THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 57 May-June 1978 Number 5

COMSTAR Experiment:

An Overview of the Bell Laboratories 19- and 28-GHz COMSTAR Beacon Propagation Experiments

By D. C. COX

(Manuscript received January 10, 1978)

Radio beacons on the COMSTAR communication satellites transmit continuously at 19 and 28 GHz to permit the long-term measurement of the properties of earth-space propagation needed in designing future high-capacity satellite communication systems. An extensive receiving facility has been established at Crawford Hill, New Jersey, for measuring attenuation, depolarization, coherence bandwidth and scatter of the beacon signals by atmospheric processes. The facility includes a precision 7-meter antenna and multichannel receiving electronics designed to obtain optimum benefit from the COMSTAR beacons. Other Bell Laboratories receiving facilities in Georgia and Illinois are accumulating statistics on signal attenuation and diversity improvements for other climatic conditions.

I. INTRODUCTION

Propagation experiments using the 19- and 28-GHz beacons on COMSTAR satellites represent a significant milestone in the quest for design information for future satellite communication systems. The experimental results also contribute to knowledge of meteorological processes. Earth-space propagation information above 10 GHz is a key component in the exploration of concepts for high-capacity domestic satellite communication systems.[1,2]

Future high-capacity satellite communication systems will probably use frequencies above 10 GHz because of the large segment of unused frequency spectrum available and because of spectrum crowding at the lower frequencies. In particular, frequency bands over 2000 MHz wide are allocated at 19 and 28 GHz, and 500 MHz bands are available at 12 and 14 GHz for common-carrier satellite communication systems; at 4 and 6 GHz the allocated frequency bands are only 500 MHz wide. Even now satellite systems at 4 and 6 GHz are severely constrained by requirements for avoiding interference with the extensive terrestrial radio networks that share these same frequency bands.

New systems will probably reuse frequencies on two orthogonal polarizations to double the usable bandwidth and may reuse frequencies among spot antenna beams covering small areas of high traffic concentration. Thus, knowledge of the decrease in cross-polarization isolation produced by rain and ice is needed and satellite antenna sidelobe control will be important. Sidelobe levels of earth station antennas and the scattering of energy from one antenna beam to another by rain will limit how close communication satellites can be spaced in the geosynchronous orbit. Other characteristics of future systems operating above 10 GHz compared to present 4 and 6 GHz systems are as follows: (*i*) they most likely will use smaller earth station antennas; (*ii*) they will experience more rain attenuation and may use transmitter power control or site diversity to cope with it; (*iii*) interaction with the ionosphere will be significantly less; (*iv*) they may use wider bandwidths and thus be more sensitive to delay dispersion, and (*v*) some systems probably will use digital modulation. Some system calculations that were used as a guide in selecting representative values of experimental parameters, such as earth station antenna size, are presented by Tillotson in Ref. 1.

It is well known that rain affects radio propagation more severely at frequencies above 10 GHz.[2] However, present knowledge of rainstorm characteristics and atmospheric processes is not adequate for predicting all earth-space propagation characteristics from terrestrial propagation measurements, surface rain measurements, or radar measurements. Thus, the complete information needed for satellite communication system design can be obtained only from measurements made along earth-space paths. The complication and expense of placing radio sources in synchronous orbit has meant that continuously transmitting beacons were not available in the 1960s for measuring earth-space propagation characteristics. Advantage was taken of a natural extraterrestrial radio source, the sun,[3] and of thermal emission from rain itself[3,4] for indirectly measuring the attenuation of earth-space radio signals above 10 GHz. The sun trackers and radiometers used for these measurements provided the only continuous attenuation data available in the late 1960s and early 1970s.[5] This data base has several limitations. The sun azimuth and elevation change continuously during the day and, of course, the sun

is not available during the night. Radiometer measurements of thermal emission from rain have a range restricted to about 12 dB because of rain scatter and uncertainty in effective rain temperature.[3,4] Also, these indirect measurement techniques cannot provide any information on depolarization or on bandwidth limitations imposed by the propagation medium. The NASA experimental ATS-5 and ATS-6 satellites carried beacons transmitting at 15 GHz and at 20 and 30 GHz, respectively. These beacons did not transmit continuously and were often not available during rainstorms because of scheduling and total power constraints on the satellites. Therefore, continuous long term statistics could not be collected using these sources. In fact, availability of the ATS-5 and -6 beacons was so limited that a significant depolarizing phenomenon went undetected until continuously transmitting satellite beacons became available.[6] Propagation information is needed for the satellite bands at 12, 14, 19, and 28 GHz; however, expense, power and weight limitations restrict the number of satellite beacon sources that can be provided reasonably. Beacons transmitting within the 19- and 28-GHz bands were a logical choice for sources for earth-space propagation experiments because attenuation and other atmospheric interaction is greater at the higher frequencies, extrapolation is reasonable from measurements made at two frequencies, and it is safer to extrapolate from large numbers to small numbers. The 19- and 28-GHz beacons were put on the COMSTAR satellites then to satisfy the fundamental need for continuous long-term measurements of propagation parameters such as attenuation, depolarization, coherence bandwidth and differential phase.

In support of this measurement effort, an extensive receiving facility has been established at Crawford Hill, New Jersey.[7] The facility is described in Section III of this paper and in the next two papers.[8,9] of this issue. An interim receiving facility at Crawford Hill is described in the fourth paper.[10] Bell Laboratories receiving facilities have been established near Atlanta, Georgia (Palmetto), and near Chicago, Illinois (Grant Park), for accumulating attenuation and diversity statistics for other climatic conditions. These measurements are described in the fifth paper.[11] Other experimenters outside Bell Laboratories are also making use of the beacons for accumulating propagation information.[19,20] This paper describes the COMSTAR beacons and the experimental measurements at Crawford Hill.

II. THE COMSTAR 19- AND 28-GHz BEACONS

The need for measurements in different climatic regions of the U.S. along with the need for continuous measurements suggest that satellite beacons providing signals for propagation experiments should have U.S. coverage antennas. Such antennas provide about 30-dB gain. Beacons placed on operational satellites such as COMSTAR logically can make use

Table I — COMSTAR satellite beacon parameters

Station-keeping tolerance		$<\pm 0.1°$ E-W and N-S
Antenna pointing tolerance		$<\pm 0.1°$
Carrier frequencies		19.0400 GHz and 28.5600 GHz
Variation:	Diurnal	$<\pm 1 \times 10^{-6}$
	Aging	$<\pm 1 \times 10^{-6}$ per year
	Maximum Rate	$<\pm 5 \times 10^{-11}$ per second
	Jitter	90% of carrier power in bandwidth $<\pm 8 \times 10^{-10}$
EIRP		
19 GHz		>+52 dBm per polarization
28 GHz		>+56 dBm
Variation*:	Diurnal	<±0.3 dB
	Aging	<±0.5 dB per year
	Maximum Rate	<±0.1 dB per minute
Polarization		
19 GHz		Switched between two orthogonal linear
28 GHz		Linear, aligned with most nearly vertical 19-GHz signal
Orientation†		4° for satellite at 128°W 21° for satellite at 95°W
Crosspolarized components‡		>32 dB below copolarized level at worst case within U.S.
Polarization switch (19 GHz only)		
Rate		1000.0 Hz ± 0.1 Hz
Stability		$<1 \times 10^{-7}$ per 10 min.
Switching time		<10 us
Asymmetry		<±5%
Phase modulation (28 GHz only)		Coherent with carrier
Frequency**		264.4 MHz
Sideband level		<7 dB below carrier

* A circuit malfunction has reduced the power output of the 19 GHz beacon at 128°W by 2 dB since launch.

† Polarization orientation is the angle (<45°) that the received polarization is rotated from vertical or horizontal at Crawford Hill.

‡ >36 dB and >41 dB below 19- and 28-GHz copolarized levels, respectively, toward Crawford Hill.

** A satellite scheduled for a later launch has a modulation frequency of 528.9 MHz.

of the excess power generated by solar cells at the beginning of satellite life. Since this power is limited, low beacon power is desirable; for COMSTAR, about +30 dBm each from the 19- and 28-GHz beacons could be produced without having a detrimental effect on the primary communication mission of the satellite. Since a reliable measuring range of over 30 dB is desirable with simple, relatively inexpensive earth stations and since a range of 60 to 70 dB at 30 GHz is desirable for extrapolating attenuation data to lower frequencies, receiving system signal margin must be obtained by narrowing receiver noise bandwidth. Narrow-band measuring systems, however, require very stable oscillators. These considerations, along with the needs for measuring depolarization and delay dispersion, resulted in the COMSTAR beacons with parameters summarized in Table I.

The spin stabilized COMSTAR satellites illustrated in Fig. 1 were built by Hughes Aircraft Corporation and are owned and controlled in orbit by COMSAT General Corporation. They are leased to the AT&T and

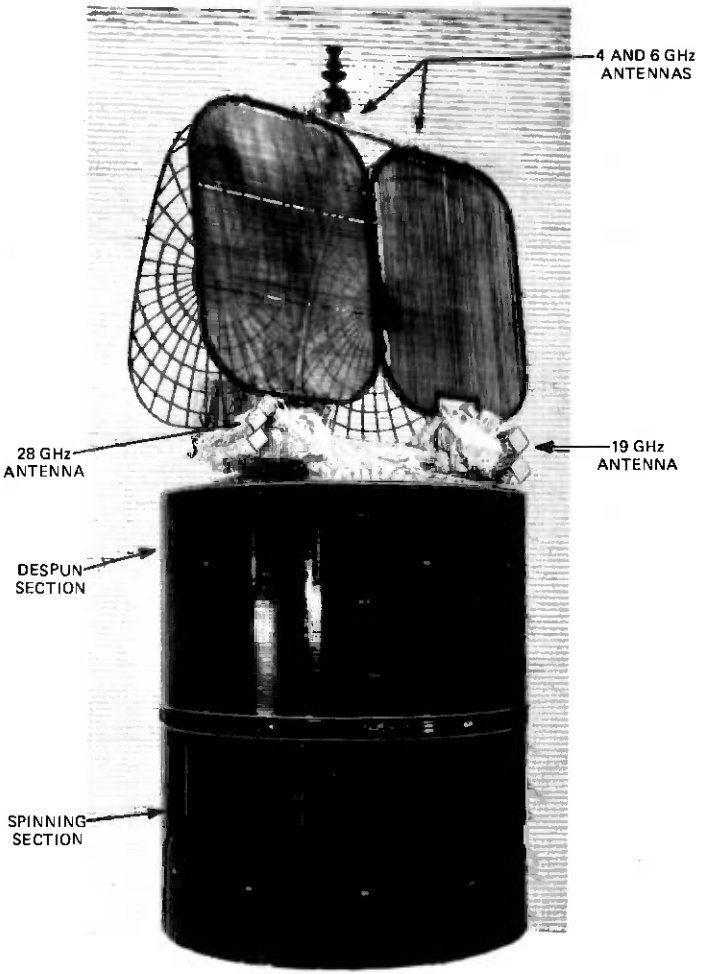


Fig. 1—COMSTAR satellite. Lower solar-cell-covered cylindrical portion is 9 feet high by 8 feet diameter. Overall height is 20 feet.

GT&E Companies for domestic U.S. communication service.[12] The lower half of the solar-cell-covered drum spins at between 50 and 60 revolutions/minute while the upper half, which supports the antenna platform, is despun to keep the antennas pointed towards the earth. The two large antennas are used by the 4- and 6-GHz communications system, one for each of two orthogonal linear polarizations. The 19- and 28-GHz beacon antennas are on opposite sides of the satellite, just below the communication antennas.

There are to be three beacon-equipped COMSTAR satellites in synchronous orbit. The first satellite was launched in May 1976 and is at

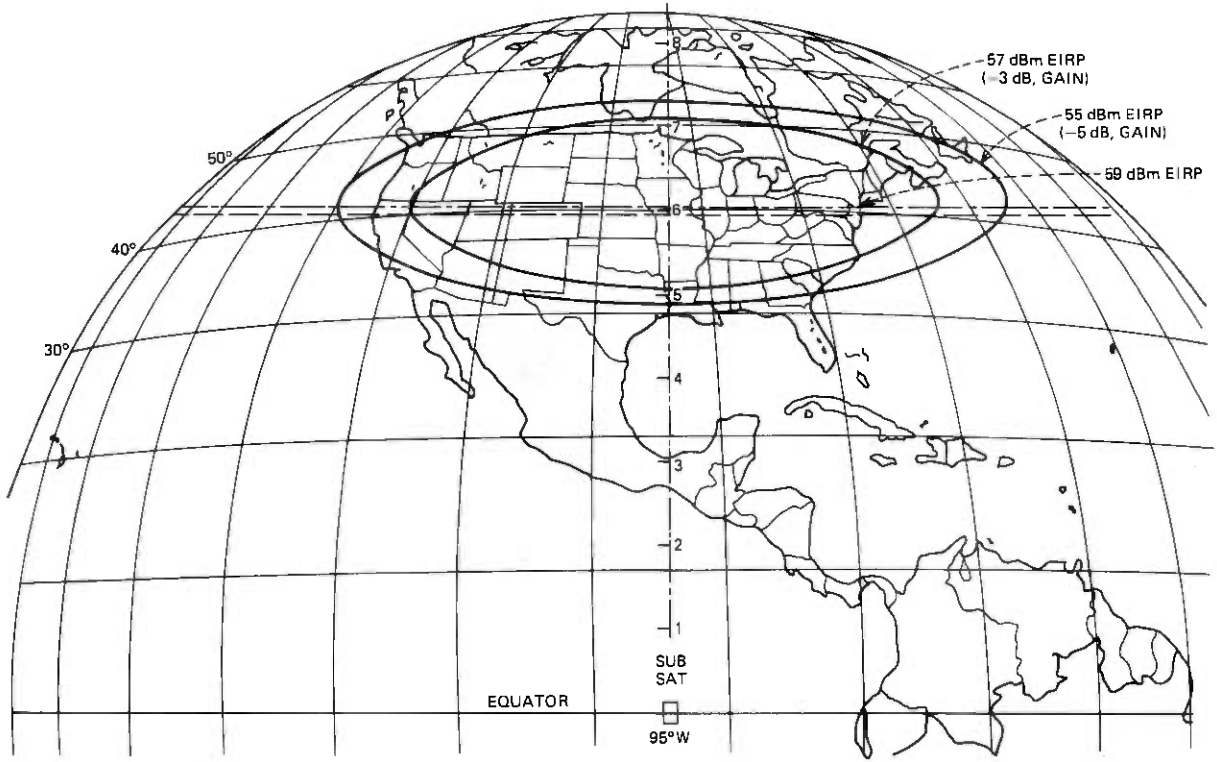


Fig. 2—28-GHz beacon coverage from satellite at 95°W longitude. Maximum antenna gain is 30 dB. Coverage for 19-GHz beacon is similar.

128°W longitude; the second was launched in July and is at 95°W; the third is to be launched in the spring of 1978.

The beacons were built by COMSAT Laboratories and are described in more detail in Ref. 13. The 19.04- and 28.56-GHz beacon signals and the 264.4-MHz signal that phase-modulates the 28.56-GHz signal are derived from a common 132.2-MHz quartz crystal oscillator. A common frequency multiplier chain multiplies the oscillator output to 2.38 GHz. The signal path splits into separate frequency multipliers after amplification at 2.38 GHz. The 28.56-GHz multiplier output is phase-modulated before it and the 19-GHz multiplier output are amplified in separate negative-resistance IMPATT amplifiers. Polarization is switched (19 GHz only) with PIN diodes driven from a separate crystal oscillator and frequency divider. The beacons make use of the surplus dc power capacity of the solar-cell array that is available until the solar cells deterioriate under the radiation environment of space. The projected availability of this surplus power is over two years (satellite lifetime is greater than seven years). The expected lifetime of the all-solid-state beacons is considerably greater than 7 years; however, due to a component failure the 19-GHZ beacon on the 128°W satellite has experienced regular daily power fluctuations since July 1976.

As shown in Fig. 1, a two-horn antenna array transmits the linearly polarized 28-GHz signal. Another two-horn array transmits two orthogonal linearly polarized 19-GHz signals. These two-horn arrays permit control of the polarization independent of the radiation pattern; polarization is determined by rotation of the horn apertures relative to the array centerline* while the pattern is determined largely by the array separation, the aperture dimension perpendicular to the array centerline, and the orientation of the centerline. The same antenna is used for both 19-GHz polarizations so that the received differential phase will not be affected by angular motion of the satellite.† If two antennas with separate phase centers were used, unavoidable residual angular motion of the satellite (~0.1°) would produce differential phase shift as in an interferometer. Although ±0.1° is a small variation, it would cause up to ±10° differential phase shift at 19 GHz between two antennas with phase centers placed as close together as possible. Figure 2 is a beacon antenna-coverage pattern constructed from pre-launch antenna-range measurements. Antenna patterns for the other frequency and polarization are similar.

* The beacon at 128°W was designed to produce horizontal and vertical polarization at Crawford Hill with the satellite at 119°W. The beacon at 95°W was designed to produce horizontal and vertical polarization at Crawford Hill with the satellite at 129°W. Orbital assignments by the Federal Communications Commission (FCC) and launch scheduling did not permit placing of the satellites in their requested orbital locations.

† M. J. Gans of Bell Laboratories collaborated with engineers at Hughes Aircraft in adapting the two-horn array for the two polarization application.

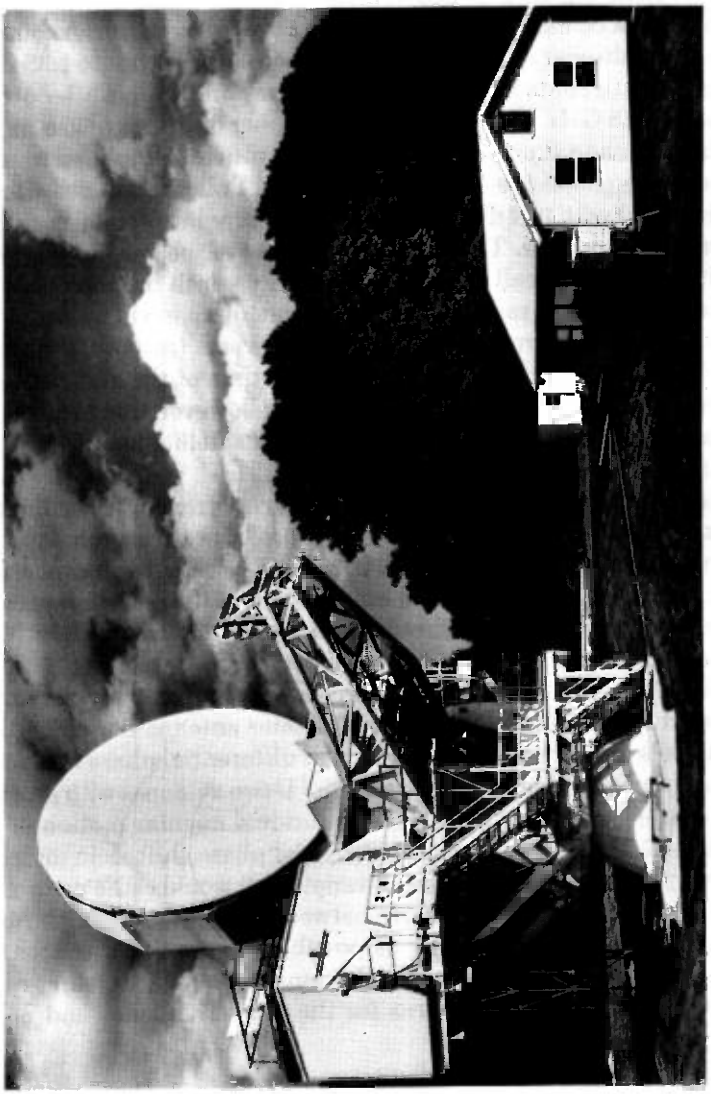


Fig. 3—Crawford Hill main receiving facility showing 7-meter diameter antenna and control building.

III. THE CRAWFORD HILL RECEIVING FACILITY

The Crawford Hill receiving facility was designed to obtain optimum benefit from the COMSTAR beacon emissions[7] while also demonstrating techniques applicable to 19- and 28-GHz earth stations. The facility is located on top of Crawford Hill, New Jersey, about 50 km south-southeast of New York City at 40.392° north latitude and 74.187° west longitude. Crawford Hill is 115 meters above sea level. The main facility shown in Fig. 3 comprises the 7-meter diameter millimeter-wave antenna, receiving electronics, antenna pointing and data collection equipment, and computer programs indicated in the simplified block diagram in Fig. 4. The antenna, antenna feed and receiver front ends could function as communication system components. For example, the antenna size and gain are representative of the needs of typical high-capacity earth stations.[1] The antenna does not have any aperture blockage because of the offset geometry. This clean aperture, combined with a good surface and heavily tapered illumination, result in the low sidelobe levels that would be required for systems working with satellites closely spaced (~1°) in orbit. These low sidelobe levels are also needed in the propagation experiment for measuring the crosstalk produced by the scattering of radio waves by rain. The good surface, long effective focal length and feed design result in the low antenna cross polarization throughout the antenna beam needed by frequency reuse dual polarization systems and required in the experiment for measuring depolarization by raindrops and ice crystals. The simple format for obtaining open loop antenna pointing information from COMSAT also should prove useful in future system operation.[16] The dual-frequency dual-polarization antenna feed uses low loss quasi-optical frequency and polarization diplexers suitable for diplexing high-power transmitters with sensitive receivers. The receiver front ends are low-noise broadband mixers and broadband IF amplifiers that provide a 1-GHz-wide channel suitable for system use from RF through first IF. Following the first IF, the receiver is narrow band and optimized to provide the sensitivity and stability needed for the propagation measurements.

An interim receiving facility shown in Fig. 5 was used before completion of the main facility. Measurements at the interim facility are continuing using the beacon at 128°W. From Crawford Hill the beacon at 95°W is at azimuth = 210.5° and elevation = 38.6°; the beacon at 128°W is at azimuth = 244.7° and elevation = 18.5°. The main receiving facility is briefly described in the following sections. More detailed descriptions of this equipment and a description of the interim facility are contained in the following papers of this issue.[8,9,10]

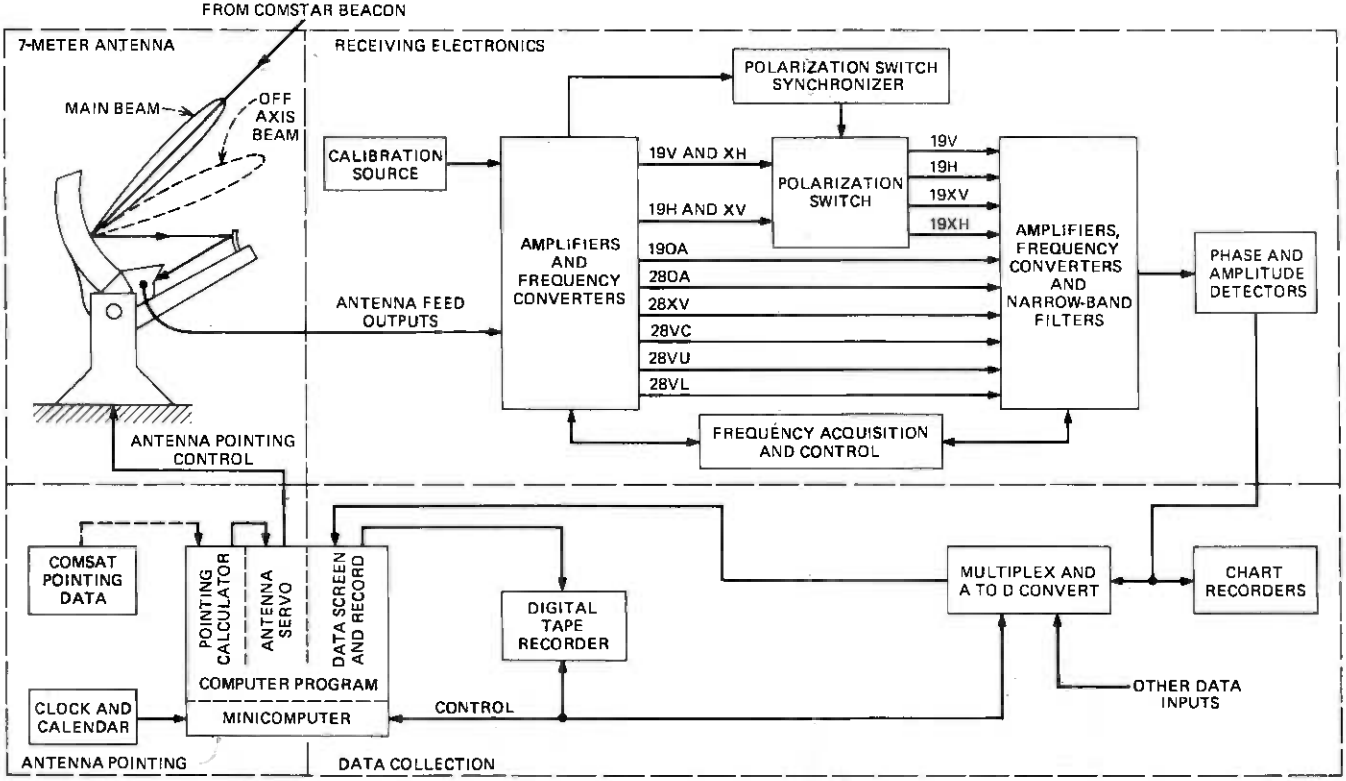


Fig. 4—Configuration of main receiving facility at Crawford Hill comprising 7-meter antenna, receiving electronics, data collection subsystem and antenna pointing components described in the text. Facility simutaneously records COMSTAR beacon signal parameters for two polarizations at 19 and 28 GHz and signals scattered off-axis from the main beam at 19 and 28 GHz.

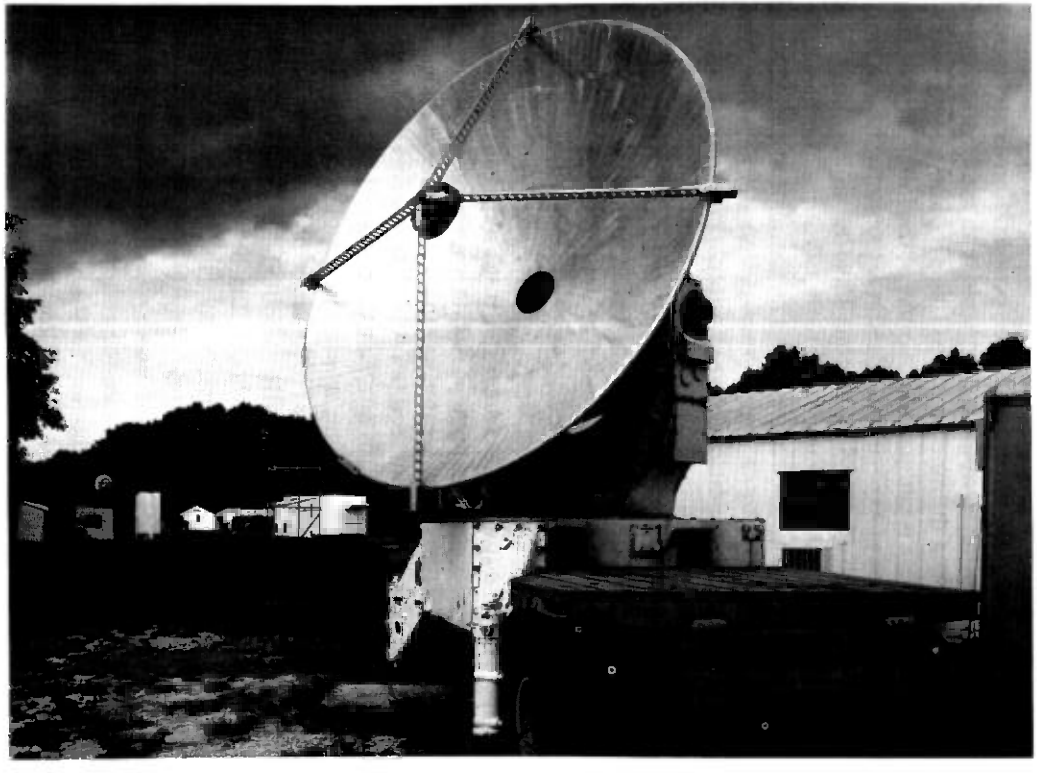


Fig. 5—Crawford Hill interim receiving facility showing 3.7-meter-diameter antenna and instrumentation building. The facility described in Ref. 10 receives beacon signals on two polarizations at 19 GHz.

3.1 7-meter antenna and antenna pointing

The 7-meter diameter millimeter wave antenna[8] shown in Fig. 3 is of novel design and extreme precision. It is designed to operate in the high winds and generally bad weather that are characteristic of summer thunderstorms (and hurricanes). These bad weather events produce most of the propagation degradation such as attenuation and depolarization that are measured in the continuing propagation experiments. The antenna and pointing equipment are shared with radio astronomers who use the facility as a millimeter wave (100 GHz and above) radiotelescope. This joint use is compatible because of the inherent requirements of the two types of experiments. Radio astronomy observations require clear atmospheric conditions that have minimum effect on millimeter wave propagation. These conditions, of course, produce little effect on the satellite beacon signals and thus no useful propagation data. When atmospheric conditions cause propagation degradation that should be measured, the conditions are not suitable for radio astronomy observations.

The millimeter wave offset Cassegrainian antenna consists of a 7-meter diameter parabolic reflector on an altazimuth mount; a convex hyperbolic subreflector; dual-frequency (19 and 28 GHz) dual-polarization quasi-optical main-beam feeds; dual-frequency (19 and 28 GHz) single-polarization off-axis feeds; drive motors, angle encoders, and pointing servos; two equipment rooms; and feeds at other frequencies for radio astronomy. The antenna does not have a radome cover.

All 19- and 28-GHz feeds are located in a small equipment room, the vertex cab, at the Cassegrainian feed point near the vertex of the main reflector, above both the azimuth and elevation axes. The main beam feeds receive simultaneously two linear orthogonal polarizations at both 19.04 GHz and 28.56 GHz; these feeds are mounted on a single frame that permits rotation of the feed polarization about the coaxial main beam axis without affecting polarization orthogonality or beam pointing. The off-axis feeds form a beam that is pointed away from the axis of the main beam. The 0.002-inch Mylar feed window in front of the vertex cab is covered with a nonwetting coating and is nearly vertical to prevent water from collecting on it. The second, somewhat larger, equipment room, the side cab, is to one side of the antenna, on the elevation axis but above the azimuth axis. The antenna has the following characteristics:

Gain	19 GHz	61 dB
	28 GHz	64 dB
Beamwidth:	19 GHz	0.17°
	28 GHz	0.11°
Cross-polarization isolation:	>35 dB throughout main beam	

Sidelobes:	<−40 dB at >1° off-axis	
Pointing error:	<0.0028°	20 mph steady wind
	<0.017°	45 mph steady wind + gusts to 60 mph
	0.001°	angle readout
Slew rates:	2°/sec azimuth; 1°/sec elevation	warm weather

The antenna is fully steerable by the drive system and servo over an elevation of 0° to 90° and an azimuth of 0° to 450° with the azimuth rotation limited by the wrap-up of cables (no slip rings are used).

The antenna is pointed by a minicomputer that compares angle encoder outputs with command pointing angles and calculates velocity commands for the drive system from the resulting error. The commanded pointing angles for the COMSTAR satellites are calculated by the computer from parameters derived from orbital predictions.[16] COMSAT General supplies via teletype a set of coefficients for equations describing antenna azimuth and elevation angles at Crawford Hill. These equations result in a pointing error of less than ±0.008° when updated every 2 weeks, assuming perfect orbital prediction. The antenna follows the ±0.1° diurnal motion of the satellite within about ±0.01°.

3.2 Receiving electronics

The propagation experiments placed strong demands on technology for the receiving electronics.[9] Continuous unattended operation is required so that all significant weather events are included in the resulting data base; thus, a very high degree of reliability is necessary and rapid automatic reacquisition of the beacon signal after dropout due to severe attenuation or momentary power outage is essential. Since relative phases of the many signal components must be precisely measured, the phase stability of all circuits and components demanded careful attention. Also, circuit arrangements had to be devised to ensure that signals to be compared in phase traverse a common path through high-gain amplifiers and other phase-sensitive equipment.

In order to obtain the maximum possible measuring range using the modest power radiated by the satellite beacons, very narrow receiver noise bandwidths are required. This in turn requires excellent stability in the source oscillators in the satellites and the local oscillators in the earth stations. The receiver includes an automatic frequency control circuit with built-in memory to facilitate reacquisition after loss of signal.* Maximum use was made of known correlations among strong and

* The feature also permits easy return to propagation measurements after use of the antenna during clear weather periods for radio astronomy observations.

weak signal components to permit detection of weak cross-polarized signals during severe fading.

The receiving electronics portion of Fig. 4 is subdivided to indicate receiving functions. This part of the facility includes the balanced mixers (frequency converters), oscillators, frequency multipliers, IF amplifiers, bandpass filters, switches, and envelope (amplitude) and phase detectors required to process the following signals: (*i*) the nearly vertically and horizontally polarized 19-GHz main beam signals, 19V and 19H, (*ii*) the corresponding crosspolarized 19-GHz main-beam signal components, 19XV and 19XH, (*iii*) the 19- and 28-GHz off-axis beam signals, 19OA and 28OA, (*iv*) the nearly vertically polarized 28-GHz main-beam carrier, 28VC, upper sideband, 28VU, and lower sideband, 28VL, and (*v*) the corresponding cross-polarized 28-GHz main-beam carrier, 28XV. Included in these functional blocks are the circuits for automatic frequency control, frequency acquisition, polarization switch synchronization and receiving system calibration. The receiving electronics are distributed among the two equipment rooms on the 7-meter antenna and the control building about 15 meters away from the antenna. The equipment distribution optimizes noise performance and phase and amplitude stability while staying within space limitations in the various equipment rooms. Since power line transients and momentary power outages are expected during heavy rain, all oscillators, filter stabilizing ovens and frequency memory registers are powered by batteries charged continuously from the power line. The receiving electronics have the following characteristics:

	19 GHz		28 GHz		
	V,H	XV,XH,OA	VC,XV	VU,VL	OA
Noise figure	≤7 dB		≤7 dB		
Noise bandwidth	16 Hz	16 Hz and 1.6 Hz	24 Hz and 2.4 Hz	24 Hz	2.4 Hz
Channel-to-channel isolation	>65 dB			>68 dB	
Amplitude instability	<0.5 dB			<0.5 dB	
Phase instability	<2°			<5°	
Frequency tracking rate	2 Hz/sec			3 Hz/sec	

Since the 19- and 28-GHz beacon signals are derived from a common oscillator, they have the same frequency fluctuations. Thus, extended measuring range is provided in the 28-GHz channels and in 19-GHz low-signal channels (off-axis and cross-polarization) by: (*i*) using common frequency sources for corresponding 28-GHz and 19-GHz receiver local oscillators (LOs), (*ii*) tracking out frequency fluctuations in the

Table II — Radio link parameters for Crawford Hill, New Jersey

	19.04 GHz		28.56 GHz	
Beacon (95°W) output power	+27.7 dBm		+30.6 dBm	
Polarization switching	−3		—	
Satellite antenna gain	+28.6 dB		+28.2 dB	
EIRP per polarization (average)	+53.3 dBm		+58.8 dBm	
Path loss 22,300 miles	−209.7 dB		−213.2 dB	
Crawford Hill antenna gain	+61 dB		+64 dB	
Clear air absorption (O_2 and H_2O)	−0.6 dB		−0.9 dB	
Signal into receiver (avg. per polarization)	−96 dBm		−91 dBm	
Noise into receiver (clear air)	−156 dBm	−166 dBm	−154 dBm	−164 dBm
Noise bandwidth* (7 dB NF)	(16 Hz)	(1.6 Hz)	(24 Hz)	(2.4 Hz)
Receiver carrier-to-noise (C/N) (clear air with noise blanking)	+60 dB (16 Hz)	+70 dB (1.6 Hz)	+63 dB (24 Hz)	+73 dB (2.4 Hz)
Oscillator stability [PLL off (below thresh.)]	−0.2 dB	—	—	—
Measurement range (in rain) (to $C = N$) (ant. temp. 273°)	59 dB	69 dB	62 dB	72 dB

* A single complex pole pair bandpass filter with 10-Hz 3-dB bandwidth has a noise bandwidth of 16 Hz.

beacon and in LOs with a common oscillator in a loop locked in phase to the 19-GHz vertically polarized signal, the signal that experiences the least attenuation, and (*iii*) using very narrow-band filters in the extended range channels.

Differential amplitude and phase stability is maintained by carefully controlling differential temperature between corresponding components in different receiver channels, by using low temperature coefficient components, by choice of IF frequencies and filter bandwidths, and by designing for good circuit linearity.

3.3 Radio link parameters

Radio link parameters affecting dynamic measuring range are summarized in Table II for the beacon at 95°W and the Crawford Hill 7-meter antenna and receiving electronics. The major contributors to the 19-GHz parameters are illustrated in Fig. 6.

Filters with 1.6-, 24-, and 2.4-Hz noise bandwidths are in the channels that receive signals attenuated by rain to lower levels than the vertically polarized (V) 19-GHz signal. The 1-dB difference between clear air carrier-to-noise ratio and measuring range reflects the difference in antenna temperature between clear air and rain.

Refraction in the atmosphere due to temperature and humidity gradients does not have a significant effect on the experiment.[14] For example, the maximum expected surface refractive index variation at Crawford Hill is $<\pm 45$ ppm for August, the month of maximum spread in refractive index. This refractivity variation corresponds to a $\pm 0.003°$ elevation angle variation for the 38.6° elevation from Crawford Hill to

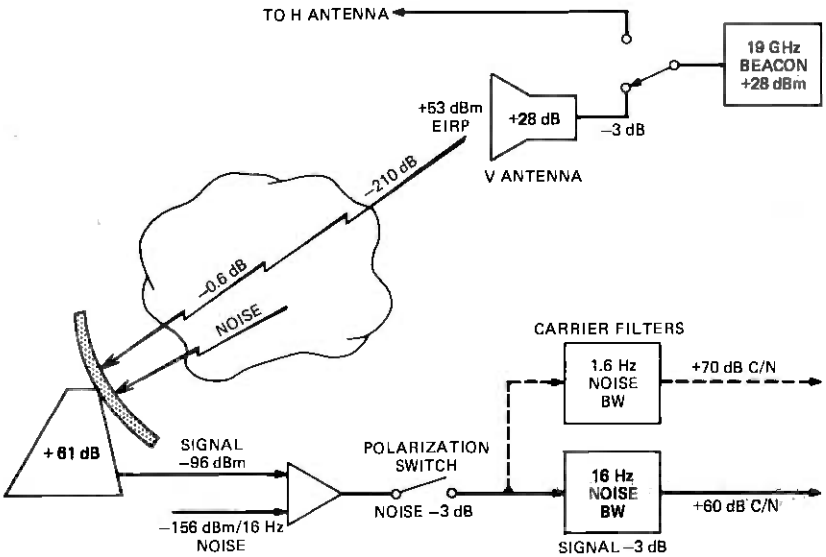


Fig. 6—Radio link summary indicating major parameters that determine the carrier-to-noise ratio of the 19-GHz COMSTAR beacon signals received in clear air with the Crawford Hill main receiving facility.

the satellite at 95°W. The elevation variation would be $<\pm 0.008°$ for the 18.5° elevation to the 128°W satellite. Such angular variations produce $<\pm 0.05$ dB amplitude variation for the 7-meter antenna and insignificant change in differential phase and cross polarization.

Faraday rotation of linear polarization by the ionosphere is insignificant at 19 and 28 GHz.[14,15] The maximum expected variation in polarization angle is $<\pm 0.15°$ for the extremes of solar flares or magnetic storms. During normal ionospheric conditions the variation is significantly less. If the cross-polarization of the instrumentation system (antennas, beacon polarization switch and receiver) were zero, the rotation variation of $\pm 0.15°$ would produce a cross-polarized component of <-50 dB. Cross-polarization variation of this magnitude is insignificant in the experiment.

3.4 Data collection

The data collection equipment is common to all receiving channels as indicated in Fig. 4. Data that are critical for maintaining continuity in the data base for long term statistics, e.g., signal attenuation and depolarization, are recorded continuously on analog ink-pen paper-chart recorders. These chart recordings provide a backup in the event of failure of the digital recording system and also provide a "quick look" at the recorded data. The logarithms of signal amplitudes are recorded on the chart recorders with a range of 50 dB.

All receiver outputs are multiplexed along with (*i*) system status indicators such as whether the frequency control loop is tracking or holding in a signal fade, (*ii*) outputs from weather instruments such as rain gauges, thermometers and wind speed recorders, (*iii*) outputs from the on-going interim experiment[10] using the COMSTAR beacon at 128°W, and (*iv*) another propagation experiment[17] using a 12-GHz beacon on the NASA/Canadian Communications Technology Satellite (CTS). These multiplexed signals are digitized, temporarily stored in the mini-computer core memory, screened by the computer for relevance, and stored on digital magnetic tape. Multiplexer and analog-to-digital converter sequencing, digital data buffering and digital tape drive control are handled by the same minicomputer that points the receiving antenna.

The objectives of the data screening procedure* are to minimize the amount of superfluous data stored while not discarding any relevant propagation data. The screening algorithm copes with the multiplicative signal fluctuations caused by the atmosphere and with the additive noise that dominates at low signal level.

All receiver outputs are digitized every ¼ second and temporarily stored as a sample set. A running mean of these samples is accumulated for each channel. This running mean is compared with the mean value last recorded for the channel. If for any channel, the running mean value becomes different from the previous mean value by an amount greater than that expected because of receiver noise and atmospheric fluctuations, the running means for all channels are recorded. These recorded running means then become the previous mean values for testing a new sequence of running means that starts with the next sample set. This procedure detects gradual changes in received signal parameters. A test for rapid or impulsive change in a received signal parameter is also made on the individual sample sets. A data set is recorded at least once each minute regardless of whether there are changes in the data.

Each data set contains the time it was recorded so the time interval spanned is available for further data processing. This data screening procedure, then, records all significant changes in data whether instantaneous or gradual, records data periodically for equipment checking, and within these constraints minimizes the amount of data recorded.

The data handling procedures also include provisions for (*i*) easily stopping and starting data collection when the facility is used for radio astronomy, (*ii*) recovering from primary power interruption,† and (*iii*)

* The computer programming required for data screening and storage was done by H. W. Arnold.

† The computer programming required for power failure recovery and antenna pointing was done by R. W. Wilson.

recording calibration signals on the data tapes. Weather data and system status are recorded on the magnetic tape periodically.

3.5 Propagation parameters measured

The propagation parameters measured and recorded at the main receiving facility are briefly described in this section. A more detailed parameter description is included in Ref. 9. The measurements are recorded for all events that produce propagation irregularities. These events are all associated with cloudy weather; the most severe are associated with precipitation. Propagation statistics for continuous time intervals spanning at least a year are being compiled from the data.

3.5.1 19-GHz attenuation and depolarization measurements

The measurements made on the polarization switched signals and illustrated in Fig. 7 are:

A19V = co-polarized vertical signal amplitude (TVRV)
A19XV = cross-polarized signal amplitude coupled from vertical to horizontal (TVRH)
A19H = copolarized horizontal signal amplitude (THRH)
A19XH = cross-polarized signal amplitude coupled from horizontal to vertical (THRV)
ϕ19V-H = phase difference between vertical and horizontal signals (TVRV and THRH)
ϕ19V-XV = phase difference between vertical signal (TVRV) and its cross-polarized component (TVRH)
ϕ19H-XH = phase difference between horizontal signal (THRH) and its cross-polarized component (THRV)

where TV indicates transmit vertical polarization from the satellite, TH indicates transmit horizontal, RV indicates receive on vertical polarization on the ground, and RH indicates receive horizontal. Attenuation is obtained by comparing amplitudes of attenuated signals with amplitudes of clear air signals. The measurement of ϕ19V-H requires holding a phase reference from one polarization switch time period to the next. This sets a lower bound on the switching frequency determined by the instabilities of both the satellite and ground station primary oscillators (phase references) and the desired accuracy of the phase measurement.

Attenuation and depolarization of signals with any transmitted polarization can be determined directly from these amplitude and phase measurements without having to assume a rain model.* This procedure

* This capability is desirable since all positions within a U.S. coverage beam cannot have their polarizations set optimally with respect to their local rainfall.

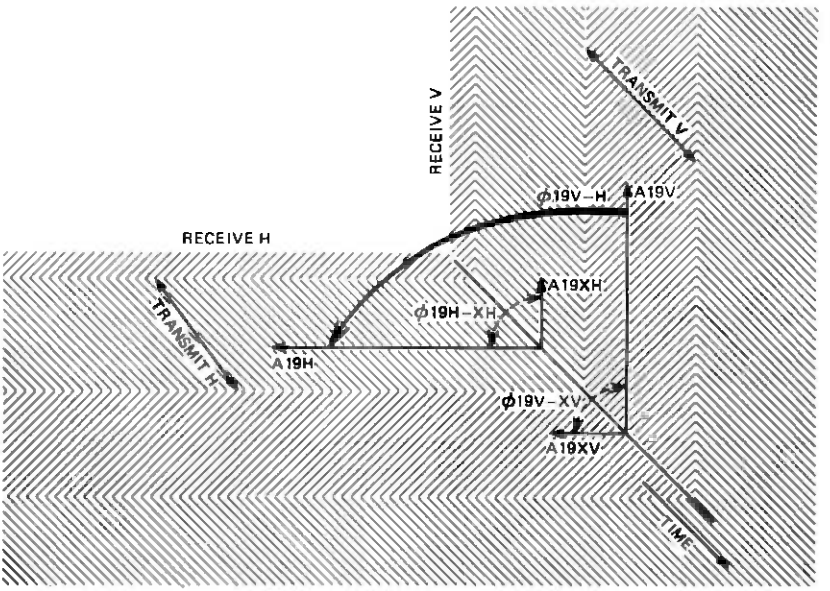


Fig. 7—Three-dimensional diagram illustrating with different crosshatching the transmitter sequencing in time for the vertically (V) and horizontally (H) polarized 19-GHz beacon signals. Also illustrated in two different planes are the reception on V and H polarizations of these signals. The actual signals received are indicated by arrows. Labels of measured signal parameters are as follows. A19V and A19H are copolarized 19-GHz signal amplitudes; A19XV and A19XH are cross-polarized 19-GHz signal amplitudes; ϕ19V-XV and ϕ19H-XH are phase differences between co- and cross-polarized signals; and ϕ19V-H is differential phase between copolarized signals. Note that the measurement of ϕ19V-H requires holding a phase reference between time slots when the beacon transmits V and when it transmits H.

is described in Refs. 7 and 18. These results will be useful for determining (*i*) if there is a fixed polarization orientation with precipitation-related crosstalk low enough to permit doubling capacity by simultaneously using the same frequency on two polarizations, or (*ii*) if (*i*) is not possible, then if a sufficiently low crosstalk can be obtained simply by tracking the linear polarization rotation, or (*iii*) if a more complicated crosstalk minimizing technique will be required, or (*iv*) if crosstalk is low enough to permit use of orthogonal circular polarizations or non-optimally oriented orthogonal linear polarizations for capacity doubling. Determining parameters for signal polarizations different from the transmitted polarizations with sufficient accuracy for use in system design requires measurement accuracy considerably greater than is necessary for using the measured parameters directly in design.[18]

3.5.2 28.56-GHz attenuation and depolarization measurement

Since only one polarization is transmitted at 28.56 GHz, only A28V, A28XV and ϕ28V-XV can be measured. These amplitudes and phases

are similar to the 19-GHz measurements made in the transmit V time slot (see Fig. 7). Thus, crosstalk can be determined directly only for the polarization orientation of the measurement. Estimates of crosstalk for other polarization orientations and for other frequencies will be made using these direct measurements, rain models and the extensive 19 GHz measurements.

3.5.3 Amplitude and delay dispersion measurements (usable bandwidth)

The satellite transmits a carrier with frequency $f_0 = 28.56$ GHz and two sidebands coherently related to $\omega_0 = 2\pi f_0$ with frequencies $\omega_\ell = [(n-1)/n]\,\omega_0$ and $\omega_u = [(n+1)/n]\,\omega_0$. These signals can be represented at the earth station by $s(t) = A(\omega)\cos[\omega t - \omega\tau(\omega)]$ where $\tau(\omega)$ is the dispersive delay at frequency ω and $A(\omega)$ is the dispersive amplitude. If there is no dispersion, $A(\omega)$ and $\tau(\omega)$ are constants; normalized values at the satellite are $A(\omega) = 1$ and $\tau(\omega) = 0$.

Two terms, A_1 and A_2 or τ_1 and τ_2, for series representations of the amplitude or delay, $A(\omega) = A_o + A_1\omega + A_2\omega^2 + \ldots$ and $\tau(\omega) = \tau_o + \tau_1\omega + \tau_2\omega^2 + \ldots$, can be obtained from measurements of differential amplitudes, $A_u = A(\omega_u) - A(\omega_o)$ and $A_\ell = A(\omega_0) - A(\omega_\ell)$ and differential phases, $\phi_u = [n/(n+1)]\omega_u\tau(\omega_u) - \omega_o\tau(\omega_0) = \omega_o[\tau(\omega_u) - \tau(\omega_0)]$ and $\phi_\ell = \omega_o\tau(\omega_0) - [n/(n-1)]\,\omega_\ell\tau(\omega_\ell) = \omega_o[\tau(\omega_0) - \tau(\omega_\ell)]$.

The coefficient A_o can be determined by an absolute attenuation measurement. The absolute delay τ_o cannot be determined since the phase at the satellite is unknown. With only the carrier and one set of sidebands, no higher order dispersion coefficients can be determined.

The measurement of differential phases ϕ_u and ϕ_ℓ requires scaling of ω_u by $n/(n+1)$ to ω_o and of ω_ℓ by $n/(n-1)$ to ω_o as indicated.

Measurement of dispersion (if it exists because of precipitation or index of refraction inhomogeneities) is made over the 528-MHz bandwidth ($n = 108$) and the 1056-MHz bandwidth ($n = 54$) permitted by the different satellites with the different sideband frequencies (± 264 MHz and ± 528 MHz). These measurement bandwidths are large enough to cover system requirements for many years since it is unlikely that transmission rates for individual channels within these bands will exceed 1 gigabit/s in the near future.

3.5.4 Rain signal-scatter measurement

The off-axis beam of the 7-meter antenna points toward a geostationary orbit position that is currently not occupied but could be occupied in the future. Beacon signal scattered by rain or other atmospheric phenomena into the off-axis beam represents a potential interference mechanism between the COMSTAR orbit position and the unoccupied orbital position along the off-axis beam. Measurement of the off-axis

beam signal amplitudes determines crosstalk levels and thus limitations on orbital spacing between satellites that may result from scattering of signal by the rain or other atmospheric phenomena.

3.6 Performance

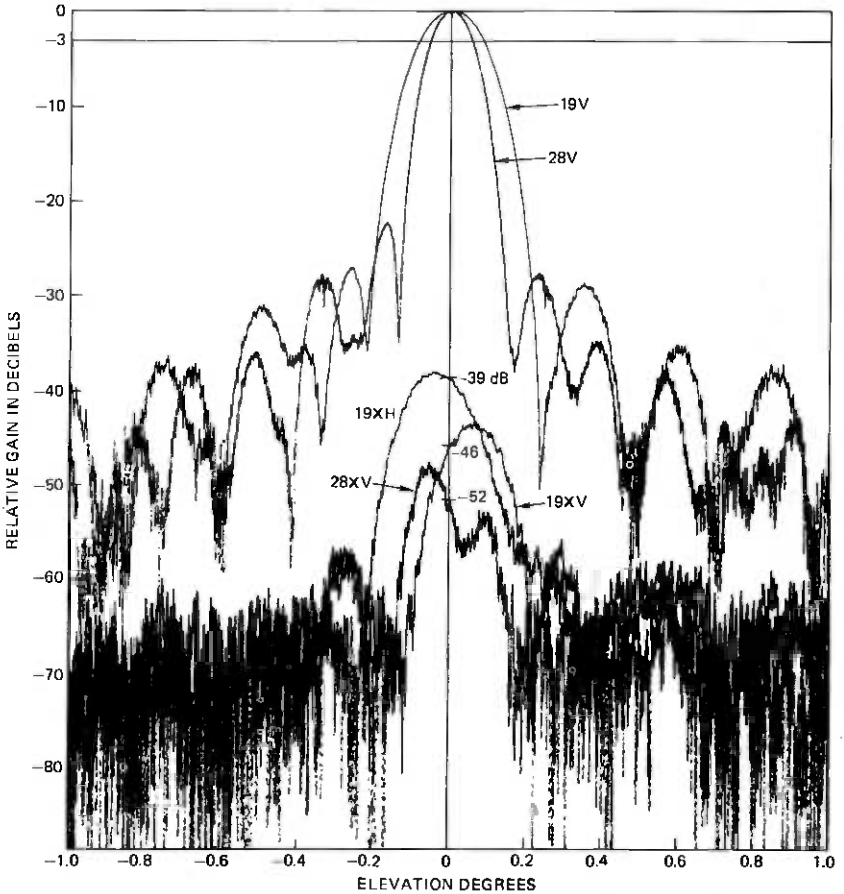
Performance of the individual subsystems in the Crawford Hill propagation experiment is described in the previous sections and in the following papers.[8,9] This section summarizes overall performance of the Crawford Hill main receiving facility working with the COMSTAR beacon at 95°W.

Clear-air signal-to-noise (S/N) ratios measured on cold low-humidity winter days are 60 and 59 dB for 19-GHz H and V polarizations and 61 dB for the 28-GHz carrier. These values are ratios of average signal power to average noise power measured at the narrow-band IF outputs.

Figure 8 shows antenna patterns made by scanning the seven-meter antenna past the beacon and recording the logarithmic amplitude detector outputs. In Fig. 8, the lowest three curves, A19XV, A19XH and A28XV, are the cross-polarized signals from receiver channels with 1.0-, 1.0-, and 1.5-Hz 3-dB IF bandwidths, respectively; the upper curves, A19V and A28V, are the copolarized signals from receiver channels with 10- and 15-Hz 3-dB IF bandwidths. The main beam of the A19H curve lies on top of the A19V curve and A19H sidelobes are all within a few dB of A19V sidelobes. The maximum values of cross-polarization discrimination, i.e., the cross-polarized signal level relative to the copolarized signal level at all points within the 3-dB beamwidths are XRV19 $\leq$ 41 dB; XRH19 $\leq$ 36 dB and XRV28 $\leq$ 45 dB. Differential phase ϕ19V-H is an indicator of the accuracy of alignment of the phase centers of the 7-meter antenna feeds. The differential phase varies $<\pm 1.2°$ over the 3-dB beamwidths. Scans in planes other than the principal planes (azimuth and elevation) are similar.

These antenna pattern measurements illustrate the combined performance of the beacon antennas and the 7-meter antenna. The measurements also demonstrate the capability of the receiver to maintain frequency lock and polarization switch synchronization through deep signal fades (antenna nulls) and to measure the low cross-polarized signal components to levels of the order of 70 dB below the clear-air copolarized signal levels.

Signal amplitudes vary less than ±0.3 dB and differential phase varies less than ±2° over several weeks. Most of this variation has a diurnal cycle. Day-to-day repeatability is within measurement resolution. Residual cross-polarized signal levels change with a diurnal cycle that also changes slowly with time, apparently due to change in the orientation of the satellite with respect to the sun. Day-to-day repeatability of cross-polarized signal levels are within ±0.2 dB.

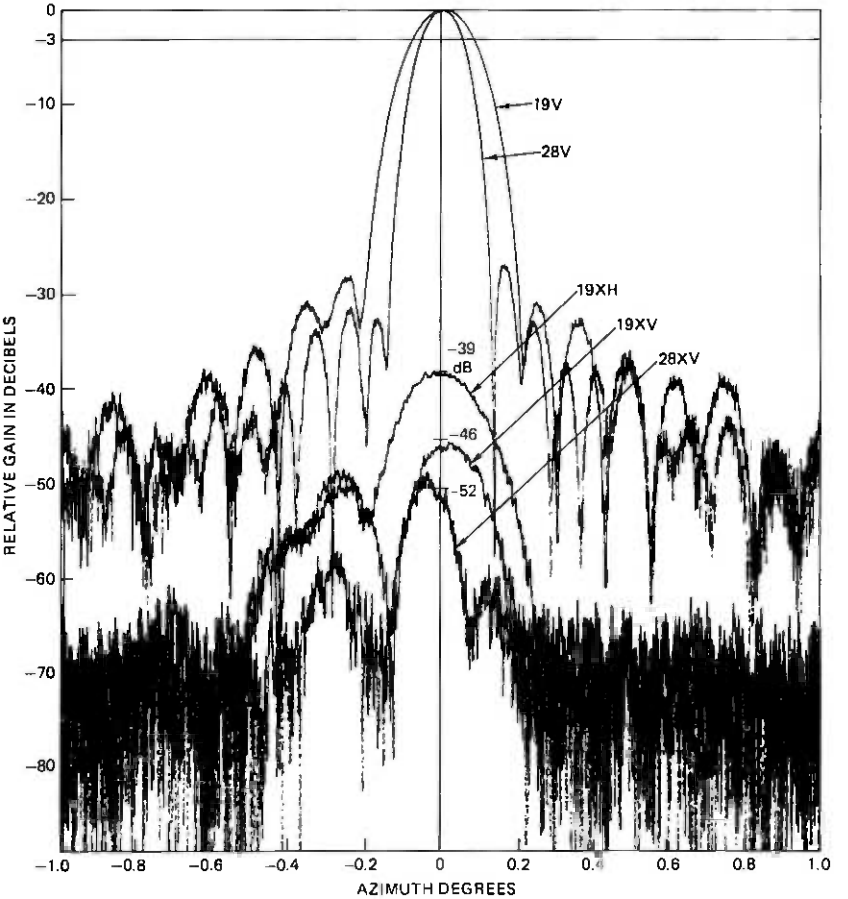


(a) Elevation scan.

Fig. 8—Crawford Hill 7-meter antenna scans of the COMSTAR beacon at 95°W made with the Crawford Hill main facility receiving electronics. Note that the worst first sidelobe is 22 dB down and that all others are at least 27 dB down. Sidelobe levels approach 40 dB down 1° off beam axis. Cross-polarized signals 19XH, 19XV, and 28XV remain more than 38 dB below on axis copolarized signals throughout. Cross-polarization discrimination is better than 36 dB throughout the 3 dB beamwidths. Receiving electronics follow signals through antenna nulls.

IV. CONCLUSIONS

Continuous transmission at 19 and 28 GHz from the COMSTAR beacons in geosynchronous orbit are providing unique opportunities for gathering propagation information needed for designing future high-capacity satellite communication systems. An extensive receiving facility at Crawford Hill, New Jersey, comprising a precision 7-meter-diameter antenna and sophisticated receiving electronics, is providing a detailed look at propagation effects such as attenuation, depolarization, coherence



(b) Azimuth scan.

Fig. 8 (*continued*)

bandwidth and signal scatter that are usually related to atmospheric precipitation. This facility receives on two orthogonal linear polarizations at the two beacon frequencies. In clear weather when atmospheric attenuation is minimal, the signal-to-noise ratio when receiving the beacon at 95°W is 60 dB. Phase differences and amplitudes are measured for all signals and for their cross-polarized components. Cross-polarization isolations are >35 dB throughout the entire antenna beams to the −3-dB beam edges.

Bell Laboratories receiving facilities in Georgia and Illinois are collecting information on signal attenuation and diversity in other climatic regions. Additional propagation information is being collected by non-Bell experiments at other locations.

V. ACKNOWLEDGMENTS

These COMSTAR beacon propagation experiments would not have been possible without the foresight, cooperation and support of a number of people at AT&T, at the Long Lines department of AT&T and at Bell Laboratories. The efforts of all these people are much appreciated by all of us who are now directly involved in these very successful experiments.

E. E. Muller very effectively coordinated the requirements of these experiments with COMSAT General, the overall satellite contractor, with Hughes Aircraft Corporation, the satellite and beacon antenna manufacturer, and with COMSAT Laboratories, the beacon manufacturer.

The contributions of many individuals concerned with particular parts of the experimental facilities are acknowledged in the following companion papers. H. W. Arnold was particularly involved in the overall design, assembly and operation of the Crawford Hill experiment. The efforts of H. H. Hoffman in these areas are also appreciated. The experiment was originally conceived and supported by L. C. Tillotson and D. C. Hogg among others. The continuing support and interest of A. A. Penzias and D. O. Reudink has been helpful. The backing of R. F. Latter and R. C. Harris of AT&T Long Lines and E. F. O'Neill of Bell Laboratories is also appreciated.

REFERENCES

1. L. C. Tillotson, "A Model of a Domestic Satellite Communication System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2137.
2. D. C. Hogg, "Millimeter-Wave Communication Through the Atmosphere," Science, *159*, January 5, 1968, pp. 39–46.
3. R. W. Wilson, "Sun Tracker Measurements of Attenuation by Rain at 16 and 30 GHz," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1383–1404.
4. R. W. Wilson, "A Three-Radiometer Path-Diversity Experiment," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1239–1242.
5. D. C. Hogg and T. S. Chu, "The Role of Rain in Satellite Communications," Proc. IEEE, *63*, September 1975.
6. D. C. Cox, H. W. Arnold, and A. J. Rustako, "Some Observations of Anomalous Depolarization on 19 and 12 GHz Earth-Space Propagation Paths," Radio Science, *12*, May–June 1977, pp. 435–440 and D. C. Cox and H. W. Arnold, "Preliminary Results from the Crawford Hill 19 GHz COMSTAR Beacon Propagation Experiment," U. S. Nat'l Committee of International Union of Radio Science (USNC/URSI) meeting, October 11–15, 1976, Amherst, Massachusetts.
7. D. C. Cox, "Design of the Bell Laboratories 19 and 28 GHz Satellite Beacon Propagation Experiment," IEEE International Conference on Communications (ICC '74) Record, June 17–19, 1974, Minneapolis, Minnesota, pp. 27E1–27E5.
8. T. S. Chu, R. W. Wilson, R. W. England, D. A. Gray, and W. E. Legg, "The Crawford Hill 7-Meter Millimeter-Wave Antenna," B.S.T.J., this issue, pp. 1257–1288.
9. H. W. Arnold, D. C. Cox, H. H. Hoffman, R. H. Brandt, R. P. Leck, and M. F. Wazowicz, "The 19- and 28-GHz Receiving Electronics for the Crawford Hill COMSTAR Beacon Propagation Experiment," B.S.T.J., this issue, pp. 1289–1329.
10. H. W. Arnold, D. C. Cox, and D. A. Gray, "The 19-GHz Receiving System for an Interim COMSTAR Beacon Propagation Experiment at Crawford Hill," B.S.T.J., this issue, pp. 1331–1339.
11. N. F. Dinn and G. A. Zimmerman, "COMSTAR Beacon Receiver Diversity Experiment," B.S.T.J., this issue, pp. 1341–1367.

12. AT&T Company Application to FCC for a Domestic Satellite Communications System, March 1, 1973 and papers in COMSAT Technical Review, 7, No. 1, Spring 1977: "The COMSTAR Program" by R. Briskman, and "The COMSTAR Satellite System" by G. Abu-Taleb, M. Kim, K. Manning, J. Phiel, Jr., and L. Westerlund.
13. Papers in COMSAT Technical Review, Vol. 7, No. 1, Spring 1977: "Centimeter Wave Beacons for the COMSTAR Satellites" by L. Pollack, "Centimeter Wave Beacon Transmitter Design" by W. J. Getsinger, "19- and 28-GHz IMPATT Amplifiers" by M. J. Barrett, and "Low-Jitter Oscillator Source for the COMSTAR Beacon" by R. E. Stegens.
14. D. C. Cox, unpublished notes based on Ref. 15 and on B. R. Bean and E. J. Dutton, "Radio Meterology," National Bureau of Standards Monograph 92, U.S. Government Printing Office, Washington, D.C., March 1, 1966.
15. A. A. M. Saleh, unpublished notes based on J. A. Ratcliffe, "The Magneto-Ionic Theory," University Press, Cambrdige, 1961; K. G. Budden, "Radio Waves in the Ionosphere," University Press, Cambridge, 1961; J. M. Kelso, "Radio Ray Propagation in the Ionosphere," McGraw-Hill, New York, 1964; and K. Davies, "Ionospheric Radio Propagation," U.S. National Bureau of Standards, Monograph 80, April 1, 1965.
16. V. J. Slabinski, "Expressions for Time-Varying Topocentric Directions of a Geostationary Satellite," *COMSAT Technical Review, 5,* No. 1 (Spring 1975), pp 1–14.
17. A. J. Rustako, Jr., "An Earth-Space Propagation Measurement at Crawford Hill Using the 12-GHz CTS Satellite Beacon," B.S.T.J., this issue, pp. 1431–1448.
18. D. C. Cox, "Some Effects of Measurement Errors on Rain Depolarization Experiments," B.S.T.J., *54,* No. 2 (February 1975), pp. 435–450.
19. R. B. Briskman, R. F. Latter and E. E. Muller, "Call for Help," IEEE Spectrum, *11,* October 1974, pp 35–36.
20. E. E. Muller, "Notes on the COMSTAR Experiment," B.S.T.J., this issue, pp. 1369–1370.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

COMSTAR Experiment:

The Crawford Hill 7-Meter Millimeter Wave Antenna

By T. S. CHU, R. W. WILSON, R. W. ENGLAND, D. A. GRAY, and W. E. LEGG

(Manuscript received January 27, 1978)

A 7-meter offset Cassegrainian antenna with a precise surface has been built and tested. Measurements using a terrestrial source were made and compared with calculations for 19, 28.5, and 99.5 GHz. Low sidelobe level ($\lesssim -40$ dB) at one degree off the main beam and low cross polarization (≤ -40 dB) throughout the main beam are achieved using a quasi-optical 19/28.5-GHz feed system that also demonstrates very low multiplexing loss (~ 0.1 dB). The prime-focus gain measurement at 99.5 GHz found the difference between the measured and calculated gains to be (0.79 ± 0.45) dB, which is consistent with the expected rms surface error (~ 0.1 mm). Multiple-beam operation accommodates both propagation experiments with the COMSTAR beacons at 19 and 28.5 GHz and millimeter wave radio astronomy observations without physical disturbance of equipment.

I. INTRODUCTION

The Crawford Hill 7-meter antenna (Fig. 1) was built for propagation measurements with the COMSTAR beacons at 19 and 28.5 GHz, and for radio astronomy at frequencies from 70 to 300 GHz. It demonstrates that low sidelobes and high cross-polarization rejection can be obtained in an earth station antenna serving several satellites simultaneously. The antenna will also serve as a test bed for future propagation and antenna measurements.

The main part of this paper is contained in the following three sections. Section II describes the 7-meter antenna starting with the initial requirements. Section III gives a detailed description of the 19/28.5-GHz

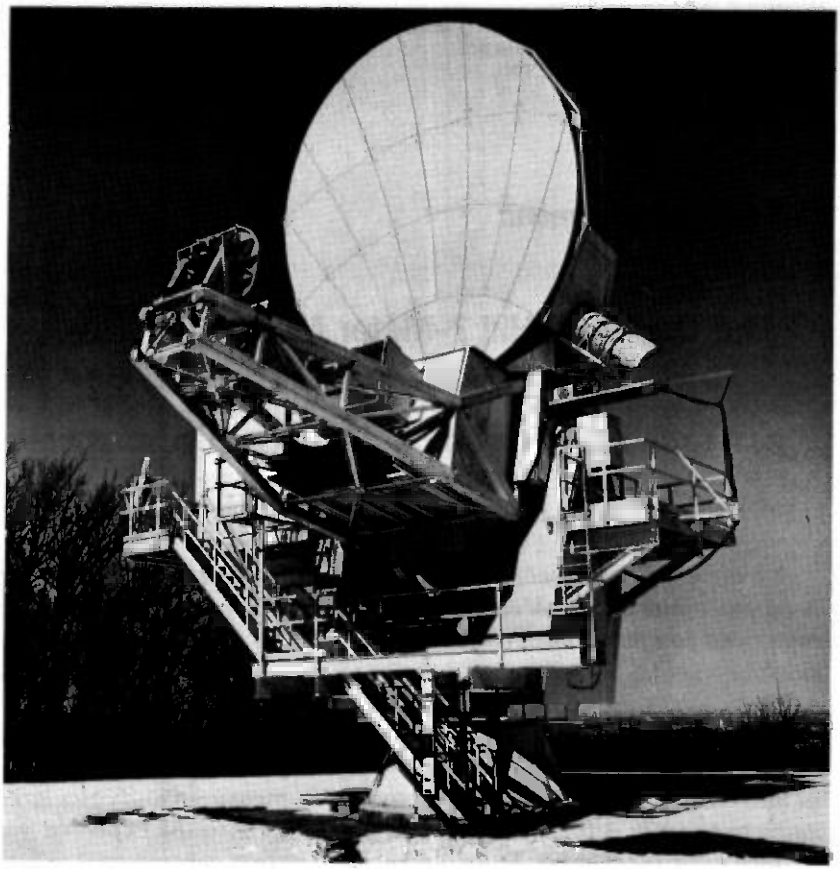


Fig. 1—The 7-meter offset Cassegrainian millimeter wave antenna. The subreflector is mounted below antenna aperture, and key structural parts are covered with insulation.

feed system for the COMSTAR beacon experiment. Section IV describes measurements of the completed antenna at 19, 28.5, and 99.5 GHz using terrestrial sources on a tower at a distance of 11 km.

II. DESCRIPTION

2.1 Requirements

The major requirements imposed by the satellite beacon experiment on the 7-m antenna are for cross polarization to be less than −35 dB within the main beam, sidelobes to be below −40 dB at 1 degree or more from the main beam, multi-beam operation to be possible, and performance in all weather to be good, especially during summer thunder

showers. The attainment of at least minimal performance at the high end of the radio astronomy band further requires a surface accuracy of 0.1-mm rms and a maximum pointing error of 10″ arc. The expectation of performing beam-feed experiments and of possibly accommodating large cryogenically cooled radio astronomy receivers made it desirable to provide for a Nesmyth focus (feed along the elevation axis) in addition to the Cassegrainian focus.

The requirements for low sidelobes imposes severe limitations on the amount of aperture blocking that can be tolerated.[1] Thus, an offset paraboloid was chosen with almost no aperture blockage. This can be made compatible with the cross-polarization and multibeam requirements by choosing a sufficiently large secondary focal ratio.[2,3,4] An additional benefit of the offset design is a large return loss. The low sidelobe requirements also place restrictions on the allowable surface errors. The magnitude of the allowable errors depends on their correlation lengths and the angles at which sidelobes can be tolerated, but a tolerance of about 0.01λ is needed to avoid degradation of the far sidelobes.[5] Thus a surface tolerance of 0.1-mm rms was chosen to simultaneously satisfy the sidelobe requirements at 30 GHz and allow radio astronomical operation in the 200 to 300 GHz band.

The main reflector has a focal length of 6.5659 m. It is offset such that its bottom is 1.2074 m above the reflector axis. The subreflector is a 1.2-m by 1.8-m oval portion of a hyperboloid offset 8.258 cm above its axis (Fig. 2). It has focal lengths of 0.8697 m and 5.2262 m giving a magnification ratio of 6.01. The central ray from the feedhorn to the subreflector makes an angle of 6.828 degrees with the axis and the illumination cone has a half angle of 5.06 degrees. The main reflector is subtended by a circular cone of 26.75-degree half angle, and the axis of the cone makes an offset angle of 37.26 degrees with the axis of the main reflector. The geometry results in a blocking of 4.4 cm of the main reflector by the subreflector and an expected contribution to the cross-polarization level of -56 dB in the main beam region, well below the required level.

The narrow beam pattern required for the feed is the main price which must be paid for the advantages of a large effective F/D ratio. Small-cone-angle corrugated horns have been rejected for both radio astronomy and propagation feeds because of their excessive length. In the radio astronomy feed, cryogenically cooled receivers are in use, and the best performance is obtained by cooling the feedhorn with the receiver and radiationally coupling out of the Dewar. Thus, a small-cone-angle horn would not only be a large mass to cool, but a large low-loss Dewar window would be difficult to make. Instead, a quasi-optical feed system[6] is used for radio astronomy to couple a small corrugated horn in the receiver Dewar to the 7-meter antenna. This feed system incorporates image rejection, local oscillator injection, and calibration.

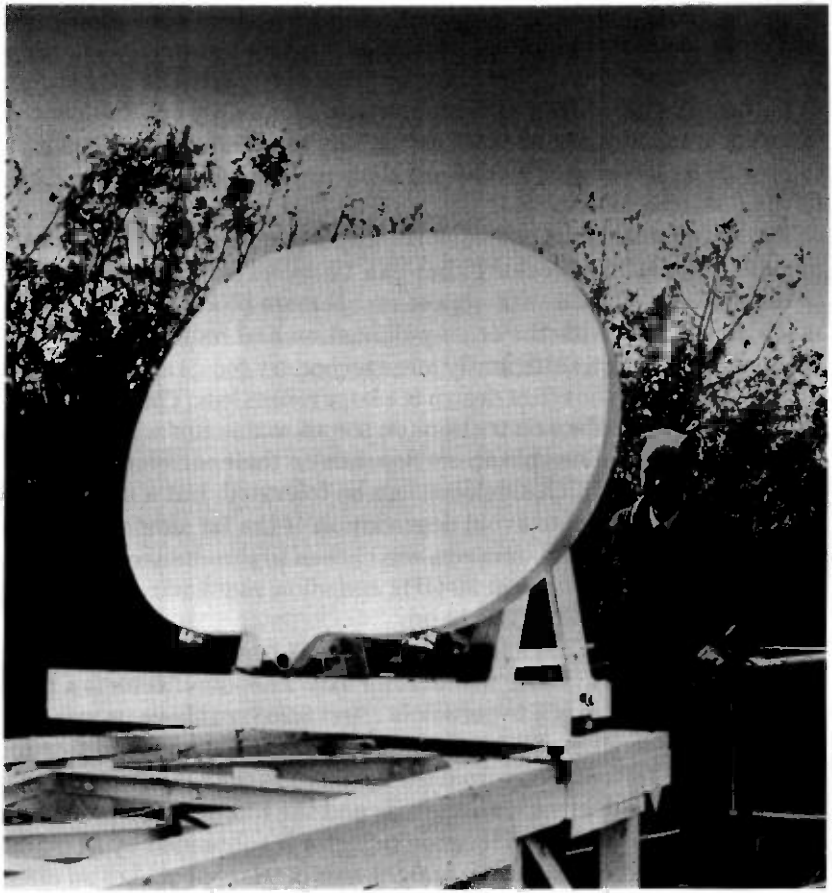


Fig. 2—The subreflector with a machined aluminum surface of 20 μm rms. The oval shape is provided for azimuthal off-axis beams.

In the case of the 19/28.5-GHz propagation feed, the two widely separated frequencies would lead to problems with frequency-dependent phase patterns of a single corrugated feed and with higher order modes generated in a low-loss waveguide polarization diplexer. These problems were circumvented by using quasi-optical methods[7] for both polarization and frequency diplexing as shown in Fig. 3. Four launchers, each consisting of an offset ellipsoid and a corrugated horn, are used, each for a single frequency and polarization. A quasi-optical frequency diplexer[8] first combines two frequencies at each polarization, then a polarization grid[9] combines and simultaneously cleans the two orthogonal polarizations. A 45-degree mirror finally reflects the combined beam to the subreflector, leaving adjacent areas in the focal plane available for other

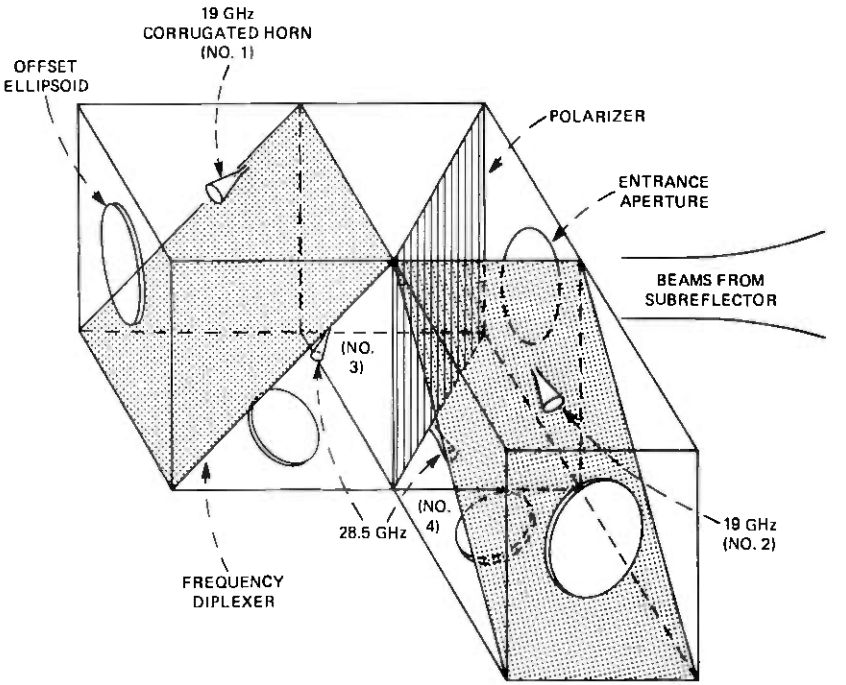


Fig. 3—Sketch of a 19/28.5 GHz dual-polarization feed system. The beam from the subreflector passing through the entrance aperture is first split by the polarizer and then subdivided by two frequency diplexers.

receivers. This feed system was first tested on axis, but operates in a position to produce a beam 0.5 degrees off-axis in azimuth, leaving the on-axis feed position free for the higher frequency radio astronomy feed. This off-axis operation results in very little degradation in the antenna pattern.

Past experience has shown that, after aging, radome surfaces hold thick water films during rain, causing large microwave attenuation.[10] The 7-m antenna was therefore designed without a radome, but with thermal insulation of critical parts of the structure to allow operation in the sun without significant degradation of pointing or surface accuracy. The support structure affords stiffness sufficient for operation at 30 GHz in winds up to 70 mph.

2.2 Mechanical description

The mount for the 7-m antenna is a conventional elevation-over-azimuth design with a single 2-m diameter cross-roller azimuth bearing and a yoke holding the elevation bearings. The elevation moving structure is built around a steel box girder. One end of the girder has a 32-cm bore

spherical roller bearing. The other end has a 1.4-m bore radial roller bearing sized to provide a 1.2-m diameter path to the side cab from the vertex area. The surface of the box girder facing the subreflector has a 1.6-m opening for the RF beam, completing the path from the subreflector to the elevation axis. A truss girder above the box girder supports the nine steel trusses that are radial to the axis of the main reflector and support its surface panels. The center truss is the longest and has been made deeper than the others by connecting its outer member to the elevation wheel support structure. The stiffness of the other trusses was adjusted using computer calculations so that gravity and wind deflections are expected to introduce mainly pointing errors with only small deviations of the antenna surface from its parabolic shape.

The 27 surface panels are arranged in four approximately equal width rings. The panels are A356 aluminum castings containing 17-cm deep ribs on the back side near the edges. These are joined by 9-cm deep ribs running in the circumferential direction with spacings of approximately 30 cm. The castings were tempered to T51, rough machined on a numerically controlled milling machine, stress relieved, and then machined to the final contour on the same machine. The manufacturer tested all the panels on his milling machines and found the surfaces to be accurate to better than 50-μm rms with an average value of about 40 μm. One of the early panels was tested on a Portage measuring machine and was found to have the same rms error (37 μm) as determined by the manufacturer. It was then subjected to 10 rapid temperature cycles from -40°C to 60°C and back. Remeasurement on the Portage machine showed that the surface error had increased to 50-μm rms.

After machining, the panels were cleaned and painted with a 30 to 50-μm coat of an Alkyd base, TiO_2 pigment, flat white paint. After final alignment of the panel corners to the alignment template (see below), five panels near the center were measured against the template on an 18-cm grid. One of these panels was found to have a much larger surface error (100-μm rms) than expected with one bad area and a general warp. A visual inspection showed this bad area to be by far the worst on the antenna. However, it is only a small contribution to the total surface error of 100-μm rms.

The panels are held to the back-up structure by 2.54-cm diameter adjusting bolts. At the panel end, these bolts attach to machined pads in the corners of the peripheral ribs with ball and socket joints to avoid transmitting torques to the panels. These adjusting bolts are perpendicular to the surface and bend to take up the differential expansion between the aluminum panels and the steel backup structure.

Alignment of the surface panels was accomplished using a sweep template (Fig. 4) for reference. This was done under a tent with the antenna in the vertical look position, before installing the subreflector

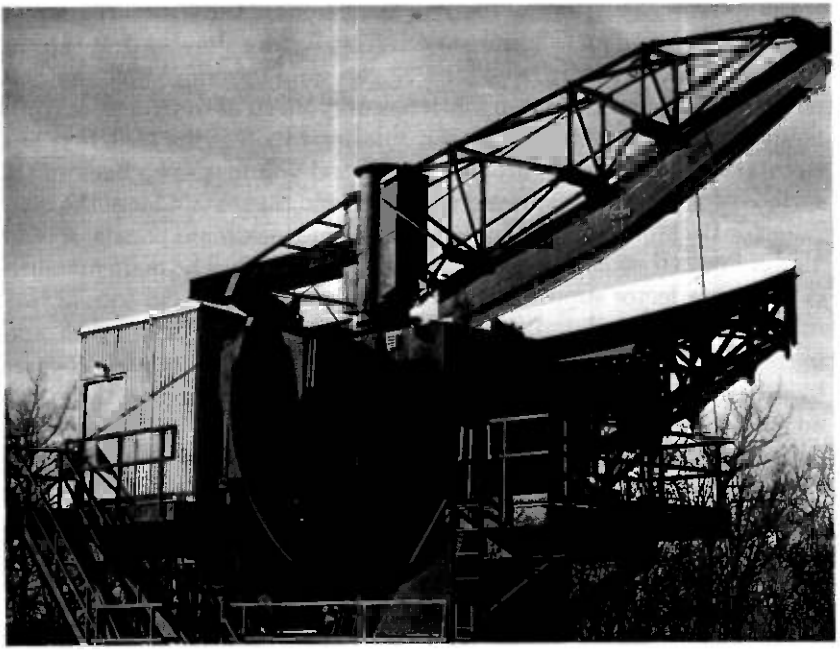


Fig. 4—Panel alignment template mounted in place of subreflector support tower above 7-meter reflector prior to installation of tent. The template consists of two sections joined by a link.

support tower and the vertex equipment room. The alignment template was machined in two parts, each of which has a reference hole at each end. The reference holes at the outer end of the inner template and the inner end of the outer template are at the same height and were joined by an accurate link. A counterweighted boom built on an adjustable vertical bearing supported the template. Another link adjusted the inner hole of the inner template to the correct radius from the rotational axis. Before each use, the vertical axis was set using an electronic tiltmeter, and the four template reference holes were set to their required relative height using gravity-oriented links on an optical level. Twelve transducers were attached to the template at radii corresponding to the panel mounting screws. They were referred to the accurate bottom surface of the template and used to measure panel corner positions.

The subreflector support structure has been kept well below the axis of the main reflector, especially near the secondary focus. This, coupled with the oversized subreflector, allow beams up to 2 degrees off axis in azimuth or 1 degree in elevation to be launched from the vertex equipment room. The subreflector itself is a machined aluminum casting of the same material and temper as the panels. It was machined on a vertical lathe and has a measured surface accuracy of 20-μm rms.

Computer calculations of thermal distortions of the structure coupled with temperature measurements on a test fixture showed that the sun shining on the structure from some angles would produce unacceptably large pointing and surface errors if the only thermal control were white paint. A strategy was adopted of insulating much of the structure and circulating ambient air through the critical volumes. The back of the surface panels has been sprayed with foam and the remainder of the main reflector backup structure enclosed by 5-cm-thick foam panels. A large blower floods this insulated volume with ambient air. The main members of the subreflector support structure are made of 10.16-cm-square steel tubing and are insulated. Ambient air is blown along them by blowers in the elevation girder. Most of the remaining steel structure is covered by 5-cm-thick panels of foam to increase thermal time constants, but no air circulation is provided.

In addition to shielding the backup structure from solar heat, the insulation of the back of the main reflector surface panels reduces the heat flux through the panels and hence the steady-state temperature difference between their surface and ribs. The penalty, however, is an increase ($\sim 2\times$) in the temperature rise of the surface panels in the full sun. This uniform temperature rise has the same effect as an absolute temperature change on the differential thermal expansion between the steel backup structure and the aluminum panels. It causes the panels to expand or contract in a direction parallel to the surface of the paraboloid defined by the backup structure. This has an almost negligible effect on the performance of the antenna. Warping of the panels due to thermal gradients perpendicular to their surface or thermal warping of the backup structure is much more important.

Two equipment rooms are on the antenna structure (Fig. 1). The vertex equipment room and the hollow elevation girder behind it provides a mounting space for receivers at the Cassegrainian focus. They move in elevation angle with the antenna. A double window of 50-μm Mylar* sheet allows RF energy to pass into this area with negligible loss (~ 0.1 dB) at 100 GHz. The side cab is provided for the Nesmyth focus. It moves only in azimuth. At present, it contains only receiver support equipment and the main elevation cable wrap. An 8-m by 6-m control building 15 meters from the antenna contains the rest of the receivers, the control computer, and a work area.

The volume under the azimuth bearing contains the azimuth cable wraps. A maypole wrap is used for most of the cables, but an auxiliary clock spring wrap carries the phase-sensitive RF cables.

The drive for each axis uses a single bull gear with two motors connected to separate speed reducers and pinions. The amplifiers for the two motors are biased so that, for low torque output, they torque in op-

* Registered trademark of E. I. Dupont.

posite directions to eliminate backlash, but when more than 15 percent of the maximum torque is required the opposing motor reverses. Each of the four dc motors is rated at 3 horsepower with phase control excitation. The gearing is chosen to give a slew speed of 1 degree/s in elevation and 2 degree/s in azimuth with 1 degree/s^2 acceleration in both axes and the ability to hold position or drive to the stow position in a 70-mph wind. This gearing results in the motor inertia dominating the antenna's inertia. Each motor has a brake with the same torque rating as the motor. When the drive system returns to standby from an active condition, the brakes are set before the motors relax so that the anti-backlash preload of the gear system is maintained.

The analog portion of the drive system is set up as a velocity control loop. Each drive motor has a tachometer generator. The sum of the tachometer signals is used in the main feedback loop, and the difference is used to damp possible oscillations in which the motors move in opposite directions.

Each axis has a direct drive Inductosyn* system which has an angular accuracy of 0.001 degree and a resolution of 21 bits. The antenna's minicomputer reads the position of each axis every 10 ms, subtracts it from the desired position, and applies the scaled difference to the drive system as a velocity command. Drive system overshoots are minimized by compressing the gain of the feedback loop within the computer by a factor of 4 for command velocities greater than $\frac{1}{4}$ of the maximum velocity. This strategy keeps the commanded velocity below the deceleration limit of the drive system when approaching the final position. When a source is tracked that moves at the sidereal rate or slower, the servo error has not been observed to exceed 0.001 degree with winds up to approximately 50 mph.

III. 19/28.5 GHz DUAL POLARIZATION FEED SYSTEM

3.1 Quasi-optical diplexers

The polarization diplexer (see Fig. 2) is simply a polarization grid made by photo-etching a copper-covered Mylar sheet. Copper strips 0.25-mm wide and 0.018-mm thick are spaced 0.25-mm apart on a Mylar sheet 0.013-mm thick. The grid is mounted on an aluminum supporting frame with an oval-shaped aperture of 33.02 cm by 48.26 cm. To achieve a flat grid, the supporting frame is made of jig plate instead of regular aluminum stock, which shows warping after machining. The plane of the grid is oriented at 45 degrees with respect to the incident beam from the subreflector. The conducting copper strips are perpendicular to the plane of incidence to avoid cross-polarized radiation[9] for both transmitting and reflecting orthogonal polarizations being diplexed.

* Registered trademark of Farrand Industries, Inc.

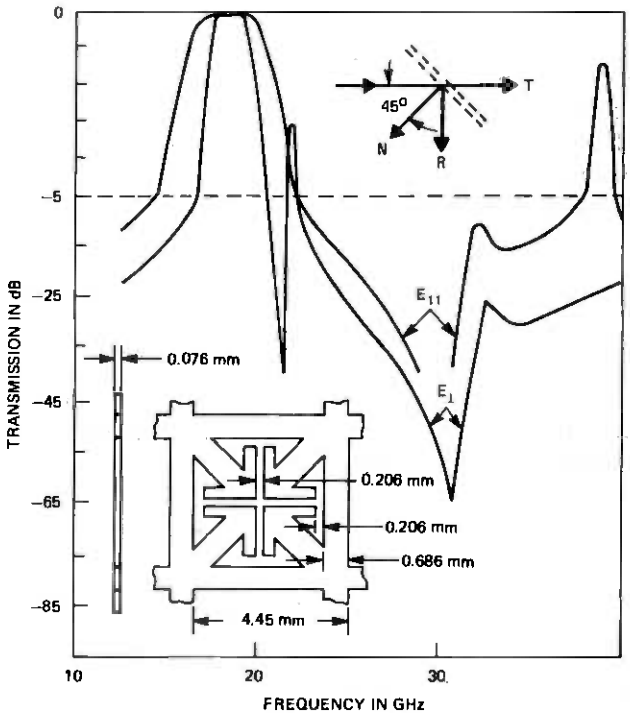


Fig. 5—Transmission of a double-self-supported frequency diplexer (see Fig. 3) with the indicated dimensions of the grid mask at 6.3 mm spacing and 45-degree angle of incidence for two polarizations.

Each of the two frequency diplexers is a pair of self-supported grids made of beryllium copper 75-μm thick. The 6.3-mm spacing between the girds is maintained by mounting on opposite sides of an aluminum frame with an oval aperture of 30.5 cm by 45.7 cm. Figure 2 shows that the polarization for Feeds No. 1 and No. 3 is perpendicular to the plane of incidence at the frequency diplexer, while that for Feeds No. 2 and No. 4 is parallel to the plane of incidence. The transmission characteristics for both polarizations and the dimensions of this double-self-supported grid are given in Fig. 5, which has been taken from Arnaud and Pelow's article.[8] The measured insertion losses for both transmitting 19.04 GHz and reflecting 28.56 GHz are less than 0.1 dB. Figure 5 also shows essentially perfect transmission and rejection frequency bands of well over 2 GHz each. This performance should be compared with a minimum insertion loss of 0.2 dB and a 2-GHz band loss of up to 1 dB for a typical waveguide diplexer at 19 and 30 GHz.

Several alternate quasi-optical frequency diplexers were tested. For a large (45-degree) angle of incidence, none of them achieved the desired transmission (−0.1 dB) and rejection (−20 dB) simultaneously for fre-

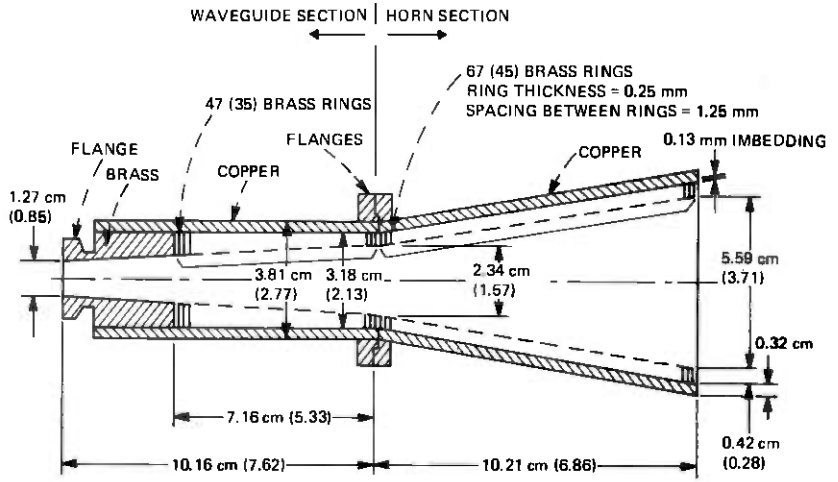


Fig. 6—The 19(28.5)-GHz conical corrugated horn and its hybrid mode launcher.

quency bands in the ratio of 1.5 to 1. A single gridded Jerusalem-cross diplexer[8] showed a transmission loss of 0.2 dB at 19 GHz for the polarization perpendicular to the plane of incidence and a rejection of only −10 dB at 28 GHz for the polarization parallel to the plane of incidence. When Mylar substrates were added to the double self-supported grid for improving its mechanical strength, transmission losses of 0.4 and 0.2 dB were observed at 19 GHz for the perpendicular and parallel polarizations.

The mechanical resonance frequency of the quasi-optical grids has been measured to be around 80 Hz, which is well above the passband of the baseband data in the COMSTAR beacon experiment.

3.2 Corrugated horns

Four corrugated horns (two each at 19 and 28.5 GHz) were constructed to illuminate the offset elliposids. The dimensions of the 19-GHz corrugated horn are illustrated in Fig. 6, where the numbers inside the parentheses are essentially scaled designs for 28.5 GHz. They are shortened versions of a very long corrugated horn built by Dragone.[11] The impedance matching between the smooth wall of the circular waveguide and the quarter wavelength slots of the corrugated horn is provided by a linear taper from half-wavelength corrugations that behave like a conducting surface. The 19-GHz transition from 1.27-cm diameter circular waveguide to 1.067 cm by 0.432 cm rectangular waveguide is hardware from the DR-18A terrestrial 18-GHz digital radio system. This transition consists of a tapered section from circular to square shape and a quarter-wavelength transformer. The 28.5-GHz transition from

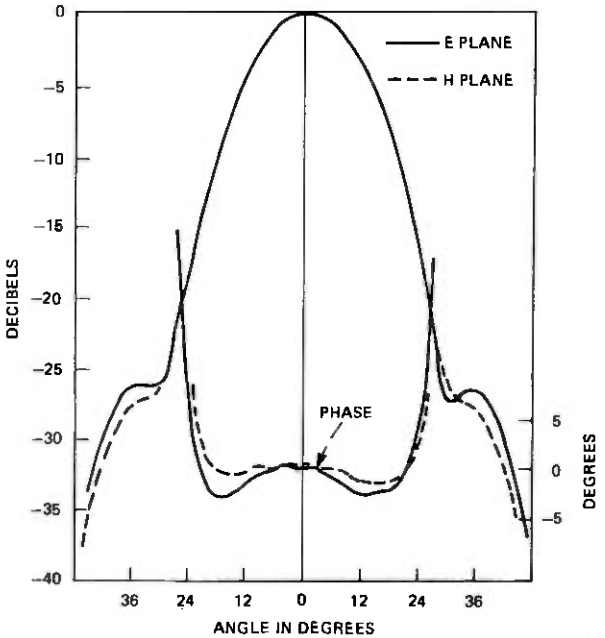


Fig. 7—Measured radiation patterns of 28.5-GHz corrugated horn. Rotation center 1.42 cm behind horn aperture.

0.846-cm diameter circular waveguide to 0.711 cm by 0.356 cm rectangular waveguide was also similarly designed. The measured return loss of all horns is better than 35 dB at the beacon frequencies 19 GHz and 28.5 GHz, and remains better than 31 dB over a 2-GHz band.

The measured far-field radiation patterns of the corrugated horns are shown for 28.5 GHz in Fig. 7, which are in excellent agreement with the calculations. The measured patterns at 19 GHz are essentially the same as those in Fig. 7, with the phase center located at 2.13 cm behind the horn aperture. The offset ellipsoid intercepts the horn radiation in the Fresnel region. It is difficult to measure the radiation pattern at a distance corresponding to the ellipsoid location. However, the calculated 20-dB half-beamwidth of the Fresnel zone radiation pattern is 2 degrees broader than those of the far-field patterns, and will be the same as the 28-degree half-cone angle of the ellipsoid subtended at the focus as shown in Fig. 8.

3.3 Offset ellipsoids

To be mirror-imaged at the Cassegrainian focal region (secondary geometrical focus of the subreflector for the on-axis case), the common phase center of the four offset launchers should be located in the middle

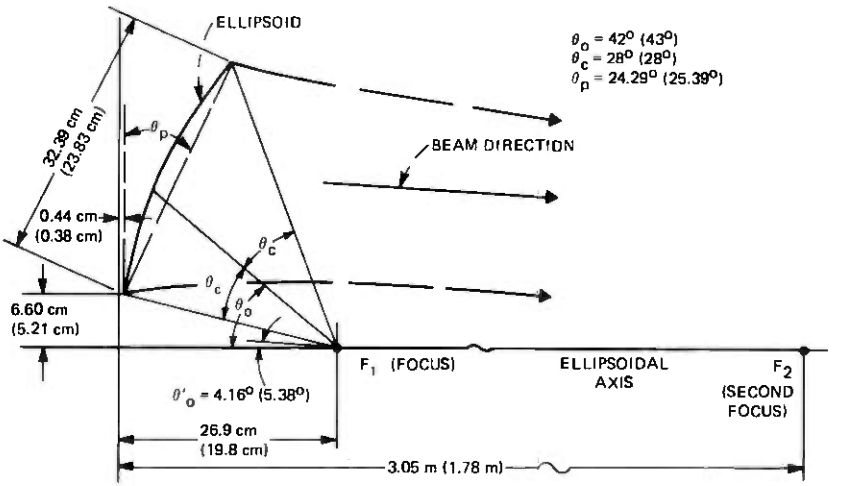


Fig. 8—Schematic diagram of the 19(28.5)-GHz offset ellipsoid.

of the feed box as shown in Fig. 9. Therefore, ellipsoids instead of paraboloids were used in the design of the feed system. One notes that a small movement of the phase center can be accomplished by the defocusing of a paraboloid. However, to place the phase center at a considerable distance in front of or behind a reflector with a small F/D ratio, an ellipsoid or a hyperboloid should be used. The design of an offset launcher

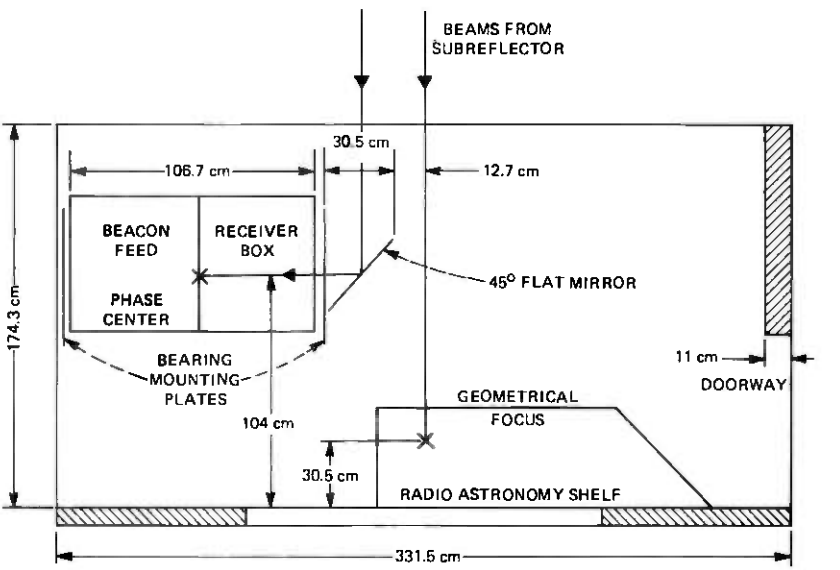


Fig. 9—Top view of vertex equipment room which has a height of 2.3 m and moves with the antenna.

was started with estimates by Gaussian beam approximation[12–14] and finalized by computer simulations which were essentially numerical integration of diffraction integrals,[2,15] assuming various design parameters.

Dimensions of both 19-GHz and 28.5-GHz offset ellipsoidal launchers are given in Fig. 8. Since the distance between the phase center and the 28.5-GHz elliposid is much greater in wavelengths than that of 19 GHz, the 28.5-GHz ellipsoid has larger size and greater curvature than a scaled version of the 19-GHz ellipsoid. The offset ellipsoid is subtended by a circular cone at the focus. The offset angle of the feedhorn axis is 3 degrees greater than that of the cone axis; thus approximately equal illumination taper can be achieved on the top and bottom edges of the reflector. The intersection of an ellipsoid and a circular cone subtended at one focus is a plane ellipse subtended by another circular cone at the other focus. The radiating beam from the ellipsoid will lie along the latter cone axis which deviates from the major axis of the ellipse by 4.16 degrees at 19 GHz and 5.28 degrees at 28.5 GHz. Aluminum jig plates 2.54-cm thick are first cut into ellipses of 32.41 by 30.43 cm and 23.85 by 22.40 cm; then they are positioned and oriented correctly with respect to ellipsoidal axes for computer-controlled numerical machining.

3.4 Mechanical mounting

All the components are mounted in an L-shape frame of 106.7 × 106.7 × 61 cm as illustrated schematically in Fig. 2. The frame is made of structural aluminum angles with sufficient diagonal bracing to insure rigidity without blocking the radiating beams. The mounting of corrugated horns and offset ellipsoids was designed to allow adjustment for positioning, orientation, and focusing.

Owing to the uncertainty of the orbital position assignment for the COMSTAR satellite, the polarizations of the beacon signals were not precisely given. Therefore physical rotation of the feed frame around the beam axis is needed to match an arbitrary pair of orthogonal linear polarizations.* This required rotation is accomplished by a thrust ball bearing with 33-cm diameter circular opening in the front and a floating bearing in the back. These bearings are mounted on an aluminum supporting frame as shown in Fig. 10. This bearing mount is directly attached to steel I-beams through holes in the floor of the vertex equipment room of the 7-m antenna. The connection between the bearing mount and the steel I-beam allows differential thermal expansion between the indoor aluminum structure and the outdoor steel structure.

* This rotation is also needed for calibrating the amplitude and phases of the cross-polarized signals received from the satellite beacons.

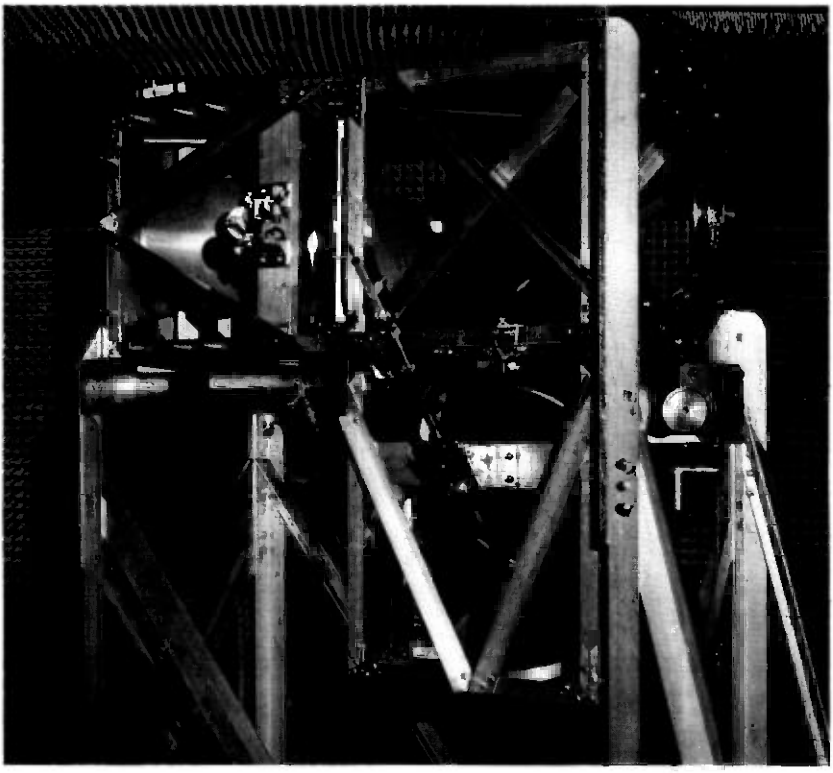


Fig. 10—A 19/28.5 GHz duo-polarization quasi-optical feed system (see Fig. 3) for the beacon measurements.

The 45-degree flat mirror, which consists of a 0.794-cm-thick aluminum jig plate, is mounted in front of the 33-cm bearing opening as shown in Fig. 9. Both azimuth and elevation orientations of the mirror can be adjusted around its center to facilitate experimental search for proper illumination of the subreflector. An oversized mirror of 38.10 by 53.34 cm oval shape was used in the initial measurents of the 7-m antenna. However, a smaller mirror of 30.48 by 43.18 cm shows very little truncation effect and is used in the beacon propagation experiment.

3.5 Measured results of the feed system

The design objective of the feed system is for each of the four offset launchers to illuminate the subreflector with a spherical wave of 15 dB taper over a 10-degree sector from a common phase center, with very low cross-polarized radiation as well as very low insertion loss of the quasi-optical diplexers. To achieve and demonstrate the desired performance, we aligned the components through pattern measurements in an ane-

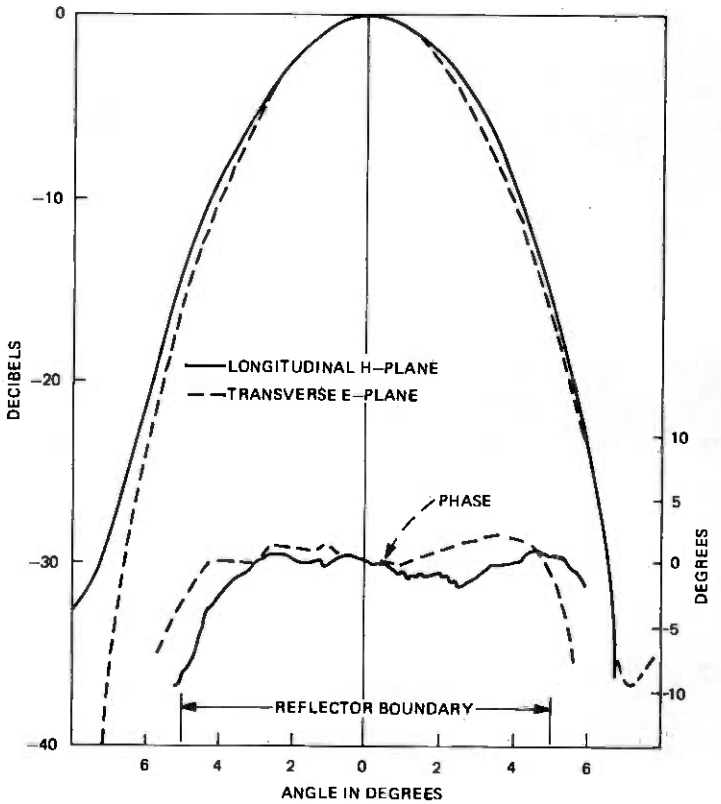


Fig. 11a—Measured radiation patterns of Feed No. 1 at 19 GHz (see Fig. 3).

choic chamber. To simulate the illumination of the 7-m offset Cassegrain antenna, the distance between the transmitting source horn used in the pattern measurements and the phase center of the receiving feed system was the same (526 cm) as that between the subreflector and its geometrical focus. Mechanical alignment was established before electrical measurements.

The measured insertion loss for a quasi-optical polarization or frequency diplexer was found to be 0.1 dB or less in each case. Amplitude and phase patterns were measured with respect to the common rotation center and the common optical boresight. After a few iterative adjustments of orientations and locations for both corrugated horn and ellipsoid, good agreement between measured and calculated patterns was achieved for each offset launcher. In particular, the measured patterns verified the calculated beamwidth, the effective phase center of the offset ellipsoids, and the approximate pattern symmetry predicted in the asymmetrical offset plane.

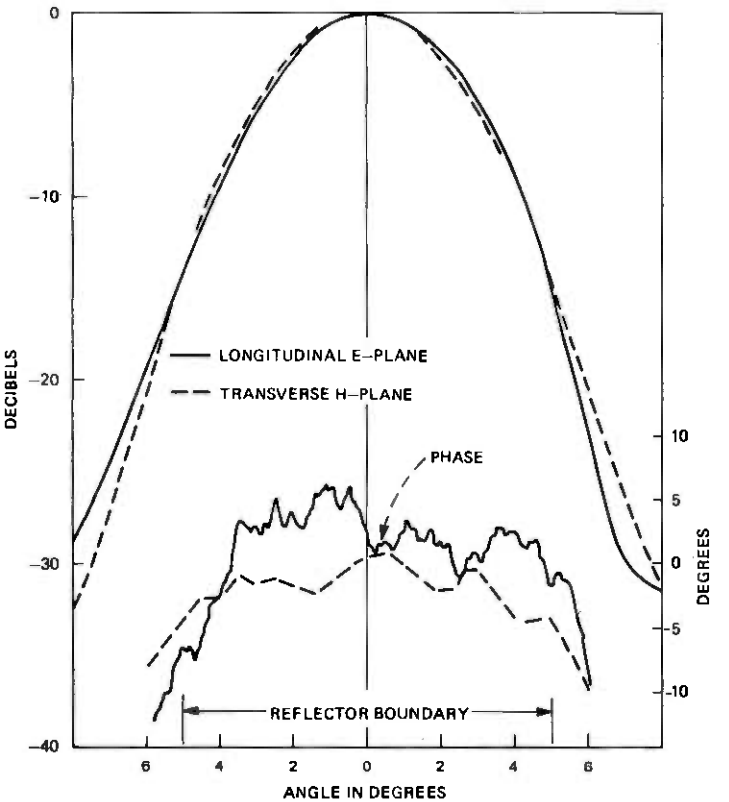


Fig. 11b—Measured radiation patterns of Feed No. 3 at 28.5 GHz (see Fig. 3).

The final measured patterns of Feeds No. 1 (19 GHz) and No. 3 (28.5 GHz) are shown in Fig. 11 for both E and H planes. Essentially identical measured patterns were obtained for the orthogonally polarized pair of Feeds No. 2 (19 GHz) and No. 4 (28.5 GHz). Following the nomenclatures used for horn reflector antennas, the longitudinal plane is the asymmetrical offset plane which divides the ellipsoid into two symmetrical halves and the transverse plane is the orthogonal principal plane. The measured phase patterns are less than ±5 degrees from a common spherical phase front while the measured amplitude patterns are within ±0.5 dB of perfect coincidence over the 15-dB pattern-width illuminating a 10-degree sector with respect to the common boresight. The cross-polarized radiation remained below −45 dB for all directions. It was found necessary to cover both the transmitting horn and the front bearing mount of the receiving feed system with absorbers to avoid excessive interactions. The ripples in the measured phase patterns of Fig. 11 were identified as the effects of the residual interactions.

The measured phase patterns of Fig. 11 imply an alignment accuracy of about $1/100$ beamwidth among the beams for the beacon receiver of the 7-m antenna. One notes that any moderate asymmetry or misalignment of the amplitude pattern of the feed system, which has little effect on the beam-pointing accuracy, can be tolerated in a communication system. However, good alignment of the feed amplitude patterns is essential to the stability of the differential phase between the beams of the 7-m antenna.

After the alignment measurements, the RF stages of the receiver[16] for the 19- and 28.5-GHz beacon experiment were mounted on the feed frame. The alignment was then checked using the beacon receiving system. Tests were made on the thermal stability of the differential phase between the two orthogonally polarized 19-GHz feeds. The change of the differential phase with respect to temperature was about 0.1 degree per 1°F and could be partially explained by the difference in waveguide lengths. Local heating of the polarization-diplexing grid shows a 0.2-degree change of the differential phase.

Since the calibration of the beacon receiver makes use of the rotation of the feed frame around the beam axis, the behavior of the differential phase during this rotation was examined. When the two orthogonal linearly polarized incident waves are of comparable magnitude (i.e., when the transmitting polarization is oriented at roughly 45 degrees with respect to the orthogonal polarizations of two 19-GHz receiving feeds), the measured differential phase remains essentially constant (within 0.1 degree) over a 10-degree rotation of the feed frame around the beam axis. When the two orthogonal linear polarizations are of vastly different magnitude, the measured differential phase can involve a substantial error because of a phase quadrature component arising from cross-polarization coupling. Even a polarization grid cannot effectively discriminate against an incident cross polarization of a magnitude much greater than that of the in-line polarization.

IV. MEASUREMENTS OF THE 7-METER ANTENNA

4.1 Transmitting sources

To measure the gain and radiation patterns of the Crawford Hill 7-m antenna, a weatherproof box, which contains transmitting sources at 19, 28.5, and 99.5 GHz was placed on an AT&T Long Lines tower at Sayreville, N.J., approximately 11 km from Crawford Hill. Two horn-lens antennas with polarization grids (providing cross-polarization discrimination better than 50 dB) are used for vertical and horizontal polarizations at the lower frequencies. Each antenna has a frequency diplexer and can transmit 19 and 28.5 GHz. Waveguide switches independently control the polarization or turn off each transmitter. The

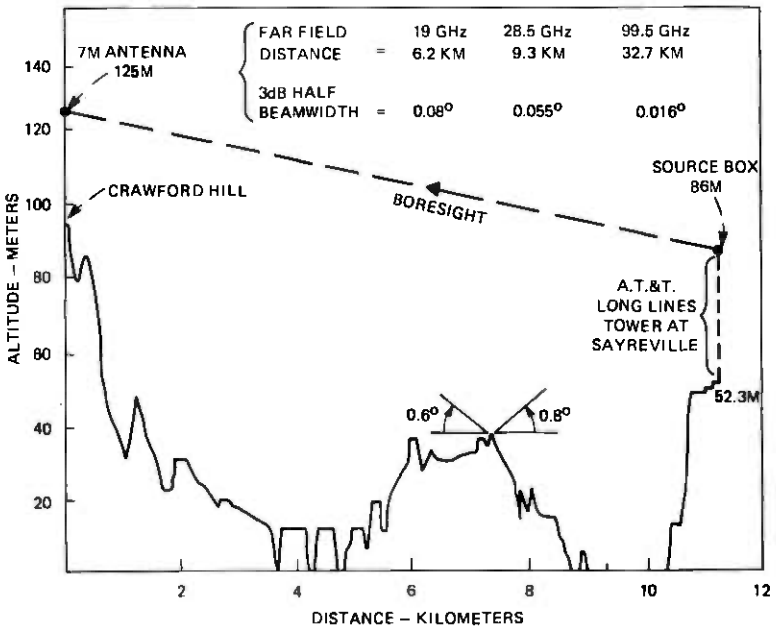


Fig. 12—Terrain profile of boresight range for 7-meter antenna.

antenna gains are 30 and 33 dB with half-power beamwidths of 6.5 and 5 degrees at 19 and 28.5 GHz, respectively. The sources at 19.04 and 28.56 GHz are 100-mw Gunn oscillators.

An antenna consisting of two cylindrical reflectors[17] with a dual-mode feed and a vertically polarized grid is used to transmit a 99.5-GHz signal from a 10-mw IMPATT source with only about 50-KHz FM noise. The oscillator is connected through an isolator directly to the feed horn. The antenna has a gain of 41.5 dB with a half-power beamwidth of 1 degree in the elevation plane.

Because of the distant location, the source box is remotely controlled from Crawford Hill. The beams from three source antennas were initially co-aligned before installation on the Sayreville tower. The expected power received by the 7-m antenna was about −30 dBm at all three frequencies.

4.2 Probing measurements of the incident field

In evaluating measured results of a large aperture antenna, it is necessary to know the field distribution incident on the aperture from a distant transmitting source. The path profile of the Sayreville to Crawford Hill measuring range is shown in Fig. 12. Although the well-elevated transmitting and receiving sites provide a clear line of sight,

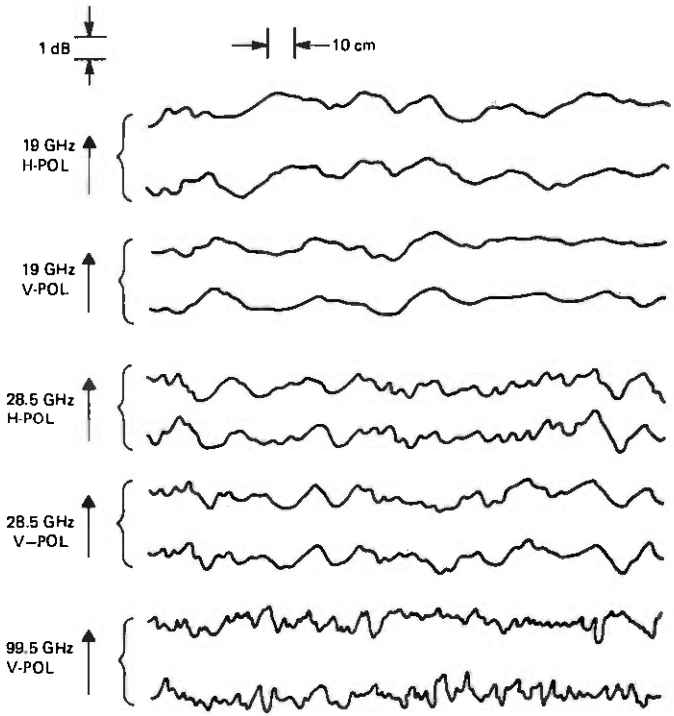


Fig. 13—Sample vertical scans for each of the five transmissions from source at Sayreville. Repeatable scans at 19 and 28.5 GHz indicate small reflections from the terrain. Uncorrelated scans at 99.5 GHz indicate atmospheric scintillations.

there still exist potential reflectors and scatterers of the millimeter-wave energy transmitted from Sayreville. A carriage and radial track mechanism was designed to permit a probe antenna to be moved along a diameter of a circular aperture to obtain field amplitude measurements. The incident field along four diameters of the 7-m aperture with 45-degree angular separation was scanned for each of the five possible states of transmission, i.e., vertical and horizontal polarizations at both 19 and 28.5 GHz and vertically polarized 99.5 GHz. Each scan was made twice to check on repeatability.

Figure 13 shows sample segments of scan pairs for all five transmissions. The good duplications of fluctuations at 19 and 28.5 GHz indicate that the systematic deviations in the field are results of specular reflections and diffractions. At 99.5 GHz, very little correlation exists between scans on a given diameter. Here the fluctuations are mostly atmospheric scintillations rather than terrain reflection. One notes that the 99.5-GHz transmitting antenna at Sayreville has a much narrower beamwidth than those of lower frequencies and the electrically rougher terrain is much less specular.

The spatial variations of the scan data suggest that most of the scattering comes at large angles with respect to the transmission path. This observation has been indeed confirmed by angular spectra obtained from a Fourier transform of the data. Hence, the terrain reflections should be negligible in the received power of the 7-m antenna pointing directly toward the transmitting source. The effect on the measured patterns will be only minor disturbance on sidelobes in the elevation plane at 19 and 28.5 GHz when the antenna is pointing below the source.

The received power in a gain-standard horn, which has a gain of the same order as that of our probe horn, will exhibit similar fluctuations across the aperture as those of Fig. 13. Since the fluctuations at 99.5 GHz are mostly atmospheric scintillations, the uncertainty arising from this fluctuation can be suppressed by taking an average of a number of comparisons between the gain standard and the 7-m antenna. However, at 19 and 28.5 GHz, the fluctuations are caused by terrain reflections; to reduce the uncertainty here, an average needs to be taken of numerous horn locations over the aperture. Analysis of the data resulting from probing measurements indicates that scanning the horn over a 1.4-m aperture segment gives a standard deviation of 0.4 dB at 19 GHz and 0.34 dB at 28.5 GHz.

4.3 Prime focus measurements

To provide an evaluation of surface tolerance as well as to locate the primary focal point of the 7-m reflector for subsequent installation of the subreflector, we first conducted 99.5-GHz prime focus measurements using a dual-mode feedhorn with 20-dB taper at the reflector boundary. The measured feed patterns are shown in Fig. 14. The focal region was probed with the feed until the best patterns were obtained. The measured patterns in the azimuth and elevation planes are shown in Fig. 15 together with the calculated pattern envelope assuming a perfect reflector surface.[2] The calculated patterns are approximately the same for azimuth and elevation. Expanded patterns not shown here indicated good agreement between measured and calculated half-power beamwidths (0.032 degree); however, the measured sidelobe levels are higher than the calculated values especially in the elevation plane. It is of interest to note that only near sidelobes in the elevation plane are higher than the corresponding lobes in the azimuth plane, whereas the far sidelobes in the two planes are essentially similar. Furthermore, the measured far sidelobe levels are consistent with an rms surface roughness of 0.1 mm. The excessive near sidelobe level in the elevation plane appears to be caused by a surface distortion of large-scale size. Since the reflector panels were set using a two-section template, a relative misalignment of these two sections could cause the elevation patterns we observed. This conjecture has been confirmed by pattern calculations

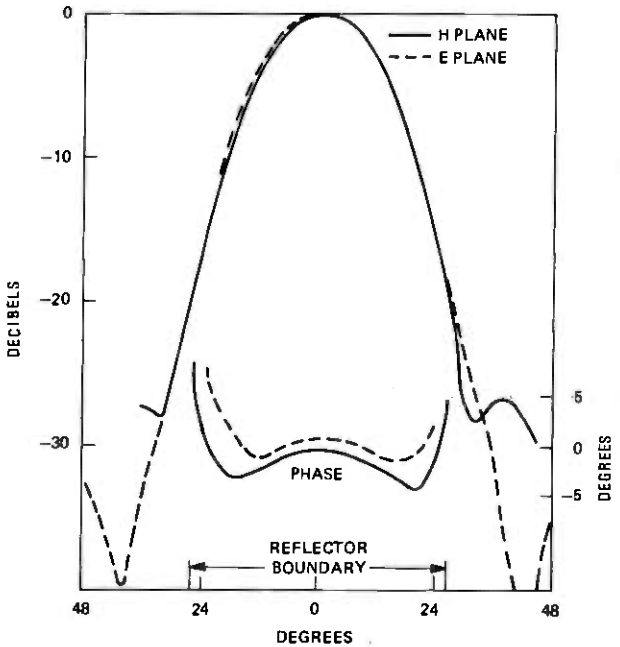


Fig. 14—Measured patterns of 99.5-GHz dual mode horn (0.98-cm diameter aperture) for prime-focus feed.

assuming a relative displacement (~0.15 mm) between two sections of the template.

Gain measurements of 99.5 GHz for the 7-m offset reflector using prime focus feed were made by comparison with a calibrated gain standard.[18] The comparisons were made at various times of day to obtain some estimate of the effects of scintillation. A time constant of about 1 s was used in the measuring set to smooth out the scintillation. Measurements indicate that gain variations due to diurnal variations of scintillations are generally less than about 0.2 dB. This observation has been also confirmed by the repeatability of measured patterns. A measured gain of 74.63 ± 0.45 dB was obtained from a sample of 10 comparisons with the gain standard taken on a clear quiet evening shortly after sunset.

The comparisons were made by measuring the difference in signal received by the gain standard (30.77 dB) and the 7-m antenna padded by a calibrated attenuator (40.05 dB). The error estimates of ±0.45 dB were obtained from the root-sum-square of the 3σ random errors: 0.14 dB for the calibrated attenuator, 0.16 dB for the gain standard, and 0.40 dB for the sample mean of 10 comparisons.

The theoretical gain of the antenna at 99.5 GHz having no roughness

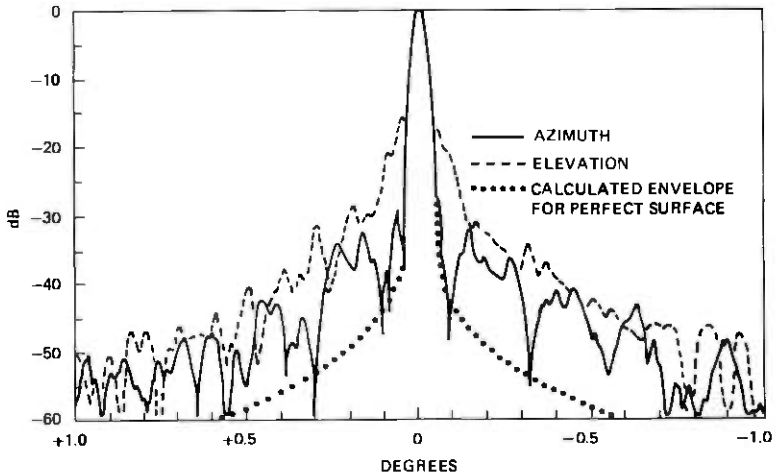


Fig. 15—99.5-GHz scan over a 2 degree range measured with prime-focus dual-mode feed of 20-dB taper. The difference between measured far sidelobe levels and the calculated pattern envelope for perfect surface is consistent[5] with 0.1 mm rms surface tolerance. The higher near-sidelobe level in the elevation plane stems from large-scale surface distortion.

is calculated as follows:

$$\begin{aligned} 77.26\ \text{dB} &= \text{Area Gain} \\ -1.56\ \text{dB} &= \text{Illumination Taper} \\ -0.08\ \text{dB} &= \text{Spillover} \\ -0.2\ \ \text{dB} &= \text{Feed Loss (Estimated)} \\ \hline 75.42\ \text{dB} &= \text{Calculated Gain for Perfect Surface} \end{aligned}$$

Using the formula $e^{-(4\pi\epsilon/\lambda)^2}$, where ϵ is the rms surface tolerance, we see the difference between measured and calculated gains, (0.79 ± 0.45) dB, corresponds to an rms roughness of (0.1 ± 0.03) mm.

Using a 20-dB taper corrugated horn (see Figs. 6 and 7) as a prime focus feed, we also measured the patterns of the 7-m offset reflector at 28.5 GHz as shown in Fig. 16. Excellent agreement between measured and calculated patterns were obtained in the azimuth plane. However, as with 99.5 GHz, there was a noticeable discrepancy between measured and calculated sidelobe levels in the elevation plane. The first sidelobe is −25 dB compared with −16 dB at 99.5 GHz.

The measured gain at 28.5 GHz was obtained by padding the 7-m reflector with a calibrated attenuator (39.93 ± 0.05) dB, and comparing with a calibrated gain standard (24.98 ± 0.08) dB.[18] Using 15 different locations for the gain standard, the measured gain was determined to be (64.7 ± 0.6) dB. The large 3σ error limit was the consequence of the data spread and is consistent with the results of the probing measurements. The expected prime focus gain at 28.5 GHz can be calculated as

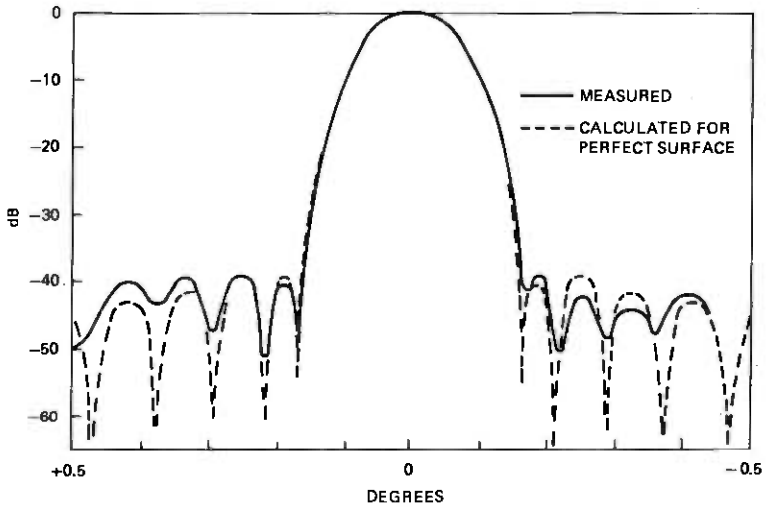


Fig. 16a—Measured 28.5-GHz azimuthal pattern with prime focus feed of 20-dB taper showing excellent agreement with calculated pattern for perfect surface.

follows:

66.42 dB	= Area Gain
−1.56 dB	= Illumination Taper
−0.08 dB	= Spillover
−0.1 dB	= Feed Loss (Estimated)
64.68 dB	= Calculated Gain for Perfect Surface

4.4 Subreflector alignment and measured results

The hyperboloidal subreflector is required to be confocal and coaxial with the paraboloidal main reflector. The primary focal point has been given by pattern measurements using prime focus feeds. Before installation of the subreflector, a laser beam was first fixed along the reflector axis. The subreflector was oriented with the aid of the laser beam reflected from a mirror attached to the bottom of the subreflector, centered on and perpendicular to its axis. The position of the subreflector was adjusted by interpreting the measured 99.5-GHz patterns until they were consistent with the prime-focus-fed patterns.

The Cassegrainian feed, used in the 99.05-GHz pattern measurements of the complete 7-m antenna including subreflector, consists of an offset ellipsoid and a dual-mode horn. The feed is essentially a scaled model of the 19-GHz offset launcher in the 19/28.5-GHz duo-polarization feed. Figure 17 shows that the measured 99.5-GHz feed patterns are almost the same as those of the 19/28.5-GHz feeds. Thus we can make direct comparison of measurements on the 7-m antenna at 99.5 GHz with the 19- and 28.5-GHz performance.

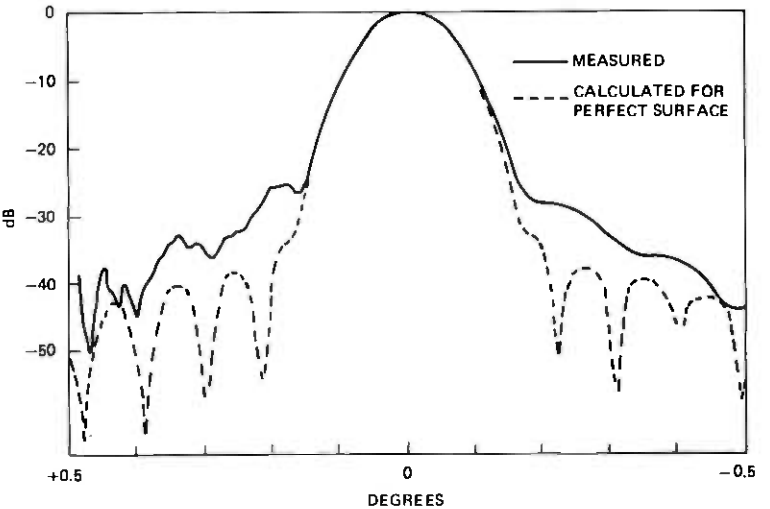


Fig. 16b—28.5-GHz elevation pattern with prime-focus feed of 20-dB taper. The measured first sidelobe level is −25 dB compared with −16 dB at 99.5 GHz in Fig. 15.

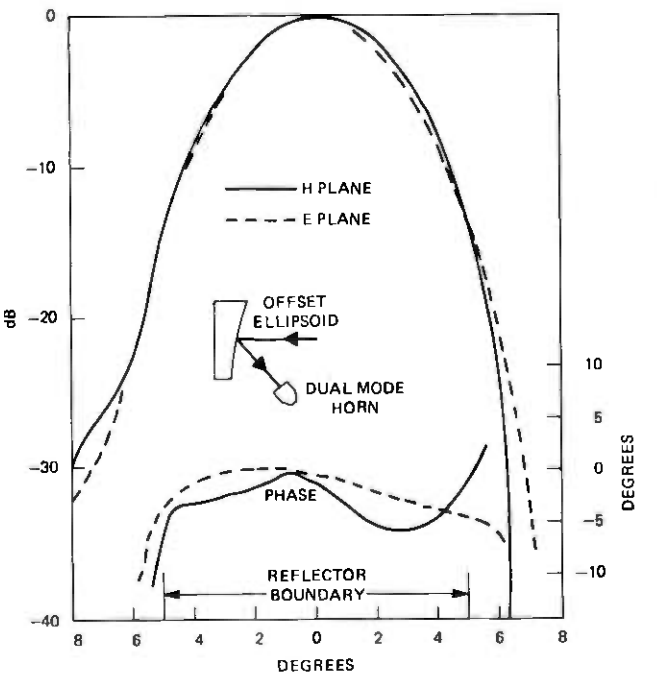


Fig. 17—Measured patterns of 99.5-GHz Cassegrainian feed consisting of an offset ellipsoid and a dual-mode horn.

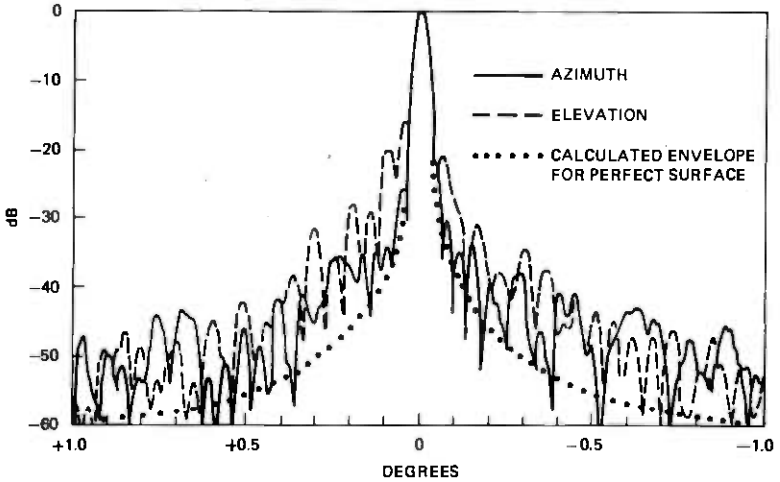


Fig. 18—99.5-GHz scan over a 2-degree range measured with Cassegrainian feed of 14-dB taper and compared with the calculated pattern envelope for perfect surface. The above patterns are consistent with those of prime-focus feed in Fig. 15.

Measured patterns are found to be insensitive to the location of the feed phase center as expected from the very long effective focal length of the antenna. With the feed phase center located on axis, the beam pointing of the Cassegrainian configuration agrees with that of the prime focus configuration to within 0.02 degree. Measured 99.5-GHz patterns are shown in Fig. 18 together with the calculated pattern envelope[15] for comparison. Measured half-power beamwidths agree with calculated values, whereas measured sidelobe levels are higher than calculated levels by about the same amount as in the prime focus measurements. Since the Cassegrainian feed has an illumination taper of 14 dB in contrast with the 20-dB taper of the prime focus feed, the sidelobe level is expected to be higher than that of the prime focus configuration. The measured first sidelobe level in the elevation plane is almost the same for prime focus and Cassegrainian configurations, because it is dominated by reflector distortion rather than illumination taper. It is seen that, as with the prime-focus case, the discrepancy between azimuth and elevation patterns is confined to the near sidelobe region, whereas the far sidelobe levels of two patterns merge together.

Pattern measurements were also taken for each of the four feeds in the duo-polarization 19/28.5-GHz quasi-optical feed assembly. The 7-m antenna was first tested with the 19/28.5-GHz feeds located at the on-axis position using both an oversized mirror of 38.10 by 53.34 cm oval shape and a smaller mirror of 30.48 by 43.18 cm. Measurements in each case showed the coincidence of four beam maxima for each polarization and frequency of the quasi-optical feed network. The smaller mirror

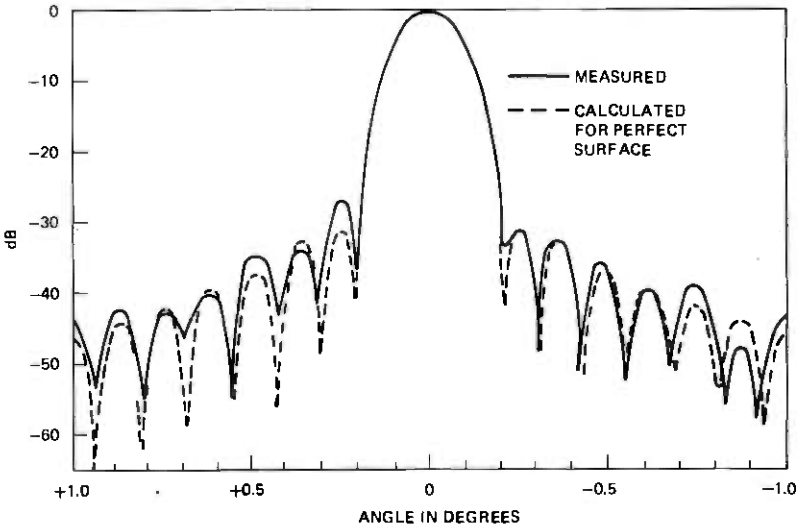


Fig. 19a—Measured 19-GHz azimuthal pattern with Cassegrainian feed of 15-dB taper and vertical polarization showing excellent agreement with the calculated pattern for perfect surface.

showed very little truncation effect in measured patterns. Following these on-axis measurements, the feed box for the beacon receiver together with the smaller mirror was moved to a position 0.5 degree off axis from the center of the vertex equipment room to allow clearance for an on-axis beam for millimeter-wave radio astronomy as shown in Fig. 9. As expected,[3,4] measurements showed very little difference between the 0.5-degree off-axis and on-axis beams for both 19 and 28.5 GHz.

Measured cross-polarized radiation for each on-axis beam remained below −40 dB in all directions with respect to the in-line polarization maximum, while that for each 0.5-degree off-axis beam is smaller than −39 dB. One notes that the optimum orientations of the polarization-grid diplexer for nulling the cross polarization are about 0.7 degree apart between two orthogonally polarized feeds. The above cross-polarization data were measured using a compromise orientation of the polarization grid.

Figures 19 and 20 show comparisons between measured and calculated patterns[15] at 19 and 28.5 GHz. The measured patterns were obtained from vertically polarized feeds at on-axis position with the smaller 45-degree mirror, and remained essentially the same for other combinations of polarization and mirror at both on-axis and 0.5-degree off-axis feed positions. Good match between calculated and measured azimuthal patterns is illustrated in Figs. 19(a) and 20(a), whereas the agreement between calculated and measured elevation patterns is, again, less sat-

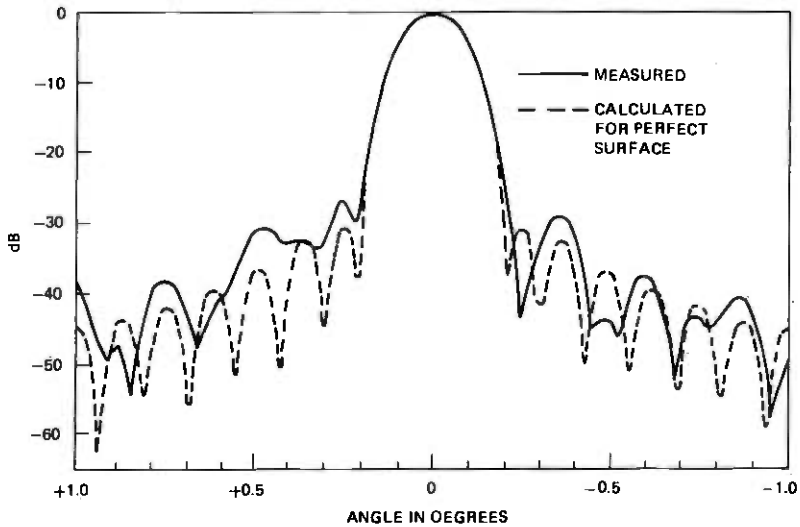


Fig. 19b—Measured 19-GHz elevation pattern with Cassegrainian feed of 15-dB taper and vertical polarization showing fair agreement with the calculated pattern for perfect surface.

isfactory as shown in Figs. 19b and 20b. The measured sidelobes of all elevation patterns have generally shown, especially in Fig. 20b, a period twice that of the calculated value. This observation supports the conjecture about a relative misalignment of two sections of the template

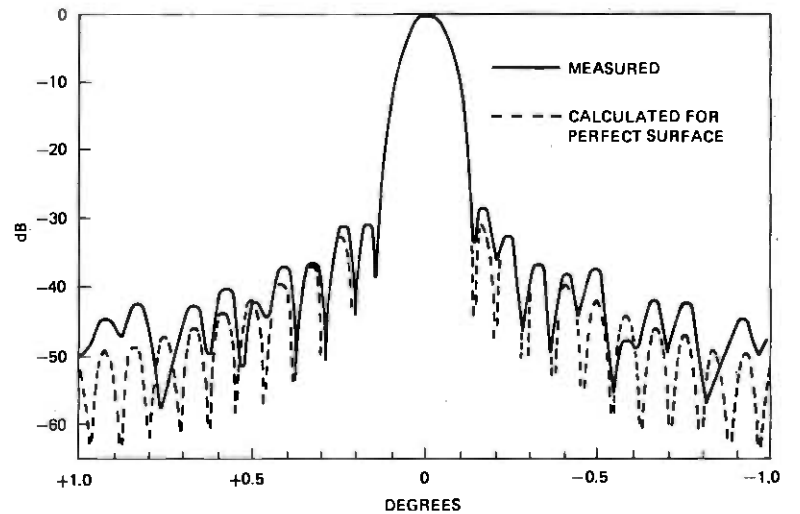


Fig. 20a—28.5-GHz azimuth pattern with Cassegrainian feed of 15 dB taper and vertical polarization. The measured pattern is consistent with that in Fig. 16a for prime focus feed of 20-dB taper.

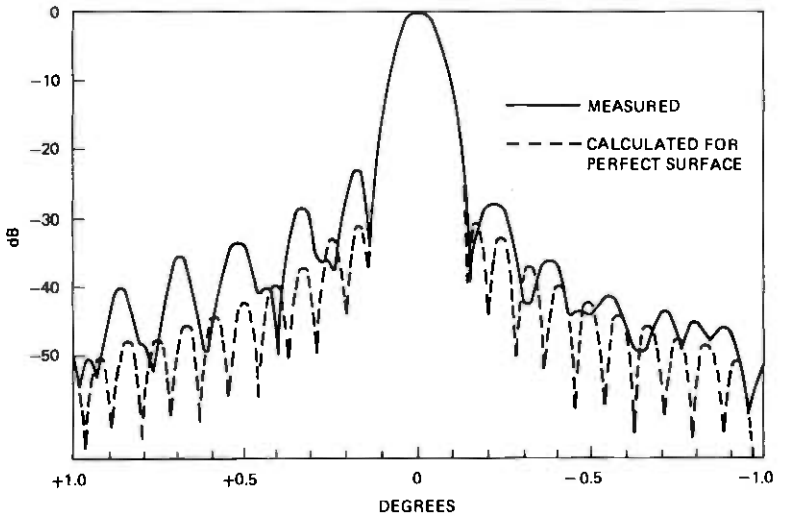


Fig. 20b—28.5-GHz elevation pattern with Cassegrainian feed of 15-dB taper and vertical polarization. The measured pattern is consistent with that in Fig. 16b for prime focus feed of 20-dB taper.

as discussed in Section 4.3. Measured 19-GHz sidelobe levels 1 degree away from the beam maximum are −43 dB in the azimuth plane and −38 dB in the elevation plane. These levels are of interest in avoiding interference from future adjacent satellites in synchronous orbits.

V. CONCLUDING REMARKS

The expected performance of the Crawford Hill 7-m antenna and its associated feed systems has been realized. This antenna is the first large offset Cassegrain in operation. Comparison between prime-focus-fed and Cassegrain-fed patterns shows very little degradation due to any surface imperfection or misalignment of the subreflector.

Comparison with a calibrated gain standard showed the difference between the 99.5-GHz measured and calculated prime-focus gains to be (0.79 ± 0.45) dB, which implies an rms surface error of about 0.1 mm. The measured 99.5-GHz azimuthal pattern appears to indicate a surface error of this magnitude, whereas the elevation pattern shows skewed shape and unexpectedly high near-in sidelobe level. Computer simulations have indicated that the distorted elevation pattern can be caused by a relative displacement (~0.15 mm) between two sections of the template used to calibrate the reflector panels. This explanation is also consistent with the measured gain because it is only accompanied by a small gain reduction (~0.15 dB).

Pattern measurements using the quasi-optical 19/28.5 GHz dual-polarization feed assembly have shown the coincidence of the four beams.

Table I — Derived Cassegrainian gain in decibels with 15-dB feed taper

Frequency (GHz)	99.5	28.5	19
Area Gain	77.26	66.4	62.88
Illumination Taper	−0.93	−0.93	−0.93
Spillover*	−0.35	−0.4	−0.45
Feed and Multiplexing†	−0.2	−0.2	−0.2
Surface Tolerance	−0.8	−0.1	−0.05
Derived Gain	74.98	64.77	61.25
Error Estimates	±0.5	±0.25	±0.25
Gain Efficiency	59.2%	68.7%	68.7%

* The estimates for spillover loss are higher at 19 and 28.5 GHz because of truncations in the quasi-optical 19/28.5 GHz duo-polarization feed.

† No multiplexing is involved at 99.5 GHz, whereas both frequency and polarization diplexing losses are included in the estimates at 19 and 28.5 GHz.

Cross-polarized radiation is −40 dB or less in all directions throughout each beam. The measured results have now confirmed the theoretical prediction[2] that there should be very little cross-polarized radiation from an offset Cassegrainian antenna with a large effective F/D ratio if the feed radiation is free of cross polarization.

Good agreement between calculated and measured patterns is shown in the azimuthal plane, whereas the comparison is less satisfactory in the elevation plane. At 1 degree away from the beam maximum, the 19-GHz sidelobe level is −43 dB in the azimuth plane and −38 dB in the elevation plane. Since the synchronous orbit is seldom close to the elevation plane of a ground station antenna, these measured sidelobe levels have practically achieved the objective of 40-dB discrimination between adjacent synchronous satellites at 1-degree spacing.

The gain measurement of the Cassegrainian configuration is hampered by the sensitivity of the harmonic mixer to the temperature difference between indoor and outdoor environments and by a lot of cable movement in addition to the terrain reflection problem at 19 and 28.5 GHz. However, having determined the main reflector surface tolerance by prime focus measurements, we can derive the Cassegrainian gains as shown in Table I.

Multiple-beam operation has been achieved with a 0.5-degree off-axis beam for beacon feed and an on-axis beam for millimeter-wave radio astronomy. Measurements showed very little difference between 0.5-degree off-axis and on-axis beams at both 19 and 28.5 GHz.

VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge the following valuable contributions to this work. Ford Aerospace and Communications Corporation carried out the mechanical design, built, and installed the main structure. M. J. Grubelich and K. N. Coyne consulted on many mechanical design

problems, including the main structure, subreflector, and feed system. M. J. Gans designed the vertex room windows and calculated the surface distortions from mechanical measurements. H. H. Hoffman collaborated on the antenna-receiver interface. D. C. Hogg proposed the antenna configuration. E. A. Ohm recognized the large beam scanning capability of the antenna configuration and modified the vertex equipment room and subreflector support structure design to facilitate launching beams at large off-axis angles while keeping the gravitational pointing error small. He also designed and supervised the fabrication of the subreflector. F. A. Pelow supplied the quasi-optical diplexers. A. Quigley and members of the shop built most of the feed hardware. H. E. Rowe examined detailed surface tolerance effects. J. Ruscio and T. Fitch took care of receiver cables in the cable wraps. R. A. Semplak provided the 99.5-GHz Cassegrainian feed. R. H. Turrin designed the carriage and radial track mechanism for probing measurements. D. Vitello programmed the pattern calculations. G. V. Whyte implemented the subreflector mount.

REFERENCES

1. C. Dragone and D. C. Hogg, "The Radiation Pattern and Impedance of Offset and Symmetrical Near-Field Cassegrainian and Gregorian Antennas," IEEE Transactions, *AP-22*, May 1974, pp. 472–475.
2. T. S. Chu and R. H. Turrin, "Depolarization Properties of Offset Reflector Antennas," IEEE Transactions, *AP-21*, May 1973, pp. 339–345.
3. E. A. Ohm, "A Proposed Multiple-Beam Microwave Antenna for Earth Stations and Satellites," B.S.T.J., *53*, No. 8 (October 1974), pp. 1657–1666.
4. E. A. Ohm and M. J. Gans, "Numerical Analysis of Multiple-Beam Offset Cassegrainian Antennas," AIAA Paper #76-301. AIAA/CASI 6th Communication Satellite Systems Conference, Montreal, Canada, April 5–8, 1976.
5. C. Dragone and D. C. Hogg, "Wide Angle Radiation Due to Rough Phase Fronts," B.S.T.J., *42*, No. 7 (September 1963), pp. 2285–2296.
6. P. F. Goldsmith, "A Quasi-Optical Feed System for Radio Astronomical Observations at Millimeter Wavelengths," B.S.T.J., *56*, No. 8 (October 1977), pp. 1483–1501.
7. T. S. Chu and W. E. Legg, "A 19/28.5 GHz Quasi-Optical Feed for an Offset Cassegrainian Antenna," IEEE/APS Symposium, Amherst, Mass., October 1976.
8. J. A. Arnaud and F. A. Pelow, "Resonant-Grid Quasi-Optical Diplexers," B.S.T.J., *54*, No. 2 (February 1975), pp. 263–283.
9. T. S. Chu, M. J. Gans, and W. E. Legg, "Quasi-Optical Polarization Diplexing of Microwaves," B.S.T.J., *54*, No. 10 (December 1975), pp. 1665–1680.
10. I. Anderson, "Measurements of 20 GHz Transmission Through a Radome in Rain," IEEE Transactions, *AP-23*, September 1975, pp. 619–622.
11. C. Dragone, "Reflection and Transmission Characteristics of a Broadband Corrugated Feed: A Comparison Between Theory and Experiment," B.S.T.J., *56*, No. 6 (July–August 1977), pp. 869–888.
12. T. S. Chu, "Geometrical Representation of Gaussian Beam Propagation," B.S.T.J., *45*, No. 2 (February 1966), pp. 287–299.
13. H. Kogelnik and T. Li, "Laser Beams and Resonators," Proc. IEEE, *54*, October 1966, pp. 1312–1329.
14. M. J. Gans and R. A. Semplak, "Some Far-Field Studies of an Offset Launcher," B.S.T.J., *53*, No. 7 (September 1974) pp. 1319–1340.
15. J. S. Cook, E. M. Elam, and H. Zucker, "The Open Cassegrain Antenna: Part 1—Electromagnetic Design and Analysis," B.S.T.J., *44*, No. 7 (September 1965), pp. 1255–1300.

16. H. W. Arnold, D. C. Cox, H. H. Hoffman, R. H. Brandt, R. P. Leck, and M. F. Wazowicz, "The 19 and 28 GHz Receiving Electronics for the Crawford Hill COMSTAR Beacon Propagation Experiment," B.S.T.J., this issue, pp. 1289–1329.
17. C. Dragone, "An Improved Antenna for Microwave Radio Systems Consisting of Two Cylindrical Reflectors and a Corrugated Horn," B.S.T.J., *53*, No. 7 (September 1974), pp. 1351–1377.
18. T. S. Chu and W. E. Legg, "Gain of Corrugated Conical Horns," IEEE/APS Symposium, College Park, Md., May 1978.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

COMSTAR Experiment:

The 19- and 28-GHz Receiving Electronics for the Crawford Hill COMSTAR Beacon Propagation Experiment

By H. W. ARNOLD, D. C. COX, H. H. HOFFMAN, R. H. BRANDT, R. P. LECK, and M. F. WAZOWICZ

(Manuscript received January 10, 1978)

This paper describes the receiving electronics built at the Bell Laboratories Crawford Hill facility at Holmdel, New Jersey to use the 19- and 28-GHz beacons on the COMSTAR satellites for propagation measurements. The receiving system accurately determines attenuation, differential phase, depolarization, bandwidth limitations and angular scatter of these signals produced by rain. This highly reliable system operates continuously and unattended; it automatically reacquires the beacon signals after dropout due to severe attenuation or momentary power outage. Correlations among strong and weak signal components are used to permit detection of weak cross-polarized signals during severe fading. Receiver noise bandwidths as low as 1.6 Hz are used. A high degree of phase stability is achieved in all circuits and components.

I. INTRODUCTION

The receiving electronics for the COMSTAR beacons has placed strong demands on technology in several areas in order to meet the requirements of the propagation experiments.[1,2,3] The receiving system built at the Bell Laboratories Crawford Hill facility includes a precision antenna[4] and the receiving electronics necessary to make maximum use of the 19- and 28-GHz beacon signals radiated by the COMSTAR satellites. The receiving electronics is the subject of this paper.

Continuous unattended operation is required so that all significant weather events are included in the resulting data base; thus, a very high

degree of reliability in the receiving electronics is required, and automatic reacquisition of the beacon signal after dropout due to severe attenuation or momentary power outage is essential. Since relative phases of the many signal components must be precisely measured, the phase stability of all circuits and components demanded careful attention. Also, circuit arrangements had to be devised to ensure that signals which were later to be compared in phase traversed a common path through high-gain amplifiers and other phase-sensitive equipment.

In order to obtain the maximum possible measuring range using the modest powers radiated by the satellite beacons, very narrow receiver noise bandwidths are required. This puts a premium on the stability of the source oscillators in the satellites and the local oscillators in the earth stations. The receiver includes an AFC circuit with built-in memory to facilitate reacquisition after loss of signal. The feature also permits easy return to propagation measurements after use of the antenna system for radio astronomy studies during clear weather periods. Maximum use was made of known correlations among strong and weak signal components to permit detection of weak cross-polarized signals during severe fading. The following is an account of how these objectives were achieved, together with a description of the resulting apparatus and its operating characteristics. General design considerations are in Section II; Sections III and IV cover 19-GHz and 28-GHz receiver channels, respectively. Local oscillators and frequency control techniques are discussed in Section V. Section VI covers polarization switch synchronization. Data collection equipment is described in Section VII; Section VIII covers the receiver calibration source. Receiver performance and some sample data are included in Section IX.

II. RECEIVER DESIGN CONSIDERATIONS

The multichannel satellite beacon receiver must measure and record the signal amplitudes and phases listed in Table I for all rain events in order to satisfy the needs of the propagation experiment. Because of beacon[1] and rain[5] characteristics the receiver must: (*i*) have narrow (1.6 to 24 Hz) final noise bandwidths, (*ii*) keep the receiver frequencies within the narrow IF filter bandwidths (BW) as the beacon frequencies vary, (*iii*) hold local oscillator (LO) frequencies within a few Hz of their last known frequency for several minutes when rain attenuates the beacon signals below the frequency tracking threshold or when the primary power drops out, (*iv*) discriminate against the 1-kHz polarization switching sidebands while reacquiring the beacon signals automatically after long periods of loss of signal, (*v*) synchronize receiver polarization switches with the beacon polarization switch, and (*vi*) perform these functions reliably and automatically to permit continuous unattended operation. During clear air conditions the earth-station antenna[4] is used for radio astronomy

Table I — Signal amplitudes and phases

Receiver output	Description
Amplitude, 19 GHz	
A19H10	Horizontal copolarized,* 10-Hz IF BW
A19V10	Vertical copolarized,† 10-Hz IF BW
A19XH10	Horizontal crosspolarized,‡ 10-Hz IF BW
A19XH1	Horizontal crosspolarized,‡ 1-Hz IF BW
A19XV10	Vertical crosspolarized,†† 10-Hz IF BW
A19XV1	Vertical crosspolarized,†† 1-Hz IF BW
A19OA1	Off-axis receiving beam, 1-Hz IF BW
Phase difference, 19 GHz	
ϕ19V-H	(Vert. copol.†)-(horiz. copol.*), 10-Hz IF BW
ϕ19V-XV	(Vert. copol.†)-(vert. crosspol.††), 10-Hz IF BW
ϕ19V-XH	(Vert. copol.†)-(horiz. crosspol.‡), 10-Hz IF BW
Amplitude, 28 GHz	
A28V15C	Vertical copolarized,† 15-Hz IF BW, carrier
A28V1.5C	Vertical copolarized,† 1.5-Hz IF BW, carrier
A28V15U	Vertical copolarized,† 15-Hz IF BW, upper sideband
A28V15L	Vertical copolarized,† 15-Hz IF BW, lower sideband
A28XV15C	Vertical crosspolarized,†† 15-Hz IF BW, carrier
A28XV1.5C	Vertical crosspolarized,†† 1.5-Hz IF BW, carrier
A28OA1.5	Off-axis receiving beam, 1.5-Hz IF BW
Phase difference, 28 GHz	
ϕ28VC-XVC	(Vert. copol.† carrier)-(vert. crosspol.†† carrier), 15-Hz IF BW
ϕ28VC-U	(Vert. copol.† carrier)-(upper sideband), 15-Hz IF BW
ϕ28VC-L	(Vert. copol.† carrier)-(lower sideband), 15-Hz IF BW
Phase difference, crossband	
ϕ19V-28VC	(19-GHz vert. copol.,† 10-Hz IF BW) −(28-GHz vert. copol. carrier, 15-Hz IF BW)

* Horizontal copolarized is transmit horizontal and receive horizontal.
† Vertical copolarized is transmit vertical and receive vertical.
‡ Horizontal crosspolarized is transmit horizontal and receive vertical.
†† Vertical crosspolarized is transmit vertical and receive horizontal.

and other studies so the receiver must be easily restarted by people not intimately familiar with it.

Signal amplitudes and phases must remain accurate over receiver temperature changes of ±15°C and for signal attenuations of >30 dB. For example, amplitude and phase differences between the two copolarized 19-GHz channels must remain within 0.5 dB and 2°.[6] This requires careful attention to differential temperature control and to linearity.

Minimizing the number of frequency conversions is desirable to minimize receiver complexity and the number of spurious mixing products.

Since the 19- and 28-GHz beacon signals are derived from a common oscillator, they have the same frequency fluctuations. Thus, extended measuring range can be provided in the 28-GHz channels and in low signal 19-GHz channels (off axis and cross-polarization) if: (*i*) the corresponding 28-GHz and 19-GHz receiver LO sources are common, (*ii*) frequency fluctuations in the beacon and in LOs are tracked out with a common oscillator in a loop locked in frequency or phase to the 19-GHz vertically polarized signal, the signal that will experience the least at-

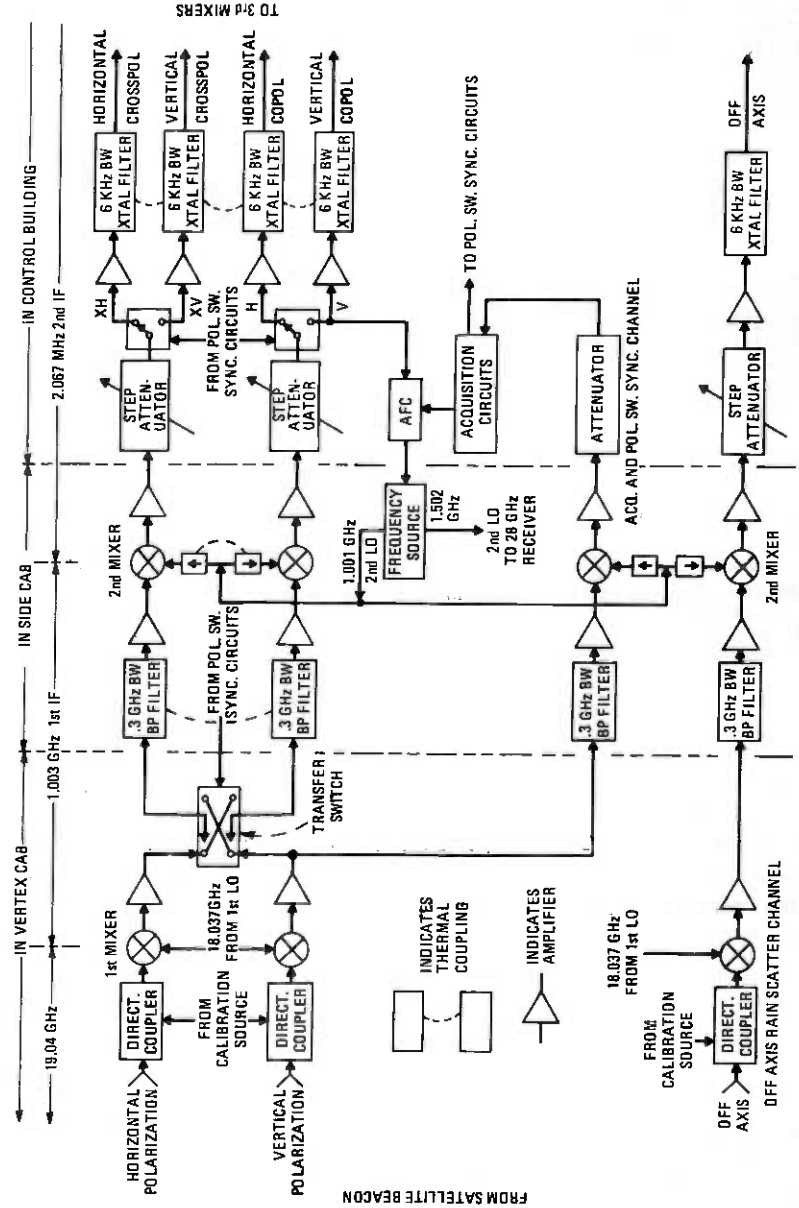


Fig. 1—19-GHz receiver channels: antenna feeds through second IF.

tenuation,[5] and (*iii*) very narrow band (~1 Hz BW) IF filters are used in the extended range channels. The provisions necessary for extending measuring range are easier to implement if all 19- and 28-GHz receiver IFs and LOs are in a 2:3 ratio.

The antenna has a small equipment room, the vertex cab, near the vertex of the main reflector, above both the azimuth and elevation axes; another room, the side cab, is to one side of the elevation axis but above the aximuth axis. The control building is about 15 meters (50 feet) away from the antenna at ground level (see Ref. 1, Fig. 3). The receiver is distributed among the three equipment locations to optimize noise performance and phase and amplitude stability, taking into account space limitations in the various equipment rooms.

The receiver is packaged by functional groups, e.g., bandpass filters, mixers and IF amplifiers for all the second frequency conversions are packaged together. This packaging approach maximizes differential phase and amplitude stability by minimizing temperature differential between similar components in different channels and also allows the receiver to be built in many separable blocks that may be "debugged" individually without complex interaction.

Since power line transients and momentary power outages are expected during heavy rain, all oscillators, filter stabilizing ovens and frequency memory registers are powered by batteries charged continuously from the power line.

III. 19-GHz RECEIVER CHANNELS

The 19-GHz receiver channels from the antenna feeds through the second IF crystal filters are shown in Fig. 1. The 19-GHz beacon signals are received with vertically and horizontally polarized feeds whose resulting antenna beams are pointed at the satellite. Another 19-GHz feed is located so its antenna beam is pointed toward an unoccupied synchronous orbit location about 0.74° off-axis from the satellite. This off-axis beam detects signals scattered by rain from the satellite beacon path into other potentially useful paths. This scattered signal is another possible source of cochannel crosstalk in multisatellite systems.[7,8]

Test signals from a calibration source described in Section VIII are fed into directional couplers between the antenna feeds and the first mixers. These calibrated test signals have adjustable amplitudes, are polarization switched and have adjustable simulated "crosspolarization" levels. The short term frequency stability of the calibration source is representative of the stabilities of the satellite beacons.

Signals are mixed with the 18.037-GHz first LO in the Schottky-diode balanced first mixers. Parametric amplifiers are not used because (*i*) they would contribute excessive amplitude and phase instability and (*ii*) adequate measuring range is more economically and easily provided by

narrow-band IF filtering. These mixers and associated low-noise broadband IF preamplifiers (~25 dB gain) are mounted on the feed horns. The first mixer and IF preamplifier single sideband noise figure (NF) of approximately 6.5 dB essentially determines the overall receiver NF. The remainder of the receiver contributes <0.1 dB. The first LO is distributed from a common source to all 19-GHz channels to preserve phase in the mixing process and to insure that the frequency fluctuations are correlated among the channels; isolating power dividers and ferrite isolators (see Fig. 5) insure >70 dB isolation between receiver channels through this common LO path. The 1.003 GHz first IF is dictated by the 1.056 GHz required bandwidth for the modulation sidebands in the 28 GHz receiver, by the need to keep 19- and 28-GHz IFs and LOs in a 2:3 ratio, and by the need to derive corresponding LOs from common sources.

After 30 dB more IF amplification the vertically and horizontally polarized channels are transfer-switched in synchronism with the beacon polarization switch. In this switch the copolarized signals (V and H) in adjacent time slots are switched into one receiver channel and the cross-polarized signals (XV and XH) are switched into the other channel. The accuracy of the differential amplitude and phase between the copolarized signals is very critical in the calculation of depolarization for other polarization angles.[6] The transfer switching insures that the phase and amplitude variations in most of the filters, amplifiers and long cable runs in the copolarized signal channel will affect the V and H signals identically. Thus, these variations will not affect the differential amplitude and phase measurements.

As indicated in Figure 1, the first mixers, first IF amplifiers and transfer switch are on the antenna feed assembly in the vertex cab. The mechanical rigidity of the feed assembly is adequate to ensure <0.5° differential phase variation (measurement limits) due to differential mechanical motion of feed components. Thermal differential phase variation for the feed assembly and associated receiver components is <0.2° per °C.

A separate 19-GHz receiver channel, used for signal acquisition and polarization switch synchronization, branches off before the transfer switch.

Cables run from the vertex cab to the side cab after the transfer switch but before 0.3-GHz BW bandpass filters (BPFs); these filters constrain the noise bandwidth to prevent noise from saturating the following broadband IF amplifiers and are mounted on an aluminum plate to minimize the temperature differential between them.

The second mixers reject image noise by more than 19 dB by phase-cancelling techniques (single sideband mixing) to permit the large fractional frequency spread between the first IF (1.003 GHz) and the

second IF (2.067 MHz). The 1.001 GHz second LO also is distributed from a common source to all 19-GHz channels. Isolating power dividers and ferrite isolators (see Fig. 6) insure >65 dB isolation between receiver channels through this common LO path. The second LO automatic frequency control (AFC) that follows average long-term frequency changes is described in Section V. The 5-MHz BW of IF amplifiers following the second mixers is narrow enough to prevent amplifier saturation on noise but wide enough both to preserve the rise time of the polarization switched signals and to render filter phase stability of little concern. The second IF amplifiers drive 50 meters of cable from the side cab to the control building.

Step attenuators match signal levels between the constant-gain low level (signals $S \leq$ noise N) front part of the receiver and the constant-gain high level ($S > N$) back part. This permits maximizing dynamic range by setting the clear air signal outputs near maximum. The second IF polarization switches separate the time-sequenced signals (V, H, XV, and XH) into separate channels. After further amplification, quartz crystal BPFs do the following: (*i*) further restrict the noise bandwidth, (*ii*) reject image frequency noise >20 dB for the third frequency conversion, (*iii*) significantly increase the rise time of the signal pulses and (*iv*) reduce second IF switch transients at the third conversion image frequency. The 6-kHz BW of these filters is as narrow as possible consistent with differential phase stability requirements. The four filters for the main beam channels (V, H, XV, and XH) are mounted together in a solid aluminum enclosure to minimize temperature differences.

The 2.067-MHz second IF frequency is a compromise restricted on the low side by LO noise close to the second LO frequency, by the need to preserve the rise time of the polarization-switched signals, and by the level of switching transients near the second IF frequency. The choice is restricted on the high side by phase stability considerations for the crystal BPFs and by the desire to minimize the number of receiver frequency conversions.

The acquisition and polarization switch synchronization channel preserves the 1-kHz polarization-switching signal from the beacon, uncontaminated by receiver switching. This channel is similar to the other receiver channels down to the crystal BPFs, except for the omission of switches. The acquisition and polarization switch synchronization circuits described in Sections V and VI contain their own automatic gain controlled IF amplifiers and crystal BPFs.

The off-axis rain scatter channel is essentially identical to the main beam channels except for the omission of polarization switches.

The 19-GHz receiver channels, continuing from the second IF crystal BPFs through the amplitude and phase detectors, are shown in Fig. 2. As indicated earlier, image rejection for the third frequency conversion

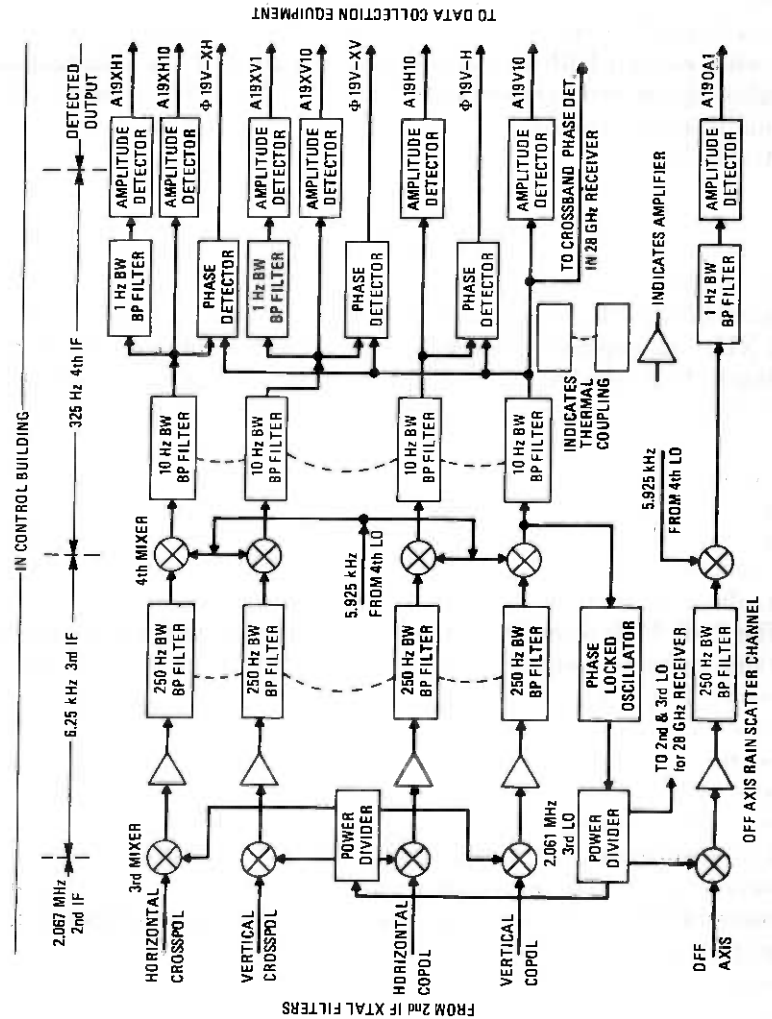


Fig. 2—19-GHz receiver channels: second IF through amplitude and phase detectors.

is done in the 6-kHz crystal BPFs. These stable high-Q filters permit a large fractional frequency spread between the second IF (2.067 MHz) and third IF (6.25 kHz). The 6.25 kHz third IF is constrained by phase stability considerations in the crystal BPFs and the need for at least 20 dB of filter rejection at the image frequency. The third mixers are double balanced (ring diode) with >30 dB signal to LO and LO to signal isolation. This isolation, along with the isolation in the LO power dividers (see Fig. 8), produces >80 dB isolation (measurement limit) between receiver channels through the common third LO path.

Because the third LO is phase-locked to the V copolarized signal channel using the phase locked loop (PLL) described in Section V, the short-term instabilities of the beacon oscillator and all receiver LOs are removed from all receiver channels at the third mixers. Thus, all third and fourth IF signals are as stable as the 325-Hz PLL reference as long as the PLL is locked.

Single resonator (single pole pair) active BPFs with 250 Hz BW follow 6.25 kHz IF amplifiers. These filters further restrict the noise BW, further reduce the polarization switching sidebands and receiver switching transients, and reject image noise and switching transients for the following fourth frequency conversion. The high input impedance of active analog multipliers used for the fourth mixers combined with the low output impedance of the operational amplifier supplying the 5.925-kHz fixed-frequency fourth LO (see Fig. 8) result in >80 dB isolation between receiver channels through the common fourth LO path.

Single resonator, 10-Hz 3-dB BW, active BPFs (~16 Hz equivalent noise BW) follow the fourth mixers. These BPFs are the final IF filters for the copolarized signal channels and for phase and amplitude measurement to moderate attenuation levels in the crosspolarized signal channels.

The specific third and fourth IFs are chosen to minimize the levels of the mixing products from the 1-kHz switching sidebands, both on the signal and the image sides of the third and fourth LOs. More than 60 dB filtering of the 1 kHz switching sidebands by the cascaded third and fourth IF BPFs insures that the phase ripple produced by the sidebands cannot exceed ±0.1°.

All of the third IF and 10-Hz fourth IF active BPFs and the fourth mixers are constructed with low-temperature-coefficient components. They also are enclosed in a temperature stabilized oven with <½°C internal temperature variation to maintain phase stability. The third and fourth IF filter Qs are nearly the same so that, assuming the same component variations, their contributions to the overall phase stability would be nearly equal. These filters were aged at oven temperature (65°C) before final alignment. The fourth IF BPF outputs are buffered through

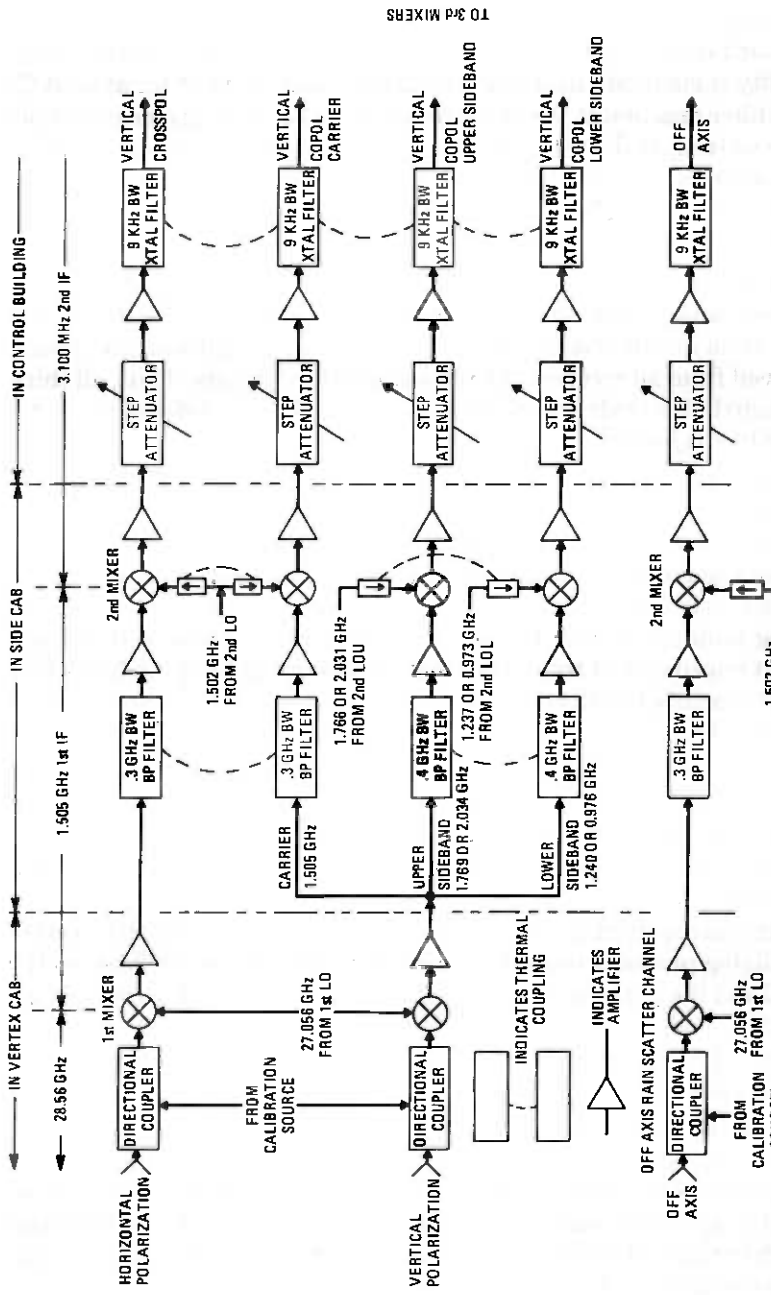


Fig. 3—28-GHz receiver channels: antenna feeds through second IF.

low-pass high-pass operational amplifiers that suppress low frequency (1/f) noise and broadband noise in the active BPF outputs.

Linear amplitude detectors following the fourth IF filters are full wave rectifiers comprising diodes in feedback paths of wideband low-drift operational amplifiers. These detectors are linear in dc output vs. signal input (voltage) to within ±0.1 dB over >60 dB signal range and a ±15°C temperature range. The phase detectors are commercial units with limiters (<±1° over 60 dB) in each channel followed by phase-to-pulse-width circuits and low-pass filter integrators. They measure phase unambiguously over 360°.

Because the PLL in the third LO removes short-term oscillator fluctuations from the fourth IF signal and because the XV, XH and rain scatter signal levels should always be less than the V copolarized signal level, the amplitude measuring range of the XV, XH, and rain scatter channels is extended by active BPFs with 1-Hz 3-dB BW. The bandwidth of these filters, which are similar to the 10-Hz BW filters and also are enclosed in an oven, is limited to about 1 Hz by the expected maximum rates of change of signal parameters.[2]

Again, the off-axis rain scatter channel is packaged separately and is essentially identical to the main beam crosspolarized signal channels except for the omission of the 10-Hz BW fourth IF BPF and the phase detector.

The outputs from the amplitude and phase detectors feed the dc signal-conditioning amplifiers in the data collection equipment in Fig. 16.

IV. 28-GHz RECEIVER CHANNELS

The 28-GHz receiver channels from the antenna feeds through the second IF crystal BPFs are shown in Fig. 3. The 28-GHz copolarized beacon signal (V) is received with a vertically polarized feed whose resulting antenna beam is coaxial with the 19-GHz beams. A horizontally polarized feed, whose beam is coaxial with the vertically polarized beam, receives the crosspolarized signal component (XV). These 28-GHz feeds share the main beam feed frame[4] with the 19-GHz main beam feeds. A 28-GHz off-axis rain scatter feed produces a beam coaxial with the 19-GHz off-axis beam.

Since many of the components and functions in the 28-GHz channels are similar to the corresponding ones in the 19-GHz channels, descriptions of similar components and functions are not repeated here. The major differences between the 19- and 28-GHz channels result because there is no polarization switching at 28 GHz but there are sidebands coherent with the 28-GHz carrier.[1] These sidebands (±264 MHz for two satellites and ±528 MHz for another) are used for measuring amplitude

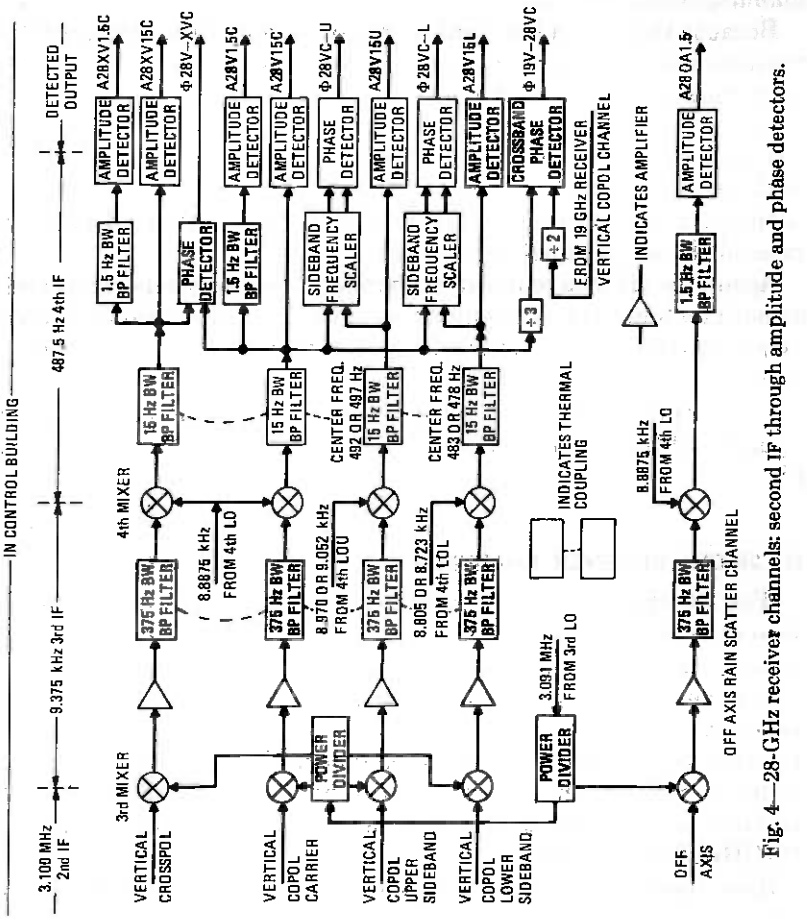


Fig. 4—28-GHz receiver channels: second IF through amplitude and phase detectors.

and delay dispersion.[1] Component locations discussed under 19-GHz receiver channels are indicated on Fig. 3.

As discussed earlier, 19- and 28-GHz receiver IFs and LOs are in a 2:3 ratio and corresponding LOs are derived from the same sources for extending measurement range and to permit measuring phase between 19- and 28-GHz carriers. This phase measurement yields the spectrum of east-west satellite pointing angle fluctuations because of the interferometer formed by the 2 meter (6.6 ft.) east-west separation between the 19- and 28-GHz satellite antennas. The measurement also yields dispersion information for a large frequency separation (9.5 GHz). In order to minimize effects of frequency changes on this phase measurement, all BPFs with significant delay (e.g., the crystal BPF) also have a 2:3 center frequency and bandwidth ratio between 19- and 28-GHz channels.

The calibration source provides calibrated test signals for all of the 28-GHz channels; sidebands coherent with the carrier and with either ±264 MHz or ±528 MHz separation can be selected.

A single fixed-frequency LO converts the entire 28.56 ± 0.528 GHz band, encompassing both sets of sidebands and the carrier, to the first IF centered at 1.5 GHz. This ensures that the short-term frequency stability of the most critical first LO is not compromised by other requirements. The first mixer and IF preamplifiers, which are similar to the 19-GHz units, have noise figures (SSB) of ≤7 dB over the entire 1.5 ± 0.528 GHz band. The first IF is a compromise that is pushed toward low frequencies by IF preamplifier noise-figure considerations and toward high frequencies both by the need to pass the wide bandwidth and by first LO noise considerations. Since many standard octave bandwidth components are available for the 1- to 2-GHz band, the 28-GHz receiver first IF band is set at 1.5 ± 0.528 GHz. This constraint, along with the 2:3 ratio IF and LO constraint, fixes the first IF of the 19-GHz receiver at 1 GHz. The specific IF frequency, 1.504 GHz, results from maximizing the frequency separation between the desired IF signals and spurious mixing products from this and later frequency conversions. Isolation between receiver channels through the common first LO path (see Fig. 5) is >65 dB. Sidebands (SBs) are split into separate channels in the first IF so the bandwidth of the remainder of the receiver can be narrowed to increase sensitivity. (With the existing radio link parameters the maximum S/N in a 1-GHz BW is <-20 dB.) The upper sideband (USB) channel uses the same 0.4 GHz BW BPF, broadband first IF amplifiers, second mixer, second IF amplifiers and crystal BPFs for both the +264 and +528 MHz USBs. The corresponding parts of the LSB channel also are the same for both sideband frequencies. The same narrowband crystal filters are usable because the third LO SB frequency corrections, discussed in the next paragraph and in Section V, are done in the second

LOs. This second LO correction keeps the frequency spread between the different SBs small (<±200 Hz) at the second IF.

The measurement of delay dispersion[1] requires measuring phase between the carrier and the USB scaled by 108/109 or 108/110 and between the carrier and the LSB scaled by 108/107 or 108/106. These scaling factors are the exact ratios between the carrier and the SBs for the different satellites. Preserving the carrier and SBs in these ratios throughout the receiver requires separate LOs for carrier and SBs scaled similarly in frequency. Simple mixing by a single LO, as is done in the first frequency conversion, does not preserve the ratios. Therefore, sideband LO frequency corrections must be made elsewhere for this conversion and any other such conversion. The SB frequency corrections for the first, second, and third LOs are made to the SB second LO frequencies. This is done by adding or subtracting the "missing" portions of these LOs [(1 or 2)/108 × f_{LO}] to the second LO used for carrier mixing. The LO implementations in Fig. 6 for making the corrections are described later. Provisions for remotely switching the SB channel LOs in the side cab from ±264 to ±528 MHz for the different satellites are provided in the control room. Isolation among the V carrier, XV and off-axis channel through their common LO path is >61 dB.

The 28-GHz off-axis rain scatter channel is identical to the 28-GHz XV channel and packaged together with the 19-GHz off-axis channel.

The 28-GHz receiver channels, continuing from the crystal BPFs through the amplitude and phase detectors, are illustrated in Fig. 4. Since the SB LO frequency correction for the third LO is included in the second LO, the third LO is common to all 28-GHz channels. Isolation among 28-GHz channels through the common third LO path (see Fig. 8) is provided in the same way as in the 19-GHz channels and is >80 dB (measurement limit).

In the USB channel the same broadband third mixer and third IF amplifier are used for both USB frequencies. The same is true for the LSB channel.

Although not shown separately in Fig. 4, the active third IF and fourth IF BPFs and fourth mixers are separate for the USB channels for the 264 and 528 MHz SBs. These components are separate also for the corresponding LSB channels. The fourth LOs for the four SB channels include the fourth LO frequency corrections (see Fig. 10). Separate channel filters are used at this point in the receiver because the different SB frequencies are spread approximately 1 percent in frequency (1/107 and 1/109) and the BPF 3-dB BWs are only 3 percent.

The 1.5-Hz BW BPFs in the XV and V carrier channels provide measuring range extension as described for 19-GHz receiver channels.

The active filters and fourth mixers for the XV and V carrier channels and for the ±264 MHz USB and LSB channels are in one oven. The ±528

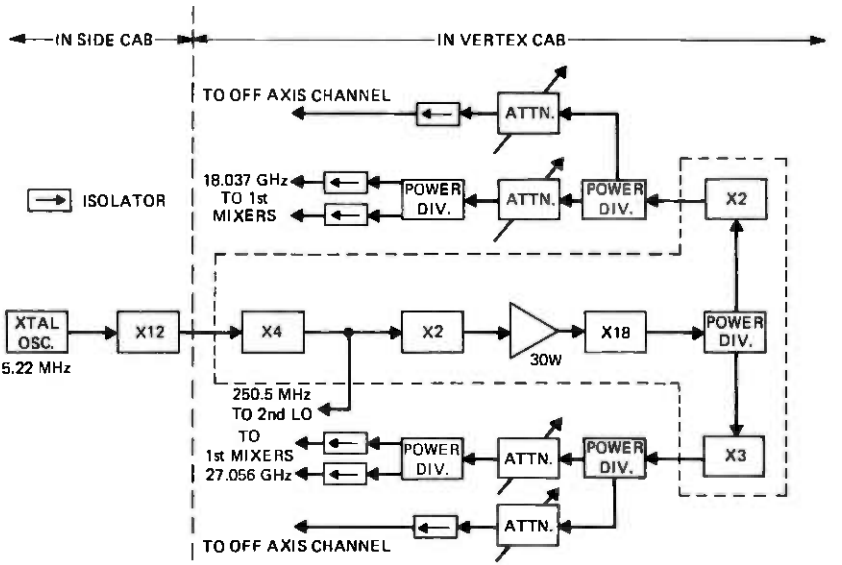


Fig. 5—Receiver first local oscillator (crystal oscillator frequency is 5.218991 MHz).

MHz USB and LSB channels and the 1.5-Hz BW active BPFs are in another oven.

In the SB frequency scalers, the SB frequencies are scaled by 107/108 and 109/108 or by 106/108 and 110/108 for the delay dispersion phase measurement. The sideband signals are limited and scaled by PLL frequency multipliers and digital dividers.

The crossband phase measurement is made between the 28.56- and 19.04-GHz V-copolarized carriers. The 325-Hz fourth IF signal from the 19-GHz receiver channel (Fig. 2) is divided by 2 and the 487.5 Hz fourth IF signal from the 28-GHz receiver channel (Fig. 4) is divided by 3 before the crossband phase detector (Fig. 4).

The off-axis rain scatter channel is similar to the XV channel. Both the 19- and 28-GHz off-axis channels are packaged together and are separate from the mainbeam channels.

V. LOCAL OSCILLATORS AND FREQUENCY CONTROL TECHNIQUES

Frequency requirements for the LOs and the receiver channel-to-channel isolation through the common LO paths are discussed in the previous sections.

5.1 First LO

The first LO is shown in Fig. 5. A 5.218991-MHz high stability (1×10^{-10}/day, 1×10^{-11}/0.1 sec) quartz crystal oscillator is multiplied first

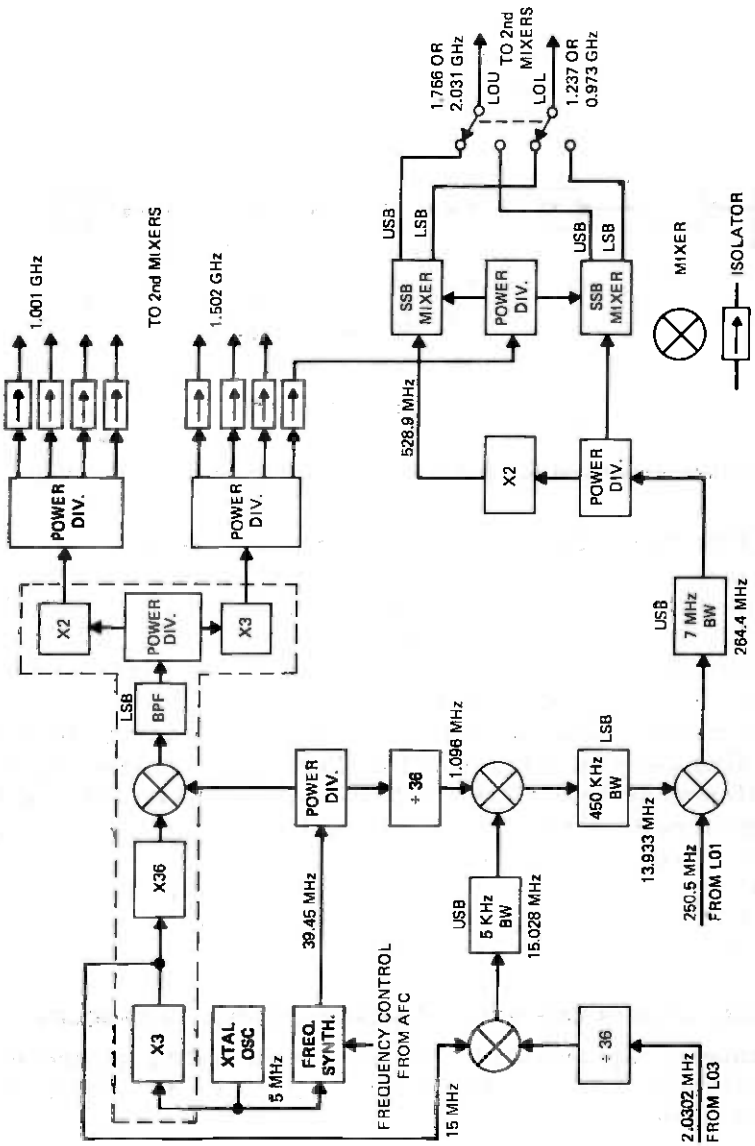


Fig. 6—Receiver second local oscillator in side cab (crystal oscillator frequency is 5.000000 MHz).

by 48 to 250.5 MHz for the second LO sideband correction frequency described earlier and then by 36 to 9.019 GHz; further multiplication by 2 or by 3 produces the coherent LO signals for the 19- and 28-GHz receiver channels. Passive varactor frequency multipliers follow the power amplifier at 501 MHz. The first LO does not include frequency control so the stability of the basic crystal oscillator is preserved. The first LO frequency multipliers from 62.6 MHz to 18 and 27 GHz, power dividers, isolators, and waveguide distribution are mounted on the antenna feed assembly in the vertex cab.

5.2 Second LO

The second LO is derived from a 5.000000-MHz quartz frequency standard as shown in Fig. 6. Multiplication by 3 produces a component of the sideband correction frequency at 15 MHz. This signal is multiplied by 36 to 540 MHz where the frequency of a 39-MHz frequency synthesizer is subtracted in a mixer. The synthesizer frequency is controlled by the automatic frequency control loop (AFC) indicated on Fig. 1 and described later in this section. Multiplication by 2 and by 3 after the mixer produces the coherent second LO signals for the 19- and 28-GHz receiver channels.

The second LO does the first, second, and third LO frequency scaling described earlier for the sideband channels of the 28-GHz receiver. This scaling starts by adding 15 MHz from the second LO frequency multiplier and 28.6 KHz from the third LO (i.e., 1.03 MHz ÷ 36) in a mixer; these are both 1/108 of the contribution of their respective sources to the second and third LOs. The synthesizer output divided by 36 to 1.096 MHz is then subtracted from the correction term since the synthesizer frequency itself is subtracted in the 1.562-GHz LO multiplier chain; this 1.096 MHz is also 1/108 of the synthesizer contribution to the 1.502-GHz LO. The resulting 13.933 MHz correction term, containing the contributions from the third LO and both components (oscillator and synthesizer) of the second LO, is then added to the 250.5 MHz = 27,056/108 MHz correction term from the first LO. The resulting 264.4 MHz correction term, $(1/108)(f_{LO1} + f_{LO2} + f_{LO3})$, is then added to the 1.502-GHz second LO frequency to produce the 1.766-GHz upper-sideband second LOU, described in the section on 28-GHz receiver channels. The correction term also is subtracted from the second LO frequency to produce the corresponding second LOL at 1.237 GHz.

The 264.4-MHz correction term is multiplied by 2 and similarly added and subtracted to provide second LOU and LOL for the beacons with ±528.8-MHz sideband separation.

The order of mixing of the correction frequencies makes the filtering easiest.

5.3 AFC loop

An AFC loop controlling the second LO is used to remove long-term frequency variations in both the satellite and the other receiver LOs. This AFC loop must meet the following requirements:

(*i*) Track over ±200 KHz.
(*ii*) Track a ±1.1 Hz/sec frequency ramp.
(*iii*) Maintain ±10-Hz accuracy while tracking.
(*iv*) Have a frequency averaging time of around 1 second.
(*v*) Hold frequency on command for 10 minutes to within ±50 Hz.
(*vi*) Contribute negligible short-term frequency instability.

Requirements (*i*) and (*ii*) are set by the expected long-term frequency behavior of the beacon.[1] Requirement (*iii*) guarantees that the phase-locked loop used for short-term frequency correction need not have a large capture range. Requirement (*iv*) constrains the AFC loop to average over most short-term frequency variations and track only long-term effects. This will improve the loop's low-SNR tracking performance and allow a better estimate of the signal's average frequency during outages. Requirement (*v*) states that the receiver frequency stability during signal outages of up to 10 minutes will be much better than the satellite's; thus, the frequency region which must be searched during reacquisition is determined only by the satellite.

These requirements impose severe stability and tunability requirements on the loop frequency generation element. A digitally controlled frequency synthesizer provides frequency memory and long-term stability equal to that of its reference oscillator and thus was chosen for this application.

It is a simple "fact of life" that a fixed-frequency oscillator has better short-term frequency stability than a variable-frequency oscillator, be it a VCXO or synthesizer. To prevent the synthesizer from adding to receiver short-term instability, a relatively low-frequency synthesizer is mixed with a high-frequency fixed source, as discussed previously. In addition, a synthesizer exhibiting low phase noise was chosen for this application. The exact synthesizer and first IF frequencies were chosen to insure that any internally-generated signal-frequency or image-frequency spurious signals would be well outside IF filter passbands.

The circuitry for controlling the synthesizer output frequency is shown in Fig. 7. The 2.067-MHz IF from the 19-GHz vertically polarized feed is filtered to remove the 12.5-kHz image response created by the following conversion. The filter output is levelled by an AGC amplifier to keep the discriminator input level constant. Since the clear-air S/N at the filter output is ~30 dB, the AGC will be noise-dominated only for

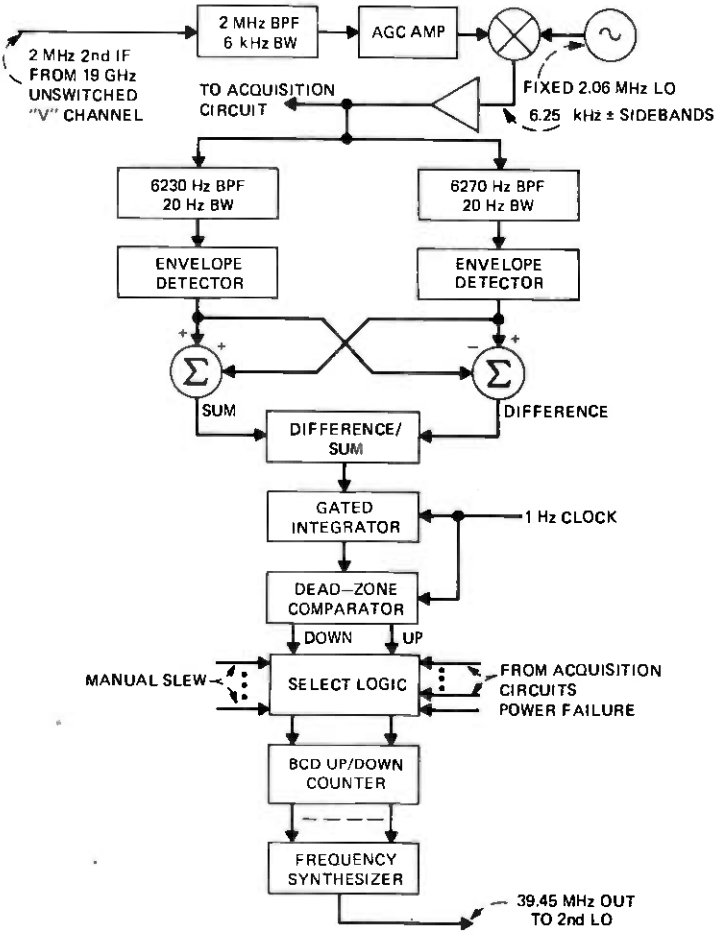


Fig. 7—Receiver second local oscillator AFC circuitry (digital portion).

fades of 30 dB or more, and will hold the signal component of its output constant for fades of lesser depth.

To obtain the desired narrow discriminator bandwidth, the signal is downconverted with a fixed-frequency LO to 6.25 kHz and fed to two 20-Hz-wide active bandpass filters tuned 20 Hz above and below center frequency. These filters are enclosed in a temperature-controlled oven to enhance their long-term frequency stability. The filter outputs are envelope-detected and the detector outputs summed and differenced. The difference is divided by the sum voltage in an analog divider. This division operation provides the effect of additional AGC and widens the capture range of the complete AFC loop by "propping up" the tails of the discriminator S curve. The 6.25-kHz signal, with 1-kHz polarization-

switching sidebands, is fed also to the acquisition detector described later.

The discriminator output is filtered by an integrator which is reset once per second. The dead-zone comparator produces an output pulse on one or the other of two lines if the integrator output indicates an average frequency error of more than 2 Hz. These pulses are used to increment or decrement a six-decade digital up-down counter; the BCD counter controls the synthesizer output frequency to 1-Hz resolution. Since the synthesizer frequency is doubled before use as a 19-GHz LO, an average signal-frequency error of greater than 2 Hz over a 1-second interval will produce a 2-Hz local oscillator frequency correction during the next second. The loop will thus track a signal moving at up to 2 Hz/second to ±2-Hz accuracy. With six-digit frequency resolution, the loop tracking range is 2 MHz.

The up-down counter may also be controlled from other sources. During the signal acquisition phase the counter is automatically stepped up or down to sweep the receiver around the expected signal frequency. The counter may also be stepped up or down manually. These operations are discussed later in this section.

Long-term frequency memory is implemented by disabling the counter inputs when the signal fades below the threshold of the tracking loop. Power-failure protection is provided by powering the counter from an uninterruptible battery power source. In both cases the long-term "frequency memory" of the loop is solely determined by the frequency stability of the synthesizer master oscillator.

5.4 Third LO and PLL

The third LO is shown in Fig. 8. The 2.06-MHz voltage-controlled crystal oscillator (VCXO) source is controlled by the phase locked loop indicated on Fig. 2. The oscillator frequency is used directly for the 19-GHz receiver channels and after division by 2 and multiplication by 3 for the 28-GHz channels. The phase locked loop removes the remaining short-term oscillator frequency variations and presents a frequency-stable signal to the final 10-Hz and 1-Hz bandwidth receiver filters. This PLL must meet the following requirements:

(*i*) Track over ±50 Hz.
(*ii*) Have a loop bandwidth selectable around 10 Hz.
(*iii*) Hold at "center" frequency on command to within ±2 Hz,
(*iv*) Exhibit no frequency "walk-off," and recover quickly from signal outages.

Requirement (*i*) allows the loop to track the expected range of short-term frequency variations. Requirement (*ii*) permits the trading of loop threshold for amplitude stability in the final signal measurement filters (narrow bandwidth extends the loop threshold at the expense of am-

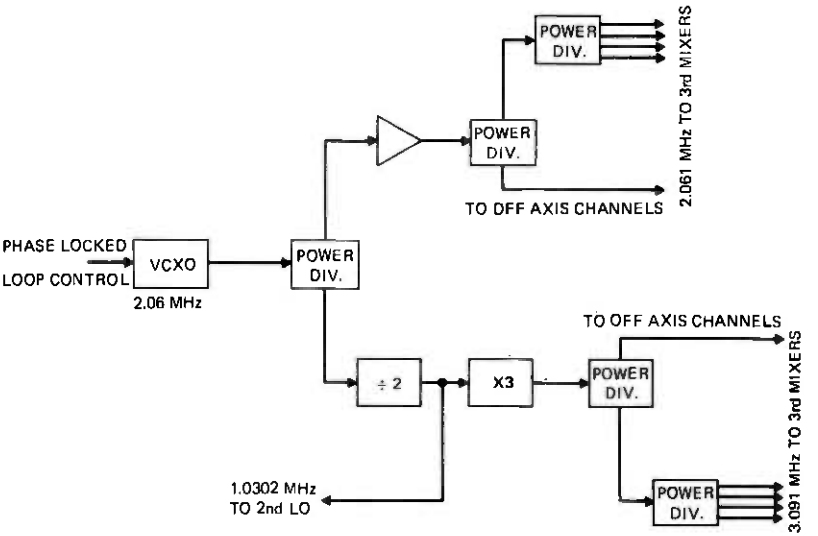


Fig. 8—Receiver third local oscillator in control room [voltage controlled crystal oscillator (VCXO) frequency is 2.0604 MHz].

plitude fluctuations due to FM-AM conversion in the measurement filters). Requirements (*iii*) and (*iv*) guarantee good frequency memory during signal fades and recovery from them without long loop pull-in times.

The loop components are shown in more detail in Fig. 9. The received 19-GHz V copolarized signal, at 325 Hz, is filtered by an active bandpass filter, limited, and phase compared with stable 325-Hz reference. The phase detector output is filtered by an active lowpass filter and used to control the third LO VCXO frequency. Thus, the 19-GHz V-copolarized signal is phase locked to the 325-Hz reference, and with it all other receiver channels.

The PLL noise bandwidth is adjustable over the range 5 to 50 Hz to encompass the range of short-term satellite oscillator stabilities expected. A loop damping factor ζ of 0.8–1 was desired to avoid phase overshoot.[9] Both of these parameters are influenced by both the loop filter and the IF bandpass filter. A test using prototype RF and PLL hardware showed that an overall loop bandwidth B could be achieved with a loop filter natural frequency ω_n (rad/sec) = 1.2 B, a loop damping factor ζ = 1.2, and an IF filter bandwidth = 1.5 B. To achieve the desired 5–50 Hz overall bandwidth range, the loop filter ω_n may be varied (through adjustment of R_1 and R_2 in Fig. 9) over 6–60 rad/sec holding ζ = 1.2, and the IF filter bandwidth varied over 9–90 Hz, holding its gain constant. This adjustment is made in 1 dB switched steps (bandwidth ratio = 1.26).

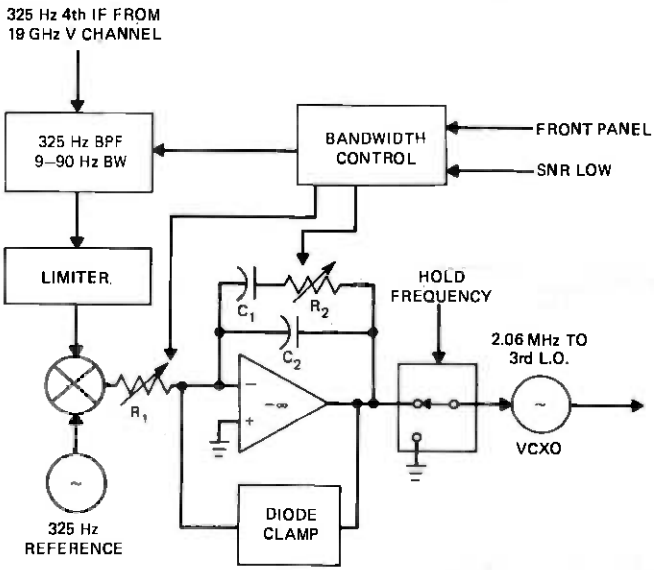


Fig. 9—Receiver third local oscillator PLL circuitry.

The capacitor C_2 in the loop filter rolls the loop off to prevent the sum-frequency (650 Hz) phase detector output from modulating the VCXO. The diode clamps limit the frequency excursion to ±50 Hz. The VCXO control line is grounded on external command, and holds the VCXO at its resting frequency during signal outages.

5.5 Fourth LO

The fourth LO shown in Fig. 10 is self-contained; it generates sideband receiver-channel local oscillators, LOUs and LOLs, which include the required fourth LO frequency corrections. Because of the low frequencies involved and the ease of frequency division using medium scale integrated circuits (ICs) all fourth LO frequencies are derived by division from a single 10.5333-MHz crystal oscillator. The divisions ($\times n/128$) that determine the final output frequency ratios are done by digital rate multipliers. These rate multipliers produce n output pulses for every 128 input pulses with n selected by pin connections on the ICs. The n output pulses are not evenly spaced because they are produced by gating the input pulse train. The uneven spacing is equivalent to a deterministic timing jitter on the output waveform which is a fixed multiple of the input pulse period. The three decade dividers ($\div$ 1000) following the rate multipliers preserve this timing jitter, which is then a much smaller fraction (1/1000) of the output waveform period. In effect this $\div$ 1000 reduces the index of the jitter phase-modulation by 1000. The jitter

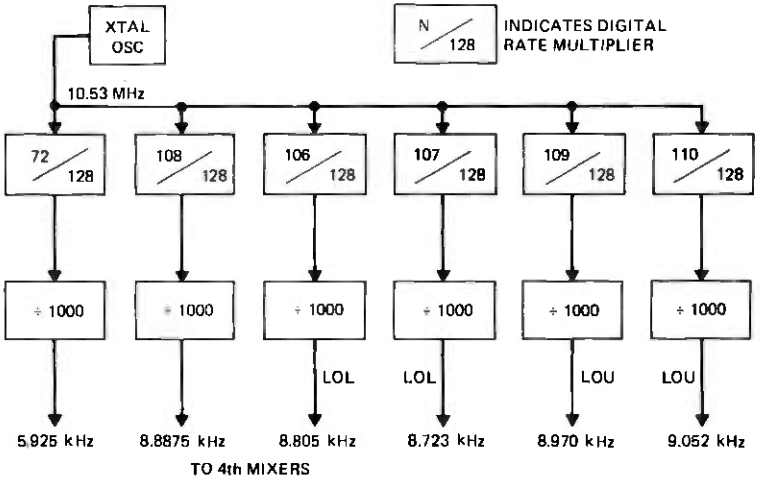


Fig. 10—Receiver fourth local oscillator in control room (crystal oscillator frequency is 10.533333 MHz).

sidebands on the worst fourth LO output (the 9.052 kHz) are 60 dB below the LO signal level; most LO outputs have sidebands >65 or 70 dB below the signal level.

5.6 Frequency acquisition

Since this receiver is designed for unattended operation, it must be capable of automatic frequency reacquisition after signal outages. This is accomplished by sweeping the receiver through the expected range of signal frequencies whenever tracking cannot be maintained. When the signal is located, this sweep stops and the receiver returns to its normal tracking mode.

The receiver's acquisition speed is increased by making use of the signal's last known frequency, its maximum drift rate, and its maximum daily frequency excursion. Let:

$$f_o = \text{last known frequency at time } t_o$$

$$\Delta = |\text{maximum drift rate}|$$

$$f_{\min} = \text{minimum observed frequency}$$

$$f_{\max} = \text{maximum observed frequency.}$$

Then the frequency region to be searched is bounded by

$$\max\{f_{\min}, (f_o - \Delta(t - t_o))\} \le f(t) \le \min\{f_{\max}, (f_o - \Delta(t - t_o))\}.$$

Acquisition speed is also improved for short-duration signal outages by holding the last known signal frequency and inhibiting the frequency

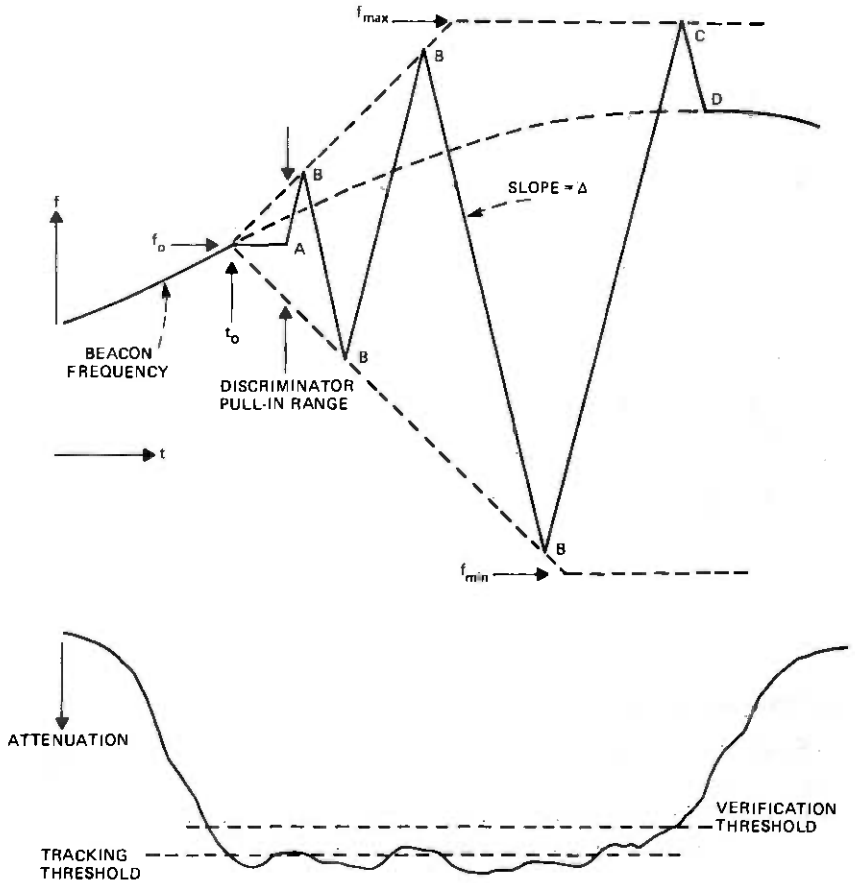


Fig. 11—Typical acquisition sequence for receiver second local oscillator. Frequency tracking is lost at time t_o. Reacquisition frequency search begins at point *A*: signal is reacquired at point *D*.

sweep until the maximum frequency uncertainty exceeds the pull-in range of the discriminator.

Figure 11 illustrates a typical receiver acquisition sequence. At time t_o the received signal level fades below the tracking threshold. The register controlling receiver frequency is held at its last value f_o and a counter is counted at a rate Δ to indicate the maximum frequency uncertainty as time progresses. When this uncertainty exceeds the pull-in range of the tracking discriminator (point *A*) the frequency-control register is stepped up or down continuously to sweep the receiver frequency. Whenever this register's excursion from f_o equals the maximum frequency uncertainty the stepping direction of the frequency control register is reversed, reversing the direction of frequency sweep (points

B). Sweep direction is also reversed at f_{max} and f_{min}, the daily frequency excursion limits (point C). These limits are set by thumbwheel switches, and are reset periodically to account for long-term aging of beacon and receiver oscillators.

During this frequency sweep the unswitched 19-GHz vertically-polarized channel is observed to detect the presence of the beacon signal. When its presence has been verified (point D) the receiver returns to its tracking mode. The procedure used to accomplish this verification is described below.

As mentioned previously, the 19-GHz received signal is used for reacquisition. However, the situation is complicated by the 1-kHz polarization switching. This switching produces sidebands separated from the carrier by multiples of the switching frequency. The receiver uses knowledge of the relative levels of the carrier and sidebands to discriminate between the two.

If the polarization modulation were symmetrical (50 percent duty cycle) the received spectrum would appear as in Fig. 12a. Only odd-order sidebands are present; the first-order (±1 kHz) sidebands are down 3.9 dB. This pattern of three signals could be detected using three narrowband filters with 1 kHz spacing to indicate the presence of carrier in the center filter. If modulation asymmetry is considered, however, these three filters are not sufficient. Figure 12b shows the received spectrum for 2 percent modulation asymmetry. This asymmetry has generated even-order sidebands about 4 dB below the level of the 17th and 19th sidebands; three-filter detection would indicate carrier acquisition at these two points. A fourth filter located 2 kHz from the carrier filter, however, resolves this ambiguity. Its relative output will be low for a true detection and high otherwise. Since the asymmetry of the beacon modulation was not known during receiver construction, four filters were used to unambiguously detect carrier acquisition.

This four-filter acquisition detector is shown in Fig. 13. Four narrowband active filters are driven by the 6.25-kHz signal (derived from the 19-GHz V channel) used by the AFC loop. These filters have a two-pole response with matched 30-Hz bandwidths, and are tuned to pass the carrier at 6.25 kHz, two ±1 kHz sidebands, and the −2 kHz sideband. These filters also are contained in a temperature-controlled oven. Since detection decisions are based on ratios of these signals, the filter outputs are envelope detected and logged. Differences of these logged signals then indicate the desired ratios.

The 30-Hz filter bandwidth was chosen as a compromise between sweep speed and SNR at the filter outputs. The filter outputs will not reach their full levels if the sweep speed exceeds an appreciable fraction of filter bandwidth per filter impulse response time. Thus, the maximum permissible sweep speed increases with the square of the detection filter

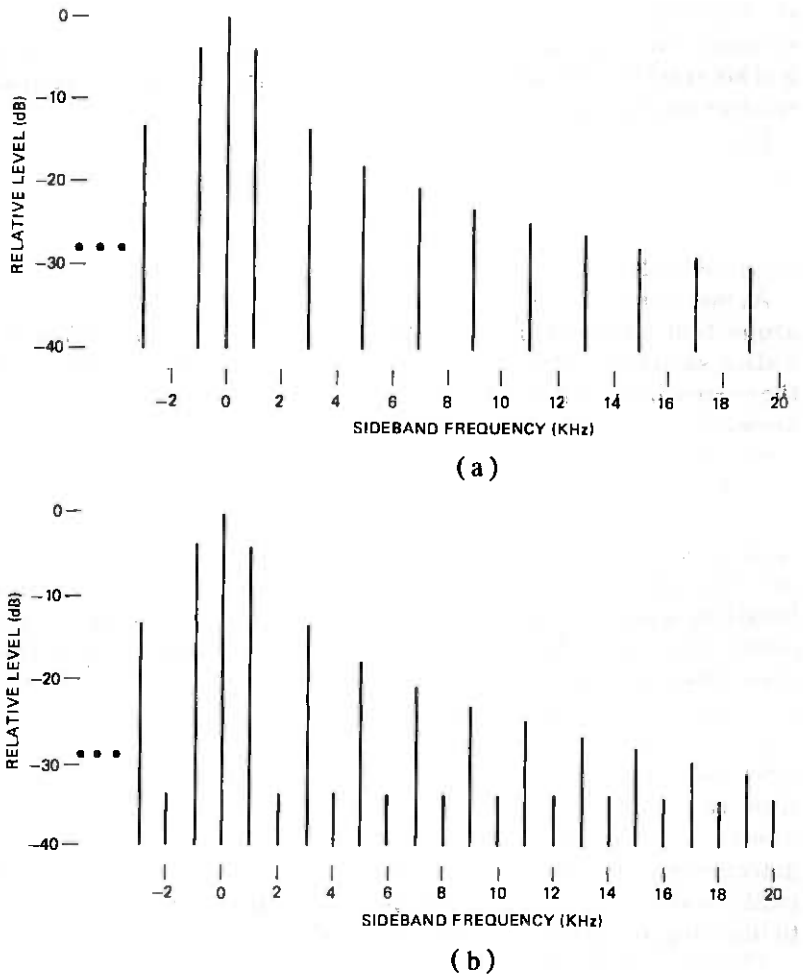


Fig. 12—19-GHz polarization-switching sideband levels, with and without switching asymmetry. Note the even-order sidebands present with switching asymmetry.

bandwidth. The signal level required to achieve a given SNR, however, increases linearly with filter bandwidth. A 10-Hz filter bandwidth (that used in the receiver's 19-GHz signal channels) was found to require a sweep speed of <50 Hz/sec for reliable detection. With the 30-Hz filters used, the sweep speed may be increased to *250* Hz/sec, speeding the reacquisition process.

Several conditions must be satisfied to indicate the presence of received carrier in the 6.25 kHz carrier filter. The carrier filter output must be >10 dB above its no-signal level to assure a reasonable false-alarm rate. Both ±1 kHz sideband filter outputs must be between 2 and 6 dB

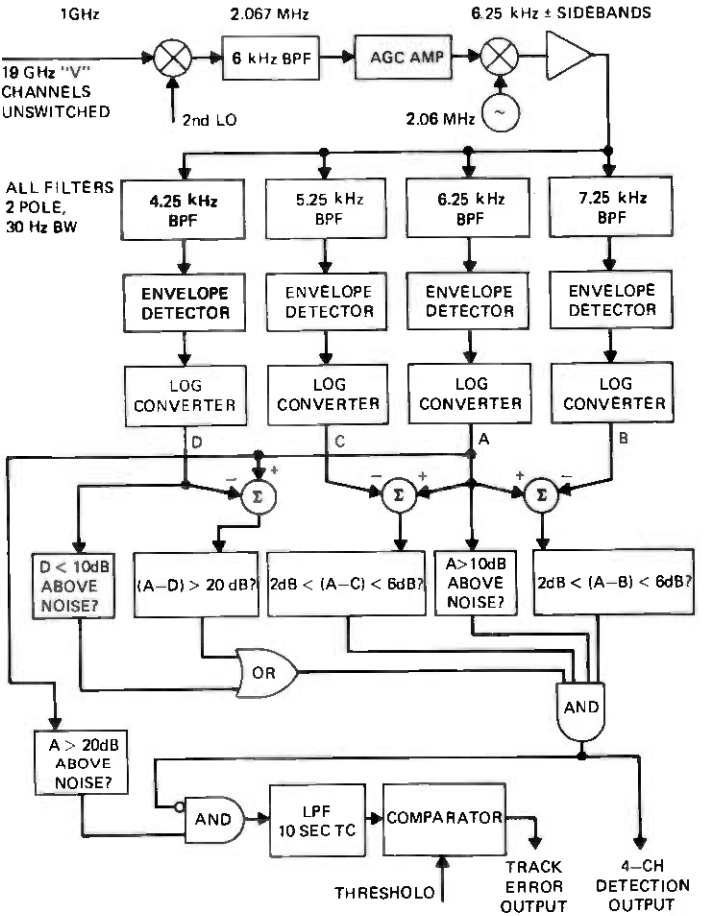


Fig. 13—Four-filter acquisition detector (part of receiver second local oscillator AFC circuitry).

down from the carrier level. This range allows for slight nonlinearities in the log converter slopes and the effect of noise at low signal levels. Finally, the −2 kHz sideband filter output must be either <10 dB above its noise level or >20 dB below the carrier filter output. If all these conditions are satisfied a four-channel detection signal is given. This action interrupts the acquisition frequency sweep and turns on the AFC to attempt to track the signal. The four-channel detection signal is observed for 15 seconds to verify the detection. If the detector output is true for >50 percent of this time, a valid signal is assumed to be present and the receiver returns to its tracking mode. Otherwise, the acquisition search is continued until the signal is found.

A second signal is generated in conjunction with the four-channel

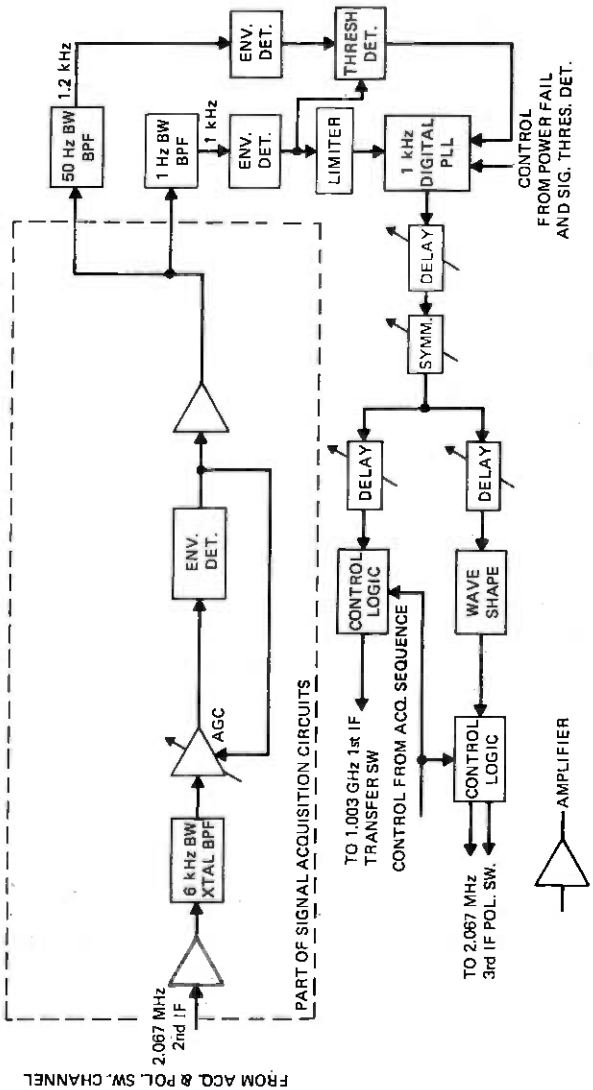


Fig. 14—1-kHz polarization switch synchronizing circuits.

detection output to further guard against tracking of a 19-GHz sideband. An error signal is generated if the carrier filter output is >20 dB above its noise level but no four-channel detection has been made. (Such a condition would occur if the receiver were tracking a sideband.) If this error signal persists for 10 seconds a frequency sweep is initiated over the full uncertainty band. This action should never be performed, but further assures the eventual acquisition of only the 19 GHz carrier.*

VI. POLARIZATION SWITCH SYNCHRONIZATION

A separate IF channel without switches is provided for synchronizing the main-channel polarization switches (see Fig. 1) with the 1-kHz polarization-switched beacon signals. Since the clear-air signal-to-noise ratio in a 2-kHz bandwidth is <+30 dB and low jitter (<5 percent of a switching period) switching is desired for rain attenuation of at least 40 dB, a very long (~100 sec) time constant (narrowband) PLL is required to recover the 1-kHz switching signal. Such a long time constant loop has an even longer pull-in time; therefore, frequency and phase memory through severe rain attenuation events (>40 or 50 dB) on the order of 10 minutes is needed to prevent loss of data for long periods after such severe events. The long PLL time constants and even longer memory are easiest to implement digitally.

The polarization switch synchronizing circuits in Fig. 14 follow the 1 GHz to 2 MHz frequency conversion in the acquisition and polarization switch synchronization channel in Fig. 1. The 1-kHz digital PLL is shown in more detail in Fig. 15.

As indicated in Fig. 14, the automatic gain controlled (AGC) 2-MHz IF amplifier holds the 1-kHz signal into the oven-enclosed 1-Hz BW active BPF constant over the range from the clear air signal level to the noise level of the 1-Hz BW BPF input. The filtered 1 kHz is limited, fed to the PLL, and is detected to provide a threshold comparison voltage. The noise voltage in a 50-Hz BW at 1.2 kHz is rectified and lowpass-filtered for the threshold detector reference.

The square-wave PLL output is phase-locked to the satellite polarization switch. However, because of filter delays, etc., the square-wave transitions and the beacon signal transitions at the receiver switches do not occur at the same instants of time. Delay and symmetry adjustments after the PLL permit adjustment of the switching waveform transitions for coincidence with the signal transitions. Further shaping of the switch driving waveform for the 2-MHz third IF polarization switches (Fig. 1) provides a "dead zone" around transitions to allow switching transients to settle.

* This circuitry was later disabled since it was often triggered by strong depolarization produced by atmospheric ice crystals.[12]

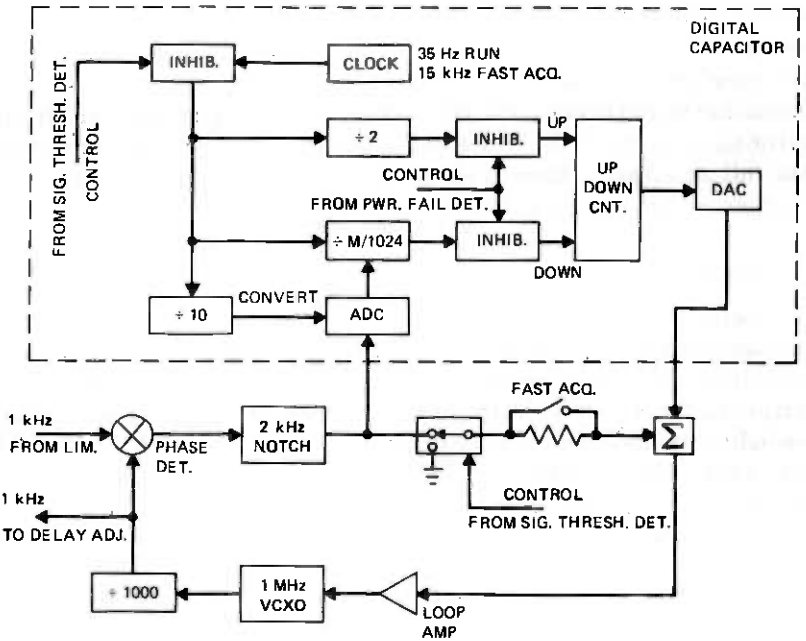


Fig. 15—1-kHz digital phase-locked loop (PLL).

6.1 Digital phase-locked loop

The digital PLL in Fig. 15 is equivalent to a conventional second order loop[9] with an active loop filter. The frequency of a 1-MHz voltage controlled crystal oscillator (VCXO) is controlled by the loop amplifier output voltage. The 1 MHz is divided in digital decade dividers to 1 kHz and compared in phase with the limited 1 kHz derived from the polarization-switched beacon signal. The notch filter following the analog-multiplier phase-detector prevents the 2-kHz sum frequency from saturating the loop amplifier. The straight-through path from the notch filter to the summer (through the switch and resistor described later) establishes the high-frequency open-loop gain. The digital capacitor output is summed with the straight-through path to determine the open-loop low-frequency response.

Operation of the digital capacitor centers on the up-down counter. A fixed clock frequency is divided by two and applied to the count up input of the counter. Input pulses are blocked by the inhibit gates when a control voltage appears on the inhibit control line; otherwise, all input pulses are passed on to the output. The analog-to-digital converter (ADC) samples the phase detector output every 10 clock cycles and converts it to a 10-bit digital word, M, that serves as the control word for a digital rate multiplier. With an ADC input of 0, $M = 512$, and the rate multiplier

passes 512/1024 = 1/2 of the input pulses to the count down input of the counter. Under this condition, the counter alternately counts up and down so there is no net change in count. The pulses into the count up and count down inputs and the convert pulses to the ADC are appropriately timed to prevent ambiguities that would arise if these pulses were nearly coincident. Positive phase detector outputs produce $M > 512$. With $M > 512$ there are more down counts than up counts in a given time period and the counter counts down at a rate proportional to the phase error. With a negative phase error, $M < 512$, there are more up counts than down, and the counter counts up. The count in the counter is analogous to charge stored in a capacitor; M is analogous to current into the capacitor; and the ADC is equivalent to a voltage to current converter. The analogy is completed by the digital-to-analog converter that converts the counter count to a voltage, the capacitor voltage, with 0 count being maximum positive voltage, maximum count being maximum negative voltage and mid-count being 0. The equivalent capacitance is a function of the clock rate, the ADC voltage-to-M factor and the DAC count-to-voltage factor. The counter saturates at maximum count and at 0 count so the contents are not "spilled" by either a continuous positive or negative phase error. The DAC output is filtered to smooth the voltage steps.

In the normal run mode of the PLL, the digital capacitor clock frequency is 35 Hz, the natural frequency,[9] ω_n, of the PLL is 6×10^{-3} and the damping is 0.5. A faster loop response is available for faster acquisition (acq) of phase lock with high signal levels. This is accomplished by increasing the clock frequency to 15 kHz to decrease "capacitance" and by decreasing the loop summing resistance to increase the high-frequency open-loop gain and thus maintain the damping at 0.5. The fast acquisition mode can be selected manually with a single switch. The digital capacitor departs from a real capacitor when capacitance is changed by changing clock frequency. Changing clock frequency does not instantly change the counter count; it changes only the rate of count. Thus, digital capacitor voltage is conserved with capacitance change and "charge" is not conserved. This is a useful property in this PLL because the VCXO control voltage and thus the VCXO frequency and phase remain continuous when the loop natural frequency is changed discontinuously. The PLL then remains in lock and does not experience a phase step in going from the fast acquisition mode to the run mode.

Phase and frequency memory during loss of signal or loss of primary power is provided by the up-down counter, loop amplifier, VCXO, and ÷1000. These components are supplied from a battery power supply that normally floats on a charger.

For loss of power all inputs to the up-down counter are inhibited to prevent start-up transients from disrupting the held count.

For loss of signal, counter memory is provided by inhibiting the clock. Continued operation at the average phase and frequency before the hold is insured by grounding the summing input from the phase detector. The VCXO is in an oven to insure that the stability of the held phase is better than the phase stability of the polarization switching oscillator in the satellite.

VII. DATA COLLECTION

The data collection equipment is common to all receiving channels as indicated in Fig. 16. Data that are critical for maintaining continuity in the data base for long term statistics, e.g., signal attenuation and depolarization, are recorded continuously on analog ink-pen paper-chart recorders. These chart recordings provide a backup in the event of failure of the digital recording system and also provide a "quick look" at the recorded data. The logarithms of signal amplitudes are recorded on the chart recorders with a range of 50 dB.

All receiver outputs are multiplexed along with (*i*) system status indicators such as whether the frequency control loop is tracking or holding in a signal fade, (*ii*) outputs from weather instruments such as rain gauges, thermometers and wind speed recorders, (*iii*) outputs from the on-going interim experiment[10] using the COMSTAR beacon at 128°W, and (*iv*) another propagation experiment[11] using a 12-GHz beacon on the NASA/Canadian Communications Technology Satellite (CTS). These multiplexed signals are digitized, temporarily stored in the minicomputer core memory, screened for relevance, and stored on digital magnetic tape. Multiplexer and analog-to-digital converter sequencing, digital data screening and buffering, and digital tape drive control are handled by the same minicomputer that points the receiving antenna.

The objectives of the data screening procedure are to minimize the amount of superfluous data stored while not losing any relevant propagation data. The screening algorithm copes with the multiplicative signal fluctuations caused by the atmosphere and with the additive noise that dominates at low signal level.

The data taking and screening procedure is outlined in the simplified flow chart in Fig. 17. Four times a second all receiver outputs are digitized and temporarily stored as a sample set (boxes 2 and 3). Each entry j in the sample set is checked for a large impulse change by testing for

$$|\text{sample value}_j - \text{old mean value}_j| < k_{ij}\sigma_{sj}$$

where the old mean value is the stored mean from the preceding time interval, k_{ij} is a scaling factor and σ_{sj} is the expected standard deviation for an individual sample in entry j. For all amplitude entries

$$\sigma_{sj}|_{\text{amplitude}} = \sigma_{aj} + (\text{old mean value}_j) \cdot C_{mj}$$

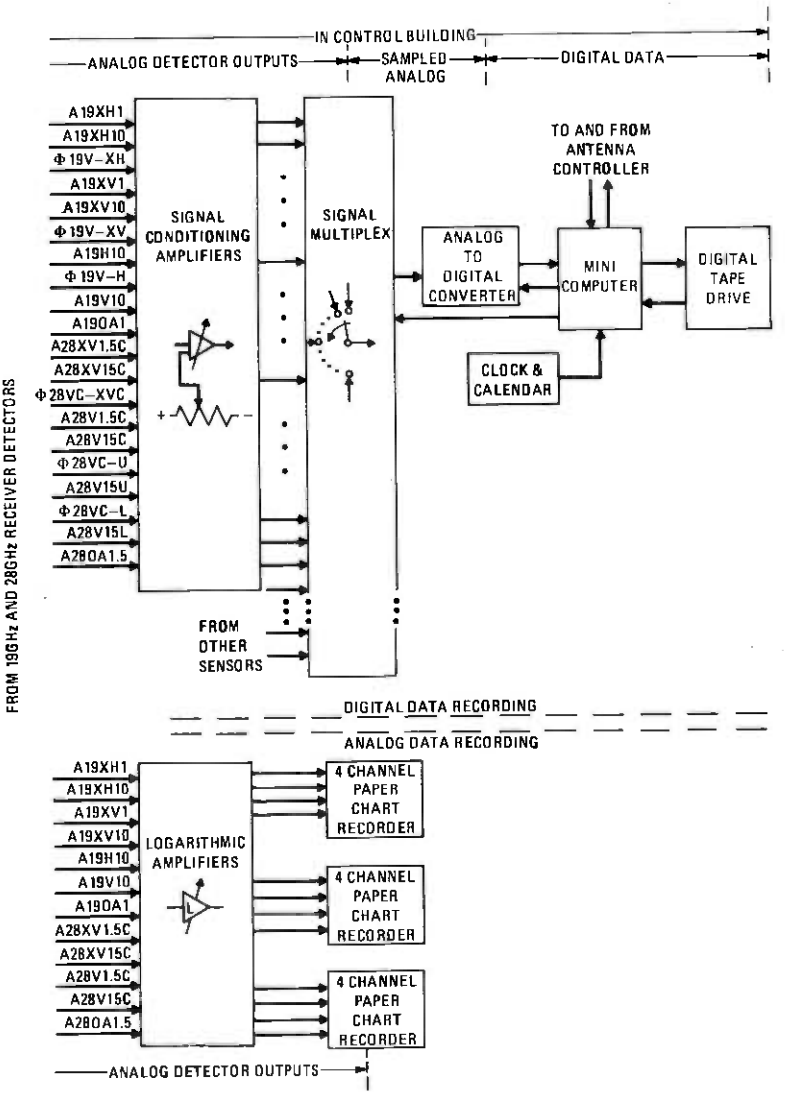


Fig. 16—Data collection equipment.

where σ_{aj} accounts for additive noise and C_{mj} weighted by the old mean value accounts for the multiplicative atmospheric effects. For phase difference entries, the multiplicative factor is inversely scaled by the old amplitude mean of the weakest signal in the phase difference pair:

$$\sigma_{sj}\,|_{\text{phase}} = \sigma_{aj} + C_{mj}/(\text{old amplitude mean}).$$

If an impulse, i.e., a large change in the data, has not occurred, the entries

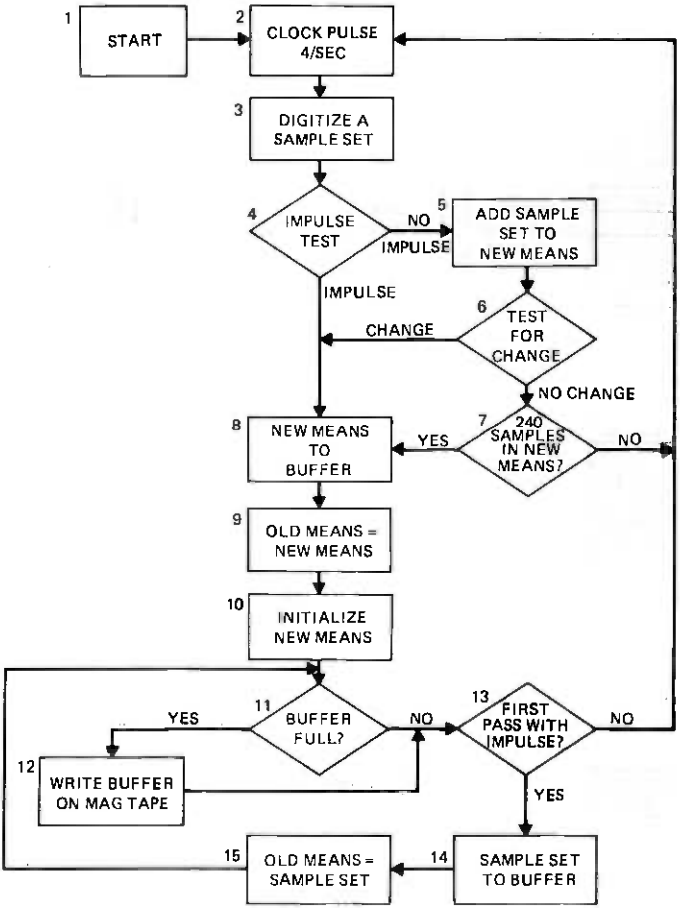


Fig. 17—Flow chart for real-time data screening software.

in the sample set are used to update a set of new mean values being accumulated (boxes 4 and 5). The set of new mean values is then tested for change by

$$|\text{new mean value}_j - \text{old mean value}_j| < k_{mj}\sigma_{sj}/\sqrt{N} \qquad (6)$$

where k_{mj} is a different (smaller) scaling constant and N is the current number of samples included in the new mean (box 6). The sample standard deviation σ_{sj} is scaled by $\sqrt{N}$ for this test since the standard deviation of the statistical fluctuations of the mean decreases by this factor (assuming independent samples). This second test quickly detects gradual signal changes. If a change is not detected and if the new mean has been accumulating for less than 1 minute, i.e., $N < 240$, then the data taking program waits for the next clock pulse (box 7). If a change is de-

tected by either of the above tests or if $N = 240$ then the new mean values are transferred to the next unused position in an output buffer (box 8), the old mean values are set equal to the new mean values (box 9) and the new mean values are initialized (box 10). If 20 entries of sets of values have been made into the output buffer, the buffer is full and it is written onto magnetic tape (boxes 11 and 12). If this pass through the procedure is not from the impulse test or is the second pass of the first impulse detected, the program waits for the next clock pulse (box 13). If this is the first pass from a detected impulse, however, the entire set of samples is transferred to the output buffer and the old mean values are set equal to the sample set (boxes 14 and 15). Thus, when an impulse occurs, both the previous running mean and the sample set containing the impulse are stored. Data screening then proceeds normally.

Each set of samples or means recorded also contains the time it was recorded so the time interval spanned is available for off-line data reduction. The data screening software includes provisions for: (*i*) stopping and starting data collection, (*ii*) recovering from primary power interruption, and (*iii*) recording calibration signals on the data tapes.

VIII. CALIBRATION SOURCE

The operation of the entire receiving system may be verified using a 19- and 28-GHz signal source with RF characteristic closely duplicating those of the beacon itself. Knowledge of the signal source amplitude also allows determination of the absolute received signal level. This source was extremely useful during prelaunch receiver construction and check out.

A block diagram of the calibration source (beacon simulator) is shown in Fig. 18. A 66.11-MHz crystal oscillator drives a multiplier chain, producing three output frequencies: 264.44 MHz, 19.04 GHz, and 28.56 GHz. Samples of the 19- and 28-GHz signals are made available to indicate the power output of the source. The 28-GHz signal is attenuated and phase-modulated by either the 264-MHz multiplier output or that output doubled to 529 MHz. The modulation index is set to produce sideband levels similar to those generated by the satellite. A precision calibrated variable attenuator allows level control from near clear-air levels to the receiver noise floor. A cross-polarized signal component is obtained from a 20-dB directional coupler.

The 19-GHz multiplier output is attenuated through fixed and variable attenuators. A cross-polarized component is generated with a 30-dB coupler. Both direct and cross-polarized signals are power-split using 3-dB quadrature hybrids. The two 19-GHz outputs are switched alternately between a "direct" and a "cross-polarized" hybrid output using PIN diode switches. A 1-MHz crystal oscillator and $\div 1000$ frequency

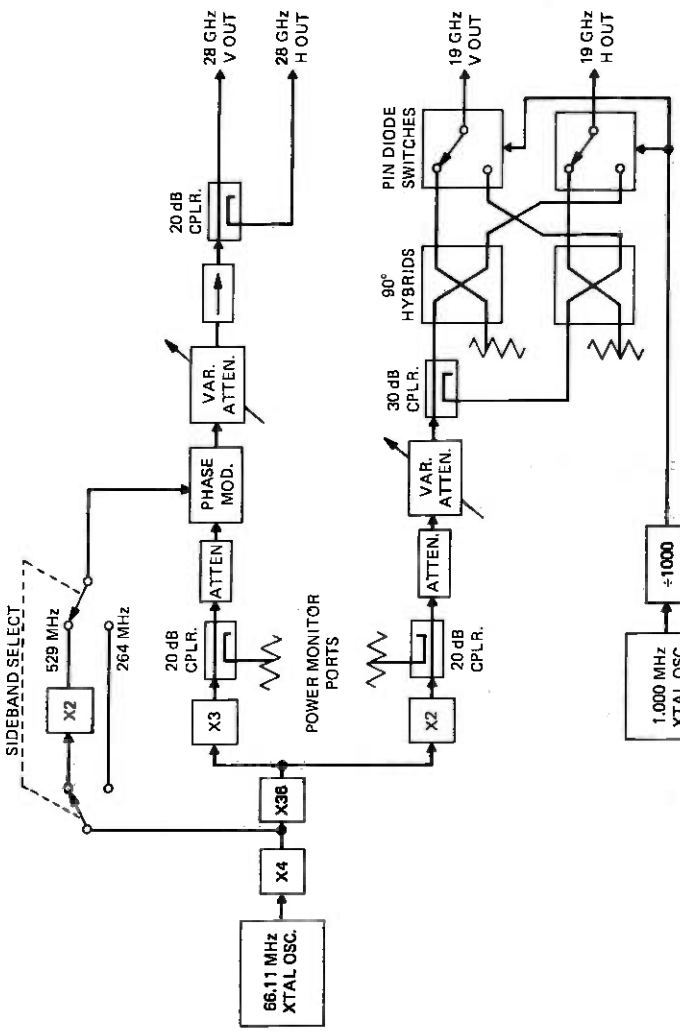


Fig. 18—Receiver calibration source (beacon simulator).

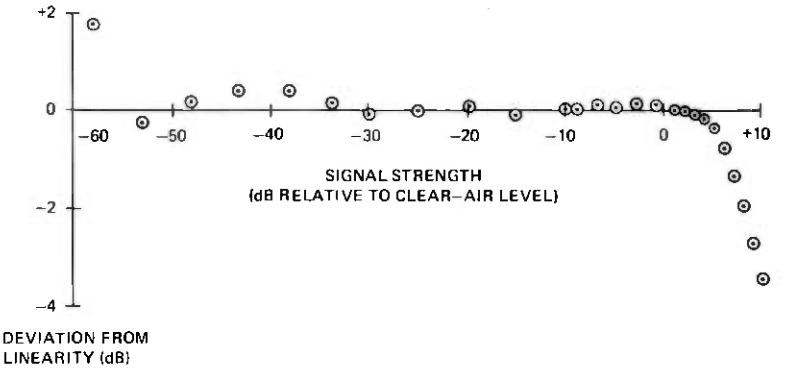


Fig. 19—Receiver input-output amplitude linearity (19-GHz vertical copolarization channel).

divider switch these switches at a 1-KHz rate. Direct and "cross-polarized" signals then appear alternately at the two 19-GHz outputs.

All four source outputs are permanently connected to their respective receiver mixer inputs through directional couplers so that the source signals can be injected with no waveguide switching. The source outputs are disabled by turning off power amplifiers in the frequency multiplier.

Since the calibration source operates in close proximity to the antenna feed frame, care was taken to hold any RF leakage from the source to <-150 dBm. Coupled with the directivity of the antenna feed assembly, this assures precise calibration signal levels down to the system noise floor.

IX. RECEIVER PERFORMANCE

The entire receiver has met all original performance requirements. Care in the design of individual assemblies has allowed integration of the entire receiving system with extremely minimal interaction or other problems. The receiver has been extremely reliable; during more than one year of operation, no significant data has been lost due to receiver failures. More detailed observations of actual receiver performance are given below.

Overall receiver linearity is nearly perfect. Figure 19 shows the deviation from perfect linearity of a typical receiver output (19-GHz vertical copol) as a function of input signal level. Other receiver outputs exhibit similar results. The calibration source was used to obtain these results; for the upper 20 dB of input level the source was temporarily connected directly to the mixer inputs. The input signal level is normalized to the clear-air satellite input level. The receiver is seen to be linear within ± 0.4 dB for all expected input levels down to 55 dB at-

tenuation. Below this level the measurement becomes noise-dominated. The 1-dB gain compression point occurs 6.5 dB above the clear-air input level. The receiver exhibits negligible differential phase shift with variations of signal amplitude. As an example, the 19-GHz copolarized differential phase is constant within 0.1° from 4 dB above the clear-air input level to the 40-dB attenuation level. Two degrees of phase error is observed at the 50-dB attenuation level.

The receiver has shown excellent long-term gain stability. Measurements over a 4-month period have shown a receiver output change of less than 0.2 dB. This figure includes both receiver and antenna gain changes and satellite power fluctuations.

Long-term measurements of 19 GHz copolarized differential phase are stable to within ±2°. This indicates the overall phase stability of both the satellite and receiver antennas and much of the receiving electronics, including all narrowband filters. Early measurements on the antenna feed frame[4] showed a temperature coefficient of <0.2°/C° at 19 GHz. This has since been improved through equalization of waveguide lengths.

The 2-MHz crystal filters have a temperature coefficient of <0.16°/C°. The two critical 19-GHz V and H copolarized filters were chosen for matched temperature coefficients, and track within ±0.2° over 15 to 40°C. The air-conditioned building environment maintains phase tracking on all other channels within ±1°.

The third and fourth IF active bandpass filters are in temperature-controlled ovens and are unaffected by ambient temperature variations. These filters were bandwidth-matched to within ±1 percent, yielding a frequency-induced phase error of <0.5° for static frequency offsets of up to 5 Hz.

As stated previously, the receiver noise floor is set by front end noise; all following amplifiers and frequency conversions contribute <0.1 dB to the overall system noise level.

The polarization switch synchronization circuitry must accurately track the 1-kHz polarization switching phase to prevent leakage of co-polarized signals into receiver crosspolarized channels. This circuitry exhibits ±5° phase jitter during a 40-dB fade and holds phase open-loop for greater attenuation. A dead band of ±9° is sufficient to maintain open-loop synchronization for at least 10 minutes.

The AFC loop will reliably track the satellite frequency excursions during a 51 dB fade at 19 GHz (copol SNR ≥ 9 dB). Below this level, the loop initiates its reacquisition sequence. Initial four-channel detection is accomplished with 50 percent probability at the 49 dB fade level, while 50 percent final verification probability is reached at the 47 dB fade level.

An example of data obtained during a typical rain event is shown in

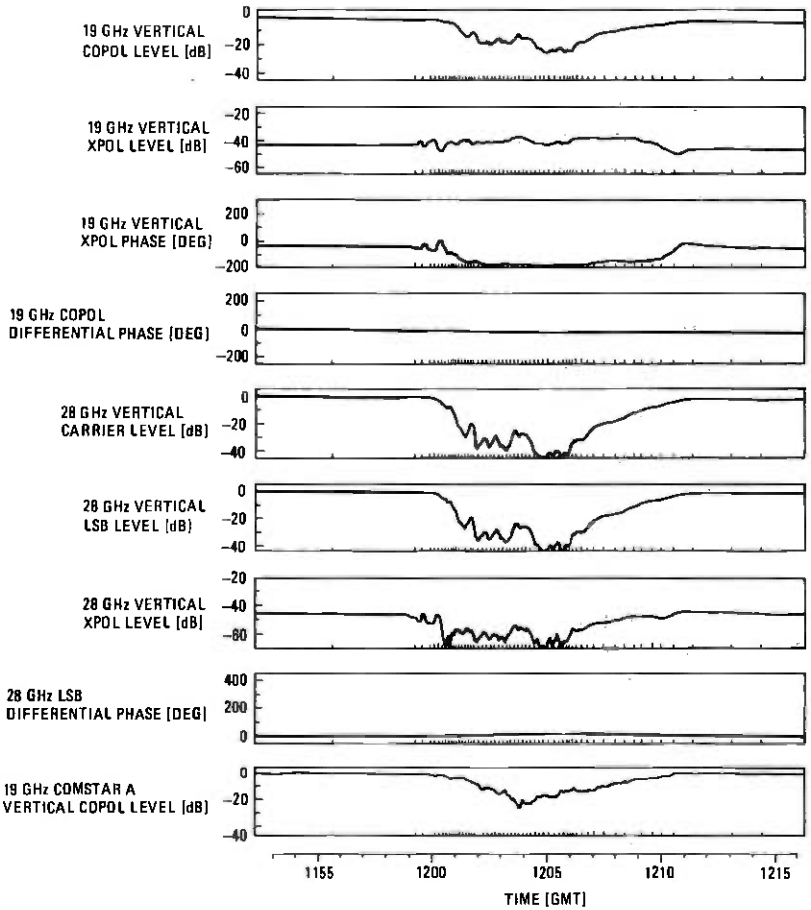


Fig. 20—Typical receiver outputs from data screening software.

Fig. 20. Only 9 of the 25 data channels recorded are shown here. The display was generated by software used for off-line data screening. The ticks along the x-axis delimit records containing 20 sets of data points (see Section VII on data collection) and indicate the increase in data-recording rate during such events.

The 4 upper traces display 19-GHz vertical copolarized (A19V) and cross-polarized (19XV) signal levels, vertical cross-polarization phase (ϕ19V-XV), and copolarization differential phase (ϕ19V-H). The cross-polarization phase zero reference was obtained by rotating the antenna feed frame slightly to leak a copolarized component into the cross-polarization channels.

The next four traces display 28-GHz vertical copolarized carrier (A28VC), lower sideband (A28VL), and cross-polarized signal level

(A28XV), and lower sideband differential phase (ϕ28VC-L). The close agreement of carrier and sideband amplitudes and lack of relative phase shift indicate little medium dispersion over the 264-MHz spacing. The phase fluctuations observed at low signal amplitudes are due to the low measurement SNR, and are not medium effects.

The final trace displays the vertical copolarized output of the colocated interim receiver[10] observing the COMSTAR beacon at 128°W longitude. The differences between this and the first trace result from the different ray paths taken through the rain.

X. SUMMARY

This paper has presented the design requirements for the Bell Laboratories Crawford Hill 19- and 28-GHz COMSTAR beacon receiving electronics, and has described the hardware and techniques used to achieve these requirements. The receiver operates unattended, continuously collecting data to determine attenuation, depolarization, differential phase shift, dispersion and angular scatter produced by precipitation. The coherence of the various transmitted beacon signals allows use of receiver noise bandwidths as narrow as 1.6 Hz to determine depolarization during severe fading. The receiving system meets all the design requirements, and has already collected data reliably for over one year.

XI. ACKNOWLEDGMENTS

Early discussions with D. M. Brady and M. J. Gans concerning satellite oscillator stability and polarization switching rates were extremely fruitful. A. W. Norris provided valuable assistance in the construction of narrowband active filters. R. W. Wilson's patient counseling and advice proved invaluable in creating the data screening software. The continued support and encouragement of D. O. Reudink has been greatly appreciated.

REFERENCES

1. D. C. Cox, "An Overview of the Bell Laboratories 19- and 28-GHz COMSTAR Beacon Propagation Experiments," B.S.T.J., this issue, pp. 1231–1255.
2. D. C. Cox, "Design of the Bell Laboratories 19 and 28 GHz Satellite Beacon Propagation Experiment," IEEE ICC '74 Record, June 1974, pp. 27E-1–27E-5.
3. R. B. Briskman, R. F. Latter, and E. E. Muller, "Call for Help," IEEE Spectrum, *11*, October 1974, pp. 35–36.
4. T. S. Chu, R. W. Wilson, R. W. England, D. A. Gray, and W. E. Legg, "The Crawford Hill 7-Meter Millimeter-Wave Antenna," B.S.T.J., this issue, pp. 1257–1288.
5. D. C. Hogg and T. S. Chu, "The Role of Rain in Satellite Communications," Proc. IEEE, September 1975, pp. 1308–1331.
6. D. C. Cox, "Some Effects of Measurement Errors on Rain Depolarization Experiments," B.S.T.J., *54*, No. 2 (February 1975), pp. 435–450.
7. L. C. Tillotson, "A Model of a Domestic Satellite System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2137.

8. D. C. Hogg, "Millimeter-Wave Communication Through the Atmosphere," Science, *159*, January 5, 1968, pp. 39–46.
9. F. M. Gardner, *Phase Lock Techniques*, New York: John Wiley and Sons, 1966.
10. H. W. Arnold, D. C. Cox, and D. A. Gray, "The 19-GHz Receiving System for an Interim COMSTAR Beacon Propagation Experiment at Crawford Hill," B.S.T.J., this issue, pp. 1331–1339.
11. A. J. Rustako, "An Earth-Space Propagation Measurement at Crawford Hill Using the 12-GHz CTS Satellite Beacon," B.S.T.J., this issue, pp. 1431–1448.
12. D. C. Cox, H. W. Arnold, and A. J. Rustako, "Some Observations of Anomalous Depolarization on 19 and 12 GHz Earth-Space Propagation Paths," Radio Science, *12*, No. 3 (May–June 1977), pp. 435–440.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

COMSTAR Experiment:

The 19-GHz Receiving System for an Interim COMSTAR Beacon Propagation Experiment at Crawford Hill

By H. W. ARNOLD, D. C. COX, and D. A. GRAY

(Manuscript received January 10, 1978)

This paper describes the antenna and receiving electronics for the Bell Laboratories Crawford Hill 19-GHz COMSTAR Interim Experiment. This experiment has collected essentially continuous 19-GHz attenuation data on the earth-space path to the COMSTAR A satellite since May 25, 1976. The receiver operates unattended, automatically reacquiring the beacon signals after deep rain-induced signal fades. A receiver bandwidth of 10 Hz allows accurate measurement of fade depth to the 40-dB level.

I. INTRODUCTION

A receiving system for the 19-GHz COMSTAR beacons[1,2] has been operating since May 25, 1976, at the Bell Laboratories Crawford Hill facility at Holmdel, New Jersey. This system is less complex than the precision receiving system[3,4] also operating there, but uses the same overall system design and many similar components. This paper will discuss the antenna and receiving electronics for this system.

This receiving system was used initially for observations of the first COMSTAR beacon as it crossed the horizon at Crawford Hill. The main receiving system electronics[3] were operated for a short period in parallel with this system, while awaiting completion of the Crawford Hill millimeter-wave antenna.[4]

The interim receiving system is presently observing the beacon located at 128°W longitude. The propagation path from Crawford Hill to the beacon has an azimuth of 244.7° and an elevation of 18.5°. The beacon characteristics are similar to those given in Table I of Ref. 1; the incident

polarizations are within 5° of vertical and horizontal at Crawford Hill. Beacon observations have been nearly continuous since the start of the experiment. Much useful information has already been obtained and has been reported elsewhere.[5,6,7] Since the main receiving system is observing the satellite at 95°W longitude, comparison of the two receiver outputs should indicate the efficacy of "satellite diversity" from a common earth terminal.

Future plans for the interim receiving system include its use as a remote space diversity site. This will allow accumulation of joint fading statistics to fade depths not achievable in previous radiometer experiments.[8]

In its basic configuration, this receiver records the amplitudes of two orthogonal components of the 19-GHz incident radiation. Two identical receiver channels are used. Narrow receiver noise bandwidths are used to maximize the measuring range. Automatic frequency tracking and reacquisition allow unattended operation. The receiving antenna beamwidth is sufficiently wide that no tracking of diurnal satellite motion is required.

The general design considerations for this receiving system are similar to those discussed for the main receiving system.[3] The receiving antenna is described in Section II. Section III discusses the receiving electronics. Receiver performance and some typical data are included in Section IV.

II. ANTENNA

The antenna and feed assembly were originally designed for a 20-GHz propagation experiment using a beacon on the ATS-6 satellite.[9] The feed assembly and antenna positioner were later modified for this experiment. The antenna is of Cassegrainian design, with a 12-foot-diameter aperture and two orthogonal linearly polarized feeds. Antenna positioning is controlled remotely from the equipment building. The complete antenna assembly is shown in Fig. 1.

The 12-foot-diameter spun-aluminum main reflector was manufactured for nominal use at C and X bands. After tensioning the reflector to correct a surface warpage, however, good performance was obtained at 20 GHz.

The 16-inch hyperboloidal subreflector is supported by four aluminum I beams, thinned for minimum depolarization. Both subreflector and support had nonessential material milled away to minimize weight.

Polarization diplexing is accomplished using a quasioptical polarization separator, as in the main antenna feed assembly.[4] This technique is shown in Fig. 2. The polarization separator is a grid of parallel copper strips on a thin mylar support membrane. The grid is oriented so that one polarization, *B* (polarized parallel to the page), is transmitted

Fig. 1—View of 12-foot antenna for COMSTAR interim experiment.

through the polarizer to aperture I. The orthogonal polarization, A, is reflected to aperture II.

Short-focal-length paraboloids at apertures I and II reflect the received energy to dual-mode feedhorns. The feedhorn size produces a 20-dB edge illumination taper at the main reflector. The feed assembly is rotatable from the equipment building to allow alignment of the feed with the incident polarization angle.

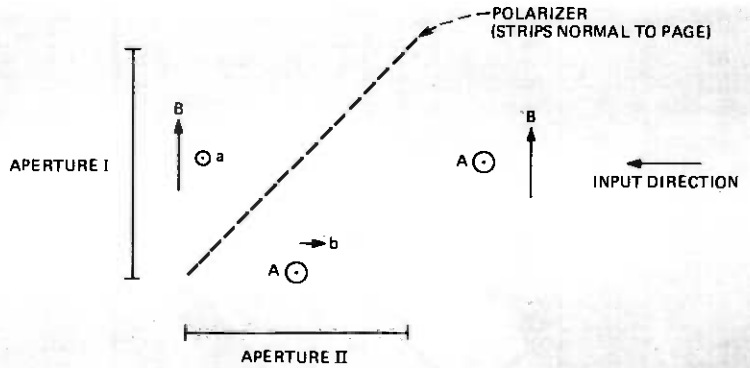


Fig. 2—Polarization separator geometry for 12-foot antenna.

The antenna is mounted on a modified Nike-Hercules elevation-over-azimuth positioner. New drive motors and position encoders allow positioning the antenna from within the equipment building to 0.01 degree precision. Since the satellite position is controlled to within 0.1 degree, continuous position tracking is not required.

The feed assembly and mylar rain window are shrouded with fiberglass weather covers. These are covered with reflective aluminum foil to minimize the "greenhouse" effect. Electric heaters and forced ventilation regulate the temperature within the enclosure.

III. RECEIVING ELECTRONICS

The electronics for the interim receiver are similar to those used in the main COMSTAR beacon receiver.[3] This section will rely heavily on the description of the main receiver electronics in this issue. The basic design philosophy and rationale behind the choice of IF frequencies, etc., is covered there.

A block diagram of the interim receiver electronics is shown in Fig. 3. The receiver consists of two unswitched 19-GHz receiver channels and somewhat simplified frequency control equipment. The first frequency conversion is performed at the antenna feed. All other equipment is located in a building alongside the antenna.

Throughout the receiver, care was taken to assure >60 dB overall isolation between receiver channels. Liberal component shielding was used, and isolators were used where necessary to avoid coupling through common local oscillator lines.

The two antenna feed outputs are down-converted to 1.003 GHz by Schottky-diode mixer-preamplifiers with 6.5-dB single-sideband noise figure. The 18.037-GHz first LO is generated at the feeds from an oscillator in the support building.

The two IF signals are fed to the support building through coaxial cable and are filtered by 0.3 GHz BW bandpass filters to avoid noise saturation of the following wideband IF stages. Down-conversion to 2.067 MHz is performed with image rejection mixers, which use phase cancellation to suppress image noise by >20 dB. Coarse frequency tracking is performed at this conversion, as will be discussed later.

After further amplification, the system noise bandwidth is further constrained with 5-MHz low-pass filters. Step attenuators set the clear-air signals to the desired levels. Amplifiers are used to isolate 6-kHz BW crystal bandpass filters, which provide image rejection at the next conversion.

Balanced mixers perform the next frequency conversion to 6.25 kHz. Short-term frequency instabilities are removed at this step with a phase-locked loop, whose operation will be described later. The 6.25 kHz signals are amplified and filtered by 250 Hz BW active bandpass filters. These and the following filters are mounted in temperature-stabilized ovens for improved gain and frequency stability.

The final frequency conversion to 325 Hz is performed by an active linear multiplier. The final predetection bandwidth is set by 10-Hz BW active bandpass filters. These filters exhibit a single-pole response with 16-Hz noise bandwidth. These filters strip off the 1-kHz polarization-switching modulation and pass only the carrier frequency.

Signal amplitudes are determined with linear amplitude detectors exhibiting ±0.1 dB linearity over 60-dB signal range. The detector outputs are processed by dc logarithmic amplifiers for better display resolution during deep rain fades. The two log amplifier outputs are recorded on a paper chart recorder operating at 4 inches/hour. These two outputs are also fed to the main receiver data recording equipment over telephone lines, using voltage/frequency and frequency/voltage converters. Log, rather than linear, recorder outputs are used to avoid dc offset problems at low signal levels.

Since a measurement of differential amplitude between the two received polarizations is desired, the two log amplifier outputs are subtracted and this difference recorded, with higher sensitivity, on the chart recorder. In addition, 1- and 10-minute timing markers are recorded from the main receiver to allow time synchronization of the two systems.

The receiver must track both the long- and short-term beacon frequency fluctuations. A digitally controlled AFC loop is used to track the thermally induced diurnal fluctuations. This loop is illustrated in Fig. 9 of Ref. 3. A sample of the 2-MHz vertically polarized signal is down-converted to 6.25 kHz and fed to a narrowband discriminator made from two stagger-tuned active filters. The discriminator output is integrated for 1 second. If the average frequency error exceeds 2 Hz, the output frequency of a digitally controlled frequency synthesizer is incremented

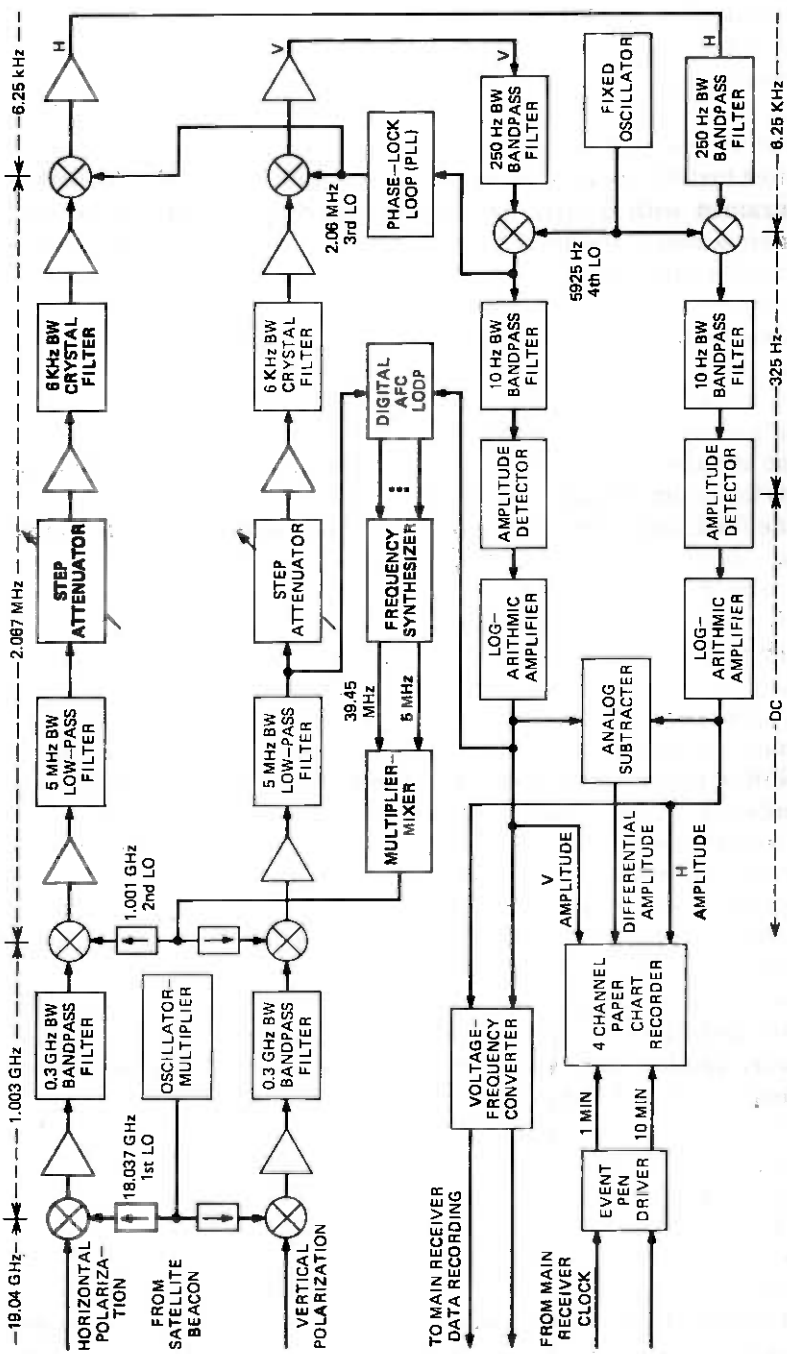


Fig. 3—Block diagram of COMSTAR interim receiver.

or decremented by 1 Hz. The synthesizer output is doubled and used to make up the second local oscillator. Thus, frequency corrections are made in increments of 2 Hz at 1 Hz rate. The loop will track a frequency excursion of ±2 Hz/sec, greater than that expected from the beacon.[1,3]

If the signal at the vertically polarized receiver output falls below a preset threshold, the AFC loop will be unable to maintain track and so initiates an ever-expanding search around the last known beacon frequency. When the signal is again detected at the receiver output, this search ceases and tracking resumes. This technique does not prohibit acquisition of a polarization-switching sideband, but performs adequately and is much simpler to implement than the technique used in the main receiver.

Short-term frequency fluctuations are tracked with a phase-locked loop (PLL). This loop locks the unfiltered 325-Hz vertically polarized received signal to a stable 325-Hz oscillator through adjustment of the 2-MHz third conversion oscillator. Since both vertical and horizontal signals are phase-coherent, the horizontally polarized signal will be locked as well. This loop has a 30-Hz bandwidth and is described in greater detail in Ref. 3.

IV. RECEIVER PERFORMANCE

The interim experiment has operated essentially continuously since May 25, 1976, and has met its design objective of collecting continuous 19-GHz amplitude statistics from the COMSTAR A satellite. Receiver failures have been minimal, and the conservative design approach taken allowed integration of the entire receiving system with no unexpected interactions between subassemblies.

Since most subassemblies for this receiver are identical to those used in the main receiver, most performance measures are identical for the two. Linearity and long-term stability are discussed in Section IX of Ref. 3.

Since this receiver operates without polarization switching using a smaller antenna aperture, the measured clear-air SNR is 50 dB. The AFC loop threshold is set at the 40-dB fade level; below this level the AFC initiates a frequency search to attempt to reacquire the beacon signal. Reacquisition is accomplished reliably to 11 dB SNR, corresponding to a fade depth of 39 dB.

Since the receiving antenna does not track the satellite motion, small diurnal amplitude variations are observed as the satellite traverses the antenna directivity pattern. These variations are generally less than 1 dB peak-to-peak. Since they are of long duration and repeatable on a day-to-day basis, they pose little problem during data reduction.

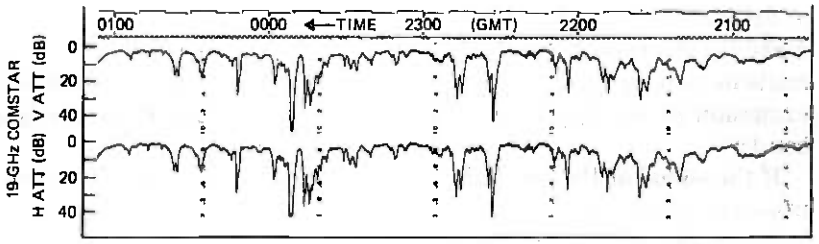


Fig. 4—Typical data obtained from COMSTAR interim experiment (Hurricane Belle, August 9, 1976).

An example of the data collected by this experiment is shown in Fig. 4. These data were taken August 8–9, 1976, during the passage of hurricane Belle 50–60 miles east of Crawford Hill. Time, indicated in GMT, runs from right to left. The upper and lower traces indicate the vertically and horizontally polarized received signal strengths and are calibrated in dB attenuation from the clear-air signal level. While this event is clearly not a typical one, the data shown indicate excellent receiver performance during periods of great environmental stress.

V. SUMMARY

This paper has presented a brief description of the antenna and receiving electronics for the Bell Laboratories Crawford Hill 19-GHz COMSTAR Interim Experiment. This equipment has collected essentially continuous amplitude data from the COMSTAR A satellite since it first became visible at the horizon on May 25, 1976. A received bandwidth of 10 Hz allows accurate measurement of fade depth to 40-dB level. These data, together with those collected by the main Crawford Hill COMSTAR propagation experiment, will be used to characterize earth-satellite propagation at 19 GHz.

VI. ACKNOWLEDGMENTS

Many people contributed to the timely operation of this experiment. The antenna and feed system were available through the encouragement of D. C. Hogg and the design and construction effort of R. H. Turrin. H. H. Hoffman contributed much to the receiver design. Assembly of the receiving electronics was done ably by R. H. Brandt, M. F. Wazowicz, and R. P. Leck; the latter has also aided in the continuing operation of the experiment. The continuing encouragement and support of D. O. Reudink has been invaluable.

REFERENCES

1. D. C. Cox, "An Overview of the Bell Laboratories 19- and 28-GHz COMSTAR Beacon Propagation Experiments," B.S.T.J., this issue, pp. 1231–1254.

2. D. C. Cox, "Design of the Bell Laboratories 19 and 28 GHz Satellite Beacon Propagation Experiment," IEEE ICC '74 Record, June 1974, pp. 27E-1—27E-5.
3. H. W. Arnold, D. C. Cox, H. H. Hoffman, R. H. Brandt, R. P. Leck, and M. F. Wazowicz, "The 19- and 28-GHz Receiving Electronics for the Crawford Hill COMSTAR Beacon Propagation Experiment," B.S.T.J., this issue, pp. 1289–1329.
4. T. S. Chu, R. W. Wilson, R. W. England, D. A. Gray, and W. E. Legg, "The Crawford Hill 7-Meter Millimeter-Wave Antenna," B.S.T.J., this issue, pp. 1257–1288.
5. D. C. Cox and H. W. Arnold, "Preliminary Results from the Crawford Hill 19 GHz COMSTAR Beacon Propagation Experiment," presented at USNC/URSI Meeting, October 11, 1976, Amherst, Mass.
6. D. C. Cox, H. W. Arnold, and H. H. Hoffman, "Differential Attenuation and Depolarization Measurements from a 19 GHz COMSTAR Satellite Beacon Propagation Experiment," presented at URSI Symposium on Propagation in Non-ionized Media, April 28, 1977, La Baule, France.
7. D. C. Cox, H. W. Arnold, and A. J. Rustako, "Some Observations of Anomalous Depolarization on 19 and 12 GHz Earth-space Propagation Paths," Radio Science, May–June 1977, pp. 435–440.
8. R. W. Wilson and W. L. Mammel, "Results from a Three Radiometer Path-diversity Experiment," IEE Conference on Propagation of Radio Waves at Frequencies Above 10 GHz, London, April 1973.
9. R. H. Turrin, "A Quasi-Optical Antenna Feed," presented at USNC/URSI Meeting, June 1975, Urbana, Ill.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

COMSTAR Experiment:

COMSTAR Beacon Receiver Diversity Experiment

By N. F. DINN and G. A. ZIMMERMAN

(Manuscript received December 9, 1977)

The design and realization of 19-GHz and 29-GHz beacon receivers for implementation of the remote site diversity reception experiment are discussed. The experiment objectives and constraints are investigated in terms of their impact on equipment realization. Data acquisition and retrieval problems associated with remote sites are also addressed. Finally, some of the results obtained from early operation are presented. These results, obtained from direct measurement of the beacons, correlate very well with earlier radiometer measurements scaled in frequency with appropriate corrections made for the impact of energy scattering due to rain.

I. INTRODUCTION

While most current-generation communication satellites operate in the same common carrier bands (4 and 6 GHz) as do terrestrial facilities, operation at significantly higher frequencies—such as the 12- to 14-GHz band, and the 18-, 30-GHz bands—offers a number of important advantages. Among these are expected reduction in interference, reduced spacecraft component sizes, and higher gain spacecraft antenna with the opportunity for independent multiple beams within the continental United States (CONUS).

On the other side of the ledger, radiation at these frequencies is more effectively scattered and attenuated by water droplets, thus threatening system operation with attenuation, depolarization and dispersion effects. These problems are under empirical investigation utilizing beacon sources carried aboard the COMSTAR satellites. This new opportunity follows a fruitful (but limited) period of measurements with radiometers.[1,2,3]

Where uncertainties in radiometer results existed in the past, due to their range limitation and their inability to account for signal loss due to scattering, they can now be overcome using the beacons. Thus the objectives of this new experimental phase are manifold but they can be summarized as follows:

(*i*) Directly obtain continuous, long-term attenuation measurements at (the higher) system frequencies (thus accounting for both absorption and scattering).

(*ii*) Increase the dynamic fade measurement range to at least 30 dB.

(*iii*) Provide for direct comparison of radiometer and beacon measurements obtained simultaneously from the same antenna, thus allowing qualification of radiometer data in hand.

(*iv*) Extend the available site diversity performance data base by operating in different meteorological environments.

Satisfaction of these objectives constrained the design of our receiving stations. In particular, the need for reliable continuous operation requires a capability not only for monitoring the system remotely but also for retrieving data to a central site. The system must provide for unmanned operation over extended periods without loss of data. Finally, the recognition that test sites would be distant (and costly to visit) required that the experiment be implemented to be self-contained and transportable with minimum expense and delays, that equipment failures be very infrequent, and that simple failures not compromise all observations. Consequently, conservative design approaches were used throughout, and provisions were made for operation alarming. The method used to achieve the experimental objectives is discussed in subsequent sections.

Section II presents a discussion of the primary factors which shaped the experimental setup, including such things as anticipated signal-to-noise ratios, the importance of remote test sites, and cost. This is followed in Section III by a detailed description of the data acquisition requirements and their realization. Section IV begins the discussion of the receiver, starting with the preliminary operating objectives. The overall receiver realization is presented in some detail in Section V and includes discussion of the antenna with its auxiliary equipment, the scanning receiver, the 19-GHz receiver, the 28.5-GHz receiver, and the frequency predictor which compensates for satellite drift in the absence of beacon frequency update information. Results obtained during the initial experiment phases are summarized in Section VI.

The Phase I experimental sites are located at Grant Park, Illinois (near Chicago), and Palmetto, Georgia (near Atlanta). Each principal site is equipped with beacon receivers (19 and 29 GHz) and a 13-GHz radiometer. Associated with these sites are remote radiometer sites which

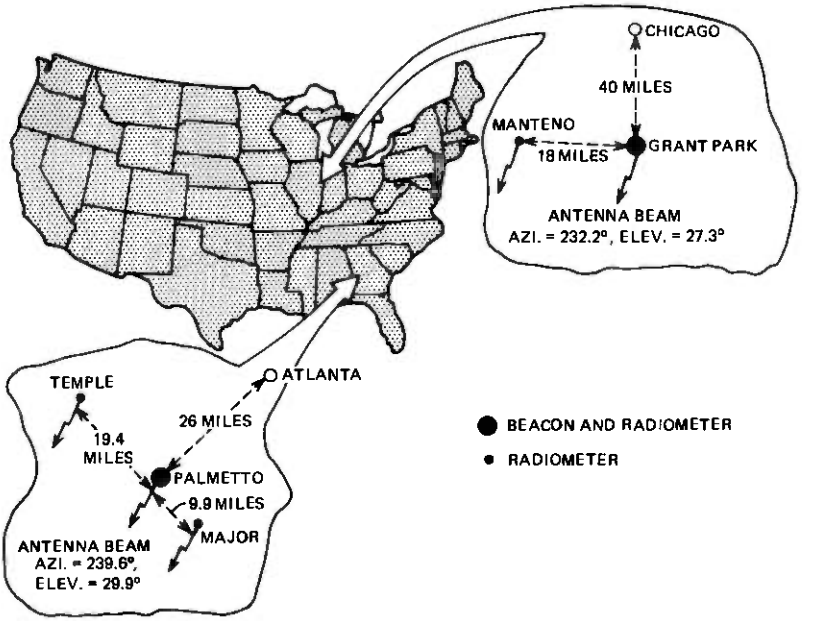


Fig. 1—Beacon/radiometer sites.

provide data for diversity performance evaluation. Figure 1 indicates Manteno, Illinois, approximately 20 miles west of Grant Park, and Temple, Georgia, approximately 20 miles northwest of Palmetto.

II. PRIMARY EXPERIMENTAL CONSIDERATIONS

Six factors were of prime importance to the design of the beacon reception equipment:

(*i*) The expected power level of the received signals and the required dynamic measurement range.

(*ii*) The spectral width, maximum frequency variations and the rate of drift of the beacon oscillators.

(*iii*) The anticipated satellite stationkeeping excursions.

(*iv*) The requirement for remote unattended and continuously operative stations.

(*v*) The need for coordinated diversity site radiometer data acquisition.

(*vi*) "Minimum" cost procurement and operation.

2.1 Preliminary signal-to-noise ratio calculations

The beacon signal levels and CONUS coverage are detailed in Ref. 4;

the relevant parameters are summarized below:

	19 GHz	29 GHz
EIRP	53 dBm	56 dBm
Polarization loss*	3 dB	—
Path loss	210 dB	213 dB
Fade range	~35 dB	~38 dB
Min. signal level	<-195 dBm $+ G_A$†	<-198 dBm $+ G_A$†
Down converter + preamp noise figure	6.5 dB	6.5 dB

The corresponding noise power in a 100-Hz‡ bandwidth is readily calculated:

$$P_a = N_F + 10 \log \text{KTB}_w + 30 \text{ dBm}$$
$$= 6.5 - 174 + 10 \log 100 \text{ dBm}$$
$$P_a = -147.5 \text{ dBm}$$

Allowing a few dB loss of sensitivity due to satellite stationkeeping variations, antenna misalignment, network loss, etc., implies that achievement of a usable signal-to-noise ratio during deep fades requires an antenna gain of 50 dB or more. In addition, to limit the noise power, it implies a final processing bandwidth significantly less than 100 Hz—which therefore requires the receiver to track the beacon frequency, maintaining lock during fades of at least 35 dB.

2.2 Beacon frequency variation

The beacon reference oscillator frequency exhibits a frequency variation less than ± 1 part in 10^6 on a diurnal basis, and ± 1 part in 10^6 per year due to aging. In addition the receiver oscillator has been allocated ± 1 part in 10^6 per year for aging, thus implying a total potential annual variation to be accommodated of ± 3 parts in 10^6 (approximately ± 57 kHz at 19 GHz). Consequently, the receiver capture and tracking range was specified to be 100 kHz.

In addition to maximum frequency excursion considerations, the short-term rate of change of frequency must be accommodated. This is a major complication since receiver sensitivity must be obtained by severe band limiting. The receiver configuration chosen utilizes parallel

* Only one polarization is processed.
† G_A, the antenna gain, is determined in Section 2.3.
‡ The beacon design objective was to provide 90 percent of the signal power in a bandwidth of 100 Hz at 19 GHz, or within 150-Hz bandwidth 29 GHz.

processing: a set of 32 narrowband comb-filters for selectivity, and an AFC loop to center the comb set about the instantaneous received frequency. The 32-filter comb set in the 19-GHz receiver spans a total of 1.6 kHz; the AFC loop drives the frequency of the down-converted signal to the center of the comb filter range, allowing only a ±800 Hz band to account for drift errors which are uncompensated by update information which, of course, would be unavailable during periods of severe fading. The maximum estimated drift rate* of approximately 1 Hz/sec implies that, even in the event of deep fades (which preclude feedback frequency updating), there would be no reacquisition delay for outages less than about 15 minutes. Still longer outages could allow drift accumulations greater than 800 Hz, with consequent delay to reacquisition. This compromises data, in that the end of fade would be uncertain. Therefore beacon frequency prediction based upon beacon behavior observed prior to the fade is included in the design. This feature extended the receiver capability for continuous operation in extremely deep fade situations (i.e., no frequency update information) from 15 minutes to over 2 hours.

2.3 Satellite stationkeeping

The satellite, while nominally stationary, actually moves within a station of ±0.1° in latitude and ±0.1° in longitude. As a consequence, the ground station antenna must either track this variation or sacrifice absolute gain to provide essentially equal response within the sector traversed by the satellite. Allowing a 1-dB maximum pattern variation due to stationkeeping implies a minimum 3-dB antenna beamwidth of about 0.3°. This corresponds to an antenna gain of about 56 dB, which is consistent with the minimum antenna gain requirement (> 50 dB) necessary to provide the required dynamic range.†

2.4 Remote test sites

Practical operation of remote unattended stations requires that equipment be conservatively designed, with broader operating margins than would be necessary if frequent adjustments could be made, and the equipment must have automatic (re-)startup features. In addition, redundant recording equipment is necessary to preclude the loss of interesting data. Finally, representative data should be remotely accessible to allow daily monitoring for both the health of the equipment and the progress of the experiment.

---

* Based on COMSAT preflight test curves.

† Antenna selection is discussed in Section 5.1.

2.5 Correlation with radiometers

A secondary objective of the experiment was to allow detailed comparison of radiometer and beacon observations, hence the requirement to derive a radiometer signal from the same antenna as that supporting the beacon receivers. A 13-GHz radiometer was chosen; this provides a reasonable match to the dynamic measurement ranges obtained with the beacons and, through frequency scaling, allows calculation of absorption at the beacon frequencies. Scattering losses may be estimated and combined with the measured absorption losses, scaled for frequency differences, and compared with the beacon losses.

2.6 Cost

The final restriction, obligatory to all operations, is cost limitation, particularly since multiple sites were to be equipped. It was this consideration that tipped the balance in favor of a fixed, limited gain antenna. This restriction also dictated that standard, readily available components be used in lieu of custom devices.

III. DATA ACQUISITION

The operation of numerous remote sites, both for beacon reception and for associated radiometer studies, places a high emphasis on data-remoting capabilities (see Fig. 1). It is necessary, of course, to transfer data between associated sites to assess diversity performance; it is equally necessary to transfer data back to Bell Laboratories, Holmdel, for monitoring purposes. The transfer between test sites is accomplished using a 12-bit analog-to-digital converter to drive an FSK telemetry system over dedicated phone lines. At the receiving end, the digital signal is reconverted to analog and processed in conjunction with signals received directly at the main site. Transmittal of summary data back to Holmdel is accomplished on a *dial-up basis* over the DDD network *once* a day.

Data are recorded in a number of different forms: At each site a real-time record of the various received signal levels is kept on stripcharts. At the main sites, Palmetto and Grant Park, stripcharts record the simultaneous levels of both the local and the diversity signals. In addition, at each main site there is a Portable Propagation Recorder (PPR),[5] which records the accumulated time during which the signal level is below various fade thresholds, as well as the number of times that a threshold level is traversed. This record is stored in a solid-state memory, the contents of which are accessible by telephone. Finally, to ensure against accessing problems, equipment failure, or holiday weekends, the data are dumped daily onto a local punched paper tape.

For several years prior to the beacon experiment, there existed at Palmetto a computer-directed data-gathering complex supporting other propagation experiments, including the ongoing radiometer tests. This complex, known as MIDAS (Multiple Input Data Acquisition System) uses magnetic tape to store a sampled time record of signal variations. Its sampling rate is five times per second on each channel. For this experiment the Palmetto MIDAS complex records the levels of not only both beacon signals and the on-site 13-GHz radiometer signal at Palmetto but also the 18-GHz radiometer signal from the remote diversity site. In addition it records the instantaneous beacon frequency and the local rain rate as sampled in a tipping bucket gauge. The magnetic tape record is then mailed to Holmdel on a weekly basis for processing.

IV. RECEIVER OPERATING OBJECTIVES

This beacon receiver was designed to the following objectives:

(*i*) Acquire the signal within 15 seconds either after turn-on or after an extended period of signal dropout. Similarly, in the event of power failure, automatically reacquire the beacon signals without external intervention.

(*ii*) Provide accurate (±0.5 dB absolute and ±0.1 dB relative) fade indications over at least a 30-dB dynamic range, at both 19 and 29 GHz.

(*iii*) For periods up to 1 hour, in the event of loss of signal due to fading, reacquire (virtually instantaneously) when the signal recovers to the fade depth at which it was lost.

(*iv*) Track frequency variations and provide output indications accurate to approximately 1 part in 10^8.

Each of the above objectives was not only achieved but the realized receiver exceeded the required performance.

The functions identified above are realized in the receiver shown in the simplified block diagram in Fig. 2. Each functional block is treated in some detail following a brief operational summary.

V. RECEIVER REALIZATION

Initial acquisition of the 19.04-GHz signal is accomplished by a scanning receiver which searches a bandwidth of 95.9 kHz for the 19-GHz tone and the two (polarization) switching sidebands, which are 1 kHz removed from the carrier.

Following acquisition, the two receivers, one at 19 GHz and one at 29 GHz, begin monitoring fades. Associated with the 19-GHz receiver is a frequency predictor which, under normal operating conditions, continuously monitors the received beacon (19 GHz) frequency and the rate

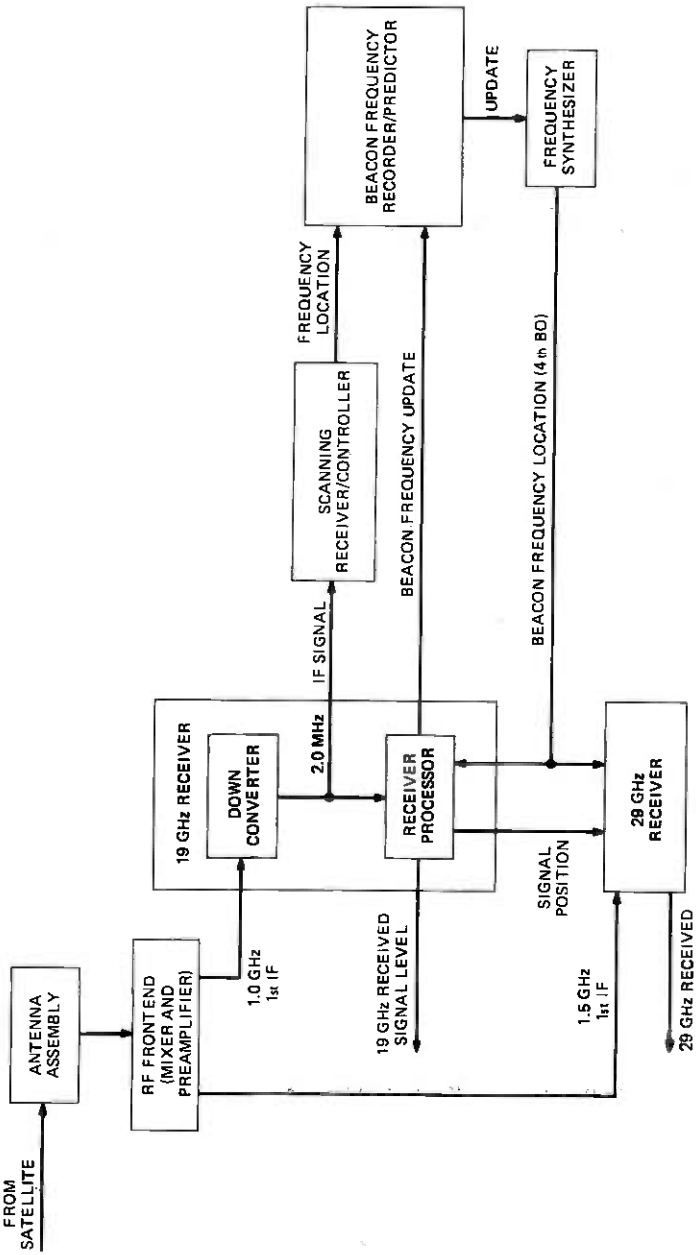


Fig. 2—Simplified receiver block diagram.

at which it is drifting. The coherent relationship between 19- and 29-GHz beacons allows slaving of the 29-GHz receiver frequency to the 19. In the event of loss of the 19-GHz signal for an extended period of time, the predictor extrapolates the last valid frequency measurement using the most recently estimated frequency derivative. This ensures that the local oscillators of the receivers track the beacon frequency and are in best condition to reacquire the signal once it reappears. Note that loss of only the 29-GHz signal has no impact on the frequency tracking of either receiver.

In addition to the basic functions identified previously, there are several other functional units which are separately identified and discussed prior to treatment of the major units. These units are: antenna assembly, frequency multiplier, frequency synthesizers, and comb filters.

5.1 Antenna assembly

The antenna assembly, shown diagramatically on Fig. 3, includes: antenna mount with azimuth and elevation adjustments, antenna, feedhorn, polarization adapter assembly, polarization coupler and a frequency diplexer.

The functions of the antenna assembly are to: intercept sufficient beacon signal energy for processing, separate a 13-GHz signal for driving a radiometer, and separate the 19- and 28.5-GHz signals for measurement by the beacon receiver.

Recall, from Section 2.3, that a minimum 3-dB bandwidth of 0.3° was necessary to meet the amplitude misalignment objective and that it was this beamwidth constraint that restricted the antenna aperture to 8 feet. The antenna selected is a CH-8 (7.5-foot aperture) conical, horn reflector antenna manufactured by Antennas for Communications Incorporated. Using a specially designed feed horn tapered to WC-65, measurements were made at 19.04 GHz and 26.01 GHz (the highest frequency available from the equipment used for tests) which indicated gains at the two frequencies of 51.0 dB and 53.6 dB for the vertical polarizations and 51.0 dB and 53.4 dB for the horizontal polarization. Scaling to 28.56 GHz implies gains of 54.5 dB for the vertical and 54.2 dB for the horizontal polarizations. Since the satellite could traverse a ±0.1° window, it was important to determine the impact of a fixed orientation on received amplitude. At 28.5 GHz, a 0.1° misalignment results in approximately 1-dB gain decrement. At the time the receiver was installed, the antenna orientation was optimized for the two beacon frequencies and then secured permanently.

The polarization adapter assembly is a rotatable framework attached to the feedhorn which permits continuous adjustment of the angular

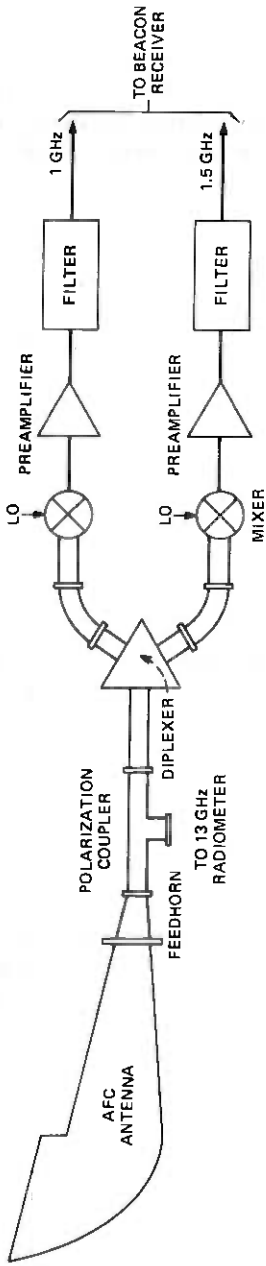


Fig. 3—Antenna assembly and RF front end.

relationship between the feedhorn (antenna) and the polarization coupler. This ensures that the nominal received signal polarization can be accounted for readily, thus placing the polarization coupler in position to couple maximum energy through the vertically polarized port for both 19 and 29 GHz.

The 13-GHz radiometer signal is obtained via a polarization coupler attached to the feedhorn. It is obtained from the "horizontally" polarized port, while the orthogonal port* provides the "vertically" polarized signals (19 and 28.5 GHz) which drive the beacon receiver. This coupler provides considerable flexibility and introduces only 0.1 dB insertion loss at 13 GHz, 0.25 dB at 19 GHz and 0.35 dB at 28.5 GHz. The return loss at each of these frequencies is greater than 30 dB. This technique provides the beacon receiver only one component of the 19-GHz signal; the horizontally polarized component is terminated.

The final item of the antenna assembly is the diplexer. Conceptually, this is a waveguide 120° Y junction with a high-frequency "short" in one output leg and a low-frequency "short" in the second output leg. This readily permits separating the two beacon signals at a loss (insertion) penalty of 0.2 dB at 28.5 GHz and 0.3 dB at 19 GHz.

5.2 Frequency multiplier

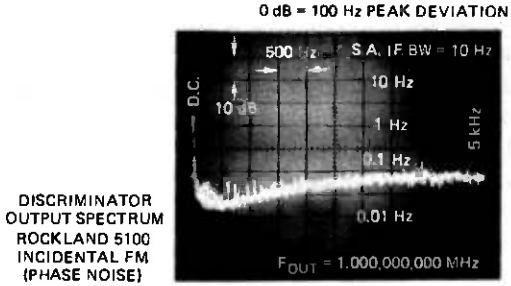
The receiver utilizes 5 IF frequencies† to obtain the required gain and selectivity prior to detection. The first IF operates at 1.0 GHz; the first beat-oscillators (BO at 18.04 GHz and 27.06 GHz) are obtained from a frequency multiplier built by RDL to Bell Laboratories specifications. This supply is basically a multiplier chain (X288 and X432) which operates upon one precision reference frequency of 62.53888 MHz which is developed within the receiver from a 10.000-MHz reference source.

5.3 Frequency synthesizer

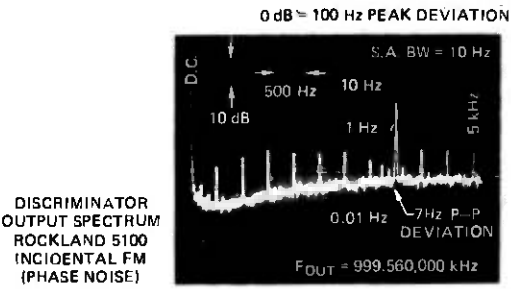
Two Rockland frequency synthesizers are used for the fourth BO, one in the main and one in the scanning receiver (see Section 5.5), to down-convert from 2 MHz to 123 kHz. In addition, the digitally controlled synthesizer provides the incrementally adjusted (100-Hz steps) frequency control needed for the scanning receiver; the second synthesizer provides the fine resolution (12.5 Hz) steps needed in the AFC loop of the tracking receiver. These synthesizers were selected because of two features which make them particularly attractive—frequency changes are essentially instantaneous, and the phase is continuous. This is ac-

* The two signals are orthogonally polarized and are nominally called horizontal and vertical polarizations but are not actually H and V polarized at the receiving sites.

† For the 19-GHz receiver they are: 1 GHz, 20 MHz, 2.0 MHz, 123 kHz, and 10.5 kHz. For the 28.5-GHz receiver the frequencies are scaled in a ratio of 3:2.



(a)



(b)

Fig. 4—(a) Discriminator output with synthesizer set to 1.000,000,000 MHz. (b) Discriminator output with synthesizer set to 0.999,560,000 MHz.

complished by using the phase as the driving variable within the synthesizer. A particular signal frequency output is achieved by controlling the rate of change of phase. Thus when a frequency change is called for, the rate at which the output signal phase accumulates is either increased, for a higher output frequency, or decreased for a lower output frequency. No discontinuities occur in the phase of the output signal, and thus there is no need to wait for filter transients to damp in the receiver whenever the synthesizer output frequency is changed. Without this capability, each change in frequency during scanning or tracking would necessitate delays for filter settling.

The synthesizer is driven from a source derived from the 10.000-MHz reference supplying the frequency multiplier discussed above. All BO frequencies in the receiver are derived from this reference, thereby eliminating many potential problems such as relative drift among the frequencies.

Although the digital frequency control made the system attractive, this approach also implies a certain amount of phase noise arising from the A/D quantization in the synthesizer. This caused some initial difficulties since the synthesizer output could not be used directly due to

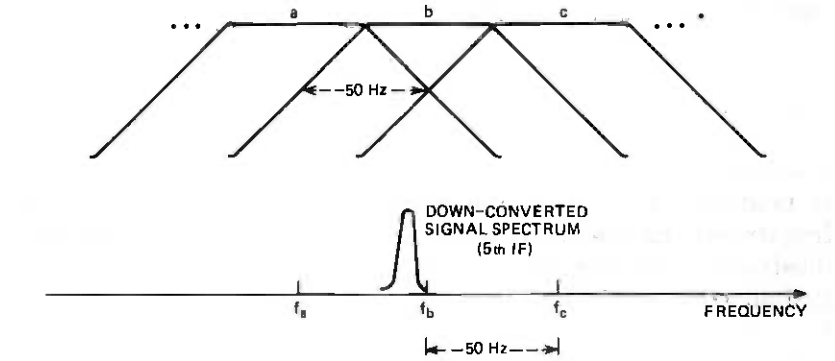
incidental FM (phase noise) it contained. For example, with the output set for 1.000,000 MHz the phase noise appears random and spectrally smooth, but with slight changes in output frequency a structured noise spectrum became apparent. For example, with an output frequency of 999.560,000 kHz, a spectral analysis revealed a number of tones (with a dominant noise tone at about 3.6 kHz offset), of sufficent amplitude to produce a 7-Hz peak-to-peak deviation (see Fig. 4). (For other frequencies, the noise tones would, of course, vary, but the problem is illustrated by this example.) To reduce this phase noise, the signal was first mixed up to 18.77 MHz (which maintains the same level of phase noise) and then divided by 10—which reduces the phase noise by 20 dB. This ensured that the residual incidental FM would be of no consequence in beacon reception.

5.4 Comb filters

In order to reduce the time needed to acquire the beacon signals, to provide a wide frequency range over which accurate signal level determination can be made (while providing narrowband processing for noise limiting), parallel signal processing was incorporated. The first stage of that parallel processing is a set of 32 high-resolution, closely matched filters of 100 Hz bandwidth and spaced by 50 Hz. These comb filters span the range from 9.7 kHz to 11.3 kHz. The output signal from each filter is detected and the frequency synthesizer is incremented to continually drive the frequency of the highest detected signal level (assumed to be the beacon) to the center of this 1600-Hz band. As the beacon frequency drifts, the detected signal level in the center filter tends to fall as the level of an adjacent filter increases. Figure 5 illustrates the condition: output $b > a > c$, which implies that the signal is within the region $f_b - 25$ Hz to f_b. If we then note the magnitude of the difference between the output of b and the output of a we can further subdivide the region. If $b - a >$ 1 dB then the signal is located within 12½ Hz of the center of filter b. Thus 12½ Hz resolution can be obtained; when the signal drifts further than 12½ Hz from the comb center, the BO is readjusted to drive it back. Parallel processing ensures that if the beacon signal fades below a detectable level (about 38 dB fade, remaining faded for an extended period), upon its recovery to a detectable level, if it lies anywhere in the band covered by the comb set, it will be detected virtually instantaneously. The actual detection process associated with the comb filter outputs is discussed in Section 5.7.

5.5 Scanning receiver

The scanning receiver is essentially a scanning spectrum analyzer. Its input is the 2.0-MHz third IF; this signal is twice down-converted to drive



*SEE FIG. 6 FOR MEASURED FILTER CHARACTERISTIC

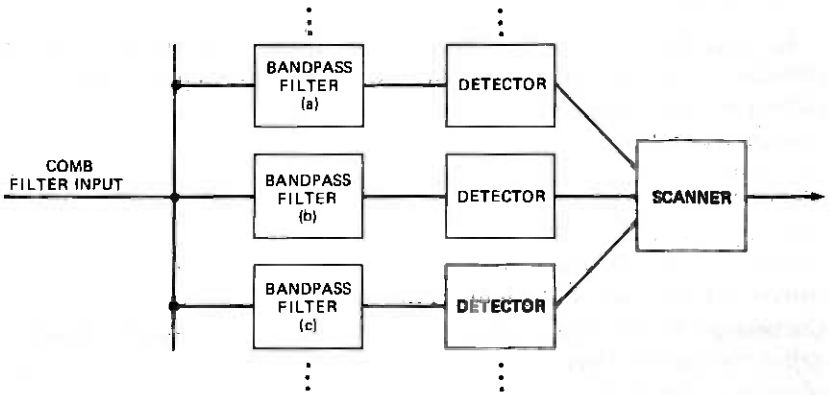


Fig. 5—19-GHz receiver comb filter characteristic.

two sets of comb filters. The two sets of filters have their outputs processed alternately, thus doubling the allowed settling time per filter while maintaining a rapid scan time (see Section 5.6). Each comb set consists of three 100-Hz filters separated by 1 kHz. The two comb sets are offset from each other by 50 Hz. The filter outputs are linearly detected to estimate power in the band and the entire frequency band of interest is scanned by causing the frequency synthesizer to shift the fourth BO. As the filter outputs are measured, the controller portion of the receiver stores observations, correlating the fourth BO frequency with the power received. At the completion of a scan, the location of the highest detected value is assumed to be the location of the carrier; however, the processor also checks to ensure that sidebands, down 4 dB from the carrier, are found in frequency slots 1 kHz above and below the carrier. If the levels are incorrect by more than 1 dB relative to the carrier, the data is not

considered valid. When a valid signal is found, i.e., a carrier of sufficient level to be a candidate and having sidebands appropriately located and energized, the frequency location is stored. This frequency is compared with that of the main receiver and the offset, if large enough, is used for correction of the fourth BO.

Although continuous scanning is desired, it is imperative that spurious responses not input false information to the basic receiver. Thus, update is inhibited without the presence of both carrier and two sidebands; additionally, the absolute signal-to-noise level must be high enough to ensure valid data. If these conditions are satisfied and the frequency difference between the scanning receiver and the main receiver is 187.5 Hz or greater, the scanning receiver output will correct the main receiver tuning. The intent is to ensure that the received signal is maintained as close as possible to the center of the comb filter set to provide maximum accommodation to drift. The 187.5-Hz threshold is somewhat arbitrary but derives from the smallest frequency increment (12.5 Hz) multiplied by 15—which is the maximum count of a four-stage counter.

Under normal operating conditions the scanning receiver continues to operate, but its output is nonfunctional. It *is* functional at initial acquisition, and because the site is remote and unmanned, it must also function in reacquisition should tracking be interrupted for two hours or more.

5.6 Equipment design considerations

Since reliable operation was desired even under worst case conditions, the maximum anticipated frequency drift range of the beacon, ±2 ppm for aging and diurnal variation, was used in determining the maximum range, the design goal was to locate the signal to within 25 Hz, using the commodate any unforeseen variations* which might cause the beacon frequency to drift beyond the receiver window. In spite of this large scan range, the design goal was to locate the signal to within 25 Hz, using the scanning receiver, and to accomplish this in a maximum of 15 seconds. The need to make the receiver relatively inexpensive and to accomplish the realization in a short time frame implied that the filters would have to be LC rather than crystal. Selection of the final IF was based on several considerations. The bandwidth desired was 100 Hz to match the beacon specification of 90 percent of the 19-GHz energy contained within a 100-Hz band (150 Hz for the 29-GHz signal). The achievable filter Q was limited by the coils used. Below about 10 kHz the coil Q fell off faster than frequency, thus setting 10 kHz as the lowest operating frequency. Higher frequency operation would be threatened by proportionately

* This includes ±1 ppm for beacon receiver aging.

greater temperature drift problems. Thus, the best engineering tradeoff between operating frequency and component availability, stability and tolerance suggested a final IF of about 10 kHz.

Conceptually, a single 25-Hz filter/detector system could then be used to find the signal (carrier and sidebands) with the processed results being stored. However, this would result in a long scanning time to allow sufficient filter settling time (60–100 msec per step) and a resultant unacceptable scan time of 3–5 minutes. As an alternative, two sets of filters of 100-Hz bandwidth were constructed. Each set was constructed to simultaneously monitor three slots separated by 1 kHz. This allows immediate recognition of the carrier and its two sidebands. The second set of filters is offset from the first by 50 Hz, permitting 25-Hz frequency accuracy and a full scan in only 15 seconds. Frequency interpolation is possible because the filter responses are well controlled and relative power levels are determined by signal frequency. As can be seen in Fig. 6, if $b > c > a$, where a, b, c represent the powers in the respective frequency bands, then the beacon signal must lie between 10.5 kHz and 10.525 kHz. The speedup in scan time (relative to the conceptual 25-Hz filter approach) is due to increasing filter bandwidth from 25 to 100 Hz.

Two critical functions in the scanning receiver are linear detection and sample-and-hold. The detectors, well matched and linear to within 0.1 dB over a 60-dB range, are built around the LM318N op amp with 458-type diodes in the feedback path. The sample-and-hold is especially critical since the peak detected sample must be stored for a full scan of 15 seconds. Stable storage with time is most critical when sampling the spectrum in the vicinity of the carrier and its sidebands, i.e., the carrier, the first and third harmonics (a total of 6 kHz) sliding through a 2-kHz window for a total of 8 kHz. With 1/64 sec allotted per step,* this implies a critical storage time of about (8 kHz/100 Hz/step)($\frac{1}{64}$ sec/step) = 1.25 seconds. The critical limiting parameter in the receiver is thermal noise. Ensuring that all other degradations (such as signal droop in the sample-and-hold circuit) contribute small degradation relative to noise requires that droop be limited to no more than 0.05 dB in 1.25 seconds. This corresponds to a maximum droop of no more than 0.6 dB over the 15-second hold time. The current drive capabilities of the op amp, coupled with the slew rate requirements of the sample-and-hold limit the size of the capacitor to about 0.02 μF. For this size capacitor, a 0.05 dB droop in 1.25 seconds implies a maximum leakage of 1/4 nA. Choice of a polystyrene capacitor for storage ensured that the primary source of this leakage would be the reverse bias of the detector diodes. Back bias on most diodes would result in several nanoamps leakage—much too

* 15 sec scan time/(960 kHz band/100 Hz steps) = $\frac{1}{64}$ sec/step.

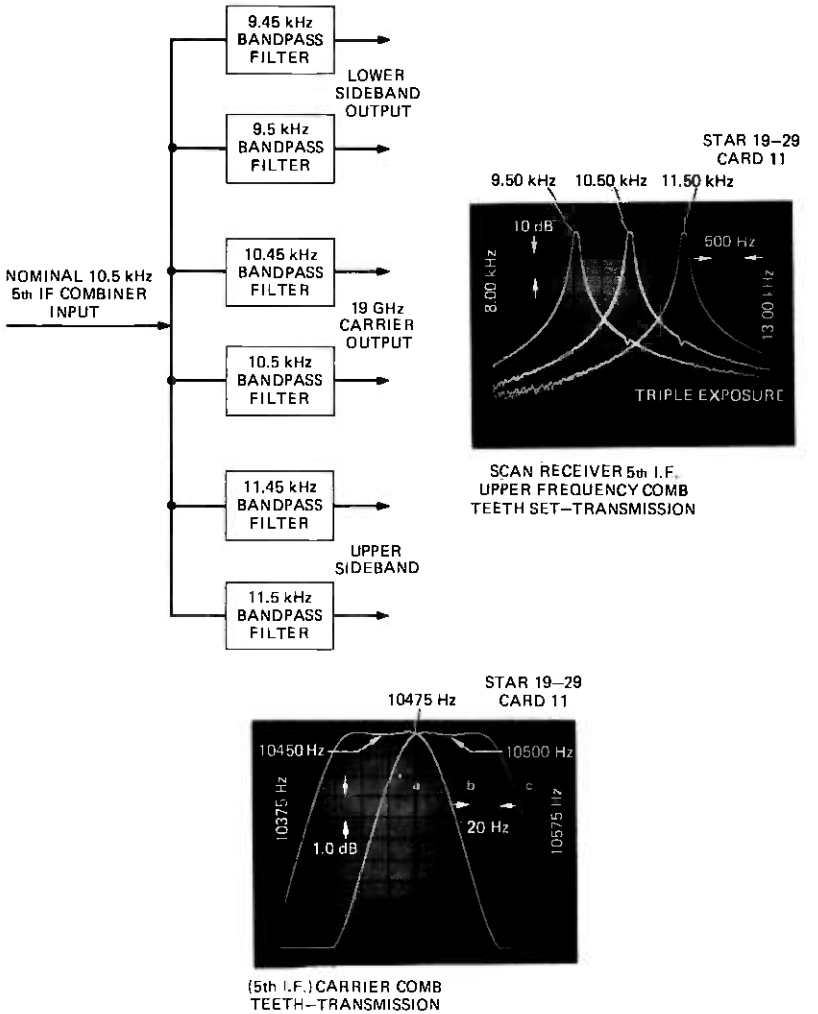


Fig. 6—Scanning receiver comb filter characteristics.

large. However, by introducing circuitry (see Fig. 7) to ensure that the bias level across the diodes is virtually zero, i.e., less than 10 mV, the leakage current reduces to less than 1 pA. In a similar fashion, another diode associated with the hold circuit was compensated. This left, as the only current of consequence, the input bias current of the isolation amplifier of the hold circuit. Using a LH0022C, which limits maximum bias current to 50 pa, ensured that droop would not be a problem.

5.7 19-GHz receiver

The 19-GHz receiver shown in Fig. 8 tracks the beacon signal, detects

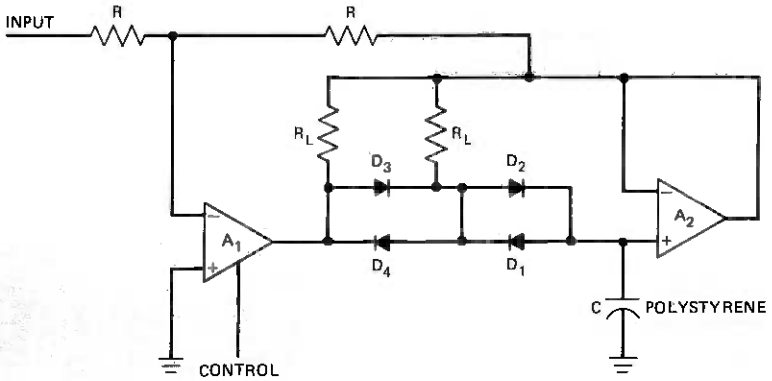


Fig. 7—Leakage compensation in sample-and-hold circuit. The feedback path around A_2 ensures that the drop across D_1 and D_2 is virtually zero, and thus there is no current leakage from C. Resistors R_L are large and serve to provide a path for residual diode leakage current.

its level, and provides the 3/2 frequency-scaled BO for the down-conversion of the 29-GHz signal.

In the interest of keeping waveguide loss to a minimum, the first down-conversion to 1 GHz and amplification of the beacon signal (by about 24 dB) takes place on the polarization adapter assembly which is mounted on the antenna feedhorn. See Fig. 9.* The Space Kom down-converter is a Schottky-diode balanced mixer which, combined with a low noise preamp, yields a combined noise figure of 6.5 dB (due primarily to the mixer). The source of the 18-GHz BO is the frequency multiplier described earlier. The resultant 1-GHz first IF signal, after bandpass filtering to 12 MHz to limit noise and reject images, is then processed by the main 19-GHz receiver. As indicated in Fig. 8 the second down-conversion, to 20 MHz, is accomplished using a double balanced mixer. This mixer requires matched impedance at each port which is constant at 20 MHz as well as 1 GHz. This was realized by following the mixer with a 2-pole Butterworth (constant resistance) LPF with a 100-MHz BW and then 15 dB of gain. To limit unwanted signals, a 2-pole Chebyshev image rejection BP filter with 46 dB of rejection is used prior to the next down-conversion.

After the third down-conversion the signal is split and used to drive the 19-GHz receiver as well as the scanning receiver discussed earlier. The signal has now been amplified from a nominal −110 dBm to −53 dBm. After the fourth conversion to 123 kHz, a 4.0-kHz BW, BP filter (Chebyshev, 0.01 dB ripple) is used to further restrict the bandwidth. Note that the source of the fourth BO used in this down-conversion is

* Note that the RF front end and temperature-controlled reference noise source for the 13-GHz radiometer are also mounted on the adapter assembly.

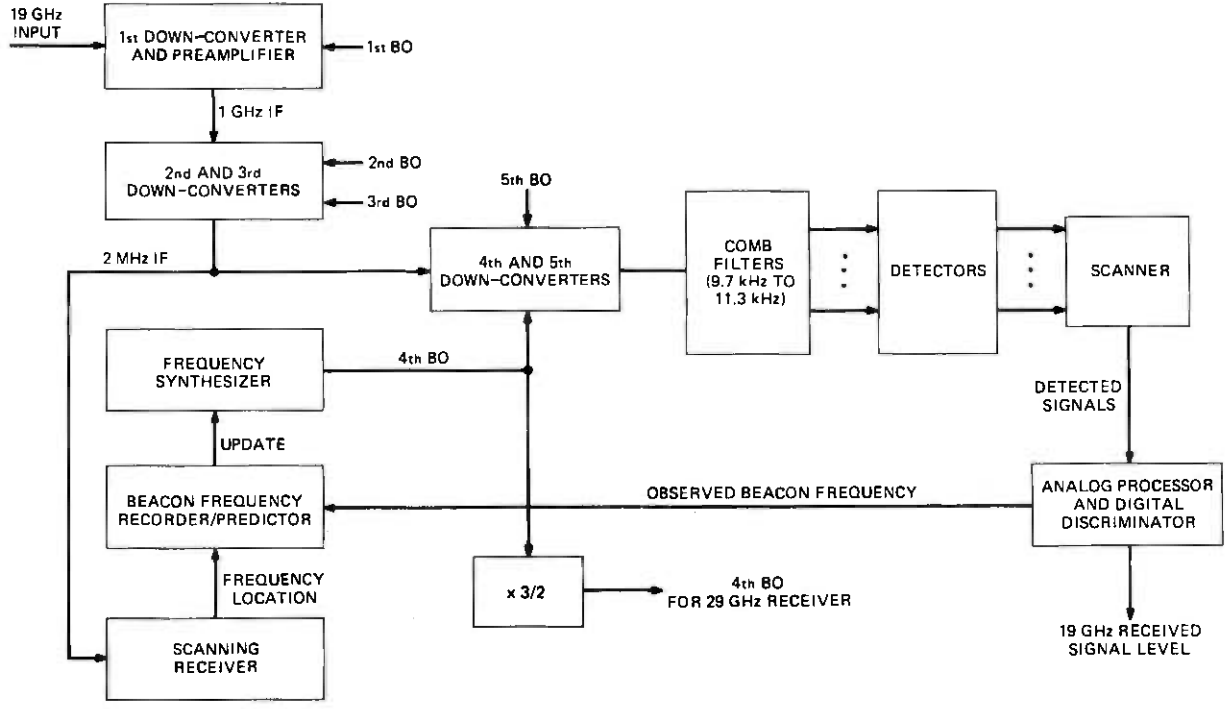


Fig. 8—19-GHz beacon receiver.

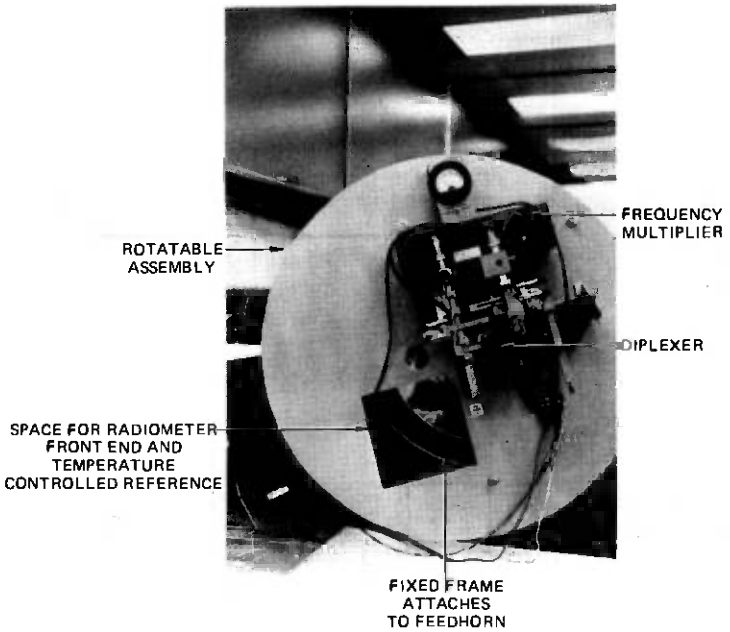


Fig. 9—Polarization adapter assembly.

the digitally controlled frequency synthesizer, which is tuned in 12.5-Hz increments to continually center the received signal at 123 kHz. After down-conversion to 10 kHz the output is passed through a 25-kHz one-pole active Chebyshev low-pass filter. At this point the signal level has been boosted to 14V P-P, and it is this signal which drives the comb filter detectors.

The five down-conversion steps were made along the way to allow progressive tightening of the noise bandwidth without requiring excessively high Q. By gradually narrowing the noise band, the filters can be designed with readily achievable Q using standard L/C technology, thus precluding the need for costly, specially designed high-Q crystal filters.

The actual detection process is illustrated in Fig. 10. The fifth IF signal is simultaneously fed to a set of 32 filters of 100-Hz bandwidth on 50-Hz centers spanning a frequency range from 9.7 kHz to 11.3 kHz. Each filter is a 2-pole Chebyshev with a 0.1-dB bandwidth of 50 Hz. Because of the well-controlled filter shape, it is possible to make an accurate determination of signal location from the relative levels in several filters. Each detector output (one detector/filter) is post-detection filtered to 1-Hz bandwidth to further reduce noise. Samples of the post-detection filter outputs are multiplexed together and scanned by a peak detector. A

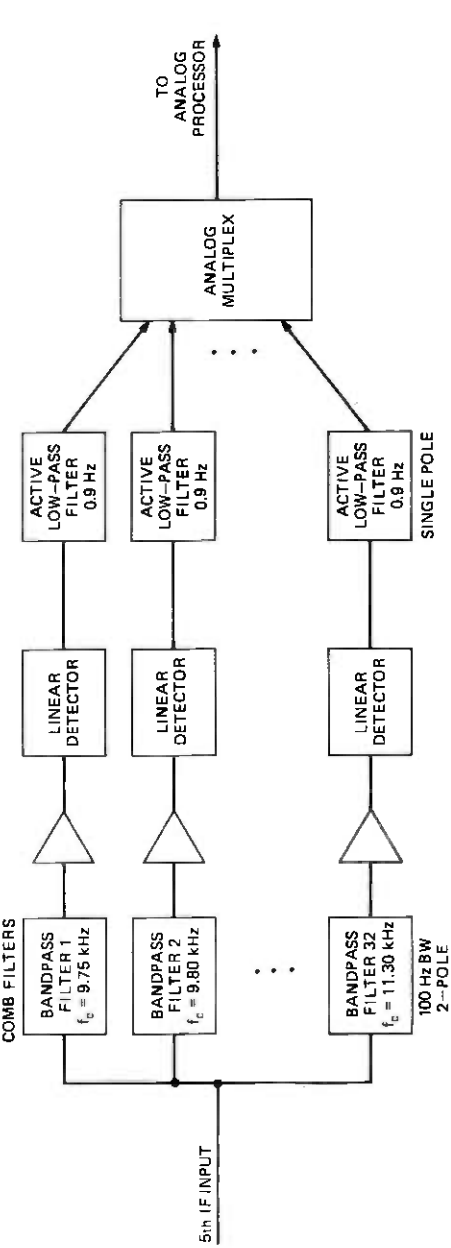


Fig. 10—Parallel processing/detection.

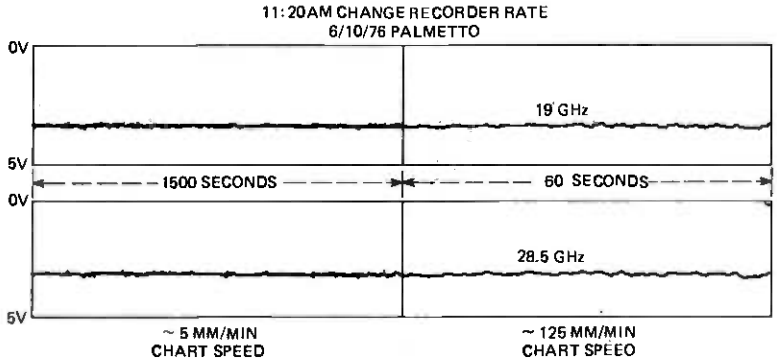


Fig. 11—Periodic amplitude variations. Amplitude scales are linear in voltage; thus the 19-GHz periodic ripple is about 0.2 dB in magnitude. The periodicity was not obvious until the chart speed was increased. At 5 mm/min the variation looked random; however, at 125 mm/min the periodicity becomes apparent.

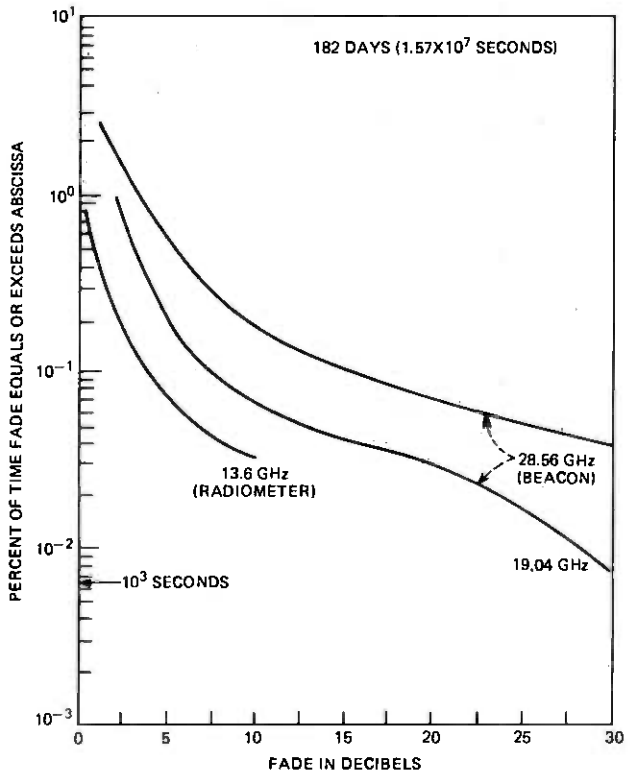


Fig. 12—Grant Park fading.

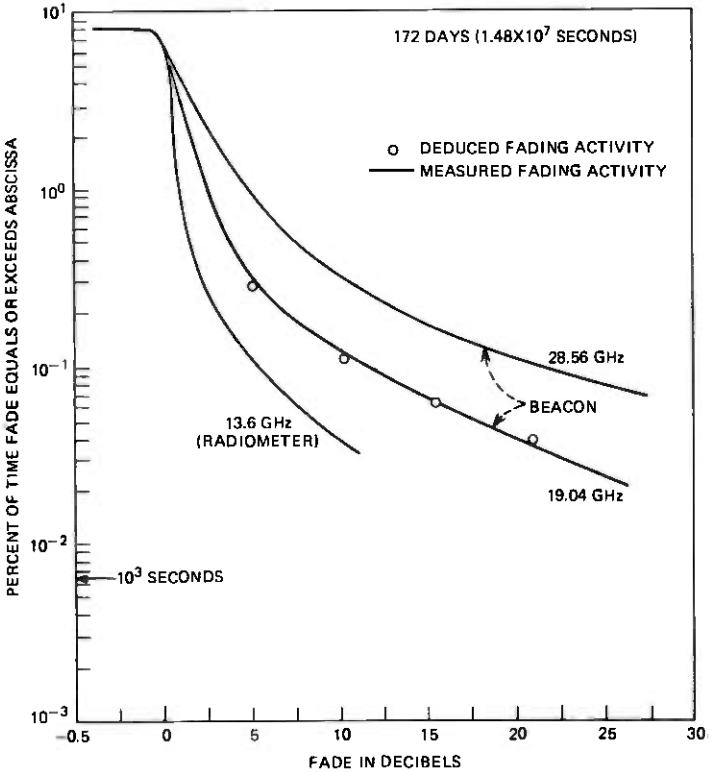


Fig. 13—Fade distribution in Palmetto, Georgia.

sample-and-hold following the peak detector always acquires the most recent sample of the received signal and is updated at each scan of the comb filter outputs. Also, by noting the location of the maximum signal within the comb set, together with the setting of the synthesizer, the beacon frequency is determined. Note that the comb filters have a well-controlled 50-Hz flat bandpass spaced on 50-Hz centers with the center frequency controlled to ±1 Hz, a 75-Hz 1-dB bandwidth, and a 100-Hz 3-dB bandwidth. As discussed in the section on comb filters, it is possible to get 12½-Hz resolution of the beacon frequency and thus to increment the frequency synthesizer so as to continually drive the beacon signal to the center comb position.

5.8 28.5-GHz receiver

Since the 28.5-GHz transmitter is scaled in frequency by 3/2 from the 19-GHz transmitter, the 28.5-GHz receiver local oscillator frequencies are also derived directly from those of the 19-GHz receiver. In this way a single AFC loop is sufficient to center both received signals in their

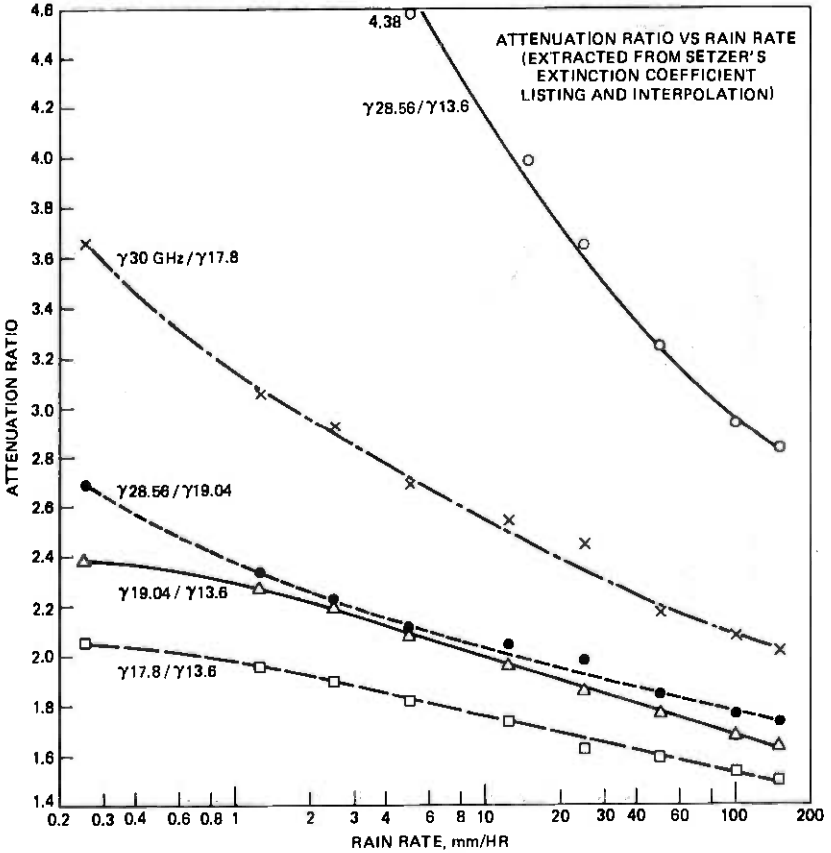


Fig. 14—Attenuation ratios vs. rain rate.

respective filters. Each step in the 28.5-GHz receiver processing is analogous with the 19-GHz receiver except that all frequencies are scaled up.

The signal is down-converted in five steps, using double balanced mixers and filtering to reject images generated in the process. The final processing is done at 12 kHz with a 150-Hz filter. This output is passed through a linear detector followed by a ½-Hz LPF to average out the noise. Note that since the 19-GHz signal will undergo less severe fading than the 28.5-GHz signal, the 19-GHz signal will always recover first and thus provide valid frequency information for detecting the 29-GHz signal before the 29-GHz signal actually recovers.

5.9 Frequency predictor

As indicated earlier, there will be occasional periods during which the

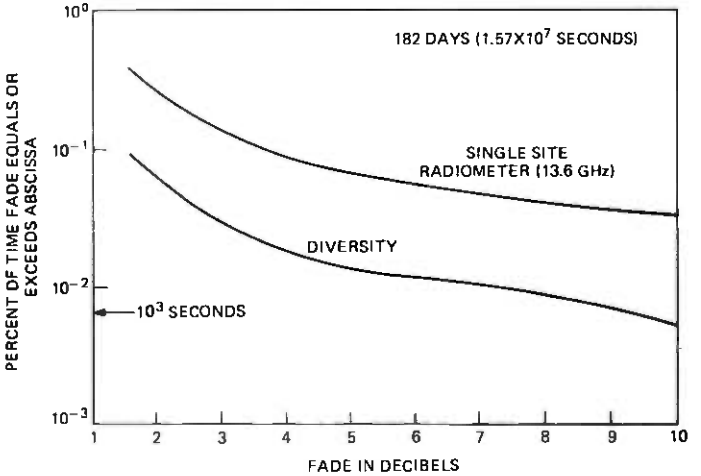


Fig. 15—Diversity results, Illinois.

beacon signal will be attenuated below the level of reliable reception. If these periods are relatively long, say 30 minutes or more, the frequency drift of the beacon may be sufficient to remove it from the range of the receiver. To avoid this potential problem, a frequency drift predictor was incorporated in the receiver. Noting the frequency of the beacon, as indicated by the settings of the synthesizer, and the location of the peak signal in the comb filter set, storing these indications each scan and averaging over a long time period (512 seconds), the average rate of change of frequency can be determined. This drift rate is continually updated as long as the received signal is strong enough for reliable tracking. Whenever the signal level falls below that level, presumably due to heavy rainfall, the drift rate update is stopped. Instead, the most recently calculated drift rate is used to increment the synthesizer and drive the local oscillators to the frequency anticipated by frequency extrapolation. This technique had been found to increase the tracking receiver's capability for immediate acquisition, extending that capability from 15 minutes to over 2 hours.

VI. EARLY RESULTS

Detailed reduction of the data is, of course, ongoing; there are, however, several items of interest already gleaned.

Immediately following installation of the beacon receiver in Palmetto, a very low level, approximately 0.25-dB peak-to-peak, periodic amplitude variation was detected in both the 19- and 28.5-GHz received signals; see Fig. 11. The frequency of this oscillation, ~54 times per minute, was found to coincide with a 0.04-degree precessing of the satellite arising

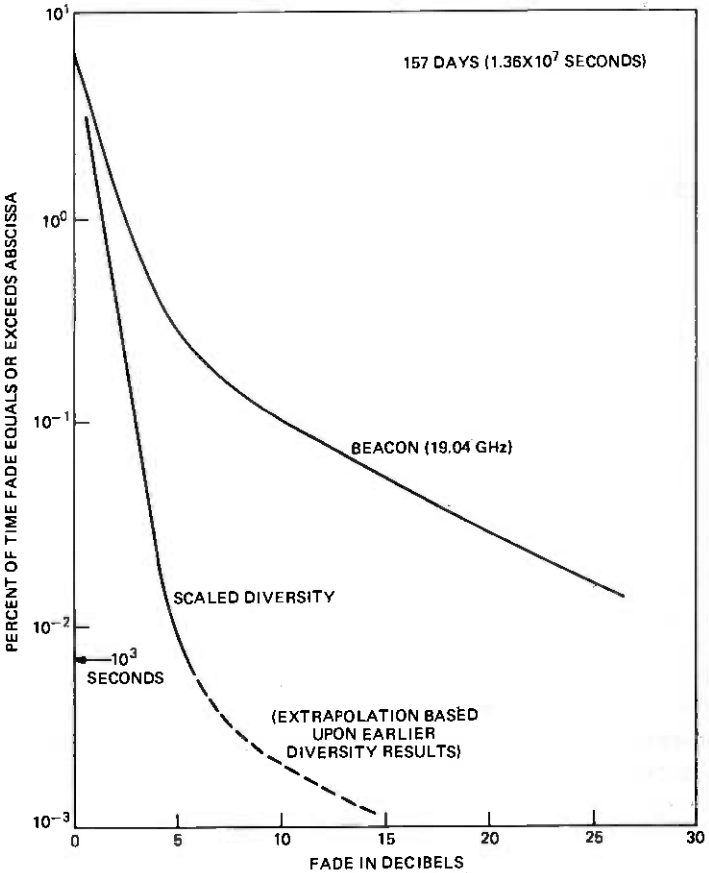


Fig. 16—Diversity results, Georgia.

from spin stabilization. This phenomenon was not detected at the other beacon reception sites in Grant Park or Crawford Hill due to the fact that Palmetto was closer to the edge of the satellite antenna pattern and thus experienced a greater amplitude change as the satellite pattern scanned.

The long-term attenuation (time faded below) of the beacons and of the 13-GHz radiometer, at both Grant Park and Palmetto, are shown in Figs. 12 and 13 respectively. Very good agreement among the three observations was demonstrated by selecting a particular level of occurrence, say 0.1 percent, and determining the attenuation ratio between two of the curves, e.g., the 28.5 GHz and the 13 GHz. From that ratio an equivalent rain rate can be deduced from Setzer's[6] work (attenuation ratios as a function of rain rate were derived from Setzer's results and are summarized in Fig. 14) and an estimate of the attenuation ratio be-

tween (say) the 19 GHz and 13 GHz can be obtained for the same rain rate. The X's in Fig. 13 correspond to 19-GHz attenuation deduced in the above manner. Similar results hold if the 13-GHz and 19-GHz signals are used to extrapolate to the 29-GHz attenuation curve.

The preliminary diversity results for Grant Park and Palmetto are given in Figs. 15 and 16. Note that the Grant Park results are biased by the fact that the only appreciable rainfall in the data base occurred during a brief period during the summer of 1976. During this one period of rain the site diversity improvement factor (the ratio of single site time faded below to diversity time faded below) was approximately 12, considerably less than the improvement factors of 50 we have seen in the past for equivalent fade levels.

VII. ACKNOWLEDGMENTS

The RF front end, as indicated, was the same as that specified by our coexperimenters at Bell Laboratories, Crawford Hill. In particular we benefited from interaction with D. C. Cox and H. W. Arnold. We are also indebted to E. A. Ohm, who designed the polarization coupler and P. Henry, who designed the frequency diplexer. Our ability to meet the design objectives and to be on-site when the satellite came on station is a tribute to the long hours of effort of all concerned. The work got off to a rapid start initially under W. T. Barnett, and the momentum was carried through with support from W. G. Ahlborn, H. J. Bergmann, J. Franzblau, L. J. Morris and E. E. Muller.

REFERENCES

1. R. W. Wilson, "A Three-Radiometer Path-Diversity Experiment," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1239–1242.
2. A. A. Penzias, "First Result From 15.3 GHz Earth-Space Propagation Study," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1242–1245.
3. H. J. Bergmann, "Satellite Site Diversity: Results of a Radiometer Experiment at 13 and 18 GHz," IEEE Trans. of AP-S, July 1977.
4. D. C. Cox, "An Overview of the Bell Laboratories 19- and 28-GHz COMSTAR Beacon Propagation Experiment," B.S.T.J., this issue, pp. 1231–1255.
5. H. J. Bergmann, "A New Tool for Gathering Statistics on Microwave Radio Fading," Bell Laboratories Record, *52* (October 1974), pp. 293–296.
6. D. E. Setzer, "Computed Transmission Through Rain at Microwave and Visible Frequencies," B.S.T.J., *49*, No. 8 (October 1970), pp. 1873–1892.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

COMSTAR Experiment:

Notes on the COMSTAR Beacon Experiment

By E. E. MULLER

(Manuscript received December 5, 1977)

Definitive empirical characterization of the transmission properties of the atmosphere has long been limited by the lack of appropriate sources radiating from beyond the atmosphere. The COMSTAR beacons provide appropriate radiation to interested experimentors throughout the continental United States.

Government, commercial, and scientific interest in transmission through the atmosphere at frequencies above 10 GHz derives from the potential for employing this portion of the spectrum for satellite communications and is basic to the attention being paid the COMSTAR beacon experiment, both within the U.S. and abroad. This opportunity for improving future generation communication satellites underlies A.T.&T.'s provision for carefully designed millimeter wave beacon sources in several earlier corporate proposals to the FCC.

Soon after the FCC granted permission to proceed with COMSTAR, specifications describing the beacon characteristics were published in technical journals along with an invitation to build and operate equipment for their reception. This announcement was received enthusiastically, and on October 1, 1975, a group of 40 interested experimenters gathered at a first "COMSTAR Experimenter's" meeting at Holmdel, New Jersey. Spacecraft development progress was discussed, along with early characterization results from prototype beacon equipments. Attendees signified their interest in participating in the experiment and outlined tentative plans. It was agreed that coexperimenters would meet from time to time, and that the data resulting from COMSTAR observations would be published in the open literature.

COMSTAR D1 has been in service now for almost two years. Beacon radiations have been observed at Bell Laboratories sites for this entire

period, beginning even before the satellite D1 came on station. Aspects of COMSTAR's behavior have been reported at professional meetings and in journals. Participation in the experiment is gratifying; at the present time observations are in progress at:

University of South Florida, Tampa, Fla.
Virginia Polytechnic Institute, Blacksburg, Va.
Johns Hopkins University, Laurel, Md.
Air Force Cambridge Research Labs, Hanson Field, Mass.
Institute for Telecommunications Science, Boulder, Colo.
COMSAT Laboratories, Clarksburg, Md.
General Telephone and Telegraph, Waltham, Mass.

In addition, several universities and government agencies have programs directed toward equipping facilities receiving the beacons at other locations.

The service lifetime of these beacons has three determinants: the lifetimes of critical devices, particularly the IMPATT amplifiers; the redundancy and fail-safe features provided in the system design; and the power budget of the spacecraft itself. These units have been designed to provide a minimum of two years of useful operation, as such a period is consistent with stable estimates for attenuation behavior. This goal would appear to have been realized in spacecrafts D1 and D2, although the 19-GHz radiation from satellite D1 has a predictable anomalous behavior during portions of the satellite thermal cycle. Long-term beacon availability depends, therefore, on continued access to surplus spacecraft power—a commodity which decreases with increased communications load and decreased solar cell efficiency. Given the staggered launch schedule of the COMSAT vehicles, the pertinent design parameters, and our present experience, it is expected that COMSTAR beacon signals should be available at least until the early 1980s.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty— V: The Discrete Case

By D. SLEPIAN

(Manuscript received July 20, 1977)

A discrete time series has associated with it an amplitude spectrum which is a periodic function of frequency. This paper investigates the extent to which a time series can be concentrated on a finite index set and also have its spectrum concentrated on a subinterval of the fundamental period of the spectrum. Key to the analysis are certain sequences, called discrete prolate spheroidal sequences, and certain functions of frequency called discrete prolate spheroidal functions. Their mathematical properties are investigated in great detail, and many applications to signal analysis are pointed out.

I. INTRODUCTION

In many branches of technology, such as sampled-data theory, time-series analysis, etc., doubly infinite sequences of complex numbers, $\{h_n\} = \ldots, h_{-1}, h_0, h_1, \ldots$ play an important role. Associated with such a sequence is its amplitude spectrum

$$H(f) \equiv \sum_{-\infty}^{\infty} h_n e^{2\pi i n f}. \tag{1}$$

In this paper we attempt to elucidate certain features of the complex relationship between $\{h_n\}$ and its amplitude spectrum $H(f)$.

Of prime importance in the analysis we present are some special sequences, here called discrete prolate spheroidal sequences (DPSS's), and some related special functions called discrete prolate spheroidal wave functions (DPSWF's). Much of the paper is devoted to a study of their mathematical properties. They are fundamental tools for understanding the extent to which sequences and their spectra can be simultaneously concentrated: they have many potential applications in communications technology.

We motivate our work by discussing a simple problem. But first some notation is needed. We adopt the abbreviation

$$E(n_1,n_2) \equiv \sum_{n=n_1}^{n_2} |h_n|^2 \tag{2}$$

and refer to this quantity as *the energy of the sequence* $\{h_n\}$ *in the index range* (n_1,n_2). Throughout this paper we restrict our attention to sequences whose *total energy* $E \equiv E(-\infty,\infty)$ is finite. Associated with a sequence is its *amplitude spectrum* defined in (1). It is periodic in f with period 1 and we shall generally consider it only for $|f| \leq 1/2$. From the theory of Fourier series, we then have the representation

$$h_n = \int_{-1/2}^{1/2} H(f)e^{-2\pi inf}df, \quad n = 0, \pm 1, \ldots \tag{3}$$

for the sequence, and from Parseval's theorem we have that

$$E = \sum_{-\infty}^{\infty} |h_n|^2 = \int_{1/2}^{1/2} |H(f)|^2 df. \tag{4}$$

If $H(f)$ is given, we say that the sequence $\{h_n\}$ defined by (3) is the sequence *belonging* to H and we write $\{h_n\} \leftrightarrow H(f)$.

Now let W be a positive number less than $1/2$. If the amplitude spectrum of $\{h_n\}$ vanishes for $W < |f| \leq 1/2$, we say that the sequence is *bandlimited* and that it has *bandwidth* W. The elements of a bandlimited sequence can be written in the form

$$h_n = \int_{-W}^{W} H(f)e^{-2\pi inf}df, \quad 0 < W < \frac{1}{2}, \quad n = 0, \pm 1, \ldots \tag{5}$$

Analogously, given two finite integers, $n_1 \leq n_2$, we shall say that a sequence $\{h_n\}$ is *indexlimited to the index interval* (n_1,n_2) if h_n vanishes whenever $n > n_2$ or $n < n_1$. It is not hard to see that, except for the trivial all-zero sequence, a bandlimited sequence cannot be indexlimited and that an indexlimited sequence cannot be bandlimited.

It is natural now to ask just how nearly indexlimited a bandlimited sequence can be. Specifically, we seek the maximum value of the *concentration*

$$\lambda \equiv \frac{E(N_0,N_0+N-1)}{E(-\infty,\infty)} = \left(\sum_{N_0}^{N_0+N-1} |h_n|^2\right) \Big/ \left(\sum_{-\infty}^{\infty} |h_n|^2\right) \tag{6}$$

for all sequences of bandwidth W, and ask for which bandlimited sequences the concentration attains this maximal value. The answers to these questions are simply stated in terms of the discrete prolate spheroidal wave functions $U_k(N,W;f)$, the discrete prolate spheroidal sequences $\{v_n^{(k)}(N,W)\}$, and their associated eigenvalues $\lambda_k(N,W)$, $k = 0,1,2, \ldots, N-1$. The bandlimited sequence $\{h_n\}$ of bandwidth W most concentrated in the sense of (6) is proportional to the DPSS $\{v_{n-N_0}^{(0)}(N,W)\}$,

its amplitude spectrum is proportional to $e^{i\pi(2N_0+N-1)f}U_0(N,W;f)$ in the interval $|f| \leq W$, and its concentration is given by $\lambda_0(N,W)$.

In earlier papers in this series[1–4] we treated the analogous problem of the maximal time-concentration of a *continuous* signal $f(t)$ of limited bandwidth. The optimal signals in that case, prolate spheroidal wave functions (PSWF's), were found to have many interesting and useful properties that were explored in related papers.[5–8] We here borrow freely from the techniques used in these earlier works and extend many of those results to the present case of discrete time series. Details of derivations that parallel closely ones to be found in Refs. 1–8 are sometimes omitted.

Part of the material presented here has been anticipated by others. As early as 1964 C. L. Mallows in an unpublished work defined versions of the DPSWF's and DPSS's. He showed that the former satisfy a second-order differential equation and that the latter satisfy a second-order difference equation, and described a number of their other properties as well. Tufts and Francis[16] in 1970 showed the importance of the DPSS's in the optimal design of digital filters. Independently, Papoulis and Bertran[12] in 1972 made a similar application. Eberhard[17] in 1973 showed that the DPSS provide optimal design of a discrete window for the calculation of power spectra under a natural criterion. All of these later authors present some numerical values of the functions and of $\lambda_0(N,W)$ for a few isolated values of N and W. None of them treat the subject as intensively as is done here. See Ref. 18 for some comments on these applications. An interesting application in optics to the theory of image formation was made by Gori and Guattari[23] in 1974.

Regarding the organization of this paper, in Section II we state without proof some of the more useful and interesting properties of the DPSWF's and the DPSS's. Included are curves and asymptotic formulae. In Section III we discuss some extremal properties and some applications of the functions. Of particular interest, perhaps, is the prediction problem of Section 3.2. In Section IV, proofs are given or outlined for the less obvious statements found in Sections II and III.

II. THE DPSWF's, THE DPSS's, AND SOME OF THEIR PROPERTIES

Throughout the remainder of this paper, unless otherwise explicitly stated, N is a positive integer and W a positive real number less than $1/2$.

2.1 The discrete prolate spheroidal wave functions

Since its kernel is degenerate, the integral equation

$$\int_{-W}^{W} \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} \psi(f')df' = \lambda\psi(f), \quad -\infty < f < \infty \tag{7}$$

has only N non-zero eigenvalues, $\lambda_0(N,W), \lambda_1(N,W), \ldots, \lambda_{N-1}(N,W)$. They are distinct, real and positive and we order them so that

$$\lambda_0(N,W) > \lambda_1(N,W) > \ldots > \lambda_{N-1}(N,W) > 0. \tag{8}$$

There are N linearly independent real eigenfunctions of (7) associated with these eigenvalues and we denote them by $U_0(N,W;f), U_1(N,W;f), \ldots, U_{N-1}(N,W;f)$. When these are normalized so that

$$\int_{-1/2}^{1/2} |U_k(N,W;f)|^2 df = 1,$$

$$U_k(N,W;0) \geq 0, \quad \frac{dU_k(N,W;0)}{df} \geq 0, \tag{9}$$

$$k = 0,1,\ldots,N-1,$$

they are the DPSWF's. Thus, *the discrete prolate spheroidal wave functions $U_k(N,W;f)$ and their associated eigenvalues $\lambda_k(N,W;f)$ are defined by*

$$\int_{-W}^{W} \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} U_k(N,W;f')df' = \lambda_k(N,W)U_k(N,W;f)$$

$$-\infty < f < \infty, \quad k = 0,1,\ldots,N-1, \tag{10}$$

along with (8) *and* (9) *and the requirement that the U_k be real.*

The DPSWF's are doubly orthogonal:

$$\int_{-W}^{W} U_i(N,W;f)U_j(N,W;f)df$$

$$= \lambda_i \int_{-1/2}^{1/2} U_i(N,W;f)U_j(N,W;f)df = \lambda_i \delta_{ij} \tag{11}$$

$$i,j = 0,1,\ldots,N-1.$$

For $k = 0,1,\ldots,N-1$, the function $U_k(N,W;f)$ is periodic in f. It has period 1 if N is odd and period 2 if N is even. In either case we have

$$U_k(N,W;f+1) = (-1)^{N-1}U_k(N,W;f), \tag{12}$$

while

$$U_k\left(N,W;\frac{1}{2}-f\right) = K(N,k)U_{N-1-k}\left(N,\frac{1}{2}-W;f\right)$$

$$\lambda_k\left(N,\frac{1}{2}-W\right) = 1 - \lambda_{N-1-k}(N,W) \tag{13}$$

$$K(N,k) = \begin{cases} (-1)^{(N-1)/2+k}, & N \text{ odd} \\ (-1)^{(N/2)-1}, & N \text{ even.} \end{cases}$$

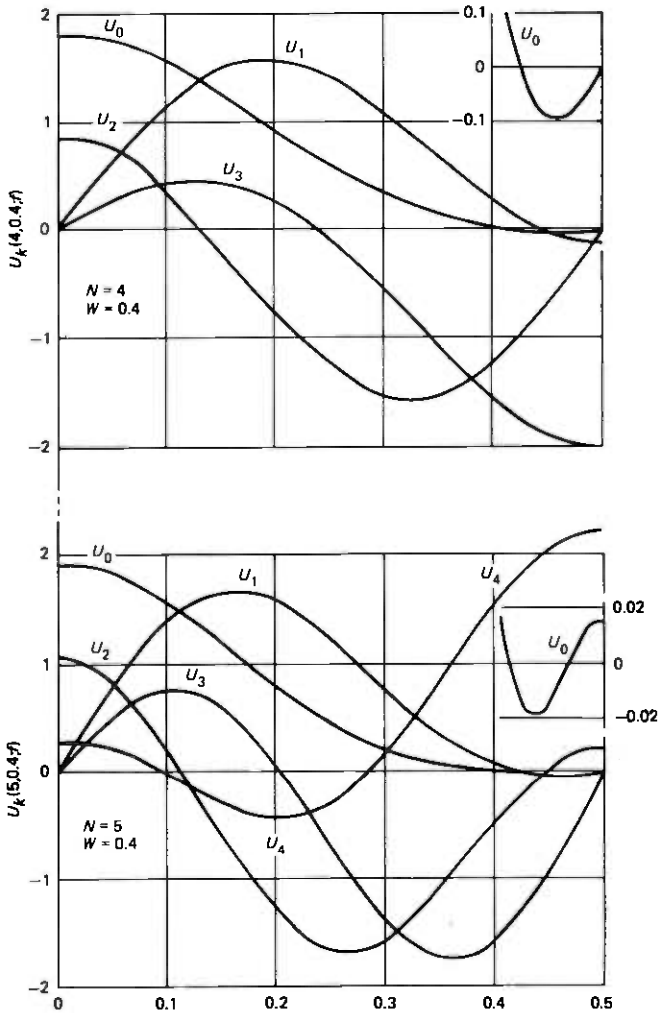


Fig. 1—$U_k(4,0.4;f)$ for $k = 0,1,2,3$ and $U_k(5,0.4;f)$ for $k = 0,1,2,3,4$ for $0 \leq f \leq 0.5$.

The DPSWF $U_k(N,W;f)$ has exactly k zeros in the open interval $-W < f < W$ and exactly $N - 1$ zeros in $-\frac{1}{2} < f \leq \frac{1}{2}$. It is an even or odd function of f according to the parity of k. Plots of some selected DPSWF's are given in Figures 1 and 2. Note the inserts with changed scales needed to show detail of U_0 for $0.4 \leq f \leq 0.5$ in Fig. 1 and for $U_4(5,0.2;f)$ in $0 \leq f \leq 0.1$ in Fig. 2. Values of some $\lambda_k(N,W)$ can be obtained from the ordinates of the curves of Figures 3, 4, 5 and 6 corresponding to integer abscissa values. Figure 7 shows the dependence of some $\lambda_k(N,W)$ on W.

Let $\sigma(N,W)$ denote the $N \times N$ tri-diagonal matrix whose element in the ith row and jth column is

$$\sigma(N,W)_{ij} = \begin{cases} \frac{1}{2}i(N-i), & j = i-1 \\ \left(\frac{N-1}{2} - i\right)^2 \cos 2\pi W, & j = i \\ \frac{1}{2}(i+1)(N-1-i), & j = i+1 \\ 0, & |j-i| > 1, \end{cases} \tag{14}$$

$$i,j = 0,1,\ldots,N-1.$$

The N eigenvalues of this matrix are real and distinct. We denote them by

$$\theta_0(N,W) > \theta_1(N,W) > \ldots > \theta_{N-1}(N,W). \tag{15}$$

Then the DPSWF's satisfy the differential equation

$$\frac{d}{d\omega}[\cos\omega - A]\frac{dU_k(N,W;f)}{d\omega} + \left[\frac{1}{4}(N^2-1)\cos\omega - \theta_k(N,W)\right] U_k(N,W;f) = 0 \tag{16}$$

where we write

$$\omega \equiv 2\pi f, \quad A \equiv \cos 2\pi W. \tag{17}$$

2.2 The discrete prolate spheroidal sequences

For each $k = 0,1,2,\ldots,N-1$, the DPSS $\{v_n^{(k)}(N,W)\}$ is defined as the real solution to the system of equations*

$$\sum_{m=0}^{N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)} v_m^{(k)}(N,W) = \lambda_k(N,W) v_n^{(k)}(N,W), \tag{18}$$

$$n = 0,\pm1,\pm2,\ldots$$

normalized so that

$$\sum_{j=0}^{N-1} v_j^{(k)}(N,W)^2 = 1, \tag{19}$$

$$\sum_{0}^{N-1} v_j^{(k)}(N,W) \geq 0, \quad \sum_{0}^{N-1} (N-1-2j) v_j^{(k)}(N,W) \geq 0. \tag{20}$$

* It is understood here, of course, that when $n = m$ the expression $[\sin 2\pi W(n-m)]/\pi(n-m)$ has the value $2W$.

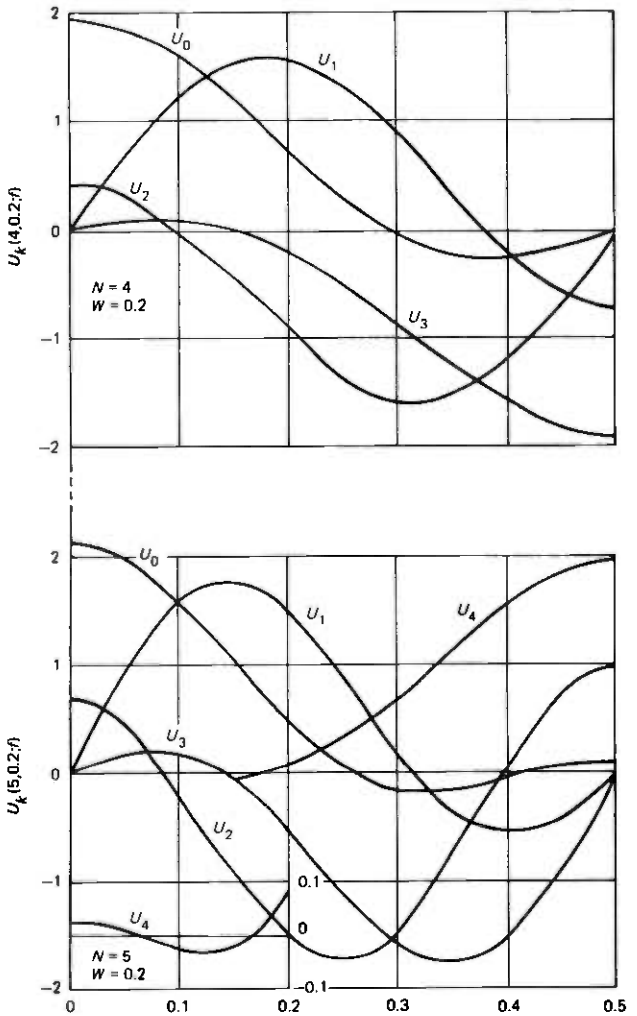


Fig. 2—$U_k(4,0.2;f)$ for $k = 0,1,2,3$ and $U_k(5,0.2;f)$ for $k = 0,1,2,3,4$ for $0 \le f \le 0.5$.

The $\lambda_k(N,W)$ here are, as before, the ordered non-zero eigenvalues of the integral equation (7). These quantities are thus seen to be also the eigenvalues of the $N \times N$ matrix $\rho(N,W)$ with elements

$$\rho(N,W)_{mn} = \frac{\sin 2\pi W(m-n)}{\pi(m-n)}, \quad m,n = 0,1,\ldots,N-1, \qquad (21)$$

and the $(N-1)$-vector obtained by indexlimiting the DPSS $\{v_n^{(k)}(N,W)\}$ to the index set $(0,N-1)$ is an eigenvector of $\rho(N,W)$.

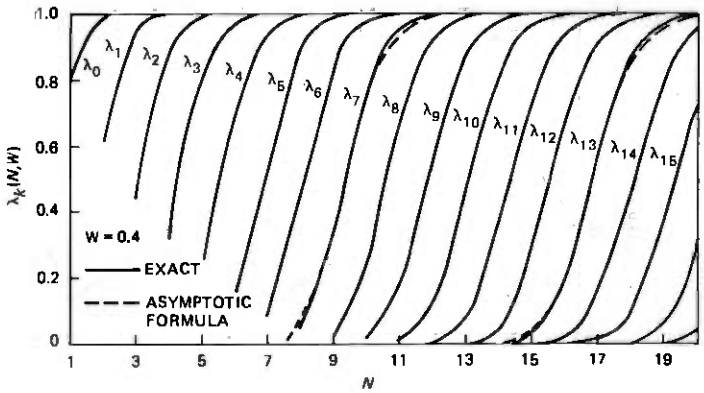


Fig. 3—Values of $\lambda_k(N,0.4)$ for $k = 0, \ldots, 15$ and $N = 0,1, \ldots, 20$.

The DPSS's are doubly orthogonal:

$$\sum_{n=0}^{N-1} v_n^{(i)}(N,W)v_n^{(j)}(N,W) = \lambda_i \sum_{-\infty}^{\infty} v_n^{(i)}(N,W)v_n^{(j)}(N,W) = \delta_{ij} \qquad (22)$$

$$i,j = 0,1, \ldots, N-1.$$

They obey the symmetry laws

$$v_n^{(k)}(N,W) = (-1)^k\, v_{N-1-n}^{(k)}(N,W) \qquad (23)$$

$$v_n^{(k)}(N,W) = (-1)^k v_{N-1-n}^{(N-1-k)}\,(N,\, 1/2 - W), \qquad (24)$$

$$n = 0,\pm 1,\pm 2, \ldots$$

$$k = 0,1, \ldots, N-1.$$

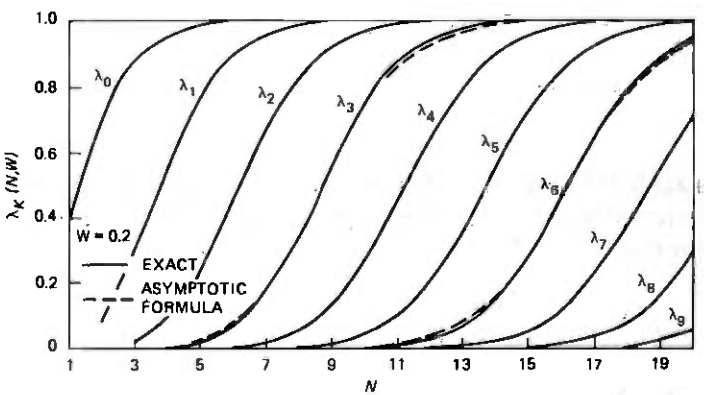


Fig. 4—Values of $\lambda_k(n,0.2)$ for $k = 0, \ldots, 9$ and $N = 0,1, \ldots 20$.

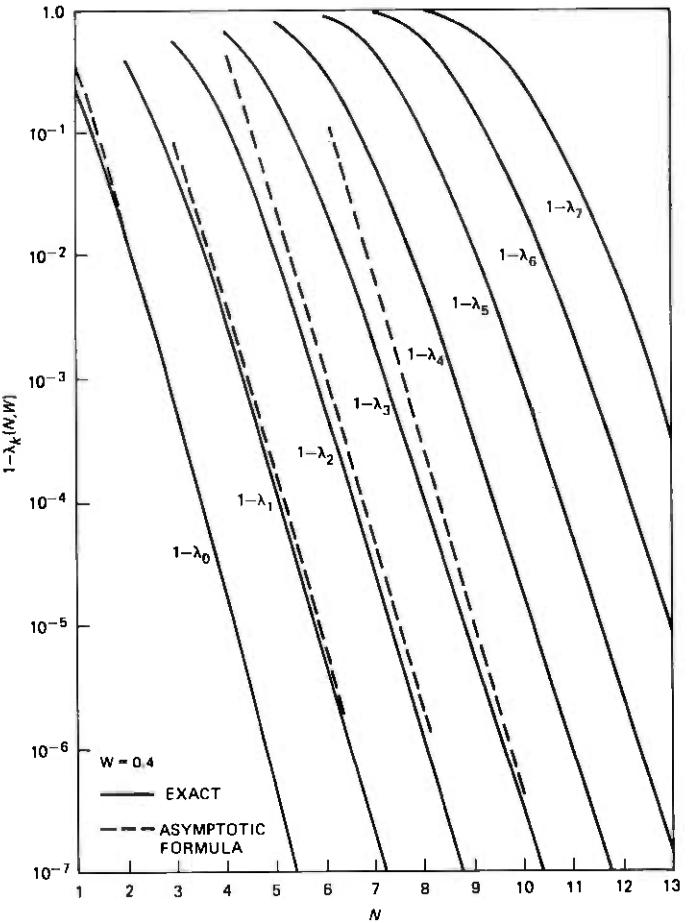


Fig. 5—$1 - \lambda_k(N,0.4)$ for $k = 0, \ldots, 7$ and $N = 1,2, \ldots, 13$.

The DPSS's indexlimited to $(0,N-1)$ satisfy the difference equation

$$\frac{1}{2}n(N-n)v_{n-1}^{(k)}(N,W) + \left[\cos 2\pi W\left(\frac{N-1}{2}-n\right)^2 - \theta_k(N,W)\right]v_n^{(k)}(N,W) + \frac{1}{2}(n+1)[N-1-n]v_{n+1}^{(k)}(N,W) = 0, \quad (25)$$

$$k,n = 0,1, \ldots, N-1.$$

Here the θ's are as in the differential equation (16).

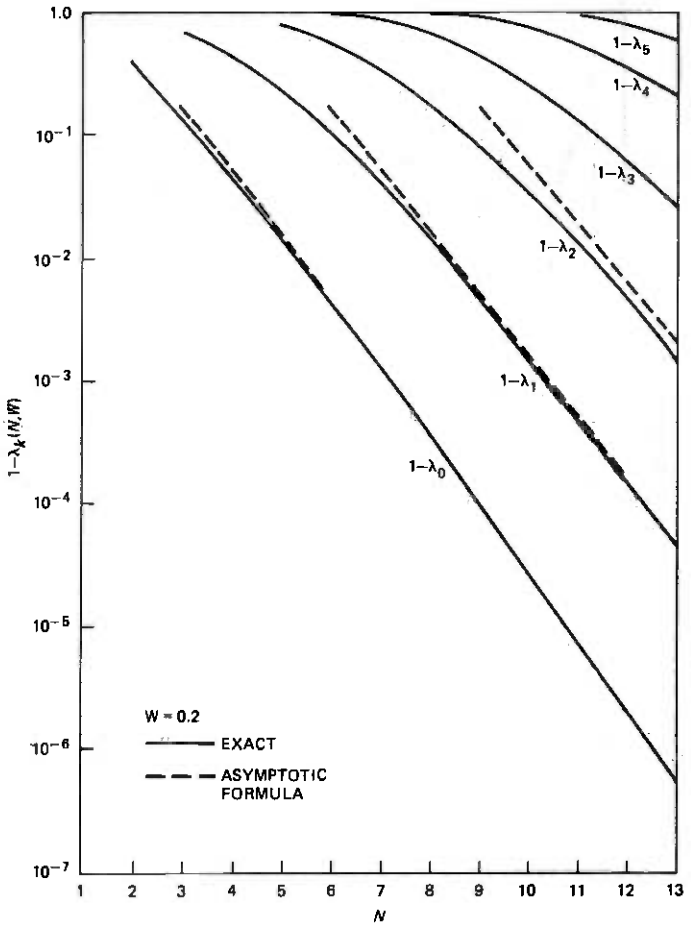


Fig. 6—$1 - \lambda_k(N,0.2)$ for $k = 0, \ldots, 5$ and $N = 1,2, \ldots, 13$.

2.3. Connections between the DPSWF's and the DPSS's

We have

$$U_k(N,W;f) = \epsilon_k \sum_{n=0}^{N-1} v_n^{(k)}(N,W) e^{-i\pi(N-1-2n)f} \qquad (26)$$

$$k = 0,1, \ldots, N-1,$$

where

$$\epsilon_k = \begin{cases} 1, & k \text{ even} \\ i, & k \text{ odd.} \end{cases} \qquad (27)$$

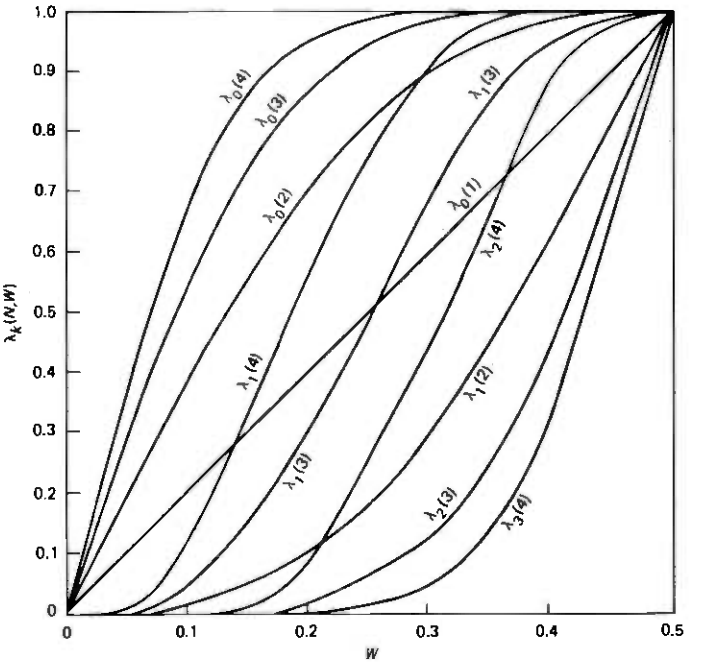


Fig. 7—$\lambda_k(N,W) \equiv \lambda_k(N)$ vs. W for several values of k and N.

Conversely,

$$v_n^{(k)}(N,W) = \frac{1}{\epsilon_k} \int_{-1/2}^{1/2} U_k(N,W;f) e^{i\pi(N-1-2n)f} df \tag{28}$$

$$n,k = 0,1,\ldots,N-1.$$

But, one also has

$$v_n^{(k)}(N,W) = \frac{1}{\epsilon_k \lambda_k(N,W)} \int_{-W}^{W} U_k(N,W;f) e^{i\pi(N-1-2n)f} df \tag{29}$$

$$k = 0,1,\ldots,N-1$$

for all values of n.

It is convenient now to introduce the *bandlimiting operator* B_W defined by

$$B_W H(f) = \begin{cases} H(f), & |f| \le W \\ 0, & |f| > W \end{cases} \tag{30}$$

and the *indexlimiting operator* $I_{N_1}^{N_2}$ defined by

$$I_{N_1}^{N_2}\{h_n\} = \{g_n\}$$

where

$$g_n = \begin{cases} h_n, & N_1 \le n \le N_2 \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

In terms of these operators, (28) and (29) can be stated

$$\epsilon_k I_0^{N-1}\{v_n^{(k)}(N,W)\} \leftrightarrow U_k(N,W;f)e^{i\pi(N-1)f} \tag{32}$$

$$\epsilon_k \lambda_k(N,W)\{v_n^{(k)}(N,W)\} \leftrightarrow B_W U_k(N,W;f)e^{i\pi(N-1)f}. \tag{33}$$

For the sequence $\{u_n^{(k)}(N,W)\}$ belonging to the DPSWF $U_k(N,W;f)$ we have

$$u_n^{(k)}(N,W) = \int_{-1/2}^{1/2} U_k(N,W;f)e^{-2\pi inf}df$$

$$= \epsilon_k \sum_{j=0}^{N-1} \frac{\sin \pi\left(\frac{N-1}{2} + n - j\right)}{\pi\left(\frac{N-1}{2} + n - j\right)} v_j^{(k)}(N,W)$$

$$n = 0, \pm 1, \pm 2, \ldots \tag{34}$$

When $N = 2M + 1$ is odd, this reduces simply to

$$u_n^{(k)}(2M+1,W) = \begin{cases} \epsilon_k v_{n+M}^{(k)}(2M+1,W), & |n| \le M \\ 0, & |n| > M, \end{cases} \tag{35}$$

a multiple of the indexlimited shifted DPSS. Equation (29) shows that conversely the spectrum of the shifted DPSS is in this case a multiple of the bandlimited DPSWF,

$$\epsilon_k \lambda_k(2M+1;W)\{v_{n+M}^{(k)}(2M+1,W)\} \leftrightarrow B_W U_k(2M+1,W;f). \tag{36}$$

2.4. Asymptotics of DPSWF's

In what follows, in addition to (17) we adopt the abbreviation

$$\alpha = 1 - A = 1 - \cos 2\pi W. \tag{37}$$

A. $U_k(N,W;f)$ for fixed k and large N

When N is large and W and k are fixed,

$$U_k(N,W;f) \sim \begin{cases} c_1 f_1(\omega), & 0 \le \omega \le N^{-1/3} \\ c_2 f_2(\omega), & N^{-1/3} \le \omega \le \arccos\left[A + N^{-3/2}\right] \\ c_3 f_3(\omega), & \arccos\left[A + N^{-3/2}\right] \le \omega \le 2\pi W \\ c_3 f_4(\omega), & 2\pi W \le \omega \le \arccos\left[A - N^{-3/2}\right] \\ c_5 f_5(\omega), & \arccos\left[A - N^{-3/2}\right] \le \omega \le \pi. \end{cases} \tag{38}$$

Here

$$f_1(\omega) = D_k\left(\left(\frac{N^2}{2\alpha}\right)^{1/4}\omega\right)$$

$$f_2(\omega) = \frac{[\sqrt{1+\cos\omega} + \sqrt{\cos\omega - A}]^N}{[(1-\cos\omega)(\cos\omega - A)]^{1/4}} \times \left[\frac{\sqrt{1-\cos\omega}}{\sqrt{\alpha(1+\cos\omega)} + \sqrt{2(\cos\omega - A)}}\right]^{k+1/2}$$

$$f_3(\omega) = I_0\left(\frac{N}{\sqrt{2-\alpha}}\sqrt{\cos\omega - A}\right) \tag{39}$$

$$f_4(\omega) = J_0\left(\frac{N}{\sqrt{2-\alpha}}\sqrt{A - \cos\omega}\right)$$

$$f_5(\omega) = \frac{\cos\left[\frac{N}{2}\arcsin\theta(\omega) + \frac{1}{2}\left(k+\frac{1}{2}\right)\arcsin\phi(\omega) + (k-N)\frac{\pi}{4} + \frac{3\pi}{8}\right]}{[(A-\cos\omega)(1-\cos\omega)]^{1/4}}$$

$$\theta(\omega) \equiv \frac{\alpha + 2\cos\omega}{2-\alpha}, \quad \phi(\omega) \equiv \frac{(2-3\alpha) - (2+\alpha)\cos\omega}{(2-\alpha)(1-\cos\omega)}$$

where $D_k(\cdot)$ is the Weber function (Ref. 9, Vol. II, Chapter VIII), and I_0 and J_0 are the usual Bessel functions. The constants in (38) are given by

$$c_i = (-1)^{[k/2]}(k!)^{-1/2}\pi^{Y_1}2^{Y_2}\alpha^{Y_3}N^{Y_4}[\sqrt{2}+\sqrt{\alpha}]^{Y_5}[2-\alpha]^{Y_6}$$

i	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	$\frac{1}{4}$	$\frac{1}{8}$	$-\frac{1}{8}$	$\frac{1}{4}$	0	0
2	$\frac{1}{4}$	$\frac{7}{4}k + \frac{7}{8}$	$\frac{k}{4} + \frac{3}{8}$	$\frac{k}{2} + \frac{1}{4}$	$-N$	0
3	$\frac{3}{4}$	$\frac{7}{4}k + \frac{11}{8}$	$\frac{k}{4} + \frac{1}{8}$	$\frac{k}{2} + \frac{3}{4}$	$-N$	$(N-k-1)/2$
5	$\frac{1}{4}$	$\frac{7}{4}k + \frac{15}{8}$	$\frac{k}{4} + \frac{3}{8}$	$\frac{k}{2} + \frac{1}{4}$	$-N$	$\frac{1}{4} + (N-k-1)/2.$

(40)

B. $U_k(N,W;f)$ for large N and $k = \lfloor 2WN(1-\epsilon)\rfloor$

When k and N become large together with

$$k = \lfloor 2WN(1-\epsilon) \rfloor, \quad 0 < \epsilon < 1 \tag{41}$$

and ϵ fixed, then

$$U_k(N,W;f) \sim \begin{cases} d_1 g_1(\omega), & 0 \le \omega \le \arccos(B + N^{-1/2}) \\ d_2 g_2(\omega), & \arccos(B + N^{-1/2}) \le \omega \le \arccos(B - N^{-1/2}) \\ d_3 g_3(\omega), & \arccos(B - N^{-1/2}) \le \omega \le \arccos(A + N^{-1}) \\ d_4 g_4(\omega), & \arccos(A + N^{-1}) \le \omega \le 2\pi W \\ d_4 g_5(\omega), & 2\pi W \le \omega \le \arccos(A - N^{-1}) \\ d_6 g_6(\omega), & \arccos(A - N^{-1}) \le \omega \le \pi. \end{cases} \tag{42}$$

Here B is determined so that

$$\int_B^1 \sqrt{\frac{\xi - B}{(\xi - A)(1 - \xi^2)}}\, d\xi = \frac{k}{N}\pi \tag{43}$$

and

$$g_1(\omega) = R(\omega) \cos\left[\frac{N}{2}\int_0^\omega \sqrt{\frac{\cos t - B}{\cos t - A}}\, dt - \frac{C}{4}\int_0^\omega \frac{dt}{\sqrt{(\cos t - B)(\cos t - A)}} - (1 - (-1)^k)\frac{\pi}{4}\right]$$

$$g_2(\omega) = Ai\left(-\frac{N^{2/3}(\cos\omega - B)}{[4(1-B^2)(B-A)]^{1/3}}\right)$$

$$g_3(\omega) = R(\omega) \exp\left[-\frac{N}{2}\int_{\arccos B}^\omega \sqrt{\frac{B - \cos t}{\cos t - A}}\, dt - \frac{C}{4}\int_{\arccos B}^\omega \frac{dt}{\sqrt{(B - \cos t)(\cos t - A)}}\right]$$

$$g_4(\omega) = I_0\left(N\sqrt{\frac{(B-A)}{(1-A^2)}}(\cos\omega - A)\right) \tag{44}$$

$$g_5(\omega) = J_0\left(N\sqrt{\frac{(B-A)}{(1-A^2)}}(A - \cos\omega)\right)$$

$$g_6(\omega) = R(\omega) \cos\left[\frac{N}{2}\int_\omega^\pi \sqrt{\frac{B - \cos t}{A - \cos t}}\, dt + \frac{C}{4}\int_\omega^\pi \frac{dt}{\sqrt{(B - \cos t)(A - \cos t)}} + \theta\right]$$

$$R(\omega) \equiv |(B - \cos\omega)(A - \cos\omega)|^{-1/4}.$$

The function $Ai(x)$ is the Airy function defined in Ref. 10, page 446. The parameter C here is given by

$$C = \frac{4}{L_2}\left[\frac{N}{2}L_1 + (2 + (-1)^k)\frac{\pi}{4}\right]_{\text{rem } 2\pi} \tag{45}$$

where $[x]_{\text{rem } 2\pi} = x - 2\pi\lfloor x/2\pi \rfloor$ is the number between zero and 2π congruent to x modulo 2π, and the parameter θ in g_6 is

$$\theta = \left[\frac{\pi}{4} - \frac{N}{2}L_5 - \frac{C}{4}L_6\right]_{\text{rem } 2\pi}. \tag{46}$$

The L's are given by

$$\begin{aligned} L_1 &= \int_B^1 P(\xi)d\xi \qquad L_2 = \int_B^1 Q(\xi)d\xi \\ L_3 &= \int_A^B P(\xi)d\xi \qquad L_4 = \int_A^B Q(\xi)d\xi \\ L_5 &= \int_{-1}^A P(\xi)d\xi \qquad L_6 = \int_{-1}^A Q(\xi)d\xi = L_2 \end{aligned} \tag{47}$$

$$P(\xi) \equiv \left|\frac{\xi - B}{(\xi - A)(1 - \xi^2)}\right|^{1/2}, \qquad Q(\xi) \equiv |(\xi - B)(\xi - A)(1 - \xi^2)|^{-1/2}.$$

The L's can be expressed simply in terms of complete elliptic integrals of the first and third kind. (See Ref. 13, pages 242 and 265.) (The integrals in (44) can also be expressed in terms of elliptic functions, but the resulting expressions shed no light on the nature of the solution.) Finally, the d's in (42) are

$$\begin{aligned} d_1 &= L_2^{-1/2}\pi^{1/2}2^{1/2} \\ d_2 &= L_2^{-1/2}\pi 2^{1/3}(1 - B^2)^{-1/12}(B - A)^{-1/3}N^{1/6} \\ d_3 &= L_2^{-1/2}\pi^{1/2}2^{-1/2} \\ d_4 &= L_2^{-1/2}\pi(1 - A^2)^{-1/4}e^{-CL_4/4}e^{-NL_3/2}N^{1/2} \\ d_6 &= L_2^{-1/2}\pi^{1/2}2^{1/2}e^{-CL_4/4}e^{-NL_3/2} \end{aligned} \tag{48}$$

C. $U_k(N,W;f)$ for large N and $k = \lfloor 2WN + (b/\pi)\log N \rfloor$

When $N \to \infty$ and

$$k = \lfloor 2WN + (b/\pi)\log N \rfloor \tag{49}$$

with b and W fixed, we have asymptotically in N

$$U_k(N,W;f) \sim \begin{cases} e_1 h_1(\omega), & 0 \le \omega \le \arccos\left[A + N^{-2/3}\right] \\ e_2 h_2(\omega), & |\cos\omega - A| \le N^{-2/3} \\ e_3 h_3(\omega), & \arccos\left[A - N^{-2/3}\right] \le \omega \le \pi. \end{cases} \tag{50}$$

Here

$$h_1^{\dagger}(\omega) = \frac{1}{\sqrt{\cos\omega - A}} \cos\left[\frac{N}{2}\omega + \frac{E\beta}{2}\log\left|\frac{\beta(1+A)\tan\frac{\omega}{2}+1}{\beta(1+A)\tan\frac{\omega}{2}-1}\right| - k\frac{\pi}{2}\right]$$

$$h_2(\omega) = e^{i(\beta/2)N(\cos\omega - A)}\Phi\left(\frac{1}{2} - i\frac{E\beta}{2},\quad 1;\ -i\beta N(\cos\omega - A)\right) \tag{51}$$

$$h_3(\omega) = \frac{1}{\sqrt{A - \cos\omega}} \times \cos\left[\frac{N}{2}\omega + \frac{E\beta}{2}\log\left|\frac{\beta(1+A)\tan\frac{\omega}{2}+1}{\beta(1+A)\tan\frac{\omega}{2}-1}\right| - (k+1)\frac{\pi}{2}\right]$$

$$\beta \equiv |\csc 2\pi W| \tag{52}$$

and

$$\Phi(a,c;x) = 1 + \frac{a}{c}\frac{x}{1!} + \frac{a(a+1)}{c(c+1)}\frac{x^2}{2!} + \ldots$$

is the confluent hypergeometric function in the notation of Ref. 9, Vol. 1, Chapter 6. The constant E is to be determined as the root of smallest absolute value of

$$N\pi W + \frac{E\beta}{2}\log\frac{2N}{\beta} + \psi(E\beta) - k\frac{\pi}{2} - \frac{\pi}{4} = 0. \tag{53}$$

Here we have written

$$\Gamma\left(\frac{1}{2} - \frac{1}{2}is\right) = r(s)e^{i\psi(s)} \tag{54}$$

where r, ψ, and s are real and Γ is the usual gamma function.

The constants in (50) are given by

$$\begin{aligned} e_1 &= (-1)^{\lfloor k/2 \rfloor}\left[\frac{3\pi}{\beta[1+e^{\pi E\beta}]\log N}\right]^{1/2} \\ e_2 &= r(E\beta)\sqrt{\beta N}e^{(\pi/4)E\beta}e_1 \\ e_3 &= e^{(\pi/2)E\beta}e_1. \end{aligned} \tag{55}$$

D. $U_k(N,W;f)$ for large N and $k = \lfloor 2WN(1+\epsilon)\rfloor$

The case of large N with $k = \lfloor 2WN(1+\epsilon)\rfloor$, $0 < \epsilon < 1/2W - 1$ can be

reduced to case B above by means of the formula (13). One finds

$$U_k(N,W;f) = K(N,k)U_{k'}\left(N,W';\frac{1}{2}-f\right) \tag{56}$$

where $U_{k'}(N,W';\frac{1}{2}-f)$ can be obtained from (41)–(48). Here

$$\begin{aligned} W' &= \frac{1}{2} - W \\ k' &= N - 1 - k \sim 2W'N(1-\epsilon') \\ \epsilon' &\sim \left(1 - \frac{1}{2W}\right)\epsilon. \end{aligned} \tag{57}$$

E. $U_{N-\ell}(N,W;f)$ for fixed ℓ and large N

Formula (13) reduces this case to case A above:

$$U_{N-\ell}(N,W;f) = K(N,N-\ell)U_{\ell-1}\left(N,\frac{1}{2}-W;\frac{1}{2}-f\right)$$

where formulas (38)–(40) can be used to obtain asymptotic values for $U_{\ell-1}(N,\frac{1}{2}-W;\frac{1}{2}-f)$.

2.5 Asymptotics of the eigenvalues $\lambda_k(N,W)$

For fixed k and large N, one has

$$1-\lambda_k(N,W) \sim \pi^{1/2}(k!)^{-1}2^{(14k+9)/4}\alpha^{(2k+1)/4}[2-\alpha]^{-(k+1/2)}N^{k+1/2}e^{-\gamma N}$$

$$\alpha = 1 - \cos 2\pi W, \quad \gamma = \log\left[1 + \frac{2\sqrt{\alpha}}{\sqrt{2}-\sqrt{\alpha}}\right]. \tag{58}$$

Some values computed from this expression are shown as dotted lines on Figs. 5 and 6. The fit with λ_0 is very good for $N \geq 2$ when $W = 0.4$ and for $N \geq 6$ when $W = 0.2$.

For large N and k with

$$\begin{aligned} k &= \lfloor 2WN(1-\epsilon) \rfloor, \qquad 0 < \epsilon < 1, \\ 1-\lambda_k(N,W) &\sim e^{-CL_4/2}e^{-L_3N}. \end{aligned} \tag{59}$$

Here the L's are given by (47) with B and C determined from (43) and (45).

For large N and k with

$$\begin{aligned} k &= \lfloor 2WN + (b/\pi)\log N \rfloor \\ \lambda_k(N,W) &\sim \frac{1}{1+e^{\pi b}}. \end{aligned} \tag{60}$$

A good approximation to $\lambda_k(N,W)$ when $0.2 < \lambda < 0.8$ is given by

$$\lambda_k(N,W) \sim [1+e^{\pi b}]^{-1} \tag{61}$$

where

$$\delta = -\frac{2\pi[NW - k/2 - 1/4]}{\log [8N|\sin 2\pi W|] + \gamma} \tag{62}$$

where $\gamma = 0.5772156649$ is the Euler-Mascheroni constant. Some values of (61)–(62) are shown on Fig. 3 for $k = 7$ and 13 and on Fig. 4 for $k = 3$ and 6. Near $\lambda = 1/2$ the discrepancy between the true value and the formula (61)–(62) cannot be seen on the scale of Figs. 3 and 4.

Asymptotic values for $\lambda_k(N,W)$ with N large and $N - k = \ell$ fixed can be obtained directly from (13) and (58). In a similar way, (13) and (59) provide an asymptotic formula for $\lambda_k(N,W)$ when $k = \lfloor 2WN(1+\epsilon)\rfloor$, $0 < \epsilon < 1/2W - 1$. One has in this case

$$\lambda_k(N,W) \sim e^{-CL_4/2} e^{-L_3 N} \tag{63}$$

where C, L_3, and L_4 are to be computed from (43), (45) and (47) with W replaced by $1/2 - W$ and k replaced by $N - k - 1$.

For fixed k and N, but W small, we find

$$\lambda_k(N,W) = \frac{1}{\pi}(2\pi W)^{2k+1} G(k,N)[1 + 0(W)] \tag{64}$$

where, for example

$$G(0,N) = N$$

$$G(1,N) = \frac{1}{36}(N-1)N(N+1)$$

$$G(2,N) = \frac{1}{8100}(N-2)(N-1)N(N+1)(N+2)$$

$$G(N-1,N) = \frac{2^{2N-2}}{(2N-1)\binom{2N-2}{N-1}^3}. \tag{65}$$

The general term is

$$G(k,N) = \frac{2^{2k}(k!)^6}{(2k+1)^2[(2k)!]^4} \prod_{j=-k}^{k} (N-j). \tag{66}$$

For fixed k and N, but W near $1/2$, i.e., $1/2 - W > 0$ small, (13) combined with (64) gives

$$1 - \lambda_k(N,W) = \frac{1}{\pi} G(N-1-k,N)[\pi(1-2W)]^{2(N-k)-1}[1 + 0(1-2W)]. \tag{67}$$

2.6 Relationship to PSWF's: $W \to 0$, $N \to \infty$, $\pi NW \to c > 0$

The prolate spheriodal wave functions (PSWF's) $\psi_i(c;x)$ and their associated eigenvalues $\lambda_i(c)$, $i = 0,1,2,\ldots$ are defined by

$$\int_{-1}^{1} \frac{\sin c(x-x')}{\pi(x-x')} \psi_i(c;x')dx' = \lambda_i(c)\psi_i(c;x), \tag{68}$$

$$-\infty < x < \infty$$

$$\lambda_0 > \lambda_1 > \lambda_2 \ldots$$

$$\int_{-\infty}^{\infty} \psi_i^2(c;x)dx = 1, \quad \psi_i(0) \geq 0, \quad \psi_i'(0) \geq 0 \tag{69}$$

$$i = 0,1,2,\ldots.$$

For each $i = 0,1,2,\ldots$ the PSWF $\psi_i(c;x)$ satisfies the differential equation

$$\frac{d}{dx}(1-x^2)\frac{d\psi_i}{dx} + [\chi - c^2x^2]\psi_i = 0 \tag{70}$$

for a special value

$$\chi = \chi_i(c) \tag{71}$$

of the parameter χ. The PSWF's and the quantities $\lambda_i(c)$ and $\chi_i(c)$ are discussed in detail in Refs. 1–6.

Now let $c > 0$ and y, a real number, be given. If, as

$$W \to 0, N = \left\lfloor \frac{c}{\pi W} \right\rfloor \quad \text{and } n = \left\lfloor \frac{N}{2}(1+y) \right\rfloor \tag{72}$$

then

$$\lambda_i(N,W) \sim \lambda_i(c)$$

$$\sqrt{W}\, U_i(N,W;Wf) \sim \psi_i(c,f) \tag{73}$$

$$\sqrt{\frac{N}{2}}\, v_n^{(i)}(N,W) \sim \frac{\pm 1}{\sqrt{\lambda_i(c)}} \psi_i(c;y) \tag{74}$$

$$N^2 - 1 - 2\theta_i(N,W) \sim \chi_i(c). \tag{75}$$

In (74) when i is even the plus sign is taken when $\int_{-1}^{1}\psi_i(c;x)dx > 0$; if i is odd, the plus sign is taken when $\int_{-1}^{1}x\psi_i(c,x)dx > 0$; otherwise, the negative sign is to be used.

III. APPLICATIONS

3.1 Extremal properties

3.1.1 Most concentrated bandlimited sequence

To maximize (6) over the bandlimited sequences, we replace h_n

there by its representation (5) and use (4) to obtain

$$\lambda = \frac{\sum_{n=N_0}^{N_0+N-1} \int_{-W}^{W} df \int_{-W}^{W} df' H(f)\overline{H}(f')e^{-2\pi in(f-f')}}{\int_{-W}^{W} |H(f)|^2 df}$$

$$= \frac{\int_{-W}^{W} df \int_{-W}^{W} df' e^{-i\pi(2N_0+N-1)(f-f')} \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} H(f)\overline{H}(f')}{\int_{-W}^{W} |H(f)|^2 df}$$

$$= \frac{\int_{-W}^{W} df \int_{-W}^{W} df' \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} \psi(f)\overline{\psi}(f')}{\int_{-W}^{W} |\psi(f)|^2 df}. \quad (76)$$

Here we have written

$$\psi(f) \equiv e^{-i\pi(2N_0+N-1)f} H(f) \quad (77)$$

and used the fact that

$$\sum_{N_0}^{N_0+N-1} e^{-2\pi in(f-f')} = e^{-i\pi(2N_0+N-1)(f-f')} \frac{\sin N\pi(f-f')}{\sin \pi(f-f')}.$$

A simple variational argument applied to (76) shows that λ is stationary when ψ satisfies (7) and hence the maximum value of λ is $\lambda_0(N,W)$, attained when $\psi(f) = cU_0(N,W;f)$, $|f| \leq W$. Equation (77) then shows that the most concentration bandlimited sequence is

$$\{h_n\} \leftrightarrow H(f) = \begin{cases} ce^{i\pi(2N_0+N-1)f} & U_0(N,W;f), \quad |f| \leq W \\ 0, & W < |f| \leq \tfrac{1}{2}. \end{cases} \quad (78)$$

For the sequence itself, we then find from (5)

$$h_n = c \int_{-W}^{W} U_0(N,W;f) e^{i\pi[N-1-2(n-N_0)f]} df$$

whence from (29)

$$\{h_n\} = d\{v^{(0)}_{n-N_0}(N,W)\} \quad (79)$$

where d is independent of n. The results (78) and (79) were stated in Section I after eq. (6).

More generally, for $k = 1,2,\ldots,N-1$ we have that $d\{v^{(k)}_{n-N_0}(N,W)\}$ is the bandlimited sequence most concentrated in (N_0, N_0+N-1) that is orthogonal to $\{v^{(i)}_{n-N_0}(N,W)\}$, $i = 0,1,\ldots,k-1$. The fraction of its energy in the range (N_0, N_0+N-1) is $\lambda_k(N,W)$.

3.1.2 Indexlimited sequence with most concentrated spectrum

If $\{h_n\}$ is indexlimited, so that

$$h_n = \begin{cases} 0, & n < N_0 \\ h_n, & N_0 \le n \le N_0 + N - 1 \\ 0, & n > N_0 + N - 1, \end{cases}$$

and if $H(f) \leftrightarrow \{h_n\}$, then

$$\mu \equiv \frac{\int_{-W}^{W} |H(f)|^2 df}{\int_{-1/2}^{1/2} |H(f)|^2 df} = \frac{\sum_{n=N_0}^{N_0+N-1} \sum_{m=N_0}^{N_0+N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)} h_n \bar{h}_m}{\sum_{N_0}^{N_0+N-1} |h_n|^2}$$

$$= \frac{\sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)} h_{n+N_0} \bar{h}_{m+N_0}}{\sum_{0}^{N-1} |h_{n+N_0}|^2}. \quad (80)$$

Here we have used (2) to replace $H(f)$ in the numerator, and have used (4) to rewrite the denominator. The quantity μ is a natural measure of the extent to which $H(f)$ is concentrated in the frequency interval $(-W, W)$. Comparison of the right member of (80) with (18) shows that μ will be a maximum when $h_{n+N_0} = cv_n^{(0)}(N,W)$, $n = 0,1,\ldots,N-1$. Thus the indexlimited sequence with most concentrated spectrum in $-W \le f \le W$ is

$$\{h_n\} = \begin{cases} 0, & n < N_0 \\ v_{n-N_0}^{(0)}(N,W), & N_0 \le n \le N_0 + N - 1 \\ 0, & n > N_0 + N - 1. \end{cases} \quad (81)$$

The concentration of its spectrum $H(f)$ is $\mu = \lambda_0(N,W)$ and

$$H(f) = dU_0(N,W;f)e^{i\pi(2N_0+N-1)f}, \quad \forall f \quad (82)$$

with d independent of f.

More generally, for $k = 1,2,\ldots,N-1$, $I_{N_0}^{N_0+N-1}\{v_{n-N}^{(k)}(N,W)\}$ is the indexlimited sequence with most concentrated spectrum in $-W \le f \le W$ that is orthogonal to

$$I_{N_0}^{N_0+N-1}\{v_{n-N}^{(i)}(N,W)\} \quad i = 0,1,\ldots,k-1.$$

The fraction of its spectral energy in $|f| \le W$ is $\lambda_k(N,W)$.

3.1.3 Simultaneously achievable concentrations

Let $\{h_n\} \leftrightarrow H(f)$ and consider the two measures of concentration

$$\alpha^2 \equiv \frac{\sum_{N_0}^{N_0+N-1} |h_n|^2}{\sum_{-\infty}^{\infty} |h_n|^2}, \qquad \beta^2 \equiv \frac{\int_{-W}^{W} |H(f)|^2 df}{\int_{-1/2}^{1/2} |H(f)|^2 df}.$$

What values of α and β are possible?

Just as in Ref. 2, pp. 74–77, one finds the attainable nonnegative values of α and β are given by the intersection of the unit square $0 \le \alpha \le 1$, $0 \le \beta \le 1$ and the elliptical region

$$\alpha^2 - 2\alpha\beta\sqrt{\lambda_0(W,N)} + \beta^2 \le 1 - \lambda_0(W,N).$$

The elliptical boundary cuts the square at $\alpha = 1$, $\beta = \sqrt{\lambda_0(W,N)}$ and $\alpha = \sqrt{\lambda_0(W,N)}$, $\beta = 1$. As either N gets large, or as $W \to 1/2$, $\lambda_0(W,N) \to 1$ as seen by (58) and (67), and the attainable region becomes the unit square.

3.1.4 Minimum energy bandlimited extension of a finite sequence

Let numbers $h_0, h_1, \ldots, h_{N-1}$ be given. There are infinitely many ways that one can choose numbers $h_N, h_{N+1}, \ldots$ and $h_{-1}, h_{-2}, \ldots$ so that the infinite sequence $\{h_n\}$ is bandlimited. Which of these sequences has least energy?

The answer is

$$h_n = \sum_{j=0}^{N-1} \alpha_j v_n^{(j)}(N,W) \tag{83}$$

$$n = 0, \pm 1, \pm 2, \ldots$$

where

$$\alpha_j = \sum_{n=0}^{N-1} h_n v_n^{(j)}(N,W) \tag{84}$$

$$j = 0, 1, \ldots, N-1.$$

The energy of this bandlimited sequence is

$$E = \sum_{-\infty}^{\infty} |h_n|^2 = \sum_{j=0}^{N-1} \frac{|\alpha_j|^2}{\lambda_j(N,W)}. \tag{85}$$

The dual to this problem is the following: Let $H(f)$ be given for $|f| \le W$. Consider extensions of H to the interval $|f| \le 1/2$ that correspond to sequences $\{h_n\} \leftrightarrow H(f)$ that are indexlimited to the index set $(N_0, N_0 + N - 1)$. Which such extension has least energy?

The situation is quite different here from the dual just discussed. Given

an arbitrary $H(f)$, $|f| \leq W$, in general there is *no* way to extend it so that the corresponding sequence will be indexlimited. The extension can be accomplished only if for $|f| \leq \frac{1}{2}$

$$H(f) = \sum_{N_0}^{N_0+N-1} h_n e^{2\pi i n f}$$

or stated another way, only if for $|f| \leq W$ we have

$$H(f) = e^{2\pi i(N_0+N-1/2)f} \sum_{j=0}^{N-1} \alpha_j U_j(N;W;f). \tag{86}$$

Then, of course,

$$\alpha_j = \frac{1}{\lambda_j} \int_{-W}^{W} H(f) e^{-2\pi i(N_0+(N-1)/2)f} U_j(N,W;f) df$$

by (11). But (86) for $|f| \leq \frac{1}{2}$ is then the extension sought of minimum energy. Its energy is

$$\int_{-1/2}^{1/2} |H(f)|^2 df = \sum_{j=0}^{N-1} \alpha_j^2. \tag{87}$$

The distinction between the two cases just treated arises, of course, because the Hilbert space of indexlimited sequences is finite dimensional while the space of bandlimited sequences is of infinite dimension.

3.1.5 Trigonometric polynomial with greatest fractional energy in an interval—optimal windows

Let $g(f)$ be a function of the form

$$g(f) = \sum_{k=0}^{N-1} g_k e^{-i\pi(N-1-2k)f}. \tag{88}$$

If $N = 2M + 1$ is odd, this can be written

$$g(f) = \sum_{n=-M}^{M} \hat{g}_n e^{2\pi i n f} \tag{89}$$

and if $N = 2M$ is even it can be written

$$g(f) = \sum_{n=-(M-1)}^{M} \hat{\hat{g}}_n e^{i\pi(2n-1)f} \tag{90}$$

where $\hat{g}_n$ and $\hat{\hat{g}}_n$ are suitably defined. In either case $g(f)$ can be called a trigonometric polynomial.

For functions of form (88) one readily computes

$$\lambda = \frac{\int_{-W}^{W} |g(f)|^2 df}{\int_{-1/2}^{1/2} |g(f)|^2 df} = \frac{\sum_{n,m=0}^{N-1} \rho(N,W)_{mn} g_m \bar{g}_n}{\sum_{0}^{N-1} |g_n|^2} \tag{91}$$

with $\rho(N,M)$ given by (21). Comparison of (91) with (18) and (26) shows that $U_0(N,W;f)$ *is the trigonometric polynomial of form* (88) *having the largest fractional concentration of energy in* $(-W,W)$.

Applications of this fact have been made to digital filtering,[12,16,18] to spectral estimation,[17] and to the definition of an essentially band-limited process by Balakrishnan[19] in 1965. In most of these applications, N is odd, and the $\hat{g}_k$ of (89) are required to be real and even in k. Thus for functions of form

$$g(f) = 2a_o + \sum_1^M a_j \cos 2\pi jf$$

with the a's real, one desires to choose the a's to maximize the fraction of the energy of g in $(-W,W)$. The answer is

$$a_j = v_{M-j}^{(o)}(2M + 1,W) \quad j = 0,1,\ldots,M$$

and

$$g(f) = cU_o(2M + 1,W;f).$$

This basic property of U_o can clearly make it of special interest in many fields.

3.2 A prediction problem

N successive samples spaced T_0 seconds apart are taken from a stationary white noise $X(t)$ of bandwidth W_0 and mean zero. The linear predictor formed from these samples that has minimum mean-squared error is used to estimate the next sample value of $X(t)$. What is the mean-squared error of this prediction, and how fast does it decrease with N?

We write $X_j = X(jT_0)$. Let the observed samples of $X(t)$ be X_0, X_1, $\ldots$, X_{N-1}. Then the predicted value $\hat{X}$ of X_N is to be of the form

$$\hat{X} = \sum_0^{N-1} a_j X_j$$

where the a's are chosen to minimize $\eta \equiv E(\hat{X} - X_N)^2$. The solution to this problem is well known. (See Ref. 11, pp. 302–305, for example.) The least value possible for η is

$$\eta_0 \equiv \min_{a's} \eta = \frac{\Delta_{N+1}}{\Delta_N} \tag{92}$$

where Δ_ℓ is the $\ell \times \ell$ determinant whose entry in row i and column j is EX_iX_j, $i,j = 0,1,\ldots,\ell - 1$. For the white noise case at hand this entry is

$$\sigma^2 \frac{\sin 2\pi W_0T_0(i-j)}{2\pi W_0T_0(i-j)} = \frac{\sigma^2}{2W_0T_0}\rho(N,W_0T_0)_{ij}$$

in the notation of (21). Here $\sigma^2 = EX(t)^2$ is the noise power. Since the determinant of a matrix is the product of its eigenvalues, it follows that

$$\eta_0 = \frac{\sigma^2}{2W_0T_0} \frac{\prod_0^N \lambda_k(N+1, W_0T_0)}{\prod_0^{N-1} \lambda_k(N, W_0T_0)}. \tag{93}$$

We can now use our knowledge (Section 2.5) of the asymptotics of the $\lambda_k(N, W)$ to find the behavior of η_0 for large N. It is shown in Appendix A that for

$$0 < W_0T_0 < \frac{1}{2},$$

$$\lim_{N\to\infty} \frac{1}{N} \log \eta_0 = \log (\sin \pi W_0T_0)^2. \tag{94}$$

Thus the mean-squared error of the best linear prediction vanishes exponentially in N when the sampling rate $1/T_0$ is greater than the Nyquist rate $2W_0$. The exponent decreases in absolute value towards the limit zero as the sampling rate is decreased to the Nyquist rate.

The situation is very different when $W_0T_0 > \frac{1}{2}$. Then η_0 approaches a limiting positive value, η_∞, as N gets large. We find (see Appendix A) that

$$\lim_{N\to\infty} \eta_0 \equiv \eta_\infty = \frac{\sigma^2 n}{2W_0T_0} \left(1 + \frac{1}{n}\right)^{2W_0T_0 - n} \tag{95}$$

$$\frac{n}{2} \leq W_0T_0 \leq \frac{n+1}{2}, \quad n = 1,2,\ldots$$

A plot of η_∞ for $\frac{1}{2} \leq W_0T_0 \leq \frac{5}{2}$ is shown in Fig. 8. Examination of (95) shows that $\eta_\infty = \sigma^2$ for $W_0T_0 = n/2$, $n = 1,2,\ldots$, and that the loops between these values shown in Fig. 8 get smaller and smaller as W_0T_0 increases. Thus, while η_∞ is zero for all sampling rates greater than the Nyquist rate, $\eta_\infty > 0.94$ for rates less than $1/2W_0$.

The foregoing is, of course, an unrealistic model of a physical prediction scheme in that it assumes perfect knowledge of the samples. If one assumes that to each sample $X(jT)$ an independent observation noise Y_j is added, then the linear predictor takes the form

$$\hat{X} = \sum_0^{N-1} a_j(X_j + Y_j). \tag{96}$$

If we assume that $\hat{X}$ is a prediction of the noisy next measurement $X(NT) + Y_N$, then all proceeds as before with the matrix ρ replaced by $\rho + (2W_0T_0\mu/\sigma^2)\, I$ where $\mu = EY_j^2$ and I is the unit matrix. By replacing

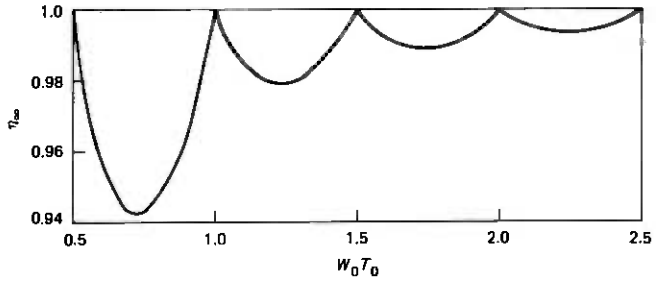


Fig. 8—Best mean squared error η_∞ vs. W_0T_0. The noise variance $\sigma^2 = 1$.

$\lambda_k(N, W_0T_0)$ by $2W_0T_0\mu/\sigma^2 + \lambda_k(N, W_0T_0)$ one finds readily that

$$\eta_\infty = \frac{\sigma^2 s}{2W_0T_0}\left(1 + \frac{1}{s}\right)^{2W_0T_0 - n}$$

$$s = \frac{\mu}{\sigma^2}2W_0T_0 + n \tag{97}$$

$$\frac{n}{2} < W_0T_0 \leq \frac{n+1}{2}, \quad n = 0,1,2,\ldots$$

When $n = 0$, so that sampling takes place faster than the Nyquist rate, η_∞ is positive. Indeed, η_∞ rises monotonically from the value μ at $T_0 = 0$ to the value $\mu + \sigma^2$ when $W_0T_0 = \frac{1}{2}$, as might be expected; perfect prediction is no longer possible.

A more satisfying model would add independent noise to the observed samples, but require $\hat{X}$ to be a best linear predictor of $X(NT)$ itself, rather than of $X(NT)$ plus noise. The asymptotic behavior of η_0 in this case seems more difficult to obtain. A related problem is readily solved, however.

Let $\hat{X}$ as given by (96) now be a minimum variance estimate of X_{N-1}, where as before the Y_i are independent identically distributed random variables that represent the imprecision of the measurement process. We are now not trying to predict X_N but rather to eliminate the noise and estimate X_{N-1} correctly. One then finds for the mean-squared error

$$\eta_0 = \mu\left[1 - \phi\,\frac{\prod_{k=0}^{N-2}[\phi + \lambda_k(N-2, W_0T_0)]}{\prod_{k=0}^{N-1}[\phi + \lambda_k(N-1, W_0T_0)]}\right] \tag{98}$$

where

$$\phi = \frac{2W_0T_0}{\sigma^2}\mu.$$

Again using the techniques of Appendix A, we find in this case that

$$\eta_\infty = \mu\left[1 - \frac{\phi}{\phi+n}\left(1+\frac{1}{n+\phi}\right)^{n-2W_0T_0}\right] \tag{99}$$

$$\frac{n}{2} < W_0T_0 \leq \frac{n+1}{2}, \quad n = 0,1,2,\ldots$$

Equation (99) can be obtained as a special case of a filtering problem solved by Viterbi.[20] He uses the result of Szegö that

$$\lim_{N\to\infty}\frac{Q_{N+1}}{Q_N} = \exp\int_{-1/2}^{1/2}\log H(f)df \tag{100}$$

where Q_N is the determinant of the $N \times N$ Toeplitz matrix having h_{i-j} as the entry in the ith row and jth column. Here, as usual, $\{h_n\} \leftrightarrow H(f)$ and we require that $h_{-n} = \bar{h}_n$, so that $H(f)$ is real. Szegö's result can indeed be applied to the ratio of determinants in (92). The Poisson summation formula, Ref. 14, p. 466, gives

$$H(f) = \frac{\sigma^2}{2W_0T_0}\sum_k \chi\left(\frac{f-k}{W}\right)$$

for this case where $\chi(f) = 1$ if $|f| \leq 1$ and zero otherwise. Carrying out the details one finds (95) again, but finds only that

$$\lim_{N\to\infty}\eta_0 = 0$$

when $W_0T_0 < \frac{1}{2}$. Our detailed knowledge of the λ's has permitted calculation of the rate at which η_0 approaches zero as expressed in eq. (94).*

3.3 The approximate dimension of signal space

The DPSS's $\{v_n^{(k)}\}$, $k = 0,1,\ldots,N-1$ are bandlimited to $(-W,W)$ (see (33)). The concentration of $\{v_n^{(k)}\}$ is given by

$$\lambda_k(N,W) = \frac{E(0,N-1)}{E(-\infty,\infty)}, \quad k = 0,1,\ldots,N-1$$

[see (22)]. From the results of Section 2.5 we have seen that as $N \to \infty$, $\lambda_k \to 1$ if $k = 2WN(1-\epsilon)$, while if $k = 2WN(1+\epsilon)$, $\lambda_k \to 0$. And this is true for any ϵ satisfying $1 > \epsilon > 0$. Thus a fraction arbitrarily close to $2W$ of the bandlimited DPSS's are confined almost entirely to the index set

* Note: After the work in Section 3.2 was completed, it was called to my attention that Widom[22] has derived an important extension of Szegö's theorem which applies to the case at hand here and gives the stronger result $\eta_0 \sim k[\sin \pi W_0T_0]^{2N}$ with k given explicitly to replace (94). The derivations of (94) and (95) given in the present paper are felt to be of interest in their own right and serve to verify the accuracy of the results given in Section 2.5.

$0 \le n \le N-1$. The remaining DPSS's have almost none of their energy in this index set.

The facts just noted can be summarized loosely in the statement that "for large N the set of sequences of bandwidth W that are confined to an index set of length about N has dimension approximately $2WN$." This basic intuitive notion can be made precise in a number of ways. We prefer the following method which treats bandlimiting and indexlimiting symmetrically.

Denote by I the index set $I = \{0,1,\ldots,N-1\}$. Now let $\epsilon > 0$ be given. Denote by G_ϵ the set of finite-energy sequences $\{h_n\} \leftrightarrow H(f)$ for which

$$E_{\bar{I}} \equiv \sum_{n \notin I} |h_n|^2 \le \epsilon \tag{101}$$

and

$$E_{\bar{W}} \equiv \int_{1/2 \ge |f| > W} |H(f)|^2 df \le \epsilon. \tag{102}$$

If ϵ is small, members of G_ϵ have little energy outside the index set $(0, N-1)$ or outside the frequency range $(-W,W)$. Now let $M = M(N,W,\epsilon,\epsilon')$ be the smallest integer such that there exist fixed sequences $\{g_n^{(1)}\}, \{g_n^{(2)}\}, \ldots, \{g_n^{(M)}\}$ such that for every $\{g\} \in G_\epsilon$ a's can be found for which

$$\sum_{n=0}^{N-1} \left[g_n - \sum_{1}^{M} a_j g_n^{(j)} \right]^2 \le \epsilon'. \tag{103}$$

In words, M is the dimension of the smallest linear space of sequences that approximates G_ϵ on the index set $(0,N-1)$ with "energy" error less than ϵ'.

With these definitions out of the way, the key theorem on the dimension of signal space can be stated as follows.

Theorem: If $\frac{1}{2} \ge W > 0$ *and* $\epsilon' > \epsilon > 0$, *then*

$$\lim_{N \to \infty} \frac{M(N,W,\epsilon,\epsilon')}{N} = 2W. \tag{104}$$

Proof of this theorem follows very closely that given in the Appendix of Ref. 7 and will be omitted here.

For applications of this theorem it is important to note that the DPSS's $\{v_n^{(k)}\}$, for $k = 0,1,\ldots,2NW(1-\eta)$ for suitable choice of η, can be used as an orthogonal basis for the M-dimensional space of sequences that best approximates G_ϵ in the sense of (103). Thus if N is large and one is dealing with sequences known to be approximately of bandwidth W and very small outside the index set $(0,N-1)$, $2WN$ numbers suffice to describe the sequence—namely, the first $2WN$ coefficients a_i in the ex-

pansion of the sequence $\{h_n\}$ on the appropriate DPSS's. We have then

$$h_n \approx \sum_0^{2WN-1} a_i v_n^{(i)}(N,W) \tag{105}$$

and

$$a_i = \sum_0^{N-1} h_n v_n^{(i)}(N,W). \tag{106}$$

Of course when N is large and $W < \frac{1}{2}$, $2WN \ll N$ so that the savings can be considerable.

Possible applications of the foregoing ideas to picture processing, cryptography, bandwidth compression and other sampled data systems should be evident. In many such applications, one starts with a signal $x(t) \in L^2(-\infty,\infty)$ defined for all time. A sequence $\{h_n\}$ is derived from $x(t)$ by sampling at rate $1/T_0$ so that

$$h_n = x(nT_0), \quad n = 0,1,\ldots. \tag{107}$$

If $X(f)$ is the spectrum of $x(t)$, so that

$$x(t) = \int_{-\infty}^{\infty} X(f)e^{2\pi ift}df, \tag{108}$$

then

$$H(f) = \sum_{-\infty}^{\infty} h_n e^{2\pi inf} = \sum_{-\infty}^{\infty} e^{2\pi inf} \int_{-\infty}^{\infty} X(f')e^{2\pi inT_0 f'}df$$

$$= \frac{1}{T_0}\sum_{-\infty}^{\infty} X\left(\frac{f-n}{T_0}\right) \tag{109}$$

by the Poisson summation formula (Ref. 14, p. 466). If now $X(f)$ vanishes for $|f| > W_0$ and if $T_0 < 1/2\ W_0$, then $H(f) = 0$ for $W' < |f| \leq \frac{1}{2}$ where $W' = W_0T_0 < \frac{1}{2}$. Thus when signals are sampled at rates greater than the Nyquist rate, the DPSS are of particular value in providing a succinct method of describing N-vectors of samples.

An interesting application of these ideas forms part of a digital transmission scheme invented by Wyner to be described in a forthcoming paper by him.

IV. DERIVATIONS

4.1 Basic facts of Section 2.1–2.3

An orderly development of this subject is facilitated by a few comments about the operators

$$L \equiv \int_{-W}^{W} df' \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} \tag{110}$$

and

$$M \equiv \frac{1}{4\pi^2}\frac{d}{df}(\cos 2\pi f - A)\frac{d}{df} + \frac{1}{4}(N^2 - 1)\cos 2\pi f \qquad (111)$$

that appear in (10) and (16). As before, we take $0 < W < \frac{1}{2}$, but now allow N to be an arbitrary real number. Operators of the type (110) and (111) have been well studied in the past and we borrow freely from the literature.

The kernel

$$K(f - f') = \frac{\sin N\pi(f - f')}{\sin \pi(f - f')} \qquad (112)$$

is real, symmetric and square-integrable over the region $|f| \leq W$, $|f'| \leq W$. The characteristic equation, $L\psi = \lambda\psi$, therefore has as solutions a set of real eigenfunctions $\psi_0, \psi_1, \psi_2, \ldots$ that are orthogonal on $|f| \leq W$ and complete in $L^2(-W,W)$. The corresponding eigenvalues are real, and those eigenvalues that are different from zero have a finite degeneracy. The eigenfunctions and eigenvalues are continuous functions of the parameter N. The kernel of the operator L in (110) is defined for all values of f. The domain of definition of eigenfunctions of L belonging to non-zero eigenvalues can then be extended to the whole line $-\infty < f < \infty$ by means of

$$\psi \equiv \frac{1}{\lambda} L\psi.$$

These eigenfunctions are readily seen to possess continuous derivatives of all orders. The eigenfunctions belonging to the eigenvalue zero can also be chosen to have derivatives of all orders.

The characteristic equation for M,

$$MU = \theta U, \qquad (113)$$

is an example of the well studied Sturm-Liouville equation (Ref. 14, p. 719). Let us denote by $\mathcal{U}$ the class of function continuous on $|f| \leq W$ and piecewise twice differentiable there. Then (113) has solutions in $\mathcal{U}$ only for a discrete set of real values of θ, the eigenvalues of M, say $\theta_0 \geq \theta_1 \geq \theta_2 \geq \ldots$ and a corresponding set of real eigenfunctions $U_0, U_1, \ldots$ can be found that are orthonormal, i.e. that satisfy

$$\int_{-W}^{W} U_i(f)U_j(f)df = \delta_{ij}. \qquad (114)$$

Furthermore the U_i's are complete in $\mathcal{L}^2(-W,W)$.

For our particular M, (111), all the eigenvalues are non-degenerate. For, suppose U_i and U_j are linearly independent continuous solutions of (113) belonging to the same eigenvalue θ. From $MU_i = \theta U_i$ and MU_j

$= \theta U_j$ we obtain $U_j M U_i - U_i M U_j = 0$ or

$$U_j \frac{d}{df}(\cos 2\pi f - A)\frac{dU_i}{df} - U_i \frac{d}{df}(\cos 2\pi f - A)\frac{dU_j}{df}$$
$$= \frac{d}{df}(\cos 2\pi f - A)\left[U_j \frac{dU_i}{df} - U_i \frac{dU_j}{df}\right] = 0.$$

Integrate this last equation from $f = -W$ to a generic point f' with $-W < f' < W$. We find that

$$U_j(f')\frac{dU_i(f')}{df} - U_i(f')\frac{dU_j(f')}{df} = 0, \quad -W < f' < W,$$

which contradicts the assumed linear independence of U_i and U_j.

The non-degeneracy of all eigenvalues permits us to write

$$\theta_0 > \theta_1 > \theta_2 > \ldots \tag{115}$$

It follows then from well-known theorems that $U_k(f)$ has exactly k zeros in the open interval $|f| < W$. That an eigenfunction U cannot vanish at either $f = W$ or $f = -W$ follows directly from the differential equation. For if U vanishes at $f = W$, for instance,

$$\frac{1}{(2\pi)^2}(\cos 2\pi f - A)\frac{d^2U}{df^2} - \frac{1}{2\pi}\sin 2\pi f \frac{dU}{df}$$
$$+ \left[\frac{1}{4}(N^2 - 1)\cos 2\pi f - \theta\right] U = 0 \tag{116}$$

evaluated at $f = W$ shows that $dU(W)/df = 0$. Differentiate (116) and evaluate at $f = W$ to see that $d^2U(W)/df^2 = 0$. Continued differentiation shows that all derivatives of U vanish at $f = W$. But U possesses a Taylor series about $f = W$ and so the assumption that $U(f) = 0$ leads to the conclusion $U \equiv 0$ which cannot be. Thus $U(W) \neq 0$.

We now know that both L and M possess orthonormal sets of eigenfunctions belonging to $\mathcal{U}$ that separately span $\mathcal{L}^2(-W,W)$. We show in Appendix C that L and M commute, i.e. for all $g(f) \in \mathcal{U}$, $LMg = MLg$. It is not hard to see then[15] that one can find a single set of orthonormal functions in $\mathcal{U}$ complete in $\mathcal{L}^2(-W,W)$ that are simultaneously eigenfunctions of L and M. Because of (115), however, the normalized eigenfunctions U_k, $k = 0,1,\ldots$ of M are unique except for sign. Thus the normalized solutions of (16) in $\mathcal{U}$, ordered by (115), are a complete set of eigenfunctions of L as well.

Any continuous solution to (113) in $|f| \leq \frac{1}{2}$ can be written as a Fourier series

$$U(f) = e^{i\pi(N-1)f}\sum_{-\infty}^{\infty} c_n e^{2\pi i n f}.$$

Substituting this form in (113) we find the 3-term recurrence for the c's

$$\frac{1}{2}n(N-n)c_{n-1} + \left[A\left(\frac{N-1}{2}-n\right)^2 - \theta\right]c_n + \frac{1}{2}(n+1)(N-n-1)c_{n+1} = 0, \quad (117)$$

$$n = 0, \pm 1, \ldots$$

Note that the coefficient of c_{n-1} here vanishes if $n = 0$ or $n = N$, while the coefficient of c_{n+1} vanishes for $n = -1$ and $n = N-1$. Thus if N *is a positive integer*, which is the case of primary importance to us, the infinite system of equation (117) uncouples and we see that a solution is possible with $0 = c_{-1} = c_{-2} = \ldots = c_N = c_{N+1} = c_{N+2} = \ldots$ provided that

$$\sum_{j=0}^{N-1} \sigma(N,W)_{ij}c_j = \theta c_i \quad (118)$$

$$i = 0,1,\ldots,N-1$$

where the real symmetric matrix $\sigma(N,W)$ is given by (14). Such a matrix has N real eigenvalues, which we now see to be eigenvalues of M as well. We denote them by $\theta_{i_1}, \theta_{i_2}, \ldots, \theta_{i_N}$. From (115) we know that $i_1, i_2, \ldots, i_N$ are N distinct non-negative integers. We denote the real eigenvector of $\sigma(N,W)$ corresponding to θ_{i_j} by

$$\underline{v}^{(i_j)} = (v_0^{(i_j)}(N,W), v_1^{(i_j)}(N,W), \ldots, v_{N-1}^{(i_j)}(N,W))^T \quad (119)$$

$$j = 1,2,\ldots,N$$

and suppose these vectors normalized so that

$$\sum_{\ell=0}^{N-1} v_\ell^{(i_j)}(N,W)v_\ell^{(i_k)}(N,W) = \delta_{jk}, \quad j,k = 1,2,\ldots,N. \quad (120)$$

We denote the corresponding eigenfunction of M by

$$U_{i_j}(N,W;f) = \sum_{n=0}^{N-1} v_n^{(i_j)}(N,W)e^{i\pi(N-1-2n)f}, \quad j = 1,2,\ldots,N. \quad (121)$$

Now again let N be a positive integer and denote by $\mathcal{G}_N$ the finite-dimensional space of functions of form

$$g(f) = \sum_{n=0}^{N-1} g_n e^{i\pi(N-1-2n)f} \quad (122)$$

where $g_0, g_1, \ldots, g_{N-1}$ are arbitrary complex numbers. We have just seen that if N is a positive integer, M leaves $\mathcal{G}_N$ invariant. Indeed, a simple

calculation shows that if g is given by (122), then

$$Mg = g'(f) \equiv \sum_{n=0}^{N-1} g_n' e^{i\pi(N-1-2n)f} \tag{123}$$

where

$$g_m' = \sum_{n=0}^{N-1} \sigma(N,W)_{mn} g_n. \tag{124}$$

With N a positive integer, L also leaves $\mathcal{G}_N$ invariant. (Indeed, in this case L projects all of $\mathcal{L}^2$ onto $\mathcal{G}_N$.) If g is given by (122), one readily finds that

$$Lg = g''(f) \equiv \sum_{n=0}^{N-1} g_n'' e^{i\pi(N-1-2n)f} \tag{125}$$

where

$$g_m'' = \sum_{n=0}^{N-1} \rho(N,W)_{mn} g_n \tag{126}$$

and the $N \times N$ symmetric matrix $\rho(N,W)$ is given by (21). This is most easily seen from the fact that for integer N the kernel (112) is degenerate. Specifically,

$$\frac{\sin N\pi(f-f')}{\sin \pi(f-f')} = \sum_{n=0}^{N-1} e^{i\pi(N-1-2n)f} e^{-i\pi(N-1-2n)f'}. \tag{127}$$

Since L and M commute, so do the matrices $\rho(N,W)$ and $\sigma(N,W)$.

We now show that for integer N the eigenfunctions of M spanning $\mathcal{G}_N$, namely $U_{i_j}(N,W;f)$, $j = 1,2,\ldots,N$, belong to the N largest eigenvalues of M, namely, $\theta_0, \theta_1, \ldots, \theta_{N-1}$. We order the integers i_j so that $\theta_{i_1} > \theta_{i_2} > \ldots > \theta_{i_N}$, so that our task is to show that $i_j = j-1$, $j = 1,2,\ldots,N$. Now, if θ' and θ'' are two eigenvalues of M with $\theta'' < \theta'$, the eigenfunction belonging to θ'' must have at least one more zero in $|f| < W$ than the eigenfunction belonging to θ' (see Ref. 14, p. 721). It follows then that U_{i_N} must have at least $N-1$ zeros in $|f| < W$, since the smallest number of zeros U_{i_1} could have in $|f| < W$ is zero. But U_{i_N} cannot possibly have more than $N-1$ zeros in this interval, since, from (121), we can write

$$U_{i_N} = e^{i\pi(N-1)f} \sum_{n=0}^{N-1} v_n^{(i_N)} z^n, \quad z = e^{-2\pi i f}$$

which shows U_{i_N} to be a function of modulus unity times a polynomial of degree at most $N-1$. It follows then that U_{i_N} has exactly $N-1$ zeros in $|f| < W$, whence U_{i_j} has precisely $j-1$ such zeros $j = 1,2,\ldots,N$. It then follows that $\theta_{i_1} = \theta_0$ the largest eigenvalue of M, $\theta_{i_2} = \theta_1$, the next largest eigenvalue, $\ldots \theta_{i_N} = \theta_{N-1}$. *Q.E.D.*

We have now shown that when N is an integer, the eigenfunctions of M that span $\mathcal{G}_N$ are

$$U_i(N,W;f) = \sum_{n=0}^{N-1} v_n^{(i)}(N,W)e^{i\pi(N-1-2n)f} \tag{128}$$

where the $\underline{v}^{(i)}$ are normalized eigenvectors of $\sigma(N,W)$:

$$\sum_{m=0}^{N-1} \sigma(N,W)_{nm} v_m^{(i)}(N,W) = \theta_i(N,W) v_n^{(i)}(N,W) \tag{129}$$

$$i,n = 0,1,\ldots,N-1.$$

These U's are also eigenfunctions of L and from (126) it then follows that

$$\sum_{m=0}^{N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)} v_m^{(i)}(N,W) = \lambda_i(N,W) v_n^{(i)}(N,W) \tag{130}$$

$$i,n = 0,1,\ldots,N-1.$$

The matrix $\rho(N,W)$ is positive definite, since

$$\sum_{n,m=0}^{N-1} \rho(N,W)_{nm}\xi_n\bar{\xi}_m = \sum \int_{-W}^{W} dt e^{2\pi i t(n-m)}\xi_n\bar{\xi}_m$$

$$= \int_{-W}^{W} \left| \sum_{0}^{N-1} \xi_n e^{2\pi i n t} \right|^2 dt$$

which is positive unless all the ξ's are zero. Thus the $\lambda_i(N,W)$ in (130) are all positive.

We have defined the U_i as eigenfunctions of M and have ordered them so that (115) is true. These same U_i are a complete set of eigenfunctions of L and we define λ_i to be the eigenvalue of L corresponding to U_i. We shall show next that the non-zero eigenvalues of L are non-degenerate and that when N is a positive integer

$$\lambda_0(N,W) > \lambda_1(N,W) > \ldots > \lambda_{N-1}(N,W) > 0. \tag{131}$$

The proof that if $\lambda \neq 0$ then λ is non-degenerate can be made exactly as in Ref. 1, equations (30)–(39). The assumption that two independent eigenfunctions of L, say U_n and U_m, belong to the same eigenvalue $\lambda \neq 0$ leads to the conclusion that $\theta_n = \theta_m$ which we have shown to be false. The reader can find details of the proof in Ref. 1.

We note next that for integer N, $U_k(N,W;fW) \to c_k P_k(f)$, $|f| \leq 1$, as $W \to 0$ where $P_k(f)$ is the Legendre polynomial of degree k. This follows directly from the differential equation (16) which for small W becomes

$$\frac{d}{df}(1-f^2)\frac{dU_k(N,W;fW)}{df} + \chi U_k(N,W;fW) + 0(W^2) \doteq 0$$

where $\chi = \frac{1}{2}(N^2 - 1) - 2\theta$. Thus $\theta_k(N,W) = \frac{1}{4}(N^2 - 1) - \frac{1}{2}k(k+1) + 0(W)$, $k = 0,1,\ldots,N-1$. Now the argument of Ref. 1, pages 61–62 holds again and it follows that for sufficiently small positive W, (131) holds. Since for integer N and $0 < W < \frac{1}{2}$ these λ's are non-degenerate and are continuous in W, it follows that (131) holds for $0 < W < \frac{1}{2}$ which is stated as (8).

Proofs of the remaining claims of Sections 2.1–2.3 are all of a more elementary nature. Most involve a straightforward calculation. We leave the details of the verification of these claims to the reader.

4.2 Asymptotics of the differential equation

We now consider solutions of the differential equation

$$\frac{d}{d\omega}[\cos\omega - A]\frac{dU}{d\omega} + \left[\frac{1}{4}(N^2-1)\cos\omega - \theta\right]U = 0 \tag{132}$$

for $0 \le \omega \le \pi$ when N is large and

$$4\theta = BN^2 + CN + \sum_{j=0}^{\infty} D_j N^{-j} \tag{133}$$

where B, C and the D's are assumed independent of N. The substitution

$$U = \frac{G}{\sqrt{\cos\omega - A}} \tag{134}$$

gives

$$\frac{d^2G}{d\omega^2} + \frac{N^2(\cos\omega - y_0)(\cos\omega - y_1)}{4(\cos\omega - A)^2}G = 0 \tag{135}$$

or

$$\frac{d^2G}{d\omega^2} + \left[N^2\frac{\cos\omega - B}{4(\cos\omega - A)} - N\frac{C}{4(\cos\omega - A)} + 0(1)\right]G = 0. \tag{136}$$

Here

$$y_0 = B + 0\left(\frac{1}{N}\right), \qquad y_1 = A + 0\left(\frac{1}{N}\right). \tag{137}$$

Case A. $1 > B > A > -1$ or $k = \lfloor 2WN(1-\epsilon)\rfloor$

If $1 > B > A > -1$, then, as seen from (135) and (137), U is oscillatory for $1 \ge \cos\omega \ge B$ and for $A \ge \cos\omega \ge -1$, but is non-oscillatory in the interval $B \ge \cos\omega \ge A$. We investigate the solutions of (132) separately in each of these regions and also in the vicinity of the turning points y_0 and y_1.

Let $\cos \omega - A = t/N^2$. Then (132) becomes

$$t\frac{d^2U}{dt^2} + \frac{dU}{dt} - \frac{B-A}{4(1-A^2)}U + 0\left(\frac{1}{N}\right) = 0$$

so that near $\cos \omega = A$ we have

$$U \sim \begin{cases} U_4 \equiv d_4 J_0\left(N\sqrt{\frac{(B-A)}{(1-A^2)}(\cos\omega - A)}\right), & \cos\omega \geq A \\ U_5 \equiv d_4 J_0\left(N\sqrt{\frac{(B-A)}{(1-A^2)}(A - \cos\omega)}\right), & \cos\omega \leq A. \end{cases} \tag{138}$$

Here I_0 and J_0 are the usual Bessel functions. We note that when $\cos \omega = A + u/N$

$$U_4 = d_4 J_0\left(N^{1/2}\sqrt{\frac{(B-A)}{(1-A^2)}}u\right)$$

$$\sim \frac{d_4}{\sqrt{2\pi}}\left[N^{1/2}\sqrt{\frac{(B-A)}{(1-A^2)}}u\right]^{-1/2}\exp\left(N^{1/2}\sqrt{\frac{(B-A)}{(1-A^2)}}u\right) \tag{139}$$

(see Ref. 9, Vol. II, eq. 7.13.5, p. 86). When $\cos \omega = A - u/N$

$$U_5 = d_4 J_0\left(N^{1/2}\sqrt{\frac{(B-A)}{(1-A^2)}}u\right)$$

$$\sim d_4\sqrt{\frac{2}{\pi}}\left[N^{1/2}\sqrt{\frac{(B-A)}{(1-A^2)}}u\right]^{-1/2}\cos\left[N^{1/2}\sqrt{\frac{(B-A)}{(1-A^2)}}u - \frac{\pi}{4}\right] \tag{140}$$

(see Ref. 9, Vol. II, eq. 7.13.3, p. 85).

Now, the *WKB* solution of

$$\frac{d^2g}{dx^2} - [n^2E^2(x) + nF(x) + 0(1)]g = 0 \tag{141}$$

for large n is

$$g(x) \sim \frac{1}{\sqrt{E}}\left[c_1 e^{-n\int E dx - 1/2\int (F/E)dx} + c_2 e^{n\int E dx + 1/2\int (F/E)dx}\right] \tag{142}$$

provided x is not a zero of $E(x)$. (See Ref. 8, Sec. 7, Lemma 2.) Applying this to (136) and taking account of (134), we find that for $B > \cos \omega > A$

an asymptotic solution of (132) is

$$U \sim U_3 = d_3 R(\omega) \exp\left(-\frac{N}{2}\int_{\arccos B}^{\omega}\sqrt{\frac{B-\cos t}{\cos t - A}}\,dt - \frac{1}{4}\int_{\arccos B}^{\omega}\frac{C\,dt}{\sqrt{(B-\cos t)(\cos t - A)}}\right), \quad (143)$$

$$R(\omega) \equiv |(B-\cos\omega)(A\cos\omega)|^{-1/4}.$$

Here we have chosen $c_2 = 0$ in (142) to obtain a matching of U_3 and U_4 at $\cos\omega = A + u/N$. Indeed, one finds in a straighforward way that

$$U_3\left(\arccos\left(A+\frac{u}{N}\right)\right) \sim \frac{d_3 N^{1/4}}{u^{1/4}[B-A]^{1/4}} \times \exp\left(-\frac{N}{2}L_3 + N^{1/2}\sqrt{\frac{B-A}{1-A^2}}\,u - \frac{1}{4}CL_4\right) \quad (144)$$

where

$$L_3 = \int_A^B \sqrt{\frac{B-\xi}{(\xi-A)(1-\xi^2)}}\,d\xi,$$

$$L_4 = \int_A^B \frac{d\xi}{\sqrt{(B-\xi)(\xi-A)(1-\xi^2)}}. \quad (145)$$

Comparison of (144) and (139) shows that

$$d_4 = \sqrt{2\pi}\,N^{1/2}(1-A^2)^{-1/4}e^{-(NL_3/2)-(CL_4/4)}d_3. \quad (146)$$

An asymptotic solution to (132) near the turning point $\cos\omega = y_0$ is obtained by substituting $\cos\omega - B = t/N^{2/3}$ to obtain

$$\frac{d^2U}{dt^2} + \frac{t}{4(1-B^2)(B-A)}U + 0(N^{-2/3}) = 0. \quad (147)$$

Thus, near $\cos\omega = B$, we find

$$U \sim U_2 \equiv d_2 Ai\left(-\frac{N^{2/3}(\cos\omega - B)}{[4(1-B^2)(B-A)]^{1/3}}\right) \quad (148)$$

(see Ref. 10, 10.4.1, p. 446). Here we have chosen the asymptotic solution of (147) that agrees with U_3 at $\cos\omega = B - u/\sqrt{N}$. Indeed, from Ref. 10, 10.4.59, page 448, we have that

$$U_2\left(\arccos\left(B-\frac{u}{\sqrt{N}}\right)\right) = d_2 Ai\left(\frac{N^{1/6}u}{[4(1-B^2)(B-A)]^{1/3}}\right) \sim \frac{d_2}{2\sqrt{\pi}}\left[\frac{N^{1/6}u}{[4(1-B^2)(B-A)]^{1/3}}\right]^{-1/4} \times \exp\left(-\frac{2}{3}\frac{N^{1/4}u^{3/2}}{[4(1-B^2)(B-A)]^{1/2}}\right). \quad (149)$$

On the other hand, from (143) we find that

$$U_3\left(\arccos\left(B-\frac{u}{\sqrt{N}}\right)\right)\sim\frac{d_3N^{1/8}}{[u(B-A)]^{1/4}}\times\exp\left(-\frac{N^{1/4}}{2}\frac{2}{3}\frac{u^{3/2}}{[(1-B^2)(B-A)]^{1/2}}\right)\quad(150)$$

so that on comparison with (149) we must have

$$d_3=\frac{2^{-5/6}}{\sqrt{\pi}}N^{-1/6}(1-B^2)^{1/12}(B-A)^{1/3}d_2.\quad(151)$$

On the other side of this turning point the solution U_2 continues as

$$U_2\left(\arccos\left(B+\frac{u}{\sqrt{N}}\right)\right)=d_2Ai\left(\frac{-N^{1/6}u}{[4(1-B^2)(B-A)]^{1/3}}\right)$$
$$\sim\frac{d_2}{\sqrt{\pi}}\left[\frac{N^{1/6}u}{[4(1-B^2)(B-A)]^{1/3}}\right]^{-1/4}\sin\left[\frac{2}{3}\frac{N^{1/4}u^{3/2}}{\sqrt{4(1-B^2)(B-A)}}+\frac{\pi}{4}\right]\quad(152)$$

as seen from Ref. 10, 10.4.60, page 448.

Applying (141)–(142) to (136) for $1\geq\cos\omega>B$, we find that

$$E=i\sqrt{\frac{\cos\omega-B}{4(\cos\omega-A)}}\,.$$

On recalling (134), we find the asymptotic formula

$$U\sim U_1\equiv d_1R(\omega)\cos\left[\frac{N}{2}\int_0^\omega\sqrt{\frac{\cos t-B}{\cos t-A}}\,dt-\frac{C}{4}\int_0^\omega\frac{dt}{\sqrt{(\cos t-B)(\cos t-A)}}+\phi\right]\quad(153)$$

with $R(\omega)$ as in (143). Near the turning point $\cos\omega=y_0$, this becomes

$$U_1\left(\arccos\left(B+\frac{u}{\sqrt{N}}\right)\right)\sim\frac{d_1N^{1/8}}{u^{1/4}(B-A)^{1/4}}\times\cos\left[\frac{N}{2}L_1-\frac{\frac{2}{3}N^{1/4}u^{3/2}}{\sqrt{4(B-A)(1-B^2)}}-\frac{C}{4}L_2+\phi\right]\quad(154)$$

where

$$L_1=\int_B^1\sqrt{\frac{\xi-B}{(\xi-A)(1-\xi^2)}}\,d\xi,$$
$$L_2=\int_B^1\frac{d\xi}{\sqrt{(\xi-B)(\xi-A)(1-\xi^2)}}.\quad(155)$$

Comparison of (154) and (152) shows that we must have

$$d_2 = \sqrt{\pi}\, N^{1/6}[4(1-B^2)]^{-1/12}(B-A)^{-1/3} d_1 \tag{156}$$

and

$$\frac{N}{2} L_1 - \frac{C}{4} L_2 + \phi = \frac{\pi}{4} \pmod{2\pi}. \tag{157}$$

Turning now to the interval $A > \cos\omega \geq -1$, we find from (141)–(142)

$$U \sim U_6 \equiv \frac{d_6}{[(B-\cos\omega)(A-\cos\omega)]^{1/4}} \times \cos\left[\frac{N}{2}\int_\omega^\pi \sqrt{\frac{B-\cos t}{A-\cos t}}\, dt + \frac{C}{4}\int_\omega^\pi \frac{dt}{\sqrt{(B-\cos t)(A-\cos t)}} + \theta\right]. \tag{158}$$

At $\cos\omega = A - u/N$ this becomes

$$U_6\left(\arccos\left(A - \frac{u}{N}\right)\right) \sim \frac{d_6 N^{1/4}}{(B-A)^{1/4} u^{1/4}} \cdot \times \cos\left[\frac{N}{2} L_5 - N^{1/2}\sqrt{\frac{B-A}{1-A^2}}\, u + \frac{C}{4} L_6 + \theta\right] \tag{159}$$

where

$$L_5 = \int_{-1}^{A} \sqrt{\frac{B-\xi}{(A-\xi)(1-\xi^2)}}\, d\xi,$$

$$L_6 = \int_{-1}^{A} \frac{d\xi}{\sqrt{(B-\xi)(A-\xi)(1-\xi^2)}}. \tag{160}$$

Comparison with (140) shows that

$$d_6 = \sqrt{\frac{2}{\pi}}\, N^{-1/2}(1-A^2)^{1/4} d_4. \tag{161}$$

$$\frac{N}{2} L_5 + \frac{C}{4} L_6 + \theta = \frac{\pi}{4} \pmod{2\pi}. \tag{162}$$

Equations (138) and (148) provide asymptotic solutions to (132) at the two turning points $\cos\omega = B$ and $\cos\omega = A$. Equations (153), (143) and (158) provide asymptotic solutions for the regions away from the turning points. Equations (146), (151), (156)–(157) and (161)–(162) insure that these solutions join together. This solution is summarized in Eqs. (42) where the regions of validity for each piece are shown explicitly, and in (43)–(48). As presented there, the constant ϕ of (153) has been

chosen as $-[1-(-1)^k]\pi/4$ as it must to satisfy the inequalities shown in (9).

To normalize the solution (42)–(44) we must compute

$$W_i \equiv d_i^2 \int g_i^2(\omega) d\omega \tag{163}$$

$$i = 1,2,\ldots,6$$

where the range of integration for each g_i^2 is the range of validity for that g given in (42). We then require that

$$\sum_1^6 W_i = \pi \tag{164}$$

since $\omega = 2\pi f$.

Asymptotic forms for the W_i are readily worked out. One finds, for example,

$$W_1 = d_1^2 \int_0^{\arccos (B+1/\sqrt{N})} g_1^2(\omega) d\omega$$

$$\sim d_1^2 \frac{1}{2} \int_{B+1/\sqrt{N}}^1 \frac{d\xi}{\sqrt{(\xi - B)(\xi - A)(1-\xi^2)}} \sim \frac{1}{2} d_1^2 L_2 \tag{165}$$

while, with $\nu \equiv [4(1-B^2)(B-A)]^{-1/3}$,

$$W_2 = d_2^2 \left[\int_{B-1/\sqrt{N}}^B dt + \int_B^{B+1/\sqrt{N}} dt \right] \frac{Ai^2(-N^{2/3}\nu(t-B))}{\sqrt{1-t^2}}$$

$$\sim \frac{d_2^2}{\sqrt{N}} \int_0^1 \frac{Ai^2(N^{1/6}\nu\xi) d\xi}{\sqrt{1-B^2}} + \frac{d_2^2}{\sqrt{N}} \int_0^1 \frac{Ai^2(-N^{1/6}\nu\xi) d\xi}{\sqrt{1-B^2}}.$$

By using the asymptotic forms for $Ai(x)$ (see Ref. 10, 10.4.59, 10.4.60, page 448) one finally finds

$$W_2 \sim \frac{d_2^2}{\pi\sqrt{\nu}\sqrt{1-B^2}} N^{-7/12} = c \frac{d_1^2}{N^{1/4}} = \frac{2c}{L_2} \frac{W_1}{N^{1/4}}$$

where c is independent of N. In a like manner, one finds that all the ratios W_i/W_1, $i = 2,3,\ldots,6$ vanish with increasing N. We omit the details here. Equations (164) and (165) now give $\pi \sim W_1 \sim \frac{1}{2} d_1^2 L_2$ so that $d_1 = [2\pi/L_2]^{1/2}$. Equations (146), (151), (156) and (161) now determine all the d's to have the values given in (48).

Recall now that $U_k(N,W;f)$ has k zeros in $-W < f < W$. For the solution we have just constructed, all zeros in $(-W,W)$ are contributed by U_1 of (153). From (154) we see that the number of zeros is given asymptotically by

$$k = \left\lfloor \frac{2}{\pi} \left[\frac{N}{2} L_1 - \frac{2}{3} \frac{N^{1/4}}{\sqrt{4(B-A)(1-B^2)}} - \frac{C}{4} L_2 \right] \right\rfloor \sim \frac{N}{\pi} L_1.$$

Thus if we set $k/N = 2W(1-\epsilon) \sim L_1/\pi$, we must have $L_1 \sim 2\pi W(1-\epsilon) = \pi k/N$ which is (43).

Finally, the two phase continuity requirements (157) and (162) must be met. The first of these is satisfied by the choice of C given in (45). This number lies between zero and $8\pi/L_2$ and hence is 0(1) as has been assumed throughout the development. Equation (162) is satisfied by choosing θ as in (46).

Case B. $1 = B > A > -1$ or $k = 0(1)$

If, in the preceding analysis, B is allowed to approach 1, the first turning point approaches $\omega = 0$ and the subinterval of $(-W, W)$ in which U can oscillate becomes vanishingly small. This suggests a separate investigation of (132)–(133) around $\omega = 0$ when $B = 1$.

Substitute

$$\omega = \frac{(2\alpha)^{1/4}}{\sqrt{N}}\, t$$

into (132), where, as before, $\alpha \equiv 1 - A$. One finds

$$\frac{d^2U}{dt^2} + \left[-\frac{C}{4}\sqrt{\frac{2}{\alpha}} - \frac{t^2}{4}\right] U + 0\left(\frac{1}{N}\right) = 0.$$

Asymptotically, then, U is a solution of Weber's equation $\ddot{D} + (\chi - \frac{1}{4}t^2)D = 0$ (see Ref. 9, Vol. II, 8.2, page 116) which has bounded solutions only if $\chi = k + \frac{1}{2}$ where k is a nonnegative integer. We are thus forced to take

$$C = -4\sqrt{\frac{\alpha}{2}}\left(k + \frac{1}{2}\right), \quad \theta = \frac{1}{4}N^2 - \left(k + \frac{1}{2}\right)\sqrt{\frac{\alpha}{2}}\,N + 0(1). \quad (166)$$

The corresponding solution is generally denoted by $D_k(t)$ and has exactly k zeros (Ref. 9, Vol. II, 8.6, page 126). Thus we are led to take

$$U_k(\omega) \sim c_1 D_k(t) = c_1 D_k\left(\left(\frac{N^2}{2\alpha}\right)^{1/4}\omega\right) \quad (167)$$

for fixed t, or $\omega = 0(N^{-1/2})$, as reported in (39). Examination of higher order terms (omitted here) shows (167) to be correct asymptotically even for $\omega = 0(N^{-1/3})$, whence the range of validity shown in (38).

Solutions of (132) near A and in the regions away from the turning points can be obtained in the present case from U_3, U_4, U_5 and U_6 of Section 4.2, Case A, by letting $B = 1$ and $C = -\sqrt{\alpha/2}(k + \frac{1}{2})$ in (138), (143), and (158). The indicated integrals in these last two equations can now be carried out explicitly. Equations (39) result. The constants c_1, $c_2, \ldots, c_5$ are then adjusted so that the solutions match asymptotically at the edges of their regions of validity. Finally, the solution is normalized. Again the oscillatory part near $f = 0$ dominates the asymptotic

behavior of $\int_{-1/2}^{1/2} U_k(N,W;f)^2 df$ and we find

$$V_1 \equiv c_1^2 \int_0^{N^{-1/3}} f_1(\omega)^2 d\omega = \left(\frac{2\alpha}{N^2}\right)^{1/4} \times c_1^2 \left[\int_0^\infty D_k^2(t)dt - \int_{N^{1/6}/(2\alpha)^{1/4}}^\infty D_k^2(t)dt \right]$$

$$\sim c_1^2 \left(\frac{2\alpha}{N^2}\right)^{1/4} \left[\sqrt{\frac{\pi}{2}} k! - \int_{N^{1/6}/(2\alpha)^{1/4}}^\infty t^{2k} e^{-t^2/2} dt \right]$$

$$\sim c_1^2 \frac{(2\alpha)^{1/4}}{\sqrt{N}} \sqrt{\frac{\pi}{2}} k! \quad (168)$$

(See Ref. 13, 7.711-1, p. 885 and Ref. 9, Vol. II, 8.4.1, page 122.) This determines c_1 and the values shown in (40) are obtained. We omit the straightforward but tedious details here.

Case C. $1 > B = A > -1$ or $k = \lfloor 2WN + (b/\pi) \log N \rfloor$

If, in the analysis of Section 4.2, Case A, the parameter B is allowed to approach the value A, the two turning points coincide at $\cos \omega = A$ and a new analysis of U in this neighborhood is now required.

With $\theta = AN^2 + CN + 0(1)$ and

$$\beta \equiv \frac{1}{\sqrt{1 - A^2}} = |\csc 2\pi W|,$$

in (133) substitute

$$\cos \omega - A = i \frac{\xi}{N\beta}, \; U = e^{-1/2\xi} F \quad (169)$$

to obtain

$$\xi F'' + (1 - \xi)F' - \frac{1}{2}(1 - iE\beta)F + 0\left(\frac{1}{N}\right) = 0 \quad (170)$$

where

$$E \equiv \frac{1}{2}(A - C). \quad (171)$$

For large N, then, $F(\xi) \sim \Phi(a,1;\xi)$ where

$$a \equiv \frac{1}{2}(1 - iE\beta)$$

and

$$\Phi(a,c;x) = 1 + \frac{a}{c}\frac{x}{1!} + \frac{a(a+1)}{c(c+1)}\frac{x^2}{2!} + \ldots$$

is the confluent hypergeometric function. (See Ref. 9, Vol. I, 6.1.1, page

248, and 6.2.6, page 250.) Thus for $\cos\omega$ near A, we have

$$U \sim e_2 e^{i(\beta/2)N(\cos\omega - A)}\Phi(a,1;-i\beta N(\cos\omega - A))$$

as reported in (51). This expression is real.

Solutions away from the double turning point $\cos\omega = A$ can be obtained from (153) and (158) by setting $B = A$. The integrations can be done explicitly. The functions h_1 and h_3 of (51) result when C is replaced by E via (171).

There now remains the task of choosing the constants so that h_1, h_2, and h_3 join properly. We indicate a few key steps.

When

$$\cos\omega = A + \frac{u}{N^{2/3}},$$

$$h_1 \sim u^{-1/2}N^{1/3}\cos\left[\frac{N}{2}\arccos A - \frac{1}{2}N^{1/3}\beta u - \frac{E\beta}{2}\log u + E\beta\log\frac{N^{1/3}2^{1/2}}{\beta} - k\frac{\pi}{2}\right]. \quad (172)$$

To develop an asymptotic expression for h_2 at this point, we avail ourselves of the formula (Ref. 9, Vol. 1, 6.13.1.2, p. 278)

$$\Phi(a,c;x) \sim \frac{\Gamma(c)}{\Gamma(c-a)}\left(\frac{e^{i\pi\epsilon}}{x}\right)^a + \frac{\Gamma(c)}{\Gamma(a)}e^x x^{a-c} \quad (173)$$

where $\epsilon = 1$ if $\operatorname{Im} x > 0$ and $\epsilon = -1$ if $\operatorname{Im} x < 0$. One finds

$$h_2\left[\arccos\left(A + \frac{u}{N^{2/3}}\right)\right] \sim \frac{e^{-\pi E\beta/4}}{r(E\beta)\sqrt{2\gamma}}\cos\left[-\gamma - \psi(E\beta) + \frac{\pi}{4} - \frac{1}{2}E\beta\log 2\gamma\right] \quad (174)$$

where $\gamma \equiv \frac{1}{2}\beta N^{1/3}u$ and where the real functions r and ψ are defined by

$$\Gamma\left(\frac{1}{2} - \frac{1}{2}i\alpha\right) = r(\alpha)e^{i\psi(\alpha)}.$$

Comparison of (172) and (174) shows that we must have

$$e_2 = \beta^{1/2}r(E\beta)e^{\epsilon\beta\pi/4}N^{1/2}e_1 \quad (175)$$

and

$$\pi WN + \frac{E\beta}{2}\log N - \frac{E\beta}{2}\log\frac{\beta}{2} + \psi(E\beta) - k\frac{\pi}{2} - \frac{\pi}{4} = 0 (\mathrm{mod}\, 2\pi). \quad (176)$$

When $\cos \omega = A - u/N^{2/3}$, from (51) one finds

$$h_3 \sim \frac{N^{1/3}}{\sqrt{u}} \cos \left[\frac{N}{2} \arccos A + \frac{1}{2} N^{1/3} \beta u - \frac{E\beta}{2} \log u + E\beta \log \frac{N^{1/3} 2^{1/2}}{\beta} - (k+1)\frac{\pi}{2} \right] \quad (177)$$

while from (51) and (173)

$$h_2 \left[\arccos \left(A - \frac{u}{N^{2/3}} \right) \right] \sim \frac{e^{\pi E\beta/4}}{r(E\beta)\sqrt{2\gamma}} \cos \left[\gamma - \psi(E\beta) - \frac{E\beta}{2} \log 2\gamma - \frac{\pi}{4} \right]. \quad (178)$$

Comparison of these last two equations yields

$$e_3 = \beta^{-1/2} r(E\beta)^{-1} e^{E\beta\pi/4} N^{-1/2} e_2 \quad (179)$$

to match the amplitudes, while matching of the cosine arguments gives (176) again.

Now the number of zeros, k_1, of (50) in the interval $(0 \leq \omega \leq \arccos (A + N^{-2/3}))$ is seen from (172) to be given asymptotically by

$$k_1 \sim \frac{1}{\pi} \left[\pi WN - \frac{1}{2} N^{1/3} \beta + E\beta \log \frac{N^{1/3} 2^{1/2}}{\beta} \right]$$

while asymptotically the number, k_2, of zeros of U in $(\arccos (A + N^{-2/3}) \leq \omega \leq \Omega)$ is obtained from (174) as

$$k_2 \sim \frac{1}{\pi} \left[\frac{1}{2} \beta N^{1/3} + \psi(E\beta) - \frac{\pi}{4} + \frac{E\beta}{2} \log \beta N^{1/3} \right].$$

Thus the number of zeros of $U(f)$ in $(-W, W)$ is given by

$$k = 2(k_1 + k_2) \sim \frac{2}{\pi} \left[\pi WN + \frac{E\beta}{2} \log N - \frac{E\beta}{2} \log \frac{\beta}{2} + \psi(E\beta) - \frac{\pi}{4} \right]. \quad (180)$$

This motivates the choice of E as a root of (53). When this is done, the matching condition (176) is also satisfied.

The constant e_1, e_2, e_3 must now be determined by (175), (179) and the normalization requirement (9). Routine calculations show that

$$X_1 \equiv e_1^2 \int_0^{\arccos [A + N^{-2/3}]} [h_1(\omega)]^2 d\omega \sim \frac{1}{2} e_1^2 \int_0^{\arccos [A + N^{-2/3}]} \frac{d\omega}{\cos \omega - A} \sim \frac{e_1^2 \beta}{3} \log N, \quad (181)$$

$$X_3 \equiv e_3^2 \int_{\arccos\,[A-N^{-2/3}]}^{\pi} [h_3(\omega)]^2 d\omega$$

$$\sim \frac{1}{2} e_3^2 \int_{\arccos\,[A-N^{-2/3}]}^{\pi} \frac{d\omega}{A - \cos\omega} \sim \frac{e_3^2 \beta}{3} \log N = \frac{e_1^2 \beta e^{E\beta\pi}}{3} \log N$$

while

$$X_2 \equiv e_2^2 \int_{\arccos\,[A+N^{-2/3}]}^{\arccos\,[A-N^{-2/3}]} [h_2(\omega)]^2 d\omega \sim e_2^2\, 0\left(\frac{1}{N}\right) \sim e_1^2\, 0(1) \quad (182)$$

which is negligibly small compared to the first two integrals. The normalization integral thus gives

$$e_1^2 \frac{\beta}{3} [1 + e^{E\beta\pi}] \log N = \pi.$$

The values (55) then follow where the factor $(-1)^{\lfloor k/2 \rfloor}$ is dictated by (9).

We have assumed throughout this analysis that $E \equiv \frac{1}{2}(A - C) = 0(1)$. It follows then from (53) that we must have $k = 2WN + (E\beta/\pi) \log N + 0(1)$. If then, we write

$$k = \left\lfloor 2WN + \frac{b}{\pi} \log N \right\rfloor$$

as in (49), it is seen that for the root of (53) we have

$$E \sim b/\beta. \quad (183)$$

Consideration of the detailed nature of ψ shows that E must be taken as the root of (53) of smallest absolute value.

4.3 Asymptotics of $\lambda_k(N,W)$ for large N

The values of $\lambda_k(N,W)$ for large N reported in Section 2.5 are obtained from the asymptotic expressions for $U_k(N,W;f)$ given in Section 2.4 by means of the basic relation

$$\lambda_k(N,W) = \int_{-W}^{W} U_k(N,W;f)^2 df \Big/ \int_{-1/2}^{1/2} U_k(N,W;f)^2 df. \quad (184)$$

Let

$$V_i = \int c_i^2 f_i^2(\omega) d\omega \qquad i = 1,2,\ldots,5$$

$$W_i = \int d_i^2 g_i^2(\omega) d\omega \qquad i = 1,2,\ldots,6$$

$$X_i = \int e_i^2 h_i^2(\omega) d\omega \qquad i = 1,2,3, \quad (185)$$

where the ranges of integration are given by the corresponding intervals of validity shown in (38), (42) and (50).

For fixed k and large N, $1 - \lambda_k \sim (V_4 + V_5)/\pi$ since $\Sigma V_i = \pi$. Now straightforward developments yield

$$V_4 = c_4^2 \int_{2\pi W}^{\arccos [A - N^{-2/3}]} J_0^2 \left[\frac{N}{\sqrt{2-\alpha}} \sqrt{A - \cos \omega} \right] d\omega$$

$$\sim \frac{c_3^2}{N^{2/3}\sqrt{1-A^2}} \int_0^1 J_0^2 \left[\frac{N^{2/3}}{\sqrt{2-\alpha}} \sqrt{t} \right] dt \sim \frac{\sqrt{2-\alpha}}{2\pi} \frac{c_3^2}{N^{4/3}}, \quad (186)$$

$$V_5 = c_5^2 \int_{\arccos [A - N^{-2/3}]}^{\pi} f_5^2(\omega) d\omega$$

$$\sim \frac{c_5^2}{2} \int_{-1}^{A - N^{-2/3}} \frac{dt}{\sqrt{(A-t)(1-t)(1-t^2)}} \sim \frac{\pi}{2\sqrt{2(1-A)}} c_5^2. \quad (187)$$

But, from (40), $c_3^2/(N^{4/3}c_5^2) = 0(N^{-1/3})$ so V_4 is negligible compared to V_5 and we have

$$1 - \lambda_k \sim \frac{1}{\pi} V_5 \sim c_5^2/2\sqrt{2(1-A)}$$

which is (58).

The formula (59) is obtained from $1 - \lambda_k \sim (W_5 + W_6)/\pi$ using (42) and (44). One finds $W_5 \sim d_4^2/(2\pi N^{3/2}\sqrt{B-A})$ and $W_6 \sim \frac{1}{2} d_6^2 L_6$. Equations (48) now show W_5 to be negligible and $1 - \lambda_k \sim W_6/\pi \sim L_6 d_6^2/2\pi$. Insertion of the value for d_6 in (48) gives (59).

Formula (60) arises from

$$\lambda = \frac{1}{\pi} \left[X_1 + \int_{\arccos [A + N^{-2/3}]}^{2\pi W} e_2^2 h_2^2(\omega) d\omega \right].$$

We have already commented in connection with (182) that the integral here is of smaller order than X_1 so that

$$\lambda = \frac{1}{\pi} X_1 \sim \frac{e_1^2 \beta}{3\pi} \log N = [1 + e^{E\beta\pi}]^{-1} \quad (188)$$

by (181) and (55). Since $E \sim b/\beta$ by (183), (60) results.

Finally, the approximation (61)–(62) arises from (188) and (53) by solving the latter approximately for $E\beta$. From the theory of the Γ function (Ref. 10, 6.1.27, p. 256, and 6.3.3, p. 258) one finds that $\psi(s) = \frac{1}{2}(\gamma + 2 \log 2)s + 0(s^2)$. Inserting this in (53) one finds

$$E\beta \approx - \frac{N\pi W - \frac{k\pi}{2} - \frac{\pi}{4}}{\frac{1}{2} \log \frac{8N}{\beta} + \frac{1}{2}\gamma}.$$

This together with (188) is (61)–(62).

4.4 Asymptotics of $\lambda_k(N,W)$ for small W

Consider the matrix eigenvalue problem

$$\sum_{j=0}^{N-1} K(cx_i,cx_j)w_j\psi_j = \mu\psi_i, \quad i = 0,1,\ldots,N-1 \tag{189}$$

which has solutions only for those values of $\nu \equiv 1/\mu$ for which the determinant $\mathcal{D}(\nu) \equiv |\delta_{ij} - \nu K(cx_i,cx_j)|_{N-1}$ vanishes. Here, as in the rest of this section, we denote by $|f(i,j)|_{N-1}$ the determinant of the $N \times N$ matrix whose element in the ith row and jth column is $f(i,j)$, $i,j = 0,1,\ldots,N-1$. In (189) we consider the function $K(\cdot,\cdot)$, the weights w_j and the points x_j, $j = 0,1,\ldots,N-1$ as given. The number c is a parameter. For the determinant we have the development in powers of ν

$$\mathcal{D}(\nu) = 1 + \sum_{n=1}^{N} (-1)^n d_n \nu^n,$$

where

$$d_n = \frac{1}{n!} \sum_{\ell_0=0}^{N-1} \cdots \sum_{\ell_{n-1}=0}^{N-1} |K(cx_{\ell_i},cx_{\ell_j})w_{\ell_j}|_{n-1}.$$

If now

$$K(x,y) = \sum_{0}^{\infty} a_{ij}x^i y^j, \quad a_{00} = 1, \tag{190}$$

the development in the appendix of Ref. 21 from equation (A4) to (A9) can be repeated step by step with all integrals replaced by sums to show that for small c we have for the eigenvalues of (189)

$$\mu_n = c^{2n}\chi_0(n)[1 + 0(c)] \tag{191}$$

$$n = 0,1,\ldots,N-1$$

where

$$\chi_0(n) = \frac{|a_{ij}|_n |h_{i+j}|_n}{|a_{ij}|_{n-1}|h_{i+j}|_{n-1}}. \tag{192}$$

Here

$$h_\gamma = \sum_{i=0}^{N-1} x_i^\gamma w_i. \tag{193}$$

To use this result to obtain asymptotics of $\lambda_k(N,W)$ for small W, divide (18) by $2W$, and write

$$c \equiv 2\pi W, \mu = \frac{\lambda}{2W}, x_j = j, j = 0,1,\ldots,N-1. \tag{194}$$

Equation (18) then becomes (189) with $w_j = 1$ and

$$K(x,y) = \frac{\sin(x-y)}{x-y} = \sum_0^\infty \frac{(-1)^n (x-y)^{2n}}{(2n+1)!}$$

$$= \sum_{n,j} \frac{(-1)^n \binom{2n}{j} x^j (-y)^{2n-j}}{(2n+1)!}. \quad (195)$$

For evaluation of (192) we thus have

$$a_{ij} = \begin{cases} 0, & i+j \text{ odd} \\ \dfrac{(-1)^{j+(i+j)/2}}{i!j!(i+j+1)}, & i+j \text{ even} \end{cases} \quad (196)$$

and

$$h_\gamma = \sum_{\ell=0}^{N-1} \ell^\gamma. \quad (197)$$

To evaluate (192) we first note that the equations

$$\sum_{j=0}^{n} a_{ij} Y_j = \delta_{in}, \quad i = 0,1,\ldots,n \quad (198)$$

yield

$$Y_n = \frac{|a_{ij}|_{n-1}}{|a_{ij}|_n}, \quad (199)$$

the reciprocal of one factor of (192). Now, from (196) we see that we can also write

$$a_{ij} = \frac{(-1)^{j+(i+j)/2}}{i!j!} \frac{1}{2} \int_{-1}^{1} t^{i+j} dt. \quad (200)$$

Insert this expression for a_{ij} into (198) and define

$$F(t) \equiv \sum_{j=0}^{n} (-1)^{3j/2} \frac{t^j}{j!} Y_j. \quad (201)$$

Equation (198) then reads

$$\int_{-1}^{1} F(t) t^i dt = \frac{2n!}{(-1)^{n/2}} \delta_{in} \quad (202)$$

$$i = 0,1,\ldots,n.$$

But $F(t)$ is a polynomial of degree n orthogonal on $(-1,1)$ to t^i, $i = 0,\ldots,n-1$. We can write therefore $F(t) = kP_n(t)$ with $P_n(t)$ the usual Le-

gendre polynomial. Now

$$\int_{-1}^{1} P_n(t)t^n dt = \frac{2^{n+1}(n!)^2}{(2n+1)!}$$

(see Ref. 10, p. 786, 22.13.8-.9) so the last of equations (202) shows that $k = (2n+1)!/2^n n!(-1)^{n/2}$. We thus have

$$F(t) = kP_n(t) = \frac{(2n+1)!}{2^n n!(-1)^{n/2}} \cdot \frac{(2n)!}{2^n (n!)^2}\left[t^n - \frac{n(n-1)}{2(2n-1)}t^{n-2} + \ldots\right]$$
$$= \frac{(-1)^{3n/2}}{n!} Y_n[t^n + \ldots]$$

on using an explicit form for $P_n(t)$ (see Ref. 10, p. 775, 22.3.8) and on recalling the definition (201). Comparing coefficients of t^n we have now established that

$$\frac{1}{Y_n} = \frac{|a_{ij}|_n}{|a_{ij}|_{n-1}} = \frac{2^{2n}(n!)^2}{(2n+1)!(2n)!}. \tag{203}$$

It is not difficult to obtain this result by direct evaluation of $|a_{ij}|_n$ which is a product of Cauchy determinants.

We use a similar technique to evaluate the second factor in (192). The equations

$$\sum_{j=0}^{n} h_{i+j}Z_j = \delta_{in}, \quad i = 0,1,\ldots,n \tag{204}$$

yield

$$Z_n = \frac{|h_{i+j}|_{n-1}}{|h_{i+j}|_n}. \tag{205}$$

Using the definition (197) of h_{i+j}, (205) becomes

$$\sum_{x=0}^{N-1} x^i G(x) = \delta_{in}, \quad i = 0,1,\ldots,n \tag{206}$$

where we have written

$$G(x) = \sum_{j=0}^{n} Z_j x^j. \tag{207}$$

Thus we seek an nth degree polynomial $G(x)$ satisfying (206). The coefficient of x^n will give the desired ratio (205).

The Tchebyshev polynomial $t_n(x)$ (see Ref. 9, vol. 2, pp. 221–223) has just the properties sought. It satisfies

$$\sum_{x=0}^{N-1} x^i t_n(x) = 0, \quad i = 0,1,\ldots,n-1. \tag{208}$$

An explicit formula for the polynomial is

$$t_0(x) \equiv 1$$

$$t_n(x) \equiv n!\Delta^n[(x)_n(x-N)_n], \quad n = 1, \ldots, N-1 \tag{209}$$

where we write $(x)_n \equiv x(x-1)(x-2)\ldots(x-n+1)$ and define the forward difference operator Δ by $\Delta f(x) = f(x+1) - f(x)$ and $\Delta^n f(x) = \Delta[\Delta^{n-1} f(x)]$. The polynomials satisfy the recurrence

$$(n+1)t_{n+1}(x) - (2n+1)(2x-N+1)t_n(x) + n(N^2-n^2)t_{n-1}(x) = 0 \tag{210}$$

$$n = 1,2,\ldots,N-2$$

(Ref. 9, vol. 2, p. 223, (6)).

From (210) by using (208) we can easily calculate $\Sigma_{x=0}^{N-1} x^n t_n(x) \equiv S_n$. To do so, multiply (210) by x^{n-1} and sum. Recalling (208), one finds $2(2n+1)S_n = n(N^2-n^2)S_{n-1}$. Since $S_0 = N$, we have

$$S_n = \frac{Nn! \prod_{k=1}^{n} (N^2-n^2)}{2^n 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n+1)}. \tag{211}$$

It follows then that

$$G(x) = \frac{1}{S_n} t_n(x) \tag{212}$$

is the nth degree polynomial satisfying (206).

We now seek the coefficient of x^n in $G(x)$. The coefficient of x^n in $t_n(x)$ is not evident from the definition (209). However, it is easy to show that $\Delta(x)_n = n(x)_{n-1}$ and that

$$\Delta^n[f(x)g(x)] = \sum_{j=0}^{n} \binom{n}{j} \Delta^j f(x) \Delta^{n-j} g(x+j).$$

Applying these rules to (209) we obtain the alternative expression

$$t_n(x) = \sum_{j=0}^{n} \binom{n}{j}^2 (x)_{n-j}(x-N+j)_j. \tag{213}$$

It follows then that the coefficient of x^n in $t_n(x)$ is

$$k \equiv \sum_{j=0}^{n} \binom{n}{j}^2 = \binom{2n}{n} \tag{214}$$

(see Ref. 13, p. 4, 0.157-1). From (207), (212) and (213) it follows that Z_n

$= k/S_n$. From (205), (211) and (214), then one has

$$\frac{|h_{i+j}|_n}{|h_{i+j}|_{n-1}} = \frac{1}{Z_n} = \frac{\prod_{j=-n}^{n}(N-j)}{(2n+1)\binom{2n}{n}^2}. \tag{215}$$

This result combined with (203), (191), (192) and (194) yields (64)–(66).

Since $|h_{i+j}|_0 = N$, (215) yields

$$|h_{i+j}|_n = N^{n+1}\prod_{j=1}^{n}\frac{(j!)^4(N^2-j^2)^{n+1-j}}{(2j+1)[(2j)!]^2}, \tag{216}$$

a formula that seems difficult to arrive at by direct manipulation of the determinant.

APPENDIX A

Asymptotic Behavior of η_0

We here investigate the product (93) for large N. We adopt the abbreviation $W' = W_0T_0$.

Suppose first that $0 < W' < \frac{1}{2}$. We write (93) in the form

$$\frac{1}{N}\log\eta_0 = P_1 + P_2 + P_3 \tag{217}$$

where

$$\begin{aligned} P_1 &= \frac{1}{N}\log\left(\frac{\sigma^2}{2W'}\lambda_0(N+1,W')\right) \\ P_2 &= \frac{1}{N}\sum_{k=0}^{2W'N-2}\log\frac{\lambda_{k+1}(N+1,W')}{\lambda_k(N,W')} \\ P_3 &= \frac{1}{N}\sum_{k=2W'N-1}^{N-1}\log\frac{\lambda_{k+1}(N+1,W')}{\lambda_k(N,W')}. \end{aligned} \tag{218}$$

In this last sum set $k = N-1-\ell$ and use (13) to obtain

$$\begin{aligned} P_3 &= \frac{1}{N}\sum_{\ell=0}^{N(1-2W')}\log\frac{\lambda_{N-\ell}(N+1,W')}{\lambda_{N-1-\ell}(N,W')} \\ &= \frac{1}{N}\sum_{\ell=0}^{2N\overline{W}}\log\frac{1-\lambda_\ell(N+1,\overline{W})}{1-\lambda_\ell(N,\overline{W})} \end{aligned} \tag{219}$$

where we have written $\overline{W} = \frac{1}{2} - W'$. Now from (59), if $\ell = sN$, with s fixed and $0 < s < 2\overline{W}$, $1-\lambda_\ell(N,\overline{W}) \sim \exp[-\frac{1}{2}C(B,N)L_4(B) - NL_3(B)]$ where C, L_4 and L_3 are given by (45) and (47) and B is determined as a

function of s by (43), namely

$$\frac{1}{\pi}\int_{B(s)}^{1}\sqrt{\frac{\xi - B(s)}{(\xi - A)(1-\xi^2)}}\,d\xi = s. \tag{220}$$

In these formulas we now have $A = \cos 2\pi\overline{W}$.

For large N, then, and fixed s, a term in (219) takes the value

$$J \equiv \log\frac{1-\lambda_\ell(N+1,\overline{W})}{1-\lambda_\ell(N,\overline{W})} \sim -\frac{1}{2}C(\hat{B},N+1)L_4(\hat{B}) - (N+1)L_3(\hat{B}) - \left[-\frac{1}{2}C(B,N)L_4(B) - NL_3(B)\right] \tag{221}$$

where

$$\frac{1}{\pi}\int_{\hat{B}}^{1}\sqrt{\frac{\xi-\hat{B}}{(\xi-A)(1-\xi^2)}}\,d\xi = \frac{k}{N+1} = \frac{k}{N} - \frac{k}{N(N+1)} = s - \frac{s}{N+1}\cdot \tag{222}$$

Thus

$$\hat{B} = B\left(s - \frac{s}{N+1}\right) = B(s) - sB'(s)\frac{1}{N} + 0\left(\frac{1}{N^2}\right)\cdot$$

Straightforward Taylor expansions of quantities on the right of (221) now give

$$J = -\left[L_3(B(s)) - s\frac{d}{ds}L_3(B(s))\right] + o(1)$$

Returning to (219), we have

$$P_3 \sim -\frac{1}{N}\sum_{\ell=0}^{2N\overline{W}}\left[L_3\left(B\left(\frac{\ell}{N}\right)\right) - \frac{\ell}{N}\frac{d}{ds}L_3(B(s))\Big|_{s=\ell/N}\right] + o(1)$$

so that

$$\begin{aligned}\lim_{N\to\infty} P_3 &= -\int_0^{2\overline{W}}\left[L_3(B(s)) - s\frac{d}{ds}L_3(B(s))\right]ds \\ &= -\int_0^{2\overline{W}} L_3 ds + sL_3\Big|_{s=0}^{2\overline{W}} - \int_0^{2\overline{W}} L_3 ds \\ &= -2\int_0^{2\overline{W}} L_3(B(s))ds. \end{aligned} \tag{223}$$

Now recalling (220) and the definition (47) of L_3, we have

$$\lim_{N\to\infty} P_3 = -2\int_0^{2\overline{W}} ds \int_A^{B(s)} \sqrt{\frac{B-\xi}{(\xi-A)(1-\xi^2)}}\, d\xi$$

$$= -\frac{1}{\pi}\int_A^1 dB \int_A^B d\xi \sqrt{\frac{B-\xi}{(\xi-A)(1-\xi^2)}}$$

$$\int_B^1 \frac{dt}{\sqrt{(t-B)(t-A)(1-t^2)}} = 2\log(\sin \pi W_0 T_0). \quad (224)$$

Details of the evaluation of the triple integral to obtain the last line here are given in Appendix B.

It is not difficult to see that both P_1 and P_2 approach zero as N increases. We omit the demonstration here. Our result thus far reads

$$\lim_{N\to\infty} \frac{\log \eta_0}{N} = 2\log(\sin \pi W_0 T_0), \quad 0 < W_0 T_0 < \tfrac{1}{2}$$

which is (94).

To study η_0 when $W' = W_0 T_0 > \frac{1}{2}$, we return to (18) and consider it now for arbitrary values of W. Since

$$\frac{\sin 2\pi\left(W+\frac{1}{2}\right)(n-m)}{\pi(n-m)} = \left[\frac{\sin 2\pi W(n-m)}{\pi(n-m)} + \delta_{nm}\right](-1)^{n-m}, \quad (225)$$

the eigenvalue equation

$$\sum_{m=0}^{N-1} \frac{\sin 2\pi\left(W+\frac{1}{2}\right)(n-m)}{\pi(n-m)} v_m^{(k)}\left(N, W+\frac{1}{2}\right)$$

$$= \lambda_k\left(N, W+\frac{1}{2}\right) v_n^{(k)}\left(N, W+\frac{1}{2}\right)$$

can be rewritten on direct substitution of (225) as

$$\sum_{m=0}^{N-1} \frac{\sin 2\pi W(n-m)}{\pi(n-m)}\left[(-1)^m v_m^{(k)}\left(N, W+\frac{1}{2}\right)\right]$$

$$= \left[-1+\lambda_k\left(N, W+\frac{1}{2}\right)\right]\left[(-1)^n v_n^{(k)}\left(N, W+\frac{1}{2}\right)\right].$$

Comparison with (18) now shows that

$$\lambda_k\left(N, W+\frac{1}{2}\right) = 1 + \lambda_k(N, W)$$

$$v_m^{(k)}\left(N, W+\frac{1}{2}\right) = C(-1)^m v_m^{(k)}(N, W). \quad (226)$$

Suppose now that

$$\frac{n}{2} < W' < \frac{n+1}{2} \tag{227}$$

where n is a positive integer. Then in virtue of (226) Eq. (93) becomes

$$\eta_0 = \frac{\sigma^2}{2W'}[n + \lambda_0(N+1,W'')] \prod_0^{N-1} \frac{n + \lambda_{k+1}(N+1,W'')}{n + \lambda_k(N,W'')} \tag{228}$$

$$W'' \equiv W' - \frac{n}{2}$$

$$0 < W'' < \frac{1}{2}. \tag{229}$$

Then

$$\log \eta_0 = Q_1 + Q_2 \tag{230}$$

where

$$Q_1 \equiv \log \frac{\sigma^2[n + \lambda_0(N+1,W'')]}{2W'} \sim \log \frac{\sigma^2(n+1)}{2W'} \tag{231}$$

and

$$Q_2 \equiv \sum_0^{N-1} \log \frac{n + \lambda_{k+1}(N+1,W'')}{n + \lambda_k(N,W'')}. \tag{232}$$

When N is large only the terms for k near $2W''N$ contribute significantly to Q_2, for if $k = 2W''N(1-\epsilon)$, λ_k approaches 1 exponentially, while if $k = 2W''N(1+\epsilon)$, λ_k approaches zero exponentially. In either event, the summand in (232) approaches zero exponentially while the number of term grows linearly with N.

Consider now

$$H(\bar{\alpha},N) = \sum_{k=2WN-(\bar{\alpha}/\pi)\log N}^{2WN+(\bar{\alpha}/\pi)\log N} \log \frac{n + \lambda_{k+1}(N+1,W'')}{n + \lambda_k(N,W'')} \tag{233}$$

where $\bar{\alpha}$ is an arbitrary positive real number. We change from the variables N, k to new variables Δ and b via the transformation

$$\Delta = \frac{\pi}{\log N} \qquad N = e^{\pi/\Delta}$$

$$b = (k - 2W''N)\Delta \qquad k = \frac{b}{\Delta} + 2W''e^{\pi/\Delta} \tag{234}$$

and write

$$\lambda_k(N,W'') \equiv g(\Delta,b). \tag{235}$$

Then $\lambda_{k+1}(N+1,W'') = g(\Delta',b')$ where

$$\Delta' = \frac{\pi}{\log(N+1)} = \Delta + 0\left(\frac{\Delta^2}{N}\right)$$

and

$$b' = [k+1-2W''(N+1)]\Delta' = b + (1-2W'')\Delta + 0\left(\frac{\Delta}{N}\right).$$

Thus

$$g(\Delta',b') = g(\Delta,b) + (1-2W'')\Delta\frac{\partial g(\Delta,b)}{\partial b} + 0\left(\frac{\Delta}{N}\right)$$

and the summand of (233) becomes

$$\log\frac{n+\lambda_{k+1}(N+1,W'')}{n+\lambda_k(N,W'')} = \log\frac{n+g(\Delta,b)+(1-2W'')\Delta\dfrac{\partial g}{\partial b}+0\left(\dfrac{\Delta}{N}\right)}{n+g(\Delta,b)}$$

$$= (1-2W'')\Delta L(\Delta,b) + 0\left(\frac{\Delta}{N}\right)$$

where

$$L(\Delta,b) \equiv \frac{d}{db}\log[n+g(\Delta,b)]. \tag{236}$$

Now write $j = k - 2W''N$. Equation (233) becomes

$$H(\bar{\alpha},N) = (1-2W'')\sum_{j=-\bar{\alpha}/\Delta}^{\bar{\alpha}/\Delta}\left[L(\Delta,j\Delta)\Delta + 0\left(\frac{\Delta}{N}\right)\right].$$

In the limit of large N, we have

$$H(\bar{\alpha},N) \sim (1-2W'')\int_{-\bar{\alpha}}^{\bar{\alpha}} L(0,x)dx$$

$$= (1-2W)\log[n+g(0,x)]\Big|_{x=-\bar{\alpha}}^{\bar{\alpha}} \tag{237}$$

from (236). But (60) and (235) show that

$$g(0,b) = \frac{1}{1+e^{\pi b}}$$

and so

$$H(\bar{\alpha},N) \to (1-2W'')\log\frac{n+[1+e^{\pi\bar{\alpha}}]^{-1}}{n+[1+e^{-\pi\bar{\alpha}}]^{-1}}.$$

Finally, since $\bar{\alpha}$ can be chosen arbitrarily large,

$$H(\bar{\alpha},N) \to (1-2W'')\log\frac{n}{n+1} = -\log\left(1+\frac{1}{n}\right)^{(1-2W'')}. \tag{238}$$

Thus if

$$\lim_{N\to\infty} Q_2 = \lim_{N\to\infty} H(\bar{\alpha},N),$$

we can write

$$\eta_0 \sim \eta_\infty = \frac{\sigma^2(n+1)}{2W'} \frac{1}{\left(1+\frac{1}{n}\right)^{1-2W''}} \tag{239}$$

in virtue of (229)–(231). This result is (95).

APPENDIX B

Evaluation of Integral (224)

$$J \equiv \lim_{N\to\infty} P_3 = -\frac{1}{\pi}\int_A^1 dB \int_A^B d\xi \sqrt{\frac{B-\xi}{(\xi-A)(1-\xi^2)}}$$
$$\times \int_B^1 \frac{dt}{\sqrt{(t-B)(t-A)(1-t^2)}} = -\frac{1}{\pi}\int_A^1 \frac{dt}{\sqrt{(t-A)(1-t^2)}}$$
$$\times \int_A^t \frac{d\xi}{\sqrt{(\xi-A)(1-\xi^2)}} \int_\xi^t dB \sqrt{\frac{B-\xi}{t-B}}$$
$$= -\frac{1}{2}\int_A^1 dt \int_A^t d\xi \frac{t-\xi}{\sqrt{(1-\xi^2)(1-t^2)(A-t)(A-\xi)}} \tag{240}$$

since under the substitution $u^2 = (B-\xi)/(t-B)$

$$\int_\xi^t dB \sqrt{\frac{B-\xi}{t-B}} = (t-\xi)\int_0^\infty u \frac{2u}{(1+u^2)^2} du$$
$$= (t-\xi)\left[-u\frac{1}{1+u^2}\Big|_0^\infty + \int_0^\infty \frac{du}{1+u^2}\right] = (t-\xi)\frac{\pi}{2}.$$

Now change from the variables of integration ξ,t to α,β via

$$\xi = \frac{1+A}{2} + \frac{1-A}{2}\sin\alpha$$

$$t = \frac{1+A}{2} + \frac{1-A}{2}\sin\beta$$

and obtain

$$J = -\frac{1}{2}\int_{-\pi/2}^{\pi/2} d\beta \int_{-\pi/2}^{\beta} d\alpha \frac{\sin\beta - \sin\alpha}{\sqrt{(x+\sin\alpha)(x+\sin\beta)}}$$

where we write

$$x = \frac{3+A}{1-A}, 1 < x < \infty.$$

Changing now to variables ϕ,ψ through the 90° rotation $\frac{1}{2}(\alpha+\beta)=\phi$, $\frac{1}{2}(\alpha-\beta)=\psi$, we find that

$$J = 2\int_{-\pi/2}^{0} d\psi \int_{-\psi-\pi/2}^{\psi+\pi/2} d\phi \frac{\cos\phi\sin\psi}{\sqrt{x^2-\sin^2\psi+2x\cos\psi\sin\phi+\sin^2\phi}}$$

$$= 2\int_{-\pi/2}^{0} d\psi \sin\psi \left[\log|x\cos\psi + \sin\phi\sqrt{x^2-\sin^2\psi+2x\cos\psi\sin\phi+\sin^2\phi}\,\right]_{\phi=-(\psi+\pi/2)}^{\psi+\pi/2}$$

$$= 2\int_{-\pi/2}^{0} d\psi \sin\psi \times \log\left|\frac{x\cos\psi+\cos\psi+\sqrt{x^2-\sin^2\psi+2x\cos^2\psi+\cos^2\psi}}{x\cos\psi-\cos\psi+\sqrt{x^2-\sin^2\psi-2x\cos^2\psi+\cos^2\psi}}\right|$$

$$= -2\int_0^1 du \log\frac{a[au+\sqrt{b^2+u^2}]}{b[bu+\sqrt{a^2-u^2}]}$$

where we have set $\cos\psi = u$ and

$$a=\sqrt{\frac{x+1}{2}},\quad b=\sqrt{\frac{x-1}{2}},\quad a^2-b^2=1. \tag{241}$$

Thus

$$J = -2\log\frac{a}{b} - 2\int_0^1 du \log[au+\sqrt{b^2+u^2}] + 2\int_0^1 du \log[bu+\sqrt{a^2-u^2}]. \tag{242}$$

Now it can be verified by direct differentiation that when (241) holds

$$\int du \log[au+\sqrt{b^2+u^2}] = (u+1)\log[au+\sqrt{b^2+u^2}] - u - \log[u+b^2+a\sqrt{b^2+u^2}]$$

and

$$\int du \log[bu+\sqrt{a^2-u^2}] = (u+1)\log[bu+\sqrt{a^2-u^2}] - u - \log[a^2-u+b\sqrt{a^2-u^2}].$$

It then follows readily that the last two terms of (242) both have magnitude $\log 2(a+b)-1$, and that equation becomes simply

$$J = -2\log\frac{a}{b} = -2\log\sqrt{\frac{x+1}{x-1}} = 2\log\sqrt{\frac{1+A}{2}}. \tag{243}$$

When $A=\cos 2\pi\overline{W}=\cos 2\pi(\frac{1}{2}-W_0T)=-\cos 2\pi W_0T$, (243) yields (224).

The simple result (243) for the integral given by the last member of

(240) was first obtained by J. A. Morrison by a route involving elliptic functions and identities among them.

APPENDIX C

Commutation of L and M

Let operators

$$L \equiv \int_a^b df' K(f,f')$$

and

$$M = \frac{d}{df} p(f) \frac{d}{df} + q(f)$$

be given. Then

$$MLg = \int_a^b df' M_f K(f,f') g(f')$$

while

$$LMg = \int_a^b df' K(f,f') M_{f'} g(f')$$

$$= \left[p(f') \left\{ K \frac{dg(f')}{df'} - \frac{\partial K(f,f')}{\partial f'} g(f') \right\} \right]_{f'=a}^{b} + \int_a^b df' g(f') M_{f'} K(f,f').$$

Now if

$$p(a) = p(b) = 0, \tag{244}$$

we have

$$MLg - LMg = \int_a^b df' g(f') [M_f - M_{f'}] K(f,f')$$

and so, if in addition

$$M_f K(f,f') \equiv M_{f'} K(f,f'), \tag{245}$$

the operators commute. In the special case when $K(f,f') = K(|f - f'|)$, (245) becomes

$$[p(f) - p(f')] \frac{\partial^2 K(|f - f'|)}{\partial f^2} + [p'(f) + p'(f')] \frac{\partial K(|f - f'|)}{\partial f} + [q(f) - q(f')] K(|f - f'|) \equiv 0. \tag{246}$$

Applying this to the operators of (110) and (111) we have $p(f) = (1/4\pi^2)[\cos 2\pi f - A]$, $q(f) = 1/4(N^2 - 1)\cos 2\pi f$ and $a = -b = W$. It is seen that (244) is satisfied. To verify (246) observe that

$$p(f) - p(f') = -\frac{1}{2\pi^2}\sin \pi(f + f')\sin \pi(f - f')$$

$$p'(f) + p'(f') = -\frac{1}{\pi}\sin \pi(f + f')\cos \pi(f - f')$$

$$q(f) - q(f') = \frac{1}{2}(N^2 - 1)\sin \pi(f + f')\sin \pi(f - f').$$

Thus every term of (246) in the present case contains a factor $\sin \pi(f + f')$. This equation then is equivalent to

$$\sin t \frac{d^2}{dt^2}\frac{\sin Nt}{\sin t} + 2\cos t \frac{d}{dt}\frac{\sin Nt}{\sin t} + (N^2 - 1)\sin t \frac{\sin Nt}{\sin t} = 0 \quad (247)$$

where we have written $t \equiv \pi(f - f')$. The reader can readily verify that (247) holds identically in t.

REFERENCES

1. D. Slepian and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—I," B.S.T.J., *40*, No. 1 (January 1961), pp. 43–64.
2. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—II," B.S.T.J., *40*, No. 1 (January 1961), pp. 65–84.
3. H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—III," B.S.T.J., *41*, No. 4 (July 1962), pp. 1295–1336.
4. D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—IV," B.S.T.J., *43*, No. 6 (November 1964), pp. 3009–3058.
5. D. Slepian, "Some Asymptotic Expansions for Prolate Spheroidal Wave Functions," J. Math. and Phys., *44*, No. 2 (June 1965), pp. 99–140.
6. D. Slepian and E. Sonnenblick, "Eigenvalues Associated with Prolate Spheroidal Wave Functions of Zero Order," B.S.T.J., 44, No. 8 (October 1965), pp. 1745–1760.
7. D. Slepian, "On Bandwidth," Proc. IEEE, 64, No. 3 (March 1976), pp. 292–300.
8. E. N. Gilbert and D. Slepian, "Doubly Orthogonal Concentrated Polynomials," SIAM J. Appl. Math., *8*, No. 2 (April 1977), pp. 290–319.
9. A. Erdélyi, *Higher Transcendental Functions*, New York: McGraw-Hill, 1953.
10. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover Publications, 1965.
11. Harald Cramér, *Mathematical Methods of Statistics*, Princeton, New Jersey: Princeton University Press, 1946.
12. A. Papoulis and M. S. Bertran, "Digital Filtering and Prolate Functions," IEEE Trans. Circuit Theory, *CT-19*, No. 6 (November 1972), pp. 674–681.
13. I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products*, New York: Academic Press, 1965.
14. P. M. Morse and H. Feshback, "Methods of Theoretical Physics," New York: McGraw-Hill, 1953.
15. J. Von Neumann, *Mathematical Foundations of Quantum Mechanics*, Princeton, New Jersey: Princeton University Press, 1955, pp. 170–178.
16. D. W. Tufts and J. T. Francis, "Designing Digital Low-Pass Filters—Comparison of Some Methods and Criteria," IEEE Trans. Audio and Electroacoust., *AU-18*, No. 4 (Dec., 1970), pp. 487–494.
17. A. Eberhard, "An Optimum Discrete Window for the Calculation of Power Spectra," IEEE Trans. Audio and Electroacoust., *AU-21* (February 1973), pp. 37–43.

18. D. W. Tufts, "Comments on 'FIR Digital Filter Design Techniques Using Weighted Chebyshev Approximation'," Proc. IEEE, *63* (November 1975), p. 1618.
19. A. N. Balakrishnan, "Essentially Band-Limited Stochastic Processes," IEEE Trans. Infor. Theory, *IT-11* No. 1 (January 1965), pp. 154–156.
20. A. Viterbi, "On the Minimum Mean Square Error Resulting from Linear Filtering of Stationary Signals in White Noise," IEEE Trans. Inform. Theory, *IT-11*, No. 4 (October 1965), pp. 594–595.
21. D. Slepian, "A Numerical Method for Determining the Eigenvalues and Eigenfunctions of Analytic Kernels," SIAM J. Numerical Analysis, *5*, No. 3 (September 1968), pp. 586–600.
22. H. Widom, "The Strong Szegö Limit Theorem for Circular Arcs," Indiana Univ. Math. J., *21*, No. 3 (1971), pp. 277–283.
23. F. Gori and G. Guattari, "Degrees of Freedom of Images from Point-Like-Element Pupils," J. Optical Soc. Am., *64*, No. 4 (April 1974), pp. 453–458.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

An Earth-Space Propagation Measurement at Crawford Hill Using the 12-GHz CTS Satellite Beacon

By A. J. RUSTAKO, JR.

(Manuscript received September 26, 1977)

This paper describes a measurement of atmospheric attenuation and depolarization, primarily due to rain, of the 11.7-GHz cw beacon signal from the Communications Technology Satellite (CTS). This beacon source made possible the first fixed-path, nearly continuous measurements of earth-space propagation at this frequency. A measurement is being made at Bell Laboratories, Crawford Hill, in Holmdel, New Jersey using a fully steerable, 6-meter-aperture, horn-reflector antenna fitted with a dual-sense, circular-polarized feed. The amplitudes of the copolarized and cross-polarized components are measured with a two-branch, stable, narrowband, frequency tracking receiver. The receiving system, which was designed to run unattended, is described. Propagation data for a greater than 1-year period beginning April 1976 are presented. The attenuation data show that an outage time of $2^1/_2$ hours per year can be expected for a 10-dB rain fade margin. Significant anomalous depolarization effects not directly related to rainfall have been observed.

I. INTRODUCTION

The launch of the CTS satellite has provided a useful signal source for earth-space propagation experiments. The interest is in measuring long-term (approximately one year) atmospheric attenuation and depolarization, primarily due to rain. The satellite carries a cw beacon at a frequency of 11.7 GHz with a nominal output power of 200 mW. The satellite beacon antenna is a right-hand sense, circularly polarized, earth-coverage horn with an axial ratio of 1.5 dB within an enclosed beam

angle of 17 degrees. The satellite is in geosynchronous orbit and is located nominally at 116°W longitude with an inclination of less than 1 degree.

A disadvantage in using the CTS satellite beacon for these measurements is that the beacon is turned off during the spring and fall eclipse periods when the solar cell panels are shadowed by the earth. In the fall 1976 eclipse, the beacon was turned off completely for seven weeks. Fortunately, few significant rain events occurred during this time. In the spring 1977 eclipse, the beacon was turned off for only short periods, approximately one hour each day over a six-week period.

The rain attenuation data for these eclipse periods were bridged by statistically scaling data from a colocated 19-GHz COMSTAR A beacon measurement. This scaling technique is described in the appendix to this paper.

The Crawford Hill receiving system is located in Holmdel, New Jersey at approximately 74°W longitude and 40°N latitude. For this location, the receiving antenna point is nominally at an azimuth angle of 234 degrees with an elevation of 27 degrees. This provides approximately a 20-km path through atmospheric rain showers. The signal parameters for this path are shown in Table I. The clear air carrier-to-noise ratio in the copolarized signal branch is 43.4 dB in a 32-Hz bandwidth. The cross-polarization signal level for clear air is typically 33 dB below the copolarized signal level. A 28.5-dB carrier-to-noise ratio in the cross-polarized signal branch is obtained by detection in a 0.5-Hz bandwidth.

The Crawford Hill receiver provides a continuous measure of both the copolarized received signal and the cross-polarized component due to rain or other atmospheric effects.[1] The entire system is normally unattended and takes data continuously on a 24-hour basis.

II. ANTENNA AND TRACKING

The receiving antenna is a fully steerable horn reflector[2] with a 6-meter aperture. Figure 1 is a photograph of the antenna. The antenna feed structure, receiver, and recording system are located within the vertex cab of the horn. The antenna has a hybrid-coupled, orthogonal feed to provide two isolated signal signal components to the receiver. One port provides the right-hand circular copolarized component and the second provides the left-hand cross-polarized component. The measured isolation of these two output ports using the satellite beacon is greater than 33 dB on axis for clear-air propagation conditions.

The receiving antenna has a gain of 57 dB and a 3-dB beamwidth of approximately 0.3 degree. The narrow beamwidth necessitates continual antenna pointing to track the diurnal satellite motion. The antenna is pointed by an open loop technique[3] using predicted azimuth and ele-

Table I — Path parameters for CTS satellite

Beacon transmitter power ≧ 200 mW	≧ −7.0 dbW
Antenna gain toward Crawford Hill	= +18.4 dB
ERP	= +11.4 dBW
Path loss (38982 km) to Crawford Hill	= 205.6 dB
Clean air attenuation	= 0.3 dB
Net loss	= 205.9 dB
Horn reflector antenna gain	= 57 dB
Received signal power	= −137.5 dBW
	= −107.5 dBm
Ground Receiver	
Receiver T_E = 1150K, kTF	= −166 dBm/Hz
Copolarized received signal C/N density	= 58.5 dB/Hz
Copolarized signal C/N ratio in 32 Hz bandwidth (clean air)	= 43.4 dB
Residual cross-polarized level below copolarized received signal (clear air)	= −33 dB
Cross-polarized signal C/N ratio in 0.5 Hz bandwidth (clear air)	= 28.5 dB

vation data supplied by NASA. The overall antenna servo pointing accuracy is estimated to be better than ±0.03 degree.

III. RECEIVER DESIGN

The receiver is composed of two branches. One branch measures the amplitude of the copolarized signal, the second measures the relatively weak cross-polarized signal. The copolarized signal branch is used to

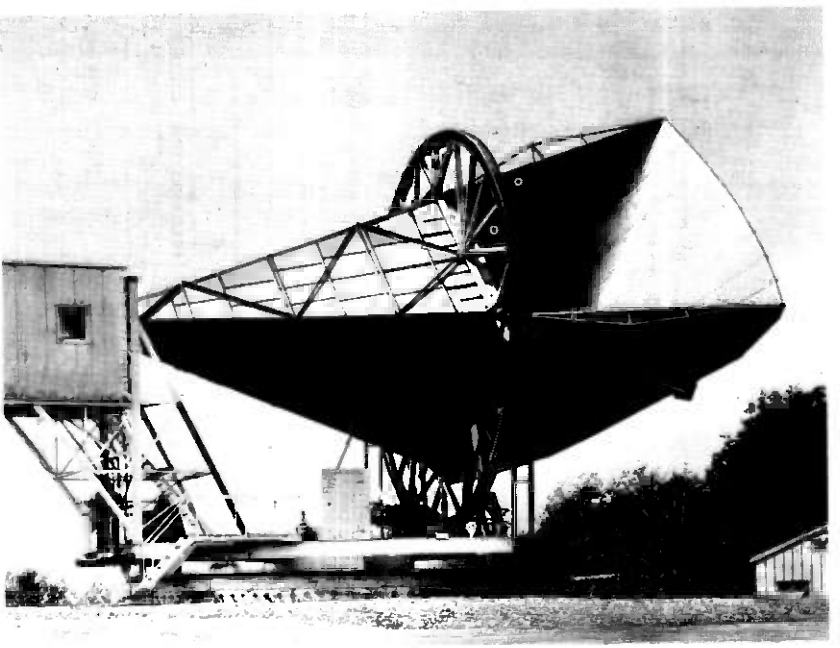


Fig. 1—Six-meter horn reflector antenna.

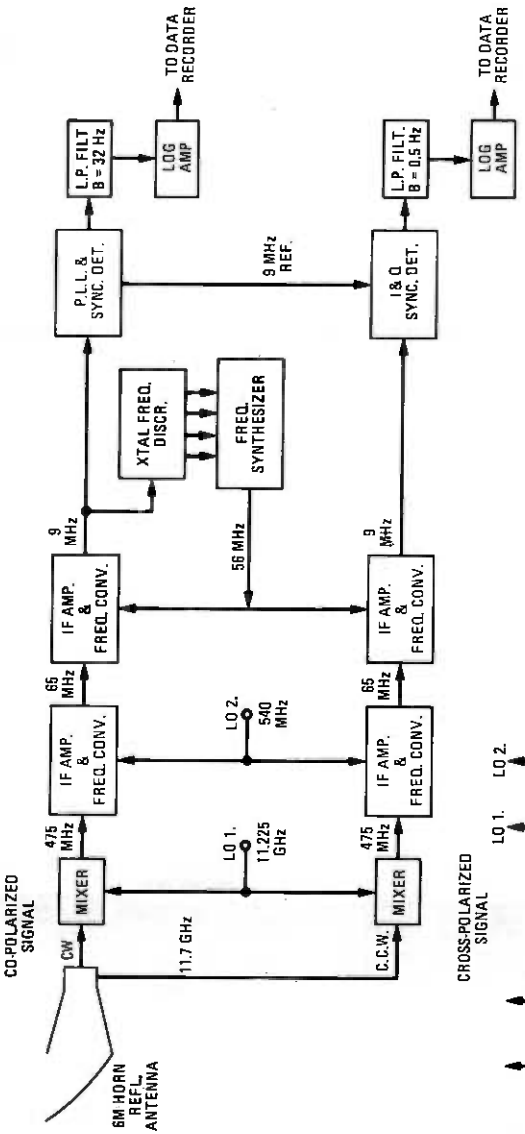


Fig. 2—CTS beacon measuring receiver.

control frequency tracking and provide a phase reference for detection in the cross-polarized signal branch. Figure 2 is a block diagram of the CTS beacon measuring receiver. Each receiver branch uses multiple heterodyning to convert the 11.7-GHz input frequency from the satellite to a final IF frequency of 9 MHz where envelope detection is carried out. The 9-MHz IF signal in each branch is bandpass-filtered through a 500-Hz wide crystal filter. The 9-MHz IF output from the copolarized signal branch is split, with one portion used for frequency control and the other envelope-detected. The frequency control component is envelope-limited to provide an input to a frequency discriminator. The second 9-MHz IF component is envelope-detected using a phase-locked synchronous detector. The detector output is low-pass filtered, its logarithm taken and recorded.

The phase-locked 9-MHz reference signal for the copolarized synchronous detector is split to provide a reference signal for in-phase and quadrature detection of the cross-polarized branch, 9-MHz IF signal. The outputs from these detectors are separately low-pass filtered and vectorially summed. The logarithm of the summed output is taken, then recorded.

A detailed block diagram of the receiver automatic frequency tracking control is shown in Fig. 3. The components in the control loop are an envelope limiter, a temperature-controlled crystal frequency discriminator, a gated integrator, a window comparator, a gated clocked up-down counter, and a BCD controlled frequency synthesizer with frequency doubler. The slow-frequency variation of the satellite beacon signal (~5 kHz/day) is tracked by measuring the change in the copolarized 9-MHz IF signal frequency and controlling the nominal 56-MHz local oscillator frequency to return the IF signal to exactly 9 MHz. A portion of the copolarized 9-MHz IF signal is envelope-limited to provide an input to a quartz crystal frequency discriminator. The discriminator output is integrated to smooth short-term variations and applied to the window voltage comparator. The window voltage comparator determines if the discriminator input frequency has changed beyond a preset allowable amount. The comparator gates a clock signal that either increases or decreases the count in a digital register. A BCD output from the digital register drives the frequency control of a programmable frequency synthesizer. The frequency synthesizer output at about 28 MHz is frequency-doubled to 56 MHz to provide a local oscillator for the 65-MHz to 9-MHz frequency conversion.

A clock frequency of 1 Hz is used to drive the digital register. With this clock frequency and the synthesizer frequency doubling, the receiver will track an input frequency rate of change of 2 Hz per second. This has been found adequate for the beacon frequency drift rates encountered.

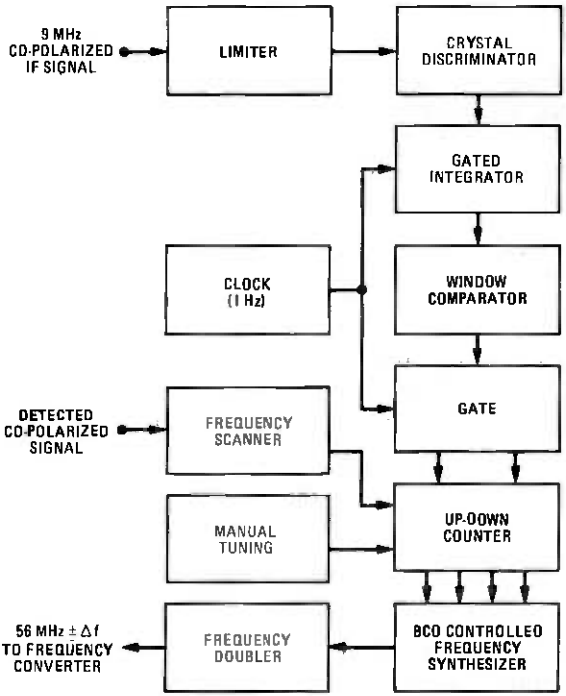


Fig. 3—Automatic frequency tracking control.

The use of the digital register-frequency synthesizer technique provides an automatic-frequency control with memory.[4] If the received beacon signal is lost because of a deep fade, beacon turnoff, or tracking malfunction, the receiver remains tuned to the beacon frequency preceding the loss and will not normally change frequency until the signal returns. The automatic frequency tracking control is provided with a frequency scan circuit to retune the receiver automatically if the beacon signal is lost and later appears outside the 500-Hz receiver bandwidth. This occurs during eclipse periods when the satellite power is shut down. The beacon oscillator cools and changes frequency by several kilohertz. The receiver frequency scan is delayed about 3 minutes on loss of signal, then proceeds to scan at a 100-Hz-per-second rate approximately 8 kHz either side of the last known beacon frequency. This scan technique has provided rapid reacquisition of the beacon signal during eclipse.

The overall frequency of the receiver is controlled by a single 10-MHz stable crystal oscillator in conjunction with several frequency multiplier chains. The crystal oscillator has a short-term (1 s) stability of 1×10^{11} and an aging rate of $<1.5 \times 10^{-7}$/year. Figure 4 is a block diagram of the oscillator and multiplier chains. The first local oscillator at 11.225 GHz

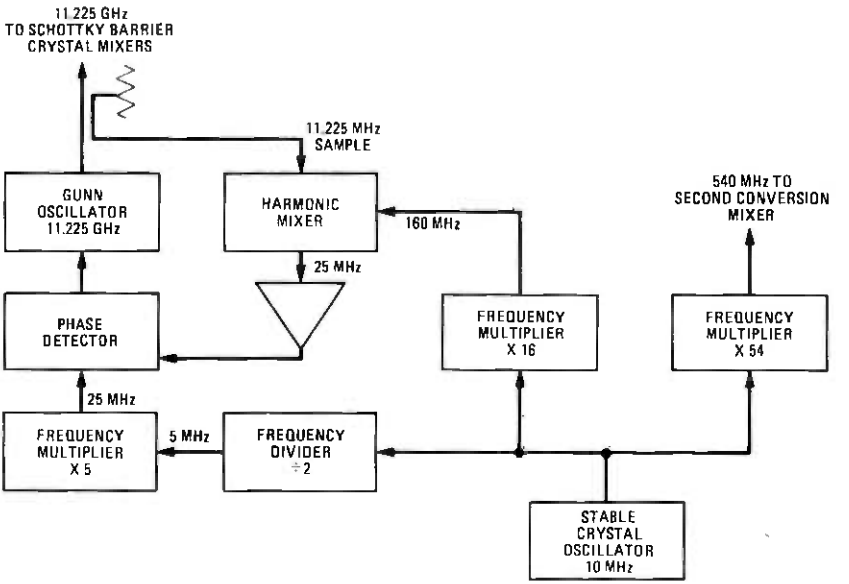


Fig. 4—Receiver local oscillator chain.

is obtained from a phase-locked, varactor-tuned, Gunn diode oscillator. The Gunn oscillator output is heterodyned with a stable reference signal at 160 MHz in a harmonic mixer. The frequency difference is limited and compared in phase with a second reference signal at 25 MHz. An error signal from the phase detector is fed back to the Gunn oscillator frequency control. The second local oscillator at 540 MHz is directly obtained from frequency multiplication of the stable crystal oscillator.

Figure 5 is a detailed block diagram of the synchronous envelope detectors in both receiver branches. A detection reference signal is provided by a 9-MHz, voltage-controlled crystal oscillator. This VCXO is phase-locked to the 9-MHz limiter output in the frequency control of the co-polarized branch. The 9-MHz IF output signal from the copolarized branch is synchronously envelope-detected and low-pass filtered. A bandwidth of 32 Hz is provided by an active 4-pole Butterworth low-pass filter. The signal envelope output from the low-pass filter is passed through a baseband log amplifier to the data recorders. The parameters of the log amplifier are set to provide a 40-dB dynamic range.

The 9-MHz IF output signal from the cross-polarized branch is envelope-detected by in-phase and quadrature detectors. The detected in-phase and quadrature components are low-pass filtered and vectorially summed to provide a single output proportional to the magnitude of the cross-polarized signal. A bandwidth of 0.5 Hz is provided by active 4-pole Butterworth low-pass filters. The combined output is passed

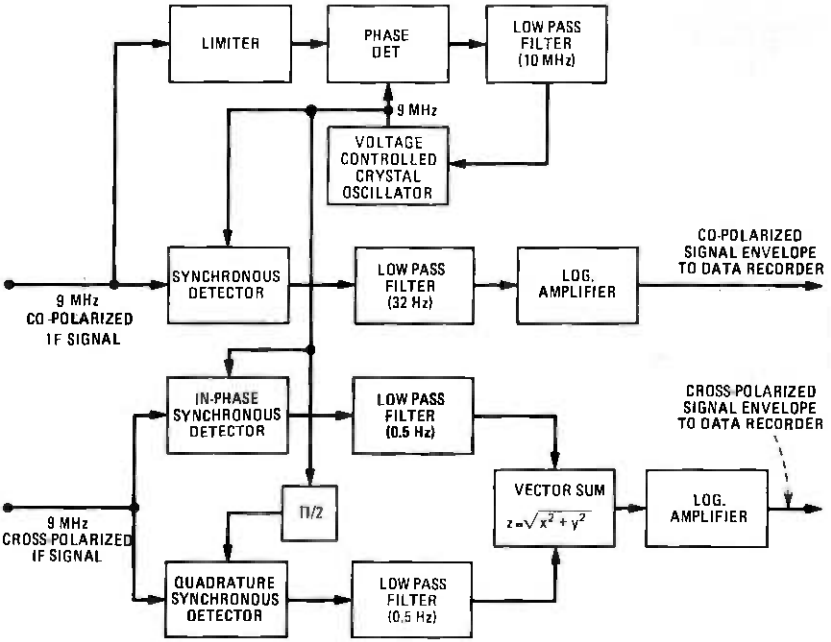


Fig. 5—Synchronous envelope detectors.

through a baseband log amplifier to the data recorders. The parameters of the log amplifier are again set to provide a 40-dB dynamic range.

The receiver and data recorder are calibrated by providing a stable input signal at 11.7 GHz to the receiver input mixers. This calibration source is obtained by frequency multiplication of a synthesized signal source which is referenced to the 10-MHz stable crystal oscillator. The calibration source was carefully shielded to prevent stray signal leakage to the receiver. The calibration source output, controlled by a precision waveguide attenuator, is coupled into waveguide switches at the receiver inputs.

IV. MEASUREMENT RESULTS

The beacon measuring system has been operational since April 1976. Nearly continuous measurements have been made of the coplarized and cross-polarized components from the satellite beacon signal. Data are presented for greater than a 1-year period beginning April 1976. During this time, some data were lost during the fall and spring eclipse periods when the satellite beacon was turned off to conserve battery power. A small amount of data was also lost due to an antenna servo malfunction. The copolarized signal data were analyzed for a one-year period to obtain attenuation statistics. Using data for attenuation greater than 1 dB, a

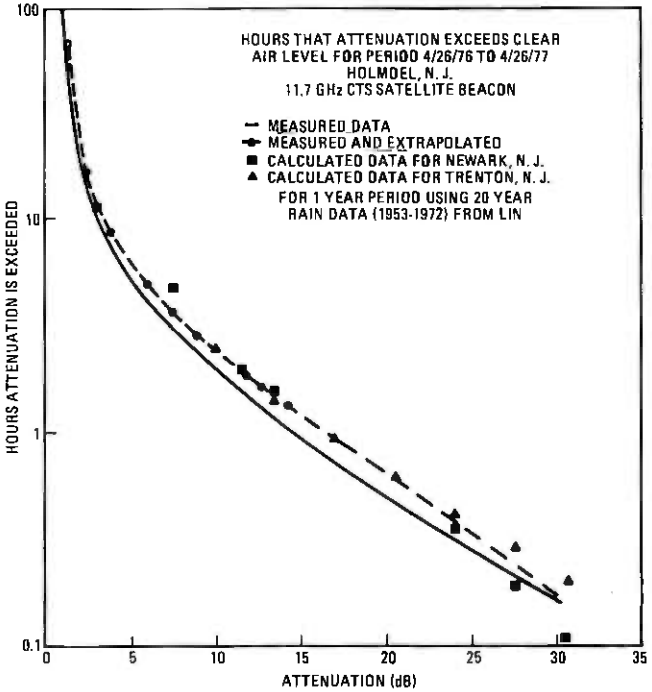


Fig. 6—Cumulative attenuation statistics for April 26, 1976 to April 26, 1977.

distribution curve showing the time an attenuation greater than clear air was exceeded versus attenuation is plotted as the solid line in Fig. 6. This curve shows data for the one-year period excluding eclipse outage time. A second curve, the broken line in Fig. 6, shows the effect of bridging the outage periods with data scaled from a simultaneous 19-GHz COMSTAR A beacon measurement. The frequency scaling, described in the appendix, relates measured attenuation statistics at the two frequencies during a common time to predict attenuation statistics at 12 GHz when the CTS beacon signal was lost. The 12-GHz statistics could be predicted up to approximately 14 dB corresponding to the 40-dB threshold in the COMSTAR A data. Beyond 14 dB, the 12-GHz distribution curve, shown dashed, asymptotically approaches the measured curve since heavy rain with attendant high attenuation did not occur during the bridged periods. In general, for practical system design, the attenuation statistics are of greatest interest for attenuation less than 14 dB.

The bridged curve represents a reasonably good estimate of the attenuation statistics for the full year. The curve therefore could be useful in estimating the outage time of a satellite communications link given

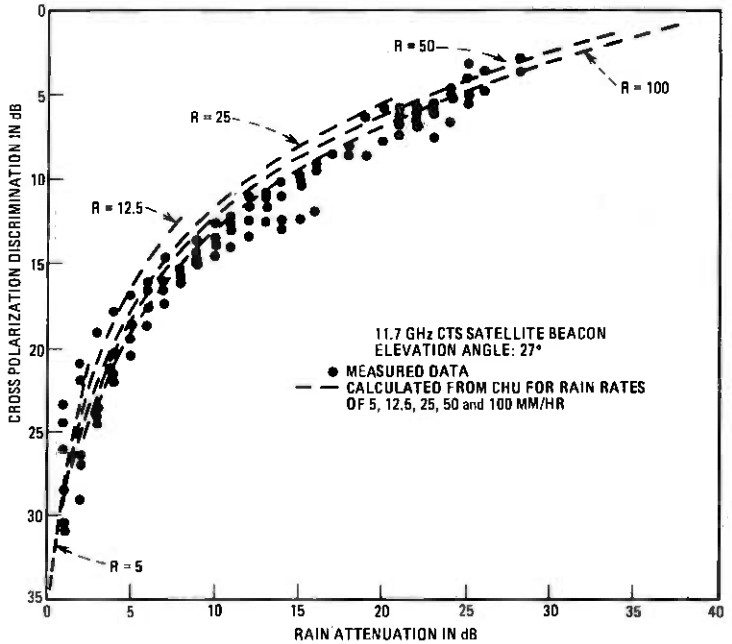


Fig. 7—Cross polarization discrimination as a function of rain attenuation.

an attenuation or fading margin. For example, a system with a 10-dB margin would have had a 2½-hour outage time over the year period.

The distribution curve shows a change in slope as attenuation increases. At low attenuation, the logarithmic change in outage time is very rapid with incremental decibel changes in attenuation. Above about 10-dB attenuation, the slope appears to be fairly constant, indicating there is a decreasing rate of return in outage time with an increase in attenuation or fading margin. Therefore, unless a much lower outage time is required for a particular service, a margin of about 10 dB may be optimum.

In addition to the experimental result, Fig. 6 shows several calculated points of an attenuation distribution obtained from Lin.[5] The calculation was based on five-minute point rain rate data for a 20-year period (from 1953 to 1972) for Newark and Trenton, New Jersey for a path elevation angle of 27 degrees. These points show the average number of hours per year that an attenuation is exceeded for a similar path at these two locations. Very close agreement can be observed between the measured and calculated data.

The cross-polarization data were analyzed for several rain events during the measurement period. The cross-polarization discrimination, defined as the decibel difference between the desired and undesired

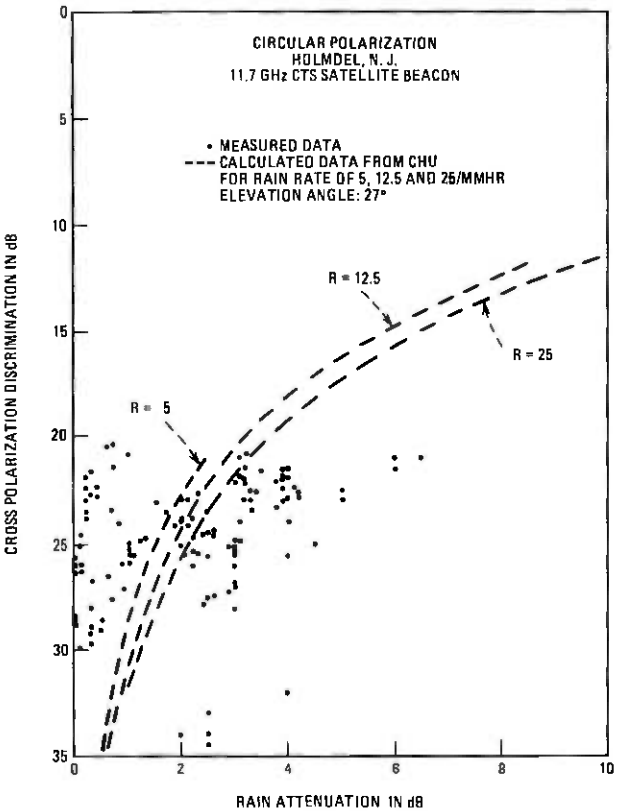


Fig. 8—Cross polarization discrimination as a function of rain attenuation.

polarization components, was measured as a function of rain attenuation. Data for two types of rain events are shown in Figs. 7 and 8. In Fig. 7, measured data shown as dots are for three events during August 1976 where high rain rates were encountered. The cross-polarization discrimination was measured for copolarized signal attenuations 1 dB or greater than clean air. The measured data show a decrease in discrimination with increasing path attenuation. For example, at a 10-dB attenuation the undesired cross-polarized components are approximately 14 dB below the copolarized signal and at 20-dB attenuation the discrimination has decreased to about 8 dB. In addition to the measured data, five calculated curves of cross-polarization discrimination are shown. These curves from Chu[6] were calculated from differential attenuation and phase shift through oblate spheroidal raindrops at 12 GHz for five rain rates. A maximum path length of 20 km, an elevation angle of 27 degrees, and a single uniform rain-canting angle were assumed. These curves represent an upper bound of the theoretical cross-polar-

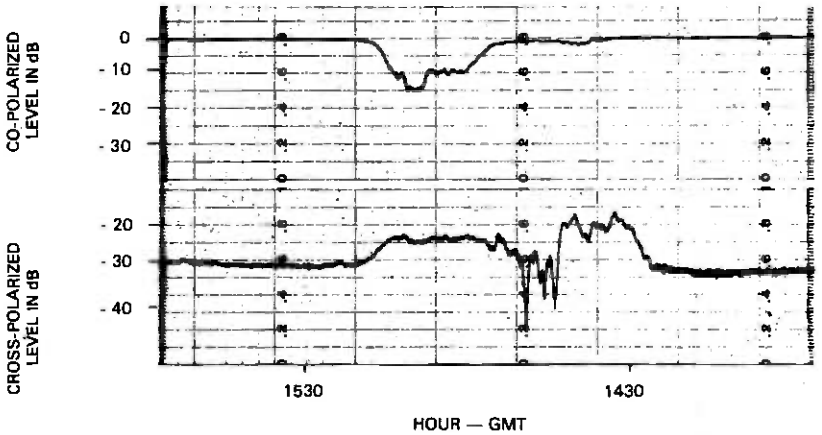


Fig. 9—Anomalous cross polarization event, 11.7 GHz CTS beacon, Holmdel, N.J., August 8, 1976.

ization discrimination. A reasonably good fit is shown between Chu's theory and the measured data for copolarized signal attenuation greater than 5 dB. The measured data generally fall below Chu's curves by a few dB, which could be caused by randomness in rain-canting angle as would be normally by expected with wind-driven rain.

Similar data are plotted in Fig. 8 for three rain events during April, May, and June 1977. The events of this period were of low rain rate compared to the August data and therefore the copolarized signal fading was small, less than 7 dB. The cross-polarization discrimination was measured for copolarized signal attenuation as low as 0.1 dB greater than clean air. The measured data show considerable scatter with little resemblance to Chu's curves. This scatter might be due to ice or atmospheric turbulence effects that are hidden in the more intense summer rains.

In addition to the normally expected strong depolarization effects with rain attenuation, some anomalous effects not predicted by contemporary rain propagation theory have been observed. One effect, which has been reported elsewhere,[1] is a large change in cross-polarization discrimination without an attendant measurable change in copolarization signal attenuation. A graphic example of one event is shown in Fig. 9. This segment of a pen recording shows a simultaneous time trace of the right-hand circular copolarized and the left-hand circular cross-polarized signal components. The decibel levels shown are relative to the clear air copolarized signal level. Significant variation in depolarization was observed in the vicinity of 1430 hours when the copolarized signal is at its clear air level. This effect has been observed both during summer rain events, often related to thunderstorms, and during the cooler months

with only an overcast sky, near-freezing temperature, and no precipitation. This effect is very likely caused by differential phase shift through ice crystals present in the propagation path.

A second observed anomalous depolarization effect has been rapid discrete steps in the cross-polarized signal during thunderstorms. This effect, shown in the cross-polarized trace in Fig. 10, is believed due to rapid orientation of ice crystals in the transmission path by lightning discharges near the path.[7]

Both these anomalous effects have been simultaneously observed over similar transmission paths using the COMSTAR A and B beacon signals at 19 and 28 GHz.

V. SUMMARY

A measurement of atmospheric attenuation and depolarization at 12 GHz over an earth-space propagation path has been described. The measurement system uses a two-branch, narrowband frequency tracking receiver in conjunction with a 6-meter-aperture, horn-reflector antenna to measure copolarized and cross-polarized components of the received cw beacon signal from the Communications Technology Satellite.

The measurement system has been operating essentially unattended since April 1976, recording the atmospheric attenuation and depolarization of the satellite beacon signal. Attenuation statistics showing a distribution of the number of hours the path attenuation exceeded the clear air level have been presented for a one-year period. The distribution was corrected for data loss when the 12-GHz beacon signal was unavailable during eclipse by bridging scaled statistical data from a colocated 19-GHz COMSTAR A beacon measurement. The distribution shows that a 12-GHz satellite communication link with an attenuation or fading margin of 10 dB would experience a $2\frac{1}{2}$-hour outage time in a one-year period.

Depolarization effects have been observed during normal rain fade events and when no measurable copolarized signal attenuation was present. The depolarization effects observed for rain attenuation greater than 5 dB appear to follow predictions based on differential attenuation and phase shift through oblate spheroidal rain drops. However, when no measurable or very low copolarized signal attenuation is present, the observed depolarization effects are very likely caused by differential phase shift through ice crystals.

VI. ACKNOWLEDGMENTS

I want to acknowledge useful discussion with H. W. Arnold, T. S. Chu, D. C. Cox, S. H. Lin, D. O. Reudink, and R. W. Wilson. I am also indebted to H. W. Arnold for providing his data analysis programs and the COM-

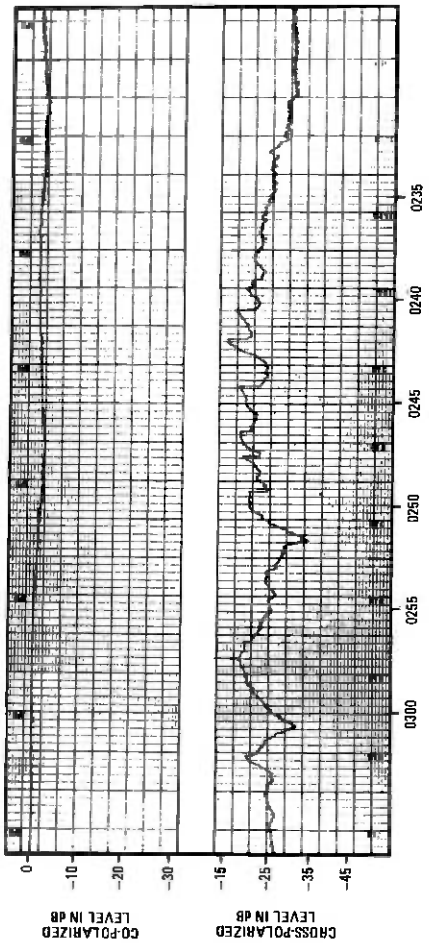


Fig. 10—Anomalous cross polarization event, 11.7 GHz CTS beacon, Holmdel, N.J., April 25, 1977.

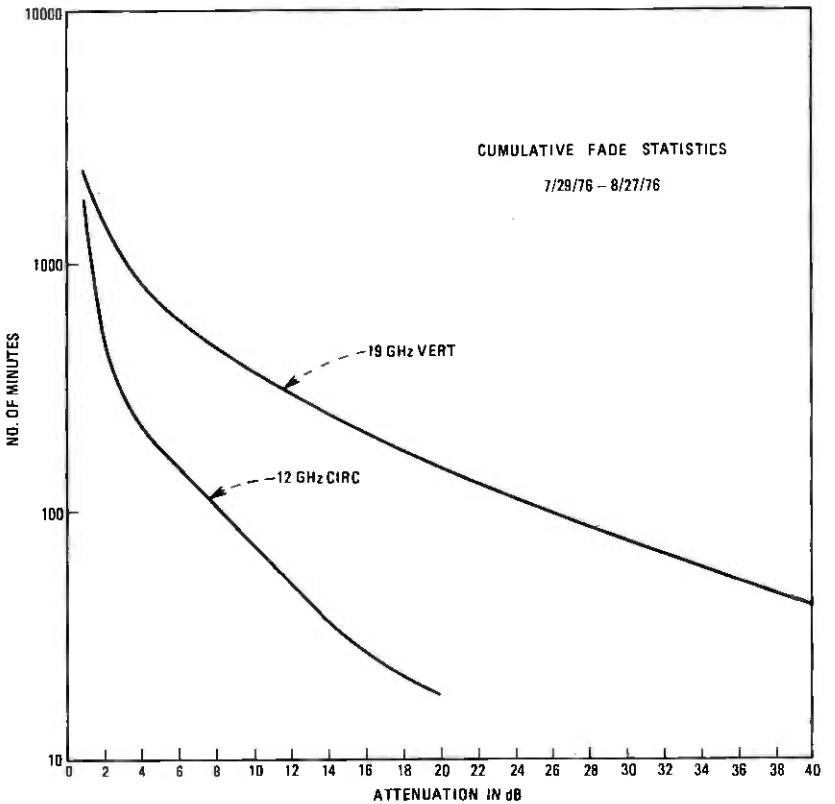


Fig. 11—Cumulative attenuation statistics at 12 and 19 GHz for July 29, 1976 to August 27, 1976.

STAR data, A. W. Norris for his continuing assistance in the operation of the receiving system, and H. R. Hunczak at NASA Lewis Research Center for providing CTS tracking data.

APPENDIX

Frequency Scaling of Attenuation Statistics

To obtain an estimate of the attenuation statistics at 12 GHz when the CTS beacon signal was lost during eclipse periods and system outage, an empirical relation was determined between the 12-GHz attenuation statistics and 19-GHz COMSTAR beacon attenuation statistics.

The propagation paths through the atmosphere from the two satellites and the signal polarizations are different. The CTS beacon signal is circularly polarized and at Holmdel requires a ground station antenna point at approximately 234 degrees azimuth and 27 degrees elevation. The

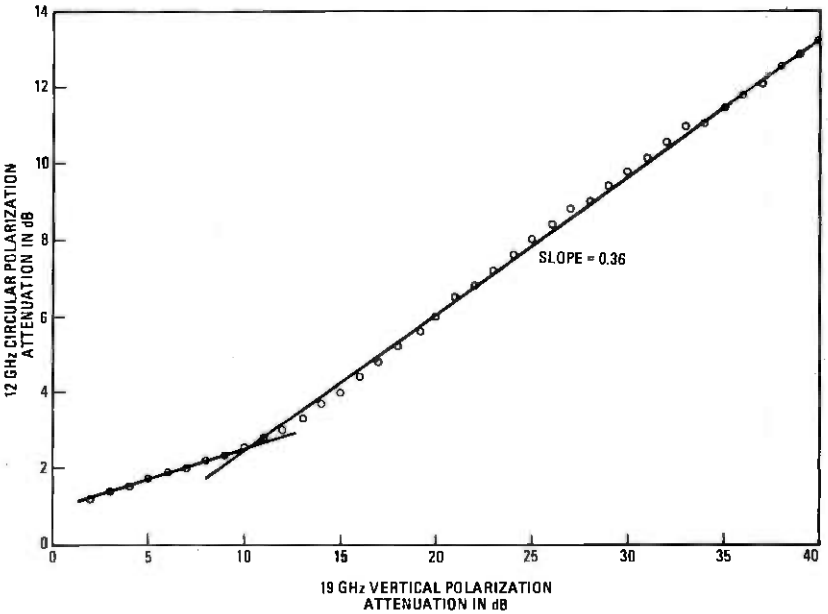


Fig. 12—Plot of 12-GHz attenuation as a function of 19-GHz attenuation at constant outage periods for July 29, 1976 to August 27, 1976.

COMSTAR A beacon signal is linearly polarized and requires an antenna point of approximately 244 degrees azimuth and 18 degrees elevation. The frequency scaling presented here, therefore, only applies to these two experiments.

The frequency scaling of the attenuation statistics was done by relating the decibel levels of the 12- and 19-GHz cumulative attenuation distribution curve at equal outage times. A common measurement period was chosen when both systems were operating and where sufficient rainfall occurred to produce significant attenuation at the two frequencies. This period, between July 29 and August 27, 1976, included several rain events, providing an ensemble average over multiple events. Figure 11 shows the measured cumulative attenuation statistics at 12 and 19 GHz. Using the data from this figure, another curve was plotted showing the attenuation at 12 GHz as a function of attenuation at 19 GHz at constant outage times. This frequency relationship is shown in Fig. 12. This curve appears to have segments with two distinctly different slopes. The change in slope at lower attenuations may be due in part to digitizing errors of the 12-GHz data. In the region above 10 dB attenuation at 19 GHz, the curve has a constant slope of 0.36. This shows, for a given outage time, the attenuation observed at 12 GHz will be 0.36 times the decibel attenuation at 19 GHz for the same measurement period.

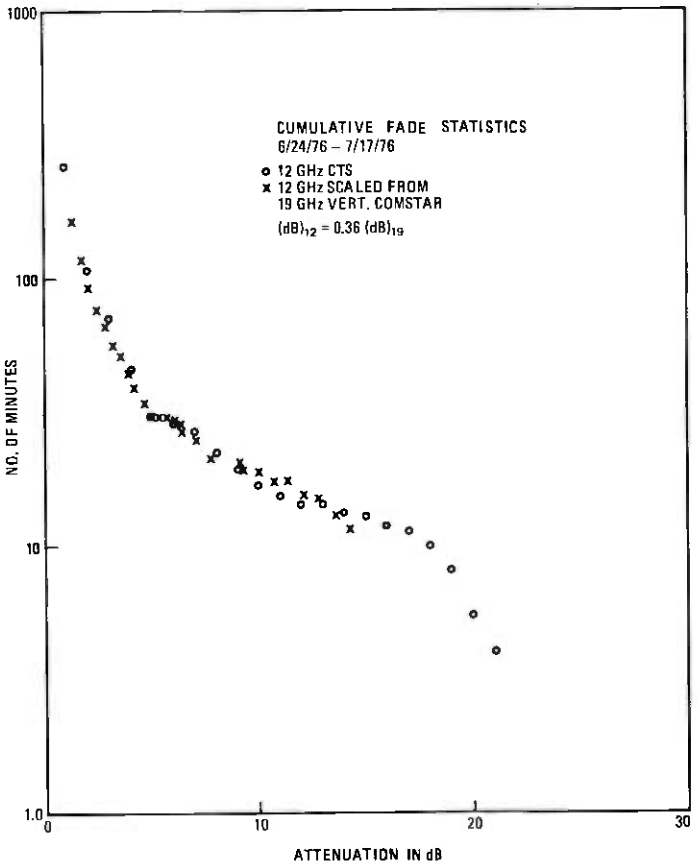


Fig. 13—Cumulative attenuation statistics at 12 GHz with data scaled from 19 GHz for June 24, 1976 to July 17, 1976.

To test the accuracy of this rule, another period of data was analyzed. The cumulative attenuation statistics for 12 and 19 GHz were calculated for the period between June 24 and July 17, 1976. The 19-GHz data were scaled to 12 GHz using the previously determined factor. Both data are plotted in Fig. 13. The two sets of data are almost indistinguishable over most of the distributions, which lends reasonable confidence to the scaling technique.

REFERENCES

1. D. C. Cox, H. W. Arnold, and A. J. Rustako, Jr., "Some Observations of Anomalous Depolarization on 19 and 12 GHz Earth-Space Propagation Paths," Radio Science, *12*, No. 3 (May–June 1977), pp. 435–440.
2. A. B. Crawford, D. C. Hogg, and L. E. Hunt, "A Horn-Reflector Antenna for Space Communications," B.S.T.J., *40*, No. 4 (July 1961), pp. 1095–1116.

3. P. V. Bradford and R. W. Wilson, "Fourier Series Representation for Tracking Inclined, Elliptical, Synchronous Satellite Orbits," private communication.
4. H. W. Arnold, private communication.
5. S. H. Lin, private communication.
6. T. S. Chu, "Rain-Induced Cross-Polarization at Centimeter and Millimeter Wavelengths," B.S.T.J., *53,* No. 8 (October 1974), pp. 1557–1579.
7. N. J. McEwan, P. A. Watson, A. W. Dissanayake, D. P. Haworth, and V. T. Vakili, "Cross-Polarisation From High-Altitude Hydrometeors on a 20 GHz Satellite Radio Path," Electronics Letters, *13,* No. 1 (January 6, 1977), pp. 13–14.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Fault Modeling and Logic Simulation of CMOS and MOS Integrated Circuits

By R. L. WADSACK

(Manuscript received October 10, 1977)

This paper addresses the simulation and detection of logic faults in CMOS integrated circuits. CMOS logic gates are intrinsically tri-state devices: output low, output high, or output open. This third, high-impedance condition introduces a new, nonclassical logic fault: the "stuck-open." The paper describes the modeling of this fault and its complement, the stuck-on, by means of gate-level networks. In addition, this paper provides a methodology for creating simulator models for tri-state and other dynamic circuit elements. The models are gate-level in structure, provide for both classical and stuck-open/stuck-on faults, and can be adopted for use on essentially any general purpose logic simulator.

I. INTRODUCTION

The challenge of testing silicon integrated circuits (ICs) is becoming more formidable with the rapidly expanding production of large-scale integrated (LSI) circuits. Increased gate-count, increased pin-count, smaller feature size, higher performance, and higher complexity all contribute to a mounting "testability" problem. Furthermore, there is considerable evidence that the economic requirements to meet that challenge will continue to grow at a rate markedly greater than that of circuit size alone.

As a further dimension to the challenge, IC tests must be specifically designed to recognize failure-mode dependence upon circuit configuration, processing parameters, and the overall technology (TTL, ECL, PMOS, CMOS, etc.). That is, a Boolean network realized in one technology can have a strikingly different implementation in another. Consequently, logic tests must be created which exercise not only the gross functional behavior of the IC but also the structure used for that function. However, for large-scale ICs, internal circuit structure and complexity are in-

creasing at a much more rapid rate than is the number of access terminals.

The rising use of MOS technology for LSI circuits has introduced a number of circuit elements whose logical behavior and faults are generally not treated by existent logic simulators.[1] These include, for example, transmission gates, tri-state inverters, and bidirectional buses. Furthermore, the failure modes of such circuits and even those of ordinary combinational logic gates can introduce nonclassic logic faults. That is, they possess a faulted behavior for which test coverage would not be verified on a conventional fault simulator.

The focus of this paper will be centered upon fault modeling and logic simulation of CMOS digital integrated circuits. The motivation for this direction is the recent emergence of CMOS as a mature technology for the design of densely-packed, low-power digital LSI circuits.[2] Secondly, CMOS is intrinsically a three-state logic technology. Consequently, it readily lends itself both to the illustration of the new, nonclassical logic faults and to a methodology for modeling the dynamic nature of MOS circuits.

II. CMOS LOGIC FAULTS

2.1 CMOS logic gates

Figure 1 shows a two-input CMOS NOR gate: the output is high if and only if $A = B = 0$. The realization of the NOR function shows the series/parallel complementary nature of CMOS logic gates: $F = \overline{A} \cdot \overline{B}$ and $\overline{F} = A + B$ where the Boolean function $\overline{A} \cdot \overline{B}$ connects the output to the "1" level and the function $A + B$ connects the output to the "0" level. Each is the complement of the other and is implemented, respectively, with p-channel FETs and n-channel FETs.

The NOR circuit is a specific example of the general CMOS characteristic of complementary pull-up/pull-down networks. The only two steady-state logic outputs are 0 and 1. The former arises when the pull-down network is conducting and the pull-up network is nonconducting. The latter, $F = 1$, occurs when the two networks reverse their conductivity states. Consequently, there is no static current path between VDD and VSS, and CMOS ICs therefore dissipate power only to charge and discharge circuit capacitance.

On the other hand, there are two common situations which can lead to a third logic state. This third condition is the "open" or high-impedance state.[3] One source of the "open" state is the presence of a logic fault which prevents one network from conducting when the other is in a nonconductive state. A second cause is the legitimate use of a high-impedance state in dynamic circuits or tri-state buffers, for example. In each instance the output retains the logic value of the previous output

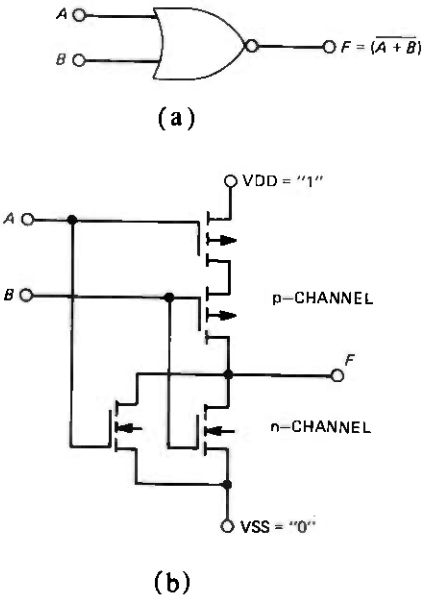


Fig. 1—The CMOS two-input NOR gate: (a) logic symbol and transfer function, and (b) FET realization.

state. This is true because the gates are loaded with capacitance only. The length of time the state is retained, however, is determined by the leakage current at the node.

Conceptually, a fourth state could exist: both networks conducting. However, this represents a logical inconsistency, i.e., the output cannot simultaneously be both high and low. Such a fault is more analog in nature because the output voltage lies somewhere between VDD and VSS. The actual value is determined by the impedance ratios of the networks and the associated fault. For the most part these failures will not be treated as logic faults.

There are two kinds of *classical* logic faults: stuck-at-one (SA1) and stuck-at-zero (SA0). These faults may occur at either an input or an output of a logic gate. On the other hand, a gate with n inputs can have only $n + 2$ distinct classical faults. Input faults must be stuck in the "nondominant" state to be distinguishable from an output stuck-at fault. For AND/NAND gates and OR/NOR gates such nondominant faults are SA1's and SA0's, respectively. These faults are sometimes called "input-open-from" faults and will be denoted as IOF faults in this paper. The two-input NOR gate of Fig. 1 has the four classical faults: SA0, SA1, IOFA, and IOFB. These are also symbolized as $F(0)$, $F(1)$, $F(A)$, and $F(B)$, respectively.

For CMOS logic gates the nonclassical "stuck-open" faults must be

Table I — CMOS two-input NOR gate: truth table

A	B	F	F (0)	F (1)	F (A)	F (B)	F (ASOP)	F (BSOP)	F (VDDSOP)
0	0	1	0	1	1	1	1	1	4
0	1	0	0	1	0	1	0	4	0
1	0	0	0	1	1	0	4	0	0
1	1	0	0	1	0	0	0	0	0
normal			classical faults				nonclassical faults		

4 = previous output state
$F(A) = F(\text{IOFA})$
$F(B) = F(\text{IOFB})$

included to represent the undesired, high-impedance state caused by a faulty pull-up or pull-down network. For the two-input NOR gate of Fig. 1, there are three such stuck-open (S-OP) faults: ASOP, BSOP, and VDDSOP. The first, ASOP, is caused by an open, or missing, n-channel A-input pull-down FET. The second, BSOP, is caused by an open, or missing, B-input pull-down FET. The third, VDDSOP, is caused by an open anywhere in the series, p-channel pull-up connection to VDD.

Table I shows the truth table for the two-input CMOS NOR gate for both the fault-free and the seven faulted conditions. For example, the fault-free gate obeys

$$F = \overline{(A + B)}$$

whereas

$$F(\text{IOFB}) = \overline{A}$$

and

$$F(\text{ASOP}) = \overline{A} \cdot \overline{B} + 4 \cdot A \cdot \overline{B},$$

where "4" denotes the previous state of F. [Using the notation of sequential circuit design, the latter equation would read $F_{n+1}(\text{ASOP}) = \overline{A} \cdot \overline{B} + F_n \cdot A \cdot \overline{B}$. The use of a "4" to symbolize F_n is a convention adopted to describe the effect of S-OP faults.]

How are the seven NOR gate logic faults related to actual physical flaws in the IC? The SA0, SA1 faults correspond to a low-impedance "short" to VSS or VDD, respectively. The IOF faults are caused by an open input to the *logic gate* as a whole. In addition to being open, the input is in a charged condition which is recognized as a logic 0. (An IOF fault in a NOR gate is an SA0, by definition; in a NAND gate the analogous fault would be an SA1, of course.) That is, both the p-channel and the n-channel FETs have a 0 applied to them. On the other hand, the nonclassical S-OP faults arise from a missing connection to the gate of *individual* FETs, for example, with the gate in a charge state such that the FET is nonconducting. Another cause of an S-OP fault is an open, or missing, connection to either the source or the drain of an FET.

Table II — CMOS two-input NOR gate: fault detection (test sequence: AB = 00,01,00,10)

	A	B	F	F (0)	F (1)	F (A)	F (B)	F (ASOP)	F (BSOP)	F (VDDSOP)
1	0	0	1	0*	1	1	1	1	1	3
2	0	1	0	0	1*	0	1*	0	1*	0
3	0	0	1	0	1	1	1	1	1	0*
4	1	0	0	0	1	1*	0	1*	0	0
	normal			classical faults				nonclassical faults		

3 = unknown output state (0 or 1).
* Vector at which simulator detects the fault.

In this context an "open" denotes an undesired high impedance at either the gate, the source, or the drain of an FET. Of course, any residual capacitive or resistive coupling must be negligibly small for a high-impedance fault to be regarded as a true "open." In addition, the actual occurrence of such faults depends on the specific topology of the logic gate.

The truth table of Table I shows that S-OP faults create sequential circuits where only a combinational circuit existed for the fault-free gate. This increases the difficulty of both testing the circuit and designing a set of input test "vectors" to achieve a high percentage of fault coverage. For example, the four input states of Table I, if applied in that order, will detect only 5 out of the 7 logic faults. ASOP and VDDSOP will be undetected. The ASOP fault is not detected because $F(10)$ = "4" = 0 which is the correct output for the "10" input vector. The VDDSOP fault will not necessarily be detected because of the chance that the gate powers up with the output in the high (but correct) state for the "00" input vector.

The primary reason that the above two faults were undetected is that the corresponding circuit paths (devices) were not tested to determine whether they fulfilled their most basic function. For example, to test the A-input n-channel *pull-down* FET, the output node must first be driven high and then that FET, and it alone, must be capable of pulling the node low. The sequence of inputs in Table I did not meet that condition.

A set of vectors which detects all 7 faults is the following: 00,01,00,10. Table II shows the response of the 8 circuits (the good circuit and the 7 faulty ones) to the above input sequence. The * symbol designates the vector at which each fault is first detected by the *simulator*. The "3" denotes an unknown state (either a 1 or a 0), caused by the VDDSOP fault as described previously. Depending upon the actual state of the circuit at power-up, the VDDSOP fault may, therefore, be detected at either vector 1 or 3.

The influence of CMOS faults on fault coverage is shown in Fig. 2. The

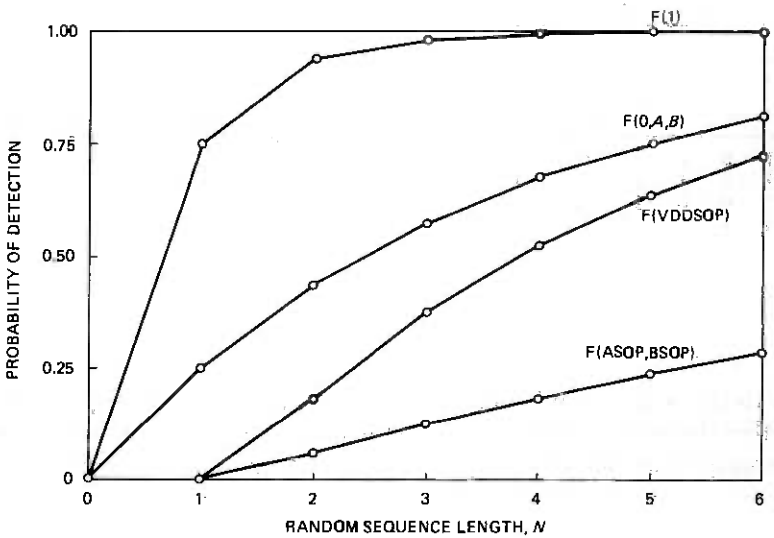


Fig. 2—The CMOS two-input NOR gate: probability of fault detection (for a simulator) versus the length N of a sequence of random input vectors.

probability of detection for random vector sequences is given for each of the 7 faults in Table I. The probability is that which would prevail for a logic simulation of a two-input NOR gate subjected to random vector sequences of length N. The output SA1 fault ($F(1)$) has a probability of detection of nearly unity for sequences of length greater than three. The two pull-down stuck-open faults have the lowest probability, reaching only 0.29 for sequences 6 vectors long (45 vectors are required to reach 0.95). Note also the similarity between the VDDSOP fault and the output SA0 fault ($F(0)$). Further, the lag for all three CMOS faults is evident.

In addition to increasing the number and complexity of CMOS logic faults, stuck-open faults are also timing-sensitive. Specifically, the above set of input vectors will detect three S-OP faults of the NOR gate only if they are applied to the gate at a rate more rapid than that associated with leakage current time-constants. A rate significantly slower than, say, 10 kHz may allow some faulty devices to charge to the correct state before the output is sampled by the test set. Truth-table testing at quasi-dc rates is inadequate. Conversely, an ill-chosen vector set, such as the binary sequence of Table I, may not detect S-OP faults no matter how fast it is applied to the device under test. Note also that the sequence of Table I is "exhaustive," but does not achieve 100 percent fault coverage.

Another aspect of stuck-open faults is that of long-term reliability. A fault caused by a missing connection to the gate of an individual FET may cause that FET to be open and yet remain undetected during production testing if the test vector set has less than 100 percent fault cov-

erage. Later, however, under actual operating conditions the FET gate may acquire charge of the opposite polarity and become conducting or "stuck-on." CMOS S-OP field failures have been observed.[4]

2.2 Modeling CMOS logic faults

The design of a set of vectors to achieve 100 percent fault coverage for small-scale integrated (SSI) circuits such as NAND and NOR gates is trivial and can be done by inspection. However, for complex high gate-count circuits such as medium- and large-scale ICs (MSI, LSI), the use of computer-based logic simulators is a necessity. To meet this need the simulation of both stuck-open faults and dynamic logic elements has been approached from the standpoint of circuit modeling. The models represent circuit elements both in their fault-free condition and in high-impedance state(s), if any. Combinational "static" gates (e.g., NAND or NOR) enter the high-impedance condition only in the presence of S-OP faults. Dynamic or tri-state logic elements, however, can intentionally be placed into a high-impedance state by auxiliary, or control, inputs.

The models presented in this paper are gate-level oriented because no other method generally exists for simulating nonclassical logic faults. On the other hand, the logical behavior of the models can be incorporated into a higher-level functional description if that capability is available on the logic simulator. On the LAMP system[1] the Function Description Language (FDL) is being modified to include such "internal faults."[5] Although the models in this paper are implemented in terms of NAND/NOR logic, they do have the advantage of being simulator independent. That is, they will correctly model fault-free and faulted logic networks regardless of the particular simulator chosen. The specific illustrations, however, are taken from the author's experience with LAMP.

Existent machine aids simulate for the most part only the classical SA0/SA1 faults and not the "stuck-open" faults. One possible solution to the problem is to use a network of conventional gates (NOT, NAND, NOR, etc.) to form a model which duplicates both the normal and the faulted behavior of a single CMOS gate. One of the basic properties of the model is that it must possess the capability of passing 0/1 data from input to output in accordance with the fault-free logic function of the gate. Second, when there is an S-OP fault, it must retain the previous state in the presence of the "provoking" input (see Table I). This suggests the use of gated latch.

The general approach to modeling either stuck-open (S-OP) faults or dynamic gates is shown in Fig. 3. The "gated latch" represents the nodal capacitance associated with the logic function. For $T = 1$, the output equals the input ($Z = D$); the $T = 0$, the output latches and stores the previous state ($Z =$ "4"). The "node faults gate" has been added to introduce the two classical SA0 and SA1 (stuck-at) faults. The gate is not

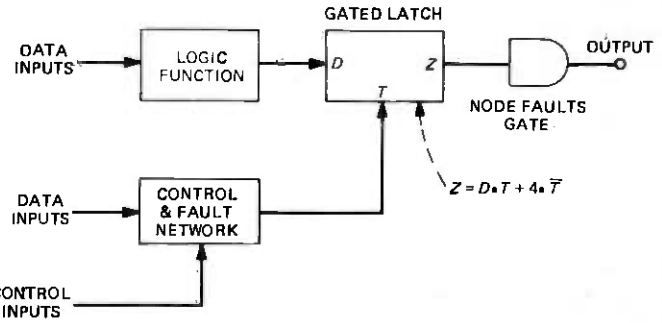


Fig. 3—Block diagram: the general approach for stuck-open/dynamic gate models.

necessary if those faults are incorporated within the "logic function" network. The latter represents the actual logical operation being modeled (NAND, NOR, etc.). The "control and fault network" establishes whether the output of the "logic function" is to be transmitted to the gates driven by the "output." In the case of static gates, there are no "control inputs" to that network. Only the presence of S-OP faults would cause $T = 0$ (i.e., no transmission). Dynamic circuit elements have "control inputs" which can cause $T = 0$ in the absence of faults. Classical faults generated in the "logic function" network at input D will, of course, propagate through the latch (and not through the S-OP fault generating network).

III. SPECIFIC MODELS

3.1 The NOR gate

The model for the two-input CMOS NOR gate is shown in Fig. 4. The NOR gate named GATE has zero delay but is faulted (i.e., the fault simulator will simulate faults for this gate) so as to introduce all the classical faults of a two-input NOR gate. Of course, GATE represents the modeled logic function. All other gates in Fig. 4 compose the fault-generating network which sets $T = 0$ in the presence of an S-OP fault *and* the "provoking" input for that fault.

The model functions as follows: If there are no faults in the circuit, then $F = \overline{(A + B)}$. If there are only classical faults (from GATE), then $T = 1$ and those faults propagate through GL. If there is an S-OP fault in the gate, than $T = 0$ for the provoking vector and $F = 4$ as required from Table I. For example, because its input is grounded the only fault that the simulator assigns to the gate ASOP is an SA1. Therefore, in the presence of the fault ASOP(1) and when $AB = 10$, then $T = 0$ and the GL latch holds the output equal to the previous value. Similarly the output F will be stuck-open in the presence of the VDDSOP(1) fault if and only if $AB = 00$.

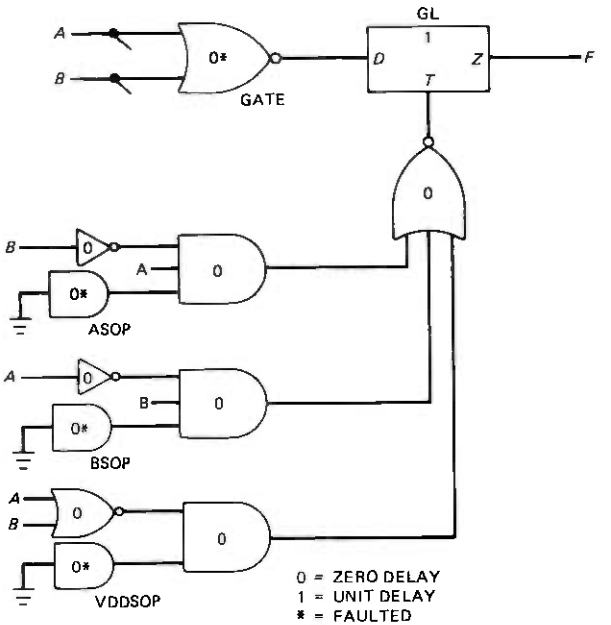


Fig. 4—The CMOS two-input NOR gate model, NR2. The block labeled GL is the gated latch circuit of Fig. 3. Only those gates marked * are faulted by the simulator. Propagation delays are denoted by either 0 or 1.

The NOR gate model of Fig. 4 has been designed to yield the correct logic behavior of the two-input CMOS NOR gate in the absence of faults. Secondly, it correctly models all 7 logic faults: SA0, SA1, IOFA, IOFB, ASOP, BSOP, and VDDSOP. Thirdly, it generates no spurious faults associated with the modeling, nor does it have a propagation delay other than the one unit that would be expected from a single NOR gate. Finally, the nature of the fault-generating network prevents the spurious propagation of faults from other gates connected to the inputs of the gate in question.

The names of the gates in the model have been chosen to provide ease of use during simulation. Specifically, if a circuit contained a two-input CMOS NOR gate with the name NOR17, then the LAMP simulator would assign the following fault list to it:

NOR17.GATE(0)
NOR17.GATE(1)
NOR17.GATE(A)
NOR17.GATE(B)
NOR17.ASOP(1)
NOR17.BSOP(1)
NOR17.VDDSOP(1)

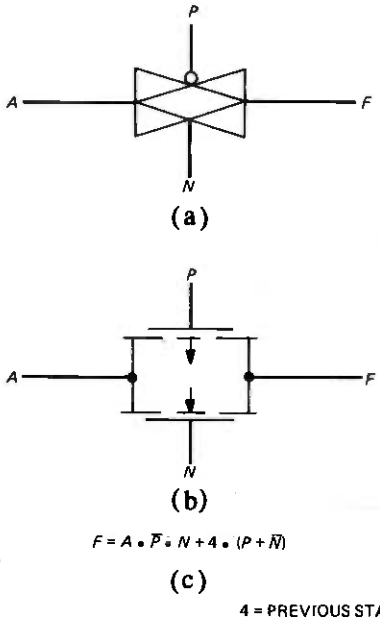


Fig. 5—The CMOS transmission gate: (a) logic symbol, (b) FET realization, and (c) the static, unilateral logic transfer function.

Consequently, the test engineer can determine at a glance the nature of a fault and where it is located. In addition, in the LAMP system all S-OP faults can be globally nonfaulted (or faulted) because they arise from gates ending in the string "SOP." This nonfaulting capability is useful in determinations of relative fault coverage.

3.2 The transmission gate

MOS technology possesses an interesting circuit element with both digital and analog capability: the transmission gate. The CMOS transmission gate is shown in Fig. 5.[6] It is a *bilateral* device with a conducting mode for PN = 01 and a nonconducting state for PN = 10. A problem occurs for the two other vectors PN = 00 and PN = 11. In each case either the p-channel or the n-channel FET will be on, but not both. Low-to-high transitions (open p-FET) or high-to-low transitions (open n-FET) will be attenuated as they pass through the transmission gate. In addition, there will be an accompanying speed degradation caused by the higher impedance of the single FET path.

Although the two input vectors 00 and 11 apply abnormal conditions to the transmission gate, the choice of whether to regard the gate as a whole as either "on" or "off" is somewhat arbitrary. For these particular

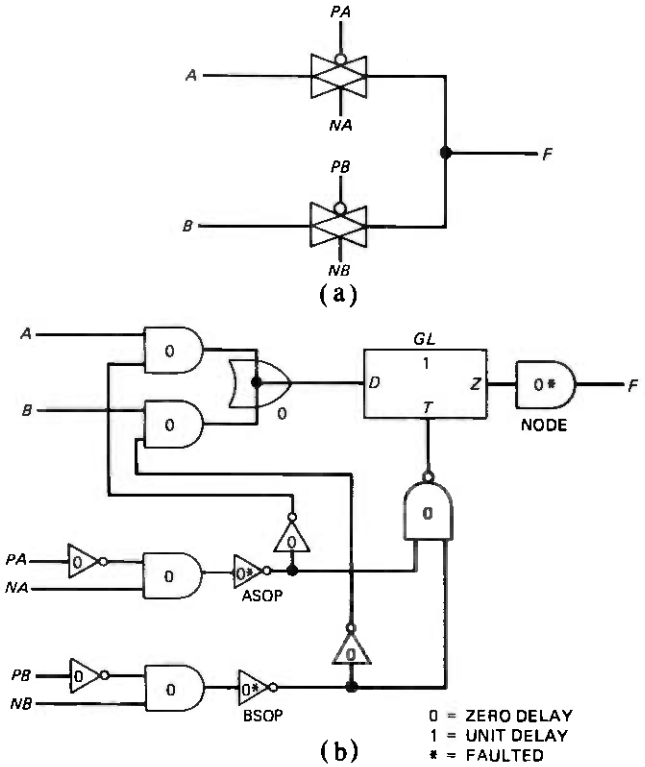


Fig. 6—The CMOS two-channel transmission gate node, XG2: (a) logic symbol and (b) a gate-level model which exhibits both the fault-free and the faulted behavior of the XG2 element. Signal flow is unilateral: from A and B to F. Normally only one channel is enabled at one time.

models the worst-case logic behavior of the gate has been taken to be

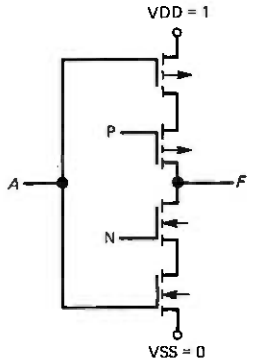
$$F = A \cdot \overline{P} \cdot N + 4 \cdot (P + \overline{N})$$

where "4" represents the previous state of the output. The above expression can be rewritten as

$$F = D \cdot T + 4 \cdot \overline{T}$$

where $T = \overline{P} \cdot N$ and $D = A$. This is just the equation for the gated latch of Fig. 3, i.e., the gated latch is equivalent to a nonfaulted transmission gate.

Functionally, transmission gates generally occur in groups of two, three, four, etc. Hence, they can be regarded as multiplexing nodes at which usually only one of the gates is conducting at a time. The model for the CMOS two-channel transmission gate node, used in the above sense, is shown in Fig. 6. The gate named SUM is a zero-delay nonfaulted



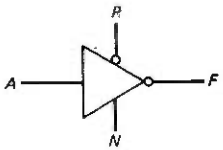
(a)

A	P	N	F
0	0	0	1
0	0	1	1
0	1	0	4
0	1	1	4
1	0	0	4
1	0	1	0
1	1	0	4
1	1	1	0

(4 = PREVIDUS STATE)

$F = \bar{A} \cdot \bar{P} + 4 \cdot (\bar{A} \cdot P + A \cdot \bar{N})$

(b)



(c)

Fig. 7—The CMOS tri-state inverter: (a) the FET realization, (b) the truth table, and (c) the logic symbol.

tied-OR node (TOR). The purpose of the gate labeled NODE is to provide the two classical faults SA0, SA1 associated with the node F. The nonclassical faults generated by this network are four in number: ASOP(0), ASOP(1), BSOP(0), and BSOP(1). As usual, ASOP(1) means that the A-channel is stuck-open and, conversely, ASOP(0) means that it is "stuck-on" (S-ON). For the latter fault, $F = A + B \cdot (\overline{PB}) \cdot (NB)$. Here, the stuck-on has been treated as a legitimate logic fault whose presence induces a "1-dominant" short, i.e., the spurious A-input is ORed with the correct B-channel response. Of course, there may be no technological reason to assign a S-ON fault as either 1-dominant or 0-dominant. In that case, the model in Fig. 6 (and others) can easily be recast to exhibit only S-OP nonclassical faults.

Stuck-open faults in a CMOS transmission gate are also timing-sensitive as in NAND and NOR gates, but in a different manner. Specifically, even with one FET of the pair open, the other can provide nearly a complete logic transition but at a higher average impedance level. Consequently, 0/1 data will propagate through the gate but at slightly slower speeds. However, the reduction in speed is much smaller than that caused by S-OP faults in NAND or NOR gates. Therefore, an S-OP fault in a transmission gate may be intrinsically undetectable even at the highest data rates that occur at a particular gate.

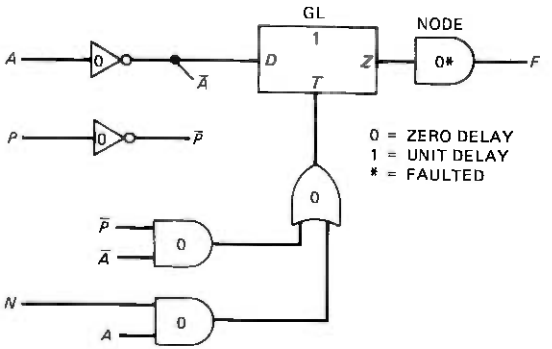


Fig. 8—The CMOS tri-state inverter model. Only the gate marked * is faulted by the simulator. Propagation delays are denoted by either 0 or 1.

3.3 The tri-state inverter

The FET implementation of the CMOS tri-state inverter is shown in Fig. 7(a). In this case P and N are control leads that determine whether the circuit inverts the input A or remains in the high-impedance ("4") state. For the latter condition, the nodal capacitance retains the value of the previous state ("0," "1," or "3"). The truth table is given in Fig. 7(b) and leads to the transfer function: $F = \overline{A}\cdot\overline{P} + 4\cdot(\overline{A}\cdot P + A\cdot\overline{N})$. The logic symbol is shown in Fig. 7(c).

The model for the tri-state inverter is given in Fig. 8. Here, only the two classical faults $F(0)$ and $F(1)$ have been assigned to the model because of the similarity of VDDSOP/VSSSOP faults to $F(0)$ and $F(1)$ faults, respectively. The S-OP and S-ON faults could be modeled, of course, but the added complexity does not warrant separate treatment.

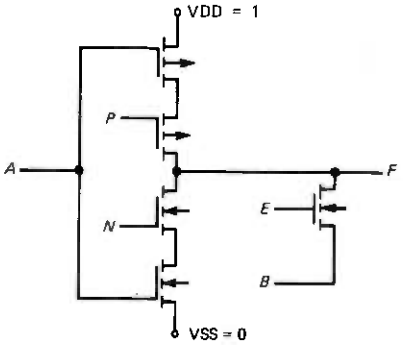
The relation of the model network to the truth table is perhaps more obvious if the transfer function is written as

$$F = 1\cdot\overline{A}\cdot\overline{P} + 0\cdot A\cdot N + 4\cdot(\overline{A}\cdot P + A\cdot\overline{N}).$$

The terms $\overline{A}\cdot\overline{P}$ and $A\cdot N$ then compose the *t*ransmit (T) function: $T = \overline{A}\cdot\overline{P} + A\cdot N$. Therefore, when $T = 0$ then $F =$ "4." That is, the output latches (or stores) the previous state.

3.4 Modified tri-state inverter

The addition of an n-channel transmission gate FET to the output of the tri-state inverter forms a "modified" inverter of Fig. 9(a). (This circuit element forms one half of a gated sense amplifier.) The resultant truth table is shown in Fig. 9(b). The "3" (unknown 0 or 1) state occurs whenever $A = B$ and both the inverter and the transmission gate are enabled. In other words, whenever each attempts to drive the node F to

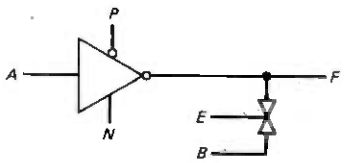


(a)

$E = 1$

			$E = 0$	$B = 0$	$B = 1$	
A	P	N	F	F	F	
0	0	0	1	3	1	
0	0	1	1	3	1	
0	1	0	4	0	1	
0	1	1	4	0	1	
1	0	0	4	0	1	
1	0	1	0	0	3	
1	1	0	4	0	1	(3 = UNKNOWN STATE)
1	1	1	0	0	3	(4 = PREVIOUS STATE)

(b)



(c)

Fig. 9—The CMOS modified tri-state inverter: (a) the FET realization, (b) the truth table where 3 indicates an unknown 0 or 1 state and 4 symbolizes the previous state of F, and (c) the logic symbol for the combination tri-state inverter and n-channel FET.

opposing logic states, the output is indeterminate ("3"). The high impedance "4" state occurs whenever both channels are disabled.

The model for the modified tri-state inverter can be developed by noting that the Boolean expression that selects the $\overline{A}$ channel is $\overline{P}\cdot\overline{A} + N\cdot A$ and the term that selects the B channel is E [see Fig. 9(a)]. Therefore, define $\text{SA} = \overline{P}\cdot\overline{A} + N\cdot A$ and $\text{SB} = E$. Then, $T = \text{SA} + \text{SB}$ or $T = \overline{\text{SA}}\cdot\overline{\text{SB}}$. Therefore,

$$F = \overline{A}\cdot\text{SA}\cdot\overline{\text{SB}} + B\cdot\overline{\text{SA}}\cdot\text{SB} + 3\cdot\text{SA}\cdot\text{SB} + 4\cdot\overline{\text{SA}}\cdot\overline{\text{SB}}$$

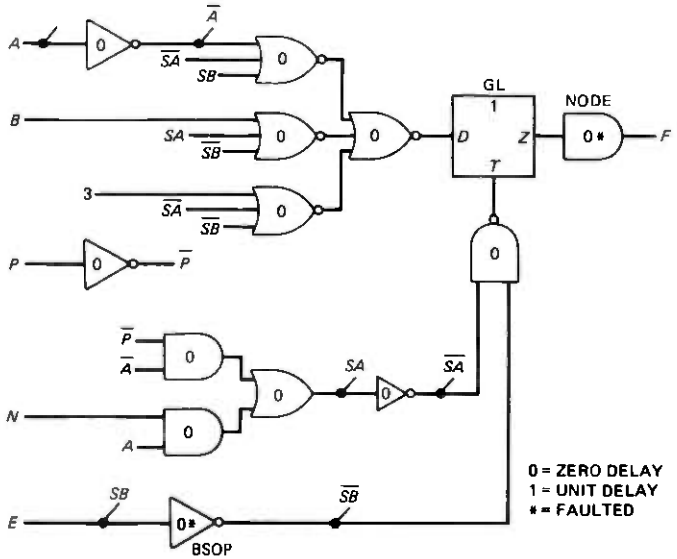


Fig. 10—The CMOS modified tri-state inverter model.

or

$$F = (\overline{A} + \overline{\text{SA}} + \text{SB}) \cdot (B + \text{SA} + \overline{\text{SB}}) \cdot (3 + \overline{\text{SA}} + \overline{\text{SB}}) \cdot (4 + \text{SA} + \text{SB}).$$

The model corresponding to the latter expression for F is shown in Fig. 10. The model has four faults assigned to it: two classical faults NODE(0,1), and two CMOS faults BSOP(0,1). For example, the fault BSOP(1) means that the B-channel transmission gate is stuck-open and nonconducting. S-OP faults in the inverter itself are ignored (see Section 3.3, above).

The model of Fig. 10 produces a "3" output whenever conflicting output conditions are generated by the simultaneous selection of both the A channel and the B channel. The "3" capability may be important during design verification of the fault-free circuit behavior. On the other hand, if it is known that mis-selection is unimportant or not possible, then the model can be reduced to that shown in Fig. 11.

It would be pointless to introduce into either model a fault whose sole effect would be the generation of a "3" output. For fault simulations "3" outputs are effectively ignored. A fault whose only result is to produce a "3" would be undetectable on a logic simulator.

3.5 Input/output port

Some integrated circuits contain pins which can serve as both input and output ports. Figure 12(a) shows the physical structure of one such

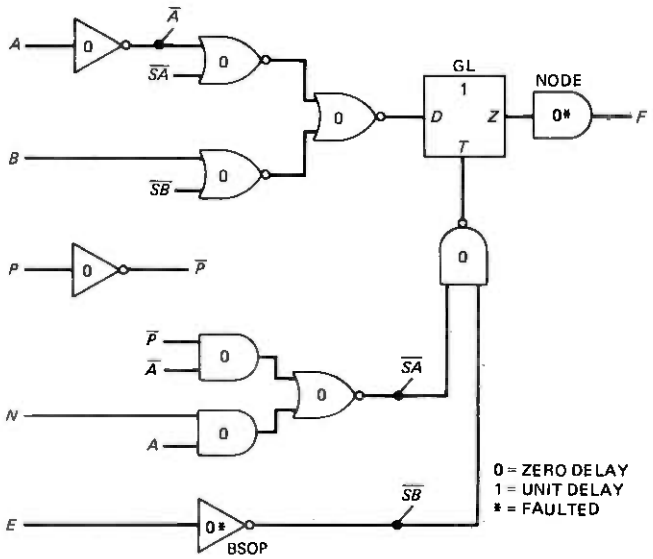


Fig. 11—The CMOS modified tri-state inverter model: reduced version. The unknown 3 condition has been ignored and replaced by a 0-dominant short assignment.

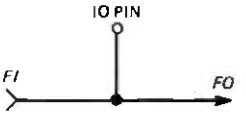
pin. IOPIN is the actual connection to the external world, FO represents all fan-out loads from the node, and FI represents all fan-in (tri-state) devices at the node. The modeled structure is shown in Fig. 12(b) where the IOPIN has been divided into separate IN and OUT functions. The D and E variables have been introduced as the control inputs which define the state of the input driver and the FI branch, respectively.

The truth table of Fig. 12(c) defines the relation between D, E, and OUT. Note that the impedance of the input driver (of the test set) is taken to be much less than that of the *FI* branch. Consequently, whenever the node is driven from an external source ($D = 1$), the logical state of the driver overrides that of FI.

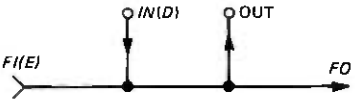
Under the above assumptions the gate-level model for the I/O port can be constructed as shown in Fig. 13. The model has been assigned four faults: the usual output SA0, SA1 faults and two S-OP/S-ON faults associated with the FI branch. Ideally each fan-in branch at the node would have two such faults. However, in the absence of a detailed knowledge of the fan-in network, just two faults have been indicated.

3.6 Tri-state bilateral buses joined by a transmission gate

The characteristics of the simple I/O port described above can be extended to more complex networks. One example is shown in Fig. 14. Two tri-state bilateral buses are connected by a bilateral transmission gate. All the data sources ("talkers") are grouped on the left and all data



(a)



(b)

D	E	OUT = FO
0	0	4
0	1	FI
1	0	IN
1	1	IN

$$OUT = IN \cdot D + FI \cdot \bar{D} \cdot E + 4 \cdot \bar{D} \cdot \bar{E}$$

(c)

Fig. 12—Input/output port: (a) physical structure, (b) modeled structure, and (c) truth table. IOPIN is the actual connection to the outside world, *FO* represents all fan-out loads, and *FI* represents all fan-in (tri-state) devices. *D* and *E* define the state of the input driver and the *FI* branch, respectively. The symbol 4 denotes the previous state of OUT.

sinks ("listeners") on the right. A device that can be both send and receive would be represented once in each group.

The models described previously in this paper have all been inde-

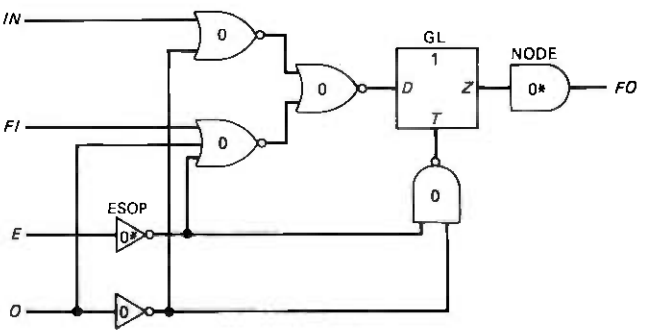


Fig. 13—Input/output port model. Only gates marked * are faulted, 0 and 1 denote delays, and *D* and *E* are enable controls for IN and FI, respectively. FI = fan-in; FO = fan-out.

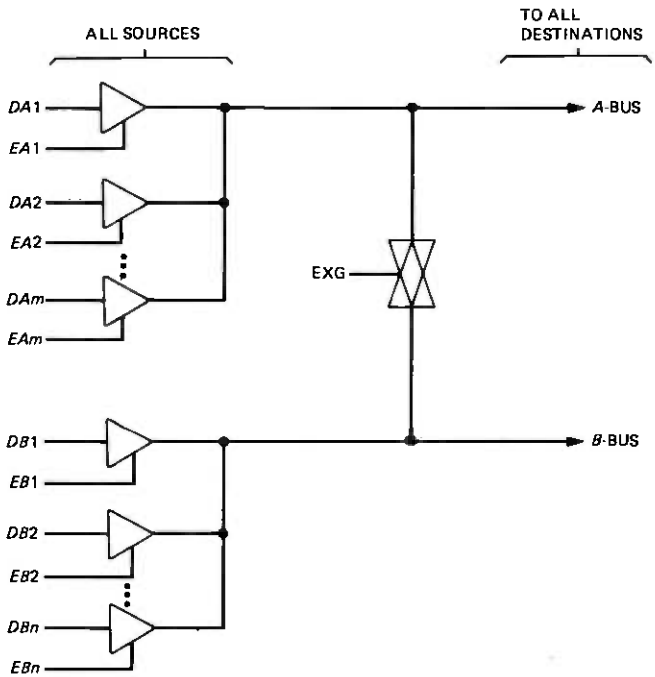


Fig. 14—Two bilateral tri-state buses connected by a transmission gate. All data sources (drivers) are grouped on the left. All data sinks (gate inputs) are to the right.

pendent of each other. Therefore, they can be organized as a library of subnetwork building blocks. Consequently, it would be tempting to model the network of Fig. 14 as an interconnection of three independent models: one for the A-bus, one for the B-bus, and one for the bilateral transmission gate. However, in this instance, that is not possible. A successful model must incorporate features from all three to correctly represent the true bilateral interactions between each bus by way of the transmission gate.

The model for the fault-free behavior of the A-bus output is shown in Fig. 15. The interbus coupling is modeled by means of the EXG·EB AND gate and the B·SB AND gate. For example, if the A-bus driving sources are in the high-impedance state (EA = 0) and the B-bus is in the low-impedance sourcing mode (EB = 1), then when the transmission gate is enabled (EXG = 1), the A-bus output takes the same value as that present on the B-bus. Conversely, information can travel from the A-bus to the B-bus. The model for the B-bus output has not been shown because it is analogous to that of the A-bus: In Fig. 15, wherever the symbol A appears substitute B and vice versa. The fault-free behavior of the bus model as compared to that of the "actual" buses is given in Table III.

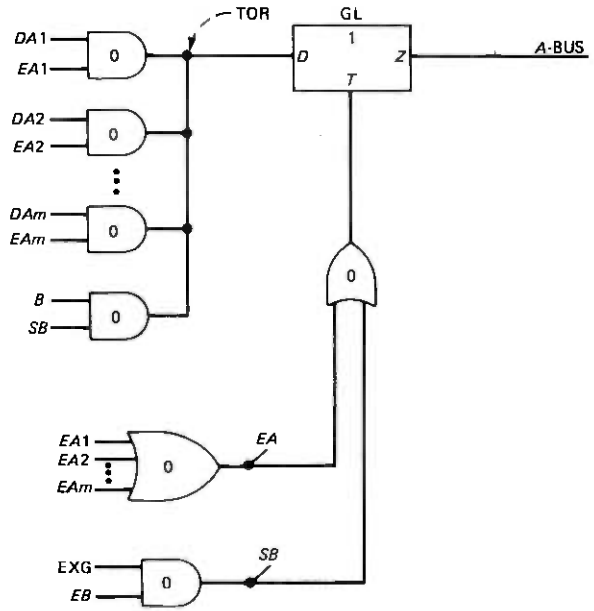


Fig. 15—The A-bus fault-free model. The unknown 3 condition is ignored here. The B-bus model is analogous.

Table III — Fault-free behavior of the coupled buses and their model

			Model		Actual Buses	
EXG	EA_i	EB_j	A	B	A	B
0	0	0	4	4	4	4
0	0	1	4	DB_j	4	DB_j
0	1	0	DA_i	4	DA_i	4
0	1	1	DA_i	DB_j	DA_i	DB_j
1	0	0	4	4	*	*
1	0	1	DB_j	DB_j	DB_j	DB_j
1	1	0	DA_i	DA_i	DA_i	DA_i
1	1	1	$DA_i + DB_j$	$DA_i + DB_j$	*	*

*

present		next	
A	B	A	B
0	0	0	0
0	1	3	3
1	0	3	3
1	1	1	1

EXG, EA_i, EB_j = enable control inputs
DA_i, DB_j = data inputs
A, B = bus outputs
3 = unknown 0 or 1

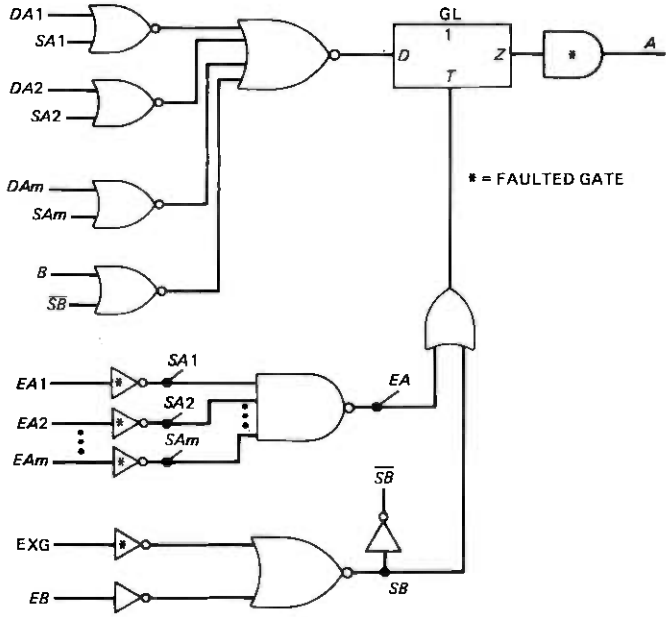


Fig. 16—The A-bus faulted model. Only gates marked * will be faulted by the simulator. Two SA0/SA1 node (pin) faults and $m + 1$ pairs of stuck-open/stuck-on faults have been assigned. The B-bus faulted model is analogous.

One possible faulted model for the A-bus is shown in Fig. 16. For multiple selections among the DAj and B the response is that of a 0-dominant short. The "3" output could be substituted, however. All elements are zero delay except the gated latch. The usual faults have been assigned to the model: two SA0/SA1 classical "pin" faults and $m + 1$ pairs of S-OP/S-ON faults. As before, the faulted model for the B-bus output is obtained by the symbolic interchange of As and Bs.

3.7 The programmable logic array (PLA)

The programmable logic array (PLA) is a simple method for implementing "random" combinational logic networks. An example of a three-variable, three output PLA is pictured in Fig. 17. The circuit is implemented in dynamic "pseudo NMOS." [7] That is, only the pull-up, or precharge, FETs are PMOS; all others are NMOS.

Clock $\Phi 1$ precharges the word and bit lines to the 1 state. Next, clock $\Phi 2$ causes the input signals x, y, z to propagate to the output terminals $W1$, $W2$, $W3$. (Although the two clock waveforms are essentially in phase, they are applied to PMOS and NMOS FETs, respectively, and the resulting conduction modes are 180° out of phase.)

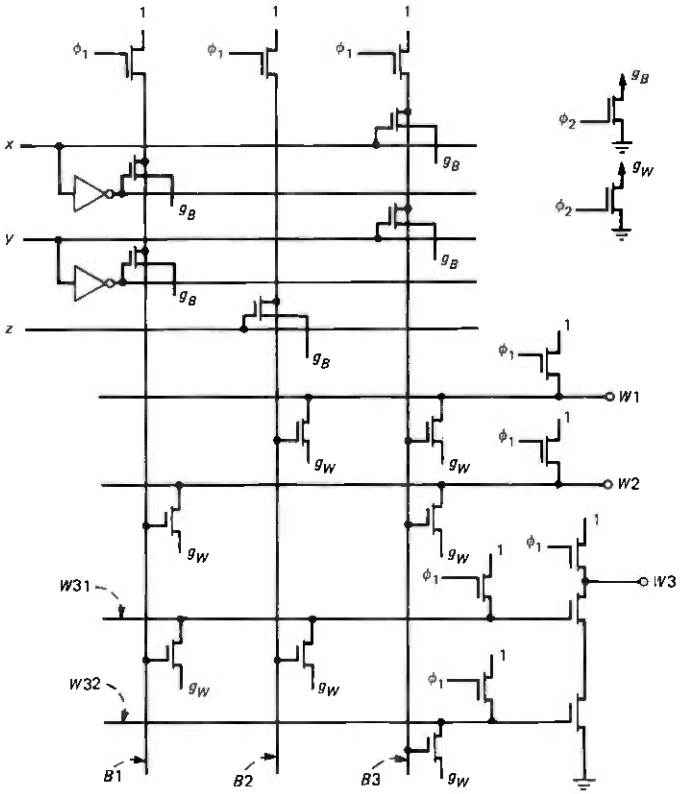


Fig. 17—Dynamic pseudo NMOS programmable logic array (PLA). There are three input variables (x, y, z) and three outputs ($W1$, $W2$, $W3$). Pull-up FETs are PMOS; all others are NMOS.[7]

The steady-state transfer functions are:

$$B1 = x \cdot y, \quad B2 = \overline{z}, \quad B3 = \overline{x} \cdot \overline{y},$$

$$\overline{W1} = B2 + B3, \quad \overline{W2} = B1 + B3, \quad \overline{W3} = W31 \cdot W32,$$

$$\overline{W31} = B1 + B2, \overline{W32} = B3.$$

Therefore,

$$\overline{W1} = \overline{z} + \overline{x} \cdot \overline{y},$$

$$\overline{W2} = x \cdot y + \overline{x} \cdot \overline{y},$$

$$\overline{W3} = \overline{x} \cdot y \cdot z + x \cdot \overline{y} \cdot z.$$

Models for the example PLA can be constructed by noting that the relation between inputs and bit lines and between bit lines and word lines is essentially that of a dynamic NOR gate. Figure 18 shows several

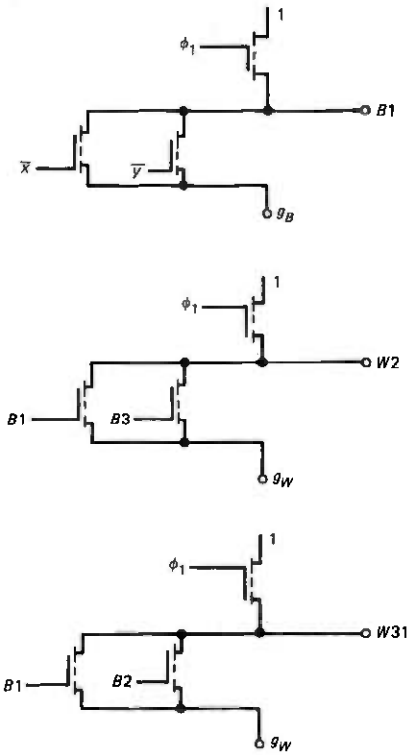


Fig. 18—Subcircuit portions of the PLA of Fig. 17.

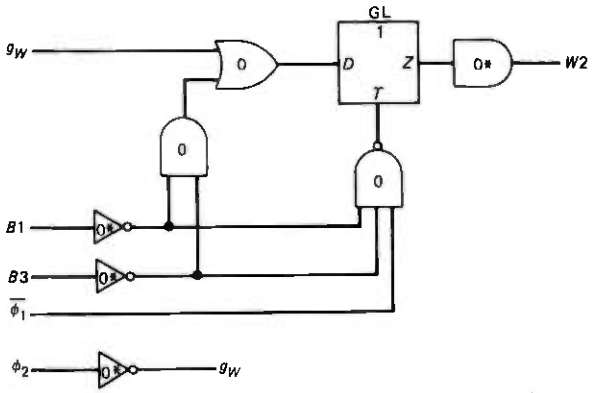
"subcircuit" portions of the PLA. Figure 19(a) gives the model for one typical subcircuit, $W2$. Figure 19(b) gives the model for the more complex $W3$ output. By this method the characteristics of any PLA, including dynamic properties, can be represented as a sum of simpler subcircuit models.

IV. THE GENERAL CASE

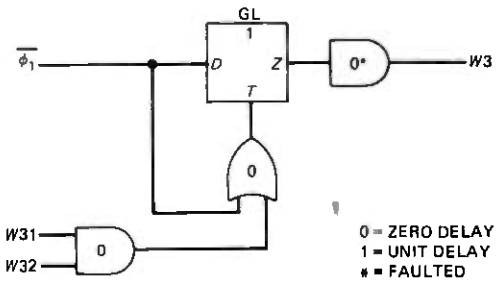
The preceding examples are particular illustrations of a more general configuration, shown in Fig. 20. The capacitance implicitly associated with the node F will be in one of four possible states: (*i*) low-impedance 0, (*ii*) low-impedance 1, (*iii*) high-impedance 0, or (*iv*) high-impedance 1. The first two conditions occur whenever one of the four input channels are enabled (conducting). The latter two cases arise when no circuit is enabled and the output is the resultant high-impedance state ($F = 4$).

In terms of the model, the logic state of the network is "transferred" to node F if and only if $T = 1$ where

$$T = \sum_i \mathrm{EX}_i \cdot \overline{\mathrm{EXSOP}}_i.$$



(a) W2 MODEL



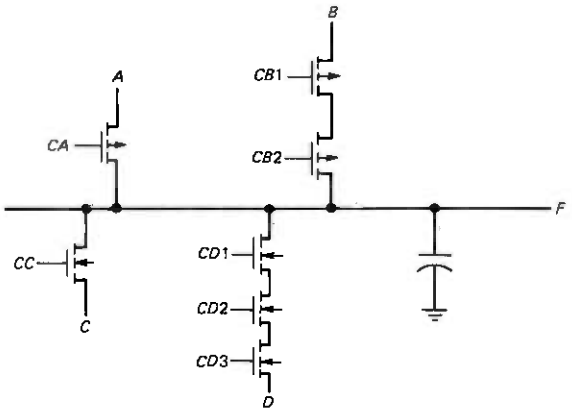
(b) W3 MODEL

Fig. 19—PLA subcircuit models. Gates marked * are faulted by the simulator, and 0 and 1 denote delays.

That is, at least one of the input circuits must be enabled (EXj = 1) and that circuit must not be stuck-open, i.e., EXSOPj = 0. Otherwise, T = 0 and F = 4.

If two channels, A and B, are simultaneously enabled ($A\&B$), then the result is said to be a 0-dominant short if $A\&B = A \cdot B$. For the 1-dominant short, $A\&B = A + B$. These two cases are shown in Figs. 21 and 22, respectively. Of course, if stuck-on faults are ignored entirely, then the network for T is represented by the summation immediately above.

Logic faults are assigned in the following manner. Simple n-input combinational gates are given the usual $n + 2$ stuck-at faults. In addition, one stuck-open fault is associated with each parallel branch of the pull-up/pull-down networks. Stuck-opens in strictly series paths to either VDD or VSS are ignored because of their similarity to SA0/SA1 faults, respectively. Unless the details of circuit technology and design dictate

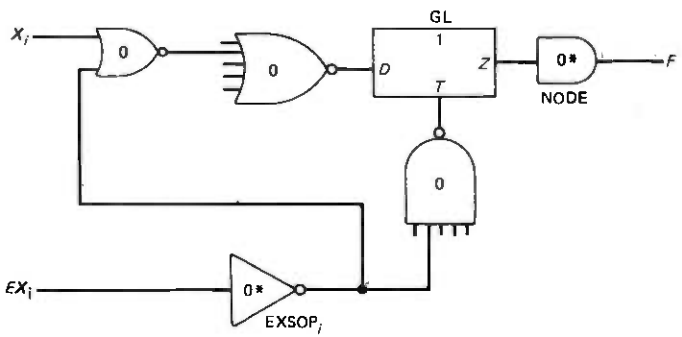


$$T = EA \cdot \overline{\text{EASOP}} + EB \cdot \overline{\text{EBSOP}} + EC \cdot \overline{\text{ECSOP}} + \cdots$$

WHERE $EA = \overline{CA}$, $EB = \overline{CB1} \cdot \overline{CB2}$
$EC = CC$, $ED = CD1 \cdot CD2 \cdot CD3$

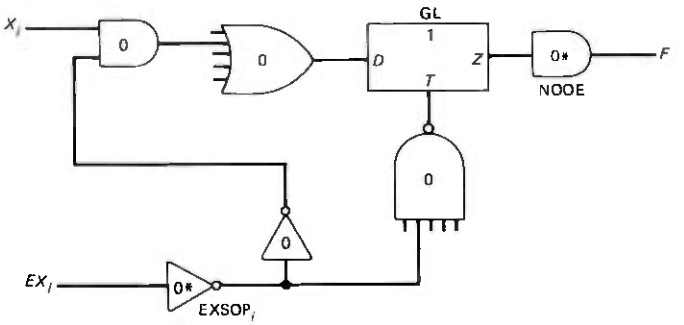
Fig. 20—The general case: multiple series/parallel combinations driving one node. FETs on the upper side are p-channel; those below are n-channel.

a clear choice, stuck-on logic faults should generally be disregarded. Nodes at which several tri-state elements connect can be assigned the two SA0/SA1 classical faults and one S-OP fault per channel. Fault conditions producing the unknown "3" state should not be modeled. On the



	T	D
NO FAULTS	$\sum_i EX_i$	$\prod_i (X_i + \overline{EX_i})$
$EXSOP_j(1)$	$\sum_{i \neq j} EX_i$	$\prod_{i \neq j} (X_i + \overline{EX_i})$
$EXSOP_j(0)$	1	$X_j \cdot \prod_{i \neq j} (X_i + \overline{EX_i})$

Fig. 21—The general case: S-OP and S-ON faults (0-dominant).



	T	D
NO FAULTS	$\sum_i EX_i$	$\sum_i X_i \cdot EX_i$
$EXSOP_j$ (1)	$\sum_{i \neq j} EX_i$	$\sum_{i \neq j} X_i \cdot EX_i$
$EXSOP_j$ (0)	1	$X_j + \sum_{i \neq j} X_i \cdot EX_i$

Fig. 22—The general case: S-OP and S-ON faults (1-dominant). This is a variant of the model of Fig. 21 except that the S-ON fault is 1-dominant.

other hand, simulation models for design verification ("true value" simulation) can include the "3" state as an indication of erroneous tri-state selection as in bus-oriented circuits.

V. SUMMARY

This paper has described a procedure for modeling faults which are a peculiarity of CMOS digital integrated circuits. Furthermore, the resultant methodology was also utilized to provide simulator models for complex MOS dynamic circuit topologies. The models are gate-level in structure and can be adapted for use on essentially any general purpose logic simulator. In addition, the models have been chosen to avoid faults which are artifacts of the model and which do not represent physically likely logic defects. The models have also been structured to preserve the distinction between classical and nonclassical faults. In addition, all models are designed to avoid races and circuit oscillations.

From the examples in this paper, it can be seen that there are a number of choices that can be taken for modeling a given logical function. In addition, a particular function may well be reduced in gate-count from the examples and the "general" realizations shown in this paper. That is, depending upon the selection of characteristics most important to the user, different models will result. In that spirit, the illustrations used above were selected to demonstrate the principles of modeling in a clear, straightforward manner.

REFERENCES

1. "LAMP: Logic Analyzer for Maintenance Planning," B.S.T.J., *53*, No. 8 (October 1974), pp. 1431–1555.
2. T. G. Athanas, "Development of COS/MOS Technology," Solid State Tech., *17*, No. 6 (June 1974), pp. 54–59.
3. H. Rombeek and P. Wilcox, "Interactive Logic Simulation and Test Pattern Development for Digital Circuitry," Electro 76, Paper 26.2, April 1976.
4. R. H. Krambeck, private communication.
5. The first successful demonstration of the FDL capability was performed for the NOR gate faults in Table I, R. E. Strebendt, private communication.
6. The logic symbol was suggested to the author by D. L. Kushler,
7. J. A. Cooper, J. A. Copeland, R. H. Krambeck, D. C. Stanzione, and L. C. Thomas, "A CMOS Microprocessor for Telecommunications Applications," ISSCC 77, Paper THPM 12.3, February 17, 1977.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Fault Coverage in Digital Integrated Circuits

By R. L. WADSACK

(Manuscript received October 10, 1977)

A theoretical expression is derived in this paper that evaluates the effectiveness of a set of logic tests for digital integrated circuits. The validity of the proposed figure of merit is examined with experimental data from CMOS integrated circuits. In addition, the importance of simulating the nonclassical stuck-open/stuck-on CMOS logic faults is also studied.

I. INTRODUCTION

The ever-growing complexity of digital integrated circuits places increasing emphasis upon the use of computerized design aids. Because no integrated circuit design is complete without an accompanying set of tests, one essential tool is the logic simulator.[1] The two principal reasons for logic simulation are (*i*) to verify the logic design and (*ii*) to develop the set of tests. A third purpose, related to the second, is that of diagnosis, i.e., identification of logic faults causing specific yield problems.

This paper will address itself to a study of the relation between fault coverage and measured yield and will consider specifically CMOS integrated circuits. The latter choice was made for two reasons. First, CMOS ICs are an attractive choice for many system designs. Second, CMOS ICs can possess nonclassical logic faults peculiar to MOS circuit elements: stuck-opens and stuck-ons.[2]

To verify the logical behavior of the IC, the test engineer usually begins with binary "vectors," or test patterns, that test the basic input/output logic functions of the circuit. For example, if the IC is a multiplexer, then multiplexing different data patterns is a natural starting point. Designing a set of vectors for high fault coverage generally represents a larger challenge than that of design verification. The difficulty arises because the logical structure of the IC must be tested and not just its generic properties, such as multiplexing. The major disadvantage of "behavioral"

tests is that they are usually too lengthy.[3,4] Consequently, the simplest approach is to begin with a sequence of representative behavioral tests and then add to them the necessary "structural" tests to bring the fault coverage to the required level. In any event, the process of developing the digital tests should start as soon as the systems logic design is formulated and before the design reaches the mask layout phase.

II. FAULT COVERAGE AND MEASURED YIELD

Perhaps one of the most important questions for test vector development is: How much fault coverage is enough? The answer must obviously be related to the intrinsic yield of the IC under test. For example, if the yield is 100 percent, then any low-coverage vector set can be used, including none at all. On the other hand, if the yield is low, then there will be many defective ICs that can potentially masquerade as "good" devices if the fault coverage is poor. In the latter case, the lower the fault coverage the more probable it will be that a chip that tests "good" contains a fault.

In any event, the measured functional yield,* ym, is the sum of two components: the actual functional yield, y, and the yield of bad ICs tested "good," ybg. Thus, $ym = y + ybg$. This leads to a second question: How is ybg related to the fault coverage f? Consider the following definitions for the functional yield problem:

y = actual functional yield (good chips).
ym = measured functional yield.
$1 - y$ = yield of bad chips.
yi = yield of chips with i faults ($i = 1,2,3 \cdots$).
ybg = yield of bad chips that test good.
$ybg(i)$ = yield of bad chips, with i faults, that test good.
f = fault coverage ($0 \leq f \leq 1$).
yr = field reject rate due to functional defects.

First, assume that $ybg(i) = (1 - f)^i \cdot yi$. For $i = 1$, this implies $ybg\ (1) = (1 - f) \cdot y1$, which is quite reasonable because it represents the basic assumption behind most logic simulation. That is, a logic simulator considers only the population of all chips in which there is only one logic fault per chip. Second, if the fault coverage is f, then $(1 - f)$ is the fraction of all faults (chips) that will be undetected by the test sequence. For $i = n > 1$, the assumption amounts to stating that the probability of n faults being undetected is $(1 - f)^n$. This is true only if multiple faults are independent of one another.

* The yield of chips that are free of logic faults irrespective of their analog voltage/current behavior.

The second assumption is that $yi = y \cdot (1 - y)^i$ $(i = 1,2,3,\cdots)$. This is the geometric distribution function and has been found to correctly describe the distribution of defective cells in static RAM chips.[5] In particular, the average yield y and the average number of fault-producing defects per chip, x_0, were measured and found to be related by the equation $y = 1/(1 + x_0)$. In the case of the RAM chips, $x_0 = 2.7$. For defect densities higher than, say, 5 per chip, a different distribution may be necessary.

Under the above assumptions,

$$ybg = \sum_{i=1}^{\infty} (1 - f)^i \cdot y \cdot (1 - y)^i$$

and

$$ybg = y\left[\frac{(1 - f)(1 - y)}{1 - (1 - f)(1 - y)}\right].$$

Therefore,

$$ym = \frac{y}{1 - (1 - f)(1 - y)}.$$

The field (or incoming inspection) reject rate yr is determined by the fraction of bad ICs that passed the functional test vector sequence, but that would have failed had the fault coverage been higher.

Therefore,

$$yr = ybg/ym,$$

which gives

$$yr = (1 - f)(1 - y).$$

(Note that in this context *undetectable* faults are not included in the statistical base for fault coverage. The most likely negative consequences of undetectable faults are long-term reliability problems or intermittents, not failures at incoming inspection.)

Inversely, for a given field reject rate (or "quality level") the fault coverage would be

$$f = 1 - \frac{yr}{1 - y}.$$

As an example, for an IC with a yield of 20 percent ($y = 0.2$), the fault coverage would have to be equal to or greater than 98.8 percent for a reject rate of 1 percent ($yr = 0.01$) or lower due to undetected logic faults.

III. FAULT COVERAGE AND LOGIC SIMULATORS

The percent fault coverage quoted for a set of test vectors is an important measure of their test effectiveness. Other things being equal,

a vector sequence with 90 percent coverage is twice as likely to identify faulty ICs as one with only 45 percent coverage. Equally important, however, is the question: Ninety percent of what? Unfortunately, the answer is usually 90 percent of what faults the simulator simulates. Therefore, in comparing different simulations and quoted fault coverage, it is essential to know the kinds of faults that were modeled. Many simulators model only classical faults (stuck-at-1 and stuck-at-0). Others may treat only gate *output* stuck-at faults. It is common in the case of printed circuit board simulations to consider only the pin faults of each IC on the board. In the experimental results of the next section, all classical faults were modeled in addition to the relevant CMOS stuck-open and stuck-on faults.[2]

Second, to lower costs, simulations generally use only a "collapsed" set of faults. That is, several faults on different gates may cause the same faulted circuit behavior as viewed from the primary circuit outputs. Consequently, they are included in a single fault equivalence class with only one fault in that class being simulated. As an example, a chain of three inverters would have six physically distinct classical faults. However, after fault collapsing, only two faults would be simulated. The obvious drawback to fault collapsing is that it distorts the relation between the predicted and the observed number of failures.

An additional factor can alter the ratio of predicted to observed failures: the probability of and relation between physical faults and simulator faults. First, not all physical faults are equally probable. Second, an individual physical fault does not necessarily produce a single logical fault, i.e., one fault may map into two or more simulator faults or vice versa. In addition, the distribution function for physical defects produces many more chips with multiple faults than chips with only one fault. Obviously, this could pose a problem in the interpretation of failure data because fault simulators simulate only singly faulted circuits, not those with multiple faults. The effect of gross physical faults and the preponderance of chips with multiple faults is to cause a higher number of failures during the initial part of the test sequence compared to that predicted by the simulator.

The simulator can also contribute its own distortions. For example, it is clear that undetected faults that caused the simulation to oscillate may well cause an actual integrated circuit to fail during testing. In the same fashion, the simulator can be overly pessimistic with respect to other faults that produce or leave unknown states in the circuit (e.g., set/reset inputs to flip-flops). Conversely, the simulator may treat a particular fault as having been detected on a specific vector, but the fault causes the IC to fail on a following vector. This can occur because of differences between the discrete delays of the simulator and the actual delays present in the IC.

Table I — Circuit characteristics

Circuit	Inputs	Outputs	Gates
D flip-flop	4	2	15
Multivibrator	8*	6†	47
MUX	14	7	238

* Includes one as I/O and another for I/O control.
† Includes one as I/O.

The above are some of the more evident reasons why fault simulator results are only approximations to those actually measured on integrated circuits. This is true even if the simulator modeled all reasonable types of logic faults.

IV. EXPERIMENTAL RESULTS

Three CMOS integrated circuits were selected for studying the relation between fault coverage and yield. Two of these circuits were studied in some detail. The two circuits are (*i*) a dual D flip-flop functionally equivalent to the RCA CD4013A and (*ii*) a monostable/astable retriggerable multivibrator similar to the RCA CD4047A.

Table I gives the circuit characteristics of the circuits. The column labeled "gates" gives the gate count for each circuit with the convention that a node to which two or more transmission gates connect is counted as one logic gate. Table II summarizes the fault characteristics of each circuit. The circuits were modeled to include the nonclassical stuck-open/stuck-on CMOS faults.[2]

Although CMOS faults double the number of total faults, it is not obvious whether they should be counted on a 1:1 basis with classical faults. If the probability of occurrence of stuck-open/stuck-on faults is markedly different than that of SA0, SA1 faults, then a weighting factor different than unity should be used to determine the total number of "effective" faults. As a second consideration, classical faults are collapsible into equivalence classes, but the CMOS nonclassical faults are individualized to single gates.

Table II — Fault characteristics

Circuit	(1) Physical Gates	(2) Faults* Classical	(3) Faults* CMOS	(4) Total	(5) Total Faults per Gate
D-FF	15	38	38	76	5.1
MULTI	47	133	134	267	5.7
MUX	238	902	536	1438	6.3

* After fault collapsing.

Table III — Fault coverage results

Circuit	(1) No. of Gates	(2) Faults per Gate	(3) No. of Vectors	(4) Vectors per Gate	(5) Fault Coverage* Classical	(6) CMOS	(7) Total
D-FF	15	5.1	15	1.0	100%	100%	100%
MULTI	47	5.7	119	2.5	95	84	89
MUX	238	6.3	5549	23	95	90?	93?

* All faults, including undetectables.

4.1 The D flip-flop circuit

Not surprisingly, the D flip-flop circuit is 100 percent testable for both classical and CMOS logic faults (see Table III). Figure 1 shows the cumulative total fault coverage versus the fraction of the test vector sequence applied to the IC. In this case, 15 vectors were used to reach 100 percent coverage. Strictly speaking, the points in Fig. 1 should have been connected by step functions that rise to meet each datum point. For the sake of clarity, however, straight line segments running directly from one point to the other were used.

Figure 2 shows the relation between the total fault coverge f(total) and the classical fault coverage f(class). The total fault coverage falls below that for classical faults because of a characteristic lag in CMOS fault

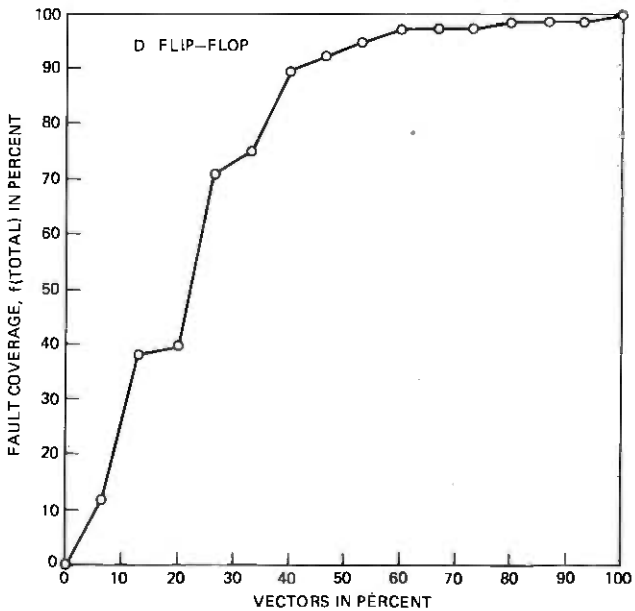


Fig. 1—The D flip-flop: total fault coverage vs. normalized vector number (15 vectors total).

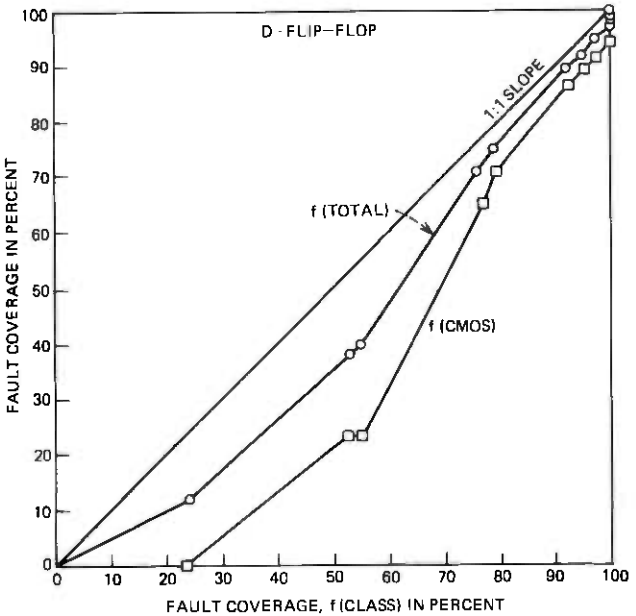


Fig. 2—The D flip-flop: total fault coverage and CMOS fault coverage as functions of the classical fault coverage for the vector sequence of Fig. 1.

coverage. The CMOS lag is more clearly shown by the second curve, marked *f*(CMOS), in Fig. 2. The lag is caused by the "history-dependence" of CMOS stuck-open faults that require at least two different vectors for detection.

Figure 3 shows a comparison between the simulation data and actual measurements for 11,150 ICs from 28 wafers. It is important to note that the curves of Fig. 3 are *reverse* cumulative distribution functions.[6] The independent variable is the vector number. Fifteen vectors were used to test this circuit. Vector 1 was the first in the sequence and is shown at the origin of the graph. The dependent variable is the running sum of the number of detected faults (or chip failures) beginning with vector 15 and proceeding to vector 1. Hence, the curves indicate the likelihood that a defective chip will fail at a specific vector.

In the simulation curve, the cumulative number of detected faults is shown. For the wafer data the cumulative number of functional, or logic test, failures is plotted. Reverse cdf's were chosen because they reveal the structure of the tail regions where the fault coverage is not changing as rapidly as it is near the beginning of the vector sequence. Of course, the tail of each curve illuminates most clearly the effects of ICs with single faults.

The simulation data are the same as those used for Fig. 1 in which all

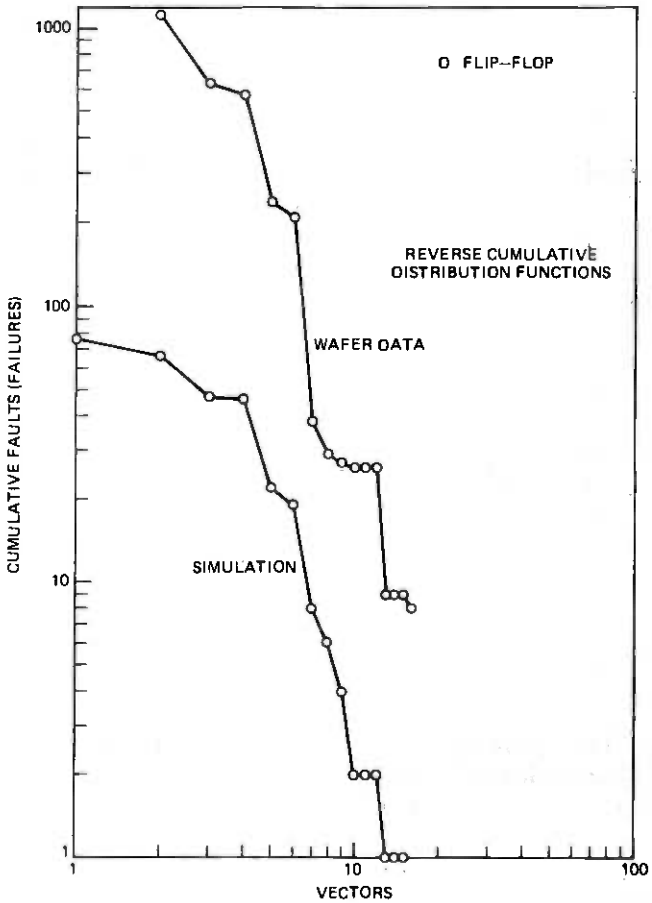


Fig. 3—The D flip-flop: reverse cumulative distribution functions for functional yield loss plotted as functions of the vector number. The simulation data are that of Fig. 1. The measured data points were obtained from 11,150 chips from 28 wafers.

faults, both classical and CMOS, are included. There is reasonable agreement between the "predicted" and the measured rcdf's. Only by coincidence would the two curves lie one upon the other because each is plotted in absolute numbers. Ideally, of course, they would be separated by a constant vertical displacement. The agreement is one of general shape. A specific quantitative comparison is treated later in this paper.

Naturally, near the beginning of the sequence there are more actual IC failures than indicated by the simulation. Recalling the factors discussed in Section III above, initial failures are probably caused by gross shorts, opens, and multiple faults. (All chips, however, were prescreened

for contact failures.) The slight dip in the wafer data for vectors 7,8,9 is caused by a 2:1 overprediction of fault coverage. Vectors 7,8,9 detect, in a worst-case sense, set and reset faults which in a real IC are more likely to fail earlier in the vector sequence.

Only one fault is detected by vector 12, the data input transmission gate stuck-on. The failure rate on that vector was 17/11150 or 0.15 percent. The only other vector that detects a single CMOS fault is vector 15. The fault is a stuck-open in the master flip-flop feedback transmission gate. The failure rate was 1/11150, or 0.009 percent, quite low compared to the stuck-on fault. Strangely, there are 8 ICs (0.07 percent) that failed at an added vector 16 where the fault coverage is already at 100 percent. Vector 16 forces the set and reset inputs active at the same time. Although the behavior of the fault-free circuit is deterministic, no structural faults (classical, stuck-open, or stuck-on) remain in the circuit to be detected at that vector. The failures may have been caused by analog effects.

The predicted field reject rate yr, as a function of fault coverage f, can be computed from the data of Figs. 1 and 3. The procedure begins first with Fig. 3 where the measured yield is obtained as a function of the vector at which the sequence was truncated. Next, the fault coverage for the truncated vector set is established by reference to Fig. 1. Of course, if the entire untruncated vector set is applied to the IC, the cumulative fault coverage at the last vector is 100 percent and the measured yield ym should equal the true yield y.

The results of the above calculations are shown in Fig. 4. The theoretical relation, $yr = (1 - f)(1 - y)$, is the solid line. The two sets of data points represent the computed yr based, first, on all faults and, second, on only classical faults. The latter lie closer to the theoretical prediction.

Several conclusions can be drawn from the results of Fig. 4. The first is that the equation is a good estimator of the reject rate, but is somewhat pessimistic at high fault coverage. Second, the difference between the predicted and measured yr at high values of f can be explained by a proportionally larger actual fault coverage used as the abscissa (in each case). Finally, even though CMOS stuck-open/stuck-on faults were identified in some of the circuits, their relative probability seems to be much less than that of the more general classical stuck-at faults. In other words, a 1-to-1 weighting factor does not appear to be warranted.

4.2 The multivibrator circuit

The multivibrator was not 100 percent testable. The cumulative fault coverage for all faults is shown in Fig. 5. The occasional abrupt jumps in fault coverage are caused by sequential portions of the circuit. That

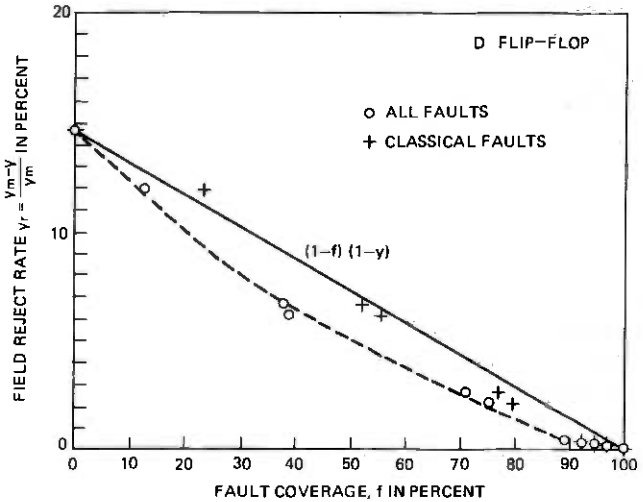


Fig. 4—The D flip-flop: field reject rate vs. total fault coverage as determined from the data of Figs. 1 and 3. The theoretical expression is indicated by the solid line.

is, several vectors are needed before a "fault effect" is generated at a gate and several more are necessary to propagate the fault to an output. When faults from that part of the circuit finally reach an output, there is a sharp increase in coverage.

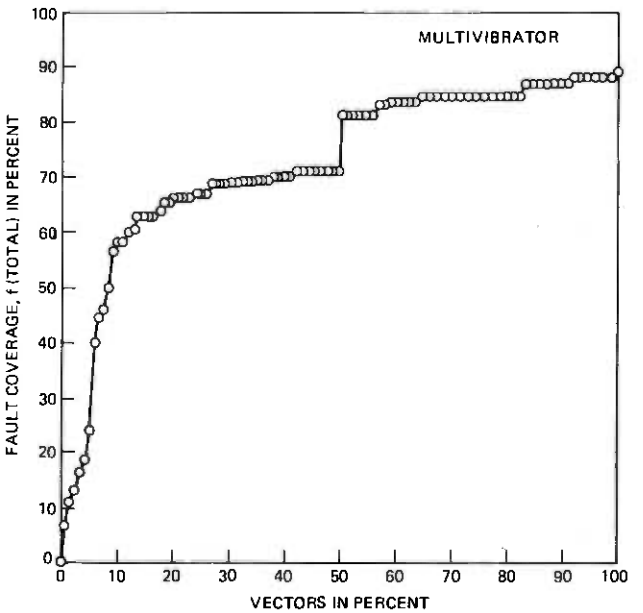


Fig. 5—The multivibrator: total fault coverage vs. normalized vector number (119 vectors total).

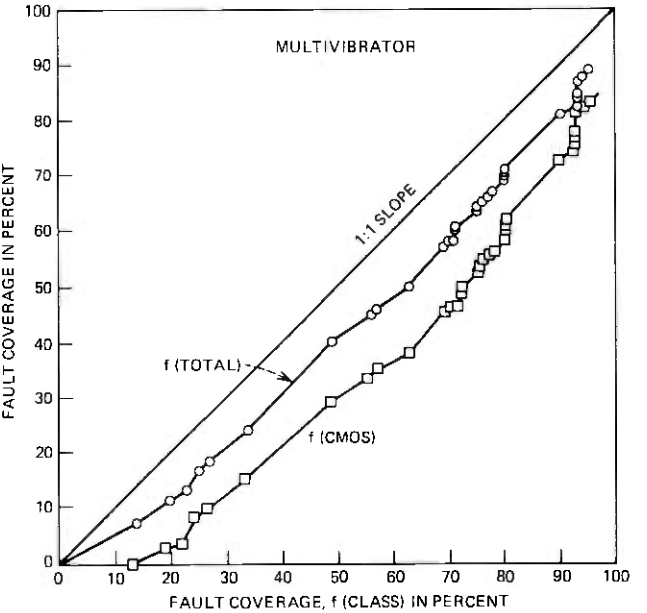


Fig. 6—The multivibrator: total fault coverage and CMOS fault coverage as functions of the classical fault coverage for the vector sequence of Fig. 5.

The final vector set provided a coverage of 89.1 percent, or 238 faults out of 267. The remaining 29 faults are all undetectable. There were two primary causes for the undetectable faults. The first was the presence of "asynchronous" circuit behavior (in the retrigger control section). The second was the use of two D flip-flops which had data inputs tied permanently to a fixed logic value (lack of controllability). Two undetectables occurred in a NOR latch: all CMOS latches formed by cross-coupled NOR or NAND gates have two undetectable faults.

Figure 6 shows the relation between the total fault coverage f(total) and the classical fault coverage f(class). Again the lag in CMOS fault coverage is evident. The total fault coverage reaches 89.1 percent. Classical coverage is 94.7 percent; CMOS coverage is 83.6 percent. Of the 29 undetectable faults, 22 were CMOS and 7 were classical.

Figure 7 compares simulation data with measurements taken from a single wafer. The wafer contained 418 chips which passed initial contact tests. Of those, 275 passed the logic tests for a gross functional yield of 66 percent. Again, the reverse cdf's are used to show the behavior in the tail regions near the end of the vector sequence. The overall agreement between the two curves is reasonable.

As in the above D flip-flop example, the field reject rate yr can be calculated for the multivibrator circuit from the corresponding data

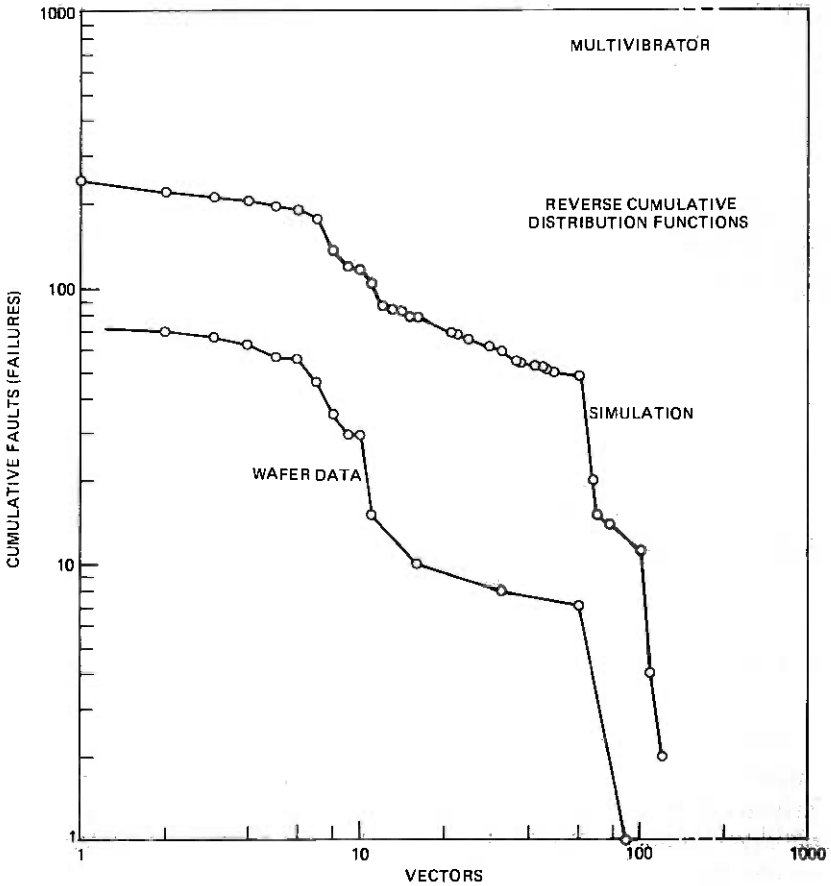


Fig. 7—The multivibrator: reverse cumulative distribution functions of the vector number. The simulation data are that of Fig. 5. The measured data points were obtained from 418 chips from one wafer.

(Figs. 5 and 7). The resultant points are shown in Fig. 8. The solid line is the predicted relation $yr = (1 - f')(1 - y)$, where f' is the fault coverage for all detectable faults.

The data of Fig. 8 suggest the same observations as for the D flip-flop: The actual fault coverage appears to be higher toward the end of the vector sequence than that predicted by the simulator. Also, the theoretical yr is best matched by the "classical faults only" data. Again, this indirectly implies that the relative frequency of CMOS faults is significantly less than that of the classical faults. In addition, for both the D flip-flop and the multivibrator the curves are similar above the 75 percent fault coverage point. In particular, each indicates that a coverage of 85 to 90 percent is needed to achieve a reject rate of 1 percent or less.

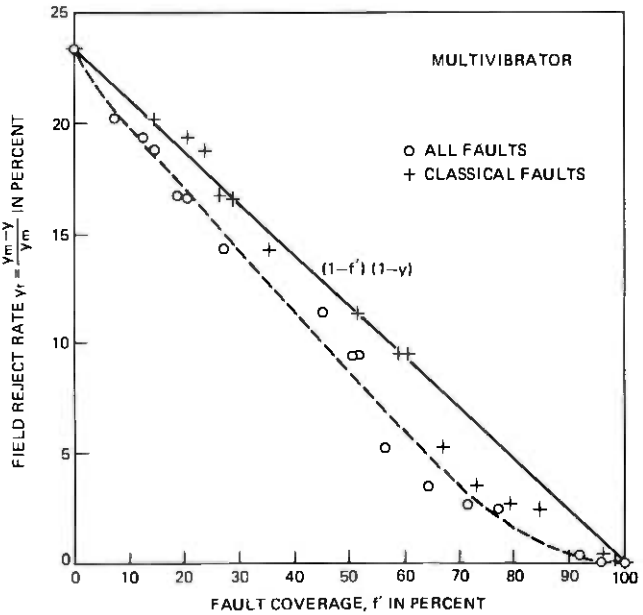


Fig. 8—The multivibrator: field reject rate vs. total (detectable) fault coverage as determined from the data of Figs. 5 and 7. The theoretical expression is indicated by the solid line.

On the other hand, the true yields for each are approximately equal (85 and 77 percent, respectively).

V. SUMMARY

The fault coverage results of the previous section have been summarized in Table III. To compare the difficulty of generating test vectors for one circuit versus another, a measure of circuit complexity is needed.[7] Gate count [column (1)] is a poor measure because it reflects only silicon area and not the interconnections that create the actual circuit. One potential measure of circuit complexity that indicates at least the magnitude of test vector generation is the ratio of the number of vectors to the number of gates [column (4)]. In that sense, the multivibrator is 2.5 times more complex than the D flip-flop, and the multiplexer (MUX) is 23 times more complex. Of course, a measure could be devised to incorporate the number of circuit inputs and outputs. However, for the three circuits the most dominant effect is that of the number of test vectors. The reader can readily interpret from Table III the magnitude of the test generation and fault coverage problems for large-scale silicon-integrated circuits with thousands of gates. In addition, for modern IC test equipment the number of circuit inputs and outputs

generally does not affect the functional test time (as long as the number is less than the test set maximum).

Nevertheless, there are two major drawbacks to using "vectors/gate" as a measure. First, it is retrospective: Only after effort has been expended to develop the test vectors does the "complexity" become known. The second reason is that it depends somewhat upon the method or skill used to generate the test vectors themselves. In particular, the test vector sequences used for the example circuits are certainly not unique. Nor is it likely that any of them is optimal in the sense of being the least number for the same level of coverage. The basic problem is that there probably isn't any simple one-dimensional measure of circuit complexity that is useful for a broad spectrum of circuit types.

The relative frequency of CMOS stuck-open/stuck-on faults appears to be significantly less than that of the classical stuck-at-0/stuck-at-1 faults. On the other hand, CMOS nonclassical faults do occur. Perhaps the best approach to resolving this quandary would be a study of many different CMOS ICs. The investigation would use vector sets of high diagnostic capability to determine which kinds of logic faults are important.

Finally, the data presented in this paper support the reject rate (quality level) concept as an answer to the question, "How much fault coverage is enough?" However, the total economic picture obviously must take into account the cost of developing the vectors and the cost of using (applying) them. Only when all three of the above factors are considered can the cost of integrated circuit testing be properly judged.

REFERENCES

1. "LAMP: Logic Analyzer for Maintenance Planning," B.S.T.J., *53*, No. 8 (October 1974), pp. 1431–1555.
2. R. L. Wadsack, "Fault Modeling and Logic Simulation of CMOS and MOS Integrated Circuits," B.S.T.J., this issue, pp. 1449–1474.
3. A. K. Susskind, "Diagnostics for Logic Networks," IEEE Spectrum, *10*, No. 10 (October 1973), pp. 40–47.
4. W. G. J. Kreuwels, "Structural Testing of Digital Circuits," Philips Tech. Rev., *35*, No. 10 (1975), pp. 261–270.
5. R. L. Wadsack, unpublished work.
6. M. B. Wilk and R. Gnanadesikan, "Probability Plotting Methods for the Analysis of Data," Biometrica, *55*, No. 1 (1968), pp. 1–17.
7. J. Stephenson and J. Grason, "A Testability Measure for Register Transfer Level Digital Circuits," FTCS-6, Proceedings (1976), pp. 101–107.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Jitter Comparison of Tones Generated by Squaring and by Fourth-Power Circuits

By J. E. MAZO

(Manuscript received October 28, 1977)

The tone-to-jitter power ratio is calculated for some conventional methods of generating a tone at the pulse repetition frequency of a PAM data signal by operating on the latter by an appropriate nonlinearity. Attention is focused on this ratio for a fourth-order nonlinearity, which will produce a tone even in the absence of excess bandwidth, and on this ratio for a square-law nonlinearity for small excess bandwidth. If the excess bandwidth is less than about 50 percent, the fourth power is superior. In particular, it yields a 10-dB improvement for 12 percent roll-off and binary data.

I. INTRODUCTION

Successful detection of the symbols in a pulse-modulated waveform requires a knowledge of the pulse repetition period T. Specifically, if the signal $s(t)$ is of the form

$$s(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT) \tag{1}$$

where a_n are independent equiprobable binary symbols having values ± 1, knowledge of T is required for proper sampling of $s(t)$ to recover the a_n. Due to small differences in transmitter and receiver oscillators, *a priori* information concerning T is not usually sufficient, and constant updating of the precise current value of T is required. Often one prefers to deduce such information directly from eq. (1), rather than directly transmitting a tone at frequency $1/T$ Hz.

A very popular method is to pass $s(t)$ through an appropriate nonlinear circuit (e.g., a square-law) so that a tone is generated at frequency $1/T$.[1] For example, using the square-law operation we have the following identity:

$$s^2(t) = \left(\sum_{-\infty}^{\infty} a_n g(t - nT) \right)^2 = \sum_{n=-\infty}^{\infty} g^2(t - nT) + \sum_{n,m}^{n \neq m} a_n a_m g(t - nT) g(t - mT) \tag{2}$$

where in writing (2) we have used $a_n^2 = 1$.† The essential features of (2) are made clear on noting the Poisson sum formula‡

$$\sum_{n=-\infty}^{\infty} f(t - nT) = \frac{1}{T} \sum_m e^{2\pi i m t/T} F\left(\frac{2\pi}{T} m\right). \tag{3}$$

Thus the first term of the last member of (2) consists of a series of sine and cosine terms given by the right member of (3), where in (3) $f(t)$ is replaced by $g^2(t)$. In our particular case, we are especially interested in the terms corresponding to frequency $\omega = 2\pi/T$ rad/sec, i.e., the terms

$$\frac{1}{T}\left[e^{2\pi i/T} F\left(\frac{2\pi}{T}\right) + e^{-2\pi i/T} F\left(-\frac{2\pi}{T}\right)\right] \tag{4}$$

where, explicitly

$$F(\omega) = \int_{-\infty}^{\infty} g^2(t) e^{-i\omega t} dt. \tag{5}$$

In the hardware, these terms are isolated by a very narrow postfilter (of bandwidth B Hz, say) approximately centered about this frequency.

In order that $F[\pm 2\pi/T] \neq 0$, "excess bandwidth" is required for the pulse $g(t)$, that is, its frequency spectrum must extend beyond the Nyquist frequency $\omega = \pi/T$. The percent of excess is usually referred to as the "rolloff."

Once $g(t)$ is given, the tone power is easily evaluated using (4) and (5). Thus, for simplicity, assume the $g(t)$ is an ideal Nyquist pulse having the real Fourier transform $G(\omega)$ shown in Fig. 1, where percent of rolloff equals $100 \times \alpha$.

Using the general formula

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega') G(\omega - \omega') d\omega' \tag{6}$$

we note how $F(2\pi/T)$ depends on $G(\omega)$ only in the rolloff region

$$\frac{\pi}{T}(1 - \alpha) \leq |\omega| \leq \frac{\pi}{T}(1 + \alpha).$$

For the special case of $G(\omega)$ given in Fig. 1 we calculate

$$F\left(\pm \frac{2\pi}{T}\right) = \frac{\alpha T}{4}, \tag{7}$$

† For a multilevel situation we would replace $a_n^2 = 1$ with $E(a_n)^2$, where E denotes statistical expectation.

‡ In (3)

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp[-i\omega t] f(t) dt$$

is the Fourier transform of $f(t)$.

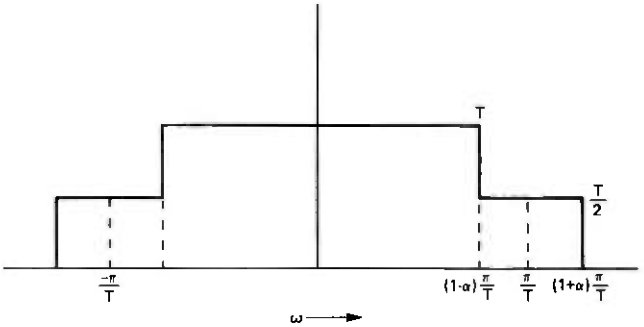


Fig. 1—Fourier transform $G(\omega)$ of the Nyquist pulse $g(t)$ used for the calculations.

allowing (4) to be rewritten as

$$\frac{\alpha}{2}\cos\frac{2\pi t}{T} \tag{8}$$

Thus from (8),

$$\text{square-law tone power} = \frac{\alpha^2}{8}. \tag{9}$$

Tone power is not a sufficient measure of performance, of course. This power must be compared to the power in the background noise. One component of this background would be additive noise on the channel. This is usually negligible, however, and the background to which we refer is self-generated. Mathematically it is given by the last term in (2). This added background will cause the zero-crossings of (8) to be perturbed, or "jitter," about their nominal values. Computing the tone power/background power ratio for various situations is the purpose of our work.† In addition to the state of affairs just described, two other proposals are of considerable practical interest. These may be conveniently and descriptively described as

(*i*) Prefiltering.

(*ii*) Fourth power law.

To motivate the first, recall that the tone power is determined by the "overlap" of the excess bandwidth regions in the integral (6). This power will not be changed if we filter out the remaining central portion of the pulse (the region $|\omega| < (\pi/T)(1-\alpha)$ in Fig. 1) before we do the squaring operation. The elimination of this portion of $G(\omega)$ does decrease the *total* power in the background term of (2). Will it improve the tone-to-jitter ratio in the neighborhood of $\omega = 2\pi/T$ as well?

† Since the tone will be "picked-off" out of the background by a narrow-band filter or phase-lock loop, only the value of the background power spectrum at $\omega = 2\pi/T$ will be needed.

The motivation of the second situation also begins by mentioning the overlap contribution to the integral in (6). We note that the overlap, and hence the tone power (9) vanish as α vanishes. Thus for a squaring circuit no tone is produced if there is no excess bandwidth. We shall see later that this is not true for all nonlinearities, and in fact a fourth-power law will produce a strong tone even when $\alpha = 0$. Again, what about the jitter? Our calculation of the latter for the fourth power is a completely new result. To appreciate the difficulties involved here note that the time averaged autocorrelation function $R(\tau)$ for the output of the fourth-power law is given by

$$R(\tau) = E\frac{1}{T}\int_0^T s^4(t)s^4(t-\tau)dt \tag{10}$$

with $s(t)$ given by (1). A straightforward evaluation of the matematical expectation in (10) would involve dealing with the eighth-order terms

$$Ea_{n_1}a_{n_2}a_{n_3}\cdots a_{n_8}. \tag{11}$$

The bookkeeping involved with (11) would be unmanageable. One novelty of our method is the introduction of a technique from the algebra of symmetric polynomials[2] for skirting the direct evaluation implied by (11).

Since we are mainly interested in knowing if prefiltering or the fourth-power law can produce large improvements in tone-to-jitter ratio, our explicit evaluations will be based on simple pulse shapes. For prefiltering, overlap is important and we stick to the pulse shape with frequency characteristic given in Fig. 1. For fourth-order we assume that small excess bandwidth produces only higher-order corrections to the effect present when $\alpha = 0$. Consequently in this case we use only the $\alpha = 0$ pulse.

We shall show that prefiltering offers no improvement at all. With or without prefiltering the output tone-to-jitter ratio is about 12 dB if an output filter 10 Hz wide is used or 22 dB for one 1 Hz wide. This assumes a 12 percent excess bandwidth as in the Bell System 209 data set. Numbers for the fourth-power law are 10 dB better than this, which is a significant improvement.

Before proceeding, a final comment is in order. This concerns a recent publication of Franks and Bubrowski[3] concerning prefiltering and the square-law nonlinearity. Their claim is that if prefiltering has a symmetrical result about π/T and the post-filter is symmetrical about $2\pi/T$, there will be no jitter about the zero-crossings of (8). This is true, but if T were known exactly so that the required symmetrization could be done exactly, then there would be no need to measure T. If, however, we symmetrize about a $T' \neq T$, then the background, using a standard

representation of passband signals, would have the form

$$y(t) \cos \frac{2\pi}{T'} t \tag{12}$$

with no quadrature component relative to $2\pi/T'$. If $T' = T$ we see why the zeros of (8) are unchanged. If $T' \neq T$ the quadrature component of (12) relative to $2\pi/T$ will come in, with a strength *independent* of how small $T' - T$ is. Only the beat frequency depends on $T' - T$. Thus the Franks-Bubrowski result might be termed unstable and not applicable.

II. BACKGROUND SPECTRUM FOR SQUARING CIRCUIT

In this section we compute $S_c(2\pi/T)$, the value of the spectrum $S_c(\omega)$ of the jitter term which appears at the output of the squaring circuit at angle frequency $\omega = 2\pi/T$.† We do this for the special pulse of Fig. 1 with and without prefiltering. In terms of this quantity the tone-jitter ratio will be, for a final filter of bandwidth B, using (9),

$$\frac{\text{tone power}}{\text{background power}} = \frac{\alpha^2}{8} \Big/ 2S\left(\frac{2\pi}{T}\right) B. \tag{13}$$

The quantity $S_c(\omega)$ is the Fourier transform of the autocorrelation function

$$R(\tau) = \frac{1}{T} \int_0^T E[s^2(t)s^2(t-\tau)]dt \tag{14}$$

$$S_c(\omega) = \int_{-\infty}^{\infty} e^{-i\omega\tau} R(\tau) d\tau - \begin{pmatrix}\text{spectral}\\ \text{lines}\end{pmatrix} \tag{15}$$

where in (14) $s^2(t)$ is given by (2). Denoting $g(t - nT)$ by g_n and $g(t - \tau - nT)$ by h_n, the first item to evaluate is

$$Es^2(t)s^2(t-\tau) = E[(\Sigma a_n g_n)^2(\Sigma a_n h_n)^2], \tag{16}$$

the expectation being taken over the i.i.d. binary variables a_n, having values ± 1 with equal probability. The expectation in the right side of (16) can be done directly, using

$$Ea_p a_q a_r a_s = \delta_{pq}\delta_{rs} + \delta_{pr}\delta_{qs} + \delta_{ps}\delta_{qr} - 2\delta_{pq}\delta_{pr}\delta_{ps} \tag{17}$$

to yield

$$E[(\Sigma a_n g_n)^2(\Sigma a_n h_n)^2] = (\Sigma g_n^2)(\Sigma h_n^2) + 2(\Sigma g_n h_n)^2 - 2\Sigma g_n^2 h_n^2. \tag{18}$$

Looking ahead to the fourth-order case when we will need the average of eight order terms, we will not be able to write the analog of (17) in any

† The subscript c on $S_c(\omega)$ emphasizes that only the continuous portion of the complete spectrum is being considered.

Table I — Summary of the evaluation of background power for a squaring circuit

Coefficient	Term	Contribution to $S(2\pi/T)$ $\left(\begin{smallmatrix}\text{without}\\ \text{coefficient}\end{smallmatrix}\right)$
1	$(11)(aa)$	Tone term
2	$(1a)^2$	$\frac{\alpha T}{16}$
−2	$(11aa)$	$\frac{\alpha^2 T}{16}$

manageable way. Thus it will be pedagogically useful to introduce the new method here first, and evaluate (18) again. We first notice from the structure of the left side of (18) that only terms of the type $(\Sigma g^2)(\Sigma h^2)$, $(\Sigma gh)^2$ and $\Sigma g^2 h^2$ can occur on the right-hand side, i.e., the answer must be of the form

$$A(\Sigma g^2)(\Sigma h^2) + B(\Sigma gh)^2 + C(\Sigma g^2 h^2) \tag{19}$$

where A, B, and C are constants independent of whatever values the g_n and h_n take. Setting $g_n = h_n = \delta_{n0}$, the left-hand side of (18) is obviously unity. Then also using (19) we obtain the result that

$$1 = A + B + C. \tag{20}$$

Likewise setting $g_n = h_{n+1} = \delta_{n0}$ yields (since $g_n h_n = 0$, all n)

$$1 = A \tag{21}$$

while the choice $g_0 = h_0 = 1$, $g_1 = h_1 = 1$, $g_k = h_k = 0$, $k \neq 0,1$ provides

$$8 = 4(A + B) + 2C. \tag{22}$$

The solution of (20)–(22) is $A = 1, B = 2, C = -2$ in complete agreement with (18).

It is convenient to introduce the following shorthand: a sum $\Sigma_{n=-\infty}^{\infty}$ will be denoted by a parenthesis ( ). If the nth term of the sum is $g_n^p h_n^q$ the notation

$$(\underbrace{1\,1\cdots1}_{p\text{ times}}\,\underbrace{a\,a\cdots a}_{q\text{ times}})$$

is used. Thus the terms in (19) are of the types $(11)(aa)$, $(1a)^2$, and $(11aa)$ and the right side of (18) is given in the first two columns of Table I.

Having now obtained the sums and coefficients in (18), the next step is to evaluate the sums by the Poisson sum formula (regarding τ as a fixed parameter). Note first that, from (2), the term $(11)(aa)$ in (18) is due to

the deterministic part of (2) and hence the background terms are only

$$2(1a)^2 - 2(11aa). \tag{23}$$

After the Poisson sum formulas are evaluated we next perform

$$\frac{1}{T}\int_0^T (\ )dt$$

to eliminate the t-dependence. Finally the Fourier transform with respect to T is taken. For example,

$$(11aa) = \sum_{n=-\infty}^{\infty} g^2(t - nT)g^2(t - \tau - nT)$$

$$= \frac{1}{T}\sum_{m=-\infty}^{\infty} e^{2\pi i m t/T} F\left(\frac{2\pi}{T}m\right)$$

where $F(\omega)$ is the Fourier transform of $g^2(t)g^2(t - \tau)$. If $G_2(\omega)$ is the Fourier transform of $g^2(t)$ then

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega'\tau} G_2(\omega')G_2(\omega - \omega')d\omega'.$$

When we time-average $(11aa)$ only the $m = 0$ term survives, giving

$$\frac{1}{T}\int_0^T (11aa)dt = \frac{1}{T}F(0) = \frac{1}{2\pi T}\int_{-\infty}^{\infty} e^{i\omega'\tau} G_2(\omega')G_2(-\omega')d\omega'.$$

The Fourier transform of this with respect to τ is simply $(1/T)G_2(\omega)$-$G_2(-\omega)$ which is to be evaluated at $\omega = 2\pi/T$.

The actual contribution of the terms in (23) is listed in the third column of Table I, the same values being obtained with or without prefiltering. Thus from (9), (13), (23) and Table I we obtain for the squaring loop, with or without a prefilter,

$$\frac{\text{tone power}}{\text{background power}} = \frac{\alpha^2/8}{2B[2(1a)^2 - 2(11aa)]}$$

$$= \frac{1}{2BT}\frac{\alpha}{1-\alpha}, \alpha < 1. \tag{24}$$

The fact that (24) becomes infinite for $\alpha = 1$ is due to the fact that the spectrum has a zero then, and higher-order terms in B would be required to estimate the background power.

Applying these results to the 209 data set where $\alpha = 0.12$, $1/T = 2400$ sec^{-1}, (24) evaluates to about a 12 dB tone/filter ratio if B = 10 Hz, or 22 dB if $B = 1$ Hz.

In order to provide some contrast between the cases which do or do not involve prefiltering, the output spectrum in a neighborhood of $2\pi/T$

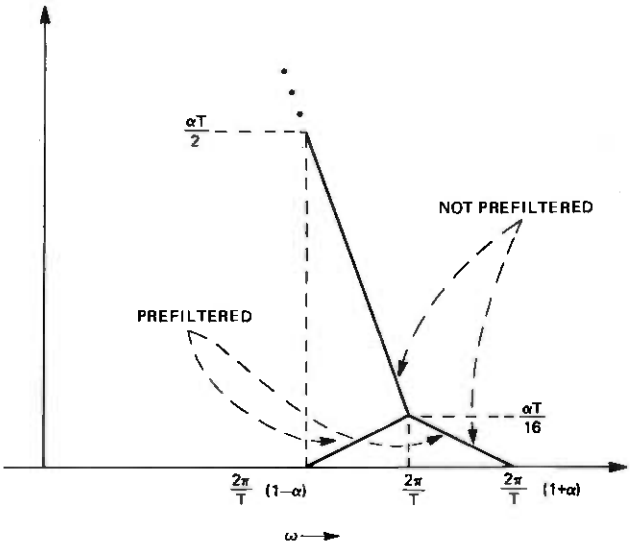


Fig. 2—Background power spectra for squaring circuit, with and without prefilter.

is given in Fig. 2 for small α, i.e. neglecting the $(11aa)$ contribution which is proportional to α^2.

The two spectra coincide above $2\pi/T$, but the divergence between the two on the lower side of the tone frequency is apparent from the curve.

III. BACKGROUND SPECTRUM FOR FOURTH-POWER CIRCUIT

The lines in the spectrum of $s^4(t)$ come from the deterministic terms, namely from $Es^4(t)$. Setting $g_n = h_n$ in (18) provides us with the evaluation

$$Es^4(t) = 3[\Sigma g^2(t - nT)]^2 - 2\Sigma g^4(t - nT). \tag{25}$$

If $\alpha = 0$ only the second term in (25) contributes a tone at $2\pi/T$. Since, when $\alpha = 0$,

$$2\Sigma g^4(t - nT) = \frac{4}{3} + \frac{2}{3}\cos\frac{2\pi t}{T},$$

the power in the tone is

$$\frac{1}{2}\left(\frac{2}{3}\right)^2 = \frac{2}{9}.$$

As before, the first step in the evaluation of the background spectrum is the calculation of

$$Es^4(t)s^4(t - \tau). \tag{26}$$

Table II — Summary of the evaluation of the background power for a fourth-power circuit

Coefficient	Term	Contribution to $S(2\pi/T)$ (without coefficient)
96	$(1111aa)(aa)$	0
256	$(111aaa)(1a)$	$T/20$
96	$(11aaaa)(11)$	0
24	$(1a)^4$	$T/6$
9	$(11)^2(aa)^2$	Tone term
72	$(11)(aa)(1a)^2$	0
4	$(1111)(aaaa)$	Tone term
64	$(111a)(1aaa)$	$T/120$
72	$(11aa)^2$	$T/30$
−6	$(1111)(aa)^2$	Tone term
−6	$(aaaa)(11)^2$	Tone term
−96	$(111a)(1a)(aa)$	0
−96	$(aaa1)(1a)(11)$	0
−72	$(11aa)(11)(aa)$	0
−144	$(11aa)(1a)^2$	$T/12$
−272	$(1111aaaa)$	$T/36$

Here the second technique introduced in Section II becomes decisive. Just as the results of the evaluation of (18) are summarized in Table I, the first two columns of Table II give the evaluation of (26).

We must next evaluate the contribution of the terms to $S_c(2\pi/T)$. Simplifications occur when $\alpha = 0$, since then

$$(11) = (aa) = 1 \text{ and } (1a) = \frac{\sin \frac{\pi\tau}{T}}{\frac{\pi\tau}{T}}.$$

The final result of applying the Poisson sum formula, averaging over t, and Fourier-transforming with respect to τ gives the results in the third column, Table II.

Collecting this we have, for a final filter of bandwidth B at $2\pi/T$, that

$$\frac{\text{signal power}}{\text{background power}} = \frac{5}{8BT}. \tag{27}$$

This is a significant improvement over the result (24) for the squaring-loop for small α. In fact, using $\alpha = 0.12$ again we calculate an improvement factor of 9.16 or close to 10 dB.

IV. ACKNOWLEDGMENT

The author wishes to thank J. Salz for pointing out the need for a comparison between the fourth power and squaring circuits, and for checking many of the calculations.

REFERENCES

1. W. R. Bennett, "Statistics of Regenerative Digital Transmission," B.S.T.J., *37*, No. 6 (November 1958), pp. 1501–1542.
2. M. Bocher, *Introduction to Higher Algebra*, New York: MacMillan, 1936, Chap. 18.
3. L. E. Franks and J. P. Bubrowski, "Statistical Analysis of PAM Timing Recovery," IEEE Trans. Commun., *COM-22*, July 1974, pp. 913–920.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

On Predictive Quantizing Schemes

By P. NOLL

(Manuscript received October 26, 1977)

*This paper analyzes the performance of various predictive quantizing schemes both for noiseless and noisy channels. The fidelity criterion used to define optimum performance is that of minimum mean-squared error. The first part of this paper compares differential pulse code modulation (DPCM) with a system that lacks the feedback around the quantizer. Such a system (that is called D*PCM in this paper) is actually a pulse code modulation (PCM) system with a prefilter and a postfilter. In the second part of this paper a noise-feedback coding structure is used as a framework for a unified analysis of predictive quantizing schemes with a frequency-weighted mean-squared error as the performance criterion. The last part of this paper extends the analysis to include the effects of channel transmission errors on the overall performance of these predictive quantizing schemes. It is shown that DPCM and D*PCM when appropriately optimized are less sensitive to channel errors than PCM, and that the performances of DPCM and D*PCM are almost identical in the case of high bit-error rates.*

I. INTRODUCTION

Predictive quantizing schemes employ prediction to exploit the inherent redundancy of input signals; the difference between the actual sample of an input signal and its estimate is quantized and transmitted to the receiver in a digital format. If the samples of the input are highly correlated, then the variance of the samples of the difference signal will be significantly less than the variance of the input samples. Hence, the overall error between input and output of the communication system will be lower than that of a conventional pulse code modulation (PCM) system. The first part of this paper compares PCM and two differential (predictive) pulse code modulation systems (see Fig. 1). A noise-feedback coding structure is then used as a framework of a unified analysis of predictive quantizing schemes (including those of Fig. 1) on the basis of a frequency-weighted error criterion. The last part of this paper ex-

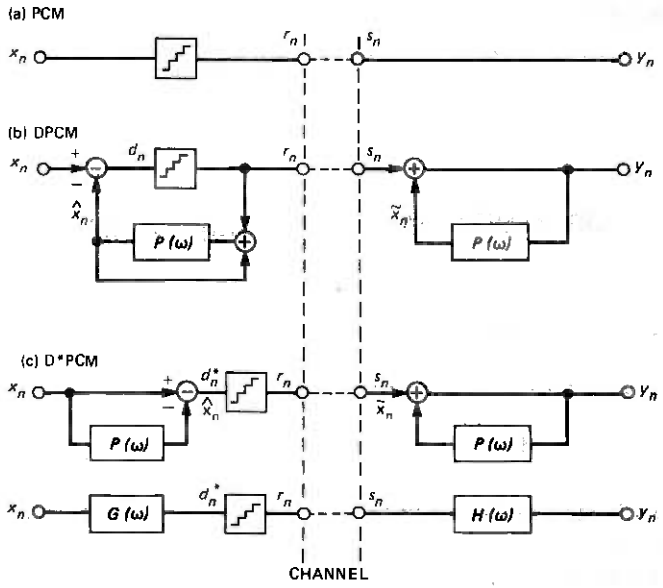


Fig. 1—Structures of PCM, DPCM, and D*PCM coders.

tends the analysis to include the effects of channel transmission errors on the overall performance of these predictive quantizing schemes.

The differential pulse code modulation (DPCM) system has a feedback around the quantizer resulting in a prediction that is based on previous *reconstructed* samples (Fig. 1b). The idea is to base the prediction on information that is also available at the receiver.[1] The D*PCM system is an open-loop quantization scheme,[2] i.e., previous samples of the input are used for predicting the actual input samples (Fig. 1c). The term D*PCM indicates that we have denoted by d^*_n the sequence of difference samples in the open-loop system. It is important to realize that coder and decoder calculate different prediction values. We also see, from Fig. 1c, that in this latter scheme the prediction network can be replaced with a more general linear network. The resulting scheme is then actually a PCM scheme with pre- and postfiltering and we shall use this latter scheme to derive bounds for the D*PCM performance.

Let us briefly discuss the differences between DPCM and D*PCM. We denote by $q_n = d_n - r_n$ and $q_n = d^*_n - r_n$ the quantization errors of DPCM and D*PCM, respectively, and by r_n the quantized versions of the difference samples.

It is easy to see that the total coding error

$$t_n = x_n - y_n \tag{1}$$

between encoder input x_n and decoder output y_n is given by

$$t_n = q_n \text{ for DPCM} \tag{2}$$

provided the channel is error-free and the decoder uses the same predictor of frequency response $P(\omega)$. On the other hand,

$$t_n = q_n + (\hat{x}_n - \tilde{x}_n) \text{ for D*PCM,} \tag{3}$$

i.e., the total error is increased by the difference between the prediction values $\hat{x}_n$ and $\tilde{x}_n$ of the predictors of encoder and decoder, respectively. Alternatively, we may write

$$t_n = q_n * h_n \text{ for D*PCM} \tag{4}$$

provided that the networks of encoder and decoder are reciprocal. In eq. (4), $*$ denotes discrete-time convolution, and h_n; $n = 0,1,2 \ldots$ is the impulse response of the linear decoder network. The result can be easily derived by recognizing that

$$\begin{aligned} t_n &= x_n - y_n \\ &= x_n - r_n * h_n \\ &= x_n - (d^*_n - q_n) * h_n \\ &= q_n * h_n + x_n - x_n * g_n * h_n \\ &= q_n * h_n \end{aligned}$$

since $g_n * h_n = \delta_n$ and $x_n * \delta_n = x_n$ (g_n is the impulse response of the encoder network and δ_n is the Kronecker delta).

Quantization errors are approximately white noise samples if the number of quantizer levels is sufficiently high. Hence in D*PCM the total error is nonwhite noise with each quantization error causing an infinite output sequence. This error propagation effect is also sometimes called error accumulation; this latter term, however, should be used cautiously, because it may seem to imply that D*PCM cannot give any improvement in signal-to-noise ratio (SNR) over PCM, or, even more strongly, that the overall performance can only be degraded.

One purpose of this paper is to analyze and explain the differences between DPCM and D*PCM. The D*PCM coding system has been analyzed by Bodycomb and Haddad[3]; their approach has been based on a predictor optimized for a minimum prediction error variance and they have shown that this predictor as a prefilter for a Gaussian process produces the same total error variance as a PCM system. We shall use the term MMSE predictor to describe such a predictor (that has been optimized for a minimum prediction error variance). We shall see very shortly that this D*PCM scheme can perform better than PCM provided that the

predictor is reoptimized; the MMSE predictor is not optimal in this context.

We have already mentioned that D*PCM can be viewed as a PCM scheme with pre- and postfilters. If we model the quantizer in the PCM scheme as an additive noise source, the optimization of this scheme is almost identical to a joint optimization of pre- and postfilters in communication systems in which channel errors are present. Many contributions have been made to this problem both for continuous-time and pulse-modulated communication systems.[4–13] One common result that can be extracted from these papers is that *half-whitening* of the input spectrum minimizes the overall mean-squared error (MSE) between input and output of a communication system with pre- and postfilters in which additive white noise (either caused by a quantization or by channel errors or by both effects) is present. Half-whitening is obtained if the magnitude-squared frequency response is inversely proportional to the square-root of the power density spectrum of the input signal.

PCM schemes with pre- and postfilters do not take into account that in systems with quantizers not only can the input spectrum be shaped to improve the overall performance but that, additionally, the quantization noise spectrum can be shaped as well to further improve the performance of the system. This problem has been discussed in detail by Kimme and Kuo[14] and later by Brainard and Candy[15] on the basis of different coder configurations. These coders have in common a feedback of filtered quantization noise to the input of the quantizer and we shall use the term *noise-feedback coding* (NFC) for this approach. Our analysis differs from earlier contributions in that a power constraint on the quantizer input is not needed. We offer a simplified solution based on well-known results in prediction theory.

It is the aim of this paper to discuss the differences between DPCM and D*PCM both for noise-free and noisy channels and to show how they relate to noise-feedback coding if the basis of the comparison is a frequency-weighted minimum mean-squared error. The organization of this paper is as follows: in Section II we calculate the differences in performance between predictive quantizing schemes with and without feedback around the quantizer. We show that D*PCM can perform better than PCM for all nonwhite input spectra, but that its performance is always below that obtainable with a DPCM scheme. A bound will be derived for the differences in performance between these two schemes, and a first-order Markov source will be used as an example to explain these differences. Section III analyzes noise-feedback coding; its structure is given by a prefilter followed by a quantizer with feedback around the quantizer.[14] We note that PCM, PCM with noise feedback, DPCM, and D*PCM are special cases of this configuration; thus a unified approach is possible. We optimize this coder with a frequency-weighted mean-

squared error as the performance criterion and show that the prefilter has to be a whitening filter for the input signal irrespective of the chosen frequency-weighting and that the feedback filter is the MMSE predictor of the weighted input spectrum. We also give frequency-weightings for which the noise-feedback scheme degenerates to DPCM or to D*PCM. Section IV extends the analysis to include the effects of channel transmission errors on the overall performance of predictive quantizing schemes. It is shown that the effects of these errors on total MSE can be significantly smaller than those in PCM. In D*PCM the optimum filters minimize simultaneously quantization and channel error variances and do not depend on the bit-error rate. In the case of DPCM a compromise is needed in order to minimize the total effect of both noise sources on the total MSE. At high bit-error rates the performances of D*PCM and DPCM are almost identical if their prediction networks are identical.

The analyses are made under certain restrictions; first of all, we use the MSE as a performance measure (in Section III a frequency-weighted MSE criterion will be taken into account). Second, the analyses are based on the assumption that the quantizer can be modeled as an additive white noise source. It is known that this model is accurate for quantizers with a sufficiently high number of levels. The model is a poor approximation, however, in the case of coarse quantization, especially if the quantizer input samples are highly correlated. The results of our analysis could be extended to these cases by modifying the model but we do not consider such an extension in order not to obscure the main results. We also assume that the variance of the quantization noise is much smaller than that of the signal to be quantized. Comparison with simulation results will show the range where our rather restrictive assumptions hold.

II. ANALYSIS OF PCM, DPCM, D*PCM

The principal aim in this section is to calculate and to compare the variance of the total errors in PCM, DPCM, and D*PCM. We assume that the input is a sample of a zero-mean stationary random sequence $\{x_n\}$ with autocorrelation function

$$R_x(k) = E[x_n x_{n+k}], \tag{5}$$

power density spectrum (pds)

$$S_x(\omega) = \sum_{k=-\infty}^{\infty} R_x(k) e^{-jk\omega}, \tag{6}$$

and variance

$$\sigma_x^2 = R_x(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) d\omega. \tag{7}$$

$S_x(\omega)$ will be assumed to be a rational function of ω. This is not really restrictive since most spectra of practical interest can be approximated by a rational function. The assumption implies that the pds can be represented by $S_x(\omega) = \eta_x^2 A(\omega)A^*(\omega)$ where $\eta_x^2 > 0$ is a scale factor and $A(\omega)$ is the ratio of two polynomials whose zeros are inside the unit circle of the z-domain (factorization theorem for rational spectra). All power density spectra are nonnegative continuous functions defined for $\omega = [-\pi,\pi]$. Since all signals are represented by stationary random sequences, all filters are necessarily of discrete-time type. The action of the quantizer is represented spectrally as white noise added to the quantizer input signal:

$$S_q(\omega) = \sigma_q^2 = R_q(0). \tag{8}$$

We are interested in the variance

$$\sigma_t^2 = E[t_n^2] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_t(\omega)d\omega \tag{9}$$

of the total error $t_n = x_n - y_n$ [eq. (1)], where $S_t(\omega)$ is the power density spectrum of the zero-mean random sequence $\{t_n\}$. A frequency-weighting of the total error can easily be added if necessary (see Section III).

2.1 PCM

The total error is identical to the quantization error. Hence we have

$$\sigma_t^2 = \sigma_q^2 = \epsilon_q^2 \cdot \sigma_x^2. \tag{10}$$

The quantizer performance factor ϵ_q^2 depends on the properties of the quantizer and on the probability density function of the signal being quantized; its value is the noise variance generated by a quantization of a unit-variance signal. Table IV in Section IV lists various values for the cases of 1-bit and 2-bit quantizers. For a given quantizer we thus find that

$$\min_{\text{PCM}} \{\sigma_t^2\} = \epsilon_q^2 \cdot \sigma_x^2. \tag{11}$$

2.2 DPCM

The total error is again identical to the quantization error but its variance is now proportional to that of the difference signal:

$$\sigma_t^2 = \sigma_q^2 = \epsilon_q^2 \cdot \sigma_d^2. \tag{12}$$

Remark. Some caution is needed if two coding schemes A and B, e.g., PCM and DPCM, are being compared. We have to take into account dif-

ferences in the quantizer performance factors ϵ_{qA}^2 and ϵ_{qB}^2 of the two schemes, because the probability density functions of the signals at the corresponding quantizer inputs may differ. For example, let A and B denote PCM and DPCM, respectively. The gain of DPCM over PCM in signal-to-noise ratio is then given as

$$\frac{\sigma_{tA}^2}{\sigma_{tB}^2} = \frac{\epsilon_{qA}^2}{\epsilon_{qB}^2} \cdot \frac{\sigma_x^2}{\sigma_d^2}. \tag{13}$$

The ratio $\epsilon_{qA}^2/\epsilon_{qB}^2$ is very close to unity if the quantizers have a logarithmic characteristic because their performance is relatively independent of signal statistics. The ratio is also close to unity if the samples of the coder input sequences are Gaussian distributed since all coders discussed in this paper employ linear networks which do not affect the Gaussian distribution.

In DPCM systems the prediction is affected by the quantization due to the error feedback, i.e., the predictor uses previous reconstruction values $y_j = x_j - t_j = x_j - q_j; j = n-1, n-2, \ldots$ instead of the corresponding input samples x_j. The influence of this feedback on the prediction error variance has been analyzed elsewhere[16–18]; it will be briefly discussed in the following example but will then be neglected in the further analyses. It is known that this simplification is valid if quantizers with at least eight levels are employed. If necessary, we shall use the terms "real" DPCM and "ideal" DPCM for analyses based on prediction with and without error feedback, respectively.

Example 1: We show the influence of the feedback on the prediction error in a DPCM scheme with a first-order predictor of value a. Let $\rho = R_x(1)/\sigma_x^2$ be the normalized mutual correlation between adjacent samples. Without feedback the prediction error variance is

$$\sigma_d^2 = (1 + a^2 - 2a\rho) \cdot \sigma_x^2 \tag{14}$$

with a minimum

$$\min\{\sigma_d^2\} = (1 - \rho^2) \cdot \sigma_x^2 \tag{15}$$

for $a = a_{opt} = \rho$. The DPCM difference signal is $d_n = x_n - ax_{n-1} + aq_{n-1}$. Its variance is given as

$$\sigma_d^2 = \frac{1 - 2a\rho + a^2}{1 - \epsilon_q^2 a^2} \cdot \sigma_x^2 \tag{16}$$

on the assumption of a vanishing correlation between input and quantization error,[16] and its minimum variance is obtained with[19]:

$$a_{opt} = \frac{1 + \epsilon_q^{-2}}{2\rho}\left[1 - \sqrt{1 - \epsilon_q^{-2}\left(\frac{2\rho}{1 + \epsilon_q^{-2}}\right)^2}\right]. \tag{17}$$

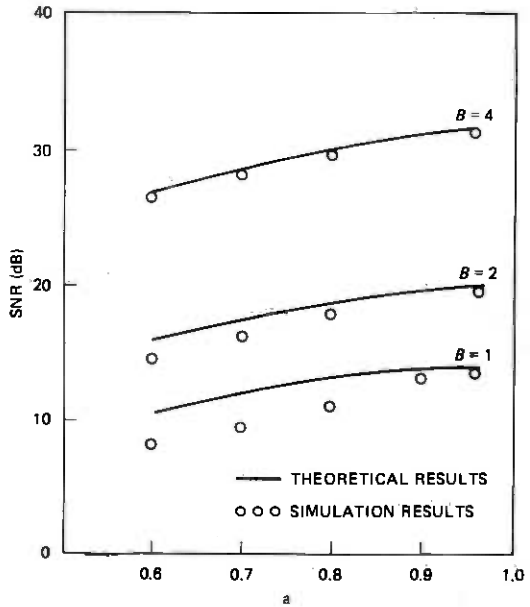


Fig. 2—DPCM performance: SNR vs. value a of the predictor coefficient. Source: Gaussian first-order Markov source with correlation $\rho = 0.9625$. Quantizer: B-bit quantizer optimized for Gaussian signals. Predictor: Previous sample prediction with coefficient a.

The value a_{opt} is not critical, however. We use $a_{opt} \approx \rho$ to determine the total MMSE:

$$\min\{\sigma_t^2\} = \epsilon_q^2 \frac{1-\rho^2}{1-\epsilon_q^2\rho^2} \cdot \sigma_x^2. \tag{18}$$

Due to feedback there is an increase in variance by a factor $(1-\epsilon_q^2\rho^2)^{-1}$. Figure 2 shows the dependence of the signal-to-noise ratio (SNR) on the value of the predictor coefficient for various B-bit quantizers and compares theoretical results obtained from eq. (18) with simulation results. It is seen that these are useful approximations in the vicinity of the optimum setting of the predictor coefficient (this is the region where the difference signal is almost white noise, and the simulations reveal that the assumption of a vanishing correlation between quantization error and input signal holds in this case).

In an "ideal" Nth order DPCM system the prediction of an input sample x_n is based on previous input samples x_{n-j}; $j = 1,2 \ldots ,N$. Hence the prediction scheme is essentially that of D*PCM (see Fig. 1c), but it is important to realize that these schemes are optimal for different frequency responses $P(\omega)$ of the predictor as will be seen later in this paper. Equation (12) shows that the optimum predictor in DPCM is the MMSE

predictor. The prediction error variance is given as

$$\sigma_d^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) |1 - P(\omega)|^2 d\omega \tag{19}$$

and its minimum is reached if the random sequence $\{d_n\}$ is a white noise sequence of variance η_x^2 (*full-whitening*). Thus we have

$$\min\{\sigma_d^2\} = \eta_x^2, \tag{20}$$

where η_x^2 is the minimum prediction error variance to be obtained from the random sequence $\{x_n\}$ by passing its samples through a prediction error filter with frequency response $1 - P_{opt}(\omega)$ such that

$$S_x(\omega) |1 - P_{opt}(\omega)|^2 = \eta_x^2. \tag{21}$$

We note that this minimum can be obtained for any stationary random sequence if the predictor impulse response is of semi-infinite length, and an Nth order predictor can be employed if the random sequence is Markovian of order N. For a given pds $S_x(\omega)$ Kolmogoroff[20] has given the minimum of the prediction error variance as

$$\eta_x^2 = \min\{\sigma_d^2\} = \exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log_e S_x(\omega) d\omega\right]. \tag{22}$$

This variance η_x^2 is a positive quantity if the process is undetermined, i.e., if $S_x(\omega)$ is zero at most at a countable set of frequencies. Otherwise the signal is perfectly predictable and the prediction error variance is zero then. The normalized prediction error variance

$$\gamma_x^2 = \frac{\eta_x^2}{\sigma_x^2} = \frac{\exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log_e S_x(\omega) d\omega\right]}{\frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) d\omega} \tag{23}$$

is called spectral flatness measure[21] and its inverse is the optimally obtainable prediction gain. The spectral flatness measure can be interpreted as the ratio of the geometric mean of the pds $S_x(\omega)$ to its arithmetic mean. It is easy to show[21] that

$$0 \leqslant \gamma_x^2 \leqslant 1. \tag{24}$$

From eq. (12) we finally find that

$$\min_{\text{DPCM}} \{\sigma_t^2\} = \epsilon_q^2 \cdot \eta_x^2, \tag{25}$$

*2.3 D*PCM*

The quantization error with pds $S_q(\omega) = \sigma_q^2$ is filtered by the linear decoder network with frequency response $H(\omega) = G^{-1}(\omega)$ (see Fig. 1c).

Thus the total error variance is

$$\sigma_t^2 = \sigma_q^2 \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(\omega)|^2 d\omega, \tag{26}$$

i.e., the quantization error variance is increased by the power transfer factor

$$\alpha = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(\omega)|^2 d\omega. \tag{27}$$

The variance of the quantization error depends on the variance of the difference signal d^*_n:

$$\sigma_q^2 = \epsilon_q^2 \cdot \sigma_{d^*}^2 = \epsilon_q^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) |G(\omega)|^2 d\omega. \tag{28}$$

Thus the total error variance is

$$\sigma_t^2 = \epsilon_q^2 \cdot \alpha \cdot \sigma_{d^*}^2, \tag{29}$$

i.e., the total error variance of D*PCM is α times that of an equivalent DPCM coding scheme (with $\sigma_d^2 = \sigma_{d^*}^2$ in the case of fine quantizing and identical prediction filters). The impulse response of the postfilter is a sequence $\{1, h_1, h_2, \cdots\}$; thus we have

$$\alpha = 1 + \sum_{k=1}^{\infty} h_k^2 \geqslant 1 \tag{30}$$

and we find that DPCM outperforms D*PCM for *any* choice of the predictor network.[22] This result also implies that the optimum performance of DPCM is better than the optimum performance of D*PCM. This fact will be shown very shortly in a slightly broader context.

Example 2: We calculate now the total MSE of a simple D*PCM scheme which employs a first-order predictor of value a. The prediction error variance has been given in eq. (14). The power transfer factor of the decoder network is

$$\alpha = 1 + a^2 + a^4 + \cdots = (1 - a^2)^{-1}. \tag{31}$$

Thus the total MSE is

$$\sigma_t^2 = \epsilon_q^2 \cdot \alpha \cdot \sigma_{d^*}^2 = \epsilon_q^2 \frac{1 - 2a\rho + a^2}{1 - a^2} \cdot \sigma_x^2. \tag{32}$$

Note that MMSE prediction ($a = \rho$) results in the same MSE as that of standard PCM,[3] and that $\sigma_t^2 \to \infty$ for any ρ if $a \to 1$. The minimum total MSE is obtained for

$$a_{opt} = \frac{1}{\rho}(1 - \sqrt{1 - \rho^2}), \tag{33}$$

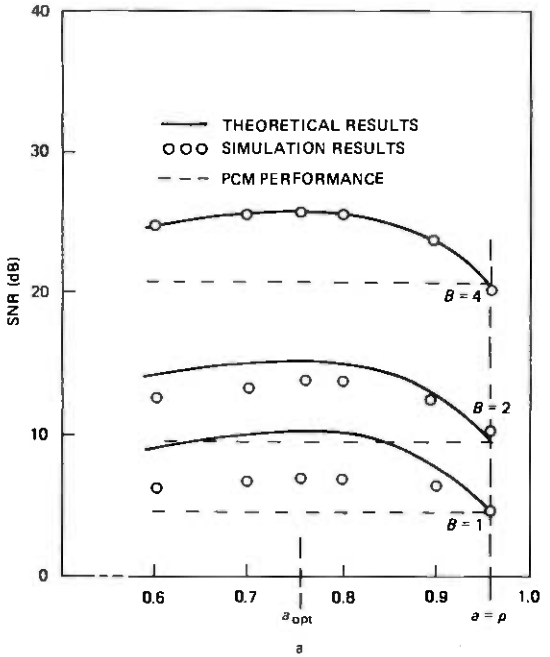


Fig. 3—D*PCM performance: SNR vs. value a of the predictor coefficient. Source and coder same as in Fig. 2.

and its value is

$$\min\{\sigma_t^2\} = \epsilon_q^2 \sqrt{1-\rho^2} \cdot \sigma_x^2. \tag{34}$$

Thus it is possible to reduce the total noise variance by a factor $\sqrt{1-\rho^2}$ instead of $(1-\rho^2)$ as in ideal DPCM (with negligible feedback). It is interesting to note that this is the same reduction that we can obtain with a block quantization scheme of blocklength 2 (this is a scheme where the sum and the difference of adjacent samples are quantized independently).[23] Figure 3 demonstrates this fact that the SNR is lower than that obtainable with DPCM (Fig. 2). In the case of quantizing with one and two bits/sample there is, as expected, a significant difference between theory and measurements especially for a-values which are not close to the correlation coefficient ρ. This difference is now explained by the fact that, in the case of coarse quantization, the white noise assumption [eq. (8)] does not hold for correlated quantizer input samples. Simulations have revealed that higher signal-to-noise ratios can be obtained by choosing a decoder coefficient that is higher than that of the coder. Note also, from Fig. 3, that no gain over PCM is obtainable for the specific case $a = \rho$.

We come back now to the total MSE of D*PCM. Equation (29) is given

explicitly as

$$\sigma_t^2 = \epsilon_q^2 \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) |1 - P(\omega)|^2 d\omega \right] \times \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} |1 - P(\omega)|^{-2} d\omega \right] \quad (35)$$

where we have used our assumptions $G(\omega) = 1 - P(\omega)$ and $H(\omega) = G^{-1}(\omega)$ (both assumptions will be reviewed in the next section). Note that the term in the first brackets reduces to η_x^2 if $P(\omega)$ is the MMSE predictor [full-whitening, see eq. (21)]. The difference signal d^*_n is white noise then, and, since the last term in eq. (35) is the power transfer factor for white noise inputs, it equals $\sigma_x^2/\sigma_{d^*}^2 = \sigma_x^2/\eta_x^2$. Therefore the total MSE is $\sigma_t^2 = \epsilon_q^2 \cdot \sigma_x^2$, and thus equals that of PCM.[3] Note that this statement is based on assumptions given in the remark in Section 2.2. We relate now the minimum total MSE obtainable with D*PCM to that of DPCM. By applying Schwarz's inequality† to eq. (35) we obtain

$$\min_{\text{D*PCM}} \{\sigma_t^2\} \geq \epsilon_q^2 \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \sqrt{S_x(\omega)}\, d\omega \right]^2 = \epsilon_q^2 \sigma_{\sqrt{x}}^4 \quad (36)$$

where $\sigma_{\sqrt{x}}^2$ defines the variance of a process with pds $\sqrt{S_x(\omega)}$.

Equality is obtained if

$$|1 - P(\omega)|^2 \sqrt{S_x(\omega)} = \text{const.} \quad (37)$$

We conclude that the squared-magnitude frequency response of the D*PCM encoder has to be inversely proportional to the square root of the pds of the input (*half-whitening*). Therefore the optimum frequency response $1 - P(\omega)$ is a MMSE prediction error filter for a pds $\sqrt{S_x(\omega)}$. The corresponding minimum $\eta_{\sqrt{x}}^2$ of the prediction error variance would be just the square root of that to be obtained from an optimum prediction of a process with pds $S_x(\omega)$ [this can be verified from eq. (22)]:

$$|1 - P(\omega)|^2 \sqrt{S_x(\omega)} = \eta_{\sqrt{x}}^2 = \eta_x \quad (38)$$

However, it is important to realize at this point that it is not certain that the equality in eq. (37) can be obtained at all in practice. Indeed, equality can only be expected if $\sqrt{S_x(\omega)}$ happens to be a rational spectrum, since such a spectrum can always be modeled by white noise passed through a purely recursive linear filter with poles inside the unit circle. It is clear that such a process can be whitened by a one-step ahead prediction error filter $1 - P(\omega)$. In general, however, the pds $\sqrt{S_x(\omega)}$ is not rational, and the minimum total error variance to be obtained by D*PCM coding is

† $\int_I f_1^2(x)dx \cdot \int_I f_2^2(x)dx \geq \left[\int_I f_1(x) f_2(x)dx\right]^2$ with equality if $f_1^2(x)/f_2^2(x) = \text{const.}$ for any square-integrable functions $f_1(\cdot)$ and $f_2(\cdot)$.

greater than the right-hand term of eq. (36). Example 2 has shown that an improvement over PCM can be obtained with D*PCM. By applying Schwarz's inequality to the right-hand term of eq. (36) we have

$$\epsilon_q^2 \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \sqrt{S_x(\omega)}\, d\omega \right]^2 \leq \epsilon_q^2 \sigma_x^2 = \min_{\text{PCM}} \{\sigma_t^2\} \tag{39}$$

with equality only if $\{x_n\}$ is a white noise sequence, i.e., an improvement over PCM can be obtained for all nonwhite processes. Equation (30) has already indicated that DPCM outperforms D*PCM for any nonwhite pds $S_x(\omega)$. We can make this statement more quantitative by comparing the total error variances of D*PCM (in the most favorable case, given if the pds $\sqrt{S_x(\omega)}$ is rational) and that of DPCM. To be more specific, we use eqs. (22)–(25) and (36), and find

$$\frac{\min\limits_{\text{DPCM}} \{\sigma_t^2\}}{\min\limits_{\text{D*PCM}} \{\sigma_t^2\}} \leq \frac{\exp\left[\dfrac{1}{2\pi} \displaystyle\int_{-\pi}^{\pi} \log_e S_x(\omega) d\omega\right]}{\left[\dfrac{1}{2\pi} \displaystyle\int_{-\pi}^{\pi} \sqrt{S_x(\omega)}\, d\omega\right]^2}$$

$$= \frac{\eta_x^2}{\sigma_{\sqrt{x}}^4} = \left[\frac{\eta_{\sqrt{x}}^2}{\sigma_{\sqrt{x}}^2}\right]^2 = \gamma_{\sqrt{x}}^4 \leq 1. \tag{40}$$

We conclude that the ratio of the total error variances of DPCM and D*PCM is upperbounded by the squared spectral flatness measure $\gamma_{\sqrt{x}}^2$ of the pds $\sqrt{S_x(\omega)}$. The validity of the right-hand inequality in eq. (40) has already been stated in eq. (24). Equality is obtained if and only if $\{x_n\}$ is a white noise sequence. In other words, DPCM outperforms D*PCM for all nonwhite spectra.

Example 3: Example 1 has shown that the total MSE of a DPCM scheme with a first-order predictor is smaller by a factor $\beta^2 = 1 - \rho^2$ than that of PCM (provided that the quantizer performance factors are identical in both cases). In D*PCM, the corresponding factor is $\beta^2 = \sqrt{1 - \rho^2}$ (see example 2). The value ρ is in both examples the one-lag normalized autocorrelation coefficient of an otherwise arbitrary (though stationary) random input sequence. Let us assume now that this sequence is a first-order Markovian sequence with autocorrelation

$$R_x(k) = \rho^{|k|} \cdot \sigma_x^2 \tag{41}$$

and pds

$$S_x(\omega) = \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos \omega}\, \sigma_x^2. \tag{42}$$

The best DPCM performance is still obtained with a first-order predictor since $\{x_n\}$ is Markovian. The D*PCM reduction factor $\beta^2 = \sqrt{1-\rho^2}$ is not optimal, however. It is interesting to compare it with the upper bound to be obtained without the constraint of the realizability of the prefilter.

Using eqs. (36) and (42) we find

$$\min_{\text{D*PCM}} \{\sigma_t^2\} \geqslant \epsilon_q^2 \cdot \beta^2 \cdot \sigma_x^2 \tag{43}$$

where

$$\beta^2 = (1-\rho^2)\left(\frac{2}{\pi}\right)^2 F^2\left(\frac{\pi}{2}, \rho\right). \tag{44}$$

$F(\pi/2, \rho)$ is the complete elliptical integral of the first kind:

$$\begin{aligned} F\left(\frac{\pi}{2}, \rho\right) &= \int_0^{\pi/2} \frac{d\phi}{\sqrt{1-\rho^2 \sin^2\phi}} \\ &= \frac{\pi}{2}\left[1 + \left(\frac{1}{2}\right)^2 \rho^2 + \left(\frac{1\cdot 3}{2\cdot 4}\right)^2 \rho^4 + \cdots\right]. \end{aligned} \tag{45}$$

We conclude that a D*PCM coding of a first-order Markov source is upperbounded by a reduction factor

$$\beta^2 = (1-\rho^2)\left[1 + \frac{1}{4}\rho^2 + \frac{9}{64}\rho^4 + \cdots\right]^2 \tag{46}$$

instead of $(1-\rho^2)$ for DPCM and $\sqrt{1-\rho^2}$ for the one-tap D*PCM. For $\rho = 0.9625$ the one-tap D*PCM coding results in a total MSE that is by a factor 3.7 greater than that of (ideal) DPCM, and it is lower-bounded by a factor 3.0 (for $\rho = 0.85$ the corresponding values are 1.9 and 1.8, respectively).

The foregoing discussion has shown that D*PCM is a suboptimal coding scheme if the performance criterion is the unweighted total error variance. The next section will demonstrate that D*PCM is an optimum coding scheme for a specific frequency-weighted error criterion, and Section IV will show that its performance bounds the overall performance of DPCM coders in the presence of channel errors. As a final remark we mention that D*PCM based on adaptive prediction with a separate transmission of the predictor coefficients has been used recently for speech coding purposes[24] to improve the subjective performance.

III. NOISE-FEEDBACK CODING

In the preceding chapter we have optimized pre- and postfilters for a quantizing scheme (D*PCM) and have shown that DPCM has a superior performance for all nonwhite processes. At a first glance this seems to

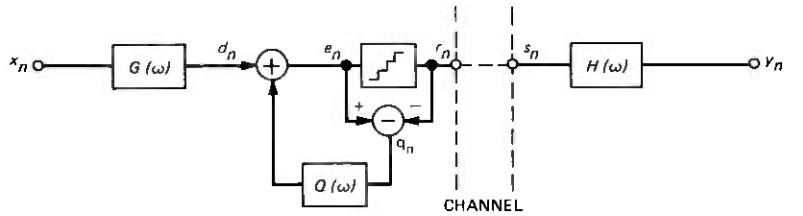


Fig. 4—Noise-feedback coding scheme.

be surprising since D*PCM coding is also based on linear filtering. In the D*PCM analysis, however, we did not take into account that not only the pre- and postfilters are under our control, but also the quantizer. To be more specific, we may additionally filter the quantization noise whereas in typical pre- and postfiltering applications it is the channel that is perturbed by noise; that noise is not under control of the designer. Figure 4 demonstrates how the filtering of the quantization noise is obtained. The quantization noise, i.e., the difference between input and output of the quantizer, is fed back through a linear filter with frequency response $Q(\omega)$ and is added to the input. $Q(\omega)$ is required to have a minimum delay of one sampling time for stability reasons. The purpose of the feedback scheme is a reshaping of the spectrum of the quantization noise such that the total error variance is minimum. It should also be mentioned at this point that there exists also another linearly equivalent scheme, the direct-feedback coder, which has been studied by Brainard and Candy.[15] It uses a prefilter of frequency response $A(\omega)$, and the output of the quantizer (not the quantization noise!) is first passed through a feedback filter of frequency response $B(\omega)$ and then added to the prefiltered signal to form the quantizer input. The equivalence to the noise-feedback coder is given by $G(\omega) = A(\omega)/(1 - B(\omega))$ and $1 - Q(\omega) = 1/(1 - B(\omega))$.

3.1 Derivation of the basic formula

Let us, as before, represent the quantizer as a device that adds signal-independent white noise of pds $S_q(\omega) = \sigma_q^2$ to the signal. Due to this assumption we can also replace the feedback-quantizer with a nonwhite noise source of pds

$$\begin{aligned} S_n(\omega) &= S_q(\omega)|1 - Q(\omega)|^2 \\ &= \sigma_q^2|1 - Q(\omega)|^2. \end{aligned} \tag{47}$$

The feedback acts as linear filter on the open-loop quantizing noise, and the effective quantization noise is colored noise then. Additionally we may introduce a subjective noise-weighting function $S_w(\omega)$ whose inverse describes the sensitivity of the sink to uncorrelated noise. A small value

of $S_w(\omega)$ indicates that a high error variance is acceptable in that specific frequency region and vice versa. It is well known that such a frequency-weighting can only serve as a first approximation of noise sensitivity since it does not take into account nonlinear effects and dependencies on local properties of the signals. Nevertheless, the frequency-weighted noise power is a quite popular design criterion partly due to the lack of better distortion measures. What is worth emphasizing is that $S_w(\omega)$ can be interpreted as the squared-magnitude frequency response of a noise-weighting filter. Therefore it is reasonable to assume that $S_w(\omega)$ is a rational function of ω. $S_w(\omega)$ can also be explained as the output pds of such a weighting filter whose input is a white noise process. This interpretation suggests to define a variance

$$\sigma_w^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_w(\omega) d\omega, \tag{48}$$

a minimum prediction error variance

$$\eta_w^2 = \exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log_e S_w(\omega) d\omega\right], \tag{49}$$

and a spectral flatness measure

$$\gamma_w^2 = \eta_w^2 / \sigma_w^2 \tag{50}$$

in accordance with eqs. (7), (22), and (23).

The MSE of a NFC coder is given by

$$\sigma_t^2 = \epsilon_q^2 \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) |G(\omega)|^2 d\omega\right] \cdot \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} S_w(\omega) |H(\omega)|^2 |1 - Q(\omega)|^2 d\omega\right] \tag{51}$$

where we have used eqs. (26), (28), and (47). Note that the MSE does not include a linear distortion term resulting from a possible mismatch between prefilter and postfilter. We are now free to choose linear filters $G(\omega)$, $H(\omega)$, and $Q(\omega)$ such that the frequency-weighted error variance is minimized. This is obtained by a proper preshaping of the signal spectrum *and* the quantization noise prior to transmission. The general design of this NFC scheme has been studied by Kimme and Kuo in the context of picture coding.[14] Our objective is to show the connection of this noise-feedback scheme with those discussed so far and to show how subjective noise-weighting functions influence the design. Our approach will also be different from that in Ref. 14 since a constraint is not needed in the optimization procedure. The MMSE design problem is to find the optimum combination of prefilter $G(\omega)$, postfilter $H(\omega)$, and feedback filter $Q(\omega)$ for a given pds $S_x(\omega)$ of the input signal, and a weighting

function $S_w(\omega)$. Let us first assume that $G(\omega)$ and $Q(\omega)$ are given and let us determine $H(\omega)$. In an optimized system the total error $x_n - y_n$ must be orthogonal to the output of the postfilter and thus orthogonal to the data r_n used in that filter (the sequence $\{r_n\}$ is the decoder input sequence). It can be shown that this condition also holds in the case of weighted total errors. If the postfilter $H(\omega)$ is not constrained to be physically realizable (its characteristics can always be approximated arbitrarily closely by allowing for a sufficient time delay), the optimum filter is given by the Wiener-Hopf condition

$$H_{opt}(\omega) = \frac{S_{rx}(\omega)}{S_r(\omega)} \tag{52}$$

where $S_{rx}(\omega) = G^*(\omega) \cdot S_x(\omega)$ is the cross-spectrum between the sequences $\{r_n\}$ and $\{x_n\}$. Thus we see that

$$H_{opt}(\omega) = \frac{G^*(\omega) \cdot S_x(\omega)}{|G(\omega)|^2 S_x(\omega) + S_q(\omega)|1 - Q(\omega)|^2}. \tag{53}$$

We shall restrict our attention to the case of a small quantization noise variance:

$$S_q(\omega) \ll \frac{|G(\omega)|^2 S_x(\omega)}{|1 - Q(\omega)|^2} \quad \text{for } \textit{all } \omega. \tag{54}$$

Equation (53) becomes

$$H_{opt}(\omega) = \frac{1}{G(\omega)}, \tag{55}$$

so that $G(\omega)$ and $H(\omega)$ are reciprocal filters for *any* given $G(\omega)$ which does not violate the assumption of eq. (54). It is seen that choosing reciprocal coding and decoding filters is not just a convenience but a requirement by the MSE criterion. By applying Schwarz's inequality to eq. (51), we find

$$\min\{\sigma_t^2\} = \epsilon_q^2 \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \sqrt{S_x(\omega) S_w(\omega) |1 - Q(\omega)|^2}\, d\omega \right]^2 \tag{56}$$

for any given $S_w(\omega)$ and $Q(\omega)$. This minimum is reached if

$$|G_{opt}(\omega)|^2 = C^2 \sqrt{\frac{S_w(\omega)|1 - Q(\omega)|^2}{S_x(\omega)}} \tag{57}$$

where C is a constant. Equations (51) and (56) can be used to calculate the error variances of various coding schemes.

3.2 Optimization of the noise-feedback coder

We shall now derive two conditions which have to be met by $Q(\omega)$ and $G_{opt}(\omega)$, respectively. We apply again Schwarz's inequality to eq. (56)

and thus have

$$\min\{\sigma_t^2\} \leqslant \epsilon_q^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) S_w(\omega) |1 - Q(\omega)|^2 d\omega \tag{58}$$

with equality if the integrand is a constant. Note that the term $1 - Q(\omega)$ describes a prediction error structure. Therefore equality is obtained by choosing $Q(\omega)$ to be the MMSE predictor of a random sequence of pds $S_x(\omega) \cdot S_w(\omega)$ and we have [in the notation of eq. (21)]:

$$S_x(\omega) S_w(\omega) |1 - Q_{opt}(\omega)|^2 = \eta_{xw}^2 = \eta_x^2 \cdot \eta_w^2 \tag{59}$$

where $Q_{opt}(\omega)$ is the MMSE predictor that whitens a random sequence of pds $S_x(\omega) \cdot S_w(\omega)$, and where $\eta_x^2 \eta_w^2$ is its MMSE. The right-hand equality in eq. (59) can be obtained from Kolmogoroff's result [eq. (22)] by substituting $S_x(\omega)$ with its frequency-weighted version $S_x(\omega) S_w(\omega)$. When comparing eqs. (56) and (59) we find

$$\min_{\text{NFC}} \{\sigma_t^2\} = \epsilon_q^2 \cdot \eta_x^2 \cdot \eta_w^2. \tag{60}$$

We conclude that the frequency-weighted total error variance of an NFC scheme with given quantizer is determined by the product of the prediction error variances of the spectra $S_x(\omega)$ and $S_w(\omega)$, and that $Q_{opt}(\omega)$ is the optimal predictor of a pds $S_x(\omega) \cdot S_w(\omega)$. We also have, from eqs. (57) and (59), that $S_x(\omega) |G_{opt}(\omega)|^2 = \text{const}$. This implies, however, that $|G_{opt}(\omega)|^2$ is a filter which whitens $S_x(\omega)$ which has been assumed to be rational. Thus a sufficient condition for an optimum NFC scheme is to choose as a prefilter an MMSE optimized prediction error filter:

$$G_{opt}(\omega) = 1 - P_{opt}(\omega) \tag{61}$$

where $P_{opt}(\omega)$ is defined by

$$S_x(\omega) |1 - P_{opt}(\omega)|^2 = \eta_x^2. \tag{62}$$

Note the important fact that the overall performance is optimized if the prefilter is an MMSE prediction-error filter that whitens the input process of pds $S_x(\omega)$ and that this result holds for any choice of the weighting function. It turns out that we have to modify the quantization noise feedback loop but not the prefilter if a weighting of the noise has to be taken into account. We finally find from eqs. (20), (47), (59), and (62) that the weighted error spectrum $S_t(\omega) = \epsilon_q^2 \sigma_d^2 S_w(\omega) |1 - Q_{opt}(\omega)|^2 \cdot |1 - P_{opt}(\omega)|^{-2}$ is constant and thus equals the total error variance [see eq. (60)]. We finally note that these optimization results can also be extracted from the Kimme and Kuo paper.[14] Musmann has recently derived equivalent results based on a more information-theoretical analysis.[25]

We shall now briefly discuss two special cases.

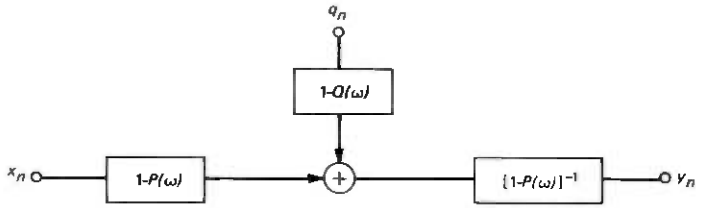


Fig. 5—Model of NFC coder with prediction error filter as input.

Special case: $S_w(\omega) = 1$ for all ω (DPCM). A comparison of eqs. (59) and (62) reveals that $Q_{opt}(\omega) = P_{opt}(\omega)$, i.e., the predictor prefilter and the feedback filter have to have identical frequency responses. It is easy to see[26] that these filters can then be combined to yield the DPCM structure of Fig. 1b. A DPCM structure is indeed defined by having $Q(\omega) = P(\omega)$ and Fig. 5, which is a model of a noise-feedback coder, reveals that no frequency-weighting of the quantization noise is possible in this case for *any* choice of the predictor (provided that pre- and postfilters are reciprocals), since the quantization noise passes both the feedback filter with frequency response $1 - P(\omega)$ and its inverse, the postfilter.

Special case: $S_w(\omega) \propto S_x^{-1}(\omega)$ for all ω (D*PCM). A weighting function that is in some sense inverse to the pds of the signal is of importance for quantizing acoustic signals since it may avoid a masking of weak signal energies in specific frequency ranges by the quantization noise.

For $S_w(\omega) \propto S_x^{-1}(\omega)$ we find from eq. (59) that $|1 - Q_{opt}(\omega)|^2$ has to be a constant. This clearly means that $Q_{opt}(\omega) = 0$ for all ω; i.e., the best coding scheme is now D*PCM. The prefilter is, of course, as for all optimal NFC coders, a whitening filter. The D*PCM scheme is suboptimum if it is used in connection with other weighting functions. The special case of $S_w(\omega) = 1$ has been discussed in detail in Section II; we have seen there that for $S_w(\omega) = 1$ and a MMSE predictor $P(\omega)$ no reduction in total error variance over PCM is obtainable. The above discussion reveals that the same coder is, however, optimal for the specific noise-weighting function $S_w(\omega) \propto S_x^{-1}(\omega)$. Indeed, subjective gains of about 6–10 dB have been reported for speech signals, for this choice of the prefilter.[26,27]

Table I lists various NFC configurations. The performance of some suboptimal coders will be compared with the optimal NFC scheme in the next part of this section. An elementary example will suggest the manner in which the NFC design influences the total error variance.

Example 4: Assume a noise-feedback coder with just one tap, i.e., $Q(\omega) = q \cdot \exp(-j\omega)$, and a prefilter with the structure of a one-tap prediction-error filter: $G(\omega) = 1 - a \cdot \exp(-j\omega)$ (see Fig. 6). Further assume a sequence with an adjacent-sample correlation ρ. We have $H(\omega) = G^{-1}(\omega)$

Table I — Performance comparison of NFC coder configurations

Noise weighting	NFC	DPCM $Q(\omega) = P(\omega)$	D*PCM $Q(\omega) = 0$
$S_w(\omega) = 1$ (no weighting)	$\hat{=}$ DPCM	Optimal coder: $\lvert 1 - P\rvert^2 \propto S_x^{-1}$	Suboptimal coder (upper bound half-whitening): $\lvert 1 - P\rvert^2 \propto S_x^{-1/2}$
$S_w(\omega) \propto S_x^{-1}(\omega)$	$\hat{=}$ D*PCM	Suboptimal coder	Optimal coder (full-whitening): $\lvert G\rvert^2 = \lvert 1 - P\rvert^2 \propto S_x^{-1}$
All other cases:	Optimal coder	Suboptimal coder	Suboptimal coder

and the unweighted total MSE [$S_w(\omega) = 1$] can be derived from eq. (51); we shall omit the intermediate steps. The final result is

$$\sigma_t^2 = \epsilon_q^2 \frac{1 + a^2 - 2a\rho}{1 - a^2}(1 + q^2 - 2aq) \cdot \sigma_x^2$$

$$a < 1;\ q \text{ arbitrary.} \tag{63}$$

For $a = 0$ and $q = 0$ we have the PCM result of eq. (10). For $a = 0$ and finite q we obtain $\sigma_t^2 = \epsilon_q^2(1 + q^2)\cdot\sigma_x^2$, i.e., noise feedback without prefiltering increases the MSE by a factor $1 + q^2$ over that of PCM. For $q = 0$ and finite $a < 1$ we have the case of D*PCM, i.e., prefiltering followed by quantization [eq. (32)]. The best choice of q is $q = a$ if a is given; the scheme reduces then to DPCM [eq. (1)] and reaches its MMSE for $a = q = \rho$. The different cases have been tabulated in Table II.

3.3 Suboptimal coding schemes

The last section has shown that noise-feedback coding is a scheme that allows for the optimal shaping of the spectra of the input signal *and* the quantization noise such that the noise-weighted overall error variance is minimized. The prefilter is in all cases a MMSE prediction error filter $1 - P_{opt}(\omega)$ which performs a decorrelation of the input signal. The optimal scheme reduces to DPCM ($Q_{opt}(\omega) = P_{opt}(\omega)$) in the case of unweighted noise ($S_w(\omega) = 1$), and to D*PCM ($Q_{opt}(\omega) = 0$) in the case of $S_w(\omega) \propto S_x^{-1}(\omega)$. In all other cases, only the general NFC scheme (with $Q(\omega) \neq P(\omega)$) is optimal.

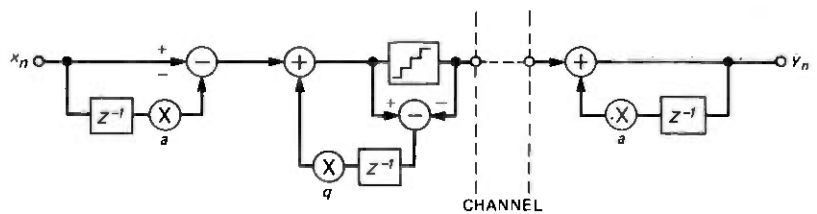


Fig. 6—One-tap NFC coder of Example 4.

Table II—Error variances of coding schemes with one predictor coefficient a_{opt} and one feedback coefficient q_{opt}

	$\min\{\sigma_t^2\}$	q_{opt}	a_{opt}
PCM	$\epsilon_q^2 \cdot \sigma_x^2$	0	0
Noise-feedback PCM	$\epsilon_q^2(1+q^2)\sigma_x^2$	$\neq 0$	0
D*PCM	$\epsilon_q^2 \sqrt{1-\rho^2}\,\sigma_x^2$	0	$\frac{1}{\rho}(1-\sqrt{1-\rho^2})$
DPCM (ideal)	$\epsilon_q^2 (1-\rho^2)\sigma_x^2$	ρ	ρ
DPCM (real, i.e., with coarse quantization)	$\epsilon_q^2 \frac{1-\rho^2}{1-\epsilon_q^2\rho^2}\sigma_x^2$	a_{opt}	$<\rho$ [see eq. (17)]

This section compares the performances of various suboptimal coding schemes which belong to the class of NFC schemes and which are derived by choosing suboptimum prefilters and feedback filters. The corresponding error variances can be derived directly from eq. (51). We shall omit the intermediate steps in the calculations and optimizations of σ_t^2 and shall present the final results only.

PCM. PCM results if $G(\omega) = H(\omega) = 1$ and $Q(\omega) = 0$. We then have

$$\min_{\text{PCM}} \{\sigma_t^2\} = \epsilon_q^2 \sigma_x^2 \sigma_w^2. \tag{64}$$

The quantity σ_w^2 as defined in eq. (48) represents the subjective gain if white quantization noise is weighted.

Noise-Feedback PCM. We have $G(\omega) = H(\omega) = 1$, and $Q_{opt}(\omega)$ is the MMSE predictor of the weighting function $S_w(\omega)$. Hence we have

$$\min_{\substack{\text{noise-feedback} \\ \text{PCM}}} \{\sigma_t^2\} = \epsilon_q^2 \sigma_x^2 \eta_w^2. \tag{65}$$

This result shows that η_w^2 as defined in eq. (49) represents the error reduction obtainable if the quantization noise is optimally shaped. Note that optimal noise-shaping reduces the weighted error variance in relation to PCM by a factor $\gamma_w^2 = \eta_w^2/\sigma_w^2$ which is the spectral flatness measure of the weighting function $S_w(\omega)$.

Remark: We have to mention at this point that eq. (65) only holds if γ_w^2 is not close to zero. This implies that the predictibility of $S_w(\omega)$ (which is bounded by the dynamic range of the spectrum[28]) should not be too high. Otherwise one could obviously achieve large reductions in error (including unbounded ones for weighting functions which are zero over a finite segment of the frequency axis; see eq. (22) and the following discussion thereof). The noise reductions are obtained by shifting the noise in frequency to a range where it is less heavily weighted by $S_w(\omega)$. However, the noise-shaping increases the total variance of the quantizer input signal since the correlated quantization noise is added to the input

signal. Therefore the quantizer step sizes have to be readjusted accordingly. As a consequence we obtain an increase in the quantizer performance factor ϵ_q^2 and subsequently an increase in total error variance. Spang and Schultheiss[29] have discussed in detail the stability problems involved. They have been interested in noise reductions in oversampled systems where the noise can be shifted into the high-frequency range and subsequently eliminated by lowpass filtering. In that application we have a weighting function that is zero over a finite segment of the frequency axis. As a means for reducing the stability problems Spang and Schultheiss propose to keep the number of feedback elements R finite. In this case the total error variance is given as

$$\sigma_l^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_w(\omega) \left| 1 - \sum_{k=1}^{R} q_k e^{-jk\omega} \right|^2 d\omega. \tag{66}$$

The feedback network can be viewed as an Rth order predictor. Its optimal coefficients q_k; $k = 1,2,\ldots,R$ can be derived from the set of R normal equations whose coefficients are just the Fourier coefficients of $S_w(\omega)$.[30]

D*PCM. We have $G(\omega) = 1 - P(\omega)$ and $Q(\omega) = 0$.

(*i*) *Lower bound (half-whitening).* The lower bound is reached if $|1 - P_{opt}(\omega)|^2 = \sqrt{S_w(\omega)/S_x(\omega)}$. We obtain

$$\min_{\text{D*PCM}} \{\sigma_l^2\} = \epsilon_q^2 \sigma^4_{\sqrt{xw}} \tag{67}$$

where

$$\sigma^2_{\sqrt{xw}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sqrt{S_x(\omega) \cdot S_w(\omega)}\, d\omega. \tag{68}$$

(*ii*) *Full-whitening.* Let $P(\omega)$ be the MMSE predictor for $S_x(\omega)$. It follows that

$$\sigma_l^2 = \epsilon_q^2 \sigma_{xw}^2 \tag{69}$$

where

$$\sigma_{xw}^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_x(\omega) \cdot S_w(\omega) d\omega. \tag{70}$$

DPCM. Let $P(\omega) = Q(\omega)$ be the MMSE predictor for $S_x(\omega)$. It follows that

$$\min_{\text{DPCM}} \{\sigma_l^2\} = \epsilon_q^2 \eta_x^2 \sigma_w^2 \tag{71}$$

Note that optimal prediction has reduced the weighted error variance

in relation to PCM by a factor $\gamma_x^2 = \eta_x^2/\sigma_x^2$, i.e., by the spectral flatness measure of $S_x(\omega)$.

Prefiltered DPCM. Let $Q(\omega)$ be the MMSE predictor for $S_x(\omega)$. DPCM results if $P(\omega) = Q(\omega)$. Prefiltered DPCM results if $P(\omega) \neq Q(\omega)$. By varying $P(\omega)$ we obtain

$$\min_{\substack{\text{Prefiltered} \\ \text{DPCM}}} \{\sigma_l^2\} = \epsilon_q^2 \eta_x^2 \sigma_{\sqrt{w}}^4 \tag{72}$$

where

$$\sigma_{\sqrt{w}}^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sqrt{S_w(\omega)}\, d\omega. \tag{73}$$

From Schwarz's inequality we have $\sigma_{\sqrt{w}}^4 \leqslant \sigma_w^2$. Hence we find the result that the performance of DPCM can be improved by employing an additional prefilter. It is worth emphasizing, however, that we have set $Q(\omega)$ to be the MMSE predictor for $S_x(\omega)$. The optimal NFC scheme results if we are also free to optimize this feedback filter (see below).

NFC. The minimum total error variance of NFC schemes has already been given in Section 3.2 and is repeated here for completeness:

$$\min_{\text{NFC}} \{\sigma_l^2\} = \epsilon_q^2 \eta_x^2 \eta_w^2. \tag{60}$$

Note that optimal prediction *and* optimal noise shaping properties are provided by NFC schemes. Accordingly the total error variance is reduced in relation to PCM by a factor $\gamma_x^2 \cdot \gamma_w^2$ which is the product of the spectral flatness measures of input spectrum and weighting function. We shall see very shortly that this scheme is very close to theoretical bounds in the case of Gaussian input sequences.

Example 5: We shall now evaluate the above derived results for the specific example of a first-order Markov source of variance σ_x^2 and normalized adjacent-sample correlation $\rho \geqslant 0$ whose power density spectrum has already been given in Example 3. Such a source is a useful first approximation for modelling the statistics of speech (with $\rho = 0.85$ in this example) and of television signals (with $\rho = 0.9625$ in this example).

Speech signals. We assume a weighting function which is inversely proportional to the pds of the signal. We have already mentioned in Section 3.2 that subjective gains of about 6–10 dB have been reported for this specific weighting. The foregoing analysis has also revealed that the optimal scheme, i.e. NFC is identical to D*PCM. We have

$$S_w(\omega) = c^2 \cdot S_x^{-1}(\omega) \tag{74}$$

where c^2 can easily be determined if $S_w(\omega)$ is normalized such that

$$\max_{\omega} \{S_w(\omega)\} = 1 \tag{75}$$

Table III — Comparison of coder design parameters

	$S_w(\omega) = 1$	$S_w(\omega) \propto S_x^{-1}(\omega)$
η_x^2	$(1 - \rho^2) \cdot \sigma_x^2$	$(1 - \rho^2) \cdot \sigma_x^2$
η_w^2	1	$\frac{1}{(1 + \rho)^2}$
σ_w^2	1	$\frac{1 + \rho^2}{(1 + \rho)^2}$
σ_{xw}^2	σ_x^2	$\eta_x^2 \cdot \eta_w^2$
$\sigma_{\sqrt{xw}}^2$	$\frac{2}{\pi} \eta_x F\left(\frac{\pi}{2}, \rho\right)$	$\frac{\eta_x}{1 + \rho} = \sigma_{xw}$
$\sigma_{\sqrt{w}}^2$	1	$\frac{2}{\pi} E\left(\frac{\pi}{2}, k\right)$

and if, in addition, use is made of the following result:

$$\min_{\omega} \{S_x(\omega)\} = \frac{1 - \rho}{1 + \rho} \sigma_x^2. \tag{76}$$

Table III lists the important quantities for $S_w(\omega) = 1$ and $S_w(\omega) \propto S_x^{-1}(\omega)$. $F(\cdot,\cdot)$ is the normal elliptic integral of the first kind [see eq. (45)], and $E(\cdot,\cdot)$ is of the second kind. We have

$$E\left(\frac{\pi}{2}, k\right) = \frac{\pi}{2}\left(1 - \frac{1}{4} k^2 - \frac{3}{64} k^4 - \ldots\right). \tag{77}$$

The quantity k is given by

$$k = \sqrt{4\rho}/(1 + \rho) \tag{78}$$

and $E(\pi/2, k)$ is close to unity for $\rho \geq 0.85$.

From Table III we find that the total error variance can be reduced by a factor $\gamma_x^2 = \eta_x^2/\sigma_x^2 = 1 - \rho^2$ by means of optimal prediction and additionally by a factor $\gamma_w^2 = \eta_w^2/\sigma_w^2 = (1 + \rho)^{-2}$ by means of optimal noise shaping. For $\rho = 0.85$ the obtainable improvements in weighted signal-to-noise ratio are 5.6 dB and 2.4 dB, respectively. It is interesting to note that a noise-shaping improvement of about 1 dB has been calculated in Refs. 25 and 31 on the basis of experimentally determined speech spectra and noise-weighting functions. The total improvement is 8.0 dB for D*PCM or NFC, and it is 6.5 dB for prefiltered DPCM.

The achievable error variances of various coding schemes can be determined by using eqs. (60) and (64)–(73):

PCM:
$$\min \{\sigma_t^2\} = \epsilon_q^2 \frac{1 + \rho^2}{(1 + \rho)^2} \sigma_x^2 \tag{79a}$$

Noise-feedback PCM: $$\min\{\sigma_t^2\} = \epsilon_q^2 \frac{1}{(1+\rho)^2}\sigma_x^2 \tag{79b}$$

D*PCM and NFC: $$\min\{\sigma_t^2\} = \epsilon_q^2 \frac{1-\rho}{1+\rho}\sigma_x^2 \tag{79c}$$

DPCM: $$\min\{\sigma_t^2\} = \epsilon_q^2 \frac{(1-\rho)(1+\rho^2)}{(1+\rho)}\sigma_x^2 \tag{79d}$$

Prefiltered DPCM: $$\min\{\sigma_t^2\} \approx \epsilon_q^2(1-\rho^2)\left(\frac{2}{\pi}\right)^2\sigma_x^2. \tag{79e}$$

Television signals. An average video spectrum is flat for frequencies below the line rate and falls at about 6 dB per octave through the rest of the band. A weighting function for such a signal has a negative slope of about 3 dB per octave.[7,31,32] We assume a weighting

$$S_w(\omega) = c^2\sqrt{S_x(\omega)} \tag{80}$$

and it is not difficult to show that the reduction factor for noise shaping is

$$\gamma_w^2 = \frac{\pi}{2F\left(\frac{\pi}{2},\rho\right)} = \frac{1}{1+\frac{1}{4}\rho^2+\frac{9}{64}\rho^4+\ldots} \tag{81}$$

for this choice of the weighting function. Using $\rho = 0.9625$, we find a performance improvement of 11.3 dB obtainable by optimal prediction and a noise-shaping gain of 2.4 dB. This latter figure is in good agreement with the results of calculations based on experimental data.[25,31]

3.4 Absolute performance bounds

In the foregoing sections we have optimized various coding schemes whose structures had been given beforehand. It is useful to compare the performance of this class of encoders-decoders with absolute performance bounds by consulting the distortion-rate function.[12] A detailed discussion of these bounds for speech and television signals has been given by O'Neal.[31] In the following we shall restrict our attention to source coding of stationary ergodic Gaussian processes. For a fixed rate R there exists a minimum possible average distortion D which is a lower bound for any coding scheme. As above we adopt a frequency-weighted mean-squared error as our distortion measure. D and R are then related parametrically as follows:[33,34]

$$D(\varphi) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \min\{\varphi, S_w(\omega)\cdot S_x(\omega)\}d\omega \tag{82}$$

$$R(\varphi) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \max\left\{0, \log_2\frac{S_w(\omega)\cdot S_x(\omega)}{\varphi}\right\} d\omega. \tag{83}$$

Next we define the weighted error spectrum by

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_t(\omega) d\omega \tag{84}$$

and, by comparing it with eq. (82), we find

$$S_t(\omega) = \begin{cases} \varphi & \varphi \leqslant S_w(\omega) \cdot S_x(\omega) \\ S_w(\omega) S_x(\omega) & \text{otherwise.} \end{cases} \tag{85}$$

Over the frequency range where $\varphi > S_w(\omega) \cdot S_x(\omega)$ we have $S_t(\omega) = S_w(\omega) \cdot S_x(\omega)$. This implies that no signal has to be transmitted over this frequency range since such a measure produces just this error spectrum. Over the frequency range where $\varphi \leqslant S_w(\omega) \cdot S_x(\omega)$ we have $S_t(\omega) = \varphi$, i.e., the weighted error spectrum must be constant. Section 3.2 has shown that this requirement is met by the NFC scheme. In this latter case of small distortions we have $D(\varphi) = \varphi = D$ and, by combining eqs. (22), (49), and (83), we obtain

$$D = 2^{-2R} \cdot \eta_x^2 \cdot \eta_w^2. \tag{86}$$

A comparison with eq. (60) indicates that optimal NFC coding is very close to the distortion-rate bound D. The difference is in the first right-hand term of these equations because $\epsilon_q^2 > 2^{-2R}$ for single-letter quantizers (see Table IV). We finally note that the three right-hand terms in eq. (86) correspond to the terms T_B, T_P, and T_S in O'Neal's paper.[31]

IV. TRANSMISSION ERRORS

Noise-feedback coding schemes (including D*PCM and DPCM) are affected differently from PCM systems by bit errors on the communication channel because the decoder loop causes an error propagation while a PCM error does not propagate in time. The objective of this section is to show the effects of transmission errors in predictive coding systems using some of the results of our above analysis. We shall concentrate on two coding schemes, D*PCM and DPCM, and we shall only use the unweighted mean-squared error criterion. Recall that DPCM and NFC are identical in this case.

4.1 PCM

Let us assume that quantizing noise q_n and channel noise c_n can be modelled as additive noise sources. Thus the total error is

$$t_n = q_n + c_n \tag{87}$$

and its variance is

$$\sigma_t^2 = \sigma_q^2 + \sigma_c^2 + 2E[q_n \cdot c_n]. \tag{88}$$

Table IV — Quantization error variances ϵ_q^2 (Max-quantizers) and channel coefficients γ

	ϵ_q^2		γ	
	1 bit	2 bit	1 bit	2 bit
Uniform pdf	0.25	0.0063	3.0	3.75
Gaussian pdf	0.363	0.118	2.55	4.65
Laplacian pdf	0.5	0.176	2.0	5.3
Gamma pdf	0.667	0.232	1.33	6.28

Totty and Clark have shown[35] that channel errors and quantization errors are uncorrelated if the quantizer structure is that of Max.[36] These quantizers minimize the variance of the quantization noise but not necessarily that of the total error. This approach is of interest if a coding scheme has to operate on noisy channels with small bit-error probabilities which additionally are unknown or changing.† It is also justified by our observation that the step-size of quantizers with a low number of levels is not critical. The channel error variance depends on the bit-error probability P, on the density function of the signal being quantized, and on σ_x^2 (because the input variance determines the quantizer step-size scaling). Thus we have

$$\min_{\text{PCM}} \{\sigma_c^2\} = \epsilon_c^2 \cdot \sigma_x^2 \tag{89}$$

and the normalized channel error variance can be written as

$$\epsilon_c^2 = \gamma \cdot P \tag{90}$$

provided that the codewords are only affected by single bit errors. The channel coefficients γ can be derived following an approach in Ref. 39. Table IV lists these values for 1-bit and 2-bit quantizers.[19] In the case of 1 bit, the channel coefficient is simply given as

$$\gamma = 4 \cdot (1 - \epsilon_q^2). \tag{91}$$

In Table IV the γ-values for 2-bit quantizers are given for the folded binary code with the exception of the uniform probability density function whose γ-value is lower in the case of a natural binary code and is given by:[40]

$$\epsilon_c^2 = 4 \cdot P(1 - \epsilon_q^2) \tag{92}$$

where

$$\epsilon_q^2 = 2^{-2B} \tag{93}$$

with B as the number of bits.

† A re-optimization of quantizers for noisy channels has been discussed in Refs. 37 and 38.

The total error variance can thus be calculated as

$$\min_{\text{PCM}} \{\sigma_t^2\} = (\epsilon_q^2 + \epsilon_c^2) \cdot \sigma_x^2 = (\epsilon_q^2 + \gamma \cdot P) \cdot \sigma_x^2 \tag{94}$$

on the assumption of a vanishing correlation between the two errors.

4.2 Noise-feedback coding

The analysis of the predictive quantizing systems in the presence of channel errors is essentially identical to that of the last section if we substitute the nonwhite noise source $S_n(\omega)$ in eq. (47) with

$$\begin{aligned} S_n(\omega) &= \sigma_q^2 |1 - Q(\omega)|^2 + \sigma_c^2 \\ &= [\epsilon_q^2 |1 - Q(\omega)|^2 + \epsilon_c^2] \cdot \sigma_d^2 \end{aligned} \tag{95}$$

and if the assumption of eq. (54) still holds. Thus it is possible to reoptimize the various coders by following the same procedure as in the last section. We shall concentrate in this section on two coding schemes, D*PCM and DPCM. The objective here is to show that the D*PCM performance provides a bound for all predictive quantizing schemes if the channel is noisy.

For both schemes the contribution of the channel errors σ_c^2 on the total error variance can directly be derived from the D*PCM results of Section 2.3 by replacing ϵ_q^2 with ϵ_c^2. We then have, from eq. (29),

$$\begin{aligned} \text{D*PCM: } \sigma_c^2 &= \epsilon_c^2 \cdot \alpha \cdot \sigma_{d*}^2 \\ \text{DPCM: } \sigma_c^2 &= \epsilon_c^2 \cdot \alpha \cdot \sigma_d^2 \end{aligned} \tag{96}$$

where α is the power transfer factor of eq. (27). Therefore the variance of white noise on the channel is increased in the decoder network by a factor $\alpha \geqslant 1$. This error accumulation does not imply that the effect of transmission errors in D*PCM and DPCM is more severe than in PCM, because the *generated* noise variance depends now on the variance of the difference signal and can thus be influenced by the prefilter. Gains over PCM even for noisy channels have indeed been reported recently.[41,42] The discussion of Section 2.3 has already shown that the total MSE can be smaller than that of PCM; it has also shown that coder and decoder should have inverse networks.

4.2.1 D*PCM

Transmission errors contribute to the total error in exactly the same way as quantization noise. The total error

$$t_n = (q_n + c_n) * h_n \tag{97}$$

has a variance

$$\sigma_t^2 = (\epsilon_q^2 + \epsilon_c^2) \cdot \alpha \cdot \sigma_{d*}^2. \tag{98}$$

An optimized D*PCM scheme minimizes at the same time both error contributions. The mean-squared errors are those given in Section 2.3 if ϵ_q^2 is replaced with $\epsilon_q^2 + \epsilon_c^2$. Note that the optimal filters in D*PCM schemes do *not* depend on the bit-error rate. We have seen that half-whitening of the input spectrum provides a D*PCM performance bound. From eq. (36) we conclude that the minimum channel error variance is given as

$$\min_{\text{D*PCM}} \{\sigma_c^2\} = \epsilon_c^2 \sigma_{\sqrt{x}}^4. \tag{99}$$

4.2.2 DPCM

Section 3.2 has revealed that D*PCM has the same total quantization error variance as PCM if the input is full-whitened (MMSE prediction). The D*PCM postfilter is then identical with that of DPCM if this latter scheme has been optimized for the noiseless channel. These observations imply that transmission errors cause the same channel error variances in DPCM as in PCM if an MMSE predictor is employed. Smaller MSE values, i.e., improvements over PCM, can be gained by *reducing* the whitening effect. The quantization noise MSE, however, increases then. The total error

$$t_n = q_n + c_n * h_n \tag{100}$$

has a variance

$$\sigma_t^2 = (\epsilon_q^2 + \alpha\epsilon_c^2)\sigma_d^2. \tag{101}$$

In the case of high bit-error probabilities ($\alpha\epsilon_c^2 \gg \epsilon_q^2$) the total MSE is minimized if the prediction network performance is close to that of an optimized D*PCM coder and eq. (99) provides a lower bound in channel error variance for DPCM (and NFC) in the case of noisy channels. For low bit-error rates the best predictor will have a characteristic between full-whitening (error-free transmission) and half-whitening (D*PCM bound for noisy channels). Therefore

$$\min_{\text{PCM}} \{\sigma_c^2\} \geq \min_{\text{DPCM}} \{\sigma_c^2\} \geq \min_{\text{D*PCM}} \{\sigma_c^2\} \tag{102}$$

Example 6: Assume a previous-sample 2-bit DPCM and a Gaussian (not necessarily Markovian) source with adjacent-sample correlation $\rho = 0.85$. Let the bit-error probability be $P = 0.05$. From eq. (90) and Table IV we have $\epsilon_q^2 = 0.118$ and $\epsilon_c^2 = 0.233$. Figure 7 shows the dependence of the signal-to-noise ratio of this DPCM scheme on the value a of its predictor coefficient both for the noiseless and the noisy channel. The theoretical results have been calculated from eq. (101) using eqs. (16) and (31).

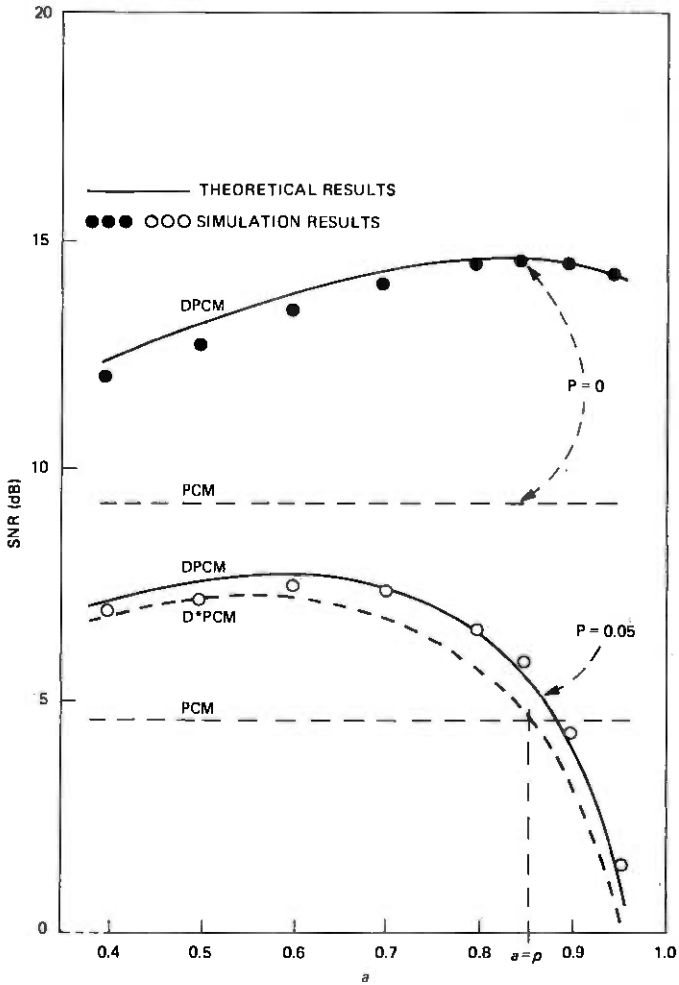


Fig. 7—DPCM performance on noisy channels. Two-bit quantization of Gaussian source with adjacent-sample correlation $\rho = 0.85$. Folded binary code with bit-error rates $P = 0$ and $P = 0.05$.

It is seen that DPCM performs better than PCM if the predictor is appropriately chosen. The differences between theory and measurements for low values of the predictor coefficient are again a consequence of the fact that the quantization error has been assumed to be not correlated with the input signal. For quantizers with a low number of levels this assumption only holds if the signal being quantized is uncorrelated, i.e. for predictor coefficients close to ρ. Notice that the crosspoint between PCM and DPCM performance is reached for a value of a which is slightly higher than ρ; this deviation from the predicted performance is a con-

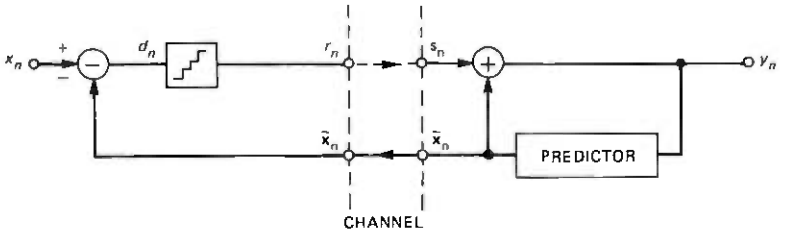


Fig. 8—Noiseless channel feedback DPCM.

sequence of the noise feedback. Figure 7 also compares the DPCM performance in the case of a noisy channel with the equivalent D*PCM performance which has been obtained from eqs. (98), (14), and (31). It is apparent that this one-tap D*PCM performance is a very useful bound of the DPCM performance. The optimal value of the DPCM predictor coefficient is also close to the optimal D*PCM coefficient a_{opt} as given in eq. (33) ($a_{opt} \approx 0.57$ for high bit-error probabilities). The choice of a_{opt} in accordance with eq. (33) for very noisy channels has first been mentioned in Ref. 41.

4.2.3 Noiseless channel feedback DPCM

In D*PCM the predictors of coder and decoder operate on slightly different signals, because there are quantizing noise and channel noise, respectively, in the intervening path. In DPCM both predictors operate on the same signal, viz. the sequence of reconstructed samples, if only the channel is noiseless. In the case of channel errors the predictions are again different, and the channel noise has the same effect on the overall MSE as in D*PCM. Let us assume that a noiseless channel from decoder to coder is available (see Fig. 8). It is then possible to ensure identical predictions via the feedback channel by retransmitting the prediction values calculated at the decoder. Notice that this scheme is identical to standard DPCM if the transmission of the encoded difference samples to the decoder is noiseless. In the case of channel errors, however, we have a total error

$$t_n = q_n + c_n \tag{103}$$

of variance

$$\sigma_t^2 = (\epsilon_q^2 + \epsilon_c^2) \cdot \sigma_d^2. \tag{104}$$

The contrast to D*PCM and standard DPCM is quite clear; the feedback is now around quantizer *plus* noisy channel, and thus error accumulation in the decoder loop has been totally avoided.

Example 7: In the case of previous-sample prediction the results of Example 1 (DPCM with analysis of the effects of noise feedback) apply. The optimum choice of the predictor-coefficient is given by eq. (17) if ϵ_q^2 is replaced with $\epsilon_q^2 + \epsilon_c^2$. To demonstrate the improvements let us use the figures of the previous example. We have $\epsilon_q^2 + \epsilon_c^2 = 0.351$, and hence $a_{opt} = 0.755$. We finally find [from eqs. (16) and (104)] a signal-to-noise ratio of 9.0 dB instead of 7.7 dB in the case of DPCM without noiseless channel feedback.

V. SUMMARY

There is a great interest in low bit-rate transmission of speech and television signals. Especially for acoustic signals it is well known that the subjective performance of a coder is strongly affected by the way in which quantizing distortion is distributed in frequency. In this paper we have compared coding schemes which employ prediction to exploit the inherent redundancy of these signals and which employ noise-shaping for optimizing the subjective performance on the basis of a frequency-weighted error criterion. First we have shown that DPCM outperforms D*PCM (a predictive scheme that lacks the feedback around the quantizer) for all nonwhite input spectra if the performance criterion is the unweighted total error variance. We have then used a noise-feedback coding structure as a framework for a unified analysis of predictive quantizing schemes. With this structure a minimum frequency-weighted error variance can be obtained by a proper shaping of the signal spectrum *and* the quantization noise prior to transmission. A comparison of this error variance with the distortion bound as given by the distortion-rate function for Gaussian signals has revealed that the performance of the noise-feedback coder is almost optimal for this class of signals. We have also shown that this coding structure degenerates to DPCM in the case of unweighted noise, and to D*PCM if the weighting function is inverse to the input spectrum. The performance results for these optimal coders have then been compared with those of suboptimal schemes including noise-feedback PCM and DPCM with prefiltering. For first-order Markov sources which often serve as a model of actual input spectra we have been able to derive simple explicit results in terms of the autocorrelation coefficient. In the last part we have examined the effects of channel transmission errors on the overall performance of these predictive quantizing schemes. We have shown that these coders when appropriately designed are less sensitive to channel errors than PCM. In D*PCM channel errors contribute to the total error in exactly the same way as quantization noise. Thus the D*PCM results provide a guideline for the optimization and a bound for the performance of DPCM.

VI. ACKNOWLEDGMENTS

The author wishes to thank A. Gersho, A. Wasiljeff, G. Wessels, and R. Zelinski for helpful suggestions. Thanks are also due to N. S. Jayant for significant contributions to the clarity of this paper.

REFERENCES

1. C. C. Cutler, "Transmission Systems Employing Quantization," U.S. Patent 2,927,962, 1960 (filed April 26, 1954).
2. J. L. Flanagan, *Speech Analysis, Synthesis, And Perception*, second edition, Berlin, Heidelberg, New York: Springer-Verlag, 1972.
3. J. V. Bodycomb and A. H. Haddad, "Some Properties of a Predictive Quantizing System," IEEE Trans. on Commun. Tech., *COM-18*, 1970, pp. 682–684.
4. J. P. Costas, "Coding with Linear Systems," Proc. IRE, *40*, 1952, pp. 1101–1103.
5. V. M. Shteyn, "On Design of Linear Predistortion and Correcting Systems," Radiotekn., *11*, No. 2, 1956, p. 60.
6. J. J. Spilker, "Theoretical Bounds on the Performance of Sampled Data Communications Systems," IRE Trans. Circuit Theory, *CT-7*, 1960, pp. 335–347.
7. R. A. Bruce, "Optimum Pre-Emphasis and De-Emphasis Networks for Transmission of Television by PCM," IEEE Trans. on Commun. Tech., *COM-12*, 1964, pp. 91–96.
8. L. M. Goodman and P. R. Drouilhet, "Asymptotically Optimum Pre-Emphasis and De-Emphasis Networks for Sampling and Quantizing," Proc. IEEE, *54*, 1966, pp. 795–796.
9. R. A. McDonald, Bell Laboratories, unpublished work.
10. B. G. Cramer, "Optimum Linear Filtering of Analog Signals in Noisy Channels," IEEE Trans. on Audio and Electroacoustics, *AU-14*, 1966, pp. 3–15.
11. T. Berger and D. W. Tufts, "Optimum Pulse Amplitude Modulation: Parts I and II," Trans. IEEE, *IT-13*, 1967, pp. 196–216.
12. T. Berger, *Rate Distortion Theory*, Englewood Cliffs, N.J.: Prentice Hall, Inc., 1971.
13. D. Chan and R. W. Donaldson, "Optimum Pre- and Postfiltering of Sampled Signals with Application to Pulse Modulation and Data Compression Systems," IEEE, Trans. on Commun. Tech., *COM-19*, No. 2, 1971, pp. 141–156.
14. E. G. Kimme and F. F. Kuo, "Synthesis of Optimal Filters for a Feedback Quantization System," IEEE Trans. on Circuit Theory, *CT-10*, 1963, pp. 405–413.
15. R. C. Brainard and J. C. Candy, "Direct-Feedback Coders: Design and Performance with Television Signals," Proc. IEEE, *57*, No. 5, 1969, pp. 776–786.
16. K. Nitadori, "Statistical Analysis of Δ-PCM," J. Inst. Electron. Commun. Eng., Japan, *48*, 1965, pp. 17–26.
17. R. A. McDonald, "Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems—Particular Applications to Voice Signals," B.S.T.J., *45*, No. 7 (September 1966), pp. 1123–1151.
18. P. Noll, "Nonadaptive and Adaptive Differential Pulse Code Modulation of Speech Signals," Polytechnisch Tijdschrift, Den Haag 1972, No. 19, pp. 623–629.
19. P. Noll, Bell Laboratories, unpublished work.
20. U. Grenander and G. Szegö, "Toeplitz Forms and their Applications," Berkeley, Calif.: University of California Press, 1958.
21. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, New York: Springer-Verlag, 1976.
22. P. Noll, "Some Properties of Differential PCM Schemes" (in German), Archiv für Elektronik und Übertragungstechnik (AEÜ), Bd. 30, 1976, pp. 125–130.
23. P. Noll, "Transform Coding of Speech," International Conference on Communications 1977, Conference Record Vol. 1, pp. 13.5-306–13.5-309.
24. D. J. Esteban and J. E. Menez, "Low Bit Rate Voice Transmission Based on Transversal Block Coding," Acoustical Society of America, 91st ASA Meeting, 1976, Washington, Paper RR 15.
25. H. G. Musmann, "Redundancy Reduction by Linear Transforms" (in German), Nachrichtentechnischer Fachbericht, NTF Band 40, 1970, pp. 13–27.
26. K. W. Cattermole, *Principles of Pulse Code Modulation*, Iliffe Books, Ltd., 1969.
27. W. Schlink, "On Source Encoding of PCM Speech Signals" (in German), Thesis, TU Braunschweig, 1976.

28. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, *63,* 1975, pp. 561–580.
29. H. A. Spang and P. M. Schultheiss, "Reduction of Quantizing Noise by Use of Feedback," IRE Trans. on Commun. Systems, 1962, pp. 373–380.
30. B. Heuser and P. Noll, "Error Feedback as a Means For Colouring Quantization Noise" (in German), Heinrich-Hertz-Institut Berlin, Tech. Rep. No. 191, 1976.
31. J. B. O'Neal, "Bounds on Subjective Performance Measures for Source Encoding Systems," IEEE Trans. on Inform. Theory, *IT-17,* 1971, pp. 224–231.
32. J. M. Barstow and H. N. Christopher, "Measurement of Random Video Interference to Monochrome and Color TV," Trans. AIEE, Commun. and Electronics, 1962, pp. 313–320.
33. R. L. Dobrushin and B. S. Tsybakov, "Information Transmission with Additional Noise," IRE Trans. Inform. Theory, *IT-8,* 1962, pp. 293–304.
34. R. A. McDonald and P. M. Schultheiss," Information Rates of Gaussian Signals under Criteria Constraining the Error Spectrum," Proc. IEEE (Correspondence), *52,* 1964, pp. 415–416.
35. R. E. Totty and G. C. Clark, "Reconstruction Error in Waveform Transmission," IEEE Trans. Inform. Theory (Correspondence), *IT-13,* 1967, pp. 336–338.
36. J. Max, "Quantizing for Minimum Distortion," IRE Trans. on Inform. Theory, *IT-6,* 1960.
37. A. J. Kurtenbach and P. A. Wintz, "Quantizing for Noisy Channels," IEEE Trans. on Communications, *COM-17,* No. 2 (1969), pp. 291–302.
38. B. R. Murthy, "Optimization in PCM, Adaptive PCM and DPCM Systems," Purdue University, Thesis, 1969.
39. P. Noll, "Effects of Channel Errors on the Signal-to-Noise Performance of Speech-Encoding Systems," B.S.T.J., 54, No. 9 (November 1975), pp. 1615–1637.
40. A. J. Viterbi, "Lower Bounds on Maximum Signal-to-Noise Ratios for Digital Communication Over the Gaussian Channel," IEEE Trans. on Communications, *COM-12* (1964), pp. 10–17.
41. K. Chang and R. W. Donaldson, "Analysis, Optimization and Sensitivity Study of Differential PCM Systems Operating on Noisy Communication Channels," IEEE Trans. on Communications, *COM-20,* No. 3 (1972), pp. 338–350.
42. J. E. Essman and P. A. Wintz, "The Effects of Channel Errors in DPCM Systems and Comparison with PCM Systems," IEEE Trans. on Communications, *COM-21,* No. 8 (1973), pp. 867–877.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

An Automatic Bias Control (ABC) Circuit for Injection Lasers

By A. ALBANESE

(Manuscript received March 18, 1977)

A totally electronic method of stabilizing the output light of an injection laser is presented. This novel method compensates the drift of the threshold current in gallium aluminum arsenide lasers by means of a feedback signal derived from the laser voltage.

I. INTRODUCTION

The threshold current of an injection laser varies from device to device and is a function of the device age and temperature. This threshold variation causes the laser output to change when the drive current is held constant. One must therefore provide a bias control circuit to compensate for the threshold variations.

Feedback circuits using photodetection have been successfully used for this purpose.[1] The output is monitored with an optical detector and compared with the input signal to generate an error signal that is fed back into the laser current.

This paper introduces a new concept of compensating the change in the laser threshold by using the electrical characteristics of junction lasers. An electrical circuit monitors the ac voltage and the ac current of the laser and generates the bias current needed to operate the laser above the threshold level independently of the laser temperature and age.

One of the benefits of using an electronic feedback scheme to stabilize the laser output is the elimination of an optical detector; this reduces the number of optical components required and may lead to a more economical and simpler solution than using a photodetector.

The electronic bias control method is based on the fact that the laser junction voltage saturates at currents above threshold.[2–8] Figure 1 shows L and V_j as functions of I. L is the laser output light at one of the faces, V_j is the laser junction voltage, and I is the laser current. Figure 1 also shows the changes in L and V_j produced by varying I in the vicinity of the laser threshold, I_t. The laser current is the sum of a bias current I_b

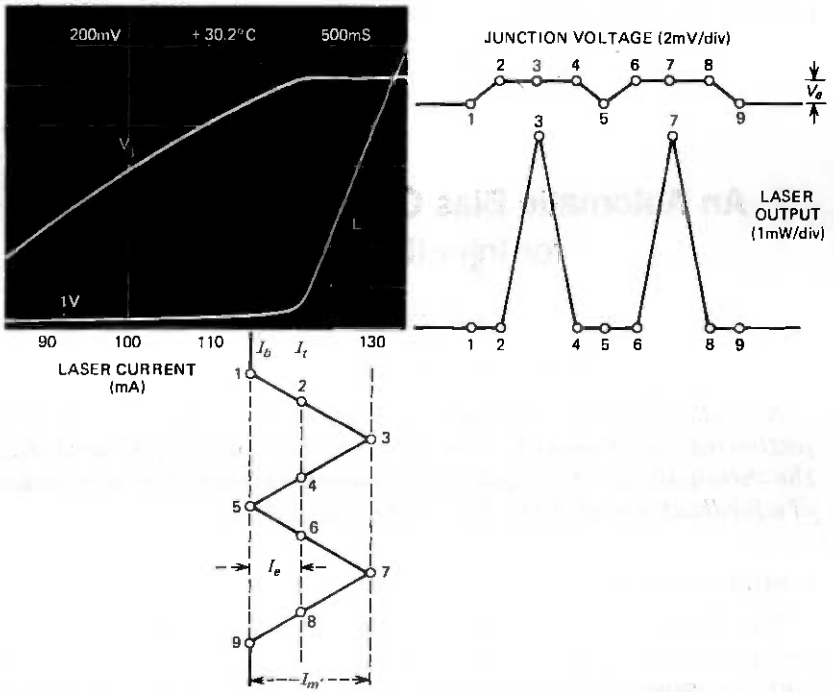


Fig. 1—Laser junction voltage and laser output light as a function of the laser current.

and a modulated current I_m. The part of I_m below I_t (called I_e) is an error current that does not modulate L but modulates V_j by a magnitude V_e. Conversely, the part of I_m above I_t modulates L but not V_j.

Laser stabilization consists of monitoring the junction voltage and increasing the laser bias current until the junction voltage is saturated. This is achieved automatically and continuously by an electronic circuit called an Automatic Bias Control (ABC) circuit. The ABC circuit monitors the laser voltage, generating an error signal proportional to the degree that the laser junction voltage is not saturated, and increases the bias current in the laser until the error voltage is minimized. The ABC circuit consists of an operational amplifier that amplifies the error voltage, a peak envelope detector that rectifies the amplifier output, and a current source that produces the current to bias the laser.

The circuit shown in Fig. 2 was built to study some of the characteristics of the electronic feedback method proposed here. Without stabilization provided by this circuit, the light output L of an injection laser varies strongly with temperature, as shown in Fig. 3 by the curve labeled "without feedback." Here, a constant bias of 124 mA was supplied, and the laser was modulated with a 70-kHz, 5-mA sinusoidal signal. One sees

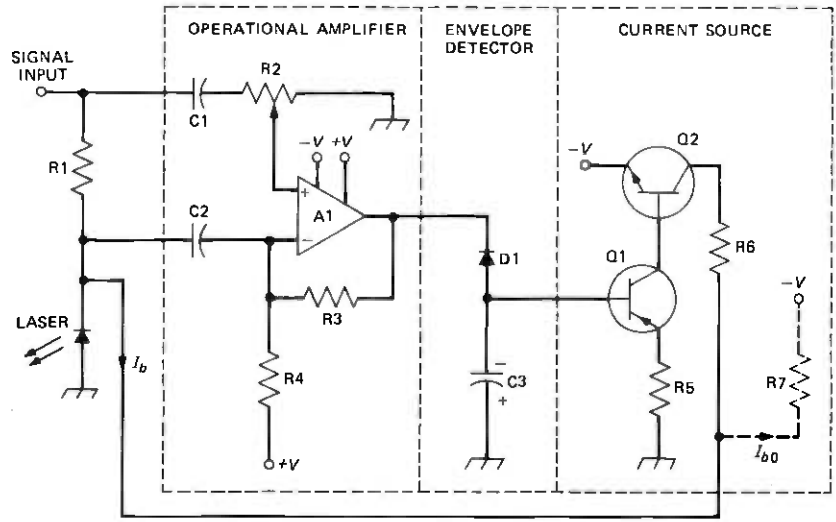


Fig. 2—Automatic bias control circuit.

immediately from the other curve, labeled "with feedback," how use of the ABC circuit improves the output stability. The variation $\Delta L/\Delta T$ is reduced from 0.37 mW/°C to 0.023 mW/°C, a factor of 16.

The following sections cover the details of the functioning and the limitations of the ABC circuit.

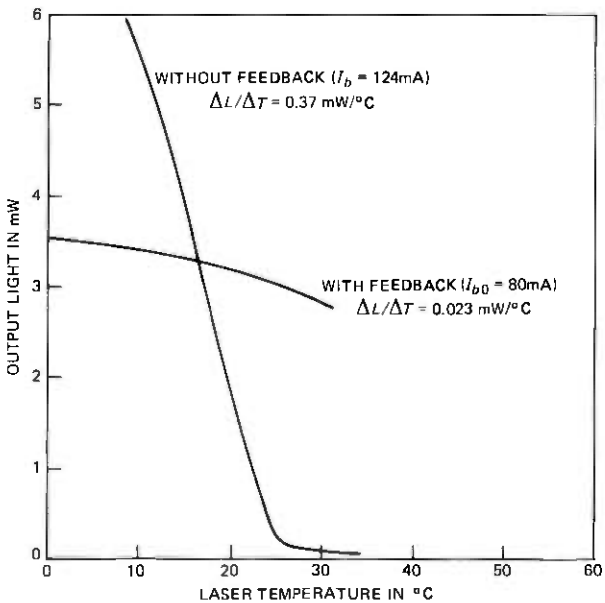


Fig. 3—Laser output as a function of the laser temperature.

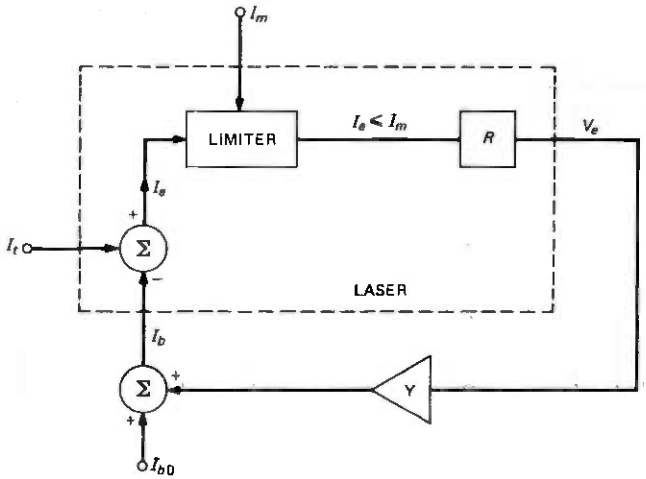


Fig. 4—System representation of the ABC circuit.

II. SYSTEM REPRESENTATION

In this section a mathematical model for the laser together with the ABC circuit is developed. This model is used to explain the function of the different parts needed to stabilize the laser output.

To analyze the operation of the ABC circuit, let us assume that the laser is biased with a direct current I_b which is smaller than the threshold current level I_t. Then, according to Fig. 1, V_e can be expressed as a first-order approximation by

$$V_e = \begin{cases} RI_e & \text{for } 0 \leq I_e \leq I_m \\ RI_m & \text{for } I_m \leq I_e \end{cases} \tag{1}$$

where

$$R = \frac{dV_j}{dI}, \text{ evaluated at } I \approx I_t^-, \tag{2}$$

and

$$I_e = I_t - I_b. \tag{3}$$

The purpose of the ABC circuit is to minimize I_e, by monitoring V_e to control I_b. Figure 4 shows a diagram of the ABC circuit and the laser. The laser is represented by the box indicated by dashed lines. It has one output, V_e, and three inputs, I_m, I_t and I_b. In principle the ABC circuit is a transadmittance amplifier, with a gain Y, that amplifies and converts V_e into I_b:

$$I_b = I_{b0} + (Y \times V_e). \tag{4}$$

I_{b0} is a fixed prebias current, which is smaller than the minimum I_t; I_{b0} is optionally provided to reduce the current in the output transistor and gain required of the transadmittance amplifier.

Equations (1) through (4) determine the operating points of the laser to be:

$$I_b = I_t - \frac{I_t - I_{b0}}{1 + A} \quad \text{for} \quad 0 \leq \frac{I_t - I_{b0}}{1 + A} \leq I_m \tag{5}$$

and

$$I_e = \frac{I_t - I_{b0}}{1 + A}. \tag{6}$$

A is the closed-loop gain given by the product $R \times Y$. Equations (5) and (6) indicate that if $1 + A \gg (I_t - I_{b0})/I_m$ then

$$I_b \approx I_t \tag{5a}$$

and

$$I_m \gg I_e \approx 0, \tag{6a}$$

that is, the laser is biased at the laser threshold, the error current becomes small, and the modulating current is above the laser threshold.

III. INSTABILITY OF THE LASER OUTPUT

It will be assumed that the instability of the laser output is mainly caused by the variation of the laser threshold. The laser output power, L, produced by a fixed current I is

$$L = S(I - I_t) \tag{7}$$

where S is the differential quantum efficiency of the laser. Then the output instability is defined as:

$$U = \frac{dL}{dI_t} = -S. \tag{8}$$

When using the ABC circuit the laser has an additional bias current I_b, and in this case the power output is:

$$L_1 = S(I + I_b - I_t). \tag{9}$$

Then, according to Eq. (8), the instability of the laser driven by the ABC circuit is:

$$U_1 = \frac{dL_1}{dI_t} = -S\left(1 - \frac{dI_b}{dI_t}\right). \tag{10}$$

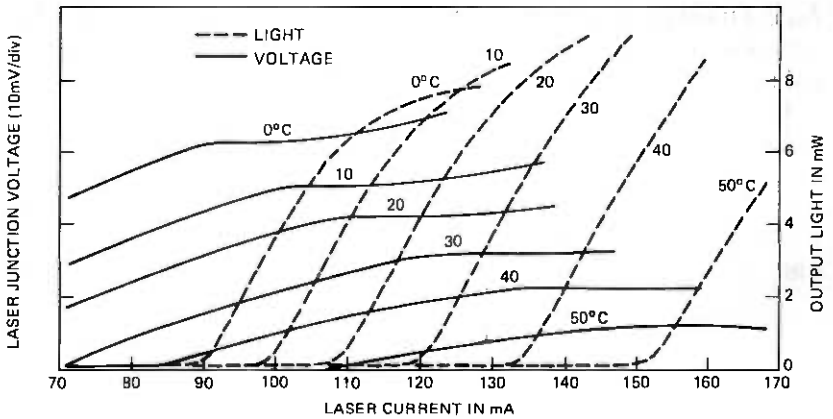


Fig. 5—Laser junction voltage and output light as a function of the bias current and the temperature.

Using Eq. (5) to calculate the derivative of I_b, one obtains

$$U_1 = \frac{-S}{A+1}. \tag{11}$$

The comparison of Eq. (11) with Eq. (8) indicates that, when the ABC circuit drives the laser, the instability of the laser output decreases $(A + 1)$ times, as would be expected for a closed-loop system of gain A.

IV. BIASING THE LASER BELOW THRESHOLD

This analysis has assumed so far that, above threshold, V_j is perfectly saturated and that V_e vanishes. This is not the physical case, because there is always a minimum V_e, called V_n, caused by the noise at the output of the operational amplifier and the lack of saturation of V_j (see Fig. 5). This deviation from the ideal situation requires one to limit the value of A such that

$$I_t - I_{b0} > \frac{A \times V_n}{R}; \tag{12}$$

otherwise the amplified noise will produce an $I_b > I_t$, and the circuit will overrespond to changes in V_e.

In order to be able to increase the value of A above that determined by eq. (12), a voltage ΔV is subtracted from V_e such that $\Delta V \gg V_n$. Then, the feedback loop increases I_b until $V_e = \Delta V$. $\Delta V \neq 0$ causes the laser to be biased below the threshold level. Figure 6 shows a diagram of the ABC similar to that of Fig. 4 but including ΔV. In this case, the new

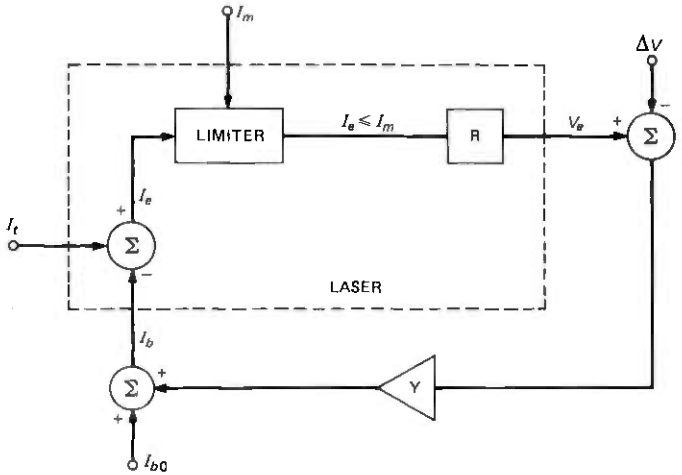


Fig. 6—System representation of the modified ABC circuit.

operating point of the laser is

$$I_e = \frac{I_t - I_{b0}}{A+1} + \frac{A}{A+1} \cdot \frac{\Delta V}{R}, \tag{13}$$

then for $A \gg \dfrac{(I_t - I_{b0})R}{\Delta V}$ and $A \gg 1$

$$I_e \approx \frac{\Delta V}{R} \tag{13a}$$

where $\Delta V > V_n$.

In eq. (13), ΔV is a constant but R is a function of I (see Fig. 7); R is the derivative of V_j, evaluated at a current below and near threshold. Because V_j is a logarithmic function of I,

$$R = \frac{dV_j}{dI} \approx \frac{V_T}{I_t} \tag{14}$$

where $V_T = nkT/q$. The constant n characterizes the semiconductor junction, k is the Boltzmann's constant, T is the junction temperature, and q is the electron charge.

By an analysis similar to that used for eq. (11), we can find a corrected value of the instability, using eqs. (14), (13), (3), (9), and (10). This new value, called U_2, considers the effect of having a $\Delta V \neq 0$:

$$U_2 = -S\left(\frac{1}{A+1} + \frac{A}{A+1} \cdot \frac{\Delta V}{V_T}\right). \tag{15}$$

The minimum and limiting value of U_2 is obtained by having an $A \gg 1$

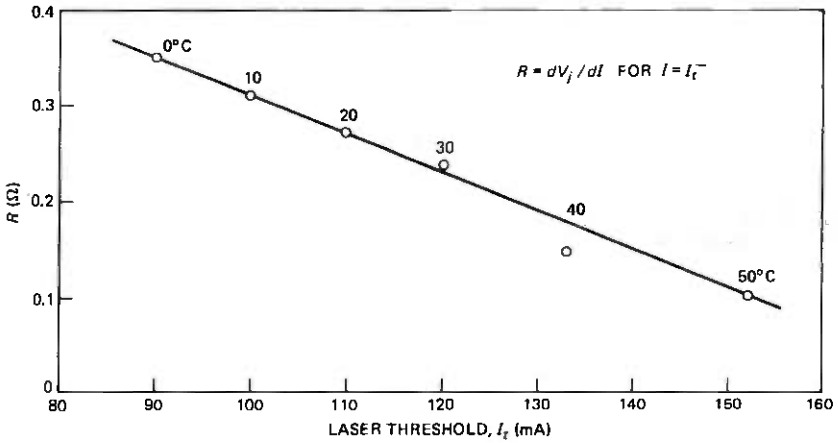


Fig. 7—R as a function of the laser threshold.

and $A \gg V_T/\Delta V$; then eq. (15) results:

$$U_2 \approx -S \frac{\Delta V}{V_T}. \tag{16}$$

According to eq. (16) the maximum improvement in the laser stability is determined by the ratio $V_T/\Delta V$. ΔV is caused by the nonsaturation of V_j, V_n. Figure 5 shows L and V_j as a function of I and T. At $L = 6$ mW and $T = 0°C$, V_n is less than 2 mV. Therefore one could make $\Delta V = 2$ mV. V_T can be computed from the data in Fig. 7 using eq. (14), $V_T = 32$ mV. For these values, one finds $V_T/\Delta V = 16$. According to eq. (15), $V_T/\Delta V$ determines the improvement in the laser instability caused by the ABC circuit.

V. ELECTRICAL CIRCUIT

The different parts of the ABC circuit are described in this section. As is shown in Fig. 2, the circuit consists of an operational amplifier, a peak detector, and a current source.

The operational amplifier has three functions: first, it determines V_j by compensating for the voltage drop in the laser series resistance; second, it filters the dc component of the junction voltage which depends on the laser temperature; and third, it amplifies the ac component to a value that can be processed by the envelope detector. In general, the bandwidth of the operational amplifier should be larger than the bandwidth of the modulating signal. This is required by the expression $AI_m \gg I_t - I_{b0}$, from eq. (5). But in certain cases, like that of video signals, one may use the synchronization signal to control the laser instead

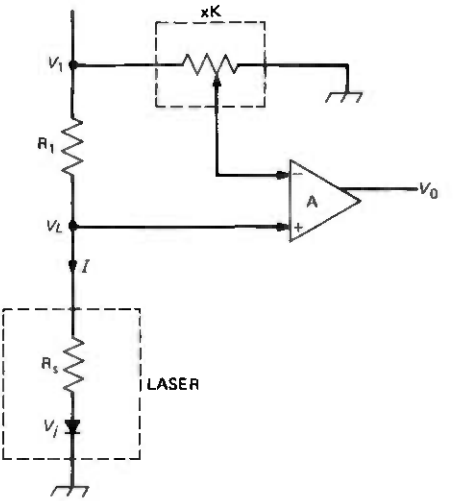


Fig. 8—Circuit to monitor the laser junction voltage.

of using the full bandwidth of the signal. In other cases one may add a low frequency signal for the ABC circuit.

Figure 8 shows a simplified portion of Fig. 2, with the two capacitors removed. The circuit was used to analyze the laser diode. It generates an output voltage V_o, which can be expressed in terms of V_j, the laser current I, the laser series resistance R_s, and other parameters of the circuit like the resistance R_1, the potentiometer ratio K, and the amplification A of the differential amplifier:

$$V_o = \{V_j(1 - K) + I[R_s - (R_s + R_1)K]\}A. \tag{17}$$

The circuit of Fig. 8, similar to a bridge circuit, compensates for the voltage drop across R_s by having K set to

$$K = \frac{R_s}{R_s + R_1} \tag{18}$$

which results in

$$V_o = \frac{A \times R_1}{R_1 + R_s} V_j \tag{19}$$

indicating that V_o is proportional to V_j. If a large A is desired, then the circuit should include an offset control to shift the output voltage so V_o is within the dynamic range of the amplifier. Figure 5 shows how the laser output light and V_o vary with I in the vicinity of threshold and at different temperatures of the laser. These data are useful to determine V_n; i.e., for $L = 6$ mW and $T = 0°C$, V_n is less than 2 mV.

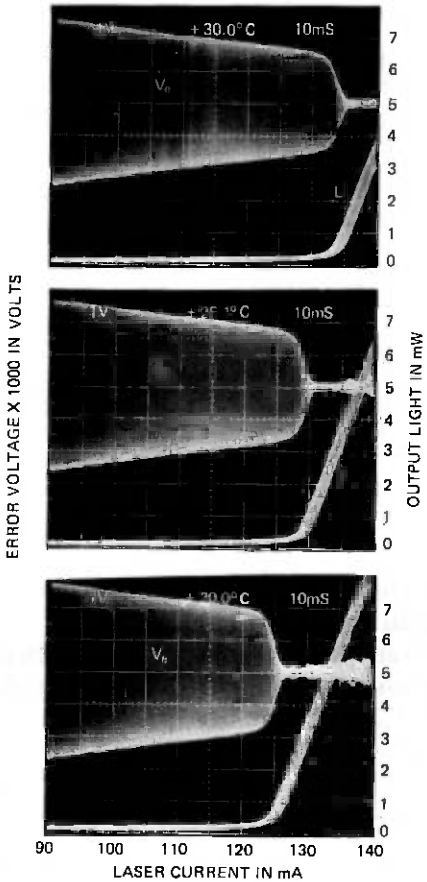


Fig. 9—Voltage at the output of the operational amplifier and laser output light as a function of the laser bias current at three different temperatures and without feedback connection.

In the circuit shown in Fig. 2, C_1 and C_2 remove the dc component of the input signal and the laser voltage, respectively; the potentiometer R_2 and the resistor R_1 compensate the voltage drop in R_s. The feedback resistor R_s determines the gain of the operational amplifier. Resistor R_4 sets the dc level of V_o, compensates the voltage drop across the diode D_1 and the emitter-base voltage of Q_1, and determines ΔV. In the circuit of Fig. 2, the peak output voltage is given by

$$V_o = \frac{R \times R_3}{R_s} \times I_e. \tag{20}$$

Figure 9 shows the open-loop output of the operational amplifier of Fig. 2 and the laser output light when the laser current is modulated by

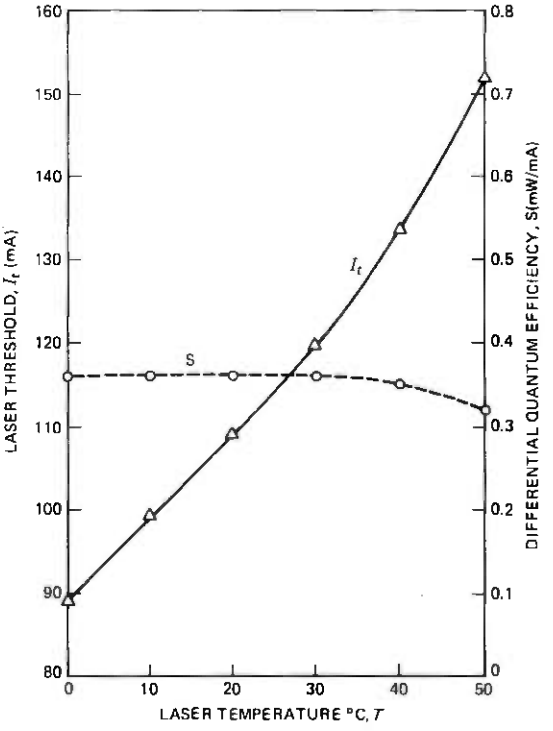


Fig. 10—Laser threshold and differential quantum efficiency as a function of the laser temperature.

a sinusoidal signal, as a function of the laser bias I_b at three different temperatures of the laser. The photographs clearly show that $V_e = V_n \approx 0$ for currents above threshold, and different laser temperatures.

The peak detector that determines the maximum error voltage consists of a rectifier and a low-pass filter. Different types may be used, and a simple one is shown in Fig. 2 which consists of a half-wave rectifier (D_1) and a low-pass filter formed by the capacitor C_3 and the input resistance of the current source.

The current source consists of an emitter-follower that converts the voltage from the peak detector into a current which is amplified by transistor Q_2. For the circuit of Fig. 2, the output of the current source is

$$I_b - I_{b0} = \frac{\beta_2 \times V_o}{R_5}, \tag{21}$$

where β_2 is the current amplification factor of Q_2. The peak output V_o of the operational amplifier $A1$ is given by eq. (20). According to Fig. 4,

one can define the current amplification A as

$$A = R \times Y = \frac{I_b - I_{b0}}{I_e} = \frac{\beta_2 \times R_3 \times R}{R_5 \times R_s}. \tag{22}$$

At room temperature, $I_t = 120$ mA, $R = 0.25\ \Omega$. $\beta_2 = 40$, $R_s = 2\ \Omega$, $R_3 = 100\ \Omega$ and $R_5 = 200\ \Omega$. One obtains an amplification $A = 2500$. This amplification satisfies the condition of eq. (13a).

VI. DISCUSSION

An electronic circuit to improve the stability of an injection laser was presented. Emphasis was placed on describing the fundamentals of the electronic feedback rather than comparing its performance and limitations with other methods of intensity control.

The ABC circuit assumes that the laser junction voltage saturates above threshold, and it did not consider laser anomalies like kinks in the $L - I$ curve nor changes in the differential quantum efficiency, S.

The changes in S were neglected because they are less important than the changes in I_t produced by temperature variation and age. This can be confirmed by looking at Fig. 10 which shows I_t and S as a function of the laser heat-sink temperature.

The circuit was operated with low-frequency signals below 100 kHz, and no effort was made to improve the frequency response of the amplifier so the ABC circuit could be used to transmit analog or digital signals at frequencies larger than 100 kHz.

The circuit described here may be useful to bias lasers near threshold during aging studies.

REFERENCES

1. P. W. Shumate, Jr., F. S. Chen, and P. W. Dorman, "GaAlAs Laser Transmitter for Lightwave Transmission Systems," to be published in B.S.T.J., July-August 1978.
2. P. G. Elisev, A. E. Krasil'nKov, M. A. Man'Ko, and V. P. Strakhov, "Investigation of DC Injection Lasers," *Physics of p-n Junctions and Semiconductor Devices*, S. M. Ryvkin and Yu V. Skmartsev, eds., New York: Plenum, 1971, p. 150.
3. D. L. Rode and L. R. Dawson, "Differential I/V of Heterostructure Correlates with Laser Threshold," Appl. Phys. Lett., *21* (August 1, 1972), pp. 90–93.
4. P. A. Barnes and T. L. Paoli, "Current-Voltage Characteristics of Double Heterostructure Injection Lasers," IEEE J. Quantum Electronics, *QE-12* (October 1976), p. 633.
5. T. L. Paoli, "Observation of Second Derivatives of the Electrical Characteristics of Double Heterostructure Junction Lasers," IEEE Trans. Electron Devices, *ED-23* (December 1976), p. 1333.
6. T. L. Paoli and P. A. Barnes, "Saturation of the Junction Voltage in Stripe-Geometry (AlGa) As Double-Heterostructure Junction Lasers," Appl. Phys. Lett., *28* (June 1976), pp. 714–717.
7. R. W. Dixon, "Derivative Measurements of Light-Current-Voltage Characteristics of (Ga,Al) As Double Heterostructure Lasers," B.S.T.J., *55*, No. 7 (September 1976), p. 973.
8. W. B. Joyce and R. W. Dixon, "Fundamental and Harmonic Response Voltages of a Sinusoidally-Current Modulated Ideal Semiconductor Laser," J. Appl. Phys., *47* (August 1976), p. 3510.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

More on Rain Rate Distributions and Extreme Value Statistics

By S. H. LIN

(Manuscript received October 14, 1977)

A new methodology is described for estimating the 5-minute rain rate distribution from yearly 5-minute maximum rain rate data and yearly accumulated rainfall data published by the National Climatic Center for U.S. locations. The method previously described gives the high rain rate portion of the distribution, whereas the extended methodology yields the complete distribution, which is assumed to be approximately lognormal. The three parameters characterizing the lognormal distribution can be calculated by application of the theory of extreme value statistics. The calculated results agree well with the 20-year data. The accuracy of the calculated results is limited by the instability of extreme rain rate data with a finite time base. Two-year rain rate data measured by a tipping bucket rain gauge at Palmetto, Georgia, are used to demonstrate that the time variation of rainfall process obeys a proportionate relationship, supporting the lognormal hypothesis.

I. INTRODUCTION

Reference 1 has described a methodology for calculating long-term distributions of high rain rates by applying the theory of extreme value statistics to the yearly maximum 5-minute rain rate data published by the National Climatic Center.[2,3] The obtained high rain rate distributions cover the range of interest to the engineering of terrestrial microwave radio links. However, for other applications, such as earth-satellite radio engineering, the rain rate distributions in the moderate and low rain rate ranges are also needed. This paper describes a methodology to obtain rain rate distributions covering the entire range (i.e., from below 5 mm/hr to greater than 200 mm/hr). The rain rate distributions are assumed to be approximately lognormal. The three parameters characterizing the lognormal distribution can be calculated from the yearly maximum 5-minute rain rate data and the yearly total accumulated

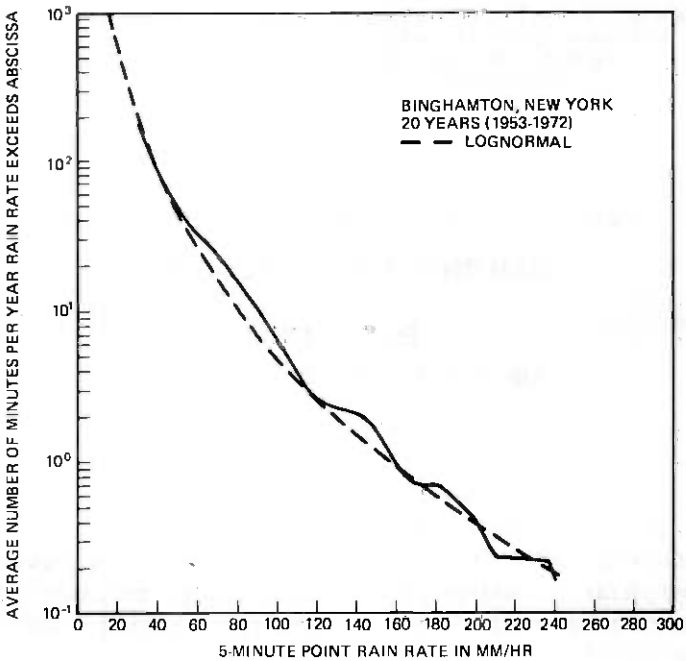


Fig. 1—Binghamton, New York: Comparison of 20-year distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year data (solid line).

rainfall data by applying the theory of extreme value statistics. The calculated results agree well with 20-year data as shown in Figs. 1 to 13 and Figs. 17 to 20.

Sections II and III describe the method and discuss the results. Section IV discusses characteristics of measured time variations of rain rates in support of the proportionate effect described by Aitchison and Brown.[10] The proportionate variation of rain rates is simply another manifestation of the lognormality of rain rate statistics.

In this paper, a "5-minute rain rate" corresponds to the average value of the randomly varying rain rate in a 5-minute interval and is calculated as $\Delta H/\tau$ where ΔH is the 5-minute accumulated depth of rainfall and $\tau = 5$ minutes or $1/12$ hour is the rain gauge integration time. The methodology is also applicable to integration times other than 5 minutes.

II. EXTREME VALUE STATISTICS AND LOGNORMAL RAIN RATE DISTRIBUTION

Many sets of rain rate data indicate that rain rate distributions can be closely approximated by the lognormal distribution (see Refs. 4 to

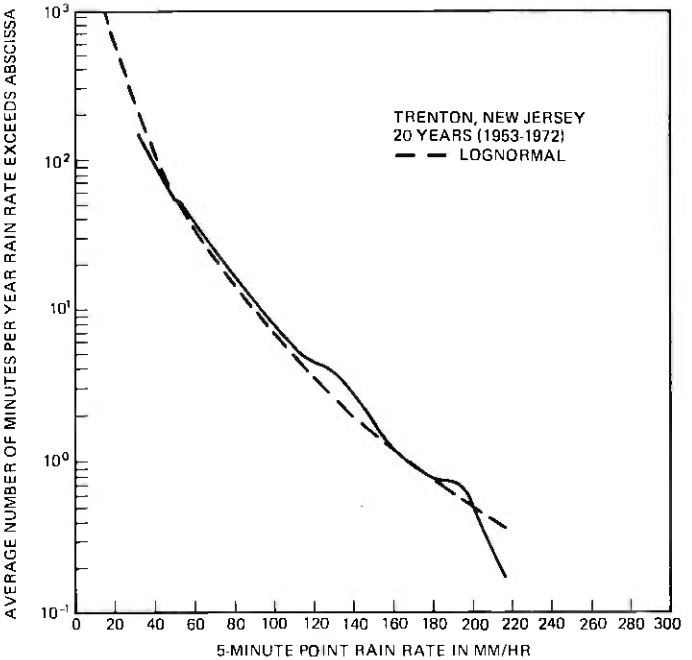


Fig. 2—Comparison for Trenton, New Jersey.

9, 23, 24 and Figs. 1 to 13 and 17 to 20 in this paper)*:

$$P(R \geq r) \approx P_0 \cdot \frac{1}{2} \operatorname{erfc}\left[\frac{\ln r - \ln R_m}{\sqrt{2}\, S_R}\right] \tag{1}$$

where R is the randomly varying 5-minute point rain rate, erfc(~) denotes the complementary error function, ln(~) denotes natural logarithm, S_R is the standard deviation of $\ln R$ during the raining time,[4] R_m in mm/hr is the median value of R during the raining time and P_0 is the probability that rain will fall at the point where the rain rate R is measured. Rain rate data usually emphasize high rain rate statistics with the result that the value of P_0, and hence the total raining time per year, are not directly available. In the following, it is demonstrated that the values of P_0, R_m and S_R, and hence the entire distribution $P(R \geq r)$, can be determined from the yearly maximum 5-minute rain rate data and the yearly total accumulated rainfall data.

Let W denote the long-term average value of the yearly accumulated depth of rainfall.† The relationship between W and the parameters in

* Figures 14, 15, 16, 21, and 22 are discussed later in Sections II and III.
† Excluding snowfall.

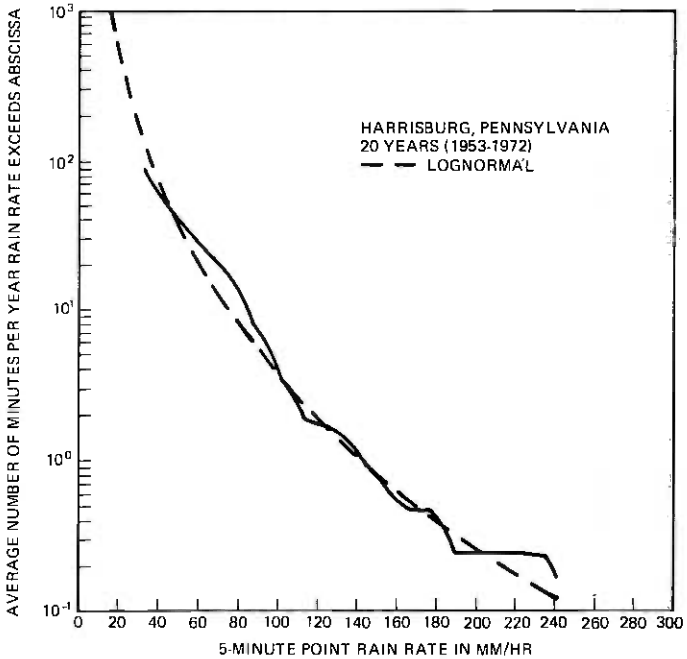


Fig. 3—Comparison for Harrisburg, Pennsylvania.

eq. (1) is

$$\begin{aligned} W &= \langle R \rangle \times \text{total raining time/year} \\ &= \langle R \rangle \times P_0 \times (8760 \text{ hours/year}) \\ &= R_m \times e^{S_R^2/2} \times P_0 \times (8760 \text{ hours/year}) \end{aligned} \tag{2}$$

where

$$\langle R \rangle = R_m \times e^{S_R^2/2} \tag{3}*$$

is the mean value of R during the raining time.[4] Long-term ($\geq$30 years) data on W for U.S. locations can be found in Refs. 2 and 11.

Let R_1 denote the yearly maximum 5-minute rain rate which varies from year to year. The distribution of R_1 is[1]

$$P(R_1 \geq r) = 1 - e^{-(e^{-y})} \tag{4}$$

where

$$y = \alpha(\ln r - U) \tag{5}$$

is called the reduced variate, α and U are scale and location parameters

* This relationship among $\langle R \rangle$, R_m and S_R holds if R is lognormal.[4,10]

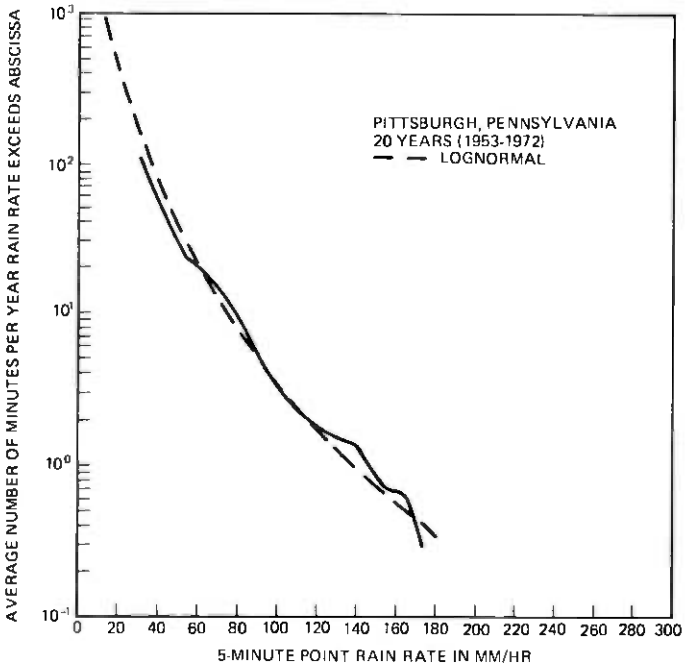


Fig. 4—Comparison for Pittsburgh, Pennsylvania.

respectively. Notice that the lognormal rain rate distribution (1) is uniquely determined by the three parameters P_0, R_m and S_R; whereas the distribution (4) of the yearly maximum 5-minute rain rate R_1 is uniquely determined by the two parameters α and U. Gumbel[12,13,22] has given the following approximate relationships among α, U and the parent distribution (1):

$$\Phi\left(\frac{U - \ln R_m}{S_R}\right) \simeq 1 - \frac{1}{P_0 \cdot N} \tag{6}$$

$$\alpha = \frac{P_0 \cdot N}{S_R}\, \phi\left(\frac{U - \ln R_m}{S_R}\right) \tag{7}$$

where

$$\Phi\left(\frac{U - \ln R_m}{S_R}\right) = 1 - \frac{1}{2}\,\mathrm{erfc}\left[\frac{U - \ln R_m}{\sqrt{2}\, S_R}\right] \tag{8}$$

is the standard unit normal distribution function,

$$\phi(z) = \frac{d}{dz}\,\Phi(z) \tag{9}$$

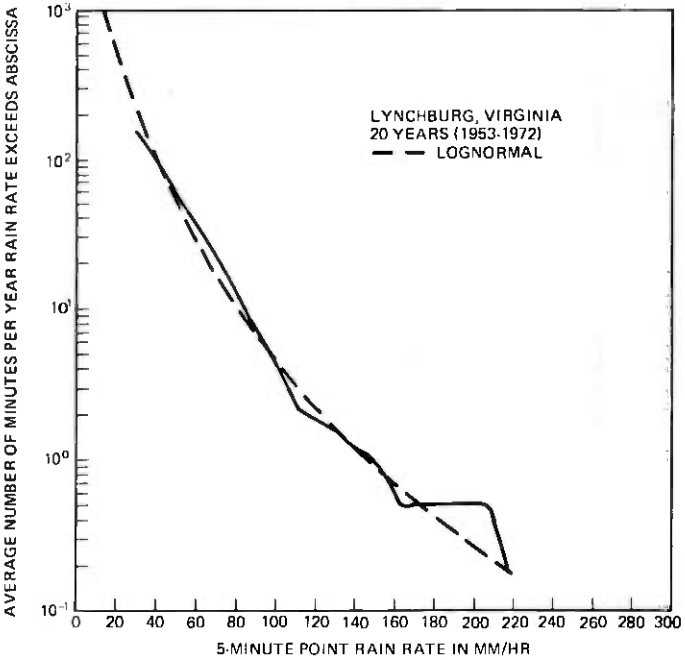


Fig. 5—Comparison for Lynchburg, Virginia.

is the normal probability density function, and

$$N = \text{total number of 5-minute intervals per year}$$
$$= (525600 \text{ minutes/year})/5 \text{ minutes}$$
$$= 105120. \tag{10}$$

From eqs. (4) and (5) it is easily shown[12,13] that U is the most probable value (i.e., the mode) of $\ln R_1$ where R_1 is the randomly varying yearly maximum 5-minute rain rate. Let us define

$$R_u = e^U. \tag{11}$$

Equation (6) states that, on long-term average, the randomly varying rain rate R will exceed R_u by approximately 5 minutes per year.* Equation (7) further specifies the slope (i.e., the derivative or probability density) of the rain rate distribution at $R \doteq R_u$. Solving eqs. (6) and (7)

* From eqs. (1) and (6), it is easily shown that

$$P(R \geq R_u) = P_0 \cdot \left\{1 - \Phi\left(\frac{U - \ln R_m}{S_R}\right)\right\} = \frac{1}{N}.$$

Multiplying this probability by the total time per year yields 5 minutes per year.

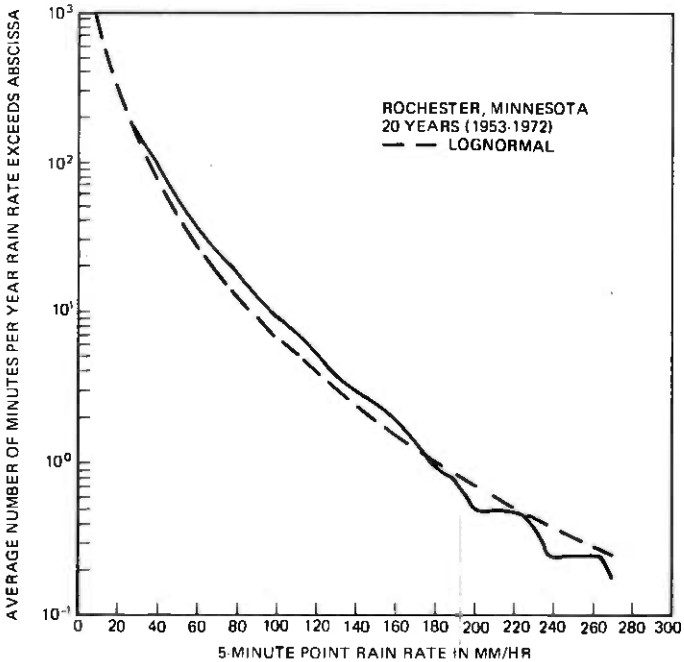


Fig. 6—Comparison for Rochester, Minnesota.

yields

$$S_R = \frac{P_0 \cdot N}{\alpha} \cdot \phi \left[\Phi^{-1} \left(1 - \frac{1}{P_0 \cdot N} \right) \right] \tag{12}$$

and

$$R_m = \exp \left[U - S_R \cdot \Phi^{-1} \left(1 - \frac{1}{P_0 \cdot N} \right) \right] \tag{13}$$

where $\Phi^{-1}(\sim)$ denotes the inverse normal probability function.

Reference 1 has given a set of formulas for calculating the parameters α and U from the yearly maximum 5-minute rain rate data. For completeness, this set of formulas is included in Appendix A. Knowing the values of W, α and U allows us to solve* the three equations (2), (12), and (13) for the three unknowns P_0, R_m and S_R. Substituting these three parameters into eq. (1) then yields the entire rain rate distribution.

For example, Table I lists the yearly maximum 5-minute rain rate R_1 measured at Binghamton, New York, for the 20-year period from 1953 to 1972.[2] Applying the formulas in Appendix A to the data in Table I

* These transcendental equations are solved numerically by a computer iteration process.

Table I — Yearly maximum 5-minute rain rates at Binghamton, New York

Year	Yearly maximum 5-minute rain rate, mm/hr
1953	103.63
1954	100.58
1955	161.54
1956	152.40
1957	91.44
1958	103.63
1959	201.17
1960	243.84
1961	112.78
1962	67.06
1963	115.82
1964	188.98
1965	91.44
1966	134.11
1967	85.34
1968	91.44
1969	106.68
1970	91.44
1971	97.54
1972	76.20

yields

$$\alpha = 3.224$$

$$U = 4.5736.$$

The 30-year (1941–1970) average value of W at Binghamton[2] is

$$W = 762 \text{ mm/year}.$$

Substituting this set of W, α and U into eqs. (2), (12), and (13) yields

$$P_0 = 0.018 \text{ (i.e., 1.8 percent)},$$

$$R_m = 2.631 \text{ mm/hr},$$

$$S_R = 1.1015 \text{ nepers}.$$

The lognormal distribution (1) of the 5-minute rain rates calculated from this set of P_0, R_m and S_R agrees closely with the 20-year data[14] as displayed in Fig. 1. Similarly, Figs. 2 to 13 show the close agreement between the calculated result and the 20-year data at 12 other locations.

However, high rain rate statistics require a very long time base to yield stable results. The sensitivity of the high rain rate distribution with respect to time base measured at Newark, New Jersey, is shown in Fig. 14. It is seen that increasing the time base from 19 years to 21 years significantly alters the distribution for rain rate beyond 150 mm/hr. J. W. King[15] and J. Xanthakis[16] have presented approximately 100 years of

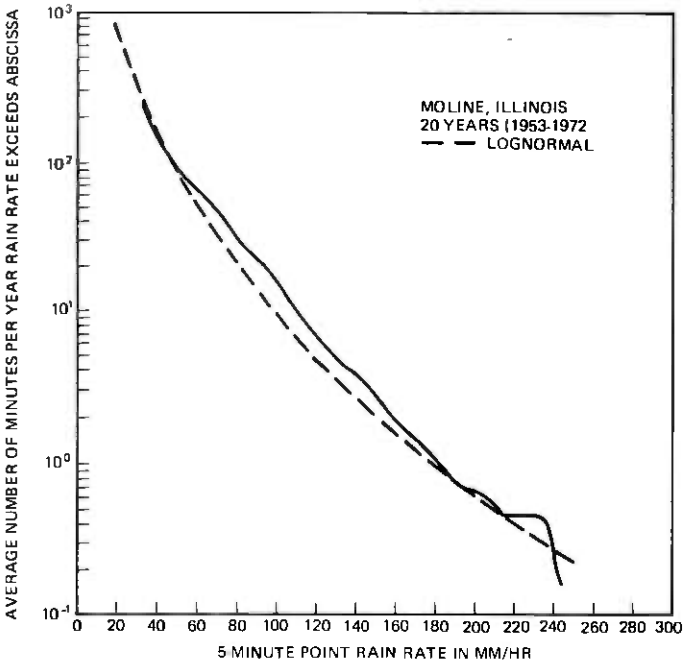


Fig. 7—Comparison for Moline, Illinois.

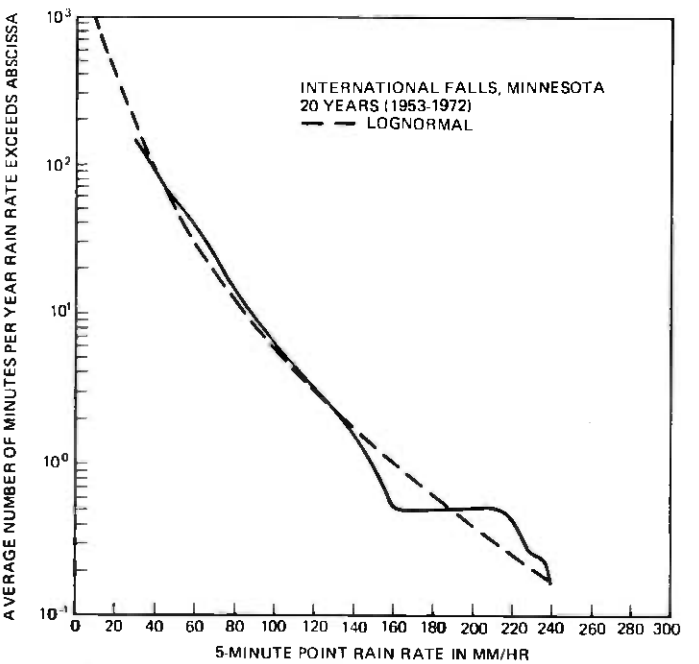


Fig. 8—Comparison for International Falls, Minnesota.

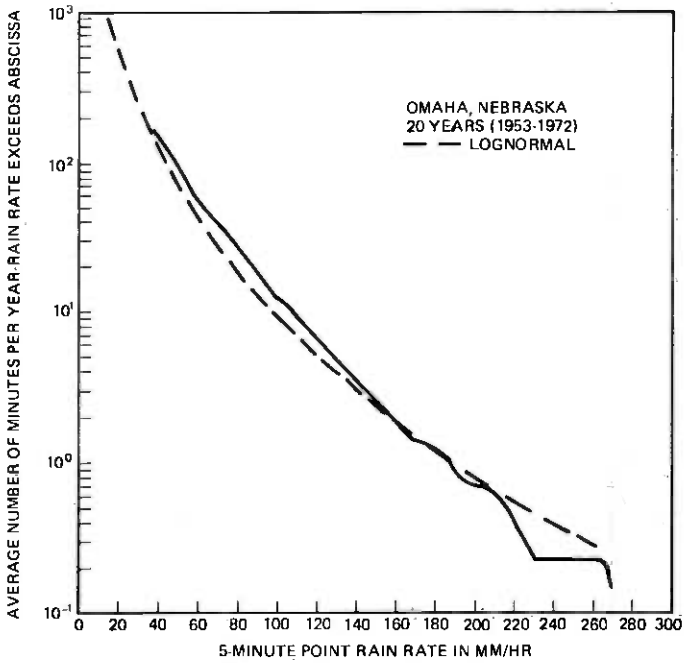


Fig. 9—Comparison for Omaha, Nebraska.

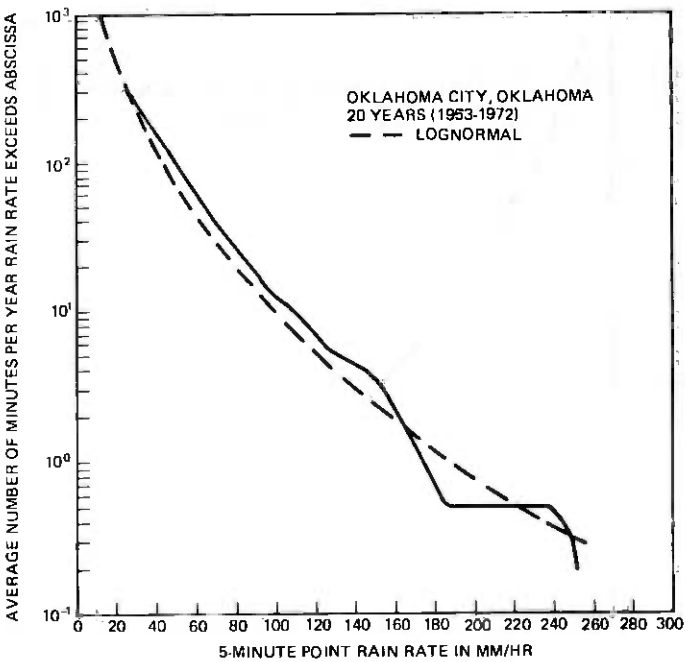


Fig. 10—Comparison for Oklahoma City, Oklahoma.

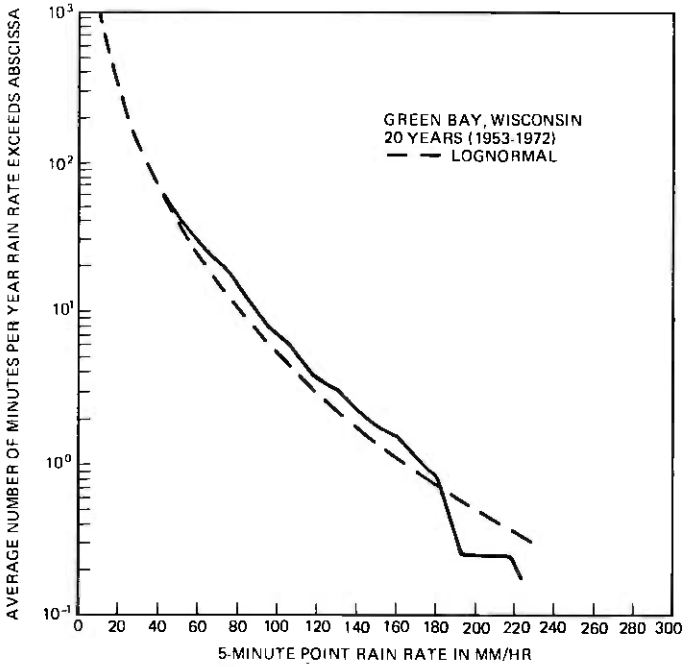


Fig. 11—Comparison for Green Bay, Wisconsin.

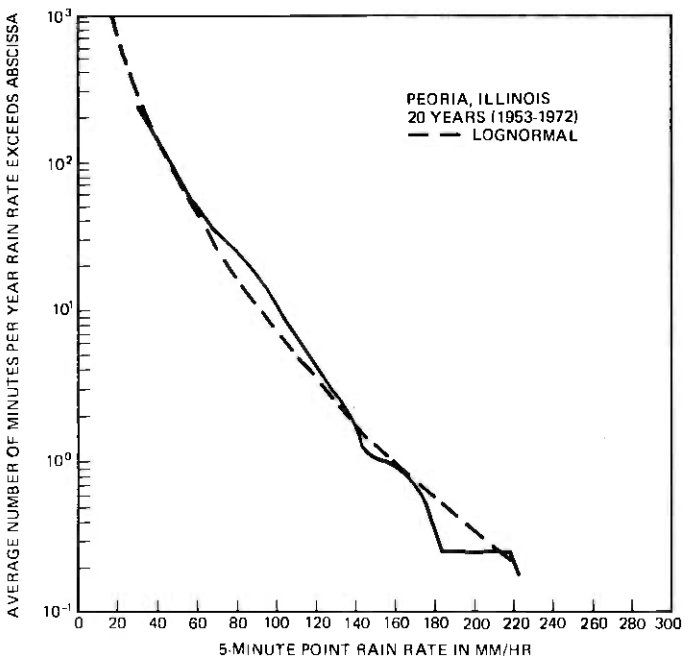


Fig. 12—Comparison for Peoria, Illinois.

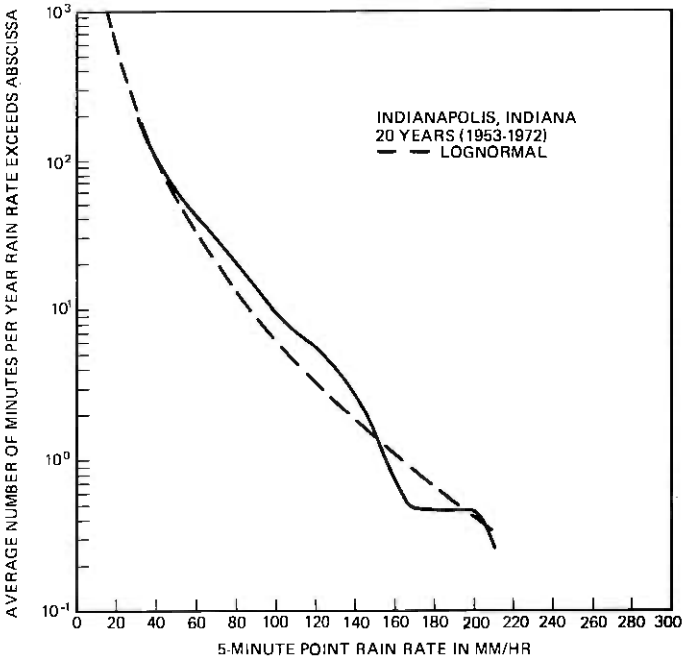


Fig. 13—Comparison for Indianapolis, Indiana.

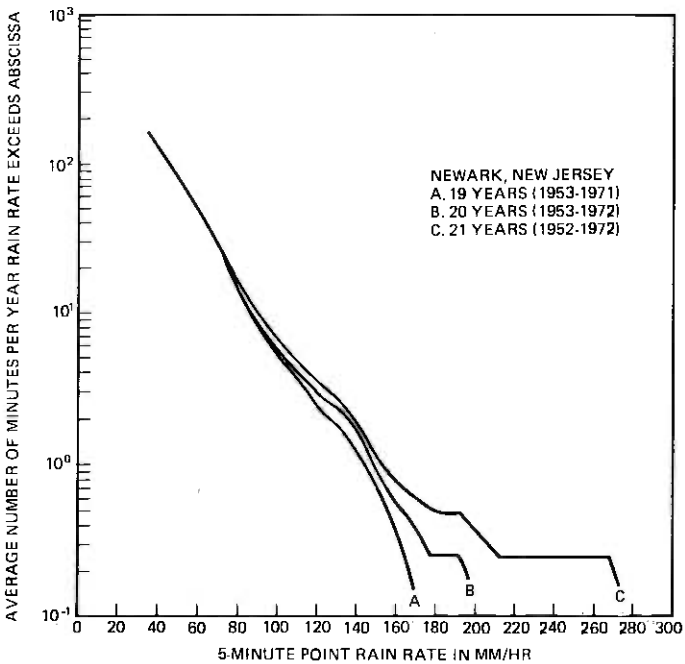


Fig. 14—The sensitivity of rain rate distribution with respect to time base.

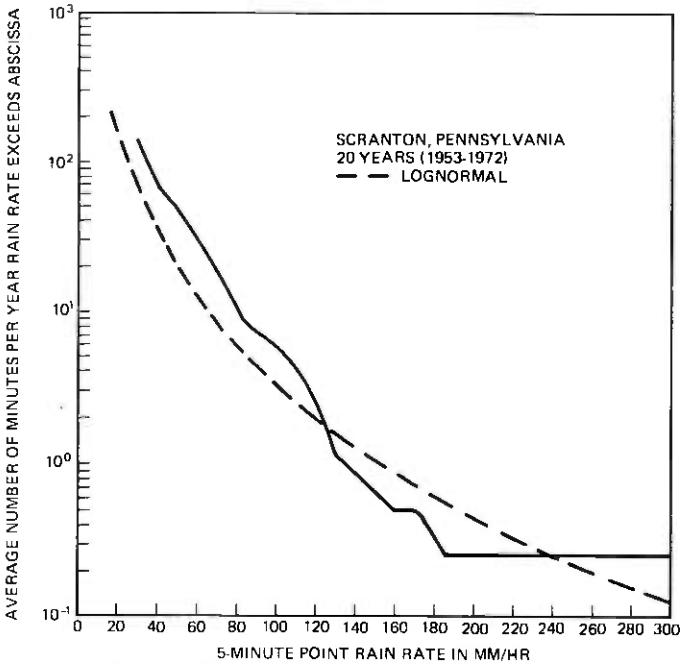


Fig. 15—Comparison for Scranton, Pennsylvania.

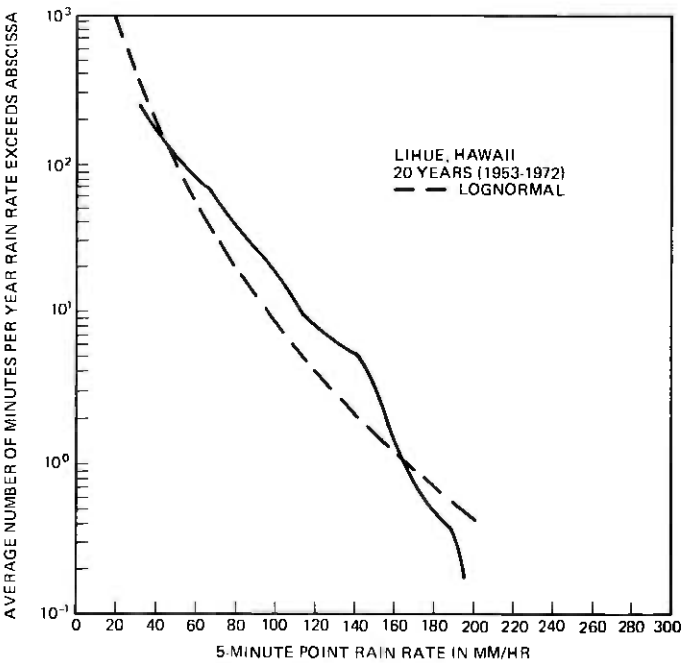


Fig. 16—Comparison for Lihue, Hawaii.

rainfall data to show that the variations of annual accumulated rainfall from year to year is correlated with the 11-year cyclic variations of sunspot numbers. R. J. Talbot and D. M. Butler[17] have discussed the climatic effects during passage of the solar system through nonuniform interstellar dust clouds. Therefore, rainfall activity is influenced not only by many terrestrial environmental factors but also possibly by extra-terrestrial sources. The required time base for stable rainfall statistics may be longer than 20 years. Since the parameters of the lognormal distribution are estimated from the yearly maximum rain rate data, the instability noted in Fig. 14 limits the accuracy of the calculated results.* Figures 15 and 16 give two examples of the effects of unstable high rain rate data on the estimated rain rate distributions.

III. FIFTY-YEAR DISTRIBUTIONS

Section IV of Ref. 1 describes a set of formulas for calculating the parameters α and U from rainfall intensity-duration-frequency curves for U.S. locations published by the Weather Bureau.[3] These curves are derived by the Gumbel method[12,13] using the theory of extreme value statistics and are based on approximately 50 years (1900–1950) of rainfall data. From this data source, we need only the following three numbers for a given location to calculate α and U:

M = the number of years of rainfall data from which rainfall-intensity-duration frequency curves are derived,

r_a = the extreme rain rate with 2-year return period, i.e., the rain rate which is exceeded once in 2 years, on the average, by the yearly maximum 5-minute rain rates,

r_b = the extreme rain rate with 10-year return period, i.e., the rain rate which is exceeded once in 10 years, on the average, by the yearly maximum 5-minute rain rates.

Therefore, in principle, long term distribution (1) of 5-minute rain rates for U.S. locations can easily be obtained by this method. The only input required are the four parameters W, M, r_a and r_b for each location read from Refs. 2 and 3.

For example, for San Francisco, California, the four numbers are

W = 115 mm/year

M = 48 years (1903–1950)

r_a = 1.9 inches/hr = 48.3 mm/hr

r_b = 3.05 inches/hr = 77.5 mm/hr.

* The accuracy of the calculated results for the very low rain rate region (i.e., $\leq$ 10 mm/hr) may be also limited because the parameters P_0, R_m and S_R are estimated from the extreme, high rain rate data.

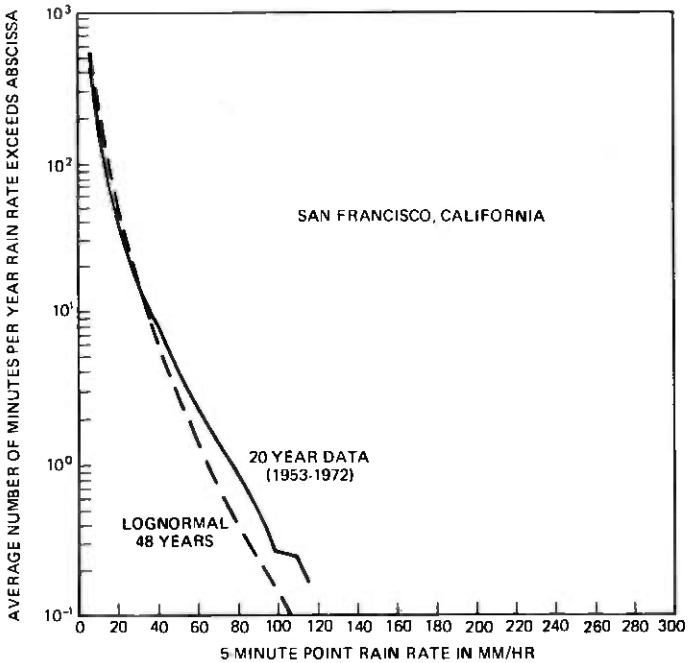


Fig. 17—Comparison of 48-year (1903–1950) distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year (1953–1972) data (solid line) at San Francisco, California.

The formulas for calculating α and U are given in Appendix B for completeness. By substituting these values of M, r_a and r_b into eqs. (26) to (31) we obtain

$$\alpha = 3.6297$$

$$U = 3.7786.$$

Substituting this set of W, α and U into eqs. (2), (12) and (13) yields

$$P_0 = 0.0016 \qquad \text{(i.e., 0.16 percent)}$$

$$R_m = 6.23 \text{ mm/hr}$$

$$S_R = 0.7771 \text{ neper.}$$

Figure 17 shows that the calculated lognormal distribution of 5-minute rain rates for the 48-year period (1903–1950) is reasonably close to the 20-year data (1953–1972). Similarly, Figs. 18, 19, and 20 show the agreement between calculated results ($\geq$43 years) and the 20-year data. On the other hand, Figs. 21 and 22 give two examples of appreciable differences due to the instability of high rain rate data.

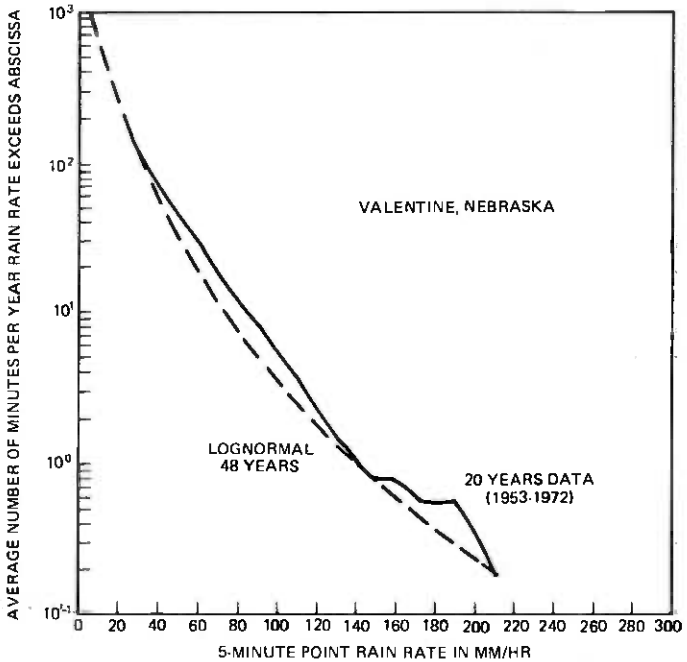


Fig. 18—Comparison of 48-year (1903–1906, 1908–1951) distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year (1953–1972) data (solid line) at Valentine, Nebraska.

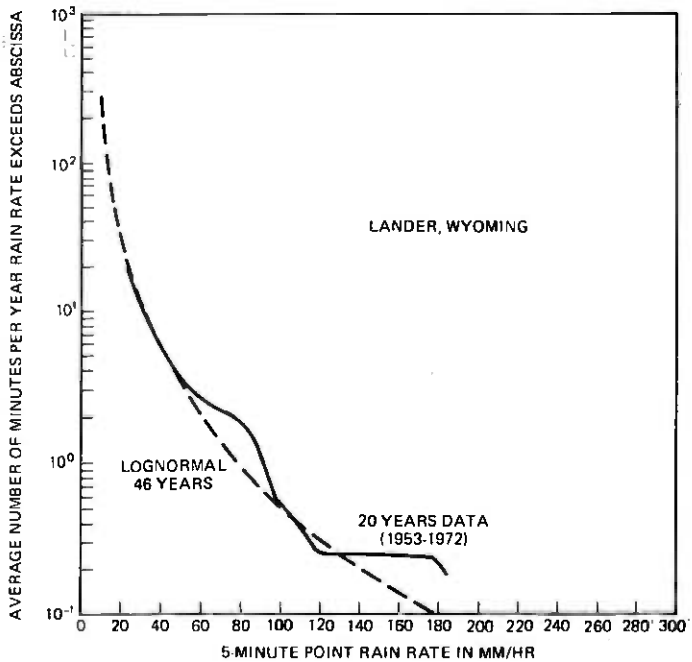


Fig. 19—Comparison of 46-year (1905–1950) distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year (1953–1972) data (solid line) at Lander, Wyoming.

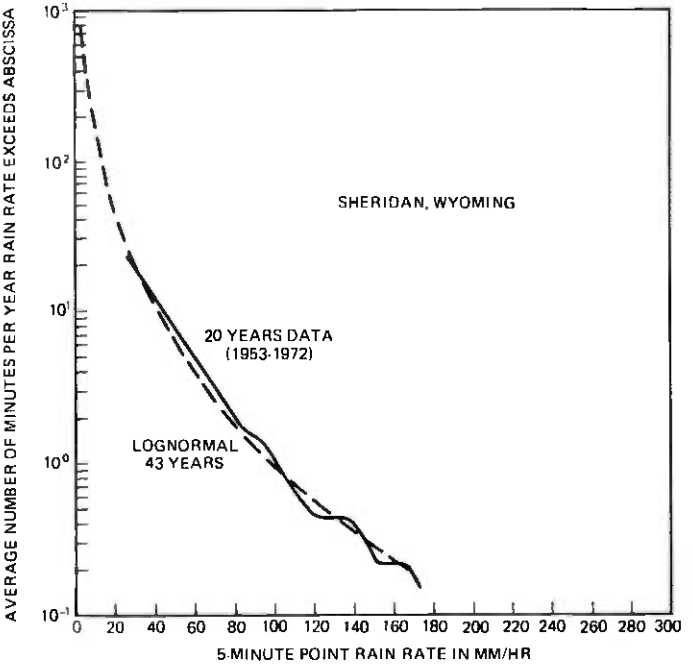


Fig. 20—Comparison of 43-year (1908–1950) distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year (1953–1972) data (solid line) at Sheridan, Wyoming.

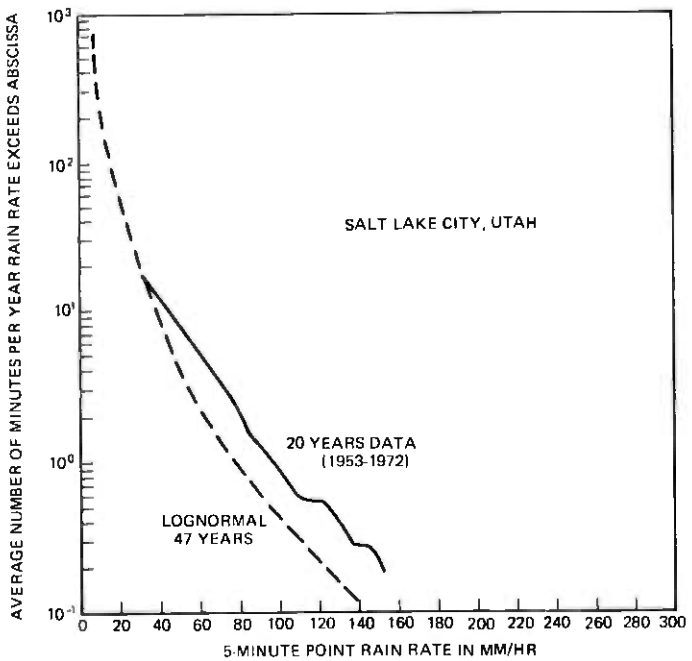


Fig. 21—Comparison of 47-year (1903–1907, 1909–1920, 1922–1951) distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year (1953–1972) data (solid line) at Salt Lake City, Utah.

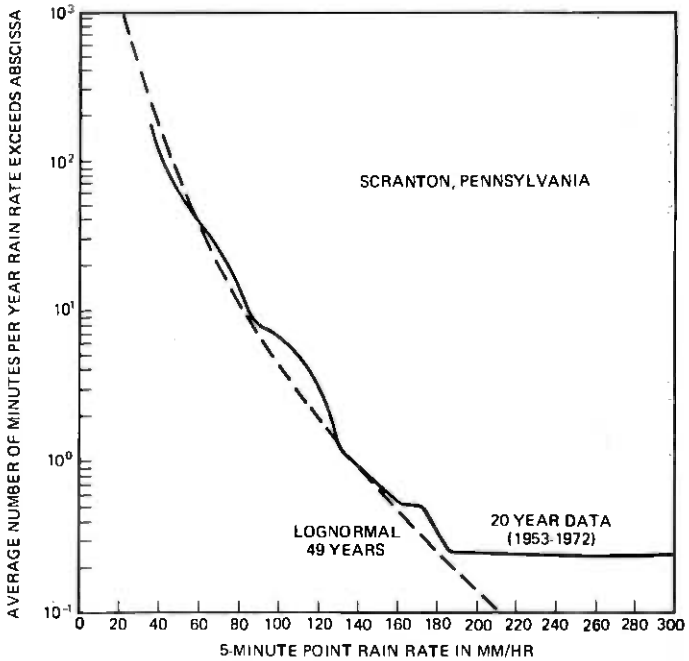


Fig. 22—Comparison of 49-year (1903–1951) distribution of 5-minute rain rate calculated by extreme value theory and lognormal hypothesis (dashed line) with 20-year (1953–1972) data (solid line) at Scranton, Pennsylvania.

IV. PROPORTIONATE EFFECT AND LOGNORMAL RAIN RATE DISTRIBUTION

The rainfall process is influenced by many environmental parameters. An important question is whether the environmental parameters affect the rain rate in a proportional fashion or in an additive fashion. It is well known[5,10,21] that a proportional fashion leads to a lognormal distribution whereas an additive fashion leads to a normal distribution. The following rain rate data will shed some light on this question.

Rain rate data measured in Illinois,[18] New Jersey,[19] and Canada[20] indicate that the short term mean rain rate $\langle R \rangle_s$ and the deviations, ΔR, from the short term mean $\langle R \rangle_s$ appear to be correlated. The subscript s in this section denotes "short term" mean value. These data indicate that the magnitude of the deviations, ΔR, tends to increase with the short term mean $\langle R \rangle_s$. In the following, we present 2 years of rain rate data measured by a tipping bucket rain gauge at Palmetto, Georgia, to confirm this correlation between ΔR and $\langle R \rangle_s$.

Figure 23 displays the time-varying rain rate, $R(t)$, in two 1-hour periods measured by the tipping bucket rain gauge at Palmetto, Georgia. In Fig. 23a, the hourly mean rain rate $\langle R \rangle_s$ is 12.6 mm/hr and the hourly

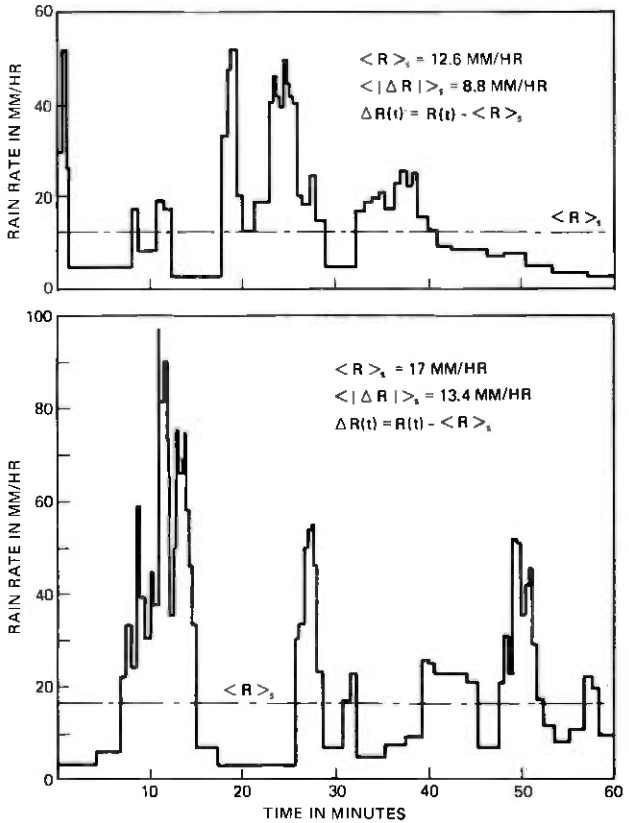


Fig. 23—Time-varying rain rate measured by a tipping bucket rain gauge at Palmetto, Georgia.

mean deviation $\langle|\Delta R|\rangle_s$ is 8.8 mm/hr where

$$|\Delta R(t)| = |R(t) - \langle R\rangle_s| \tag{14}$$

In Fig. 23b, the values of $\langle R\rangle_s$ and $\langle|\Delta R|\rangle_s$ are 17 and 13.4 mm/hr, respectively. Two years of rain rate data at Palmetto have been processed in this fashion and all the hourly $\langle R\rangle_s$ and $\langle|\Delta R|\rangle_s$ pairs are plotted in Fig. 24. It is seen that $\langle R\rangle_s$ and $\langle|\Delta R|\rangle_s$ are indeed correlated and the average relationship is approximately a straight line with a 45 degree slope on the log × log graph paper. To examine this proportional relationship more closely, let

$$X(t) = \ln R(t), \tag{15}$$

$$|\Delta X(t)| = |X(t) - \langle X\rangle_s| \tag{16}$$

The relationship between the hourly mean value $\langle X\rangle_s$ and the hourly

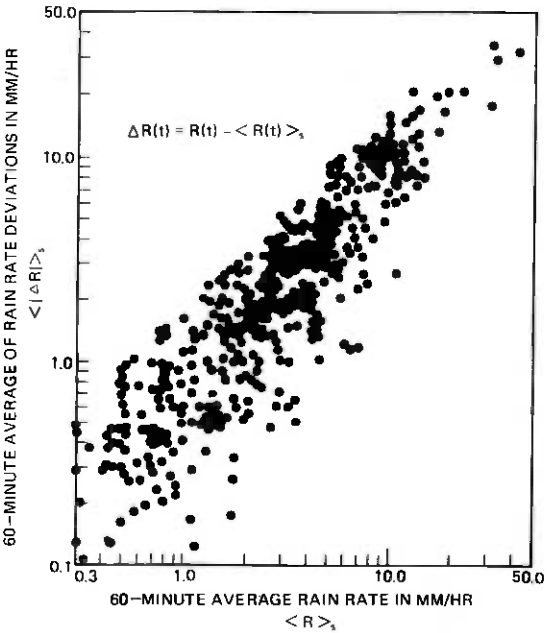


Fig. 24—Two-year data on correlation between hourly mean rain rate, $\langle R \rangle_s$, and hourly mean deviation, $\langle |\Delta R| \rangle_s$, measured by the tipping bucket rain gauge at Palmetto, Georgia.

mean deviation $\langle |\Delta X| \rangle_s$ processed from the same 2-year rain rate data are plotted in Fig. 25. It is seen that $\langle |\Delta X| \rangle_s$ is practically independent of $\langle X \rangle_s$. In Figs. 23 to 25, we use 1-hour period for short-term mean only as an example. The 2-year data were processed by several different "short-term periods" ranging from 5 minutes to 1 hour and showed essentially the same correlation between ΔR and R. Figures 24 and 25 indicate that ΔR is approximately linearly proportional to R:

$$\Delta R = h \cdot R \qquad (17)^*$$

where h is a proportional parameter. The scattering of the data in Fig. 24 and the random variations of ΔR in Fig. 23 indicate that the proportional parameter, h, is not a constant, but is a time-varying random variable. Equation (17) can be interpreted in that the change, ΔR, in the rain rate is proportional to the product of the rain rate R and the intensity of the cause, h. In other words, the environmental parameters affect the rain rate in a proportional fashion. Therefore, the data of Figs. 24 and 25 are another manifestation of the lognormal rainfall process and support the lognormal hypothesis (1). Readers interested in the

* Equations (15) and (17) imply that ΔX is independent of X and is consistent with the data in Fig. 25.

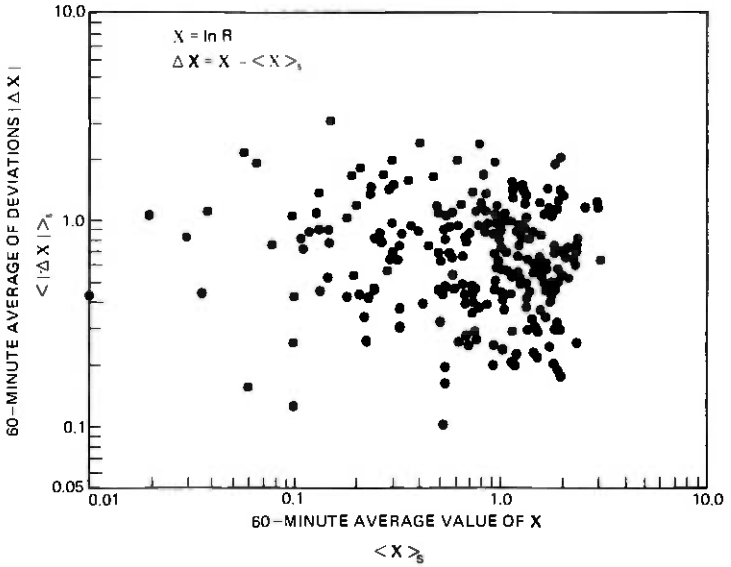


Fig. 25—Two-year data measured by the tipping bucket rain gauge at Palmetto, Georgia, demonstrating that the hourly mean deviation, $\langle |\Delta X| \rangle_s$, and the hourly mean value, $\langle X \rangle_s$ are uncorrelated, where $X = \ln R$.

derivation of the lognormal distribution from the proportionate relationship (17) are referred to Refs. 10 and 21.

V. CONCLUSION

A new method has been described for calculation of 5-minute rain rate distributions from yearly maximum 5-minute rain rate data and yearly total accumulated rainfall data which are available from National Climatic Center[2,3] for U.S. locations. By applying the theory of extreme value statistics and the lognormal hypothesis, the obtained rain rate distribution covers the entire range of rain rates (i.e., from below 5 mm/hr to greater than 200 mm/hr) for wide application. The calculated results agree well with long term (20 to 50 years) data as shown in Figs. 1 to 13 and 17 to 20. The accuracy of the calculated results is limited by the instability of the high rain rate distribution with finite time base.

VI. ACKNOWLEDGMENTS

W. C. Y. Lee,[25] R. A. Semplak,[26] P. L. Rice, and N. R. Holmberg[27] have separately described three different approximate methods for obtaining rain rate distribution from rainfall data published by the National Climatic Center. In an unpublished work, W. Y. S. Chen and R. L. Lahlum applied the theoretical distribution of yearly maximum 5-minute rain rates and an empirical extrapolation to obtain the distribution of high

rain rates. The methods described in Ref. 1 and this paper are inspired from the pioneer work of Lee, Semplak, Rice, Holmberg, Chen, and Lahlum. The work of W. C. Y. Lee provided an impetus for more sophisticated approaches taken by Semplak, Chen, Lahlum, and the author.

APPENDIX A

Formulas for Calculating Extreme Value Parameters α and U from Yearly Maximum Five-Minute Rain Rates

Let $R_1(j), j = 1,2,3, \ldots, M$ be the measured yearly maximum 5-minute rain rate in M years of measurements, and let

$$x_1(j) = \ln [R_1(j)] \tag{18}$$

The formulas for calculating α and U are:

$$\alpha = \frac{\sigma_z}{\sigma_x}, \tag{19}$$

and

$$U = \bar{x}_1 - \frac{\bar{z}}{\alpha} \tag{20}$$

where

$$\bar{x}_1 = \frac{1}{M}\sum_{j=1}^{M} x_1(j) \tag{21}$$

is the sample mean of x_1,

$$\sigma_x = \left\{\frac{1}{M-1}\sum_{j=1}^{M} [x_1(j) - \bar{x}_1]^2\right\}^{1/2} \tag{22}$$

is the sample standard deviation of x_1,

$$z(j) = -\ln\left(-\ln\frac{j}{M+1}\right), \tag{23}$$

$$\bar{z} = \frac{1}{M}\sum_{j=1}^{M} z(j), \tag{24}$$

and

$$\sigma_z = \left\{\frac{1}{M-1}\sum_{j=1}^{M} [z(j) - \bar{z}]^2\right\}^{1/2}. \tag{25}$$

APPENDIX B

Formulas for Calculating α and U from Rainfall Intensity-Duration-Frequency Curves

$$\alpha = \alpha_\infty \cdot \sigma_z \cdot \frac{\sqrt{6}}{\pi}, \tag{26}$$

$$U = U_\infty + \frac{1}{\alpha_\infty}\left[\gamma - \frac{\bar{z}}{\sigma_z} \cdot \frac{\pi}{\sqrt{6}}\right], \tag{27}$$

$$\alpha_\infty = \frac{A_a - A_b}{\ln r_a - \ln r_b}, \tag{28}$$

and

$$U_\infty = \frac{A_a \ln r_b - A_b \ln r_a}{A_a - A_b} \tag{29}$$

where

$$A_a = -\ln\left[\ln \frac{Q_a}{Q_a - 1}\right], \tag{30}$$

$$A_b = -\ln\left[\ln \frac{Q_b}{Q_b - 1}\right], \tag{31}$$

$$Q_a = 2 \text{ (years)}, \tag{32}$$

$$Q_b = 10 \text{ (years)}, \tag{33}$$

$$\gamma = \text{Euler's constant} \approx 0.5772, \tag{34}$$

$\bar{z}$ and σ_z are defined by eqs. (24) and (25).

REFERENCES

1. S. H. Lin, "Rain-Rate Distributions and Extreme Value Statistics," B.S.T.J., *55*, No. 8 (October 1976), pp. 1111–1124.
2. "Climatological Data, National Summary," annual issues since 1950, U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Climatic Center, Federal Building, Asheville, North Carolina 28801. The Excessive Short Duration Rainfall Data prior to 1950 are published in the Monthly Weather Review, the U.S. Meteorological Yearbook (last published for the period 1943 to 1949), and the Report of the Chief of the Weather Bureau (last published for 1931).
3. "Rainfall Intensity Duration Frequency Curves," U.S. Department of Commerce, Weather Bureau, Technical Paper No. 25, Washington, D.C., December, 1955. Available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.
4. S. H. Lin, "A Method for Calculating Rain Attenuation Distributions on Microwave Paths," B.S.T.J., *54*, No. 6 (July–August 1975), pp. 1051–1086, Figs. 10–13.
5. S. H. Lin, "Statistical Behavior of Rain Attenuation," B.S.T.J., *52*, No. 4 (April 1973), pp. 557–581.
6. L. Hansson, "General Characteristics of Rain Intensity Statistics tn the Stockholm Area," Tele, No. 1, Sweden, 1975, pp. 43–48.
7. L. Hansson, "General Characteristics of Rain Intensity Statistics in the Gothenburg Area," Report USR 75 012, Central Administration of Swedish Telecommunications, January 1975, S-12386, Farsta, Sweden.
8. B. N. Harden, D. T. Llewellyn-Jones, and A. M. Zavody, "Investigations of Attenuation by Rainfall at 110 GHz in Southeast England," Proc. IEE (London), *122*, No. 6, June 1975.
9. P. T. Schickedanz, "Theoretical Frequency Distributions for Rainfall Data," International Symposium on Probability and Statistics in the Atmospheric Sciences, June 1–4, 1971, Honolulu, Hawaii. Sponsored by American Meteorological Society and cosponsored by the World Meteorological Organization, 45 Beacon Street, Boston, Massachusetts 02108, U.S.A. Preprints of Symposium Papers, pp. 131–135.

10. J. Aitchison and J. A. C. Brown, *The Lognormal Distribution,* London: Cambridge University Press, 1957, Chap. 3, pp. 20–27.
11. H. M. Conway, Jr., S. L. May, Jr., and E. Armstrong, Jr., "The Weather Handbook," Atlanta: Conway Publications, 1963.
12. E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications,* National Bureau of Standards. Applied Mathematics Series No. 33, February 12, 1954, pp. 15, 16, and 28.
13. E. J. Gumbel, *Statistics of Extremes,* New York: Columbia University Press, 1958.
14. S. H. Lin, "Dependence of Rain Rate Distribution on Rain Gauge Integration Time," B.S.T.J., *55,* No. 1 (January 1976), pp. 135–141.
15. J. W. King, "Sun-Weather Relationships," Astronautics and Aeronautics, *13,* No. 4 (April 1975), pp. 10–19.
16. J. Xanthakis, "Solar Activity and Precipitation," *Solar Activity and Related Interplanetary and Terrestrial Phenomena,* Vol. 1 of the Proceedings of the First European Astronomical Meeting, Athens, September 1, 1972. Printed in Germany by Springer-Verlag, Berlin Heideberg, 1973, pp. 20–47.
17. R. J. Talbot, Jr. and D. M. Butler, "Climatic Effects During Passage of the Solar System Through Interstellar Clouds," Nature, *262,* August 12, 1976, pp. 561–563.
18. E. A. Mueller and A. L. Sims, "The Influence of Sampling Volume on Raindrop Size Spectra," Proc. of the Twelfth Conference on Radar Meteorology, October 1966, p. 135.
19. D. C. Hogg, private communication.
20. G. Drufuca and I. I. Zawadzki, "Statistics of Rain Gauge Records," preprints of the papers at the Inter-Union Commission on Radio Meteorology (I.U.C.R.M.) Colloquium on "The Fine Scale Structure of Precipitation and EM Propagation," Nice, France, October 31, 1973, Vol. 2.
21. A. Hald, *Statistical Theory With Engineering Applications,* New York: John Wiley & Sons, Inc., 1952, Section 8.3, pp. 195–197.
22. S. B. Gershwin, R. V. Laue, and E. Wolman, "Peak-Load Traffic Administration of Rural Multiplexer With Concentration," B.S.T.J., *53,* No. 2 (February 1974), pp. 261–280.
23. G. I. Pozdnyakov, "Fluctuation in the Attenuation of Radio Waves Along a Path in Rain," Telecommun. Radio Eng., *30/31,* N6, 1976, pp. 88–93.
24. R. R. Rogers, "Statistical Rainstorm Models: Theoretical and Physical Foundations," IEEE Trans. Ant. Prop., July 1976, pp. 547–566.
25. W. C. Y. Lee, private communication.
26. P. L. Rice and N. R. Holmberg, "Cumulative Time Statistics of Surface Point Rainfall Rates," IEEE Trans. Comm., *COM-21,* No. 10, October 1973, pp. 1131–1136.
27. R. A. Semplak, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Nonlinear Analysis of a Photovoltaic Optical Telephone Receiver

By D. A. KLEINMAN and D. F. NELSON

(Manuscript received February 2, 1977)

The general problem is considered of calculating the voltage when a sinusoidal current generator is connected to a parallel combination of a linear and a nonlinear resistance load. A practical algorithm is described for computing any desired moment and any Fourier component of voltage. An alternative approximate treatment is also presented which avoids numerical integrations and is valid when the nonlinear characteristic is rapidly varying. Both methods are applied to a photovoltaic optical telephone receiver employing a silicon $n\pi p$-photodiode and a conventional ring armature telephone receiver coupled by a transformer. Harmonic distortion is presented for several illustrative cases. A clipping level is defined for the receiver, and it is proposed that the receiver clipping level should be matched to the clipping level of the analog optical channel bringing the signal. On the basis of this principle a simple procedure is given, along with the necessary curves, for determining the required optical power at the source and the optimum transformer ratio for any value of transmission loss between source and receiver. An illustrative example is given for an analog dynamic range of 18 dB that requires a peak source power in the lightguide of 0.9 mW. This type of receiver may find limited application if lightguides ever serve customers directly.

I. INTRODUCTION

There is now strong indication[1,2] that optical transmission using lightguides[3] and optical cables is technologically approaching a readiness for use in telecommunications. While we are not here suggesting that any extensive application of optical telephones is foreseeable, we have nevertheless found the prospect of limited special applications sufficiently interesting to undertake a study of optical telephone receivers from a device point of view. The receiver is only one of a number of devices, some of which perhaps have not even been invented yet, that would

be required in an optical telephone. In addition, further devices would be required in the electrical-optical interface between the lightguides and the metallic network. The receiver, however, determines one key property of the system, the optical power required at the interface to transmit speech to the human ear with an acceptable volume and quality.

We presuppose that an optical telephone receiver is required to convert analog-modulated light power to sound pressure at the ear with no other power available. There are two mechanisms that might be employed to do this using analog modulation: the optoacoustic effect in which sound is directly produced when power-modulated light is absorbed, and the photovoltaic effect in which an intermediate electrical signal is produced which produces sound by way of an electrical earphone. We have previously completed a theoretical[4,5] and experimental[6] study of optoacoustic receivers in which it was necessary to solve a variety of linear acoustical problems to establish the feasibility of the device and to obtain its response. Nonlinear distortion was not considered and is not believed to be very important for optoacoustic receivers. Optimization in that case is a matter of maximizing the response subject to the requirement of a flat response over the telephone voice band, 300–3300 Hz.

Subsequently we have attempted to define an optimum photovoltaic receiver in a similar way on the basis of maximizing the linear response. It might at first appear that this is a simple problem, because a suitable photodiode and earphone are already available, the required frequency response has already been engineered into the earphone, and one might expect that it is only required to match the impedances of the photodiode and the earphone by a simple two-port network (e.g., a transformer). However, we have now concluded that no optimum of this type exists for the photovoltaic receiver, and a different principle of optimization is required which is based on the nonlinearity of the photodiode and the quality of speech reproduction required in the system. Thus it has not been possible to keep system concepts completely out of the discussion.

We propose here a very simple principle of optimization, which we call "quality matching," which leads to important conclusions about the optical power at the interface, the dynamic range of the system, sound levels in the system, and of course, the receiver design.

We base our discussion on the circuit of Fig. 1. Modulated light power u is conducted by a lightguide to the photodiode, which may be regarded as a current generator g in parallel with a junction current $j(v)$. The photodiode assumed is a specific silicon $n^+\pi p^+$ structure[7] designed and packaged for lightguide use with a light-sensitive area of diameter 80 μm and a quantum efficiency at 900 nm of about $\eta = 0.8$. It is ideal for

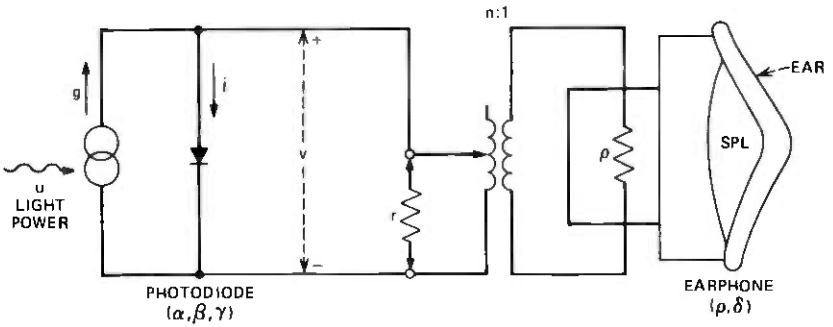


Fig. 1—Circuit of photovoltaic receiver. The earphone resistance ρ is transformed to r.

our purpose because of its low series resistance ($\approx 1\ \Omega$). The earphone is the ring armature telephone receiver[8] which provides good sensitivity with essentially flat response over the band 300–3300 Hz. The photodiode and earphone are coupled by an ideal transformer having a turn ratio n adjustable by virtue of primary taps. The size and cost of the transformer would be approximately proportional to the maximum value of n. The effective load resistance seen by the photodiode is zero at dc and $r = n^2\rho$ at all signal frequencies, ρ being the earphone resistance. The value of n is to be selected at the time of installation in accordance with our optimizing principle.

The first result to be described in this paper is the nonlinear analysis itself, in Section II. Mathematically our problem is the following: Find the periodic voltage response $v(t)$ to a source current $g_1 \cos \omega t$ in Fig. 1 assuming r is linear and independent of frequency (except dc), $j(v)$ is nonlinear and monotonic, and $v(t)$ has zero average value. What makes the problem awkward to treat by textbook methods[9] is the restraint on the average value. The analysis, so far as we know, is not covered in texts on nonlinear circuits, and may be applicable to a variety of situations. In Section III is given an approximate method called the clipping model which we have found quite reliable and especially appropriate for the photovoltaic receiver.

The main body of the paper consists of Sections IV, V, VI, and VII, devoted to the sensitivity, harmonic distortion, clipping level, and quality matching of the receiver, respectively. The sensitivity is an inverse measure of response (the smaller the better!) defined here in the same way as in our previous work as the amplitude of sinusoidal (power) modulation of u required to produce at the ear the average speech power level (81 dB SPL) found in the telephone network surveys. It is interesting to note that the minimum sensitivity is achieved for $r \approx 3 \times 10^5\ \Omega$, which is considerably smaller than the small signal resistance of the photodiode, $\approx 8 \times 10^9\ \Omega$. This shows the essential importance of the nonlinear analysis

presented here since a linear analysis would lead to the equality of these two resistances. The minimum sensitivity is not, in general, the optimum because of nonlinear distortion.

We have made extensive calculations of the second, third and fourth harmonic distortions and the total harmonic distortion. Some curves of total harmonic distortion for a few selected cases are presented in Section V. We have found it difficult, however, to draw concrete conclusions from a consideration of the harmonic distortion that could be used as the basis for an optimizing principle. Rather, we have turned to clipping as the most convenient, relevant, and useful way of specifying the nonlinear distortion.

The clipping model, Section III, is based on the assumption that the distortion of the waveform is an abrupt one-sided clipping of the peak. This assumption is shown to be an adequate modeling of the nonlinear effects of an exponential junction characteristic. It is shown in Section VI that a clipping level can be defined which is analogous (except for being one-sided) to the clipping level of an analog-modulation channel. Our optimizing principle, "quality matching," then follows in Section VII in an obvious way, namely that the clipping levels of the receiver and the analog light channel feeding the receiver should be set equal to each other. The receiver then has the minimum sensitivity consistent with the requirement that the quality of the channel not be degraded by the receiver. This means that the system could operate at the minimum power consistent with a given dynamic range. By using the curves given in Section VII, quality matching can be carried out to determine for any desired dynamic range the optical power required and the correct transformer ratio for the receiver needed for the particular value of the transmission loss at that receiver.

A summary with discussion and conclusions is given in Section VIII. This is followed by two appendixes containing useful background material on speech quality and the electrical-optical interface.

II. NONLINEAR ANALYSIS

In the circuit of Fig. 1 ignore the light, the transformer, and the earphone; assume that the effective load resistance r is constant at all signal frequencies of interest and zero at dc. Let the current generator be

$$\begin{aligned} g &= g_o + g_1 \cos\theta \\ \theta &\equiv \omega t \\ g_o &\geq g_1 \geq 0. \end{aligned} \tag{1}$$

Define averages over time by

$$\langle f(\theta)\rangle_\theta \equiv \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)d\theta \equiv f_o. \tag{2}$$

The equation of the circuit is

$$v - rj_o = rg_1 \cos\theta - rj(v), \tag{3}$$

which implies the relation

$$v_o = 0. \tag{4}$$

Concerning $j(v)$ we only assume

$$j(v)' \equiv dj/dv \geqslant 0, \qquad j(0) = 0 \tag{5}$$

which rules out negative-resistance instabilities. We call $v(\theta)$ the solution and $\theta(v)$ the inverse solution.

Define the parameters

$$v(\pi) \equiv a, \qquad v(0) \equiv b, \tag{6}$$

and

$$q_\pm \equiv \frac{1}{2}[(b \pm a) + r(j(b) \pm j(a))]. \tag{7}$$

We can assume v and θ are restricted to the domains

$$a \leqslant v \leqslant b, \qquad 0 \leqslant \theta \leqslant \pi \tag{8}$$

and the inverse solution $\theta(v)$ is unique. It follows that

$$v_o = a + \frac{1}{\pi}\int_a^b \theta(v)\, dv \tag{9}$$

and

$$j_o = j(a) + \frac{1}{\pi}\int_a^b j(v)'\theta(v)\, dv. \tag{10}$$

From (3) we obtain the inverse solution

$$\theta(v) = \cos^{-1}[(v + rj(v) - q_+)/q_-] \tag{11}$$

and the implicit relations which determine a,b

$$q_+/j_o = r \tag{12}$$

$$(q_-/q_+)j_o = g_1. \tag{13}$$

Our algorithm proceeds as follows: (*i*) choose initial values for a,b; (*ii*) compute $q_\pm$ from (7) and j_o from the integral (10); (*iii*) test (12) and (13); (*iv*) iterate this procedure with adjusted values of a,b until the desired precision is achieved; (*v*) the inverse solution is now given by (11) with the final values of a,b; (*vi*) as a check, evaluate v_o from (9) and test (4). Since (9) and (10) involve numerical integrations, this algorithm requires a large computer.

A function of v such as $j(v)$ can be expanded in a Fourier cosine series

$$j(v(\theta)) = j_o + \sum_{k=1}^{\infty} j_k \cos k\theta \tag{14}$$

$$j_k = 2\,(j(v(\theta)) \cos k\theta)_\theta. \tag{15}$$

Write (15) in the form

$$j_k = \frac{2}{\pi k} \int_a^b j(v)' \sin k\theta \, dv. \tag{16}$$

It follows that the Fourier coefficients of v are

$$v_k = \frac{2}{\pi k} \int_a^b \sin k\theta \, dv. \tag{17}$$

From (3) it follows that v_k is also given by

$$v_k = q_- \delta_{k1} - r j_k \qquad (k = 1,2,\ldots), \tag{18}$$

which is preferable to (17) for numerical work because it automatically becomes exact as $j_k \to 0$. The moments of $v(\theta)$ can be obtained in the same way as (10)

$$\langle v(\theta)^k \rangle_\theta = a^k + \frac{k}{\pi} \int_a^b v^{k-1}\, \theta(v) dv. \tag{19}$$

Define a *normalized waveform*

$$\phi(\theta) = (v(\theta) - q_+)/q_- \tag{20}$$

which has the property of reducing to the input waveform $\cos\theta$ whenever $j(v) = 0$. If $j(v)$ is a rapidly increasing function of v (such as the exponential characteristic of a junction), it may be that the nonlinearity of $j(v)$ can be neglected over part of the cycle while over the rest of the cycle the nonlinearity effectively clamps v at a constant upper limit. This is one-sided abrupt clipping. It is customary to define the *clipping factor* Φ in terms of the ratio of the true analog peak to the clipped peak as follows

$$\Phi = -20 \log \phi_{\max}, \tag{21}$$

where from (20)

$$\phi_{\max} = \phi(0) = (b - q_+)/q_-. \tag{22}$$

The power delivered to the load at frequency $k\omega$ is

$$p_k = v_k^2/2r. \tag{23}$$

A common way of specifying nonlinear distortion is in terms of the ratios

$$d_k = p_k/p_1 \qquad (k = 2,3,\ldots) \tag{24}$$

and

$$d = \sum_{k=2}^{\infty} d_k = (2\langle \nu^2 \rangle_\theta / \nu_1^2) - 1. \tag{25}$$

We define the kth *harmonic distortion* by

$$D_k = 10 \log d_k, \tag{26}$$

and the total harmonic distortion by

$$D = 10 \log d. \tag{27}$$

III. CLIPPING MODEL

Assume that the distortion may be represented as abrupt clipping of the positive peaks. Then

$$\nu(\theta) = w \cos(\theta,\tau) - w\langle \cos(\theta,\tau)\rangle_\theta, \tag{28}$$

where w and τ are parameters to be determined, and

$$(\theta,\tau) = \begin{cases} \tau & 0 \leqslant \theta \leqslant \tau \\ \theta & \tau \leqslant \theta \leqslant \pi \end{cases}. \tag{29}$$

A simple calculation gives

$$\langle \cos(\theta,\tau)\rangle_\theta = (\tau \cos \tau - \sin \tau)/\pi. \tag{30}$$

The clipping factor is

$$\Phi = -20 \log(\cos \tau). \tag{31}$$

The Fourier coefficients are now obtained without numerical integration from the relations

$$\begin{aligned} \nu_1 &= (w/\pi)(\pi - \tau + \frac{1}{2} \sin 2\tau) \\ &= w + \ldots \end{aligned} \tag{32}$$

$$\begin{aligned} \nu_k &= [2w/\pi(k^2 - 1)](\cos k\tau \sin \tau - k^{-1} \cos \tau \sin k\tau) \qquad (k = 2,3,\ldots) \\ &= -(2w/3\pi)\tau^3 + \ldots (k < \tau^{-1}), \end{aligned} \tag{33}$$

and the total distortion from

$$\begin{aligned} \langle \nu^2 \rangle_\theta &= (w^2/2\pi)(\pi + \tau \cos 2\tau - \frac{1}{2} \sin 2\tau) - w^2\langle \cos(\theta,\tau)\rangle_\theta^2 \\ &= (\nu_1^2/2)[1 + (4/15\pi)\tau^5 + \ldots]. \end{aligned} \tag{34}$$

For the determination of w,τ consider (11) in the form

$$\nu(\theta) = q_-\cos\theta + q_+ - rj(\nu). \tag{35}$$

Comparing (28) with (35) shows that

$$q_- = w, \qquad q_+ = -w\langle\cos(\theta,\tau)\rangle_\theta, \tag{36}$$

and that $j(\nu)$ is being approximated by

$$j = \begin{cases} r^{-1}(\cos\theta - \cos\tau)q_- & 0 \leqslant \theta \leqslant \tau \\ 0 & \tau \leqslant \theta \leqslant \pi \end{cases}. \tag{37}$$

The normalized waveform defined in (20) is being approximated by

$$\phi(\theta) = \cos(\theta,\tau). \tag{38}$$

When the model is valid, it is also valid to neglect $j(a)$ in (7); it then follows from (7), (12) and (13) that w,τ must obey the relations

$$w = rg_1 \tag{39}$$

$$j(b)(1 - \cos\tau)^{-1} = g_1, \tag{40}$$

where

$$b = w[\cos\tau + (\sin\tau - \tau\cos\tau)/\pi]. \tag{41}$$

Equation (40) requires that (37) be exact at $\theta = 0$.

The clipping model is suitable for use with a minicomputer since no numerical integrations are required. If $j(b)$ is easy to evaluate, the iteration of (40) is easy to do by trial and error. The accuracy of the model is best determined by comparison with the results calculated by the method of Section II. However, it is possible to obtain a corrected waveform in the region of clipping by solving the implicit relation

$$j(\nu(\theta)) = (\cos\theta - \cos\tau)g_1 \quad (0 \leqslant \theta \leqslant \tau). \tag{42}$$

obtained from (37).

IV. RECEIVER SENSITIVITY

In Fig. 1 let the light power be

$$u = u_o + u_1\cos\theta$$

$$\theta = \omega t, \qquad u_o \geqslant u_1 \geqslant 0. \tag{43}$$

The photodiode assumed here is a specific $n^+\pi p^+$ silicon unit having the dark current characteristic shown in Fig. 2. It is typical of a class of photodetectors developed by H. Melchior[7] for lightguide applications not requiring the sensitivity of avalanche photodiodes.[10] It has a quantum efficiency at 900 nm of $\eta = 0.8$ and a very low series resistance

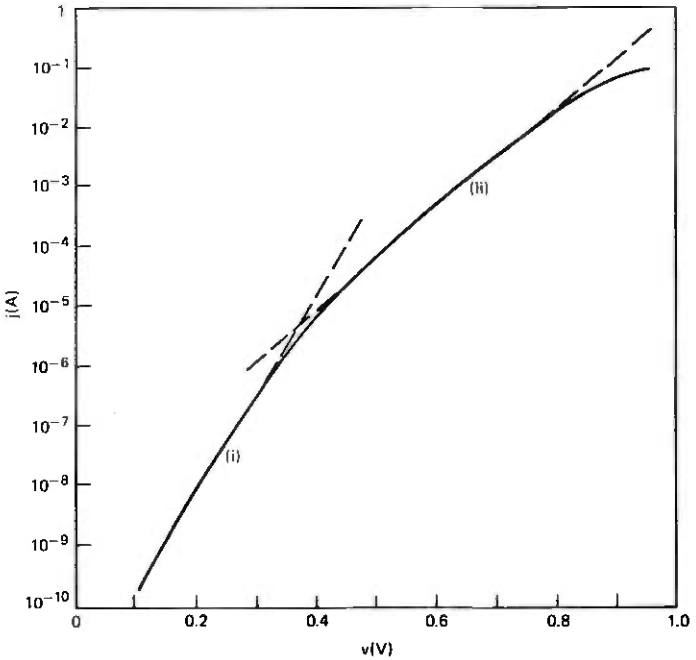


Fig. 2—Dark current-voltage characteristic of the photodiode. Parameters of the characteristic (44) for regions (*i*) and (*ii*) are listed in Table I.

$\approx 1\ \Omega$. The characteristic of Fig. 2 has two exponential regions of the form

$$j(v) = \alpha(e^{\beta v} - 1). \tag{44}$$

The current generator g in Fig. 1 can be written

$$g = \gamma u. \tag{45}$$

The parameters α,β,γ of the photodiode are listed in Table I.

The earphone is the ring armature telephone receiver[8] which provides an essentially flat response over the voice band, 300–3300 Hz. We shall specify this earphone by a resistance ρ and power sensitivity δ, both assumed independent of frequency over the band. The response may be

Table I — Photodiode parameters,
$\eta = 0.80$, $\gamma = e\eta/h\nu = 0.58\ V^{-1}$ (at 900 nm)

	(*i*) $v < 0.37$ V	(*ii*) $v > 0.37$ V
α(A)	3.1×10^{-12}	3.8×10^{-9}
$\beta(V^{-1})$	38.70	19.35
$(\alpha\beta)^{-1}$ (Ω)	8.3×10^{9}	1.36×10^{7}

characterized by the relation

$$\text{SPL} = 81 + 10 \log (p/\delta) \qquad \text{(dB)}, \tag{46}$$

where SPL is the sound pressure level produced in a closed volume of 6 cm^3 and p is the inband power available from a matched generator of resistance ρ. Equation (46) shows that, when p equals the power sensitivity δ, the telephone receiver produces a speech pressure level (SPL) of 81 dB, which is the average level found in the telephone network by surveys.[11,12,13] Representative values for ρ,δ may be taken as

$$\rho = 128\ \Omega$$
$$\delta = 2.5 \times 10^{-7}\ \text{W}. \tag{47}$$

The transformer is assumed to be ideal over the band with primary taps to give a variable turns ratio n. The dc resistance is assumed negligible; this reduces the dc bias on the junction to zero, the most advantageous value. The inductance of the secondary must exceed 0.07 H determined by ρ and the low-frequency cutoff. It follows that the size, weight, and cost of the transformer would be approximately proportional to the maximum required value of n. In the following we specify the receiver by n; the effective ac load resistance of the photodiode is

$$r = n^2\rho. \tag{48}$$

The sensitivity of a receiver of given n will be defined as the value of u_1 which produces SPL = 81 dB at the ear; thus the *sensitivity* s is defined by the relations

$$s = u_1, \qquad p_1 = \delta \tag{49}$$

where p_1 is defined by (23). There is a range of optical powers u_1 (and hence a range of sensitivities s) allowed by (49) because an increasing turns ratio n can be used to compensate (until nonlinearities become significant) for a decreasing optical power u_1. Figure 3 shows s plotted versus n. At the minimum we find by the method of Section II the values

$$n = 48, \qquad r = 0.29\ \text{M}\Omega\ (s = \chi) \tag{50}$$

and

$$\chi \equiv s_{\min} = 2.6\ \mu\text{W}. \tag{51}$$

The clipping model of Section III gives the same asymptotic straight line and the minimum value 2.5 μW.

The asymptotic straight line represents the sensitivity $\hat{s}$ of the *distortion-free receiver* that would result if we could take $j(\nu) \equiv 0$; the

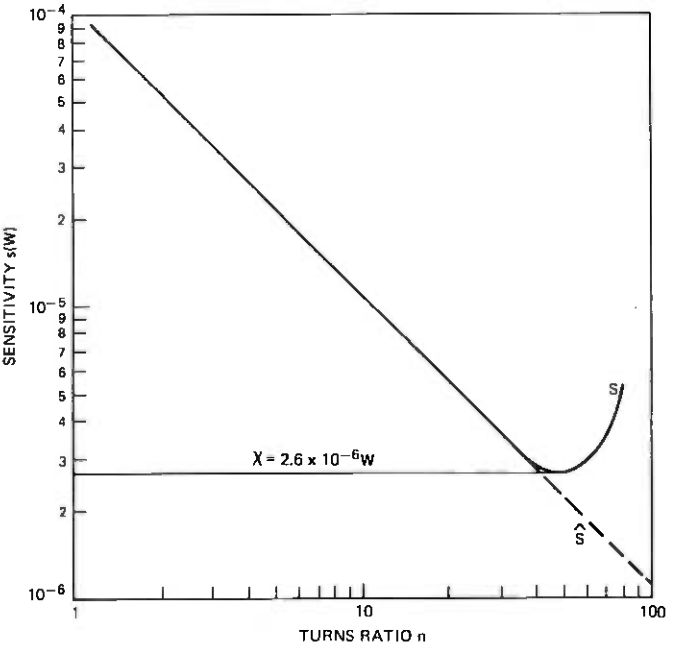


Fig. 3—Sensitivity s defined in (49) versus turns ratio n of transformer. The straight line shows the distortion-free receiver (52).

equation of the line is

$$\delta = \frac{\gamma^2 r \hat{s}^2}{2} = \frac{\gamma^2 n^2 \rho}{2} \hat{s}^2 \qquad \text{(distortion–free receiver).} \qquad (52)$$

V. RECEIVER HARMONIC DISTORTION

We define the *amplitude level* U_1 by

$$U_1 = 20 \log (u_1/\chi). \qquad (53)$$

The need for the factor of 20 in this equation in comparison to the factor of 10 in (46) results from (*i*) the photoelectric effect in the photodiode by which the electrical signal power produced is proportional to the square of the optical power modulation u_1 and (*ii*) the desire to express U_1 on the same logarithmic scale as SPL. In analogy to the present metallic telephone network we assume that the level reaching any receiver can be regarded as a random variable with a distribution[11,12,13] that is approximately normal with a variance of 7.8 dB. We define the *dynamic range* Γ of the optical channel

$$\Gamma \equiv U_o - \langle U_1 \rangle_N, \qquad (54)$$

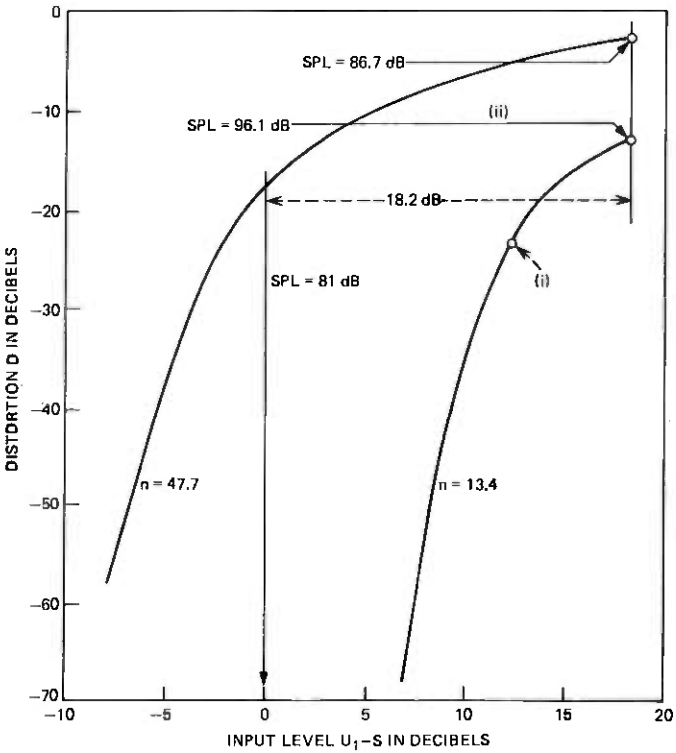


Fig. 4—Total harmonic distortion D defined in (27) versus input signal level U_1 defined in (53) relative to sensitivity level S defined in (56) for two values of n. Various sound pressure levels SPL are indicated; for all n, $U_1 = S$ corresponds to SPL = 81 dB. Points (i) and (ii) are chosen for waveform examination in Fig. 5.

where $\langle U_1 \rangle_N$ is the average over the level distribution N (see Appendix A), and

$$U_o \equiv 20 \log (u_o/\chi) \tag{55}$$

is the *clipping level* of the channel. This will always be a bottom-side clipping level; we will also assume for simplicity that it is a top-side clipping level. We assume Γ is a system constant maintained by the electrical-optical interfaces, whereas U_o and $\langle U_1 \rangle_N$ fall off with the transmission distance x of the particular lightguide. The dynamic range determines the probability that clipping will not occur, which we here call the *quality*. A more general discussion of Γ and speech quality is given in Appendix A.

Harmonic distortion $D_2, D_3, \ldots$ and total harmonic distortion D are defined in (26) and (27) respectively. These quantities are functions of U_1 and are only meaningful up to $U_1 = U_o = \Gamma + \langle U_1 \rangle_N$. We define a

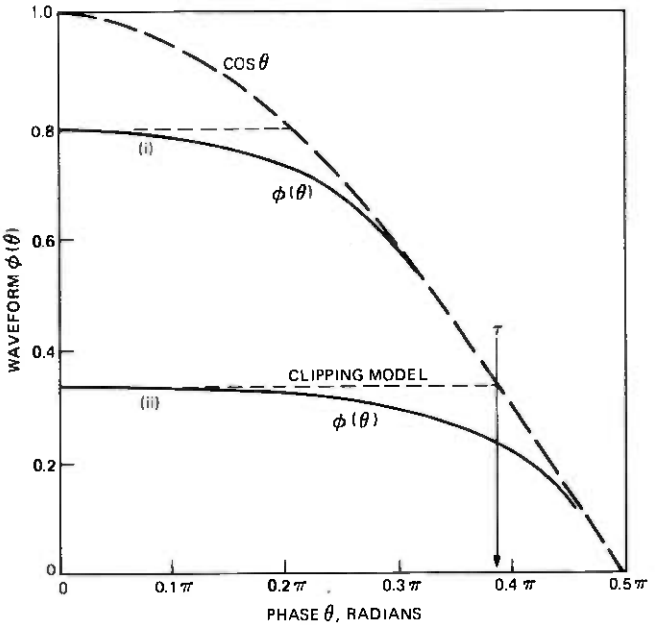


Fig. 5—Normalized waveform (20) for points (*i*) and (*ii*) of Fig. 4 for half of fundamental range of θ. The clipping model approximation is shown dashed and the parameter τ is indicated on (*ii*).

sensitivity level

$$S = 20 \log (s/\chi) \geq 0 \tag{56}$$

with a similar definition for $\hat{S}$ of the distortion-free receiver (52). For our calculations we have chosen to consider the receiver at the *reference point* $\bar{x}$ defined by

$$\langle U_1(\bar{x}) \rangle_N = S, \tag{57}$$

or the distortion-free receiver at the reference point $\bar{x}_{DF}$ defined by

$$\langle U_1(\bar{x}_{DF}) \rangle_N = \hat{S}. \tag{58}$$

Since we will presently conclude that significant distortion in the receiver will be avoided in practice, the distinction between $\bar{x}$ and $\bar{x}_{DF}$ need not concern us further. For illustrative purposes we choose the value $\Gamma = 18.2$ dB. The reasonableness of this value in terms of speech quality is discussed in Appendix A. Figure 4 shows D versus $U_1 - S$ for two receivers $n = 47.7$ and $n = 13.4$ out to $U_1 - S = 18.2$. The steep rise of these curves and the absence of any extensive straight portions show that D is not dominated by the second harmonic but involves a large number of harmonics. At several points the SPL at the fundamental is indicated to show

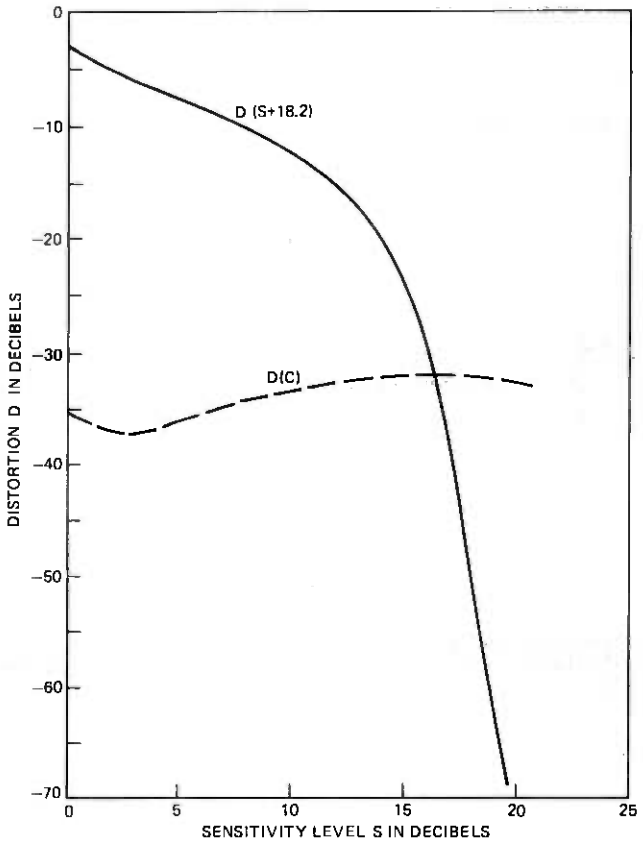


Fig. 6—Total harmonic distortion $D\ (S + 18.2)$ at the channel clipping level for the case $\Gamma = 18.2$ dB versus sensitivity level S defined in (56). Dashed curve shows $D(C)$ at receiver clipping level C defined in (61).

that an 18-dB rise in level does not produce a corresponding rise in SPL when D is high. For a distortion-free receiver, (46) becomes with the help of (52), (53), and (56)

$$\widehat{\text{SPL}} = 81 + U_1 - \hat{S} \qquad \text{(distortion–free receiver).} \qquad (59)$$

The normalized waveform (20) is shown in Fig. 5 for two cases identified as points (i), (ii) in Fig. 4. The distortion of the waveform is shown to be a gradual clipping of the positive peak. In the clipping model this is approximated by abrupt clipping out to $\theta = \tau$ as indicated for curve (ii). We have obtained good results in calculating D_2, D_3 and D by the clipping model when $D > -40$ dB; higher harmonics are given with diminishing accuracy.

Figure 6 shows the total distortion at the channel clipping level

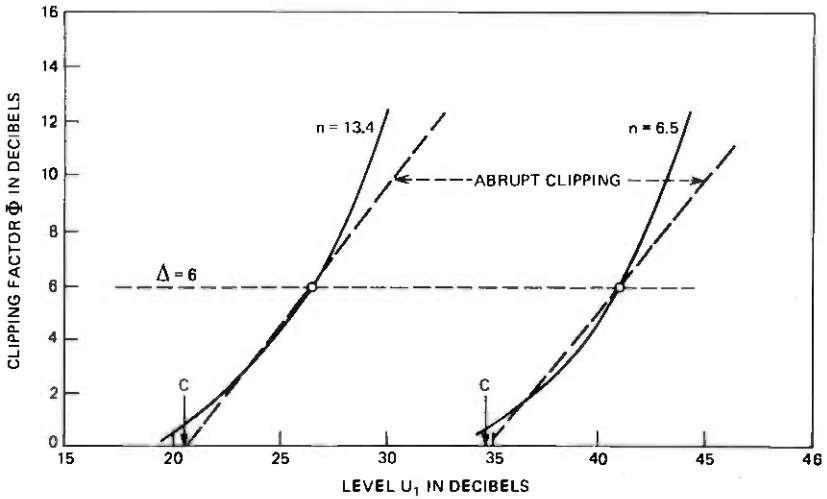


Fig. 7—Clipping factor (21) versus level (53) for two illustrative values of n. Dashed lines represent abrupt clipping approximation (61) which defines the clipping level C of the receiver. The clipping discount Δ (see Appendix A) is chosen as $\Delta = 6$ dB.

$D(S + 18.2)$ versus S for a receiver at the reference point. Conceivably an optimization could be based on an upper-limit objective for D at the channel clipping level. The dashed curve shows $D(C)$ at the receiver clipping level (to be defined in the next section). Notice that $D(C)$ is approximately constant, so in the present instance the quality matching principle is in effect equivalent to requiring $D < -32$ dB at the clipping level.

VI. RECEIVER CLIPPING LEVEL

The clipping factor has been defined in (21) for the general analysis and in (31) for the clipping model. The validity of the clipping model has been confirmed from the waveform and from calculations of D. Using the clipping model, we have calculated $\Phi(U_1)$ for various receivers n. (For brevity, n is not explicitly indicated in writing Φ.) Figure 7 shows the results for $n = 13.4$ and $n = 6.5$. It is known[14] that telephone speech quality, as determined in subjective listening tests, is not degraded appreciably by small clipping, $\Phi < \Delta$, where we call Δ the *clipping discount* and adopt the value

$$\Delta = 6 \text{ dB}. \tag{60}$$

(This value has been deduced by us from an examination of the unpublished work of A. M. Noll.[14]) The discount is shown as a dashed line in Fig. 7. At each intersection of the discount with Φ we draw a line of

unit slope as shown. This represents *abrupt clipping*

$$\Phi(U_1) \to \begin{cases} 0 & U_1 < C \\ U_1 - C & U_1 > C \end{cases} \tag{61}$$

at the *receiver clipping level* C. Thus C (as a function of n) is defined by the relation

$$\Phi(C + \Delta) = \Delta. \tag{62}$$

Notice that C is not very sensitive to the value chosen for Δ; any value of Δ in the range 2 to 8 dB would give about the same $C(\pm 1$ dB).

If the concept of a clipping level is valid, the actual $\Phi(U_1)$ for the receiver can be replaced with the abrupt clipping approximation (61). A test of the validity of the concept is the distortion D at $U_1 = C$; for abrupt clipping D would be zero up to $U_1 = C$. Figure 6 shows $D(C)$ versus S; it is approximately constant around -35 dB. The departure of $\Phi(U_1)$ from abrupt clipping above Δ is not of great significance, because the important question in quality determination is whether degradation occurs, not how much degradation has occurred.

VII. QUALITY MATCHING

The *optimum sensitivity* is the smallest value consistent with the requirement that the quality of the channel not be degraded by the receiver. This gives the principle of *quality matching* expressed by the relation

$$C = U_o \qquad \text{(quality matching)}, \tag{63}$$

that is, the equality of receiver and channel clipping levels. This matching is to hold at all points x in the optical loop system.

Figure 8 shows a schematic diagram of levels in a system versus the distance x from the optical source at the electrical-optical interface. We define the power *transmission loss* of the lightguide in the usual way

$$\text{TL}(x) = 10 \log [u(0)/u(x)] \text{ (dB)}. \tag{64}$$

At $x = 0$ the interface injects the optical power

$$u(0) = h_o + h_1 \cos \theta \tag{65}$$

at the levels

$$\begin{aligned} H_0 &= 20 \log (h_o/\chi) = U_o(0) \\ H_1 &= 20 \log (h_1/\chi) = U_1(0). \end{aligned} \tag{66}$$

It follows that

$$\begin{aligned} U_o(x) &= H_0 - 2\text{TL}(x) \\ U_1(x) &= H_1 - 2\text{TL}(x). \end{aligned} \tag{67}$$

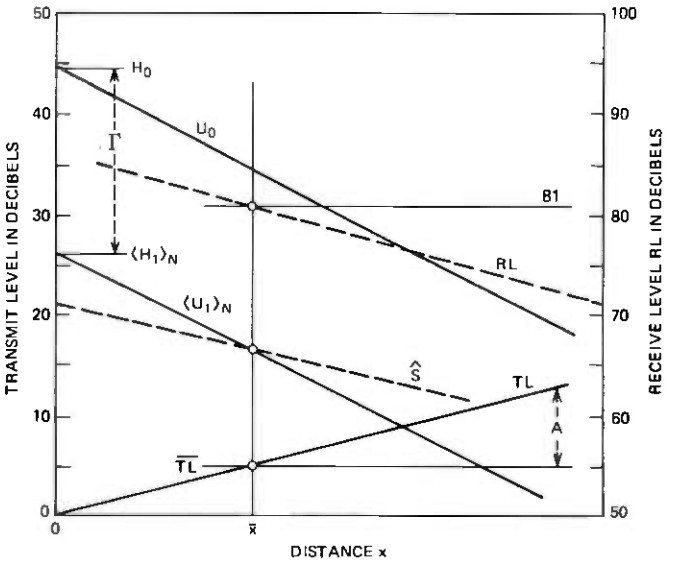


Fig. 8—Levels versus distance in a quality matched lightguide loop system (receiver portion only). The transmit levels (referring to the optical source) are the channel clipping level U_o, mean amplitude level $\langle U_1 \rangle_N$, and transmission loss TL defined in (55), (54), and (64) respectively. Receiver sensitivity level $\hat{S}$ is also shown on the transmit scale. Receive level (68) is also shown. Reference point $\bar{x}$ is defined in (57). The optical source is characterized by H_o; $\langle H_1 \rangle_N$ is defined in (66).

In Fig. 8, TL, U_0, and $\langle U_1 \rangle_N$ are called transmit levels and are referred to the scale on the left. Also shown referred to the left scale is the receiver sensitivity level S. The *receive level* $\mathrm{RL}(x)$ defined by

$$\mathrm{RL}(x) \equiv \langle \mathrm{SPL}(x) \rangle_N \tag{68}$$

is shown referred to the scale on the right. The reference point defined by (57) is denoted by $\bar{x}$. The dynamic range Γ defined by (54) is a constant of the system.

From (54), (57), and (63) we obtain the quality matching relations

$$C(\bar{x}) - S(\bar{x}) = \Gamma \tag{69}$$

$$C(\bar{x}) - C(x) = 2A(x), \tag{70}$$

where

$$A(x) \equiv \mathrm{TL}(x) - \mathrm{TL}(\bar{x}). \tag{71}$$

Figure 9 shows C, $C - S$, and $\hat{S}$ plotted versus n. The implementation of quality matching is illustrated for the case $\Gamma = 18.2$ dB. We find point (i) on the $C - S$ curve according to (69), which determines point (ii) on $\hat{S}$ and (iii) on C. The optical power at the source is then determined by

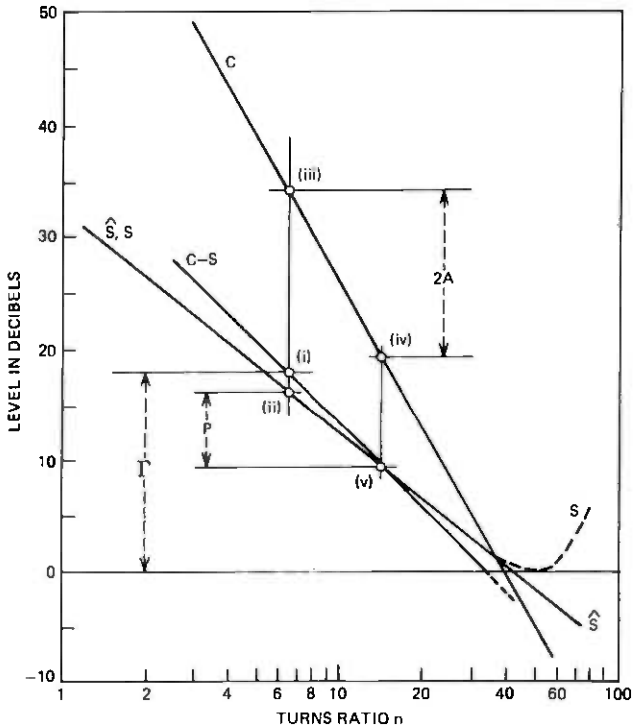


Fig. 9—Receiver clipping level C defined in (61) versus n. Also shown are $C - S$, $\hat{S}$, and S, where S is the sensitivity level (56) and $\hat{S}$ is the sensitivity level for the distortion-free receiver defined by (52). Points (i) to (v) trace the quality matching procedure based on (54), (63), (69), (70), and (74).

(iii) through the relation

$$H_o = C(\bar{x}) + 2\,\overline{\mathrm{TL}} \tag{72}$$

in terms of $\overline{\mathrm{TL}} \equiv \mathrm{TL}(\bar{x})$. The choice of $\overline{\mathrm{TL}}$ determines the average RL of the system, which is a system objective (not necessarily 81 dB) we must leave open at this time. We find point (iv) on C according to (70), which determines the optimum value of n for a receiver at x. It also determines point (v) on $\hat{S}$. (The intersection with S, when it differs from $\hat{S}$, is not of interest.) From (22) and (v) we determine the quantity

$$P(x) \equiv \hat{S}(\bar{x}) - \hat{S}(x). \tag{73}$$

From (59), (67), and (68) the receive level for a distortion-free receiver is

$$\widehat{\mathrm{RL}}(x) = 81 + P(x) - 2A(x). \tag{74}$$

This is an excellent approximation for RL, because quality matching guarantees that distortion is too small to have any effect on response.

In Fig. 9 the lines C and $\hat{S}$ are given by the relations

$$C = 70.5 - 44.3 \log n \tag{75}$$

$$\hat{S} = 32.6 - 20 \log n. \tag{76}$$

If $\Gamma > 5$ dB, it is justified to approximate

$$C - S = 37.9 - 24.3 \log n. \tag{77}$$

It follows that

$$C(\bar{x}) = 1.4 + 1.82\,\Gamma \tag{78}$$

$$P(x) = 0.90\,A(x) \tag{79}$$

$$\log n = 1.56 - 0.041\,\Gamma + 0.045\,A. \tag{80}$$

Thus the receive level is

$$\mathrm{RL} \approx \widehat{\mathrm{RL}} = 81 - 1.1\,A. \tag{81}$$

and the clipping level at the optical source is

$$H_o = 1.4 + 1.82\,\Gamma + 2\overline{\mathrm{TL}}. \tag{82}$$

VIII. SUMMARY AND DISCUSSION

We have obtained a practical algorithm suitable for a large computer for solving the nonlinear integral equation (3) referring to Fig. 1. The equation implies an integral restraint (4) on the solution which is an unusual feature that removes it from the types found discussed in texts on nonlinear networks. The load r is taken as zero at dc and a constant resistance at all signal frequencies of interest. This implies that the load is dispersive at frequencies below a certain cutoff frequency (300 Hz in the receiver problem). Ordinarily a nonlinear dispersive circuit requires nonlinear differential equations to describe it. Here we have avoided the differential equation and obtained instead an integral equation (3) by: (*i*) asking only for the periodic solution, and (*ii*) treating separately the ac and dc voltages with different values of r. The assumption of zero dc resistance is convenient and usually appropriate in practice, but the analysis presented in Section II can easily be generalized to any value of dc resistance. The nonlinear conductor $j(\nu)$ is passive ($j(0) = 0$) and monotonically increasing ($j(\nu)' \geq 0$) but otherwise arbitrary. The conditions on $j(\nu)$ rule out negative-resistance instabilities and guarantee a unique solution.

The method presented in Section II is exact in principle; it is based on the fact that any functional of the solution (e.g., a Fourier coefficient) can be calculated explicitly by integration once two parameters (a,b) have been determined from the implicit relations (12), (13). Standard

routines are available for solving simultaneous implicit relations to any desired precision. In using this method we have usually obtained satisfactory convergence with no special precautions. When convergence problems are encountered, the answer is to start the iteration with better estimates for a,b.

The clipping model described in Section III was originally worked out to provide initial estimates of a,b for use in the exact method. It soon became apparent, however, that it is sufficiently accurate in the receiver problem for all calculations of sensitivity, total distortion (when $D > -40$ dB), and clipping factor. The reason for this is that the exact waveform comes close to the clipped waveform assumed in the model. The clipping model is not recommended for the calculation of specific harmonics higher than the second. Generally the clipping model is expected to be useful whenever $j(\nu)$ is a rapidly increasing function. In this model distortion (clipping) is represented by a single parameter τ which is determined from the implicit relation (40). No numerical integrations are involved in using the model, which makes it convenient for use with a minicomputer.

The analysis of the photovoltaic receiver in the remainder of the paper is based on a sinusoidal input waveform. The calculation of the sensitivity s, defined in (49) as a measure of response based upon producing a certain reference sound level at the fundamental of 81 dB, is presented in Section IV. The photodiode and the earphone assumed in this calculation are the best presently available for the purpose. The transformer secondary must have an inductance of at least 70 mH, so the turns ratio n is adjusted by means of taps on the primary. By adjusting n, any value of s down to the minimum 2.6 μW can be obtained. However, the size and cost of the transformer are expected to increase approximately as the maximum value of n.

The minimum $s = \chi$ shown in Fig. 3 is a nonlinear effect having nothing to do with the impedance matching concept of linear circuit theory. The small signal resistance of the junction is given in Table I, $(\alpha\beta)^{-1} = 8.3 \times 10^9\ \Omega$ which is larger than $r(\chi) = 2.9 \times 10^5\ \Omega$ by a factor $\approx 3 \times 10^4$! This shows the necessity for a nonlinear analysis. For $n > 48$ the distortion increases very rapidly at the reference level assumed for the calculation of s. This does not mean, however, that $n > 48$ is ruled out for receivers operating at much lower levels. In a properly designed system the receiver must respond almost like the distortion-free receiver defined in (52) at the levels to which it is subjected. The reference level SPL = 81 dB is the overall average level for receivers in the existing telephone system.

At a particular point, the reference point, in a loop system the receive level defined in (68) will equal 81 dB, and at other points its value will depend on the transmission loss from the reference point to that point.

For a discussion of total harmonic distortion D we have considered a receiver at the reference point in Section V. We find (Fig. 4) that D rises very rapidly with level up to about $D \approx -20$ dB and then begins to bend over. The value of D of interest is the value at the channel clipping level shown versus sensitivity in Fig. 6 for an assumed dynamic range of 18.2 dB. The results confirm that D is much too high in the minimum sensitivity receiver for use at the reference point. It is conceivable that some criterion on D (e.g., $D < -20$ dB) could be used as the basis for optimizing the receiver (choosing n). However, we believe that clipping provides a more objective and less arbitrary basis for optimization. The existence of clipping is shown by the waveform of Fig. 5 and by the good agreement between the clipping model and the exact method of D.

The clipping factor defined in (21) has been calculated as a function of level in Section VI. At a certain value Δ, called here the clipping discount, abrupt clipping begins to degrade speech quality.[14] Therefore an abrupt clipping approximation has been fitted to the receiver clipping factor at the value Δ. This defines the clipping level C as illustrated in Fig. 7 for the choice $\Delta = 6$ dB. Actually C is not very sensitive to Δ. The total distortion at level C would be zero for abrupt clipping. We find $D(C) \approx -35$ dB for all receivers (Fig. 6); in our opinion this is small enough to confirm the validity of the clipping level concept.

Quality matching as an optimizing principle is defined in Section VII. It amounts to setting the channel and receiver clipping levels equal as in (63). The quality of speech transmitted in an analog channel in which the only nonlinearity is abrupt clipping is determined by the dynamic range, defined for sinusoidal signals in (54). The dynamic range Γ is independent of lightguide transmission loss so we consider it a constant of the optical loop system. In Fig. 8 we pass from a strictly device viewpoint to a system viewpoint. The quality matching relations (69) determine the optical power needed at the reference point as well as the optimized receiver and receive level throughout the optical loop system. The various levels are shown schematically in Fig. 8 as a function of distance x from the optical source assuming the same source power for all loops. The curves required to obtain the solution are shown in Fig. 9, and approximate equations for the solution are given in (78) through (82).

In the existing system the mean loop insertion loss[15] is about 5 dB. The receive level defined in (68) is 81 dB at a point 5 dB from the central office, and 81 dB is approximately the mean loop receive level. This is not necessarily an objective for future loop planning, so we here emphasize that our analysis contains no such assumption. The use of 81 dB as a reference is only a convenience and involves no loss of generality. The reference point $\bar{x}$ in Fig. 8 with the transmission loss $\overline{TL}$ need not be the "average point" for the loop system. The mean loop receive level

is from (81)

$$\langle \mathrm{RL} \rangle_L = 81 + 1.1\,(\overline{\mathrm{TL}} - \langle \mathrm{TL} \rangle_L) \tag{83}$$

where $\langle\ \rangle_L$ denotes an average over loops. This can be set at any desired level by properly choosing $\overline{\mathrm{TL}}$.

The quality matching concept requires a receiver with adjustable turns ratio which probably involves some added cost compared with a fixed receiver. However, any fixed receiver would give a receive level varying as $-2A$,

$$\mathrm{RL} = 81 - 2A \qquad \text{(fixed receiver)}, \tag{84}$$

whereas quality matching gives (81) varying approximately as $-A$. This effect is shown in Fig. 8 by the different slopes of RL and $\langle U_1 \rangle_N$. Clearly this is a desirable effect which utilizes the capabilities of the receiver to the fullest extent and permits serving a radius approximately twice that which would be possible with a fixed-sensitivity receiver having the sensitivity of an optimized receiver at the reference point. The actual range of the loop system would probably be limited by objectives on the minimum RL which we are not discussing here. Another limitation which we can only mention is that of transformer cost. From (80) we find that the cost, as measured by n, doubles for an increase of 6.7 dB in TL.

To illustrate the principles being discussed, we have chosen a realistic case specified by

$$\begin{aligned} \Gamma &= 18.2\ \mathrm{dB} \\ \overline{\mathrm{TL}} &= 5\ \mathrm{dB} \qquad \text{(illustrative)} \\ A(x) &= 7.5\ \mathrm{dB}. \end{aligned} \tag{85}$$

Points (i) to (v) in Fig. 9 trace the solution for this case. At the reference point $\bar{x}$ we find

$$\begin{aligned} C(\bar{x}) &= U_o(\bar{x}) = 34.6\ \mathrm{dB} \\ u_o(\bar{x}) &= 0.14\ \mathrm{mW} \\ s(\bar{x}) &= 17\ \mu\mathrm{W}. \end{aligned} \tag{86}$$

At the point x we find

$$\begin{aligned} C(x) &= 19.6\ \mathrm{dB} \\ n(x) &= 14.1 \\ \mathrm{RL} &= 72.8\ \mathrm{dB}. \end{aligned} \tag{87}$$

At the source we find

$$H_o = 44.6 \text{ dB}$$
$$\langle H_1 \rangle_N = 26.4 \text{ dB}$$
$$h_o = 0.45 \text{ mW}$$
$$\text{peak source power} = 2\, h_o = 0.9 \text{ mW}. \tag{88}$$

The fact that the peak source power is $2\, h_o$ follows from our assumptions of equality of top and bottom side clipping and of occurrence of bottom side clipping at $h = 0$.

The sensitivity at the reference level $s(\bar{x}) = 17\ \mu$W is 50 times smaller (better) than that of the best optoacoustic receiver, the xenon photophone[4] ($s = 0.9$ mW). Furthermore, the photophone is a fixed receiver. Therefore the photovoltaic receiver is clearly superior for loop applications.

The "system" referred to here should be regarded as a relatively small subsystem of the loop plant serving a special class of customers who have lightguides running to their premises primarily to provide high capacity services. At the central office or other junction point there must be an electrical-optical interface containing the optical source for the receiver. The peak power of this source according to (88) should be 0.9 mW. This power into the lightguide is within the capabilities of present day heterojunction laser diodes[16] but about an order of magnitude above the capabilities of luminescent diodes.[16] To a limited extent the power at the source might be varied to compensate for the transmission loss; one could imagine all adjustment being done at the source instead of the receiver. Although we have described a system with fixed source power in Fig. 8 and believe that this is the most likely type of system, no change is required in the equations to treat the variable source. The dynamic range, however, would be determined by the interface circuitry and held to a uniform value to maintain transmission quality.

It may be objected that our nonlinear analysis has been based on a sinusoidal input signal whereas a telephone is required to transmit speech. Appendix A contains a discussion of the extension of the theory to a general waveform and the general definition of dynamic range. It is argued that the clipping level is valid for any waveform. A quality Q is objectively defined on the basis of the approximately normal distribution of speech levels in the telephone system. Figure 10 shows Q versus Γ for speech and for a hypothetical system having sinusoidal signals. Essentially, Q is a probability that speech is received without degradation caused by clipping. The illustrative case $\Gamma = 18.2$ dB used in this paper corresponds to $Q = 86$ percent and $Q_{\cos} = 99$ percent. An objectively defined quality is not equivalent to a grade of service determined sub-

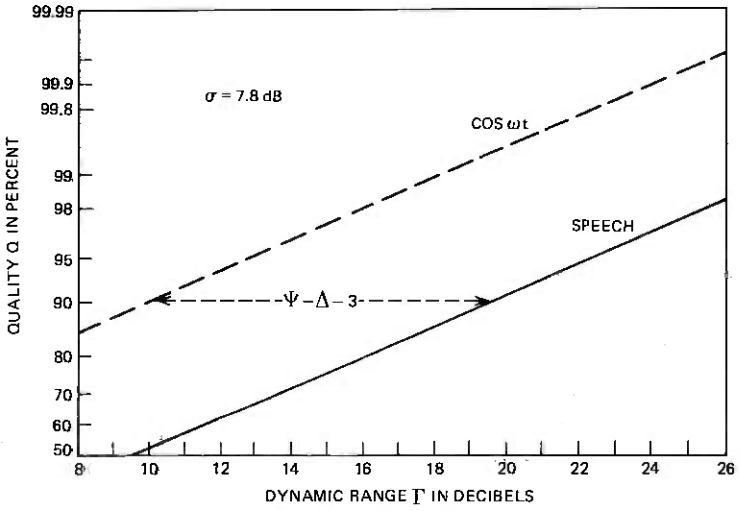


Fig. 10—Objective speech quality Q defined in (104) versus dynamic range Γ defined in (93) plotted on probability scale. The corresponding quantity for sinusoidal waveform is shown dashed.

jectively from listening tests, although we would expect a strong correlation between the two.

We shall not speculate on the circuitry of the interface, but we offer an interesting observation in Appendix B. On the basis of the light-emitting characteristic of laser diodes and the signals from the metallic network existing at the central office, it is shown that the direct application of the metallic network signal current (no amplifiers or transformers) to the laser would produce a somewhat greater level than $\langle H_1 \rangle_N$ called for in (88).

Let us suppose that $\overline{\text{TL}} = 5$ dB in the illustrative case (85) is the average loss and the lightguide attenuation constant[17] is 2 dB/km. Then the reference point, which is also the average point, is at the distance

$$\bar{x} = 2.5 \text{ km} = 8.3 \text{ kft (illustrative)}. \tag{89}$$

This distance may be compared to the average loop length[15] 10.3 kft in the present metallic loop plant. A more conservative estimate of attenuation, like 6 dB/km, reduces (89) by a factor of three. Thus, the optical loop system seems to be limited to a somewhat smaller radius of service than the metallic loop system.

IX. ACKNOWLEDGMENTS

We acknowledge with thanks a crucial suggestion from D. A. Berkley on an earlier version of this work which caused us to change from the circuit we had first analyzed to the circuit of Fig. 1. We also thank H.

Melchior for valuable discussions and the use of unpublished data. We also thank W. M. Hubbard, D. Gloge, A. M. Noll, R. W. Dixon, D. Schinke, and S. D. Personick for important information and suggestions.

APPENDIX A

Speech Quality and Dynamic Range

We continue to let θ denote a normalized time variable, but in (43) and (65) we replace $\cos\theta$ with a more general normalized waveform $\psi(\theta)$

$$\max \psi(\theta) = -\min \psi(\theta) = 1$$
$$\langle \psi(\theta) \rangle_\theta = 0. \tag{90}$$

Define the *peak factor*

$$\Psi = -20 \log \psi_{rms} \tag{91}$$

and *signal level*

$$U = 20 \log (u_1 \psi_{rms}/\chi)$$
$$= U_1 - \Psi \tag{92}$$

where U_1 is given by (53). Define the *dynamic range*

$$\Gamma \equiv \langle 20 \log (u_o/2^{1/2} u_1 \psi_{rms}) \rangle_N$$
$$= U_o - \langle U \rangle_N - 3 \tag{93}$$

where the *channel clipping level* U_o is given by (55) and where

$$\langle U \rangle_N \equiv \int_0^1 U(N) dN, \tag{94}$$

N being a distribution function. For $\psi(\theta) = \cos\theta$ we have $\Psi = 3$ dB and (93) reduces to the definition first given in (54). For the distortion-free receiver the $\widehat{\text{SPL}}$ is

$$\widehat{\text{SPL}} = 81 + U - S + 3, \tag{95}$$

which reduces to (59) when $\Psi = 3$. We assume (61) and (63) remain valid but write (61) in terms of U

$$\Phi(U) = \begin{cases} 0 & U < U_o - \Psi \\ U + \Psi - U_o & U > U_o - \Psi. \end{cases} \tag{96}$$

The levels U are distributed[11,12,13] such that the probability that $U < X$ is

$$P(U < X) = N[(X - X_o)/\sigma] \tag{97}$$

where

$$X_0 = \langle U \rangle_N, \qquad \sigma = 7.8 \text{ dB} \tag{98}$$

and $N(x)$ is the normal distribution

$$N(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-t^2/2}\, dt. \tag{99}$$

It is known from subjective studies[14] of the effect of clipping on telephone speech quality that no degradation occurs for $\Phi < \Delta$, where we take

$$\Delta = 6 \text{ dB}. \tag{100}$$

(This is a deduction from the work of A. M. Noll for which we assume responsibility.) The probability that $\Phi < \Delta$ is

$$\begin{aligned} P(\Phi < \Delta) &= N[(U_o + \Delta - \Psi - \langle U \rangle_N)/\sigma] \\ &= N[(\Gamma + \Delta + 3 - \Psi)/\sigma]. \end{aligned} \tag{101}$$

The speech waveform is characterized by the value[18]

$$\Psi = 18.6 \text{ dB} \qquad \text{(speech)}, \tag{102}$$

so that

$$\Delta + 3 - \Psi = -9.6 \text{ dB} \qquad \text{(speech)}. \tag{103}$$

The objective definition of quality is

$$\begin{aligned} Q &= 100 \times P(\Phi < \Delta) \qquad \text{(speech)} \\ &= 100 \times N[(\Gamma - 9.6)/7.8]. \end{aligned} \tag{104}$$

A corresponding definition can be given for sinusoidal signals having the same level distribution assuming $\Delta = 0$ and $\Psi = 3$

$$Q_{\cos} = 100 \times N(\Gamma/7.8). \tag{105}$$

Figure 10 shows Q and $Q_{\cos}$ versus Γ on a probability scale.

APPENDIX B

Comments on the Optical Source

A laser diode has a threshold current for lasing and at higher currents a steeply rising emission h as a function of current i. In the lasing region we assume[19]

$$\begin{aligned} dh/di &= \epsilon \\ \epsilon &= 2.3\, W/A. \end{aligned} \tag{106}$$

If a sinusoidal current of amplitude i_1 flows through the laser (superposed on a suitable bias current), the optical signal level H_1 defined in

(66) is

$$H_1 = 20 \log (\epsilon i_1/\chi). \tag{107}$$

Suppose that the current i_1 is that supplied to a matched load by a generator of resistance r and available power p; then (107) can be written

$$H_1 = 20 \log [2^{1/2}(\epsilon/\chi)(p/r)^{1/2}]. \tag{108}$$

Finally, suppose that the generator is the central office and p corresponds to the mean signal level; representative values are[20]

$$p = 2\,\mu\text{W}, \qquad r = 1166\ \Omega. \tag{109}$$

It then follows that the mean optical amplitude level at the source is

$$\langle H_1 \rangle_N = 34\ \text{dB} \qquad (h_1 = 0.13\ \text{mW}). \tag{110}$$

This compares very favorably with the level called for, 26 dB, in (88). This shows that the current flowing in a matching resistor (1166 Ω) combined with the "gain" of the laser emission characteristic provides more than enough signal without a matching transformer in the circuit.

If a matching transformer is used, r in (108) is replaced by the dynamic laser diode resistance[19,21] $r_d \approx 1.5\ \Omega$, giving $h_1 \approx 3.8$ mW. For a luminescent diode an appropriate value of ϵ in (106) is[16] $\epsilon \approx 5 \times 10^{-4}$ W/A giving for the transformer matched case with $r_d = 1.5\ \Omega$ the power $h_1 \approx 0.8\,\mu$W, $\langle H_1 \rangle_N = -10$ dB.

REFERENCES

1. I. Jacobs, "Lightwave Communications Passes Its First Test," Bell Laboratories Record, *54*, No. 11 (December 1976), pp. 290–297.
2. A. R. Meier, "Bell Labs Unveils a Practical Optical Communications System," Telephony, August 9, 1976, pp. 48–54.
3. S. E. Miller, E. A. J. Marcatilli, and T. Li, "Research Toward Fiber Transmission Systems," Proc. IEEE, *61*, No. 12 (December 1973), pp. 1703–1751.
4. D. A. Kleinman and D. F. Nelson, "The Photophone—An Optical Telephone Receiver," J. Acoust. Soc. Amer., *59*, No. 6 (June 1976), pp. 1482–1494.
5. D. A. Kleinman and D. F. Nelson, "The Photophone—Physical Design," J. Acoust. Soc. Amer., *60*, No. 1 (July 1976), pp. 240–250.
6. D. F. Nelson, K. W. Wecht, and D. A. Kleinman, "Photophone Performance," J. Acoust. Soc. Amer., *60*, No. 1 (July 1976), pp. 251–255.
7. H. Melchior, private communication.
8. E. E. Mott and R. C. Miner, "The Ring Armature Telephone Receiver," B.S.T.J., *30*, No. 1 (January 1951), pp. 110–140.
9. Thomas A. Stern, *The Theory of Nonlinear Networks and Systems*, Reading, Mass.: Addison-Wesley, 1965, pp. 95–136.
10. H. Melchior, "Photodetectors for Optical Communications Systems," Proc. IEEE, *58*, No. 10 (October 1970), pp. 1466–1486.
11. K. L. McAdoo, "Speech Volumes on Bell System Message Circuits," B.S.T.J., *42*, No. 5 (September 1963), pp. 1999–2012.
12. F. T. Andrews and R. W. Hatch, "National Telephone Network Transmission Planning in the American Telephone and Telegraph Company," IEEE Trans. on Commun. Tech., *COM-19*, No. 3 (June 1971), pp. 302–314.

13. Staff of Bell Laboratories, *Transmission Systems forCommunications*, Winston-Salem, N.C.: Western Electric Company Technical Publications, 1970, 4th ed., Fig. 4–5, p. 74.
14. A. M. Noll, private communication.
15. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J. *48*, No. 10 (December 1969), pp. 3337–3385.
16. H. Kressel, I. Ladany, M. Ettenberg, and H. Lockwood, "Light Sources," Physics Today, *29*, No. 5 (May 1976), pp. 38–47.
17. A. G. Chynoweth, "The Fiber Lightguide," Physics Today, *29*, No. 5 (May 1976), pp. 28–37.
18. Staff of Bell Laboratories, *Transmission Systems for Communications*, Winston-Salem, N.C.: Western Electric Company Technical Publications, 1970, 4th ed., p. 222.
19. R. W. Dixon, private communication.
20. Staff of Bell Laboratories, *Transmission Systems for Communications*, Winston-Salem, N.C.: Western Electric Company Technical Publications, 1970, 4th ed., Fig. 4–12, p. 84, and Fig. 4–5, p. 74.
21. R. L. Hartman and R. W. Dixon, "Reliability of DH GaAs Lasers at Elevated Temperatures," Appl. Phys. Lett., *26*, No. 5 (March 1975), pp. 239–242.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Perceptual and Objective Evaluation of Speech Processed by Adaptive Differential PCM

By B. McDERMOTT, C. SCAGLIOLA, and D. GOODMAN

(Manuscript received November 9, 1977)

An experiment has been performed to study the perceptual characteristics of speech processed by adaptive differential PCM. We created 18 three-bit and four-bit coders spanning a wide range of quantizer adaptation parameters. Subjects judged differences between coders and rated the quality of each coder individually. The difference data reveal three important perceptual characteristics: overall clarity, signal vs. background degradation, and rough vs. smooth impairment. These characteristics are strongly correlated with coder design parameters and objective performance measures. Overall subjective quality is well predicted by segmental signal-to-noise ratio and even better by a linear combination of measures of granular distortion and overload distortion.

I. INTRODUCTION

Speech signal processing systems are susceptible to a variety of audible impairments often classified with words like "distortion," "noise," "echo," and "sidetone." These categories are themselves subdivided: for example, "linear" and "nonlinear" distortion, "white" noise, "impulsive" noise, "speech-dependent" noise, etc. When the type of system is familiar to a large body of listeners, the application of these names becomes standardized and a language exists for describing the quality of specific implementations. With new systems, however, the types of degradation are often not known *a priori*, and special effort is required to identify them and to relate them to physical characteristics of the system.

For example, experiments on PCM (pulse code modulation) have identified peak clipping, granular quantizing noise, and bandlimiting as important audible degradations.[1,2] In PCM, there are relatively few design parameters, and each of these impairments can be related to one of them: peak clipping to quantizer overload point, granular noise to step size, and bandlimiting to sampling rate.

In ADPCM (adaptive differential PCM), a coding method that appears promising for a number of practical applications, the situation is more complicated. Here each design parameter has interrelated effects on several types of degradation, and the perceptual correlates of a particular design are hard to predict. In ADPCM, the step size and overload point vary with time, producing a dynamic mixture of overload and granularity that depends on the adaptation mechanism. One can identify in the quantized waveform two types of overload: overload that causes clipping of stationary inputs and in addition, overload due to slow quantizer response to increases in short term (syllabic) signal level. Moreover, a mathematical study[3] has identified two separate aspects of adaptive quantizer performance: static (response to constant-level inputs) and dynamic (response to changes in input level). However, it is by no means evident or even likely *a priori* that these mathematically separable characteristics are perceived separately.

To investigate perceptual characteristics of speech processed by ADPCM we conducted an experiment that is summarized in the next section. The following five sections provide details of the coding method, objective performance measures, the experimental design, and analyses of subjective and objective measurement data. Section VIII discusses the implications of the principal findings.

II. SUMMARY

High-quality digital recordings of speech samples from four talkers (two male and two female) were processed according to 18 different ADPCM coding schemes on a digital computer. The coders incorporate all combinations of two bit rates, three load constants, and three time constants. The data obtained from the experiment consisted of two types: objective measurements and subjective judgments of the processed speech. With these two types of data we addressed the following questions:

(*i*) What are the perceived characteristics of speech processed by ADPCM?

(*ii*) How are these characteristics related to subjective judgments of quality?

(*iii*) What is the relationship between objective performance measures and the perceptual features?

(*iv*) What is the relationship between objective performance measures and judgments of circuit quality?

(*v*) What combinations of design parameters produce coders within a given quality range?

The analyses of the data indicate the following answers to each of the above questions:

(*i*) Listeners perceive three distinct characteristics of the processed speech: (*a*) the overall clarity, (*b*) the kind of degradation that reduces the clarity, namely, whether the degradation is signal distortion or noise or both, and (*c*) the nature of the signal distortion and/or the nature of the noise.

(*ii*) Quality judgments are correlated with all of these subjective variables. The overall clarity is by far the strongest correlate.

(*iii*) Signal-to-noise ratio, measured segmentally, SNR_{seg}, is a good predictor of the overall clarity. The log of the ratio of attack to recovery speed, $\log A/R$, is a good predictor of the mixture of signal distortion and background noise. The log of the attack time, $\log T_a$, predicts the kind of signal distortion and/or the kind of noise.

(*iv*) SNR_{seg} is a very good predictor of quality judgments, while SNR measured in the traditional manner is a very poor predictor of quality. A linear combination of probability of overload, P, and segmental signal-to-granular-noise ratio, SNRG_{seg} is an even better predictor of quality than SNR_{seg}.

(*v*) By applying the prediction equations to coders with design values intermediate to those of the experiment, we show the combinations of load constant and time constant at each bit rate that would be judged about equal in quality. Those that would be rated almost as highly as the best coder cover a wide range of design parameters.

III. CODER DEFINITIONS

Figure 1 is a block diagram of an ADPCM coder-decoder. In the absence of transmission errors, the sequence of received samples $r'(k)$, is identical to the quantized approximation sequence $r(k)$. In our experiment $s(k)$ was a digital speech signal represented in a 12 bit, 8 kHz format and the coders were realized in software on a Data General Eclipse computer.

The conversion from 12-bit PCM to 3-bit or 4-bit ADPCM is performed according to the algorithm described by Castellino et al.[4] In all of the coders the predictor is a two tap transversal filter with coefficients 1 and -0.5 so that the relationship of approximation signal, $r(k)$, to quantizer output, $d(k)$, is

$$r(k) = d(k) + r(k-1) - 0.5r(k-2). \tag{1}$$

Signal-level estimation. The step size, $\Delta(k)$, which is derived from the sequence of quantized prediction error samples $d(k)$ or equivalently from the transmitted code words, $I(k)$, is proportional to an estimate of the mean absolute value of the quantizer input, $e(k)$. The estimate at time k, $\sigma(k)$, is an exponentially weighted sum of quantizer output magnitudes. It is computed recursively as

$$\sigma(k) = \alpha\sigma(k-1) + (1-\alpha)|d(k-1)|. \tag{2}$$

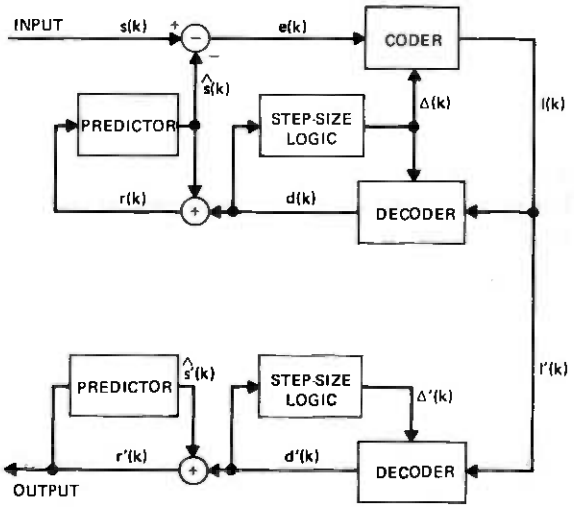


Fig. 1—Block diagram of ADPCM coder-decoder.

Here, the parameter α ($0 < \alpha < 1$) determines the speed of response of the quantizer to changes in input level. A low value of α provides a fast response; $\alpha \approx 1$ is associated with a slow response. A compromise between static and dynamic performance objectives is required in selecting α. A high value ($\alpha \approx 1$) provides more accurate "steady-state" tracking of a constant signal level than does a low value.

If the signal level suddenly increases, the estimate σ increases with an initial slope proportional to $1 - \alpha$ volts/sample (if σ is measured in volts). In this paper we shall refer to an adaptation time constant, τ sec, that is the reciprocal of this initial slope in response to a unit step. It is defined by

$$\tau = \frac{T}{1 - \alpha} \text{ sec}$$

where T is the sample period. For 8 kHz sampling

$$\tau = \frac{0.125}{1 - \alpha} \text{ msec.}$$

To incorporate a perceptibly wide range of signal conditions, we have selected, after informally listening to a large number of coders, 3 values of α for the experiment. They are $\alpha = 1/2, 31/32, 255/256$ with corresponding time constants: $\tau = 0.25, 4, 32$ msec.

Quantizer loading. The quantization step-size $\Delta(k)$ is proportional to the signal-level estimate, $\sigma(k)$:

$$\Delta(k) = C\sigma(k) \tag{3}$$

where the load constant, C, determines in steady-state (fixed signal level) the mixture of granular noise and overload distortion in $d(k)$. A relatively high value of C produces a large average step size and causes granularity to be the principal distortion component. With a very low value of C, overload predominates.

For a given number B bits/sample, we define a nominal load constant, C_0. C_0 is the step size which produces minimum mean square error in a fixed quantizer processing a signal that has Laplacian probability density with unity average magnitude. In fact, it has been noted that the shape of the PDF of the compressed signal $e(k)/\sigma(k)$ is between a Gaussian and a Laplacian function, and more near to the Laplacian one for higher time constants.[5] For $B = 2, 3, 4, 5$ bits, $C_0 = 1.53, 1.03, 0.65, 0.40$, respectively. We define a relative load constant for each quantizer to be

$$L = 20 \log_{10} \frac{C}{C_0} \text{ dB}.$$

After listening informally to speech processed by a variety of coders, we chose for the experiment three relative load constants, $L = -10, -4, 4$ dB.

It has been shown[6] that this adaptive quantizer is a special case of the one with multiplicative step size changes[7]:

$$\Delta(k + 1) = M[I(k)]\Delta(k).$$

The multipliers associated with code words $I(k) = \pm 1, \pm 2, \ldots \pm 2^{B-1}$ are

$$M[I(k)] = \alpha + (1 - \alpha)C(|I(k)| - 0.5).$$

Dynamic and static behavior. The dynamic behavior of the coder can be described by two characteristics: the attack and recovery speeds. The attack speed is defined as the step size increase (in dB) per unit time when the signal level suddenly changes from a very low value to a very high value. The recovery speed is defined as the step size decrease (in dB) per unit time when the signal level suddenly falls. The attack and recovery speeds can be computed from the largest and smallest multipliers[3]:

$$v_a = \frac{20}{T} \log M(I_n) \frac{\text{dB}}{\text{sec}}$$

$$v_r = \frac{-20}{T} \log M(I_1) \frac{\text{dB}}{\text{sec}}$$

where T is the sampling time, and $n = 2^{B-1}$. A small attack speed will produce slope overload distortion, while a small recovery speed will result

in greater granular distortion. The static behavior is also affected by attack and recovery speed: a high attack and low recovery speed result in a step size that is higher on the average than that resulting from a slow attack and fast recovery. A very good indicator of static performance is the attack to recovery ratio:

$$A/R = \frac{\nu_a}{\nu_r} = \frac{\log M(I_n)}{-\log M(I_1)}.$$

Attack time, which we have found to be strongly correlated with the type of distortion or the type of noise produced by a coder, is the reciprocal of attack speed:

$$T_a = \frac{1}{\nu_a}.$$

Summary of conditions. The experiment includes coders with 3 variable design parameters: B bits/sample, τ msec response time, and L dB relative load constant. The 18 coders comprise all combinations of $B = 3, 4$; $\tau = 0.25, 4, 32$; $L = -10, -4, 4$.

IV. OBJECTIVE MEASURES

Our aims include exploration of the relationships between perceived characteristics of the processed speech and objectively measurable quantities. To investigate these relationships, we have computed several objective performance indices for each processed utterance. The measures are defined as follows:

Total signal-to-noise ratio.

$$\text{SNR} = 10 \log \frac{\Sigma s^2(k)}{\Sigma[s(k) - r(k)]^2}.$$

Here k ranges over all samples in the utterance, and $r(k)$ is defined as the best estimate of $s(k)$.

Granular signal-to-noise ratio.

$$\text{SNRG} = 10 \log \frac{\Sigma s_g^2(k)}{\Sigma[s_g(k) - r_g(k)]^2}$$

where k ranges over all samples in the utterance and the signals $s_g(k)$ and $r_g(k)$ are defined only when the quantizer is not overloaded; that is, when the quantization error is less than one-half the step size:

$$s_g(k) = s(k); r_g(k) = r(k) \quad \text{if} \quad |s(k) - r(k)| \leqslant \frac{\Delta(k)}{2}$$

$$s_g(k) = r_g(k) = 0 \quad \text{if} \quad |s(k) - r(k)| > \frac{\Delta(k)}{2}.$$

Percent of samples overloaded.

$$P = 100\left[1 - \frac{\sum_{k=1}^{N} s_g(k)/s(k)}{N}\right]$$

where N is the total number of samples in the utterance.

Total segmental signal-to-noise ratio. This is a measure proposed by Noll[8] as a more relevant index of speech quality than SNR:

$$\text{SNR}_{seg} = \frac{1}{M}\sum_{m=0}^{M-1} 10 \log \frac{\sum_{j=1}^{128} s^2(j + 128m)}{\sum_{j=1}^{128} [s(j + 128m) - r(j + 128m)]^2}$$

Here the utterance is divided into segments each containing 128 samples (16 msec) and the signal-to-noise ratio in each segment is measured in dB. The average of these measures over the M segments in the utterance is SNR_{seg}.

Granular segmental signal-to-noise ratio.

$$\text{SNRG}_{seg} = \frac{1}{M}\sum_{m=0}^{M-1} 10 \log \frac{\sum_{j=1}^{128} s_g^2(j + 128m)}{\sum_{j=1}^{128} [s_g(j + 128m) - r_g(j + 128m)]^2}$$

Here the same procedure as for SNR_{seg} is applied to samples $s_g(k)$ and $r_g(k)$.

V. TESTING PROCEDURE

Digital recordings* of 10 sentences spoken by each of 4 talkers (2 male and 2 female) were processed by each of the 18 coders. The processed sentences were equalized to the same mean power to eliminate level differences due to quantizer overloading and thereby minimize differences in subjective loudness. Four analog test tapes, each containing the 153 possible pairs of coders, were prepared from these recordings. The speech samples in a pair of conditions were the same sentence by the same talker. The talkers and sentences were assigned to coder pairs so that they occurred as equally as possible on a tape. A pair of coders processed a different talker and sentence on each tape and the order of

* The source speech was the set of digital tape recordings used in a previous experiment on PCM.[1]

presentation within the pair was reversed on half of the tapes. The coder pairs appeared in a different (random) order on each tape.

Students from the junior and senior classes of local high schools served as paid subjects. They listened to the processed speech over Pioneer SE700 earphones at 80 dBSPL while seated in a double-walled sound booth with frequency-weighted room noise introduced at a level of 50 dBA.

Dissimilarity judgments. In the first experiment, 17 subjects (3,4,5,5 per random order) judged the pairs of conditions. They were told to use the numbers from 0 to 9 to indicate how different the speech sounded over each pair of coders, using a 0 for no difference, a 9 for very different, and the numbers between 0 and 9 for intermediate differences. Before the test session began they judged 6 pairs, for practice, that represented the expected range of differences.

Preference judgments. In the second experiment, 16 different subjects (4,5,4,3 per random order), also junior and senior high-school students, listened to the same tapes containing the pairs of coders. However, this time the subjects were instructed to indicate which condition of each pair they would find more acceptable for listening to speech.

Rating judgments. In the third experiment, subjects judged the quality of the coders individually. Eight audio tapes were prepared, each containing 36 sentences. Tapes 1–4 had the processed speech (played through the 18 coders) of one male and one female talker. The other two talkers appeared on tapes 5–8. The stimuli on tapes 1–4 appeared in different (randomized) orders. The same 4 orderings were used for tapes 5–8. The sentences occurred as equally as possible on each tape.

In this experiment the subjects were asked to rate the quality of the 36 conditions according to the adjectives: excellent, good, fair, poor, unsatisfactory. Their answer sheets contained 36 rows of short lines separated into 9 columns. The odd columns were labeled with the adjectives and the even ones unlabeled, allowing the subjects to check intermediate ratings if they chose to do so. On half the answer sheets the order of the labels were reversed. Tapes 1–4 were presented to the 17 listeners of the first experiment in a short session that took place 5 minutes after the completion of the difference judgments. Tapes 5–8 were presented to the 16 listeners of the second experiment 5 minutes after the completion of the preference judgments.

VI. INITIAL DATA REDUCTION

The experiment was performed to provide information about relationships between ADPCM coders and it is expected that for the most part differences in listener responses are due to coder differences. The experiment was designed to cause other sources of variability to be mutually cancelling in the average data for each coder. Sources of extraneous

variability are: differences in the way listeners use the response scales, differences in speech material, and effects of presentation order. Before aggregating the data for individual coders it was necessary to assess the importance of each of them.

Difference judgments. The variability due to listener differences is revealed by the correlation coefficients of pairs of subjects who heard the same tape. These coefficients have a mean value of 0.61 and standard deviation of 0.09, indicating substantial agreement. The effects of presentation order and talker were tested by means of an analysis of variance which showed that the responses to the 4 random orders were not significantly different. The variability due to the different talkers was significant at the 0.05 level, but accounted for only 1 percent of the total variance. This variability is due to the fact that speech of female talkers was rated differently from speech of male talkers. There was no significant difference in the ratings of talkers of the same sex.

The important variability in the difference data can therefore be attributed to coder differences, and to assess these differences, we normalized the 153 responses of each subject to zero mean and unity standard deviation. The averages, across the 17 subjects, of the normalized responses were the elements of a dissimiliarity matrix which was analyzed according to the MDSCAL[9–13] procedure.

MDSCAL locates points, representing the stimuli, in a multidimensional space so that the distances between the points are monotonically related to the judged differences. Because the dimensionality of a solution is specified as input, successive solutions of increasing dimensionality are usually computed. Then, the stress values (essentially the root mean square error) and the interpretability of each solution are used as criteria for deciding upon the smallest number of dimensions that are needed to explain the data. The stress values give a measure of how well the distances in the solution spaces correspond to the reported differences among the coders. Solutions in 1, 2, and 3 dimensions for the difference judgments among the 18 coders had stress values of 0.25, 0.11, and 0.07, respectively. The large decrease in stress between the 1 and 2 dimensional solutions indicates that at least 2 dimensions are needed to account for the data. Although a 3-dimensional solution accounted for only a small additional decrease in stress, it offered an enhanced interpretation of the subjective space. (See Section VII.)

Preference judgments. The preference data were analyzed according to MDPREF,[14,15] a factor analytic procedure that measures the variability in preference among the subjects. The proportion of the total variance contributed by each factor is related to the agreement among the subjects on the relative importance of different characteristics of the stimuli. In the solution for the preference judgments of the 18 coders, the first factor accounted for 0.89 of the variance and the second accounted for only an

additional 0.02, indicating strong agreement among the listeners and a single factor solution. Therefore, the values of the points from the one factor solution were used for the scale of preference.

Rating judgments. We computed the correlations of the ratings of each subject with those of each of the other subjects who listened to the same tape. The distribution of the correlation coefficients had a mean of 0.78 and standard deviation of 0.08, again showing a high degree of agreement. Therefore, the mean across subjects of the individual responses, normalized so that the ratings of each subject had zero mean and unit variance, were used in an analysis of variance due to coder parameters, talkers, and random orders. The 3 design variables, load constant, time constant, and bits, were all significant at the 0.05 level. The variability due to the random orders was not significant, but the variability due to the different talkers was significant. As in the difference experiment, the significant talker variability was due to differences between the male and female talkers, accounting for only 1 percent of the variance.

Rating vs. preference. The two types of quality judgments, preference and rating, were obtained so that the two testing methods could be compared. The paired comparison tapes were designed to balance many of the sources of variability that are artifacts of the testing procedure. Each coder was heard an equal number of times with each talker and approximately an equal number of times with each sentence. The order of presentation was reversed on half of the trials and, of course, the relative merit of each coder was ultimately determined by comparing it with every other coder. In the rating judgments, the merit of a coder was determined by one presentation per talker. Ratings assume that the quality represented by the five adjectival categories are not only well defined for each individual, but are essentially the same for all individuals. Although the ratings were normalized before computing the analysis of variance, the more customary procedure is to simply average the original judgments across subjects. Therefore, to compare the results of the rating study with those of the more critical paired-comparison study, the mean across subjects of the original unnormalized ratings were used.

Figure 2 shows a scatter plot of these ratings vs. transformed quality measures from the one factor MDPREF solution of the paired-comparison judgments. (The linear transformation scales the maximum and minimum measures to one and nine, respectively.) As this plot shows, the agreement in ratings for the two methods is extremely high: the correlation is 0.99. Thus it appears that the uncontrolled sources of variability that could contaminate simple rating judgments did not have a strong influence on the variability of these data. The additional experimental effort involved in collecting paired-comparison judgments in order to control this variability did not increase the accuracy.

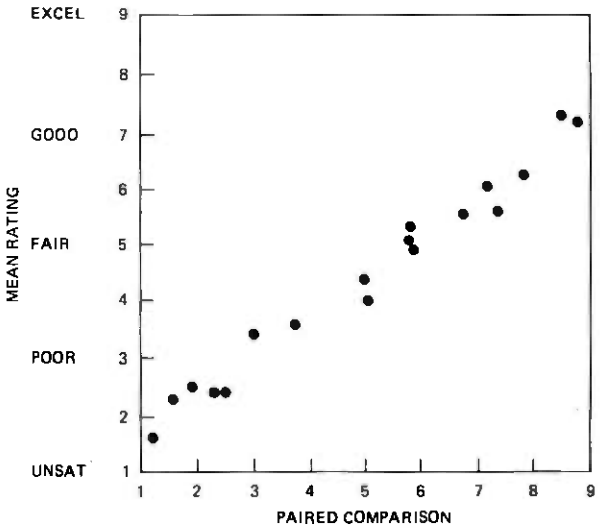


Fig. 2—Relationship between overall evaluation of coders by category ratings and paired comparison preference methods.

VII. RESULTS

The output of MDSCAL is a set of points, representing the coders, in Euclidean space. The inter-point distances are related to the judged differences between coders and the orientation of the points in this space is supposed to reveal the underlying perceptual characteristics of the coders. However, the distances are invariant under orthogonal rotation and it is often the case that a rotation of the MDSCAL coordinates is necessary to interpret the configuration in terms of known coder properties.

In evaluating MDSCAL analyses of varying dimensionality, we concluded that a 3-dimensional geometry would be most informative. We approached the rotation problem by using multiple linear regression procedures to locate vectors in the 3-dimensional space on which the projections of the points are maximally correlated with various objective measures and design parameters. We also located the vector on which the projections of the points are maximally correlated with the average subjective ratings obtained in the third experiment. The vector for each measure of the 18 coders was located independently. Table I displays some of the measures for which vectors were derived and Table II gives the correlations between measurements and the corresponding projections on vectors in the MDSCAL space. These vectors are an aid to the interpretation of the subjective space because they make it possible to relate directions in space to changes in design parameters and performance measures. As a visual aid, the coordinate axes were rotated so that they nearly or exactly coincide with meaningful directions.

Table I

Coder	B	τ ms	L_r db	T_a, ms	T_r, ms	A/R	SNR, dB	SNR_{seg}, dB	SNRG, dB	$SNRG_{seg}$, dB	P, %	R
1	3	0.25	4	0.01	0.15	12.63	10.5	8.8	10.5	8.8	0.07	3.4
2	3	0.25	−4	0.03	0.04	1.20	14.8	15.7	20.3	19.0	6.69	5.7
3	3	0.25	−10	0.21	0.03	0.13	2.9	4.5	24.9	23.7	58.28	2.4
4	3	4.00	4	0.11	2.51	23.99	11.1	9.0	11.1	9.2	0.11	3.2
5	3	4.00	−4	0.37	0.68	1.84	13.1	14.6	20.9	19.6	6.15	5.6
6	3	4.00	−10	3.27	0.54	0.17	2.1	4.2	24.3	23.0	57.32	2.5
7	3	32.00	4	0.79	20.13	25.54	11.1	5.5	11.2	5.9	0.60	2.3
8	3	32.00	−4	2.89	5.45	1.89	7.9	11.2	20.7	17.8	13.03	4.4
9	3	32.00	−10	26.16	4.39	0.17	0.7	3.6	22.4	22.9	63.79	1.6
10	4	0.25	4	0.01	0.05	5.32	15.9	14.8	16.0	14.8	0.07	4.9
11	4	0.25	−4	0.02	0.03	1.41	19.2	20.4	24.5	23.1	2.75	7.4
12	4	0.25	−10	0.06	0.02	0.40	8.9	11.7	29.9	28.1	25.74	5.2
13	4	4.00	4	0.08	0.94	12.53	17.1	15.5	17.2	15.7	0.11	5.2
14	4	4.00	−4	0.23	0.57	2.50	17.0	19.7	25.5	23.9	2.59	7.2
15	4	4.00	−10	0.85	0.51	0.59	7.4	10.3	29.8	27.9	24.91	6.2
16	4	32.00	4	0.55	7.60	13.73	17.1	13.0	17.7	13.5	0.44	4.0
17	4	32.00	−4	1.78	4.63	2.60	11.8	16.2	25.5	22.3	6.81	6.1
18	4	32.00	−10	6.78	4.10	0.60	3.3	9.4	28.9	27.1	34.90	3.5

Table II

Objective measure	Corr. with vector values
SNR	0.95
SNR_{seg}	0.96
log A/R	0.96
SNRG	0.94
$SNRG_{seg}$	0.95
P overload	0.98
log T_a	0.92
log T_r	0.85
Rating	0.99

As a further step in interpreting the space, we listened to one of the tapes used in the rating experiment. After hearing each sentence, the three of us independently wrote adjectives to describe the processed speech. Examples of coder descriptions are: "clear, some noise," "slightly muffled, medium noise," "crackling noise," "very hoarse."

After considering several other rotations, we chose the solution displayed in Figs. 3 and 4 as most interpretable because the coordinate axes nearly or exactly coincide with vectors of measurable quantities and the coder descriptions cluster in a meaningful way. With this rotation, the proportions of the total variance accounted for by dimensions I, II, and III are 0.62, 0.19, and 0.19, respectively.

Subjective variables. When the descriptive adjectives were related to the configuration of points on the plane of the first two dimensions, shown in Fig. 3, an interpretation emerged that was reminiscent of a similar analysis in a study of analog circuits.[16] The interpretation of the space in that study indicated that listeners distinguish among the pro-

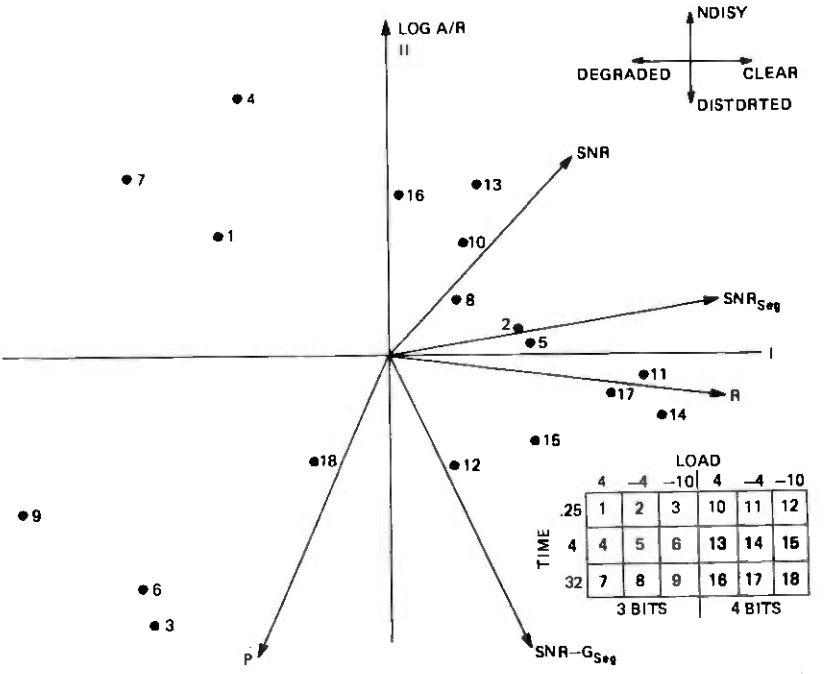


Fig. 3—Projections of points, representing coders, and vectors, representing measures, on the plane of dimensions I and II.

cessed speech samples according to whether the speech is clear or degraded. When degradation is present, they further distinguish between noise in addition to the signal and distortion of the signal itself. The same interpretation applies to the first 2 dimensions in the ADPCM coder space. The coders we described as having clear speech and little noise are high on the first dimension. The coders that were described as noisy, muffled, and hoarse are low on the first dimension, and intermediate amounts of each type of reduction in overall clarity are distributed between these two extremes. Thus, the first dimension appears to represent the overall clarity of the speech.

The plane of dimensions II and III, shown in Fig. 4, identifies the characteristics of the speech that reduce the clarity. The coders that we described as noisy are high on the second dimension and those that we described as muffled or hoarse are low on the second dimension. Thus, the second dimension represents the two kinds of degradations that reduce the overall clarity: background noise and distortion of the speech signal itself. The conditions we described as very muffled sounded as though the speaker had his hand, or some other material object, in front of his mouth, and these conditions are high on dimension III. Those conditions we described as hoarse sounded as though the speaker had

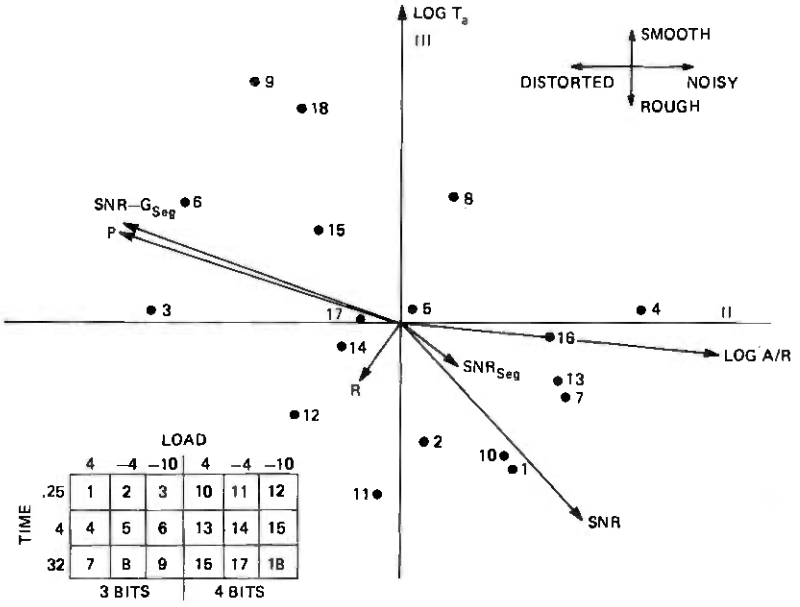


Fig. 4—Projections of points, representing coders, and vectors, representing measures, on the plane of dimensions II and III.

laryngitis and, in general, they are low on the third dimension. Two kinds of background noise were also identified: one that is described as crackling, and one that is more like the familiar white random noise. In general, the coders with crackling noise have low values on the third dimension and those with white random noise are in an intermediate position. Thus, the third dimension appears to represent a further distinction between each kind of degradation that could be described as rough vs. smooth. Hoarse speech and crackling noise are rough or irregular in character, muffled speech is smooth or uniform, while speech corrupted by white noise is intermediate between these two extremes.

Objective measures. Figures 3 and 4 also show the vectors corresponding to various objective measures. The vector SNR_{seg} is very close to the coordinate axis of dimension I and is therefore a good indicator of the overall clarity of the processed speech. Log A/R predicts the distribution of points on the second dimension, interpreted as the prevalent kind of degradation, signal distortion or background noise. A low A/R produces a low step size on average, leading to slope overload, perceived as signal distortion. On the other hand, a high value of A/R results in a high average step size and high granular noise. The locations of the vectors $SNRG_{seg}$ and P are also consistent with our interpretation of the coordinate axes. They both have high negative weighting on dimension II because both reflect the predominant impairment category. High P

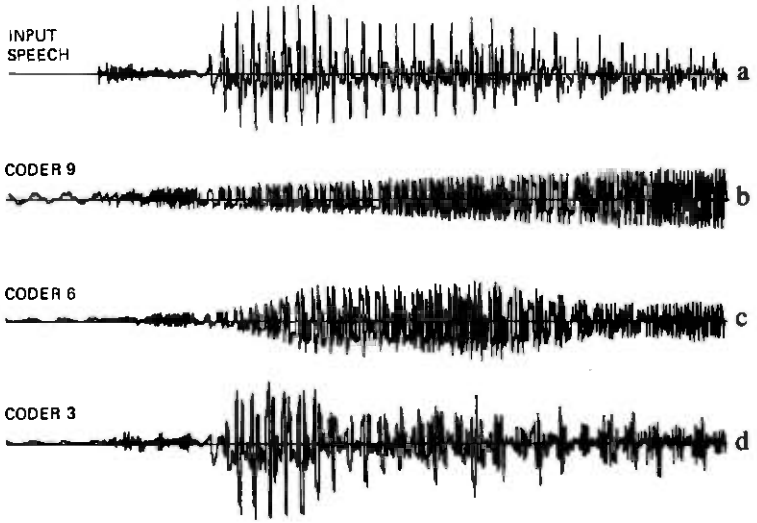


Fig. 5—Waveforms of the word "tool" for 3-bit ADPCM coders low on dimension II (signal distortion), with different values on dimension III, showing the relationship of attack time to the smooth or rough subjective descriptions.

means high overload and substantial distortion; high SNRG_{seg} means low noise. Their non-zero weight on dimension I indicates their influence on speech clarity. A high value of SNRG_{seg} indicates low background noise and enhanced clarity. Conversely, a high value of P is correlated with high distortion and thus with low clarity. The small angle between P and SNRG_{seg} reflects the fact that overload and granularity usually vary reciprocally in coders with a given number of bits per sample.

The coordinate axis of dimension III is highly correlated with $\log T_a$. When the attack time is very high, the step size is very slow in following fluctuations in input level and the speech sounds muffled. When the attack time is very low, the step size frequently overshoots its target value at the beginning of pitch periods, causing irregularity in the periodicity of the processed speech. This irregularity makes the speech sound hoarse.

These properties are apparent in Figs. 5 to 7 which show waveforms that are representative of coder locations in the II–III plane. All of them display the word "tool" processed by 3-bit ADPCM coders with substantially impaired clarity. Figure 5 shows the waveforms of distortion-producing coders (low weighting on dimension II). With −10 dB relative load factor they all produce substantial slope overload in steady state. Coder 9 (Fig. 5b) with $T_a = 32$ msec and high weight on dimension III is the most muffled; fluctuations in the signal envelope are very heavily smoothed. Coder 6 (Fig. 5c), $T_a = 4$ msec, lower on dimension III, reproduces long-term envelope fluctuations, but smoothes out in-

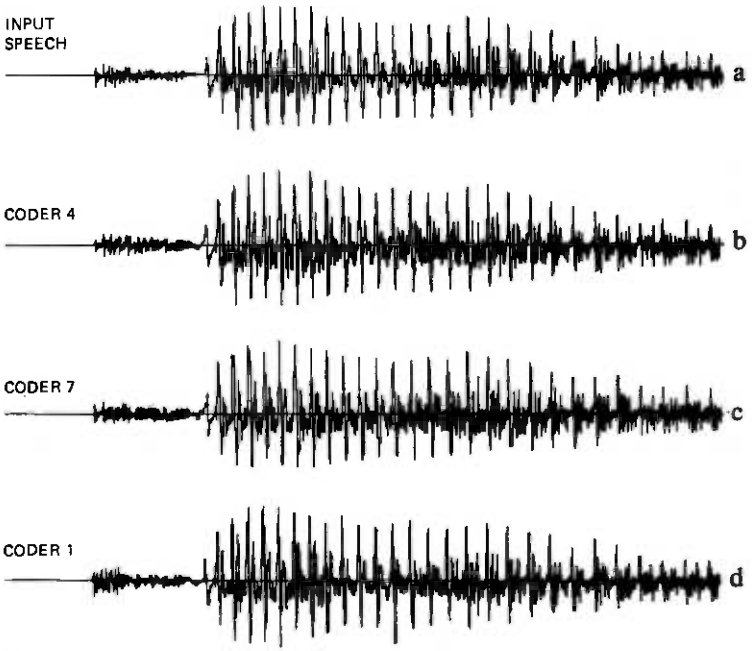


Fig. 6—Waveforms of the word "tool" for 3-bit ADPCM coders high on dimension II (background noise), with different values on dimension III.

dividual pitch periods. Coder 3 (Fig. 5d), T_a = 0.25 msec, moderate loading on dimension III, reproduces pitch contours but with substantial time and amplitude distortion. Figure 6 shows the waveforms of the very noisy coders, 4, 7, and 1, with high relative load factors (4 dB) and high weighting on dimension II. With low distortion they all preserve the general envelope and time structure of the original, so that the nature of their impairment is best seen in oscillograms of noise voltages (coder output minus input). The two extreme types of noise are displayed in Fig. 7. Coder 4, with relatively high weight on dimension III, has "smooth" noise which is shown in Fig. 7b to be correlated with the long term envelope of the speech. In Fig. 7c, coder 1, with crackling noise impairment, low weight on dimension III, is seen to produce impulsive-type noise correlated with the pitch contours of the signal.

Quality prediction. The vector labeled R corresponds to the mean ratings on the 9-point response scale and is very close to the first dimension. As indicated by Fig. 2, a vector corresponding to quality derived from the paired-comparison preference judgments would be in essentially the same location. Multiple regression procedures were used to derive linear relationships between the objective measures and subjective quality. Table III lists the formulas for predicting the ratings from several

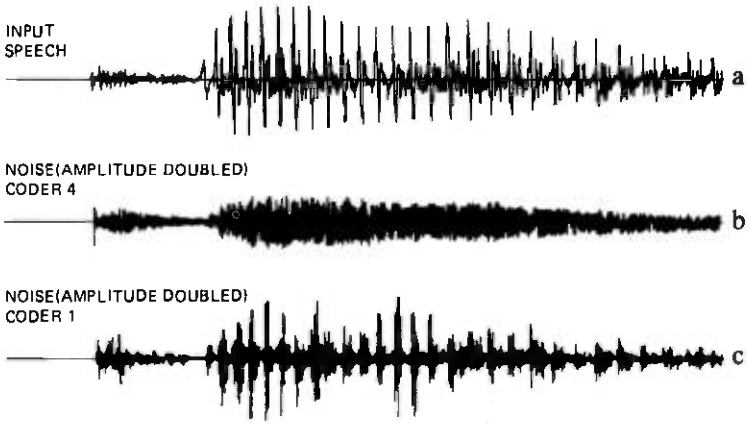


Fig. 7—Noise waveforms relative to the word "tool" processed by two coders of Fig. 6, with the amplitude scale doubled, showing the influence of attack time on the types of noise.

of the objective measures, singly and in combination. Prediction accuracy is indicated by correlations between actual and estimated ratings (1.0 would be perfect agreement), and by the rms error expressed as a fraction of a point on the 9-point scale. Table III shows that, consistent with the locations of their vectors relative to the R vector, SNR_{seg} (formula 2) is a very good predictor of subject quality and SNR (formula 1) is not a good predictor. Among the formulas that contain more than one objective measure, the most accurate predictors of subjective quality are 7 and 8 which include separate measures of granular and overload impairments. Although the location of the vector corresponding to SNRG is essentially the same as that of the SNRG_{seg} vector, prediction accuracy is higher when the measurement is made segmentally.

Other coders. Since formula 8 proved an accurate estimator of the subjective quality of the 18 coders in the experiment, we used it to estimate subjective quality of other ADPCM coders with a wide range of

Table III

	Formula for predicting rating	Corr.	RMS error
1	$0.21\ \text{SNR} + 2.27$	0.69	1.20
2	$0.31\ \text{SNR}_{seg} + 0.89$	0.93	0.63
3	$0.33\ \text{SNR}_{seg} - 0.45 \log A/R + 0.79$	0.95	0.54
4	$0.25\ \text{SNRG} - 0.067\ P - 0.39 \log T_a + 0.22$	0.94	0.55
5	$0.24\ \text{SNRG} - 0.077\ P + 0.68$	0.93	0.63
6	$0.16\ \text{SNRG} - 1.11 \log T_a + 0.47$	0.69	1.21
7	$0.24\ \text{SNRG}_{seg} - 0.078\ P - 0.22 \log T_a + 1.00$	0.96	0.45
8	$0.25\ \text{SNRG}_{seg} - 0.084\ P + 1.19$	0.96	0.48
9	$0.14\ \text{SNRG}_{seg} - 1.03 \log T_a + 1.38$	0.65	1.27
10	$-0.036\ P - 0.26 \log T_a + 4.97$	0.56	1.38

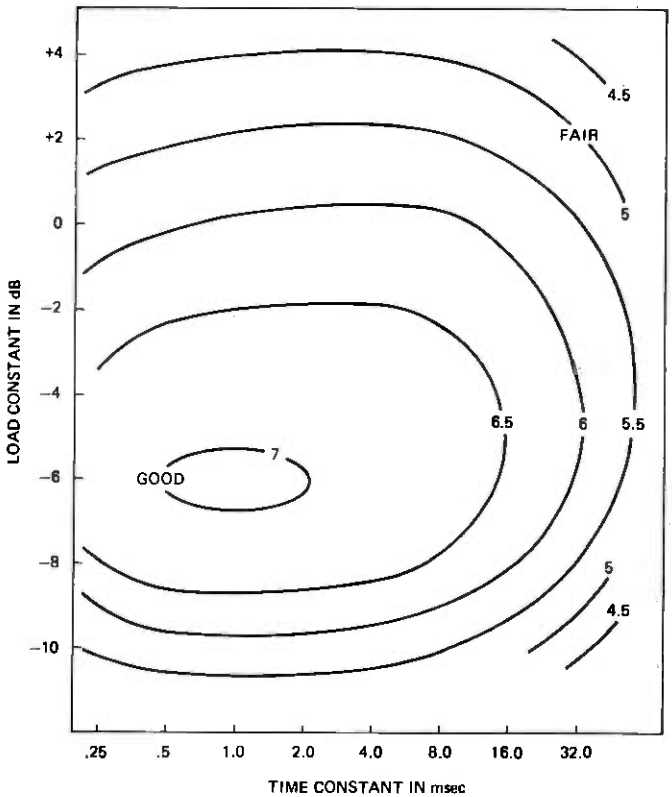


Fig. 8—Equi-rating contours predicted by formula 8, Table III, for 4-bit ADPCM coders.

design parameters. To do so, for each bit rate, we simulated 64 coders that comprised all combinations of 8 load factors and 8 time constants. The 8 values of each parameter included the 3 tested in the original experiment and 5 intermediate values. SNRG_{seg} and P were measured on 4 sentences, one by each talker, processed through each of the 64 coders. The quality ratings, predicted using formula 8 with the averages of the measures on the 4 sentences, are displayed in Figs. 8 and 9, which pertain to 4-bit and 3-bit coders, respectively. The equi-rating contours show that near-optimum quality can be expected over a surprisingly wide range of circuit conditions. For instance, with 4-bit coding, a rating of 6.5 ($\frac{1}{2}$ point from optimum on the 9-point scale) is maintained over a 7 dB range of load factors and a 32:1 range of time constants.

VIII. DISCUSSION

Perceptual characteristics. Our interpretation of the 3-dimensional subjective space is consistent with previous work[16] on analog speech

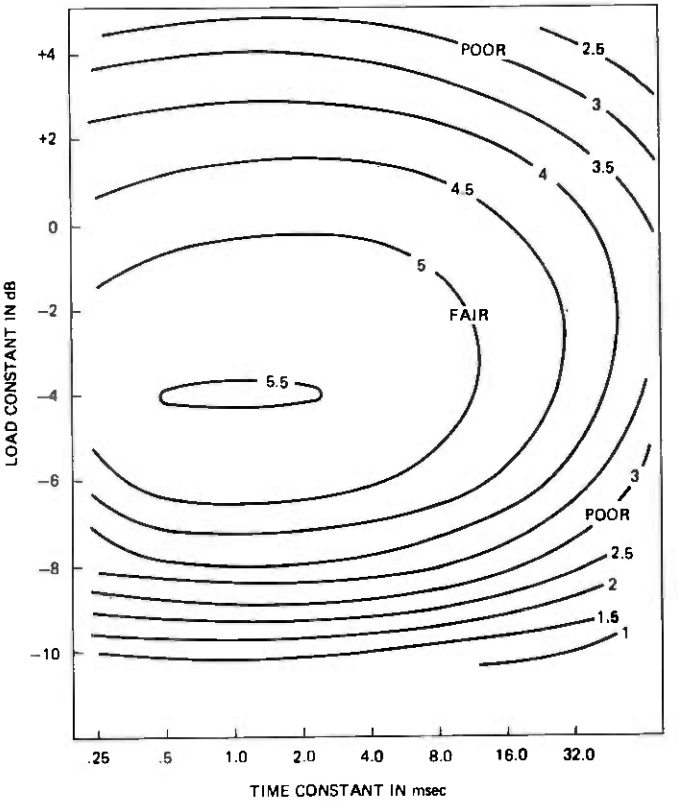


Fig. 9—Equi-rating contours predicted by formula 8, Table III, for 3-bit ADPCM coders.

impairments. In both cases, the first two dimensions have the same meaning. The third dimension in Ref. 16 was related to loudness. In the present experiment, the stimuli were equalized in level, so that subjects could attend to less obvious differences, like the "rough" or "smooth" character of the impairment. The plane of dimensions II and III, Fig. 4, provides perhaps the most interesting view of the subjective space. Accounting for 38 percent of the variance in the average difference judgments, it represents the *kind* of degradation, independent of the *amount* of degradation. In this plane, the perceptually meaningful classifications of ADPCM impairments are the categories, "speech distortion" and "background noise" (dimension II) and, in addition, the types of distortion, "muffled" (smooth) and "hoarse" (rough) and the types of noise, "continuous" (smooth) and "crackling" (rough).

This geometric representation also confirms that the mathematical separation of ADPCM performance into static and dynamic response categories[3] is perceptually meaningful. Dimension II is highly correlated

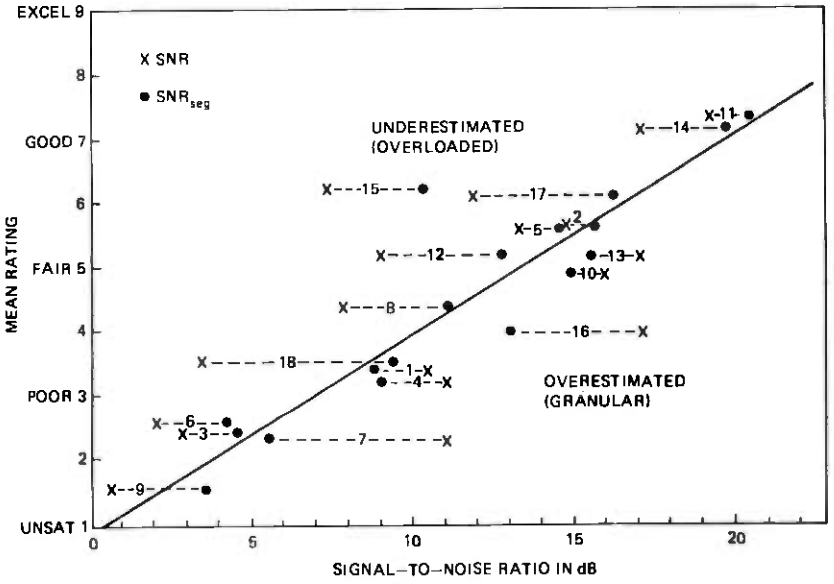


Fig. 10—Mean ratings for the 18 coders vs. SNRs showing the improvement in prediction by measuring SNR segmentally. The solid line is the graph of formula 2, Table III, the regression of rating on SNR_{seg}.

with log A/R, a measure of static performance, and dimension III is highly correlated with log T_a, a measure of dynamic performance.

Objective measures. Table III indicates that, as in the case of PCM,[1] ADPCM quality is accurately predicted by a linear combination of overload and granularity measures. Formulas 4, 5, 7, and 8, all containing separate measures of overload and granularity, are among the 6 good predictors of average rating. The table also demonstrates the value of measuring signal-to-noise ratio segmentally. Formula 2, which contains the single measurement, SNR_{seg}, is also one of the 6 good predictors. Segmental measures give equal importance to strong and weak components of speech, while non-segmental SNR is essentially a measure of the quality of the high-level components. The strong correlation of SNR_{seg} with average rating indicates that subjective quality judgments are influenced by weak sounds as well as strong sounds.

These properties of SNR and SNR_{seg} are revealed by Fig. 10 which is a scatter plot of average rating vs. both measures for the 18 coders. SNR points are labeled with crosses and SNR_{seg} points are labeled with circles. The line is the graph of formula 2, the regression of R on SNR_{seg}. The coders to the left of the line are, for the most part, those with low A/R, in which overload distortion is the predominant impairment. This distortion affects only the strong sounds which are the ones that determine SNR. The good reproduction of weak sounds by overloaded coders is not

reflected by SNR and the crosses for these coders tend to be far to the left of the regression line. That is, they give an unduly poor indication of quality. By contrast, the points to the right of the line tend to be those with high A/R, coders with mainly background noise impairment. This degradation is particularly harmful to weak sounds and therefore its effect is less on SNR than on SNR_{seg}. Consequently, SNR gives an unduly good indication of the quality of these coders.

SNR_{seg} apparently resembles Q, the objective measure of coder quality proposed by D. L. Richards.[17] Q is an average of SNR measures performed with different input levels of a stationary signal. As such it apparently fails to take into account the dynamic response of a coder, which is an important aspect of adaptive quantization. We therefore speculate that as an estimator of subjective quality, the accuracy of Q is intermediate between that of SNR and SNR_{seg}.

Coder design. The relatively large distances between equi-rating contours in Figs. 8 and 9 show that a designer has very substantial latitude in choosing a coder with a prescribed quality rating. This finding is contrary to quality predictions based on conventional SNR measures, which indicate that only restricted sets of design parameters offer near-optimum performance. This newly discovered design flexibility could be valuable in finding coders that simultaneously satisfy criteria in addition to the quality of the coding-decoding process. Examples of such criteria are quality of tandem connections of codecs, resistance to transmission errors, ability to communicate voiceband data, compatibility with other code formats, and economy of implementation.

IX. ACKNOWLEDGMENT

We thank Ann Quinn for recruiting the subjects and running the experiment.

REFERENCES

1. D. J. Goodman, B. J. McDermott, and L. H. Nakatani, "Subjective Evaluation of PCM Coded Speech," B.S.T.J., *55*, No. 8 (October 1976), pp. 1087–1109.
2. D. J. Goodman, J. S. Goodman, and M. Chen, "Intelligibility and Subjective Quality of Digitally Coded Speech," IEEE Trans. on Acoustics, Speech & Signal Processing (in press).
3. D. J. Goodman and A. Gersho, "Theory of an Adaptive Quantizer," IEEE Trans. on Commun., *COM-22*, No. 8 (August 1974), pp. 1037–1045.
4. P. Castellino, G. Modena, L. Nebbia, and C. Scagliola, "Bit Rate Reduction by Automatic Adaption of Quantizer Step Size in DPCM Systems," Int. Zurich Seminar, Zurich, Switzerland, April, 1974.
5. C. Scagliola, "An Adaptive Quantizer with Channel Error Recovery," CSELT Rapporti Technici, *IV*, No. 4 (December 1976), pp. 177–184.
6. D. J. Goodman and R. M. Wilkinson, "A Robust Adaptive Quantizer," IEEE Trans. on Comm., *COM-23*, No. 11 (November 1975), pp. 1362–1365.
7. N. S. Jayant, "Adaptive Quantization with One Word Memory," B.S.T.J., *52*, No. 7 (September 1973), pp. 1119–1144.
8. P. Noll, "Adaptive Quantizing in Speech Coding Systems," Int. Zurich Seminar, Zurich, Switzerland, April, 1974.

9. R. N. Shepard, "Metric Structures in Ordinal Data," J. Math. Psych., *3*, No. 2 (July 1966), pp. 287–315.
10. R. N. Shepard, "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function. I," Psychometrika, *27* (1962), pp. 125–140.
11. R. N. Shepard, "The Analysis of Proximities: Multidimensional Scaling With an Unknown Distance Function. II," Psychometrika, *27* (1962), pp. 219–246.
12. J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika, *29* (March 1964), pp. 1–27.
13. J. B. Kruskal, "Nonmetric Multidimensional Scaling: a Numerical Method," Psychometrika, *29* (June 1964), pp. 115–129.
14. P. Slater, "Analysis of Personal Preferences," Brit. J. Stat. Psychol., *13* (November 1960), pp. 119–135.
15. J. D. Carroll, "Individual Differences and Multidimensional Scaling," in *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences—Vol. I: Theory*, Shepard, Romney, Nerlove (eds.), New York: Seminar Press, 1972, pp. 105–155.
16. B. J. McDermott, "Multidimensional Analyses of Circuit Quality Judgments," J. Acoust. Soc. Am., *45*, No. 3 (March 1969), pp. 774–781.
17. D. L. Richards, "Speech Transmission Performance of PCM Systems," Electronics Letters, *1*, No. 2 (April 1965), pp. 40–41.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Evaluation of a Word Recognition System Using Syntax Analysis

By S. E. LEVINSON, A. E. ROSENBERG, and J. L. FLANAGAN

(Manuscript received May 18, 1977)

A speech recognition system has been implemented which accepts reasonably natural English sentences spoken as isolated words. The major components of the system are a speaker-dependent word recognizer, a programmed grammar, and a syntax analyzer. The system permits formulation of complete sentences from a vocabulary of 127 words. The set of sentences selected for investigation is intended for use as requests in an automated travel information system. Results are presented of evaluations for speakers using their own stored reference patterns, the reference patterns of other speakers, and composite reference patterns averaged over several speakers. For speakers using their own reference patterns the median error rate for acoustic recognition of the individual words is 11.7 percent. When syntax analysis is applied to the complete sentence, word recognition errors can be corrected and the error rate reduced to 0.4 percent.

I. INTRODUCTION

A speech recognition system composed of a programmed syntax analyzer and a speaker-dependent word recognizer has been evaluated. The system accepts complete sentences in which the successive words are spoken distinctly and in isolation. The purpose of the experiment is to determine the capability of syntax analysis for improving the accuracy of word recognition and for expanding the command ensemble of a voice-actuated system.

The word recognition system, designed by Itakura,[1] is based on representing speech utterances by equally spaced frames of LPC coefficients. Recognition ensues from a comparison of a sample input pattern of LPC coefficients with an ensemble of stored reference patterns previously established by the designated speaker. The comparison consists of a frame-by-frame scan of a sample pattern against each reference pattern. A distance metric (or measure of dissimilarity) is calculated and accu-

mulated by a dynamic programming technique as the scan proceeds. The vocabulary item corresponding to the reference pattern with the lowest accumulated distance is designated the recognized item. In addition, a distance rejection threshold is imposed. If the accumulated distance exceeds the threshold at any frame during a reference scan, that particular reference comparison is aborted. If all reference comparisons for a sample pattern are aborted, the result is said to be "no match" or "reject."

II. EVALUATION OF THE ACOUSTIC ANALYZER

An earlier evaluation of the automatic word recognition system was carried out over a five-month period over dialed-up telephone lines.[2] Thirteen speakers participated in that test. Each dialed the system once a day and provided utterances of words selected from an 84-word vocabulary. The 84-word vocabulary was designed to provide one-word responses to questions asked by a computer-controlled digital voice response system. The computer was programmed to provide airline flight information requested by a caller. In this system the question-answer dialog that takes place between the caller and the computer results in the specification of a category of flights for which information is desired. Because of the nature of this dialog, 50 of the 84 vocabulary items were the names of North American cities. Other entries were digits, days of the week, etc. In the evaluation using this vocabulary, with approximately 750 trials per speaker, the median word error rate was 8.4 percent. This figure is composed of 5.7 percent rejections and 2.7 percent actual mismatches.*

III. EXPERIMENTAL DESCRIPTION

The vocabulary selected for the present evaluation was designed to fulfill a similar function as that of the earlier one, namely to request flight information and to make reservations using an automated system with word-recognition capabilities. The difference is that in the present system the requests are made in the form of complete sentences rather than as one-word responses to queries. The 127-word vocabulary for this purpose is shown in Table I together with some sample sentence requests. The vocabulary contains many auxiliary and function-type words so that reasonably natural English sentences may be formed. The vocabulary includes 10 city names. In the earlier mode using the question-answer dialog, depending on the complexity of the task, a long series of questions may be necessary to specify a complete request. In the present mode,

* Therefore, the term "word error rate," as used in this paper, is more appropriately defined as the rate of nonrecognition, since it includes both outright errors (mismatches) and rejects.

Table I — 127-word vocabulary for requesting flight information and reservations and two sample sentences constructed from this vocabulary

1 Evening	33 To	65 Reservation	97 Card
2 Nine	34 Charge	66 A	98 Saturday
3 October	35 Make	67 Fare	99 Pay
4 Douglas	36 Home	68 BAC	100 By
5 DC	37 Five	69 Departure	101 Ten
6 Arrival	38 Does	70 Of	102 March
7 Seattle	39 Go	71 Meal	103 Cash
8 Eleven	40 Seat	72 Flights	104 Miami
9 Los Angeles	41 From	73 What	105 Thursday
10 Friday	42 Time	74 I	106 American
11 January	43 On	75 When	107 Plane
12 AM	44 December	76 Sunday	108 Eight
13 April	45 June	77 Boston	109 Club
14 May	46 Would	78 Arrive	110 Master
15 Morning	47 Some	79 Twelve	111 Office
16 Detroit	48 Many	80 Leave	112 My
17 Do	49 In	81 August	113 Class
18 New York	50 Please	82 For	114 Six
19 At	51 Will	83 November	115 Three
20 Tuesday	52 Lockheed	84 Philadelphia	116 Washington
21 Oh	53 Want	85 February	117 Night
22 Wednesday	54 Flight	86 Are	118 Phone
23 Need	55 Four	87 There	119 Area
24 Chicago	56 Depart	88 Return	120 Two
25 September	57 Repeat	89 Coach	121 Code
26 Is	58 Take	90 O'clock	122 Nonstop
27 PM	59 Number	91 How	123 Seats
28 Boeing	60 Denver	92 Much	124 Seven
29 Information	61 Diners	93 Served	125 Times
30 Afternoon	62 Prefer	94 Credit	126 Stops
31 Express	63 July	95 The	127 First
32 Like	64 Monday	96 One	

Sample test sentences:
"I would like some information please."
"I would like one first-class seat on flight number four four to Los Angeles on Saturday the oh one January."

the efficiency of natural English is approached by combining several commands in a single sentence input. The second sample test sentence in Table I is a good example. A more complete description of the task domain and the grammar is found in a companion paper by Levinson.[3]

Seven speakers—five male, two female—participated in the evaluation. The system programs resided in a Data General Nova 840 computer. Speech was input to the system via dialed-up telephone lines from an ordinary handset adjacent to the computer console. The speakers spoke their utterances after a prompt from the console. A display scope provided an intensity curve for their current input, together with end-point markers. The speakers had the option of repeating an utterance if they felt it was botched or corrupted by external noise disturbances.

Two sessions per speaker were devoted to establishing reference patterns. In each of these sessions speakers provided a single utterance

of each of the 127 words in the vocabulary. Reference patterns were computed from these utterances. Each speaker therefore ended up with a reference containing two distinct reference patterns for each word in the vocabulary. Speakers could also provide additional optional pronunciations for the articles "a" and "the."

Finally, each speaker provided one or more test sessions in which a total of 51 specified sentences were input as strings of isolated words, for a total of 444 words. There was thus an average of 8.7 words per sentence. Every word in the vocabulary and every production rule in the grammar were represented in the sentence set at least once. The word utterances composing the sentence strings for each speaker were stored on disk files.

Recognition was carried out off-line with acoustic recognition followed by parsed recognition accomplished by the syntax analyzer. For each test sentence the acoustic recognizer provided to the parser a matrix of distances or scores $[d_{ij}]$, $i = 1,2,\ldots,N$, $j = 1,2,\ldots,M$, where i represents the ith word in a sentence string of $N \leq 22$ words, and j represents the jth vocabulary item in the vocabulary of size $M = 127$ words. The smaller the score d_{ij}, the closer is the acoustic match for the ith word in the sentence to the jth word in the vocabulary. Recognition was carried out under four different experimental conditions. Two conditions were examined in which the utilized references were those for the designated speaker. In the first of these a fixed rejection threshold was imposed. In the second there was no rejection threshold. With a rejection threshold, an arbitrarily large number was assigned to the distance score for each rejected candidate. In the third experimental condition each speaker was compared against an arbitrarily selected reference. In the fourth condition each speaker was compared against a reference which was a composite of individual references from four arbitrarily selected male speakers.

IV. RESULTS

The overall results are shown in Table II as median error rates over the seven speakers.

"Word error—acoustic best candidate" refers to the rate at which the specified test word was not the best acoustic candidate. "Word error—acoustic five best candidates" refers to the rate at which the specified test word was not included among the five best acoustic candidates. The recognition scores for word error—acoustic best candidate are quite comparable to those obtained in the earlier evaluation. Given the larger size of the vocabulary, and especially the greater frequency of common, more easily confused words, the 11.7 percent word error performance*

* Compared to 8.4 percent for an initial trial in the earlier study with an 84-word vocabulary.

Table II — Median error rates over five speakers with 51 test sentences per speaker

	Condition: Reference:	1 Designated speaker	2 Designated speaker	3 Arbitrary male speaker*	4 Composite of four male speakers*
Word error	Rejection threshold:	Fixed	None	Fixed	Fixed
	Acoustic best candidate:	11.7%	10.8%	45.5%	34.9%
	Acoustic five best candidates:	1.8%	1.1%	20.0%	20.0%
	Parsed:	0.4%	1.6%	5.6%	6.5%
	Sentence error parsed:	3.9%	5.9%	35.3%	37.2%

* Scores include female speakers using male references.

seems reasonable. As anticipated, the syntactic constraints imposed by the task language have a powerful correcting influence on acoustic word errors. For example, for condition 1 the median number of word errors was reduced from 52 to 2 out of a total of 444.† Since a single word error creates a sentence error and since the number of sentences in the sample is relatively small, the parsed sentence error rate is not as reliable an indicator of the improvement gained by parsing as the parsed word error rate. It is interesting to note that although the acoustic word error rates are about the same, with or without a rejection threshold, the parsed word and sentence error rates are somewhat larger for the no-rejection-threshold condition. We attribute this result to the following possible situation. If a specified word in a sentence string is poorly recognized acoustically, in the condition with a rejection threshold it will have the same arbitrarily large distance score as other rejected candidates. Without a rejection threshold, however, the true word may have a score which is considerably worse than other candidates resulting in a greater chance of misleading the parser.

The recognition system was not designed to be speaker-independent. We did, however, try a naive experiment in that mode. Table II also shows the results of comparing all speakers against an arbitrary reference, condition 3, as well as a comparison of all speakers against a composite reference, condition 4. It was anticipated that, although acoustic recognition would be considerably poorer for these conditions than for the speaker-dependent condition, the parser would be able to compensate to some extent this poor performance, resulting in a reasonable overall recognition performance. This seems to be true with a 5 or 6 percent parsed word error rate and 35 or 37 percent parsed sentence error rate. Comparing speakers against an arbitrary reference does not seem significantly different from comparing them against a composite reference. The performance of the two female speakers in both these condi-

† I.e., a reduction of word error rate from 11.7 percent to 0.4 percent!

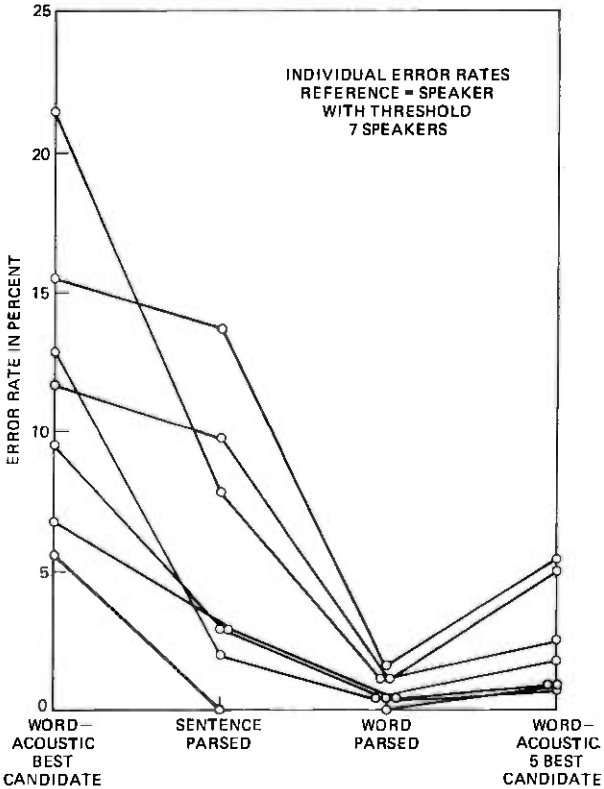


Fig. 1—Individual error rates for the speaker-dependent condition.

tions was considerably worse than that of the male speakers. The parsed word error rate for the women, for example, was approximately 30 percent.*

Individual error rates for the speaker-dependent condition are shown in Fig. 1. Most striking is the contraction of a large range of acoustic word error rates (best candidate) to a very tight range of parsed word error rates all below 2 percent. Parsed sentence error rates vary over a wide range and are sensitive functions of parsed word error rates. An additional indicator of acoustic word performance is acoustic word error rate, five best candidates. These rates occupy a considerably reduced and overall lower range than the standard acoustic word error rates. This measure may be a more reliable predictor of parsed error rates, as shown by the monotonic character of the lines that connect these individual

* The female speakers, therefore, contribute substantially to the error scores for conditions 3 and 4. This is not surprising in that the references for 3 and 4 were derived from male speakers only.

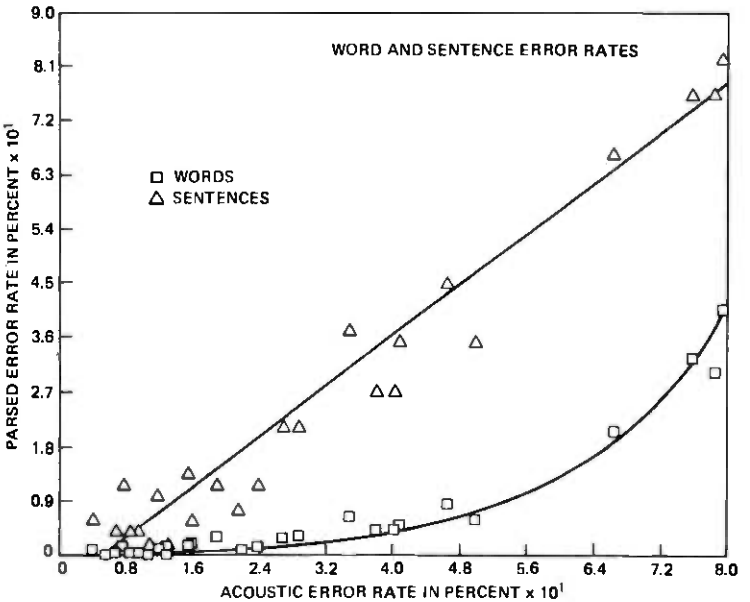


Fig. 2—Parsed error rate as a function of acoustic error rate for both words and sentences.

rates with the parsed rates. This seems reasonable since the five-candidate rate gives a good measure of the quality of the acoustic recognizer in the sense of indicating whether the true word has a good chance of having a low score.

Finally, individual parsed error rates collected from all speakers and conditions are plotted versus individual acoustic word error rates (best candidate) in Fig. 2 to characterize the effectiveness of the parser over the widest possible range of performance. This figure is analogous to the one in the companion paper by Levinson[3] which shows simulated results. The trends in both figures are the same, but the parsed error rates are significantly greater functions of acoustic error rate for the actual recognizer than for the simulation. The solid curves drawn have been fitted by eye. Note that although there is considerable scatter among the individual parsed sentence error rates the trend is almost linear. It is evident that even though the parser is a highly effective corrector of acoustic word errors this beneficial effect is neutralized to some degree by the highly sensitive dependence of sentence error on word error.

V. CONCLUSION

The command ensemble for an automatic word recognizer can be greatly expanded by forming complete sentences from a relatively modest word vocabulary. For applications where speaking discipline can

be exercised, complete sentences can be input on a word-by-word basis. The present study demonstrates that realistic syntactic constraints can dramatically compensate acoustic errors by the use of a well-constructed parser.

REFERENCES

1. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23*, 67–72, 1975.
2. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System Over Dialed-up Telephone Lines," talk presented at the 92nd meeting of the Acoustical Society of America, San Diego, November 1976.
3. S. E. Levinson, "The Effects of Syntactic Analysis on Word Recognition Accuracy," B.S.T.J., this issue, pp. 1627–1644.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

The Effects of Syntactic Analysis on Word Recognition Accuracy

By S. E. LEVINSON

(Manuscript received May 18, 1977)

In this paper we examine the effects of an algorithm for syntactic analysis on word recognition accuracy. The behavior of the algorithm is studied by means of a computer simulation. We describe the syntactic analysis technique, the problem domain to which it was applied, and the details of the simulation. We then present the results of the simulation and their implications. We find, for example, that an acoustic word error rate of 10 percent is reduced to 0.2 percent after syntactic analysis, resulting in a sentence error rate of 1 percent. These figures are based on a 127-word vocabulary and an average of 10.3 words per sentence for 1000 sentences. We expect that these results are indicative of the performance which will be attained by a real speech recognition system which uses the syntactic analysis algorithm described herein.

I. INTRODUCTION

The utility and flexibility of a speech recognition system can be substantially expanded if it can accept sentence length utterances rather than single words. Simultaneously, accuracy can be greatly improved by exploiting the grammatical constraints of language on the input sentences.[1,2,3]

The purpose of this investigation is to establish, by means of a computer simulation, how much improvement in reliability can be obtained by using a particular optimal method for syntactic analysis in conjunction with an isolated word recognition system.

This paper is in five sections. First we give a description of the method of syntax analysis under consideration. In the second section we describe the content and the semantic and grammatical structure of the problem domain to which we intend to apply the analysis. The third section is devoted to a description of the simulation, particularly of the acoustic recognizer and the procedure for generating random sentences. In the

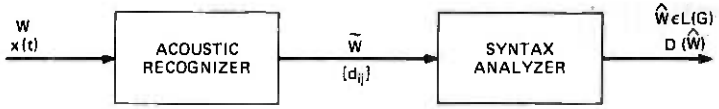


Fig. 1—Block diagram of the speech recognition system.

fourth section we present the results of the simulation. We conclude with an evaluation of the results and a brief discussion of directions for further investigations.

It should be said at the outset that the results of this experiment are encouraging. We find that, by using the grammatical constraints for our problem domain, the syntax analyzer reduces the acoustic error rate from 10 percent to 0.2 percent. This results in a sentence error rate of 1 percent. The sentences were composed from a 127-word vocabulary and contained an average of 10.3 words per sentence over 1000 randomly generated sentences.

We believe that these figures, with the more detailed results given in Section IV, are indicative of those which will be attained when the method of syntax analysis described herein is incorporated into a real speech recognition system.

II. AN OPTIMAL ALGORITHM FOR SYNTAX ANALYSIS

The type of speech recognition system we are evaluating is shown in Fig. 1, and its operation may be formally described as follows.

Let the language, L, be the subset of English used in a particular speech recognition task. Sentences in L are composed from the vocabulary, V, consisting of the M words $v_1, v_2, \ldots v_M$. Let W be an arbitrary sentence in the language. Then we write $W \in L$ and

$$W = w_1 w_2 \cdots w_k \tag{1}$$

where each w_i is a vocabulary word which we signify by writing $w_i \in V$ for $1 \leqslant i \leqslant k$. Clearly W contains k words, and we will often denote this by writing $|W| = k$. Similarly the number of sentences in L will be denoted by $|L|$.

The sentence W of eq. (1) is encoded in the speech signal $x(t)$ and input to the acoustic recognizer from which is obtained the probably corrupted string

$$\tilde{W} = \tilde{w}_1 \tilde{w}_2 \cdots \tilde{w}_k \tag{2}$$

where $\tilde{w}_i \in V$ for $1 \leqslant i \leqslant k$ but $\tilde{W}$ is not, in general, a sentence in L.

The acoustic recognizer also produces the matrix $[d_{ij}]$ whose ijth entry, d_{ij}, is the distance, as measured by some metric in an appropriate pattern space, from the ith word, $\tilde{w}_i$ to the prototype for the jth vocabulary word, v_j, for $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant M$.

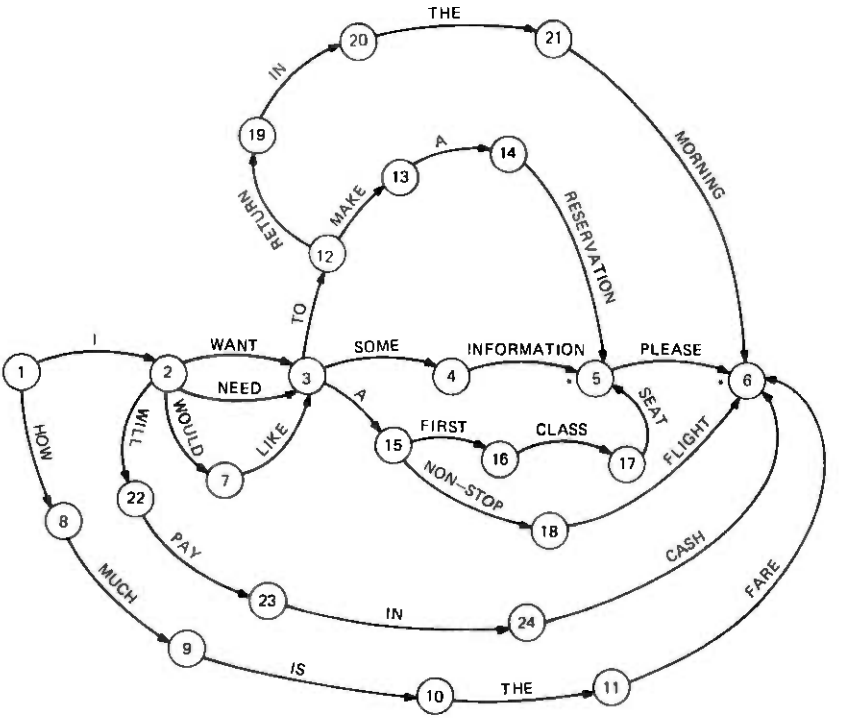


Fig. 2—Example state transition diagram.

The syntax analyzer then produces the string

$$\hat{W} = \hat{w}_1 \hat{w}_2 \cdots \hat{w}_k \tag{3}$$

for which the total distance, $D(\hat{W})$, given by

$$D(\hat{W}) = \sum_{i=1}^{k} d_{ij_i} \; 1 \leqslant j_i \leqslant M \tag{4}$$

is minimized subject to the constraint that $\hat{W} \in L$. Thus the syntax analysis is optimal in the sense of minimum distance.

Since, in general, $\hat{W} \notin L$, whereas W was assumed to be grammatically well-formed (i.e. $W \in L$), the process should correct word recognition errors.

In principle one could minimize the objective function of eq. (4) by computing $D(W) \; \forall \; W \in L$ and choosing the smallest value. In practice, when $|L|$ is large, this is impossible. One must perform the optimization efficiently. It has been shown by Lipton and Synder[4] that for a particular class of languages one can minimize $D(W)$ in time proportional to $|W|$. In fact one can optimize any reasonable objective function in time linear in the length of the input.

The particular class of languages for which the efficiency can be attained is called the class of Regular languages. For the purposes of this discussion we shall define the class of Regular languages as that class for which each member language can be represented by an abstract graph called a state transition diagram.

A state transition diagram consists of a finite set of vertices or states, Q, and a set of edges or transitions connecting the states. Each such edge is labeled with some $v_i \in V$. The exact manner of the interconnection of states is specified symbolically by a transition function, δ, where

$$\delta:(Q \times V) \rightarrow Q \tag{5}$$

That is, if a state $q_i \in Q$ is connected to another state $q_j \in Q$ by an edge labeled $v_m \in V$ then

$$\delta(q_i,v_m) = q_j \tag{6}$$

We also define a set of accepting states, $Z \subset Q$, which has the significance that a string $W = w_1w_2 \cdots w_k$, where $w_i \in V$ for $1 \leqslant i \leqslant k$, is a well-formed sentence in the language, L, represented by the state transition diagram if and only if there is a path starting at q_1 and terminating in some $q_j \in Z$ whose edges are labeled, in order, $w_1,w_2, \cdots w_k$.

Alternatively we may write $W \in L$ iff

$$\begin{cases} \delta_1(q_1,w_1) = q_{j_1} \\ \delta_2(q_j,w_2) = q_{j_2} \\ \quad\vdots \\ \delta_k(q_{j_{k-1}},w_k) = q_{j_k} \in Z \end{cases} \tag{7}$$

We may then define the language, L, as the set of all W satisfying eq. (7). An example of these concepts is shown in Fig. 2. The accepting states are marked by asterisks.

While the definition given above of a Regular language is mathematically rigorous, it is not the standard one used in the literature on formal language theory but rather has been specifically tailored to the notational requirements of this paper. The interested reader is urged to refer to Hopcroft and Ullman[10] for a standard and complete introduction to formal language theory.

In the following discussion we shall restrict ourselves to finite Regular languages, i.e., those for which $|L|$ is finite. This restriction in no way alters the theory but its practical importance will become obvious in what follows. The finiteness of the language implies that its state transition diagram has no circuits, i.e., no paths of any length starting and ending at the same state. Thus there is some maximum sentence length which we shall denote, $l_{\max}$.

We now turn to the problem of efficiently solving the minimization problem of eq. (4). To do this we shall define two data structures Φ and Ψ which will be used to store the estimates of $D(\hat{W})$ and $\hat{W}$, respectively.

The first stage of the algorithm is the initialization procedure in which we set

$$\Phi_i(q) = \begin{cases} 0 & \text{for } q = q_0 \text{ and } i = 0 \\ \infty & \text{otherwise} \end{cases}$$

$$\Psi_i(q) = 0 \qquad 1 \leqslant i \leqslant |W| = k;\ \forall q \in Q \tag{8}$$

The data structures have two indices. The subscript is the position of the word in the sentence and the argument in parentheses refers to the state so that the storage required for each array is, at most, the product of $l_{\max} + 1$ and $|Q|$, the number of states in the set Q.

After initialization we utilize a dynamic programming technique defined by the following recursion relations:

$$\Phi_i(q) = \min_{\Delta} \{\Phi_{i-1}(q_p) + d_{ij}\} \tag{9}$$

where the set Δ is given by:

$$\Delta = \{\delta(q_p, v_j) = q\} \tag{10}$$

Then

$$\Psi_i(q) = \Psi_{i-1}(q_p)\hat{w}_i \tag{11}$$

where $\hat{w}_i$ is just the v_j which minimizes $\Phi_i(q)$. Equation (11) is understood to mean that the word $\hat{w}_i$ is simply appended to to the string $\Psi_{i-1}(q_p)$.

Unfortunately the concatenation operation is not easily implemented on general purpose computers so we change the recursion of eq. (11) by making Ψ into a linked list structure of the form:

$$\Psi_{1i}(q) = q_p$$

$$\Psi_{2i}(q) = \hat{w}_i \tag{12}$$

Then when $i = k$ we can trace back through the linked list of eq. (12) and construct the sentence $\hat{W}$ as follows: First find $q_f \in Z$ such that

$$\Phi_k(q_f) = \min_{q \in Z} \{\Phi_k(q)\} \tag{13}$$

set $q = q_f$ and then for $i = k, k-1, k-2, \ldots, 1$

$$\hat{w}_i = \Psi_{2i}(q)$$

$$q = \Psi_{1i}(q) \tag{14}$$

Table I — Example $[d_{ij}]$ matrix

Code	Vocabulary word	$I = 1$	$I = 2$	$I = 3$	$I = 4$	$I = 5$
1	Is	9	9	1	8	4
2	Fare	2	2	5	3	1
3	I	7	3	2	3	2
4	Want	2	9	4	7	3
5	Would	1	5	4	8	2
6	Like	2	5	2	6	5
7	Some	2	1	9	8	3
8	Information	7	7	4	7	8
9	Please	2	3	2	4	9
10	To	4	5	8	1	7
11	Make	6	3	9	8	5
12	A	4	7	6	9	8
13	Reservation	3	6	7	8	9
14	Return	9	7	6	4	8
15	The	8	6	5	2	3
16	Morning	3	4	5	6	7
17	First	8	6	8	7	5
18	Class	5	5	4	3	9
19	Seat	9	9	8	7	3
20	Non-stop	3	3	4	5	8
21	Flight	9	8	3	5	6
22	Will	6	7	7	6	5
23	Pay	4	4	4	4	3
24	In	3	3	3	6	9
25	Cash	5	4	3	7	6
26	How	2	9	8	7	5
27	Much	6	2	8	4	9
28	Need	7	6	5	4	3

Thus the sentence $\hat{W}$ is computed from right to left.

The operation of the above algorithm is illustrated in Tables I, II, and III. Table I shows the vocabulary words of the language diagrammed in Fig. 2 along with numerical codes and a sample $[d_{ij}]$ matrix. Table II shows the details of the operation of the algorithm for $i = 0,1,2$. Table III shows the results after the sentence has been completely analyzed.

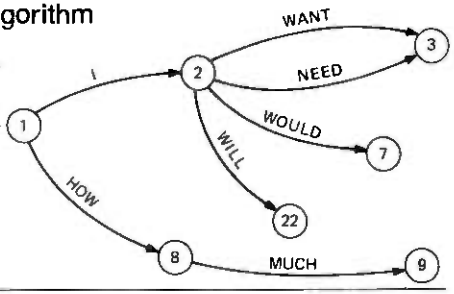
By locating the smallest entry in each column of the sample $[d_{ij}]$ matrix of Table I, it can be seen that the acoustic transcription of the sentence from which this matrix was produced is: WOULD SOME IS TO FARE. Clearly this is not a valid English sentence nor is there any path through the state transition diagram of Fig. 2 whose edges are so labeled.

Following Table II the reader can trace the operation of the algorithm as it computes the valid sentence having the smallest total distance. First the Φ and Ψ arrays are initialized according to eq. (8). To make the figure easier to read, this has been shown only for $i = 0$.

Note that there are two transitions from state 1; one to state 2 labeled I and the other to state 8 labeled HOW. Accordingly $\Phi_1(2)$ is set to 7, the metric for I; $\Psi_{11}(2)$ is set to 1, the state at the beginning of the transition and Ψ_{21} is set to 3, the code for the transition label I. Similarly $\Phi_1(8)$ is

Table II — Detailed operation of the first three stages of the algorithm

Code	Word	Position 1	Position 2
3	I	7	3
4	Want	2	9
28	Need	7	6
5	Would	1	5
22	Will	6	7
26	How	2	9
27	Much	6	2



$i\backslash q$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
ϕ_0	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞		∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
ψ_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ψ_{20}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ϕ_1		7						2																
ψ_{11}		1						1																
ψ_{21}		3						26																
ϕ_2			13				12		4													14		
ψ_{12}			2				2		8													2		
ψ_{22}			28				5		27													22		

Table III — Final results of the algorithm

$i\backslash q$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	22	23	24
ϕ_0	0	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
ψ_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ψ_{20}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ϕ_1		7						2															
ψ_{11}		1						1															
ψ_{21}		3						26															
ϕ_2			13				12		4												14		
ψ_{12}			2				2		8												2		
ψ_{22}			28				5		27												22		
ϕ_3			14	22						5		21			19							18	
ψ_{13}			7	3						9		3			3							22	
ψ_{23}			6	7						1		10			12							23	
ϕ_4				22	29						7	15	29		23	26		24	25				24
ψ_{14}				3	4						10	3	12		3	15		15	12				23
ψ_{24}				7	8						15	10	11		12	17		20	14				24
ϕ_5					30	8							20	37		28	35	31	23	34			
ψ_{15}					4	11							12	13		15	16	15	12	19			
ψ_{25}					8	2							11	12		17	18	20	14	24			

set to 2, the metric for HOW; $\Psi_{11}(8)$ is set to 1 as before and $\Psi_{21}(8)$ is set to 26, the code for HOW. All other entries remain unchanged.

In the next stage more transitions become possible. Note in particular that there are two possible transitions from state 2 to state 3. In accordance with eq. (9), the one labeled NEED is chosen since it results in the smallest total distance, 13, which is entered in $\Phi_2(3)$; $\Psi_{12}(3)$ is set to 2, the previous state, and $\Psi_{22}(3)$ is set to 28, the code for NEED. Transitions to state 7, 9, and 22 are also permissible and thus these columns are filled in according to the same procedure.

The completion of this phase of the algorithm results in Φ and Ψ as shown in Table III. We can now trace back according to eqs. (13) and (14) to find $\hat{W}$. From Fig. 2 we see that there are two accepting states, 5 and 6. $\Phi_5(6)$ is 8 which is less than 30, the value of $\Phi_5(5)$, so we start tracing back from state 6. The optimal state sequence is, in reverse order, 6,11,10,9,8,1. The corresponding word codes which, when reversed, decode to the sentence: HOW MUCH IS THE FARE.

One final note: from the operation of the algorithm it should be clear that it is not necessary to retain $\Phi_i(q)$ for $0 \leqslant i \leqslant |W|$. At the ith stage one needs only $\Phi_{i-1}(q)$ to compute $\Phi_i(q)$. Thus the storage requirements are nearly halved in the actual implementation.

In closing we should note that this scheme is formally the same as (though conceptually different from) the Viterbi[5] algorithm and similar to methods used by Baker[6] and stochastic parsing techniques discussed in Fu[7] and Paz.[8] The crucial difference is that in the cited references, estimates of transitions probabilities are used whereas in this method the transitions are deterministic and the probabilities used are only those conditioned on the input $x(t)$.

III. THE SEMANTIC AND GRAMMATICAL STRUCTURE OF THE RECOGNITION TASK

In this section we shall discuss the application of the abstractions of the previous section to a particular speech recognition problem domain. The task which was finally selected was that of an airline information and reservation system. The choice was made for three reasons. First, the problem is difficult enough so that even under some artificial constraints, it is a significant test of the above described techniques. Second, previous work by Rosenberg and Itakura[9] which used single words rather than sentences composed of isolated words as input was available for purposes of comparison. Third, it affords the opportunity to add modes of human/machine communication such as speaker verification and voice response.

The semantics of the language we designed limits a user to the following types of messages. First, one may state the desired kind of transaction (i.e., requesting flight information or making a reservation). Then, one may make a reservation either by providing all necessary information in one sentence or by giving answers to questions as required. The user may select arrival and/or departure dates, times and cities, number of stops, number and class of seats, specific flight numbers and aircraft types. Alternatively, the user may ask questions about arrival and/or departure dates, times and locations of specific flights, the type of aircraft, the number of stops, the fare, the number of meals served and the flight time. Finally, he may request a repeat of any information or supply telephone numbers and method of payment.

For most of the above messages there are several acceptable grammatical structures of sentences conveying the same semantic information.

In order to limit the complexity of the syntactic analysis, certain arbitrary constraints were imposed:

(*i*) Dates consist of two digits followed by the name of the month.

(*ii*) Flight numbers are limited to one or two digits.

(*iii*) The vocabulary includes the names of only ten cities.

(*iv*) The length of the longest sentences is 22 words.

It was felt that these constraints could all be relaxed if so desired without making major system modifications.

Next we give an informal specification of the syntactic structure. In this description phrases enclosed in curly brackets are alternatives. Those enclosed in square brackets are optional, while those within angle brackets represent a class of words of the indicated type.

$$\text{I} \left\{ \begin{matrix} \text{WANT} \\ \text{WOULD LIKE} \end{matrix} \right\} \left\{ \begin{matrix} \text{SOME INFORMATION} \\ \text{TO MAKE A RESERVATION} \end{matrix} \right\} [\text{please}]$$

$$\text{I} \left\{ \begin{matrix} \text{WANT} \\ \text{WOULD LIKE} \end{matrix} \right\} \text{TO} \left\{ \begin{matrix} \text{GO} \\ \text{LEAVE} \\ \text{RETURN} \\ \text{DEPART} \end{matrix} \right\} [\text{FROM } \langle \text{city} \rangle][\text{TO } \langle \text{city} \rangle]$$

$$[\text{ON}][\langle \text{day} \rangle] \left[\left\{ \begin{matrix} \text{MORNING} \\ \text{AFTERNOON} \\ \text{EVENING} \\ \text{NIGHT} \end{matrix} \right\} \right] [\text{THE } \langle \text{date} \rangle]$$

$$\left\{ \begin{matrix} \text{AT WHAT TIME} \\ \text{WHEN} \end{matrix} \right\} \text{DO FLIGHTS LEAVE } [\langle \text{city} \rangle] \text{ FOR } \langle \text{city} \rangle$$

$$\text{HOW MANY FLIGHTS} \left\{ \begin{matrix} \text{ARE THERE} \\ \text{GO} \end{matrix} \right\} [\text{FROM } \langle \text{city} \rangle]$$

$$\text{TO } \langle \text{city} \rangle \text{ ON [day]} \left[\left\{ \begin{matrix} \text{MORNING} \\ \text{AFTERNOON} \\ \text{EVENING} \\ \text{NIGHT} \end{matrix} \right\} \right] [\text{THE } \langle \text{date} \rangle]$$

$$\text{WHAT PLANE IS ON FLIGHT } \langle \text{flightnumber} \rangle \text{ [TO } \langle \text{city} \rangle]$$

$$\left\{ \begin{matrix} \text{WHAT} \\ \text{HOW MUCH} \end{matrix} \right\} \text{IS THE FARE [FROM } \langle \text{city} \rangle] \text{ [TO } \langle \text{city} \rangle]$$

$$\text{IS A MEAL SERVED ON [THE] FLIGHT } [\langle \text{flightnumber} \rangle][\text{TO } \langle \text{city} \rangle]$$

I { WANT / WOULD LIKE / WILL TAKE } FLIGHT [NUMBER] ⟨flightnumber⟩

[TO ⟨city⟩] [ON ⟨day⟩ { MORNING / AFTERNOON / NIGHT / EVENING }] [THE ⟨date⟩]

I { WANT / NEED / WOULD LIKE } ⟨digit⟩ [{ FIRST CLASS / COACH }] SEAT [s].

I PREFER THE [⟨manufacturer⟩] ⟨aircraft type⟩.

I { WANT / WOULD LIKE } TO GO AT ⟨hour⟩ { a.m. / p.m. / O'CLOCK }

PLEASE REPEAT THE { ⟨flightnumber⟩ / FARE / ARRIVAL TIME [s] / DEPARTURE / PLANE / NUMBER OF MEALS / FLIGHTS }

{ AT WHAT TIME / WHEN } DOES FLIGHT ⟨flightnumber⟩

[{ FROM / TO } ⟨city⟩] { ARRIVE / DEPART }

I WILL PAY BY { CASH / AMERICAN EXPRESS / DINERS CLUB / MASTER CHARGE }

MY { HOME / OFFICE } PHONE [NUMBER] IS [AREA CODE]
[⟨area code number⟩] ⟨phone number⟩

I { WANT / WOULD LIKE } A NON-STOP FLIGHT

HOW MANY STOPS ARE THERE ON FLIGHT ⟨flight number⟩

$$\left[\left\{\begin{matrix}\text{FROM}\\ \text{TO}\end{matrix}\right\}\ \langle\text{city}\rangle\right]\left[\text{ON}\ \langle\text{day}\rangle\ \left\{\begin{matrix}\text{MORNING}\\ \text{AFTERNOON}\\ \text{EVENING}\\ \text{NIGHT}\end{matrix}\right\}\right]\ [\text{THE}\ \langle\text{DATE}\rangle]$$

WHAT IS THE FLIGHT TIME [FROM ⟨city⟩] [TO ⟨city⟩]

The above description is too informal to define every detail of the Flight Information Language. It should, however, give the reader a feeling for the basic syntax and semantics. This specification of the language is quite useless for the purpose of the syntax analysis algorithm. For that purpose we have produced a formal specification of the language, the state transition diagram for which is shown in Fig. 3. From this graph it may be seen that $|V| = 127$; $|\delta| = 450$; $|Q| = 144$ and $|Z| = 21$ with the accepting states being designated by asterisk.

IV. DETAILS OF THE SIMULATION

The simulation of the speech recognition system based upon the analysis described above may be treated as three separate problems. They are: (*i*) Random generation of many well-formed sentences, (*ii*) computing a $[d_{ij}]$ matrix for each in such a way that the number of acoustic errors resulting from a nearest-neighbor decision rule is controllable, and (*iii*) syntactically analyzing the sentences and tabulating the appropriate statistics automatically. We shall now discuss these in order.

The method for generating random sentences in the language is just the following algorithm:

(*i*) $W \leftarrow O; q_i \leftarrow q_1$ (W gets the null string) at the ith stage,

(*ii*) Chose a word, $w_{k_i} \in V$

(*iii*) If $\delta(q_i, w_{k_i}) \neq q_j$ for some j go to (2) else $W \leftarrow W w_{k_i}$ (W gets itself concatenated with W_{k_i}) $q_i \leftarrow q_j$

(*iv*) If $q \in Z$ and $\rho < \theta$ where ρ is a pseudorandom number and θ is some threshold, STOP; else go to (2)

When the procedure terminates, $W = w_{k_1} w_{k_2} \ldots w_{k_l}$; $l \leq l\max$. Obviously by changing the threshold, θ, one can vary the average length of the sentences produced. In the actual simulation, ρ was uniformly distributed on $(\frac{1}{2}, -\frac{1}{2})$ and θ was set to -0.25 producing an average sentence length of 10.3 words.

In all, 42,000 such sentences were generated. In addition we made up

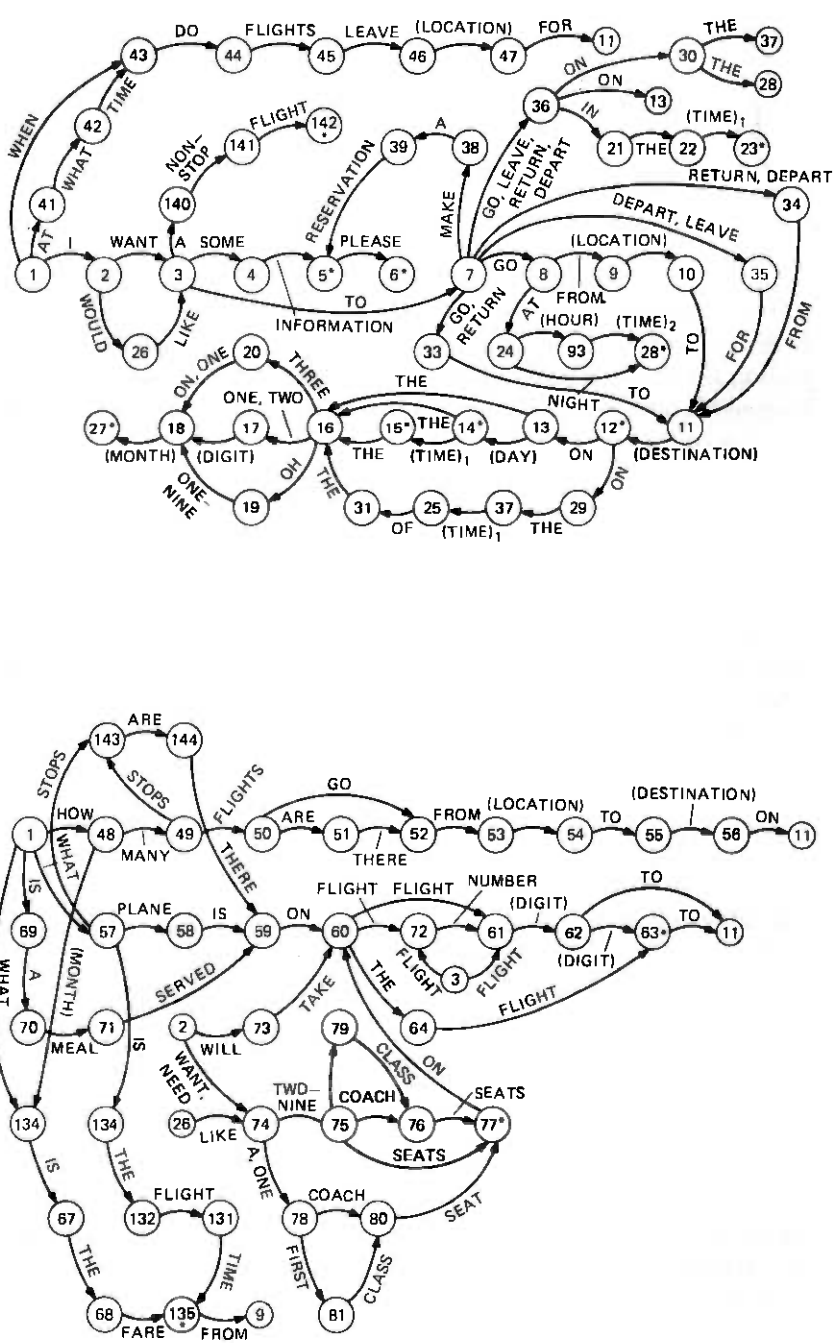


Fig. 3—State transition diagram of the Flight Information and Reservation Language.

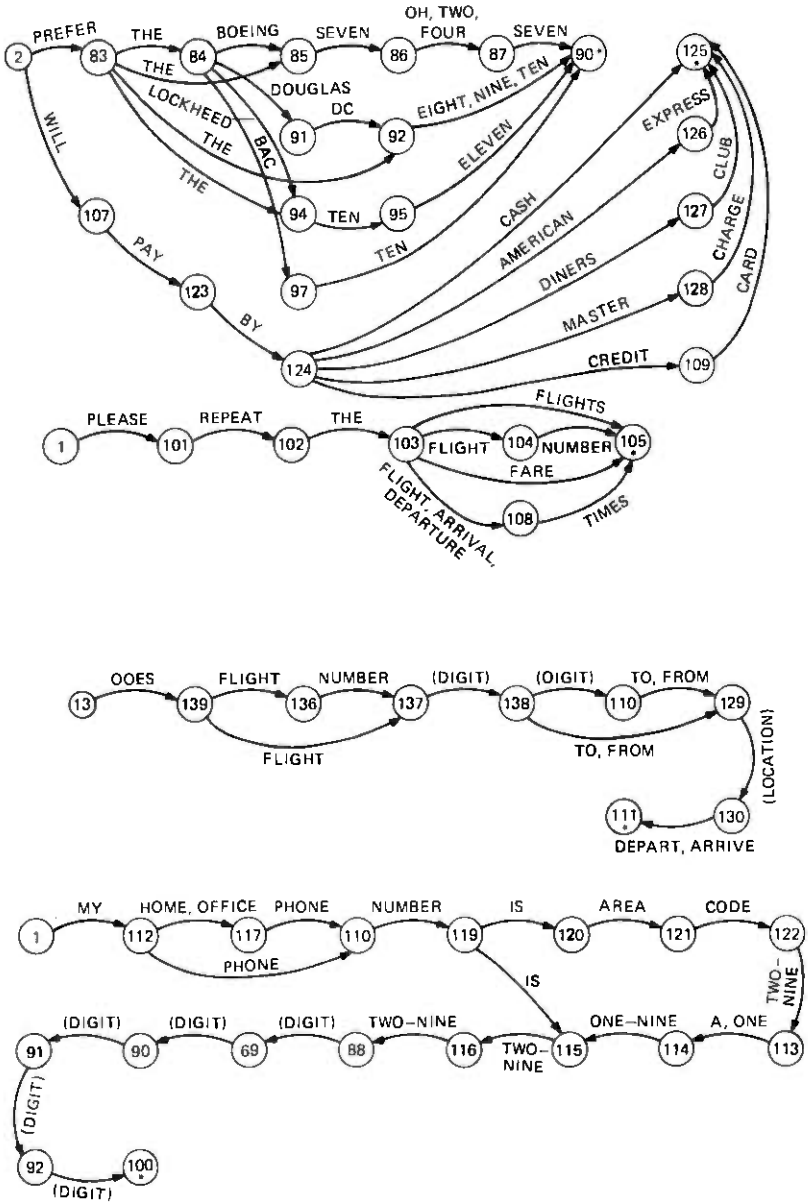


Fig. 3 (*continued*)

an additional 171 sentences averaging 9.7 words in length to use as a check against the randomly generated sentences. The additional sentences comprised several realistic transactions between airline customers and an automated flight information and reservation system.

The procedure for generating a $[d_{ij}]$ matrix simulating one which might be produced by acoustic recognition for the randomly selected sentence $W = w_1 w_2 \cdots w_k$ is as follows. Corresponding to each $v_j \in V$ we assign the five-dimensional Gaussian density function

$$p_j(\vec{x}) = (2\pi)^{-5/2}|U|^{-1/2} e^{-1/2(\vec{x}-\vec{m}_j)U^{-1}(\vec{x}-\vec{m}_j)^T} \tag{15}$$

where the covariance matrix, U, was chosen, for convenience, to be

$$U = \begin{bmatrix} \sigma^2 & & & & \\ & \sigma^2 & & 0 & \\ & 0 & \sigma^2 & & \\ & & & \sigma^2 & \\ & & & & \sigma^2 \end{bmatrix} \tag{16}$$

for selected values of σ^2. The mean vectors

$$\vec{m}_j = (m_{1j}, m_{2j}, m_{3j}, m_{4j}, m_{5j})$$

were fixed by selecting the m_{ij} at random from

$$m_{ij} = \begin{cases} 2 \\ 1 \\ 0 \end{cases} \quad \text{for } 1 \leq l \leq 5;\ 1 \leq j \leq |V| \tag{17}$$

until the 127 mean vectors were defined. Then for $w_i = v_j$, a random vector $\vec{y}$ was drawn from $p_j(\vec{x})$ and distances were computed according to:

$$d_{ij} = ||\vec{y} - \vec{m}_j|| \text{ for } 1 \leq j \leq |V| \tag{18}$$

where the norm is the simple Euclidean distance. Equation (18) was evaluated for $1 \leq i \leq k$ thus all entries in the $[d_{ij}]$ matrix were computed for each sentence W.

The acoustic recognition was simulated by a nearest neighbor rule so that $\bar{w}_i = v_j$ if

$$d_{ij} \leq d_{in} \text{ for } 1 \leq n \leq |V| \tag{19}$$

Again, eq. (19) was applied for $1 \leq i \leq k$ and ties were arbitrarily broken. Clearly by changing the value of σ^2 in eqs. (15) and (16) the simulated acoustic error rate can be varied with small values of σ^2 producing low error rates.

Given the foregoing discussion, description of the simulation is quite simple. A set of 1000 random sentences was generated and its distance matrices computed. The sentences were syntactically analyzed and errors counted. This was done for $0.05 \leq \sigma \leq 2.1$ with σ being incremented by 0.05 for each set of 1000 sentences.

In addition, the specially formulated 171 sentences were typed in and processed. In this case σ was fixed at a value of 0.245 which resulted in

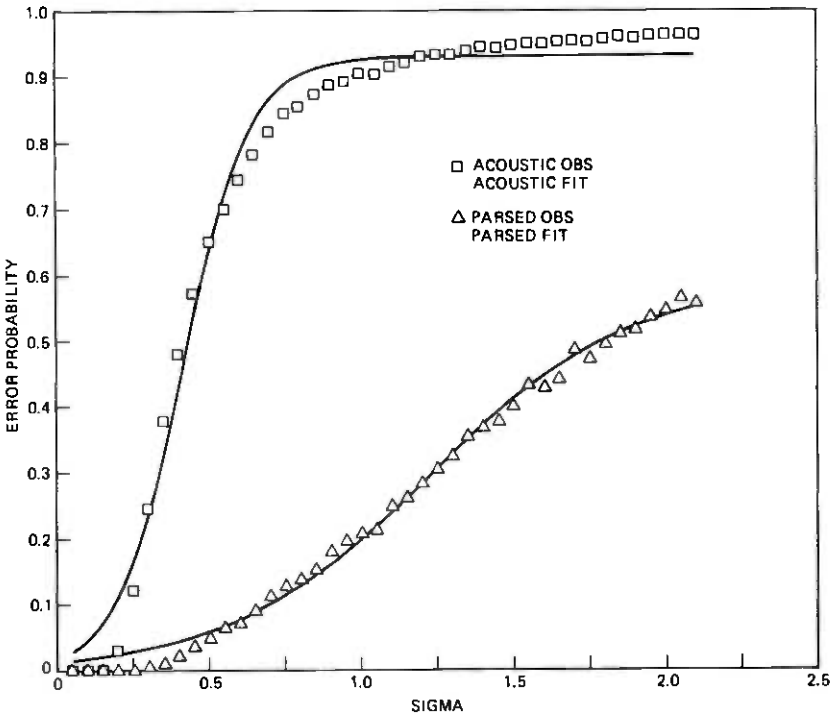


Fig. 4—Error rates as a function of σ.

an acoustic error rate of 11 percent which is close to the value observed by Rosenberg and Itakura[9] for a similar word recognition task. Error rates were measured for these sentences as well.

V. SIMULATION RESULTS

The overall results of the simulation are encouraging, showing considerable improvement in word and sentence recognition accuracy. Given an acoustic error rate of 10 percent on 1000 randomly generated sentences averaging 10.3 words per sentence, syntactic analysis reduces the word error rate to 0.2 percent resulting in a sentence error rate of 1 percent. A sentence is in error if even one word is improperly classified. For the test set of 171 sentences containing 1662 words, an 11 percent acoustic word error rate was lowered to 0.2 percent after syntactic analysis resulting in a 1.2 percent sentence error rate. The actual time required to analyze a 22-word sentence on the Data General Nova 840 is a small fraction of a second.

Details of the results are best given in the accompanying figures. Figure 4 is a plot of acoustic and syntactic word error probabilities as a

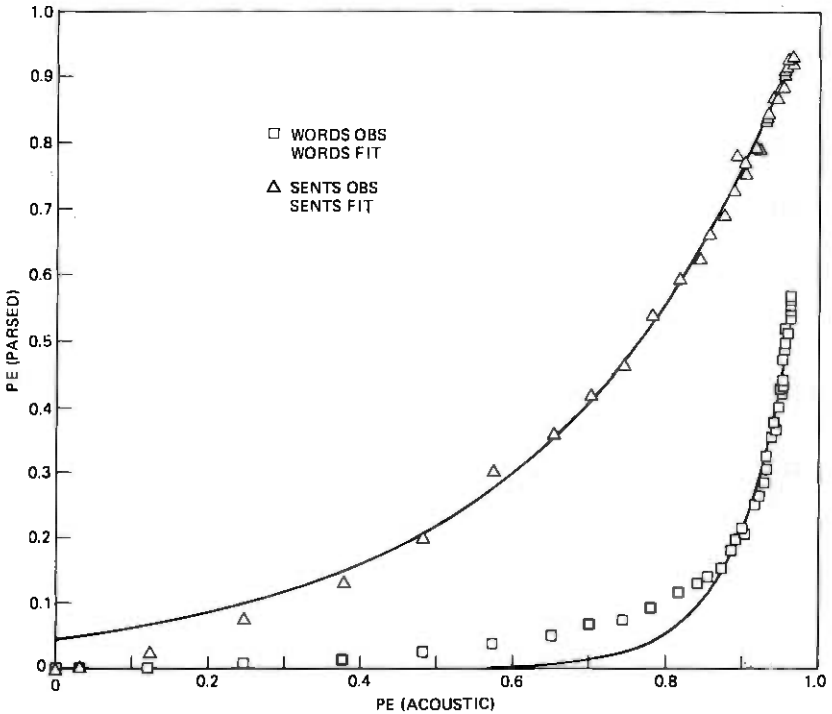


Fig. 5—Word and sentence error rates as a function of acoustic error rate.

function of σ of eqs. (15) and (16). Each point marked with a symbol is an observed data point based on the results obtained from 1000 sentences. It should be noted that the results for each point are based upon different sentences. The solid lines are obtained from a nonlinear least-squares fit of the data to the equation

$$P_e = \frac{\alpha_1\alpha_2}{\alpha_3\alpha_2 + (\alpha_1 - \alpha_3\alpha_2)e^{-\alpha_1\sigma}}. \tag{20}$$

The method used in fitting the data is described by McCalla.[11] The curve of eq. (20) is called a logistic curve; its significance is discussed in detail by Braun.[12] For the simulated data the standard deviation of the actual data from the fitted curve was <0.005.

Figure 5 shows the word and sentence error probabilities as a function of the acoustic word error probability. Once again the marked points are derived from sets of 1000 randomly generated sentences while the solid curves are obtained by fitting the data to an exponential curve of the form

$$P_{ep} = \alpha_1 e^{\alpha_2 P_e} \tag{21}$$

Once again the resulting fits were good having a standard deviation of <0.004.

VI. CONCLUSIONS

It is clear from the results of the simulation that the syntax analysis algorithm is very fast and effective in eliminating word recognition errors which occur at the acoustic level. It is expected that the performance of the algorithm for real speech input will depend on the characteristics of the acoustic recognizer and the task language. However, we believe that our results are indicative of the performance which can be attained in real speech recognition systems.

There are several areas for further research which are immediately suggested by this work. Although this entire presentation has been oriented toward the grammatical structure of sentences, the method described is certainly not restricted to that area. For example, the phonemic structure of words can be specified by a formal language as can the composition of acoustic features into phenomes. We therefore feel that optimal syntax analysis methods will be useful in more difficult speech recognition tasks than the one described here.

Another useful extension of the technique would be achieved by retaining the same optimality criterion while relaxing the restriction that $|\tilde{W}| = |\hat{W}|$. In other words, the algorithm would be allowed to insert and delete words. This could be an important aid in the solution of the segmentation problem in continuous speech.

On the theoretical side, it would be enlightening to derive analytical expressions for the average probability of error for the syntax analyzer, given the properties of the language and a characterization of the acoustic recognizer. Perhaps for this purpose the entropy or redundancy of the language might be sufficient, while the acoustic recognizer might be viewed as a noisy channel and characterized by its equivocation or capacity. In any event, it seems obvious that an information theoretic analysis would provide insights into the behavior of speech recognition systems.

Finally we note that Regular languages are the most simple syntactic structures. One naturally wonders whether efficient, optimal methods exist for formal languages of much greater complexity which would be better models of Natural Language.

In summary, we may say that optimal syntactic analysis techniques are useful and powerful tools to be used in tractable speech recognition tasks as well as being interesting mathematical objects.

REFERENCES

1. S. E. Levinson, "An Artificial Intelligence Approach to Automatic Speech Recognition," Proc. IEEE Conf. on Systems, Man and Cybernetics, Boston, November, 1973.

2. S. E. Levinson, "The VOCAL Speech Understanding System," Proc. 4th IJCAI, Tbilisi, U.S.S.R., September 1975.
3. S. E. Levinson, "Cybernetics and Automatic Speech Understanding," Proc. IEEE ICISS, Patras, Greece, August 1976.
4. R. J. Lipton and L. Snyder, "On the Optimal Parsing of Speech," Yale Univ. Dept. of Comp. Sci. Res. Report No. 37, New Haven, Conn., October 1974.
5. A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimal Algorithm," IEEE Trans. on Inf. Theor., *IT-13*, March 1967.
6. J. K. Baker, *The DRAGON System: An Overview," IEEE Trans. on Acoust., Speech, and Signal Processing, ASSP-23*, February 1975.
7. K. S. Fu, *Syntactic Pattern Recognition*, New York: Academic Press, 1974.
8. A. Paz, *Introduction to Probabilastic Automata*, New York: Academic Press, 1971.
9. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System over Dialed-up Telephone Lines," J.A.S.A., *60*, Supp. 1, Fall 1976.
10. J. E. Hopcroft and J. D. Ullman, *Formal Languages and Their Relations to Automata*, Reading, Mass.: Addison-Wesley, 1969.
11. T. R. McCalla, *Introduction to Numerical Methods and FORTRAN Programming*, New York: Wiley, 1967.
12. M. Braun, *Differential Equations and their Applications as an Introduction to Applied Mathematics*, New York: Springer Verlag, 1975.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

A Combinatorial Lemma and Its Application to Concentrating Trees of Discrete-Time Queues

By J. A. MORRISON

(Manuscript received November 11, 1977)

Concentrating rooted tree networks of discrete-time single server queues, all with unit service time, are considered. Such networks occur as subnetworks connecting remote access terminals to a node in a data communications network. It is shown that the network of queues may be replaced by a single queue, with prescribed input, which has the same output as the queue at the root of the tree. The result is applied, in particular, to the case of several queues in tandem, and it is shown how this problem may be reduced to that of just two queues in tandem. The latter problem was analyzed earlier by the author.

I. INTRODUCTION

In this paper we consider concentrating rooted tree networks of discrete-time single server queues, all with unit service time. Such networks occur as subnetworks connecting remote access terminals to a node in a data communications network.[1] Our purpose is to show that the rooted tree network of queues may be replaced by a single queue, with prescribed input, which has the same output as the queue at the root of the tree. In particular, the result is applied to the case of queues in tandem.

In Section II we consider the pooling of data from M buffers into a single buffer, which also receives data from another source, as depicted in Fig. 1. We establish a combinatorial lemma which shows that there is a single equivalent buffer, with prescribed input, and the same output as the buffer in which the data is pooled. It is then pointed out how this result may be applied to a concentrating rooted tree network of queues, such as the one depicted in Fig. 3. A related observation was made by Kaspi and Rubinovitch[2] in connection with networks of continuous time queues involving the pooling of data from inputs with idle periods that are exponentially distributed.

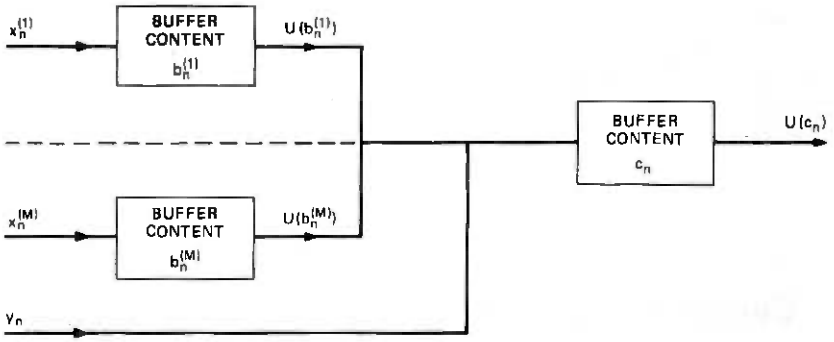


Fig. 1—Schematic of pooling of data from several buffers into a single buffer.

In Section III we consider the repeated application of the lemma to the case of several queues in tandem, as depicted in Fig. 4. It is shown how this problem may be reduced to that of just two queues in tandem, so that the results obtained for that problem[3] may be applied. Specifically, in the case that the input processes $z_n^{(i)}$, $i = 1, \ldots, I$, are mutually independent, and each process is a sequence of independent identically distributed nonnegative integer valued random variables, the generating function of the steady state distribution of the content of each buffer in Fig. 4 may be determined. Also, under the assumption that all arrivals take place at the end of a unit time interval, the average waiting time in each queue may be obtained.

II. COMBINATORIAL LEMMA

We first consider the pooling of data from M buffers into a single buffer, which also receives data from another source, as depicted in Fig. 1. It is assumed that a buffer transmits one packet, the basic unit of data, in a unit time interval, provided that it is not empty, and that the buffers are of unlimited size. Let $b_n^{(j)}$, $j = 1, \ldots, M$, denote the contents of the M buffers at time n, and let $x_n^{(j)}$ denote the corresponding number of packets entering the buffers in the time interval $(n, n+1]$. We define

$$U(\ell) = \begin{cases} 1, & \ell = 1,2,\ldots, \\ 0, & \ell = 0. \end{cases} \tag{1}$$

Then the contents of the buffers at time $(n+1)$ are given by the equations

$$b_{n+1}^{(j)} = b_n^{(j)} - U(b_n^{(j)}) + x_n^{(j)}, \quad j = 1, \ldots, M, \tag{2}$$

for $n = 0,1,2,\ldots$. It is assumed that the initial contents $b_0^{(j)}$, as well as the inputs $x_n^{(j)}$, are nonnegative integers.

The outputs of the M buffers enter another buffer, the content of

which at time n is denoted by c_n. Also, the number of packets entering this other buffer in the time interval $(n,n+1]$ from another source is denoted by y_n. Then the content of this buffer at time $(n+1)$ is given by the equation

$$c_{n+1} = c_n - U(c_n) + \sum_{j=1}^{M} U(b_n^{(j)}) + y_n, \tag{3}$$

for $n = 0,1,2,\ldots$. It is assumed that the initial content c_0, as well as the inputs y_n, are nonnegative integers. We now show that there is a single equivalent buffer, with prescribed input, which has the same output.

Let e_n denote the content of the equivalent buffer at time n, and define

$$\begin{aligned} e_0 &= c_0, \\ e_n &= \sum_{j=1}^{M} [b_n^{(j)} - x_{n-1}^{(j)}] + c_n, \quad n = 1,2,\ldots. \end{aligned} \tag{4}$$

Further, we define the inputs

$$\begin{aligned} w_0 &= \sum_{j=1}^{M} b_0^{(j)} + y_0, \\ w_n &= \sum_{j=1}^{M} x_{n-1}^{(j)} + y_n, \quad n = 1,2,\ldots. \end{aligned} \tag{5}$$

Then we have the following

Lemma 1. Subject to (1)–(5), e_n is a nonnegative integer, and

$$\begin{aligned} U(e_n) &= U(c_n), \\ e_{n+1} &= e_n - U(e_n) + w_n, \quad n = 0,1,2,\ldots. \end{aligned} \tag{6}$$

Proof: It follows from (2)–(4) that

$$e_{n+1} = \sum_{j=1}^{M} b_n^{(j)} + c_n - U(c_n) + y_n, \quad n = 0,1,2,\ldots. \tag{7}$$

But, from (1), $\ell - U(\ell) \geq 0$. Hence e_{n+1} is a nonnegative integer for $n = 0,1,2,\ldots$, and so is $e_0 = c_0$, by assumption. Moreover, $e_{n+1} = 0$ implies that $b_n^{(j)} = 0$, $j = 1,\ldots,M$, $c_n = U(c_n)$ and $y_n = 0$, and hence, from (3), that $c_{n+1} = 0$. On the other hand, $c_{n+1} = 0$ also implies that $b_n^{(j)} = 0$, $j = 1,\ldots,M$, $c_n = U(c_n)$ and $y_n = 0$, and hence, from (7), that $e_{n+1} = 0$. Therefore $U(e_{n+1}) = U(c_{n+1})$, $n = 0,1,2,\ldots$, and $U(e_0) = U(c_0)$ since $e_0 = c_0$. Finally, from (7), with the help of (4) and (5),

$$e_{n+1} = e_n - U(c_n) + w_n, \quad n = 0,1,2,\ldots. \tag{8}$$

Since we have just shown that $U(e_n) = U(c_n)$, $n = 0,1,2,\ldots$, this completes the proof of the lemma.

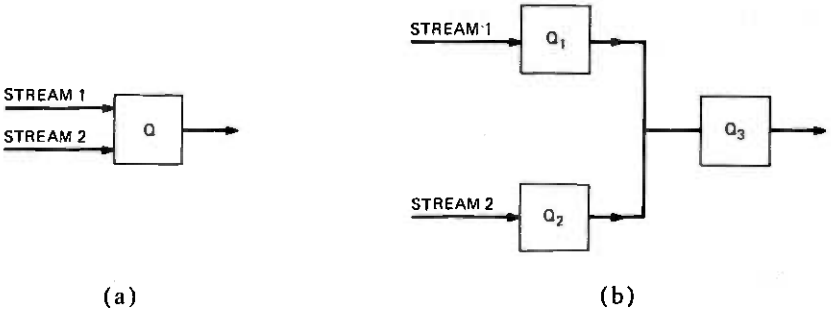


Fig. 2—(a) Queue, Q, fed by two independent input streams, and (b) Queue, Q_3, fed by the outputs of two queues, Q_1 and Q_2, which are fed separately by the two input streams.

In the particular case $M = 2$ and $y_n \equiv 0$, Ziegler and Schilling[4] obtained a related result, not restricted to discrete-time queues. They considered single server queues with identical constant service times, but assumed that the interarrival times between packets for each of the two independent input streams were governed by some general probability distribution. They compared a queue, Q, fed directly by the two input streams, and a queue, Q_3, fed by the outputs of two queues, Q_1 and Q_2, which are fed separately by the two input streams, as depicted in Fig. 2a and b. They established that the number of packets serviced at Q during its jth busy period is equal to the number serviced at Q_3 during its jth busy period, and hence that the jth idle periods at Q and Q_3 have the same duration. Note that in the discrete-time case we have shown that $U(e_n) = U(c_n)$, so that the corresponding buffers are empty at the same times.

Returning to our lemma, the result may be applied to concentrating rooted tree networks of discrete-time single server queues with unit service time, such as the network depicted in Fig. 3. The queues Q_1, Q_2 and Q_3 may be replaced by a single equivalent queue, $\hat{Q}_3$ say, which has a prescribed input sequence, $\hat{z}_n^{(3)}$ say, and the same output as Q_3. Then, by a second application of the lemma, the queues $\hat{Q}_3$, Q_4, Q_5 and Q_6 may be replaced by a single equivalent queue, $\hat{Q}_6$ say, which has a prescribed input sequence, $\hat{z}_n^{(6)}$ say, and the same output as Q_6. Thus the rooted tree network of Fig. 3 may be replaced by a single queue with the same output and prescribed input. In the next section we consider the repeated application of the lemma to several queues in tandem.

III. TANDEM QUEUES

We now consider I discrete-time single server queues, with unit service times, in tandem, as depicted in Fig. 4. The output of buffer i enters buffer $i + 1$, for $i = 1, \ldots, I - 1$. Let $d_n^{(i)}$, $i = 1, \ldots, I$, denote the content

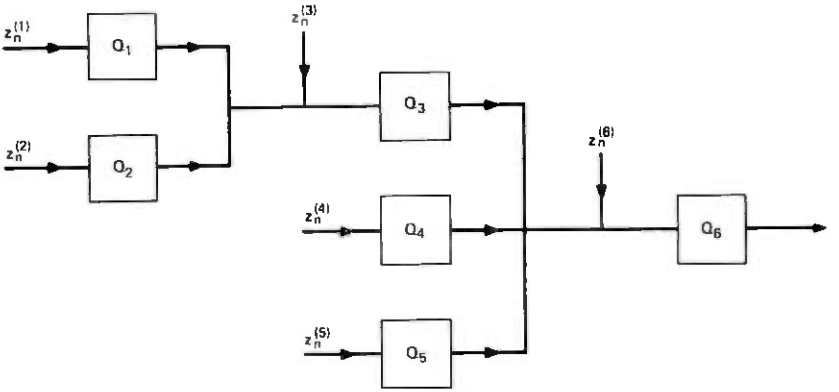


Fig. 3—Example of a concentrating rooted tree network of queues.

of buffer i at time n, and let $z_n^{(i)}$ denote the corresponding number of packets entering the buffer from a source in the time interval $(n,n+1]$. For convenience, we define $d_n^{(0)} \equiv 0$. Then the content of buffer i at time $n+1$ is given by the equation

$$d_{n+1}^{(i)} = d_n^{(i)} - U(d_n^{(i)}) + U(d_n^{(i-1)}) + z_n^{(i)}, \tag{9}$$

for $n = 0,1,2,\ldots$, and $i = 1,\ldots,I$. It is assumed that the initial contents $d_0^{(i)}$, as well as the inputs $z_n^{(i)}$, are nonnegative integers.

Let

$$e_0^{(i+1)} = d_0^{(i+1)}, \quad i = 1,\ldots,I-1, \tag{10}$$

and

$$e_n^{(1)} = d_n^{(1)}, \quad n = 0,1,2,\ldots. \tag{11}$$

Moreover, define

$$v_n^{(i)} = \begin{cases} \sum_{k=1}^{i} z_{n-i+k}^{(k)}, & n = i-1,i,\ldots, \\ d_0^{(i-n-1)} + \sum_{k=i-n}^{i} z_{n-i+k}^{(k)}, & n = 0,\ldots,i-2, \end{cases} \tag{12}$$

for $i = 1,\ldots,I$, and let

$$e_n^{(i+1)} = e_n^{(i)} + d_n^{(i+1)} - v_{n-1}^{(i)}, \tag{13}$$

for $n = 1,2,\ldots$, and $i = 1,\ldots,I-1$. Then we have the following

Lemma 2. Subject to (9)–(13), $e_n^{(i)}$ is a nonnegative integer, and

$$\begin{aligned} U(e_n^{(i)}) &= U(d_n^{(i)}), \\ e_{n+1}^{(i)} &= e_n^{(i)} - U(e_n^{(i)}) + v_n^{(i)}, \end{aligned} \tag{14}$$

for $n = 0,1,2,\ldots,$ *and* $i = 1,\ldots,I.$

Proof: Since $d_n^{(0)} \equiv 0$, it follows by definition, from (9), (11), and (12), that the lemma holds for $i = 1$. We proceed by induction on i. We assume that the lemma holds for some $i < I$, and will show that it holds for $i + 1$. We identify $e_n^{(i)}$, $d_n^{(i+1)}$, $e_n^{(i+1)}$, $v_n^{(i)}$ and $z_n^{(i+1)}$ with $b_n^{(1)}$, c_n, e_n, $x_n^{(1)}$ and y_n, respectively, for $n = 0,1,2,\ldots$. The induction hypothesis then implies (2), with $M = 1$, and, from (9),

$$d_{n+1}^{(i+1)} = d_n^{(i+1)} - U(d_n^{(i+1)}) + U(e_n^{(i)}) + z_n^{(i+1)}, \tag{15}$$

and hence (3), with $M = 1$. Moreover, (10) and (13) imply that (4) holds, with $M = 1$. Hence, from Lemma 1, with $M = 1$, it follows that $e_n^{(i+1)}$ is a nonnegative integer, and

$$U(e_n^{(i+1)}) = U(d_n^{(i+1)}), \quad n = 0,1,2,\ldots. \tag{16}$$

Also, using (5),

$$e_1^{(i+1)} = e_0^{(i+1)} - U(e_0^{(i+1)}) + e_0^{(i)} + z_0^{(i+1)}, \tag{17}$$

and

$$e_{n+1}^{(i+1)} = e_n^{(i+1)} - U(e_n^{(i+1)}) + v_{n-1}^{(i)} + z_n^{(i+1)}, \tag{18}$$

for $n = 1,2,\ldots$.

But, from (10)–(12),

$$e_0^{(i)} + z_0^{(i+1)} = d_0^{(i)} + z_0^{(i+1)} = v_0^{(i+1)}. \tag{19}$$

Also, for $i \geq 2$ and $n = 1,\ldots,i-1$,

$$v_{n-1}^{(i)} + z_n^{(i+1)} = d_0^{(i-n)} + \sum_{k=i-n+1}^{i+1} z_{n-1-i+k}^{(k)} = v_n^{(i+1)}. \tag{20}$$

Finally, for $n = i, i+1,\ldots,$

$$v_{n-1}^{(i)} + z_n^{(i+1)} = \sum_{k=1}^{i+1} z_{n-1-i+k}^{(k)} = v_n^{(i+1)}. \tag{21}$$

Hence, from (17)–(21),

$$e_{n+1}^{(i+1)} = e_n^{(i+1)} - U(e_n^{(i+1)}) + v_n^{(i+1)}, \tag{22}$$

for $n = 0,1,2,\ldots$. In view of (16), this completes the proof by induction.

From (9) and (14) we have the following

Corollary. For $n = 0,1,2,\ldots,$ *and* $i = 1,\ldots,I-1,$

$$\begin{aligned} e_{n+1}^{(i)} &= e_n^{(i)} - U(e_n^{(i)}) + v_n^{(i)}, \\ d_{n+1}^{(i+1)} &= d_n^{(i+1)} - U(d_n^{(i+1)}) + U(e_n^{(i)}) + z_n^{(i+1)}. \end{aligned} \tag{23}$$

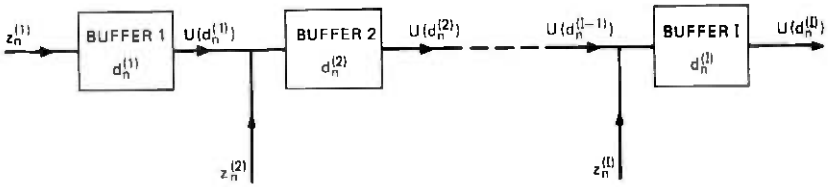


Fig. 4—Schematic of several queues in tandem.

Thus, by replacing the first i queues in Fig. 4 by a single equivalent queue, with the same output as the ith queue, we have reduced the problem of several queues in tandem to that of just two in tandem.

Suppose now that the input processes $z_n^{(i)}$, $i = 1, \ldots, I$, are mutually independent, and that each is independently and identically distributed (i.i.d.), with

$$E(s^{z_n^{(i)}}) = \phi_i(s), \quad i = 1, \ldots, I. \tag{24}$$

Then, from (12),

$$E(s^{v_n^{(i)}}) = \prod_{k=1}^{i} \phi_k(s), \quad n = i - 1, i, \ldots, \tag{25}$$

and the input processes $v_n^{(i)}$ and $z_n^{(i+1)}$ are mutually independent, and each is i.i.d. The problem of two queues in tandem was investigated recently,[3] and the results are applicable to (23). The generating function of the steady state distribution of the contents of the two buffers was calculated, under the assumption that the mean combined input rate from the two sources is less than unity. Accordingly, we assume that

$$\sum_{i=1}^{I} E(z_n^{(i)}) < 1. \tag{26}$$

Then we may use (23) to calculate the generating function of the steady state distribution of the content of each buffer in Fig. 4. The initial values $v_0^{(i)}, \ldots, v_{i-2}^{(i)}$, for $i \geq 2$, do not affect the steady state distributions.

A particular example was considered,[3] in which the input to the first queue is geometrically distributed, while the input from the source into the second queue is either 0 or 1, with fixed probabilities. The steady state probability that the content of the second buffer exceeds m was calculated, and asymptotic results were derived for $m \gg 1$. It would be of interest to carry out an analogous derivation for the case of Poisson inputs to both queues. The results would be applicable to the case of Poisson inputs into I queues in tandem, corresponding to $\phi_i(s) = \exp[\lambda_i(s - 1)]$ in (24). Then, from (25), the input process $v_n^{(i)}$ is also Poisson, for $n = i - 1, i, \ldots$, with parameter $\Sigma_{k=1}^{i} \lambda_k$.

Formulas were derived[3] for the average waiting times in two queues in tandem, under the assumption that all arrivals take place at the end

of a unit time interval. The average waiting time in the second queue was taken over all arrivals to that queue, both from the source and from the first queue. The results may be applied to (23), to obtain the average waiting times in each of the I queues in Fig. 4. The averages are over all arrivals to each queue.

REFERENCES

1. L. Kleinrock, *Queueing Systems, Volume II: Computer Applications*, New York: Wiley, 1976, p. 292.
2. H. Kaspi and M. Rubinovitch, "The Stochastic Behavior of a Buffer with Non-Identical Input Lines," Stoch. Proc. and Their Appl., *3* (1975), pp. 73–88.
3. J. A. Morrison, "Two Discrete-Time Queues in Tandem," IEEE Trans. Commun., to be published.
4. C. Ziegler and D. L. Schilling, "Delay Decomposition at a Single Server Queue with Constant Service Time and Multiple Inputs," IEEE Conference Record, Vol. I, International Conference on Communications, Chicago, Illinois, June 12–15, 1977, pp. 284–287; IEEE Trans. Commun., *COM-26*, No. 2 (February 1978), pp. 290–295.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Pulse Dispersion Properties of Fibers with Various Material Constituents

By L. G. COHEN, F. V. DiMARCELLO, J. W. FLEMING, W. G. FRENCH, J. R. SIMPSON, and E. WEISZMANN

(Manuscript received November 11, 1977)

Intermodal dispersion properties are compared for high silica fibers with borosilicate (B_2O_3-SiO_2) and germania borosilicate (GeO_2-B_2O_3-SiO_2) graded-index profiles. Pulse transmission measurements were systematically correlated with profile shapes so that new fibers could be fabricated with closer-to-optimal profile gradients at a wavelength of 907.5 nanometers. Germania borosilicate fibers with power law profile exponents ($\alpha \approx 2.03$) lowered intermodal dispersion 50 times from the result expected for comparable step-index fibers with N.A. ≈ 0.19. By contrast, borosilicate fibers with $\alpha \approx 1.78$ caused a 100-fold pulse width reduction in fibers with N.A. ≈ 0.14, corresponding to a $2\sigma = 0.13$ ns/km pulse-broadening rate.

I. INTRODUCTION

Dispersive refractive index differences between material constituents (GeO_2, B_2O_3, and SiO_2) for germania borosilicate core fibers cause modal group velocity differences to depend on the source wavelength (profile dispersion). Therefore, nonparabolic profile gradients are generally required to minimize pulse dispersion.

Previous time-domain transmission measurements[1,2] were used to direct the fabrication of a nearly optimal borosilicate fiber. This paper compares more recent and extensive data for germania borosilicate fibers and borosilicate fibers at $\lambda = 907.5$ nm wavelength corresponding to a GaAs injection laser. We have also characterized small profile undulations about a nearly optimal power law shape which degrade intermodal dispersion characteristics. Another paper[3] describes how profile dispersion influences fiber bandwidth properties at other wavelengths.

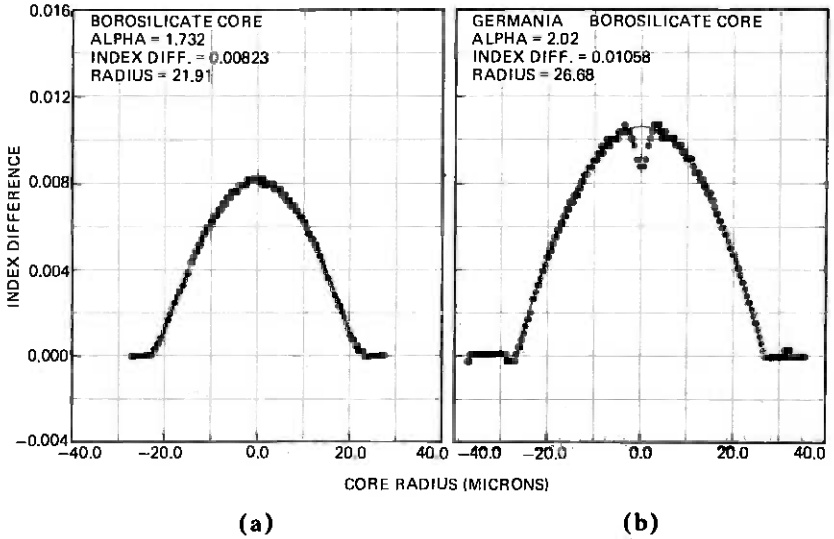


Fig. 1—Refractive-index profiles derived from interference micrographs: (a) a borosilicate graded-core fiber; (b) a germania borosilicate core fiber. The fitted curves are determined from a weighted [eq. (1)] nonlinear least-squares fit to the data ⊕.

II. FIBER FABRICATION AND PROFILE EVALUATION

The borosilicate fibers in this study have a uniform 1 B_2O_3–6 SiO_2 cladding composition and a core in which the B_2O_3 concentration decreases from 14 mole percent to 0 percent at the center.[4] Germania borosilicate fibers have a uniform SiO_2 cladding, a thin (2 μm) 1 B_2O_3–9 SiO_2 barrier layer and a core in which the GeO_2 concentration increases from 0 percent to 8.5 mole percent at the center where the material composition is 2 GeO_2–1 B_2O_3–21 SiO_2. The modified chemical vapor deposition process[5] is used to deposit the appropriate glass compositions by the reaction of BCl_3, $GeCl_4$, and $SiCl_4$ with oxygen at a temperature of 1400–1700°C inside a fused quartz substrate tube. In the case of the borosilicate fibers, the borosilicate cladding is first deposited followed by the graded borosilicate core. The germania borosilicate fibers are typically prepared by depositing 2 borosilicate barrier layers followed by 40–50 graded core layers. The fused quartz support tube is the cladding in this case. The radial index profile is graded by a programmed variation of the chloride dopant concentrations in the reaction stream. After the composite substrate tube is collapsed into a solid preform structure and then drawn into fiber, the resultant index profiles are determined by interference microscopy of thin fiber cross sections.[6,7]

Figure 1b shows a profile for a typical germania borosilicate fiber. The dip in the center is caused by GeO_2 dopant burn-off during the collapse stage. This type of distortion does not appear in boron graded fibers (Fig.

1a) because the dopant concentration is very small at the core center. The dip at either edge of the germania borosilicate profile corresponds to the borosilicate layer at the core-cladding interface. The index profile parameter, α, is determined by fitting power law profiles to the measured curve. The least-mean-square fit profile is determined from weighted differences between the measured and optimal curves according to:

$$\text{Weighted \% dev.} = \left[\sum_{i=1}^{j} \frac{(\Delta N_{\text{meas}} - \Delta N_{\text{power law}})^2}{j} W(r_i)\right]^{1/2} \times \frac{100}{\Delta N_{\max}} \tag{1}$$

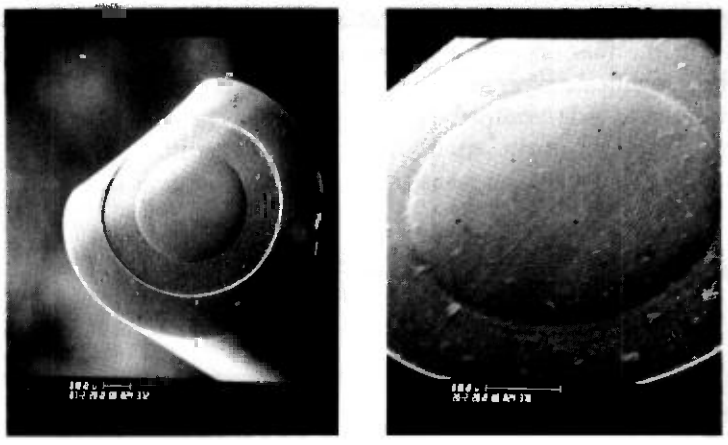
where

$$W(r_i) = r_i^2 (1 - r_i^2)^{3/2}$$
$$\Delta N_{\text{power law}} = N_{\max} (1 - r_i^{\alpha}) . \tag{2}$$

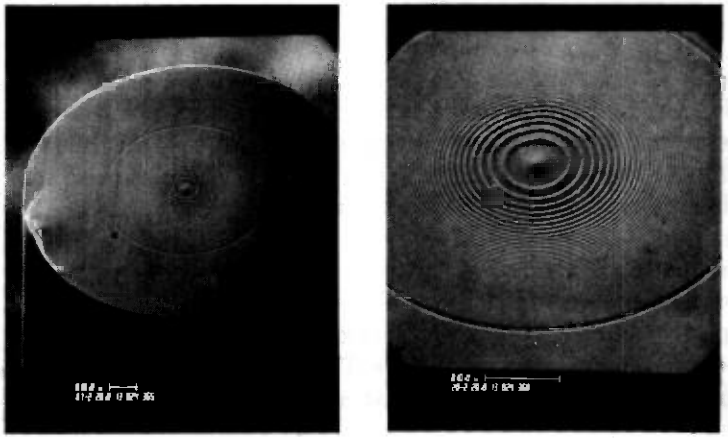
The ΔN's are index differences and r_i is the distance from the core center divided by the core radius. The weighting function $W(r_1)$ [eq. (2)] was determined[8] by computing the approximate number of modes that are confined to various sections of the profile. This reduces the importance of the core center, $r_i = 0$, since only the lowest-order modes are confined there and of the profile tail, $r_1 = \pm 1$, which is only important to high-order modes. Maximum weight is placed on the radial region midway between the core and cladding. After obtaining a best fit α, using the weighting function $W(r)$, we determine the unweighted deviations of the data from the calculated profile. Typical unweighted deviations range from 1 to 3 percent from the least-mean-square fit power law profile. However, recent theoretical calculations[9] have shown that 1 percent profile deviations can cause order-of-magnitude pulsewidth increases from the optimal 2σ (min).

Profile distortions can also be illustrated through scanning electron photomicrographs of etched fiber cross sections as in Fig. 2. The ridged structure midway between core and cladding is in the vicinity of the maximum gradient slope. This ridged distortion is observed in both boron and germania doped fibers, but it appears more prominent in the germania doped fibers due to the high doping in the center where the layers are farther apart. These distortions are due to differences in volatility of the glass components which cause concentration variations in each layer deposited by MCVD.

The concentration of core dopants may be measured directly by the use of electron-beam x-ray microanalysis techniques.[10] An ETEC scanning electron microscope equipped with a KEVEX energy dispersive x-ray spectrometer has been used to measure the germania concentration in



(a) B_2O_3 DOPED FIBERS



(b) GeO_2 DOPED FIBERS

Fig. 2—Scanning electron photomicrographs of etched fiber cross sections illustrate dopant concentration profiles for a boron graded profile and a germania graded profile.

germania borosilicate core fibers as shown in Fig. 3. The high resolution of this technique generates data which more accurately represent the central germania depletion region than the index profile data (Fig. 1b) measured using thin section interference microscopy. The ridged structure seen in the germania borosilicate fiber is coincident with regions of GeO_2 concentration fluctuation. The germania concentration profile characterizes a fiber core in the same way as the refractive index profile, provided the effect of the boron is either independent of radius or negligible. Fitting the weighted power law function to the concentration profile shown in Fig. 3, one obtains an α of 2.12.

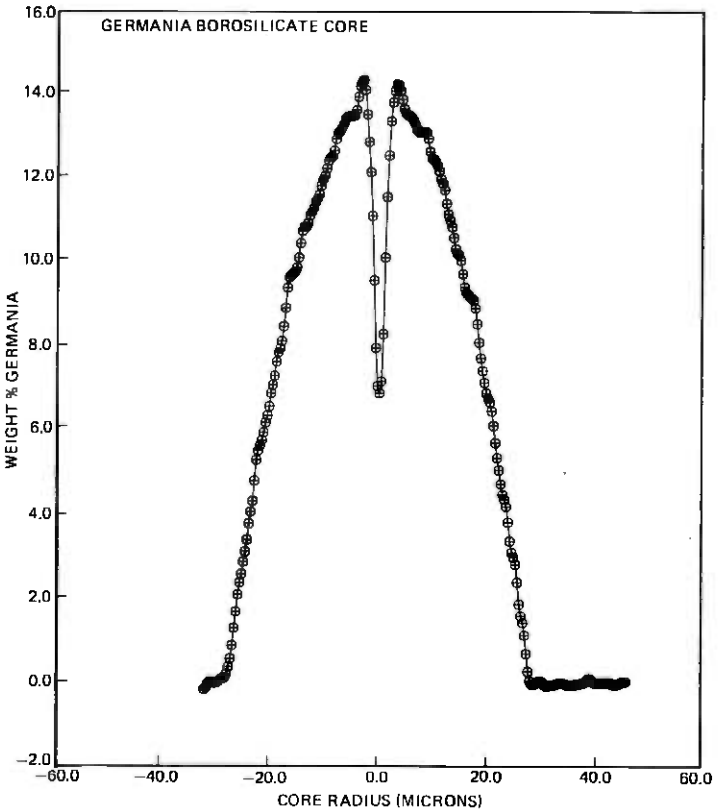


Fig. 3—Germania composition profile for a typical germania borosilicate fiber (the refractive index profile for this same fiber is shown in Fig. 1b).

III. PULSE TRANSMISSION MEASUREMENTS AT λ = 907.5 nm

Pulse dispersion is characterized at $\lambda = 907.5$ nm by injecting impulses of light (3 dB width = 0.3 nsec, full rms pulsewidth $2\sigma = 0.4$ nsec) from a GaAs laser and measuring the broadened fiber output pulsewidth. The optical shuttle pulse technique[2,11] is used to make length-dependent pulsewidth measurements by reflecting propagating light back and forth between partially transparent mirrors at the ends of a fiber. Mode-mixing effects were relatively small in all the tested fibers since pulsewidths increased with an almost linear dependence for multi-kilometer path lengths. Therefore, output pulse broadening was primarily caused by intermodal dispersion, profile dispersion, and material dispersion effects due to relative time delays between the source spectral components within its 2.5–3 nm bandwidth.

Material dispersion effects were reduced by a narrowband interference filter which has a 1.4 nm bandwidth approximately centered about the

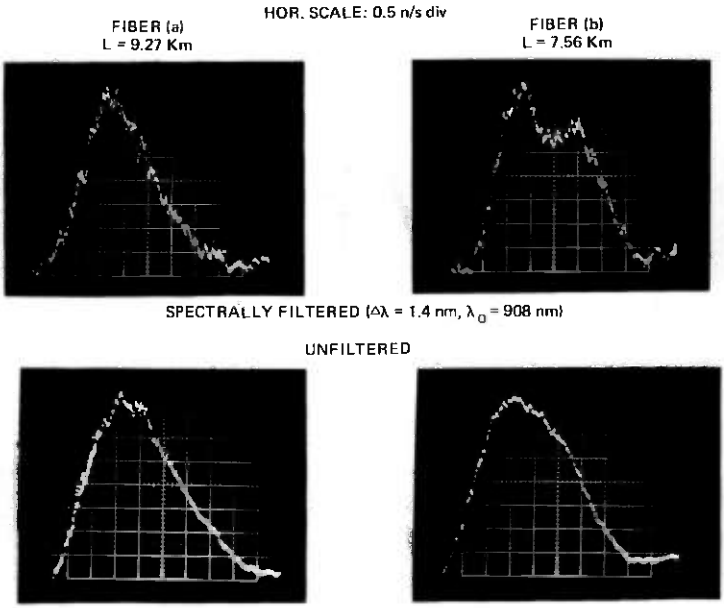


Fig. 4—Material dispersion effects on output pulsewidths from borosilicate fibers (a) and (b). The bottom row of photographs represents pulse propagation of unfiltered GaAs laser light, and the top row represents the propagation of spectrally filtered laser light with bandwidth $\Delta\lambda \approx 1.4$ nm centered about $\lambda_0 \approx 907.5$ nm.

laser line peak at $\lambda = 907.5$ nm. The filtered source spectral bandwidth[12,13] should cause $2\sigma = 0.09$ nsec/km full rms width pulse spreading in borosilicate fibers and 0.11 nsec/km pulse spreading in germania borosilicate fibers. These kinds of effects are clearly illustrated in Fig. 4 for shuttle pulse extrapolated lengths of 9.27 km for fiber (a) ($L = 1.03$ km) and 7.56 km for fiber (b) ($L = 1.08$ km). The bottom row of photographs shows output pulses due to unfiltered laser light, and the top row of photographs shows pulses due to spectrally filtered light. Results for fiber (a) show that when the source spectral bandwidth is cut in half, the output pulsewidth is reduced by 20 percent. Pulse outputs from fiber (b) show how material dispersion effects mask intermodal effects by smoothing the impulse response. A multipeak pulse structure is recovered by narrowing the laser linewidth.

Far-field spatial filters[2] (circles and annular rings) are used to measure time-of-flight differences between high- and low-order modes arriving at the fiber output. If high-order modes arrive before low-order modes, the fiber profile is overcompensated because $\alpha < \alpha(\text{opt})$. When high-order modes arrive last, the profile is undercompensated because $\alpha > \alpha(\text{opt})$. Output pulses from fibers with nearly optimal profiles are not altered in shape by spatial ray filters or changed launch conditions. This

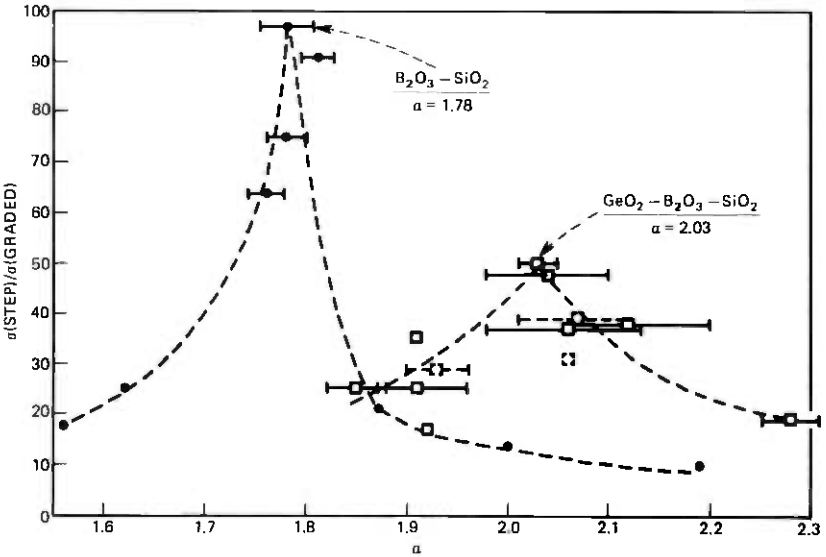


Fig. 5—Root-mean-square pulsewidth reduction factor $\sigma(\text{step})/\sigma(\text{graded}) = [\Delta NL/(12)^{1/2}c]\ 1/\sigma(\text{graded})$ is plotted vs. α for germania borosilicate and borosilicate fibers. Data points (□ for GBS fibers; • for BS fibers) were obtained by deconvolving expected material dispersion effects (0.12 ns/km for GBS fibers, 0.09 ns/km for BS fibers) from the measured fiber output pulsewidths.

type of profile diagnosis has proved to be a very useful guide for fabricating new fibers with close-to-optimal profile gradients.

Figure 5 summarizes our pulse dispersion data with a plot of the rms pulsewidth reduction factor, $\sigma(\text{step})/\sigma(\text{graded})$, relative to comparable step-index fibers, as a function of α for fibers with different profile gradients. Pulsewidth measurement precision is $2\sigma < 0.07$ ns/km because the optical shuttle pulse technique is used to extrapolate 1 km fiber sample lengths by an order of magnitude. A germania avalanche diode with a response time (2σ) of 0.65 nsec was used as the detector, and all such system broadening was deconvolved from the measured pulsewidths. Data points • were obtained from borosilicate fibers with graded B_2O_3–SiO_2 cores and uniform B_2O_3–SiO_2 claddings. The three peak data points, which correspond to pulse dispersions of 0.13, 0.14, and 0.15 nsec/km, are results which show that nearly optimal borosilicate fibers can be repeatably fabricated with $2\sigma < 0.2$ ns/km. The optimal profile at $\lambda = 907.5$ nm is characterized by $\alpha(\text{opt}) \approx 1.78$, and the minimum measured pulse dispersion, $2\sigma = 0.13$ ns/km, represents a 100-fold reduction from the theoretical rms pulse spreading in a step-index fiber with $\Delta N \approx 0.0067$ core-to-cladding index difference.

Data points □ were obtained from germania borosilicate fibers with a graded GeO_2 concentration and a nearly uniform B_2O_3 concentration

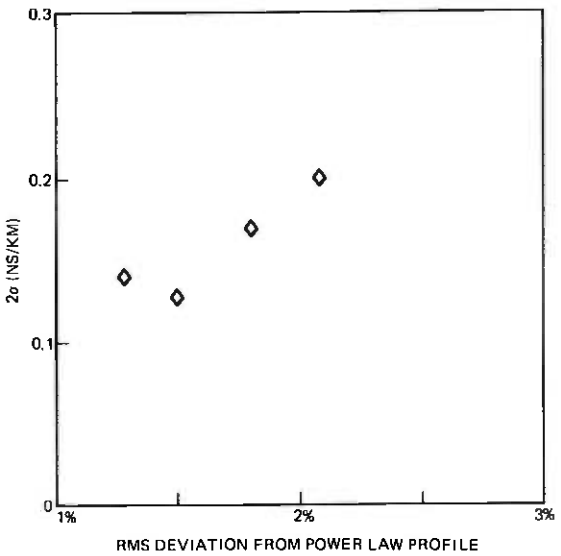


Fig. 6—A plot of the rms deviation, of ΔN vs. r data, from a power law profile vs. pulse dispersion, 2σ, for fibers with near-optimum α values.

across the core. Small amounts of B_2O_3 are added to reduce the viscosity of the core glass, thereby aiding the elimination of bubbles. The two peak □ data points occur when $\alpha \approx 2.03$ and correspond to $2\sigma = 0.4$ ns/km and 0.48 ns/km dispersion values. They represent a 50- and 48-fold reduction from the theoretical rms pulse broadening in equivalent step-index fibers with $\Delta N \approx 0.013$. The optimal α values for each type of fiber are consistent with the values predicted by Fleming[12] on the basis of refractive index measurements.

The theoretical minimum pulse broadening in an optimally graded-index fiber is characterized by:[14]

$$2\sigma(\text{min}) \approx 141(\Delta N)^2 \tag{3}$$

The theoretical maximum pulsewidth reduction factor is given by:

$$\left(\frac{\sigma(\text{step})}{\sigma(\text{graded})}\right)_{\text{max}} = \frac{1}{\sqrt{12}} \frac{\Delta NL/c}{\sigma(\text{min})L} = \frac{14}{\Delta N} \tag{4}$$

which is inversely proportional to ΔN, the maximum core-to-cladding index difference. Therefore, the ratio of 2 between the peak pulsewidth reduction factors for the borosilicate and germania borosilicate fibers in Fig. 5 is consistent with the fact that $\Delta N(\text{GBS}) \sim 2\Delta N(\text{BS})$. However, the measured pulsewidths are 18 to 30 times greater than the minimum values predicted by eq. (3) [$2\sigma(\text{min}) \sim 0.007$ ns/km for borosilicate fibers with $\Delta N \approx 0.007$; $2\sigma(\text{min}) \sim 0.024$ ns/km for germania borosilicate fibers with $\Delta N \sim 0.013$].

Two possible reasons for the measured dispersion being so much larger than the theoretical minimum value are (*i*) the profile α value varies along the length of the fiber, or (*ii*) the profile is not exactly a power law shape. Each of these differences probably applies to some degree. We have attempted to correlate the rms deviations from a power law profile with the pulse dispersion for fibers which have nearly optimal α values.[9] The influence of the refractive index dip in the germania-doped fiber profile was too large for us to obtain any meaningful comparison. In the case of the borosilicate fibers, which have a smoother profile, a rough correlation can be found. Figure 6 demonstrates this possible correlation for fibers with $\alpha = 1.8 \pm 0.05$. These data are far from conclusive, but it is clear that further significant pulsewidth reduction will require improved control of the refractive index profile so that it conforms to the correct power law function more exactly than the present fibers. Development of these improvements will require substantial improvements in MCVD deposition control and in fiber profile measurement.

IV. CONCLUSIONS

Pulse transmission properties have been compared for two of the most common types of high silica graded-index fibers. A spectral filter was used to reduce material dispersion effects caused by a GaAs injection laser at $\lambda = 907.5$ nm and the optical shuttle pulse technique was used to make precise intermodal pulse dispersion measurements.

Nearly optimal profile gradients are characterized by $\alpha \approx 2.03$ in germania borosilicate fibers and by $\alpha \approx 1.78$ in borosilicate fibers. Resultant pulsewidth reduction factors are approximately 50 for germania graded fibers with $\Delta N \approx 0.013$ (N.A. ≈ 0.19) and 100 for boron graded fibers with $\Delta N \approx 0.0067$ (N.A. ≈ 0.14). The factor-of-2 ratio between the peak pulsewidth reduction factors for the two kinds of fiber is consistent with the fact that $\Delta N(\text{GBS}) \approx 2\Delta N(\text{BS})$. However, the measured pulsewidths are 18 to 30 times greater than the minimum values predicted by $2\sigma(\text{min}) \sim 300\Delta^2$ ns/km. Further significant pulsewidth reductions will require improved profile control to reduce by an order of magnitude current 1–3 percent rms deviations between fabricated profiles and optimum power law shapes.

The results in this paper apply at a design wavelength, $\lambda = 907.5$ nm, corresponding to a GaAs injection laser. A companion paper[3] describes how profile dispersion affects fiber transmission bandwidths at other wavelengths.

REFERENCES

1. L. G. Cohen, G. W. Tasker, W. G. French, and J. R. Simpson, "Pulse Dispersion in Multimode Fibers with Graded B_2O_3-SiO_2 Cores and Uniform B_2O_3–SiO_2 Cladding," Appl. Phys. Lett., *28* (April 1976), pp. 391–393.
2. L. G. Cohen, "Pulse Transmission Measurements for Determining Near Optimal

Profile Gradings in Multimode Borosilicate Optical Fibers," Appl. Opt., *15* (July 1976), pp. 1808–1814.
3. L. G. Cohen, I. P. Kaminow, H. W. Astle, and L. W. Stulz, "Profile Dispersion Effects on Transmission Bandwidths in Graded Index Optical Fibers," IEEE J. Quant. Electron., *14* (January 1978), pp. 37–41.
4. W. G. French, G. W. Tasker, and J. R. Simpson, "Graded Index Fiber Waveguides with Borosilicate Composition: Fabrication Techniques," Appl. Opt., *15* (July 1976), pp. 1803–1807.
5. J. B. MacChesney, P. B. O'Connor, F. V. DiMarcello, J. R. Simpson, and P. D. Lazay, "Preparation of Low Loss Optical Fibers Using Simultaneous Vapor Deposition and Fusion," Proc. 4th Int. Cong. on Glass, Kyoto, Japan, *6-40* (July 1974).
6. H. M. Presby, W. Mammel, and R. M. Derosier, "Refractive Index Profiling of Graded Index Optical Fibers," Rev. Sci. Instrum., *47* (March 1976), pp. 348–352.
7. B. C. Wonsiewicz, W. G. French, P. D. Lazay, and J. R. Simpson, "Automatic Analysis of Interferograms: Optical Waveguide Refractive Index Profiles," Appl. Opt., *15* (April 1976), pp. 1048–1052.
8. D. Gloge, private communication.
9. E. A. J. Marcatili, "Modal Dispersion in Optical Fibers with Arbitrary Numerical Aperture and Profile Dispersion," B.S.T.J., *56*, No. 1 (January 1977), pp. 49–63.
10. J. W. Fleming and Luis Soto, "SEM Microanalysis of Germanium Borosilicate Optical Waveguides," Proc. of Pittsburgh Conf. on Analytical Chemistry and Applied Spectroscopy, Paper No. 197, February 1977.
11. L. G. Cohen, "Shuttle Pulse Measurements of Pulse Spreading in an Optical Fiber," Appl. Opt., *14* (June 1975), pp. 1351–1356.
12. J. W. Fleming, "Measurements of Dispersion in GeO_2–B_2O_3–SiO_2 Glasses," J. Amer. Cer. Soc., *59* (November–December 1976), pp. 503–507.
13. L. G. Cohen and C. Lin, "Transmission Measurements of Zero Material Dispersion in Optical Fibers," CLEA Conf., Washington, D.C., Paper PD 5.12, June 1977, abstract in IEEE J. Quant. Electron., *13* (September 1977), pp. 91D–92D; "Pulse Delay Measurements in the Zero Material Dispersion Wavelength Region for Optical Fibers," Appl. Opt., *16* (December 1977), pp. 3136–3139.
14. J. A. Arnaud and J. W. Fleming, "Pulse Broadening in Multimode Optical Fibers with Large $\Delta n/n$. Numerical Results," Electron. Lett., *12* (April 1976), pp. 167–169.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Analytical Foundation for Low-Frequency Power-Telephone Interference

By J. C. PARKER, JR.

(Manuscript received June 30, 1977)

The mechanisms of interference at voice-band frequencies from a power distribution system which adversely affect the telephone loop plant are systematically described. A unified derivation is presented of simple lumped-element circuit models for telephone plant coupling, shielding, and longitudinal-to-metallic conversion. This approach establishes both qualitative understanding and quantitative analytical tools for characterizing the effects of low frequency interference. A glossary is included which represents a consensus evaluation of the best contemporary relationship between historical terminology and modern analytical viewpoints.

I. INTRODUCTION

This paper explores a systematic approach for understanding the electromagnetic interaction between power and telephone systems. Historically, the various mechanisms of coupling, shielding, and longitudinal-to-metallic conversion have evolved into separately addressed concerns. This has led to useful insight but somewhat narrow understanding, since the interdependence of the various concepts has received limited consideration. Moreover, since some of the classical treatments extend back over five decades, they are sometimes difficult to read owing to variations in terminology and basic units. We wish to provide a cohesive overview that emphasizes the interrelationship among these topics within a modern analytical setting. This approach, using concepts familiar to recent engineering graduates, unifies historical developments and provides a basis for understanding current viewpoints toward reducing power-telephone interaction.

Although transmission line theory is briefly touched upon as a starting point, the basic framework consists of an analytical model that utilizes only lumped-element circuit theory. This lumped-element general analytical model serves the following purposes. First, it ties together within

one framework a description of the various physical mechanisms that have previously been treated separately. Second, it readily lends itself to systematic computer evaluation of the electromagnetic interaction between these mechanisms. Finally, from this general model is derived several specialized circuit representations of a sufficiently simple nature to furnish maximum physical insight. These specialized circuits highlight the specific and well-known physical mechanisms of inductive, capacitive, and dissipative coupling, inductive shielding, and longitudinal-to-metallic conversion. A practioner may utilize the more comprehensive analytical model, or he may wish to adopt the specialized circuit representations as his basis for understanding, depending on the extent of his concerns. The general analytical model retains its conceptual value as the origin of a single unified approach since it assures that the specialized models provide a consistent description. The usefulness of the various models is indicated in the summary section.

II. GENERAL ANALYTICAL MODEL

Although electromagnetic theory forms a fundamental core from which all macroscopic electrical behavior can be derived, it would be a rather remote starting point for the purposes of this discussion. On the other hand, the overall generality of our useful circuit models might well go unappreciated or else be questioned in the absence of a clear understanding of their origin. To strike a balance, this paper will adopt transmission line theory as a starting point, then move quickly to more familiar lumped-element circuit equations and models.

The multiconductor transmission line theory is derivable from Maxwell's equations with remarkably few restrictions,[1] although in common textbooks generality is often swapped for expediency. Many intricate details involving specialized electromagnetic analysis can be succinctly summarized as transmission line parameters. The presence of all skin effect phenomena,[2] both within the metal conductors and more importantly the resistive earth, can be rigorously accounted for by the transmission line parameters. Since the circuit theory equations and associated models are evolved from transmission line theory, they too can account for all skin effect phenomena. Such phenomena manifest themselves in the form of circuit parameters that are no longer frequency-independent; i.e., the R, L, G, and C can take on frequency dependencies in accordance with rigorous solutions to electromagnetic boundary value problems. In this way, full advantage is taken of the relative simplicity and usefulness of lumped-element circuit theory while, at the same time, maintaining substantial generality.

2.1 Transmission line synopsis

A rather concise physical interpretation is stated here to aid in the

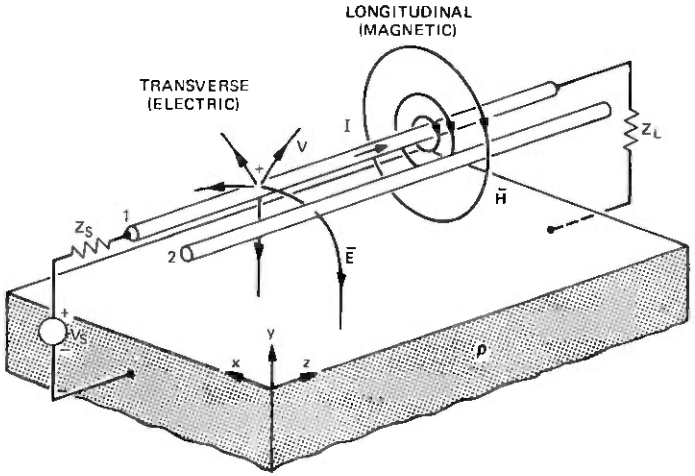


Fig. 1—Induction by electric and magnetic fields from disturbing into disturbed conductor.

understanding of transmission line equations and to instill confidence in their usefulness. While an understanding of these differential equations is highly desirable, their physical content carries over to the simpler algebraic equations to be introduced in the next section.

Consider two conductors located at fixed heights in relation to the surface of a resistive medium as illustrated in Fig. 1. Electromagnetic induction or coupling among parallel conductors may be classified as either transverse or longitudinal with respect to the conductor axes. Transverse coupling characterizes an electric force acting at right angles to the conductor and medium. This perpendicular force arises from an excess of charge that is proportional to conductor potential, V. Since the force acts to drain off and thereby deplete conductor current in the amount of $-dI$ within an incremental distance of dz, this coupling mechanism is described quantitatively by

$$-\frac{dI}{dz} = \mathcal{Y} V. \tag{1}$$

The proportionality factor $\mathcal{Y}$ is called an incremental transverse admittance (mhos/meter). Longitudinal coupling, on the other hand, characterizes an electric force in the direction of the conductor axes. This axial force arises from the movement of charge that is proportional to conductor current, I. Since the axial electric force tends to decrease conductor potential with an increasing z, this coupling mechanism is described by

$$-\frac{dV}{dz} = \mathcal{Z} I. \tag{2}$$

The proportionality factor Z is called an incremental longitudinal impedance (ohms/meter).

Equations (1) and (2) may be recognized as the simple transmission line equations found in basic textbooks.[3] Such treatments often define V as the voltage difference between conductors 1 and 2 and take the I of conductor 1 to return entirely through conductor 2. This convention is appropriate when the two conductors are energized to form just one (metallic) circuit. More generally, a second (longitudinal) circuit is able to coexist on the two conductors. All possible circuits can be systematically taken into account by using the following reference convention. The voltage on each separate conductor is defined with respect to a common reference, in this case the finitely conducting earth which forms a "remote ground," as implied by the electric field lines in Fig. 1. Moreover, the current in each conductor is defined as having total "earth return," quite irrespective of the circuit's actual completion path. The relationship between this systematic "earth-return" reference convention and the useful longitudinal and metallic convention is more fully explored in Appendix B.

Equations (1) and (2) apply implicitly to multiconductor circuits.[4] In this case V and I are taken as column vectors whose components associate with each individual conductor, while $\mathcal{Y}$ and Z are taken as square symmetric matrices. The off-diagonal matrix elements represent "mutual coupling" effects, whereas the "self-reaction" of each conductor is characterized by the diagonal matrix elements. To illustrate this point, observe the matrix differential equations that characterize the two conductors in Fig. 1. The explicit matrix form of eq. (1) is

$$-\frac{d}{dz}\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} \mathcal{Y}_{11} & \mathcal{Y}_{12} \\ \mathcal{Y}_{21} & \mathcal{Y}_{22} \end{bmatrix}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}; \tag{3}$$

for eq. (2) it is

$$-\frac{d}{dz}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}\begin{bmatrix} I_1 \\ I_2 \end{bmatrix}. \tag{4}$$

Each matrix equation is simply a compact notational expedient for representing a system of M coupled individual equations, where M (two in this case) is the number of conductors in the multiconductor transmission line. For instance, expanding this last matrix equation yields

$$\begin{aligned} -\frac{dV_1}{dz} &= Z_{11}I_1 + Z_{12}I_2 \\ -\frac{dV_2}{dz} &= Z_{21}I_1 + Z_{22}I_2, \end{aligned} \tag{5}$$

where Z_{12} (equal to Z_{21} from reciprocity) represents the longitudinal

mutual impedance/unit length which couples conductors 1 and 2. Matrix notation is exceedingly useful in subsequent developments and well worthy of the effort required to acquire familiarity. With this notation it becomes straightforward to systematically account for the interaction of neutral and multiple phase wires of the power system with the strand, sheath, and metallic circuit twisted pair conductors of the telephone system.

2.2 Segment model

A fortunate simplification arises owing to the telephone loop plant not being "long" when measured in relation to the wavelength of voiceband interference frequencies. A typical loop can be subdivided into a minimal number of electrically short segments, each of which is characterized by a fixed geometrical configuration. This allows the transmission line differential eqs. (1) and (2) to be replaced by much simpler lumped-element circuit equations. In addition, an equivalent circuit representation can be identified that will form the basis for all subsequent analyses. The more important details of this simplification are outlined below.

Consider an exposure segment of power and telephone system conductors with a uniform geometrical configuration and extending a length, $\Delta\ell$, between two locations identified as j and $j+1$. The need for $\Delta\ell$ to be electrically short is generally not restricting at voiceband interference frequencies; this point is addressed more quantitatively in Appendix C. A general segment consisting of M individual conductors is illustrated schematically in Fig. 2a. Some conductors represent the power system neutral and phase wires; the remaining conductors can characterize such telephone system wires as a support strand, cable sheath, and twisted voice-circuit pairs. The nature of coaxial or cable-sheath-enclosed conductors is totally characterized by the numerical value of individual elements within the incremental impedance and admittance matrices. Moreover, these matrix elements also reflect spacing and height information. Since the conductor configurations may change from one segment to another, this variation will be identified by the superscript j, $j+1$, on the incremental matrices that are applicable between locations j and $j+1$. The column vectors representing voltage and current variables will also carry an appropriate superscript. The integer subscripts continue to represent specific conductors within matrices and column vectors.

In terms of the matrix notation described above, eq. (2) may be accurately approximated as

$$-\frac{(V^{j+1}-V^{j})}{\Delta\ell}=Z^{j,j+1}I^{j,j+1}, \tag{6}$$

where $I^{j,j+1}$ is the current in the center of segment $j,j+1$. Similarly, using an average value of voltage, eq. (1) becomes

$$\frac{(I_a^j + I_b^{j+1})}{\Delta\ell} = \mathcal{Y}^{j,j+1}\frac{(V^j + V^{j+1})}{2}. \tag{7}$$

In the above equation, the decrease in longitudinal current, $-dI$, between j and $j+1$ has been identified with the transverse currents, I_a^j and I_b^{j+1}, which flow *a*fter and *b*efore the location identified by superscript. (These are the so-called charging currents associated with distributed capacitance of aerial cable, although they might also represent current flow due to distributed conductance of direct buried cable.) It is convenient to define total impedance and admittance matrices such that all elements of the incremental matrices are multiplied by the segment length, $\Delta\ell$:

$$\mathcal{Z}^{j,j+1} \equiv \mathcal{Z}^{j,j+1}\Delta\ell \tag{8}$$

$$\mathcal{Y}^{j,j+1} \equiv \mathcal{Y}^{j,j+1}\Delta\ell. \tag{9}$$

Then, upon algebraic rearrangement eq. (6) becomes

$$V^j - V^{j+1} = \mathcal{Z}^{j,j+1}\, I^{j,j+1} \qquad \begin{pmatrix}\text{longitudinal}\\ \text{impedance}\\ \text{coupling}\end{pmatrix}. \tag{10}$$

Similarly, the two independent terms on the right side of eq. (7) yield

$$\left.\begin{aligned} I_a^j &= Y_a^j\, V^j \\ I_b^{j+1} &= Y_b^{j+1}\, V^{j+1} \end{aligned}\right\} \qquad \begin{pmatrix}\text{transverse}\\ \text{admittance}\\ \text{coupling}\end{pmatrix}, \tag{11a}$$

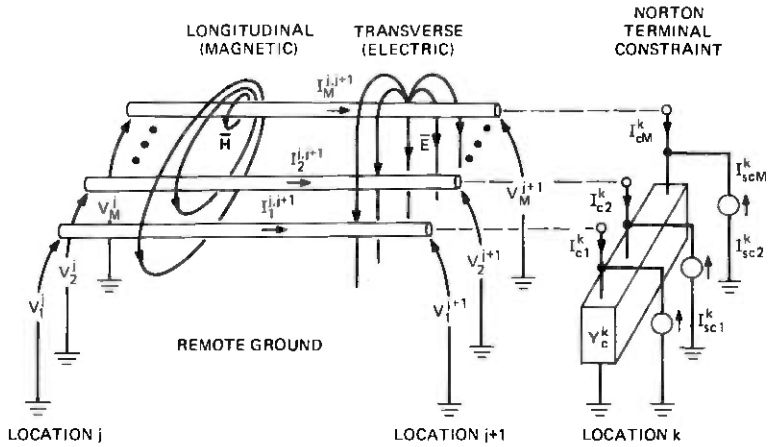
(11b)

where

$$Y_a^j = Y_b^{j+1} \equiv \tfrac{1}{2}\, Y^{j,j+1} \tag{12}$$

has been defined in this decomposition. The circuit theory eqs. (10) and (11) are the desired replacements for the transmission line differential eqs. (1) and (2).

The foregoing basic circuit relationships completely characterize the physical coupling mechanisms, both longitudinal and transverse, within each segment of a power and telephone exposure. These equations lead directly to the equivalent circuit represented in a compact matrix form in Fig. 2b. The voltage and current variables are represented by column vectors, whose elements correspond to variables shown in Fig. 2a. The impedance and admittance elements represent square matrices. Observe that half the total segment admittance has been associated with the centers of each half-segment, occurring *a*fter location j and *b*efore



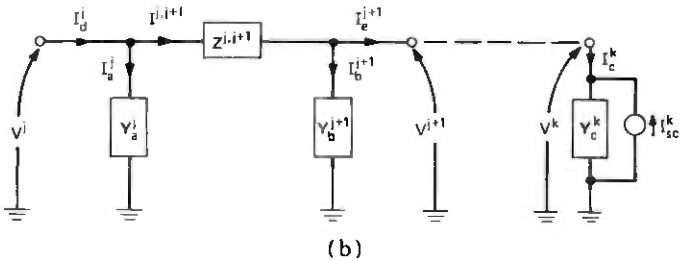


Fig. 2—Interaction model for multiconductor segments and general terminal constraint. (a) Nomenclature for matrix representation. (b) Equivalent circuit in compact matrix form.

location $j + 1$, in accordance with eq. (12). The total current which *d*eparts location j and flows into segment $j, j + 1$ is identifiable as

$$I_d^j = I^{j,j+1} + I_a^j, \tag{13}$$

whereas the total current which *e*nters location $j + 1$ from this segment is given by

$$I_e^{j+1} = I^{j,j+1} - I_b^{j+1}. \tag{14}$$

In summary, all pertinent equations are implied by the equivalent circuit in Fig. 2b when analyzed with matrix algebra.

2.3 Constraint characterization

Now that a multiconductor exposure segment has been completely characterized in quantitative analytical terms starting from transmission

line concepts, the terminations or boundary conditions imposed at each end of a segment must be considered. The variety of terminations encountered in practice is rather substantial. For instance, the power system conductors will have load impedances and discrete grounding impedances attached at various locations. Generators will feed the power line from at least one end and possibly at two or more locations. The telephone system shielding conductors may be either independently grounded or resistively coupled to the power line ground. This resistive coupling may arise from the interaction of closely spaced ground rods or from direct bonding to the power system neutral conductor.

Each of these termination constraints could be implemented on a case-by-case basis for those relatively simple configurations that involve just a few conductors and one or two segments. Once the boundary conditions are characterized with individual circuits, the conglomerate network including the segment representation(s) must then be analyzed. There is a wide variety of network analysis techniques from which an efficient method can be selected. Generally, either mesh or node analysis is convenient for most simple networks.

Rather involved network configurations arise when considering the interactive effects of multiple shielding conductors and several cascaded exposure segments. For these cases it becomes highly desirable to invoke systematic network analysis techniques that lend themselves to straightforward computer implementation. It is possible to handle the wide variety of segment terminations (assumed linear) in a single versatile model by utilizing a Norton equivalent circuit to represent the general terminal constraint. When cast in matrix notation, these boundary conditions take the form:

$$I_c^k = Y_c^k V^k - I_{sc}^k \qquad \begin{pmatrix} \text{general} \\ \text{terminal} \\ \text{constraint} \end{pmatrix}. \tag{15}$$

This approach has been illustrated in Fig. 2 by including a Norton terminal constraint for an arbitrary location k. With a Norton representation, the ideal short circuit current sources, I_{sc}^k, simply vanish for the special case of passive terminations, while the off-diagonal mutual terms in the admittance constraint matrix, Y_c^k, readily account for any resistive coupling from nonindependent grounding. Details concerning the explicit form of the termination circuit have been suppressed into an implicit equivalent characterization. This will serve to systematize the subsequent network analysis procedure. The motivation for choosing a Norton instead of a Thevenin equivalent circuit will be clarified in the following section.

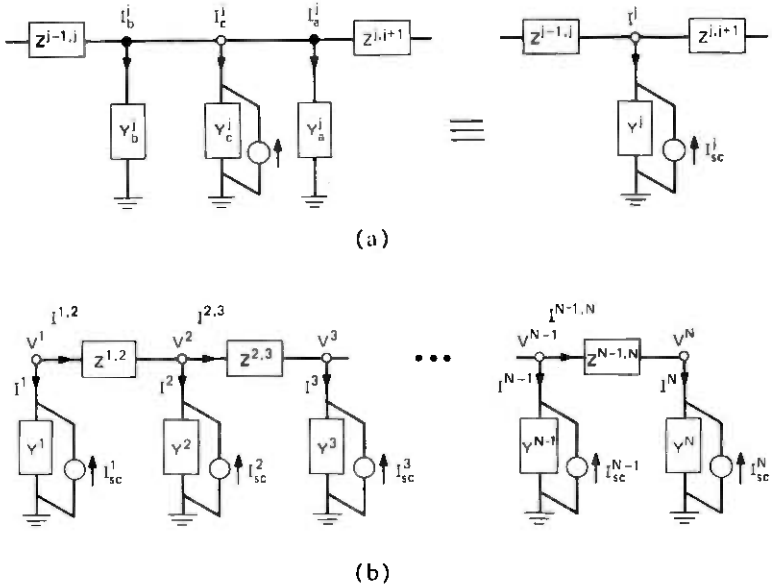


Fig. 3—Matrix representation of network model for many cascaded segments. (a) Simplification at adjoining segments. (b) Composite circuit for N − 1 segments.

2.4 Cascaded configuration

Circuit models have been derived for a multiconductor exposure segment and for a general terminal constraint in the preceding sections. With this building-block approach in mind, these circuits can now be combined to form the general network illustrated in Fig. 3. This network enables analyzing the effects of many exposure segments connected in cascade, while incorporating various grounding and bonding constraints at the segment interfaces. Thus, it becomes a fundamental tool from which to assess electromagnetic interaction between the power distribution system and telephone loop plant. This interaction varies with the configuration changes that occur along the loop plant. Physical parameters amenable to analysis include power-line geometry and separation, size and type of shielding conductors, intervals between grounding and bonding points, quality of grounds and bonds, and degradation of sheath continuity. The presence of such devices as drainage reactors and neutralizing transformers can also be accounted for. Since this network model has basic importance for both the derivation of several specialized circuits in the next section, as well as computerized algorithms that allow detailed parametric studies, a brief description follows.

A useful simplification is contained within the network model at adjoining exposure segments as indicated in Fig. 3a. Note that the two segment admittances, Y_a^j and Y_b^j adjacent to location j, can be combined

with the constraint admittance, Y_c^j at location j, by simple matrix addition to form a total admittance, Y^j.

$$Y^j = Y_c^j + Y_b^j + Y_a^j. \tag{16}$$

With this reduction the total transverse current, I^j, which is "bled-off" the longitudinal current flow, constitutes the column vector addition of currents associated with the individual transverse admittance matrices.

$$I^j = I_c^j + I_b^j + I_a^j. \tag{17}$$

A knowledge of the node voltage, V^j, is sufficient to reconstruct the individual transverse current contributions via multiplication with the appropriate admittance matrix, as detailed in eqs. (11) and (15). Hence, the network analysis can now focus primarily upon determining V^j and $I^{j,j+1}$ for each location and segment. This network simplification is a direct result of having chosen the Norton equivalent circuit to characterize a general terminal constraint.

The foregoing simplification allows the overall network to be represented as shown in Fig. 3b. Assumed known in this model are the parameters for the Norton termination constraints, as well as the segment admittance and impedance matrices. Further discussion of these matrices can be found in Appendix A. An analytical solution for the network model will determine the unknown voltages and currents in terms of the remaining known circuit parameters. An algorithm that provides an efficient solution can be obtained by analyzing the composite circuit using ladder network techniques.[5] A computer program centered around such an algorithm has been developed and utilized extensively to examine the parametric dependence of shielding.

III. SPECIALIZED CIRCUIT REPRESENTATIONS

Particular physical mechanisms associated with just certain conductors of a larger network will be the focal point of this discussion. The influence of the remaining parts of the network upon these chosen conductors can usually be characterized by dependent (i.e., controlled) sources. These sources contain information about the remaining circuitry and succinctly characterize its interaction with the chosen conductors. This approach makes it easier to understand the interaction with the remaining circuitry resulting from specific physical mechanisms. Moreover, it points the way to certain measurements that can be used to characterize a complex network. With these objectives in mind, this section develops specialized circuit models for coupling, shielding, and longitudinal-to-metallic conversion.

This approach furnishes only specialized characterizations of various physical mechanisms; it does not remove the overall system interde-

pendence among these mechanisms. Numerical values for the dependent sources, which represent the effects of the remaining network, may be partly influenced by parameters associated with the few chosen conductors. This interactive behavior can best be accounted for by an analysis of the total network. Such a solution may be simply formulated through the use of mutually interacting specialized circuits. These circuit representations serve as building blocks to construct the total network, and hence, its total interaction. This approach facilitates a direct analysis without the need of general matrix algebra formulations described in the previous section.

On the other hand, the interaction between the dependent sources and parameters of the chosen conductor(s) may sometimes be weak, or of "second-order" effect. In these instances simplifying approximations are appropriate, and quite limited measurements may suffice to characterize specific physical mechanisms. It is difficult, of course, to know when such approximations are valid, short of either actually performing the complete analysis or collecting extensive measurement data. In simplifying approximations, practical experience and/or intuition gained from prior analyses should be of value in establishing sound engineering judgment.

3.1 General coupling model

To simplify the discussion, a single conductor within an exposure segment can be selected for examining the longitudinal and transverse coupling to the remaining conductors within the segment. Let the chosen conductor be labeled i to denote any one of the several conductors numbered 1 through M within the segment. Moreover, consider the segment that connects arbitrary locations j and $j + 1$.

The longitudinal impedance type of coupling within segment $j, j + 1$ is characterized by matrix eq. (10). In its explicit form, this equation reads

$$\begin{bmatrix} V_1^j \\ V_2^j \\ \cdot \\ \cdot \\ \cdot \\ V_i^j \\ \cdot \\ \cdot \\ \cdot \\ V_M^j \end{bmatrix} - \begin{bmatrix} V_1^{j+1} \\ V_2^{j+1} \\ \cdot \\ \cdot \\ \cdot \\ V_i^{j+1} \\ \cdot \\ \cdot \\ \cdot \\ V_M^{j+1} \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1i} & \cdots & Z_{1M} \\ Z_{21} & Z_{22} & \cdots & Z_{2i} & \cdots & Z_{2M} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ Z_{i1} & Z_{i2} & \cdots & Z_{ii} & \cdots & Z_{iM} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ Z_{M1} & Z_{M2} & \cdots & Z_{Mi} & \cdots & Z_{MM} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ \cdot \\ \cdot \\ \cdot \\ I_i \\ \cdot \\ \cdot \\ \cdot \\ I_M \end{bmatrix}. \quad (18)$$

The superscript notation has been omitted from the elements of $Z^{j,j+1}$ and $I^{j,j+1}$, since only one segment is under discussion. The difference

in voltage to remote ground between locations j and $j+1$ for the ith conductor is given by multiplying the ith row of matrix $Z^{j,j+1}$ with column vector $I^{j,j+1}$:

$$V_i^j - V_i^{j+1} = Z_{i1}I_1 + Z_{i2}I_2 + \cdots + Z_{ii}I_i + \cdots Z_{iM}I_M. \tag{19}$$

This equation for longitudinal voltage drop is made up of two kinds of terms.

(i) It is easy to recognize the voltage drop due to the self-impedance of conductor i as $Z_{ii}I_i$.

(ii) There are several mutual impedance terms, each of which accounts for longitudinal voltage induced in conductor i because of current flowing in some other conductor, e.g., k, as $Z_{ik}I_k$.

Hence, rewriting eq. (19) to keep these terms separate gives

$$V_i^j - V_i^{j+1} = Z_{ii}I_i + V_{si}, \tag{20a}$$

where

$$V_{si} \equiv \sum_{\substack{k=1 \\ (k \neq i)}}^{M} Z_{ik}I_k. \tag{20b}$$

The summation term above accounts for all effects associated with longitudinal coupling of the remaining conductors within the segment. When these conductors are relatively close, the mutual impedances, Z_{ik}, consist dominantly of positive reactance, and the coupling is referred to as inductive. For conductors having large spacing, the Z_{ik} parameters are both frequency-independent and dominantly resistive, which constitutes one form of dissipative coupling. The numerical dependence of Z_{ik} upon conductor spacing, frequency, and earth composition is more fully described elsewhere.[6] What is important here is that these Z_{ik} parameters quantify both inductive coupling and resistive coupling.

In a comparable manner, the transverse admittance type of coupling within segment $j, j+1$ is characterized by matrix eqs. (11a) and (11b). With the aid of eq. (12), the explicit form of eq. (11a) reads

$$\begin{bmatrix} I_{a1}^j \\ I_{a2}^j \\ \cdot \\ \cdot \\ \cdot \\ I_{ai}^j \\ \cdot \\ \cdot \\ \cdot \\ I_{aM}^j \end{bmatrix} = \tfrac{1}{2} \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1i} & \cdots & Y_{1M} \\ Y_{21} & Y_{22} & \cdots & Y_{2i} & \cdots & Y_{2M} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ Y_{i1} & Y_{i2} & \cdots & Y_{ii} & \cdots & Y_{iM} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ Y_{M1} & Y_{M2} & \cdots & Y_{Mi} & \cdots & Y_{MM} \end{bmatrix} \begin{bmatrix} V_1^j \\ V_2^j \\ \cdot \\ \cdot \\ \cdot \\ V_i^j \\ \cdot \\ \cdot \\ \cdot \\ V_M^j \end{bmatrix}. \tag{21}$$

The superscript notation has again been omitted from the elements of $Y^{j,j+1}$, but must be retained for the column vectors I_a^j and V^j to distinguish between eqs. (11a) and (11b). The transverse current, I_{ai}^j, which flows after location j from the ith conductor, is obtained by multiplying the ith row of matrix Y_a^j with column vector V^j:

$$I_{ai}^j = \tfrac{1}{2}(Y_{i1}V_1^j + Y_{i2}V_2^j + \cdots + Y_{ii}V_i^j + \cdots + Y_{iM}V_M^j). \tag{22}$$

This equation for transverse current flow is composed of two kinds of terms.

(*i*) The current flow due to the self-admittance of conductor i for the half of the segment closest to location j is identified as $(1/2)Y_{ii}V_i^j$.

(*ii*) Each of several mutual admittance terms accounts for the transverse current flow induced in conductor i caused by voltage on some other conductor, e.g., k, as $(1/2)Y_{ik}V_k$.

Rewriting eq. (22) to keep these terms separate gives

$$I_{ai}^j = \tfrac{1}{2}Y_{ii}V_i^j - \tfrac{1}{2}I_{si}^j, \tag{23a}$$

where

$$I_{si}^j \equiv -\sum_{\substack{k=1 \\ (k\neq i)}}^{M} Y_{ik}V_k^j. \tag{23b}$$

The choice of a negative sign with the summation is partially motivated by the fact that Y_{ik} is generally negative. A completely analogous development may be pursued starting with eq. (11b). Equation (12) is again used to obtain the admittance matrix for the half of the segment occurring before location $j+1$. Separating out the two types of contributors to transverse current flow produces

$$I_{bi}^{j+1} = \tfrac{1}{2}Y_{ii}V_i^{j+1} - \tfrac{1}{2}I_{si}^{j+1}, \tag{24a}$$

where

$$I_{si}^{j+1} \equiv -\sum_{\substack{k=1 \\ (k\neq i)}}^{M} Y_{ik}V_k^{j+1}. \tag{24b}$$

The summation terms of eqs. (23b) and (24b) account for all effects associated with transverse coupling of the remaining conductors within the segment. When all the conductors are above ground, the mutual admittances Y_{ik} consist of appropriately signed susceptance, and the coupling is referred to as capacitive. For buried conductors having direct contact with the soil, the Y_{ik} parameters are both frequency-independent and dominantly conductive, which constitutes a second form of dissipative coupling. In actual buried installations, some conductors may be in direct contact with the soil, while others are insulated with dielectric

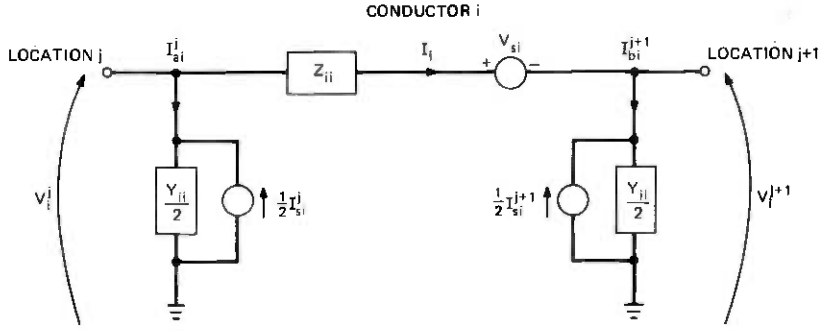


Fig. 4—Inductive, capacitive, and dissipative coupling model for ith conductor.

coatings. Hence, the Y_{ik} transverse admittance parameters quantify capacitive coupling and conductive coupling, both of which may occur simultaneously within a single exposure segment.

A general coupling model can now be obtained based upon the foregoing development. All pertinent coupling mechanisms are contained within eqs. (20), (23), and (24), and an equivalent circuit based upon these equations will portray all the relevant physical concepts. Such a circuit is evolved in the following manner.

The summation terms appearing in eqs. (20b), (23b), and (24b) may be modeled as dependent sources, in accordance with the substitution (or compensation) theorem of circuit theory.[7] In particular, the induced longitudinal voltage term of eq. (20b) is modeled by an ideal dependent voltage source. This voltage source is termed ideal because it contains zero internal impedance. Moreover, it is dependent because its voltage value is determined by currents that exist on the other conductors within the segment, precisely in accord with eq. (20b). On the other hand, the compensation theorem permits modeling the induced transverse current terms of eqs. (23b) and (24b) by ideal dependent current sources. These current sources are termed ideal because they contain zero internal admittance (i.e., infinite impedance). Moreover, they are dependent because their current values are determined by voltages that exist on the other conductors within the segment, precisely in accord with eqs. (23b) and (24b).

Having ascribed dependent sources to account for the summation terms, eqs. (20a), (23a), and (24a) lead directly to the general coupling model shown in Fig. 4. This model applies to each single conductor within a segment; i.e., $i = 1, \cdots, M$. Two conditions must be fulfilled for this equivalent circuit to constitute a valid coupling model.

(*i*) The equivalent circuit when subjected to standard circuit analysis techniques must yield precisely eqs. (20a), (23a), and (24a), since these equations characterize the pertinent coupling mechanisms.

(*ii*) The circuit analysis techniques must yield *only* these equations, i.e., no extraneous information may be falsely implied. This second requirement is quite important if the model is to avoid artifacts suggesting physical behavior that is not actually present.

Careful reflection should reveal that both these conditions have been satisfied in the illustrated coupling model. Hence, this equivalent circuit can be relied upon to furnish physical insight regarding the detailed interactions of all coupling mechanisms.

3.2 Various shielding models

With a reliable circuit model for coupling as a foundation, it is easy to evolve circuit models that describe various forms of shielding. In a rather fundamental way, shielding is nothing more than a wise utilization of available coupling mechanisms. These mechanisms are appropriately constrained to minimize an undesired signal that could contribute to interference. For instance, a low-impedance shunting path (such as a ground on an aerial cable sheath) is commonly used as a constraint upon capacitive coupling. Such a low-impedance terminal constraint serves to decrease the voltage that is supplied to an inherently high-impedance capacitive coupling mechanism and thereby renders the capacitive coupling mechanism ineffective. This type of shielding requires only one low-impedance shunting path and may be correctly termed electric shielding. (Historically, the unfortunate misnomer of "electrostatic" shielding has been used to describe electric shielding.) A second type of shielding requires at least two low-impedance shunting paths. These constraints allow a "shielding current" to flow which, in turn, induces a "shielding voltage" via magnetic induction. Owing to its enormous practical importance, this magnetic (or inductive) type of shielding will be emphasized in the following circuit representations.

The simplest form of inductive shielding is illustrated by the classical shielding model shown in Fig. 5. An understanding of the basic shielding phenomenon will be evident from a straightforward analysis of this simple circuit. However, a derivation of this circuit representation is first necessary to make clear its inherent limitations as caused by various simplifying assumptions.

Consider an exposure segment in which attention is focused upon two individual conductors: one conductor that is to be shielded and a second conductor that will furnish the inductive shielding. The coupling of each conductor to all remaining conductors within the segment is obtained from the inductive coupling model of Fig. 4. To apply this model to the shield (conductor 2), the admittances to ground $Y_{22}/2$ are taken to be zero, since they are shunted by the (assumed) low-resistance grounding terminations R_a and R_b. Moreover, the dependent current sources

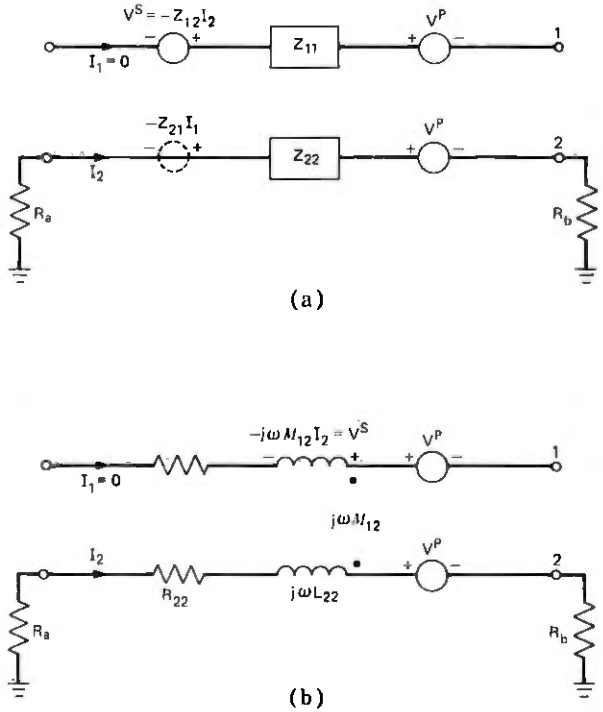


Fig. 5—Classical inductive shielding model. (a) Dependent-source circuit. (b) Complex transformer circuit.

$(1/2)I_{s2}$ are absent since transverse coupling is assumed negligible. Similarly, the coupling model for the shielded wire (conductor 1) also has zero admittances. This is in anticipation of a terminating impedance to ground that is located external to the segment and is low in relation to the impedance of the distributed capacitance of the conductors. By so suppressing extraneous detail wherever possible, it is easy to obtain a clear representation of the inductive shielding mechanism. In fact, no additional complication arises if the shielded wire is actually a twisted pair. In this event, the parameters for conductor 1 are simply replaced by those in the longitudinal circuit of Fig. 6a, which is developed in the next section.

The dependent voltage source within each coupling model is now decomposed. Recall that these voltage sources are determined by eq. (20b) with i given the values of 1 and 2. Those contributions to the voltage sources which arise from conductors 1 and 2 are first separated out as

$$V_{s1} = Z_{12}I_2 + \sum_{k=3}^{M} Z_{1k}I_k \qquad (25a)$$

and

$$V_{s2} = Z_{21}I_1 + \sum_{k=3}^{M} Z_{2k}I_k. \tag{25b}$$

The summation terms account for coupling to all remaining conductors within the segment. If conductors 1 and 2 are physically near each other in comparison to the distance from each remaining current-carrying conductor, then $Z_{1k} \simeq Z_{2k}$, $k = 3, \cdots, M$, as an examination of the appropriate mutual impedance equations will verify. In other words, the coupling to conductors 1 and 2 from each "outside" conductor becomes essentially identical. Under this condition, the summation terms in eqs. (25a) and (25b) are practically equal, and their contribution to the induced voltage within each conductor is the primary voltage, V^p:

$$\sum_{k=3}^{M} Z_{1k}I_k \simeq V^p \simeq \sum_{k=3}^{M} Z_{2k}I_k. \tag{26}$$

Although primary voltage is dependent upon the currents, I_k, ($k = 3, \cdots, M$) which flow elsewhere within the segment, these currents are usually assumed to be unaffected by the presence of the shield, conductor 2. Hence, the V^p appearing in conductors 1 and 2 of Fig. 5 is often taken to be independent, or fixed, just as with a true applied source.

The remaining term in both eqs. (25a) and (25b) accounts for interaction between the shielding and shielded conductor. Consider the first term in eq. (25a) which is the voltage induced into the shielded conductor as a result of current flow in the shielding circuit formed by conductor 2. Since I_2 can be substantially 180 degrees out of phase from its assumed reference direction, this first term is rewritten as

$$Z_{12}I_2 = -[Z_{12}(-I_2)] \tag{27a}$$

$$\equiv -V^s, \tag{27b}$$

where the expression inside brackets is the shielding voltage, V^s. Shielding voltage as defined will often be of similar phase angle to V^p (owing to the negative sign within the brackets), and appears in the circuit representation of Fig. 5a as an opposing voltage source (owing to the second negative sign outside the brackets). It will be convenient to view the difference between the primary and shielding voltage sources as a remnant voltage, V^r. Hence,

$$V^r = V^p - V^s \tag{28}$$

is a useful measure of the shielding mechanism's effectiveness since, by eqs. (25a), (26), and (27), V^r is simply the total induced longitudinal voltage in the presence of shielding. This induced voltage is the driving

mechanism for longitudinal current flow in the shielded circuit and, consequently, for the accompanying longitudinal-to-metallic conversion process. The longitudinal current flow is usually small, however, in comparison to that in the shielding circuit, since the shielded circuit termination impedances are generally large compared to R_a and R_b. Hence, it is assumed that $I_1 = 0$ for purposes of the classical shielding model. This idealization allows the first term in eq. (25b) to be discarded, although its would-be presence is indicated in Fig. 5a by a "dashed-in" dependent voltage source.

By way of summarizing the physical mechanisms contained in the dependent source circuit of Fig. 5a, a simple calculation of shielding effectiveness will be carried out utilizing this classical shielding model. A measure of shielding effectiveness is the shield factor, η, defined as remnant voltage normalized to the exciting primary voltage

$$\eta \equiv \frac{V^r}{V^p} \tag{29a}$$

or

$$\eta = 1 - \frac{V^s}{V^p}, \tag{29b}$$

where the last relation follows from eq. (28). The shielding voltage depends upon longitudinal mutual impedance and shielding current as

$$V^s = Z_{12}(-I_2). \tag{30}$$

The latter quantity obtains from the shielding conductor loop equation

$$-I_2 = \frac{V^p}{Z_{22} + R_T}, \tag{31a}$$

where

$$R_T \equiv R_a + R_b. \tag{31b}$$

The amplitude and phase of the shielding current (and thereby the shielding voltage) is fundamentally dependent upon both longitudinal self-impedance and total termination resistance. Utilizing the above relations with the definition of eq. (29) yields the shield factor as

$$\eta = 1 - \frac{Z_{12}}{Z_{22} + R_T}. \tag{32}$$

Thus, the circuit representation leads quite directly to a simple expression for shielding effectiveness.

It is sometimes enlightening to visualize the inductive shielding mechanism in the context of a single-turn transformer. The circuit

formed by the shielding conductor and its earth-return path is viewed as one side of a transformer. The shielded conductor is viewed as an open-circuited secondary winding. A primary field is presumed to excite both windings, owing to its associated magnetic flux cutting both circuits equally. The resultant current flow in the primary winding then couples a shielding voltage into the secondary winding as a result of mutual inductance between the windings. These notions have been highlighted in the alternate shielding model shown in Fig. 5b.

The validity of the circuit representation in Fig. 5b rests upon its direct equivalence to that of Fig. 5a. Basically, the longitudinal coupling as characterized by dependent sources in Fig. 5a has been alternatively characterized as a complex-valued mutual inductance, with an appropriate transformer dot convention, in Fig. 5b. A conceptual generalization is required here in that the new mutual inductance must now represent energy dissipation as well as energy storage. That is, both the resistive and reactive parts of longitudinal mutual impedance are to be characterized by a mutual inductance term,

$$Z_{12} = (R_{12} + j\omega L_{12}) \equiv j\omega \mathcal{M}_{12}. \tag{33}$$

Hence, this new mutual term must assume a complex value given by

$$\mathcal{M}_{12} = \left(L_{12} + \frac{1}{j\omega} R_{12}\right). \tag{34}$$

It turns out for close conductor spacings that $\mathcal{M}_{12}$ is dominantly real-valued (about 90 percent) as given via L_{12}, and the remaining imaginary term, $-jR_{12}/\omega$, characterizes dissipative coupling associated with the earth's nonzero resistivity. Although for most practical conductor spacings the coupling mechanism itself remains dominantly inductive, the phase relationship between V^s and V^p, and thereby the shielding, can be significantly affected by R_{12}.

3.3 Longitudinal-to-metallic conversion model

The general analytical model and all subsequent models described so far have been developed with the earth-return reference convention for voltage and current variables. This choice of reference convention was motivated primarily from the standpoint of consistency, such that various segments and termination constraints could be systematically cascaded to facilitate easy computer evaluation. Sometimes physical insight into a particular aspect of the overall problem can best be enhanced by adapting voltage and current variables which have a reference convention relating more directly to actual operating conditions. For instance, the effects of unbalanced operation in telephone circuits (which are intended to operate in basically a balanced circuit mode) are most

easily understood when analyzed as a so-called longitudinal and metallic circuit.

The relationship between the earth-return reference convention and the longitudinal and metallic reference convention is covered in detail in Appendix B. This relationship constitutes a change in variables, or a transformation, when stated in mathematical terms. This change further requires an appropriate modification of the previous circuit relations. The transformed equations, also detailed in the appendix, basically relate impedance and admittance quantities as used in a longitudinal or metallic circuit back to those appropriate to an earth-return circuit. The nomenclature and definitions introduced in Appendix B will be utilized in the following development.

To focus attention upon the electrical behavior of conductors 1 and 2 utilizing longitudinal and metallic reference conventions, the transformed circuit relations will be mathematically "partitioned." As in the specialized models already developed, one portion of the matrix equations remains unaltered, and the influence of other portions are lumped together into a single dependent term. This allows specialized circuit models for the first two conductors to be derived, in which the effects of all other conductors 3 through M are characterized as dependent voltage and current sources. From eq. (48) in Appendix B, the longitudinal interaction involving conductors 1 and 2 is given by

$$\begin{bmatrix} V_m^j \\ V_\ell^j \end{bmatrix} - \begin{bmatrix} V_m^{j+1} \\ V_\ell^{j+1} \end{bmatrix} = \begin{bmatrix} Z_m & \tfrac{1}{2}\Delta Z \\ \tfrac{1}{2}\Delta Z & Z_\ell \end{bmatrix} \begin{bmatrix} I_m \\ I_\ell \end{bmatrix} + \begin{bmatrix} V_{om} \\ V_{o\ell} \end{bmatrix}. \tag{35}$$

Here, the voltage contribution due to other conductors within the segment is denoted as

$$V_{om} \equiv \sum_{k=3}^{M} (Z_{1k} - Z_{2k}) I_k \tag{36a}$$

for the metallic component, whereas for the longitudinal component,

$$V_{o\ell} \equiv \sum_{k=3}^{M} \tfrac{1}{2}(Z_{1k} + Z_{2k}) I_k. \tag{36b}$$

Hence, excitation of the longitudinal mode depends upon the average longitudinal mutual impedance between the pair conductors (1 and 2), and each of the others ($k = 3, \cdots, M$). On the other hand, excitation of the metallic mode depends upon the difference in mutual impedances, as based upon the earth-return reference convention. This difference is typically minimized by twisting the pair conductors, causing them to effectively occupy the same position in relation to the other conductors. The transverse interaction, which corresponds to eq. (11), follows using

eq. (49) as

$$\begin{bmatrix} I^j_{am} \\ I^j_{a\ell} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} Y_m & \tfrac{1}{2}\Delta Y \\ \tfrac{1}{2}\Delta Y & Y_\ell \end{bmatrix}\begin{bmatrix} V^j_m \\ V^j_\ell \end{bmatrix} - \frac{1}{2}\begin{bmatrix} I^j_{om} \\ I^j_{o\ell} \end{bmatrix}. \tag{37a}$$

The transverse current flow due to other conductors within the segment is given by

$$I^j_{om} \equiv -\sum_{k=3}^{M} \tfrac{1}{2}(Y_{1k} - Y_{2k})V^j_k \tag{38a}$$

and

$$I^j_{o\ell} \equiv -\sum_{k=3}^{M} (Y_{1k} + Y_{2k})V^j_k. \tag{38b}$$

As in eq. (36), excitation of the longitudinal mode depends upon a sum of transverse mutual admittances, whereas excitation of the metallic mode depends upon a difference of terms characterizing coupling to other conductors. The seemingly illogical appearance or omission of a 1/2 multiplicative factor is just a peculiarity of the longitudinal and metallic reference convention stated in eq. (43), Appendix B. A second eq., (37b) (not given), is obtained by replacing j with $j + 1$ and the subscript a with b. All the physical phenomena between locations j and $j + 1$ is encompassed within eqs. (35) and (37). If $V_3, \cdots, V_M$ and $I_3, \cdots, I_M$ are assumed known, either by previous analytical solution or by direct measurement, the above equations yield the response for conductors 1 and 2.

Physical insight is gained by constructing equivalent circuit models whose behavior is controlled by eqs. (35) and (37). It is convenient to derive two specialized circuit models: one to characterize the metallic variables and another for the associated longitudinal variables. Consolidating similar variables from eqs. (35) and (37) by simple rearrangement yields the pairs

$$\begin{aligned} V^j_m - V^{j+1}_m &= Z_m I_m + V_{sm} \\ I^j_{am} &= \tfrac{1}{2}Y_m V^j_m - \tfrac{1}{2}I^j_{sm} \\ I^{j+1}_{bm} &= \tfrac{1}{2}Y_m V^{j+1}_m - \tfrac{1}{2}I^{j+1}_{sm} \end{aligned} \tag{39}$$

and

$$\begin{aligned} V^j_\ell - V^{j+1}_\ell &= Z_\ell I_\ell + V_{s\ell} \\ I^j_{a\ell} &= \tfrac{1}{2}Y_\ell V^j_\ell - \tfrac{1}{2}I^j_{s\ell} \\ I^{j+1}_{b\ell} &= \tfrac{1}{2}Y_\ell V^{j+1}_\ell - \tfrac{1}{2}I^{j+1}_{s\ell}, \end{aligned} \tag{40}$$

where the new subscript s denotes source. Two terms are contained in

each dependent voltage source:

$$V_{sm} \equiv \tfrac{1}{2}\Delta Z I_\ell + V_{om}$$

$$= \tfrac{1}{2}(Z_{11} - Z_{22})I_\ell + \sum_{k=3}^{M} (Z_{1k} - Z_{2k})I_k \tag{41a}$$

and

$$V_{s\ell} \equiv \tfrac{1}{2}\Delta Z I_m + V_{o\ell}$$

$$= \tfrac{1}{2}(Z_{11} - Z_{22})I_m + \sum_{k=3}^{M} \tfrac{1}{2}(Z_{1k} + Z_{2k})I_k. \tag{41b}$$

The first term in each source represents mode coupling due to impedance unbalance and the second term represents the coupling to other conductors as defined in eq. (36). Similarly, the dependent current sources become

$$I^j_{sm} \equiv -\tfrac{1}{2}\Delta Y V^j_\ell + I^j_{om}$$

$$= -\tfrac{1}{2}(Y_{11} - Y_{22})V^j_\ell - \sum_{k=3}^{M} \tfrac{1}{2}(Y_{1k} - Y_{2k})V^j_k \tag{42a}$$

and

$$I^j_{s\ell} \equiv -\tfrac{1}{2}\Delta Y V^j_m + I^j_{o\ell}$$

$$= -\tfrac{1}{2}(Y_{11} - Y_{22})V^j_m - \sum_{k=3}^{M} (Y_{1k} + Y_{2k})V^j_k. \tag{42b}$$

Here the first term in each source represents mode coupling due to admittance unbalance and the second term represents coupling to other conductors as defined in eq. (38). A corresponding set of current sources is obtained from eq. (37b) simply by replacing j with $j + 1$ in the above. The metallic circuit eqs. (39) can now be characterized by the equivalent circuit in Fig. 6b, and the associated longitudinal circuit eqs. (40) similarly in Fig. 6a.

It should be noted that in general these two circuits are "coupled" via the dependent sources; that is, the dependent voltage and current sources for the metallic circuit require knowledge of V_ℓ and I_ℓ from the longitudinal circuit, and vice versa. In certain situations, however, the coupling parameters are such that the excitation from a dependent source becomes independent of the other equivalent circuit. In this situation the metallic and longitudinal segment models are said to be "decoupled"; i.e., they function independently in the absence of termination unbalances. Furthermore, it may even occur that mutual coupling parameters to the partitioned conductors $3, \cdots, M$ are such as to suppress excitation of a dependent source. To illustrate these points, consider the case of

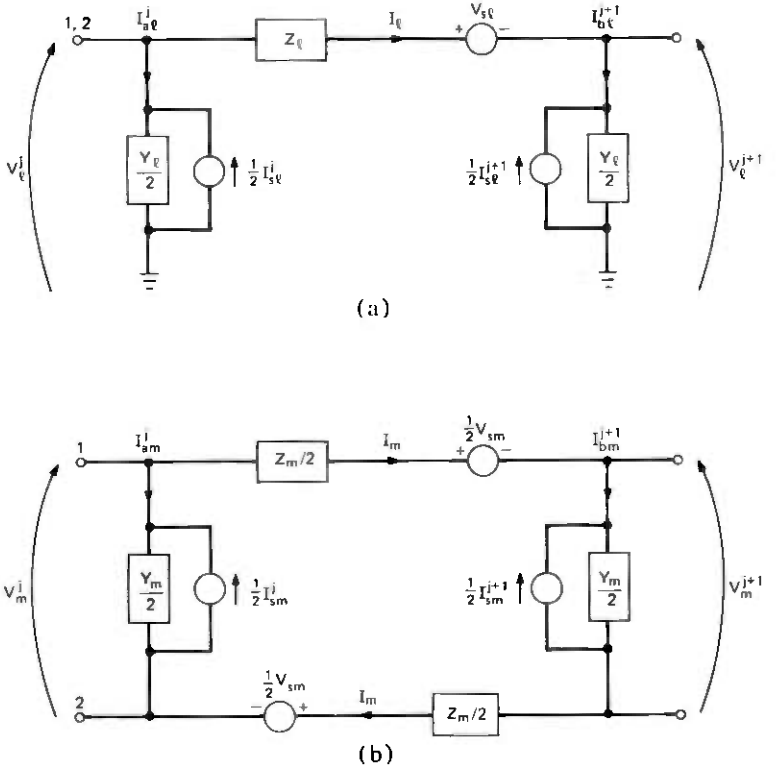


Fig. 6—Longitudinal-to-metallic conversion model. (a) Segment of a longitudinal circuit. (b) Segment of a metallic circuit.

a twisted pair with small admittance and impedance unbalances enclosed within a conducting sheath. The twisting results in $Z_{22} \simeq Z_{11}$, i.e., small impedance unbalance, and generally $|I_m| \ll |I_k|$ for some $k = 3, \cdots, M$. This allows the first term in eq. (41b) to be dropped since, as the product of two small quantities, it is negligible compared to the summation term. Similar reasoning applies to the first term in eq. (42b), which also can be dropped since the admittance unbalance is assumed small. Let us assume for illustrative purposes the remaining pair conductors are bonded to the sheath, and denote this aggregate collection as conductor 3. Then, within the summation, terms Y_{1k} and Y_{2k} are zero for $k > 3$, owing to capacitive shielding furnished by the enclosing sheath (conductor 3). Hence, the current sources can be removed from the longitudinal circuit model provided the sheath conductor is at ground potential, i.e., $V_3 = 0$, while the voltage source becomes dependent only upon current flow in conductors 3 through M. This allows the longitudinal equivalent circuit to be solved first, subject only to various terminal

constraints (i.e., termination impedances). The results are then applied to the metallic equivalent circuit to assess longitudinal-to-metallic conversion, arising from coupling parameter and termination impedance unbalances. This simplified longitudinal circuit and the decoupled metallic circuit can be explored quite effectively to illustrate the longitudinal-to-metallic conversion process.

IV. SUMMARY

A systematic approach has been described for evolving lumped-element circuit models appropriate to low frequency interference analysis. By straightforward computer implementation one can reliably assess the intricate interaction of various physical mechanisms which occur in real life interference situations. Care has been taken to follow a strictly deductive approach in the modeling procedure. This ensures an accurate characterization of all relevant physical mechanisms, while preventing the occurrence of any extraneous modeling artifacts.

The specialized circuit representations have highlighted the individual effects of coupling, shielding, and longitudinal-to-metallic conversion on telephone cable facilities. Their development systematically identifies all underlying assumptions and thereby offers clarification on the conditions which must be satisfied to permit confident reliance on these models. Three distinct types of coupling are identified: inductive, capacitive, and dissipative (the latter taking both longitudinal and transverse forms). Electric and magnetic types of shielding have been motivated as just wise utilizations of available coupling mechanisms. Finally, the longitudinal-to-metallic conversion model is unique in its ability to concisely characterize an arbitrary multi-conductor power and telephone environment.

A carefully prepared glossary has been included as Appendix D. Particular attention is given to resolving jargonistic ambiguities by referring back to fundamental principles. It is the author's experience that unnecessary confusion generally results from the use of "loose" terminology. In instances of conflicting historical usage, the present analytical framework is relied upon to furnish definitiveness.

V. ACKNOWLEDGMENTS

The author wishes to acknowledge many enlightening discussions, particularly concerning the glossary, with colleagues W. A. Reenstra, C. W. Anderson, D. V. Batorsky, G. A. DeBalko, G. H. Estes, D. N. Heirman, P. M. Lapsa, and D. W. McLellan, along with S. W. Guzik, formerly from the Electrical Coordination and Protection Group of the AT&T Company.

APPENDIX A

Perspective on Admittance Matrices

Transverse coupling can occur either within an individual exposure segment or from mutual interaction within a terminal constraint, as indicated by the model for cascaded segments shown in Fig. 3. In all but a few actual cases, the terminal constraint phenomena are generally dominant by manifesting high admittances, Y_c^j, compared to those for distributed effects in individual segments, Y_a^j and Y_b^j. In these situations, it is acceptable to assume Y_a^j and Y_b^j are zero for computational purposes. Basically, this implies that within a segment the radial flow of displacement current (associated with distributed capacitance) and of conduction current (associated with leakage conductance) are both negligible, either in themselves or with respect to current entering conductor nodes of the terminal constraint at location j. These conditions are typically met for nodes to which practical grounding configurations have been applied. However, for nodes at which grounding is either relatively poor or intentionally absent, all contributing admittances need to be retained. For instance, the conductance to ground and to other conductors may require consideration with direct burial noninsulated cable, such as concentric neutral power line. Conductance values can be obtained, either by means of direct measurement or by existing analytical techniques.[6] Moreover, the susceptance portion of Y_a^j and Y_b^j arising from self-capacitance and mutual capacitance effects may be important under either very high voltage or exceptionally low current situations. The former situation can exist in the proximity of EHV power transmission systems.[8,9] The latter low-current situation manifests itself in affecting actual open-circuit voltage-to-ground (or sheath) on pairs within long segments of telephone cable.

Treating terminal constraint phenomena in the mathematical framework of admittance parameters is somewhat arbitrary, since a Thevenin representation might also have been chosen to characterize a general terminal constraint in terms of impedance parameters. When the terminal constraints are simply independent grounds, it is logical to consider the associated ground potential raise (GPR) a longitudinal impedance phenomenon as did Sunde,[6] since its primary effect is to restrict longitudinal current flow. On the other hand, when closely spaced nonindependent grounds interact strongly through mutual GPR effects, it is reasonable to view longitudinal currents as being partially excited through a localized transverse interaction with adjacent grounds acting as primary sources. Aside from whichever type of excitation actually prevails, it is computationally expedient to model mutual GPR interaction as a transverse admittance phenomenon, primarily because of the

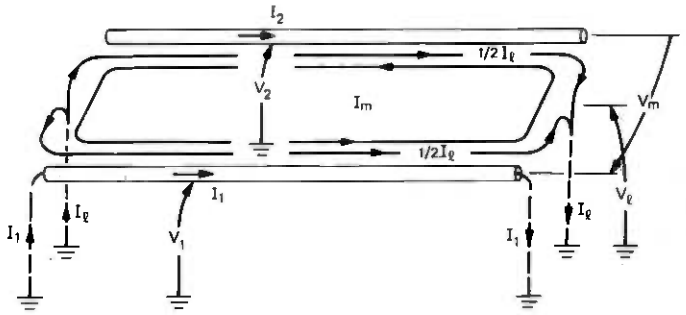


Fig. 7—Relationship of longitudinal and metallic variables to earth-return variables.

analytical simplification of combining parallel admittances through addition as illustrated in Fig. 3a.

APPENDIX B

Longitudinal and Metallic Reference Convention

A portion of the general analysis pertaining to multiconductor segments is modified by explicitly singling out two conductors for an alternate characterization in terms of the longitudinal and metallic reference convention. This "change of variables" necessitates that a transformation be applied to the basic circuit relations. The modified equations for use with this alternate reference convention are summarized below.

Let us identify conductors 1 and 2 for this new representation. (By appropriate choice of numbering, these two conductors might represent a twisted pair within a conducting sheath, or even more simply an open wire line. Alternatively, they could represent a phase and neutral conductor from a single-phase power system.) The basic approach is to define four new variables, V_m, V_ℓ, I_m, I_ℓ, in terms of the previous voltages-to-ground V_1, V_2 and earth-return currents I_1, I_2 as follows:

$$\begin{aligned} V_m &= V_1 - V_2 && \text{(metallic voltage)}, \\ V_\ell &= \tfrac{1}{2}(V_1 + V_2) && \text{(longitudinal voltage)}, \\ I_m &= \tfrac{1}{2}(I_1 - I_2) && \text{(metallic current)}, \\ I_\ell &= I_1 + I_2 && \text{(longitudinal current)}. \end{aligned} \tag{43}$$

It may be helpful to note the last two current relations derive from

$$\begin{aligned} I_1 &= \tfrac{1}{2}I_\ell + I_m, \\ I_2 &= \tfrac{1}{2}I_\ell - I_m. \end{aligned} \tag{44}$$

The relationships between old and new variables are illustrated in Fig. 7. These new longitudinal and metallic variables differ quite notably

from the previous earth-return variables in that they encompass effects occurring on two conductors. Consequently, all four new variables quantify effects that are "shared" between conductors 1 and 2 as they function together. (It is interesting to note that standard noise measuring sets automatically record V_ℓ and V_m.) In summary, the foregoing change of variables can be compactly represented by transformations T_v and T_i in the following matrix form

$$V_T \equiv \left[\begin{array}{c} V_m \\ V_\ell \\ \hline V_3 \\ \cdot \\ \cdot \\ \cdot \\ V_M \end{array}\right] = \left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \cdots & 0 \\ \tfrac{1}{2} & \tfrac{1}{2} & 0 & \cdots & 0 \\ \hline 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{array}\right] \left[\begin{array}{c} V_1 \\ V_2 \\ \hline V_3 \\ \cdot \\ \cdot \\ \cdot \\ V_M \end{array}\right] \equiv [T_v]V \qquad (45a)$$

and

$$I_T \equiv \left[\begin{array}{c} I_m \\ I_\ell \\ \hline I_3 \\ \cdot \\ \cdot \\ \cdot \\ I_M \end{array}\right] = \left[\begin{array}{ccc|ccc} \tfrac{1}{2} & -\tfrac{1}{2} & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \hline 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{array}\right] \left[\begin{array}{c} I_1 \\ I_2 \\ \hline I_3 \\ \\ \cdot \\ \cdot \\ I_M \end{array}\right] \equiv [T_i]I. \qquad (45b)$$

The new transformed voltage and current column vectors are denoted as V_T and I_T, respectively. Moreover, for ease of examination dashed lines are used to partition each matrix equation into regions having a particular form. Within the transformation matrices, the upper left-hand corner represents eqs. (43), whereas the lower right-hand corner preserves the earth-return reference convention for conductors 3 through M.

Recall the transmission line eqs. (1) and (2) in their matrix form, such that they characterize the complete behavior within a multiconductor exposure segment. Interjecting the new transformed variables of eq. (45), the transmission line relations become

$$-\frac{dI_T}{dz} = [T_i \mathcal{Y} T_v^{-1}]V_T \equiv \mathcal{Y}_T V_T \qquad (46)$$

and

$$-\frac{dV_T}{dz} = [T_v Z T_i^{-1}]I_T \equiv Z_T I_T, \qquad (47)$$

where $\mathcal{Y}_T$ and Z_T are the new transformed incremental admittance and impedance matrices, as obtained from the indicated matrix operations. On the basis of reasoning identical to that preceding eqs. (10) and (11), a corresponding set of circuit relations is evolved in terms of the transformed variables. In particular, eq. (10) becomes eq. (48) shown on page 1691, where the following notation has been introduced:

$$Z_m \equiv Z_{11} + Z_{22} - 2Z_{12} \quad \text{(metallic circuit impedance)},$$

$$Z_\ell \equiv \tfrac{1}{4}(Z_{11} + Z_{22} + 2Z_{12}) \quad \text{(longitudinal circuit impedance)},$$

$$\Delta Z \equiv Z_{11} - Z_{22} \quad \text{(impedance unbalance)}.$$

Individual Z_{ik} values are the matrix elements of $Z^{j,j+1}$ defined in eq. (10). The superscripts on $Z_T^{j,j+1}$ and $I_T^{j,j+1}$, which identify the specific segment under consideration, have been omitted on the matrix and column vector elements in eq. (48) to minimize notational congestion. The transformed version of eqs. (11) follow simply upon noting eq. (49), shown on page 1691, where

$$Y_m \equiv \tfrac{1}{4}(Y_{11} + Y_{22} - 2Y_{12}) \quad \text{(metallic circuit admittance)},$$

$$Y_\ell \equiv Y_{11} + Y_{22} + 2Y_{12} \quad \text{(longitudinal circuit admittance)},$$

$$\Delta Y \equiv Y_{11} - Y_{22} \quad \text{(admittance unbalance)}.$$

Both $Z_T^{j,j+1}$ and $Y_T^{j,j+1}$ are symmetric matrices since all $Z_{ik} = Z_{ki}$ and $Y_{ik} = Y_{ki}$ from reciprocity.

APPENDIX C

Circuit Model Frequency Restrictions

It is worthwhile from both an applicational and theoretical point of view to clarify the frequency range over which the equivalent circuit models are applicable. Recall that in starting from the transmission line eqs. (1) and (2), and going to the lumped-element circuit theory eqs. (10) and (11), an electrically short segment of the loop plant was specifically considered. The meaning of this earlier assumption is now examined in some detail.

Note the electrically short assumption is likewise implicit in arriving at the metallic and longitudinal circuit eqs. (39) and (40), to which the equivalent circuit models of Fig. 6 apply. The initial discussion will focus upon these two models and later be expanded to include all conductors within an exposure segment. Whereas these circuit models are exact characterizations of the lumped-element circuit equations, it is important to know how well they approximate the behavior associated with the original transmission line differential equations. Fortunately, this

$$\begin{bmatrix} V_m^j \\ V_\ell^j \\ \hline V_3^j \\ V_4^j \\ \cdot \\ \cdot \\ \cdot \\ V_M^j \end{bmatrix} - \begin{bmatrix} V_m^{j+1} \\ V_\ell^{j+1} \\ \hline V_3^{j+1} \\ V_4^{j+1} \\ \cdot \\ \cdot \\ \cdot \\ V_M^{j+1} \end{bmatrix} = \left[\begin{array}{cc|ccc} Z_m & \tfrac{1}{2}\Delta Z & (Z_{13}-Z_{23}) & (Z_{14}-Z_{24}) & \cdots \\ \tfrac{1}{2}\Delta Z & Z_\ell & \tfrac{1}{2}(Z_{13}+Z_{23}) & \tfrac{1}{2}(Z_{14}+Z_{24}) & \cdots \\ \hline (Z_{31}-Z_{32}) & \tfrac{1}{2}(Z_{31}+Z_{32}) & Z_{33} & Z_{34} & \cdots Z_{3M} \\ (Z_{41}-Z_{42}) & \tfrac{1}{2}(Z_{41}+Z_{42}) & Z_{43} & Z_{44} & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ & & Z_{M3} & \cdots & Z_{MM} \end{array}\right] \begin{bmatrix} I_m \\ I_\ell \\ \hline I_3 \\ I_4 \\ \cdot \\ \cdot \\ \cdot \\ I_M \end{bmatrix} \tag{48}$$

$$Y_T^{j,j+1} \equiv \mathcal{Y}_T^{j,j+1}\,\Delta\ell = \left[\begin{array}{cc|ccc} Y_m & \tfrac{1}{2}\Delta Y & \tfrac{1}{2}(Y_{13}-Y_{23}) & \tfrac{1}{2}(Y_{14}-Y_{24}) & \cdots \\ \tfrac{1}{2}\Delta Y & Y_\ell & (Y_{13}+Y_{23}) & (Y_{14}+Y_{24}) & \cdots \\ \hline \tfrac{1}{2}(Y_{31}-Y_{32}) & (Y_{31}+Y_{32}) & Y_{33} & Y_{34} & \cdots Y_{3M} \\ \tfrac{1}{2}(Y_{41}-Y_{42}) & (Y_{41}+Y_{42}) & Y_{43} & Y_{44} & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ & & Y_{M3} & \cdots & Y_{MM} \end{array}\right] \tag{49}$$

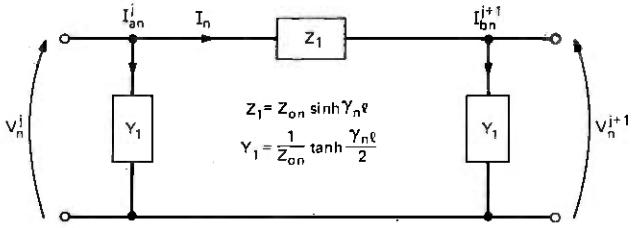


Fig. 8—Equivalent network for a general transmission line.

question is rather easily resolved, since an exact pi-network model is available for the general transmission line case.[10] Specifically, the exact behavior of the terminal voltage and terminal current is characterized by the equivalent network shown in Fig. 8. The particular choice of voltage and current variables defines a form of "mode" which will be designated as n. Here, $n = \ell$ or m, for either the longitudinal or metallic type mode response of the transmission line differential equations. The characteristic impedance for mode n is defined by

$$Z_{on} \equiv \sqrt{\frac{Z_n}{Y_n}}, \tag{50}$$

and the all-important electrical length is

$$\gamma_n \ell \equiv \sqrt{Z_n Y_n}\,, \tag{51}$$

where the Z_ℓ, Z_m and Y_ℓ, Y_m were specified in Appendix B by eqs. (48) and (49), respectively, for a segment length of $\Delta\ell \equiv \ell$.

The general transmission line equivalent network of Fig. 8 reduces to the impedances and admittances within the circuit models of Fig. 6 for sufficiently small values of $\gamma_n \ell$. To see this, note that

$$\sinh \gamma_n \ell \to \gamma_n \ell \tag{52a}$$

and

$$\tanh \frac{\gamma_n \ell}{2} \to \frac{\gamma_n \ell}{2} \tag{52b}$$

for $|\gamma_n \ell| \ll 1$. Hence, for the series impedance branch of the pi-network,

$$Z_1 = Z_{on} \sinh \gamma_n \ell \to Z_{on} \gamma_n \ell = \sqrt{\frac{Z_n}{Y_n}} \sqrt{Z_n Y_n} = Z_n; \tag{53}$$

for the two shunt admittance elements,

$$Y_1 = \frac{1}{Z_{on}} \tanh \frac{\gamma_n \ell}{2} \to \frac{\gamma_n \ell}{2 Z_{on}} = \sqrt{Z_n Y_n} \Big/ 2\sqrt{\frac{Z_n}{Y_n}} = \frac{Y_n}{2}, \tag{54}$$

where $n = \ell$ or m. The restriction $|\gamma_n \ell| \ll 1$ can be relaxed substantially without incurring appreciable error. For instance, with $|\gamma_n \ell| \leq 0.5$, the circuit elements of Fig. 6 will be within 5 percent of their true values as given by transmission line theory in Fig. 8. This condition serves as a practical guide in deciding the upper frequency limit for which just one segment of the longitudinal and metallic circuit models furnish reasonable accuracy.

A few observations of more general nature will now be made. The metallic and longitudinal variable equivalent circuits derived previously focus attention upon the circuit variables and parameters contained in the upper part of the partitioned eqs. (48) and (49). Dependent sources were utilized to account for the coupling to other conductors within the segment. One can similarly focus attention upon the lower part of the partitioned eqs. (48) and (49). In this case, each distinct conductor will have just a single earth-return equivalent circuit, corresponding to that shown in Fig. 4. Moreover, each conductor will give rise to one distinct mode n, wherein the impedance and admittance parameters of the transmission line solution and lumped-element circuit solution relate as $Z_n = Z_{nn}$, $Y_n = Y_{nn}$, $n = 3, \cdots, M$. Hence, we see that *every* conductor contained in the segment has associated with it an electrical length, $\gamma_n \ell$, the largest of which will be denoted as $\gamma_N \ell$. Evidently a segment is electrically short, and thus characterizable by a correspondingly simple circuit model, whenever $|\gamma_N \ell|$ is sufficiently small in the sense discussed above. The presence of dependent sources does not alter this conclusion, since they too characterize only electrically short behavior. That is, the sources are composed of mutual admittance and mutual impedance terms, for which it can be shown that $|Y_{ik}| < |Y_{kk}|$ and $|Z_{ik}| < |Z_{kk}|$ for all k (or n) of interest. Hence, if the self-admittance and self-impedance terms, Y_{nn} and Z_{nn}, satisfy an electrically short criterion, so too must the mutual terms.

It should be stated that these modes and associated propagation constants, γ_n, are somewhat unorthodox; they differ from the customary "eigenmodes" of the M-conductor system. Eigenmodes are ordinarily defined[4] from the n eigenvalue roots, γ_n^2, of the matrix equation $|YZ - \gamma^2 U| = 0$, where U is a unity matrix. Each distinct root then constitutes two eigenmodes, $e^{\pm \gamma_n z}$, having the property that they can propagate independently without undergoing distortion. To the extent the largest $\gamma_N \equiv \max \gamma_n$ can be easily estimated without extensive numerical procedures, this more orthodox approach furnishes a completely rigorous method for testing a multiconductor segment of length ℓ to determine compliance with being electrically short. The values of γ_n associated with the eigenmode approach differ from those described above in accordance with the levels of mutual coupling, as characterized by the dependent sources in the circuit models of Figs. 4 and 6.

APPENDIX D

Glossary

Conduction Current	The current flow that is associated solely with the finite conductivity, σ, or resistivity, ρ, of a medium; e.g., the current flow through a resistance.
Conductive Coupling	The contribution to transverse coupling arising from Ohm's law and the flow of conduction current.
Controlled Sources	Current and voltage generators (sources) controlled by voltage and current signals, respectively. Since controlled sources are dependent on a control signal, they are often referred to as dependent sources.
Dependent Sources	See controlled sources.
Displacement Current	The current flow associated solely with the permittivity, ϵ, of a medium; e.g., the current flow through a capacitance.
Dissipative Coupling	A general term encompassing the physical mechanism of resistive or conductive coupling.
Earth-Return Reference Convention	A voltage variable whose reference potential is that of remote ground; a current variable whose return path is assumed to be through the ground.
Effective Mutual Impedance	The ratio of the total induced open-circuit voltage on the disturbed circuit to the disturbing power system phase current with the effects of all conductors taken into account.
Electric (or Capacitive) Coupling	The contribution to transverse coupling arising from Gauss' law of electric induction and the flow of displacement current.
Electromagnetic Coupling	A general term encompassing primarily electric and magnetic coupling and, in principle, dissipative coupling too.
External Impedance	The (total) longitudinal impedance less that contribution due to internal impedance.
Internal Impedance	The sum of the conductor resistance and internal reactance (from the magnetic field inside the conductor).

Longitudinal and Metallic Reference Convention	Voltage and current variables that are related to the earth-return reference convention as shown in eq. (43).
Longitudinal Circuit	A circuit utilizing longitudinal reference conventions for the voltage and current variables.
Longitudinal Coupling	The force exerted on a charge by an electric field in a longitudinal or axial direction of a conductor.
Longitudinal Current	The directed current flow along the axis of a conductor; this flow may be measured using either earth-return or longitudinal reference conventions.
Longitudinal Impedance	A quantitative measure of longitudinal coupling within a segment, accounting for both inductive and resistive types of coupling (dependent upon chosen reference convention).
Longitudinal Input Impedance	The Thevenin impedance looking into the longitudinal circuit formed by the wire pair and terminating impedances.
Longitudinal Voltage	The change in voltage occurring along the axis of a conductor; this change may be measured using either longitudinal or earth-return reference conventions. (Also the voltage variable in the longitudinal reference convention).
Magnetic (or Inductive) Coupling	The contribution to longitudinal coupling arising from Faraday's law of magnetic induction and the flow of conduction current.
Metallic Circuit	A circuit utilizing metallic reference conventions for the voltage and current variables.
Mutual Coupling	A transverse admittance or longitudinal impedance which relates the stimulus on one conductor to its response upon another conductor.
Phasor	A harmonically time-dependent complex number. Phasors are added, subtracted, multiplied, and divided in the same way as

	complex numbers. Often phasor additions are referred to as vector additions.
Power Line Balance	The relative absence of residual current at a particular frequency for a group of power line conductors; balance at one frequency, in general, does not imply balance at another (harmonic) frequency.
Power Line Balance Factor	The ratio of the sum of the phase current magnitudes divided by the residual current magnitude; this quantity or its logarithm is a figure of merit for the degree of balance of a power line at a particular frequency.
Primary Voltage	The voltage induced in the disturbed circuit by current in the disturbing circuit, in the absence of the intended shielding circuit.
Remnant Voltage	The voltage induced in the disturbed circuit by current in the disturbing circuit with the intended shielding circuit present.
Remote Ground Potential	The potential of a region several skin-depths into the earth immediately beneath some specified conductor location. This reference potential is usually assumed to equal zero.
Residual Current	The instantaneous or phasor sum of the conductor currents for a group of power line conductors; this current returns to the source through a path other than those conductors.
Resistive Coupling	The contribution to longitudinal coupling arising from Ohm's law and the flow of conduction current.
Self-Reaction	A transverse admittance or longitudinal impedance which relates the stimulus to its associated response upon the same conductor.
Shield Factor	The ratio of remnant voltage to primary voltage for a constant value of disturbing current; a measure of shielding effectiveness, i.e., the smaller the ratio or shield factor, the better the shield in reducing remnant voltage.
Skin Depth	The depth into a conductor at which the magnitude of the electromagnetic field is reduced to approximately $1/e$ of its surface value; its thickness is given in terms of resistivity, radian frequency, and permeability as $\delta = \sqrt{2\rho/\omega\mu}$.

Skin Effect	A phenomenon associated with time-varying fields which tends to concentrate currents toward the surface of conductors that are nearest to the field sources which produce the currents. The skin-depth is a measure of this effect.
Transverse Admittance	A quantitative measure of transverse coupling, accounting for both capacitive and conductive types of coupling (dependent upon chosen reference convention).
Transverse Coupling	The force exerted on a charge by an electric field in a direction transverse or perpendicular to the axis of a conductor.
Transverse Current	The outward current flow perpendicular to the axis of a conductor which accompanies a decrease in longitudinal current.

REFERENCES

1. P. I. Kuznetsov and R. L. Stratonovich, *The Propagation of Electromagnetic Waves in Multiconductor Transmission Lines*, Oxford, England: Pergamon Press, 1964.
2. J. R. Whinnery and S. Ramo, *Fields and Waves in Modern Radio*, New York: John Wiley & Sons, 1953, Chap. 6.
3. S. R. Seshadri, *Fundamentals of Transmission Lines and Electromagnetic Fields*, Reading, Mass.: Addison-Wesley, 1971.
4. C. R. Paul, "On Uniform Multimode Transmission Lines," IEEE Trans. on Microwave Theory and Techniques, *MTT-21*, No. 8 (Aug. 1973).
5. Franklin F. Kuo, *Network Analysis and Synthesis*, 2nd ed., New York: John Wiley & Sons, 1966.
6. Erling D. Sunde, *Earth Conduction Effects in Transmission Systems*, New York: Dover Publications, 1968.
7. E. J. Angelo, Jr., *Electronic Circuits*, New York: McGraw-Hill, 1964.
8. Eric T. B. Gross and M. Harry Hesse, "Electrostatically Induced Voltages Above High Voltage Lines," Journal of the Franklin Institute, *295*, No. 2 (February 1973).
9. L. O. Berthold, Chrm., Working Group, et al., "Electrostatic Effects of Overhead Transmission Lines, Part II–Methods of Calculation," Paper 21 TP 645-PWR, IEEE Summer Meeting and International Symposium on High Power Testing, Portland, Oregon, July 18–23, 1971.
10. Walter C. Johnson, *Transmission Lines and Networks*, New York: McGraw-Hill, 1950.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 5, May–June 1978
Printed in U.S.A.

Idle Channel Noise Suppression by Relaxation of Binary ADM-Encoded Speech

By S. V. AHAMED

(Manuscript received November 3, 1977)

Techniques of identifying the location of silence periods in the binary data of prerecorded telephone messages from Adaptive Delta Modulation (ADM) encoders are discussed. Two algorithms for detecting and replacing such periods by absolute silence are investigated. In the first method one or two words (each 16 bits long) spanning one or two msecs, respectively, at a 16 kHz ADM-sampling frequency, are searched. They are either replaced or not in their entirety by a perfect silence sequence based upon a computed decision. In the second method, blocks of any prespecified number (1 to 512) of words spanning 1 to 512 msec at 16 kHz are searched to detect the authenticity of a silence period within a smaller block (typically between 1 to 31 words) spanning 1 to 31 msec embedded within the larger block. The first method has yielded unsatisfactory results, and the second method with larger window scan of 80–200 msecs and a smaller window duration of 1 to 5 msec has yielded nearly perfect results by eliminating all of the idle channel noise without degrading the quality of the message. This technique of locating the silence periods in ADM encoded speech is compared with that for ADPCM encoded speech. Also the optimized parameters (for making the computed decision) are shown to be satisfactory for preventing false silence clues for genuine speech data over the wide variations of frequencies and amplitudes. Further, these parameters may be converted for different clock rates, and some are traced back to the specifics of the encoder and the others to the character of speech.

I. INTRODUCTION

The source of idle channel noise in most ADM (Adaptive Delta Modulation) speech encoders is the nonuniformity of the data stream generated by the encoder during silence periods. From a relaxation* study

* In the context of this paper, relaxation implies changing the binary data to a state such that any further change will not yield an improvement in the quality of the speech or of silence.

on the computer it becomes obvious that if the decoder is forced into a repetitive input bit pattern synchronized with the main ADM clock, the decoder idle channel noise can be suppressed. The decoder, responding only to the incoming data stream, cannot generate any noise on its own, and a perfect encoder, to assure silence, would provide a repetitive pattern to the decoder. In the absence of such an encoder, it is possible to determine the silence periods during speech by a computed decision and force the bit pattern to be repetitive during such periods.

The silence periods are particularly conspicuous in exposing the imperfection of the encoder. Being devoid of any meaningful information, they can become annoying if the encoder does not produce a bit pattern which forces a complete silence for the decoder. For instance, the most commonly used bit pattern is 0101. . ., even though other patterns such as 001100. . ., 01001101. . ., etc., all yield a type of semisilence or humming silence with intertwined frequencies. The nature of decoder silence is also influenced by the companding algorithm and the output filter characteristics. In this application we have obtained the best silence by a sequence of 0101. . . which offers two distinct advantages:

(*i*) The compandor having been rendered inoperative by this sequence, the step size decays to its minimum value.

(*ii*) The frequency generated by this sequence, being half the clock rate, is beyond the bandwidth of the audio filter.

II. SILENCE CLUES

2.1 Cluster and transitions clues for ADM data

When the only available data is the encoded binary stream from an ADM encoder, three clues to judge the authenticity of the silence period may be used: (*i*) the absence of at least one cluster of ones whose width exceeds that of a prechosen minimum threshold cluster, (*ii*) the absence of at least one cluster of zeros whose width exceeds that of a prechosen minimum threshold cluster, and (*iii*) the excess of transitions between zeros and ones and vice versa over a prechosen number of transitions during a prechosen interval of time. The first two clues together have been termed "cluster clues," and the last one is termed "transitions clue." There are thus five variables necessary to implement the scheme: (*a*) the number of ones in the minimum threshold cluster of ones, (*b*) the number of zeros in the minimum threshold clusters of zeros, (*c*) the number of transitions in the duration to seek the clues, (*d*) the duration of time to seek the clues and finally, (*e*) the duration for substituting the ideal silence sequence 0101. . . instead of the imperfect silence data from the encoder.

2.2 Code word energy clue for ADPCM data

In Ref. 1, the concept of "code word energy" which is defined as an integrated energy for 16 msec (8 msec forward, 8 msec backward) around a preselected code word of ADPCM data is used. The concept has been used to locate the beginning and end of utterances. Individual discrete code words which convey one of the sixteen levels of amplitude information of the speech wave are scanned for amplitude deviation from a mean value. If the code word energy consistently exceeds that obtained during silence periods at the ADPCM encoder for 50 msec, then the beginning of the utterance is traced back to the instant at which the code word energy first started to exceed the threshold. Conversely, if the code word energy falls below that recorded during speech consistently for 160 msec, the end of the utterance is traced back to the instant at which the code energy first receded from the threshold.

2.3 Differences between clues for ADPCM and ADM data

Code word energy clue is well suited for ADPCM data where amplitude information is discretely coded in each word. For ADM data, only the sign information is conveyed to the decoder and there is no distinguishable boundary between any neighboring bits. For this reason, the concept of code word energy is inapplicable for ADM-encoded data. The cluster clues contain the information that not once during the search interval did the encoder experience an unidirectional change in amplitude lasting for a minimum number of clock cycles. The transitions clue contains the information that the frequency of change of direction of the output from the decoder exceeds what one would expect it to be during low level speech signals, and very close to what one would expect during the silence periods. The cluster clue is based on the fact that speech contains the dominant portion of energy in lower formant frequencies and the transition clue is based on the fact that speech energy concentration tapers off at higher formant frequencies.

Further, the computed decision for ADPCM data is a process of only looking forward in time to determine the increase of energy over the silence threshold energy (for locating the beginning of an utterance) or the decrease of energy below that of the speech threshold energy (for locating the end of the utterance). In ADM data we have found it neessary to look forward and backward for clues to make a computed decision regarding the authenticity of silence during the intermediate interval. As shown in the following sections, if the five variables [(a) through (e) in Section 2.1] on which the concept is implemented, are carefully chosen, the cluster and transitions clues work very dependably.

III. IMPLEMENTATIONAL DETAILS

Implementing the concept has been attempted by two distinct techniques: (*i*) word relaxation schemes and (*ii*) block relaxation schemes. In word relaxation schemes the duration for seeking the cluster and transitions clues [i.e., (*d*) in Section 2.1] is made the same as the duration for replacing the ideal silence sequence 0101 . . . instead of the imperfect silence data [i.e., (*e*) in Section 2.1]. In block relaxation schemes the duration for replacement has been made 2.3 to 13 percent of the duration for searching for the clues to judge the authenticity of the silence.

3.1 Word relaxation schemes

In this method the durations of search and replacement have both been made typically 1 or 2 msec. The width of the minimum threshold cluster for ones (hereafter referred to as "ones cluster threshold") and the width of the minimum threshold cluster for zeros (hereafter referred to as "zeros cluster threshold") are both typically set at two. The threshold for the transitions (hereafter referred to as "transitions threshold") is typically set at ten (16 kHz clock rate) for a one msec search and replacement window. The concept has been programmed in assembly language on a minicomputer in a conversational mode of input and output. Large blocks of data have been processed. The results have been analyzed both by subjective tests and oscillographic studies. Whenever the computer does replace the imperfect silence by a perfect silence the replaced data generates a limit cycle with respect to the input frequency and it appears as a stationary pattern on an oscilloscope. The scheme, even though successful about 90 percent of the time, fails to satisfy a perceptive listener during the silence period. When the three threshold parameters are made (3,3,8) respectively,* the quality of the silence becomes perfect but the speech tends to become slightly choppy and a happy compromise of the three threshold parameters has not proved possible.

Similar results have been encountered with two msec duration for seeking the clues and replacing the silence bit-pattern. A slight improvement gained in the quality of silence with perfect speech still does not satisfy a perceptive listener, and for this reason the block relaxation procedure has been developed.

3.2 Block relaxation schemes

The limited performance of the word relaxation scheme is due to a very narrow slice of time used to determine whether a word contains silence

* These parameters are always written in the same order: ONES, ZEROS, TRANSITIONS.

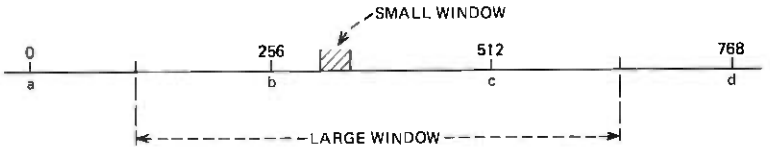


Fig. 1—Data management during the detection and replacement of silence periods. *a* to *d*: Memory buffer refilled from the disk after processing the middle 256 words such that *b* is reread at *a*, *c* at *b*, etc. *b* to *c*: Limits of excursion of the small window and storage of processed data.

or genuine data. The block relaxation scheme expands the search for clues over a much larger interval (large window) to compute the decision for a much shorter interval (small window, see Fig. 1). Typical durations for the large and small windows are from 39 to 131 msec and from 1 to 5 msecs, respectively. The clusters and transitions clues are verified by the same method discussed in Section 2.1. The large window is defined during program execution by two parameters: (*i*) the duration for which the search for the clues should start prior to the beginning of the small window and (*ii*) the duration for which the search for the clues should continue after the end of the small window. Parameters (*i*) and (*ii*) have been made distinct because the choppiness of plosives, fricatives, siblants and nasals at the ends of the utterances can be eliminated by making (*i*) longer than (*ii*) while simultaneously eliminating the slightest crackle before an utterance. The effects of changing the durations of the large window, the small window and the threshold parameters are tabulated in Table I.

IV. DISCUSSION OF RESULTS

Dependable determination of the silence periods and complete suppression of idle channel noise is possible with the right value of the threshold parameters and window sizes. On one hand if the threshold parameters are (1,1,16) for ADM data at 16 kHz with a large window and small window of 1 msec each, then the block relaxation scheme becomes identical to the word relaxation scheme and no change is effected between the original data and the processed data. On the other hand, if the threshold parameters are made (16,16,0) with one msec small and large window, all the data is changed to that of a perfect silence. Other variations of threshold parameters and window widths, if not *all* optimal, may fail one way (by retaining slight choppiness in the words) or the other (by imperfect idle channel noise suppression). Hence the search for optimal threshold parameters and window widths has led us to a minimum cluster threshold parameter of three* and to a transition

* This defines that the presence of an "instantaneous" frequency generated with a cluster whose width exceeds three (i.e., 2 kHz and below at 16 kHz clock rate) within the large window qualifies the small window to be classified as genuine speech data.

Table I — Control parameters and quality

File Name	Durations in msec: Before	Small window	After	Threshold parameters: Ones	Zeros	Transitions	Remarks, subjective
A	3	3	3	3	3	80	Noisy silence, message OK.
B	32	17	32	2	2	896	Noisy silence, message OK.
C	3	3	3	3	3	88	Slightly noisy silence, slightly choppy words.
D	3	3	3	3	3	72	Perfect silence, choppy words.
T	64	3	64	3	3	1856	Noisy silence, message OK.
U	64	3	64	3	3	1664	Perfect silence, message OK.
V	64	3	64	3	3	1536	Perfect silence, very slight choppiness of words.
P	64	3	64	3	3	1728	Trace of crackle in the silence periods, message OK.
Y	18	3	18	3	3	480	Very slight crackle, very slight choppiness.
Z	18	3	18	3	3	512	One crackle in 8 sec of silence, message OK.
P*	18	3	18	3	3	480	One crackle in 16.2 sec of silence.
O*	18	3	18	3	3	544	Three crackles in 16.2 sec of silence.

Notes: 1. The clock frequency is 16 kHz.
2. The message is a typical standard telephone announcement.
3. Large window width is sum of three columns Before, Small window and After.
* The imperfect silence of the encoder is processed here.

threshold parameter of about 0.783† times the maximum number of transitions that can occur in the large window; the large and small windows themselves being 100 to 131 and 1 to 5 msecs respectively. This transition threshold parameter is consistent with our estimation of the frequency of 01 or 10 transitions and 00 or 11 and 000 or 111 cluster formations during silence periods. Sixty percent of the time the encoder generates a 01 or 10 transition, thirty percent of the time the 00 or 11 cluster appears and ten percent of the time the 000 or 111 cluster is generated. These statistics yield the transitions threshold parameter as 0.783‡ times the maximum transitions one may expect.

Computationally optimized values of the transitions threshold verify the result. For instance, in case of File U (Table I) processed with a large window width of 131 msec, the maximum number of transitions possible are 2096 at a clock rate of 16 kHz and the estimated transitions for the silence period are thus 1634. The computationally optimized value is 1664. A similar assertion of the threshold parameters may be made for File P (Table I) with an estimated value at 489 transitions and a computationally optimized value of 480 transitions.

† This defines that if the average frequency generated in the large window is above 6.26 kHz (i.e., 0.783 × 8 kHz), then in the absence of the cluster clues, the data in the small window is the imperfect silence of the encoder.

‡ 0.783 = 0.6 + 0.3/2 + 0.1/3.

V. RELATION BETWEEN OPTIMIZED PARAMETERS AND DATA ON SPEECH-SILENCE STATISTICS

The computationally optimized control parameters are consistent with published speech statistics.[2]

Low level consonant sounds (some plosives "p", "t" and fricative "θ"), which are short (30–50 msec) are not mistaken as silence because of neighboring higher level signals which precede or follow them. The cluster clues in the large window (up to 131 msecs wide) fail to be triggered. Longer duration (up to 230 msecs) low level consonant sounds especially sibilant "s" and fricative "z" may tend to trigger the cluster clues but fail to trigger the transitions clue because of the larger value of the transitions threshold parameter. Other plosives, fricatives, sibilant-fricatives, nasals and semivowels lasting between 80 to 200 msec falling between the two earlier cases fail to trigger either the cluster clue (because of neighboring clusters in the large window) at the lower limit of 80 msec or the transitions clue (because of high value of transitions expected in the large window) at the higher limit. The fricative "f" and the semivowel "r" which have some spectral energy at about 3500 Hz and also last between 200 and 160 msecs, still fail the transitions clue since the average limit for the frequency of the imperfect encoder silence is about 6.26 kHz. Higher level consonants and key word vowels fail to trigger either cluster clue or the transitions clue because of their higher amplitude. As has been determined by the variation of the large window width on the computer, the results obtained by longer larger windows are more satisfactory than those from much shorter ones.

The nature of the compandor (syllabic or instantaneous) also plays a role here. Consider a syllabic compandor whose encoded data is being scanned by a large window width of about 70 msec and a long sibilant or fricative that is at the end of a word. If the step size is large due to earlier letters in the word, then a false trigger of both the cluster and transitions clues is possible. Such a process yields a choppy ending of the word. However, a larger window of about 300 msec will eliminate this condition. Conversely, data from an instantaneous compandor would fail to trigger either of the two clues even with a 70 msec large window.

Further, we have noticed that an exact central spacing of the smaller window in the larger window is less desirable than spacing the smaller window towards the end (shifted 10–15 percent). Such a spacing has two advantages: (*i*) false clues by long sibilants and fricatives at the end of words (when the step size of the ADM encoder is already large due to earlier letters in the word) are eliminated and (*ii*) crackles in the silence periods located just at the beginning of words are correctly eliminated (because the step size is low due to the preceding silence, a cluster is most likely formed at the beginning of the utterance). A similar observation

has also been made in Reference 1 while detecting the end of an utterance. The critical duration after which code word energy falls below the threshold is 160 msec as against 50 msec for the energy to be above the threshold while detecting the beginning of the utterance. Large windows, inconsistently long in relation to the type of companding of the encoder face the risk of imperfectly suppressing the idle channel noise, and conversely large windows too narrow for the type of companding in the encoder face the risk of leaving behind traces of choppiness in the words.

From the study presented in the paper for the data from an ADM encoder* with an attack time constant of 3 msec and a decay time constant of about 9 msec, the large window should be about 120–150 msec encompassing a small window of 3–5 msec located at 60 percent from the start of the large window duration. The cluster threshold constants are each 3 and the transitions threshold is approximately 0.783 times the maximum number of transitions possible in the large window.

VI. CONCLUSIONS

Idle channel noise may be virtually eliminated from stored ADM messages by the block relaxation techniques presented in this paper. With a proper selection of the control parameters, the silence periods can be accurately located and replaced by a perfect silence. At very low clock rates (about 16 kHz) where the compromise between idle channel noise and intelligibility is very real, the intelligibility can be enhanced by recording with a low step size, which implies a disturbing component of the idle channel noise. This can be completely suppressed, however, by block relaxation techniques.

The flexibility offered by this technique also permits the detection of silence over a wide range of clock frequencies and amplitudes of the recording message. Typically it is necessary to scan about 70–200† milliseconds before judging the authenticity of a silence period and to base the judgment on the number of transitions and consecutive ones and zeros in that interval of time.

REFERENCES

1. L. H. Rosenthal, R. W. Schafer, and L. R. Rabiner, "An Algorithm for Locating the Beginning and End of an Utterance Using ADPCM coded Speech," B.S.T.J., *53*, No. 6 (July–August 1974), pp 1127–1135.
2. D. L. Richards, "Telecommunications by Speech," New York: John Wiley, 1973, Sec. 2.1.3.3.

* With syllabic companding.

† H. Seidel and C. H. Bricker have implemented the concepts in real time hardware, arriving at similar values.

Contributors to This Issue

Syed V. Ahamed, B.E., 1957, University of Mysore; M.E., 1958, Indian Institute of Science; Ph.D., 1962, University of Manchester, U.K.; Post Doctoral Research Fellow, 1963, University of Delaware; Assistant Professor, 1964, University of Colorado; Bell Laboratories, 1966—. At Bell Laboratories, Mr. Ahamed has worked in computer-aided engineering analysis and design of electromagnetic components. He has designed and implemented minicomputer software and hardware interfacing. He has applied algebraic analysis to the design of domain circuits and investigated computer aids to the design of bubble circuits. He has investigated new character designs for microwave power in the ϵ-band. He has developed hardware and software interfacing for audio frequency codecs. Since 1975, he has been optimizing codec designs, encoding techniques, and speech encoded data storage and manipulation by minicomputers.

Andres Albanese, Ingeniero Electricista, 1970, Universidad Central De Venezuela; M.Sc., 1972, University of Texas at Austin; Ph.D., 1976, Stanford University. Instituto Venezolano De Investigaciones Cientificas, 1969–1970; Bell Laboratories, 1975—. Mr. Albanese's current research interests are systems and components for lightwave communications.

H. W. Arnold, B.A., 1965, Occidental College; M.A., 1967, Sc.D., 1971, Columbia University; Bell Laboratories, 1971—. Mr. Arnold has conducted millimeter wave mobile radio propagation experiments and has investigated advanced communications satellite systems. He was involved in the design of the Crawford Hill COMSTAR beacon propagation receivers and is presently performing data analysis from that experiment. Member, IEEE.

Robert H. Brandt, Stevens Institute of Technology; Bell Laboratories, 1944–1977. Mr. Brandt's work in the radio research group was concerned with components for micro-wave radio relay systems, antenna impedance measurements, multiplexing system for light route microwave relay, phase correction and pulse timing circuits, the Project Echo sat-

ellite communication experiments, and mobile radio propagation and equipment. Immediately before retiring he was associated with the COMSTAR Satellite Beacon Propagation Experiment.

Ta-Shing Chu, B.S., 1955, National Taiwan University; M.S., 1957, and Ph.D., 1960, Ohio State University; Research Associate, Courant Institute of Mathematical Sciences, New York University, 1961–1963; Bell Laboratories, 1963—. Mr. Chu has been engaged in research on microwave antennas and tropospheric wave propagation for satellite communication and terrestrial microwave network. Fellow, IEEE; member, International Scientific Radio Union, Sigma Xi, Pi Mu Epsilon.

Leonard G. Cohen, B.E.E., 1962, City College of New York; Sc.M., 1964, and Ph.D. (Engineering), 1968, Brown University; Bell Laboratories, 1968—. At Brown University, Mr. Cohen was engaged in research on plasma dynamics. At Bell Laboratories, he has concentrated on optical fiber transmission studies. Member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu; senior member, IEEE.

Donald C. Cox, B.S. (EE), 1959, and M.S. (EE), 1960, University of Nebraska; Ph.D. (EE), 1968, Stanford University; U.S. Air Force Research and Development Officer, Wright-Patterson AFB, Ohio, 1960–1963; Bell Laboratories, 1968—. After coming to Bell Laboratories from Stanford where he was engaged in microwave transhorizon propagation research, Mr. Cox was engaged in microwave propagation research in mobile radio environments and in high-capacity mobile radio systems studies until 1973. He is now doing millimeter wave satellite propagation and systems research. Senior Member, IEEE and member, Commissions B, C and F of USNC/URSI, Sigma Xi, Sigma Tau, Eta Kappa Nu, and Pi Mu Epsilon; Registered Professional Engineer.

F. V. DiMarcello, B.S. (Geochemistry), 1960, Pennsylvania State University; M.S. (Ceramics), 1966, Rutgers, The State University; Bell Laboratories, 1960—. Mr. DiMarcello has worked on various glass and ceramic materials problems including substrates for thin film circuitry, microwave windows for hardened antennae, and glass and ceramic-to-metal seals. He is currently involved in materials and processing aspects of optical waveguides.

N. F. Dinn, B.S.E.E, 1967, Northeastern University; M.S.E.E., 1969, MIT; Bell Laboratories 1967—. Mr. Dinn's initial work was concerned with design and development of adaptive equalization and automatic timing control for digital systems. He subsequently performed the initial systems engineering for T1C. He currently supervises the Radio Characterization Studies Group which has responsibility for designing specialized measurement and data acquisition equipment for characterizing radio propagation and for evaluating general trade digital radio systems. Member, Phi Kappa Phi, Tau Beta Pi, Sigma Xi, and Eta Kappa Nu.

Robert W. England, B.S., 1973, Capitol Institute of Technology; Bell Laboratories, 1973—. Mr. England has worked on microwave antennas for satellite communication. He was involved in the measurement of the Crawford Hill 7-meter antenna and the development of an earth station receiver for the ATS-6 satellite.

James Flanagan, Sc.D. Electrical Engineering, 1955, Massachusetts Institute of Technology; Bell Laboratories, 1957—. Mr. Flanagan has worked in voice communications, acoustics, and digital techniques for signal coding and transmission. He is head of the Acoustics Research Department. Fellow, IEEE; fellow, and currently president-elect, Acoustical Society of America, Sigma Xi, Tau Beta Pi.

James W. Fleming, B.S., 1970 and M.S. 1971 in Cer. E., University of Missouri at Rolla; Research Associate University of Missouri, 1971–1972; Bell Laboratories 1972—. Mr. Fleming has worked on the design and properties of PTCR thermistors and other polycrystalline materials for communication applications. Since 1973, he has been developing techniques for the preparation of high melting oxide glass compositions such as those used in lightguides and examining the properties of these glasses. He is currently involved in analysis of dispersion in optical materials and characterizing lightguide core composition profiles. Mr. Fleming is pursuing a Ph.D. in Cer. Sci. at Rutgers University and is a member of the American Ceramic Society.

William G. French, B.A., 1965, University of California, Riverside; Ph.D., 1969, University of Wisconsin; Bell Laboratories, 1969—. Mr. French has worked on fundamental studies of glass as well as glass purification techniques and the development of low loss optical fiber materials. His present interests are concerned with vapor deposition methods for the fabrication of low loss fibers with low dispersion char-

acteristics. Member, Optical Society of America, American Chemical Society, and American Ceramic Society.

David J. Goodman, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., New York University; Ph.D. (E.E.), 1967, Imperial College, London; Bell Laboratories, 1967—. Mr. Goodman has studied various aspects of digital communications, including analog-to-digital conversion, digital signal processing, assessment of the quality of digitally coded speech, and error mechanisms in digital transmission lines. He is Head, Communications Methods Research Department. In 1974 and 1975, he was a Senior Research Fellow at Imperial College, London, England. Member, IEEE.

D. A. Gray, B.S.E.E., 1963, Tufts University; M.S.E.E., 1965, and Ph.D. (E.E.), 1969, Stanford University; Bell Laboratories, 1969—. Mr. Gray has worked on the transmission of microwaves and millimeter waves through rainfall on terrestrial and earth-space paths, and on the measurement of millimeter wave antennas. He is presently working on satellite transmission systems. Member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Commission F of URSI/USNC.

Harold H. Hoffman, New York University; Bell Laboratories. Mr. Hoffman has worked on micro-wave radio relay systems, cordless telephone, satellite orientation, mobile radio and millimeter wave propagation. As a member of the Satellite Systems Research Development, he is presently concerned with the COMSTAR Satellite Beacon Propagation Experiment.

David A. Kleinman, Sc.B. (chemical engineering), 1946, and S.M. (mathematics), 1947, Massachusetts Institute of Technology; Ph.D. (Physics), 1952, Brown University; Brookhaven National Laboratory 1949–1953; Bell Laboratories 1953—. Mr. Kleinman has worked on semiconductor devices, the infrared optical properties of semiconductors, lasers and nonlinear optics, optical telephone receivers, and most recently on electron spin polarization and optical pumping of semiconductors.

Robert P. Leck, A.A.S.E.E., 1968, Middlesex County College; 1969, Rutgers College of Engineering; 1972—, Monmouth College; Bell Laboratories, 1972—. From 1969 to 1972 Mr. Leck was engaged in the design of both digital and analog measurement systems. Since joining Bell Laboratories, he has participated in mobile radio experiments, has done

work on experimental linear amplifiers, and was involved with the design and assembly of the electronics used in the COMSTAR Satellite Beacon Propagation Experiments. He is presently involved in the reduction of data obtained from that experiment, microcomputer-based measurement and control systems, and phase-locked-loop frequency multiplier design. Member, Eta Kappa Nu.

W. E. Legg, Rutgers University, 1945–1949; Bell Laboratories, 1945—. Mr. Legg has worked on dielectric lenses and microwave antennas for radio relay systems. He was involved in the development of Project Echo and Telstar tracking equipment and participated in mobile radio telephone experiments. He is presently engaged in antenna measurements for satellite communication and terrestrial microwave repeaters in the Radio Research Laboratory.

Stephen E. Levinson, B.A. Engineering Sciences, Harvard, 1966; M.S.E.E., Ph.D., University of Rhode Island, 1972 and 1974, respectively; Design Engineer, Electric Boat Division, General Dynamics 1966–1970. J. Willard Gibbs Instructor of Computer Science, Yale University, 1974–1976. Joined Acoustics Research Department, Bell Laboratories, August 1976. Research interests: speech recognition, pattern recognition, theory of computation. Member, IEEE, ACM.

Sing-Hsiung Lin, B.S.E.E., 1963, National Taiwan University; M.S.E.E., 1966, and Ph.D., 1969, University of California, Berkeley; Bell Laboratories, 1969—. At the Electronics Research Laboratory, University of California at Berkeley, Mr. Lin was engaged in research on antennas in plasma media and numerical solutions of antenna problems. Mr. Lin is presently working on wave propagation problems on terrestrial radio systems and earth-satellite radio systems. He was an invited lecturer on microwave radio communication systems in the 1976 Modern Engineering and Technology Seminar in Taipei, Taiwan, Republic of China, sponsored by the Chinese Institute of Engineers. Member, IEEE, Sigma Xi.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. (Physics), 1963, Syracuse University; Research Associate, Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964–. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now

working in the Mathematical Research Center. Member, American Physical Society, IEEE.

Barbara J. McDermott, B.A. (Psychology), 1949, University of Michigan; M.A. (Psychology), 1963, Columbia University; Haskins Laboratories, 1950–1959; Bell Laboratories, 1959—. Ms. McDermott has worked on speech quality evaluation and multidimensional scaling analysis. Member, Acoustical Society of America.

John A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954, and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has done research in various areas of applied mathematics and mathematical physics. He has recently been interested in queuing problems associated with data communications networks. He was a Visiting Professor of Mechanics at Lehigh University during the fall semester of 1968. Member, American Mathematical Society, SIAM, IEEE, Sigma Xi.

Erwin E. Muller, B.S., 1952, Stevens Institute of Technology; M.S., 1954, University of California; Bell Laboratories, 1954—. Mr. Muller has worked on design of ballistic missile-guidance computers, single-sideband long-haul radio systems, and satellite communications systems. He is head of the Transmission Systems Characterization Department, concerned with describing the operational environment of radio and wire-pair transmission systems. Senior member, IEEE.

Donald F. Nelson, B.S., 1952; M.S., 1953; Ph.D., 1959, all in physics, University of Michigan; University of Michigan, 1958–1959; University of Southern California, 1967–1968; Princeton University, 1976; Bell Laboratories, 1959–1967, 1968—. Mr. Nelson's research has included study of coherence properties of the ruby laser, making the first continuously operating ruby laser, determining the emission mechanism and kinetics of electroluminescence in light emitting diodes of gallium phosphide, study of electrooptic modulation of light within optical waveguides at pn junctions, study of Auger-type nonradiative recombination of bound excitons in semiconductors, prediction and experimental confirmation of a new symmetry for the elastooptic interaction, studies of four- and five-wave acoustooptic interactions in crystals, and formulation of a basic theory of nonlinear electroacoustics of dielectric crystals. Fellow American Physical Society; member, Optical Society of America, Acoustical Society of America, Sigma Xi, Phi Beta Kappa, Phi Kappa Phi, and Phi Eta Sigma.

Peter Noll, Dipl.-Ing., 1964, Dr.-Ing. (Electrical Communication Engineering), 1969, Habilitation, 1974, Technical University of Berlin, Germany; Heinrich-Hertz-Institut Berlin-Charlottenburg, 1964–1976. Mr. Noll was initially concerned with the development of electronic telephone exchanges. Since 1970, he has been engaged in research on speech coding and communication theory. During the summers of 1974 to 1977 he was on the Technical Staff of Bell Laboratories. Since 1976, he has been a member of the University of Bremen, Germany, as a Professor of Electrical Engineering and Statistical Communication Theory. Member, Nachrichtentechnische Gesellschaft (NTG), and Verein Deutscher Elektrotechniker (VDE), Germany. Senior member, IEEE.

James C. Parker, Jr., B.S.E., 1965, M.S.E., 1967, and Ph.D., 1970 (all E.E.), University of Michigan. Electronic Countermeasures organization and Radiation Laboratory at University of Michigan, 1965–1968; Electromagnetics Department at Conductron Corporation, 1968–1970; Bell Laboratories, 1970—. Until 1976 Mr. Parker was with the Electromagnetic Interference Department, investigating wideband rf and low-frequency EMC problems. He is presently working on stochastic modeling techniques in the Electrical Protection and Interference Department. Member, IEEE, Eta Kappa Nu, Phi Kappa Phi, Sigma Xi. Registered Professional Engineer in the state of New Jersey.

Aaron E. Rosenberg, S.B. (E.E.) and S.M. (E.E.), 1960, Massachusetts Institute of Technology; Ph.D. (E.E.), 1964, University of Pennsylvania; Bell Laboratories, 1964—. Mr. Rosenberg is presently engaged in studies of systems for man-machine communication-by-voice in the Acoustics Research Department at Bell Laboratories. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi; fellow, Acoustical Society of America; member IEEE and IEEE Acoustics, Speech, and Signal Processing Group Technical Committee on Speech Processing.

A. J. Rustako, Jr., B.S.E.E., 1965, M.S.E.E. 1969, New Jersey Institute of Technology; Bell Laboratories, 1957—. Mr. Rustako, a member of the Satellite Systems Research Department, has been engaged in system and propagation studies in both multipath mobile radio and earth-space satellite propagation media. He is presently concerned with scanning spot beam phased array techniques for satellite communications.

Carlo Scagliola, Dr. Ing. (Electronic Engineering), 1970, University of Pisa, Italy. Mr. Scagliola has been with CSELT (Centro Studi e Laboratori Telecommunicazioni), Turin, Italy since 1970. He has been engaged in adaptive speech coding, assessment of the quality of digitally coded speech and in studies on automatic synthesis of the Italian language. Mr. Scagliola served as a consultant at Bell Laboratories from January 1977 through January 1978.

Jay R. Simpson, B.S. (Glass Science), 1972, Alfred University; Bell Laboratories, 1972—. Mr. Simpson is a member of the materials research laboratory and has been engaged in providing optical fiber transmission measurements for the development of fiber preform fabrication. He has also pursued an interest in inelastic electron tunneling spectroscopy.

David Slepian, M.A., 1946, Ph.D., 1949 (theoretical physics), Harvard University; Bell Laboratories, 1950—. Mr. Slepian is Head of the Mathematical Studies Department within the Mathematics and Statistics Research Center. He has worked in a variety of areas of applied mathematics and has served as a consultant on many Bell System projects. His main fields of applied interest are information theory and applications of probability theory to communication engineering. He is currently also Professor of Electrical Engineering at the University of Hawaii, Honolulu, where he periodically spends time on leave from the Laboratories. Member, National Academy of Sciences, National Academy of Engineering, and Society for Industrial and Applied Mathematics. Fellow, IEEE, the Institute of Mathematical Statistics, and American Association for the Advancement of Science.

Ronald L. Wadsack, B.S.E.E., 1962, University of Washington; M.S.E.E., 1964, New York University; M.S., 1968, Yale University; Ph.D. (Applied Quantum Physics), 1971, Yale University; Bell Laboratories, 1962–1966, 1971—. Mr. Wadsackhas worked in the areas of magnetic and semiconductor memory systems, laser PCM systems, and automated integrated circuit test systems. His current interests are logic simulation, fault modeling, test vector generation, and microprocessors. He is currently an Adjunct Visiting Professor of Electrical Engineering at Tennessee State University. National Science Graduate Fellow, 1966–1969. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Pi Mu Epsilon.

Martin F. Wazowicz, RCA Institute; Western Electric; Bell Laboratories, 1951—. Mr. Wazowicz has worked on such projects as Mark IV, Essex, "X" Band radio propagation experiments, mobile radio, and satellite receiver projects.

Ecaterina Weiszmann, B.S. (Engineering), 1977 Cornell University; M.S. (E.E.), 1978, Cornell University; Engineering Research Center, Western Electric, 1977—. Miss Weiszmann is presently studying laser light interaction with semiconductor materials. Member, IEEE.

Robert W. Wilson, B.A. (Physics), 1957, Rice University; Ph.D. (Physics), 1962, California Institute of Technology; Bell Laboratories, 1963—. At Bell Laboratories Mr. Wilson has made radio astronomical and propagation measurements. In radio astronomy his work includes measurements of the disk component of the galaxy and absolute fluxes of radio sources, discovery of the cosmic background temperature, and discovery and measurement of carbon monoxide and other molecules in interstellar clouds. His propagation measurements include measurements of 10μ and the short centimeter region. He is presently working in both fields. Member, American Astronomical Society, American Physical Society, International Scientific Radio Union, Sigma Xi, Phi Beta Kappa.

G. A. Zimmerman B.S.E.E., 1958, University of Wisconsin; M.S.E.E., 1960, New York University; Bell Laboratories, 1958—. Mr. Zimmerman's work has centered on design of specialized measurement and data acquisition equipment for characterizing radio transmission phenomena and determining their impact on existing and proposed radio telephone equipment. In particular, the detailed data regarding multipath induced amplitude and phase variations obtained from this equipment have impacted FM, single-sideband AM, and digital radio design, engineering, and protection. Member, Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi, and IEEE.