THE BELL SYSTEM TECHNICAL JOURNAL

MARCH 1983
VOL. 62, NO. 3, PART 1

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL (ISSN0005-8580) is published by the American Telephone and Telegraph Company, 195 Broadway, N. Y., N. Y. 10007; C. L. Brown, Chairman and Chief Executive Officer; W. M. Ellinghaus, President; V. A. Dwyer, Vice President and Treasurer; T. O. Davis, Secretary.

The Journal is published in three parts. Part 1, general subjects, is published ten times each year. Part 2, Computing Science and Systems, and Part 3, single-subject issues, are published with Part 1 as the papers become available.

The subscription price includes all three parts. Subscriptions: United States—1 year $35; 2 years $63; 3 years $84; foreign—1 year $45; 2 years $73; 3 years $94. Subscriptions to Part 2 only are $10 ($12 foreign). Single copies of the Journal are available at $5 ($6 foreign). Payment for foreign subscriptions or single copies must be made in United States funds, or by check drawn on a United States bank and made payable to The Bell System Technical Journal and sent to Bell Laboratories, Circulation Dept., Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

Comments on the technical content of any article or brief are welcome. These and other editorial inquiries should be addressed to the Editor, The Bell System Technical Journal, Bell Laboratories, Room 1J-319, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078. Comments and inquiries, whether or not published, shall not be regarded as confidential or otherwise restricted in use and will become the property of the American Telephone and Telegraph Company. Comments selected for publication may be edited for brevity, subject to author approval.

Printed in U.S.A. Second-class postage paid at Short Hills, N. J. 07078 and additional mailing offices. Postmaster: Send address changes to The Bell System Technical Journal, Room 1E-335, 101 J. F. Kennedy Parkway, Short Hills, N. J. 07078.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 62 March 1983 Number 3, Part 1

Measurements of Selective Near-In Sidelobe Reduction of a Pyramidal, Horn-Reflector Antenna

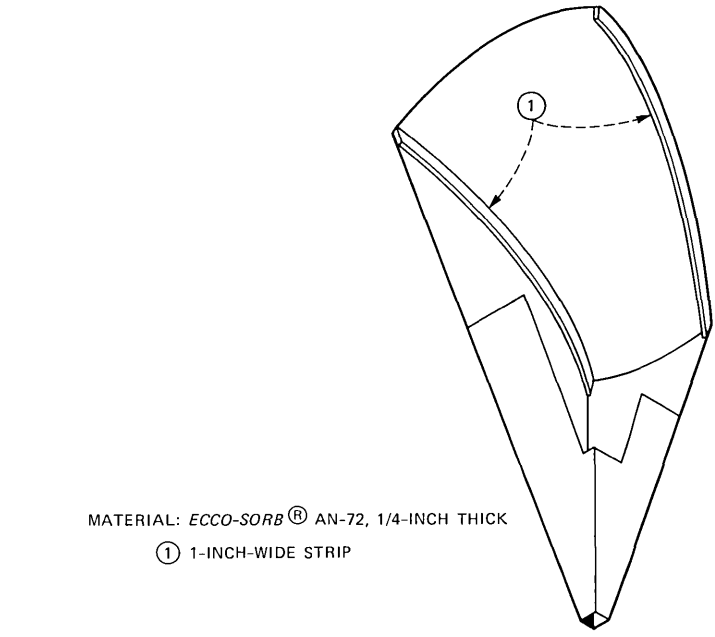
By R. A. SEMPLAK

(Manuscript received January 18, 1982)

This paper describes measurements of a simple but effective means for reducing selective near-in sidelobe levels of a pyramidal horn-reflector antenna by using microwave absorber to modify the electric field across the reflector surface of the horn-reflector antenna. Examples of several modifications (made in the transverse plane for transverse polarization) are discussed and compared with data obtained before modification. Good reductions are achieved in the region of 2 to 6 degrees from the main beam. For example, an improvement was obtained on the order of 4 dB in the first sidelobe level with corresponding improvements of 8 and 10 dB in the angular region of the second and third sidelobe levels. One would expect some of the far-out sidelobe regions to increase; however, the reductions obtained in the levels of the near-in sidelobes could warrant a trade-off.

I. INTRODUCTION

It is well known that radio interference from adjacent paths limits the number of converging routes of a common carrier microwave radio system, and in recent years demands have been made to improve the sidelobe performance of the pyramidal, horn-reflector antenna. The use of blinders[1] (stepped extensions to the side walls of the antenna aperture) provide a degree of far sidelobe reduction, i.e., lobes beyond



(a)

Fig. 1a—A cutaway view of the pyramidal, horn-reflector antenna showing the placement of the 1-inch strips of microwave-absorbing material.

35 degrees from the axis of the main beam. Two methods now exist for dealing with the troublesome reflections from the flat weather cover of the horn-reflector antenna. One eliminates the reflections by using a focused weather cover.[2] The other uses a bottom-edge blinder to direct the reflections upward. An experimental investigation by C. A. Siller of Bell Laboratories showed that lining the sidewalls of the horn-reflector antenna with microwave absorber would reduce the sidelobe levels for angles greater than 40 degrees from the axis of the main beam, but this degraded the near-in sidelobes. A proposal by D. C. Hogg of Bell Laboratories for apodizing the horn-reflector antenna by applying a graded microwave absorber on the weather cover of the antenna might accomplish the desired sidelobe reduction. However, to achieve reduction in the sidelobe level by this technique one would have to suffer not only a reduction in antenna gain of at least 3 dB, but also an increase in the 3-dB beamwidth of the antenna. Such an increase in the beamwidth of the main beam could further degrade system performance. An analysis of available data on path interference, conducted by R. H. Turrin of Bell Laboratories, indicates that the most severe levels of interference occur at small angles (i.e., within ten degrees of the main beam), where the antenna discrimination is less than at large angles.

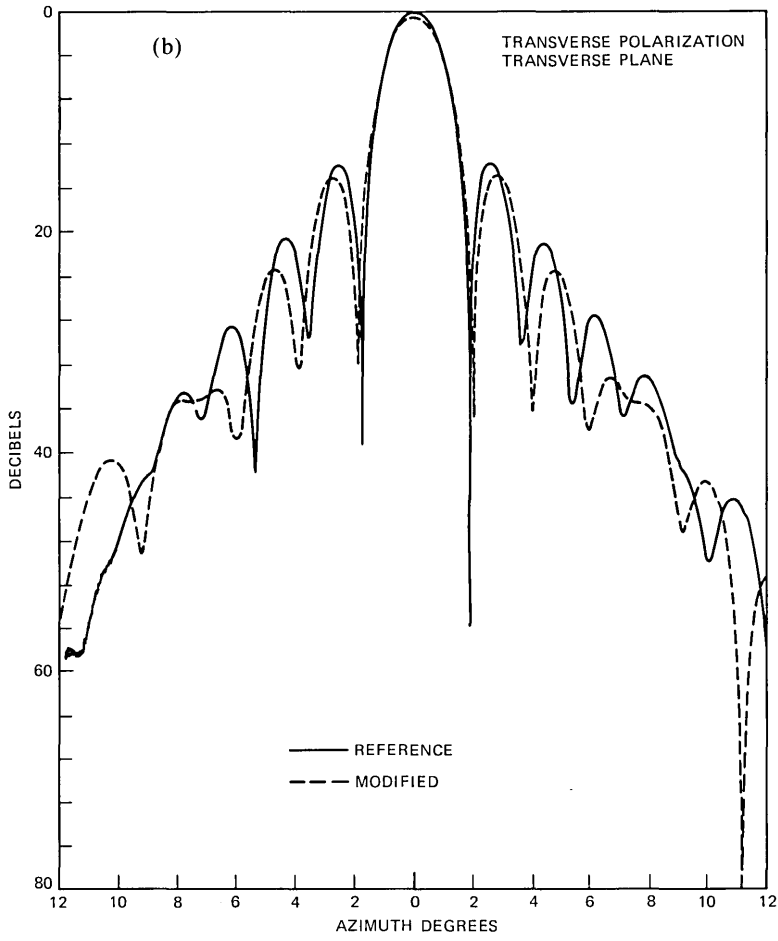
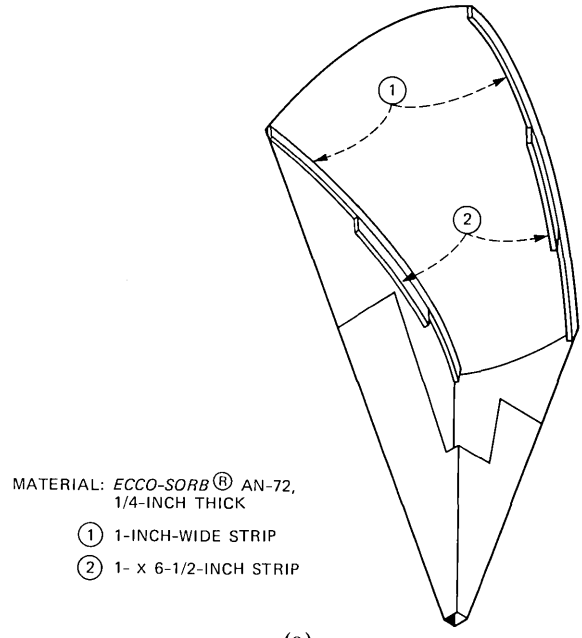


Fig. 1b—Comparison of the radiation patterns for transverse polarization of the modified (with absorber) antenna, the dashed curve, and the solid curve of the unmodified antenna. The gain loss owing to the absorber is on the order of 0.6 dB.

To continue the investigation for ways of reducing the near-in sidelobe levels of the horn-reflector antenna, a scale-model was built. The model has a numerically machined, precision reflector, and the scaling factor is 7.5, which means that measurements made at a frequency of 30 GHz will represent the performance of a full-size antenna measured at a frequency of 4 GHz. The discussion of the measurements made on the scaled model at a frequency of 30 GHz and the comparisons with data obtained by others on full-sized antennas are presented elsewhere.[3]

Inherent in the design of the pyramidal, horn-reflector antenna is a problem that results from illuminating the aperture with a dominant



(a)

Fig. 2a—A cutaway view of the pyramidal, horn-reflector antenna showing an additional strip of absorbing material added to the sidewall of the horn.

waveguide mode.[4] The theoretically obtainable off-axis radiation levels in the transverse plane for transverse polarization are therefore considerably higher than one would like, i.e., they are essentially the equivalent of an aperture with constant illumination.

It should be remembered that longitudinal polarization and longitudinal plane of antenna rotation, for radiation measurements, are aligned with the pyramidal horn axis, i.e., vertically polarized in a terrestrial radio system. Transverse polarization and transverse plane indicate that the electric field and the plane of antenna rotation are perpendicular to the horn axis, i.e., horizontally polarized in a terrestrial radio system.

In view of the information obtained from Siller and Hogg, it appeared that a more fruitful approach to the reduction of the near-in sidelobes would be to take advantage of the reflection occurring at the surface of the paraboloidal section and attempt to modify the electric-field distribution across the surface of the reflector. This could be accomplished by introducing a microwave-absorbing material directly on the surface of the reflector.

In the discussion that follows, the microwave absorbing material* is

* *ECCOSORB* AN-72, a product of Emerson and Cuming, Inc.

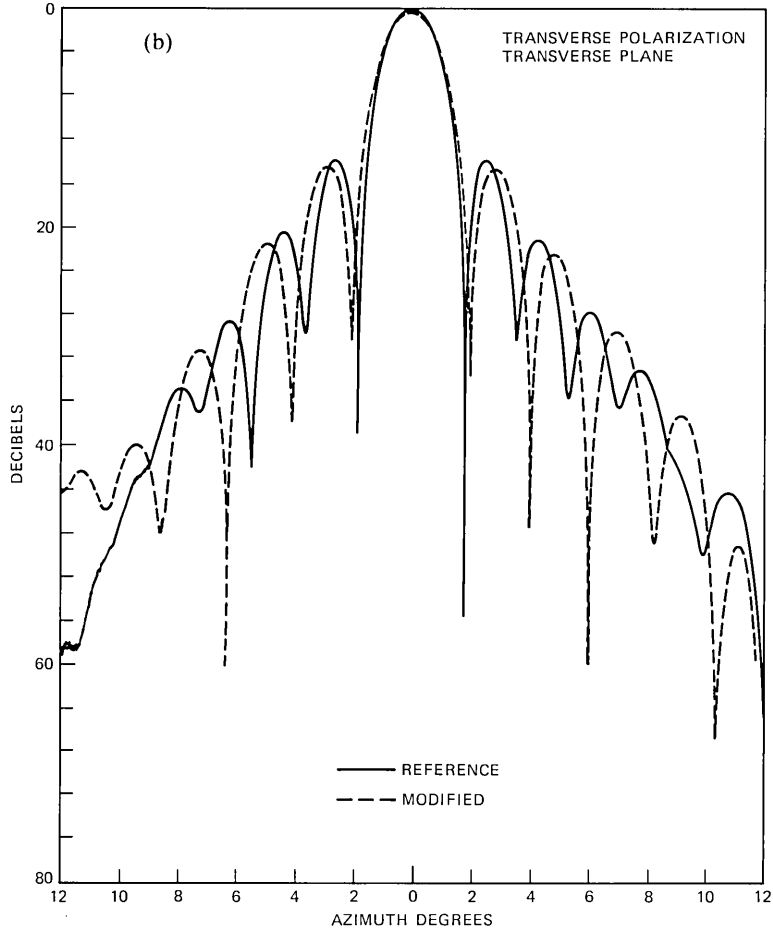


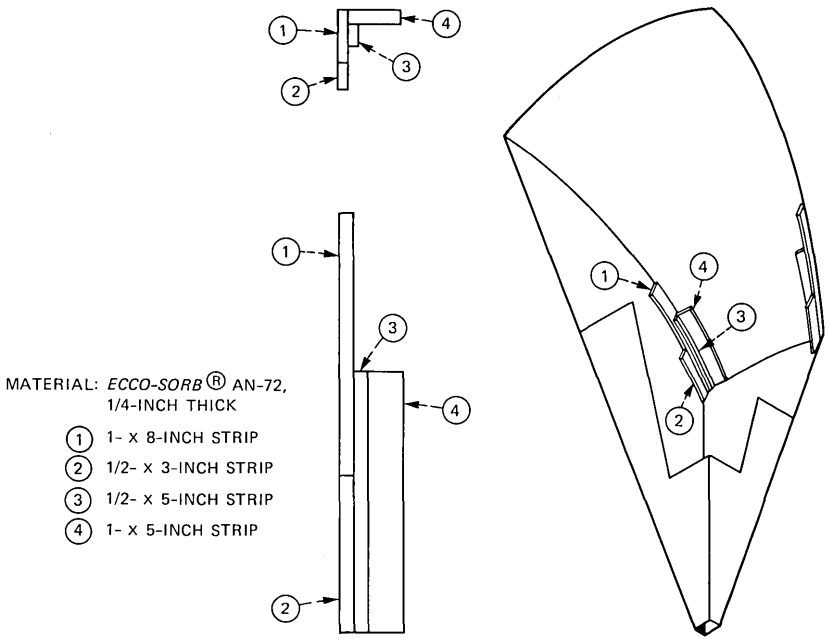
Fig. 2b—Comparison of the radiation patterns for transverse polarization of the modified (with absorber) antenna, the dashed curve, and the solid curve of the unmodified antenna.

attached to the surface of the reflector with a soluble floor-tile cement. The radiation patterns were made at a frequency of 30 GHz in the transverse plane for transverse polarization.

II. DISCUSSION

From the many combinations (of absorbing-material configuration and placement on the surface of the reflector) that were measured, several were selected for discussion here.* For example in Fig. 1a, the

* The examples in the order cited serve to illustrate the evolution of these combinations.



(a)

Fig. 3a—A cutaway view of the pyramidal, horn-reflector antenna showing the placement of strips of microwave absorber in the rear corners on the reflector.

placement of the absorbing material on the reflector is shown in the cutaway view of the horn-reflector antenna. Here, the 1/4-inch-thick edge of the absorbing material is fastened to the reflector surface with the one-inch width of the material fastened to the sidewall of the horn. The absorbing material extends from the back wall to the edge of the aperture. The radiation pattern for this combination is shown by the dashed curve of Fig. 1b; the solid curve is the reference radiation pattern of the horn-reflector antenna, i.e., with nothing introduced into the antenna. In spite of the obvious lowering of the near-in sidelobes, which in this instance were 0.6, 3.0, and 6.0 dB for the first, second, and third sidelobes, respectively, there was no measurable increase in the 3-dB beamwidth. Although there is a gain reduction of 0.5 dB, this has been accounted for when stating the sidelobe reductions. One should note that the modified radiation pattern has very good symmetry and the angular displacements of the maxima for the sidelobes of the modified antenna have increased.

One might say that the observed results are typical of what would be expected when the aperture area is reduced. But consider Fig. 2a.

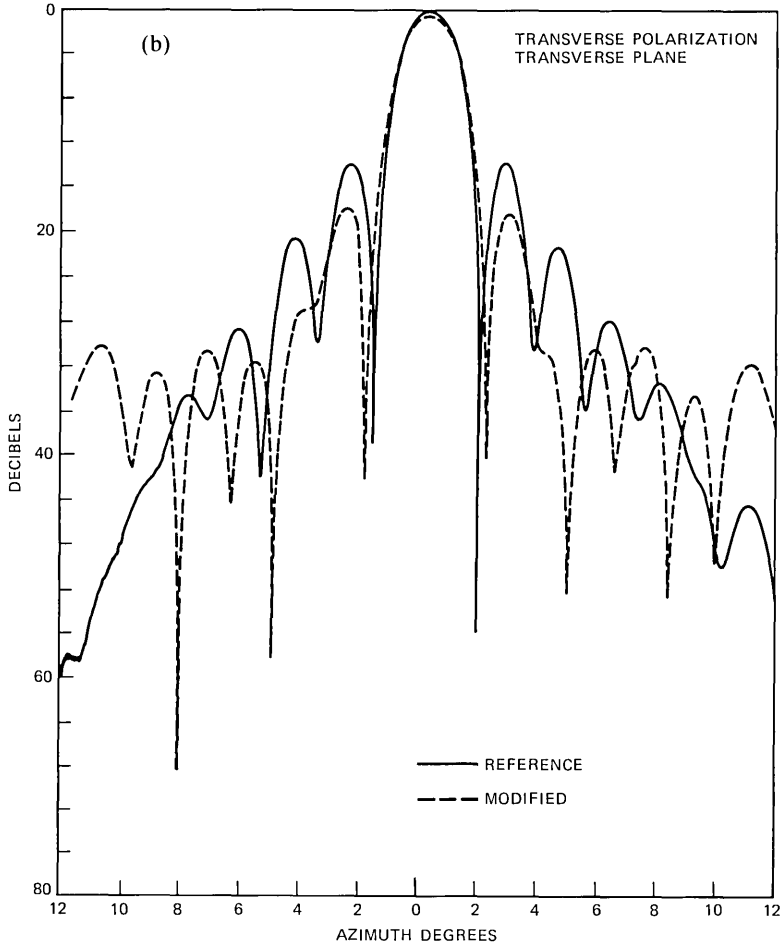
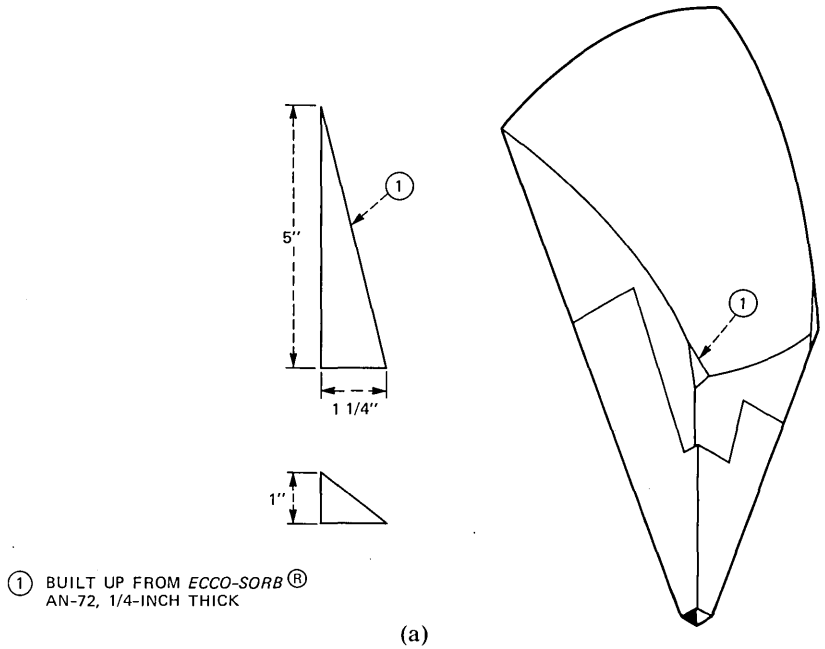


Fig. 3b—Comparison of the radiation patterns for transverse polarization of the modified (with absorber) antenna, the dashed curve, and the solid curve of the unmodified antenna. For this configuration the gain loss is 0.5 dB.

Here, an additional piece of absorber was placed on the sidewalls of the horn. No further reduction in the effective aperture area has taken place. However, the electric field distribution has again been modified. This is clearly seen by the dashed curve Fig. 2b, which shows a slight decrease in the near-in sidelobe levels of the modified antenna. The gain loss is essentially the same, but now there is a just discernible increase in the 3-dB beamwidth of the main beam. There is a further displacement on the sidelobe maxima for the modified antenna. Again the modified radiation pattern shows good symmetry. This configura-



(a)

Fig. 4a—A cutaway view of the pyramidal, horn-reflector antenna showing the placement of a triangular wedge of absorbing material in the rear corners of the reflector.

tion could reduce specific interferences (i.e., at about 6 degrees), but the pattern envelope is slightly degraded.

It soon became apparent that one could achieve very good results by working in areas on the reflector that have the highest electric fields. The cutaway view of Fig. 3a shows the absorbing material confined to the rear corners of the reflector. The rather dramatic effects for this combination are shown by the dashed curve of Fig. 3b. Here, the first sidelobes are reduced by 4 dB, and the reduction out to about 5 degrees is remarkable. Again, the radiation pattern of the modified antenna shows good symmetry. The disconcerting feature of the radiation pattern for this configuration is the increase in the sidelobe levels beyond 8 degrees. In view of the number of edges that were introduced by the absorber, it seems that the increase in the sidelobe level beyond 8 degrees was due in part to scattering from these multiple edges.

An example, for another configuration, where an attempt was made to minimize the number of edges is shown in the cutaway view of Fig. 4a. Indeed, as one can see in Fig. 4b (when compared to Fig. 3b), there is an improvement in the region beyond 8 degrees. Here the loss in gain was of the order 0.2 dB and one can easily see the reduction in

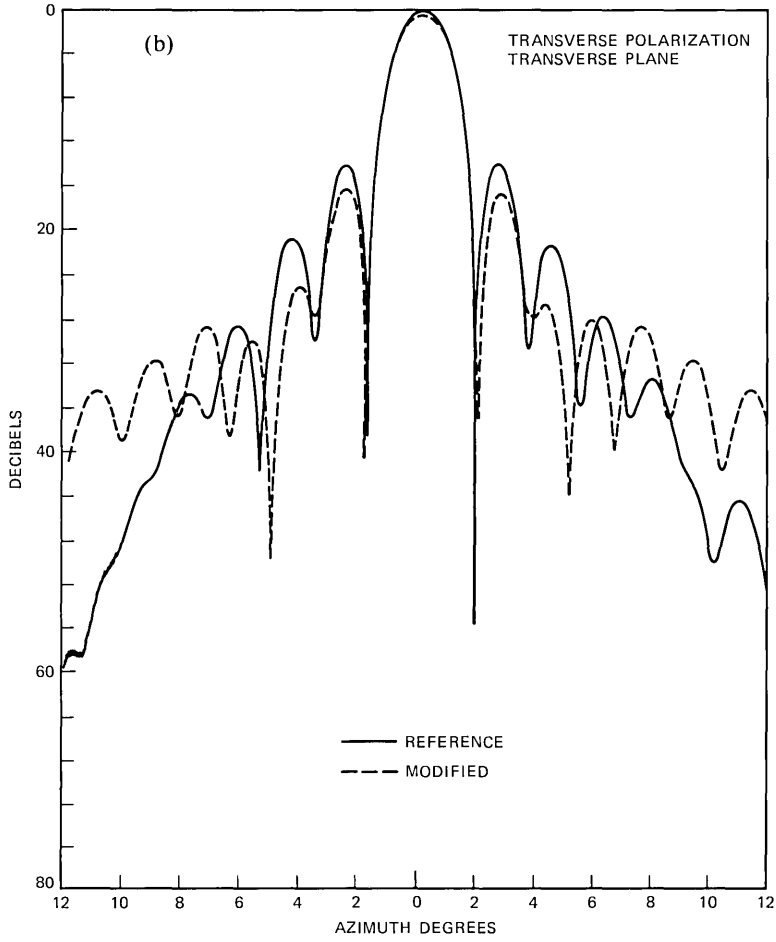
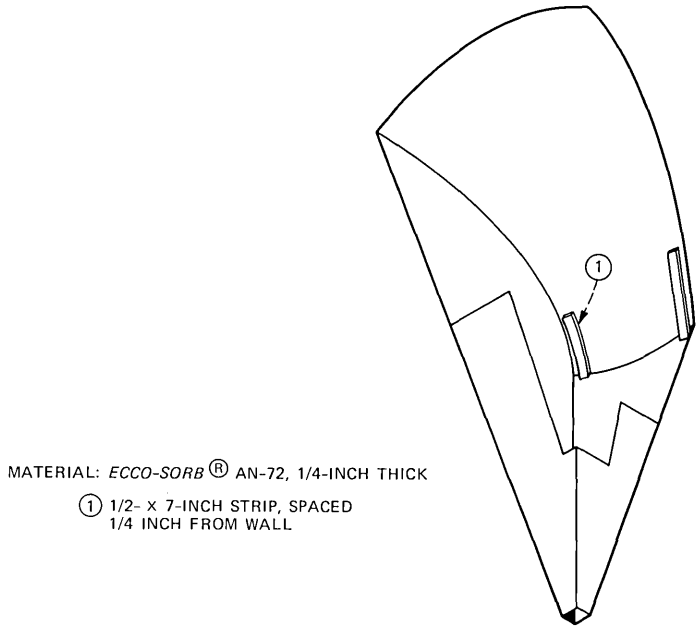


Fig. 4b—Comparison of the radiation patterns for transverse polarization of the modified (with absorber) antenna, the dashed curve, and the solid curve of the unmodified antenna. The gain loss is 0.3 dB.

the first and second sidelobe levels. Again, the symmetry of the pattern is preserved.

The last configuration to be discussed is shown by the cutaway view of Fig. 5a. In this configuration, a simple strip of absorbing material is placed on the reflector and spaced one-quarter inch from the sidewall. As shown by the dashed curve of Fig. 5b, the loss in gain is of the order 0.2 dB, and there is an appreciable reduction in the second and third sidelobe levels. Note the improvement obtained by this configuration in the region beyond 8 degrees when compared with the previously discussed modifications.



(a)

Fig. 5a—A cutaway view of the pyramidal, horn-reflector antenna showing the placement of a strip of absorbing material on the reflector surface.

Radiation patterns made for longitudinal polarization (in the transverse plane) indicate that the modifications introduced on the reflector were transparent for this polarization. Therefore, these patterns are not included here. Measurements in the longitudinal plane were not made, but one would expect to obtain similar performance. Further, the effects on far-out sidelobes were not evaluated.

III. CONCLUSIONS

A simple but effective means for reducing selected near-in sidelobe levels has been developed and the measurements are presented and discussed. For example, if an interferer were located at the maxima of a third sidelobe, then by modifying the electric field distribution on the reflector one could reduce the amplitude at the angle of the third sidelobe at least 10 dB (as shown in Figs. 3b and 5b). Better results can be obtained for reduction in the angular region of the second sidelobe. Even the first sidelobe can be reduced by more than 4 dB. One would expect some of the far-out sidelobe regions to increase; however, the reductions obtained in the levels of the near-in sidelobes could warrant a trade-off.

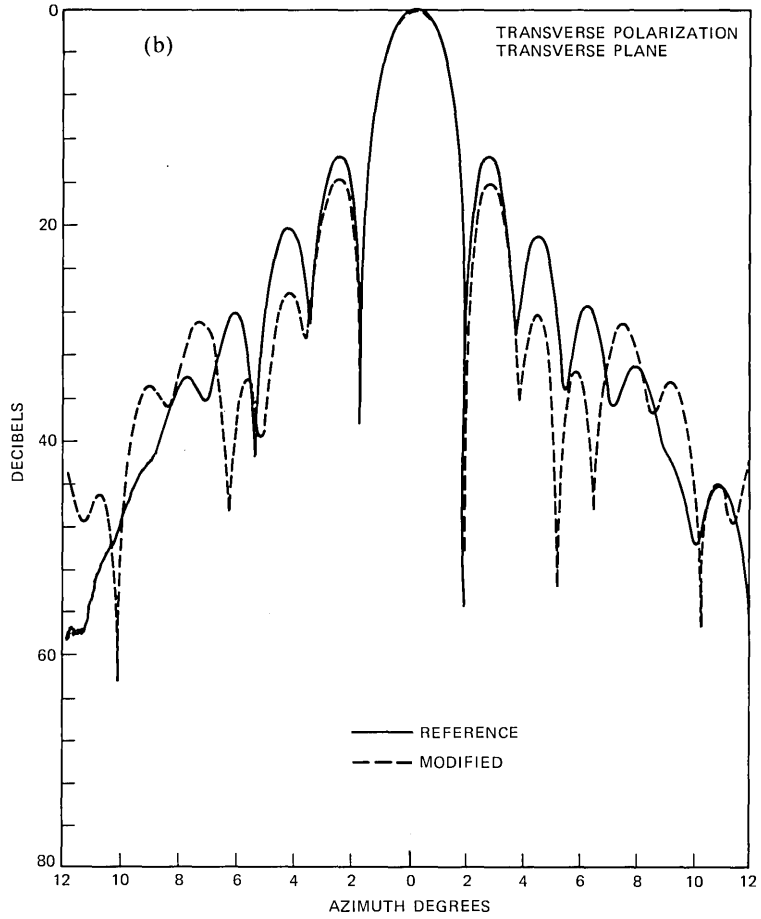


Fig. 5b—Comparison of the radiation patterns for transverse polarization of the modified (with absorber) antenna, the dashed curve, and the solid curve of the unmodified antenna. The gain loss is 0.3 dB.

REFERENCES

1. D. T. Thomas, "Analysis and Design of Elementary Blinders for Large Horn Reflector Antennas," B.S.T.J., *50,* No. 8 (November 1971), pp. 2979–95.
2. R. A. Semplak, "Horn Reflector Antenna-Eliminating Weather Cover Reflections," B.S.T.J., *59,* No. 8 (October 1980), pp. 1333–42.
3. R. A. Semplak, "A 30-GHz Scale-Model, Pyramidal, Horn-Reflector Antenna," B.S.T.J., *58* (July–August 1979), pp. 1551–56.
4. 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. 1095–116.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

An Experimental Study of Atmospheric Optical Transmission

By B. G. KING, P. J. FITZGERALD, and H. A. STEIN

(Manuscript received August 5, 1982)

This paper reports measurements made on a 23-mile, experimental, atmospheric, optical-transmission link for possible use as a standby substitute for microwave radio when the radio suffers severe multipath or obstruction fading. To allow comparison of transmission on a microwave and on an optical path, we used two parallel systems. One, a microwave system at 11 GHz, allowed frequency-selective fading to be measured, and the other, an optical system at 6328Å, allowed amplitude changes of the received optical signal to be obtained. The measured clear-air loss on the optical path is 27 dB. This measurement is made up of 17 dB of atmospheric scattering and 10 dB due to the receiving antennas intercepting only 10 percent of the beam at the receiver. The signal-to-noise ratio, calculated using measured background sky-noise and measured received power, is about 60 dB for a 100-MHz band. The beam diameter was measured to be 32 feet where the signal is down 20 dB. On the single occasion when frequency-selective microwave fading was observed, there was no fading of the optical signal. We find that it is necessary to control the transmitter elevation angle with a servo error signal from the receiver; the azimuth angle needs only occasional manual correction. The optical beam can be automatically reacquired after severe atmospheric attenuation, and that scintillation is usually several decibels, and occasionally as much as 10 dB.

I. INTRODUCTION

We report system parameters measured on a 23-mile, atmospheric, optical-transmission path. The object of the investigation is to determine whether a modulated optical signal transmitted through the atmosphere could be used as a stand-by substitute for a microwave radio link when transmission over the radio path is impaired by clear-air multipath or obstruction fading.

So far, we have learned that:

(*i*) The control of the transmitter elevation angle can be accomplished with an error signal from the receiver. This compensates for atmospheric refractive index gradient changes, which change the curvature of the optical beam.

(*ii*) There is very little need for horizontal beam correction.

(*iii*) The transmitter can be scanned up and down when severe path loss has attenuated the signal below the detection level so as to reestablish the link after the path loss is reduced.

(*iv*) The clear-air path loss is 27 dB. Of this, 17 dB is due to scattering and 10 dB is due to the receiver intercepting only 10 percent of the beam.

(*v*) The received beam is 20 feet in diameter where the received signal is 10 dB below its value at the beam center and 32 feet in diameter where the signal is down 20 dB.

(*vi*) The calculated clear-air peak-signal-to-average-noise ratio, including the background light, is between 57 and 63 dB, using a 20-mW laser.

It has been determined[1] that rain and fog can cause as much as 0.1-dB/ft optical-transmission loss, and, at such times, optical transmission on a 20- to 30-mile path is very difficult. However, it is believed that most microwave multipath fading occurs in relatively clear weather, and to date, our measurements have shown this to be true. Thus, when microwave clear-air fading occurs, the attenuation of the optical signal should be low.[2] Recently, Schiavone[3] has shown that 7 percent of the 30-dB clear-air microwave fade-time at Palmetto, Georgia, occurs jointly with visibilities of 4 miles or less. The remaining 93 percent occurs during times of better visibility.

Further, because of the very narrow beam that can be obtained with light,* there should be no multipath fading of the optical signal.* Again, our limited measurements to date have shown no multipath fading at optical frequencies while multipath fading was occurring at 11 GHz.

II. LOCATION OF THE EXPERIMENT

Figure 1 shows the location of the experiment and also a vertical profile of the path. About three miles of the 23-mile path are over tidal water.

The transmitter is mounted in a cab at the top of a 100-foot steel tower at Murray Hill. The tower is shown in Fig. 2. The optical beam

* For a 1-foot antenna, at 6328Å, the ratio, D/λ, is 4.8×10^5, whereas for a 6-foot antenna at 4 GHz, D/λ is 2.4×10^1. This results in a 3-dB width at the receiver of 0.25 feet for the optical beam and 4800 feet for the 4-GHz beam.

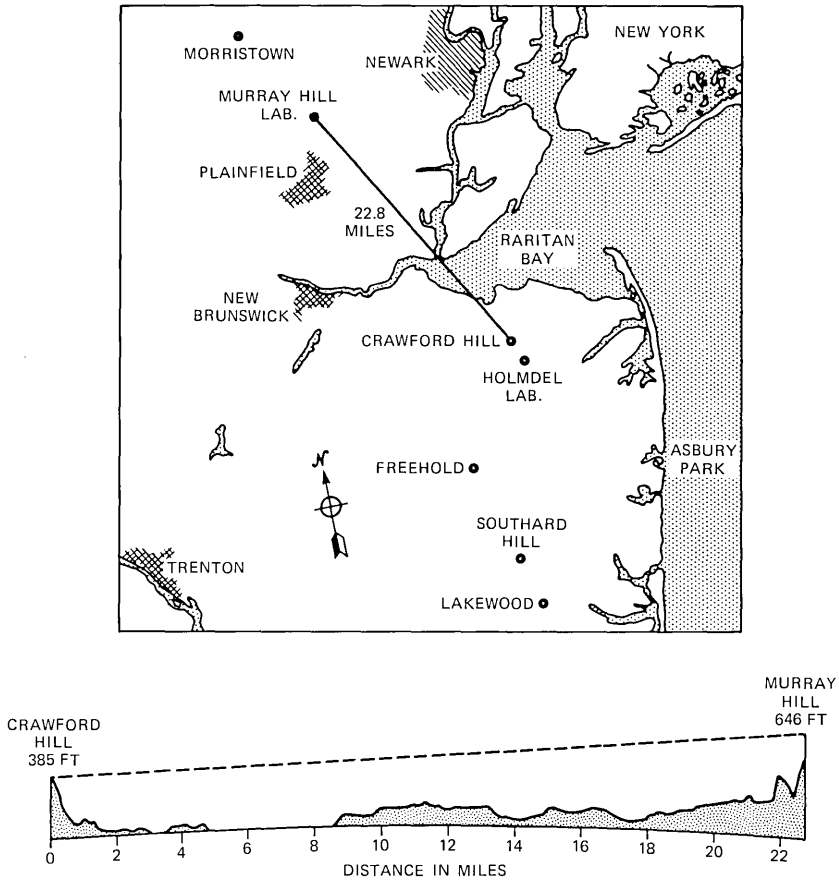


Fig. 1—Map and profile of the transmission path.

is transmitted through the cylinder protruding from the flat surface at the top of the tower. Two microwave antennas on the tower are used to measure multipath fading on the same path (11 and 30 GHz).

The optical receiver at Crawford Hill is shown in Fig. 3. The two microwave antennas on the left of the photograph are part of the microwave multipath-fading experiment.

III. THE TRANSMITTER

Figures 4 and 5 show a diagram and a photograph of the transmitter, respectively. It is composed of a rigid but movable telescope framework, 10-feet long and 1-1/2-feet wide and high, mounted at one end on an orthogonal pair of gimbals and supported on a fulcrum at the other. The telescope gimbals are mounted on one end of a rigid *L*-shaped base, while the other end of the base provides a mount for a

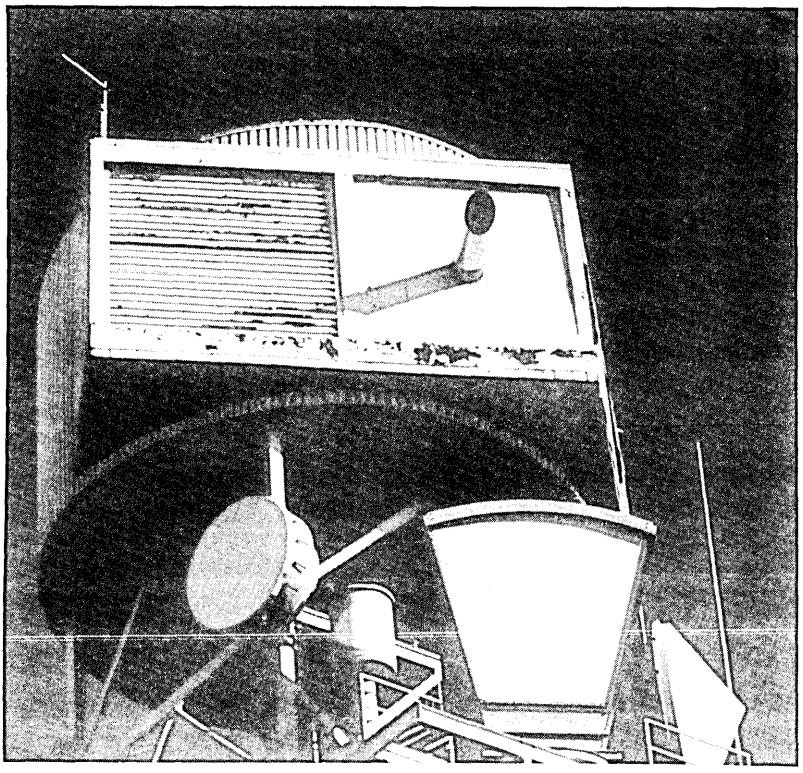


Fig. 2—Transmitter tower at Murray Hill.

second fulcrum. The fulcrum of the telescope and that of the base are about three inches apart. A steel lever pivots on the stationary fulcrum of the base and supports the fulcrum of the telescope. The length of the lever is such that a one-mil (0.001-inch) movement at its free end ("A") causes the telescope to move 0.46 μrad, or a beam motion of about 6.4 feet at the receiver. The lever and fulcrum allow both vertical and horizontal adjustments of the telescope direction.

The position of the free end of the lever is controlled by two small motors. Two tones (900 and 1100 Hz), delivered over a pair of telephone lines from the receiver, control the vertical and horizontal position of the transmitter telescope. The response is relatively slow; a one-second burst of one of the tones moves the optical beam 13.5 inches at the receiver 23 miles away (9.3 μrad/s).

The laser power is 15 mW at 6328Å. The beam is modulated by a chopper wheel, which produces 450 pulses per second, with equal off and on time. To derive a synchronizing signal, light reflected from the chopper is detected and the 450-Hz tone is transmitted over a third

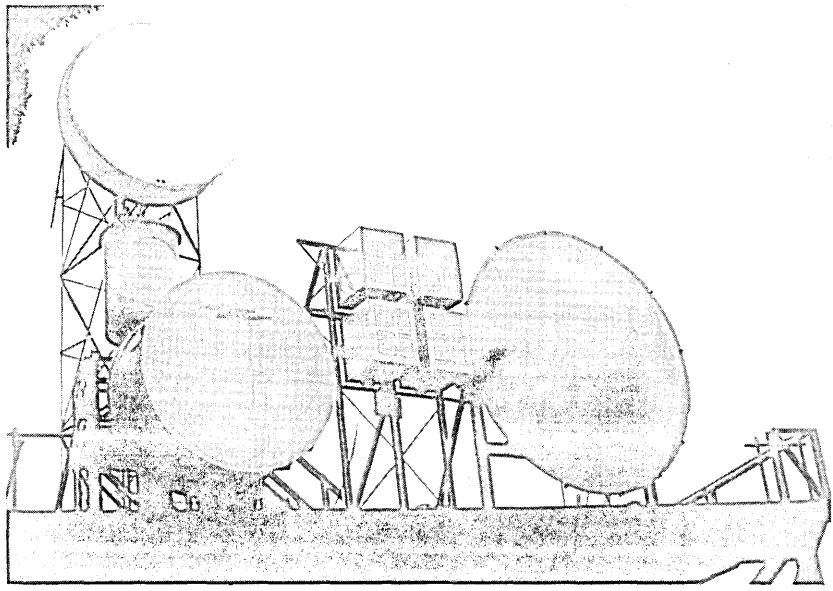


Fig. 3—Receiver assembly at Crawford Hill.

telephone line to the receiver. There, the tone is used to phase-lock the local oscillators of the homodyne detectors.

The rays of the optical signal at the transmitter are shown in Fig. 4. The focal point of the 12-inch mirror is made to fall, by the angle (45 degrees) and position of the small (2-inch diameter) plane mirror inside the telescope, just outside the top of the telescope. To allow sighting through this transmitting telescope for alignment on the receiver, the beam is collimated with a 10-diopter lens. A prism inserted just above this lens bends the beam outward, and allows precise alignment by eye. When used as a transmitter, the collimated beam is focused by a second 10-diopter lens at the focus of a 10-power microscope objective.

IV. THE RECEIVER

The configuration of optical receiving antennas is shown in the photograph of Fig. 3 and the sketch of Fig. 6. Each element is composed of a 24- × 18-inch plastic Fresnel lens mounted in a weathertight, sheet-metal cone, with a photomultiplier at the focus. To reduce background noise, the optical signal passes through a 6328Å filter. This filter reduces the background current 22 dB and the signal 3 dB, for a net 19-dB improvement. In addition, the aperture (0.5 inch) of the photomultiplier is reduced to 0.1 inch with an iris. This reduces the background current 14 dB. Finally, to narrow the part of the sky seen by the photomultiplier, a rectangular bundle of thirty rectangular

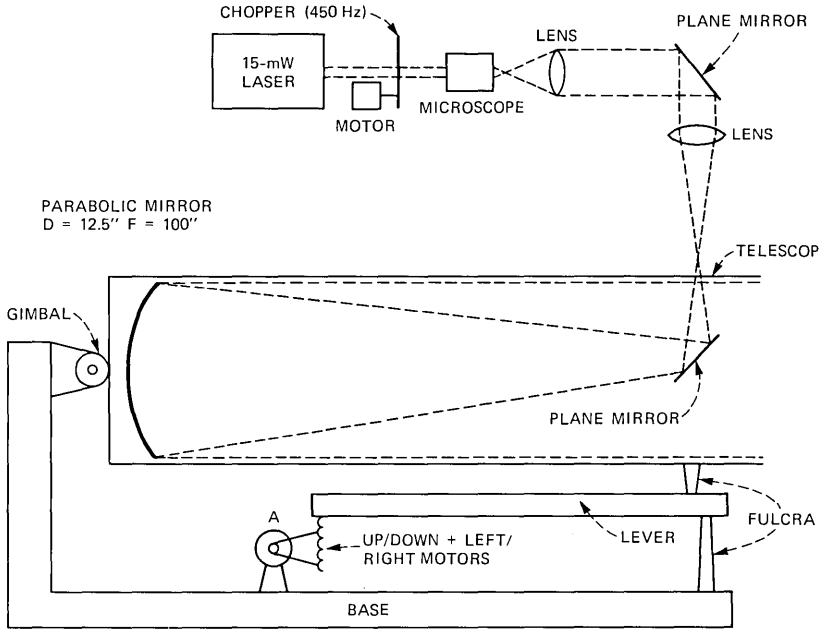


Fig. 4—A diagram of the transmitter.

(4 × 4 inches) tubes is mounted in front of the Fresnel lens. This reduces the background current 4 dB. There are four optical antennas in the receiver mount.

As we see in Fig. 7, the 450-Hz signals from the four photomultipliers are passed through an automatic gain control (AGC) step attenuator (0 to 40 dB in four steps of 10 dB). Each of the four signals is then amplified and band-limited in a homodyne detector. The output signals of the four homodyne detectors are indicated as *A*, *B*, *C*, and *D* in Fig. 7. These outputs are dc voltages linearly proportional to the power of the incoming 450-Hz optical signal. An input signal of 100 μV rms to each homodyne detector produces 10V dc at its output.

The four signals, *A*, *B*, *C*, and *D* in Fig. 7, are combined in a summing amplifier to give an indication of the optical signal strength and the integrated sum is recorded on a chart recorder. The sum is also used to control the AGC and to put the system into a search mode when the sum signal falls below a fixed threshold.

The four signals, *A*, *B*, *C*, and *D* in Fig. 7, are used to control the elevation angle of the transmitter. The signals *A* and *B* from the top two antennas are added, as are the signals *C* and *D* from the bottom. The sum *A* plus *B* is subtracted from the sum *C* plus *D* in a comparator. If the difference is positive, the lower half of the receiver is receiving a stronger signal, and if the difference exceeds a threshold, an 1100-Hz

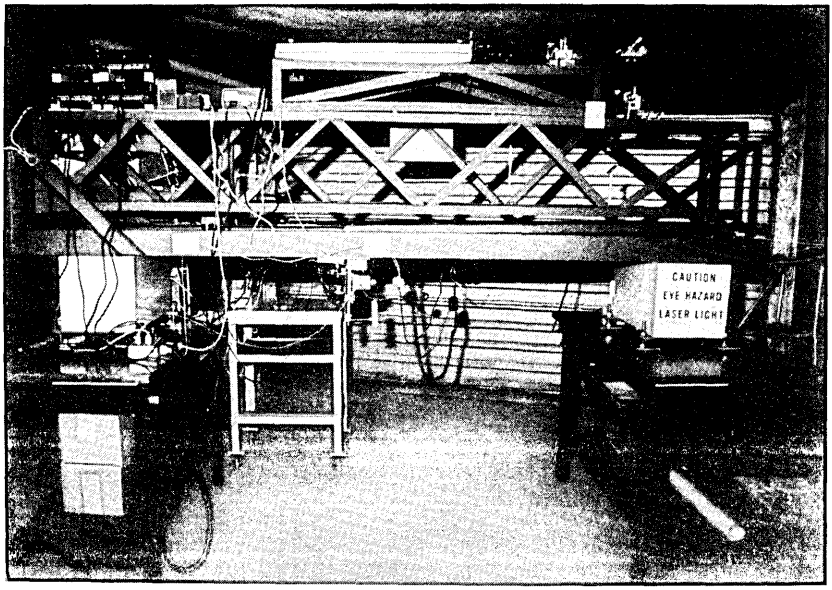


Fig. 5—Transmitter assembly.

tone is applied to the phone line. This tone causes the transmitter to elevate the beam-launch angle. A negative difference causes a 900-Hz tone, which lowers the beam-launch angle. In this way, the transmitter keeps the beam centered on the receiver.

The search control is designed to bring the optical beam back onto the receiver after a deep fade of the optical signal. The search control goes into action when:

(*i*) The sum signal falls below a threshold, and also

(*ii*) All of the attenuators are out of the AGC (zero AGC loss).

In the search mode, a sequence of 200 up commands (1100 Hz) (1 second on, 6 seconds off) is followed by a similar sequence of 200 down commands (900 Hz).* To assure that the search control is in sole command, when it is in operation the up/down servo is disabled. This vertical scanning continues until the sum signal exceeds a threshold, at which time control is handed off to the vertical servo controller. All but one of the thresholds are compared with the sum voltage ($A + B + C + D$). Figure 8 shows the thresholds and the action taken.

First, the summing amplifier saturates at 9.5 volts, which sets the upper limit. The attenuators are inserted, 10 dB at a time and with a

* Two hundred seconds of command takes the transmitter from top to bottom of its range of motion (1495 μrad at the transmitter or 182 feet vertical motion at the receiver).

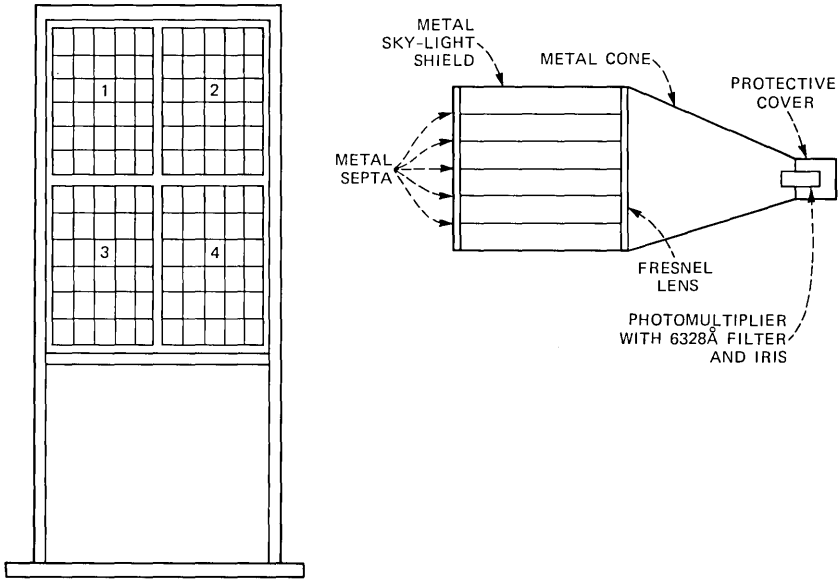


Fig. 6—A diagram of the receiver assembly.

10-second delay between insertions, when the sum voltage is 5 volts or more. The attenuators are taken out, 10 dB at a time and at 10-second intervals, when the sum voltage is 50 millivolts or less. The 100/1 threshold ratio is required to minimize attenuator switching owing to scintillation.

One of the thresholds in Fig. 8 is not compared with the sum voltage but rather with the difference $(A + B)$-$(C + D)$. This difference voltage is the up/down servo command and the minimum threshold is 20 mV. As the sum voltage increases, part of the sum is used to increase this threshold, which minimizes rapid hunting at high signal levels.

The last pair of thresholds turns the search control on and off. The search is turned on when the sum voltage is less than 200 mV and off when it exceeds 500 mV. These rather high thresholds are determined by the clear-air off-axis signal of 150 mV. It was found that, in very clear weather, when the transmitter was moved as far as it could be off the receiver, there was still a sum signal of 150 mV.

V. THE BEAM

5.1 Received beam diameter

The shape of the optical beam at the receiver was determined by incrementally scanning the transmitter and recording the received signal. The distance of motion of the lever in the transmitter (A in Fig.

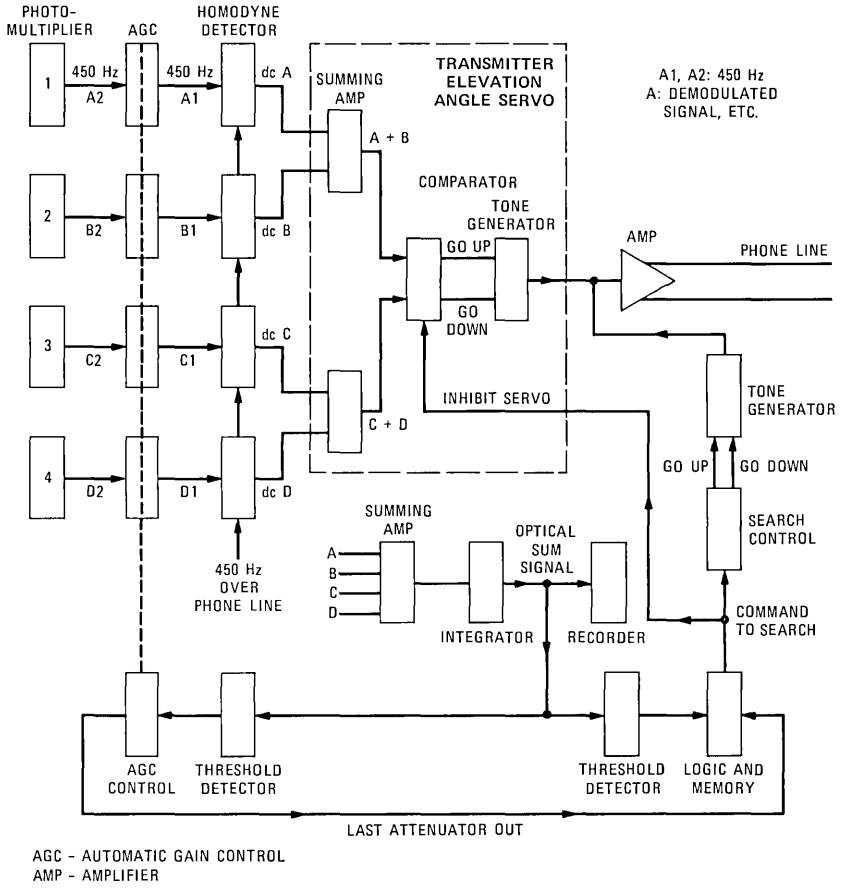


Fig. 7—An electrical diagram of the receiver.

4) was measured and converted to feet of optical-beam motion at the receiver (0.46 μradians of beam motion for each 0.001-inch motion at *A*). This is the abscissa in Fig. 9. The ordinate is simply the sum voltage converted to dB. At the 10-dB point, the beam is about 20 feet in diameter, and at 20 dB, the beam is about 30 feet in diameter.

5.2 Path loss

The receiving antenna is about 3 × 4 feet, and the beam decreases only about 0.5 dB over this area. If all of the power at the receiver is assumed to lie within a 32-foot diameter circle (signal down 20 dB), and the power distribution over this area is calculated using the data of Fig. 9, the power into the 12-square-foot receiver is down about 10 dB below the total power. So, what might be considered the geometric efficiency, that is to say, the fraction of the total optical power in the beam at the receiver that is intercepted by the receiver, is 10 percent.

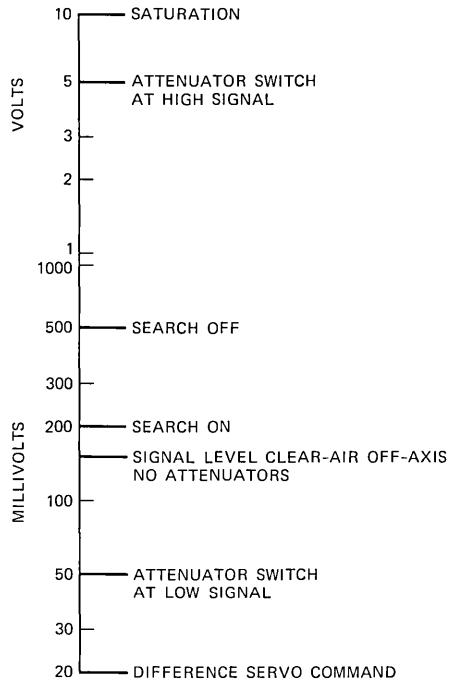


Fig. 8—Optical-system-control voltage levels.

However, there is another source of loss. Energy is scattered by the atmosphere. To determine the amount of this loss, the signal was measured at the end of the transmitter telescope, using a Fresnel lens and photomultiplier identical to those at the receiver. Twenty-seven dB of optical attenuation were inserted in the narrow collimated beam between the two 10-diopter lenses (at the transmitter) to bring the measured signal to the value observed at the receiver. Thus, we conclude that the scattering causes some 17 dB of path loss, and the total path loss, using four receiving antennas, is 27 dB. The *Handbook of Optics* indicates a 16-dB loss on a 23-mile path.[4]

5.3 Vertical beam motion

The transmitter is capable of 1495 μrad of vertical motion. This was consistent with the beam movement measured by Ochs and Lawrence,[5] who reported a maximum of 1023 μrad of vertical motion on a 28-mile path. The path from Murray Hill to Crawford Hill is 23 miles. This 1495 μrad of vertical motion has been found to be adequate to compensate for atmospheric changes on this path. Figure 10a shows a typical early-morning decrease of the transmitter elevation angle. The movement is 478 μrad at a 0.07-μrad/s rate. The largest early-morning

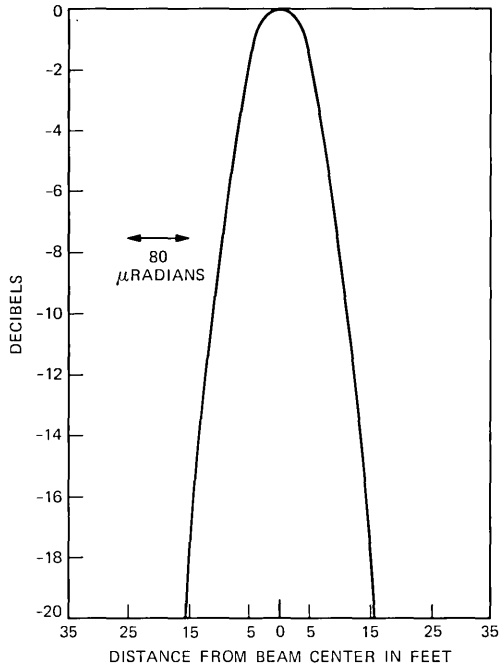


Fig. 9—Optical beam power vs distance from beam center.

movement recorded is shown in Fig. 10b. Here the transmitter moved 990 μrad at a rate of 0.18 μrad/s. The fastest transmitter movement observed was 0.89 μrad/s, well below the 9.3-μrad/s capability of the system. Ochs and Lawrence show a rate of 0.12 μrad/s on a 28-mile path.

The operation of the search control is shown in Fig. 10c. A heavy rainstorm brought the signal below the control threshold at about 16:20 and the transmitter began to search. It takes 39 minutes to complete an up/down scan. At 20:50 the signal was strong enough to turn off the search control, and the system returned to normal operation.

5.4 Attenuator operation

Fog and rain can attenuate the optical signal severely. It was found that the up/down servo control was operable when the signal had faded 40 dB below its clear-air value. To keep the control voltages within a reasonable range, an AGC was used, and is shown in Fig. 7. When the sum-signal exceeds 5 volts a 10-dB attenuator is switched into the signal path, and when it falls below 50 mV a 10-dB attenuator is removed. Figure 11 shows a signal fade which, at 17:25, caused a 10-dB attenuator to be removed.

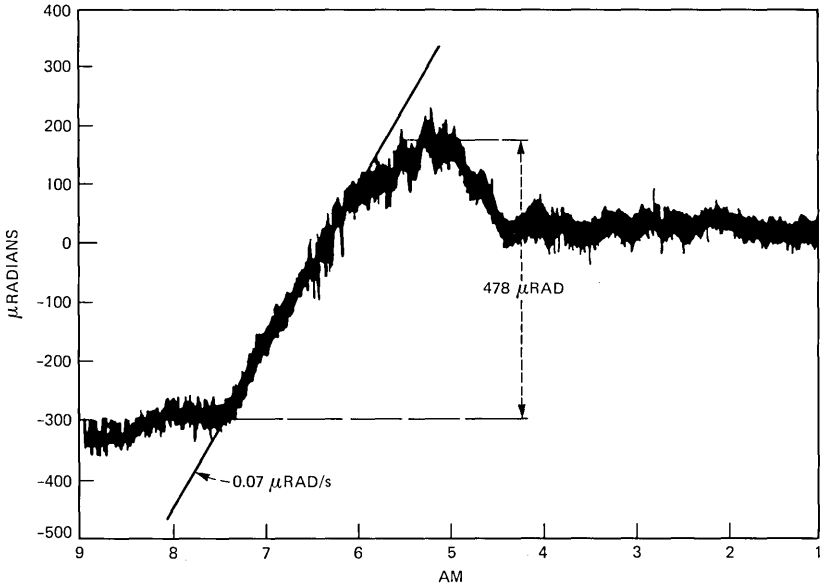


Fig. 10a—Transmitter elevation angle vs time (sunrise).

5.5 Beam microstructure

In addition to the effect of fog and rain, which causes the optical signal to be attenuated by about the same amount for all the receiving antennas, there is fine structure in the path, which causes small variations of the signal in each of the four antennas.[6] A measurement of this phenomenon is shown in Fig. 12a, which shows the output signal of the homodyne detectors with a 6-dB/octave, 0.1-second, time-constant output filter.

Small cells of heated and cooled air cause the velocity of the light to change slightly and differently in each ray. At the receiver, the signal is brighter where the preponderance of rays reaching the antenna are in phase, and dimmer where some of the rays to the antenna are out of phase with the remainder.

Ochs[6] reports that the delayed correlation between horizontal detectors is used to measure wind velocity. If this is done with the data of Fig. 12a, there is about a 0.1-second delay, left to right, in a distance of some 18 inches. This indicates a 10-mph east to west component of wind.

The photomultiplier in Antenna 1, which is seen in Figure 12a, was about 2 to 3 dB less sensitive than the other three. The signal from any one antenna shown in Fig. 12a varies as much as 5 dB in the 2.4-second record, and shows a maximum rate of change of about 18 dB/s. In Fig. 12b, the sums of the top ($A + B$) and bottom ($C + D$)

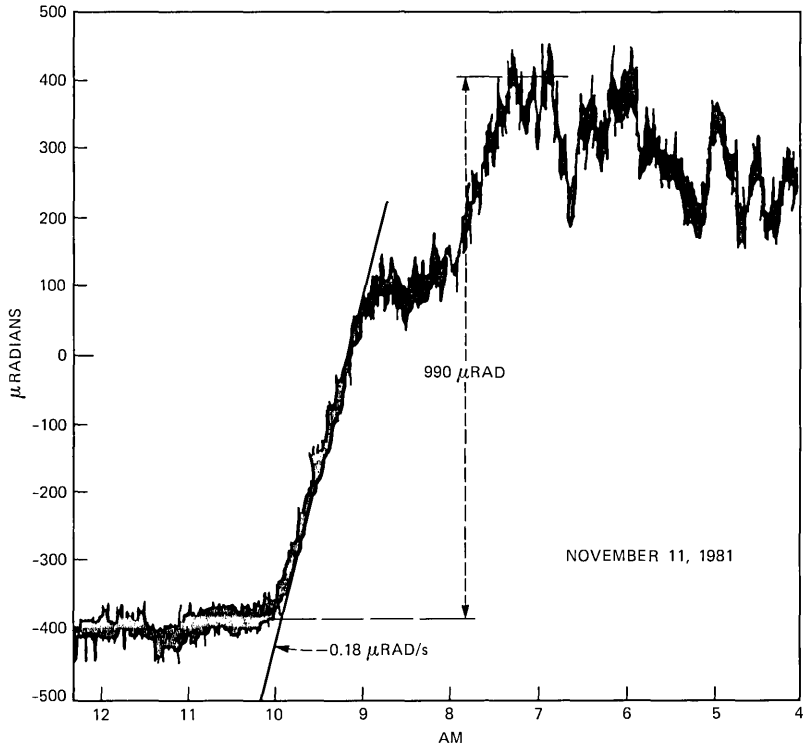


Fig. 10b—Transmitter elevation angle vs time (sunrise).

antennas are shown. At 0.5 second, the up/down servo would command the transmitter to go up, at 1.3 seconds to go down, and at 2 seconds to go up. It is to be observed that the two sums do not change as much (about 3 dB) nor as rapidly (about 11 dB/s) as the separate signals. Finally, the sum of all four signals is shown in Fig. 12c. The sum signal changes about 2 dB with a maximum rate of change of 7 dB/s. It is clear that increasing the size of the receiving antenna reduces the magnitude and rate of rapid signal fluctuations.

5.6 Scintillation

The gross effect of beam microstructure on the sum signal is usually called scintillation. Scintillation is present in microwave radio signals, and is known to increase with frequency, all other things being equal. As was shown in Section 5.5, the optical scintillation decreased with increasing antenna area. Because of the large optical antennas the scintillation measured on the optical path is no worse than that measured on the parallel 30-GHz path, though larger than that on the parallel 11-GHz path. Hannan et al.[7] report optical scintillation as severe as 10 dB.

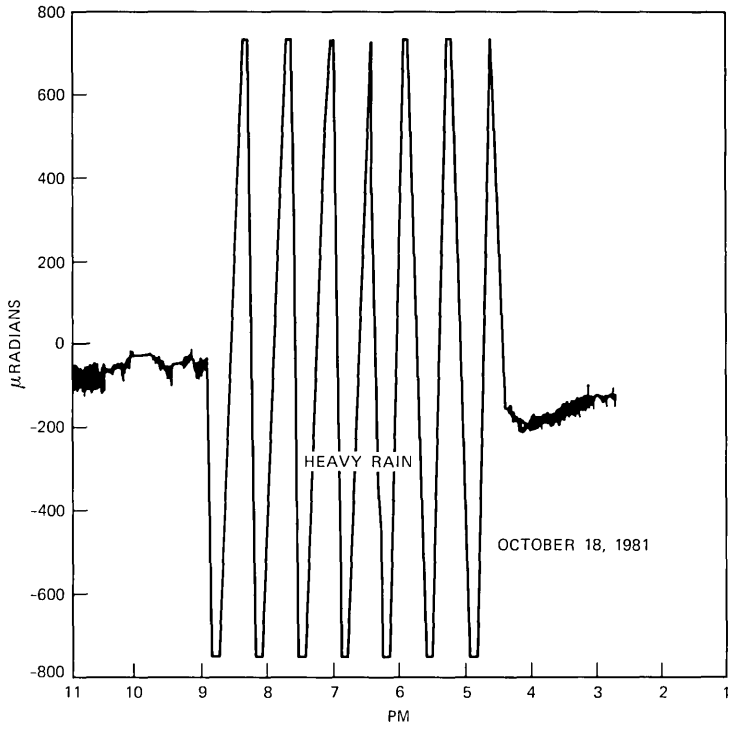


Fig. 10c—Transmitter elevation angle vs time (search mode).

Figure 13 shows the 450-Hz modulation signal, obtained directly from one of the photomultipliers on a clear, cool, sunny day. In Fig. 13a the maximum signal is about 500 mV (across 10 kilohms) and the minimum about 380 mV, a ratio of about a 1.2 dB. Of interest here is the structure seen at the maximum of the negative-going signal. It is clear that even with this large antenna the beam microstructure has components above one kilohertz. In Fig. 13b, taken at a ten times slower rate, the maximum signal is about 500 mV and the minimum about 180, a change of 4 dB. There are several instances where the rate of change is very rapid. For example, the signal drops from 450 mV to 300 mV in two pulses, or a change of 1.8 dB in 4.4 ms (or about 400 dB/s).

5.7 Optical signal during an 11-GHz multipath fade

Figure 14a shows an 11-GHz, clear-air, multipath-fading and enhancement event. The fading is frequency selective with a maximum slope of about 2 dB over the 40-MHz band.

Figure 14b shows the same event along with the optical signal. The normal, clear-air signal level of the 11-GHz event was 2 volts. There is

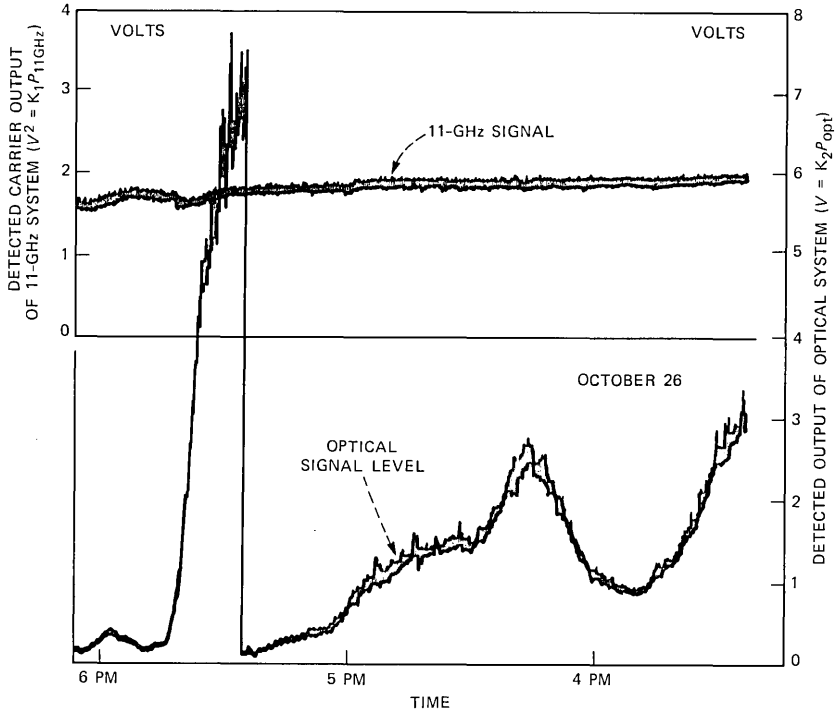


Fig. 11—Microwave and optical signal vs time (attenuator removed).

almost 6 dB of enhancement and about 12 dB of fading. The optical signal shows no fading and very little scintillation.

VI. NOISE

In the preceding sections it has been shown that a beam of light can be transmitted 23 miles. By using feedback from the receiver to control the transmitter elevation angle, the rather narrow beam (2.4×10^{-4} radians between −20 dB levels) can be kept centered on the receiver. And finally, by scanning the transmitter vertically, transmitter elevation control can be reestablished after a deep fade.

There is still the question of the ability of the optical system to satisfactorily carry information at the rates necessary. As is shown in the appendix, there are modulators available that will impress the information of a microwave intermediate frequency (IF) signal onto the light beam. The remaining question is, can the signal be received with sufficient accuracy to provide a satisfactory replica of the impressed signal?

The photomultipliers are fast enough for either frequency modulation (FM) or pulse code modulation (PCM), having a 3-ns rise time. If

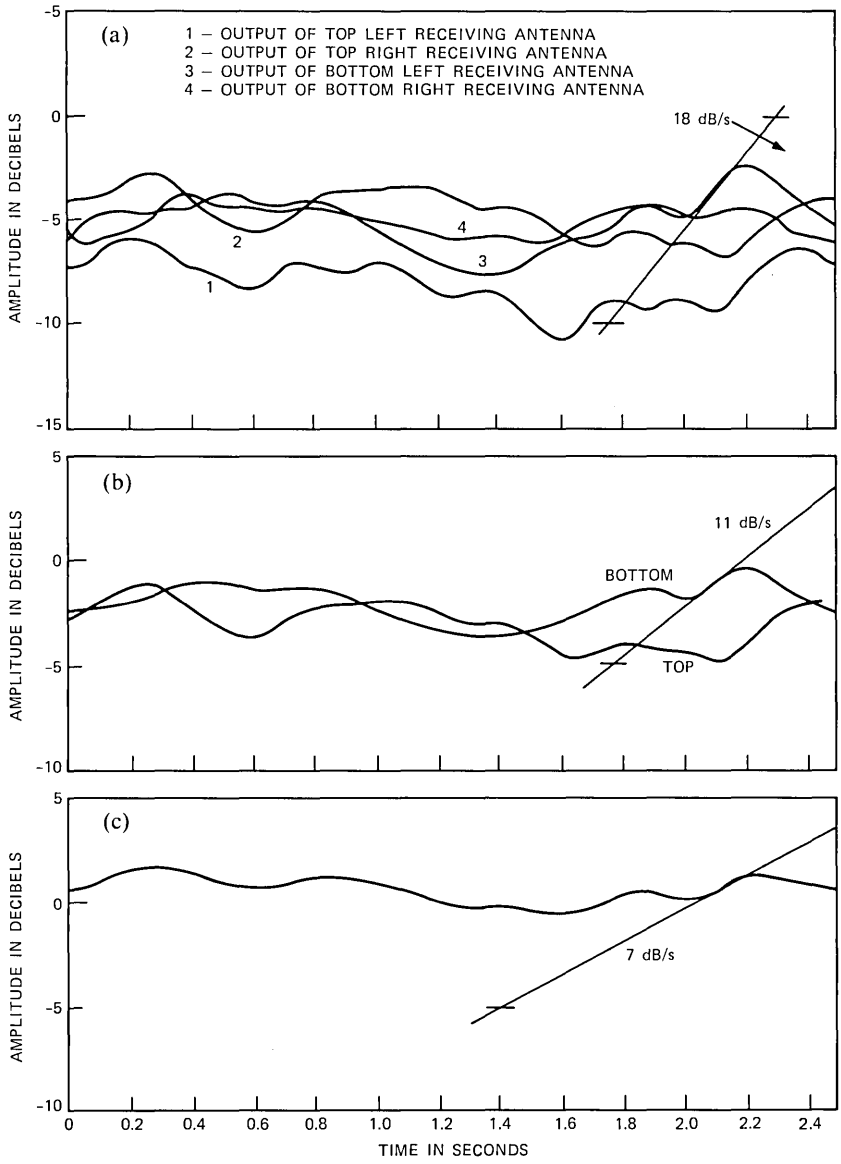


Fig. 12—(a) Demodulated output of four individual photomultipliers vs time. (b) Demodulated output of upper and lower photomultiplier pairs vs time. (c) Demodulated output of sum signal of four photomultipliers vs time.

the 4-MHz baseband signal is sampled at 8 MHz and encoded into 8-bit PCM,*[8] the rate is only 64 MHz. This is well within the 100-MHz capability of the modulator and photomultiplier. In a commercial

* Because noise is spread over the entire band, 8-bit PCM is adequate.

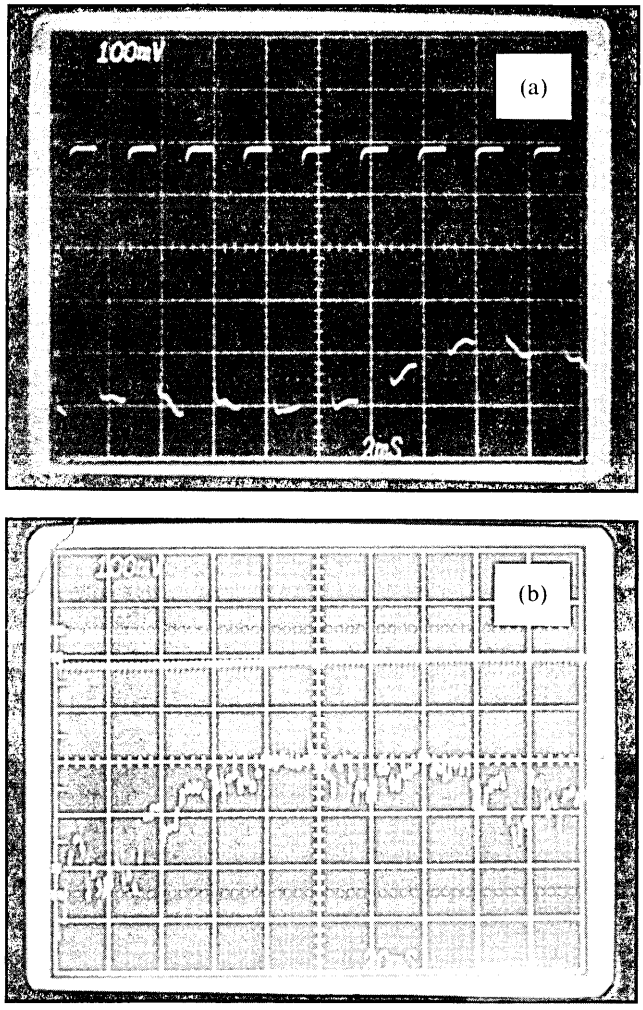


Fig. 13—(a) The ac of one photomultiplier vs time (2 ms/div). (b) The ac of one photomultiplier vs time (20 ms/div).

system, the photomultipliers would almost certainly be replaced with p-i-n diodes, and they are even faster.

So the only question is the noise generated in the receiver. There are four sources of noise in an atmospheric receiver:

(*i*) Thermal noise
(*ii*) Shot noise of the signal current
(*iii*) Shot noise of the current because of background illumination
(*iv*) Shot noise of the dark current.

The thermal noise has a mean-square value of

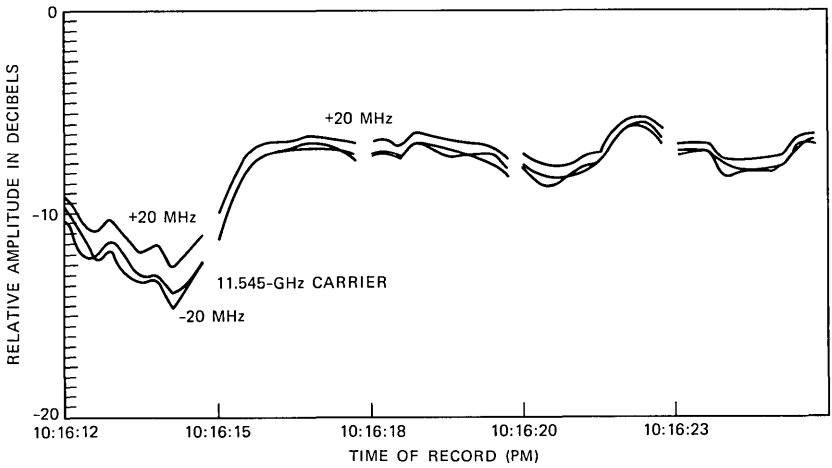


Fig. 14a—Detected amplitude of 11-GHz tones vs time.

$$\overline{i^2} = \frac{4KTB}{R},$$

where K is Boltzman's constant (1.38×10^{-23}), T is temperature (300°K), B is the bandwidth (10^8 Hz), and R is the amplifier input resistance.

The thermal noise can be made very small by using a transimpedance amplifier. But we will calculate the noise using an input resistance ranging from 50 to 1000Ω, which with the interelectrode capacitance of 3 pF gives a time constant 0.15 to 3 ns.

The average current squared ($\overline{i^2_{sig}}$) of the shot noise of the signal current (I_{sig}) is simply

$$\overline{i^2_{sig}} = 2e\frac{I_{sig}}{2}B,$$

where e is the electron charge (1.6×10^{-19}) coulombs.

The noise owing to the background illumination is

$$\overline{I^2_{back}} = 2eI_{back}B,$$

where I_{back} is the dc current due to background illumination.

By the use of the hood in front of the Fresnel lens and the iris and red filter at the photomultiplier, the background illumination, under the worst condition, causes three times as much current as the average signal current if the anode voltage is limited to 800 volts (maximum anode voltage is 1500). Therefore, $I_{back\,max} = 3\,I_{sig}/2$. Most of the time, the two currents are about equal (i.e., $I_{back} = I_{sig}/2$). Of course, at night, there is virtually no background current. The peak signal current is

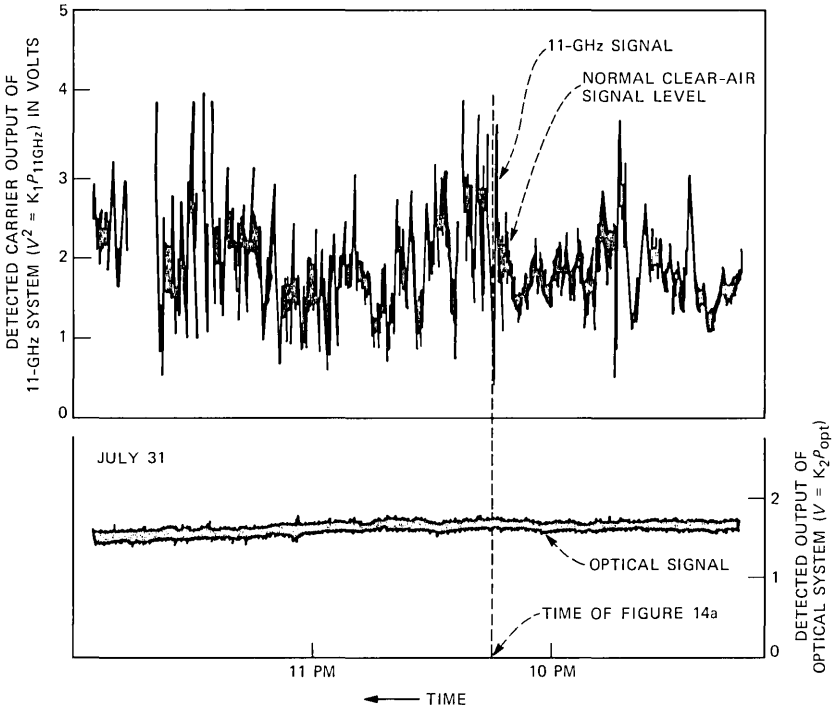


Fig. 14b—Detected amplitude of 11-GHz carrier and optical signal vs time.

about 50 μA and the highest background current was 75 μA. The dark current is three orders of magnitude smaller than the signal current, and can be ignored.

A useful measure of a receiver is the signal-to-noise ratio, and we can calculate it for this receiver with the information we have.

Figure 13b shows the output of one photomultiplier on a typical day. The voltage across a 10-kilohm resistor varies between 500 mV and 175 mV. The average, without scintillation, is 340 mV or a current of 34 μA, and from the four antennas the average sum is 136 μA. The background current was about equal to the average signal current at the time of the measurement. The peak-signal-to-average-noise ratio for an on/off PCM signal can be written as

$$\frac{S}{N} = \frac{(I_{sig})^2}{2e\left(\frac{1}{2} I_{sig}\right) B \left(1 + \frac{2I_{back}}{I_{sig}}\right) + \frac{4KTB}{R}}.$$

The calculated peak-signal-to-average-noise ratio S/N, using an average signal current of 136 μA, is shown in Table I.

Scintillation can cause a considerable variation of received signal. It

Table I—Calculated peak-signal-to-average-noise ratio

$I_{back}/I_{sig\ avg}$	R (Ω)	Signal to noise (dB)
0	50	58
0	100	60
0	1000	67
1	50	57
1	100	60
1	1000	65
3	50	57
3	100	59
3	1000	63

usually amounts to plus or minus several decibels. Under severe conditions it has been seen to cause ±10 dB variation of received signal.[7]

VII. CONCLUSIONS

We have shown that:

(*i*) The elevation angle of an optical transmitter must be continuously adjusted to compensate for beam-curvature changes in the atmosphere, and that this can be done by a servo signal from the receiver.

(*ii*) The azimuth angle of an optical transmitter need not be corrected on a continuous basis. Occasional manual adjustment (every month or so) is adequate.

(*iii*) Heavy fog or rain will cause a total loss of optical signal, but the optical system can be reestablished by scanning the transmitter vertically.

(*iv*) The clear-air path loss is 27 dB on a 23-mile path.

(*v*) The received beam diameter is 32 feet, where the power is 20 dB below the level at the center of the beam.

(*vi*) The sky light can cause as much as three times as much current in the receiver as the average signal current, and this gives a system peak-signal-to-average-noise ratio of about 60 dB.

(*vii*) Modulators exist that should comfortably accommodate either the microwave IF signal (centered at 70 MHz) or a digitized version of the 4-MHz baseband.

(*viii*) The received signal scintillates several decibels most of the time, and can get as severe as 10 dB. The frequency rate of scintillation contains components up to several kilohertz.

(*ix*) During a clear-air microwave fade, the optical signal did not fade, and in fact, had much less scintillation than usual.

REFERENCES

1. H. J. Schulte and B. G. King, unpublished work.
2. B. G. King and L. J. Greenstein, unpublished work.
3. A. Schiavone, unpublished work.
4. *Handbook of Optics*, New York: McGraw-Hill, 1978, pp. 14.16–17.
5. G. K. Ochs and R. S. Lawrence, "Measurements of Laser Beam Spread and Curvature Over Near Horizontal Atmospheric Paths," U.S. Department of Commerce/Environmental Sciences Service Administration, February 1969, ERL 106-WPL 6.
6. G. K. Ochs, private communication.
7. W. J. Hannan, J. Bordogna, and D. Karlsons, "Practical Aspects of Injection Laser Communication Systems," RCA Rev., *27*, No. 4 (December 1967), pp. 609–19.
8. J. C. Candy, private communication.

APPENDIX

Modulation of Light

There are two well-developed optical-modulation techniques. Both involve transparent solids that allow an electrical signal applied to the solid to modify the transmission of light through the solid. In one, the transparent body is mechanically driven by a piezoelectric driver and the pressure standing-wave is employed to form an optical diffraction grating in the body of the solid. This grating causes part of the light to be bent out of its unmodified path and this deflected part of the beam is used as the modulated optical signal. The frequency of the signal that causes this grating is relatively high, on the order of 100 MHz, and is resonant in the crystal. Digital or analogue modulation is impressed on the light beam by either keying the high-frequency signal off and on or linearly modulating its magnitude. Because of the time required for the resonance to build up and to decay, the modulating signal is usually much smaller than the resonant frequency, on the order of a few megabits or megacycles.

The other type of modulator uses a crystal in which the polarization angle of the light is rotated by the electrical signal. The crystal tends to be square in cross section (about 1/2 mm on a side) and about 1 cm long. Metal is plated onto two opposite long surfaces, and the signal voltage is applied to these. The light is passed through the long axis of the crystal in a path parallel to the plated surfaces. The light coming out of the crystal has some polarization, and is passed through a polarizer plate with its surface normal to the optical beam.

When voltage is applied, the polarization of the light out of the crystal rotates in proportion to the voltage applied, and after passing through the output polarizer, the output optical power is changed. The process is continuous, that is to say, the light can be brought back to the original polarization by applying enough voltage. The power out of the output polarizer varies as

$$P_{\text{out}} = \mathrm{K}_1 P_{\text{in}} \sin^2 \mathrm{K}_2 V_{\text{in}}, \tag{1}$$

where K_1 is the maximum through-transmission, $K_2 = 2\pi/V_{360}$, and V_{360} is the voltage needed to cause 360° of phase rotation. If the voltage applied is a sinewave $V_{in} = V_0 \sin \omega t$, and the crystal is biased with a dc voltage, V_B, then the output power is

$$\begin{aligned} P_{out} &= K_1 P_{in} \sin^2(K_2 V_B + K_2 V_0 \sin \omega t) \\ &= \frac{K_1 P_{in}}{2}[1 - \cos 2(K_2 V_B + K_2 V_0 \sin \omega t)] \\ &= \frac{K_1 P_{in}}{2}[1 - \cos(2K_2 V_B)\cos(2K_2 V_0 \sin \omega t) \\ &\quad + \sin(2K_2 V_B)\sin(2K_2 V_0 \sin \omega t)]. \end{aligned} \tag{2}$$

It is well known that

$$\cos(a \sin\theta) = J_0(a) + 2J_2(a)\cos 2\theta + 2J_4(a)\cos 4\theta + \cdots$$

and

$$\sin(a \sin\theta) = 2J_1(a)\sin\theta + 2J_3(a)\sin 3\theta + 2J_5(a)\sin 5\theta + \cdots.$$

Let $2K_2V_B = b$ and $2K_2V_0 = a$. Then

$$\begin{aligned} P_{out} = \frac{K_1 P_{in}}{2} &\{1 - \cos b[J_0(a) + 2J_2(a)\cos 2\omega t + \cdots] \\ &+ \sin b[2J_1(a)\sin \omega t + 2J_3(a)\sin 3\omega t + \cdots]\}. \end{aligned} \tag{3}$$

To avoid all even harmonics, set $b = \pi/2$

$$\begin{aligned} b = 2K_2 V_B &= \frac{\pi}{2} = 2\frac{2\pi}{V_{360}} V_B \\ V_B &= \frac{V_{360}}{8}. \end{aligned} \tag{4}$$

With this bias, only the fundamental and its odd harmonics are present in the modulated signal.

The fraction of the optical power going into the harmonics is determined by the amplitude of the impressed signal, V_0. The argument of the Bessel functions is

$$a = \frac{\pi}{2}\frac{V_0}{V_B}. \tag{5}$$

Because of the nonlinearity of the modulator there is a relatively strong (−27 dB) third harmonic at only 20-percent modulation of the fundamental [$m = 2J_1(4\pi V_0/V_{360})$].

However, such nonlinearity has little effect on an FM signal so long as the amplitude of the FM signal is constant. The fundamental and its sidebands can be filtered from the higher harmonics, and with no

second harmonic present, this is easily done with a microwave intermediate frequency (IF) signal. If the IF signal is centered at 70 MHz, the deviation is ±4 MHz and the baseband signal bandwidth is 4 MHz. The highest sideband of the fundamental, using Carson's rule, is $f_c + \Delta f + f_B = 70 + 8 + 4 = 82$ MHz. The lowest sideband of the third harmonic is $3f_c - 3\Delta f - f_B = 210 - 24 - 4 = 182$ MHz. This should be easily taken care of by the filter in a microwave receiver.

All of the above assumes that the modulation is FM. This requires that the bias be set at $V_{360}/8$. However, binary PCM transmission calls for setting the bias so that no light is transmitted when the signal is a zero and maximum light is transmitted when the signal is a one. Thus, for PCM the bias V_B is set at zero and the digital signal "one" is set at $V_B = V_{360}/4$. The harmonic generation during the transition is of little concern.

Acousto-optic and electro-optic modulators are commercially available. The acousto-optic modulators allow modulation up to about 3.5 MHz with rise and fall times of 120 ns, which is inadequate for most microwave IF signals. The modulation efficiency is about 85 percent, and the extinction ratio is about 1000:1. The electro-optic modulators allow modulation up to 100 MHz with rise and fall times of 3.5 ns. The depth of modulation at 6328Å is 85 percent and the extinction ratio is 500:1. The electro-optic modulators require a feedback bias control.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

Star Network With Collision-Avoidance Circuits

By A. ALBANESE

(Manuscript received July 15, 1982)

This paper describes the implementation of a multiple-access star network that uses a new collision-avoidance circuit to avoid collision of packets and to take full advantage of wideband transmission systems. The analysis shows that the resulting network has a channel capacity equal to 1, has good stability under heavy traffic conditions, allows transmission of packets shorter than the network round trip time, and provides distributed switching. Collision-avoidance circuits have been built and operated up to 50 Mb/s.

I. INTRODUCTION

Carrier-Sense Multiple-Access local-area networks with collision detection (CSMA-CD)[1] have been implemented using lightwave technology,[2] but the use of lightguides is questioned because the networks do not benefit from the full bandwidth of the lightguides. In these networks, the lightguides are used to reduce ground-loop voltage, electromagnetic interference (EMI), size, and weight of the cable.

CSMA-CD networks have the disadvantage of requiring a large packet size when they operate at high bit rates. For example, a 1-km-long (2-km round trip) passive star network,[2] operating at 150 Mb/s, has an efficiency of 5 percent for 256 bits/packet, 44 percent for 4096 bits/packet, and 93 percent for 65,536 bits/packet. In addition to the low efficiency there is an associated instability problem that appears when the network reaches the channel capacity.[3] Without traffic restrictions or under heavy traffic conditions the network becomes unstable and crashes.

This paper describes a new method to eliminate the collisions caused by simultaneous transmission of packets in a star network. The method consists of placing collision-avoidance circuits in the node at the center of the star. This collision-avoidance circuit does the work of a "traffic cop," in that it lets pass the packets that arrive while the node is idle and blocks those that arrive while the node is busy. The packets that

make it through the "traffic cop" are broadcast to all users including the sender, and the blocked ones are retransmitted (by the originating users) until they succeed in getting through the node.

The insertion of collision-avoidance circuits in a star network results in a network that has the combined advantages of a zero-length CSMA-CD (high throughput and good stability) with those of ALOHA (transmission of packets shorter than the network round trip time). The resulting network is more efficient and stable than CSMA-CD networks and takes full advantage of high-bit-rate transmission. These characteristics make the network attractive for transmission at high bit rates and make lightwave networks suitable for local area networks.

The "traffic cop" requires active components at the center of a passive star, which poses a disadvantage in terms of reliability. This problem is minimized by using a small number of components in the circuit, conserving the distributed-switching feature of multiple-access networks, and providing centralized maintenance.

The sections that follow describe the architecture, protocol, implementation, and analysis of the network.

II. NETWORK ARCHITECTURE

Figure 1 shows the case of a local-area network where all users are connected in a star configuration to a central node (N) by a user-interface circuit (UIC). This interface consists of a transmitter (T), a receiver (R), and a logic circuit (C) that enable the receiver that first receives a packet after the node is idle. The node consists of a short bus with a maximum length equal to the length that makes the time of a bus round trip equal to the period of one bit. Each user is connected to the UIC by a user's link.

III. PROTOCOL

The protocol between the user and the network can be summarized as follows: the user-interface circuit blocks those packets arriving at the central node before they could cause a collision, and the users keep retransmitting unsuccessful packets until they succeed in getting through the node.

The description of the protocol is divided into the user and node protocols.

3.1 User protocol

The user has input and output buffers to store the received and transmitted packets. When a user end has a packet in the output buffer to be transmitted:

(*i*) It transmits the packet at once.

(*ii*) It waits the user's round trip time, T_{RT}.

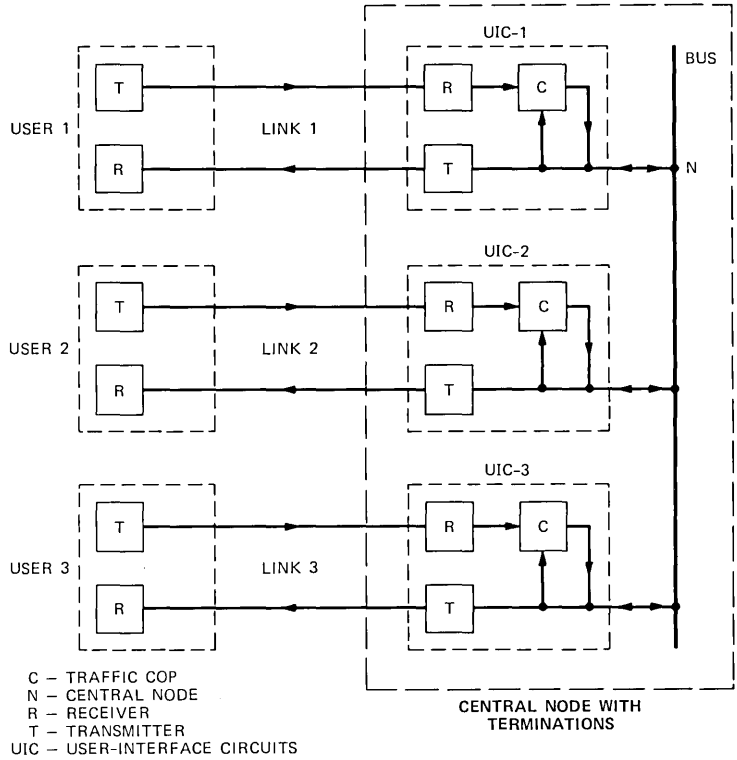


Fig. 1—A local-area network showing how all users are connected in a star configuration to a central node by independent user-interface circuits.

(*iii*) It checks the input buffer, and if a packet was received, it analyzes the origination and destination.

(*iv*) If a packet was not received, or if the received packet is not the same as the transmitted one, the packet was blocked at the node, and the whole process is repeated until the packet succeeds in getting through the node.

3.2 Node protocol

The transmitters and receivers at the central node are controlled in the following way:

(*i*) All the transmitters are connected to the node N at all times. Therefore, all the users are continuously receiving the packets broadcast by the node.

(*ii*) All the receivers are disconnected from the node.

(*iii*) The receivers and the node are continuously monitored for "idle" or "busy" status.

(*iv*) When a packet arrives at the receiver, the receiver becomes

"busy." A transition of the receiver from "idle" to "busy" while the node is "idle" connects the receiver to the node, and changes the node status to "busy." A transition of the receiver from "idle" to "busy" while the node is "busy" is ignored and the receiver remains disconnected. In this way, only the receiver with the first arriving packet is connected to the node, and all the others are ignored or disconnected.

(*v*) The node status returns to "idle" when the broadcast packet ends and the receiver is disconnected.

This protocol is performed by a logic circuit (C) whose functions are shown in Fig. 2. The receiver (R) and the transmitter (T) are part of the user's link and they vary according to the type of link. The logic circuit consists of an AND gate and four flip-flops. The four flip-flops are wired to perform the receiver-monitor, node-monitor, arbiter, and hold-on functions. The AND gate connects and disconnects the receiver to the node (N) following the status of the hold-on flip-flop. Figure 3 shows the timing diagram for the different parts of the circuit.

The receiver monitor and the node monitor are retriggerable single-shot circuits that detect the presence of a packet by sensing the carrier. They go "high" at the arrival of a packet and return "low" at the end of the packet.

The arbiter is a D-type flip-flop that produces an "I was first" pulse when the receiver monitor goes "busy" while the node monitor is "idle." The "I was first" pulse indicates that the receiver has the first arriving packet and sets the hold-on flip-flop "high." All "idle" to "busy" transitions of the receiver monitor that occur while the node monitor is "busy" are ignored and they do not set the hold-on flip-flop.

The hold-on flip-flop goes "high" following the "I was first" pulse

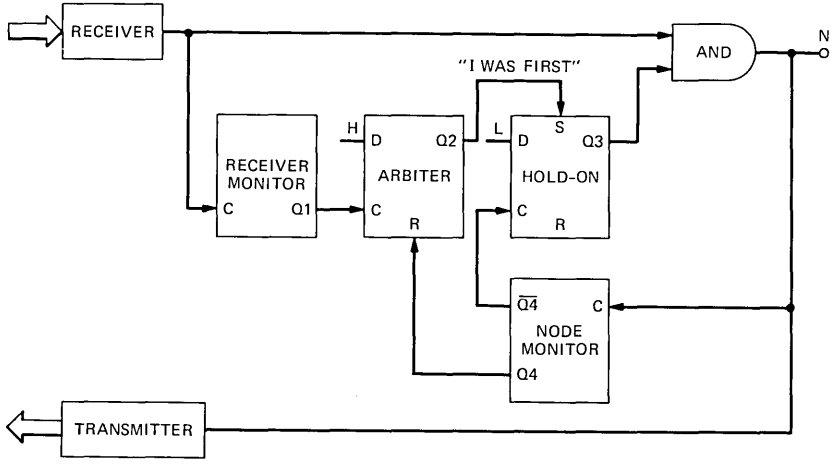


Fig. 2—User-interface circuit showing four flip-flops wired to perform the receiver-monitor, node-monitor, arbiter, and hold-on functions.

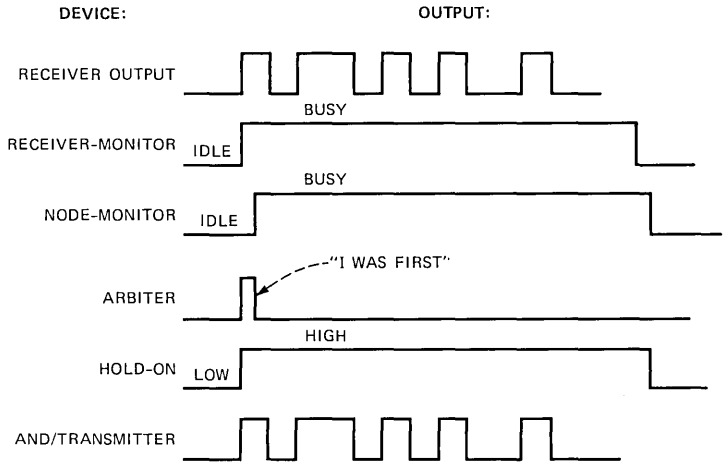


Fig. 3—A timing diagram of the user-interface circuit. The hold-on goes high only when the receiver monitor goes high before the node monitor does.

and returns "low" when the node monitor goes "idle." The output of the hold-on flip-flop activates the AND gate that connects and disconnects the receiver to the node (N).

The reader may wonder what happens when two or more packets arrive simultaneously. This is a rare event because it takes less than 20 ns to determine which was the first packet arriving. But suppose anyway that two packets arrive within a period of 20 ns; then a collision is possible because two receivers will be simultaneously connected to the node. An exclusive-OR circuit can be installed in the way shown by Fig. 4 to handle this rare event. This additional circuit resets the hold-on flip-flop ("low") any time that the receiver signal is different from that of the transmitter and while the node is "busy." The two receivers remain connected until the signal for one of the receivers differs from its transmitter signal. The receiver with the first low bit stays connected and the others are disconnected. Figure 5 shows the timing diagram of the different components when two users are randomly transmitting words.

IV. ANALYSIS

To study the system let us consider a network of K users each successfully transmitting λ packets/second, each τ seconds long through the node. Then one defines $S = K\lambda\tau$, which is the average channel utilization, also known as the traffic intensity or traffic throughput.

S is also the probability of a user encountering a busy node in any attempt at transmitting a packet. The probability of successfully

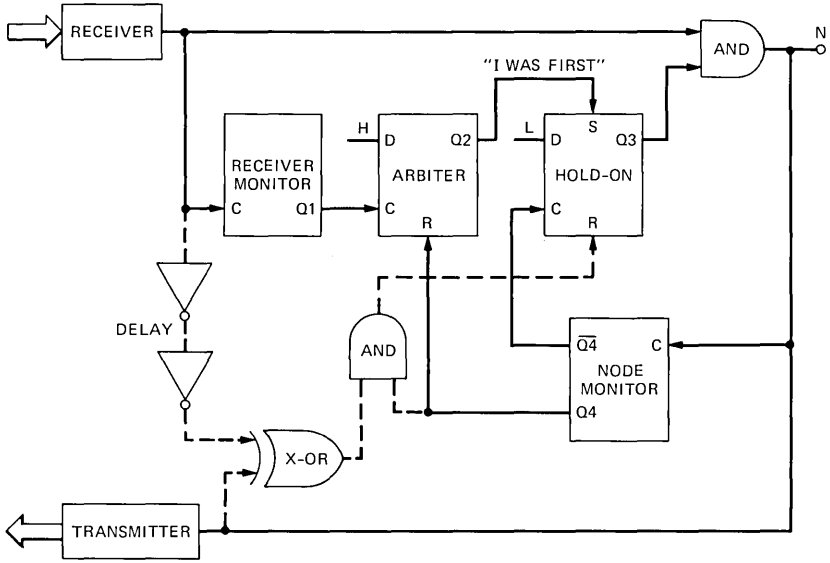


Fig. 4—Addition to the circuit to handle simultaneous events. The dashed lines indicate the circuit added to Fig. 3 to determine the priority of packets that arrive within the circuit response time.

transmitting a packet is $1 - S$. Therefore, the average number of transmissions required to get a packet through the node is

$$N = \frac{1}{1 - S} \tag{1}$$

and the average traffic offered by the users to the network is

$$G = SN = \frac{S}{1 - S}; \tag{2}$$

consequently, S can be expressed in terms of G as

$$S = \frac{G}{G + 1}. \tag{3}$$

Another parameter of importance in the network is the average user's transmission delay. This delay is computed adding all the possible delays weighted by their probabilities:

$$\begin{aligned} D &= T_1(1 - S) + (2T_1 + T_2)S(1 - S) \\ &\quad + (3T_1 + 2T_2)S^2(1 - S) + \cdots \\ &= T_1N + T_2SN, \end{aligned} \tag{4}$$

where $T_1 = T_{RT} + \tau$, T_2 is the retransmission delay in addition to T_1,

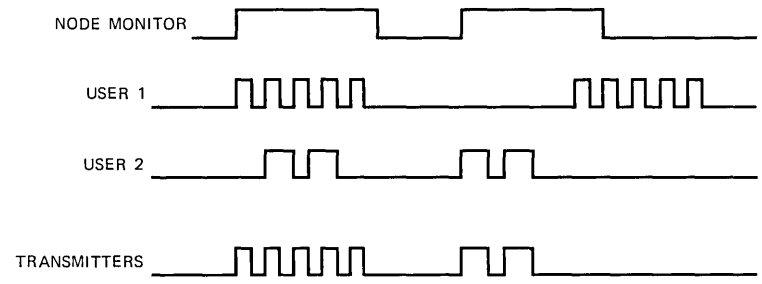


Fig. 5—Timing diagram showing two users randomly transmitting. The circuit broadcasts the packets of user 1 and user 2 that arrive only while the bus is idle.

and T_{RT} is the user's round trip time. The delay is reduced by making T_2 small comparable to T_1, which leaves only the first term in eq. 4.

The above equations are analogous to those of a zero-length CSMA-CD network where the ratio $T_{RT}/\tau = 0$ makes the maximum value of S equal to 1.

More efficient modes of operation are possible but they will increase the complexity of the stations. For example, a clock may be installed at the node N to broadcast a burst to indicate the beginning of a frame, and have reserved channel assignment. This arrangement decreases the average delay but it requires a reservation protocol.

V. CONCLUSION

A new arrangement for a star network was proposed and a central-node configuration, as shown in Fig. 1 with UIC, was built and operated with packets up to 50 Mb/s. The node eliminates collisions by resolving the right-of-way when several packets arrive at the node. The node lets pass the packet that arrives first and blocks all other packets that would collide with the first one. Meanwhile, the users keep retransmitting packets until they get through the node.

The analysis shows that the average number of retransmissions depends on the traffic intensity at the node, and the retransmission of unsuccessful packets does not degrade the traffic throughput of the node. Also, the channel capacity has the maximum value 1, and it is not a function of the ratio between the packet size and the network-round-trip time.

The analysis shows that at moderate traffic intensities (50 percent) the average number of retransmissions is two. Traffic intensities approaching 100 percent may cause a large number of retransmissions. A 75-percent traffic intensity requires an average of four retransmissions, 80 percent requires five, and 90 percent requires ten.

High-bit-rate services can be provided simultaneously with low-bit-rate services without changing the traffic intensity simply by increasing

the transmission bit rate and holding the packet duration and the packet frequency constant.

REFERENCES

1. R. M. Metcalfe and D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. ACM, *19,* No. 7 (July 1976), pp. 395–404.
2. E. G. Rawson, R. M. Metcalfe, R. Norton, A. B. Nafarrete, and D. Cronshaw, "Fibernet: A Fiber Optic Computer Network Experiment," IEEE Trans. Commun., *COM-26,* No. 7 (July 1978), pp. 983–90.
3. L. Kleinrock and F. A. Tobagi, "Packet Switching in Radio Channels: Part I—Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics," IEEE Trans. Commun., *COM-23,* No. 12 (December 1975), pp. 1400–16.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

Pressure-Volume-Temperature Behavior in the System H_2O-NaOH-SiO_2 and Its Relationship to the Hydrothermal Growth of Quartz

By E. D. KOLB, P. L. KEY, R. A. LAUDISE, and E. E. SIMPSON*

(Manuscript received September 21, 1982)

We have measured the pressure-volume-temperature relations in the high-pressure solutions used to grow electronic quartz and used this data to establish safe operating conditions for commercial production. High-temperature aqueous solution (hydrothermal) quartz growth, because of the importance of its product to electronics, must be ranked as one of the more important crystal-growth processes. We report here a convenient laboratory method for hydrothermal p-V-T measurements and give pressure data in 1.0-mol NaOH and in 1.0-mol NaOH saturated with quartz as a function of temperature up to 450°C and 30,000 psi. These results are compared with pressures measured on production-sized equipment. The results are used to establish the temperature at which the gas phase disappears under various conditions. The steels used for construction of high-pressure production autoclave equipment are brittle below a specific temperature, which increases slowly with service. Our p-V-T data can be used to assure that high pressures are avoided at temperatures where the autoclave is brittle. Finally, the depressions of pressure are used to glean information about the nature of the solute species present during growth, and can ultimately be of use in quartz rate and perfection studies.

I. INTRODUCTION

Commercial hydrothermal growth of single-crystal α-quartz has been practiced for more than twenty years. It is arguably second only to Si

* Western Electric, Merrimack Valley Works, No. Andover, Massachusetts 01845.

in its importance to electronics. The hydrothermal *mineralizer* or solvent used is either NaOH or Na_2CO_3 and, particularly in the case of the more generally used NaOH, both the physical chemistry and crystal-growth aspects have been quite extensively studied, including phase relations,[1] solubility,[2] growth kinetics,[3] distribution of impurities (especially OH),[4] perfection,[5] and electrical properties[6] of grown quartz. However, the pressure-volume-temperature (p-V-T) relations for neither Na_2CO_3 nor NaOH for mineralizer solutions saturated with quartz have been determined. Data on p-V-T for aqueous solutions of NaOH[7-9] and Na_2CO_3[7] are available, but their relationship to the silica-saturated solution used in growth is tenuous. For our purposes the system H_2O—NaOH—SiO_2 is of particular interest since it is most used and probably most studied. We have directed our studies to it and have adopted it for production.[10] Indeed, except for recent p-V-T measurements in the system H_2O—H_3PO_4—$AlPO_4$,[11] under conditions of $AlPO_4$ saturation like those used for $AlPO_4$ growth, no p-V-T measurements of mineralizers saturated with the solutes used in hydrothermal crystal growth have been made. In the paper of Kolb et al.,[11] we described a technique and equipment for rather rapid p-V-T measurements that can be easily applied to other saturated hydrothermal systems, and reviewed the literature summarizing p-V-T measurements in hydrothermal mineralizers. As a result of this work and our research and factory experience with quartz, it is our belief that p-V-T measurements in the system H_2O—NaOH—SiO_2 would be particularly useful for the following reasons:

(*i*) It has been shown[12] that the brittle-ductile transition temperature of low-carbon steels of the sort used in hydrothermal autoclaves gradually increases when such steels are aged in the temperature range of quartz growth. For safety reasons it is important to be sure no autoclave is exposed to substantial pressure while in the brittle regime. To assure this, accurate p-V-T data for the hydrothermal solution are essential. The need for such data becomes particularly important when autoclaves have been in service for periods of years.

(*ii*) Pressures are regularly measured, monitored, and used for control in commercial growth. In such growth a temperature differential, ΔT, between dissolving-nutrient zone and seed-growth zone, is necessary to produce supersaturation. It would be especially useful to have p-V-T data for both isothermal *and* ΔT conditions for comparison with production data. Pressure changes during growth could possibly be used to monitor internal ΔT changes during growth.

(*iii*) Pressure depression in comparison with H_2O and H_2O—NaOH may be useful in gleaning information concerning the species present during hydrothermal growth.

(*iv*) Pressure data may provide insights into kinetics, distribution

of impurities, and perfection, which could further improve the speed, economics, and perfection of commercial growth.

For the foregoing reasons we decided to study and here report p-V-T measurements in the system H_2O—NaOH—SiO_2 under isothermal conditions and with temperature differentials over the range of conditions relevant to quartz growth. For comparison purposes we also studied the system H_2O—NaOH.

II. EXPERIMENTS

This section describes procedures used in laboratory measurements. Measurements in production vessels are described later. The apparatus and procedures used are similar to those reported in our study of H_2O—H_3PO_4—$AlPO_4$,[11] except that Pt-lined Morey vessels were not required when quartz was present. This is due to the fact that low-carbon steel autoclaves are relatively inert in the system H_2O—NaOH—SiO_2 because of the formation of insoluble sodium-iron silicates that protectively coat the vessel.[5] Thus a 1-1/4-in ID x 12-inch IL autoclave with a modified Bridgman closure made by Autoclave Engineers (Erie, PA) fabricated from Timken 17-22-AS steel was used. A series of runs at fills of 65, 75, 80, 83, 86, and 89 percent were made as were calibration runs using pure water at fills of 70, 80.1, and 80.9 percent. Pressure was measured with 0- through 5000-psi (350-bar), 0- through 20,000-psi (1380-bar), and 0- through 40,000-psi (2760-bar) Bourdon gauges (Heise Co., Newtown, Conn.), precalibrated on a dead-weight tester. Pressures above the coexistence curve were measured with gauges with a resolution of ±20 psi (~1 bar), while those along the coexistence curve (where liquid and vapor are in equilibrium) were obtained using the 0- through 5000-psi gauge whose resolution is ±5 psi (~0.3 bar). The pressure gauges were connected to the cover of the autoclave by 0.017-inch ID stainless capillary tubing brazed to a 1/4-inch diameter high-pressure cone seal.

Separate measurements in the system H_2O—NaOH were made using a Pt-lined Morey autoclave and pressure take-off of the sort described in our aluminum phosphate p-V-T work.[11] Autoclave volume for both autoclaves was measured by filling with H_2O as previously described.[11] The volumes of the steel plunger and steel coupling were calculated from dimensions. The Bourdon-gauge tube volumes were determined by forcing H_2O through the tube with a calibrated syringe that was fitted to the inlet and measuring the water collected at the exit (gauge-bleed port). The volume of H_2O in the tube is arrived at as a difference that is taken only when bubble-free water is collected at the exit. The volumes of the steel and Pt-lined pressure take-offs and the stainless steel capillary were determined as described previously.[11] Table I lists typical volumes.

Table I—Typical volumes

Bridgman Autoclave	250.0 cc	Morey Pt-Lined Autoclave	30.0 cc
Steel Take-off	1.00 cc	Pt-lined Take-off	0.15 cc
Steel Capillary	0.30 cc	Steel Capillary	0.30 cc
Gauge 40,000 psi	18.0 cc		
20,000 psi	5.0 cc		
5,000 psi	13.0 cc		

The placement of heaters and thermocouples for the Pt-lined autoclave was as described previously,[11] while for the Bridgman autoclave heating was provided by a hot plate on the bottom and band heaters appropriately spaced so as to obtain minimum ΔT or, if desired, a particular ΔT. An improved temperature-control system using an Electronix III controller with silicon-controlled rectifiers (Leeds and Northup, N. Wales, PA) was employed in the present work. Temperatures are estimated to be controlled within ±0.5°C.

To assure saturation and protect the steel autoclave walls from attack, quartz plates 1- x 2- x 0.04-inches were mounted in the Bridgman autoclave, along with 75 gm of small-particle quartz nutrient. Standard NaOH solutions of 1.000 ± 0.003 mol concentration were purchased from Fisher Scientific.

III. RESULTS AND DISCUSSION

3.1 Calibration

As a check of our procedures we elected to investigate how well our system reproduced the p-V-T data of Kennedy[13] for pure water. Three runs were made at fills of 70, 80.1, and 80.9 of the autoclave. The p-T behavior at 80.1- and 80.9-percent fill was measured in the unlined Bridgman vessel up to pressures of ~25,000 psi (~1725 bar) to simulate our measurements in H_2O—NaOH—SiO_2. The p-T behavior at 70-percent fill was measured in the Pt-lined Morey vessel with a Pt lead through to the stainless steel capillary coupling so as to simulate our measurements in H_2O—NaOH. Of necessity we were limited to the pressure capability of Morey autoclaves (~10,000 psi ~700 bar) for all work in lined vessels.

Table II gives representative pressures obtained in these runs and the equivalent percent fill for H_2O from Kennedy's data.[13] Furthermore, in Table II the experimental percent fills are corrected for: (*i*) temperature expansion of the vessel, (*ii*) pressure dilatation of the vessel, and (*iii*) compression of the water in the lines and gauge. The temperature expansion of the vessel was obtained using eq. (1) of Kolb and Laudise.[11] The Morey and Bridgman vessels were made of Timken 17-22-AS steel so that the coefficient of expansion in Fig. 5 of Kolb and Laudise[11] was used. The pressure dilatation of the vessels was calculated using eq. (2) of Kolb and Laudise[11] and the values Y and ν

Table II—Representative pressures and percent fills for H_2O

Experimental Percent Fill	80.9	80.9	80.9	80.9	80.1	80.1	80.1	70	70
Measured Temperature (°C)	250	300	350	400	300	350	400	350	400
Measured Pressure (bar)	46.9	676	1307	2060	572	1172	1765	455	903
Experimental Percent Fill Corrected for Temperature Expansion and for Initial Fill on a Volume Basis (a)	80.0	79.8	79.6	79.4	79.0	78.8	78.7	68.8	68.6
Experimental Percent Fill Corrected for (a) + Pressure Dilatation (b)	79.9	79.7	79.5	79.2	78.9	78.7	78.5	68.8	68.6
Experimental Percent Fill Corrected for (a) + (b) + Compressibility of H_2O in Lines, Leads and Pressure Gauge (c)	79.9	79.5	79.1	78.7	79.7	78.4	78.0	68.5	68.0
Kennedy[13] Percent Fill for Measured *p*-T	79.7	79.8	79.3	79.8	78.5	78.2	77.7	68.4	67.6

for Timken 17-22-AS suggested.[11] The correction owing to the compressibility of H_2O in the tubing connecting the high-pressure gauges and to water in the Bourdon tubes of the gauges was made using the specific volumes of water as a function of p and T, as tabulated in Landolt-Börnstein.[14]

In addition, it should be pointed out that Kennedy's p-V-T data[13] are given in terms of specific volume, $\bar{V}$, cm^3/g. Our experimental percent fill (as reported in the first row of Table II) is based on the quantity: volume H_2O/free volume of autoclave. The corrected fills shown in the fourth, fifth, and sixth rows of Table II are multiplied by the density of water at room temperature so that the fill in these rows is reported as gm H_2O/free volume of autoclave and can be directly compared with Kennedy's data (last row of Table II). As can be seen, the agreement between our corrected data and Kennedy's data is excellent, suggesting that our procedures are reasonable. For the convenience of those wishing to apply corrections at other conditions, Table III summarizes some useful data.

3.2 p-V-T behavior of H_2O-NaOH-SiO_2

Figures 1 and 2 show the p-T behavior of 1.0-mol NaOH saturated with SiO_2 at fills of 65, 75, 80, 83, 86, and 89 percent. The concentration of 1.0-mol NaOH was chosen as comparable to conditions used in commercial growth. Figure 1 shows the low-pressure region, while Fig. 2 covers higher pressures. The line A-B is the coexistence curve. The fills shown on these figures are based on the initial cold volumes of solution and autoclave and no correction for temperature or pressure dilatation of the vessel is made to them. This is consistent with the definition of fill that is conventional in crystal growth. Such corrections could be made if desired using the data of Table III and the procedures outlined in the previous section and in Kolb and Laudise.[11]

Equilibrium was assured by holding at temperature until pressure stabilized, usually after a few hours. Pressures were not recorded until they were invariant for 12 or more hours. Pressures were observed to reproduce regardless of whether the equilibration temperature was approached from above or below. As a further check to ensure the 12-hour equilibration time was sufficient, a 1.0-mol NaOH run at 89-percent fill was equilibrated with quartz for one week at 214°C and then the temperature was abruptly raised to 303°C. The dependence of pressure on time in this experiment is shown in Fig. 3. As we can see, pressure equilibrium is again established after 250 min. It should be noted that the pressure (23,350 psi, 1610 bar) for 89-percent fill at 303°C for 250 min in Fig. 3 is the same as the pressure at 303°C in Fig. 2. Thus, 12-hour equilibrium times are more than sufficient.

The points in Figs. 1 and 2 at 65-percent fill were obtained in a

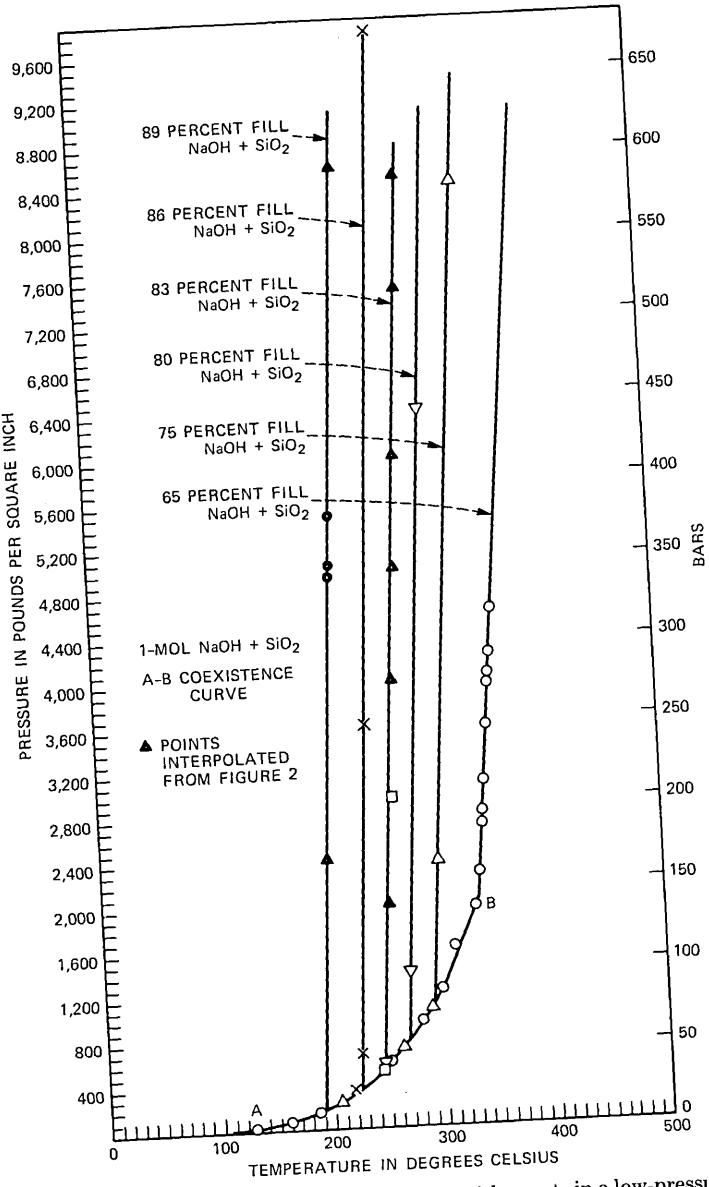


Fig. 1—p-T behavior of 1.0-mol NaOH saturated with quartz in a low-pressure region at several percent fills.

Morey autoclave using a 5000-psi gauge with a resolution of ±5 psi. These data are used to define the curve *A-B* where liquid and vapor saturated with quartz coexist. As we can see, at a particular fill once the autoclave has filled with liquid, the pressure departs from the coexistence curve and is linear with temperature; the slope of the *p-T*

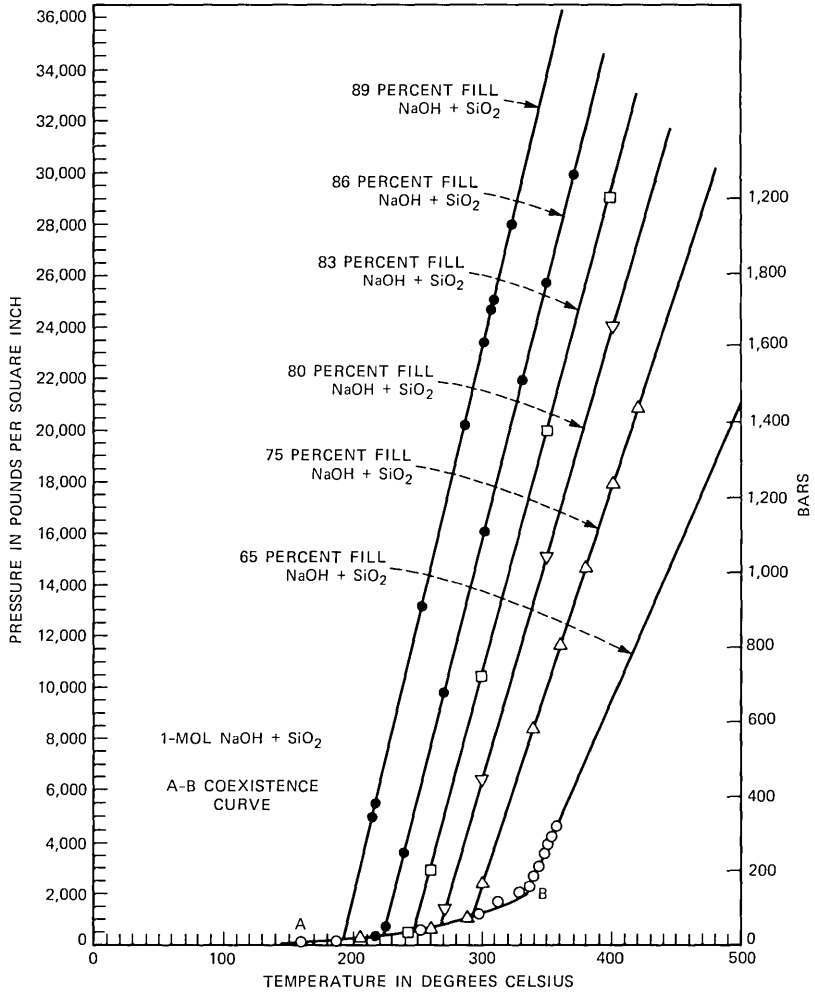


Fig. 2—p-T behavior of 1.0-mol NaOH saturated with quartz in a high-pressure region at several percent fills.

curves at constant-percent fill, $(\partial p/\partial T)_{\%f}$ is constant, and greater at higher-percent fills. Indeed, the data of Fig. 1 may be used to show the temperature at which the system departs from the coexistence curve as shown in Fig. 4.

The data of Figs. 1 and 2 may be displayed as in Fig. 5 to show the dependence of pressure on fill at constant temperatures. Similar to results on H_2O[13] and H_2O—H_3PO_4—$AlPO_4$,[11] the slope of these curves is not a constant.

3.3 p-V-T behavior of H_2O-NaOH

The Pt-lined Morey vessel was used to determine pressures in NaOH solutions. Figure 6 shows typical results, which are compared to

Table III—Data for temperature and pressure expansion of autoclaves

Temperature (°C)	Expansion Coefficient Timken 17-22-A (in/in°F)	V/V_0^*	Young's Modules Timken 17-22-A* (psi)
250	6.92×10^{-6}	1.0086	2.84×10^7
300	7.13×10^{-6}	1.0108	2.78×10^7
350	7.31×10^{-6}	1.0131	2.74×10^7
400	7.44×10^{-6}	1.0153	2.68×10^7
450	7.53×10^{-6}	1.0175	

* V = Volume at temperature, V_0 = volume at 25°C, ν (Poisson's ratio) = 0.3, independent of temperature over this range.

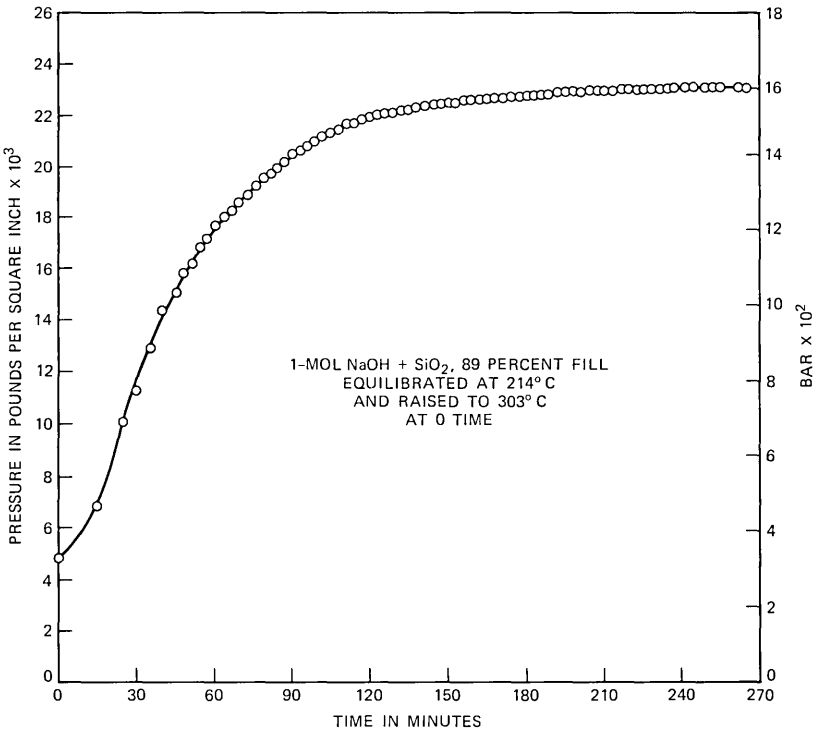


Fig. 3—Dependence of pressure on time when temperature of equilibrated vessel is increased from 214°C to 310°C (89-percent fill 1.0-mol NaOH initially saturated with quartz).

pressures of similar solutions saturated with quartz and with pure water. Some data from Samoilovich[15] and Kijama[16] are also plotted. However, the data of Kanahara, Yamasaki, and Matsuoka[9] for NaOH-H_2O are not plotted. Their pressures for 1-mol NaOH are much lower than ours or those of Refs. 15 and 16. Liebertz[7] reports pressures for 4-mol NaOH, which are much higher than those of Kanahara et al.[9] but

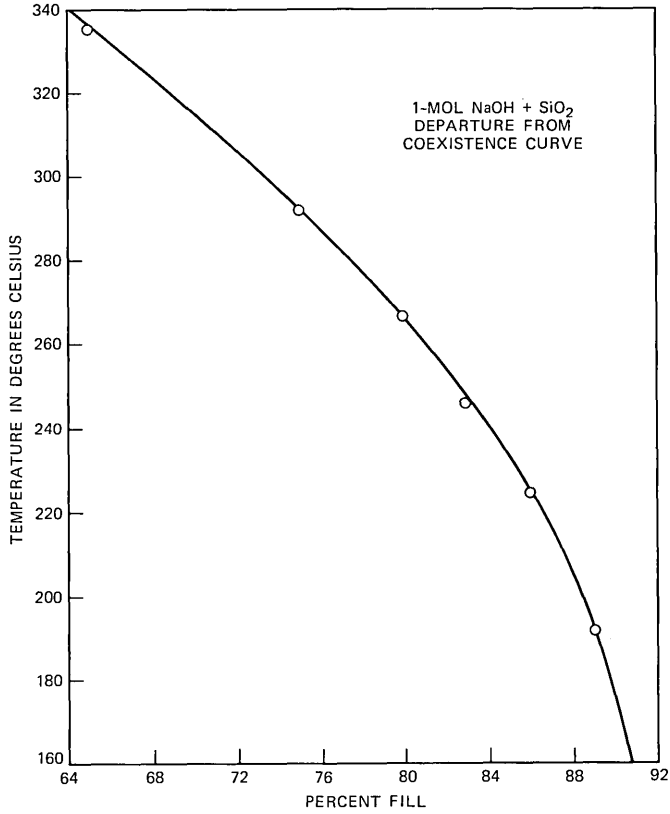


Fig. 4—Temperature at which autoclave fills for 1.0-mol NaOH saturated with quartz as a function of initial fill.

consistent with those of Refs. 15 and 16 at that concentration, suggesting a systematic error in (9).

As we can see in Fig. 6 our NaOH data are consistent with Refs. 15 and 16 obtained at slightly different NaOH concentrations. NaOH depresses the pressure along the coexistence curve and beyond, and $(\partial p/\partial T)_{\%f}$ is a constant in NaOH solutions above the coexistence curve.

When quartz is added to 1-mol NaOH, as shown in Fig. 6, the pressure is increased about 600 psi (40 bar) at 83-percent fill and more than 2000 psi (140 bar) at 65-percent fill. In previous work[2] we suggested silicate formation such as

$$2(OH)^- + SiO_2 \leftrightarrows SiO_3^{-2} + H_2O \quad (1)$$

for dissolving in $(CO_3)^{-2}$, and

$$2(OH)^- + 3SiO_2 \leftrightarrows Si_3O_7^{-2} + H_2O \quad (2)$$

for dissolving in $(OH)^-$. In eq. (1) the $(OH)^-$ is produced by $(CO_3)^{-2}$

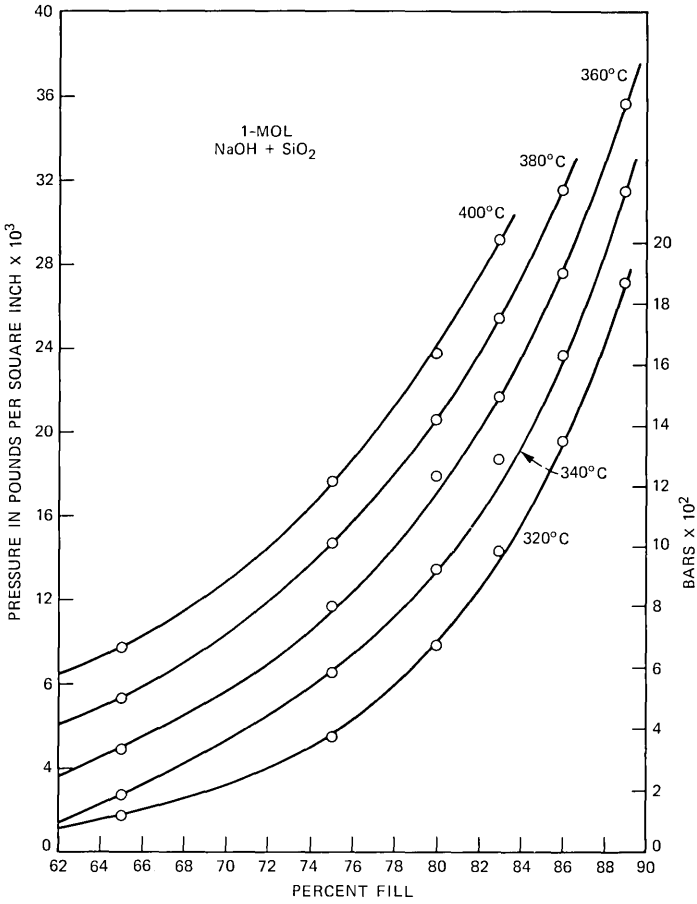


Fig. 5—*p*-percent-fill behavior of 1.0-mol NaOH saturated with quartz at several temperatures.

hydrolysis. Regardless of the details of such reactions it is likely that more than one $(OH)^-$ is consumed for each silicate ion formed. Consequently, the number of negative ions is decreased. If we make the reasonable assumption that the number and/or strength of the bonds of water molecules held in the first coordination shell by $2(OH)^-$ is greater than by a single $(SiO_3)^{-2}$, $(Si_3O_7)^{-2}$ or similar silicate then, as we observe in Fig. 6, the pressure will increase when SiO_2 dissolves in $(OH)^-$. The effect might be larger at lower fills because the H_2O shell about an $(OH)^-$ is compressed more relative to bulk H_2O when the H_2O density is less.

IV. RELATIONSHIP TO AUTOCLAVE EMBRITTLEMENT

Figure 7 shows the brittle-ductile transition temperature of some autoclave steels as a function of exposure time at quartz production

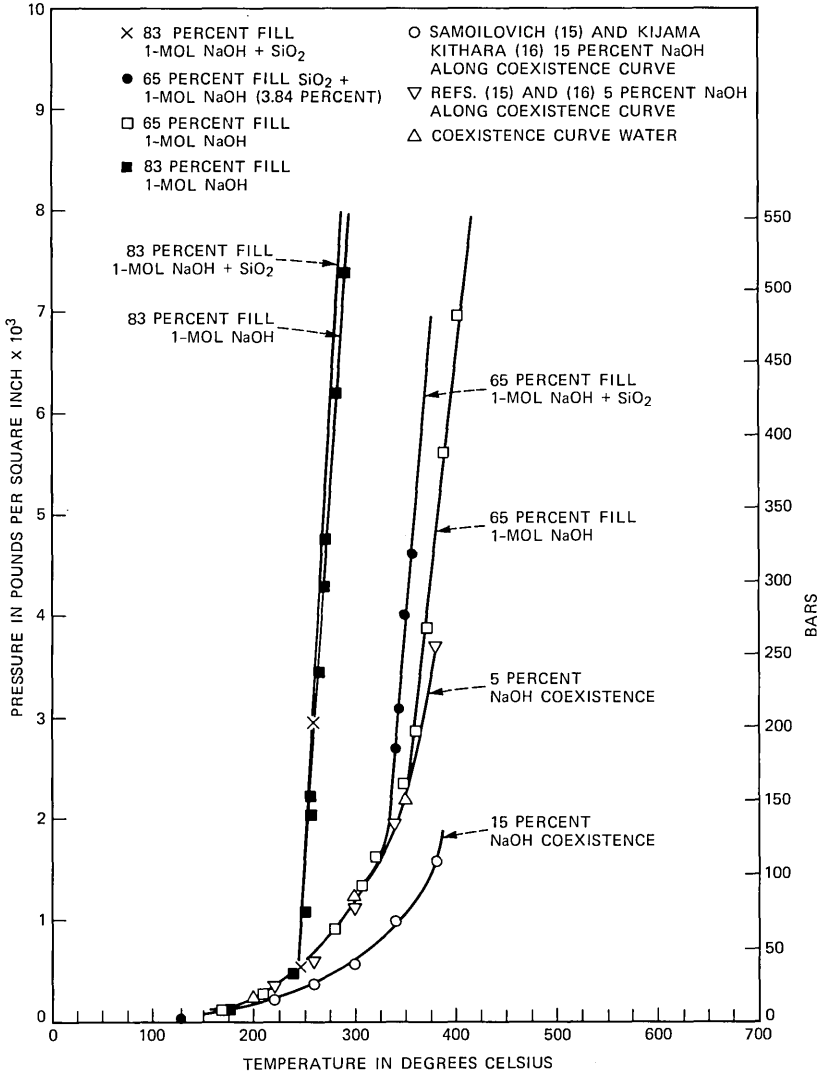


Fig. 6—p-T behavior of 1.0-mol NaOH compared to pure water and 1.0-mol NaOH saturated with quartz.

conditions (350°C crystallization temperature, 400°C nutrient temperature with the usual ~18-hr warm-up and ~24-hr cool-down and typical run time of ~30 days). Further identification of the steels is given in Ref. 12. The data of Fig. 7 were obtained by Charpy impact measurements.[12]

The autoclave operating conditions, primarily fill, should be chosen such that high pressure is not developed until the temperature is above the brittle range. On a practical basis, this requires that conditions be

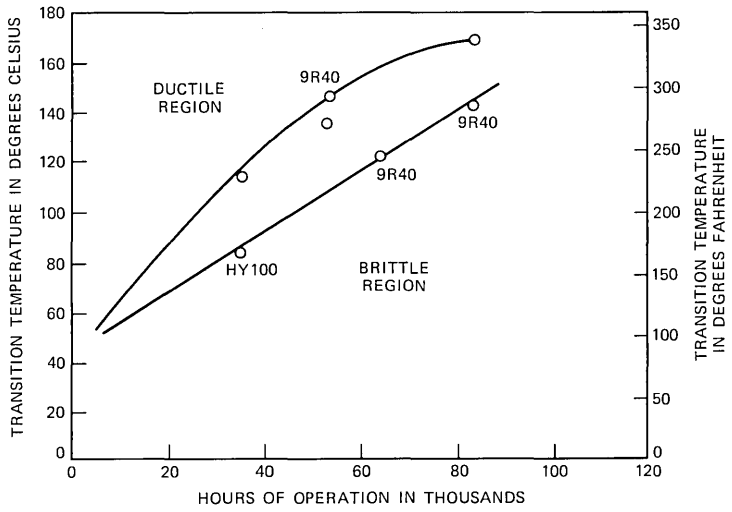


Fig. 7—Brittle-ductile transition of autoclave steels as a function of time at 350 to 400°C.

chosen using data in Figs. 4 and 7 so that pressures do not depart from the coexistence curve until the autoclave temperature is in the ductile region. For example, from Fig. 7, an autoclave with 80,000 hours of exposure (about 10 years at a typical 90-percent duty cycle) will be brittle for temperatures below 140 to 175°C. From Fig. 4, this suggests that fill should be limited to 90 percent. Thus, p-V-T data are very useful in establishing safe operating conditions and should be obtained for other mineralizers and conditions used in commercial hydrothermal processes.

V. EFFECT OF TEMPERATURE DIFFERENTIALS

For practical crystal growth saturation occurs in a hotter, lower region of the autoclave containing relatively finely divided quartz nutrient, and growth takes place in an upper, supersaturated, cooler region containing seeds. To obtain information about the effect of temperature differentials on measured pressures a series of experiments in non-isothermal vessels were conducted. All experiments were done in 1.0-mol NaOH saturated with quartz at various fills. Figure 8 summarizes these results. We designate temperatures as follows:

T_1 = upper, cooler temperature region of autoclave

T_2 = midpoint temperature of autoclave

T_3 = lower, hotter temperature region of autoclave

$\Delta T_1 = T_3 - T_1$ ($T_2 \sim$ constant)

$\Delta T_2 = T_2 - T_1$ ($T_3 \sim$ constant).

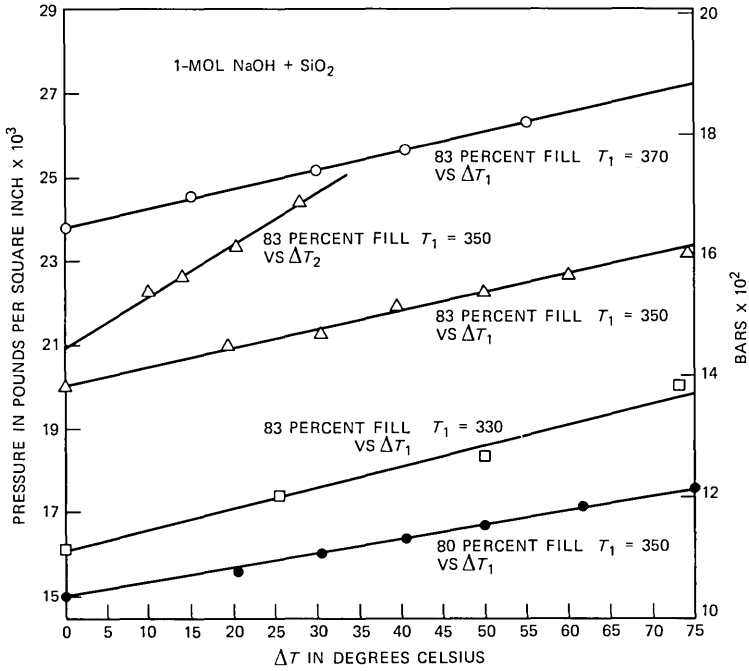


Fig. 8—p-T behavior of 1.0-mol NaOH saturated with quartz as a function of temperature as defined:

$$\Delta T_1 = T_{\text{Upper cooler region}} - T_{\text{Lower hotter region}} \quad \text{and}$$
$$\Delta T_2 = T_{\text{Mid}} - T_{\text{Upper cooler region}}.$$

As we can see, pressure is approximately linear with ΔT_1 and ΔT_2. Figure 8 can be used to estimate the correction necessary for converting isothermal data to ΔT data. It is interesting to point out that neither the hotter, lower region, upper, cooler region, nor average temperature determines pressure, but a fair fit can be obtained using an average weighted in favor of the upper, cooler temperature.

For example, consider Table IV, made from the data of Fig. 2 and Fig. 8, for 83-percent fill, 1-mol NaOH saturated with SiO_2. Row (a) is the upper, cooler temperature region of the autoclave (T_1); Row (b) the measured pressure; Row (c) the temperature difference, ΔT_1; Row (d) the temperature, T_4, which gives the same pressure in an isothermal autoclave; Row (e) $\Delta T_3 = T_4 - T_1$; and Row (f) $\Delta T_3/\Delta T_1$. If the average temperature determined the pressure in a non-isothermal autoclave, then $\Delta T_3/\Delta T_1 = 0.5$. As we can see from Row (f), the average is weighted toward the upper, cooler temperature region and the weighting factor is reasonably constant.

The pressure is of course uniquely determined by a knowledge of local density and temperature anywhere in the vessel. For example, in

Table IV—Effect of temperature differentials

(a) Upper Cooler Temperature Region of Autoclave (T_1) (°C)	350	350	350	350	350
(b) Pressure P (psi)	23,250	23,000	22,000	21,000	20,000
(c) ΔT_1 (°C)	75	67.5	45.0	22.5	0
(d) T_4 to Give P in Isothermal Autoclaves (°C)	368	367	361	356	350
(e) ΔT_3 (°C)	18	17	11	6	0
(f) $\Delta T_3/\Delta T_1$	0.240	0.252	0.244	0.267	

a vessel whose average percent fill is 83 percent, if the ΔT is 50°C with a top temperature of 350°C, the observed pressure is 22,180 psi (1530 bar). Using isothermal data for percent fill and taking density as percent fill/100 (Fig. 5), this suggests a local density of 0.85 g/cc for the region at 350° and 0.79 g/cc for the region at 400°. These densities are clearly subject to some errors since they are based on initial percent fills (Fig. 5) and do not include contributions attributable to dissolved solutes. Procedures of this sort might be used to calculate density gradients and hence be used to calculate the driving force for convective circulation in the system.

VI. COMPARISON OF LABORATORY AND FACTORY MEASUREMENTS

Measurements of p-T were made on a production crystal-growth vessel filled with nutrient, 5-percent baffle, and seeds, using the heater placement and power inputs normal for commercial growth. The volume of the vessel was determined by filling with water to the corrosion mark indicating where the seal ring had rested in the previous run. Volume corrections were made for the leads in the cover. Capillary lead tubing and Bourdon pressure gauge were filled with water using procedures similar to those of the laboratory. The vessel was filled to a nominal fill of 83 percent with 1-mol NaOH and warmed from room temperature to operating conditions: 350°C (crystallization temperature), 400°C (nutrient region), ΔT 50°C in a period of 18 hours. During the warm-up temperatures and pressures were recorded and are shown in Fig. 9.

The results are compared with isothermal laboratory data for 83- and 86-percent fills. Careful examination of the ring-seating region at the conclusion of the run and calculations of the volume from the autoclave dimensions lead us to conclude that the true percent fill in the vessel was 84 ± 0.5 percent. On this assumption we can see that the pressures obtained are in reasonable agreement with isothermal laboratory data.

VII. CONCLUSIONS

A reasonably rapid and convenient method for the measurement of the pressures of saturated hydothermal solutions of the sort used in

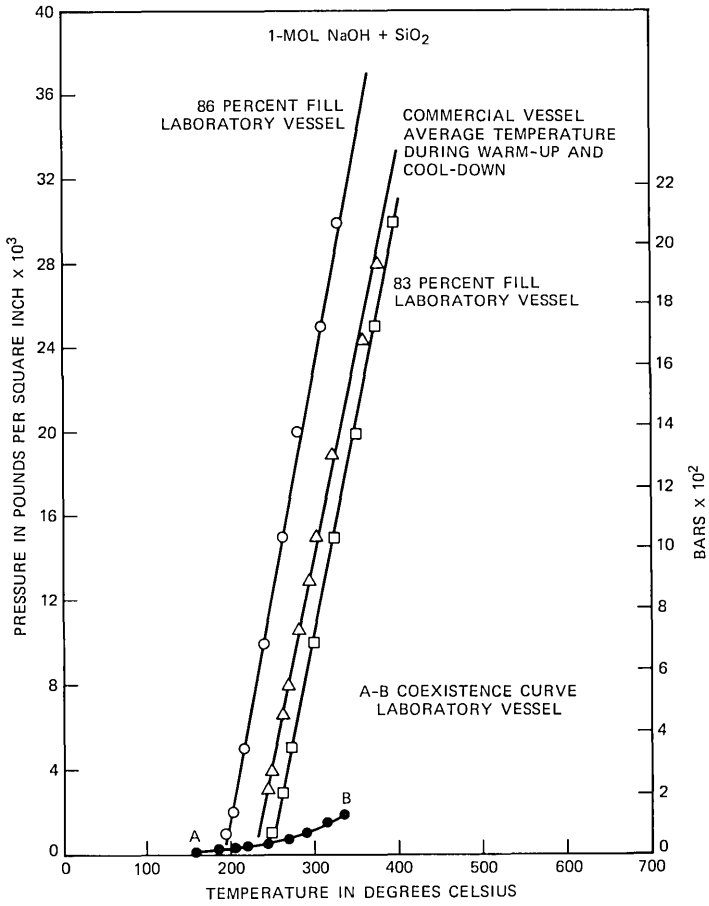


Fig. 9—Pressure-temperature of 1.0-mol NaOH plus quartz in commercial autoclaves during warm-up and cool-down.

crystal growth has been developed and applied to the determination of p-V-T relations in 1.0-mol NaOH and 1.0-mol NaOH saturated with quartz at degrees of fill from 65 to 89 percent, pressures up to ~30,000 psi (2070 bar), and temperatures up to ~450°C. The accuracy of our experimental procedures was tested in measurements on pure water where published pressure data were reproduced. Temperature and pressure dilatation effects are estimated.

For both 1-mol NaOH and NaOH saturated with quartz, $(\partial p/\partial T)_{\%f}$ is constant above the coexistence curve. Along the coexistence curve and above, the pressure of 1-mol NaOH is less than the pressure of water but the pressure of 1-mol NaOH saturated with quartz is greater than the pressure of 1-mol NaOH. The increase of pressure when quartz dissolves in 1-mol NaOH is explained on the basis of a decrease

in the number of ions in solution when $(OH)^-$ reacts with SiO_2 to form silicates.

After about 10 years of service the brittle-ductile transition temperature of autoclave steel increases to ~170°C so that substantial pressures should be avoided below that temperature. The p-V-T data are used to show the relationship between initial fill and temperature at which the pressure departs from the coexistence curve and $(\partial p/\partial T)_{\%f}$ becomes large. These data may be used as a guide for the choice of safe operating conditions.

Pressure measurements were made on autoclaves with temperature differentials and the data were related to pressure measurements on isothermal vessels. A weighted average temperature (weighted in favor of the upper, cooler region) for the temperature differential reproduced the isothermal measurements. On the assumption that local density and temperature determine pressure, p-V-T data can be used to estimate density differences associated with typical temperature differences (ΔT). A ΔT of 50° at 350°C growth temperature for 83-percent fill 1-mol NaOH saturated with quartz produces a density differential of ~0.06 g/cc.

Finally, pressure measurements were made on large-scale factory autoclaves under conditions of commercial growth and these measurements are compared to laboratory results.

VIII. ACKNOWLEDGMENTS

We would like to thank A. J. Caporaso for suggestions and assistance and especially for help with the design and fabrication of Pt equipment.

REFERENCES

1. G. W. Morey and J. M. Hesselgesser, Amer. J. Sci., Bowen Volume, *343* (1952); see also, "Phase Diagrams for Ceramists," Ed. by E. M. Levin, H. F. McMurdie, and F. P. Hall, Columbus, Ohio: American Ceramic Society, 1956, p. 259.
2. R. A. Laudise and A. A. Ballman, "The Solubility of Quartz Under Hydrothermal Conditions," J. Phys. Chem., *65* (1961), pp. 1396–1400.
3. R. A. Laudise, "Kinetics of Hydrothermal Quartz Crystallization," J. Am. Chem. Soc., *81* (1959), p. 562.
4. N. Lias, E. Grudenski, E. Kolb, and R. A. Laudise, "The Growth of High Acoustic Quartz at High Growth Rates," J. of Crystal Growth, *18* (1973), pp. 1–6.
5. R. L. Barns, P. E. Freeland, E. D. Kolb, R. A. Laudise, and J. R. Patel, "Dislocation-Free and Low-Dislocation Quartz Prepared by Hydrothermal Crystallization," J. of Crystal Growth, *43* (July 1978), pp. 676–86.
6. Proc. 35th Annual Symp. on Frequency Control, Electronic Industries Assoc., 2061 Eye St., Washington, D.C., May 27–29, 1981.
7. J. Libertz, "P, T, f-Werte von 4 n K_2CO_3-, Na_2CO_3-und NaOH-Losungen unter Hydrothermalen Bedingungen," Chem. Eng. Tech., *41* (1969), p. 1231.
8. L. A. Samoilovich, "The Pressure of Hydrothermal Solutions," Trudy VIII, Soveschaniya, Experimental. Tekhnich. Mineral. i Petrografi, Moscow (1971), p. 195 (See also Refs. 15 and 16).
9. S. Kanahara, N. Yamasaki, and K. Matsuoka, "Effect of Various Solute Concentration on Hydrothermal P-T Curves. II. NaOH," Rept. Res. Lab. Hydrothermal Chemistry, Kochi, Univ. Japan, *2* (8–15), (1978), p. 30.

10. R. A. Laudise and R. A. Sullivan, "Pilot Plant Production Synthetic Quartz," Chemical Engineering Progress, *55,* No. 5 (May 1959), pp. 55–9.
11. E. D. Kolb and R. A. Laudise, "Pressure-Volume-Temperature Behavior in the System H_2O-H_3PO_4-$AlPO_4$ and Its Relationship to the Hydrothermal Growth of $AlPO_4$," Journal of Crystal Growth, *56,* No. 1 (January 1982), pp. 83–92.
12. P. L. Key, E. D. Kolb, R. A. Laudise, and E. E. Bresnahan, "Embrittlement of Steels Used in Autoclaves for Hydrothermal Crystallization," J. of Crystal Growth, *21* (1974), pp. 164–6.
13. G. C. Kennedy, "Pressure-Volume-Temperature Relations in Water at Elevated Temperatures and Pressures," American Journal of Science, *248* (1950), p. 540.
14. Landolt-Börnstein, New Series, K. H. Hellwege, Ed., Group IV, Vol. 4 "High Pressure Properties of Matter," Kl. Schafer, New York: Springer, 1980, p. 89.
15. L. A. Samoilovich, "Relationship Between Pressure, Temperature, and Density of Aqueous Sodium Hydroxide Solutions," Acad. of Sci USSR, Inst. of Geol. of Ore Deposits, Petrography, Mineralogy and Geochemistry, VIII Conf. on Exptl. and Appl. Mineralogy and Petrography, Proceedings, Moscow: Nauka Publishing House, 1971, pp. 195–200.
16. R. Kijama and S. Kitahara, "S - P-F-T Relations of Sodium Hydroxide and Its Causticity at High Temperatures and Pressures," Review Phys. Japan, *27* (1957).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

Variable Rate ADPCM Based on Explicit Noise Coding

By N. S. JAYANT

(Manuscript received July 15, 1982)

This paper discusses a variable bit rate speech coding system based on explicit coding of the reconstruction noise in ADPCM (differential pulse code modulation with adaptive quantization). If the ADPCM bit rate is R *bits/sample, PCM coding of its noise using an average bit rate of* R_n *bits/sample provides the receiver with the possibility of operating at any bit rate in the range* R *to* $R + \max\{R_n\}$. *Using* R *values in the range 2 to 5, and* R_n *values in the range 0 to 3, we compare the performance of the* $(R + R_n)$*-bit system with that of conventional* $(R + R_n)$*-bit ADPCM. If noise coding is based on instantaneous* R_n*-bit quantization of its samples with an optimized step size, the signal-to-noise ratio performance is comparable to that of conventional ADPCM for* $R_n = 1$, *but it deteriorates significantly for* $R_n > 1$. *With non-instantaneous noise coding, the performance can exceed that of conventional ADPCM for any* $R_n > 1$, *if* $R > 2$. *This is due to a variable bit allocation algorithm that quantizes noise samples with differing resolutions, while maintaining a constant total bit rate in every block of 4 ms. The algorithm does not require the transmission of any extra side information. It can also be regarded as a way of improving the performance of ADPCM coding at a single bit rate of* $R + R_n$ *bits/sample.*

I. INTRODUCTION

Multiple-stage coding, where the reconstruction noise from an initial stage is itself coded for transmission in a subsequent stage, is known to provide substantial gains over single-stage coding in the context of deltamodulation using oversampled inputs.[1,2] In this paper, we consider two-stage systems for multibit differential pulse code modulation with adaptive quantization (ADPCM) coding of Nyquist-sampled speech inputs. Unlike systems that permit oversampling, signal-to-noise ratio

(s/n) gains in our systems will be seen to be either slightly negative, or positive but nondramatic. However, the proposed systems have a feature that is common to all noise-coding systems, the property of embedded coding: the output bit sequence of the coder contains a subsequence that can be used in a straightforward manner to provide lower bit-rate operation with an output speech quality very close to that of conventional operation at the lower bit rate; as a result, the channel or receiver can switch, as needed, between low-rate and high-rate modes. The possibility of *variable-rate* operation is a very desirable feature in digital communication systems such as packet-switched voice networks.[3] A PCM coder is inherently an embedded coder. Least significant bits in a PCM codeword can be progressively dropped, with a graceful loss of quality that is no greater than about 6 dB/bit. Conventional differential pulse code modulation (DPCM) is not an embedded coding system in a similar sense because of the presence of a feedback loop in coder and decoder.

Explicit noise coding is not the only way of designing an embedded ADPCM system. Coarse feedback in the DPCM predictor loop[4,5] is known to provide a very robust basis for embedded DPCM, with very little s/n degradation compared to conventional DPCM at a given bit rate; and the results are also expected to extend to DPCM with an adaptive quantizer. In the coarse-feedback approach, the encoder performs an appropriate quantization of predictor input in anticipation of a similar quantization that may be forced at the receiver as a result of bit-dropping. The coarse feedback embedded system can also drop more than one bit, in a progressive fashion, to provide a wide range of bit rates. The noise-coding approach provides zero degradation of quality at the lower bit rate, R. More important, explicit noise coding offers the possibility of complex versions of $(R + 1)$-bit ADPCM that can provide positive performance gains over conventional $(R + 1)$-bit coding. ADPCM with variable bit allocation (Section V) is one example of such a complex system. The noise-coding system with variable bit allocation can also be used as a *single-rate* coder in which the coding process is split into two steps (conventional ADPCM followed by noise coding) to permit a simple form of time-domain bit allocation for the improvement of ADPCM performance.

Figure 1 is a block diagram of a variable-rate coder employing optional coding of ADPCM noise, with an average noise-coding rate of R_n bits/sample. The special case of $R_n = 1$ is treated at length in Sections IV through VII. When the dashed boxes for bit allocation are eliminated, instantaneous noise-coding results, with a coding rate of exactly R_n bits for every noise sample. When the parts of the system within boxes A or B are eliminated, $R_n = 0$, and conventional single-rate ADPCM results, with a total bit rate of R bits/sample. The

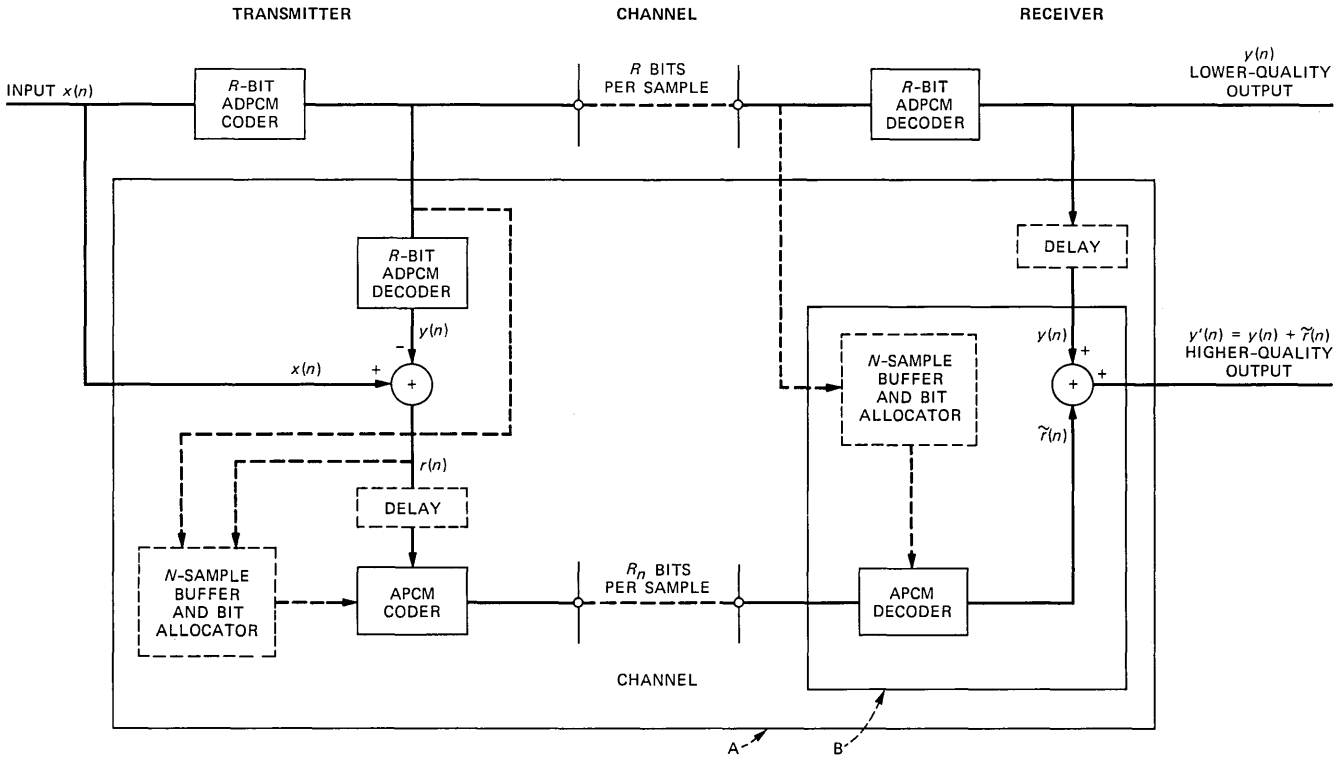


Fig. 1—Block diagram of a variable-rate ADPCM system. Variation of the average noise-coding bit rate R_n results in a variable-rate system with a total bit rate that ranges from R to $R + \max\{R_n\}$ bits/sample.

extreme upper part of the figure (outside of box A) shows a conventional R-bit ADPCM coder-decoder.[6] The rest of the diagram (the part included in box A) shows the blocks that perform R_n bit/sample coding of the reconstruction noise samples

$$r(n) = x(n) - y(n), \tag{1a}$$

where $x(n)$ and $y(n)$ are the input and the output of an R-bit/sample ADPCM system. When the system of Fig. 1 includes noise coding, the final decoded value is $y'(n)$, a refinement of the conventional value $y(n)$:

$$y'(n) = y(n) + \tilde{r}(n). \tag{1b}$$

The total bit rate of the system in Fig. 1 is $R + R_n$ bits/sample. Variable-rate coding results from the use of different values of R_n. Examples in this paper cover the range of $0 \leq R_n \leq 3$. The case of $R_n = 1$ is discussed at length before generalization to $R_n > 1$. With $R_n = 1$, the variable-rate system of Fig. 1 reduces to a dual-rate system, with a total bit rate of either R or $R + 1$ bits/sample.

The noise information can be altogether eliminated by the system ($R_n = 0$) to provide conventional R-bit operation. Alternatively, the noise information may be eliminated, as necessary, by the channel or receiver. If the receiver does the elimination, the part of the system that is eliminated is that within box B.

The results of this paper are based on simulations with three sentence-length utterances: "The *c*hairman cast three votes" (*f*emale speaker); "A *l*athe is a big tool" (*f*emale speaker); and "A *l*athe is a big tool" (*m*ale speaker). These speech inputs are identified in the rest of this paper as *CF*, *LF*, and *LM*. All inputs are band-limited to the frequency range 200 to 3200 Hz.

II. SUMMARY OF THE ADPCM AND APCM CODERS

The ADPCM coder in this paper uses first-order prediction with a time-invariant prediction coefficient of 0.85. It also uses an adaptive quantizer with a one-word memory.[7] As Fig. 2 shows for the examples of $R = 2$ and $R = 3$, the (uniform mid-rise) quantization characteristic $Q(x)$ is multiplicatively expanded or compressed at every sampling instant by a factor (step-size multiplier M) that depends only on the magnitude of the most recent quantizer output $y(n-1)$. If $\Delta(n)$ is the quantizer step size at time n,

$$\Delta(n) = M(|y(n-1)|) \cdot \Delta(n-1). \tag{2}$$

The function M takes on one of 2^{R-1} values in R-bit ADPCM. Recommended multiplier sets for $R = 2$ and $R = 3$ are included in Fig. 2. Recommended multipliers for $R = 4$ and $R = 5$ are tabulated in Ref.

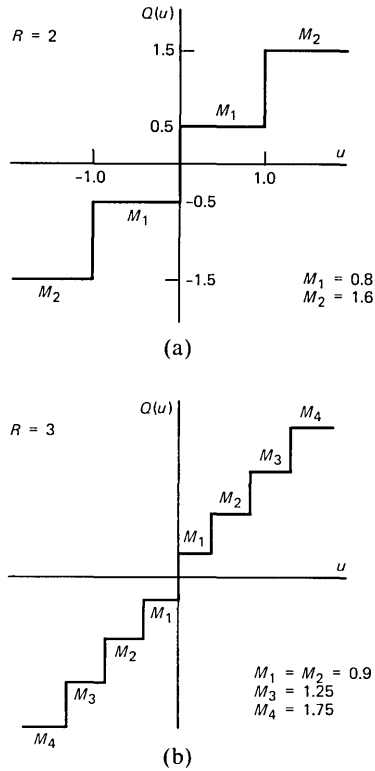


Fig. 2—Step-size multipliers used in (a) 2-bit and (b) 3-bit adaptive quantizers.

7. In the examples of Fig. 2, the use of the largest step-size multiplier also indicates the use of the outermost quantizer levels.

The adaptive PCM (APCM) coder for noise samples $r(n)$ will be described in detail in Sections IV, V, and VIII. The adaptive step size for the APCM coder will be seen to follow that of the quantizer in R-bit ADPCM. The purpose of the N-sample buffers in Fig. 1 is to permit a variable-bit-allocation procedure (Sections V and VIII) that provides a higher quality of noise quantization than what is possible with instantaneous quantization, the case of $N = 1$ (Section IV). When variable-bit allocation is employed, R_n will be interpreted as the *average* bit rate for noise coding. But the total number of noise-coding bits will be guaranteed to be a constant value, NR_n, for every block of N noise samples. The variable-bit allocation is first explained for the case of $R_n = 1$, implying noise coding with an average bit rate of 1 bit/sample (Section V). Extension to the case of $R_n > 1$ is straightforward (Section VIII).

III. RECONSTRUCTION ERROR $r(n)$

Figure 3a is a 16-ms-long speech segment from *CF*, and Fig. 3b illustrates the reconstruction-error waveform $r(n)$ in ADPCM for the example of $R = 4$. An important property is that $r(n)$ has occasional impulsive components. These are the slope-overload error bursts typical of DPCM with non-adaptive prediction. The extent of slope overload increases with coarseness of quantization. But, as seen in Fig. 3b, slope overload is quite evident even with $R = 4$ and adaptive quantization. In voiced speech, the time separation between slope overload bursts corresponds very closely to the pitch period. In the example of Fig. 3, this separation is about 40 samples (at 8 kHz), corresponding to a pitch period of about 200 Hz. Figures 3c and 3d will be discussed in Sections IV and V.

During slope overload, the noise samples $r_O(n)$ will have magnitudes in the range

$$0 \leq |r_O(n)| < \infty. \tag{3}$$

The limit ∞ can be replaced by a more meaningful finite value if the input is bounded, as in band-limited speech. But this won't be necessary for the purposes of this paper.

The non-impulsive background in the $r(n)$ waveform is associated with input samples that do not cause slope overload. In this granular noise region, the maximum magnitude of noise sample $r_g(n)$ is simply half the ADPCM step size:

$$0 \leq |r_g(n)| \leq \Delta(n)/2. \tag{4}$$

IV. THE $(R + 1)$-BIT CODER WITH INSTANTANEOUS ONE-BIT QUANTIZATION OF $r(n)$

From the theory of one-bit quantization, the reconstruction level $\delta(n)$ that provides the minimum mean square error with a one-bit noise quantizer is given by the mean absolute value of quantizer input:

$$\delta(n)_{\text{opt}} = E[|r(n)|]. \tag{5}$$

Ignoring slope-overload samples $r_O(n)$, and assuming that the magnitudes of the $r_g(n)$ samples are uniformly distributed in the range 0 to $\Delta(n)/2$,

$$\delta(n)_{\text{opt}} \sim E[|r_g(n)|] = \frac{1}{2} \cdot \frac{\Delta(n)}{2} = \frac{\Delta(n)}{4}. \tag{6}$$

Simulations have shown that the probability of slope overload is small enough for the above design to be indeed very close to the optimum. This is illustrated by the s/n versus $\delta(n)$ plots in Fig. 4 for $R = 4$, 3, and 2 bits/sample. The signal-to-noise ratio is maximum

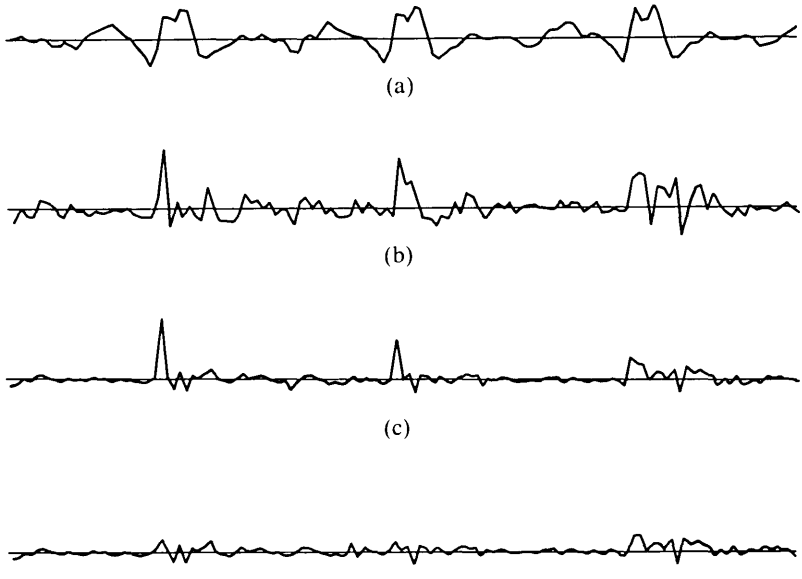


Fig. 3—(a) Input speech $x(n)$ (taken from *CF*) and reconstruction-error waveforms (b) (c) (d) in three ADPCM systems. All waveforms are 128 samples (16-ms) long. All error amplitudes are magnified by a factor of 10.

when the reconstruction level magnitude $\delta(n)$ of the 1-bit noise quantizer equals one-fourth the corresponding step-size $\Delta(n)$ in the R-bit ADPCM coder. When $\delta(n) = 0$, the system degenerates to the original R-bit ADPCM. Values at $\delta(n) = 0$ show the s/n of R-bit ADPCM.

Figure 3c shows the residual error after the $r(n)$ waveform (Fig. 3b) has been instantaneously quantized with a 1-bit/sample quantizer with reconstruction levels of $\pm\delta(n)_{\text{opt}}$. Note that the granular background components in $r(n)$ are uniformly reduced, but slope overload components are not.

The above step-size design implies that the noise coder is an instantaneous adaptive PCM (APCM) device that derives its step size from information that is already available in the R-bit ADPCM part of the $(R + 1)$-bit system. The N-sample buffer in Fig. 1 is not necessary for the operation of the instantaneous APCM coder.

The performance of the $(R + 1)$-bit system with instantaneous quantization is discussed at length in Section VI.

V. THE $(R + 1)$-BIT CODER WITH NON-INSTANTANEOUS ONE-BIT QUANTIZATION OF $r(n)$

Elimination of the impulsive components in $r(n)$ requires finer quantization. We now propose an algorithm that indeed allocates

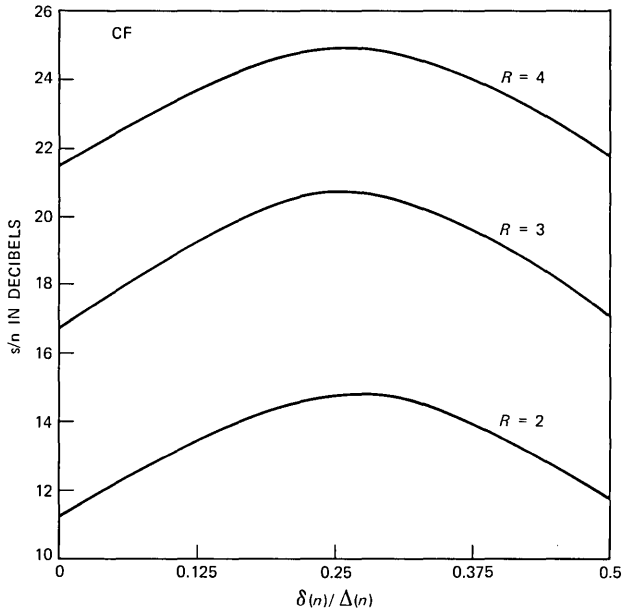


Fig. 4—Signal-to-noise ratio of $(R + 1)$-bit ADPCM system with instantaneous 1-bit quantization of reconstruction noise $r(n)$ from R-bit ADPCM.

$R_O > 1$ bit for slope-overload components, but still maintains an average bit rate of exactly 1 bit/sample in every block of N samples of $r(n)$. This is accomplished by assigning $R_g = 0$ bit/sample for granular noise components of very low magnitude in the block. The N-sample buffers and N-sample delays in Fig. 1 will be used to effect the above variable-bit assignment.

The location of slope-overload noise samples $r_O(n)$ and that of the low-magnitude granular noise samples $r_g(n)$ are both based on information that is already available to the R-bit ADPCM receiver, and therefore require no further side information to be transmitted.

The slope-overload samples are determined as those for which the quantizer output in R-bit ADPCM reaches the highest possible values for the given value of R (for example, levels associated with multiplier M_2 with $R = 2$ and levels associated with multiplier M_4 with $R = 4$; see Fig. 2).

The low-magnitude granular noise samples are located by rank-ordering $\Delta(n)$ values in the N-block, and by assigning zero bits to as many of these samples as necessary, in order of increasing $\Delta(n)$, until the total number of bits in the block is exactly N. While picking these zero-bit samples, it is very important to exclude samples associated with the use of highest output level. This precaution is needed because slope-overload errors can be associated with small values of $\Delta(n)$ as

well as large ones. In fact, as mentioned in the last paragraph, a defining cue for slope overload is not the value of the current *step size*, but rather the value of the current ADPCM quantizer *output level* (or current step-size *multiplier* if output levels and multiplier values have a one-to-one mapping, as in Fig. 2).

The net result of the above procedure will be to assign $R_O > 1$ bits where noise magnitudes are guaranteed to be the highest, and to assign $R_g = 0$ bits where noise magnitudes are guaranteed to be the smallest. The remaining samples are assigned 1 bit/sample as in Section IV.

The variable-bit-rate algorithm follows the constraint that the total number of bits per block is N:

$$N_O \cdot R_O + (N - N_O - N_g) \cdot 1 + N_g \cdot 0 = N; \qquad N_g = N_O(R_O - 1), \quad (7)$$

where N_O and N_g are the numbers of slope-overload and low-magnitude granular samples in N samples of $r(n)$. Note that the constraint above also implies that

$$N_O \cdot R_O = N_O + N_g \leq N; \qquad N_O \leq N/R_O. \quad (8)$$

This latter constraint on N_O is explicitly enforced even in those cases where the number of maximum multiplier samples may exceed N/R_O, for a chosen R_O.

The design of R_O should reflect the probability of use of the maximum reconstruction level in the R-bit ADPCM coder. This probability controls the fraction N_O/N. As shown in Ref 7, this probability is a decreasing function of R; consequently, the maximum allowable value of R_O that does not violate (8) is an increasing function of R. In fact, in our experiments, we have found that for N values of interest, the s/n maximizing values of R_O happen to be very close to the number of bits/sample R in the basic ADPCM coder. Thus, for example, the slope-overload bursts in 3-bit ADPCM are quantized with a second stage of APCM coding with an appropriately designed 3-bit quantizer.

5.1 Design of noise-quantizer characteristic

Figure 5 illustrates quantizer characteristics that were experimentally found to provide nearly minimum mean square error in noise quantization. The smallest outputs in each of these characteristics are the $\pm\Delta(n)/4$ levels used in the instantaneous noise quantizer of Section IV. The largest output levels are $\pm\Delta(n)$ and $\pm 3\Delta(n)$ in the non-instantaneous quantizers for $R = 2$ and 3. For $R = 4$, the largest output levels in the noise quantizer will be $\pm 7\Delta(n)$. All these numbers obviously depend only on $\Delta(n)$, a value already available to the R-bit ADPCM receiver.

In one experiment with $N = 32$ and input *CF*, the number N_O of $r(n)$ samples coded with $R_O > 1$ bits/sample were 2, 3, 9, and 15,

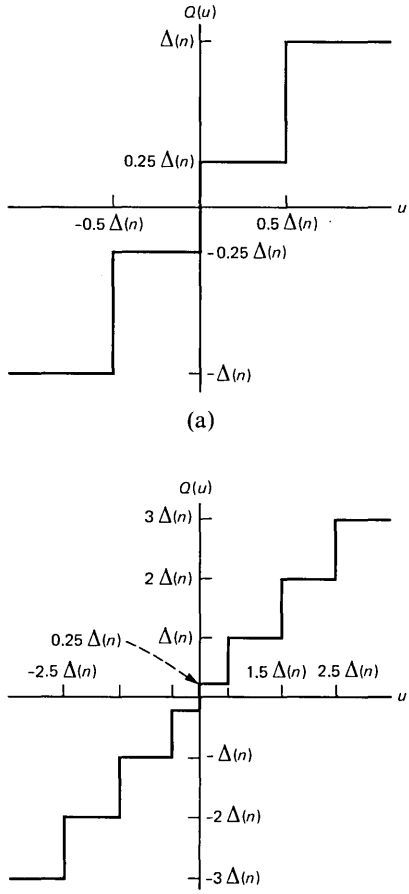


Fig. 5—Quantization characteristics used for overload noise samples $r_O(n)$ in the non-instantaneous coding of ADPCM noise when the average rate for noise coding is R_n = 1 bit/sample. The ADPCM bit rate R is 2 in (a) and 3 in (b).

respectively, with 5, 4, 3, and 2-bit DPCM. These numbers reflect the much higher probability of using the maximum quantizer output level as R decreases. With the recommended design $R_O = R$, note that $N_O \cdot R_O < N = 32$ in all the four examples above, as required in (8). With $N = 128$ and the same input CF, values of N_O were 5, 9, 19, and 32, respectively.

5.2 Design of block length N

The buffer length N should be large enough so that for every noise sample coded with $R_O > 1$, there is an adequate selection of noise samples for which bit stealing ($R_g = 0$) is appropriate. However, the

quasi-periodic nature of slope-overload bursts (Fig. 3b) indicates that N need be no greater than the pitch period. This is indeed demonstrated in Fig. 6, which plots the signal-to-noise ratio of the $(R + 1)$-bit system as a function of N. Note that the performance at $N = 32$ (which is close to the pitch period 40 in Fig. 3 for CF) is very close to that at $N = 128$. Note also that the gain with $N = 32$, over the instantaneous quantization scheme of Section IV (the case of $N = 1$), is over 2 dB. Gains over $N = 1$ are less in the case of $R = 2$.

Figure 3d shows the residual error after $r(n)$ has been quantized with an average rate of 1 bit/sample, with $N = 32$ (a buffer length of 4 ms, with 8-kHz samples). Note that unlike the instantaneous quantization scheme of Fig. 3c, even the impulsive components in $r(n)$ have been nearly eliminated in Fig. 3d. This is a result of quantizing these components with $R_O > 1$ bits/sample; $R_O = 4$ in this example. Since the impulsive components of the noise waveform $r(n)$ tend to occur predominantly during pitch-period onset, the system with non-instantaneous quantization can also be regarded as a form of "pitch-compensated" quantization.[8]

VI. SIGNAL-TO-NOISE RATIO RESULTS FOR R-BIT AND $(R + 1)$-BIT CODERS

Figures 7 and 8 compare the performance of the coders of Sections IV and V with that of conventional single-stage ADPCM.

The signal-to-noise ratios are averages over the entire length of a given utterance. The segmental s/n is obtained by obtaining the signal-

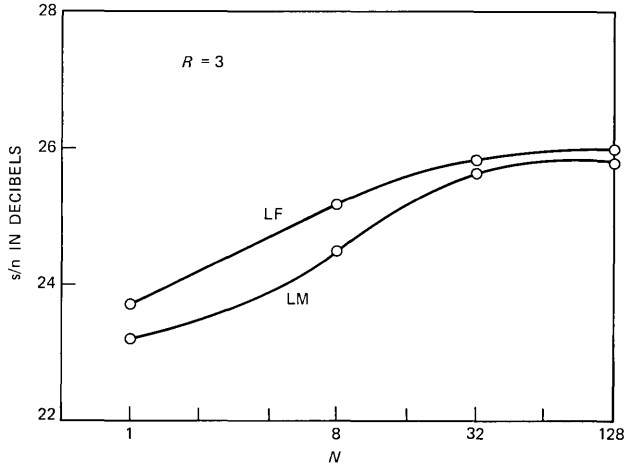


Fig. 6—Signal-to-noise ratio $(R + 1)$-bit ADPCM system with variable-rate quantization of reconstruction noise $r(n)$ from R-bit ADPCM. The signal-to-noise ratio reaches a value close to the maximum with a noise-buffer length of $N = 32$ (encoding delay of 4 ms). The gain over instantaneous noise quantization $(N = 1)$ is in excess of 2 dB.

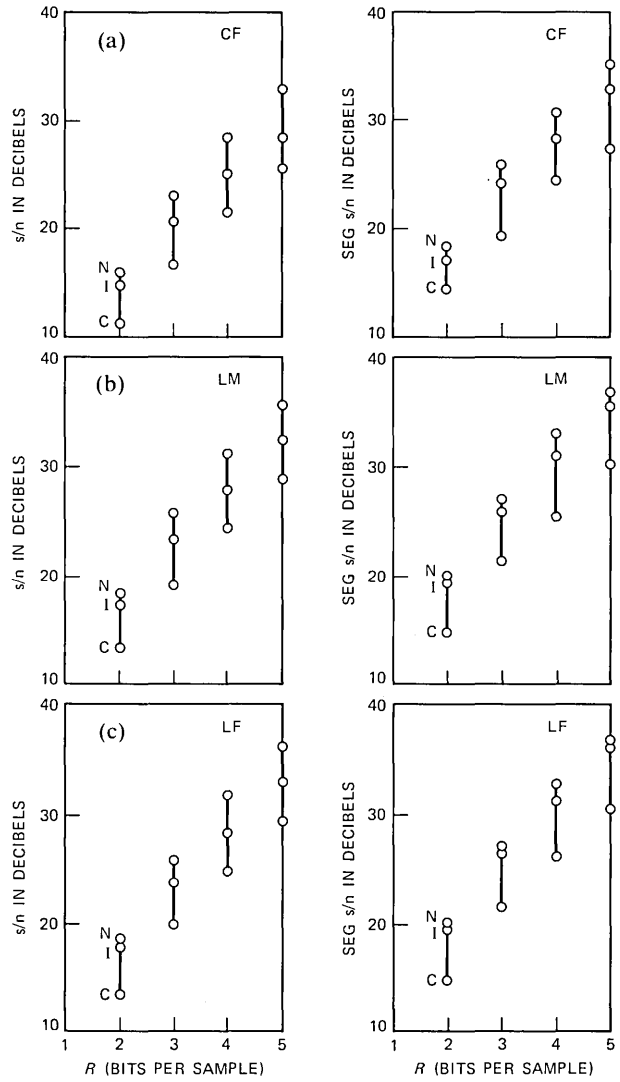


Fig. 7—Signal-to-noise ratio (s/n) and segmental signal-to-noise ratio (SEG s/n) in ADPCM systems, as a function of bit rate R of the basic coder in Fig. 1. For each value of R, there is an ordered set of three s/n or SEG s/n values. Plots in (a), (b), and (c) refer to speech inputs *CF*, *LM*, and *LF*.

to-noise ratio in dB for each 16-ms segment of an input, and by averaging such dB values over the entire length of a given utterance.

Figure 7 shows, for each bit rate R of the *c*onventional ADPCM system (C), signal-to-noise ratio gains in $(R + 1)$-bit systems with *i*nstantaneous (I) and *n*on-instantaneous (N) quantization of $r(n)$, with a total of 32 bits of quantization in every 32-sample block of $r(n)$.

Note that except in the case of $R = 2$, the performance of the $(R + 1)$-bit system with instantaneous quantization (the middle point on each vertical bar) is very close to conventional $(R + 1)$-bit ADPCM (the lowest point on the next vertical bar to the right), with an s/n gap of no more than 1 dB. Note also that for $R > 2$, the $(R + 1)$-bit system with non-instantaneous quantization (the topmost point on each vertical bar) is always better than conventional $(R + 1)$-bit ADPCM, with an s/n gain of as much as 3 dB. The substantial gains at $R = 4$ and 5

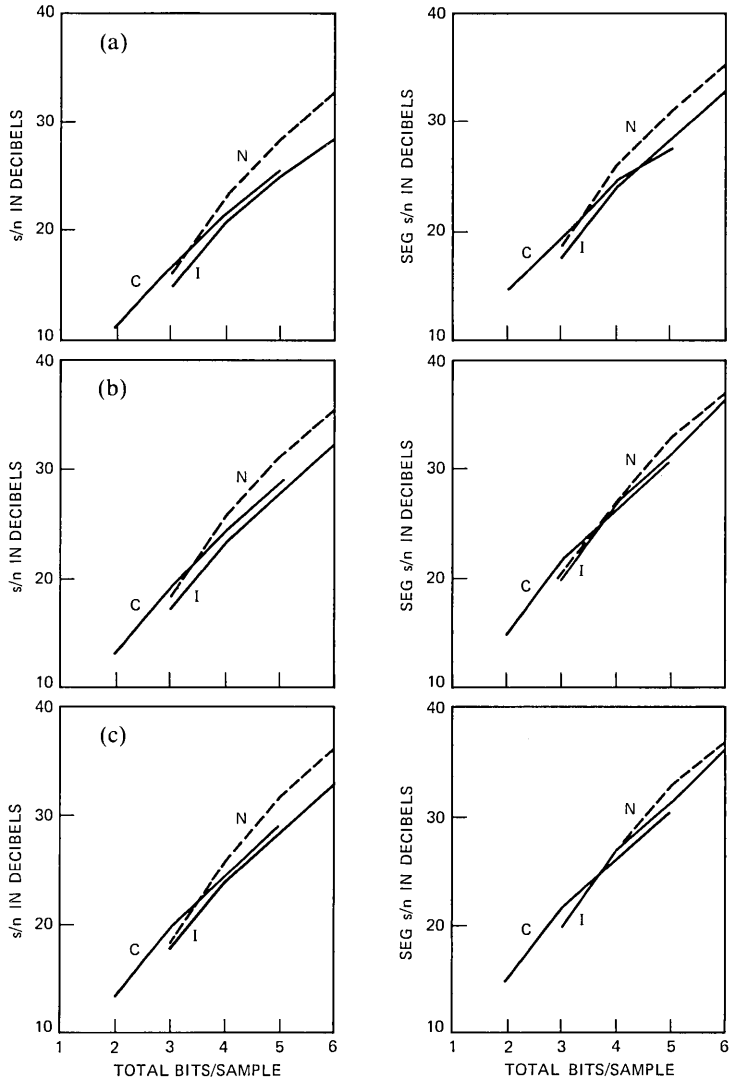


Fig. 8—Results of Fig. 7 replotted as a function of total bit rate.

may be due partly to the fact that the ADPCM quantizer in these cases is somewhat suboptimal; as R increases, optimal design of the 2^{R-1} step-size multipliers (Fig. 2) becomes increasingly difficult, and the s/n of the conventional ADPCM coder increases by less than the expected 6 dB per additional bit.

Figure 8 replots the results of Fig. 7, and compares the three coders discussed above, for given fixed values of total bit rate. Note once again that if the overall bit rate is at least 4 bits/sample, the $(R + 1)$-bit coder with instantaneous quantization is very close to conventional $(R + 1)$-bit ADPCM; while the $(R + 1)$-bit coder with non-instantaneous quantization is consistently better than $(R + 1)$-bit ADPCM.

VII. PERCEPTUAL EVALUATIONS OF THE CODERS OF SECTION IV AND V

Critical headphone listening reinforces the results suggested in Section VI. As expected, with $R = 2$, the outputs of 3-bit systems of Sections IV and V are both slightly worse than those of conventional 3-bit ADPCM. But with $R \geq 3$, even the output of the simpler $(R + 1)$-bit system (the system with instantaneous quantization) sounds extremely close to that of conventional $(R + 1)$-bit ADPCM. The very good perceptual performance of the instantaneous noise quantizer is very likely because much of its residual error (Fig. 3c) may be masked by the high-level speech activity in its temporal vicinity. In fact, the main motivation for the use of a non-instantaneous quantizer is not merely the increased performance with $(R + 1)$-bit coding, as demonstrated in Section VI, but also the fact that with more general $(R + R_n)$-bit coding $(R_n > 1)$, the performance of the instantaneous quantizer deteriorates rapidly, while that of the non-instantaneous quantizer maintains a 6-dB-per-bit behavior (Section VIII).

VIII. VARIABLE-RATE CODING WITH $R_n \geq 1$ BIT/SAMPLE

Sections IV through VII discussed the design and performance of a dual-rate system with $R_n = 0$ or 1, and a total bit rate of either R or $R + 1$ bits/sample. In this section, we consider a generalization to $R_n > 1$. Specifically, the average noise-coding bit rate R_n will range from 0 to 3, the ADPCM bit rate R will range from 2 to 5, and combinations of R and R_n will be such that the total bit rate $R + R_n$ will range from 2 to 6 bits/sample, the range used earlier in Fig. 8. We will note that the performance of an instantaneous noise-coding system deteriorates rapidly when $R_n > 1$, while that of a non-instantaneous noise-coding systems maintains an approximate 6-dB-per-additional-bit behavior.

8.1 Instantaneous noise coding

When $R_n = 1$, the recommended output levels for the APCM noise coder were $\pm 0.25\ \Delta(n)$. These levels are in fact centered in the ranges

0 to $\pm 0.5\ \Delta(n)$, the defining ranges for granular noise amplitude. As Fig. 4 shows, the design of the instantaneous quantizer is hardly affected by the occasional incidence of overload noise magnitudes much greater than $0.5\ \Delta(n)$. Generalizations to $R_n > 1$ therefore call for sets of 2^{R_n} APCM output levels that are uniformly spaced in the regions $-0.5\ \Delta(n)$ to $+0.5\ \Delta(n)$. For example, with $R_n = 2$ and 3, the output levels will be

$$R_n = 2\text{: } [\pm 0.125\Delta(n),\ \pm 0.375\Delta(n)]$$

and

$$R_n = 3\text{: } [\pm 0.0625\Delta(n),\ \pm 0.1875\Delta(n),\ \pm 0.3125\Delta(n),\ \pm 0.4375\Delta(n)]. \quad (9)$$

Experiments with $R_n = 2$ and 3 show that the above design is indeed nearly optimal for instantaneous coding. However, the performance of the instantaneous coder deteriorates badly as R_n increases, as we will see in Fig. 10. This is to be expected from the illustrative residual noise waveform of Fig. 3c, which shows that instantaneous coding is characterized by residual errors of very significant amplitude during periods of ADPCM overload. The situation does not improve with increasing R_n because the additional output levels that become available are simply used up for finer quantization in the granular noise region, shown in eq. (9).

8.2 Non-instantaneous noise coding

As we saw in the residual noise waveform of Fig. 3d for the example of average noise bit rate $R_n = 1$, non-instantaneous coding of the noise waveform can reduce the extent of granular noise as well as that of overload distortion in ADPCM coding. Slope-overload bursts are still visible in the residual noise waveform of Fig. 3d, but the waveform is much less impulsive than the original noise waveform of Fig. 3b. With $R_n > 1$, both the overload and granular components in Fig. 3d can be made increasingly smaller, provided that the bit allocation and quantizer design of Section V are properly generalized.

Recall that for an average noise bit rate of $R_n = 1$, the bit allocation (7) of Section V was as follows:

$$\begin{aligned} &R_O \text{ bits for } N_O \text{ overload noise samples} \\ &0 \text{ bits for } N_g = N_O(R_O - 1) \text{ low-amplitude noise samples} \\ &1 \text{ bit for } N - N_O - N_g \text{ remaining noise samples.} \end{aligned} \quad (10)$$

The total number of bits is then N for every block of N samples, as in (7). As noted in (8), the above constraint also implies that $N_O \leq N/R_O$. This condition has to be explicitly enforced even when the number of actual overload noise samples exceeds N/R_O for a chosen R_O. A simple

generalization of (10) that works very well with $R_n > 1$ is shown below:

$$\begin{aligned} &R_O + (R_n - 1) \text{ bits for } N_O \text{ overload noise samples} \\ &(R_n - 1) \text{ bits for } N_g = N_O(R_O - 1) \text{ low-amplitude noise samples} \\ &R_n \text{ bits for } N - N_O - N_g \text{ remaining noise samples.} \end{aligned} \quad (11)$$

The total number of bits is now NR_n for every block of N samples. Furthermore, (11) is a straightforward generalization of (10); and as in the case of (10), it requires no transmission of any side information for bit-allocation purposes, but only encoding and decoding delays in the order of $N = 32$ (4 ms, assuming 8-kHz samples).

Critical to the success of the bit-allocation algorithm (11) is a proper design of individual quantization characteristics. Unlike the instantaneous design of (9), the variable-bit allocations in (11) permit finer quantizer resolutions both in the overload range, $|r(n)| > 0.5\Delta(n)$, and in the granular noise range, $|r(n)| < 0.5\Delta(n)$. A systematic way to design these quantizers is to start with the designs in Section V (for a given R, and for an average-noise-bit rate of $R_n = 1$). Recall that each such design involves a set of three characteristics, for 0-bit, 1-bit, and $R_O = R$-bit quantization, as in (10). As the value of R_n increases, each of these sets evolves into corresponding sets of three characteristics, for $(R_n - 1)$-bit, R_n-bit, and $(R_O + R_n - 1)$-bit quantization, as in (11). Resolutions improve by a factor of two for each stage of increase of R_n, and this improvement benefits the overload as well as granular regions of coding noise. Figure 9 illustrates the quantizer evolution for the example of $R = 3$ and $R_n = 1$ and 2. The illustration includes only one of the set of three quantizers involved in the coding process. This is the R_O-bit characteristic (Fig. 9a, which is the same as Fig. 5b) used for quantizing the N_O overload noise samples in the $R_n = 1$ system. When $R_n = 2$, the above R_O-bit (in this case, 3-bit) characteristic evolves into a $R_O + R_n - 1 = 4$-bit characteristic (Fig. 9b).

Figure 10 shows the benefits of increasing R_n in a non-instantaneous noise-coding system, for the example of $R = 4$ and for average-noise-coding rates of $R_n = 1$, 2, and 3 bits/sample. All error waveforms are magnified by a factor of 50. The waveform in (b) is the same as that in Fig. 3d, but is magnified by a factor of 5. In Fig. 10 we see a significant reduction in residual noise level for each stage of increasing R_n. We will note presently that the improvement is very close to 6 dB per additional bit in R_n.

8.3 Signal-to-noise ratios

Figures 11a and b show s/n and segmental s/n results for explicit noise-coding systems with $R_n \geq 1$. The range of total bit rate $R + R_n$ is 2 to 6, the same as that in Fig. 8. The solid curves show the

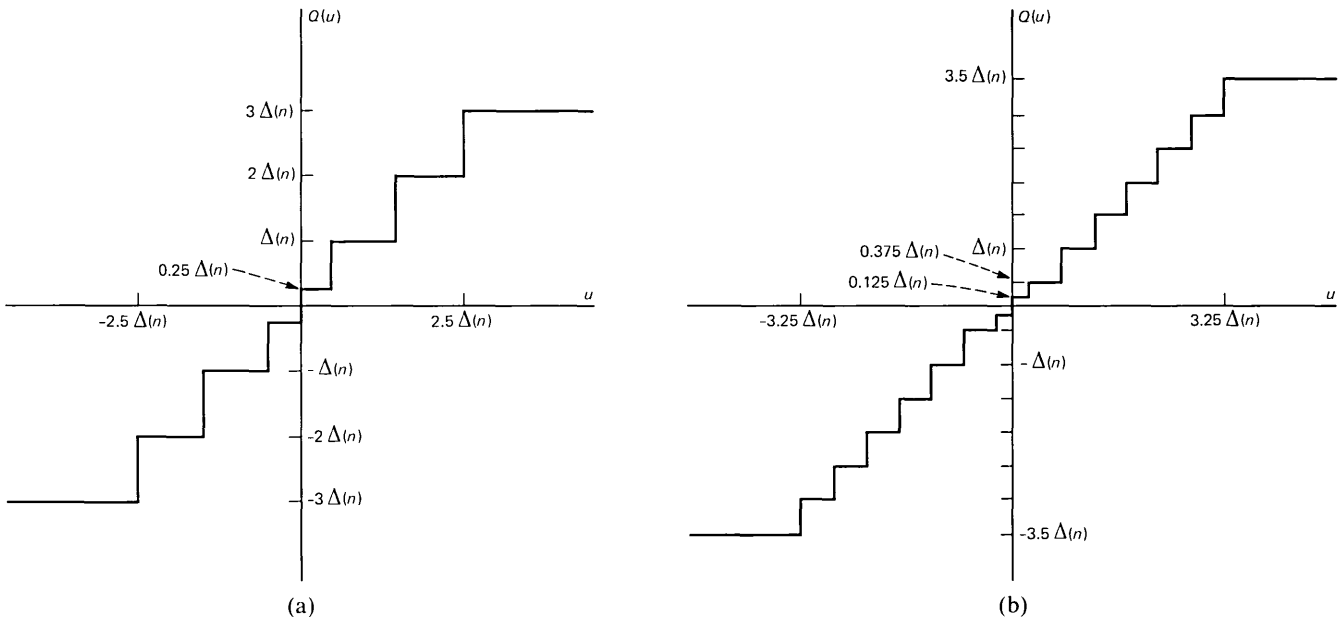


Fig. 9—Examples of quantization characteristics used for overload noise samples $r_O(n)$ in the non-instantaneous coding of 3-bit ADPCM noise with an average rate of (a) R_n = 1 bit/sample and (b) R_n = 2 bits/sample.

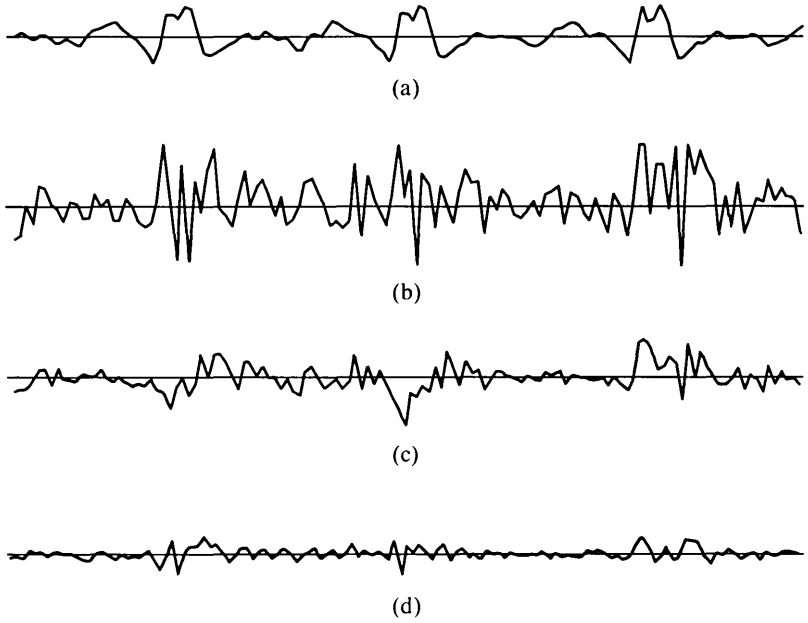


Fig. 10—(a) Input speech $x(n)$ and reconstruction error waveforms in variable-rate coding with 4-bit ADPCM, non-instantaneous noise coding and average-noise-coding rates of (b) $R_n = 1$, (c) $R_n = 2$, and (d) $R_n = 3$ bits/sample. All error waveforms are magnified by a factor of 50. The waveform in (b) is the same as that in Figure 3(d), but magnified by a factor of 5.

performance of conventional ADPCM. The circles labelled 3 show the performance of a variable-rate system based on instantaneous coding of ADPCM noise, for the example of $R = 3$. We can see that with $R_n > 1$, the s/n performance of the instantaneous coding system deteriorates fairly rapidly, compared with that of $(R + R_n)$-bit ADPCM, with increasing bit rate, but its segmental s/n performance is competitive with that of conventional ADPCM at all bit rates. Non-instantaneous coding systems, on the other hand, maintain a 6-dB per additional bit behavior, provided only that $R > 2$. This is shown by the sets of solid black dots labelled 3, 4, and 5. The performances of these systems also exceed that of conventional ADPCM at any given total bit rate, a result already noted in Section VI for the special case of $R_n = 1$.

IX. EFFECTS OF TRANSMISSION ERRORS

Bit errors in transmission can affect the noise-coding system in two ways: they may produce effects attributable to errors in the transmission of the bits from the basic DPCM coder, and effects attributable to errors in the bits from the noise coder. Effects of both types are expected to be more severe in the case of the non-instantaneous coder.

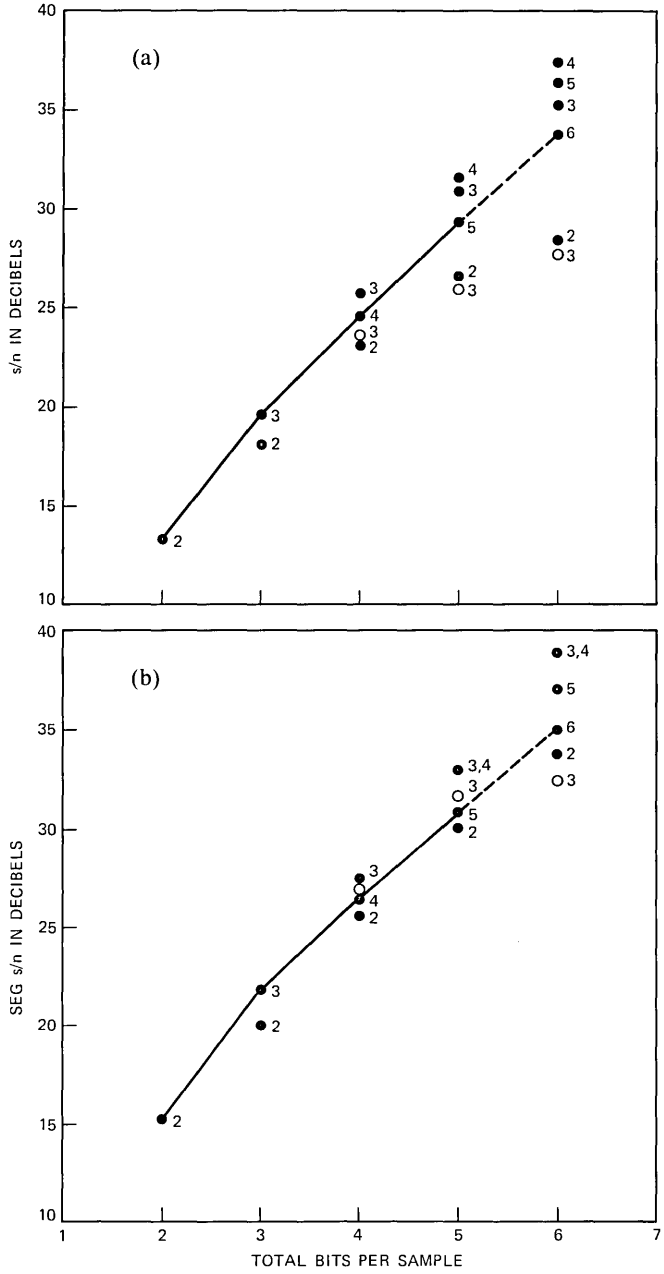


Fig. 11—(a) Signal-to-noise ratio (s/n) and (b) segmental signal-to-noise ratio (SEG s/n) in ADPCM systems, as a function of total bit rate. The solid curves refer to conventional ADPCM ($R_n = 0$) and the black dots refer to non-instantaneously quantized noise-coding systems with $R = 2$, 3, 4, and 5 and $R_n \geq 1$ bits/sample. The circles refer to an instantaneously quantized noise-coding system with $R = 3$, and $R_n = 1$, 2, and 3 bits/sample.

The greater error sensitivity of this system is due first to the presence of quantizers with larger step sizes, and hence, proportionally larger channel-error effects. More important, the increased error sensitivity of the non-instantaneous system is a result of the variable-bit-allocation algorithm, which will be miscalculated at the receiver if one or more bits from the basic ADPCM coders are received in error. Errors of this type do not propagate beyond a given N-sample block, but their effects can be severe enough to warrant the complete disabling of the noise-coding part of the system when errors are detected. A simple example of an error-detection system is one where the odd-even parity of the number N_O of overload samples is explicitly transmitted to the receiver. A change in the parity of N_O, as computed at the receiver, is a good detector of perceptually significant single-bit errors in the given block. The single bit needed to transmit the parity information, or the multiplicity of bits needed to transmit the information in an error-protected format, can be incorporated in the coder output by a bit-stealing procedure based on increasing the number of zero-bit noise samples from N_g to an appropriately greater number.

Irrespective of the noise-coding method and the procedures that may be used to protect the noise-coding system from errors, the basic ADPCM coder can be made error-robust, at least for independent error rates of up to 10^{-3}, by using robust adaptive-quantizer algorithms such as the leaky-adaptation algorithm in Ref. 9.

X. CONCLUSIONS

We have demonstrated simple systems for variable-rate, embedded ADPCM coding of speech based on explicit coding of reconstruction noise. These systems do not require the transmission of any side information other than what is already available in a conventional ADPCM decoder. The simpler of two systems proposed in this paper uses instantaneous coding of the noise, and provides a performance very close to that of the conventional ADPCM at any given value of total bit rate $R_T = R + R_n$, for the simple but non-trivial case of dual-rate operation ($R_n = 0$ or 1 bit/sample). But its s/n performance deteriorates significantly with more widely variable operation ($R_n > 1$ bits/sample). The more complex system uses non-instantaneous noise-coding, with coding and decoding delays in the order of 4 ms to realize positive gains over conventional ADPCM at any given total bit rate $R + R_n$ bits/sample. The performance of this system has been demonstrated for $R_n = 0$, 1, 2, and 3 bits/sample, and for $R = 2$, 3, 4, and 5 bits/sample. The system with non-instantaneous noise coding can also be regarded as an $(R + R_n)$-bit ADPCM coder with a quantizing system that is better than conventional adaptive quantization with a one-word memory.

REFERENCES

1. J. Das and P. K. Chatterjee, "Optimized Δ − Δ Modulation System," Elect. Lett., *3,* No. 6 (June 1967), pp. 286–7.
2. R. Steele, "Peak Signal-to-Noise Formulas for Multistage Delta Modulation with *RC*-Shaped Gaussian Input Signals," B.S.T.J., *61,* No. 3 (March 1982), pp. 347–62.
3. T. Bially, B. Gold, and S. Seneff, "A technique for adaptive voice flow control in integrated packet networks," IEEE Trans. Commun. (March 1980), pp. 325–33.
4. Y.-C. Ching, "Differential Pulse Code Communications System Having Dual Quantization Schemes," U.S. Patent 3,781,685, December 25, 1973.
5. D. J. Goodman, "Embedded DPCM for Variable Bit Rate Transmission," IEEE Trans. Commun. (July 1980), pp. 1040–46.
6. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52,* No. 7 (September 1973), pp. 1105–18.
7. N. S. Jayant, "Adaptive Quantization With a One-Word Memory," B.S.T.J., *52,* No. 7 (September 1973), pp. 1119–44.
8. D. L. Cohn and J. L. Melsa, "The Residual Encoder—An Improved ADPCM System for Speech Digitization," IEEE Trans. Commun. (September 1975), pp. 935–41.
9. D. J. Goodman and R. M. Wilkinson, "A Robust Adaptive Quantizer," IEEE Trans. Commun. (November 1975), pp. 1362–5.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

Random Processes With Specified Spectral Density and First-Order Probability Density

By M. M. SONDHI

(Manuscript received August 6, 1982)

The procedure for generating a Gaussian process with a specified spectral density is well known. It is harder to generate a process with a specified spectral density and a specified first-order probability distribution. In this paper we explore, by simulation, the possibility of generating a process with such a dual specification by passing a Gaussian process with an appropriately chosen spectral density through an appropriately chosen zero-memory nonlinearity. Several applications are cited where such a dual specification is desirable.

I. INTRODUCTION

The procedure for generating a Gaussian random process whose power spectral density (psd) is a specified function of frequency, $S(\omega)$, is well known. As we see in Fig. 1, let $H(j\omega)$ be the transfer function of a linear time-invariant filter, and let the input to the filter be a Gaussian random process $\{x(t)\}$ with psd $\Phi_x(\omega)$. Then the psd of the output process $\{y(t)\}$ is given by

$$\Phi_y(\omega) = H(j\omega)H^*(j\omega)\Phi_x(\omega) \tag{1}$$

where * denotes complex conjugation. Since the filter H is linear, $\{y\}$ is also a Gaussian random process, and if $\{x\}$ is a white noise[†] with unit psd, then $\{y\}$ has the desired psd provided

$$H(j\omega)H^*(j\omega) = S(\omega). \tag{2}$$

H is then called a shaping filter (see Ref. 1) for the psd $S(\omega)$. The spectral factorization for eq. (2) can be accomplished analytically when

[†] A high degree of mathematical rigor is not intended here. For our purpose we define white noise as a noise whose psd is constant over a wide bandwidth $(-W, W)$ and zero outside. The bandwidth is assumed wide enough so that any desired psd is negligible outside it.

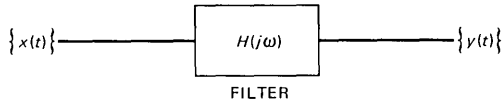


Fig. 1—Generating a Gaussian random process with specified-power spectral density.

$S(\omega)$ is a rational function of ω^2.[1] In such cases the resulting shaping filter too can be synthesized by a standard procedure as a lumped parameter filter (see Ref. 2). The appropriate shaping filter can of course be derived for more general spectral densities as well as for various nonstationary processes. Numerical approximations to the shaping filter can be obtained for any reasonably well-behaved spectral density, as we shall see in Section III.

This method of generating the random process, however, leaves no choice as to the first-order probability density function (pdf) of the process, i.e., the pdf of the random variable $Y = y(t)$. Since $\{y(t)\}$ is a Gaussian process, all joint probabilities of the random variables $Y_i = y(t_i)$, $i = 1, 2, \cdots$ are Gaussian. Hence, the pdf of Y too is Gaussian. (Of course, if *only* the pdf is specified, then one does not have to follow the above procedure. It is always possible to generate a white noise with any specified pdf by passing, for instance, a uniform white noise or a Gaussian white noise through a zero-memory nonlinearity. The procedure is quite analogous to the one discussed in Section 2.1.)

Frequently it is desirable to specify both the spectral density and the first-order pdf of the process. One situation where such a specification would be useful is in the simulation of speech-waveform coders. The performance of such coders can depend significantly upon both the spectral density and the pdf of the signal to be coded. Measurements have shown[3] that the pdf for speech signals is markedly different from Gaussian, and is in fact much better represented by a "gamma" distribution[4] (see Section II). At present, simulations with such coders are carried out on Gaussian processes with appropriately shaped spectra, or on sequences of uncorrelated samples with a gamma pdf. The behavior of the coders on speech signals is not well predicted by either of these; hence, tests are also performed on a variety of speech sentences. Perhaps these tests could be standardized and their predictive value improved by the use of random processes with a gamma pdf and a selection of typical spectral shapes.

Another area where this dual specification can be important is in perceptual studies. One such application is to Julesz's experiments on texture perception.[5] The independent control of spectral density and pdf of random dot patterns would enable one to decide between competing theories of texture discrimination.

Finally, such control of pdf and spectral density would be useful in

studying input-output properties of nonlinear systems, which can be represented by a zero-memory nonlinearity followed by a linear filter.

The problem of synthesizing a random process to approximate the pdf and power spectral density of a given process has been addressed in the literature.[6] However, to the best of our knowledge no exact procedure is known at present for generating a random process with specified pdf and spectral density. In the next two sections we explore, by simulation, the capabilities and limitations of one simple attack on the problem.

II. GENERATION OF THE RANDOM PROCESS

The method of generation that we wish to study is shown schematically in Fig. 2. For simplicity we will assume throughout that the desired random process has zero mean and that the desired pdf has even symmetry. These restrictions are by no means essential for our analysis, but only simplify our presentation.

The basic idea of the proposal is as follows: We start with a "white" Gaussian noise [see the footnote in Section I] and pass it through a filter $H(j\omega)$ and a scale factor K such that the random process $\{x(t)\}$ is a zero-mean-Gaussian process with unit variance. Let $q(.)$ be the desired pdf of the random variable $Y = y(t)$. Then it is straightforward to find the zero-memory nonlinearity $F(.)$ such that Y has the desired pdf. The problem then is to find H such that *after the nonlinear distortion by* $F(.)$ the spectral density at the output is the desired function $S(\omega)$. It is easy to come up with examples for which this problem has no solution. For instance it can be shown that $\{v(t)\}$ cannot be a strictly band-limited process for *any* choice of limiting nonlinearity $F(.)$.[7] Nevertheless, as we will show, in a variety of cases of interest the problem has a solution, or an approximate solution.

Before proceeding to a detailed description of the method, we may emphasize the reason for the choice of a Gaussian process for the input. This is the property of the Gaussian process that it stays Gaussian after a linear transformation. The Gaussian process is the only well-behaved process that has this property. The same reason also dictates the order of operations in Fig. 2. Thus, interchanging the filter and zero-memory nonlinearity would be equivalent to generating the output process by a linear transformation of a non-Gaussian white

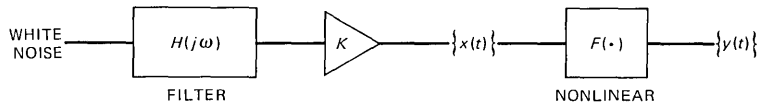


Fig. 2—Method for generating a random process with specified-power spectral density *and* specified first-order probability density.

noise. While it is possible to generate processes with non-Gaussian pdf's this way, it is not easy to compute (or control) the pdf of the output process.

2.1 Derivation of the function F(.)

Let us begin by deriving the required function $F(.)$. It is convenient to think of $F(.)$ as a composition of two functions, $P(.)$ and $f(.)$, as shown in Fig. 3. In view of the assumptions about the process $\{x(t)\}$, the random variable $X = x(t)$ has the pdf

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{3}$$

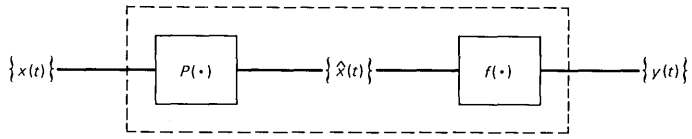
If $\hat{X} = P(X)$ where $P(x)$ is the cumulative distribution

$$P(x) = \int_{-\infty}^{x} p(\lambda) d\lambda, \tag{4}$$

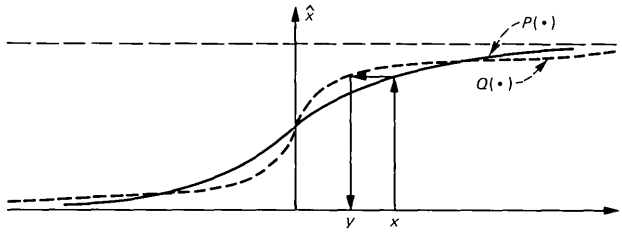
then the pdf of the random variable $\hat{X}$ is uniform on the interval (0, 1). Similarly, if $Y = f(\hat{X})$ and Y has the desired pdf $q(y)$, then the cumulative distribution $Q(y)$ is related to the function f^{-1} by the equation

$$Q(y) = \int_{-\infty}^{y} q(\lambda) d\lambda = f^{-1}(y). \tag{5}$$

Thus, in order for Y to have the desired pdf, $f(.)$ must be chosen to



(a)



(b)

Fig. 3—Computing the nonlinearity as the composition of a Gaussian cumulative probability and the inverse of the desired cumulative distribution.

satisfy eq. (5). As long as $q(y)$ is not identically zero over an interval, Q is an invertible function and eq. (5) defines f. The technical difficulty arising when this condition is not met is resolved in an obvious manner. Thus, if $q(y)$ is nonzero everywhere except on the interval $a \leq y \leq b$, then $Q(y)$ is constant from a to b. Hence, $\hat{X}$ as a function of Y has a jump at $Y = Q(a)$ of magnitude $b - a$. The limit of the function f can be defined from above and below this value. At $Q(a)$ the function may be specified arbitrarily.

Finally, the function $F(.)$ is given by

$$F(x) = f[P(x)]. \tag{6}$$

From the assumed symmetry of $q(y)$ it is evident that $F(.)$ must turn out to be an odd function of its argument.

By way of illustration in this paper we will consider three output pdf's:
the uniform pdf:

$$q(y) = \frac{1}{2\sqrt{3}}, \qquad |y| \leq \sqrt{3}; \tag{7a}$$

the gamma* pdf

$$q(y) = \frac{3^{1/4}}{\sqrt{8\pi}} \frac{1}{\sqrt{|x|}} e^{-|x|\sqrt{3}/2}; \tag{7b}$$

and the binary pdf

$$q(y) = \tfrac{1}{2}[\delta(y - 1) + \delta(y + 1)]. \tag{7c}$$

In each case the constants have been chosen to normalize the variance of the pdf to be 1. Our choice of examples is of course arbitrary. However, the uniform and the binary are obvious simple examples one expects to be useful, and the gamma has the relevance to speech signals mentioned above. The function $F(.)$ turns out to be quite simple for each of these cases:
for the uniform pdf

$$F(x) = 2\sqrt{3}[P(x) - \tfrac{1}{2}], \tag{8a}$$

for the gamma pdf

$$F(x) = \frac{1}{\sqrt{3}} x^2 \text{sign}(x), \tag{8b}$$

and for the binary pdf

$$F(x) = \text{sign}(x). \tag{8c}$$

* Strictly speaking, a gamma 1/2 distribution.

2.2 The covariance function of $\{y\}$

Having derived the nonlinearity F, we must now find the relationship between the spectral properties of $\{x\}$ and $\{y\}$. Because of the nonlinear transformation, the relationship is best derived in the time domain, i.e., between the autocovariance functions of the two processes. [Recall that the autocovariance function (acf) is just the Fourier inverse transform of the spectral density.] Let $\rho(\tau)$ be the acf of the process $\{x\}$, i.e., $\rho(\tau) = E[x(t)x(t + \tau)]$. Let $g(u, v, \rho)$ be the unit variance, zero-mean, two-dimensional Gaussian pdf given by

$$g(u, v, \rho) = \frac{1}{2\pi\sqrt{1 - \rho^2}} e^{-\frac{1}{2}(u^2+v^2-2\rho uv)}. \tag{9}$$

Then the acf of the $\{y\}$ process is given by

$$R = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(u)F(v)g(u, v, \rho)dudv, \tag{10}$$

where, of course, ρ (and hence R) is a function of the lag, τ. A general method of evaluating the integrals in eq. (10) is to use Mehler's expansion

$$g(u, v, \rho) = \frac{1}{2\pi} e^{-\frac{1}{2}(u^2+v^2)} \sum_0^{\infty} \frac{\rho^n}{n!} \mathbf{He}_n(u)\mathbf{He}_n(v) \tag{11}$$

of the two-dimensional Gaussian pdf in terms of Hermite polynomials.[8] Then if $F(.)$ has an expansion in terms of the same polynomials

$$F(x) = \sum_0^{\infty} f_n \mathbf{He}_n(x)/\sqrt{n!}, \tag{12}$$

it can be shown[7] that

$$R = \sum_1^{\infty} f_n^2 \rho^n. \tag{13}$$

When F is an odd function, as is the case here, the even coefficients in eqs. (12) and (13) vanish, and it is seen that R is an odd function of ρ. [This can of course be seen directly from eq. (10) by noting that $g(u, v, -\rho) = g(u, -v, \rho)$, and changing the sign of one of the variables of integration.]

For the examples that we have chosen, there is no need to use this general procedure. In these cases the integration is easily carried out in closed form. As shown in the appendix, the relationship between R and ρ is as follows:

$$\text{Uniform: } R = \frac{6}{\pi} \sin^{-1} \frac{\rho}{2}. \tag{14a}$$

$$\text{Gamma: } R = \frac{2}{3\pi}\left[(1 + 2\rho^2)\sin^{-1}\rho + 3\rho\sqrt{1 - \rho^2}\right] \tag{14b}$$

$$\text{Binary: } R = \frac{2}{\pi}\sin^{-1}\rho. \tag{14c}$$

The three functions are shown in Fig. 4 for $0 \leq \rho \leq 1$; the odd symmetry gives the function for negative ρ as well.

2.3 Derivation of the Filter H

The functions in Fig. 4 are monotonic increasing functions over the whole range $-1 \leq \rho \leq 1$. [From eqs. (12) and (13) this property is seen to be true whenever F has odd symmetry.] Therefore, the relationship between ρ and R is invertible.

Suppose now that $R(\tau)$ is the Fourier transform of the desired spectral density $S(\omega)$. Then the function

$$\phi(\tau) = \rho[R(\tau)] \tag{15}$$

can be computed, where $\rho(R)$ is the function corresponding to the desired pdf. For the examples chosen we get

$$\text{Uniform: } \phi_u(\tau) = 2\sin\left[\frac{\pi}{6}R(\tau)\right] \tag{16a}$$

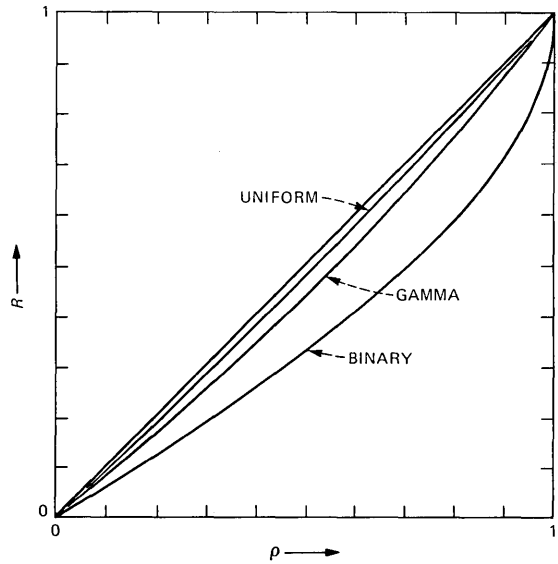


Fig. 4—The relationship between the autocovariance (R) at the output and (ρ) at the input of the nonlinearity appropriate to the generation of the three desired probability distributions. The unmarked line is $R = \rho$. Since the functions have odd symmetry, only the range of positive covariance values is shown.

$$\text{Gamma: } \phi_g(\tau) = \rho_g[R(\tau)] \tag{16b}$$

$$\text{Binary: } \phi_b(\tau) = \sin\left[\frac{\pi}{2} R(\tau)\right], \tag{16c}$$

where ρ_g is the inverse of the relation (14b). We are unable to give ρ_g explicitly; however, its numerical computation is trivial.

The function $\phi(\tau)$ is continuous and symmetric so that its Fourier transform, $\Phi(\omega)$, is real. For many specified acf's $R(\tau)$, $\Phi(\omega) \geq 0$ for all ω. In such cases we can obviously obtain an exact solution to the problem. All we need to do is to synthesize a filter, $H(j\omega)$, such that

$$K^2 H(j\omega) H^*(j\omega) = \Phi(\omega) = \mathbf{F}[\phi(\tau)], \tag{17}$$

where $\mathbf{F}$ denotes the Fourier transform. Unfortunately, $\Phi(\omega)$ is not guaranteed to be nonnegative at all frequencies for every specified acf $R(\tau)$. This is because when $\rho(\tau)$ in eq. (10) ranges over all possible acf's, $R(\tau)$ ranges only over a subset of possible acf's.

If the desired spectral density is such that $\Phi(\omega)$ becomes negative at some frequencies, the best we can do with our method is to give an approximate solution as follows: Define the function $\Phi^+(\omega)$ such that

$$\begin{aligned} \Phi^+(\omega) &= \Phi(\omega) \quad \text{if } \Phi(\omega) \geq 0 \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{18}$$

We then synthesize $H(j\omega)$ such that

$$K^2 H(j\omega) H^*(j\omega) = \Phi^+(\omega). \tag{19}$$

In the next section we will use eq. (19) [or its special case, eq. (17)] to generate random processes with a variety of pdf's and spectral densities.

III. SIMULATIONS

In this section we will describe the numerical procedures required to generate a random process with a pdf $q(.)$ and a spectral density $S(\omega)$, based on the theoretical discussion of the previous section. We have already shown how the nonlinearity $F(.)$ is to be computed. It remains to be shown how to numerically approximate the shaping filter.

3.1 The shaping filter

We will approximate the shaping filter as a transversal filter (FIR filter). Let $R(\tau)$ be the desired acf. Unless $R(\tau)$ happens to be of finite duration, it must first be truncated. To ensure that the truncated function $\hat{R}(\tau)$ is a legitimate acf, the truncation must be done by multiplying the given acf by an acf of finite duration. (Recall that the

product of two acf's is an acf). We chose a Hamming window convolved with itself as the truncating window. Thus let

$$w(\tau) = \int_{-\infty}^{\infty} h(t)h(t-\tau)dt, \tag{20}$$

where $h(t)$ is the Hamming window defined as

$$h(t) = 0.54 + 0.46 \cos \frac{\pi|t|}{T}, \; |t| \leq T \tag{21}$$

$$= 0 \qquad \text{otherwise.}$$

Then we define the truncated acf as

$$\hat{R}(\tau) = w(\tau)R(\tau). \tag{22}$$

For $R(\tau)$ decaying rapidly enough as $\tau \to \infty$, $\hat{R}(\tau)$ can be made to approximate $R(\tau)$ as closely as desired by choosing T sufficiently large. For the rest of the paper, therefore, we will regard $\hat{R}(\tau)$ as the desired acf, and $\hat{S}(\omega) = \mathbf{F}[\hat{R}(\tau)]$ as the desired spectral density.

The filter is synthesized from $\hat{R}(\tau)$ by the following sequence of steps:

(*i*) Discretize $\hat{R}(\tau)$ to give the sequence $\hat{R}_n$, $-N + 1 \leq n \leq N$.*

(*ii*) Compute the sequence $\phi_n = \phi(\hat{R}_n)$, where ϕ is the function defined by eq. (15) for the appropriate desired pdf.

(*iii*) Find the FFT of the sequence ϕ_n and set any negative values to zero. Let μ_n denote the resulting, adjusted Fourier transform.

(*iv*) The desired impulse response is then the inverse FFT of the sequence $\sqrt{\mu_n}$.

3.2 Examples

For each of the pdf's (a), (b), and (c) we have generated several examples of random processes with acf's selected from the following types:

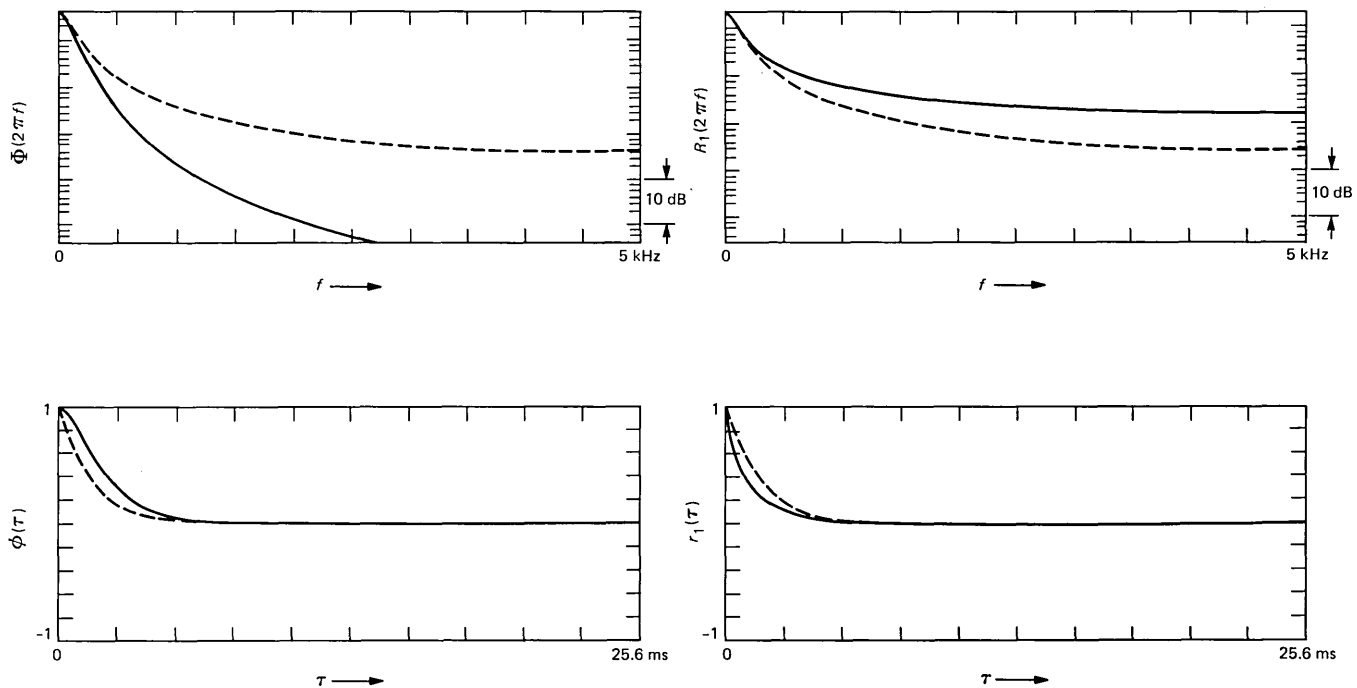
$$R(\tau) = e^{-\alpha\tau} \tag{23a}$$

$$R(\tau) = \sum_{1}^{4} a_i e^{-\alpha_i \tau} \cos(\beta_i \tau), \tag{23b}$$

where $a_i \geq 0$ in the last equation. By proper choice of parameters in eq. (23b), the spectrum can be made to approximate that of any vowel sound up to about 4 kHz. The β_i's are the formant frequencies and the α_i's the half bandwidths.

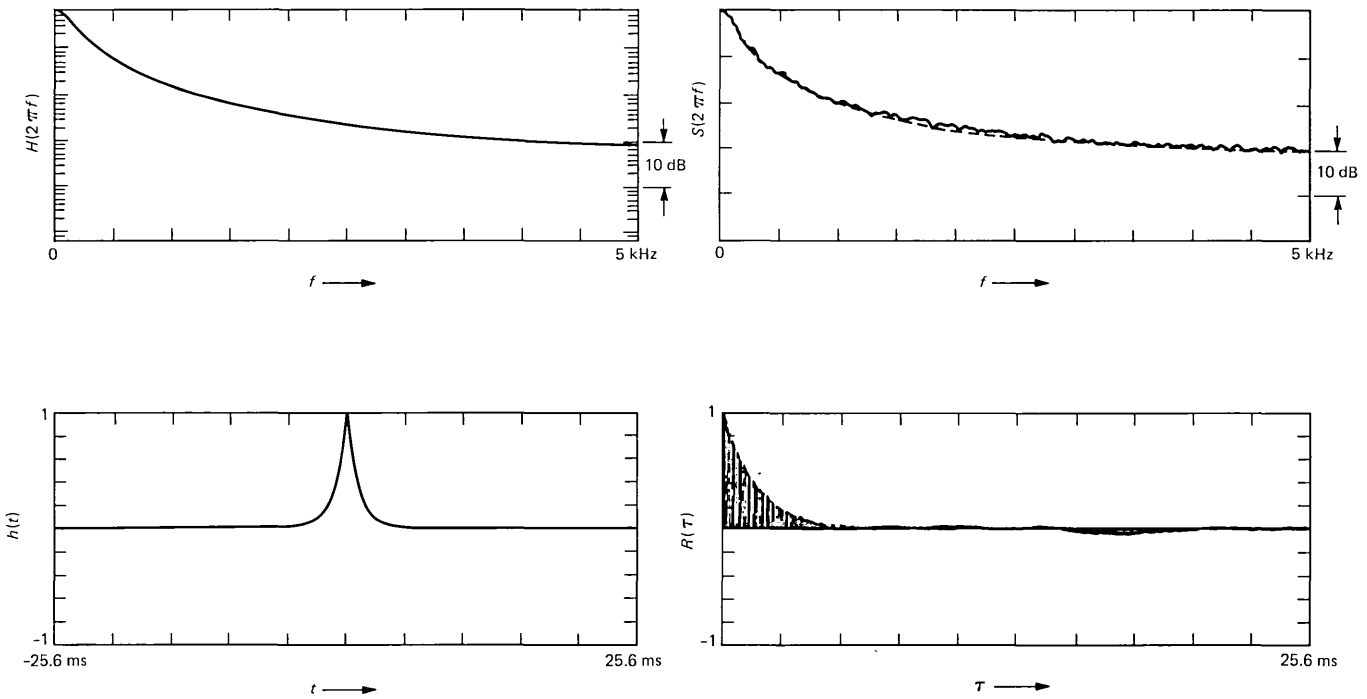
---

* The asymmetry of the range of values for n makes the number of values even. Choosing N to be a power of 2 allows the use of efficient FFT (Fast Fourier Transform) routines.



(a)

Fig. 5a—Generating a binary process with an exponential acf with 100-Hz cutoff frequency. On each graph the dotted line is the desired or prescribed function for the output process. The acf and spectral density required for the $\{x\}$ process are $\phi(\tau)$ and $\Phi(\omega)$, respectively. $r_1(\tau)$ and $R_1(\omega)$ are what the corresponding functions would be for the output process $\{y\}$ if the input $\{x\}$ had the specified acf.



(b)

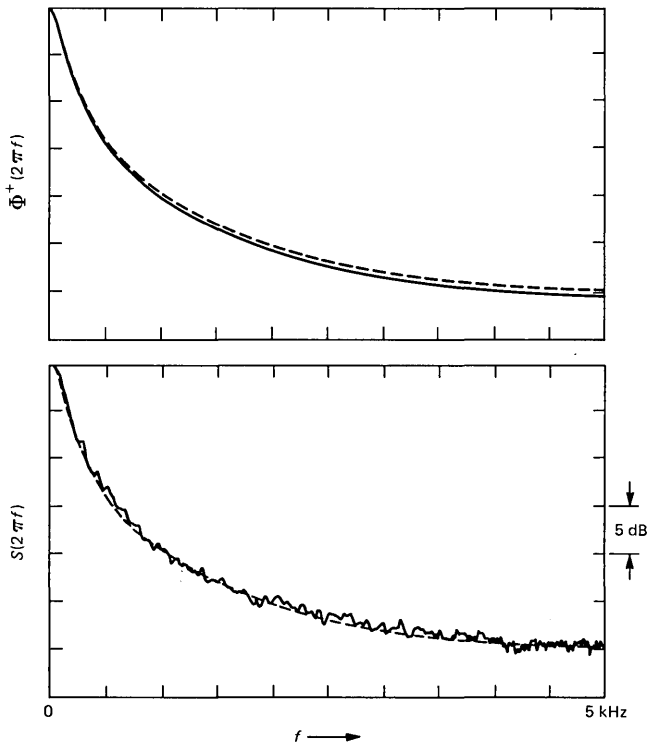
Fig. 5b—On the left are shown the time and frequency domain plots of the shaping filter required to produce the $\{x\}$ process of Fig. 5a. On the right are the acf and spectral density of an actual binary process generated by the method described in the text.

The white noise input to the shaping filter was just a sequence of independent, identically distributed Gaussian random variables. The Gaussian distribution was truncated to ±6 standard deviations.

In Figs. 5 through 9 we show several examples of acf's and spectra (both as predicted theoretically and as measured from the actual outputs) that can be generated by our method.

Figure 5 shows in detail various covariances and spectra associated with the generation of a binary process with an exponential covariance function. Whenever a dotted graph is displayed it is the desired or specified function.

In Fig. 5a the left side shows the acf and spectral density $\phi(\tau)$ and $\Phi(\omega)$ of the $\{x\}$ process, which when passed through the clipping nonlinearity gives the process $\{y\}$ with the specified properties. The right side of Fig. 5a shows the same plots for the process $\{y\}$ that would result if $\{x\}$ were a Gaussian process that had the *specified* acf.



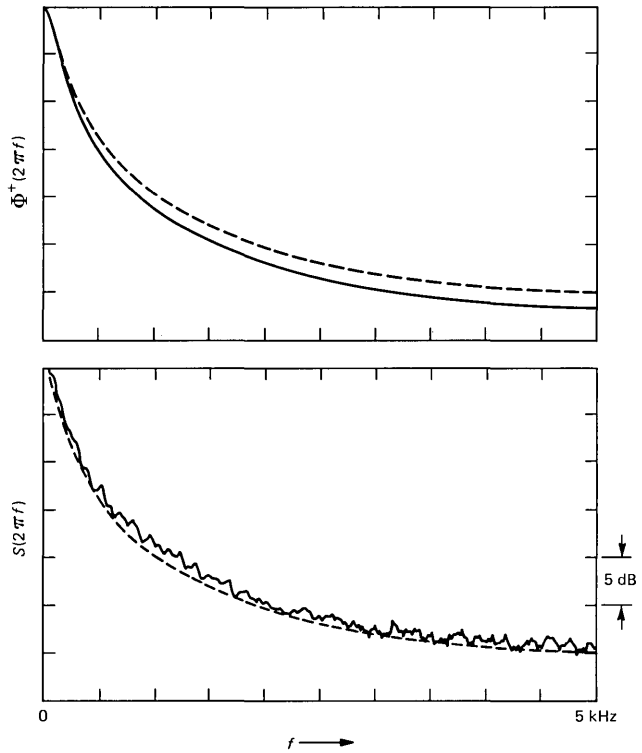
(a)

Fig. 6a—On the top the spectral density $\Phi(\omega)$ required for the $\{x\}$ process and on the bottom the spectral density estimated from a 25,000-sample sequence. The spectral density specified is the same as in Fig. 5.

In Fig. 5b the left side shows the time and frequency domain plots of the shaping filter to produce the required $\{x\}$, and the right side shows the acf and spectral density estimated from a 25,000-sample sequence generated by the method discussed above for the binary case.

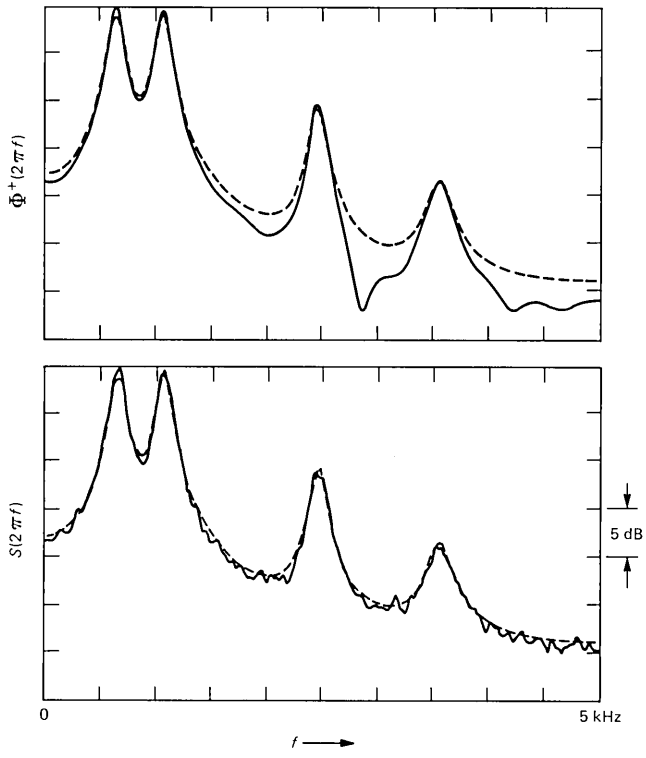
In Figs. 6 through 9 we show spectral density plots only. Further, we show only the plot for $\Phi^+(\omega)$, as defined in eq. (19), and the plot of the output spectral density estimated from 25,000-sample sequences generated by our method. (Of course, Φ^+ is often identical to Φ.) Again, in each case the dotted graphs represent the desired or specified function. The selections shown are an exponential acf for the uniform and gamma distributions in Fig. 6; spectra of vowels /a/ and /e/ with a uniform distribution in Fig. 7; the same vowels with a gamma distribution in Fig. 8; the same vowels with a binary distribution in Fig. 9.

The probability distributions of the generated processes are, of course, not approximated; they are therefore exact except for fluctuations because of finite sample size. Figure 10 shows the actual and



(b)

Fig. 6b—Same as Fig. 6a but for the gamma distribution.



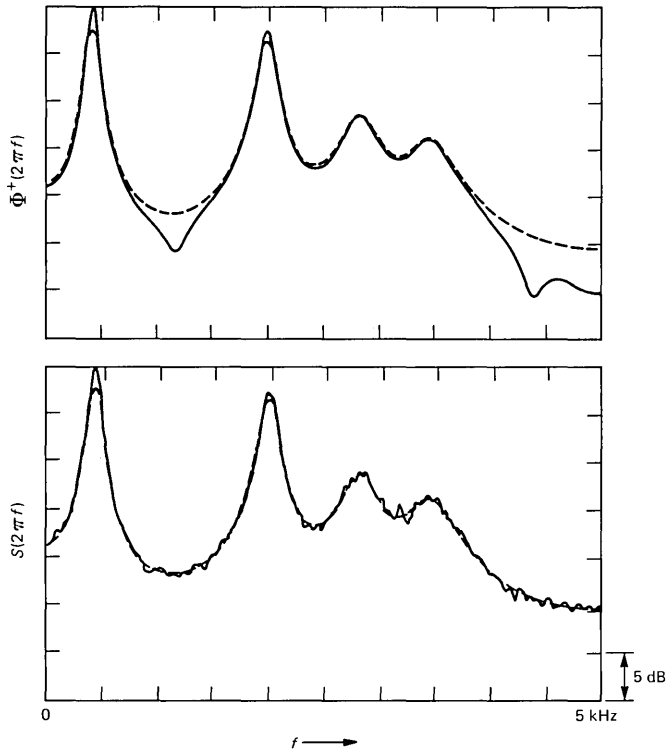
(a)

Fig. 7a—The specified spectral density is a typical spectral shape for the vowel /a/. On the top is $\Phi(\omega)$ for the required $\{x\}$ process. On the bottom is the plot estimated from a 25,000-sample sequence of the generated process. The dotted line in each case is the specified spectral density.

expected distributions for a 25,000-sample sequence with a uniform and a gamma distribution.

Figures 5 through 10 are self-explanatory and demonstrate the capabilities and limitations of the method. We may summarize the main features as follows:

(*i*) When the specified acf is given by eq. (23a), the problem can be solved exactly for any α for each of our examples of pdf. For the uniform and the binary pdf this can be proved analytically, by expanding the sin function in eqs. (16a) and (16c) in powers of R. If we group terms of these expansions in pairs, we can show that the Fourier transform of $\phi(\tau)$ is nonnegative in each of these cases. For the gamma pdf we cannot prove this analytically; however by simulation over a wide range of α's we conclude that $\phi(\tau)$ has a nonnegative Fourier transform in this case too.



(b)

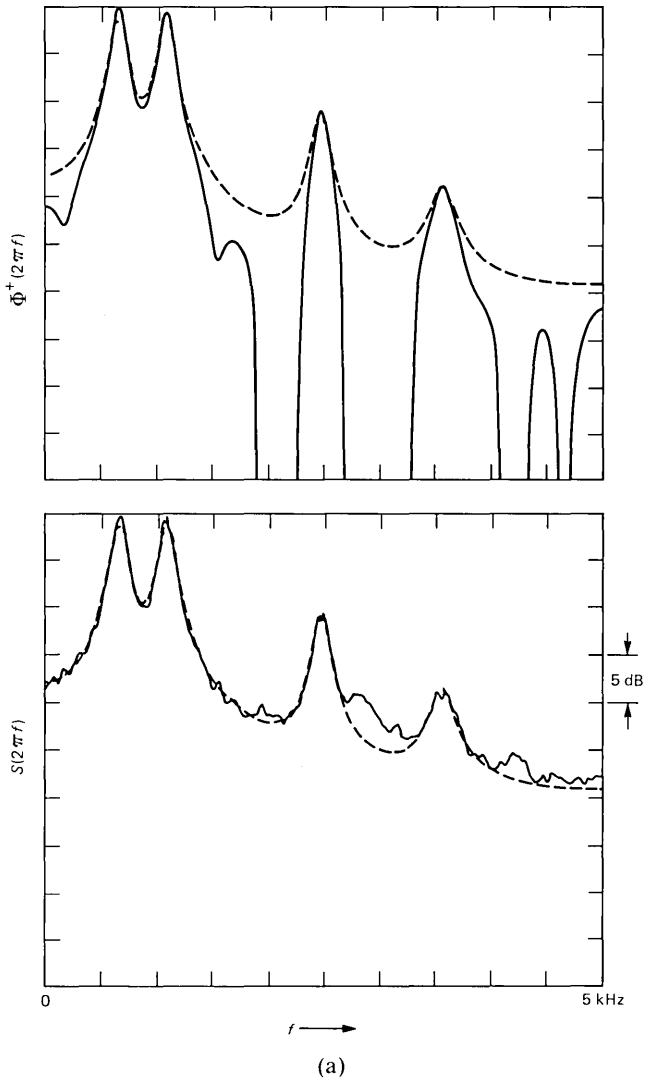
Fig. 7b—Same as Fig. 7a but for the vowel /e/. The plots are for a process with a uniform probability density.

(*ii*) Vowel spectra can be well approximated by random processes with a uniform or gamma distribution. (In the two cases shown in Figs. 7 and 8, the spectrum is realised exactly with the uniform pdf, but only approximately with the gamma pdf; however, the error in the spectrum in the latter case is not much larger than the statistical fluctuations in a 25,000-sample segment. So from a practical point of view the approximation error might not be serious.)

(*iii*) The nonlinearity required for the binary pdf is too severe to allow a well-defined formant structure, especially at high frequencies.

IV. CONCLUDING REMARKS

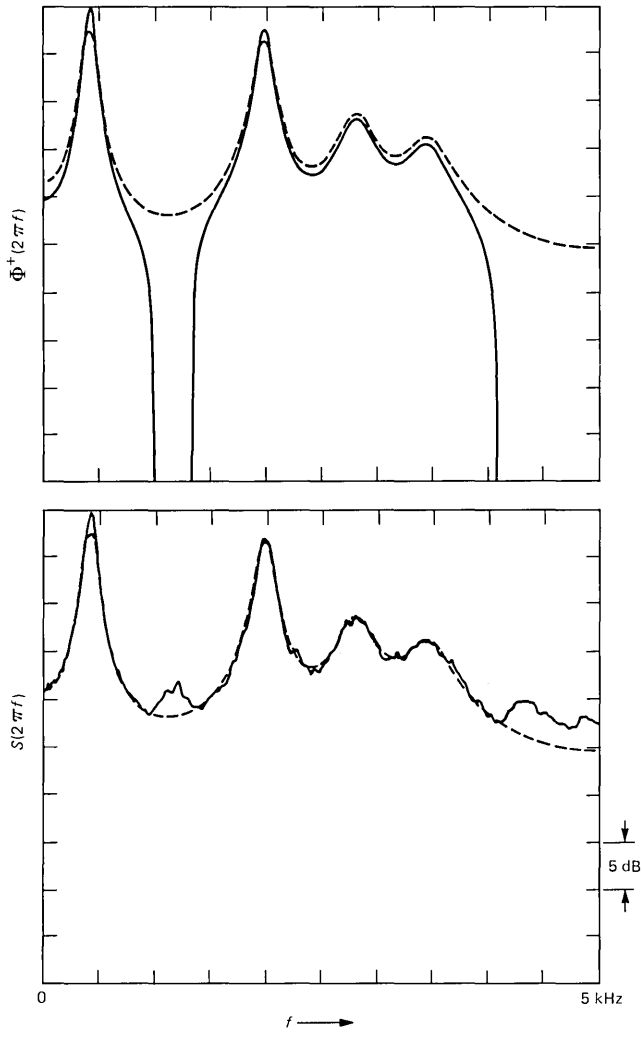
We have described a method for generating a random process with specified spectrum and first-order probability density. The method is successful for many combinations of spectrum and pdf of interest. However, with a given probability density, the method cannot achieve



(a)

Fig. 8a—The specified spectral density is a typical spectral shape for the vowel /a/ for a process with a gamma probability distribution. On the top is $\Phi(\omega)$ for the required $\{x\}$ process. On the bottom is the plot estimated from a 25,000-sample sequence of the generated process. Note that $\Phi(\omega)$ had to be corrected to $\Phi^+(\omega)$, as shown in eq. (15).

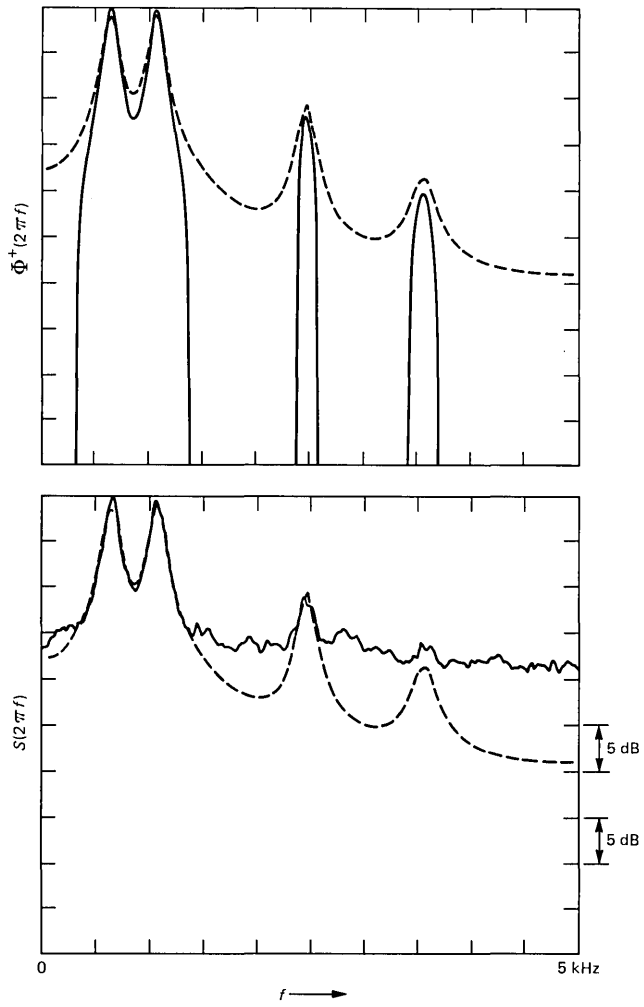
every arbitrarily specified spectral shape. One general way to characterize an achievable spectrum is to say that the corresponding covariance function must be representable in the form given in eq. (13). Another way is to say that the function $\phi(\tau)$ in eq. (15) must have a nonnegative Fourier transform.



(b)

Fig. 8b—Same as Fig. 8a but for /e/.

The class of achievable spectra may be extended by other methods of generating the random process. One interesting method has been suggested by E. N. Gilbert. Rather than nonlinearly distorting a Gaussian process, the suggestion is to *modulate* a Gaussian process by an appropriately chosen nonnegative random process. This method has its own limitations. For instance, it cannot generate a process with a uniform pdf. On the other hand it appears to be very promising for

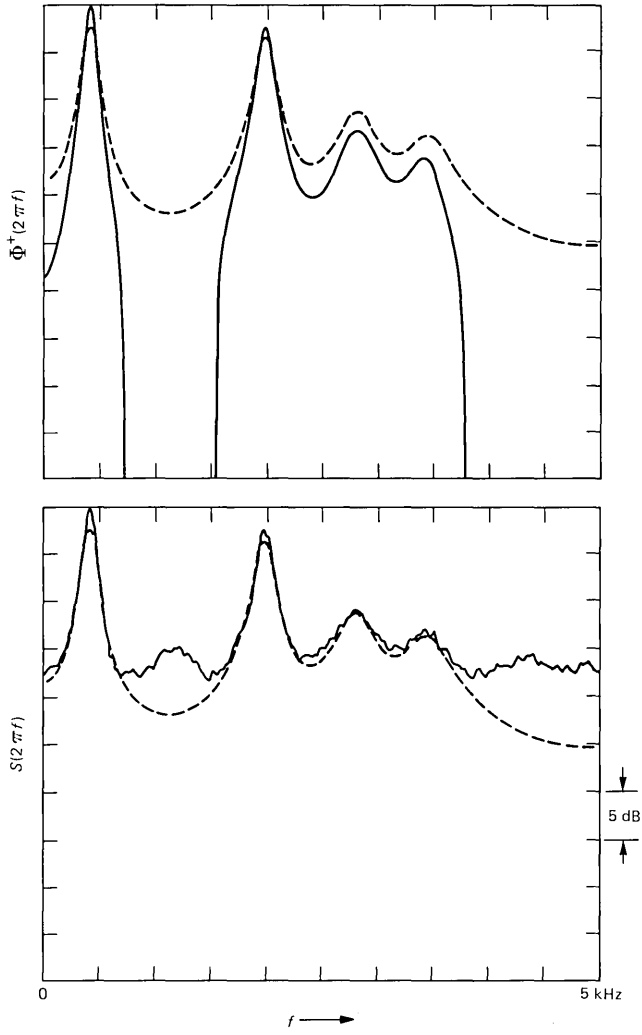


(a)

Fig. 9a—The specified spectral density is a typical spectral shape for the vowel /a/ for a process with a binary distribution. On the top is $\Phi(\omega)$ for the required $\{x\}$ process. On the bottom is the plot estimated from a 25,000-sample sequence of the generated process. Note that $\Phi(\omega)$ had to be corrected to $\Phi^+(\omega)$, as shown in eq. (15).

generating processes with speech-like pdf's (e.g., the gamma pdf discussed above). We are currently investigating this possibility.

To some extent the pdf and covariance function constrain each other regardless of the method used for generating the random process. For example, the covariance function of a time-continuous binary process must have a cusp at the origin. If the specified acf does not



(b)

Fig. 9b—Same as Fig. 9a but for the vowel /e/.

have this property, then the specifications are inconsistent and the problem has no solution. Unfortunately, to the best of our knowledge, no tractable procedure is known to decide whether a covariance function is consistent with a given pdf. L. A. Shepp has drawn our attention to some of his unpublished work in which he investigated the class of covariance functions of processes with a given pdf. He showed that this class is convex and that any such acf is a convex

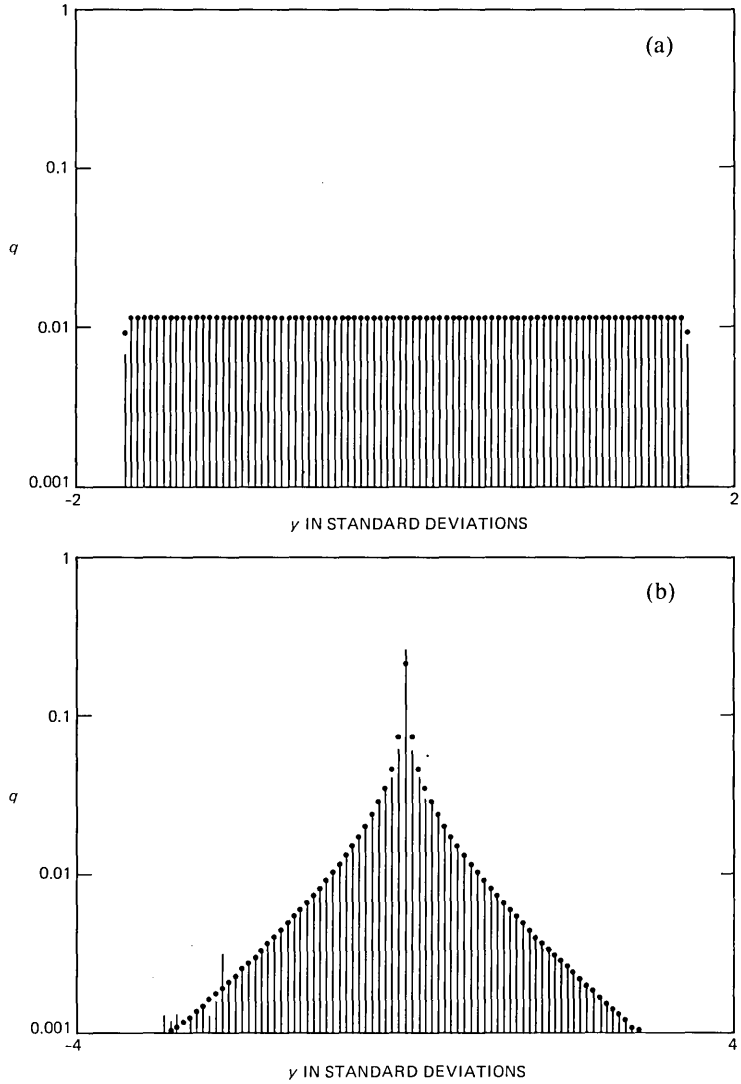


Fig. 10—A check on the probability distributions of the output processes. (a) The expected (·) and measured (bars) pdf estimated from a 25,000 sample when the specified distribution was uniform. The fit is approximately as good as this for every uniform-pdf process generated. (b) Same as for Fig. 10a but for the gamma distribution.

combination of the extremal acf's of the class. Unfortunately, no general method is known to determine the extremal covariances. Another relevant work is a paper[9] by Brockway McMillan, in which he considers covariances of binary processes, and gives a test for such

a covariance in terms of the class of "corner-positive" matrices. This test too is very difficult to apply in practice.

In some situations (for example, the experiment on texture perception mentioned above) it might not be important to specify the spectrum and the pdf precisely. Instead, it might suffice to be able to generate random processes with (*i*) different spectra and exactly the same pdf, or (*ii*) different pdf's and exactly the same spectrum. The first of these objectives is accomplished by keeping the nonlinearity fixed and varying the spectrum of the shaping filter. Inspection of eqs. (12) and (13) shows how the second objective may be accomplished.[7] Since the acf in eq. (13) depends only upon the squares of the expansion coefficients, it is evident that two nonlinearities for which one or more coefficients differ in sign (but not in magnitude) will yield identical spectra. The pdf will in general be quite different in the two cases.

REFERENCES

1. J. H. Laning and R. H. Battin, *Random Processes in Automatic Control*, New York: Mc-Graw Hill Book Company, 1956, p. 217.
2. E. A. Guillemin, *Synthesis of Passive Networks*, New York: John Wiley and Sons, Inc., 1957, Chapter 3.
3. W. B. Davenport, "An Experimental Study of Speech-Wave Probability Distributions," J. Acous. Soc. Am., *24* (July 1952), pp. 390–9.
4. D. L. Richards, "Statistical Properties of Speech Signals," Proc. IEE, *111*, May 1964.
5. Bela Julesz, "Textons, the Elements of Texture Perception, and Their Interactions," Nature, *290*, No. 5802 (12 March 1981), pp. 1–7.
6. Y. Linde and R. M. Gray, "A Fake Process Approach to Data Compression," IEEE Trans. Commun., *COM-26* (June 1975), pp. 840–7.
7. G. L. Wise, A. P. Traganitis, and J. B. Thomas, "The Effect of a Memoryless Nonlinearity on the Spectrum of a Random Process," IEEE Trans. Information Theory, *IT-23* (January 1977), pp. 84–8.
8. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C., Chapter 22.
9. Brockway McMillan, "History of a Problem," Jour. SIAM, *3* (September 1955), pp. 119–28.

APPENDIX

Derivation of R(τ)

In this appendix we will derive the three equations (14a), (14b), and (14c) of the text. All three derivations depend on a well-known property of the bivariate Gaussian pdf $g(u, v, \rho)$, namely,

$$\frac{\partial g}{\partial \rho} = \frac{\partial^2 g}{\partial u \partial v}. \tag{24}$$

This follows trivially from the representation of g as the Fourier transform of its characteristic function, i.e.,

$$g(u, v, \rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{j(\omega u + \sigma v)} e^{-\frac{1}{2}(\omega^2 + \sigma^2 + 2\rho\omega\sigma)} d\omega d\sigma. \tag{25}$$

Differentiating this expression with respect to ρ and then with respect to u and v verifies eq. (24).

[In passing we note that Mehler's expansion eq. (11) also follows trivially from eq. (25). For if

$$g(u, v, \rho) = \sum_0^\infty a_n(u, v) \frac{\rho^n}{n!} \tag{26}$$

it follows that

$$\begin{aligned} a_n(u, v) &= \frac{\partial^n g}{\partial \rho^n}\bigg|_{\rho=0} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (-\omega\sigma)^n e^{j(\omega u+\sigma v)} e^{-\frac{1}{2}(\omega^2+\sigma^2)} d\omega d\sigma \\ &= \frac{\partial^n p(u)}{\partial u^n}\frac{\partial^n p(v)}{\partial v^n}. \end{aligned} \tag{27}$$

Here $p(.)$ is the univariate Gaussian pdf, eq. (3). However, the Hermite polynomials are *defined* by the relation

$$\frac{d^n p(x)}{dx^n} = (-1)^n p(x) \mathrm{He}_n(x), \tag{28}$$

which thus gives Mehler's expansion.]

Let us differentiate eq. (10) of the text with respect to ρ and use eq. (24) to evaluate the right-hand side. Thus,

$$\frac{dR}{d\rho} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u)F(v) \frac{\partial^2 g}{\partial u \partial v} du dv. \tag{29}$$

Integrating eq. (29) by parts and assuming that $g(u, v, \rho)F(u)F(v)$ vanishes when $u, v \to \pm\infty$, we get

$$\frac{dR}{d\rho} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F'(u)F'(v)g(u, v, \rho) du dv. \tag{30}$$

For the binary case (8c), $F'(u) = 2\delta(u)$. In this case

$$\frac{dR}{d\rho} = 4g(0, 0, \rho) = \frac{2}{\pi\sqrt{1-\rho^2}}. \tag{31}$$

Since $R = 0$ when $\rho = 0$ (because F is antisymmetric) this immediately gives

$$R = \frac{2}{\pi} \sin^{-1} \rho, \tag{32}$$

which is eq. (14c).

For the nonlinearity in the case of the gamma pdf,

$$F'(u) = \frac{2}{\sqrt{3}} |u| \tag{33a}$$

$$F''(u) = \frac{2}{\sqrt{3}} \operatorname{sign}(u). \tag{33b}$$

Using eq. (24) twice we get

$$\frac{dR}{d\rho} = \frac{4}{3} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |u||v| g(u, v, \rho) du dv \tag{34a}$$

and

$$\frac{d^2R}{d\rho^2} = \frac{4}{3} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \operatorname{sign}(u)\operatorname{sign}(v) g(u, v, \rho) du dv.$$

$$= \frac{8}{3\pi} \sin^{-1}\rho, \tag{34b}$$

where the last step follows from the result just derived for the binary case. Now at $\rho = 0$, $R = 0$ and, from eq. (34a) $dR/d\rho = 8/3\pi$. With these initial values eq. (34b) can be integrated easily to give eq. (14b) of the text.

Finally, for the nonlinearity (8a) $F'(u) = 2\sqrt{3}p(u)$ and, therefore,

$$\frac{dR}{d\rho} = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(u)p(v)g(u, v, \rho) du dv$$

$$= \frac{3}{\pi^2} \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{(2-\rho^2)u^2 + (2-\rho^2)v^2 - 2\rho uv}{2(1-\rho^2)}} du dv. \tag{35}$$

The integrand in eq. (35) is of the same form as $g(u, v, \rho)$ with a somewhat different quadratic form in the exponent. It is a standard integral whose value is $2\pi/\sqrt{\Delta}$, where Δ is the determinant of the quadratic form. Simple algebraic manipulation then gives

$$\frac{dR}{d\rho} = \frac{6}{\pi} \frac{1}{\sqrt{4-\rho^2}}. \tag{36}$$

Again since $R = 0$ when $\rho = 0$, this gives

$$R = \frac{6}{\pi} \sin^{-1} \frac{\rho}{2}, \tag{37}$$

which is eq. (14a).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

A Method to Characterize the Mechanical Properties of Undersea Cables

By T. C. CHU

(Manuscript received January 22, 1982)

A method has been developed to evaluate the mechanical properties of cables containing stranded-strength members in both linear elastic and nonlinear plastic regions. The method extends Cannon and Santana's general system of equations describing the cable mechanical characteristics. In the formulation of the method, the fundamental assumptions made by Cannon and Santana are first examined and justified. Next, instead of using elastic constants for the constituent cable materials in the system of equations, the regression analysis is applied to the tensile and torsional test data of dominating high-strength cable components to obtain least-squares-fit polynominals approximating the stress versus strain and shearing-stress versus shearing-strain curves. By differentiating the polynominals, the tensile and torsional moduli of these cable components as functions of their axial strain and twist are derived. The relations describing the mechanical properties of the cable in both elastic and plastic regions are obtained by substituting the tensile and torsional moduli of the high-strength cable components and the constant moduli for the low-strength cable components into the system of equations in differential form and integrating them. Application of the method to the present experimental undersea-lightguide cable yields excellent agreement with the tensile test results of the cable.

I. INTRODUCTION

Because of its long suspended length in deep ocean, undersea cable normally experiences high tension and strain during its installation and recovery.[1] High-strength stranded members are employed in the undersea cable design to support the tension and keep the cable strain below a prescribed level. The problems of high tension and strain cause even greater concerns for the undersea lightguide cable because of the

static and dynamic fatigue of optical fibers.[2] Adequate protection of the fibers is essential to system reliability and requires accurate evaluation of the mechanical properties of the cable in development.

Mechanical characterization of cable containing helically wrapped or stranded members has been done by Cannon and Santana[3] using linear elasticity theory, which assumes that the tensile modulas E and Poisson's ratio ν are constant and independent of the cable load. Using two fundamental assumptions relating the external applied force and moment at the cable ends to the tension developed in individual helically stranded-strength members, a system of equations was developed that describes the mechanical characteristics of a cable having one or more layers of helically stranded members. Application of this theory to the present undersea lightguide cable yields excellent agreement with the experimental data up to a strain level of approximately 0.5 percent. Beyond this strain level, the theory predicts lower strain than the experimental data. This is not surprising because, at a strain level higher than 0.5 percent, the cable has been stretched beyond its elastic region to the plastic region, where the linear elasticity theory does not apply. Since undersea cable normally experiences high tension and strain, an understanding of the mechanical properties in both regions is essential to the design of undersea lightguide cable.

In this paper, a method is presented to accurately predict the mechanical properties of the cable in both elastic and plastic regions. Since the present method uses the system of equations developed in Ref. 3, the fundamental assumptions on which the system of governing equations is based will be examined and justified in Section II. The formulation of the new method is presented in Section III. Then the method is applied to the experimental undersea lightguide cable in Section IV. The results of the theory are compared with experimental data in Section V.

II. FUNDAMENTAL ASSUMPTIONS

The mechanical properties of cables containing helically stranded members have been analyzed in Ref. 3 using a model consisting of n identical wires of radius b parallel to one another in a circular array and twisted into helices with a common helix angle α and radius r, as shown in Fig. 1. The fundamental equations relating the applied force T_t and moment M_t along the cable axis to the tension T developed in individual wire were assumed to be

$$T_t = nT \sin \alpha \tag{1}$$

and

$$M_t = nTr \cos \alpha. \tag{2}$$

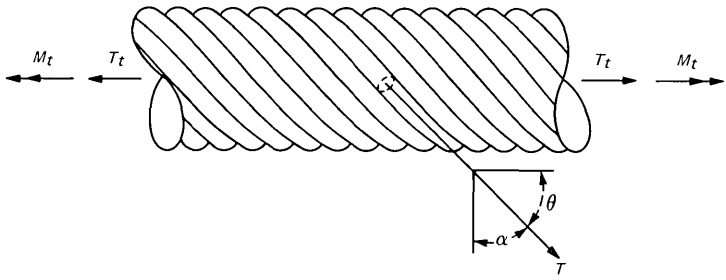


Fig. 1—Cable loads and wire tension.

From these assumptions, a system of equations was developed, which describe the mechanical properties of a cable containing helically wrapped elements.

However, from the equations of equilibrium of a bent and twisted thin rod, Costello and Phillips[4] have obtained a complete solution for the same problem illustrated in Fig. 1. The solution relates the applied tension and torque to the individual wire tension as

$$T_t = nT \sin \alpha + n \frac{\cos^3 \alpha}{r} \left| C\tau - A \frac{\sin \alpha \cos \alpha}{r} \right| \tag{3}$$

and

$$M_t = n \left| T\tau \cos \alpha - \cos^2 \alpha \sin \alpha \left| C\tau - A \frac{\sin \alpha \cos \alpha}{r} \right| + C\tau \sin \alpha + A \frac{\cos^3 \alpha}{r} \right|, \tag{4}$$

where A and C are constants, depending on the elastic properties of the material and the shape and dimensions of the rod cross-sectional area. E is the tensile modulus, ν is Poisson's ratio, and, for a circular cross section of radius b,

$$A = \frac{\pi E b^4}{4} \quad \text{and} \quad C = \frac{\pi E b^4}{4(1 + \nu)}.$$

Notice that eqs. (3) and (4) can be reduced to (1) and (2) if the first term on the right of the equations is retained while neglecting the other terms. To neglect the second and third terms on the right-hand side of eq. (3) requires that

$$\frac{b^2 \tau \cos^3 \alpha}{4(1 + \nu) r \epsilon_s \sin \alpha} \ll 1 \tag{5}$$

and

$$\left|\frac{b^2\cos^2\alpha}{2r\sqrt{\epsilon_s}}\right|^2 \ll 1, \tag{6}$$

where ϵ_s is the strain of helical wire along its own axis. Similarly, to neglect the four terms on the right-hand side of eq. (4) requires that

$$\frac{\tau b^2\cos\alpha\sin\alpha}{4(1+\nu)rE_s} \ll 1, \tag{7}$$

$$\left|\frac{b\cos\alpha\sin\alpha}{2r\sqrt{\epsilon_s}}\right|^2 \ll 1, \tag{8}$$

$$\frac{\tau b^2\sin\alpha}{4(1+\nu)r\epsilon_s\cos\alpha} \ll 1, \tag{9}$$

and

$$\left|\frac{b\cos\alpha}{2r\sqrt{\epsilon_s}}\right|^2 \ll 1. \tag{10}$$

Inequalities (5), (6), (8), and (10) can easily be satisfied if $\alpha \simeq \pi/2$, i.e., for the large helix angle or small lay angle. However, to satisfy inequalities (7) and (9), the additional condition $\tau \simeq 0$ or $b^2/r \simeq 0$ is required.

In undersea cable design and manufacture, the helix angle α of the stranded members is usually close to $\pi/2$, and the members are actually made twist-free as they are formed into helices by a planetary-stranding machine. The condition $\tau \simeq 0$ can usually be satisfied. Therefore, the approximate relations (1) and (2) assumed in Ref. 3 and the resulting system of equations are justified.

III. FORMULATION OF THE NEW METHOD

For a cable containing multiple layers of helically stranded strength members plus the cable core, the total axial load carried by the cable is the sum of the core load and the strand load, according to the following equation[3]

$$T_t = T_c + \sum_{i=1}^{m} n_i T_{si}\sin\alpha_i, \tag{11}$$

and the total cable moment equals the sum of the moment contributions from each of the strand layers plus that of the core, as shown by

$$M_t = -\psi J_c G_c + \sum_{i=1}^{m} |n_i T_{si} r_i\cos\alpha_i - \psi(n_i J_{si} G_{si}\sin^2\alpha_i)|, \tag{12}$$

where n_i is the number of stranded wires in the ith layer; T_c is the load experienced by the cable core; T_{si} is the tension supported by the ith strand layer; m is the number of strand layers in the cable; α_i is the common helix angle of the wires in the ith layer; ψ is the untwist turns experienced per unit length of the cable under applied tension; J_c and G_c are the cable core's moment of inertia and torsional rigidity; r_i is the radial location of the ith layer; J_{si} and G_{si} are the ith strand's moment of inertia and moment of rigidity. Note that eqs. (11) and (12) use the approximate relations (1) and (2) for the stranded wires.

Equations (11) and (12) can be transformed to the abbreviated forms

$$T_t = C_3\epsilon_c - C_4\psi \tag{13}$$

and

$$M_t = C_1\epsilon_c - C_2\psi, \tag{14}$$

where the coefficients C_i are defined by

$$\mathrm{C}_1 = \sum_{i=1}^{m} |\sin^2\alpha_i - \nu_i^*\cos^2\alpha_i\| n_i E_{si}\mathrm{A}_{si} r_i \cos\,\alpha_i|, \tag{15}$$

$$\mathrm{C}_2 = J_c G_c + \sum_{i=1}^{m} |n_i J_{si} G_{si}\sin^2\alpha_i + \pi r_i^2 \sin\,2\alpha_i| n_i E_{si}\mathrm{A}_{si}\cos\,\alpha_i\|, \tag{16}$$

$$\mathrm{C}_3 = E_c A_c + \sum_{i=1}^{m} |\sin^2\alpha_i - \nu_i^*\cos^2\alpha_i\| n_i E_{si}\mathrm{A}_{si}\sin\,\alpha_i|, \tag{17}$$

and

$$\mathrm{C}_4 = \sum_{i=1}^{m} |\pi r_i \sin\,2\alpha_i| n_i E_{si}\mathrm{A}_{si}\sin\,\alpha_i. \tag{18}$$

Here, ν_i^* is a pseudo-Poisson's ratio for the ith layer and is a measure of its diametric contraction; E_{si} and A_{si} are the tensile modulus and cross-sectional area of the stranded member in the ith layer. For a cable with its ends twist-restrained, $\psi = 0$, and eqs. (13) and (14) become

$$\epsilon_c|_{\psi=0} = \frac{T_t}{\mathrm{C}_3} \tag{19}$$

and

$$M_t|_{\psi=0} = \mathrm{C}_1\epsilon_c. \tag{20}$$

For a cable with its ends free to rotate, $M_t = 0$, and eqs. (13) and (14) become

$$\epsilon_c|_{M_t=0} = \frac{T_t}{\mathrm{C}_3 - \dfrac{\mathrm{C}_1\mathrm{C}_4}{\mathrm{C}_2}} \tag{21}$$

and

$$\psi|_{M_t=0} = \frac{C_1}{C_2}\epsilon_c. \tag{22}$$

If the cable characterization parameters C_1, C_2, C_3, and C_4 are known, the cable strain ϵ_c, cable moment M_t, and cable twist can be evaluated from eqs. (19) through (22) for a given cable tension, T_t. If the elastic properties of cable components are all constant and the twist ψ is zero or very small, the cable characterization parameters C_1, C_2, C_3, and C_4 are all constant and the relations among the cable strain, moment, twist, and tension from eqs. (19) to (22) are all linear. This is the case illustrated in Ref. 3.

In reality, however, the tensile moduli E's and moduli of rigidity G's of the cable components, either stranded or unstranded, are nonlinear functions of the cable strain ϵ_c and twist ψ, i.e., $E = E(\epsilon_c, \psi)$, $G = G(\epsilon_c, \psi)$, and thus the cable characterization parameters C_1, C_2, C_3, and C_4 are also functions of the cable strain ϵ_c and twist ψ.

Thus, eqs. (13) and (14) can only be applied in differential forms

$$dT_t = C_3 d\epsilon_c - C_4 d\psi \tag{23}$$

and

$$dM_t = C_1 d\epsilon_c - C_2 d\psi. \tag{24}$$

The above differential equations are meaningful only if the coefficients C_1, C_2, C_3, and C_4 can be expressed in terms of the cable strain ϵ_c and twist ψ. This can be accomplished in two steps. First, apply the regression analysis to the tensile and torsional test data of dominating cable components to obtain least-squares-fit polynominals, which approximate the stress versus strain and shearing-stress versus shearing-strain curves. By differentiating the polynominals, we obtain the tensile moduli and the moduli of rigidity of the cable components in terms of axial strain and twist:

$$\begin{aligned} E_{si} &= E_{si}(\epsilon_{si}) \\ E_c &= E_c(\epsilon_c) \\ G_{si} &= G_{si}(\psi_{si}) \\ G_c &= G_c(\psi), \end{aligned} \tag{25}$$

where ϵ_{si} and ψ_{si} are the strain and twist of the stranded member in the direction of its own axis. Second, transform the independent variables ϵ_{si} and ψ_{si} in (25) into the cable strain ϵ_c and twist ψ using the following relations:[3]

$$\epsilon_{si} = |\sin^2\alpha_i - \nu_i^* \cos^2\alpha_i|\epsilon_c - |l_i \cos^2\alpha_i|\psi \tag{26}$$

$$\psi_{si} = \psi \sin \alpha_i, \tag{27}$$

where l_i is the lay length of the ith stranded layer. Substitution of the above equations into (23) and (24) yields

$$dT_t = C_3|\epsilon_c, \psi|d\epsilon_c - C_4|\epsilon_c, \psi|d\psi \tag{28}$$

and

$$dM_t = C_1|\epsilon_c, \psi|d\epsilon_c - C_2|\epsilon_c, \psi|d\psi. \tag{29}$$

For a cable with its ends twist-restrained, $\psi = 0$, and eqs. (28) and (29) become

$$T_t = \int C_3(\epsilon_c)d\epsilon_c \tag{30}$$

and

$$M_t = \int C_1(\epsilon_c)d\epsilon_c. \tag{31}$$

Since $C_1(\epsilon_c)$ and $C_2(\epsilon_c)$ are polynominals, the above equations can readily be integrated. For a cable with its ends free to rotate, $M_t = 0$, and eqs. (28) and (29) become

$$T_t = \int \left| C_3(\epsilon_c, \psi) - \frac{C_1(\epsilon_c, \psi)C_4(\epsilon_c, \psi)}{C_2(\epsilon_c, \psi)} \right| d\epsilon_c \tag{32}$$

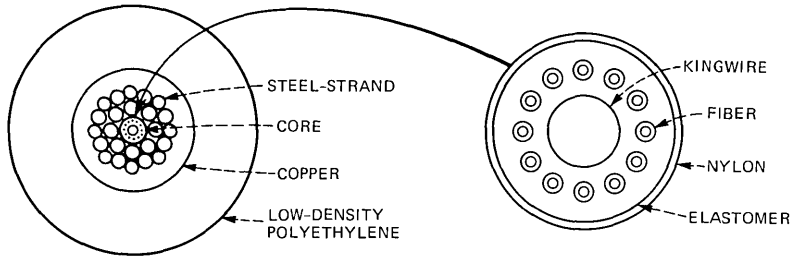
and

$$\frac{d\psi}{d\epsilon_c} = \frac{C_1(\epsilon_c, \psi)}{C_2(\epsilon_c, \psi)}. \tag{33}$$

In this case, the differential eq. (33) should be solved first to obtain a relation between ψ and ϵ_c. This relation is applied to eq. (32) to eliminate ψ. Then, eq. (32) can be integrated to give a relation between T_t and ϵ_c.

IV. APPLICATION

The recovery operation induces higher tensions in undersea cable than any other handling operation. During a steady-state recovery, the cable tension is highest at the overboard sheave, where the cable twist is practically zero. Therefore, only the case of twist restraint, $\psi = 0$, will be studied here. The method is applied to an experimental undersea lightguide cable to characterize its tensile and torsional properties. The cable structure and component dimensions are shown in Fig. 2. The cable consists of a lightguide core protected by two layers of high-strength steel-strand wires, a copper conductor, and a low-density polyethylene (LDPE) jacket for high-voltage insulation and environmental protection. The lightguide core consists of a copperplated-steel



CABLE STRUCTURE:
STRAND DIAMETER = 7.9 mm
CONDUCTOR OD (COPPER) = 10.5 mm
INSULATION OD = 21.0 mm

CABLE CORE:
CENTER WIRE OD = 0.8 mm
NUMBER OF FIBERS = 12
FIBER OD (COATED) = 250 m
SHEATH THICKNESS = 0.1 mm
CORE OD = 2.6 mm

Fig. 2—Undersea cable design.

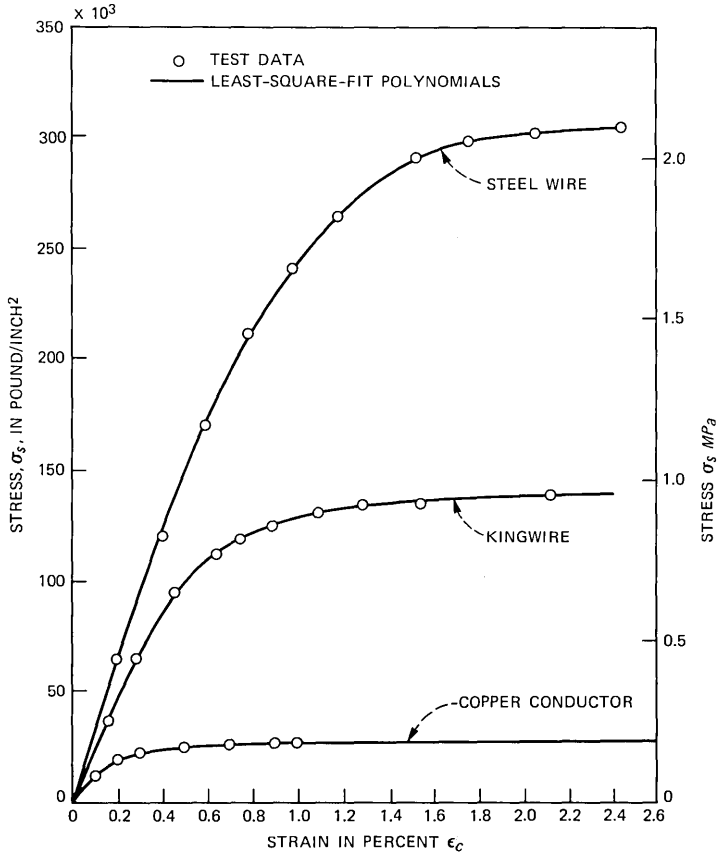


Fig. 3—Tensile testing results and the least-squares-fit polynomials of the high-strength steel-wire, copper sheath, and kingwire.

conductor called the kingwire, up to 12 helically wound fibers embedded in an elastomer material, and a thin outer nylon sheath.

From the cable structure, the metallic components, i.e., the steel wires, the copper conductor, and the kingwire, will carry most of the tensile load. Several tensile tests were conducted on individual steel wires of different sizes and on the kingwire. The test results of a representative 1.6-mm-diameter steel wire and a 0.81-mm-diameter kingwire are shown in Fig. 3. A thorough experimental study on the tensile behavior of the copper conductor was previously done in an unpublished work by R. C. Mondello of Bell Laboratories. His result on newly made copper conductors is also shown in Fig. 3. The nonlinear behavior of each component is evident at high load. Application of the regression analysis to each set of test data in Fig. 3 yields a least-squares-fit polynominal of sixth degree, approximating the stress versus strain relation for each material. The polynomials with the stresses σ_s in MP_a and ϵ_c in mm/mm are given below.

4.1 Steel wire

The polynomial for the steel wires is

$$\sigma_s = 0.663587607686 + 214542.346398\ \epsilon_s + 3198478.98934\ \epsilon_s^2 - 1617068166.73\ \epsilon_s^3 + 111346531878\ \epsilon_s^4 - 3.45797334075 \times 10^{12}\ \epsilon_s^5 + 4.21128175626 \times 10^{13}\ \epsilon_s^6\ (MP_a), \quad (34)$$

where the standard error for estimate "CHI" is CHI = 5.523 (MP_a).

4.2 Kingwire

The polynomial for the kingwire is

$$\sigma_s = 1.15682945607 + 162675.763595\ \epsilon_s + 7250437.88515\ \epsilon_s^2 - 4367711589.93\ \epsilon_s^3 + 459662005052\ \epsilon_s^4 - 2.03575240495 \times 10^{13}\ \epsilon_s^5 + 3.3234268429 \times 10^{14}\ \epsilon_s^6\ (MP_a), \quad (35)$$

where the standard error for estimate "CHI" is CHI = 5.871 (MP_a).

4.3 Copper conductor

The polynomial for the copper conductor is

$$\sigma_s = 0.299934161316 + 110226.149362\ \epsilon_s - 28633487.4233\ \epsilon_s^2 + 3792414452.98\ \epsilon_s^3 - 264983411326\ \epsilon_s^4 + 9.25874279974 \times 10^{12}\ \epsilon_s^5 - 1.27105877316 \times 10^{14}\ \epsilon_s^6\ (MP_a), \quad (36)$$

where the standard error for estimate "CHI" is CHI = 1.780 (MP_a).

Table I—Mechanical properties of cable components

Cable Components	Tensile Modulus, $E(GP_a)$	Shear Modulus, $C(GP_a)$	No. of Strand Members, N	Helix Radius, r (mm)	Helix Angle, α (degree)	Cross-Sectional Area (mm^2)	Polar Moment of Inertia (mm^4)
Kingwire	158.58	61.02	—	—	—	16.58×10^{-2}	2.14×10^{-2}
Elastomer	0.05	0.03	—	—	—	3.87	2.19
Glass fiber	71.71	31.51	12	0.79	88.8	0.95×10^{-2}	7.2×10^{-6}
Nylon	1.24	0.44	—	—	—	26.32×10^{-2}	0.68
1st layer steel wires (1.6-mm dia.)	206.84	79.29	8	2.08	86.8	2	0.32
2nd layer steel wires (1.48-mm dia.)	206.84	79.29	8	3.27	83.9	1.74	0.24
2nd layer steel wires (1.135-mm dia.)	206.84	79.29	8	3.51	83.9	1.03	0.16
Copper sheath	117.21	44.13	—	—	—	12.64	116.54
LDPE	0.24	0.08	—	—	—	2.82×10^2	9.24×10^3

The tensile moduli $E(\epsilon_s)$'s of each component can be obtained by taking the derivative of the corresponding polynomial

$$E(\epsilon_s) = \frac{d\sigma_s}{d\epsilon_s}. \tag{37}$$

For the stranded-steel wires and fibers, the independent variables ϵ_s can be transformed to ϵ_c using eq. (28); for the copper conductor and kingwire, ϵ_s can be directly replaced by ϵ_c since their axial direction coincides with that of the cable. Substitutions of the tensile moduli of the components into eq. (15) and (17) give

$$\begin{aligned} C_1(\epsilon_c) &= |\sin^2\alpha_f - \nu_f^* \cos^2\alpha_f| n_f E_f A_f r_f \cos\alpha_f \\ &\quad + \sum_{i=1}^{4} |\sin^2\alpha_i - \nu_i^* \sin^2\alpha_i| n_i E_{si}(\epsilon_c) A_{si} r_{si} \cos\alpha_i| \\ C_3(\epsilon_c) &= E_k(\epsilon_c) A_k + E_e A_e + E_n A_n + E_c(\epsilon_c) A_c + E_p A_p \\ &\quad + |\sin^2\alpha_f - \nu_f^* \cos^2\alpha_f| n_f E_f A_f \sin\alpha_f|, \\ &\quad + \sum_{i=1}^{4} |\sin^2\alpha_i - \nu_i^* \cos^2\alpha_i| n_i E_{si}(\epsilon_c) A_{si} \sin\alpha_i|, \end{aligned}$$

where the subscripts k, e, n, c, p, and f refer to the kingwire, elastomer, nylon, copper, LDPE, and fibers, respectively. The tensile moduli of the fibers, elastomer, nylon, and LDPE are all assumed constant and are listed in Table I. Substitution of the above expressions for $C_1(\epsilon_c)$ and $C_3(\epsilon_c)$ into eqs. (29) and (30) and integrating give the tensile and torsional properties of the cable under twist-restraint condition. These functions are calculated and the results are shown as solid lines in Figs. 4 and 5. The dashed lines in these figures are the results using constant-elastic moduli listed in Table I.

V. EXPERIMENTAL RESULTS

Tensile tests were conducted for five cable specimens, 100.5-meter-length each. The cable ends of the cable specimen were potted into tapered-plug epoxy terminations. The cable specimen with the terminations was installed in the tensile test bed in a straight line configuration with one end fixed to the anchor and another end attached to the hydraulic ram. Both ends were twist-restrained. The cable tension was increased in steps from 8.9 to 53.4 kN, then in steps from 4.5 to the maximum load of 62.3 kN. The cable tension and moment were measured by a load cell. The cable elongation was measured by using a 10-turn, 10-kΩ linear potentiometer. All data were monitored and displayed in digital form on a control panel and then recorded.

The experimental results are also shown in Figs. 4 and 5. The

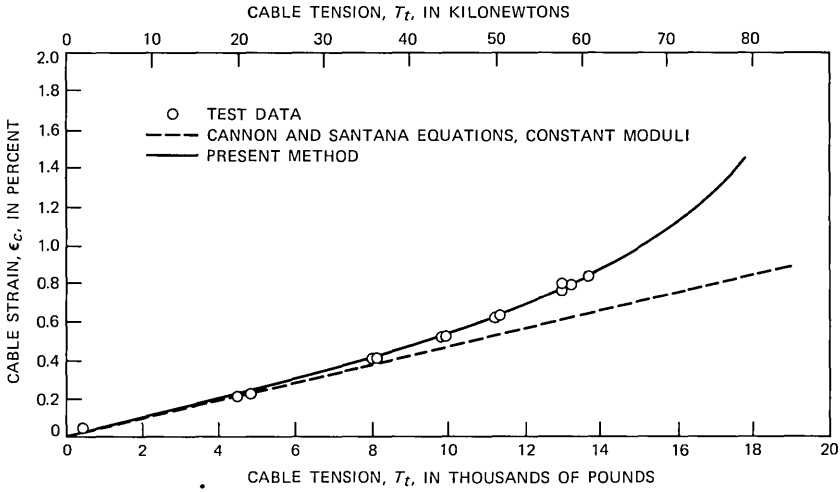


Fig. 4—Comparison of theoretical and experimental results for cable strain versus cable tension.

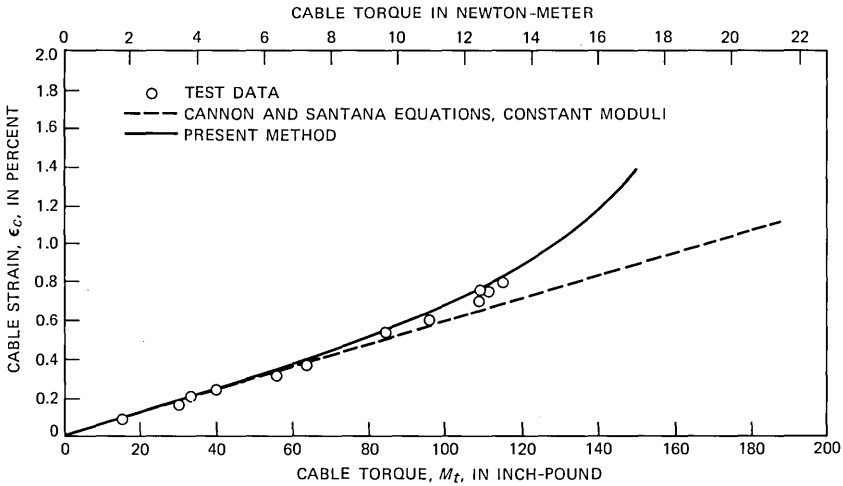


Fig. 5—Comparison of theoretical and experimental results for cable strain versus cable torque.

calculated results from the present method and the experimental results are in excellent agreement.

VI. CONCLUSION

A method has been formulated to evaluate the mechanical properties of cables containing stranded-strength members, in both linear elastic and nonlinear plastic regions. Application of the method to the exper-

imental undersea lightguide cable shows excellent agreement to the tensile test results of the cable. This method can be applied to future cable design to evaluate the cable breaking strength, cable tension versus strain, cable moment versus strain, and cable twist versus strain in both linear elastic and nonlinear plastic regions.

VII. ACKNOWLEDGMENTS

The author would like to thank his former teacher, Professor W. R. Sears, for his guidance during the graduate school years at Cornell University.

The author also thanks T. C. Cannon, M. R. Santana, and R. D. Tuminaro for useful discussions during the course of the study, and H. M. Brinser and D. A. Meade, for conducting the tests.

REFERENCES

1. E. E. Zajac, "Dynamics and Kinematics of the Laying and Recovery of Submarine Cable," B.S.T.J. *35,* No. 5 (September 1957), pp. 1129–207.
2. *Optical Fiber Telecommunications,* S. E. Miller and A. G. Chynoweth, eds., New York: Academic Press, 1979, Chapter 12.
3. T. C. Cannon and M. R. Santana, "Mechanical Characterization of Cable Containing Helically Wrapped Reinforcing Elements," Twenty-fourth Int. Wire and Cable Symp., Cherry Hill, New Jersey, 1975.
4. G. A. Costello and J. W. Phillips, "Contact Stresses in Thin Twisted Rods," J. Appl. Mech., *40* (June 1973), pp. 629–30.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 3, March 1983
Printed in U.S.A.

Performance of a Fast Algorithm for FIR System Identification Using Least-Squares Analysis

By S. L. MARPLE, JR.,* and L. R. RABINER

(Manuscript received August 31, 1982)

A wide variety of procedures have been proposed for identifying a finite impulse response (FIR) linear system from the input and output of the system. Most recently, a fast, efficient, least-squares method was proposed by Marple, and was shown to require less computation and storage than any other known procedure for identifying moderate to large FIR systems. In this paper we measure the actual performance of the newly proposed fast system identification algorithm by using it to estimate a variety of FIR systems excited by either white noise or a speech signal. It is shown that essentially theoretically ideal performance is achieved for white noise inputs; however, for speech signals poor performance was obtained because of the lack of certain frequency bands in the excitation. A simple modification to the estimation procedure is proposed and is shown to provide substantial performance improvements. Using the spectrally modified speech signal, the performance of the fast system identification algorithm was found to be acceptable for a wide variety of applications.

I. INTRODUCTION

In previous papers,[1-4] two system identification methods, the classical least-squares analysis (LSA) and a short-time spectral analysis (SSA) procedure, had their performance compared and contrasted in the presence of high noise levels and in situations where the input signal was band-limited (nonwhite noise and speech). This earlier work found that, while the LSA method produced better performance than the SSA procedure, the computational burden of the Cholesky solution of the equations of the LSA method became prohibitive when com-

* Schlumberger Well Services, Houston, Texas.

pared to the SSA method as the system order became large (as it can be in speech processing, where the filter order can be on the order of 1000). This factor weighed in favor of the SSA method. However, the development of fast algorithms for the solution of the LSA normal equations[5-8] has greatly reduced the computational complexity. This paper evaluates the performance of the LSA procedure in the context of these fast algorithms.

The fast algorithm that we will consider for solving normal equations for the least-squares FIR system identification has computational complexity proportional to M^2, where M is the filter order, and storage that grows linearly (rather than quadratically) with M. A byproduct of the computation is an estimate of the linear prediction coefficients of the input process. The fast algorithm also has simple, built-in, numerical ill-conditioning checks. If the model order is uncertain, the fast algorithm recursively provides all optimum least-squares solutions from filter order 1 up to some user-selected maximum order, M, thereby providing a built-in search capability for finding the appropriate system order without having to start over.

The outline of this paper is as follows. In Section II we review the normal equations of the least-squares system identification and show how a fast algorithm can be derived to solve these equations. In Section III we present an evaluation of the performance of the fast algorithm for several different FIR systems with both broadband noise and speech inputs. In Section IV we review the results and compare them to those obtained previously using short-time spectral analysis methods.

II. REVIEW OF THE LEAST-SQUARES NORMAL EQUATIONS

Figure 1 shows a block diagram of the finite-impulse-response system identification model. The discrete input signal, $x(n)$, drives the unknown system to produce the discrete sequence, $y(n)$, where n is an integer index. The unknown system was modeled as an FIR filter, assumed to have an impulse response duration of $M + 1$ samples, so that the estimated impulse response, $\hat{h}(n)$, is zero for $n < 0$ and $n > M$. The order M of the FIR system is defined here as the highest index of the impulse response, $\hat{h}(M)$.

The approach used is to determine the impulse-response coefficients, $\hat{h}(n)$, and the system order M that minimize the squared error based on a finite number of measurements of the input process, $x(n)$, and the output process, $y(n)$. Denoting the linear estimate of $y(n)$ by $\hat{y}(n)$, then

$$\hat{y}(n) = \sum_{m=0}^{M} \hat{h}(m)x(n-m), \tag{1}$$

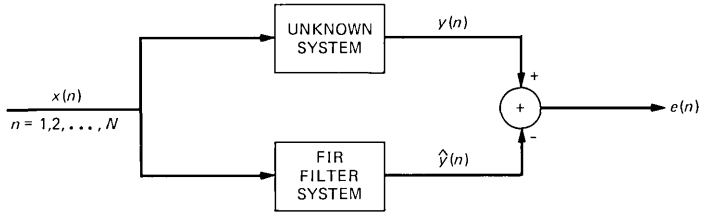


Fig. 1—Block diagram of the FIR system identification.

which is the familiar convolution expression of a finite-impulse-response filter. The estimation error, $e(n)$, is then

$$e(n) = y(n) - \hat{y}(n). \tag{2}$$

We wish to minimize the total squared error, P_M, of an Mth-order model

$$P_M = \sum_n e^2(n) \tag{3}$$

with respect to $\hat{h}(0), \cdots, \hat{h}(M)$ and based on the block of finite data sets $x(1), \cdots, x(N)$ and $y(1), \cdots, y(N)$. Note that the index range for n in eq. (3) has purposely been left unspecified. To operate only on the available data samples, the range must be selected to be $n = M + 1$ to N, which is the index range selected for this paper because it provides the best performance for relatively short data segments. However, by defining the unobserved data to be zero for $n \leq 0$, then a so-called "prewindowed" index range of $n = 1$ to N can be used [the data $x(n)$ and $y(n)$ for $n \leq 0$ is "windowed" to zero]. These two cases are illustrated in eq. (4) using a matrix structure to describe the error terms:

$$\begin{Bmatrix} 2 \\ 1 \end{Bmatrix} \left\{ \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{M+1} \\ \vdots \\ e_N \end{bmatrix} \right. = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{M+1} \\ \vdots \\ y_N \end{bmatrix} - \begin{Bmatrix} 2 \\ 1 \end{Bmatrix} \left\{ \begin{bmatrix} x_1 & 0 \\ x_2 & x_1 \\ \vdots & \vdots \\ x_{M+1} & x_M \cdots x_1 \\ \vdots & \vdots \\ x_N & x_{N-1} \cdots x_{N-M} \end{bmatrix} \right. \begin{bmatrix} \hat{h}_0 \\ \hat{h}_1 \\ \vdots \\ \hat{h}_M \end{bmatrix} \tag{4a}$$

or

$$E_M = Y_M - X_M \hat{H}_M \tag{4b}$$

and

$$P_M = E_M^T E_M, \tag{5}$$

where E_M, Y_M, $\hat{H}_M$ are column vectors and X_M is a rectangular Toeplitz data matrix. The T denotes matrix transpose. If X_M includes only that portion within the 1 brace, this corresponds to the index range $n = M + 1$ to N. If X_M includes that portion indicated within the 2 brace, this corresponds to the prewindowed index range $n = 1$ to N.

If P_M is now minimized by setting the derivatives with respect to $\hat{h}(0), \cdots, \hat{h}(M)$ to zero, then the resulting least-squares solution can be expressed in matrix form as

$$\Phi_M^{xx} \hat{H}_M = \Phi_M^{yx}, \tag{6}$$

where

$$\Phi_M^{xx} = X_M^T X_M = \text{an } (M+1) \times (M+1) \text{ matrix}$$

$$\Phi_M^{yx} = X_M^T Y_M = \text{an } (M+1) \text{ column vector.}$$

This is the discrete Wiener-Hopf equation. The minimum squared error is

$$\min P_M = Y_M^T Y_M - (\Phi_M^{yx})^T \hat{H}_M. \tag{7}$$

Note that this solution is applicable no matter what index range is selected. Also note that while the matrix Φ_M^{xx} is not, in general, Toeplitz, it is the product of two rectangular Toeplitz data matrices. This will prove to be a key factor in developing a fast algorithm solution for eq. (6). For the index range of interest here, individual elements of Φ_M^{xx} are

$$\phi_M^{xx}(i, j) = \sum_{n=M+1}^{N} x(n-j)x(n-i) \qquad \text{for } 0 \le i, j \le M. \tag{8}$$

The elements of Φ_M^{yx} for the unwindowed index range are

$$\phi_M^{yx}(i) = \sum_{n=M+1}^{N} y(n)x(n-i) \qquad \text{for } 0 \le i \le M. \tag{9}$$

Also, we have

$$Y_M^T Y_M = \sum_{n=M+1}^{N} y^2(n). \tag{10}$$

Note that the number of data samples must be at least twice the system order plus one, $N \ge 2M + 1$, in order for Φ_M^{xx} to be nonsingular.

It is also possible to perform a linear-phase FIR system identification by forcing symmetry in the filter. The linear-phase estimate is then

$$\hat{\hat{y}}(n) = \hat{\hat{h}}(0)x(n) + \sum_{m=1}^{M} \hat{\hat{h}}(m)[x(n+m) + x(n-m)] \tag{11}$$

with error

$$\epsilon(n) = y(n) - \hat{\hat{y}}(n). \tag{12}$$

Note that the total impulse-response duration of this filter is $2M + 1$. Forming the sum of squared errors over all valid error terms

$$Q_M = \sum_{n=M+1}^{N-M} \epsilon^2(n) \tag{13}$$

and minimizing leads to the normal equations

$$\Psi_M^{xx} \hat{\hat{H}}_M = \Psi_M^{yx},$$

where

$$\hat{\hat{H}}_M = \begin{bmatrix} \hat{\hat{h}}(M) \\ \vdots \\ \hat{\hat{h}}(1) \\ \hat{\hat{h}}(0) \\ \hat{\hat{h}}(1) \\ \vdots \\ \hat{\hat{h}}(M) \end{bmatrix}, \qquad \Psi_M^{yx} = \begin{bmatrix} r^{yx}(M) \\ \vdots \\ r^{yx}(1) \\ r^{yx}(0) \\ r^{yx}(1) \\ \vdots \\ r^{yx}(M) \end{bmatrix},$$

$$\Psi_M^{xx} = \begin{bmatrix} r_M^{xx}(0,0) & \cdots & r_M^{xx}(0,2M) \\ \vdots & & \vdots \\ r_M^{xx}(2M,0) & \cdots & r_M^{xx}(2M,2M) \end{bmatrix}$$

$$r_M^{xx}(j,k) = \sum_{n=1}^{N-2M} [x(n+j)x(n+k) + x(n+2M-j)x(n+2M-k)]$$

$$r_M^{yx}(k) = \sum_{n=M+1}^{N-M} y(n)[x(n+k) + x(n-k)] \qquad 0 \le k \le M$$

with minimum squared error

$$Q_M = \sum_{n=M+1}^{N-M} y^2(n) - \frac{1}{2}\hat{\hat{h}}(0) r_M^{yx}(0) - \sum_{m=1}^{M} \hat{\hat{h}}(m) r_M^{yx}(m).$$

As before, in order for the normal equations to be nonsingular, the number of data samples must be at least twice the system order ($2M$ in this case) plus one, or $N \ge 4M + 1$.

2.1 Basis for a fast algorithm solution

In this section, a brief outline of the basis for a fast algorithm that solves eq. (6) with a number of operations proportional to M^2 is presented. For details of the full algorithm, consult Ref. 5. A fast algorithm also exists for the linear-phase FIR system identification,[8]

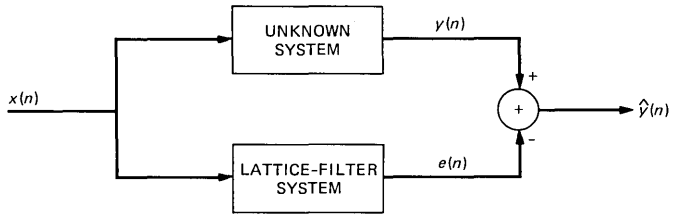


Fig. 2—Block diagram of the lattice system identification.

but will not be discussed here since the focus is on the general FIR system identification algorithm.

The fast algorithm presented here processes the available data as a block. The algorithm recursively generates all solutions from order $m = 1$ to M, where M is the maximum order. This algorithm is just one of a wider class of fast algorithms for solutions of least-squares prediction problems. If processing data on a sequential sample-by-sample basis is preferred to block processing of the data (such as for adaptive equalizer applications), then an alternative, but numerically equivalent, solution to eq. (6) is to use a lattice filter in lieu of the FIR filter. The lattice-filter output is $e(n)$ rather than $\hat{y}(n)$ (the FIR output), as shown in Fig. 2. An example of such a sequential fast algorithm for the "prewindowed" data range is presented in Ref. 6.

The key to a fast algorithm for solution of eq. (6) is the recognition that the matrix Φ_M^{xx} appears in the context of the linear prediction problem. The linear prediction filter tries to make the best estimate of the current value of $x(n)$ based on past samples of $x(n)$,

$$\tilde{x}(n) = -\sum_{m=1}^{M} a(m)x(n-m). \tag{14}$$

The prediction error $e^f(n)$ is then

$$e^f(n) = x(n) - \tilde{x}(n) = \sum_{m=0}^{M} a(m)x(n-m), \tag{15}$$

where $a(0) = 1$ and the superscript f denotes this as the "forward" linear prediction filter error. A "backward" estimate $\tilde{\tilde{x}}(n)$ can also be defined as

$$\tilde{\tilde{x}}(n-M) = -\sum_{m=1}^{M} b(m)x(n+m-M) \tag{16}$$

with "backward" prediction error

$$e^b(n) = x(n-M) - \tilde{\tilde{x}}(n-M) = \sum_{m=0}^{M} b(m)x(n+m-M),$$

where $b(0) = 1$. If we now minimize the forward squared prediction error

$$P_M^f = \sum_{n=M+1}^{N} [e^f(n)]^2$$

and the backward squared prediction error

$$P_M^b = \sum_{n=M+1}^{N} [e^b(n)]^2;$$

then we obtain the usual least-squares linear-prediction normal equations

$$\Phi_M^{xx} A_M = \begin{pmatrix} P_M^f \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \Phi_M^{xx} B_M = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ P_M^b \end{pmatrix}, \tag{17}$$

where vectors A_M and B_M are defined as

$$A_M = \begin{bmatrix} 1 \\ a(1) \\ \vdots \\ a(M) \end{bmatrix}, \qquad B_M = \begin{bmatrix} b(M) \\ \vdots \\ b(1) \\ 1 \end{bmatrix}.$$

Equation (17) is the so-called "covariance" equation of linear-prediction speech analysis. An efficient, recursive algorithm for their solution has been previously presented[7] and is incorporated as part of the FIR fast algorithm without further discussion. Equation (17) can be rewritten as

$$\Phi_M^{xx} \mathscr{A}_M = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \Phi_M^{XX} \mathscr{B}_M = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \tag{18}$$

where

$$\mathscr{A}_M = \begin{bmatrix} 1/P_M^f \\ a(1)/P_M^f \\ \vdots \\ a(M)/P_M^f \end{bmatrix}, \qquad \mathscr{B}_M = \begin{bmatrix} b(M)/P_M^b \\ \vdots \\ b(1)/P_M^b \\ 1/P_M^b \end{bmatrix}.$$

Thus, $\mathscr{A}_M$ and $\mathscr{B}_M$ form the first and last columns of the inverse of the Φ_M^{xx} matrix. Since the solution to eq. (12) involves the inverse $[\Phi_M^{xx}]^{-1}$,

$$\hat{H}_M = (\Phi_M^{xx})^{-1} \Phi_M^{xy}, \tag{19}$$

then we would suspect that the linear-prediction solution is an integral part of the system identification solution [eq. (19)]. This is indeed the case. In fact, only the first and last columns of the inverse (or alternatively, the vectors A_M and B_M) of Φ_M^{xx} are required to obtain a recursive solution for $\hat{H}_M$.

Three fundamental relationships govern the ability to obtain a recursive algorithm. To illustrate these relationships, it is first necessary to combine eqs. (5) and (6) into a single augmented expression as follows:

$$\bar{\Phi}_M \bar{H}_M = \bar{P}_M, \tag{20}$$

where

$$\bar{\Phi}_M = \underbrace{\left[\begin{array}{c|c} \phi_M^{yy}(0) & (\Phi_M^{yx})^T \\ \hline \Phi_M^{yx} & \Phi_M^{xx} \end{array}\right]}_{M+2} \Bigg\} M+2$$

$$\bar{H}_M = \left[\begin{array}{c} 1 \\ -\hat{H}_M \end{array}\right] \Bigg\} M+2$$

$$\bar{P}_M = \left[\begin{array}{c} P_M \\ 0 \\ \vdots \\ 0 \end{array}\right] \Bigg\} M+2, \qquad \phi_M^{yy}(0) = Y_M^T Y_M.$$

An alternative relationship for $\bar{\Phi}_M$ is

$$\bar{\Phi}_M = \sum_{n=M+1}^{N} \bar{X}_M(n) \bar{X}_M^T(n), \tag{21}$$

where $\bar{X}_M(n)$ is the vector

$$\bar{X}_M(n) = \left[\begin{array}{c} y(n) \\ x(n) \\ \vdots \\ x(n-M) \end{array}\right].$$

We may obtain the three basic partitions:

$$\bar{\Phi}_M = \underbrace{\left[\begin{array}{c|c} \phi_M^{yy}(0) & (\Phi_M^{yx})^T \\ \hline \Phi_M^{yx} & \Phi_M^{xx} \end{array}\right]}_{\begin{array}{cc} 1 & M+1 \end{array}} \begin{array}{l} \}1 \\ \}M+1 \end{array} \tag{22}$$

$$\bar{\Phi}_M = \underbrace{\left[\begin{array}{c|c} \bar{\Phi}'_{M-1} & W_M \\ \hline W_M^T & \phi_M^{xx}(M,M) \end{array}\right]}_{\begin{array}{cc} M+1 & 1 \end{array}} \begin{array}{l} \}M+1 \\ \}1 \end{array} \tag{23}$$

$$\bar{\Phi}'_{M-1} = \bar{\Phi}_{M-1} - \bar{X}_{M-1}(M) \bar{X}_{M-1}^T(M), \tag{24}$$

where the definition of W_M is

$$W_M = \sum_{n=M+1}^{N} x(n-M)\bar{X}_{M-1}(n).$$

All the recursive relationships in the algorithm have their roots in eqs. (22) through (24). Using these equations, it is possible to derive the following key recursive relationships for the system identification parameters

$$\bar{H}'_{M-1} = \bar{H}_{M-1} + \frac{e(M)}{(1-\delta_{M-1})}\begin{pmatrix} 0 \\ D_{M-1} \end{pmatrix} \text{ time update} \tag{25}$$

$$\bar{H}_M = \begin{pmatrix} \bar{H}'_{M-1} \\ 0 \end{pmatrix} + \frac{\alpha_M}{P_M^b}\begin{pmatrix} 0 \\ B_M \end{pmatrix} \text{ order update,} \tag{26}$$

where α_M, δ_M are scalar gain values, B_M is the vector of backward linear-prediction coefficients, and $D_M = (\Phi_M^{xx})^{-1}X_M(M+1)$ is a vector, all of which are obtained as part of the linear-prediction-algorithm solution. Equations (25) and (26) highlight the intertwined relationship of the linear-prediction fast algorithm with that of the fast algorithm for the system identification solution. One also can see how all lower-order solutions are obtained recursively along the way to the final selected order.

Counting only second-order terms, the number of multiplications required in the fast algorithm is $2NM + 12M^2$, the number of adds is $2NM + 9M^2$, the number of divides is $8M$, and a total storage of $2N + 7M + 20$ parameters is required (including input and output data sequences). Here N is the number of data values and M is the system order. Roughly NM^2 operations are required to directly solve eq. (6) by Cholesky decomposition if the fast algorithm described here is not used. The Cholesky technique also requires storage proportional to M^2, whereas the fast algorithm requires storage that only increases linearly with increasing M. Exact operation counts for the Cholesky method are given in the appendix. A simple numerical ill-conditioning check in the algorithm is to verify that the scalar variable δ_M in eq. (26) is in the range $0 \leq \delta_M < 1$, which is required by the mathematics of the solution. If it is not in this range, then the fast algorithm recursion has become ill-conditioned.

FORTRAN computer code of the general FIR system identification algorithm may be found in Ref. 5. Code for the linear-phase FIR algorithm may be found in Ref. 8.

III. PERFORMANCE EVALUATION OF THE FAST ALGORITHM

To evaluate the performance of the fast, least-squares, FIR system identification algorithm that solves eq. (6), the system of Fig. 3 was

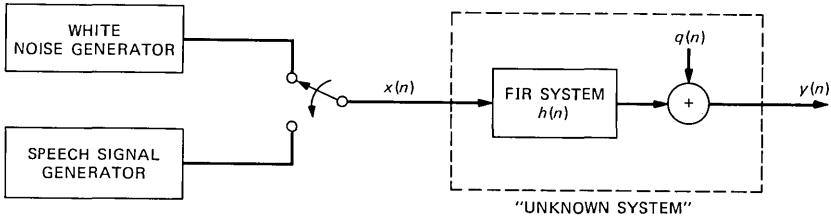


Fig. 3—Block diagram of excitation generation and signal generation for identifying an unknown system.

used to generate the required input and output signals. The signal source was either a white noise signal or a speech signal. The signal source was passed through a given Mth-order FIR system with known impulse response, $h(n)$, $n = 0, 1, \cdots, M$. An additive, random, white noise $q(n)$ was added to the output of the FIR system, at a specified signal-to-noise (s/n) ratio, giving the output signal, $y(n)$. The system was assumed to be in steady state whenever $y(n)$ and $x(n)$ were sampled for system identification purposes [i.e., no initial transients were present in $y(n)$].

The performance measure used in this study was the logarithmic misalignment error defined in Ref. 1 as

$$Q(N, \text{s/n}) = 10 \log_{10} \left[\frac{\sum_{m=0}^{M} [h(m) - \hat{h}(m)]^2}{\sum_{m=0}^{M} h^2(m)} \right], \tag{27}$$

where $\hat{h}(m)$ is a function of N and s/n, and M is the true FIR system order. Note that $h(m)$ are the true FIR coefficients and $\hat{h}(m)$ are the estimated FIR coefficients obtained from the fast algorithm. For nonwhite input signals, a weighted Q function can also be defined using the estimated linear-prediction coefficients, $a(n)$, available as part of the fast algorithm. See Ref. 2 for details.

To fully evaluate the performance of the fast algorithm, five different FIR filters were used, including:

(*i*) Filter 1—A 7-point, nonlinear-phase filter with about 10 dB of spectral deviation across the frequency range. Figure 4 shows plots of the impulse and frequency responses of this filter.

(*ii*) Filter 2—A 25-point, linear-phase, equiripple low-pass filter with about 54 dB of stopband rejection for frequencies above 0.2 times the sampling frequency. Figure 5 shows the impulse and frequency responses of this filter.

(*iii*) Filter 3—A 64-point, nonlinear-phase, reverberation filter with about 30 dB of spectral variation. This filter had an impulse response

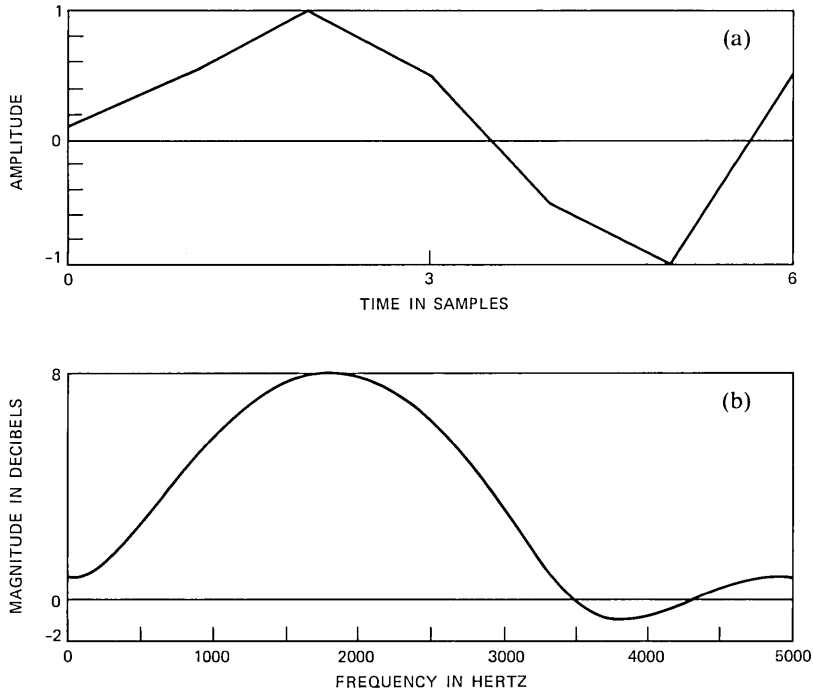


Fig. 4—(a) Impulse response and (b) frequency response of 7-point FIR filter.

that was zero except for $n = 0, 1, 3, 7, 15, 31$, and 63, at which $h(n)$ was 1.0. Figure 6 shows the impulse and frequency responses of this filter.

(*iv*) Filter 4—A 255-point, linear-phase, bandpass filter with 48 dB of rejection in both stopbands. Figure 7 shows the impulse and frequency responses of filter 4.

(*v*) Filter 5—A 256-point, nonlinear-phase, reverberation filter with impulse response that was zero except for $n = 0, 1, 3, 7, 15, 31, 63, 127, 255$, at which $h(n)$ was 1.0.

This set of five filters spans a broad range of impulse-response durations, spectral properties, and temporal properties, and it was felt that it would provide an adequate test of the fast identification algorithm.

3.1 Performance with noise excitation signals

The first set of tests used as the excitation signal the white noise source of Fig. 3. Figure 8 shows plots of the long-time average autocorrelation and power spectrum of the source. The noise spectrum is essentially flat to within ±3 dB.

The white-noise signal was used to drive the system of Fig. 3 for each of the five FIR filters discussed above. For filter 1, data lengths of N from 50 to 1950 were used, and values of s/n from −6 dB to ∞ (no

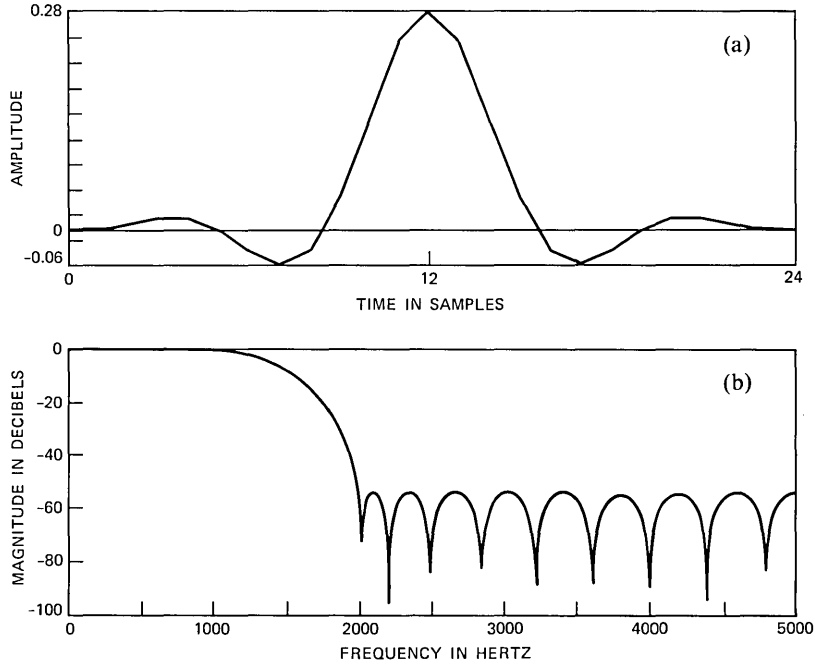


Fig. 5—(a) Impulse response and (b) frequency response of 25-point FIR low-pass filter.

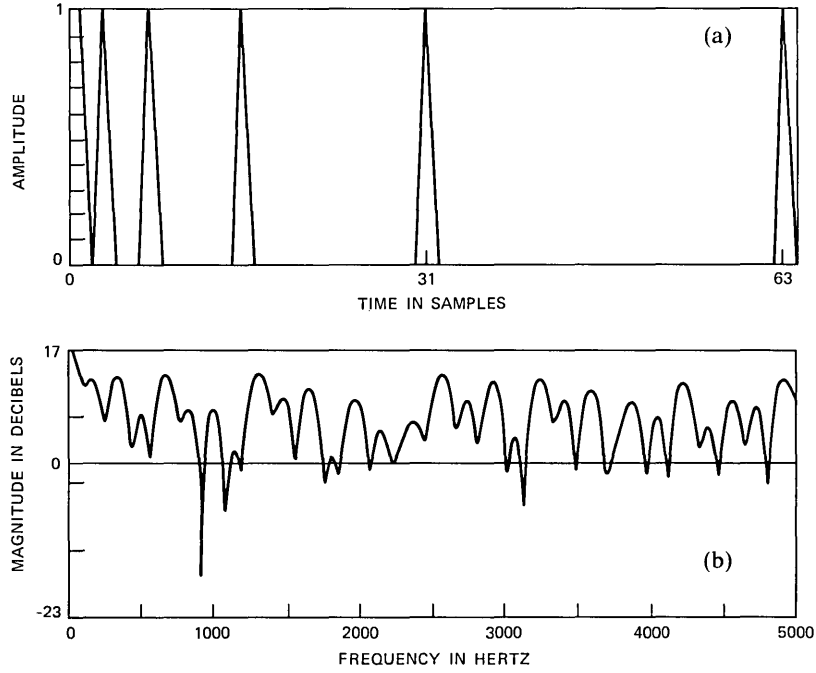


Fig. 6—(a) Impulse response and (b) frequency response of 64-point FIR echo filter.

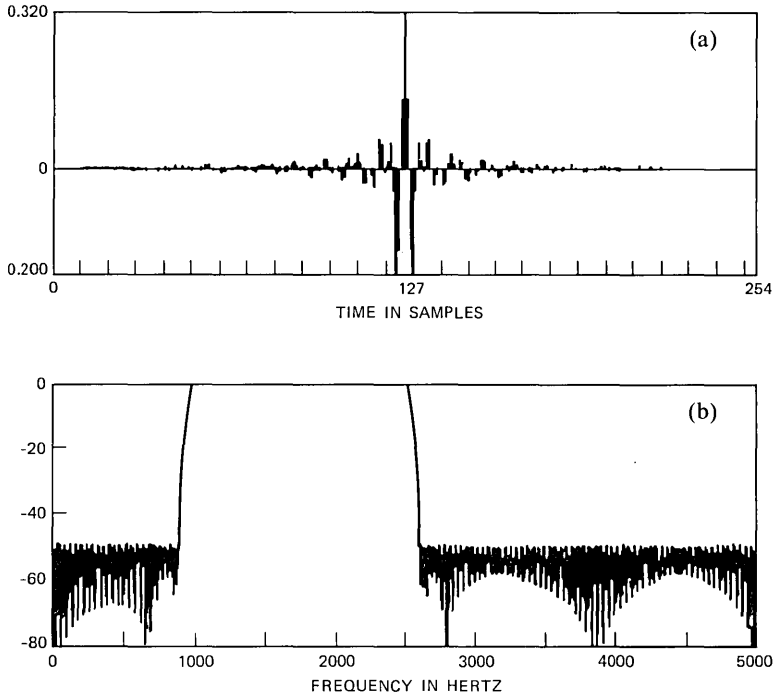


Fig. 7—(a) Impulse response and (b) frequency response of 255-point FIR bandpass filter.

additive noise) were used. For each of the other filters, values of N from the minimum possible $(2M + 1)$ to 1950 were used, along with s/n = ∞ and s/n = 30 dB.

The results for filter 1 performance scores are shown in Fig. 9, which gives a series of curves of $Q(N, \text{s/n})$ versus log N for several values of s/n. Also shown in the figure are the theoretical expected values for $Q(N, \text{s/n})$ for a white-noise input,[1] which are of the form

$$Q(N, \text{s/n})|_{\text{white input}} = 10 \log_{10}(M/N) - \text{s/n(dB)}. \quad (28)$$

It can be seen from Fig. 9 that the measured values of Q are very close to the theoretical expectations for s/n's in the range −6 to 42 dB, and for all N. For s/n = ∞, the measured values of Q (from −88 to −103 dB) reflect the obtainable single-precision accuracy of the computation.

Figures 10 and 11 show similar results for filters 3 and 4, the 64-point reverberation filter and the 255-point bandpass filter. For s/n = ∞, the algorithm had Q values of from −100 dB to −109 dB for the 64-point filter, and from −98 dB to −104 dB for the 255-point bandpass filter. The small degradation in performance is due to the higher roundoff errors for the longer impulse responses. For the case where

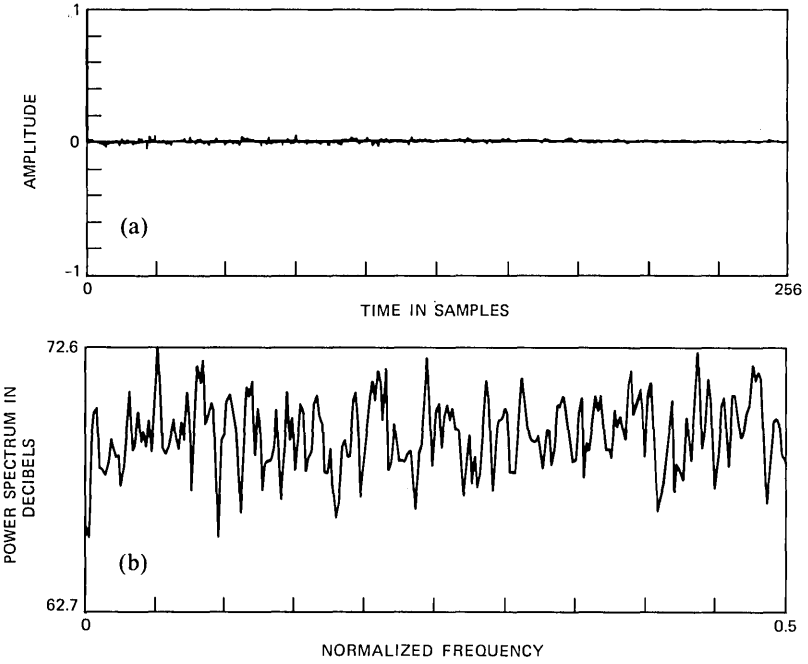


Fig. 8—Long-time average (a) autocorrelation and (b) power spectrum of white-noise source.

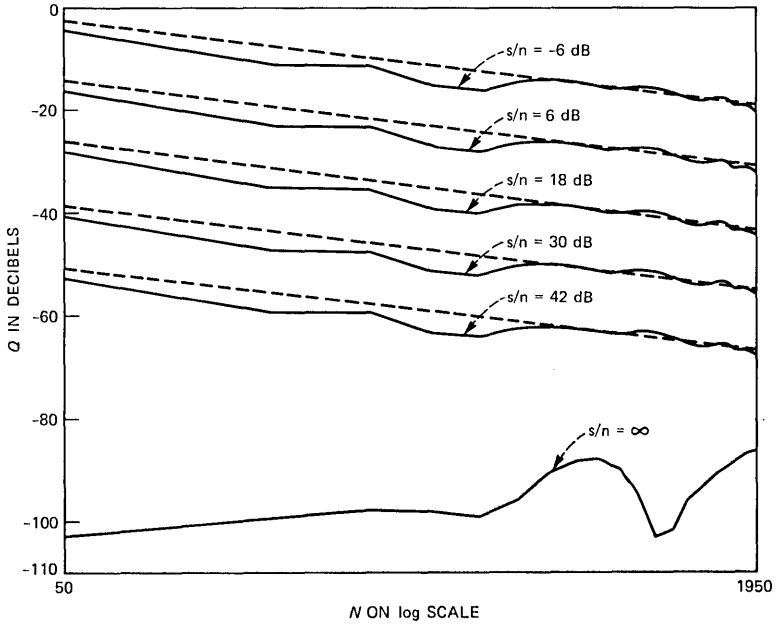


Fig. 9—Plots of Q versus N for several values of s/n for noise input and 7-point FIR filter.

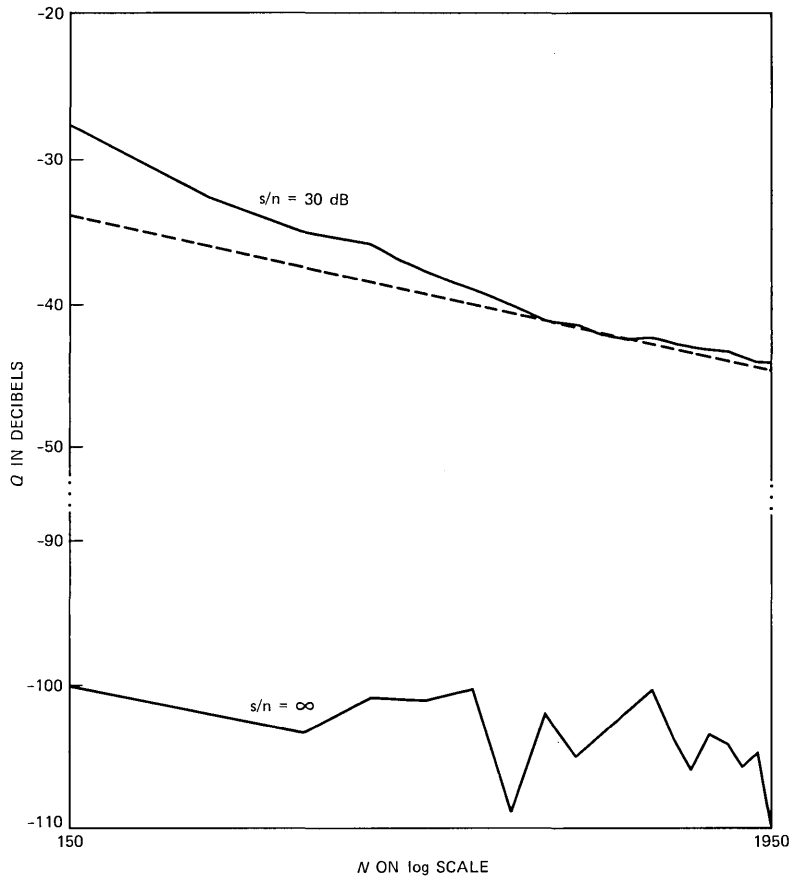


Fig. 10—Plots of Q versus N for s/n = 30 dB and s/n = ∞ for noise input and 64-point FIR echo filter.

s/n = 30 dB the Q curves of Figs. 10 and 11 show about 10-dB worse performance than the theoretical average when N is on the order of $2M + 1$ (the minimum N required to ensure nonsingularity), whereas the performance of the fast algorithm approaches the theoretical estimate as N becomes much larger than $2M + 1$.

The performance of the fast, least-squares estimation algorithm for filters 2 and 5 was essentially identical to that for filters 3 and 4.

As a final example of the noise-excited results, Fig. 12 shows a plot of $Q(N, \text{s/n})$ versus log N for filter 2 (the 25-point low-pass filter) cases, where N is varied from 30 to 70 in steps of 1 and s/n = ∞. The values of Q are very poor for $N \leq 2M$; however, once N exceeds this critical value, the values of Q fall below -75 dB, indicating excellent algorithm performance.

In summary, the results on the white Gaussian noise excitation show

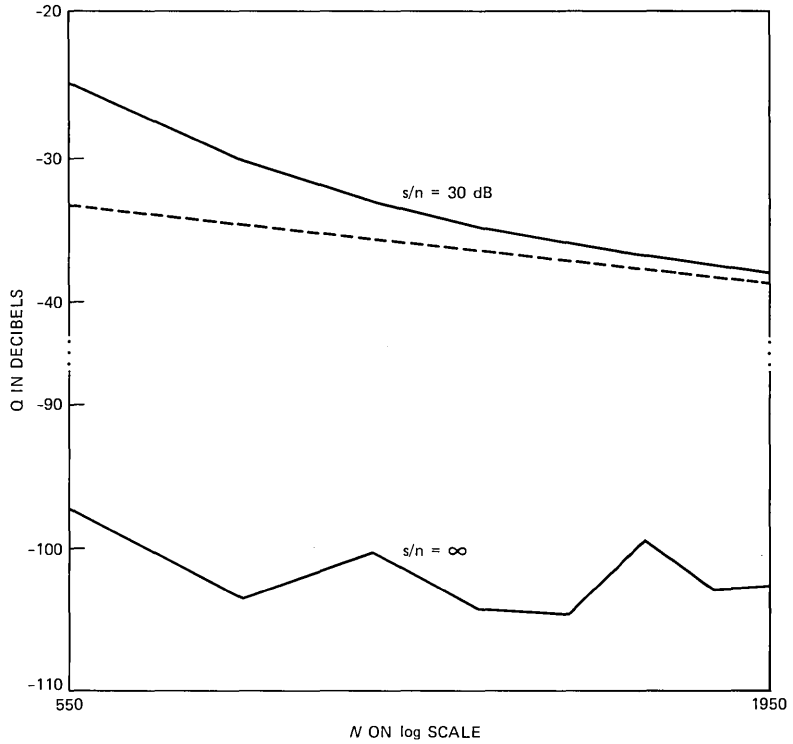


Fig. 11—Plots of Q versus N for s/n = 30 dB and s/n = ∞ for noise input and 255-point FIR bandpass filter.

that the fast, least-squares system identification algorithm performed exceedingly well for all FIR filters, signal-to-noise ratios, and data lengths (so long as $N \geq 2M + 1$).

3.2 Performance on speech excitation

The second series of tests of the performance of the fast system identification algorithm used speech samples as the excitation. Figure 13 shows plots of the long-term average autocorrelation and power spectrum of the speech signal used in our tests. It can be seen that the long-term average power spectrum exhibits a 60- to 70-dB variation in spectral magnitudes, and noticeable gaps in energy throughout the frequency range. Thus, system identification based on speech inputs and outputs is significantly more difficult than it was for white-noise inputs.

The speech excitation was used as input to each of the five FIR filters of Section III. In this section we will concentrate on results obtained using filter 2, the 25-point linear-phase low-pass filter. Results on the other filters were more or less comparable, considering the problems that were encountered.

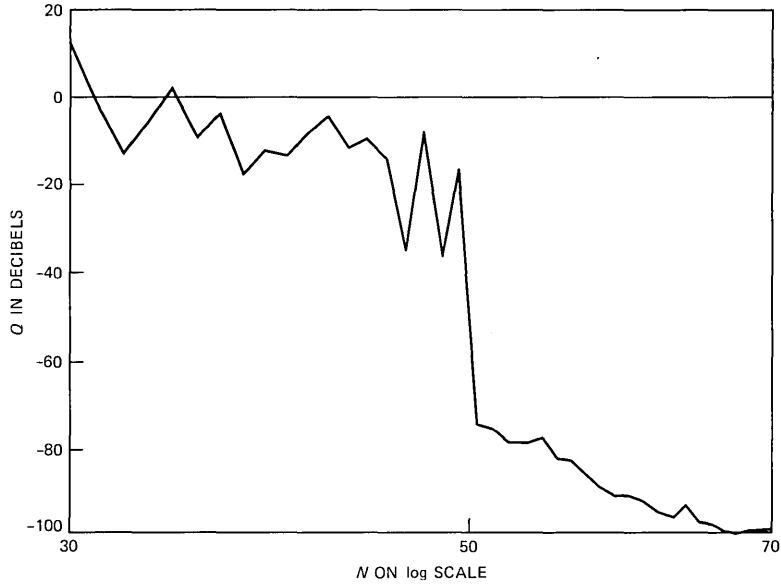


Fig. 12—Plot of Q versus N for 25-point FIR low-pass filter showing abrupt change in Q for $N = 50$.

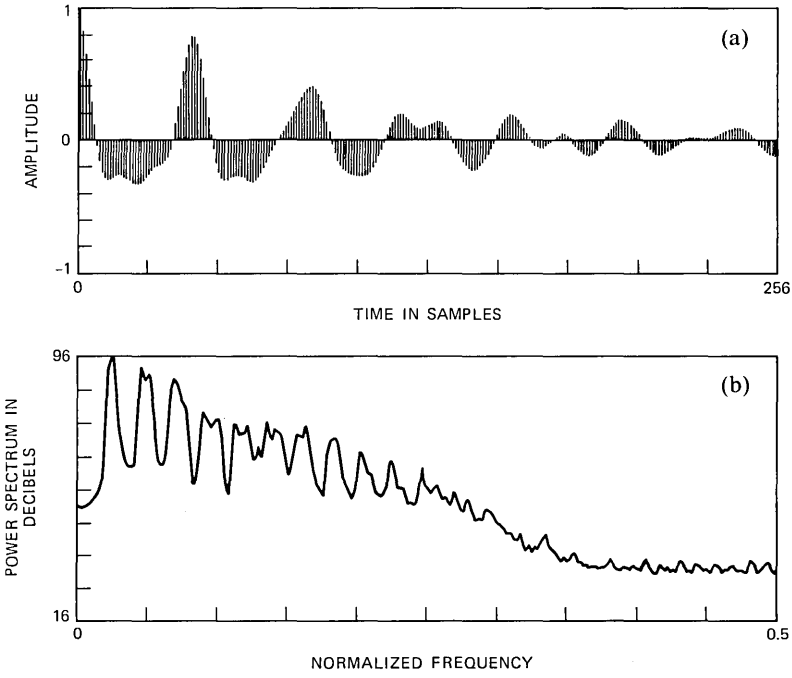


Fig. 13—Long-time average (a) autocorrelation and (b) power spectrum for speech input signal.

Figure 14 shows plots of Q versus log N for s/n = ∞ and s/n = 40 and 60 dB for the filter 2, and for values of N from 100 to 1000. [These results were obtained using the fast algorithm designed for linear-phase systems. Results using the nonlinear-phase system algorithm of eq. (6) were consistently worse than for the linear-phase algorithm, and will not be presented here.] For s/n = ∞ the values of Q range from −30 to −46 dB. These results, for the case with no additive noise,

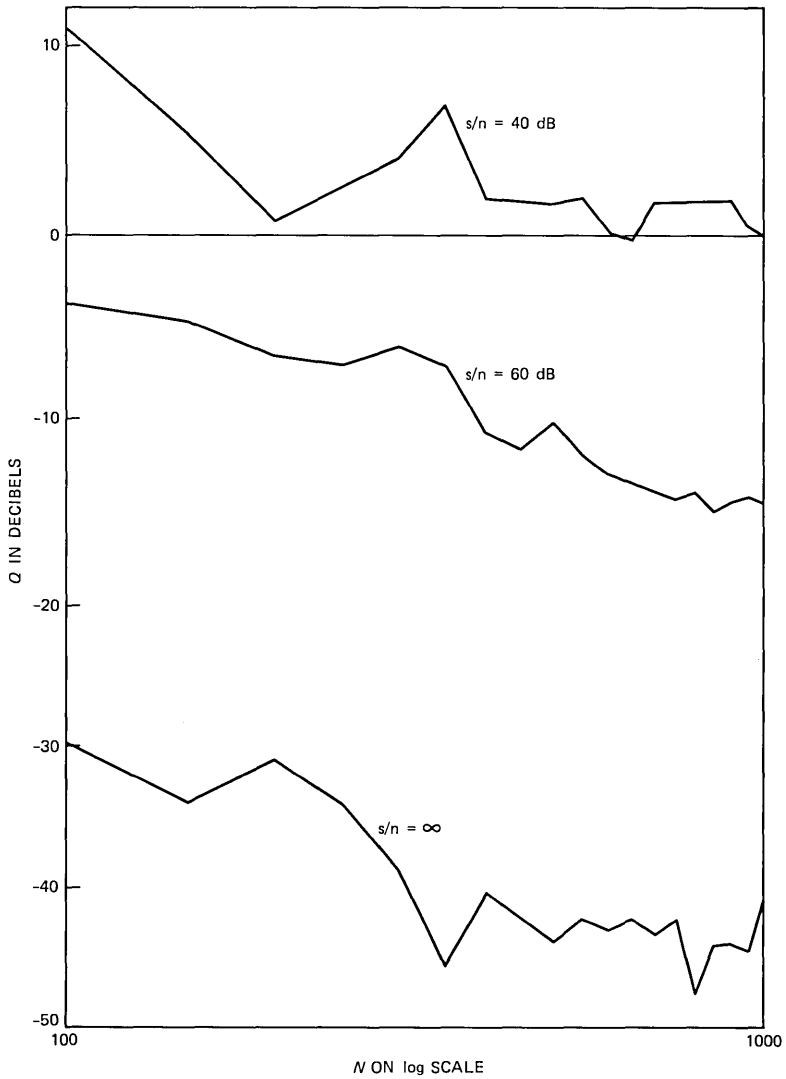


Fig. 14—Plots of Q versus N for several values of s/n for speech input and 25-point FIR low-pass filter.

are significantly worse than those for the white-noise inputs. In the s/n = 40-dB case, all values of Q were about 0 dB or higher, indicating very poor performance. For s/n = 60 dB, the results showing values of Q from −43 to −14 dB indicate marginal performance at best.

The results shown in Fig. 14 are anticipated from previous studies[1-4] and our understanding of the mechanisms of the system identification algorithm. The problem is illustrated in Fig. 15, which shows an estimated impulse response and the resulting frequency response for one set of conditions. It can be seen that the frequency-response estimate is quite good for frequencies in the passband of the low-pass filter (i.e., frequencies less than 0.2 times the sampling frequency). However, the combination of the nonwhite source and spectral gaps with the 54-dB rejection in the stopband leads to extremely poor spectral estimates in the stopband.

To eliminate the spectral gaps in the speech spectrum, a small modification was made to the system block diagram of Fig. 3. A random white noise was added to the speech signal prior to the linear filtering operation to guarantee that all frequencies were present (to some extent) at the input to the linear system. The white noise was

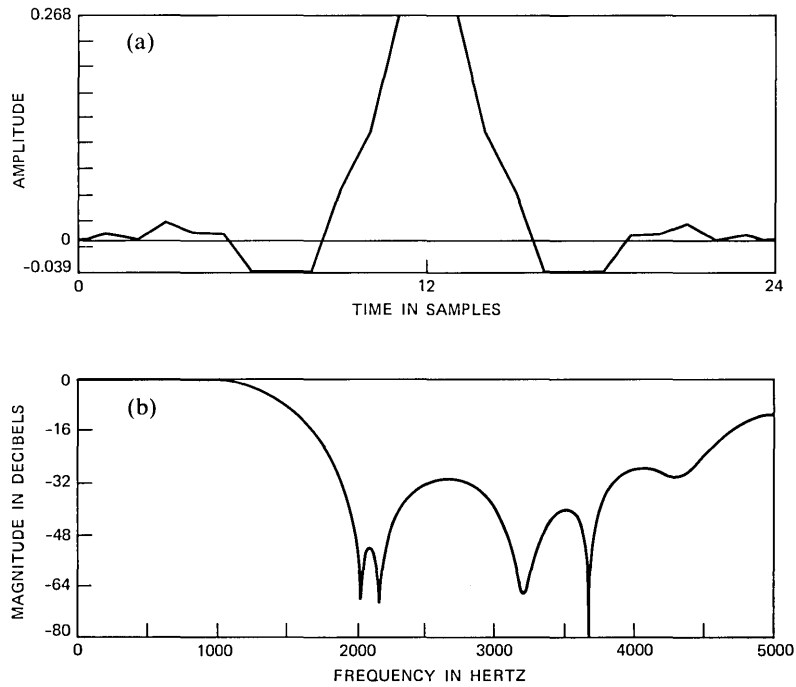


Fig. 15—(a) Plots of estimated impulse response and (b) frequency response for speech input to a 25-point FIR low-pass filter.

added to give an effective signal-to-noise ratio (at the source) of 40 dB. The speech plus noise signal was considered the new input to the system identification procedure.

Figure 16 shows the results obtained using the modified-speech signal input. Shown in this figure are plots of Q versus log N for s/n = ∞, 80, 60, and 40 dB, and for values of N from 100 to 1000. For s/n = ∞ values of Q as high as −78 dB are obtained, showing the vast improvement in system estimation. Similarly, for s/n = 40, 60, and 80 dB vast improvements in Q values are obtained, leading to useful system estimates in most cases.

In summary, the results of system identification using the fast, least-squares algorithm on speech signals indicate that without some spec-

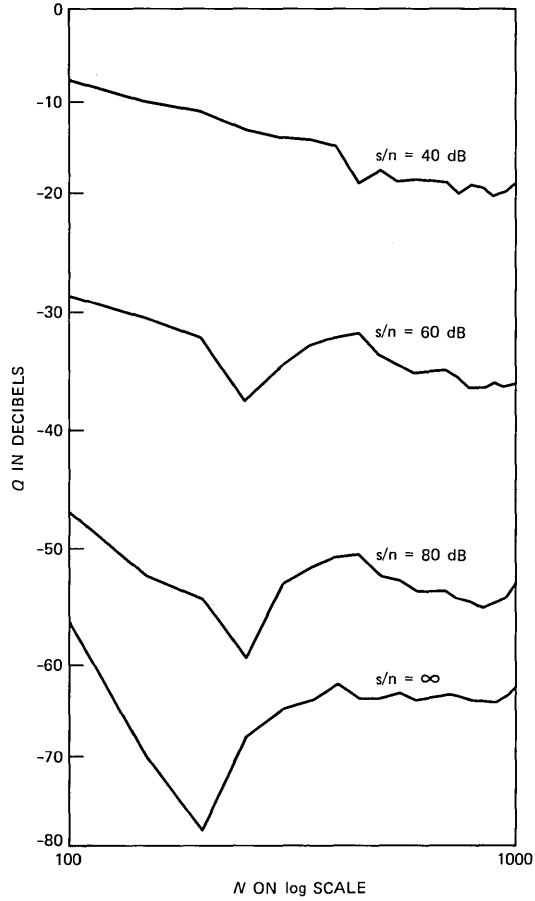


Fig. 16—Plots of Q versus N for several values of s/n for frequency-stabilized speech input to a 25-point FIR low-pass filter.

tral stabilization to guarantee some excitation at all frequencies of interest, the performance of the system identification algorithms is unacceptable in most cases of interest. However, it was shown that even a fairly trivial form of spectral stabilization produced greatly improved system estimates in all cases, and hence made the estimation procedure viable.

IV. DISCUSSION

The results presented in Section III show the following:

(*i*) For white-noise input signals the fast, nonlinear-phase, least-squares system identification procedure performed extremely well over all filter types, filter durations, and signal-to-noise ratios. For such cases the proposed algorithm appears to have significant computational and storage advantages over all other proposed methods.

(*ii*) For nonwhite input signals with spectral gaps (e.g., speech signals) the fast, linear-phase, least-squares system identification procedure was at best barely adequate (for infinite signal-to-noise ratio) and inadequate for signal-to-noise ratios in the practical range of interest (15 to 50 dB). This poor performance was shown to be due primarily to the spectral gaps in the source and a simple modification was proposed whereby a white noise was added to the speech signal to provide a stabilized spectral magnitude at all frequencies of interest.

(*iii*) When the spectral stabilization procedure was used on speech signals the fast, linear-phase, least-squares system identification procedure performed fairly well over a broad range of signal-to-noise ratios and is useful for a wide range of applications.

To appreciate just how well the fast, least-squares algorithm performed, it is worthwhile comparing the results of Section III with those obtained for alternative FIR system identification algorithms. The two main alternative procedures are the short-time spectral analysis (SSA) methods of Rabiner and Allen[1–4] and the classical "slow," least-squares analysis (LSA) solution obtained by the Cholesky method. For the SSA methods it has been found that for white-noise inputs one can approach the theoretical bounds for sufficiently long data sequences (large N). Hence the new fast, least-squares method performs significantly better than the SSA method for noise inputs, except when N becomes very large. In such cases the performances are similar, but the SSA method is still computationally more expensive than the fast, least-squares method.

In the case of speech inputs, the SSA method (which is essentially doing a spectral divide on two power spectrum estimates) runs into extremely bad sensitivity problems because of the missing frequency bands in the input signal. The spectral stabilization method proposed here does not entirely solve the problem for the SSA method because

of the large dynamic range variations in both the input and output signals. The SSA method tends to need a fairly constant spectral input level to perform at its best.

The classical least-squares method, on the other hand, performs as well as the new, fast, least-squares method. However, the computation grows as NM^2 (rather than NM) and thus the classical LSA method is greatly restricted to small values of M and N. In addition to the computational advantage, the fast least-squares algorithm has a significant storage advantage over the classical least-squares approach, growing linearly with the assumed filter order rather than quadratically. Furthermore, the classical Cholesky decomposition used for the LSA method is notoriously sensitive to numerical ill-conditioning and often yields invalid (improper) solutions for cases such as the speech signal input. Hence, except for small values of N and M, the fast, least-squares algorithm appears to be preferable to the classical LSA method.

Another important advantage of using the fast, least-squares algorithm is that it provides information about the squared prediction and identification errors [linear-prediction coefficient (LPC) and FIR] as a function of filter order. These squared-error parameters, which are computed directly in the algorithm at no additional computational cost, are important to monitor for the following reasons:

(*i*) They provide good indications of the required filter order.

(*ii*) The linear prediction and FIR system identification squared errors are monotonically decreasing functions of the filter order (at least for the nonlinear-phase algorithm).

(*iii*) When the linear prediction squared errors fall off more rapidly than the FIR system identification squared error, then the input to the filter is not spectrally rich, and therefore the solution may be ill-conditioned, and the FIR quality, Q, may be poor.

This latter case is illustrated in Fig. 17, which shows plots of FIR system identification squared error on a log scale, and LPC forward and backward squared errors (also on log scales). For this example the 25-point low-pass filter was excited by nonspectrally rich, voiced speech. The LPC prediction errors decrease rapidly for low filter orders, indicating strong spectral coloration in the input signal. In this case a very poor FIR estimate was obtained.

Figure 18 shows a similar set of squared prediction and identification error plots for the case of a white-noise excitation of the same 25-point low-pass filter. In this case the LPC squared errors are essentially flat (to within 0.6 dB) for all FIR filter orders, indicating a spectrally rich input signal. A very good estimate of the FIR filter was obtained in this case.

It should be noted that the squared error of the fast, linear-phase, system identification estimate is not guaranteed to be a monotonically

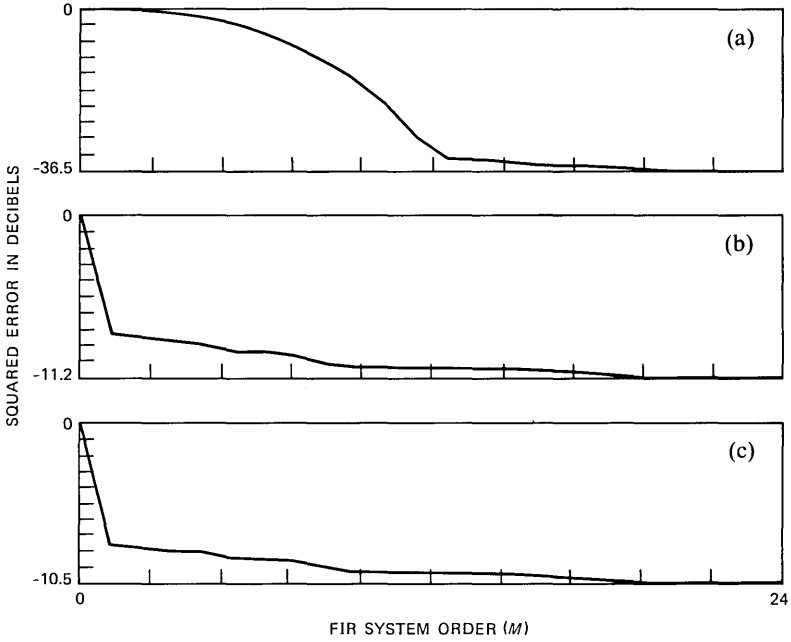


Fig. 17—Plots of squared prediction error versus FIR filter order, M, for a speech signal excitation of a 25-point low-pass filter. (a) FIR error. (b) LPC forward error. (c) LPC backward error.

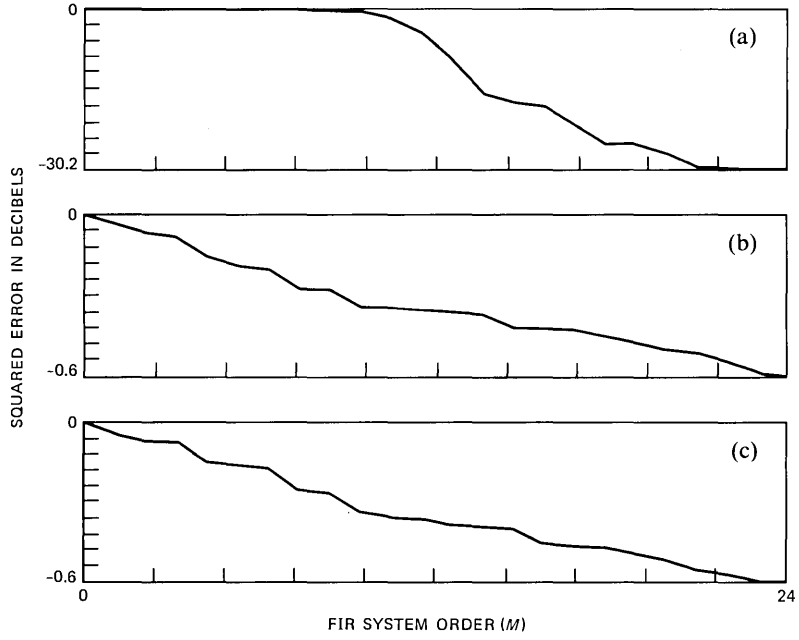


Fig. 18—Plots of squared prediction error versus FIR filter order, M, for a noise-signal excitation of a 25-point low-pass filter. (a) FIR error. (b) LPC forward error. (c) LPC backward error.

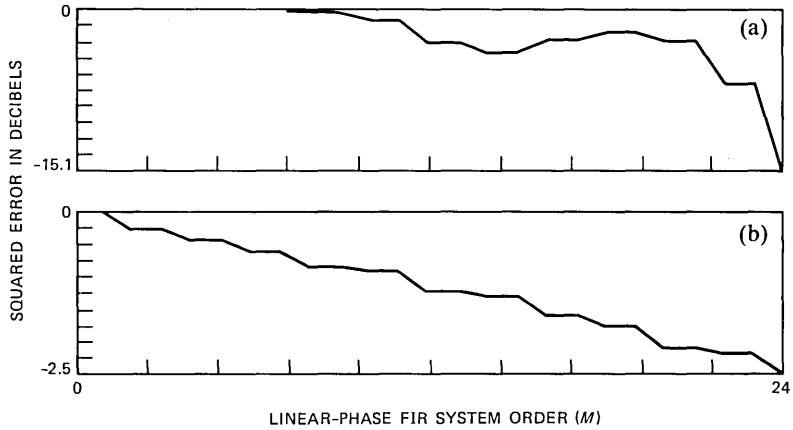


Fig. 19—Plots of squared prediction error versus FIR filter order, M, for the linear-phase algorithm with noise excitation of a 25-point low-pass filter. (a) Linear-phase FIR error. (b) LPC forward and backward error.

decreasing function of filter order. This point is illustrated in Fig. 19, which shows the normalized squared error versus filter order for a noise-signal input to the 25-point low-pass filter. The reader will note that for M in the neighborhood of 19, an increase in squared error occurs. The non-monotonicity of the squared error in the linear-phase algorithm is due to the fact that the algorithm is solving a linear *smoothing* problem (taking future and past samples) rather than a linear *prediction* problem (taking only past samples). Also, separate forward and backward linear prediction squared errors are *not* obtained in the linear-phase, fast algorithm; only a combined forward plus backward linear prediction squared error is obtained.

One final point is worth reiterating. In all cases where the FIR system to be estimated is known to be a linear-phase system, the linear-phase, fast, least-squares method is preferable to the general (nonlinear-phase) procedure in that it requires less computation and yields more accurate results.

V. SUMMARY

We have shown that the fast, least-squares FIR system identification algorithm originally proposed by Marple performs essentially perfectly for white-noise inputs for a wide range of FIR system responses and signal-to-noise ratios. For input signals whose spectrum is highly colored (e.g., speech signals) it was shown that the simple expedient of adding a low-level white noise to the input of the linear system provided a high degree of spectral stabilization and enabled the fast, least-squares algorithm to work well over a wide variety of conditions.

REFERENCES

1. L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR System Modeling and Identification in the Presence of Noise and with Band-Limited Inputs," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-26* (August 1978), pp. 319–33.
2. J. B. Allen and L. R. Rabiner, "Unbiased Spectral Estimation and System Identification Using Short-Time Spectral Analysis Methods," B.S.T.J., *58* (October 1979), pp. 1743–63.
3. L. R. Rabiner and J. B. Allen, "On the Implementation of a Short-Time Spectral Analysis Method for System Identification," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-28* (February 1980), pp. 69–78.
4. L. R. Rabiner and J. B. Allen, "Short-Time Fourier Analysis Techniques for FIR System Identification and Power Spectrum Estimation," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-27* (April 1979), pp. 182–92.
5. S. L. Marple, Jr., "Efficient Least Squares FIR System Identification," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-29* (February 1981), pp. 62–73.
6. E. Shichor, "Fast Recursive Estimation Using the Lattice Structure," B.S.T.J., *61* (January 1982), pp. 97–115.
7. M. Morf, B. Dickinson, T. Kailath, and A. Vieira, "Efficient Solution of Covariance Equations for Linear Prediction," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-25* (October 1977), pp. 429–33.
8. S. L. Marple, Jr., "Fast Algorithms for Linear Prediction and System Identification Filters with Linear Phase," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-30*, No. 6 (December 1982), pp. 942–53.

APPENDIX

Count of Operations to Solve Classical LSA Normal Equations Using Cholesky Decomposition

Solve $\Phi\alpha = \Psi$ for vector α.

Elements of matrix Φ:

$$\phi_{ij} = \sum_{t=M+1}^{N} x(t-j)x(t-i) \qquad \begin{matrix} 0 \le i \le M \\ 0 \le j \le i. \end{matrix}$$

Elements of vector Ψ:

$$\psi_i = \sum_{t=M+1}^{N} y(t)x(t-i) \qquad 0 \le i \le M.$$

In Step 1 we form elements of Φ and Ψ:

+	$\frac{1}{2}NM^2 + \frac{5}{2}NM + 2N - \frac{1}{2}M^3 - 3M^2 - \frac{9}{2}M - 2$
×	$\frac{1}{2}NM^2 + \frac{5}{2}NM + 2N - \frac{1}{2}M^3 - \frac{5}{2}M^2 - 2M$
Storage	$\frac{1}{2}M^2 + \frac{5}{2}M + 2$ (does not include x and y data).

In Step 2 we let $\Phi = VDV^T$. Solve for V and D:

+	$\frac{1}{6}M^3 + \frac{1}{2}M^2 + \frac{1}{3}M$
×	$\frac{1}{6}M^3 + \frac{1}{2}M^2 - \frac{2}{3}M$
Storage	M.

In Step 3 we solve for vector Y in $VY = \Psi$:

+	$\frac{1}{2}M^2 + \frac{1}{2}M$
×	$\frac{1}{2}M^2 + \frac{1}{2}M$
Storage	Store in place.

In Step 4 we solve for α in $DV^T\alpha = Y$

+	$\frac{1}{2}M^2 - \frac{1}{2}M$
×	$\frac{1}{2}M^2 + \frac{1}{2}M + 1$
Storage	Store in place.

In Step 5 we compute residual squared error $E_m = \sum_{t=M+1}^{N} y^2(t) - \alpha^T\Psi$

+	$N + 1$
×	$N + M + 3$
Storage	2.

As a result our total computations are:

Adds	$\frac{1}{2}NM^2 + \frac{5}{2}NM + 3N - \frac{1}{3}M^3 - \frac{3}{2}M^2 - \frac{25}{6}M - 1$
Multiplies	$\frac{1}{2}NM^2 + \frac{5}{2}NM + 3N - \frac{1}{3}M^3 - M^2 - \frac{2}{3}M + 4$
Storage	$\frac{1}{2}M^2 + \frac{7}{2}M + 4$ (not including x and y data).

The fast, least-squares algorithm requires

Adds	$2NM + 9M^2$
Multiplies	$2NM + 12M^2$(+$8M$ divides)
Storage	$7M + 20$ (not including x and y data),

which only counts terms of squared powers.

Running some various numbers for filter order M and data lengths N shows that the Cholesky method is more efficient than the "fast" algorithm when $N < 25$ and $M < 10$. If the Cholesky decomposition must be applied many times to determine the correct order to select, then the fast algorithm is more efficient since all lower solutions are obtained recursively.

CONTRIBUTORS TO THIS ISSUE

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, 1975–1982; American Bell, 1983—. At Bell Laboratories, Mr. Albanese did research in systems and components for lightwave communications and local-area networks. His current research interest is architectures and applications for local-area networks. He is Supervisor, Local Networks Application Group.

T. C. Chu, B.S., 1964, Cheng Kung University (Taiwan); M.S., 1967, Syracuse University; Ph.D., 1971 (Aerospace Engineering), Cornell University; Bell Laboratories, 1972—. Mr. Chu has worked on the physical design of T4M digital transmission system and the development of an optical fiber connector. Since 1981 he has been working on the design of undersea lightguide cable.

Patrick J. Fitzgerald, B.S. (Electronic Engineering Technology), Capitol Institute of Technology, 1973; Bell Laboratories, 1973—. Mr. Fitzgerald has been involved in various research studies including atmospheric microwave and optical propagation and digital modulation techniques. He has also been responsible for several computer-controlled, real-time, data-acquisition systems.

Nuggehally S. Jayant, B.Sc. (Physics and Mathematics), 1962, Mysore University; B.E., 1965, and Ph.D. (Electrical Communication Engineering), 1970, Indian Institute of Science, Bangalore; Research Associate at Stanford University, 1967–1968; Bell Laboratories, 1968—. Mr. Jayant was a Visiting Scientist at the Indian Institute of Science from January–March 1972 and August–October 1975. He has worked in the field of digital coding and transmission of waveforms, with special reference to speech communications. He is also editor of an IEEE Reprint Book, *Waveform Quantization and Coding.*

P. Leland Key, B.A., B.S. (Mechanical Engineering), 1960, Rice University; United States Navy, 1960–1964; M.S. (Engineering Science), 1964, George Washington University; Ph.D. (Materials Science), 1969, University of California, Berkeley; Bell Laboratories, 1969—. Mr. Key spent his naval service working on materials for nuclear reactors for naval ship propulsion. Since joining the Metallurgy and Engineering Science Department of Bell Laboratories, Mr. Key has been

involved in a broad range of materials activities from whisker growth on electrodeposited films to steels for pressure vessels. Currently, Mr. Key is studying reliability aspects of optical fibers. Member, American Society for Metals, Tau Beta Pi, Sigma Xi, Sigma Tau.

Bernard George King, B.S. (Electrical Engineering), 1944, USC; Ph.D., 1955, University of Wisconsin; Bell Laboratories, 1955—. Mr. King taught electrical engineering at USC from 1944 through 1948, and at the University of Wisconsin from 1948 through 1955. He is presently a Supervisor in the Communications Methods Research department, and since 1981 he has been the Editor of The Bell System Technical Journal.

Ernest D. Kolb graduated from Curtiss-Wright Technical Institute in 1939 and worked for the Wright Aeronautical Corporation in the Flight Research Section in Caldwell, New Jersey, until 1944. Mr. Kolb then joined the National Advisory Committee for Aeronautics in their Flight Research Section and was concerned with research on aircraft engines until entering the Air Force in 1945. He joined Bell Laboratories in 1951 and was initially involved in the Czochralski growth of single crystals of germanium and silicon from the melt. This work led to J. A. Burton, R. C. Prim, and W. P. Slichter's theory of the effective distribution coefficients of donor and acceptors in the growth of semiconductor materials. He then became involved in the hydrothermal synthesis of inorganic compounds and the growth of single crystals of compounds such as zinc oxide, zinc sulphide, yttrium iron garnet, lithium gallate, and alpha quartz. More recently, he has resumed work with semiconducting compounds with the growth of single crystals from the melt.

Robert A. Laudise, B.S. (Chemistry), 1952, Union College, Schenectady, New York; Ph.D. (Inorganic Chemistry), MIT, 1956; Bell Laboratories, 1956—. Mr. Laudise has done research on crystal growth and the preparation and properties of electronic materials. He is particularly interested in the physical chemistry of high-pressure synthesis and in pizeoelectric, semiconductor, and optical materials. He is presently Director of the Physical and Inorganic Chemistry Research Laboratory. He was a recipient of the Sawyer Award for Pizeoelectric Materials and of the International Crystal Growth Award in 1981. President, International Organization for Crystal Growth; member, National Academy of Engineering.

S. Lawrence Marple, Jr., B.A. and M.E.E., 1970 (Electrical Engineering), Rice University, Houston, TX; Engr. Deg., 1976 (Electrical

Engineering), Stanford University, Stanford, CA. From 1971 to 1977, Mr. Marple was employed by ArgoSystems, Inc., Sunnyvale, CA, where he was involved in hardware design for and digital processing of radar and communication signal analysis. From 1978 to 1979, he was a senior member of the technical staff of both Signal Science, Inc. and Advent Systems, Inc., of Mountain View, CA. From 1980 to 1982 he was with the McLean Operations of The Analytic Sciences Corporation (TASC), McLean, VA. He is presently with Schlumberger Well Services in Houston, Texas. His research interests include applications of fast, digital signal processing techniques to spectrum analysis, communications, radar, and well logging problems. Member, Tau Beta Pi.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D. (Electrical Engineering), The Massachusetts Institute of Technology; Bell Laboratories, 1962—. From 1962 through 1964, Mr. Rabiner participated in the cooperative plan in electrical engineering at Bell Laboratories. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975) and *Digital Processing of Speech Signals* (Prentice-Hall, 1978). Former President, IEEE, ASSP Society; former Associate Editor, ASSP Transactions; former member, Technical Committee on Speech Communication of the Acoustical Society, ASSP Technical Committee on Speech Communication; Member, IEEE Proceedings Editorial Board, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.

R. A. Semplak, B.S. (Physics), 1961, Monmouth College; Bell Laboratories, 1955—. Mr. Semplak's main research interest is in studies of atmospheric effects on micro- and millimeter-wave propagation. Currently, he is involved in experimental studies of astigmatic feeds. Member, Sigma Xi and Commission F of the International Union of Radio Science (URSI/USNC).

Earle E. Simpson, B.S. (Electronic Engineering), 1957, University of Maine; Western Electric, 1957—. Mr. Simpson was involved in engineering quartz crystal units, crystal filters, and crystal oscillators from 1963 to 1973. Since 1973, he has been responsible for the quartz-growing operation at the Merrimack Valley Works. He is a member of the American Association for Crystal Growth, and is a registered Professional Engineer in Massachusetts.

Man Mohan Sondhi, B.Sc. (Physics), Honours degree, 1950, Delhi University, Delhi, India; D.I.I.Sc. (Communications Engineering), 1953, Indian Institute of Science, Bangalore, India; M.S., 1955, Ph.D., 1957 (Electrical Engineering), University of Wisconsin, Madison, Wisconsin; Bell Laboratories, 1962—. Before joining Bell Laboratories, Mr. Sondhi worked for a year at the Central Electronics Engineering Research Institute, Pilani, India and taught for a year at the University of Toronto. At Bell Laboratories his research has included work on speech signal processing, echo cancellation, adaptive filtering, modeling of auditory and visual processes, and acoustical inverse problems. From 1971 to 1972 Mr. Sondhi was a guest scientist at the Royal Institute of Technology, Stockholm, Sweden.

PAPERS BY BELL LABORATORIES AUTHORS

COMPUTING/MATHEMATICS

Calderbank, R., **On Uniformly Packed [n,n-k,4] Codes Over GF (q) and a Class of Caps in PG(k-1,q).** J Lond Math 26(Oct):365–384, 1982.

Egan, D. E., Gomez, L. M., **Characteristics of People Who Can Learn to Use Computer Text Editors—Hints for Future Text Editor Design and Training.** P Am S Info 19:75–79, 1982.

Lagarias J. C., Odlyzko, A. M., Shearer, J. B., **On the Density of Sequences of Integers the Sum of Number Two of Which is a Square. 1. Arithmetic Progressions.** J Comb Th A 33(2):167–185, 1982.

Lund, R. E., Linnell, M. G., **Description and Evaluation of a Nested Cube Experimental-Design.** Comm Stat A 11(20):2297–2313, 1982.

ENGINEERING

Alferness, R. C., et al., **Efficient Single-Mode Fiber to Titanium Diffused Lithium-Niobate Wave-Guide Coupling for Gamma=1.32-μm.** IEEE J Q El 18(10):1807–1813, 1982.

Anthony, P. J., **Alteration of Diffusion Profiles in Semiconductors Due to p-n-Junctions.** Sol St Elec 25(10):1003–1009, 1982.

Boenke, M. M., Wagner, R. E., Will, D. J., **Transmission Experiments Through 101 km and 84 km of Single-Mode Fiber at 274 mbit/s and 420 mbit/s.** Electr Lett 18(21): 897–898, 1982.

Bosik, B. S., Kartalopoulos, S. V., **A Time Compression Multiplexing System for a Circuit Switched Digital Capability.** IEEE Commun 30(9):2046–2052, 1982.

Boyer, G. R., **Administration of New Feeder Technology.** IEEE Commun 30(9):2029–2033, 1982.

Byrne, T. P., Coburn, R., Mazzoni, H. C., Aughenbaugh, G. W., Duffany, J. L., **Positioning the Subscriber Loop Network for Digital Services.** IEEE Commun 30(9):2006–2011, 1982.

Campbell, L. W., Cline, R. R., Litofsky, B., Martz, L. M., Varma, G. K., **Maintenance Operations for the Circuit-Switched Digital Capability.** IEEE Commun 30(9):2053–2056, 1982.

Clarke, R. H., **Theoretical Performance of an Antireflection Coating for a Diode-Laser Amplifier.** Int J Elect 53(5): 495–499, 1982.

Cohen, L. G., Marcuse, D., Mammel, W. L., **Radiating Leaky-Mode Losses in Single-Mode Lightguides With Depressed-Index Claddings.** IEEE J Q El 18(10):1467–1472, 1982.

Coldren, L. A., Furuya, K., Miller, B. I., Rentschler, J. A., **Etched Mirror and Groove-Coupled GaInAsP/InP Laser Devices for Integrated-Optics.** IEEE J Q El 18(10):1679–1688, 1982.

Dale, O. B., Rubin, H., Vetter, R. W., **A Highly Distributed Mechanized Loop Testing System.** IEEE Commun 30(9):2038–2045, 1982.

Dutta, N. K., Wright, P. D., Nelson, R. J., Wilson, R. B., Besomi, P. R., **InGaAsP Laser With High T_0 (Letter).** IEEE J Q El 18(10):1414–1416, 1982.

Ebeling, K. J., Coldren, L. A., Miller, B. I., Rentschler, J. A., **Generation of Single-Longitudinal-Mode Sub-Nanosecond Light-Pulses by High-Speed Current Modulation of Monolithic Two-Section Semiconductor-Lasers.** Electr Lett 18(21):901–902, 1982.

Furuya, K., Miller, B. I., Coldren, L. A., Howard, R. E., **A Novel Deposit Spin Wave-Guide Interconnection (DSWI) for Semiconductor Integrated-Optics.** IEEE J Q El 18(10):1783–1789, 1982.

Joy, D. C., **Deconvolution for ELS Quantitation.** Ultramicros 9(3):289–294, 1982.

Khanarian, G., **Non-Linear Dielectric Effect of Flexible Polar Polymers.** Macromolec 15(5):1429–1432, 1982.

Korotky, S. K., Minford, W. J., Buhl, L. L., Divino, M. D., Alferness, R. C., **Mode Size and Method for Estimating the Propagation Constant of Single-Mode Ti-$LiNbO_3$ Strip Wave-Guides.** IEEE J Q El 18(10):1796–1801, 1982.

Lee, T. P., Burrus, C. A., Marcuse, D., Dentai, A. G., Nelson, R. J., **Measurement of Beam Parameters of Index-Guided and Gain-Guided Single-Frequency In-GaAsP Injection-Lasers.** Electr Lett 18(21):902–904, 1982.

Liu, P. L., **$LiNbO_3$ Wave-Guide Modulator With 1.2-μm Thick Electrodes Fabricated by Lift-Off Technique.** IEEE J Q El 18(10):1780–1782, 1982.

Liu, P. L., Lee, T. P., Burrus, C. A., Kaminow, I. P., Ko, J. S., **Observation of Transient Spectra and Mode Partition Noise of Injection-Lasers.** Electr Lett 18(21):904–905, 1982.

Marcuse, D., **Quantum-Mechanical Explanation of Spontaneous Emission K-Factor.** Electr Lett 18(21):920–922, 1982.

Minford, W. J., Korotky, S. K., Alferness, R. C., **Low-Loss Ti-$LiNbO_3$ Wave-Guide Bends at Gamma=1.3-μm.** IEEE J Q El 18(10):1802–1806, 1982.

Philen, D. L., White, I. A., Kuhl, J. F., Mettler, S. C., **Single-Mode Fiber OTDR—Experiment and Theory.** IEEE J Q El 18(10):1499–1508, 1982.

Prakash, M. N., Fam, A. T., **The Theory of Stable Minimal Partial Realizations—The Scalar Case (Letter).** IEEE Circ S 29(9): 646–647, 1982.

Santana, M. R., Yanizeski, G. M., **Bending Performance of Fiberoptic Cable Sheaths—Application of New Apparatus and Techniques.** J Mech Des 104(3):578–586, 1982.

Wyndrum, R. W., **Special Issue on Subscriber Loops—The Evolution of Electronic Loop-Systems in the 1980s (Editorial).** IEEE Commun 30(9):2005, 1982.

MANAGEMENT/ECONOMICS

Jovanovic, B., **Truthful Disclosure of Information.** Bell J Econ 13(1):36–44, 1982.

Perry, M. K., **Oligopoly and Consistent Conjectural Variations.** Bell J Econ 13(1):197–205, 1982.

Sharkey, W. W., **Suggestions for a Game-Theoretic Approach to Public Utility Pricing and Cost Allocation.** Bell J Econ 13(1):57–68, 1982.

PHYSICAL SCIENCES

Anderson, P. W., **The 1982 Nobel-Prize in Physics.** Science 218(4574):763–764, 1982.

Antler, M., **Fretting of Electrical Contacts—An Investigation of Palladium Mated to Other Materials.** Wear 81(1):159–173, 1982.

Bachelet, G. B., Hamann, D. R., Schluter, M., **Pseudopotentials That Work—From H to Pu.** Phys Rev B 26(8):4199–4228, 1982.

Bally, J., **Energetic Activity in a Star-Forming Molecular Cloud Core—A Disk Constrained Bipolar Outflow in NGC-2071.** Astrophys J 261(2):558–568, 1982.

Bjorkholm, J. E., Liao, P. F., Wokaun, A., **Distortion of On-Resonance 2-Photon Spectroscopic Line-Shapes Caused by Velocity-Selective Optical-Pumping.** Phys Rev A 26(5):2643–2655, 1982.

Bokor, J., Freeman, R. R., Cocke, W. E., **Photo-Autoionization Pumped Ba Ion Laser.** Aip Conf PR1982(90):153–162, 1982.

Bondybey, V. E., English, J. H., Shiley, R. H., **Pyridine Radical Cation and Its Fluorine Substituted Derivatives.** J Chem Phys 77(10):4826–4831, 1982.

Bowden, M. J., Thompson, L. F., Robinson, W., Biolsi, M., **Thermal-Degradation of Poly(1-Butene Sulfone).** Macromolec 15(5):1417–1422, 1982.

Bridges, T. J., Chraplyvy, A. R., Bergman, J. G., Hart, R. M., **Broad-Band Infrared Generation in Liquid-Bromine-Core Optical Fibers.** Optics Lett 7(11):566–568, 1982.

Broughton, J. Q., Gilmer, G. H., Jackson, K. A., **Crystallization Rates of a Lennard-Jones Liquid.** Phys Rev L 49(20):1496–1500, 1982.

Brown, W. L., Lanzerotti, L. J., Johnson, R. E., **Fast Ion-Bombardment of Ices and Its Astrophysical Implications.** Science 218(4572):525–531, 1982.

Chan, K., **A Simple Proof That the Unstable (CO-)Homology of the Brown-Peterson Spectrum Is Torsion Free.** J Pure Appl 26(2):155–157, 1982.

Chin, A. K., Zipfel, C. L., Dutt, B. V., **The Reliability of Schottky-Barrier Restricted GaAs/GaAlAs LEDs.** Jpn J A P 1 21(9):1308–1312, 1982.

Chu, S., Wong, S., **Pulse-Propagation in an Absorbing Medium—Reply (Letter).** Phys Rev L 49(17):1293, 1982.

Cohen, R. L., Feldman, L. C., West, K. W., Kincaid, B. M., **Displacement Around a Dissolved Impurity Atom in a Metal-Sn in Cu.** Phys Rev L 49(19):1416–1419, 1982.

Corruccini, L. R., **Second Sound in Superfluid ^{3}He.** Physica B&C 110(1–3):1590–1595, 1982.

Cross, M. C., **Magnetic-Properties of Solid ^{3}He—What Do We Know and What Do We Learn.** Physica B&C 110(1–3):1796–1804, 1982.

DuBois, L. H., **Vibrational-Spectra of Atomic Adsorbates—Carbon, Oxygen, and Sulfur on Rh(100).** J Chem Phys 77(10):5228–5233, 1982.

Even, U., Magen, J., Jortner, J., Friedman, J., Levanon, H., **Isolated Ultracold Porphyrins in Supersonic Expansions. 1. Free-Base Tetraphenylporphyrin and Zn-Tetraphenylporphyrin.** J Chem Phys 77(9):4374–4383, 1982.

Even, U., Magen, J., Jortner, J., Friedman, J., **Isolated Ultracold Porphyrins in Supersonic Expansions. 2. Zn-Tetrabenzoporphyrin.** J Chem Phys 77(9):4384–4390, 1982.

Fisher, D. S., **Shear Moduli and Melting Temperatures of Two-Dimensional Electron Crystals—Low-Temperatures and High Magnetic-Fields.** Phys Rev B 26(9):5009–5021, 1982.

Fox, D. L., Golding, B., Haemmerl, W. H., **Optically Induced Gap in the Atomic Tunneling Spectrum of AS_2S_3 Glass.** Phys Rev L 49(18):1356–1360, 1982.

Freeman, R. R., Jopson, R. M., Bokor, J., **Generation of Light Below 100nm in Hg Vapor.** AIP Conf PR1982(90):422–430, 1982.

Friedman, J. M., Rousseau, D. L., Ondrias, M. R., **Time-Resolved Resonance Raman Studies of Hemoglobin (Review or Bibliog.).** Ann R Ph Ch 33:471–491, 1982.

Geerk, J., Gurvitch, M., McWhan, D. B., Rowell, J. M., **Electron-Tunneling into Nb/Al Multilayers and into Nb with Al Overlayers.** Physica B&C 110(1–3):1775–1784, 1982.

Gibson, J. M., Bean, J. C., Poate, J. M., Tung, R. T., **Direct Determination of Atomic-Structure at the Epitaxial Cobalt Disilicide on (111) Si Interface by Ultrahigh Resolution Electron-Microscopy.** Appl Phys L 41(9):818–820, 1982.

Greywall, D. S., Paalanen, M. A., **Second Sound in Spin-Polarized ^{3}He-^{4}He Solutions.** Physica B&C 110(1–3):1575–1582, 1982.

Haque, C. A., Antler, M., **Atmospheric Corrosion of Clad Palladium and Palladium Silver Alloys—Film Growth and Contamination Effects.** Corros Sci 22(10):939+, 1982.

Hohenberg, P. C., **Critical Phenomena in -^{4}He.** Physica B&C 110(1–3):1436–1446, 1982.

Hutton, R. S., Roth, H. D., Schilling, M. L., Trozzolo, A. M., Leslie, T. M., **Magnetic-Field Dependence of F-19 Nuclear-Spin Polarization—The System Tetrafluorobenzoquinone Tetrafluorohydroquinone.** J Am Chem S 104(22):5878–5883, 1982.

Johari, G. P., Goodby, J. W., **Dielectric Relaxations in a Supercooled Liquid and Glassy Smectic Phase.** J Chem Phys 77(10):5165–5172, 1982.

Johari, G. P., **Effect of Annealing on the Secondary Relaxations in Glasses.** J Chem Phys 77(9):4619–4626, 1982.

Jopson, R. M., Freeman, R. R., Bokor, J., **Laser-Produced Plasmas in Hg—A Source of Pulsed XUV Radiation.** AIP Conf PR1982(90):303–304, 1982.

Kuo, C. Y., Vieira, M. M. F., Patel, C. K. N., **Infrared Spectroscopy of Phonons in Hexagonal-Phase Solid Hydrogen.** Phys Rev L 49(17):1284–1287, 1982.

Lee, P. A., et al., **Magnetoresistance of Weakly Disordered Electrons.** Phys Rev B 26(8):4009–4012, 1982.

Lindsay, S. M., Halawith, B., Patterson, G. D., **Observation of Hypersonic Shear-Waves in Polystyrene and Poly(Methyl Methacrylate) by Brillouin-Scattering.** J Pol Sc Pl 20(11):583–588, 1982.

Littlewood, P. B., Varma, C. M., **Amplitude Collective Modes in Superconductors and Their Coupling to Charge-Density Waves.** Phys Rev B 26(9):4883–4893, 1982.

McBrierty, V. J., Douglass, D. C., Wudl, F., Aharon-Shalam, E., **Nuclear-Resonance and Relaxation in Ditetramethyltetraselenafulvalenium Salts.** Phys Rev B 26(9):4805–4809, 1982.

McCrory, J. C., Rosamilia, J. M., **Differential Pulse Polarography of Germanium(IV), Tin(IV), Arsenic(V), Antimony(V), Selenium(IV) and Tellurium(VI) at the Static Mercury Drop Electrode in Catechol Perchlorate-Media.** Analyt Chim 142(OCT):231–238, 1982.

Miller, T. A., **Light and Radical Ions (Review or Bibliog.).** Ann R Ph Ch 33:257–282, 1982.

Mitchell, J. W., **Purification of Analytical Reagents.** Talanta 29(11):993–1002, 1982.

Nakada, Y., **Metallurgical Characterization of Arc-Conditioned Carbon Coating on Copper Electrodes.** Thin Sol Fi 95(2):185–193, 1982.

Nassau, K., Chadwick, D. L., **Class-Forming Systems Involving GeO_2 With Bi_2O_3, Ti_2O, and PbO.** J Am Ceram 65(10):486–491, 1982.

Osheroff, D. D., **Spin Ordering in Solid ^3He—A Current Perspective.** Physica B&C 110(1–3):1461–1473, 1982.

Perriere, J., Siejka, J., Chang, R. P. H., **Observation of Short-Range Oxygen Migration and Oxygen-Exchange During Low-Temperature Plasma Anodization of Silicon Through Thin ZrO_2 Films.** Thin Sol Fi 95(4):309–314, 1982.

Sammon, M. J., **Numerical Three-Dimensional Relaxation of Liquid-Crystal Director Fields.** Molec Cryst 89(1–4):305–317, 1982.

Santoro, A., Cava, R. J., Murphy, D. W., Roth, R. S., **Use of the Pearson Type-VII Distribution in the Neutron Profile Refinement of the Structures of $LiReO^3$ and Li_2ReO_3.** AIP Conf PR1982(89):162–165, 1982.

Shepp, L. A., **Citation Classic—The Fourier Reconstruction of a Head Section.** CC/Eng Tech 1982(46):14, 1982.

Turner, D. R., **Modern Techniques for Electroplating Gold Contacts.** Thin Sol Fi 95(2):143–149, 1982.

Vanbenthem, M. H., Gillispie, G. D., Haddon, R. C., **Intramolecular Hydrogen-Bonding. 2. Enormous Deuterium-Isotope Effect on the Phosphorescence of 6-Hydroxybenzanthrone (Letter).** J Phys Chem 86(22):4281–4283, 1982.

Venkatesan, T., **Species and Deposition Angle Dependence of Ion-Beam Induced Densification of Germanium Selenide Films.** Appl Phys L 41(9):839–841, 1982.

Walstedt, R. E., **Numerical-Simulation Studies of RKKY Spin-Glasses.** Physica B&C 110(1–3):1924–1935, 1982.

White, J. C., Henderson, D., **Threshold and Dispersion Effects in the Anti-Stokes Raman Laser.** Optics Lett 7(11):517–519, 1982.

Young J. S., et al., **Physical Conditions and Carbon-Monoxide Abundance in the Dark Cloud B5.** Astrophys J 261(2):513–531, 1982.

SOCIAL AND LIFE SCIENCES

Finkel, R. A., Fishburn, J. P., **Parallelism in Alpha-Beta Search.** Artif Intel 19(1):89–106, 1982.

Patel, D. J., **Antibiotic-DNA Interactions—Intermolecular Nuclear Overhauser Effects in the Netropsin-d (C-G-C-G-A-A-T-T-C-G-C-G) Complex in Solution.** P Nas Biol 79(21):6424–6428, 1982.

Seldner, M., Uson, J. M., **The Cross-Correlation of the Zwicky and Shane-Wirtanen Catalogs of Galaxies.** Astrophys J 261(2):L65–L69, 1982.

CONTENTS, APRIL 1983

THE BELL SYSTEM TECHNICAL JOURNAL is abstracted or indexed by *Abstract Journal in Earthquake Engineering, Applied Mechanics Review, Applied Science & Technology Index, Chemical Abstracts, Computer Abstracts, Current Contents/ Engineering, Technology & Applied Sciences, Current Index to Statistics, Current Papers in Electrical & Electronic Engineering, Current Papers on Computers & Control, Electronics & Communications Abstracts Journal, The Engineering Index, International Aerospace Abstracts, Journal of Current Laser Abstracts, Language and Language Behavior Abstracts, Mathematical Reviews, Science Abstracts (Series A, Physics Abstracts; Series B, Electrical and Electronic Abstracts; and Series C, Computer & Control Abstracts), Science Citation Index, Sociological Abstracts, Social Welfare, Social Planning and Social Development,* and *Solid State Abstracts Journal.* Reproductions of the Journal by years are available in microform from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.


Bell System