{-1}(x) = M(x)$, and $r_0(x) = F'(x)$; $q_{-1}(x)$ and $q_0(x)$ are not defined.

The following lemma and theorem [10] show that the Euclidean algorithm can be applied to the key Eq. (A-10) to solve for the information polynomial $F(x)$.

Lemma. Given two non-negative integers μ and ν with $\nu \ge \deg[r_m(x)]$ satisfying $\mu + \nu = \deg[M(x)] - 1$, there exists a unique index j, $0 \le j \le m$, such that

$$\deg[p_j(x)] \le \mu$$

and

$$\deg[r_j(x)] \le \nu$$

For the proof, see [10].

Using the above lemma, the following theorem can be proved [10]:

Theorem. Suppose $p(x)$, $s(x)$, and $r(x)$ are nonzero polynomials satisfying

$$-M(x)s(x) + F'(x)p(x) = r(x)$$

and

$$\deg[p(x)] + \deg[r(x)] < \deg[M(x)]$$

There then exists a unique index j, $0 \le j \le m$, and a polynomial $\lambda(x)$ such that

$$p(x) = \lambda(x)p_j(x)$$

and

$$r(x) = \lambda(x)r_j(x)$$

Now let $n - k = 2T$, where T is the maximum number of errors in an RS code which can be corrected. If the number of errors t in a received RS codeword is less than or equal to T, then $\deg[D(x)] \le T$ and $\deg[F(x)D(x)] \le k + T - 1 = n - T - 1$. Thus,

$$\deg[D(x)] + \deg[F(x)D(x)] < \deg[M(x)] = n \qquad \text{(A-13)}$$

Thus, let $n-k \geq 2t$, $u = T$, and $v = \deg[M(x)] - 1 - \mu = n-T-1$. By the proof of the above theorem and Eqs. (A-10), (A-11), (A-12), and (A-13), there exists a unique index j in the Euclidean algorithm such that $D(x) = \lambda(x)p_j(x)$ and $F(x)D(x) = \lambda(x)r_j(x)$, where $\deg[p_j(x)] \leq T$ and $\deg[r_j(x)] \leq n-T-1$. Thus, $F(x)$ can be reconstructed as:

$$F(x) = \frac{r_j(x)}{p_j(x)} \tag{A-14}$$

Application of Adaptive Least-Squares Algorithm to Multi-Element Array Signal Reconstruction

R. Kumar
Communications Systems Research Section

This article presents some results in terms of the performance improvement of a multi-feed array configuration over the usual single feed system when an adaptive least-squares algorithm is applied for the signal reconstruction. The article presents two novel versions of the least-squares algorithm, one of which is based on the maximization of the signal-to-noise ratio while the other is based on the deconvolution of the received signal field. These algorithms have been developed for the purpose of minimizing degradations arising from various sources, which can severely limit the performance (gain) of a single-feed system.

I. Introduction

The development of multi-element array processing techniques has many potential applications for the Deep Space Network (DSN). These include signal reconstruction for both X-band and Ka-band communications [1–4] and electronic pointing to augment existing mechanical pointing techniques. For all of these applications, multi-element array processing can generally provide significant performance improvements over single-feed antenna configurations, which are predominantly used in the DSN.

This article presents some results in terms of the performance improvement provided by a linear multi-feed array incorporating the proposed adaptive least-squares algorithm over a single-feed array system, as applied to the signal reconstruction problem. While the assumed linear array geometry is idealized, the results of this analysis provide an indication of performance improvements that can be achieved with adaptive, multi-element array processing. This article describes two versions of the least-squares algorithm, one of which is based on the maximization of the signal-to-noise ratio while the other is based on deconvolution of the received signal field. Here, instead of trying to model the signal degradations in terms of deterministic equations in evaluating the performance of the algorithm, it is assumed that these degradations are "unknown" to the algorithm and vary with time. The algorithm tries to implicitly estimate these degradations in an adaptive manner from the samples of the noisy received signal. On the basis of these measurements, it computes a set of weights for combining signals at the outputs of various feeds in order to maximize the signal-to-noise ratio of the combined signal.

For the purposes of illustrating the basic concepts involved with adaptive array processing, this article presents

the results for a 16-element linear-array feed system. The performance of the least-squares algorithm is to a first order determined by the signal-to-noise ratio of the received signal, the number of feeds in the configuration, and the time constants at which the received signal field is varying with time.

In practice, it may be possible to simultaneously correct for multiple degradations arising from different sources and having different time constants. These degradations may be induced by wind, gravitational loading, or antenna pointing errors. Simultaneous correction for such degradations could be achieved by adjusting the time constants of the algorithm to track the fastest mode, in which case the slower modes would be estimated sub-optimally. Alternatively, one could track the most significant mode, thereby essentially ignoring the faster but less significant modes. More sophisticated techniques could also be used for separation of the modes and tracking of them separately.

II. Array Configuration

The specific array configuration of interest in this article corresponds to a linear-feed array distributed across the focal plane of an antenna. The array outputs are then fed to a parallel receiver bank as indicated in Fig. 1. As shown in the schematic diagram of Fig. 1, the signal outputs from the feed elements are amplified by r f amplifiers. Assuming that all of the amplifiers have equal gain and noise temperature, the output of the ith amplifier can be written as

$$r_i(t) = A_i(t)\cos\Big(\omega_c t + \theta_i(t)\Big) + n_i(t) \qquad (1)$$

where $A_i(t)$ and $\theta_i(t)$ are the signal amplitudes and phases, ω_c denotes the signal carrier frequency, and $n_i(t)$ is a zero-mean white Gaussian noise of one-sided spectral density N_0. The noise is also assumed to be spatially uncorrelated, i.e., $E\left[n_i(t)n_j(t)\right] = 0$ for $i \neq j$ and $i, j = 1, 2, .., N$. Under ideal conditions and assuming a plane-wave normally incident on the antenna aperture, the amplitude of the center feed would be equal to $\sqrt{2P}$ (P denotes the normalized power received by the antenna), while the remaining feeds will have nearly zero amplitude. However, array degradations can disperse the signal amplitude (and phase) spatially over N feeds. These degradations can arise due to various sources, such as gravity, thermal fields, wind, and atmospheric turbulence.

In the presence of such degradations, the adaptive signal processing algorithm then combines the N feed outputs in a coherent manner so as to optimise some performance index, such as the signal-to-noise power ratio of the combined signal. This processing can occur either at r f or can be equivalently done at the baseband. Alternatively, in some possible imaging applications, it may be desired to reconstruct the complete focal plane field, i.e., obtain N output signals that are close to the outputs of the focal plane fields in the ideal antenna case.[1] Note that in the more general case, more than one feed may have significant amplitude if the source is not a point source and the antenna is capable of resolving such a composite source. In the limiting case, one may simply use the center output of the reconstructed field and ignore the others, thus achieving an alternative combination of the input signals.

For the purposes of signal processing, the N received r f signals $r_i(t)$; $i = 1, 2, ..., N$ are down-converted and quadrature sampled to obtain the sampled version of the complex baseband envelope $g_i(t)$ of the r f signal $r_i(t)$ with

$$\begin{aligned} r_i(t) &= Re\Big\{g_i(t)e^{j\omega_c t}\Big\} \\ g_i(t) &= A_i(t)e^{j\theta_i(t)} + \nu_i(t) \\ g_i(t) &= \Big\{r_i(t) + \hat{r}_i(t)\Big\}e^{-j\omega_c t} \end{aligned} \qquad (2)$$

In Eq. (2) above, $\hat{r}_i(t)$ denotes the Hilbert transform of $r_i(t)$, and $\nu_i(t)$ is the complex envelope of the bandpass noise $n_i(t)$.

III. Signal Combining Via Adaptive Least-Squares Algorithm I

As shown in Fig. 1, the adaptive algorithm determines the time-varying complex-valued weighting coefficients $w_1(k), ..., w_N(k)$ on the basis of signal samples $g_i(j)$; $i = 1, 2, ..., N$; and $j = 1, 2, ..., k$ according to some appropriate optimization criterion. The algorithm is adaptive in the sense that if the signal amplitudes and phases (A_i and θ_i) remain relatively constant with time, then with increasing value of k, the algorithm achieves increasingly accurate estimates of these parameters, and the weighting coefficients converge asymptotically to their theoretically optimum values with an exponential convergence rate. On the other hand, if these parameters are time-varying, then

[1] V. Vilnrotter, "Ka-Band Array Signal Processing Progress Report," JPL Interoffice Memorandum No. 331-88.5-047 (internal document), Jet Propulsion Laboratory, Pasadena, California, November 1988.

the algorithm tracks these variations and the weighting coefficients are truly time-varying (there is no tendency for $w_i(k)$ to converge to some constant value).

Denoting by $\underline{w}(k)$ and $\underline{g}(k)$ the weighting coefficient vector $[w_1(k)w_2(k)\ldots w_N(k)]'$ and the measurement vector $[g_1(k)g_2(k)\ldots g_N(k)]'$ respectively, then the familiar least-square optimization criterion is to select $\underline{w}(k)$ so as to minimize the following index

$$J_k = \sum_{j=1}^{k} |1 - \underline{w}^H(k)\underline{g}(j)|^2 \tag{3}$$

with respect to the weight vector $\underline{w}(k)$ for $k = 1, 2, \ldots$. In the above, the superscript H denotes conjugate transpose while $'$ represents just the transpose of a matrix. The optimal solution, termed least-squares estimate of $w(k)$, is given by (assuming $k > N$)

$$\underline{\hat{w}}_{LS}(k) = \left\{\sum_{j=1}^{k} \underline{g}(j)\underline{g}^H(j)\right\}^{-1} \sum_{j=1}^{k} \underline{g}(j) \tag{4}$$

If the distortion process is time-varying, then it is more appropriate to replace the index J_k by the one obtained by multiplying the summand in Eq. (3) by λ^{k-j} for some $0 < \lambda < 1$, minimization of which yields the following exponentially data-weighted least-squares estimate for $\underline{w}(k)$.

$$\underline{\hat{w}}_{ELS}(k) = \left\{\sum_{j=1}^{k} \lambda^{k-j}\underline{g}(j)\underline{g}^H(j)\right\}^{-1} \sum_{j=1}^{k} \lambda^{k-j}\underline{g}(j) \tag{5}$$

Note that in the adaptive algorithm's present non-recursive form, Eq. (4), it is required to invert an $(N \times N)$ matrix for every time instance k in the computation of $\hat{w}(k)$, which is somewhat computationally intensive. This problem can be overcome by replacing the estimate in Eq. (5) with its recursive form, which is obtained as follows.

Denoting by $\underline{P}(k)$ the matrix inverse in Eq. (5), then the matrix $\underline{P}^{-1}(k)$ has the following update.

$$\underline{P}^{-1}(k) = \lambda\underline{P}^{-1}(k-1) + \underline{g}(k)\underline{g}^H(k); \quad k = 1, 2, \ldots \tag{6}$$

Application of the matrix inversion lemma [5] to Eq. (6) yields the following desired recursion for $\underline{P}(k)$.

$$\underline{P}(k) = \lambda^{-1}\Big\{\underline{P}(k-1) - \left[\lambda + \underline{g}^H(k)\underline{P}(k-1)\underline{g}(k)\right]^{-1} \times \underline{P}(k-1)\underline{g}(k)\underline{g}^H(k)\underline{P}(k-1)\Big\} \tag{7}$$

One may note that the entity to be inverted in Eq. (7) is only a scalar. Decomposing the sum in Eq. (5) as

$$\lambda \sum_{j=1}^{k-1} \underline{g}(j)\lambda^{k-1-j} + \underline{g}(k)$$

and substituting Eq. (7) for $\underline{P}(k)$, we obtain the following expression for $\underline{\hat{w}}_{ELS}(k)$.

$$\begin{aligned}\underline{\hat{w}}_{ELS}(k) = \underline{P}(k-1) &\left\{\sum_{j=1}^{k-1} \underline{g}(j)\lambda^{k-1-j}\right\} \\ &- \left[\lambda + \underline{g}^H(k)\underline{P}(k-1)\underline{g}(k)\right]^{-1} \\ &\times \underline{P}(k-1)\underline{g}(k)\underline{g}^H(k)\underline{P}(k-1) \\ &\times \left\{\sum_{j=1}^{k-1} \lambda^{k-1-j}\underline{g}(j)\right\} + \underline{P}(k)\underline{g}(k)\end{aligned} \tag{8}$$

By noting that the first term in Eq. (8) and the product of the last two factors in the second term both are equal to $\underline{\hat{w}}_{ELS}(k-1)$, Eq. (8) may be rewritten as

$$\begin{aligned}\underline{\hat{w}}_{ELS}(k) = \underline{\hat{w}}_{ELS}(k-1) &+ \underline{P}(k)\underline{g}(k) \\ &- \left[\lambda + \underline{g}^H(k)\underline{P}(k-1)\underline{g}(k)\right]^{-1} \\ &\times \underline{P}(k-1)\underline{g}(k)\underline{g}^H(k)\underline{\hat{w}}_{ELS}(k-1)\end{aligned} \tag{9}$$

Post multiplying both sides of Eq. (7) by $\underline{g}(k)$ and with a simple algebraic manipulation, it follows that,

$$\left[\lambda + \underline{g}^H(k)\underline{P}(k-1)\underline{g}(k)\right]^{-1} \underline{P}(k-1)\underline{g}(k) = \underline{P}(k)\underline{g}(k) \tag{10}$$

With the substitution of Eq. (10) in the last term of Eq. (9), one obtains the recursive version of the algorithm given below.

$$\begin{aligned}\underline{\hat{w}}_{ELS}(k) = \underline{\hat{w}}_{ELS}(k-1) &+ \underline{P}(k)\underline{g}(k) \\ &\times \left[1 - \underline{g}^H(k)\underline{\hat{w}}_{ELS}(k-1)\right] \qquad (11) \\ \underline{P}(k) = \lambda^{-1}\Big\{\underline{P}(k-1) &- \left[\lambda + \underline{g}^H(k)\underline{P}(k-1)\underline{g}(k)\right]^{-1} \\ &\times \underline{P}(k-1)\underline{g}(k)\underline{g}^H(k)\underline{P}(k-1)\Big\}; \\ &k = 0, 1, 2, \ldots\end{aligned}$$

If the complex field $\underline{g}(k)$ is a wide-sense stationary process, a considerable simplification in computations may be achieved by replacing $\underline{P}(k)$ with an appropriate constant matrix in the first recursion of Eq. (11) and dropping the second recursion.

IV. Maximization of Signal-to-Noise Ratio via Modified Least-Squares Algorithm I

In some applications, it may be more appropriate to maximize the signal-to-noise power ratio at the combiner output. The noise variance at the output of the combiner is equal to $\sum_{i=1}^{N} |\hat{w}_i|^2 E[|\nu_i|^2] = \sigma^2 \|\underline{\hat{w}}\|^2$ with σ^2 denoting the variance of the sampled version of the complex baseband process $\nu_i(t)$ in Eq. (2). Thus, as shown in the Appendix, an effective maximization of the output signal-to-noise ratio can be achieved by minimization of the index given in Eq. (3) subject to the equality

$$\|\underline{w}\|^2 = K \tag{12}$$

for some constant K, or by simply minimizing the following index,

$$\sum_{j=1}^{k} |1 - \underline{w}^H(k)\underline{g}(j)|^2 + \gamma(k)(\|\underline{w}\|^2 - K) \tag{13}$$

with respect to $\underline{w}$ and $\gamma(k)$, where $\gamma(k)$ is the Lagrangian multiplier. Differentiation of Eq. (13) with respect to $\underline{w}$ yields the following constrained least-squares estimate for $\underline{w}$ in terms of $\gamma(k)$ as

$$\underline{\hat{w}}_{CLS}(k) = \left\{ \sum_{j=1}^{k} \underline{g}(j)\underline{g}^H(j) + \gamma(k) I \right\}^{-1} \sum_{j=1}^{k} g(j) \tag{14}$$

Substituting Eq. (14) into Eq. (12) yields an equation for the unknown $\gamma(k)$, which can be solved for $\gamma(k)$. Substituting $\gamma(k)$ back in Eq. (14) provides the constrained optimum solution for the weighting coefficient vector. Note, however, that there is no close-form solution for $\gamma(k)$ and, thus, some numerical optimization techniques may need to be applied to obtain $\underline{\hat{w}}_{CLS}$. A simplified solution is obtained by selecting some appropriate value for $\gamma(k)$ in Eq. (14) and then normalizing the estimate to have its norm equal to one. In an exponentially data-weighted version of Eq. (14), both the summands are multiplied by λ^{k-j} where λ is the exponential data-weighting coefficient. With these modifications, Eq. (14) has the following equivalent form.

$$\begin{aligned}
\underline{P}^{-1}(k) &= \lambda \underline{P}^{-1}(k-1) + \gamma_0 I + \underline{g}(k)\underline{g}^H(k) \\
\underline{\psi}(k) &= \lambda \underline{\psi}(k-1) + \underline{g}(k) \\
\underline{\hat{w}}(k) &= \underline{P}(k)\underline{\psi}(k) \\
\underline{\hat{w}}_{NLS} &= \underline{\hat{w}}(k)/\|\underline{\hat{w}}(k)\| \\
\gamma(k) &= \gamma_0(1 - \lambda^k)/(1 - \lambda)
\end{aligned} \tag{15}$$

Note that Eq. (15) requires a matrix inversion for each value of k for which $\underline{\hat{w}}(k)$ is desired. An approximate recursive form for Eq. (15), which does not require matrix inversion, may also be derived by applying the matrix inversion lemma. Note that in Eq. (15), the higher value of γ_0 results in the higher relative weight assigned to the noise variance at the combiner output. The initial values for $\underline{P}^{-1}$ and $\underline{\psi}$ at $k = 0$ may simply be selected equal to zero.

V. Least Squares Algorithm II

In an alternative solution, it may be assumed that the received focal plane signal is the result of the spatial convolution of the ideal signal (in the absence of any distortion) with an unknown filter response representing various distortions from all sources, including antenna surface deformations due to gravity, wind, antenna pointing errors, turbulence, etc., i.e.,

$$\underline{g}(k) = \underline{B}\,\underline{X}(k) + \underline{\nu}(k) \tag{16}$$

where $\underline{B}$ is a Toeplitz matrix, $\underline{X}$ is the focal plane signal vector that under appropriate sampling is equal to $[0 \ldots 010 \ldots 0]$ in the ideal case of plane-wave with no distortion, and $\underline{\nu}$ is the additive noise vector. The matrix $\underline{B}$ includes any distortion effects, pointing errors, etc. For the linear array case under consideration, $\underline{B}$ can be approximated by a circular matrix for large N or becomes identical to a circular matrix provided the vector $\underline{X}(k)$ and the signal vector $\underline{g}(k)$ are zero-padded. Thus, it is assumed that B is circular. In the absence of noise, it is observed that $\underline{X} = \underline{B}^{-1}\underline{g}$ where $\underline{B}^{-1}$ is also a circular matrix. However, $\underline{B}$ is unknown, and it needs to be estimated from the noisy observations. Or, more directly, $\underline{B}^{-1}$ is estimated as follows. Rewriting the model Eq. (16) as

$$\underline{X}(k) = \underline{F}^* \underline{g}(k) + \underline{v}(k); \quad k = 1, 2, \ldots \tag{17}$$

where $\underline{F}$ is an unknown circular matrix to be estimated from the given measurement $\underline{g}(k)$; $k = 1, 2 \ldots, n$. Letting

$\underline{f}^T = [f_1,\ f_2 \ldots f_N]$ denote the first row of the matrix $\underline{F}$, and $\underline{g}_\ell(k)$ denote the vector obtained by cyclically shifting $\underline{g}(k)$ right ℓ times, the following equivalent signal model is obtained.

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \underline{g}^T(k) \\ \underline{g}_1^T(k) \\ \vdots \\ \underline{g}_{N-1}^T(k) \end{bmatrix} \begin{bmatrix} f_1^* \\ f_2^* \\ \vdots \\ f_N^* \end{bmatrix} + \begin{bmatrix} v_1(k) \\ v_2(k) \\ \vdots \\ v_N(k) \end{bmatrix}; \quad k = 1, 2 \ldots \tag{18}$$

A least-squares estimate for $\underline{f}$ can be obtained from the signal model Eq. (18) and in its non-recursive form is given by

$$\underline{\hat{f}}(k) = \left\{ \sum_{j=1}^{k} \left(\sum_{\ell=0}^{N-1} \underline{g}_\ell(j) \underline{g}_\ell^H(j) \right) \right\}^{-1} \sum_{j=1}^{k} \underline{g}_{N/2}(j) \tag{19}$$

A recursive form for the estimate $\underline{\hat{f}}(k)$ may also be derived following the steps used in obtaining Eq. (11). Here recursion is over both the signal sample vector $\underline{g}(k)$ and its circular shifts. With appropriate initial values $\underline{P}(0,0)$ and estimate $\hat{f}(0,0)$, one has the following recursion in terms of the indices k and j, with k denoting time and j denoting the cyclic shift of the received signal vector,

$$\begin{aligned} \underline{P}(k,j) &= \lambda^{-1} \Big\{ \underline{P}(k,j-1) \\ &\quad - \left[\lambda + \underline{g}_{j-1}^H(k) \underline{P}(k,j-1) \underline{g}_{j-1}(k) \right]^{-1} \\ &\quad \times \underline{P}(k,j-1) \underline{g}_{j-1}(k) \underline{g}_{j-1}^T(k) \underline{P}(k,j-1) \Big\}; \\ &\qquad j = 1, 2, \ldots \\ \underline{P}(k+1,0) &= \underline{P}(k,N); \quad k = 1, 2, \ldots \\ \underline{\hat{f}}(k,j) &= \underline{\hat{f}}(k,j-1) + \underline{P}(k,j) \underline{g}_{j-1}(k) \\ &\quad \times \left[\xi_j - \underline{g}_{j-1}^H(k) \underline{\hat{f}}(k,j-1) \right]; \\ &\qquad j = 1, 2, \ldots \\ \underline{\hat{f}}(k+1,0) &= \underline{\hat{f}}(k,N); \quad k = 1, 2, \ldots \end{aligned} \tag{20}$$

where

$$\begin{aligned} \xi_j &= 1, \quad j = \lfloor N/2 \rfloor \\ &= 0, \quad j \neq \lfloor N/2 \rfloor \end{aligned}$$

where $\lfloor x \rfloor$ denotes the least integer greater than or equal to x for any real x. The circular spatial convolution of $\underline{\hat{f}}^*$ with the received signal $\underline{g}(k)$ yields the reconstructed signal vector $\underline{h}(k)$. In the case of perfect reconstruction (deconvolution), the central element of the vector $\underline{h}$ is the combined signal, while the remaining elements would be zero.

VI. Simulations

The performance of the least-squares algorithms of the previous section is presented here in terms of simulations. In the following simulations, a linear feed array is considered. For the purposes of these simulations, the received signal focal field is generated by spatially Fourier transforming a simulated linear aperture plane array. The signals in the simulated aperture are assumed to be of equal amplitude but with completely independent phase processes. Also, for the purposes of simulations, each of these phase processes is assumed to be a moving average process of a specified correlation interval K and variance (steady-state). The effectiveness of the different least-squares algorithms is measured in terms of the power ratio (in dB) of the reconstructed (combined) signal to the total received power (which would be concentrated in the central feed element under ideal conditions). In addition, the performance of the least-squares multi-element combining algorithm is compared against traditional single-element processing. It should be noted that for all of these simulation experiments, the noise variance at the combiner output is matched to the noise variance at the output of a single feed (the weighting parameter vector $\underline{\hat{w}}$ is normalized to have unit norm).

In the simulations reported here, it is assumed that the received signal is unmodulated with the signal field amplitude equal to 1 in the aperture plane, corresponding to a signal amplitude A_0 equal to 16 at the center feed in the focal plane (ideal case). In case of data modulation, a decision-directed approach may be used to remove the data modulation from the received signal. For the purposes of simulations, the sampling period is normalized to 1 sec. Thus, the variance σ^2 of the sampled complex envelope of the noise at any of the feed outputs used in the simulations is given by

$$\sigma^2 = A_0^2 / N_0 f_s \tag{21}$$

where f_s is the sampling frequency; N_0 is the one-sided noise spectral density; $P = A_0^2/2$ is the received-signal power; and the algorithm's performance is plotted as a function of the carrier-to-noise spectral density ratio: CNR $= (P/N_0)$ (in dB-Hz).

In Fig. 2, the reconstructed signal amplitude and phase estimates as a function of time measured in number of samples for the least-squares algorithm for a CNR of 10 dB-Hz are plotted. It may be observed from Fig. 2(a) that there is a considerable signal loss compared to the received signal amplitude equal to 16 for the ideal case. In the simulations, the initial estimate for the parameter vector $\hat{w}$ is selected to be $[\alpha \quad \alpha \ldots \alpha]$ with $\alpha = 1/16$. As is apparent from Fig. 2(b), the phase-estimation error of the reconstructed signal is much smaller compared to the rms phase error of 1 rad introduced in the simulated received signal field. Figure 3 plots the corresponding results for the modified least-squares algorithm for the same set of signal parameters and for $\gamma_0 = 100$. Comparison of Figs. 2(a) and 3(a) shows a very significant performance improvement due to the proposed modification of the least-squares algorithm. Results are plotted in Fig. 4 for the case of 20 dB CNR with the parameter γ_0 equal to 200. For this case, the signal amplitude and phase of the reconstructed signal are quite close to their respective values for the ideal case. The signal amplitude loss for this case is only 1.09 dB relative to the ideal case, and the rms phase error (after adaptive combining) is 0.1 rad.

A. Simulation Results for Least-Squares Algorithm I

Figure 5(a) plots the sample estimates of the signal power loss (compared to the ideal case) for the standard least-squares algorithm I. The signal loss P_L is simply computed as $P_L = 20 \log_{10}(\hat{A}_{rms}/16)$ with

$$\hat{A}_{rms} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \hat{A}_i^2}$$

where $\hat{A}_i$ is the reconstructed signal amplitude at the ith sampling instance, and M is the number of sample values selected to be equal to 200 for these simulations. Figure 5(b) plots the signal estimation rms phase error $\hat{\Theta}_{rms}$ computed as

$$\hat{\Theta}_{rms} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_i^2}$$

where $\hat{\theta}_i$ is the phase of the combined signal at the ith sampling instance. It is apparent from these figures that although the least-squares algorithm is optimal with respect to the prediction error criterion, it is not satisfactory in terms of the signal-to-noise ratio of the combined signal. For the case of simulated phase dynamics, there is an asymptotic signal loss of 4 dB for the high CNR ($\sim$ 20 dB) case.

Figure 6 plots the performance of the modified least-squares algorithm with $\gamma_0 = 100$ and with the same set of signal parameters as for the case of Fig. 5. The results are computed for three different values of the weighting coefficient: λ equal to 0.925, 0.95, and 0.975. Comparison of Figs. 5(a) and 6 shows a dramatic improvement in performance due to the proposed modification. Thus, the asymptotic signal loss for this case and with $\lambda = 0.925$ is only 1.3 dB compared to a 4 dB loss for the standard least-squares algorithm. Increasing the value of γ_0 or reducing λ may further reduce the signal loss.

It is not difficult to understand this marked performance difference between the two algorithms. By minimizing the sum of the norm square of the estimated signal error and the noise variance of the combined signal output, the standard least-squares algorithm produces a relatively large noise variance at the output. This is so because at high CNR the contribution of noise to the total error is relatively much smaller than the contribution due to signal error (measured as a fraction of total signal), and, thus, the former is essentially ignored by the algorithm. In the constrained least-squares algorithm, by increasing the relative weighting attached to the noise variance, the signal-to-noise ratio at the combiner output is increased. In fact, as is shown in the Appendix, whereas the standard least-squares algorithm effectively maximizes the signal plus noise power output (subject to a near-orthogonality constraint), the constrained algorithm actually maximizes the signal-to-noise power ratio (subject to a similar near-orthogonality constraint) at the combiner output.

The signal loss results presented in Figs. 6(a)–(c) can also be used to infer the performance gains provided by the multi-element least-squares technique over center-feed processing. In particular, we can define the array processing gain as the difference (in dB) between the combining losses of the least-squares and center-feed curves. These data are plotted in Fig. 6(d).

As seen in Fig. 6(d), the array processing gain is a function of both CNR and λ. Best results occur for large CNR ($>$ 15 dB) and low values of λ ($\lambda = 0.925$). In par-

ticular, a maximum array gain of 3 dB (Fig. 6(a)) can be achieved for this simulation example. For lower values of CNR (below 0 dB), the estimation variance inherent in the least-squares algorithm degrades the array processing gain to such an extent that center-feed processing actually provides better performance.

However, it must be stressed that the simulation example considered here corresponds to a relatively high dynamic scenario, e.g., as compared with typical mechanically-induced array degradations. Lower dynamic scenarios permit longer time constants for the least-squares algorithm ($\lambda \rightarrow 1$ and/or longer sampling period), thereby reducing the algorithm estimation variance. Thus, for low dynamic scenarios, positive array gains can be achieved over a wider range of CNRs (including CNR $\ll 0$ dB) by utilizing time constants that are essentially matched to the dynamics. For sampling periods with T different from 1 sec, the results of Fig. 6 are applicable with the CNRs scaled by T. For instance, with $T = 10$ sec (typical for slower processes), the normalized CNR of 0 dB in Fig. 6 will correspond to the actual CNR of -10 dB.

It should also be noted that for a given scenario, increasing λ to such an extent that the algorithm cannot track the dynamics will lead to degraded system performance. This can be clearly seen from Fig. 6, where it is observed that the array gain for $\lambda = 0.975$ is approximately 1 dB less than for $\lambda = 0.925$. Finally, it can be observed from Fig. 7 that in contrast to the array gain results, there is little difference in rms phase error between the least-squares and center-feed outputs.

B. Simulation Results for Least-Squares Algorithm II

Figures 8(a) and (b) plot the performance of least-squares algorithm II in terms of signal reconstruction. In these figures, the dashed graphs depict the amplified signal amplitude A_i at the output of various feeds, while the solid-lined graphs represent the amplitude of the reconstructed field, i.e., the magnitude of various components of $\underline{h}(k)$ obtained by the circular convolution of the weight vector $\underline{\hat{f}}(k)$ obtained from Eq. (19) or Eq. (20), and the received signal vector $\underline{g}(k)$, for two different time indices equal to 30 and 185, respectively. As for the case of least-squares algorithm I, it is assumed that the weight vector $\underline{\hat{f}}$ is normalized to have its norm equal to 1. This ensures that the noise variance at various points of the reconstructed field is equal to the variance of the input noise field and, thus, the comparison in terms of signal amplitudes is equivalent to comparing the reconstructed signal-to-noise power ratio. The results in Fig. 8 correspond to the same set of signal parameters as for least-squares algorithm I and a CNR of 10 dB. As is apparent from Fig. 8, the least-squares algorithm II focuses most of the signal power that is originally dispersed in 16 taps into the center tap. A more appropriate measure of the effective focusing is obtained by the signal amplitude of the center tap of the reconstructed field, which is plotted versus time in Fig. 9(a). In this case, only -5.2 dB of the received signal power is scattered in the other taps for the reconstructed field. Figure 9(b) plots the phase error of the center tap signal and shows an rms phase error of 0.12 rad. Figure 10 plots the corresponding results for the case of $\gamma_0 = 100$ and differs insignificantly from the corresponding results in Fig. 9. Thus, the least-squares algorithm II simultaneously optimizes the signal-to-noise ratio. Figure 11(a) plots the signal loss in the center feed of the reconstructed signal for the least-squares algorithm II. As shown in the figure, with the parameter $\lambda = 0.925$, a loss of 1.25 dB can be achieved for the high CNR case, which is similar to that obtained for modified least-squares algorithm I. Figure 11(b) plots the corresponding results for the rms phase error of the reconstructed signal, showing an rms phase error of about 0.12 rad at CNRs higher than 10 dB.

Figure 12 plots the results when the algorithm is applied to a configuration of feeds connected to amplifiers with different noise figures. For this example, the case wherein one-half of the total number of amplifiers have 6.0 dB higher noise temperature than others is considered. The CNR in the figure is still measured with reference to the amplifier with the lower noise temperature. As may be inferred from Fig. 12(a), the asymptotic signal loss (CNR > 10 dB) in this case is 2.2 dB as opposed to 1.2 dB for the case in which all of the amplifiers have low noise temperature, thus resulting in only 1 dB additional degradation. Note that if all the amplifiers were replaced by ones with higher noise temperature, then the degradation would be about 3 dB with reference to this lower CNR, thus resulting in an effective loss of 9 dB.

It may be remarked that in the above presentation, the sampling period T has been normalized to 1 sec, but the results are also applicable to different sampling periods by a simple normalization. As is apparent from Eq. (21), while increasing the sampling rate by a factor K, one should correspondingly reduce the actual CNR by the same factor to obtain the algorithm's performance for this case.

VII. Conclusions

From the simulations presented in the article, it can be observed that for the relatively fast distortion process

with a moderate dispersion and the array geometry considered, the multi-element array configuration provides an improvement of about 3 dB over a single-feed system. For a slower process with possibly higher spatial dispersion, the improvement is expected to be higher and to a certain extent will also be influenced by a match between the array geometry and the pattern of the received signal power dispersion.

Acknowledgments

The simulation model and the subroutine to generate the distorted phase process used in the simulation of the proposed algorithm were written by Victor Vilnrotter. The author thankfully acknowledges this and the discussion concerning the concept of deconvolution for this problem held with Victor Vilnrotter and George Zunich during this work.

References

[1] W. A. Imbriale et al., "Ka-Band (32-GHz) Performance of 70-Meter Antennas in the Deep Space Network," *TDA Progress Report 42-88*, vol. October–December 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 126–130, February 15, 1987.

[2] J. W. Layland and J. G. Smith, "A Growth Path for Deep Space Communications," *TDA Progress Report 42-88*, vol. October–December 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 120–125, February 15, 1987.

[3] S. J. Blank and W. A. Imbriale, "Array Feed Synthesis for Correction of Reflector Distortion and Vernier Beamsteering," *TDA Progress Report 42-86*, vol. April–June 1986, Jet Propulsion Laboratory, Pasadena, California, pp. 43–55, August 15, 1986.

[4] P. D. Potter, "64-Meter Antenna Operation at Ka-Band," *TDA Progress Report 42-57*, vol. March and April 1980, Jet Propulsion Laboratory, Pasadena, California, pp. 65–70, June 15, 1980.

[5] N. Mohanty, *Random Signals Estimation and Identification*, Van Nostrand Reinhold, New York, New York, 1986.

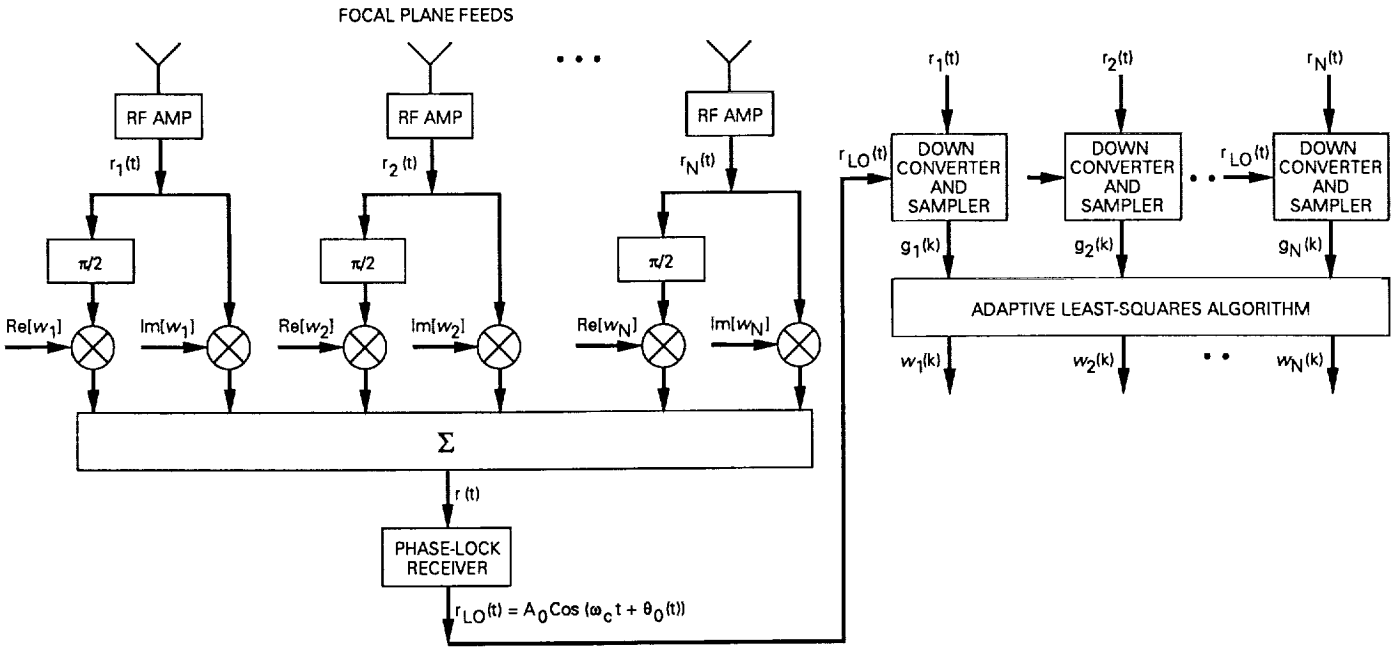


Fig. 1. RF signal combining in focal plane.

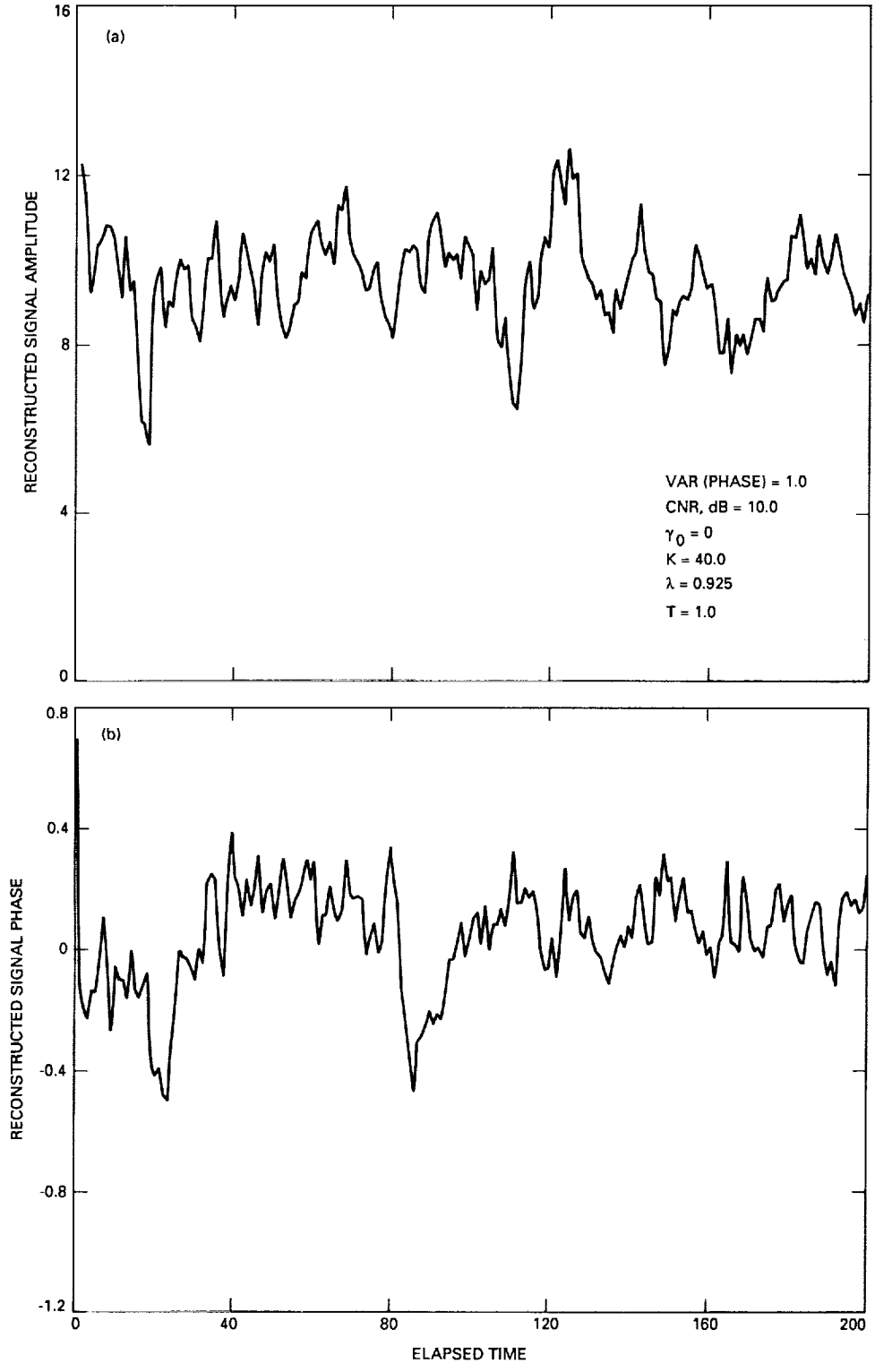


Fig. 2. Least-squares algorithm I: (a) reconstructed signal amplitude; (b) reconstructed signal phase.

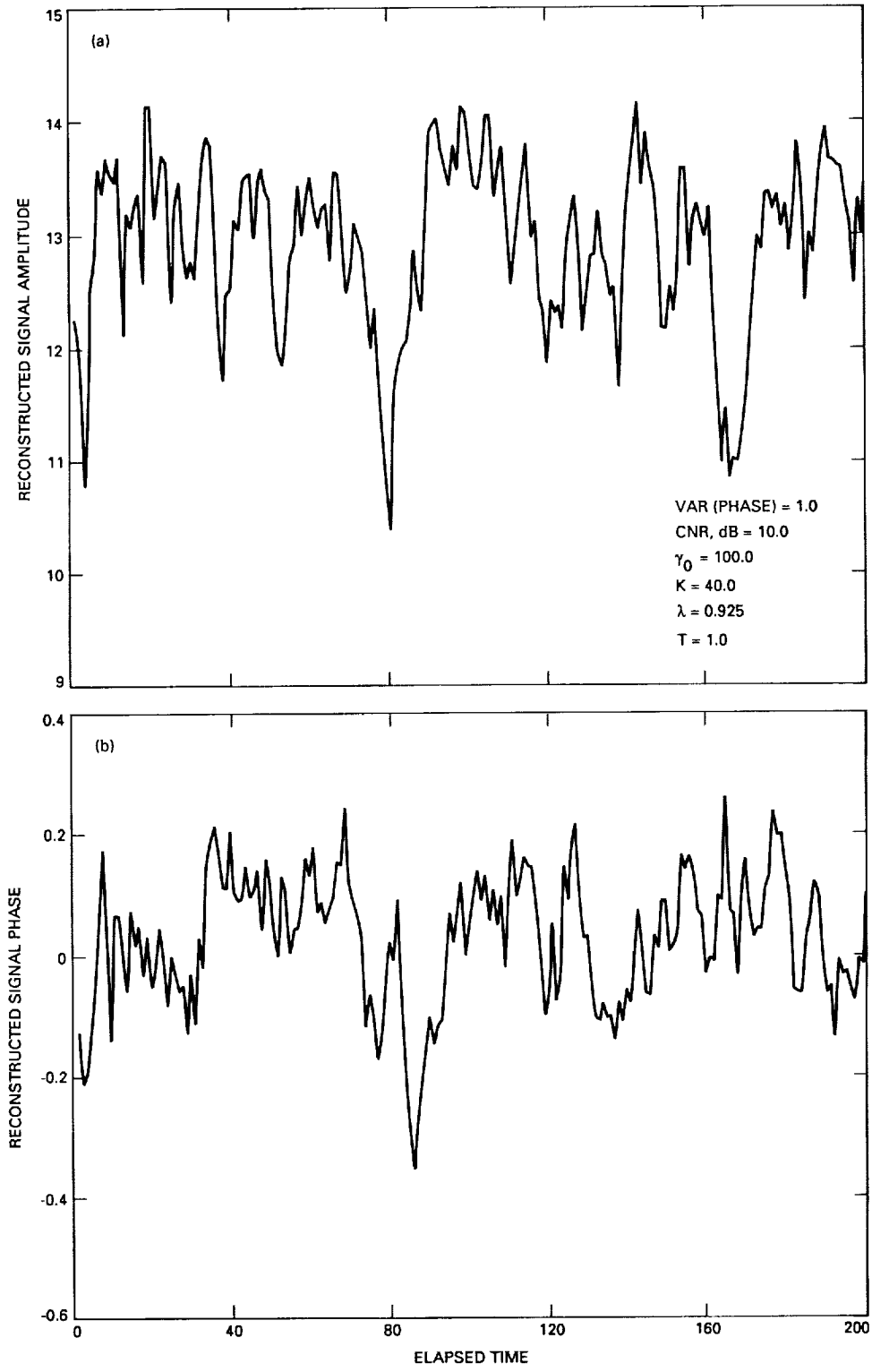


Fig. 3. Modified least-squares algorithm I: (a) reconstructed signal amplitude versus time; (b) reconstructed signal phase versus time.

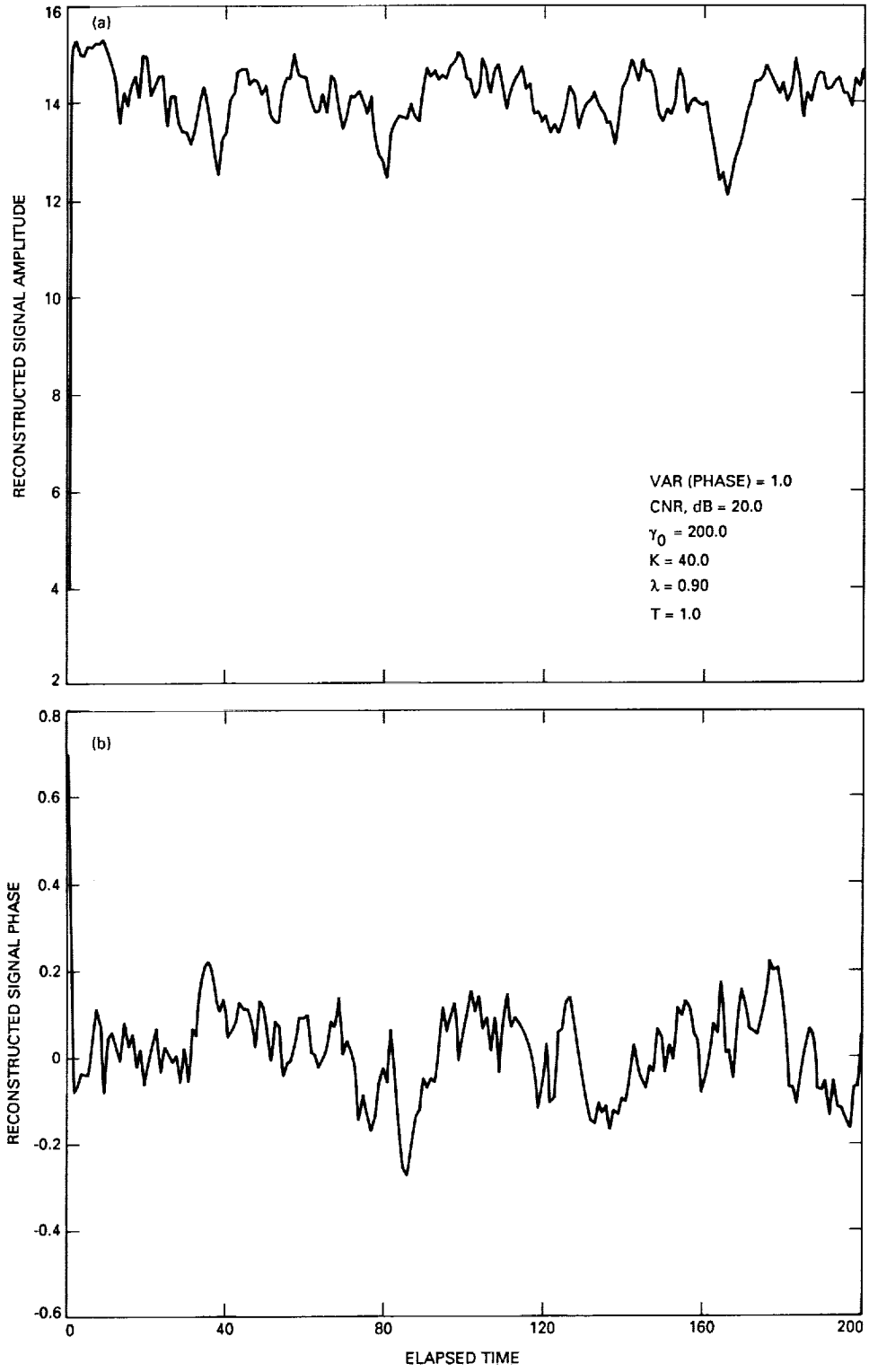


Fig. 4. Modified least-squares algorithm I: (a) reconstructed signal amplitude versus time at 20 dB CNR; (b) reconstructed signal phase versus time at 20 dB CNR.

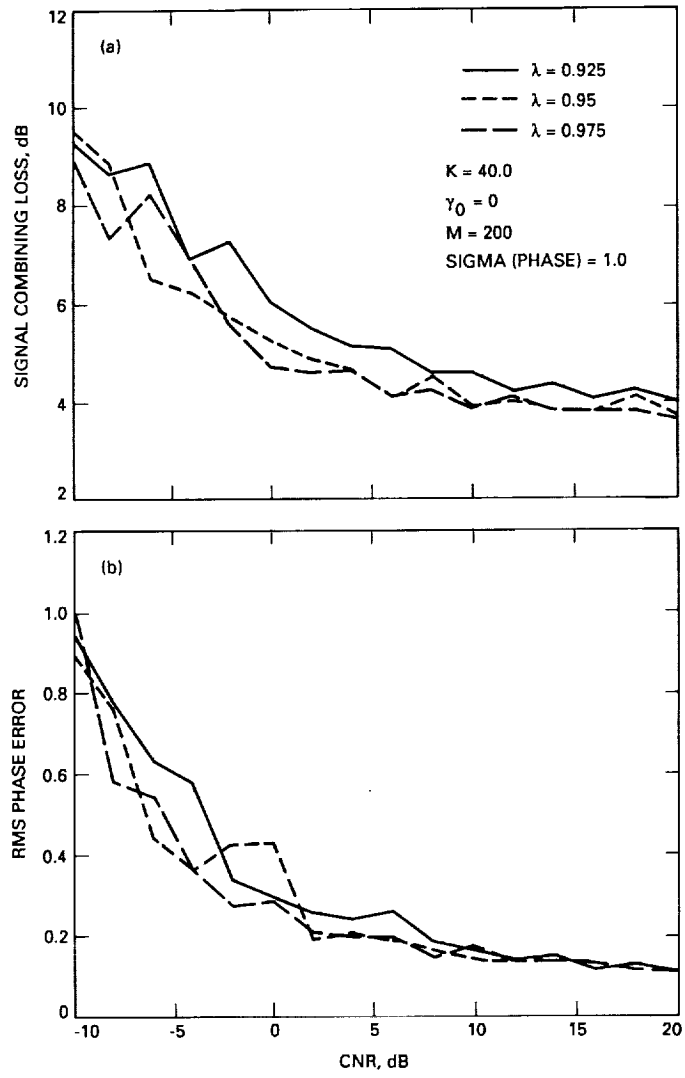


Fig. 5. Least-squares algorithm I: (a) signal combining loss versus CNR for reconstructed signal; (b) rms phase error versus CNR for reconstructed signal.

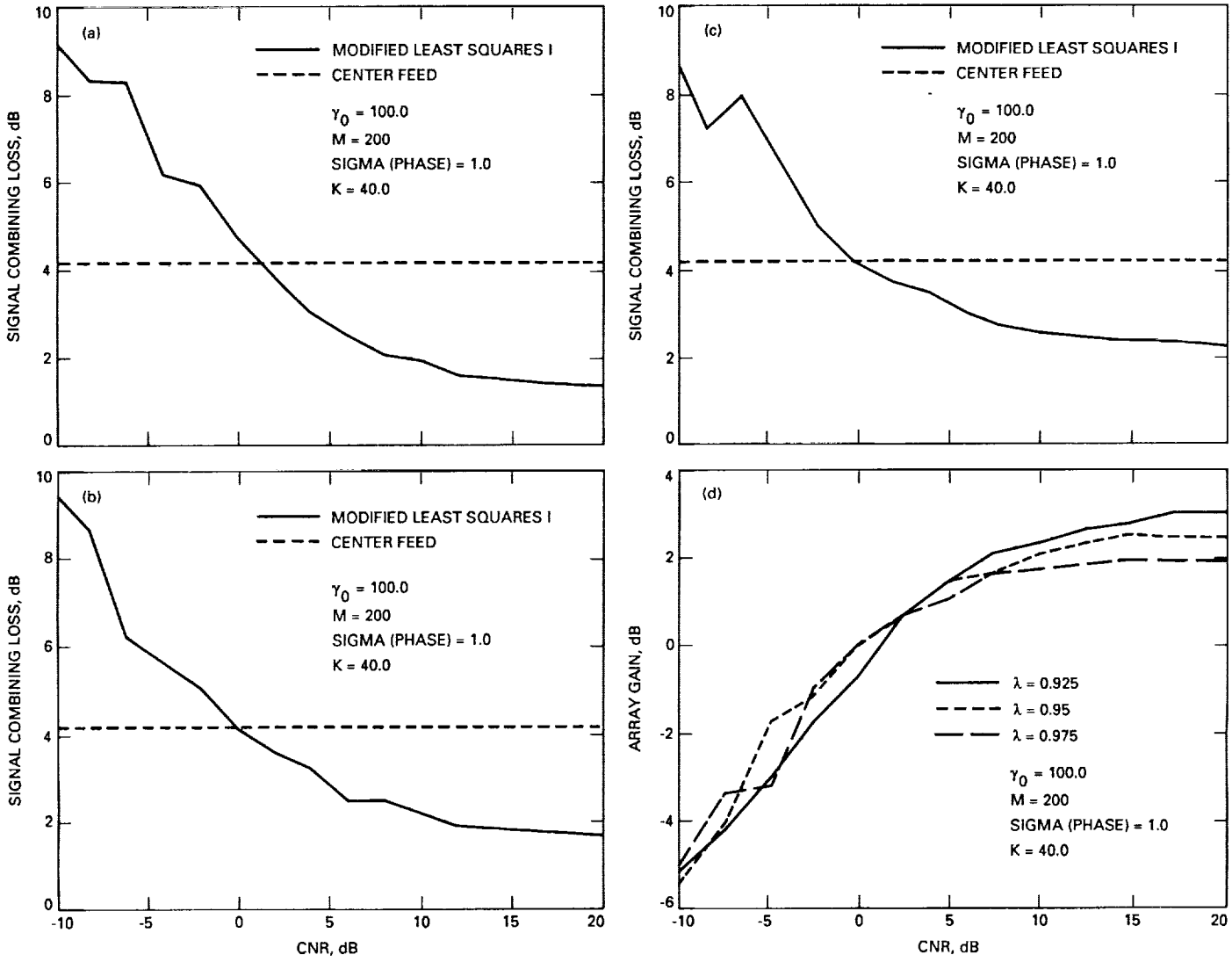


Fig. 6. Signal combining loss versus CNR for reconstructed signal: (a) $\lambda = 0.925$; (b) $\lambda = 0.95$; (c) $\lambda = 0.975$. Also (d) modified least-squares array gain versus CNR.

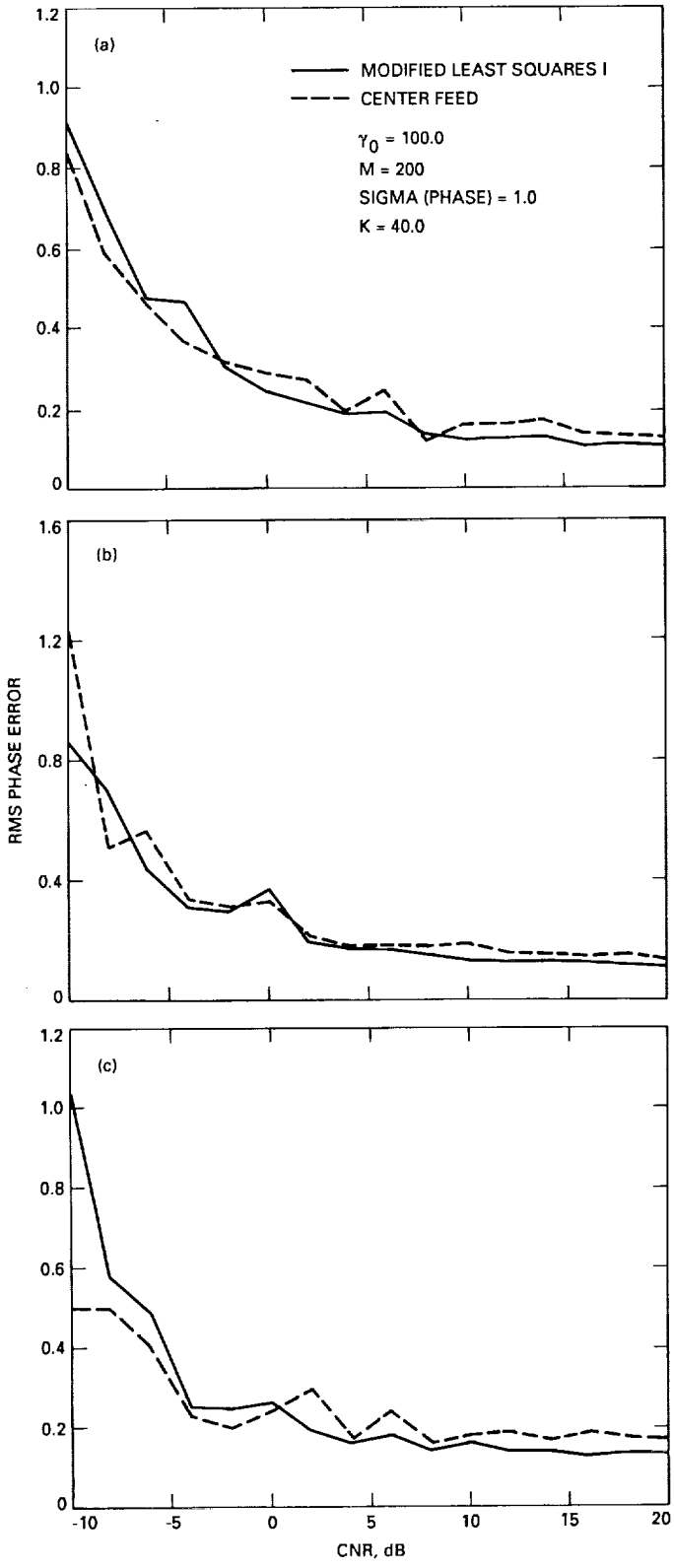


Fig. 7. rms phase error versus CNR for reconstructed signal: (a) $\lambda = 0.925$; (b) $\lambda = 0.95$; (c) $\lambda = 0.975$.

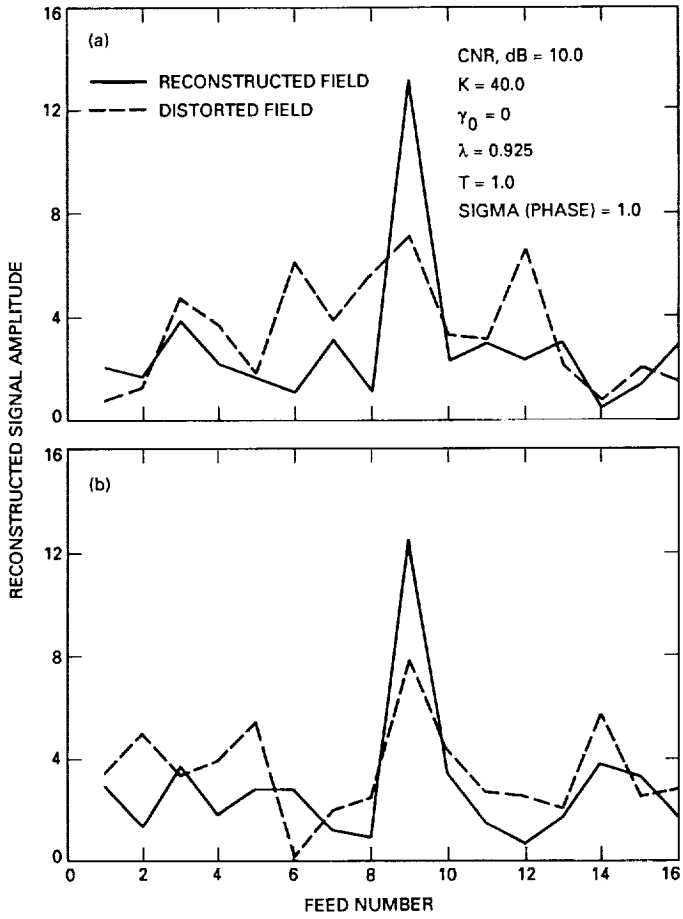


Fig. 8. Focal plane reconstructed field amplitude, least-squares algorithm II: (a) after 30 samples; (b) after 185 samples.

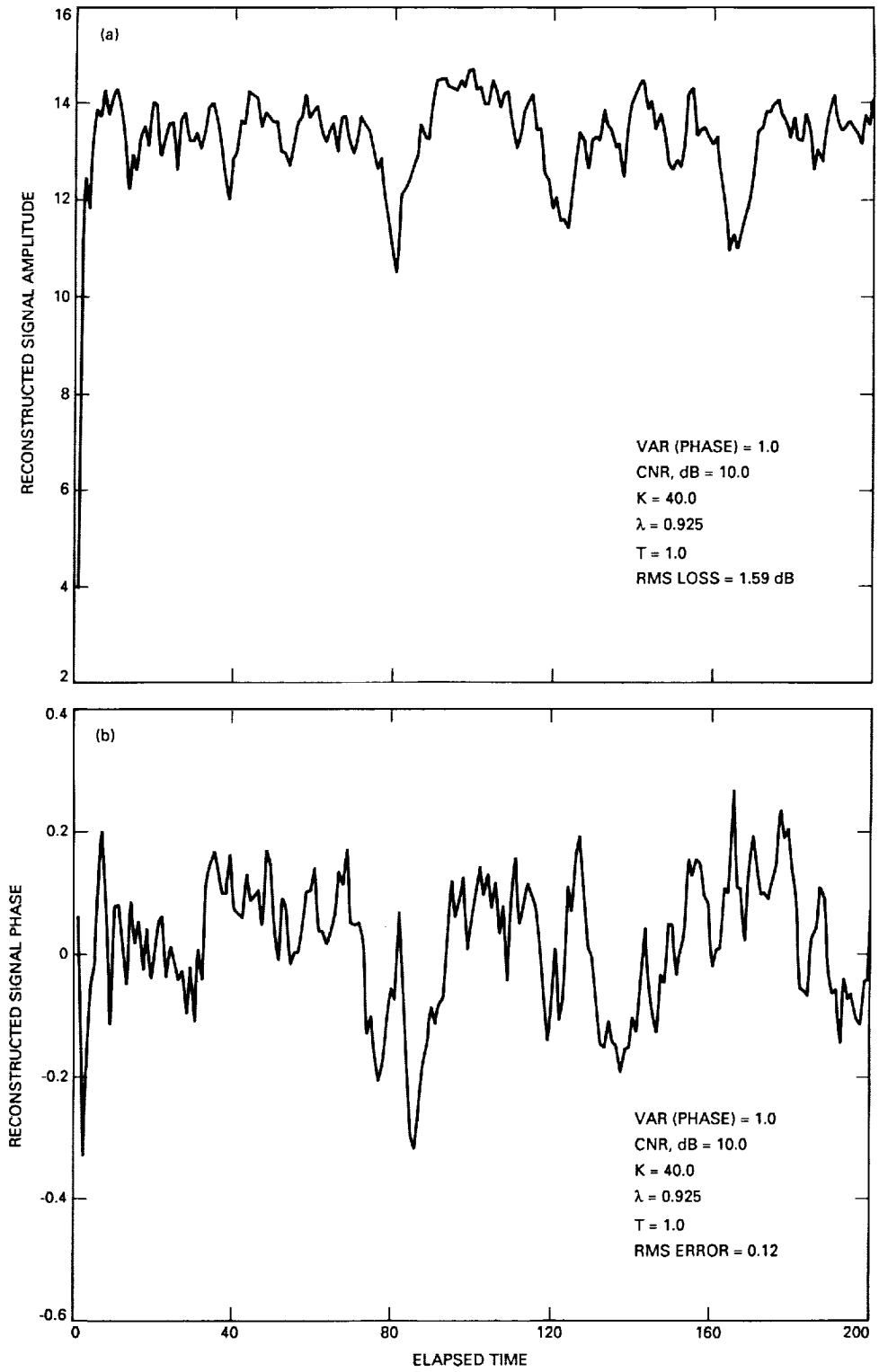


Fig. 9. Least-squares algorithm II: (a) focal plane reconstructed signal amplitude versus time, $\gamma_0 = 0$; (b) focal plane reconstructed signal phase versus time, $\gamma_0 = 0$.

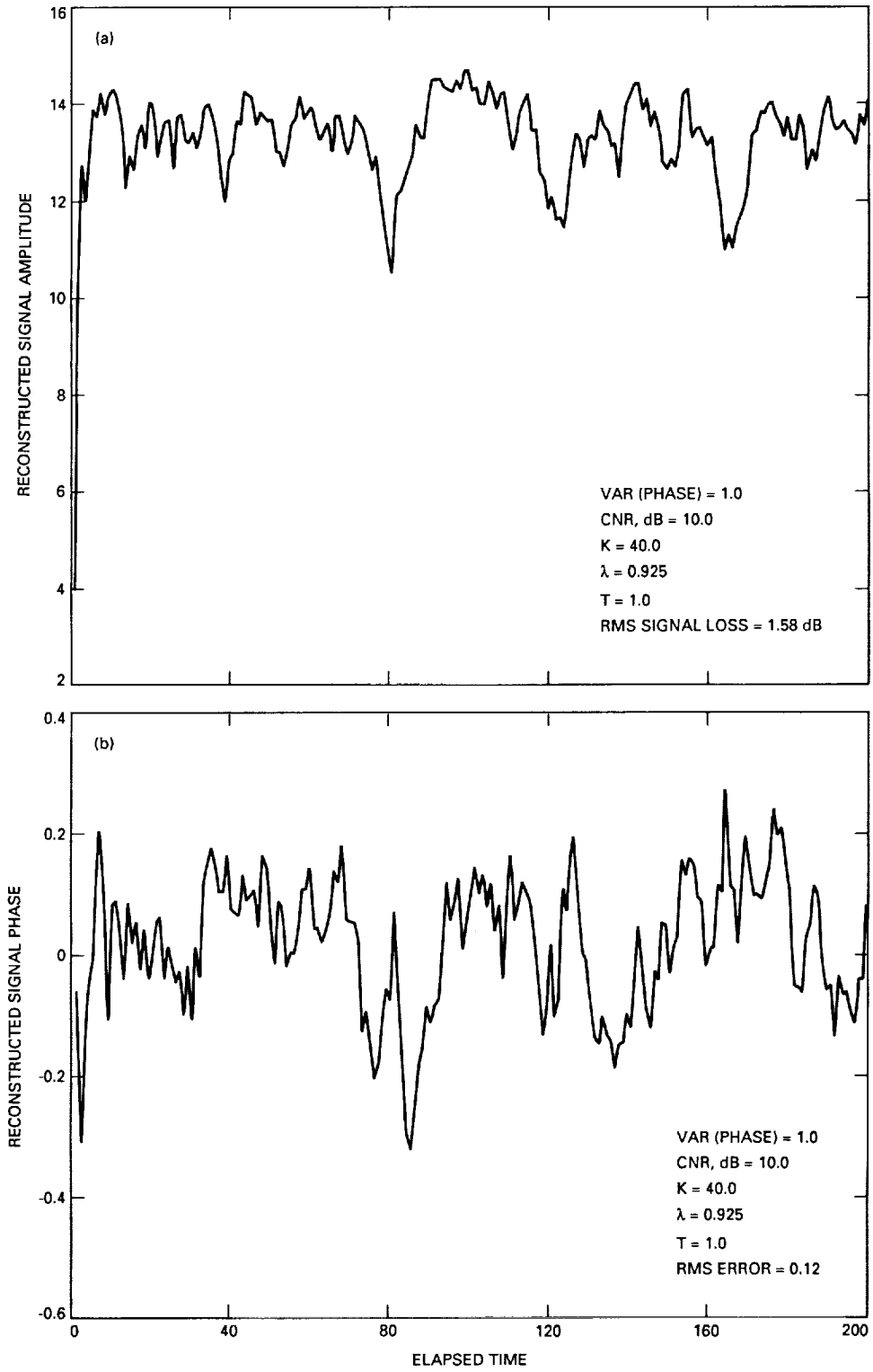


Fig. 10. Least-squares algorithm II, $\gamma_0 = 100$: (a) focal plane reconstructed signal amplitude versus time; (b) focal plane reconstructed signal phase.

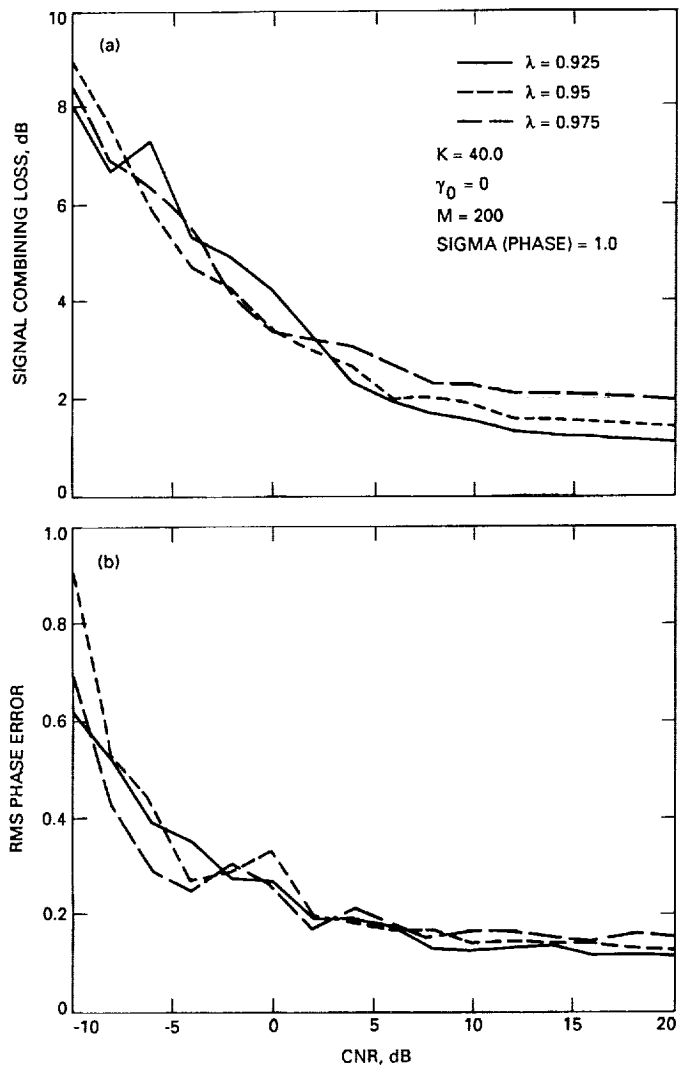


Fig. 11. Least-squares algorithm II: (a) signal combining loss versus CNR for reconstructed field; (b) rms phase error versus CNR for reconstructed field.

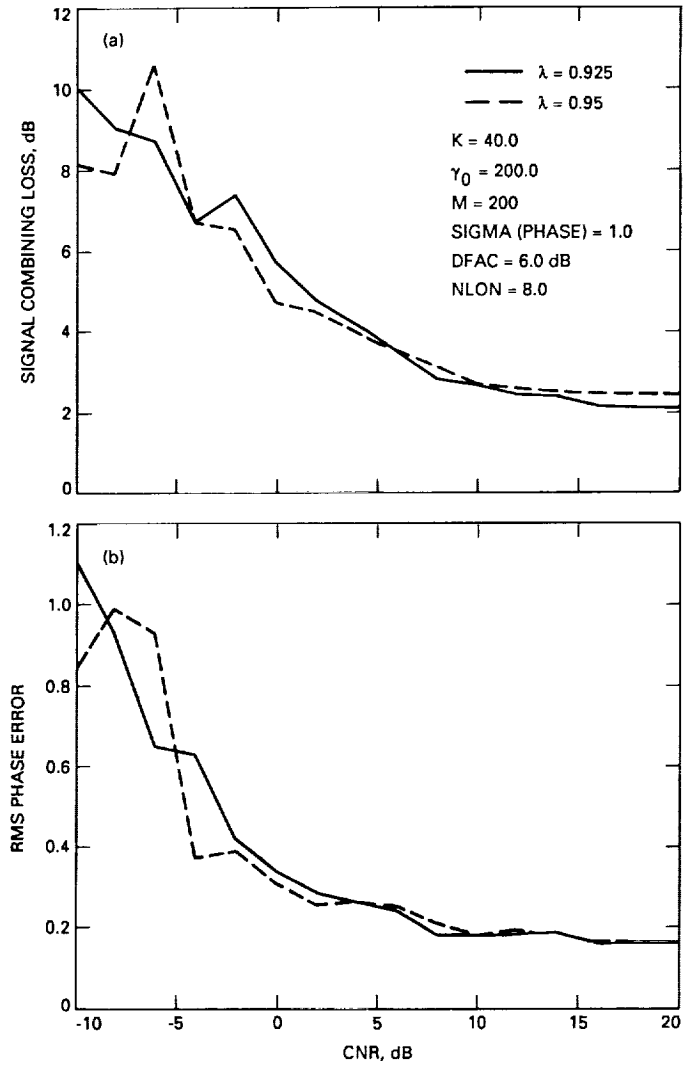


Fig. 12. Modified least-squares algorithm I, different noise figures: (a) signal combining loss versus CNR for reconstructed field; (b) rms phase error versus CNR for reconstructed field.

Appendix

The following shows that the modified least-squares algorithm I of Section IV achieves constrained maximization of the signal-to-noise ratio. In the first instance, the time averages are replaced by the ensemble averages.

Denoting by $s_i(k)$ the signal component of the ith array element output, consider the problem of minimizing

$$H = E\left[|S - \sum_{i=1}^{N} w_i^* s_i|^2\right] \tag{A-1}$$

with respect to $w_i, i = 1, 2, \ldots N$. Setting the partial derivative of H w.r.t.w_i to zero yields

$$E\left[\left(S - \sum_{i=1}^{N} w_i^* s_i\right) s_i^*\right] = 0 \tag{A-2}$$

Now with $\widehat{S} \triangleq \underline{w}^H \underline{s}$ and $\underline{s} \triangleq [s_1, s_2, \ldots, s_N]'$, the index H may be written as

$$E\left[|S - \widehat{S}|^2\right] = E\left[|S|^2 + |\widehat{S}|^2 - S\widehat{S}^* - S^*\widehat{S}\right] \tag{A-3}$$

At the optimal point, the following is obtained from Eq. (A-2).

$$E\left[(S - \widehat{S})\widehat{S}^*\right] = 0 \tag{A-4}$$

Adding the left-hand side of Eq. (A-4) and its complex conjugate to Eq. (A-3) yields the following form for the optimization index, subject to the constraint, Eq. (A-4).

$$H = |S|^2 - E\left[|\widehat{S}|^2\right] \tag{A-5}$$

Thus, the algorithm that minimizes Eq. (A-1) also maximizes $E\left[|\widehat{S}|^2\right]$, subject to the constraint, Eq. (A-4), i.e., it is also a signal maximization algorithm. There may also exist solutions that optimize $E\left[|\widehat{S}|^2\right]$ without the constraint, Eq. (A-4), but these may result in large phase error with respect to S, the desired signal, i.e., $\widehat{S}$ and $\widehat{S}e^{j\phi}$ (for any random phase ϕ) both have the same value of the index $E\left[|\widehat{S}|^2\right]$ but only one of these would simultaneously minimize the error function, Eq. (A-1). It may be remarked that there are, in general, an infinite number of solutions that satisfy Eq. (A-4), and in effect these orthogonalize the estimate $\widehat{S}$ and the "error" $(S - \widehat{S})$. Among these solutions, the one maximizing $E\left[|\widehat{S}|^2\right]$ is selected. Equation (A-4) is termed the orthogonality constraint.

The optimization index, Eq. (A-1), does not include the noise variance at the output of the array combiner, which is given by $\sigma^2||\underline{w}||^2$ where σ^2 is the variance of $\nu_i(k)$, the noise at the input of the combiner. Thus, now Eq. (A-1) is minimized subject to the constraint

$$||w||^2 = K \tag{A-6}$$

for some constant K. Or, one can simply minimize

$$E\left[|S - \underline{w}^H \underline{s}|^2\right] + \beta\left(||w||^2 - K\right) \tag{A-7}$$

where β is the Lagrangian multiplier. An analysis similar to derivation of Eq. (A-5) shows that with a constraint similar to Eq. (A-4), the index is given by

$$|S|^2 - E\left[|\widehat{S}|^2\right] - 2\beta K \tag{A-8}$$

for some constants β and K. Thus, again the algorithm maximizes $E\left[|\widehat{S}|^2\right]$ subject to the constraint that the output noise variance is equal to a constant $K\sigma^2$, and thus effectively maximizes the output signal-to-noise ratio.

Now from the independence of the received signal s_i and noise ν_i it follows that

$$E\left[|S - \underline{w}^H \underline{g}|^2\right] = E\left[|S - \underline{w}^H \underline{s}|^2\right] + ||\underline{w}||^2\sigma^2 \tag{A-9}$$

Minimization of Eq. (A-9) subject to the constraint, Eq. (A-6), is thus identical to the minimization of Eq. (A-1) subject to the constraint, Eq. (A-6), and thus effectively maximizes the signal-to-noise ratio under the near orthogonality constraint. Now for the large value of k, the index $k^{-1}J_k$ with J_k given by Eq. (3) approaches the left-hand side of Eq. (A-9) under appropriate ergodicity assumptions, and the algorithm of Section IV thus achieves constrained optimization of the signal-to-noise power ratio.

It may be remarked that the least-squares algorithm in the absence of the constraint, Eq. (A-6), effectively minimizes the following index

$$|S|^2 - E\left[|\widehat{S}|^2\right] - \sigma^2 ||\underline{w}||^2 \qquad \text{(A-10)}$$

subject to a constraint similar to Eq. (A-4). Since the last term in Eq. (A-10) represents the noise power at the combiner output, it can be observed that the standard least-squares algorithm effectively maximizes the sum of the signal plus noise power rather than the signal-to-noise power ratio.

The Effects of Sinusoidal Interference on the Second-Order Carrier Tracking Loop Preceded by a Bandpass Limiter in the Block IV Receiver

C. J. Ruggier
Telecommunications Systems Section

Drop-lock relationships for the second-order phase-locked loop are derived when the carrier and a sinusoidal interference signal lie within the predetection filter bandwidth of the Block IV receiver. Limiter suppression factors are calculated when a bandpass hard limiter is used to maintain constant total power at the loop. The parameters of interest are the interference-to-signal power ratio (ISR), the input signal-to-noise power ratio (SNR), and the interference signal frequency offset from carrier Δf. Limiter suppression caused by the combined effects of the noise and the interference signal accounts for the variability in the drop-lock threshold for given values of the input SNR and ISR parameters. This article goes beyond earlier published work that focused on the limiter's effect on the drop-lock threshold; it accounts for the limiter action in the interference mode and provides an overall improvement in the prediction accuracy of the drop-lock model.

I. Introduction

One major application of the phase-locked loop in a DSN receiver is tracking the carrier of the received signals [1]. The receiver phase-locks to the carrier and loses lock when the carrier margin drops below the lock threshold, or when an interfering signal is received at the critical amplitude and frequency offset from the carrier. Although telecommunications links are designed with sufficient margins to ensure performance requirements for the lifetime of the mission, interference can occur at any time. If the interfering signal power and frequency exceed the threshold limit, the carrier tracking loop drops the weaker carrier signal and locks up to the stronger interference signal. This jump phenomenon is due to the inherent nonlinearities present in the phase-locked loop for conditions when the interference-to-noise power ratio (INR) is sufficiently high. As the INR decreases, the signal-to-noise ratio (SNR) becomes the dominant factor, which can cause the loop to lose lock when it decreases below the noise threshold level. This article investigates the effect of the bandpass limiter when drop-lock of this type occurs.

The carrier tracking loop employed in the Block IV receiver consists mainly of a second-order phase-locked loop

preceded by a bandpass limiter. The hard limiter provides constant power at the input to the loop and effectively minimizes the total mean-square error of the loop over a wide range of the SNR. If, in addition to the noise, a sinusoidal interference signal is inserted into the limiter along with the carrier signal, the interference signal will also contribute to the limiter suppression. The limiter's effect on the drop-lock threshold becomes evident from its impact on the loop gain and loop interference-to-signal power ratio (ISR) loop input. Limiter suppression factors for these parameters are calculated and incorporated into the basic drop-lock model to improve its prediction accuracy for large variations in the loop SNR.

II. Carrier Drop-Lock Model

Figure 1 shows a representative second-order phase-locked loop preceded by a bandpass limiter. Bruno [2] derived the loop equations for the case of a strong signal and a sinusoidal interferer, without the limiter. The voltage-controlled oscillator (VCO) output is equal to

$$2\cos\Big[\omega_1 t + \phi_0(t)\Big]$$

where ω_1 is the VCO frequency (rad/sec) and $\phi_0(t)$ is the phase modulation due to the input through the loop action. The phase detector is assumed to be a perfect multiplier, and the loop filter has a transfer characteristic described as $F(s)$.

Ignoring the effects of narrow-band Gaussian noise, the input to the loop consists of the sum of two sinusoidal components:

$$A\sin(\omega_c t) + B\sin(\omega_c + \Delta\omega)t \qquad (1)$$

where the first term of Eq. (1) is the wanted signal component with frequency ω_c having constant amplitude A volts when the limiter reaches a constant output. The interference component has an amplitude equal to B volts and is offset in frequency from the signal component by an amount equal to $\Delta\omega$. Defining $\sqrt{\text{ISR}}$ as B/A, Eq. (1) can be rewritten as

$$A\sin(\omega_c t) + \sqrt{\text{ISR}}\,A\sin(\omega_c t + \Delta\omega)t$$

The output modulation $\phi_0(t)$ is given by

$$\phi_0(t) = \frac{\text{KF}(p)}{p}\left[-\sin\phi_0(t) + \text{ISR}\sin\Big(\Delta\omega t - \phi_0(t)\Big)\right]$$

where p represents the operator d/dt, $F(p)$ is the loop filter, and K (1/sec) is the open-loop gain, which includes the VCO and the phase-detector loop gain. This nonlinear differential equation cannot be solved analytically; however, using the method of perturbations, solutions with best-approximation trial functions can be obtained. A steady-state trial solution is assumed to be

$$\phi_0(t) = \lambda + \theta\sin(\Delta\omega t + \nu)$$

where λ represents the static phase error, θ is the phase deviation, and ν is the phase shift. Bruno [2] derived the lock constraints as

$$\sin\lambda = \frac{-\theta^2\delta\cos\psi}{2J_0(\theta)} \qquad (2)$$

$$\sin(\lambda - \nu) = \frac{-\theta^2\delta\cos\psi}{2\sqrt{\text{ISR}}\,J_1(\theta)}$$

$$\left[\frac{\theta\delta\sin\psi + 2J_1(\theta)\cos\lambda}{J_0(\theta) - J_2(\theta)}\right]^2 + \left[\frac{\theta^2\delta\cos\psi}{2J_1(\theta)}\right]^2 = \text{ISR} \qquad (3)$$

where J_0, J_1, and J_2 are Bessel functions of the first kind, ψ is the phase angle of $F(s)$, and δ is the normalized offset frequency

$$\delta = \frac{\Delta\omega}{K|F(s)|}$$

where $s = j\Delta\omega$, and K = the open-loop gain.

Restricting the second-order loop filter with transfer characteristics,

$$F(s) = \frac{1 + \tau_2 s}{1 + \tau_1 s}$$

where $\tau_1 \gg \tau_2$. With $\Delta\omega \gg 1/\tau_2$, one obtains the reasonable approximation

$$|F(s)| \approx \frac{\tau_2}{\tau_1} \qquad \text{for } \psi \approx 0$$

The loop is expected to drop lock when the static phase error approaches 90 deg. Applying this condition and using Eqs. (2) and (3) with the condition that $J_1(\theta) \approx \theta/2$ and $J_0(\theta) \approx 1$ for small phase deviations, the critical ISR can be given as

$$(\text{ISR})_c = \frac{4\pi\Delta f}{K\,(\tau_2/\tau_1)}$$

This describes the drop-lock threshold for critical values of ISR and offset frequency Δf without the limiter action.

III. Calculation of the Limiter Suppression Factors

The effect of the bandpass limiter also needs to be taken into account, when the carrier, interfering sinusoidal signal, and narrow-band Gaussian noise are present at the input to the limiter. Jones [3] calculated the autocorrelation function of the ideal hard limiter under similar conditions. The interaction of the two signals s_1 and s_2 with noise generates a filtered output with autocorrelation function given by

$$R(\tau) = \sum_{i=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} \sum_{k=|i|,|i|+2}^{\infty} 2\frac{b^2 k|\ell|\left|\ell - |i+1|\right|}{\left(\frac{k+|i|}{2}\right)!\left(\frac{k-|i|}{2}\right)!} \rho^k(\tau)$$

$$\times \cos\left[|i|\omega_c - |i+1|\omega_2 + \ell(\omega_2 - \omega_1)\right]\tau$$

where ω_1 and ω_2 represent the frequencies of s_1 and s_2 respectively, ω_c is the bandpass filter center frequency, and ρk is the normalized noise-power envelope function containing both discrete (due to the period terms) and continuous components (associated with the output noise). The total power contained in these discrete components at the output is then given by

$$R^{(\tau)}_{s_1 x s_2} = \sum_{\ell=-\infty}^{\infty} 2b_0^2|\ell\,||\,\ell - 1|\cos\left[\ell(\omega_2 - \omega_1) - \omega_2\right]\tau$$

and the output signal powers are given as

$$(s_1)_0 = 2b_{010}^2 = \frac{2}{\pi^2}\left(\frac{s_1}{N}\right)$$

$$\left[\sum_{i=0}^{\infty} \frac{(-1)^i (s_1/N)^i}{i!(i+1)!}\,\Gamma\left(i+\frac{1}{2}\right)\,{}_2F_1\left(-i, i-1; 1; \frac{s_2}{s_1}\right)\right]^2$$

and

$$(s_2)_0 = 2b_{001}^2 = \frac{2}{\pi^2}\left(\frac{s_2}{N}\right)$$

$$\left[\sum_{i=0}^{\infty} \frac{(-1)^i (s_1/N)^i}{(i!)^2}\,\Gamma\left(i+\frac{1}{2}\right)\,{}_2F_1\left(-i, -i; 2; \frac{s_2}{s_1}\right)\right]^2$$

where Γ and ${}_2F_1$ are the gamma and confluent hypergeometric function, respectively. For the case where both the carrier and interference power are much greater than the noise power, the convergence properties of these equations become unstable. Then it becomes necessary to use the asymptotic forms

$$(s_1)_0 = \frac{2}{\pi^2}\left(\frac{s_1}{s_2}\right)\left[{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 2; \frac{s_1}{s_2}\right)\right]^2$$

and

$$(s_2)_0 = \frac{2}{\pi^2}\left(\frac{s_1}{s_2}\right)\left[\frac{\Gamma(1/2)}{\Gamma(3/2)}\,{}_2F_1\left(\frac{1}{2}, -\frac{1}{2}; 1; \frac{s_1}{s_2}\right)\right]^2$$

for

$$\frac{s_2}{N} \to \infty;\ \frac{s_1}{s_2} < 1;\ k = 0$$

These relationships can be used to calculate the limiter suppression on the carrier power and the power ratio of the interference and carrier signal. For the case where interference is not present, the limiter suppression factor reduces to that calculated by Davenport [4] for a sinusoid and noise only. With interference present, the limiter suppression becomes affected by changes in both the ISR and SNR power ratios. Limiter suppression of the carrier signal from the limited strong signal peak level, which results in a corresponding suppression of the loop gain, is given by

$$\left[\frac{s}{8/\pi^2}\right]^{1/2}$$

Alternately, the limiter suppression of the output ISR with respect to the input ISR becomes

$$\frac{(\mathrm{ISR})_0}{(\mathrm{ISR})_i}$$

Together these factors allow adjustment of the model's critical values of ISR and frequency offset, at which carrier drop-lock occurs. The combined effect of the limiter can now be given as the product of these two ratios. The threshold limiter suppression product is defined as

$$S_t = \left[\frac{s}{8\pi^2}\right]^{1/2} \frac{(\mathrm{ISR})_0}{(\mathrm{ISR})_i}$$

showing that S_t is the suppression of loop gain times the suppression of interference-to-signal power ratio. Figure 2 illustrates the threshold limiter suppression product for varying carrier margin values.

IV. Effect of the Bandpass Limiter on the Drop-Lock Model

The limiter suppression product can now be used to determine the limiting effect caused mainly by the interference signal. The basic drop-lock equation can be written as

$$\mathrm{ISR} = \frac{4\pi\Delta f}{S_t K(\tau_2/\tau_1)}$$

Figures 3 through 8 show the critical ISR and frequency offset for the possible tracking loop modes of the Block IV receiver, with values of input SNR necessary to cause carrier drop-lock. Table 1 lists the various loop mode parameters used to generate the curves. Note that as the interference power increases, the limiter suppression approaches a constant level. The slope of the drop-lock equation is affected only slightly by the increasing interference power for high SNR, and unaffected for low SNR. This observation corroborates well with the trend shown in the measured data and indicates that the limiter action, overall, produces a linear translation of the drop-lock threshold.

V. Experimental Validation

Drop-lock threshold tests were conducted independently at the Telecommunications Development Lab (TDL) and at the Compatibility Test Area (CTA-21). The purpose of the tests was to validate the drop-lock model with the Block IV receiver under the conditions of interference described above. Only one tracking loop mode was tested for $2B_{LO} = 10$ Hz. Figures 9 and 10 show the comparison of the measured data obtained from TDL and CTA-21, respectively, to the theoretical curves for varying levels of SNR at the limiter input. The measurements were restricted to frequency offsets from 10 to 1000 Hz. Initial ISR power ratio was set 5 dB higher than the loop threshold SNR ($C/2B_{LO}N_O$), where C is the carrier power in watts, $2B_{LO}$ is the two-sided loop threshold noise bandwidth in Hz, and N_O is the noise-power density in watts per Hz.

VI. Conclusions

It has been shown that when a bandpass hard limiter precedes the carrier tracking loop, limiter effects on the drop-lock threshold can be calculated for variations in the input SNR and ISR power ratios. Limiter suppression of the loop gain and ISR is evidenced in the responses of the drop-lock threshold for strong input signal variations, with varying combinations of levels in the input SNR and ISR. However, as the input ISR increases, the limiter suppression stabilizes, exhibiting less sensitivity to varying levels in the input SNR.

Drop-lock calculations for the possible loop modes of the Block IV receiver indicate that the loop becomes more susceptible to interference as the loop gain and loop bandwidth increase. The predetection filter noise bandwidth is also a factor. For example, the filter provides no attenuation of the interference signal when its frequency lies within the filter passband; for this particular case, the narrow bandwidth modes only tend to reduce the drop-lock threshold. For larger interference frequency offsets outside the predetection filter passband, the calculations become more conservative due to the approximation of the filter's transfer characteristics in the basic loop model.

Experimental data from tests conducted at TDL and the CTA-21 facility show good agreement with the theoretical calculations. The model shows a tendency to underpredict at the low SNR values, which is inherent in the assumption of negligible noise in the derivation of the basic model. On the other hand, the model overpredicts for higher SNR values between 100- and 1000-Hz frequency offsets, attributable largely to the limiter model which was assumed to be ideal. The model shows the best agreement with the measured data (within 1 dB) for smaller frequency offsets and loop SNR of 10 dB or greater.

Acknowledgments

The author wishes to thank C. Westfall for his assistance with the computer programming and H. Olson, who performed the carrier drop-lock threshold tests at the TDL and CTA-21 test facilities.

References

[1] J. H. Yuen, editor, *Deep Space Telecommunications Systems*, New York: Plenum Press, 1983.

[2] F. Bruno, "Tracking Performance and Loss of Lock of a Carrier Loop Due to the Presence of a Spoofed Spread Spectrum Signal," *Proceedings of the 1973 Symposium on Spread Spectrum Communications*, vol. I, ed. M. L. Schiff, Naval Electronics Laboratory Center, San Diego, California, pp. 71–75, March 13-16, 1973.

[3] J. J. Jones, "Hard-Limiting for Two Signals in Random Noise," *IEEE Transactions on Information Theory*, IT-9, pp. 34–42, January 1963.

[4] W. B. Davenport, "Signal-to-Noise Ratios in Bandpass Limiters," *J. Appl. Phys.*, vol. 24, pp. 720–727, June 1953.

Table 1. Block IV receiver tracking loop modes[a]

$2B_{LO}$, Hz	Mode	K Open-loop gain, 1/sec	$\frac{\tau_2}{\tau_1}$	BW, kHz Predetection Filter
1	Narrow	9.6089×10^5	4.4340×10^{-5}	0.200
3	Narrow	9.6009×10^5	7.7057×10^{-5}	0.200
10	Narrow	9.6089×10^5	4.4343×10^{-4}	2.0
10	Wide	9.6089×10^5	4.4343×10^{-5}	2.0
30	Wide	9.6009×10^6	7.7057×10^{-5}	2.0
100	Wide	9.6089×10^6	4.4343×10^{-4}	20.0
300	Wide	9.6009×10^6	7.7065×10^{-4}	20.0

[a] From "Receiver/Exciter Block IV Equipment, Subsystem Detail Specifications," Doc. ES 505736, Rev. A (internal document), Jet Propulsion Laboratory, Pasadena, California, October 1974.

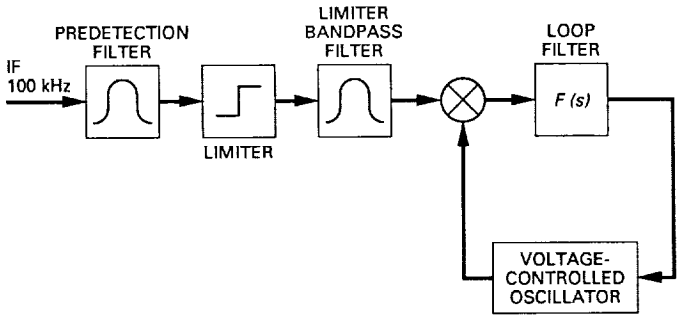


Fig. 1. Carrier tracking loop drop-lock model.

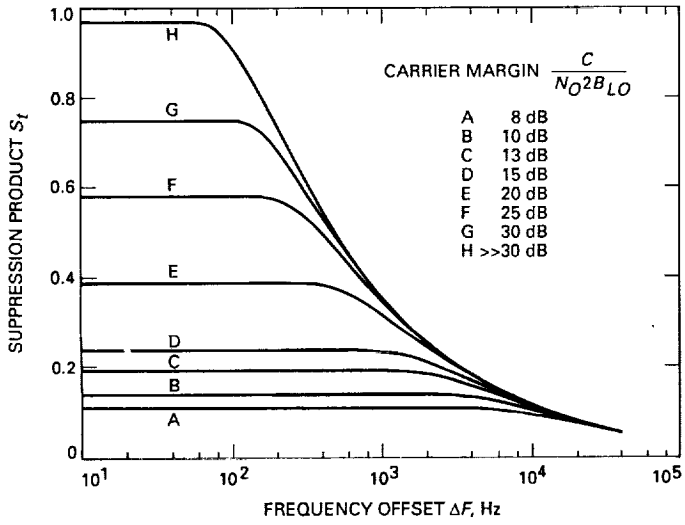


Fig. 2. Threshold limiter suppression product for various carrier margin values.

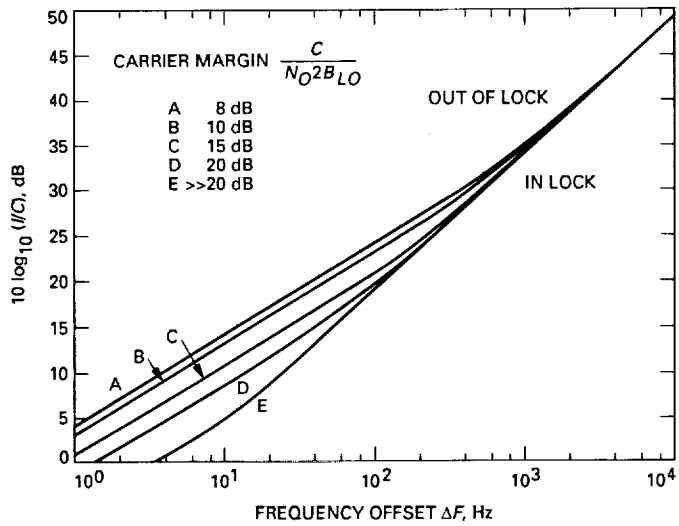


Fig. 3. Critical ISR versus frequency offset with input SNR values causing carrier drop-lock, tracking loop Mode 1, $2B_{LO} = 1$ Hz.

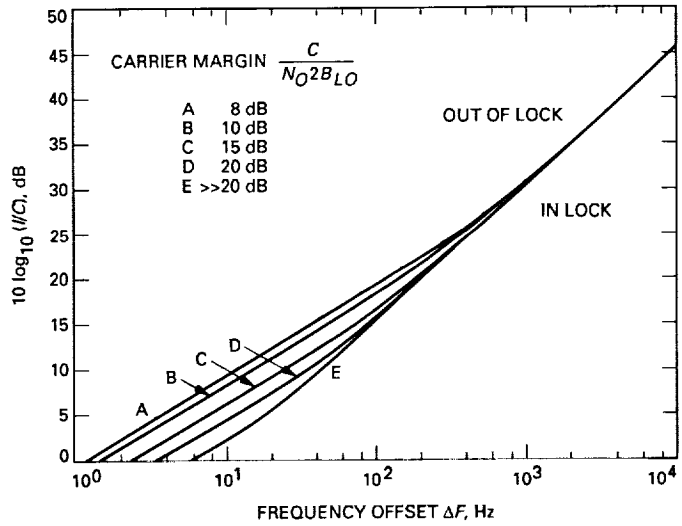


Fig. 4. Critical ISR versus frequency offset with input SNR values causing carrier drop-lock, tracking loop Mode 2, $2B_{LO} = 3$ Hz.

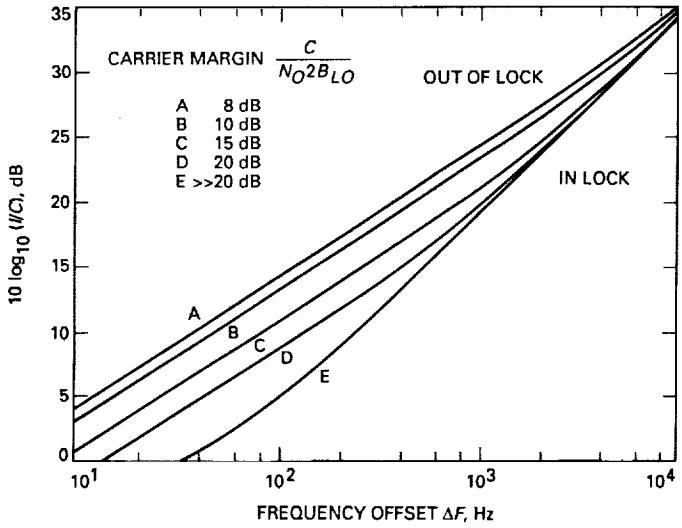


Fig. 5. Critical ISR versus frequency offset with input SNR values causing carrier drop-lock, tracking loop Mode 3, $2B_{LO} = 10$ Hz.

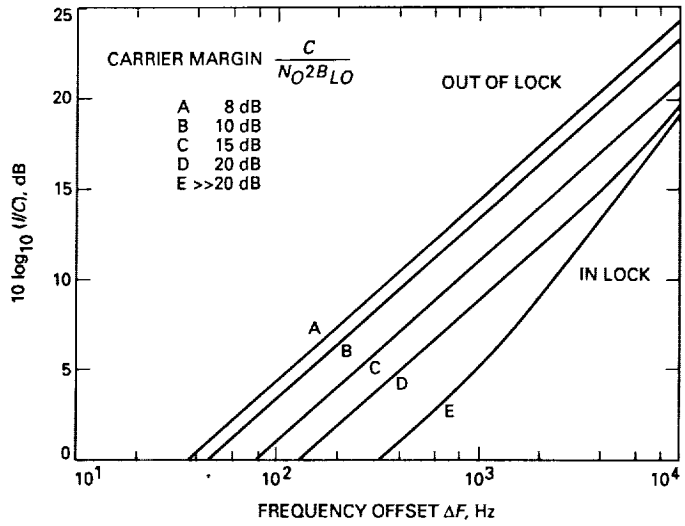


Fig. 7. Critical ISR versus frequency offset with input SNR values causing carrier drop-lock, tracking loop Mode 5, $2B_{LO} = 100$ Hz.

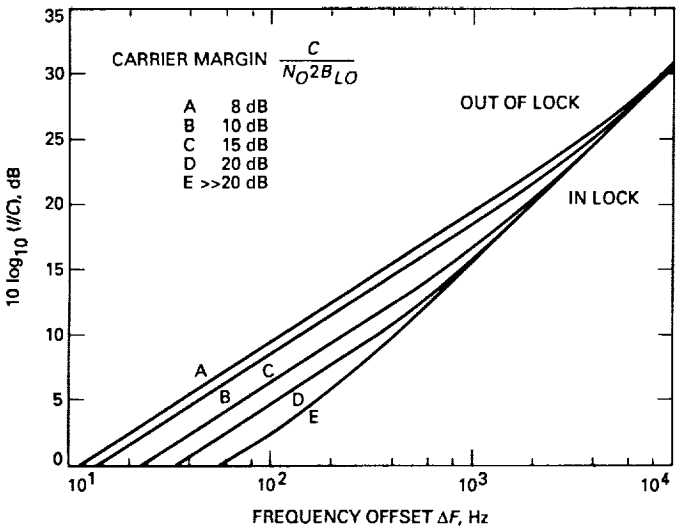


Fig. 6. Critical ISR versus frequency offset with input SNR values causing carrier drop-lock, tracking loop Mode 4, $2B_{LO} = 30$ Hz.

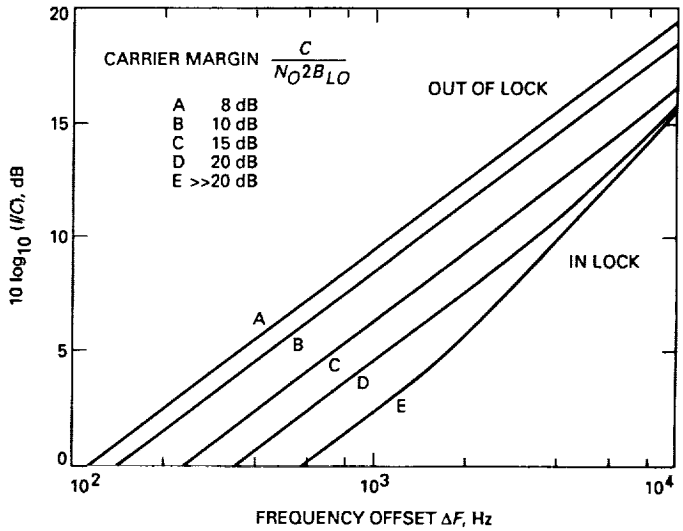


Fig. 8. Critical ISR versus frequency offset with input SNR values causing carrier drop-lock, tracking loop Mode 6, $2B_{LO} = 300$ Hz.

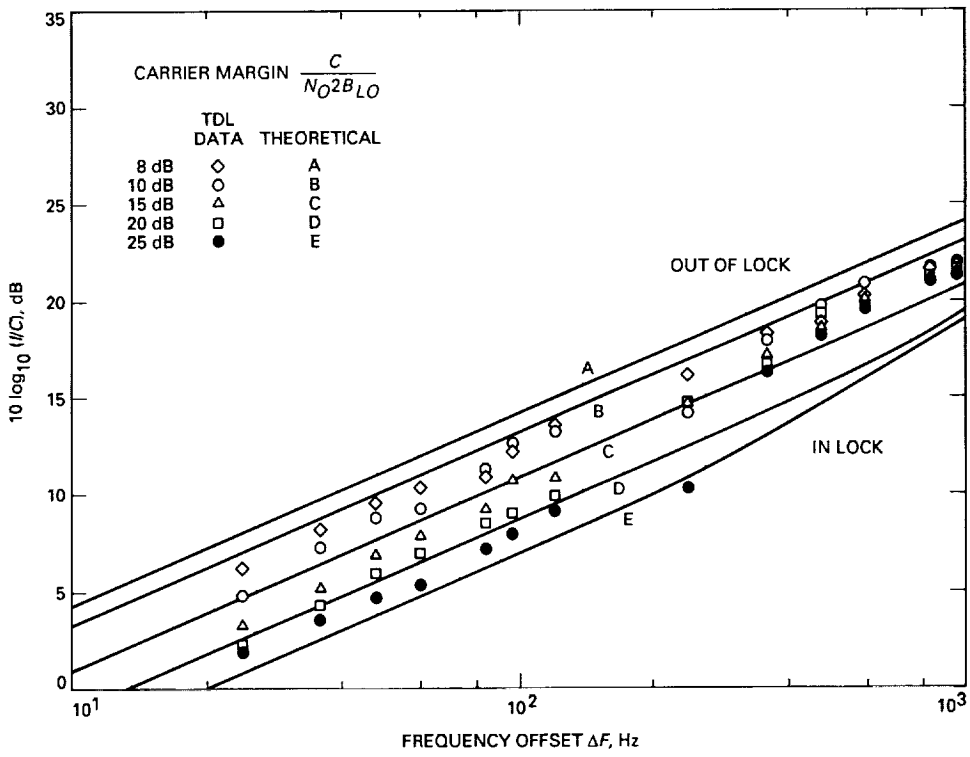


Fig. 9. Comparison of TDL data to theoretical levels of SNR at the limiter input, tracking loop Mode 3, $2B_{LO} = 10$ Hz.

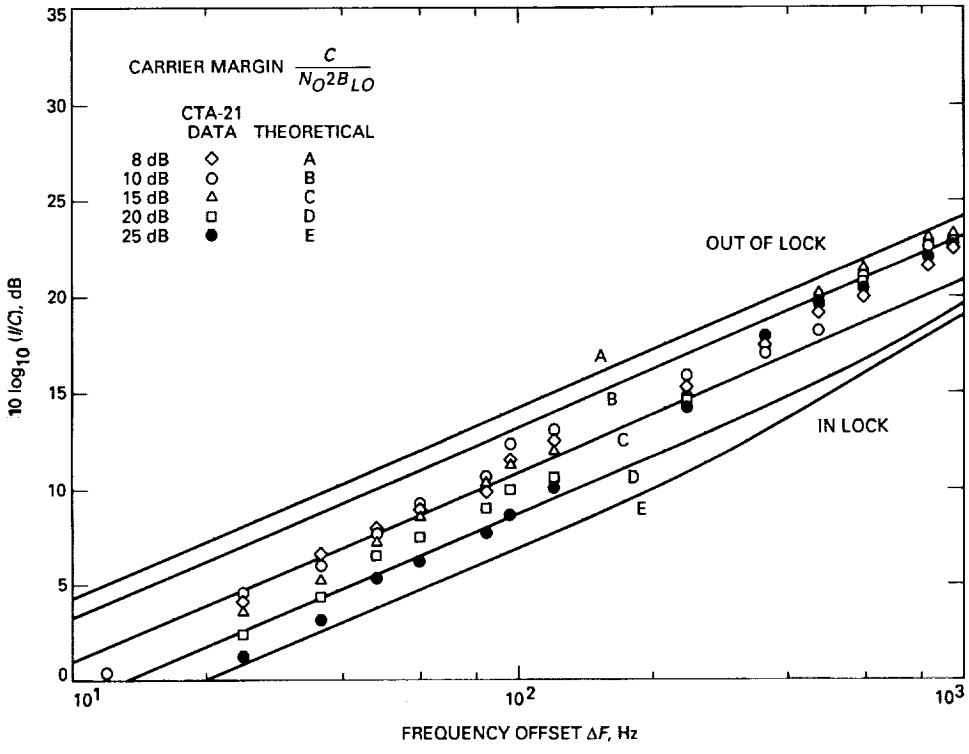


Fig. 10. Comparison of CTA-21 data to theoretical levels of SNR at the limiter input, tracking loop Mode 3, $2B_{LO} = 10$ Hz.

TDA Progress Report 42-99

November 15, 1989

Disturbance Torque Rejection Properties of the NASA/JPL 70-Meter Antenna Axis Servos

R. E. Hill
Ground Antenna and Facilities Engineering Section

Analytic methods for evaluating pointing errors caused by external disturbance torques are developed and applied to determine the effects of representative values of wind and friction torques. The expressions relating pointing errors to disturbance torques are shown to be strongly dependent upon the state estimator parameters, as well as upon the state feedback gain and the flow-versus-pressure characteristics of the hydraulic system. Under certain conditions, when control is derived from an uncorrected estimate of integral position error, the desired Type II servo properties are not realized and finite steady-state position errors result. Methods for reducing these errors to negligible proportions through the proper selection of control gain and estimator correction parameters are demonstrated. The steady-state error produced by a disturbance torque is found to be directly proportional to the hydraulic internal leakage. This property can be exploited to provide a convenient method of determining system leakage from field measurements of estimator error, axis rate, and hydraulic differential pressure.

I. Introduction

Recent studies of mechanisms contributing to limit-cycle behavior have led to a need for more accurate modeling of the disturbance torque response characteristics of the 70-m antenna axis servos. Of particular interest in the limit-cycle studies is the transient behavior of the various plant and estimator states that occur during friction-induced limit cycling. The traditional assumption that the estimator states accurately track those of the plant becomes inaccurate when the plant is subjected to external disturbance torques. This shortcoming necessitated the development of the present multivariable axis servo model where the plant and estimator states are distinct.

Precise pointing of the 70-m antenna is accomplished through the use of a two-axis, azimuth-elevation, bullgear-driven servo system. Control torques are produced by fixed-displacement, axial-piston hydraulic motors, which are coupled to the axis bullgears through spur-gear reducers. Four such motor/gear reducers are employed for each axis. Backlash is eliminated by separate countertorque motors which apply a constant torque bias to the output pinion of each control motor. The hydraulic connections to the countertorque motors are arranged as shown in Fig. 1 so as to preload all four gear reducers and apply zero net torque to the bullgear. Torque modulation is accomplished by four port-hydraulic servo valves. Servo control consists of a hardware rate loop with tachometer feedback and a computer-based position servo employing state-variable feedback.

The servo model is based on the nonlinear orifice-flow equation of the valve, along with the motor-flow equation

as described in [1]. A piecewise linear representation of the valve is obtained by partial differentiation of the valve equation (Eq. 5 of [1]) to yield the flow Q_v as a linear function of the valve input current I_v and load pressure P_L:

$$Q_v = K_P I_v - D_H P_L \tag{1}$$

with

$$K_P = K_v P_v^{1/2}$$

$$D_H = \frac{Q_v}{P_v}$$

and

$$P_v = P_S - P_L - P_R$$

where P_S and P_R are the regulated system supply pressure and return pressure, respectively. As shown later, the load pressure is sufficiently small relative to the supply pressure such that the range of the valve pressure remains within a 2 to 1 ratio. This justifies the use of a constant value approximation for the flow gain K_P for control dynamic analysis. In contrast, the equivalent damping D_H is linearly proportional to the flow, which varies in proportion to antenna rate. Therefore, D_H varies between a minimum value corresponding to leakage flow at the zero antenna rate, and a maximum corresponding to the maximum tracking rate of the antenna.

The hydromechanical system model incorporates the linearized valve of Eq. (1) along with a motor coupled to a rigid-body inertia load representation of the antenna structure, as shown in block diagram form in Fig. 2, where C_H, J_M, and V_M represent the hydraulic compressibility, motor inertia, and motor displacement, respectively.

The block diagram of the equivalent plant (servo-loop hardware) in Fig. 3 incorporates the model of Fig. 2, the tachometer feedback and control amplifier with associated compensation networks, and two additional integrations which produce the angular position and position integral states. The two inputs represent the electrical rate command input to the plant and an equivalent external disturbance torque; K_R represents a constant with value proportional to the rate loop gain, and P_2 and Z_2 are the pole and zero frequencies of the rate loop compensation network. Figure 3 also includes a simplified equivalent of the tachometer feedback network obtained by neglecting the network pole, which is at a relatively high frequency. This approximation introduces the acceleration feedback branch shown in Fig. 3 where the parameter Z_1 corresponds to the negative real frequency of the network zero.

To provide additional insight into the effects of rate and position loop parameters on system compliance, compliance equations are developed separately for the open-position loop case, for the hardware position proportional, integral, and derivative (PID) feedback, and for the closed-loop state variable controller. It is shown that the combination of plant-state and estimator-state feedback in the precision mode leads to compliance characteristics different from those of the computer mode, even when both modes have identical plant and estimator dynamics.

II. Compliance Equations for the Open Position Loop Case

The open position loop compliance can be derived from the angular position response to a unit-step torque input. By application of the Mason transmittance rule to the block diagram of Fig. 3, the compliance transfer function becomes:

$$\frac{\Theta_M}{T_X} = \frac{(1/J_M s^2)(1 + D_H/C_H s)(1 + P_2/s)}{(1 + P_2/s)(1 + D_H/C_H s + K_R/C_H J_M Z_1 s + (V_M^2 + K_R)/J_M C_H s^2) + (Z_2 - P_2)(K_R/C_H J_M s^2)(1/s + 1/Z_1)}$$

which leads to

$$\frac{\Theta_M}{T_X} = \frac{(1/s)(s + D_H/C_H)(s + P_2)}{J_M(s + P_2)(s^2 + (D_H/C_H)s + V_M^2/J_M C_H) + (K_R/C_H Z_1)(s + Z_1)(s + Z_2)} \tag{2}$$

Application of the final value theorem to Eq. (2) indicates a constant steady-state rate in response to a step function torque disturbance. The steady-state compliance is given by

$$\frac{\Theta_M}{T_X} = \frac{D_H}{s(V_M^2 + K_R Z_2/P_2)}$$

III. Compliance Equations for Hardware PID Feedback

The compliance of the closed-loop PID feedback system can be calculated by the same method used earlier for the open-loop case. The addition of position, integral, and rate feedback branches produces three additional denominator terms in the compliance expressions. Thus,

$$\begin{aligned}\frac{\Theta_M}{T_X} = {} & \frac{(1/J_M s^2)(1 + D_H/C_H s)(1 + P_2/s)}{(1 + P_2/s)(1 + D_H/C_H s + K_R/C_H J_M Z_1 s + (V_M^2 + K_R)/J_M C_H s^2)} \\ & + (Z_2 - P_2)(K_R/C_H J_M s^2)(1/s + 1/Z_1 + K_1/s^3 + K_2/s^2 + K_3/s) \\ & + (1 + P_2/s)(K_R/C_H J_M s)(K_1/s^3 + K_2/s^2 + K_3/s)\end{aligned}$$

which leads to

$$\begin{aligned}\frac{\Theta_M}{T_X} = {} & \frac{s(s + D_H/C_H)(s + P_2)}{J_M s^2(s + P_2)(s^2 + (D_H/C_H)s + V_M^2/J_M C_H)} \\ & + (K_R/C_H)(s + Z_2)(s^2 + s^3/Z_1 + K_1 + K_2 s + K_3 s^2)\end{aligned} \tag{3}$$

and the steady-state compliance becomes

$$\frac{\Theta_M}{T_X} = \frac{D_H s}{K_1 K_R (Z_2/P_2)}$$

The compliance transfer functions of Eqs. (2) and (3) include numerator zeros corresponding to the network pole P_2 and to the effective hydraulic damping D_H/C_H. The denominators are seen to contain the poles of the respective open/closed-loop system. From the dependence of D_H on hydraulic flow, the steady-state compliance properties are seen to be antenna-rate dependent. For typical closed-position loop parameter values, the damping D_H has relatively little effect on the closed-loop poles. Therefore, since the shape of the compliance transient is determined by the locations of the transfer function zeros relative to the poles, the closed-loop transient properties are also seen to vary as a function of D_H.

IV. Compliance Equations for the State Variable Controller

Earlier methods of disturbance torque effects analysis [2] were based on the assumption of negligible errors in the estimator states in relation to the corresponding plant states. This permitted a simplification of the system model by substituting plant-state feedback in place of the estimator-state feedback, thereby reducing the number of states required in the model. This assumption, widely used in evaluating command input transient responses, was found to produce finite errors in the determination of disturbance transient responses of the axis servos. It was subsequently replaced by a superior method employing full modeling of the estimator states as well as the plant states.

The model for disturbance torque response of the state-variable axis servo controller is based on a linearized multi-input state-variable representation of the system of Fig. 3, with the addition of the state-estimator and control-feedback gain. The plant state is represented by the generalized state equations, where $\dot{x}$ and Y are the state and output by

$$\dot{x} = \mathbf{A}x + \mathbf{B}U$$

$$Y = \mathbf{C}x + \mathbf{D}U$$

respectively.

The corresponding **A**, **B**, **C**, **D** matrices from Fig. 3 are:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1J_M & 0 \\ 0 & 0 & -V_M^2/C_H - K_R/C_H & -D_H/C_H - K_R/C_H J_M Z_1 & K_R(Z_2 - P_2)/C_H \\ 0 & 0 & -1 & -1/J_M Z_1 & -P_2 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1/J_M \\ K_R/C_H & -K_R/C_H J_M Z_1 \\ 1 & -1/J_M Z_1 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Representing the estimator state by $\hat{x}$, and estimator output by $\hat{Y}$, the estimator equations become:

$$\dot{\hat{x}} = \mathbf{A}\hat{x} + B_1 U_1 + L(Y - \hat{Y})$$

$$\hat{Y} = \mathbf{C}\hat{x} + D_1 U_1$$

where B_1, B_2, D_1, D_2, U_1, and U_2 represent the first and second columns of **B** and **D** and the first and second elements of **U**, respectively. This distinction is essential because both inputs couple directly into the plant, while only the rate command input U_1 couples directly into the estimator.

V. Computer Control Mode

Because the rate command U_1 is formed differently for the precision mode, the two control modes are addressed separately. Incorporating the control-feedback gain K into the expression for the rate command input U_1 for computer mode, and substituting into the plant and estimator equations leads to

$$U_1 = -K\hat{x}$$

$$\dot{x} = \mathbf{A}x - B_1 K\hat{x} + B_2 U_2 \tag{4}$$

$$\dot{\hat{x}} = LCx + (\mathbf{A} - B_1 K - LC - LD_1 K)\hat{x} + LD_2 U_2$$

C-3

With the estimator error,

$$\tilde{x} = x - \hat{x}$$

$$\dot{\tilde{x}} = (\mathbf{A} - LC)\tilde{x} + LD_1K\hat{x} + (B_2 - LD_2)U_2$$

Since for the present case both D_1 and D_2 are zero, the estimator and estimator-error equations simplify to

$$\dot{\hat{x}} = LCx + (\mathbf{A} - B_1K - LC)\hat{x} \tag{5}$$

and

$$\dot{\tilde{x}} = (\mathbf{A} - LC)\tilde{x} + B_2U_2 \tag{6}$$

The equations for the plant and estimator are illustrated in the state-space block diagram in Fig. 4. Combining Eqs. (4) and (5) to form a single state vector comprising the plant and estimator states leads to the matrix differential equation form

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & -B_1K \\ LC & \mathbf{A} - B_1K - LC \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} U_2 \tag{7}$$

which is compatible with existing linear system analysis and simulation software tools. The steady-state disturbance torque compliance properties of the combined plant and estimator system can be determined from the steady-state solution of Eq. (7). The general expressions for the steady-state values of the individual plant and estimator states can be determined by a symbolic expansion of the determinants resulting from application of Simpson's Rule to Eq. (7).

Since, in the steady state, the derivatives represented by the left-hand side of Eq. (7) equal zero, steady-state position estimate $\hat{x}_2$ becomes

$$\frac{\hat{x}_2}{U_2} = \frac{-\det \left[\begin{array}{c|c|c|c} & & & \\ \mathbf{A} & -B_1k_1 & B_2 & B_1\left[k_3k_4k_5\right] \\ & & & \\ \hline & & 0 & \\ & \mathbf{A}_1 & 0 & \\ LC & -B_1k_1 - LC_1 & 0 & \mathbf{A}_{3,5} - B_1\left[k_3\,k_4\,k_5\right] \\ & & 0 & \\ & & 0 & \end{array}\right]}{\det \begin{bmatrix} \mathbf{A} & -B_1K \\ LC & \mathbf{A} - B_1K - LC \end{bmatrix}} \tag{8}$$

where the numerator matrix is obtained from the 10 × 10 denominator matrix by replacing the column corresponding to $\hat{x}_2$ (column 7) with the right-hand-side vector, in accordance with Simpson's Rule. The subscripts applied to capital-letter symbols designate the respective columns of the associated matrix, and $\mathbf{A}_{3,5}$ denotes a matrix comprised of the third through fifth columns of $\mathbf{A}$.

A row by row examination of the numerator matrix reveals that the sixth row is comprised of the five elements of the first row of LC, a single element $(a_{11} - b_{11}k_1 - l_1c_1)$, and the last three elements of the first row of $[\mathbf{A} - B_1K - LC]$. It will be seen that, since the first element of L is zero in the current parameter set, and because of the sparseness of $\mathbf{A}$ and $\mathbf{B}$, the elements of the sixth row are all zeros and the numerator determinant vanishes. This indicates that, in the presence of a constant disturbance torque, the steady-state position estimate $\hat{x}_2$ is identically equal to zero. This in turn indicates that the final values of the position x_2 and the position estimation error $\tilde{x}_2$ are identical, and x_2 can

thus be evaluated using Eq. (6). This approach avoids the complexity of evaluating the determinant of the 10 × 10 matrices of Eq. (8).

Thus, in the steady state, using Eq. (6),

$$\dot{\tilde{x}} = (\mathbf{A} - LC)\tilde{x} + B_2 U_2 = 0$$

and

$$\tilde{x} = \frac{-\det\begin{bmatrix}\mathbf{A}_1 & B_2 & \mathbf{A}_3 & \mathbf{A}_4 & \mathbf{A}_5\end{bmatrix}}{\det\begin{bmatrix}\mathbf{A} & -LC\end{bmatrix}} U_2$$

Note that L and C are absent from the numerator because all elements of C, except for the second, are zero. Note also that the first row of $\mathbf{A}$ and the first element of C are zero, causing both determinants to vanish. However, since the plant integral state is uncoupled from both the plant and the estimator, the corresponding first rows and columns can be deleted from both the numerator and denominator with no loss in generality. It should be noted that if the first element of C is assigned a nonzero value, the denominator determinant remains finite as long as L is nonzero. The implications of this integral position feedback will be discussed later.

Substituting the values of $\mathbf{A}$, B_2, and C into the expression for $\tilde{x}_2$,

$$\frac{\tilde{x}_2}{U_2} = \frac{\det\begin{bmatrix} 0 & 1 & 0 & 0 \\ -1/J_M & 0 & 1/J_M & 0 \\ K_R/C_H J_M Z_1 & -V_M^2/C_H - K_R/C_H & -D_H/C_H - K_R/C_H J_M Z_1 & K_R(Z_2 - P_2)/C_H \\ 1/J_M Z_1 & -1 & -1/J_M Z_1 & -P_2 \end{bmatrix}}{\det\begin{bmatrix} -l_2 & 10 & 0 & \\ -l_3 & 0 & 1/J_M & 0 \\ -l_4 & -V_M^2/C_H - K_R/C_H & -D_H/C_H - K_R/C_H J_M Z_1 & K_R(Z_2 - P_2)/C_H \\ -l_5 & -1 & -1/J_M Z_1 & -P_2 \end{bmatrix}}$$

Expanding the determinants leads to the nonzero steady-state position estimate error $\tilde{x}_2$, and since the steady-state plant position x_2 equals $\tilde{x}_2$,

$$\frac{x_2}{U_2} = \frac{D_H}{l_2(V_M^2 + K_R Z_2/P_2) + l_3(D_H J_M + K_R Z_2/Z_1 P_2) + l_4 C_H + l_5(Z_2/P_2 - 1)}$$

which, using current 70-m antenna servo parameter values, can be approximated by

$$\frac{x_2}{U_2} \cong \frac{D_H P_2/Z_2}{K_R l_2}$$

This result invalidates the original premise that control feedback of estimated integral error is equivalent to feedback of plant integral error in imparting Type II servo performance. That premise is valid only for inputs, such as the position command, which are coupled equally to the estimator and the plant. From Eqs. (3) and (4) and the block diagram of Fig. 4, it is seen that the plant is influenced by the control input U_1 and a disturbance input U_2, while the estimator is influenced by the same control input and by the estimator feedback error $LC\tilde{x}$. In equilibrium, the plant disturbance U_2 is compensated by the control input U_1; thus, the identical U_1 input to the estimator must be counteracted by the error $LC\tilde{x}$ in order to obtain estimator equilibrium. This implies that a finite estimator error $\tilde{x}$ will always result from a plant disturbance input, while the individual components of $\tilde{x}$ will be determined by the product of L and C. This line of reasoning explains the absence of the control gain K from the steady-state compliance expression.

In the present case, where there is no feedback of x_1, the integral estimate error $\tilde{x}_1$ is allowed to grow unbounded. The addition of a small amount of plant integral feedback to the estimator by assigning a small value to c_1 would bound the error $\tilde{x}_1$, thereby forcing $\tilde{x}_2$, x_2, and $\hat{x}_2$ to zero. The required nonzero $\tilde{x}$ would then result from nonzero values of the components other than $\tilde{x}_2$ and $\tilde{x}_3$. This additional feedback imparts the desired disturbance accommodating control [3] properties to the system.

VI. Precision-Control Mode

In the precision (autocollimator feedback) mode, the control input U_2 is derived from the autocollimator, an electro-optical device that senses plant position error directly. This error signal is filtered to remove noise and then integrated. The resulting plant integral and position errors are then combined with additional damping terms provided by the position estimate to form the plant input according to

$$U_1 = -k_1 x_1 - k_2 x_2 - k_3 \hat{x}_3 - k_3 \hat{x}_4 - k_5 \hat{x}5$$

The resulting system equations therefore take a slightly different form from that of Eq. (7) due to the mixed plant and estimator feedback. Thus, for precision mode,

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{A} - B_1 K_P & -B_1 K_E \\ LC - B_1 K_P & \mathbf{A} - B_1 K_E - LC \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + \begin{bmatrix} B_2 - B_1 K_P D_2 \\ 0 \end{bmatrix} U_2 \qquad (9)$$

where K_P and K_E are the gain vectors associated with the plant and estimator states respectively. For the precision mode,

$$\begin{bmatrix} K_P \\ K_E \end{bmatrix} = \begin{bmatrix} k_1 & k_2 & 0 & 0 & 0 \\ 0 & 0 & k_3 & k_4 & k_5 \end{bmatrix}$$

The $B_1 K_P D_2$ term is included in Eq. (9) to maintain generality, even though the $K_P D_2$ product is always zero in the absence of plant acceleration feedback.

From Eq. (9), it is seen that Eq. (6) for the estimator error is also applicable to the precision mode case, and the steady-state estimator error $\tilde{x}$ is therefore identical in both the computer mode and precision mode. However, the altered form of the control gains K_P and K_E in the precision mode change the coupling of the integral position x_1 and its estimate $\hat{x}_1$. Accordingly, when solving Eq. (9) for final values, the row and column corresponding to the integral estimate $\hat{x}_1$ are deleted, and those corresponding to the integral x_1 are retained. As a result, the final value solution of Eq. (9) yields the expected zero value for the position x_2 and a nonzero position estimate $\hat{x}_2$. This result implies a nonzero estimator error $\tilde{x}$ consistent with Eq. (6).

These observations demonstrate an interesting duality between the two control modes, whereby the substitution of plant states for their corresponding estimator states in the control gain law results in the interchange of the final values of plant and estimated position variables. This tends to confirm the validity of the general expressions of Eqs. (7) and (9), and indicates that in the precision-control mode, a zero steady-state error always results from a disturbance torque input.

VII. Compliance Transient Properties

The shape of the transient response to disturbance torques can be inferred from the relative locations of the poles and zeros of the compliance transfer functions for the various system configurations. For the open-loop and hardware PID feedback systems, the poles and zeros are defined by Eqs. (2) and (3), respectively, where it is seen that the poles are the same as those associated with the respective command input-output transfer functions.

For the state variable controller configurations, the poles are the respective eigenvalues of the square system matrices in Eqs. (7) and (9). The corresponding zeros are the complex frequencies of zero response of $\dot{x}$, $\dot{\hat{x}}$, and Y for any disturbance input U_2 and initial condition. Thus, for the computer-mode configuration of Eq. (7), the response zeros are the roots of

$$\begin{bmatrix} (sI - \mathbf{A}) & -B_1 K & -B_2 \\ LC & (sI - \mathbf{A} + B_1 K + LC) & 0 \\ C & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

To avoid the longhand expansion of the above determinant, the zeros were determined by a numerical evaluation of the transfer function zeros of the system of Eq. (7) with coefficient values representative of the 70-m azimuth servo.

The results coincide precisely with the two nonzero real roots of the determinant

$$\begin{bmatrix} [sI - \mathbf{A}] & B_2 \\ C & 0 \end{bmatrix}$$

at $-D_H/C_H$ and $-P_2$ (the root at zero is absent), and with the five eigenvalues of $[\mathbf{A} - B_1K - LC]$. A numerical computation of the eigenvalues of the square matrix of Eq. (7) shows that the accompanying poles are the five eigenvalues of $[\mathbf{A} - B_1K]$ and the five eigenvalues of $[\mathbf{A} - LC]$.

For the precision-mode configuration of Eq. (9), the response zeros are the zeros of

$$\begin{bmatrix} [sI - \mathbf{A} + B_1K_P] & -B_1K_E & -B_2 \\ LC - B_1K_P & [\mathbf{sI} - \mathbf{A} + B_1K_E + LC] & 0 \\ C & 0 & 0 \end{bmatrix} \times \begin{bmatrix} x \\ \hat{x} \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Numerical evaluation of the transfer function zeros indicates two real zeros at $-D_H/C_H$ and $-P_2$, two zeros at the origin (one of which is canceled by a pole), and four additional real and complex zeros which are near, but noncoincident with, four of the poles of the square matrix of Eq. (9). This result indicates that unlike the computer-mode case of Eq. (7), the precision mode exhibits zero steady-state disturbance torque compliance. In addition, the low-frequency zero from the eigenvalues of $[\mathbf{A} - B_1K - LC]$ arising from Eq. (7) is absent in this case, resulting in a faster disturbance recovery transient in the precision mode. The system poles computed from Eq. (9) are identical with those from Eq. (7) for the computer mode.

From the foregoing and from Eqs. (2) and (3), it is seen that the real zeros at $-D_H/C_H$ and $-P_2$ appear in the compliance expressions for the open-loop and PID feedback systems, as well as for the computer and precision-mode state variable controllers. The differing steady-state compliance properties of these four configurations arise from the presence of poles or zeros at the origin. This commonality implies that all four configurations exhibit identical disturbance transient characteristics for a short time interval following the transient. This in turn implies that the initial transient characteristics of the state variable controllers are governed solely by the plant parameters, and the departure from this initial characteristic is governed by the controller.

VIII. Numerical Results

Numerical evaluations of the differential equations incorporated the physical parameter values of Table 1, the control coefficients of Table 2, and the disturbance torque parameters of Table 3. Physical parameter values of Table 1 were derived from component specifications according to [1] with the rate loop gain and network parameters based on [4]. The K_R value of 5.107 corresponds to a rate loop dc gain product of 40. The values shown for damping D_H are calculated using Eq. (1) and correspond to the axis rate range of 1 to 40 times the sidereal rate with added leakage Q_{HL}. Aerodynamic parameters used in determining wind torques were extrapolated by McGinness[1] from earlier wind-tunnel tests performed on scale models of the 64-m antenna.

The position feedback control gain K and estimator gain L shown in Table 2 were calculated according to the optimal control criteria of Alvarez and Nickerson [5]. The baseline values L_1 and K_4 in the table were calculated to duplicate the dominant closed-loop poles of [5] when applied to the model of Fig. 3. Alternate parameter sets were calculated in a similar manner with weighting adjusted to shift the dominant poles inward or outward from the origin. Those estimator gains that include integral terms were calculated using an alternate output vector $H = [1 \quad 1 \quad 0 \quad 0 \quad 0]$, which included equal weighting of integral and position.

The wind torque moments in Table 3 were calculated according to McGinness[2] using the torque relationship

$$\text{torque} = \frac{C_T q \pi D_A^3}{4} \tag{10a}$$

where C_T is the aerodynamic lift coefficient, the dynamic wind pressure is

$$q = \frac{\rho v^2}{2} \tag{10b}$$

with air density, $\rho = 0.00238$ lb/ft^3, and wind velocity, v.

[1] H. McGinness, "Effects of Wind Loading on 64- And 72-Meter Diameter Antenna" (JPL internal report), Reorder No. 84-2, May 1984.

[2] McGinness, ibid.

From Eqs. (10a) and (10b), the disturbance torque is seen to be a quadratic function of the wind velocity, so the disturbance imparted by a step change in velocity depends on the initial and final velocities rather than the step amplitude. Thus, if v_m is the mean of the initial and final velocities and v_d is their difference, the torque disturbance is proportional to the product $v_m v_d$. The worst-case disturbance then results when a large step-velocity change is superposed on a large mean velocity such that their product is a maximum. Assuming 30 mph as the maximum average wind and 12 mph as the maximum gust, the worst-case disturbance corresponds to a step change between 24 and 36 mph.

Breakaway friction torque levels were extracted from acceptance-test data recorded when the overseas antennas were built in 1972. The range of values shown is consistent with more recent informal reports of observations at the Goldstone site.

The differential hydraulic pressures corresponding to the disturbance torques are included in Table 3. Because of its linear relationship to torque and its ease of direct measurement, the differential pressure has become a familiar unit of torque measurement to those working with the antenna.

The disturbance torque transient responses of the open-loop, computer-mode, and precision-mode configurations obtained from the time solutions of Eqs. (7) and (9) with $K = K_4$ and $L = L_1$ are shown in Fig. 5. The disturbance input corresponds to a 12-mph wind step combined with a 30-mph mean wind. The open-loop case was obtained from Eq. (7) with zero value of control gain K. The comparatively rapid recovery of the precision mode and the finite steady-state error of the computer mode are clearly visible in the time responses. The slower recovery of the computer mode is attributed to a low-frequency transfer function zero arising from Eq. (7), which is absent from the precision-mode case.

Figure 6 shows the computer-mode disturbance torque transient response for various antenna rates. The responses were generated from Eq. (7), where the D_H/C_H term in the **A** matrix was adjusted according to Eq. (1) for each antenna rate. The effect of the changing D_H/C_H zero in the transfer function is evident from the change of the transient shape and of the final value. Figure 6 can also be used to judge the effects of increased hydraulic leakage by converting a known volumetric leakage to an equivalent antenna rate through the hydraulic displacement and gear ratio. The characteristic corresponding to the sum of the leakage equivalent rate and actual rate then represents the system behavior.

Figure 7 shows the computer-mode transient response for the various values of control gain K listed in Table 2. As indicated earlier from the general properties of Eqs. (6) and (7), the gain K is seen to influence the speed of recovery, but has no effect on the final value.

Figure 8 shows the computer-mode transient response for the various values of estimator gain L listed in Table 2. The effect of those L vectors having finite integral error correction terms l_1 is seen in the faster recovery time, as well as reduction of the final value. The final value for those cases is still finite due to the absence of position integral feedback.

Figure 9 shows the precision-mode transient response for the various values of control gain K listed in Table 2. Increasing values of K are seen to decrease the peak error and speed the recovery transient.

Figure 10 shows the precision-mode transient response for the various values of estimator gain L listed in Table 2. As expected, the estimator gain L has little effect on the compliance transient when the response of the estimator is faster than that of the control gain.

IX. Summary and Conclusions

The transient disturbance-torque rejection properties of the 70-m axis servos are shown to be governed by physical hardware parameters, as well as by the properties of the software-control algorithm. In particular, the effective hydraulic damping D_H/C_H, which is strongly affected by hydraulic leakage as well as by axis rate, determines the peak transient error, and in the case of the computer mode it also produces a finite steady-state error. The peak and steady-state disturbance-torque errors are essentially unaffected by the control gain K, but can be reduced by increased estimator gain L at the expense of noise rejection.

The precision-control mode, by virtue of the direct feedback of hardware (as opposed to estimated) position and integral errors, is shown to have disturbance torque rejection properties superior to those of the computer mode. This deficiency of the computer mode results from the omission of the position integral error from the estimator equations, and can be corrected by modifying the equations to include the integral. Significant improvement can also be obtained by modifying the estimator gain parameter L to include a coefficient corresponding to the integral position estimate.

References

[1] R. E. Hill, "A New State Space Model for the NASA/JPL 70-Meter Antenna Servo Controls," *TDA Progress Report 42-91*, vol. July-September 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 247–264, November 15, 1987.

[2] R. E. Hill, "A New Method for Analysis of Limit Cycle Behavior of the NASA/ JPL 70-Meter Antenna Axis Servos," *TDA Progress Report 42-97*, vol. January–March 1989, Jet Propulsion Laboratory, Pasadena, California, pp. 98-111, May 15, 1989.

[3] C. D. Johnson, "Disturbance Accommodating Control: A History of Its Development," *Recent Advances in Engineering Science: Proc. 15th Annual Meeting Soc. of Eng. Science*, R. L. Sierakowski, editor, pp. 331–336, 1978.

[4] R. E. Hill, "Dynamic Models for Simulation of the 70-Meter Antenna Axis Servos," *TDA Progress Report 42-95*, vol. July–September 1988, Jet Propulsion Laboratory, Pasadena, California, pp. 32–50, November 15, 1988.

[5] L. S. Alvarez and J. Nickerson, "Application of Optimal Control Theory to the Design of the NASA/JPL 70-Meter Antenna Axis Servos," *TDA Progress Report 42-97*, vol. January–March 1989, Jet Propulsion Laboratory, Pasadena, California, pp. 112–126, May 15, 1989.

Table 1. The 70-m antenna axis servo parameter values

Symbol	Parameter	Units	Value
J_{MA}	Inertia moment, azimuth	in.-lb-sec	12.0
J_{ME}	Inertia moment, elevation	in.-lb-sec	8.0
V_H	Hydraulic displacement	$in.^3$/radian	1.53
C_H	Hydraulic compressibility	$in.^3$/psi	0.00314
D_H	Hydraulic damping	$in.^3$/psi-sec	0.00188 to 0.0754
Q_{HL}	Hydraulic leakage	$in.^3$/sec	1.67
N	Reduction gear ratio	dimensionless	28730
P_1	Tach network negative pole frequency	sec^{-1}	80
P_2	Lag network negative pole frequency	sec^{-1}	0.24
Z_1	Tach network negative zero frequency	sec^{-1}	5.0
Z_2	Lag network negative zero frequency	sec^{-1}	4.4
K_R	Rate loop gain constant	$in.^6$	5.11
K_V	Valve flow constant	$in.^3$/mA-sec/$psi^{0.5}$	0.731
P_V	Valve pressure drop	psi	2750
C_{TA}	Aerodynamic lift coefficient, azimuth	dimensionless	0.1217
C_{TE}	Aerodynamic lift coefficient, elevation	dimensionless	0.1231
D_A	Reflector aerodynamic diameter	m	70

Table 2. System, control gain, and estimator parameters

(a) System matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 1.0000 & 0 & 0 & 0 \\ 0 & 0 & 1.0000 & 0 & 0 \\ 0 & 0 & 0 & 8.3333 & 0 \\ 0 & 0 & -23.7194 & -27.7072 & 13.2484 \\ 0 & 0 & -5.1070 & -8.5117 & -0.2400 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0.0833 \\ 16.2643 & -0.2711 \\ 5.1070 & -0.0851 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1.0000 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 0 \end{bmatrix}$$

(b) Control gains K

Control gain K	Numerical value	Control poles from eigenvalue $(\mathbf{A} - BK)$
K_1	0.5200	$-12.2372 + 12.5297i$
	1.3417	$-12.2372 - 12.5297i$
	0.5712	-2.5455
	0	$-0.4636 + 0.4268i$
	0	$-0.4636 - 0.4268i$
K_2	0.3162	$-12.8664 + 9.5595i$
	0.9097	$-12.8664 - 9.5595i$
	0.1685	-2.4141
	0.0162	$-0.3937 + 0.3861i$
	0.1416	$-0.3937 - 0.3861i$
K_3	0.4472	$-12.9520 + 9.4886i$
	1.1113	$-12.9520 - 9.4886i$
	0.2087	-2.4404
	0.0279	$-0.4668 + 0.4539i$
	0.1716	$-0.4668 - 0.4539i$
K_4	0.6325	$-13.1214 + 9.3446i$
	1.3684	$-13.1214 - 9.3446i$
	0.2642	-2.4911
	0.0500	$-0.5509 + 0.5291i$
	0.2106	$-0.5509 - 0.5291i$
K_5	0.8944	$-13.4526 + 9.0474i$
	1.7032	$-13.4526 - 9.0474i$
	0.3450	$-0.6445 + 0.6078i$
	0.0916	$-0.6445 - 0.6078i$
	0.2629	-2.5861
K_6	1.2649	$-14.0877 + 8.4129i$
	2.1515	$-14.0877 - 8.4129i$
	0.4690	$-0.7426 + 0.6822i$
	0.1699	$-0.7426 - 0.6822i$
	0.3339	-2.7556

Table 2. System, control gain, and estimator parameters (contd)

(c) Estimator gains

Estimator gain L	Numerical value	Estimator poles from eigenvalue $(\mathbf{A} - BK_4 - LC)$	Compliance zeros from eigenvalue $(\mathbf{A} - BK_4 - LC)$
L_1	0	$-12.7800 + 9.6298i$	$-13.1060 + 9.3382i$
	0.4546	$-12.7800 - 9.6298i$	$-13.1060 - 9.3382i$
	0.0033	-2.3859	-2.1842
	-0.0020	-0.4559	-1.4267
	0.0026	0	-0.4674
L_2	0.8638	$-12.4988 + 9.7652i$	$-12.6999 + 9.4375i$
	3.1261	$-12.4988 - 9.7652i$	$-12.6999 - 9.4375i$
	4.8336	-4.8845	-5.0267
	-7.6308	-1.1912	-2.5190
	-9.1392	0	-0.0162
L_3	0.8075	$-12.7656 + 9.6345i$	$-13.0283 + 9.3205i$
	1.4625	$-12.7656 - 9.6345i$	$-13.0283 - 9.3205i$
	1.1140	$-1.9392 + 0.7044i$	$-2.6041 + 1.3095i$
	-0.6472	$-1.9392 - 0.7044i$	$-2.6041 - 1.3095i$
	0.5716	0	-0.0333
L_4	0.6204	$-12.7686 + 9.6345i$	$-13.0829 + 9.3379i$
	0.5545	$-12.7686 - 9.6345i$	$-13.0829 - 9.3379i$
	0.1357	-2.5083	$-2.0488 + 0.4207i$
	-0.3804	-0.4562	$-2.0488 - 0.4207i$
	-0.5582	0	-0.1267

Table 3. Typical disturbance torque levels

	Azimuth	Elevation
Wind moments for 30-mph wind	2.67×10^6 ft-lbs 729 psi	2.70×10^6 ft-lbs 738 psi
Breakaway friction torque	$1.03 - 1.28 \times 10^6$ 280 – 350 psi	$0.84 - 1.46 \times 10^6$ ft-lbs 230 – 400 psi

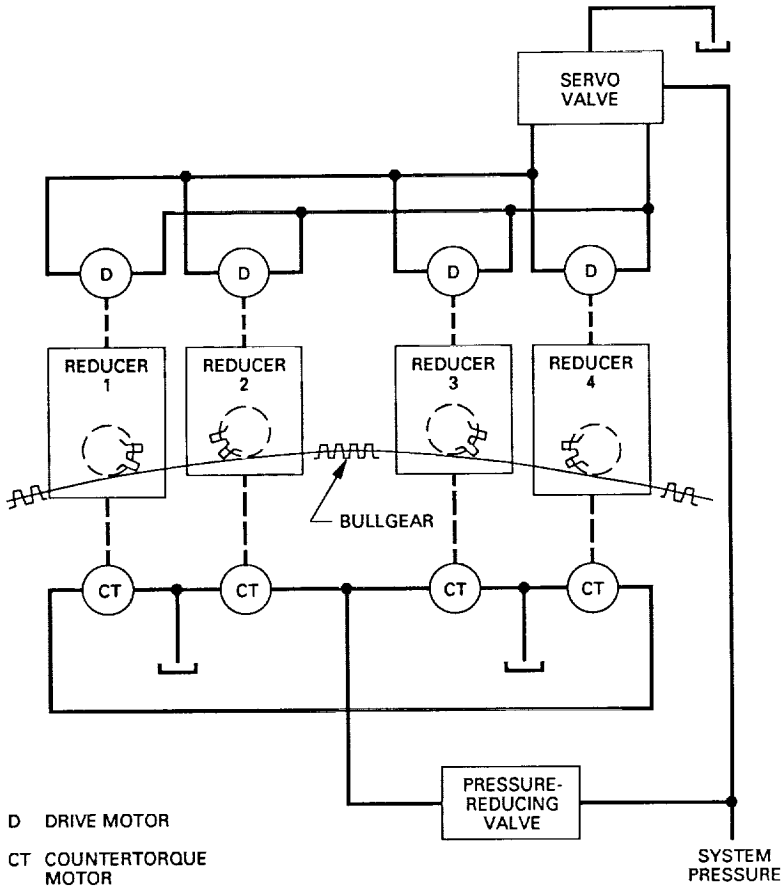


Fig. 1. The 70-m antenna axis servo countertorque system.

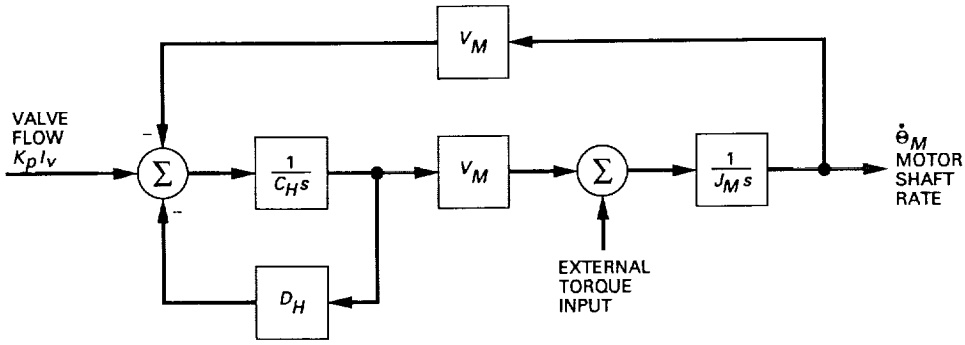


Fig. 2. Servo hydromechanical system linearized model.

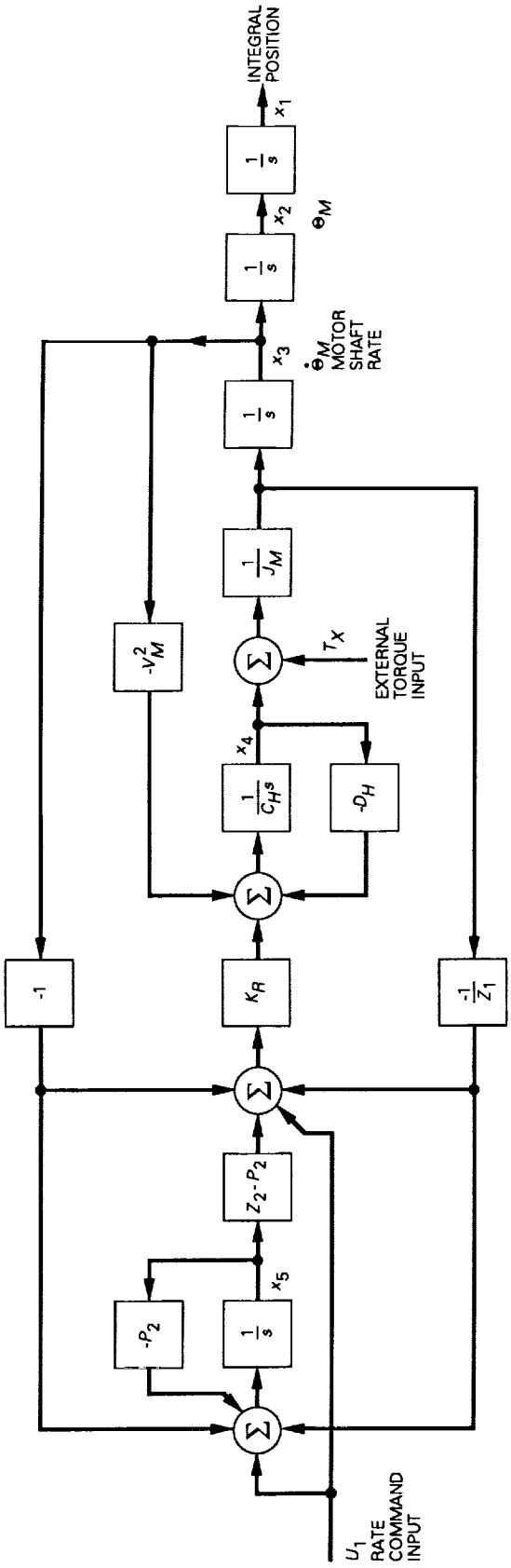


Fig. 3. Servo condensed plant block diagram.

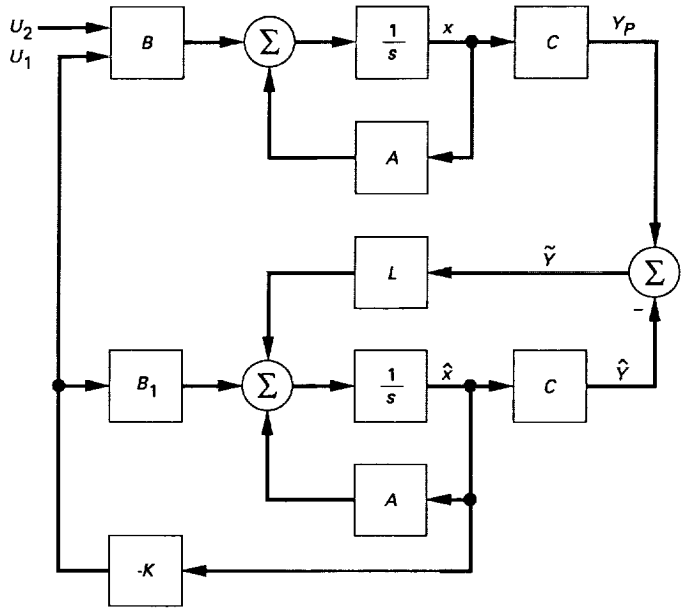


Fig. 4. Servo position controller block diagram.

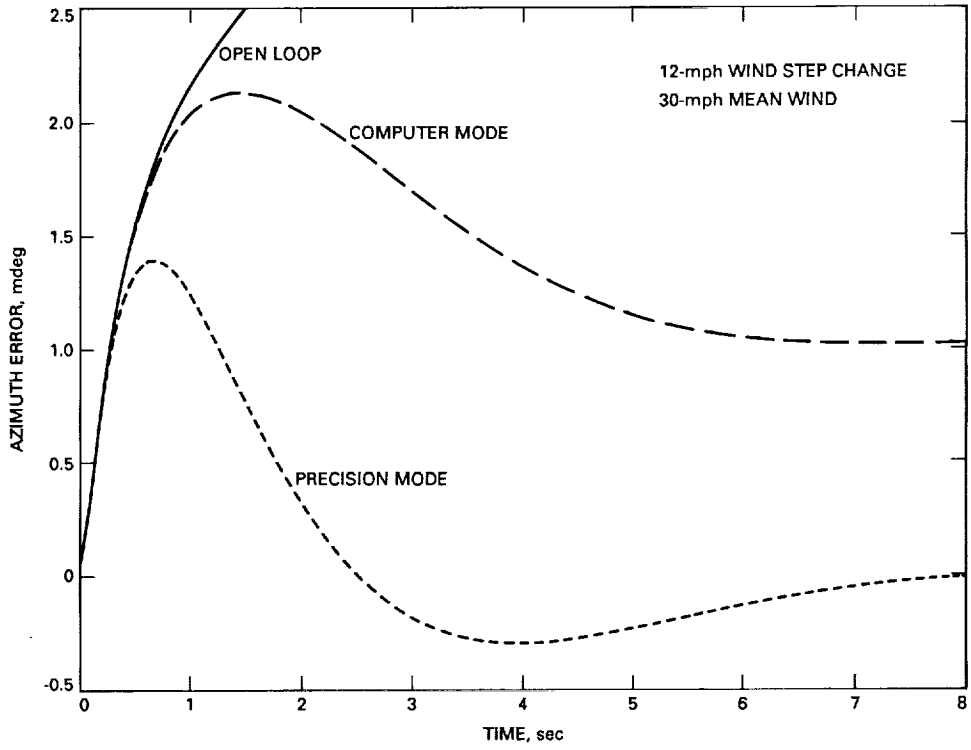


Fig. 5. Disturbance torque transient response of open-loop, computer-mode, and precision-mode configurations.

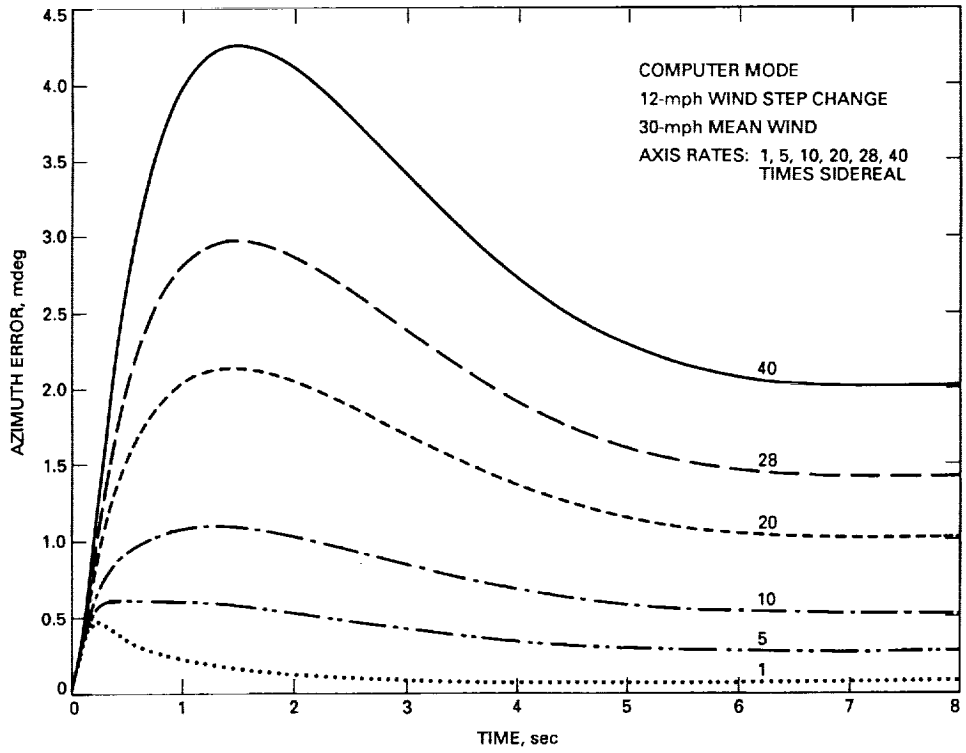


Fig. 6. Disturbance torque transient response for various axis rates.

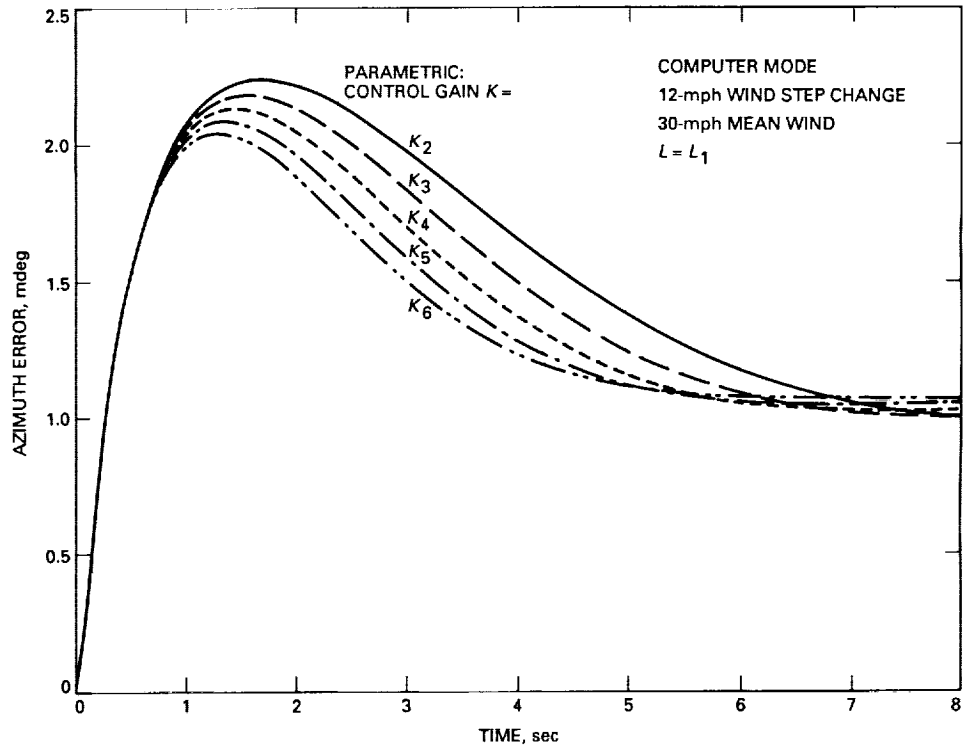


Fig. 7. Computer-mode disturbance torque transient response for various *K*.

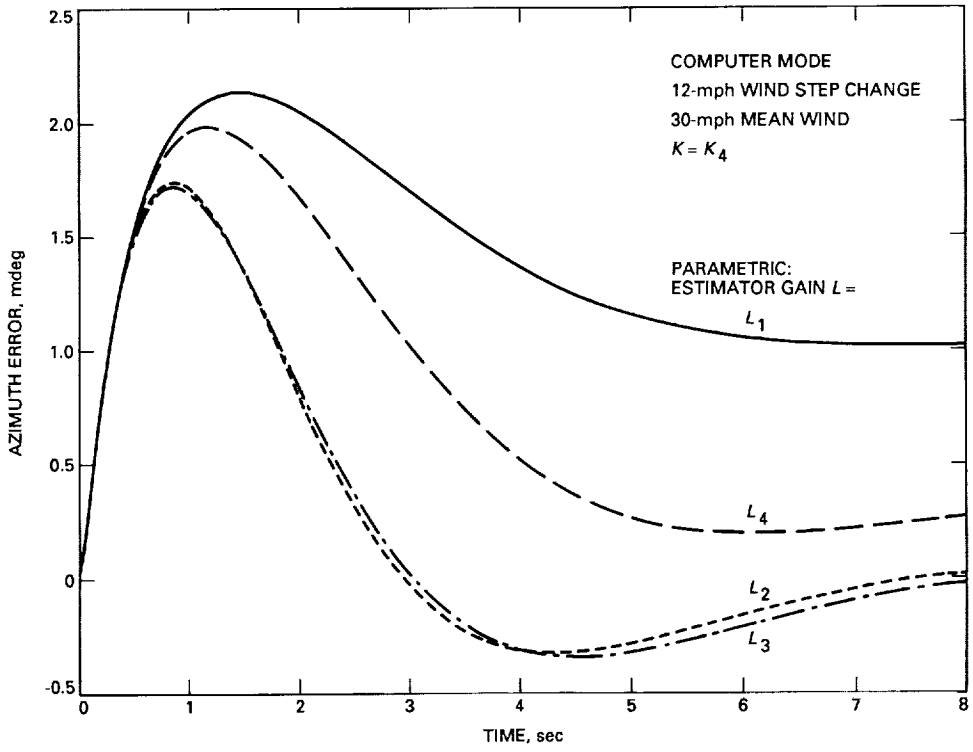


Fig. 8. Computer-mode disturbance torque transient response for various L.

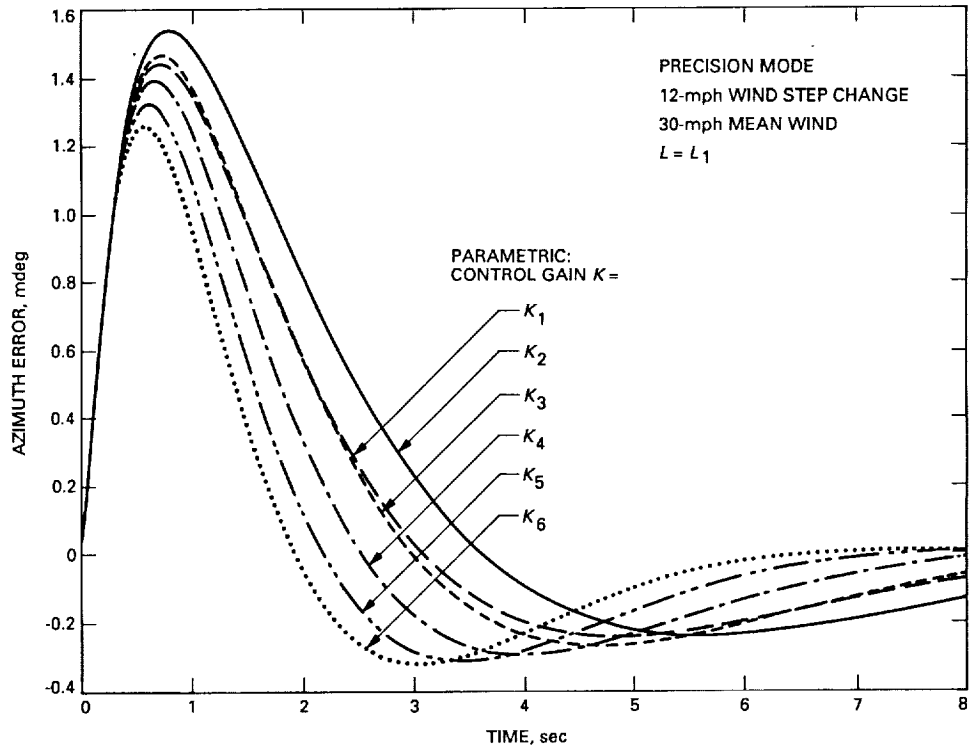


Fig. 9. Precision-mode disturbance torque transient response for various K.

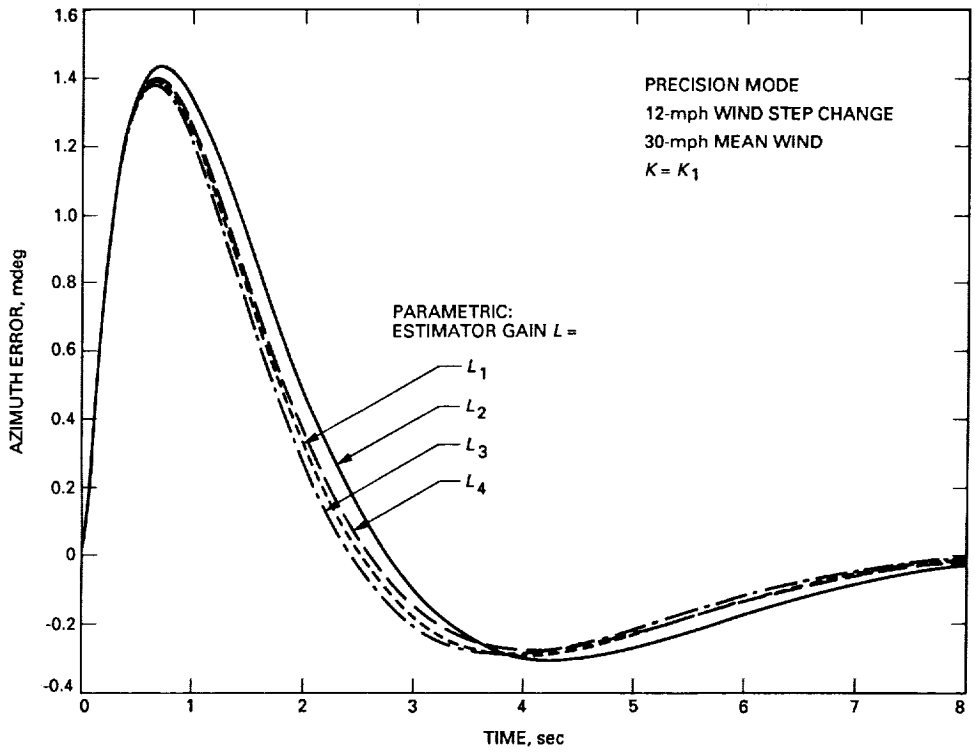


Fig. 10. Precision-mode disturbance torque transient response for various *L*.

Parkes Radio Science System Design and Testing for Voyager Neptune Encounter

T. A. Rebold and J. F. Weese
Telecommunications Systems Section

The Radio Science System installed at Parkes, Australia for the Voyager Neptune encounter was specified to meet the same stringent requirements that were imposed upon the DSN Radio Science System. This article describes the system design and test methodology employed to meet these requirements at Parkes, and presents data showing the measured performance of the system. The results indicate that the system operates with a comfortable margin on the requirements. There was a minor problem with frequency-dependent spurious signals which could not be fixed before the encounter. Test results characterizing these spurious signals are included.

I. Introduction

In conjunction with efforts to upgrade the DSN in support of the Voyager Radio Science experiments at Neptune, a Radio Science receiving system was temporarily installed at the Commonwealth Scientific Industrial Research Organization (CSIRO) Radio Telescope located in Parkes, New South Wales. The primary motivation for the Parkes installation was to allow improved signal reception at X-band through the application of signal arraying techniques between the Parkes 64-m antenna and the Canberra Deep Space Communications Complex (CDSCC) 70-m antenna located some 300 km away.

This article presents an overview of the Radio Science System at Parkes. Section II covers the performance requirements imposed by the Voyager Neptune encounter, and is followed in Section III by a description of the design chosen to meet these requirements. The test methodology for verifying system performance is discussed in Section IV, with the results from tests performed in March 1989 presented in Section V.

The results in Section V show a system that is meeting its performance requirements to an outstanding degree. Both the long- and short-term frequency stability measurements show performance margins of about an order of magnitude better than the requirements. The only exceptions are the frequency-dependent spurious signals generated by an internal frequency synthesizer which is used to tune the carrier into a narrow bandwidth. This represents a subsystem problem that was identified at an early stage, but was not deemed significant enough to warrant a major subsystem change prior to encounter. In order to understand and characterize these spurious signals, however, a calibration test was run, and the results are also included in Section V.

Although this article presents a complete overview of the Parkes system, it supplements a previous article documenting the associated Radio Science Systems installed at CDSCC and at the Usuda, Japan station [1]. Interested readers will also find in this reference a description of the Radio Science experiments, an elaboration on the phenomenon of phase noise and how it degrades the performance of the Radio Science System, and a fuller account of the system test hardware and methodology.

II. System Performance Requirements

Spacecraft Radio Science is concerned primarily with two types of measurements: (1) occultation measurements in which the spacecraft signal passes through a planetary atmosphere, ionosphere, or ring system as it propagates to Earth, and (2) gravitation measurements in which the planet's gravity field perturbs the spacecraft trajectory and hence the downlink Doppler frequency [2]. Radio Science as performed at Parkes, however, is concerned only with the occultation measurements.

For these measurements, there are certain aspects of the spacecraft carrier, namely the fluctuations in its amplitude and phase, from which the properties of the planetary medium can be determined. The Radio Science System is required to record the spacecraft carrier with minimal degradation to these parameters. However, there are many noise processes in the system that may contribute toward their degradation. These include thermal noise, short-term phase noise, receiver-generated spurious signals, long-term phase drifts due to diurnal temperature changes, and gain instabilities.

Of these noise sources, the most critical are those relating to phase stability, both long- and short-term, including spurious signals. Although amplitude, or gain, instabilities are important, they typically exist at levels well below comparable levels of the phase instabilities. This is because any phase-noise components and line spectra spurious signals that enter the first local oscillator (LO) chain at a low frequency level, say 5 MHz, are severely exaggerated by the frequency multiplication performed to generate the X-band LO. This property does not apply to amplitude noise.

Long-term phase fluctuations on the order of 1000 sec in the receiver LOs can prevent the correlation of data between two receiving stations, which in turn reduces the effective signal-to-noise ratio (SNR) of the combined data. Short-term phase-noise processes generated in the receiver, on the other hand, corrupt the power spectrum of the received signal more directly by masking the phase fluctuations placed upon the carrier by the planetary media. Spurious signals generated within the receiver also corrupt the received signal spectrum in such a way as to interfere with the filtering that is performed when the data are processed.

Since the phase stability of the Radio Science System is important over such a wide range of time scales (from 1000 sec to 0.1 msec), the requirement for it is divided into two requirements over two different time scales. The long-term phase stability is specified from 1 to 1000 sec using the Allan variance parameter, which is a universally accepted measure of fractional frequency deviation [3, 4]. The Voyager Neptune requirement for the system Allan deviation is 6×10^{-13} for 1-sec integration times and 3×10^{-13} for 10- to 1000-sec integration times. The short-term phase stability is specified using $\mathcal{L}(f_m)$, the single-sideband spectral density of the phase noise at a carrier offset frequency of f_m, where f_m ranges from 1 Hz to 10 kHz. The required level of the single-sideband phase noise is -53 dBc/Hz at an f_m of 1 Hz, and -60 dBc/Hz from 10 Hz to 10 kHz.

The spurious signals mentioned previously refer to discrete sinusoidal components within the sideband spectrum which are generated when line currents from AC power supplies or other electromagnetic interference (EMI) sources modulate the LO signals. The Voyager Neptune requirement for the system spurious signals is the same level as the phase-noise requirement using a 1-Hz measurement bandwidth (although the spurious signals maintain a level independent of the measurement bandwidth): -53 dBc single-sideband at 1 Hz offset from the carrier, and -60 dBc from 10 Hz to 10 kHz.

III. System Configuration

Figure 1 shows a block diagram of the Radio Science System at Parkes. The antenna is a 64-m dish which gathers and focuses incoming signal flux density to a feed in the aerial cabin, a small room situated at the apex of the tripod structure built upon the dish. Inside the aerial cabin are two traveling-wave masers (TWMs) which can be configured to look at the sky or an ambient load termination. The noise-adding radiometer (NAR) injects its noise diodes immediately prior to the TWMs. Also in the aerial cabin are two RF-IF converters which convert the X-band received signal of 8415 MHz to a 315-MHz IF. Each TWM has its own dedicated RF-IF converter. Taken together, the two TWM and RF-IF channels constitute a primary and backup subsystem. Following conversion, the two IF outputs are sent down a cable run to the Parkes Canberra Telemetry Array (PCTA) van. Inside the van is a switch-

ing assembly that selects the prime or backup IF channel and sends it to both the telemetry receiver and the Radio Science Data Acquisition Subsystem (DAS).

The DAS contains dual parallel IF-VF converter channels and recording assemblies. Two channels are needed to prevent loss of data when the spacecraft switches between the coherent mode of transmission in which the downlink is phase-locked to the uplink, and the noncoherent mode in which the downlink is derived from the onboard Ultra-Stable Oscillator (USO). This is because a single-channel DAS is not able to continuously track the 5.4-MHz shift in received X-band carrier frequency that results from the mode change; it would have to be halted and restarted with a different frequency-tuning predict set. Thus with a dual-channel DAS, each of the two channels is assigned one of the modes and is preprogrammed to track the time-frequency profile of this mode. The two channels are run in parallel so that when a mode change occurs, the signal disappears from one channel and appears immediately on the other channel.

The basic channel in the DAS consists of a Radio Science IF-VF Converter (RIV) followed by a Mark III Occultation Data Assembly (ODA) that digitizes the video signal, tape records the signal, and controls the LO frequency in the RIV. The ODA is made up of a Mark III Narrow-Band Occultation Converter (NBOC), a Modcomp II computer with an IBM PS/2 computer to provide terminal input, and two Wanco tape drives for storing digitized data. For frequency tuning purposes, the Modcomp II computer accepts and stores predict sets of up to 42 points. The predict sets are entered through the PS/2 terminal. The local oscillator used in the RIV is tunable by the ODA computer so that the received spacecraft carrier containing Doppler shift can be kept within the 35-kHz VF-filter bandwidth. The LO mixer injection frequency is derived by frequency multiplying the output of the computer-controlled Dana synthesizer by a factor of seven in order to achieve the desired LO frequency range. The tuning capability is provided by the Programmed Oscillator Control Assembly (POCA), which takes frequency and ramp-rate commands from the Modcomp II computer and controls and monitors the Dana synthesizer accordingly.

The Parkes IF-VF converter is different from the type used in the DSCC Radio Science System in terms of how the bandwidth limiting is accomplished and how the adjacent sideband noise spectrum is rejected from folding over into the passband. The DSCC design uses a quartz-crystal bandpass filter with a center frequency that is offset from the final LO, while the Parkes configuration uses a single-sideband (SSB) mixer for converting the IF directly to VF. The SSB mixer passes the upper sideband, which contains the carrier, and rejects the lower sideband, which contains only noise. The LO for the SSB mixer is offset below the carrier by one-half of the desired video bandwidth. A VF lowpass filter (LPF) following the SSB mixer band limits its output so that the combination forms an equivalent 35-kHz bandpass-type filter, which acts as an anti-aliasing filter for the digital sampling process. A VF amplifier follows the LPF to drive the digitizer in the NBOC.

The video signal is monitored on a spectrum analyzer at a point just prior to its digitization. Also, an RMS voltmeter measures the level of the input signal to the analog-to-digital converters (ADCs) in the NBOC. During system precalibrations, the RIV attenuator is adjusted to achieve a video level, which is largely thermal noise power, of about 1 volt RMS in order to avoid saturating the ADCs.

After sampling the signal, the NBOC sends the digitized data to the Modcomp II for magnetic tape recording. The Modcomp II reads each byte immediately after it has been written to the tape and sends the data back to the NBOC. The NBOC then converts the digital samples back to an analog signal that can be monitored with a spectrum analyzer to verify that the desired data are being recorded to tape.

The Frequency and Timing Subsystem (FTS) at Parkes uses two hydrogen-maser frequency standards located in a temperature-stabilized room that is remote from the PCTA van for magnetic isolation purposes. The frequency standards provide 5-MHz references to the PCTA van, each sent through a quartz-crystal oscillator clean-up loop (CUL) to improve the phase noise at low signal offset frequencies near 1 Hz. The output of the prime CUL is distributed throughout the van using two HP5087 distribution amplifiers. The HP5087 amplifiers supply reference signals to all of the time-code generators (TCGs) in the van, which in turn provide timing signals to the ODA and telemetry computers.

The redundant receiver first local oscillators are derived from the 5-MHz CUL output. First, a $\times 20$ multiplier in the PCTA van multiplies the CUL output up to 100 MHz. Two 100-MHz outputs are sent via hardline coaxial cables to the aerial cabin. There the 100-MHz signals are further multiplied by a factor of 81 to provide the 8.1-GHz first LOs to the RF-IF downconverters.

IV. System Test Methodology

The system testing at Parkes was aided by the test methodology and hardware that had been developed for

the CDSCC. Both of the instruments developed for Radio Science System stability testing—the test transmitter and the digital stability analyzer (DSA)—were designed to be transportable between the Canberra and Parkes facilities.

The block diagram of the Radio Science System at Parkes in Fig. 1 shows this test instrumentation. The methodology for Radio Science System stability testing involves injecting a highly stable RF test signal into the TWM input and measuring the stability of the test signal after it has been downconverted to video-frequency, digitized by the NBOC, and then reconstructed back to an analog waveform. This procedure is a direct means of measuring the amount of degradation the entire system imposes upon the received signal.

The RF test signal is generated by the portable test transmitter. The transmitter contains a phase-stable quartz-crystal oscillator for performing phase-noise measurements, and is phase-locked to an FTS 5-MHz reference for performing Allan variance measurements. To reliably measure the Radio Science System's performance, the transmitter was designed with performance requirements about an order of magnitude better than the Radio Science System requirements.

There are two places where the RF test signal can be injected into the system. The test transmitter is equipped with an RF horn so that the signal can be radiated into the antenna feed from a strategic point on the ground. However, the proximity to the antenna of most convenient locations on the ground tends to give rise to multipath effects which distort the test results. For all of the system tests described in Section V, the multipath problem is avoided by injecting the test signal into the system through a coupler at the TWM input.

In order to improve the system SNR, a 30-dB attenuator is inserted into the signal path immediately following the TWM. This allows injecting a 30-dB stronger signal into the TWM without saturating the following receiver subsystem. However, the resulting increase in SNR is only about 19 dB because the thermal noise of the receiving subsystem now becomes the dominant component of the system noise temperature.

For signal analysis, the digital stability analyzer accepts the system's video-frequency output either prior to or following digitization and processes it for phase noise and Allan variance information. The DSA is essentially a computer containing its own ADC for signal acquisition, as well as digital signal processing hardware and software to measure the power spectrum of a video test signal and decompose it into the quadrature component amplitude and phase spectra. The system phase-noise response can be read directly from the phase spectrum. For Allan variance tests, the DSA contains hardware which downconverts the video signal to 1 Hz. A time-interval counter samples the zero crossing of this 1-Hz signal and sends it to the DSA's computer for processing [1, 3, 4].

During the phase-noise test, the test transmitter is ideally configured to run on its internal oscillator so that there is no coherence between the system's reference frequency and the test signal. With this configuration the entire system performance is measured, including that of the hydrogen-maser frequency standard. On the other hand, for Allan variance testing the test transmitter must be phase-locked to the station FTS because the long-term stability of its quartz-crystal oscillator is inadequate for measuring system performance over long test periods. Because the transmitter is phase-locked to the station FTS, the FTS effects cancel out. Therefore, the stability of the FTS must be measured separately with the DSA by comparing the two similar hydrogen-type masers.

V. System Test Results

The testing of the Radio Science receiver at Parkes helped to uncover and remove many problems that were not apparent at the subsystem level. However, the results shown here include only typical performance curves taken after most of the problems had been removed. In addition, measurements taken with the NAR operating in encounter configuration are also presented. The last section summarizes a special calibration test made on the Dana LO synthesizers in the IF-VF converters. This test was performed to help characterize the behavior of certain frequency-dependent spurious signals that are generated by these synthesizers.

A. Allan Variance Test Results

Figure 2 shows the Allan variance results of the system measured with the DSA. These data were taken using the test transmitter, the prime 5-MHz CUL, the hydrogen maser, and the backup (channel 1) ×81 multiplier feeding into the prime (channel 2) RF-IF downconverter. The unusual ×81 configuration was implemented to bypass a malfunctioning unit. These results represent the system at its best performance. The Allan deviation at 1 sec is almost an order of magnitude below the 6.0E-13 specification. The only concern is the peak at about 1000 sec which represents a long-term drift in the system phase. The source of the drift was not determined with confi-

dence, so it could be due to the test transmitter, system cable temperature effects, or the LO multiplier chain. In any event, the peak is still an order of magnitude better than the system specification.

B. Phase Noise and Power Spectrum Tests

Figure 3 shows the system phase-noise spectral density measured with the DSA. The hydrogen-maser frequency standard was used. The system configuration was not optimal since a ×81 frequency-multiplier unit that later proved to be defective was used, although it was not malfunctioning during this period. The frequency span of the plots is 200 Hz around the carrier and shows detail on a number of spurious signals. Of particular interest is a set at 1 Hz, approximately −59 dBc below the carrier. The source of these spurs was not determined; however, they were not present during the later testing period and do not show up in the final data (for example, see Fig. 4).

Note that the level of the discrete sinewave components in Figs. 3 through 7 is adjusted to correct for the resolution bandwidth of the spectrum measurement. A spectrum analyzer typically measures the level of the discrete components of a signal correctly, but the continuous, broadband components need to be scaled by one over the resolution bandwidth of the measurement to indicate spectral density in units of dBc/Hz. For Figs. 3 through 7, the normalization has already been applied to the vertical axis so that the broadband noise components can be read directly off the plots in units of dBc/Hz. However, the normalization of the vertical axis makes the discrete component levels incorrect, so an adjustment is made to indicate their correct levels in units of dBc. Stated simply, the spectrum plots are corrected to show what the spectrum would look like if it was measured with a resolution bandwidth of exactly 1 Hz, and both the discrete (after correction) and continuous components are read off these plots using the same vertical scale.

For Figs. 8 through 11, only the discrete components are of interest, so the vertical axis is not adjusted. The discrete components are read off as they are plotted in units of dBc without any need for adjustment. The broadband components are plotted in units of dBc per the resolution bandwidth of the measurement (0.38 Hz, 0.38 Hz, 2.4 Hz, and 2.4 Hz respectively).

The plot in Fig. 3 shows that neither the discrete noise components nor the broadband noise components violate the phase-noise specification. However, it should be noted that the test transmitter in this case is phase-locked to the station FTS 5-MHz reference, so that the effects of FTS have cancelled out.

In order to measure the performance of the entire Radio Science System, including the FTS standard, the above test was repeated ten days later with the system again running on the hydrogen-maser frequency standard but with the test transmitter free running on its internal quartz-crystal oscillator. The uncertainty in the output signal frequency from the test transmitter made the use of the digital stability analyzer impractical for making system measurements, so the HP3561 spectrum analyzer was used to measure the power spectrum. Since the power spectrum is the root-sum-square of the phase and amplitude spectra, a power spectrum which meets the phase noise requirement implies a phase-noise spectrum which also meets the requirement.

Figures 4, 5, 6, and 7 show the single-sideband power spectrum around the carrier over four different frequency spans: 20 Hz, 400 Hz, 2 kHz, and 20 kHz. The close-in single-sideband noise power at 1 Hz is seen in Fig. 4 to be −57.3 dBc/Hz. This compares well with previous measurements of the test transmitter which showed a single-sideband phase noise at X-band of −58 dBc/Hz. The level at 10 Hz is less than −65 dBc/Hz. From 100 Hz to 10 kHz, the noise is less than −75 dBc/Hz. The power line spurs are less than −60 dBc, while the Dana spurs at 2.85 kHz are −52 dBc.

Figure 5 shows two sets of spurious signals that are generated by the Modcomp II computers in the ODAs. For comparison, Fig. 6 shows the power spectrum taken with the Modcomp IIs turned off. One set of spurs shows up (in Fig. 5) at about 109 Hz and is generated by the prime ODA. The other set of spurs at about 57 Hz is generated by the backup ODA. Over time both spurs were seen to drift randomly in frequency by several hertz or so.

The source of these spurs was unknown until the ODAs were switched off. Further investigation by the personnel at Parkes revealed that the spurs were generated by clock signals internal to the ODAs, which were not phase-locked to the FTS and hence slightly off-frequency. The clock signals were phase-modulating the FTS 5-MHz reference at the beat note frequency between the sixth harmonics of the two timing signals. The interaction was made possible through a faulty hardline connector that coupled the ODA clock radiation into the FTS 5-MHz signal. Even though the resulting spurious phase modulations were quite low on the 5-MHz line, they were exaggerated to a level of −45 dBc by the frequency multiplication factor needed to generate the 8.1-GHz first LO. The replacement

of the faulty connector reduced the power of the spurious signals to an acceptable level.

Figure 8 shows the close-in power spectrum of the 20-kHz VF carrier with 3.5-Hz and 10.5-Hz offset spurs present. This test was run with the NAR injecting a 50-K noise diode modulated at 20 Hz into the signal path before the low-noise amplifier (LNA) maser. Since the spurs are only 48 dB below the carrier, and there was no time to investigate and reduce the interaction, it was recommended that the 50-K noise diode not be used during the prime Radio Science data-gathering time periods. Fortunately, the 1-K noise diode, which is currently planned for use during the encounter, does not contribute any measurable effect above −65 dBc, as shown in Fig. 9. Note that the test configuration was run without the 30-dB attenuator after the maser that is normally installed during system testing to improve the test signal's SNR, so the noise floor with this 100-Hz span is only −67 dBc. This was necessary because the 30-dB attenuator in the front end changes the system noise floor and renders the NAR inoperable.

C. Dana Calibration Tapes

A series of calibration tapes was run at Parkes to characterize the spurious signals generated by the Dana synthesizers, which are used to provide the tuned local oscillator in the IF-VF converter. During the early phase of subsystem testing at JPL, these synthesizers had exhibited two types of spurious signals which violated specifications. The first type was induced by power supplies in the Dana chassis and nearby instruments which modulated the Dana's crystal oscillator at the power-line frequency via magnetic field coupling. These spurs were reduced by shielding the synthesizers from magnetic flux and moving their internal power supplies into a separate chassis. The second type are frequency-dependent modulations on the synthesizer output that are generated by the internal digital circuitry of the synthesizer. These spurs have been measured on the video carrier as high as about −52 dBc in some cases and may actually shift across the carrier signal as the Dana frequency changes. Unfortunately, they could not be removed from the device without a fundamental change in its design.

In order to characterize the expected behavior of these spurs during the encounter, predict sets were generated that stepped the POCA-controlled Dana across the important parts of the encounter frequency range. Energy level is examined to characterize the effect of a spur close to the carrier when the Dana frequency is changed. A total of three predict sets of 21 frequency steps each was constructed. One set was made to cover the one-way mode prior to the Neptune occultation, a second set was made for the two-way mode prior to occultation, and a third set was made to cover the one-way mode after the occultation. The frequency range of the predict sets is defined in Table 1. The actual frequency of each point can be determined by the equation:

$$F(i) = F\text{start} + \frac{i(F\text{stop} - F\text{start})}{20} \quad \text{for } i = 0 \text{ to } 20$$

where Fstart and Fstop are taken from Table 1.

The tapes were made by simulating the spacecraft signal with an HP8662 synthesizer injected into the IF input at the RIV. During recording, the HP8662 was stepped in frequency with an HP85 computer so that the resultant video remained at 20 kHz for each step. The duration of each step was about 18 sec, with 1 sec allowed for frequency changes between steps.

Both of the Danas were calibrated with the three predict sets for a total of six tapes. The tapes were analyzed on the JPL Radio Occultation Data Analysis (RODAN) computer. Some of the results are presented below.

In general, the Dana with NASA serial #106227 is superior to Dana #118430 because it has much fewer spurs. Figures 10 and 11 compare #106227 with #118430 at a Dana frequency of 45578013.78 Hz taken from the one-way pre-occultation predict set. The pair of spurs at about 1.5 kHz from the carrier is at −57 dBc. When the Dana frequency is in this neighborhood, the 1.5-kHz spurs drift at a rate equal to ten times the change in Dana frequency, so it is natural to assume that they will cross the carrier at about the same level. In other words, if the Dana frequency increases by 100 Hz, the spurs will move closer to the carrier by 1 kHz, or will be about 500 Hz away.

Not all of the spurs obey this 10-to-1 dependency upon the Dana frequency. Other plots not shown here reveal a 1-to-1 relationship, and in one case a relationship much higher than 10 to 1 was observed.

In order to calculate the real effect of the Dana spurs upon the system, one must consider the rate at which they will move during encounter. Since a moving spur spreads its power across a finite frequency range, its level at one discrete frequency point in a measured power spectrum will be much lower than that of a stationary spur with the same power level. How much lower depends upon the rate at which the spur moves.

For example, the spectra in Figs. 10 and 11 are obtained by averaging successive fast-Fourier transforms (FFTs) of the sampled video signal. The data were sampled at 80 kilosamples/sec. A total of 32768 data points is used for each FFT. This gives an FFT bin spacing of approximately 2.44 Hz, with a data acquisition time of 0.4096 sec for one buffer of data. If instead of being a fixed frequency, the LO had tracked the carrier at a rate of about 30 Hz/sec due to Doppler (about the highest rate predicted for the Neptune encounter), the carrier would remain steady at the video frequency, but the spurs would change frequency. During the time it took to acquire the data for one FFT, the LO would move by 12.288 Hz. The Dana frequency itself would only change by 1.75 Hz because of the $\times 7$ frequency multiplier placed after the Dana. If the spurs were moving at a rate ten times the Dana frequency change, they would have shifted by 17.5 Hz or about seven FFT bins. Assuming a uniform ramp rate, the signal power would be divided evenly among seven FFT bins instead of one, with each bin showing only about one-seventh the original power of the spur. The peak power in this smeared spur would then be 8.5 dB lower than if the spur had not been smeared. If, however, the spur obeyed a 1-to-1 dependency upon the Dana frequency, it would move only 1.75 Hz and would likely spend most of the time within a single FFT bin, so no reduction in level would be observed.

VI. Conclusions

This article has outlined the installation of the Radio Science System at the Parkes Radio Telescope in New South Wales, Australia. The Voyager Neptune encounter performance requirements and system design were discussed, as well as the techniques used to measure the performance of the system and the actual data taken on site during performance verification. The results shown in Section V indicate that in general the system is performing with a comfortable margin on the desired characteristics. The problem areas remaining in the system are listed below.

In the area of Allan variance performance, the system is operating at about an order of magnitude better than specification. The only concern lies in a long-term phase drift that makes the Allan deviation rise up to about 1E-14 at 1000 sec. This is not in violation of specification, but could be a problem if it gets any worse.

In the area of phase-noise performance, the system is also operating at a level better than specification. An initial problem with the ODA computers generating unwanted modulations on the first LO was corrected with the replacement of a faulty hardline connector. The only remaining violations are the discrete frequency components generated by the Dana synthesizers described in Section V, which were considered too minor to warrant a major subsystem change to remove them.

Finally, a problem is noted concerning which noise diode is used in the NAR. A 50-K noise diode generates interfering spurious signals that show up at 3.5 and 10.5 Hz on the carrier at video frequencies, at about -48 dB below the carrier. However, no spurs above -65 dBc are measured when the 1-K noise diode is used.

The system is considered adequate and ready to receive and record the Voyager spacecraft signal during the upcoming Neptune encounter. It meets the basic system requirements and contains the additional functional capability to receive the downlink signal during the abrupt frequency shifts associated with spacecraft mode changes without loss of any data.

Acknowledgments

The authors would like to acknowledge the collaboration and guidance provided by N. Ham throughout the duration of this project.

References

[1] N. C. Ham, T. A. Rebold, and J. F. Weese, "DSN Radio Science System Design and Testing for Voyager-Neptune Encounter," *TDA Progress Report 42-97*, vol. January–March 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 252–284, May 15, 1989.

[2] G. L. Tyler, "Radio Propagation Experiments in the Outer Solar System with Voyager," *Proceedings of the IEEE*, vol. 75, no. 10, pp. 1404–1430, October 1987.

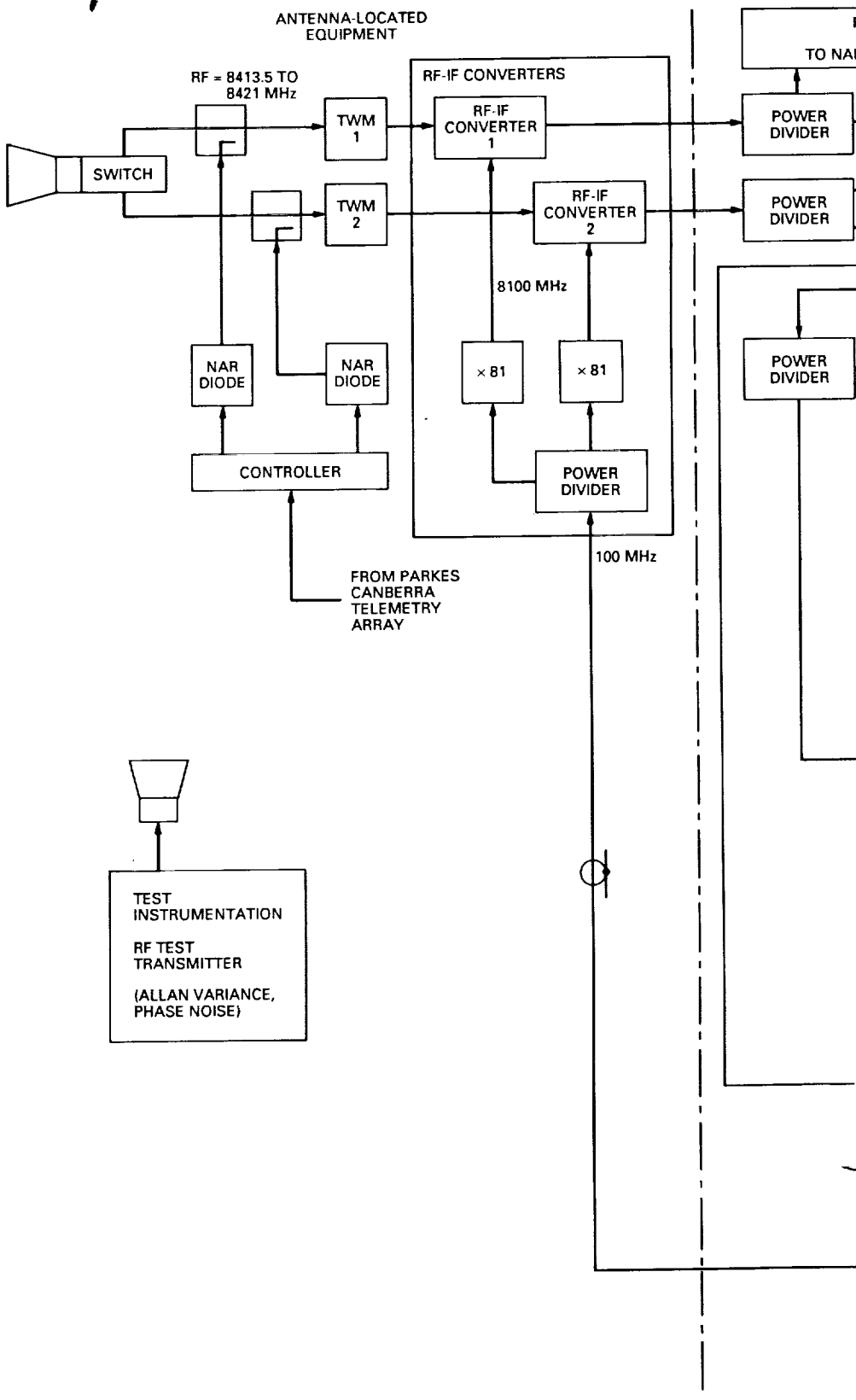
[3] D. W. Allan, "Time and Frequency (Time-Domain) Characterization, Estimation, and Prediction of Precision Clocks and Oscillators," *IEEE Trans. on Ultrasonics, Ferro-electrics and Frequency Control*, vol. UFFC-34, no. 6, November 1987.

[4] C. A. Greenhall, "A Method for Using a Time Interval Counter to Measure Frequency Stability," *TDA Progress Report 42-90*, vol. April–June 1987, Jet Propulsion Laboratory, Pasadena, California, pp. 149–156, August 15, 1987.

Table 1. Predict sets used for Dana calibration

Set type	X-band range (GHz)		POCA range (MHz)	
	*F*start	*F*stop	*F*start	*F*stop
One-way Pre-occultation	8.419059	8.419124	45.577085	45.586371
Two-way Pre-occultation	8.41365255	8.41371755	44.804735	44.814021
One-way Post-occultation	8.419340	8.419404	45.617228	45.626371

FOLDOUT FRAME 1.



198.P

FOLDOUT FRAME 2

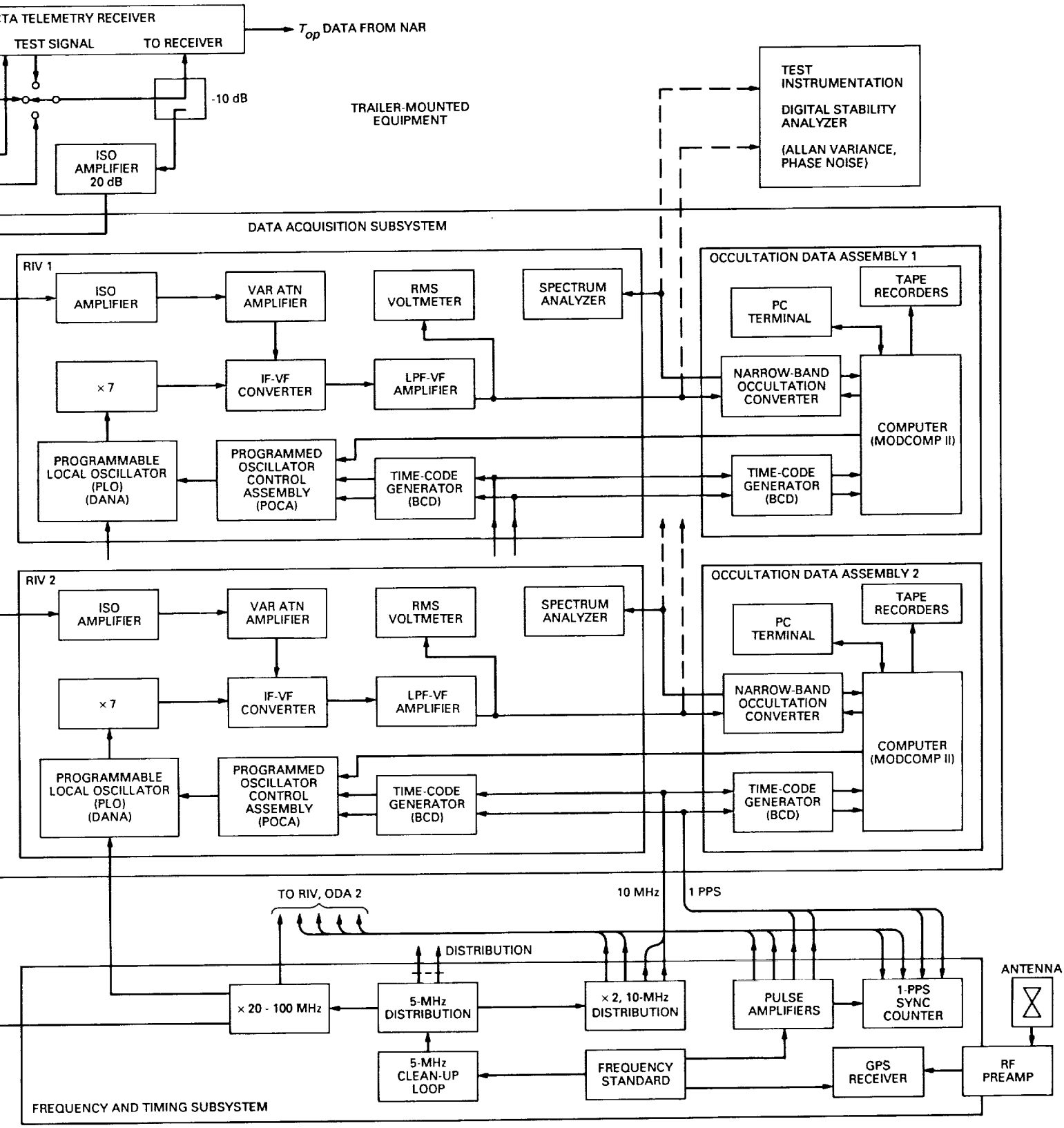


Fig. 1. Block diagram of the Parkes Radio Science System configuration.

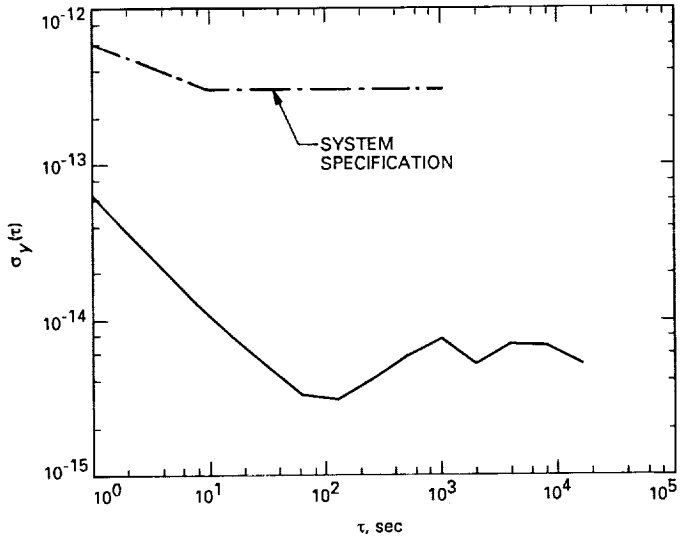


Fig. 2. Allan variance of the Parkes Radio Science System.

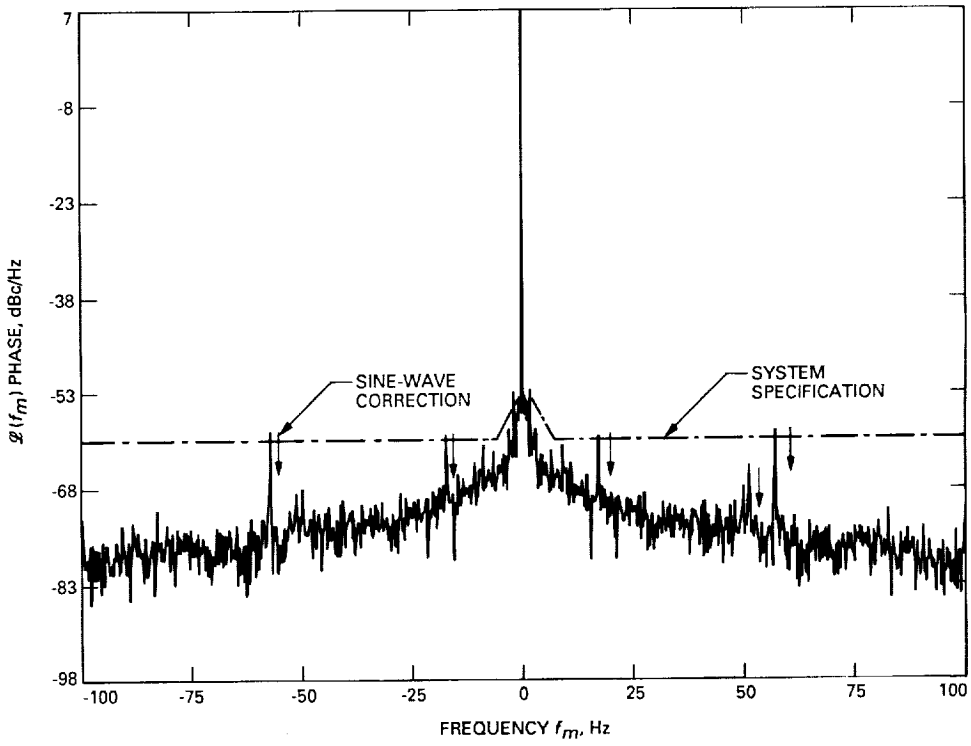


Fig. 3. Phase-noise spectral density of the Parkes Radio Science System, test transmitter phase-locked to FTS, −100 to +100 Hz (center frequency = 20 kHz).

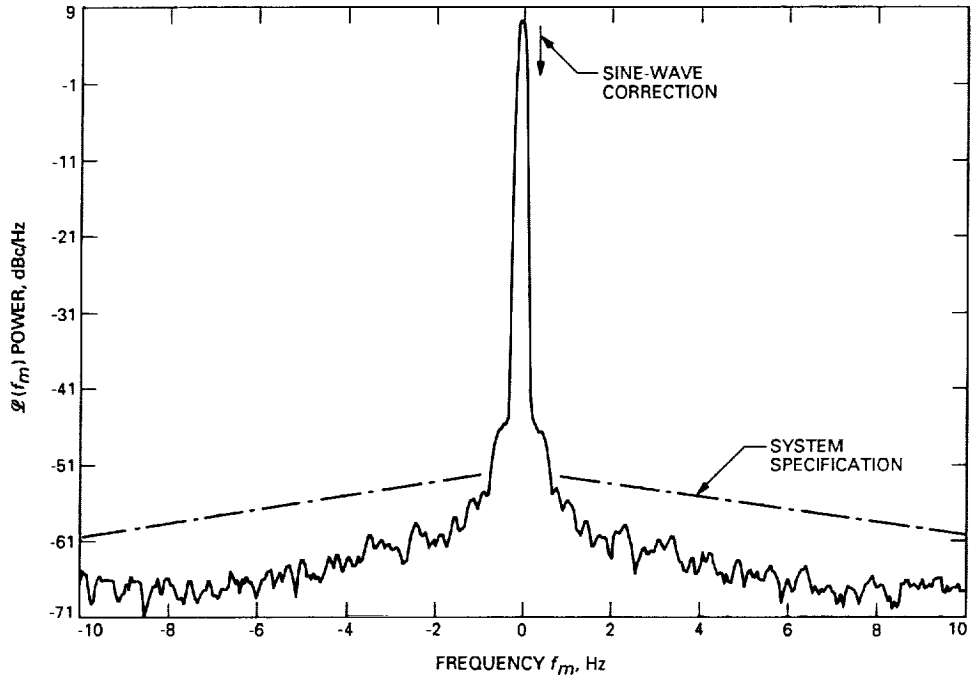


Fig. 4. Noise-power spectral density of the Radio Science System, test transmitter free-running, −10 to +10 Hz (center frequency = 19.95 kHz).

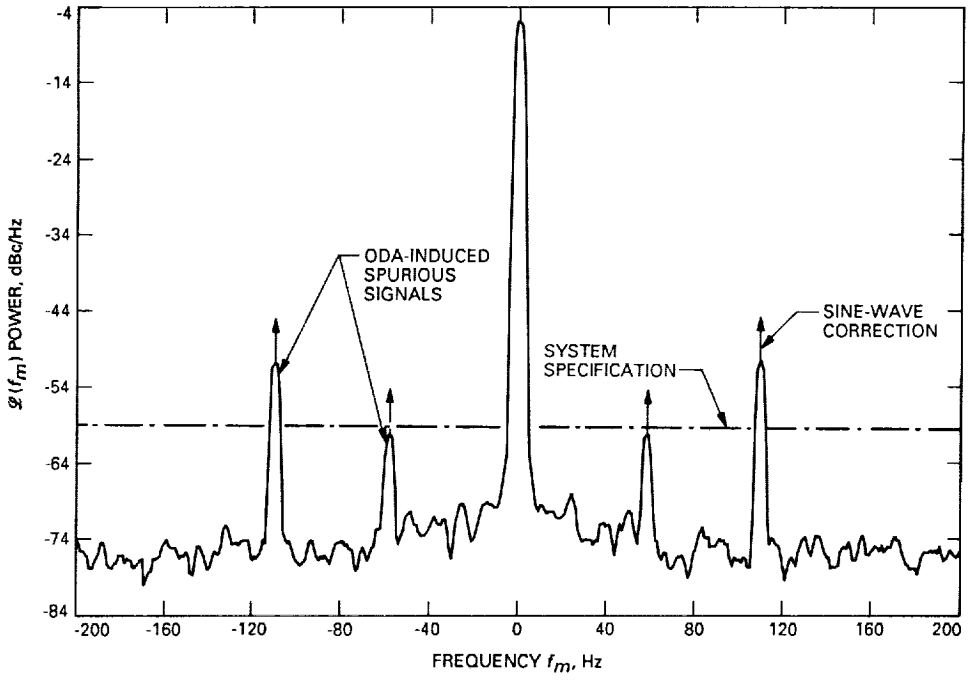


Fig. 5. System noise-power spectral density showing ODA-induced spurious signals, −200 to +200 Hz (center frequency = 19.95 kHz).

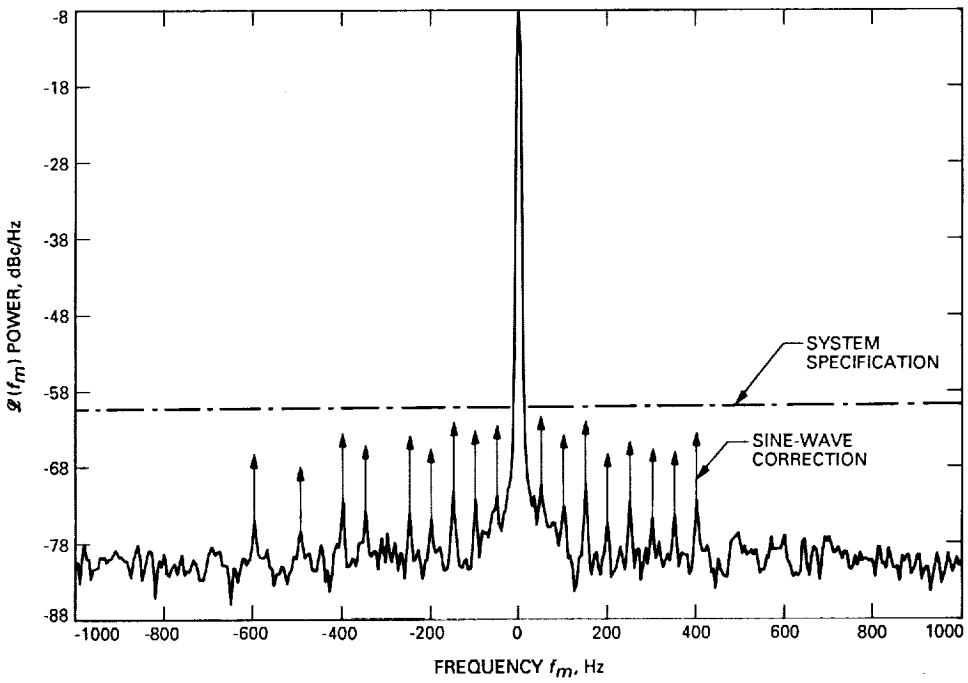


Fig. 6. System noise-power spectral density with ODAs turned off, −1000 to +1000 Hz (center frequency = 19.95 kHz).

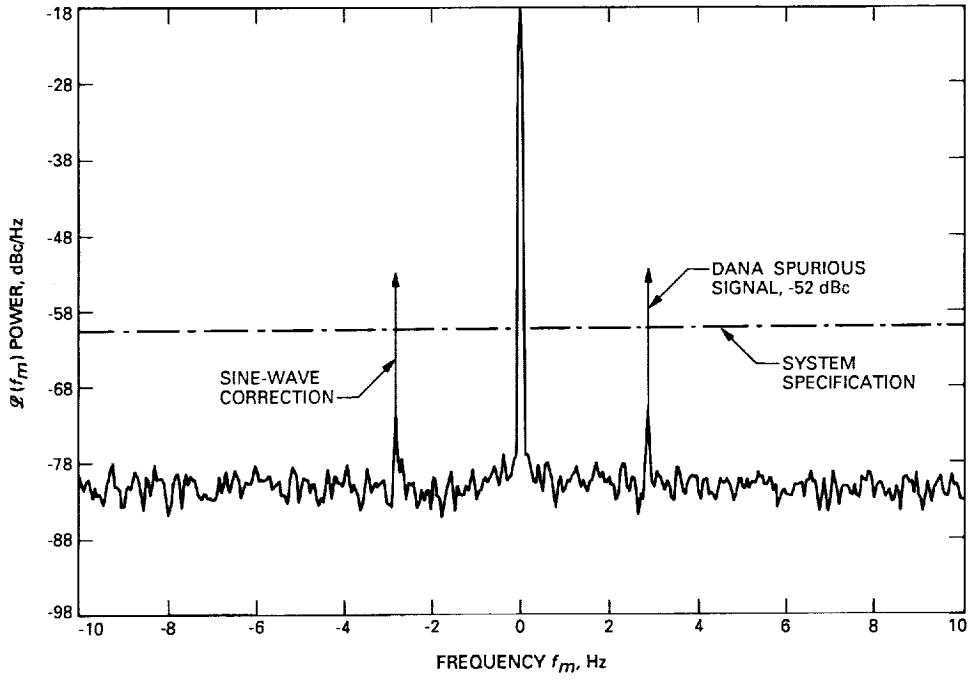


Fig. 7. System noise-power spectral density showing Dana spurious signals, −10 to +10 kHz (center frequency = 19.95 kHz).

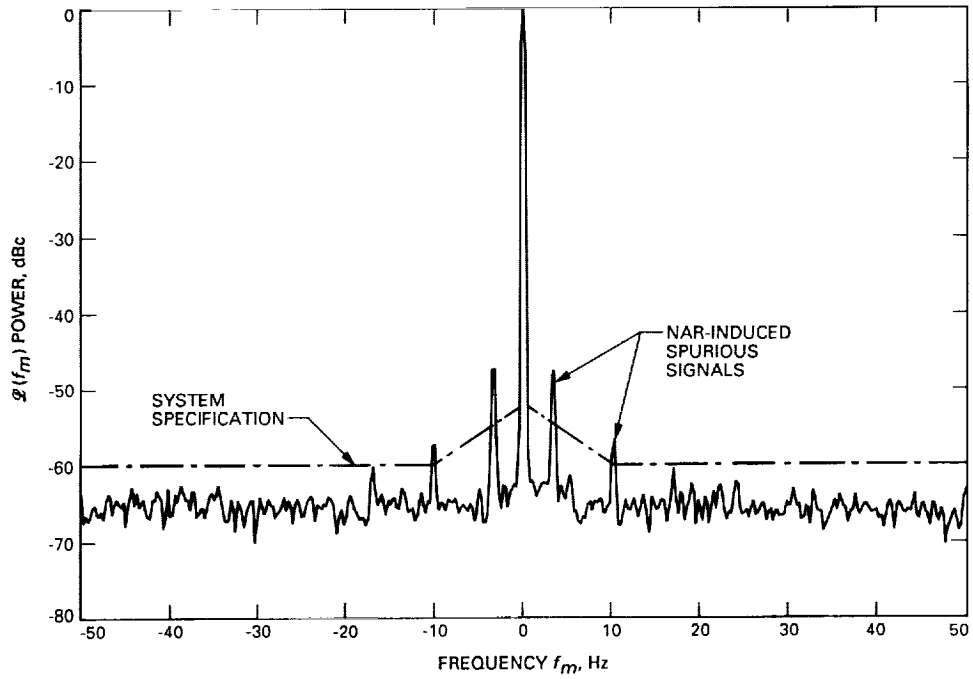


Fig. 8. System power spectrum showing spurious signals, NAR operating with 50-K noise diode.

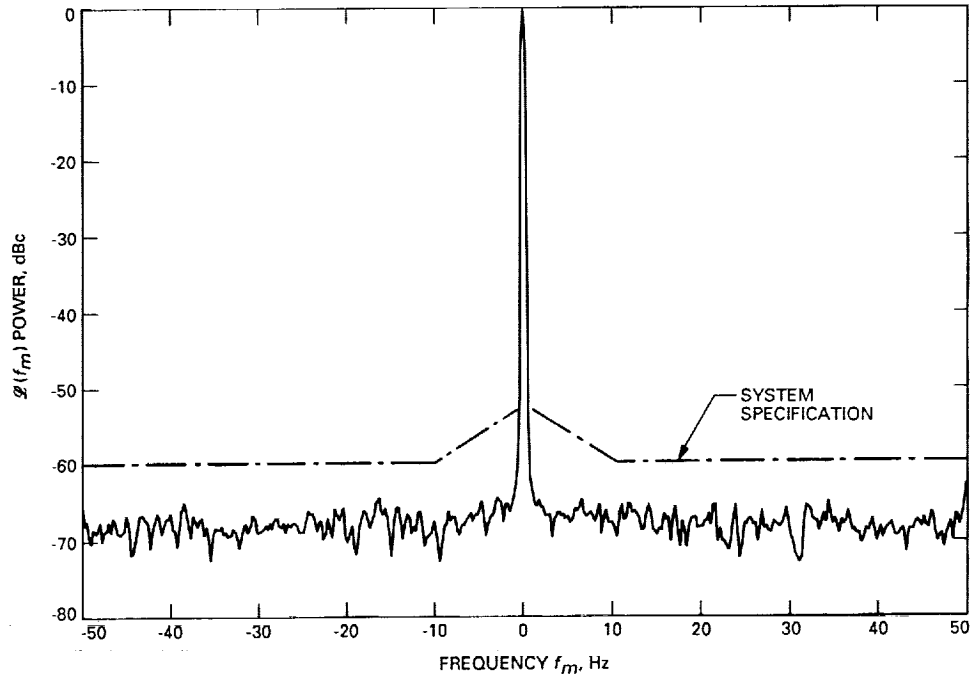


Fig. 9. System power spectrum showing no spurious signals, NAR operating with 1-K noise diode.

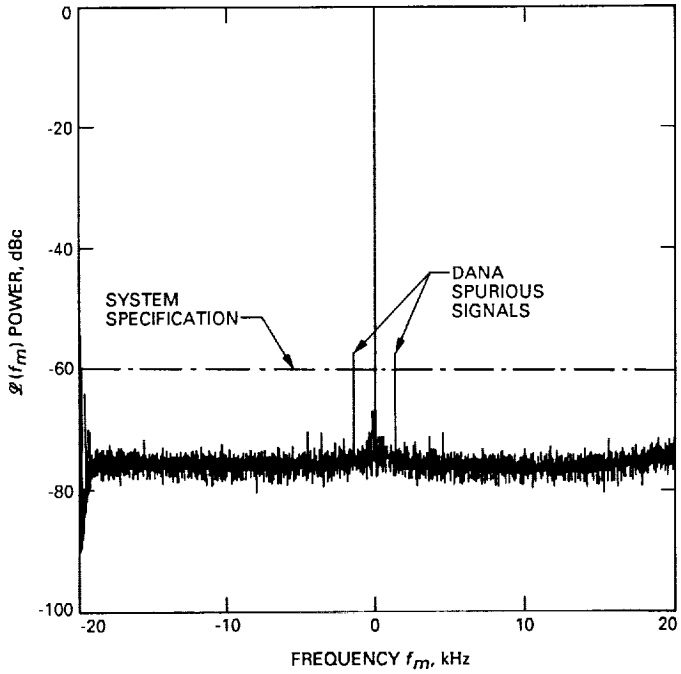


Fig. 10. Power spectrum showing Dana #106227-generated spurious signals, Dana frequency = 45.578013 MHz. Plot computed by JPL RODAN computer from tape-recorded samples.

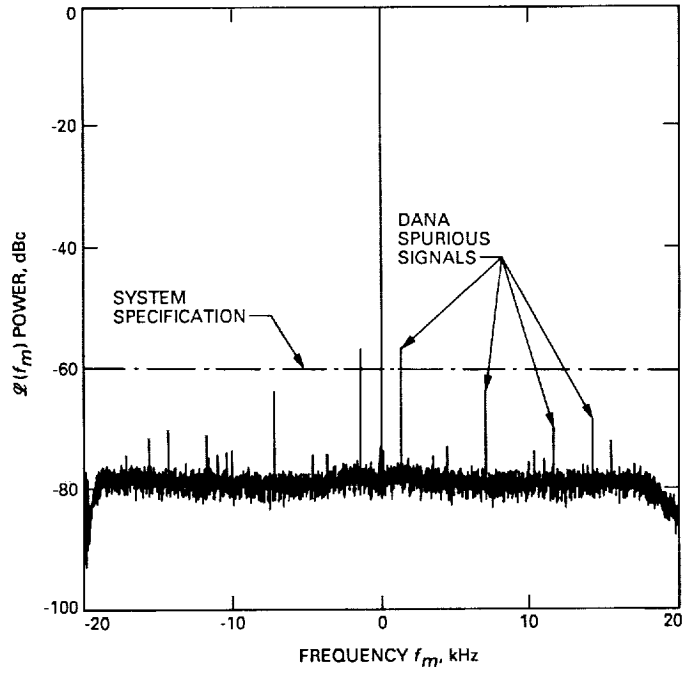


Fig. 11. Power spectrum showing Dana #118430-generated spurious signals, Dana frequency = 45.578013 MHz. Plot computed by JPL RODAN computer from tape-recorded samples.

32-GHz Performance of the DSS-14 70-Meter Antenna: 1989 Configuration

M. S. Gatti
Ground Antenna Facilities and Engineering Section

M. J. Klein and T. B. H. Kuiper
Space Physics and Astrophysics Section

The results of preliminary 32-GHz calibrations of the 70-meter antenna at Goldstone are presented. Measurements were done between March and July 1989 using Virgo A and Venus as the primary efficiency calibrators. The flux densities of these radio sources at 32 GHz are not known with high accuracy, but were extrapolated from calibrated data at lower frequencies. The measured value of efficiency (0.35) agreed closely with the predicted value (0.32), and the results are very repeatable. Flux densities of secondary sources used in the observations were subsequently derived. These measurements were performed using a beamswitching radiometer that employed an uncooled high-electron mobility transistor (HEMT) low-noise amplifier. This system was installed primarily to determine the performance of the antenna in its 1989 configuration, but the experience will also aid in successful future calibration of the Deep Space Network (DSN) antennas at this frequency.

I. Introduction

NASA/JPL is planning to use the 32-GHz frequency band for communications and navigation of future deep-space missions. Performance estimates of existing and future ground station capabilities are needed for both mission planning and technology development. A research and development radiometer has been placed on the DSS-14 70-meter antenna as a precursor to the implementation of 32-GHz systems on the 70-meter antenna network. The radiometer will be used to determine the current performance characteristics of the large antennas at this frequency and to determine what upgrades may be advisable for future systems. Our experience with the 70-meter calibrations will also contribute to successful calibrations of the new 34-meter antennas, which will be built in the future.

This article describes the measurements performed at 32 GHz between March and July, 1989, on the recently upgraded (64-m) 70-meter antenna at Goldstone (DSS-14). These measurements were done on seven separate days by using radiometric observations of six astronomical radio sources: 3C274 (Virgo A), 3C273, 3C84, 3C286, $P2134+004$, and Venus.

It is important to note that the flux densities of natural radio sources are not accurately known at 32 GHz. Very few attempts to calibrate radio source intensities on an absolute scale have been reported above 10 GHz, so that microwave spectra of even the most intense radio sources are increasingly uncertain at high frequencies. Estimates of the uncertainty in the absolute flux density scale at 32 GHz are typically 10 percent or more. The precision of the measurements reported in this article is sufficiently high that error estimates for the results are clearly dominated by the absolute flux error and not by random errors.

II. Radiometer Description

A. Description

At 32 GHz, radiometry is complicated by tropospheric effects. Specifically, the system operating temperature varies due to clouds, moisture, and other weather phenomena. The measurement accuracy of the increase in system temperature due to a radio source is, therefore, significantly impacted for all but the best of observing conditions. There are techniques to eliminate these adverse effects; the most common one is to use a beam-switching radiometer. This radiometer is an adaptation of the one proposed by Dicke [1] wherein the difference in temperature between a reference signal and the unknown signal is measured by the system. For our radiometer, the reference signal is derived from an antenna beam pointing slightly off-axis of the desired antenna beam. The reference is the cold sky, and the unknown is the source temperature plus the cold sky. The difference in detected power is, therefore, proportional to the source temperature.

The reference signal is detected when the system is switched to the second feedhorn, which is laterally displaced from the Cassegrain focus of the 70-meter reflector. For this application, the displacement is 3.375 inches. This displacement produces a beam for this feed that is offset approximately 30 millidegrees lower in elevation than the beam produced by the on-axis feed (boresight) of the antenna. This represents approximately 4 beamwidths of the pattern defined by the on-axis feed. This technique is similar to that proposed by Slobin, et al. [2]. The off-axis reference antenna beam is used to provide an "off-source" reference temperature for the Dicke switched radiometer. For this purpose, the reference beam need not be precisely matched to the on-axis beam, but the cancellation of variations in sky emission is improved if the two beams yield nearly identical system temperatures when each in turn is directed at the sky.

Figure 1 shows a block diagram of the radiometer system. Figure 2 is a photograph of the system as assembled for test. The block diagram shows the two separate feedhorns, each of which produces an antenna beam when placed on the 70-meter antenna. The feedhorn that produces beam 1 (as designated on the block diagram) is placed at the focal point of the reflector system. The signals from each beam are alternately sampled by a circulator switch that operates between 2 and 20 Hz. The low-noise amplifier (LNA) consists of a high-electron mobility transistor (HEMT) device. The system requires no cryogenic devices and is quite compact. The remaining portions of the system consist of follow-up amplifiers, downconverters, detectors, and a Stanford Research SR-510 lock-in amplifier.

The signal from the Dicke switch is a modulated square wave that is the difference between the signals in the two beams. The lock-in amplifier demodulates this signal and produces the desired difference signal. The strip chart recorder plots this difference signal as a function of time. If each antenna beam is observing the same target, such as the sky, and if the system is perfectly balanced, a zero signal from the lock-in amplifier is recorded. When beam 1 is pointed to a radio source and beam 2 is off the source, the signal is proportional to the increase in system temperature due to the radio source. The output from a diode noise source is coupled into the waveguide of beam 1 to calibrate the system. The diode calibration is routinely switched on several times each hour as the radio sources are observed. With this procedure, the increase in system temperature produced by each radio source is calibrated, and the adverse effects of gain variations in the HEMT and receiver are virtually eliminated. The aperture efficiency of the antenna is then derived from the measured source temperatures and the assumed values of the source flux densities at 32 GHz (see Section IV).

The radiometer was placed in the XKR feedcone of the 70-meter antenna at Goldstone. This feedcone contains other hardware, including a 22-GHz Dicke beamswitching radiometer and the X-band planetary radar. Figure 3 shows the top of this feedcone, including the variety of feeds. The front-end portion of the radiometer, consisting of the feeds, waveguides, switches, amplifiers, and downconverter, is located with the feeds in the feedcone. The intermediate frequency (IF) signal is sent via coaxial cable from the feedcone to the pedestal room where the lock-in amplifier, square-law detector, recorder, and pulse generator are located.

B. Operating Characteristics

The operating characteristics of the radiometer are defined by several quantities, the most important of which are

the stability and the operating system temperature, T_{op}. These characteristics limit the sensitivity of the system and determine the configuration of the actual observations, i.e., switching rates, bandwidths, follow-up amplifier gain, etc. The individual components for this system were measured and their contributions to T_{op} were determined. Table 1 shows the individual contributions to the system temperature (see Section III for definitions). Also, the system temperature was measured by the calibration techniques discussed in the next section. The measured value and predicted value of T_{op} agreed remarkably well. For these and all subsequent measurements, the reference plane for T_{op} was at the input to the feedhorn (due to the calibration technique). The tolerances reported in this article have 2σ (95 percent) confidence. The final result is that the effective receiver temperature T_e equals 391 (±8) kelvin. When the effects of the atmosphere and galactic background are included (by estimates), it is expected that T_{op} will be 410 (±10) kelvin. Actual measurements at DSS-14 indicate a T_{op} of 415 (±7) kelvin. The next section provides the details of the calculation and calibration.

The radiometer settings for our observations are given in Table 2. Included in this table are the time constants for the power integration, switching rates for the Dicke switch, IF bandwidth, and HEMT gain.

III. Calibration Techniques

A. Definitions

In this section, the quantities used in the calculation of the system temperatures and the quantities used to calibrate the system are defined. Following sections show the equations used to perform the calibration and calculations. The definitions (all temperatures in kelvin) are as follows:

T_{op} = operating system temperature of radiometer

T_a = effective antenna temperature

T_e = effective receiver temperature

T_{sky} = total temperature due to the sky contribution

T_{atm} = temperature due to the atmosphere

T_{gal} = 3.0-K cosmic background radio emission

T_{so} = temperature contribution due to feed spillover

T_{qs} = temperature contribution due to quadrupod scattering

T_{tl} = temperature due to the transmission line between amplifiers and feed (including switches)

T_p = physical temperature of transmission line (300 K)

T_m = temperature due to the LNA

T_{etc} = temperature due to the follow-up equipment, including the downconverter and IF equipment

T_f = temperature of T_{etc} referenced to the input of the LNA, called "follow-up" temperature

G_{tl} = gain of the transmission line (< 1)

L_{tl} = loss of the transmission line $= 1/G_{tl}$

G_m = gain of the LNA

For any measurement, the reference plane location for calibration must be explicitly stated since the resulting temperatures depend on reference location. In the calibrations performed, the measurement reference plane was located at the aperture of the feedhorn. Therefore, this position was chosen for reporting the various temperatures and noise diode values.

B. Calculation of System Parameters

For the reference plane chosen in this work, the calculation of T_{op} was performed using the following equations:

$$T_{op} = T_a + T_e \tag{1}$$

$$T_a = T_{sky} + T_{so} + T_{qs} \tag{2}$$

$$T_{sky} = T_{atm} + T_{gal} \tag{3}$$

$$T_e = T_l + \frac{T_m}{G_{tl}} + \frac{T_{etc}}{G_{tl}G_m} \tag{4}$$

$$T_l = (L_{tl} - 1)T_p \tag{5}$$

$$T_f = \frac{T_{etc}}{G_m} \tag{6}$$

Combining Eqs. (2) through (6) into Eq. (1) yields

$$T_{op} = \left[(T_{atm} + T_{gal} + T_{so} + T_{gs})\right] + \left[(L_{tl} - 1)T_p + L_{tl}(T_m + T_f)\right] \tag{7}$$

Using Eq. (7) and the information in Table 1, the value of T_{op} is expected to be 409.5 K, which may be compared to the measured value given earlier as 415 K.

C. Aperture Load and Noise Diode Calibrations

The calibrations to determine the system temperature, sky temperature, effective receiver temperature, and noise diode temperature were performed using an aperture loading technique [3]. This technique uses one microwave absorber (a blackbody source) at ambient temperature, $\approx$ 300 K, and another that is also soaked in a bath of liquid nitrogen, temperature $\approx$ 77 K. The two absorbers were placed in turn over the aperture of the feed for successive measurements of power. This was done by carrying the absorber, a styrofoam container, and liquid nitrogen to the feedcone and manually holding the absorber over the aperture. Noise power measurements were made on each aperture load with the noise diode turned on and off. A power meter was used to measure the IF signal in total power mode, i.e., the radiometer was not Dicke switching. The measurement was done repeatedly for approximately 5 minutes in order to minimize the time that the absorber was in the nitrogen bath for the cold load case. In this way, the effects of ice crystallization in the styrofoam and absorber were reduced. The actual temperature of the ambient load is measured by a thermometer, whereas the cold load was assumed to be 77 K. Typically, eight to ten measurements of the quantities defined below were obtained in the 5 minute period. Before calculating the quantities of interest, the following symbols are defined:

Let

$$Y_1 = \frac{P_h}{P_{sky}} = \frac{\text{Power looking at hot load}}{\text{Power looking at sky}} \tag{8}$$

$$Y_2 = \frac{P_c}{P_{sky}} = \frac{\text{Power looking at cold load}}{\text{Power looking at sky}} \tag{9}$$

$$Y_3 = \frac{P_h}{P_c} = \frac{\text{Power looking at hot load}}{\text{Power looking at cold load}} \tag{10}$$

$$Y_4 = \frac{P_{sky/nd}}{P_{sky}} = \frac{\text{Power looking at sky noise diode on}}{\text{Power looking at sky}} \tag{11}$$

From Eqs. (8) through (11) it can be shown that the following parameters result:

$$T_{op} = \frac{T_h - T_c}{Y_1 - Y_2} \tag{12}$$

$$T_e = \frac{T_h - Y_3 T_c}{Y_3 - 1} \tag{13}$$

$$T_{nd} = T_{op}(Y_4 - 1) \tag{14}$$

$$T_a = T_{op} - T_e \tag{15}$$

where T_h is the temperature of the hot load, T_c is the temperature of the cold load, and T_{nd} is the noise diode temperature at the measurement reference plane. The measured values for these parameters are found in Table 3.

A total of 31 noise diode calibrations were done on three separate days during this observation period. Using Eqs. (12) through (15), it was determined that the noise diode was quite steady regardless of the outside temperature conditions, probably because the XKR feedcone is temperature controlled. The outside temperature varied from 14.5 degrees C to 31 degrees C, while the cone temperature range was only 18.5 degrees C to 22 degrees C. This is mentioned because the noise diode for this experiment was not in an environmentally controlled box or oven that would improve the stability of the noise diode signal. Table 3 shows the normal standard deviation of the 31 measurements for each of the given quantities.

From Tables 1 and 3 it can be seen that there is excellent agreement between the measured value of T_{op} (Table 3) and the predicted value (Table 1). Also, there is excellent agreement between the measured value of T_e (Table 3) and the predicted value (Table 1). The temperature contribution due to the atmosphere is often important. For each of our calibrations, the weather was clear and calm with low humidity. From Eq. (15) T_a can be calculated, which from Eq. (2) allows estimates of T_{atm}. If T_{gal} is assumed to be 3 K, the quadrupod and spillover are assumed to be 6.5 K [4]. Thus, T_{atm} can be estimated to be 6.6 K.

The linearity measurement of the system is a by-product of the noise diode and system temperature calibrations. By injecting the noise diode into the signal path during the hot load, cold load, and sky measurements, the effects of errors due to system linearity may be calculated. This is done by comparing the noise diode value calculated from Eq. (14) using different Y_4 factors; specifically, by using observations made looking at the hot and cold load with and without the noise diode, as is done when observing the sky. Observations indicated the difference in measured noise diode values is small enough to indicate a linear system. It is usually thought that if the linearity error is small, no correction should be performed since the correction is frequently based on a model that may have errors of the same order of magnitude as the linearity error itself.

IV. Radio Source Data

A. Radio Sources

Several criteria were used to select appropriate radio sources for the antenna performance measurements. First, it was very important to observe the radio galaxy Virgo A (3C274) so that the antenna gain measurements could be traced to the Virgo A microwave spectrum, which is consistently referenced in the radio astronomy literature. It is one of the few sources whose spectrum has been calibrated over a wide range of the microwave frequencies (e.g., see Baars et al. [5]). Although the solid angle of the Virgo A source is not especially small compared to the half-power beamwidth of the 70-m antenna operating at 32 GHz, it is by far the most compact of the sources that have been used to calibrate the absolute scale of the microwave flux density spectrum.

A second source selected for gain calibration purposes was Venus, which is an intense radio source at short centimeter wavelengths. The planet's dense atmosphere exhibits a thermal emission spectrum with an equivalent black body temperature of 475 K at 32 GHz [6]. The planet was located near the far side of its orbit during the observations, and this fortuitous location meant that the angular diameter was near its minimum value and much smaller than the antenna beam. Venus also has an advantage in that models of the radio emission spectrum have been computed by applying radiative transfer theory to the physical and chemical properties of the planet's atmosphere. The properties of the atmosphere on a global scale are quite well-known from the in situ data returned by U.S. and Soviet probes and orbiting spacecraft. In addition, years of ground-based astronomical studies have added other critical information to the theoretical work.

Four other sources were included. The variable radio sources 3C84 and 3C273 were observed because they are intense emitters at 32 GHz and they have very small angular diameters (< 0.0005 degrees ≈ 1.8 arcseconds), which makes them very useful for measuring the antenna beam parameters. Both are known to be variable on timescales of weeks or months, so they can only be used for relative antenna gain measurements. The sources 3C286 and P2134+004 also were observed as a consistency check to see if the new flux density results agreed with extrapolations of their published spectra to 32 GHz.

Several other sources that would have been useful are not included in this article. Some were simply not available during the time scheduled for the experiments, and others were eliminated because of time constraints. One of these was the planet Mars, which was not favorably positioned at the time. Two others were DR21 and NCG7027, whose flux densities are thought to be known within reasonable tolerances. In fact, DR21 was observed, but the measurements appeared to be contaminated by one or more nearby sources. DR21 lies in a highly complex region of thermal emission sources in the plane of the galaxy, and the presumption is that the angular separation of the two antenna beams is small enough that the reference antenna beam is detecting neighboring sources of emission. Additional work on this problem is planned.

Table 4 lists the six sources that were used in the calibration and summarizes the measured flux density, S; the associated increase in antenna temperature at the reflector-set angle of 45 degrees, T_s, of the four secondary sources; the source size correction factors, C_r; and the increase in antenna temperature for a 100 percent efficient antenna. This increase in antenna temperature is given by:

$$T_{100} = \frac{SA_p}{2kC_r} \tag{16}$$

where k is the Boltzmann constant, $k = 1.38062 \times 10^{-23} Ws^{-1}K^{-1}$; and A_p is the physical aperture size of the antenna, $A_p = \frac{\pi D^2}{4}$. The efficiency is thus given by

$$\eta = \frac{T_s}{T_{100}} \tag{17}$$

B. Flux Densities

The flux densities for Virgo A and Venus were used for the aperture efficiency measurements. From the microwave spectrum of Virgo A published by Baars et al. [5], the flux density at 32 GHz should be $14.68 J_y \pm 0.8 J_y$ (1σ), where 1 J_y = 1 Jansky = 10^{-26} Wm^{-2} Hz^{-1}. The uncertainty is larger than those usually quoted at longer wavelengths because the spectrum (thought to be linear) must be extrapolated to 32 GHz from 25 GHz, which marks the upper-frequency limit of the absolute measurements that have been published for this source.

The flux density, S, for Venus is calculated from the disk-averaged brightness temperature and the solid angle of the planet. The expression for the flux density, using the Rayleigh-Jeans approximation to the blackbody radiation equation, is

$$S = \frac{(2kT)\Omega}{\lambda^2} \tag{18}$$

where $\lambda = 0.009375$ m is the wavelength of the observations, and $T = 475 \pm 25$ K (1σ) is the brightness tempera-

ture of Venus at 32 GHz [6]. The solid angle of the planet, $\Omega = \pi r^2/d^2$, was calculated with the known distance, d, of Venus for each observation. The effective radius, r, of the thermally emitting atmosphere of Venus is 6120 km [7]. Table 4 shows the flux density of Venus for the two separate days it was observed.

C. Radio Source Size Correction Factors

It is an unfortunate fact that radio sources bright enough for microwave calibration purposes are either time variable or they are partially resolved by the compact beamwidths of modern antennas operating at short centimeter wavelengths. It is possible to correct for the partial resolution of sources if their angular dimensions are smaller than the antenna beamwidth. Only two of the observed sources, Venus and Virgo A, have angular dimensions large enough to require correction. Venus was not a problem because it was so far from Earth at the time that the diameter of the disk was only 0.0027 degree, while the angular beamsize is about 0.009 degree at 32 GHz. A correction factor to account for the partial resolution of the planet's disk by the antenna beam was calculated; the result is 1.03 ± 0.006 (2σ).

Unfortunately, the source structure of Virgo A is neither compact nor very well-known at 32 GHz. Maps of Virgo A have not been published at frequencies above 23 GHz, but relatively good estimates of frequency dependence of the source structure have been made by M. J. Klein. These estimates, derived from maps of the source made at 10 frequencies between 400 MHz and 23,000 MHz, show that the Virgo A size correction for our measurements should be near $C_r = 1.80$. While the structure of the source is known to become more compact at successively higher frequencies, the source structure at 32 GHz is poorly understood. For this reason, we obtained an empirical estimate of C_r by scanning the antenna beam through the source position several times in two orthogonal planes (azimuth and elevation). The procedure was duplicated for the point source 3C273 at very similar azimuth and elevation angles. By comparing the response functions of the scans from the two sources, the value of C_r can be estimated from the apparent beam-broadening produced by the scans across Virgo A, whose angular structure is only slightly smaller than the beam. With the information currently available, the best estimate of the value of C_r is 1.57 ± 0.2 (2σ).

V. Atmospheric Effects

In order to provide antenna characteristics that are independent of the atmospheric attenuation, the atmospheric loss factor needs to be removed from the measurements. One technique (see [3]) to perform this correction uses an efficiency factor given by:

$$L = L_z^{sec(z)} \tag{19}$$

where L_z is the loss factor (L and $L_z \leq 1$) at the zenith direction, and z is the zenith angle, equal to 90 degrees minus the elevation angle. For these calculations, $L_z = -0.193$ dB $= 0.9565$, which is a typical value for the clear atmosphere present during the observations [3]. To determine the efficiency of the antenna for an ideally transparent atmosphere, the measured efficiency is divided by Eq. (19). This technique is sufficiently accurate for the measurements reported here.

VI. 70-Meter Performance Measurement

A. Observations

After several sessions of calibrations and setup work, the observations were performed on May 3, 4, 6, 8, 11, 15, and July 9, 1989 during both daylight and nighttime conditions. During all but one of these observations, there was exceptionally good weather. The sky was clear for most of the observations, and the effects of the high clouds that occasionally appeared were not detectable in the data. On one date, extremely high winds ($\approx$ 17.8 m/sec) precluded accurate calibrations, but useful information about the performance of the antenna in "near stow" wind conditions was obtained. Given the good observation conditions, there is high confidence in the results described below.

B. Results

1. Aperture efficiency. The efficiency as a function of the elevation angle as directly measured with the radiometer is shown in Fig. 4. This figure shows the discrete data that were measured during the observations for all the sources as well as a second-order fit to the data set. Figure 5 shows the efficiency as a function of elevation for the antenna with the effects of the atmosphere removed. Also shown in this figure is a third-order fit to the data which clearly shows the effects of the exponential secant(z) in the lower elevation angles. The aperture efficiency, as calculated using the curve fit at the reflector set angle of 45 degrees, is found to be 35 percent.

For each data point on the figures, the measurement time is on the order of 8 minutes, which is not optimal for a beam-switching radiometer. This time constraint arose due to a known cyclic oscillation in the hour-angle tracking of the 70-meter antenna that causes the beam to be pointed alternately ahead or behind the source by approx-

imately 0.004 degree, which is approximately equal to the 3 dB beamwidth. This oscillation has a period of 4 minutes; therefore, to perform an observation, one has to chart the data for 4 minutes to select the peak signal, and then run the noise diode calibrator. This operation thus requires 8 minutes. This condition is an example of one use of this radiometer: to determine advisable upgrades for future systems. The problem has been extensively studied by the Ground Antenna and Facilities Engineering Section, and solutions will soon be implemented.

2. Beam shape as a function of elevation. The beam shape of the antenna was measured by performing scans in elevation and cross-elevation. The 3 dB beamwidth at the reflector set angle is 0.009 degree. The beamwidth changes as a function of the antenna elevation angle. Wider beamwidths are seen at the lower and higher elevation angles, whereas the narrowest beamwidths are found near the set elevation angle of 45 degrees. For the data shown in Figs. 4 and 5, source size correction factors are used that are constant as a function of the antenna elevation angle; however, some error is incurred because the antenna beamwidth changes as a function of the angle. This is because the correction factor is related to the antenna beamwidth. Further work must be done to quantify the proper function for the corrections due to finite source sizes; this is another reason to calibrate relative performance using secondary very small, but perhaps variable, sources.

3. Pointing. Focus and pointing of the 70-meter antennas of the DSN are achieved by rotating the subreflector to position the system focus at the feed position on the focal ring. This technique allows a multitude of feeds to be used on the same reflector system. However, the feed system of this radiometer is placed at a position where there is no mechanical pin stop for exact positioning of the subreflector, as there is typically. Analysis showed that the likely error in subreflector positioning due to the lack of this stop would yield a beam direction error of approximately 0.002 degree. This guarantees that the source will always be within the main beam of the antenna. Furthermore, the operational technique for positioning the subreflector includes positioning it to 0 degree, setting the pin that exists for that position, resetting the synchro output, and then moving the subreflector to the 32-GHz feed position. In this way, the subreflector always approaches the desired position from the same direction with the same velocity. For each observation, a pointing exercise is performed that locates the beam bias for the remainder of the observation period. Beam "peaking" is continued during the observation period to assure maximum signal. The pointing model that is used for any feed in the XKR feedcone is accurate enough to ensure that 32-GHz sources always remain within the main beam of the antenna, which is an excellent starting position. Unfortunately, the pointing data that are recorded from observation period to observation period are usually not repeatable.

4. Focus. The focus of the reflector changes as a function of elevation, as it does with pointing. However, the focus change is repeatable from day to day. The focus is defined as the subreflector location along the axis normal to the aperture plane of the reflector that maximizes the RF signal. Figure 6 shows a focus curve of signal versus z-axis position at 27 degrees elevation angle.

C. Comparison to Prediction

The measured value of peak efficiency (35 percent at 45.5 degrees elevation) and the measured efficiency with elevation function were compared to predicted values for the 70-meter antenna. With knowledge of the surface roughness of the 70-meter antenna at Goldstone and a "long-wave" efficiency, i.e., efficiency for a perfectly smooth reflector surface, the 32-GHz performance was predicted.[1] This was done using Ruze's equation for gain degradation due to random surface errors [8]. The expected efficiency at the set angle of 45 degrees elevation was 0.313. The measured value, which was based on assumed flux densities, was 0.35 ± 0.5 (2σ). The difference, 0.04, was less than the uncertainty of the measurements.

VII. Error Sources

The repeatability of the data suggests that the precision of the measurements reported here is very high. The noise diode data are tightly clustered about the mean value; the system response to radio sources varies little from observation to observation on a relative scale; and the system temperature and sensitivity are highly repeatable from day to day. In contrast, the absolute accuracy is much less precise. Several factors that might degrade the accuracy have been identified by category: radio source parameters, noise diode calibration, atmospheric attenuation, and instrumentation.

As discussed above, the flux densities of the radio sources are known with uncertainties of approximately 10 percent (2σ). This source of error might explain why the aperture efficiency values used on Venus are so different from those derived from Virgo A. The peak efficiency (near

[1] D. A. Bathker, "DSS 14 70M Efficiencies Above X-Band Frequencies," JPL Interoffice Memorandum No. 3328-89-0109 (internal document), Jet Propulsion Laboratory, Pasadena, California, May 23, 1989.

45 degrees elevation) from Venus is about 0.38, whereas the corresponding value for Virgo A is only 0.32. The mean of the values is 0.35, which represents our best estimate of the peak efficiency.

It is difficult to determine which source might be more reliable for calibration purposes. The flux density of Virgo A may be incorrect and the correction for source structure (C_r) is poorly known. The flux density spectrum is based primarily on observations below 20 GHz [5]. As measurements at higher frequencies are added, the spectral slope of this multi-component radio source is expected to depart from the linear approximation that is currently in use.

Venus, on the other hand, is expected to exhibit a smoothly variable, featureless spectrum at short centimeter wavelengths [6]. However, the temperature of the Venusian atmosphere, on an absolute scale, is not tightly constrained by models calculated from Earth-based and spacecraft data.

Since neither source is preferable, it was assumed that the mean value is the best estimate of the true efficiency. The Venus data were multiplied by 0.92, and the Virgo A data by 1.08 to remove the adverse effects of the discrepancy on the data plotted in Fig. 5.

Another contributor to error is the noise diode used to set the calibration scale. Errors in the calibration of this noise diode level may be present due to several things: the actual temperature of the cold load is not known (it is assumed to be 77 K); the effect of the liquid nitrogen on the system (reflection, absorption, etc.) may not be negligible at this frequency; ice may have formed in or on the styrofoam tub used to hold the absorber and nitrogen; and there are always errors in the measurement of the various Y-factors [Eqs. (8) through (11)]. The random effect of the atmosphere has been minimized using a beamswitching technique, and the zenith attenuation was assumed, not measured. The system linearity is good but not perfect, and the detectors and power meters also introduce errors.

Finally, the attenuation of the signal due to the axial focus of the subreflector affects the measurements. This effect was investigated during the observations by systematically moving the subreflector in the axial direction (Fig. 6). It was found that the autofocus program that is operating on DSS-14 is near correct, and the signal was frequently maximized in order to minimize this error source.

Tables 5 and 6 list contributors to the calibration errors for the noise diode and system efficiency, respectively. The amount of uncertainty for each contributor is an estimate based on typical measurements of this kind on this present undertaking.

An attempt was made to minimize the error sources over which there was control. Diode calibrations were performed in 5 minutes or less to minimize ice formation problems. Thirty-one diode calibrations were performed using different experimenters to obtain a statistical average of the diode level. This made it possible to quantify random errors. The system was fine-tuned to be as linear as possible. Because of these measures and because observations were performed carefully, it is possible that the absolute error in the efficiency measurements is greater than 10 percent (2σ). Given the excellent agreement between the measured and the predicted values of efficiency for this antenna, it is also possible that sensible estimates of the flux density of the radio sources have been obtained, at least for preliminary performance estimating purposes.

VIII. Concluding Remarks

A series of radiometer calibrations and observations on the 70-meter antenna at Goldstone have been performed in order to characterize its performance at 32 GHz. These observations were done using a beamswitching radiometer with an uncooled HEMT, placed in the XKR feedcone. The calibrations show highly repeatable results indicating a high precision. However, several error sources that will affect the absolute accuracy of the measurements are known. It is calculated that this accuracy is on the order of 10 percent, which in any event is probably less than the absolute accuracy of the natural source flux densities.

In order to perform calibrations of higher accuracy, a program is being developed between Caltech and JPL to perform observations of 100 or more radio sources at the Owens Valley Radio Observatory. The combination of a wideband cooled HEMT amplifier on a high-efficiency (1.5-meter clear aperture) gain standard reflector antenna to measure the very strongest sources and a larger 5-meter antenna to measure numerous weaker sources should allow for significantly increased confidence in natural radio source flux densities at 32 GHz. In particular, the apparent discrepancy between the aperture efficiency values based on Venus measurements and those based on Virgo A will be investigated.

The gain standard reflector has demonstrated good agreement between theoretical and measured estimates and will be very accurately calibrated with National Institute of Standards and Technology traceability. This antenna will be used to accurately calibrate the gain of the 5-meter antenna and thus calibrate the flux density of the 100 or more sources.

Acknowledgments

The authors gratefully appreciate the assistance given by Dave Girdner in solving the many pointing problems associated with the 70-meter antenna at 32 GHz. Thanks also to Dudley Neff for the radiometer packaging and to Carl Ellston and Paul Dendrenos for site support and noise diode calibration.

References

[1] R. H. Dicke, "The Measurement of Thermal Radiation at Microwave Frequencies," *The Review of Scientific Instruments*, vol. 17, pp. 268–275, July 1946.

[2] S. D. Slobin, W. V. T. Rusch, C. T. Stelzried, and T. Sato, "Beam Switching Cassegrain Feed System and its Applications to Microwave and Millimeterwave Radioastronomical Observations," *The Review of Scientific Instruments*, vol. 41, no. 3, pp. 439–443, March 1970.

[3] C. T. Stelzried, "The Deep Space Network—Noise Temperature Concepts, Measurements and Performance," JPL Publication 82-33, Jet Propulsion Laboratory, Pasadena, California, September 15, 1982.

[4] R. M. Dickinson, "A Comparison of 8.415-, 32-, and 565646-GHz Deep Space Telemetry Links," JPL Publication 85-71, Jet Propulsion Laboratory, Pasadena, California, p. 31, October 15, 1985.

[5] J. W. M. Baars, R. Genzel, I. I. K. Pauliny-Toth, and A. Witzel, "The Absolute Spectrum of Cas A: An Accurate Flux Density Scale and a Set of Secondary Calibrators," *Astronomy and Astrophysics*, vol. 61, pp. 99–106, 1977.

[6] P. G. Steffes, M. J. Klein, and J. M. Jenkins, "Observations of the Microwave Emission of Venus from 1.3 to 3.6 cm," *Icarus*, 1989 (in press).

[7] D. O. Muhleman, G. S. Orton, and G. L. Berge, "A Model of the Venus Atmosphere from Radio, Radar, and Occultation Observations," *Astrophysical Journal*, vol. 234, pp. 733-745, 1979.

[8] J. Ruze, "Antenna Tolerance Theory—A Review," *Proc. IEEE*, vol. 54, pp. 633–640, April 1966.

Table 1. System temperature contributions for 32-GHz radiometer[a]

Item	Loss,[b] dB	Temperature, K Contribution	Temperature, K at Reference Plane	Comments
1. Waveguide	−0.75	56.55	56.55	Measured
2. LNA	+50.0	281.5	334.56	Measured
3. T_{atm}	—	—	7.87	Estimated[3]
4. $T_{qs} + T_{so}$	—	—	6.5	Estimated [4]
5. T_{gal}	—	—	3.0	Assumed
6. T_f	—	0.84	1.0	Estimated
7. T_a	—	—	17.37	Lines3 + 4 + 5
8. T_e	—	—	392.11	Lines 1 + 2 + 6
9. T_{op}	—	—	409.48	Lines 7 + 8

[a] Reference plane at the aperture of the feed horn.
[b] Positive value means gain; negative value means attenuation.

Table 2. 32-GHz radiometer operating settings

Item	Setting	Comments
Dicke switching rate	20 Hz	Square-wave rate
Post-detection time constant	1 sec	On lock-in amplifier
Pre-detection time constant	30 msec	On lock-in amplifier
Lock-in amp phase for peak	−23 degrees	On lock-in amplifier
IF bandwidth	32 MHz	IF filter
IF center frequency	64 MHz	IF filter
Level set	10.5 dB	Square-law detector

Table 3. 32-GHz system temperature measurements at DSS-14

Item	Temperature, K	Standard Deviation, K
T_{nd}	10.3	0.2
T_a	16.1	0.5
T_e	399.6	13.5
T_{op}	415.7	13.8

Table 4. 32-GHz calibration sources

IAU Designation	Other names	Measured T_s (el = 45 degrees), K	C_r	Flux density S, J_y[a]		$T_{100\ percent}$, K
				Measured[b]	Assumed	
1228 + 126	Virgo A, 3C274	4.39	1.57	***	14.68	13.03
***	Venus (4 May 89)	14.13	1.03	***	26.70	36.13
***	Venus (8 May 89)	14.48	1.03	***	26.90	36.40
0316 + 412	3C84	21.11	1.00	43.71	***	60.92
1226 − 023	3C273	14.27	1.00	29.54	***	41.17
1328 + 307	3C286	0.98	1.00	1.94	***	2.70
2134 + 004	Parks 2134 + 004	1.87	1.00	3.87	***	5.39

[a] Jy = Jansky = $10^{-26} W m^{-2} Hz^{-1}$
[b] Epoch 1989.4

Table 5. Error contributors to the noise diode calibration (all errors are estimated with confidence of 2σ)

Error Contributor	Error Amount	ΔTnd
ΔTh	±1 K	±0.05 K
ΔTc	±1 K	±0.05 K
$\Delta Y4$	±0.1 percent	±0.45 K
LN2 mismatch	±0.1 percent	±0.45 K
Linearity	±2 percent	±0.05 K
RSS error		±0.64 K

Table 6. Error contributors to efficiency (all errors are estimated with confidence of 2σ)

Error contributor	Error Amount		$\Delta\eta$	
	Venus	Virgo A	Venus, percent	Virgo A, percent
Absolute flux	±11 percent	±12 percent	±4.4	±3.7
C_r	±0.6 percent	±7 percent	±0.2	±2.5
Source Measurement	±0.45 Jy	±0.61 Jy	±0.7	±1.5
T_{nd}	±0.64 K	±0.64 K	±2.0	±2.0
T_{atm}	±2 percent	±2 percent	±1.0	±1.0
Z-focus	±1 inch	±1 inch	±0.8	±0.8
RSS error			±5.05	±5.28

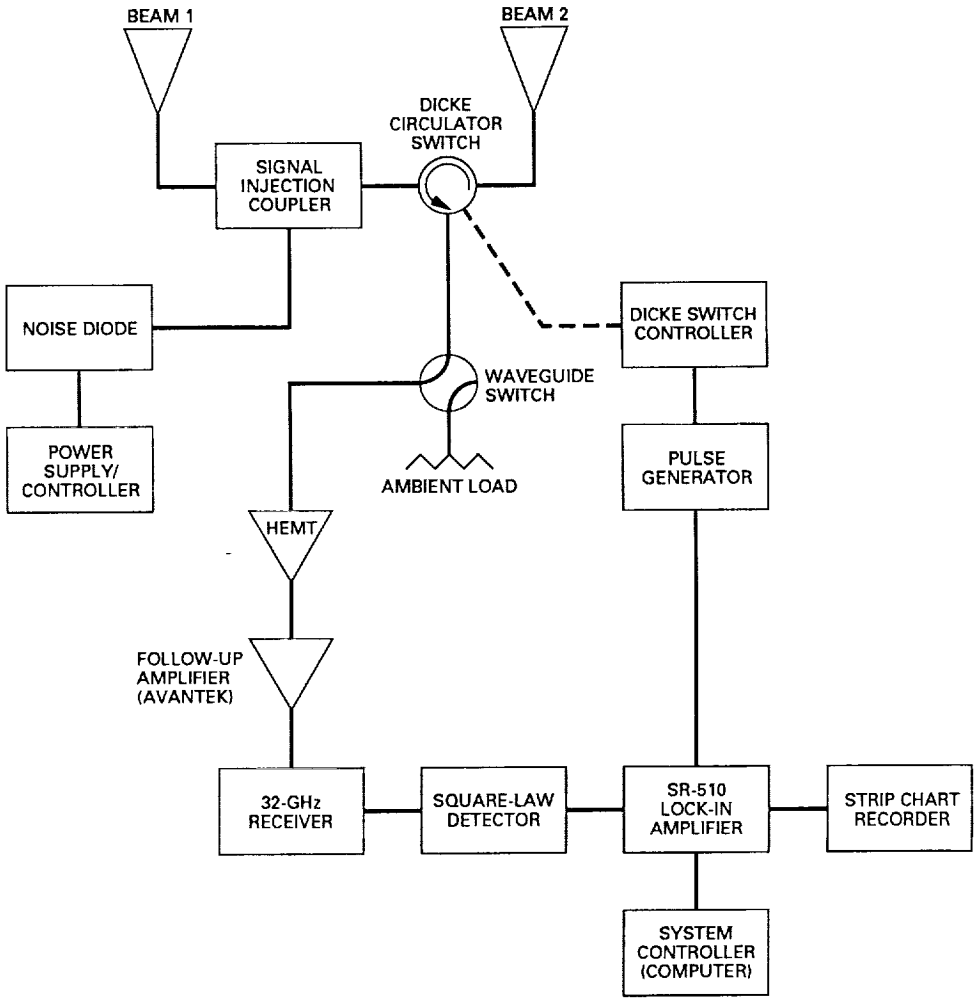


Fig. 1. Block diagram of the 32-GHz beamswitching radiometer.

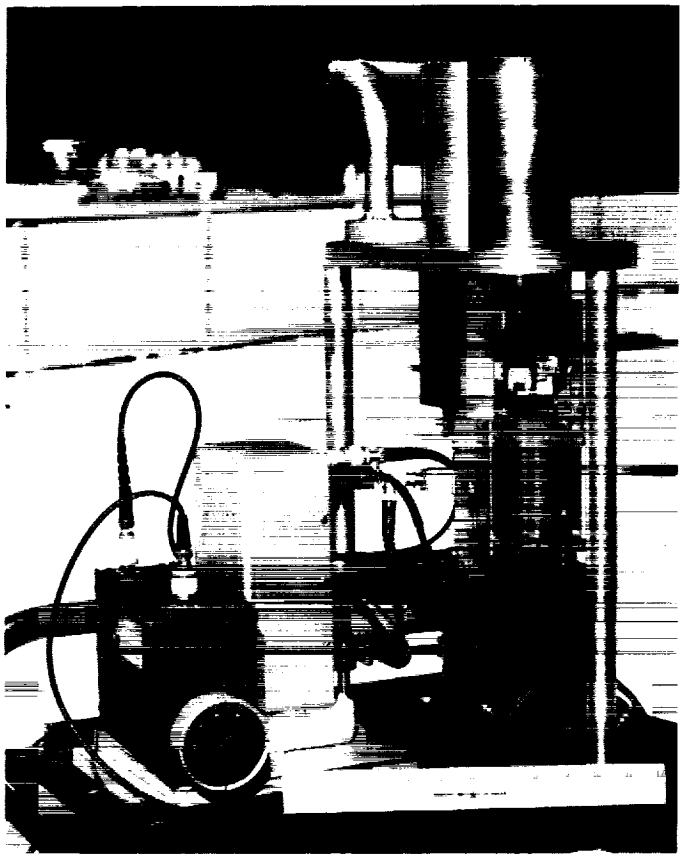


Fig. 2. Dicke radiometer front end.

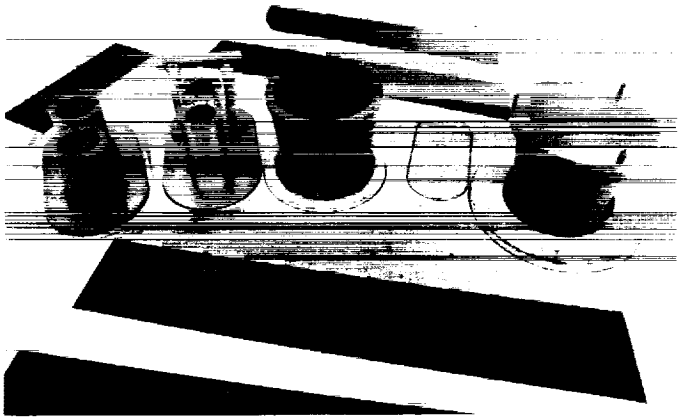


Fig. 3. Top view of XKR feedcone showing various feed systems. (32-GHz radiometer is second from left.)

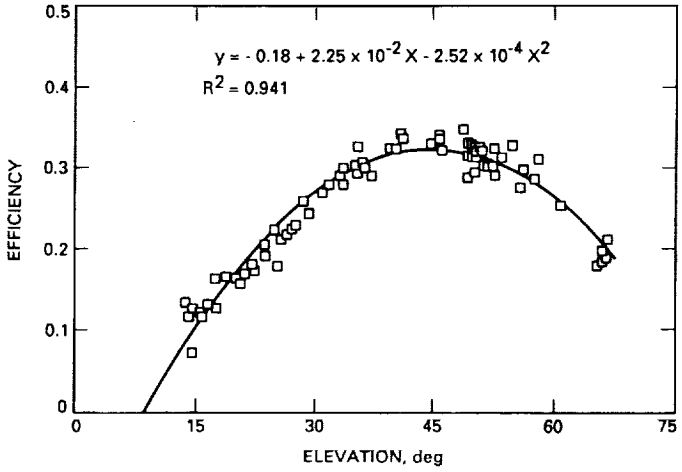


Fig. 4. DSS-14: 32-GHz efficiency including the atmospheric effects (includes second-order fit).

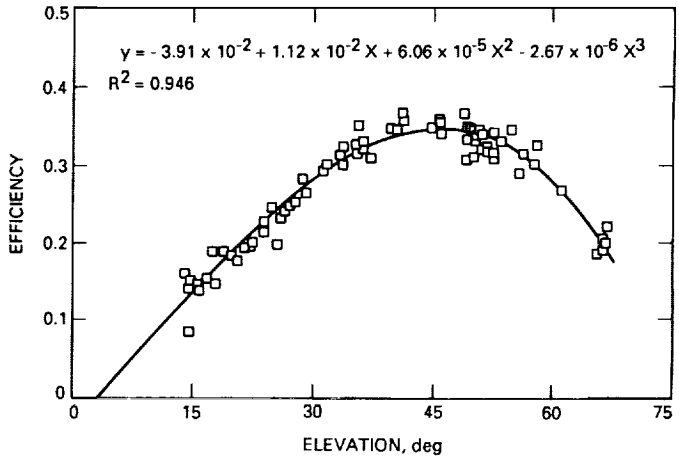


Fig. 5. DSS-14: 32-GHz efficiency; atmospheric effects removed from the data (includes third-order curve fit).

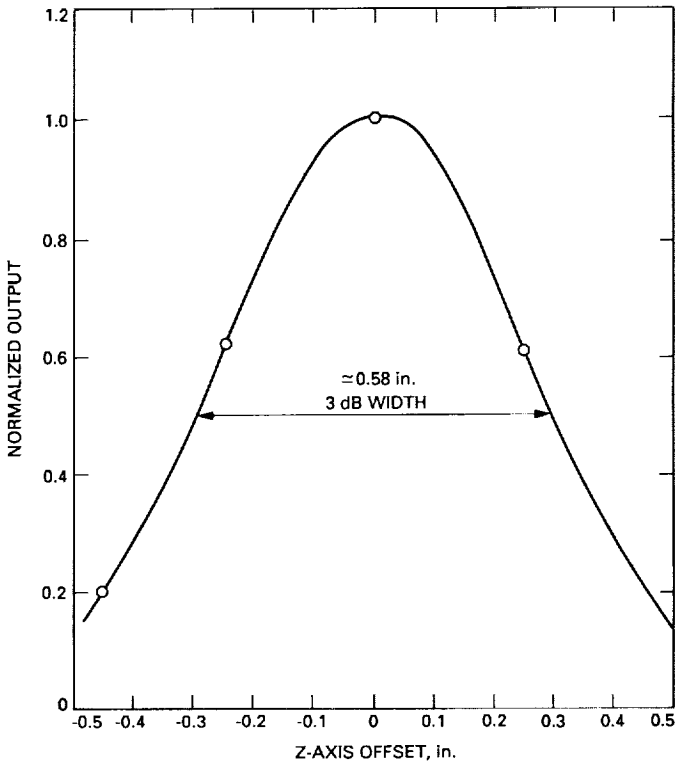


Fig. 6. Z-axis curve. Elevation = 27 degrees. Shows difference from the 70-m standard offset of 2 inches.

Errata

P. H. Richter and S. D. Slobin (Telecommunications Systems Section) have submitted the following errata to their article "DSN 70-Meter Antenna X- and S-Band Calibration Part I: Gain Measurements," that appeared in the *Telecommunications and Data Acquisition Progress Report 42-97*, Vol. January–March 1989, June 15, 1989:

As the result of a somewhat too severe round-off, the coefficients appearing at the top of page 319 do not reproduce the X-band flux density value of $S_0 = 45.2037$ Jy for the radio calibration source Virgo A (3C274) given in the article.

The values given

$a_0 = 3.9964$

$a_1 = 0.1733$

$a_2 = -0.3341$

$a_3 = 0.0352$

should be changed to

$a_0 = 3.9964$

$a_1 = 0.17327$

$a_2 = -0.33415$

$a_3 = 0.035172$

The X-band Virgo A flux density given on page 319 should be rounded to 45.20 Jy and the S-band value of 138.48 Jy listed on page 323 should be reduced to the value 137.93 Jy.

These changes result in less than 0.001 dB of gain increase at X-band and a 0.017-dB gain increase at S-band.