

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XLVI

NOVEMBER 1967

NUMBER 9

Copyright © 1967, America Telephone and Telegraph Company

An Automatic Transmission Measuring System for Telephone Trunks

By J. F. INGLE, J. J. KOKINDA, and G. E. McLAUGHLIN

(Manuscript received July 19, 1967)

An Automatic Transmission Measuring System (ATMS) has been designed to provide means for making rapid and accurate transmission measurements on telephone trunks. The system consists of a control unit (director) in one office and one or more responding units (responders) at distant locations. Trunk selection (not treated in detail in this paper) is accomplished either by a specially designed test frame in electromechanical offices or by special programming in electronic offices.

All measurement sequences are under command of the director which in turn receives its information from a teletypewriter tape or punched cards. New measurement techniques utilized in ATMS which permit rapid and accurate measurements are discussed. System accuracy of ± 0.1 dB for loss and ± 1 dB for noise is achieved using these techniques. Total measurement time (excluding trunk seizure and printout time) for loss measurements in both directions and noise at both ends is less than five seconds.

Although the ATMS is presently capable of making only loss and noise measurements, additional measurements can be added conveniently because of its modular design and construction.

Where responders are not installed at the distant offices, other kinds of existing Bell System transmission test lines may be utilized by the director to make whatever measurements the test line permits.

Two schemes are described whereby measurements may be made on trunks between two remote central offices and the results sent to a controlling director in a third office.

I. INTRODUCTION

The problem of trunk maintenance in the Bell System is magnified by the number of trunks which must be considered. A typical central office has more than a thousand trunks and there are about 2.7 million trunks in the Bell System. Proper maintenance of these trunks requires routine measurements at monthly or shorter intervals. In addition, at least four different transmission measurements are required to insure proper operation.

In view of the large number of measurements required, fully mechanized testing appears not only economical but necessary. The Automatic Transmission Measuring System (ATMS) discussed in this paper was developed to meet this need.

II. BACKGROUND AND PRIOR ARRANGEMENTS

In order to appreciate some of the intricacies involved in automatically measuring telephone trunks, a brief description of the telephone plant is in order. Fig. 1 is a simplified illustration of the telephone plant. A connection between customers is made up of two customer loops and 0, 1, 2 or more trunks. The loops shown connecting customers to the central offices are generally passive, i.e., without amplifiers. The trunks connecting central offices, on the other hand, frequently have active devices associated with them.

Fig. 2 depicts the make-up of a hypothetical trunk. A particular trunk contains some of the elements shown. They will be discussed only to the extent to which they affect measurements.

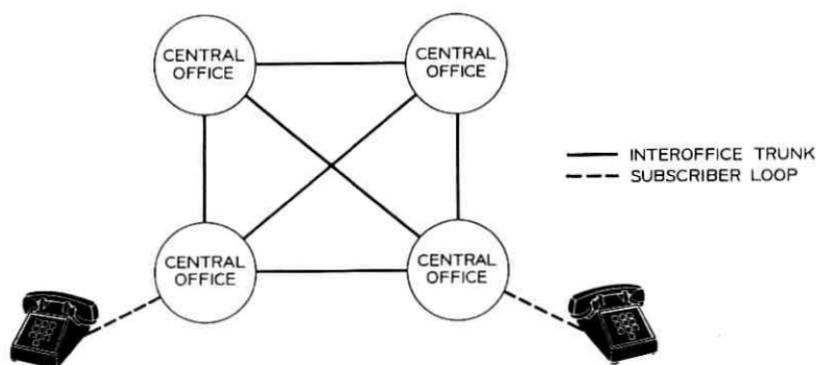


Fig. 1—Simplified telephone plant.

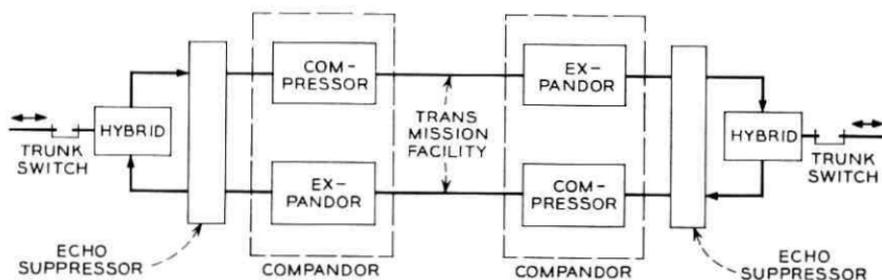


Fig. 2 — Hypothetical trunk.

(i) Systems with Amplifiers

Trunks utilizing carrier systems or hybrid-type repeaters complicate transmission measurements since the transmission in the two directions is affected by different elements. Thus, measurements must be made at both ends of the trunk.

(ii) Echo Suppressors

Echo suppressors, devices utilized in long haul trunks, affect remote automatic measurements because they prohibit simultaneous transmission in both directions.

(iii) Compandors

Compressors and expandors (known collectively as compandors) if functioning perfectly, are of no concern to transmission measurements. However, imperfect compandor action implies that the insertion loss of the trunk at one transmitting level differs from that at another transmitting level.

The use of active devices demands closer surveillance of the trunk to detect changes in their characteristics due to aging, maladjustment, etc. Until a few years ago, an operator using cords made the connection between loops and trunks. Since the operator was required to complete any call, she could perform rudimentary tests to determine the suitability of the connection. By listening to the person placing the call and to the person to whom the call was placed, the operator could note excessive loss or noise if present, and if necessary, establish an alternative connection. With the introduction of Direct Distance Dialing (DDD), even this minimal transmission check has been lost.

Trunks may be classed as incoming, outgoing, or two-way depending upon whether the trunk may be seized only from the distant office,

only from the near-end office, or from either office. These terms do not imply the transmission is limited to a single direction. It does mean that (in the case of manual measurements) if a transmission test is to be made on incoming trunks, the distant office must be requested to originate the test call.

When a connection has been established, manual effort is required at each end of the trunk if measurements at both ends are desired. Perhaps the most serious disadvantage of such manual measurements is the time involved in coordinating the efforts at each end and the number of tests involved.

Semiautomatic measurements are defined for the purposes of this paper as two-way measurements which are made manually using automatic far-end equipment. In this case, manual effort is required only at one end to make the measurements in both directions with the automatic far-end equipment. More will be said of this later.

Automatic transmission measuring systems for measuring telephone trunks are not new. At least two other systems have been developed prior to ATMS. One of these, which was developed by the Bell Telephone Laboratories in the early 1950's, is known as the automatic transmission test and control circuit (ATTC).¹ The other system was developed in the early 1960's by the Swedish Company Telefonaktiebolaget L. M. Ericsson.^{2, 3} Both systems utilize the measurement technique illustrated in Fig. 3. After amplification and rectification, the resulting voltage is compared with a fixed reference voltage. The attenuator is then adjusted in discrete steps until the output voltage is equal to the reference voltage within the limits of the attenuator's step granularity.

Here the similarity between the two systems ends. The far-end equipment associated with the ATTC adjusts a second set of attenuators to equal the near-to-far loss. The far-end equipment then sends a test tone first without and then with the additional loss stored in the second attenuator. The near-end equipment measures the far-to-near loss under both conditions and then computes the loss of the trunk

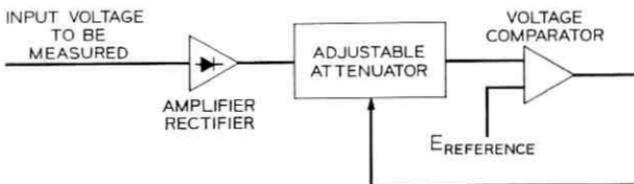


Fig. 3 — A transmission measurement scheme.

in the near-to-far direction as the difference of the levels of the two received tones. The Ericsson system sends the information contained in the state of the relays controlling the adjustable attenuator to the near-end in digital form by means of a multifrequency signaling system.

Each of these systems has its advantages and disadvantages. One of the attractions of the ATTC system is its adaptability to semi-automatic measurements. The results can be decoded by manual measurements. No special equipment is needed. On the other hand, imperfect companders and other nonlinearities in the trunk being measured can cause appreciable errors. The system used by Ericsson circumvents this problem by coding. Specialized decoding equipment, however, must be used at the near-end for both automatic and semi-automatic measurements.

Both the Ericsson system and the ATTC system are comparatively slow due to the process of adjusting discrete step attenuators. Both also make only noise checks, not noise measurements. In addition, the Ericsson system does not contain some of the self-checking features provided in the ATTC.

III. REQUIREMENTS

The following is a summary of the requirements upon which the ATMS design is based.

3.1 *Operational Requirements*

- (i) The system must make measurements at both ends of trunks.
- (ii) Measurements at both ends must be made automatically without manual assistance.
- (iii) Measurements at the far-end should be controlled by signals sent by the near-end equipment.
- (iv) Measurements should be made of insertion loss in both directions of transmission and background noise at both ends of trunks. Noise measurement results should indicate the amount of "background" noise present such as thermal noise, crosstalk, steady tone, etc. These are kinds of noise most disturbing to a human listener. Impulse noise, although its effects may be very serious on trunks used for data transmission, should be included only to the extent that it disturbs the human listener.
- (v) Measurements should be made as rapidly as practical.
- (vi) All measurement results should be made available at the near-end.

(vii) Modular construction should be used to allow for future expansion to include other tests and to facilitate maintenance.

(viii) The near-end equipment should be capable of making tests to existing far-end measuring equipment to whatever extent practical.

(ix) Measurements should be possible on trunks where echo suppressors prevent simultaneous transmission in both directions.

(x) All measurement results of loss and noise should be displayed in logarithmic (decibel) units as deviations from reference values.

3.2 Accuracy and Range Requirements

(i) Overall system accuracy should be ± 0.1 dB for loss measurements and ± 1 dB for noise measurements.

(ii) Loss measuring circuit should accept +5 to -15 dBm signals.

(iii) Noise measurement range should extend from +15 to +50 dBrnC.

IV. GENERAL SYSTEM DESCRIPTION

4.1 System Functions

In discussing a system which will automatically test the transmission performance of all outgoing trunks in a telephone office, it is necessary to include associated equipment both in the originating and terminating offices. In addition to the measurement, the functions of gaining access to the desired trunk and of establishing a connection to the far-end test equipment must be considered. The complete measurement process can be broken down into a number of relatively simple steps which can then be related to specific equipment. These steps are:

(i) accept priming information on trunks to be tested: i.e., tests to be made, the transmission requirements of the trunks, and so on;

(ii) seize the trunk to be tested and dial up the far-end test equipment;

(iii) coordinate the operations of the near-end and far-end test equipment;

(iv) make transmission measurements at both ends of the trunk;

(v) transmit the far-end test results to the near-end;

(vi) display both the near-end and far-end results, in appropriate units, at the near-end; and

(vii) release the connection.

In addition, to increase the reliability of the system operation, two more steps may be added. These are:

(viii) repeat the test on a trunk when a transmission impairment is detected to determine if it is momentary or continual;

(ix) periodically make internal system checks of the measuring circuits to insure accuracy.

4.2 *System Equipment*

The overall system is comprised essentially of four units: the director, the responder, an automatic trunk test frame, and a test line (designated the 105-type test line). The director is used in conjunction with the automatic trunk test frame in the office in which the tests (and the trunks) originate (hereafter referred to as the near-end office). The responder, which is accessed through the 105-type test line, is far-end test equipment.

4.2.1 *ATMS Director*

The director was designed primarily to make measurements with the aid of the far-end responder and its associated test line. However, it is also capable of making limited measurements with other far-end arrangements. The director performs the following functions:

- (i) Receives instructions from the automatic trunk test frame.
- (ii) Sends commands to the responder.
- (iii) Send test tones.
- (iv) Provides a termination for far-end noise measurements.
- (v) Receives data signals from a responder.
- (vi) Makes far-to-near trunk loss and near-end noise measurements.
- (vii) Converts the trunk loss and noise measurements made by the director and data signals from a responder into numerical readings and cues (indications that limits have been exceeded).
- (viii) Provides these results to the automatic trunk test frame which causes them to be printed on the readout device.
- (ix) Performs a self-check of its operation and a check on the operation of the responder when commanded to do so.

4.2.2 *ATMS Responder*

In addition to the basic measurement functions, which are similar to those of the director, the responder receives commands from the director over the trunk under test, converts the received level of the test tone or noise into data signals, and transmits the data signals to the director.

4.2.3 *Automatic Trunk Test Frames*

Several different automatic trunk test frames are used in the various types of telephone offices. ESS offices have special programs which provide the equivalent of a test frame. These test frames provide arrangements for seizing the trunks to be measured and for pulsing forward the codes of various test lines at the distant end of the trunk. Information necessary for the director to make appropriate transmission loss and noise measurements and to evaluate the results is also supplied by the test frame. In addition, it provides facilities for printing the measurement results.

Another test frame function is to maintain trunk supervision, which includes the ability to send and/or receive on-hook, off-hook, busy, and reorder signals. In addition, most test frames make operational tests on trunks (indeed, this may be their primary function) in conjunction with an operational test line in the terminating office. These tests check the trunk's ability to pass supervision and signaling and are made independently of the transmission tests.

4.2.4 *105-Type Test Line*

The responder must be accessed through a 105-type test line. This test line provides holding and supervision, connects the responder through the switching system to the trunk being measured, and supplies transmission measuring information to the responder. A group of these test lines provides a parking arrangement which enables incoming calls to wait and be served in turn if the responder is engaged.

4.3 *ATMS Operation*

A typical transmission measurement setup using the ATMS is shown in Fig. 4. A near-end connection to the trunk is made through the office switching equipment or special test connectors. The code of the far-end test line (in this case a 105-type test line) is then pulsed forward, and the distant switching machine makes the connection. The 105-type test line terminates in a responder. The automatic trunk test frame feeds the test conditions and trunk transmission requirements to the director, connects the trunk to the director, and instructs the director to perform certain measurement sequences.

As the director makes the measurements, it provides measurement data to the automatic trunk test frame which records the results on a Teletype printout or other readout device. At the end of the test sequence, the trunk and measuring equipment are released and control

either during the measurement or during data transmission of test results from the responder to the director.

It was mentioned earlier that a self-checking feature is desirable to insure system accuracy and reliability. This is true particularly in a system like ATMS, for hundreds, even thousands, of trunks may be tested without manual intervention.

The ATMS makes a self-check of both the director and responder when it advances to a new responder (a self-check command is included in the priming information supplied by the test frame). This self-check includes practically all of the measuring and data transmission circuits.

A loss and noise deviation register panel is available, which accumulates statistical data on the measurements made by the ATMS. This is accomplished by dividing the complete measurement range into intervals and counting, or "scoring," the number of measurements that fall within each interval. This provides information which can be used in compiling measurement results statistics.

Modular construction facilitates both maintenance and addition of new measurements.

4.5 *Other Transmission Test Lines*

Before completing a description of the ATMS, it is worthwhile to mention other transmission test lines that are currently in use. The ATMS director was designed with the capability to test to a number of existing Bell System transmission test lines and make all the measurements that are within the test line's capability. The exact measurements to be performed are dependent upon the capability of the far-end test line and the test requirements. Table I summarizes Bell System Test Lines to which the ATMS will test and the measurement capability they provide.

V. MEASUREMENT TECHNIQUE

In a sense the ATMS may be considered a very specialized digital voltmeter. However, the measurement technique employed by the ATMS must provide a number of features not generally imposed upon a digital voltmeter. Because measurements cannot be performed and the results transmitted simultaneously on the same trunk (a condition precluded by the use of echo suppressors on some trunks) some storage element must be used at least in the far-end equipment. Results of measurements made at the far-end must be in a form suitable for transmission to the near-end by a means essentially inde-

TABLE I—BELL SYSTEM TRANSMISSION TEST LINES

Test line type*	Measurements		Description
	Loss	Noise	
100†	Far-to-near	Near-end	5 seconds of milliwatt followed by quiet termination
102	Far-to-near	No	Milliwatt, interrupted at 10-second intervals
104	Both ways	Near-end	Transmission measurement and noise checking circuit
105	Both ways	Both ends	ATMS responder

* This includes both toll test lines accessed by 10x codes and local and tandem types accessed by other than 10x codes.

† This test line will be available soon.

pendent of the transmission characteristics of the trunk over which the results are sent. Finally, all loss and noise measurement results should be presented to the user in logarithmic units to conform with the universal use of the decibel in the Bell System.

One of the simplest schemes which accomplishes the above is used in the ATMS and takes advantage of the logarithmic character of an RC discharge. After the signal to be measured is amplified, rectified, and filtered, a capacitor is charged to the resulting voltage. The capacitor provides a needed means for temporary storage of the measurement. After the capacitor is charged, relay contacts remove the charging source, leaving the capacitor with a charge proportional to the signal voltage being measured.

The length of time that the capacitor is connected to the amplifier-rectifier is about 0.4 second in a typical case for ATMS. If the trunk being measured has a so-called beating problem (a condition associated with some carrier systems which causes the gain to fluctuate slowly with time) and if the beating period is long, an ATMS measurement will give a result equivalent to the average value of the received voltage during that time.

The capacitor is now removed from the amplifier-rectifier for a brief period during which it will retain its charge. A resistor is then connected across the capacitor and the charge on the capacitor decays in a known exponential manner as illustrated in Fig. 5. A voltage comparator monitors the voltage on the capacitor and generates a pulse from the time the capacitor begins to discharge until the voltage on the capacitor reaches the reference voltage E_R . The duration of the

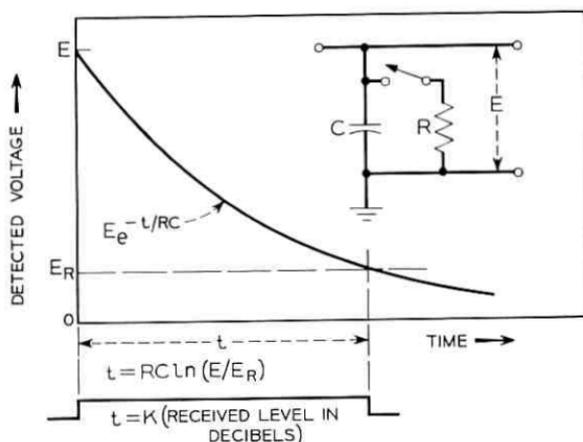


Fig. 5 — Voltage-to-time conversion.

pulse so generated is

$$t = RC \ln (E/E_R), \quad (1)$$

where E is the initial voltage on the capacitor C and R is the value of the discharging resistor.

This time interval is proportional to the signal voltage (measured in decibels) with the addition of a constant which is composed of known factors. By means of a frequency shift data transmitter and receiver the pulse length information can be sent to the director.

Errors at the director in determining the length of the pulse sent from a responder may occur due to large impulse noise and the finite bandwidth of the trunk over which the information is sent. Impulse noise, if sufficiently large, may produce "holes" or additional error pulses in the received pulse. Either of these conditions will cause errors in determining the length of the pulse. Finite bandwidth implies that the beginning and ending of the received pulse cannot be precisely determined due to the finite rise and fall times of the pulse. Both the effect of impulse noise and the effect of bandwidth limitations can be mitigated by lengthening the RC time constant which in effect increases the period of the pulse transmitted per dB. It is possible then, at the expense of increased time spent in measuring, to achieve any practical degree of error desired.

It remains now to determine the length of the pulse generated in order to ascertain the magnitude of the voltage being measured. Examination of (1) shows that equal intervals of time represent equal decibel increments. Therefore, a gated oscillator and counter circuit

such as shown in Fig. 6 may be used to determine how many decibels above the reference voltage the voltage on the capacitor was before discharge. For example, the oscillator's frequency may be set such that one cycle is equivalent to 1 dB. If then the voltage placed on the capacitor is 10 dB above the reference voltage, 10 cycles of the oscillator output will be gated to the counter. This is shown pictorially in Fig. 7. As a practical matter it can be shown that if the oscillator and gate are not synchronized, a one count ambiguity can occur. Since the exact time a pulse arrives cannot be arranged to correlate with the phase of the oscillator and since turning on a precision oscillator and obtaining full accuracy instantaneously is very difficult, synchronization is not feasible. An alternative solution used in ATMS is to employ a free-running oscillator whose frequency is much higher (36:1 in the case of ATMS) than necessary. Then a gated divider is used which divides the oscillator frequency down to the desired frequency and gates the output to the counter. In this way the ambiguity is reduced by the ratio of the oscillator frequency to the gating frequency. In the case of an ATMS loss measurement, for example, where a tenth of a dB is represented by 2 milliseconds the ambiguity is reduced to $2/36$ milliseconds which corresponds to 0.0028 dB.

The counter output in Fig. 6 is the difference in dB between the unknown voltage and the reference voltage, provided the counter was set initially to zero. In practice, the quantity of interest is the difference between the measured voltage and the expected voltage. This can be obtained by the use of a presettable reversible counter. Fig. 7 shows its operation. Assume that the readout is in dB's and that the expected voltage is 10 dB above the reference voltage. From this information the counter is preset to 10 and set to count down. Further, assume that the measured voltage is 10 dB above the reference voltage. When the capacitor discharges, 10 cycles of the oscillator output are gated into the counter causing it to count down to 0. The resulting

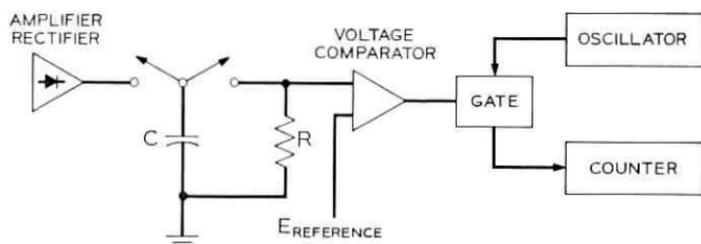


Fig. 6—Block schematic for measurement of time interval.

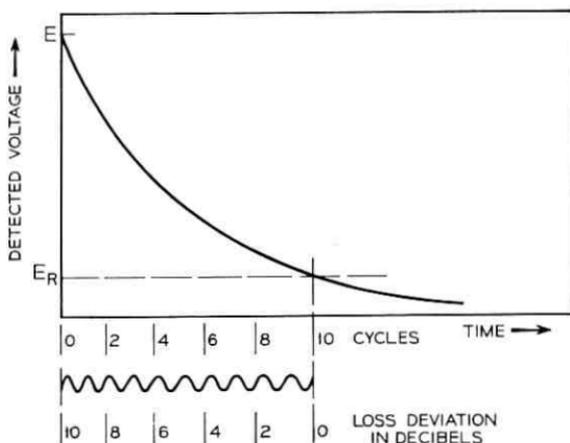


Fig. 7 — Loss deviation computation.

digital display on the counter will therefore be 0, corresponding to the difference between the measured and expected voltages.

If, however, the voltage on the capacitor was 11 dB above the reference voltage, the counter would count down from 10, through 0, reverse and count up to 1. The difference or deviation is then read out as 1 dB; the reversal of the counter indicates that the measured voltage exceeded the reference voltage.

To change the precision to which the results are displayed, it is only necessary to change either the oscillator frequency or the RC discharge time constant. In the ATMS, loss measurements are displayed to the nearest 0.1 dB and noise measurements are displayed to the nearest 1 dB. The change is made by decreasing the RC time constant by a factor of 10 during noise measurements.

VI. NOISE MEASUREMENT

The previous section discussed how a loss measurement was made and the results displayed. The ATMS noise measurement is discussed in more detail because it is the first widespread application in which a noise reading made in a fraction of a second is taken as a measure of the disturbing effect of noise to a telephone customer.

6.1 General

The fundamental objective of message circuit noise measurement is to give the same reading on various kinds of noise that are judged to be equally interfering to a telephone customer. The accepted noise measuring set in the Bell System for measuring message circuit noise

is the 3A Noise Measuring Set.⁴ The ATMS noise measurement circuit will give approximately the same results as a 3A Noise Measuring Set.

6.2 Frequency Weighting

The ATMS noise measurement circuit employs the same C-message weighting filter as the 3A set. This characteristic was determined during tests⁴ in which listeners were asked to adjust the loudness of 14 different frequencies between 180 and 3500 hertz until the sound of each was judged to be equal in annoyance to a 1000-hertz reference tone. The results of these tests were averaged at each frequency, combined and smoothed to obtain the C-message weighting as shown in Fig. 8.

6.3 Quasi-rms Detector Circuit

The ATMS quasi-rms detector employs the same kind of detector as the 3A Noise Measuring Set. The appendix of Ref. 4 explains in detail the principle of operation of this circuit (see Fig. 9).

Briefly, the quasi-rms detector is somewhere between a peak and an average detector. Since the rms value of a positive function lies between the average and peak value, it is instructive to investigate the action of a detector which gives a dc voltage corresponding to something between average and peak.

6.3.1 3A Quasi-rms Detector Circuit

Consider the capacitor of Fig. 9 to be large enough such that the circuit time constants are much longer than any associated with the input signal. The diodes conduct only when the input voltage is higher than the voltage across the capacitor. If R_1 is zero ohms, then e_{out} is equal to the peak value of the input signal minus any diode voltage

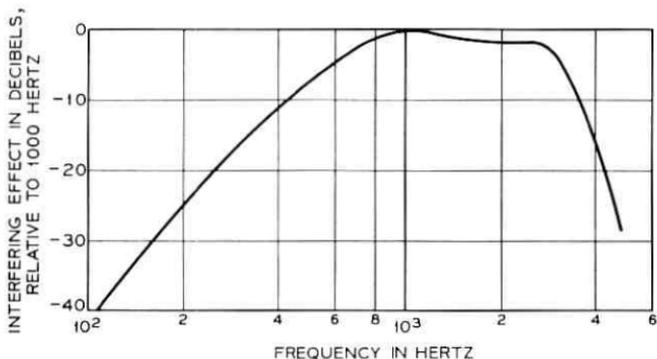


Fig. 8—Response vs frequency of C-message weighting characteristic.

drop in D_1 or D_2 . If R_1 is made very large compared to R_2 , then e_{out} is a measure of the average value of the input signal. By selecting the proper ratio of R_1 and R_2 , the circuit can be made to produce equal e_{out} for any two input signal waveforms of equal rms value.

Thus, if one wishes that sine waves and white noise of equal powers produce the same e_{out} then a ratio of

$$\frac{R_2}{R_1 + R_2} = 0.796.$$

should be chosen.

6.3.2 ATMS Quasi-rms Detector Circuit

The ATMS quasi-rms detector circuit⁵ is shown in Fig. 10. The operation of this circuit may best be understood by first considering that diode D_5 is an open circuit. Let the input be a sine wave. The gain of the amplifier from the input to point H will be very large until one pair of diodes D_1, D_2 or D_3, D_4 is broken down. D_1, D_2 will conduct on the positive swing at point H and D_3, D_4 on the negative swing. Resistors R_{1i} are much higher than the forward resistance of a conducting diode. When the diodes are conducting the gain of the amplifier is determined by the feedback resistors ($R_{1i}, R_{f1},$ and R_{f2}) and R_s .

Thus, the signal at point H will appear to be a magnified replica of the input signal sliced through at the zero voltage point with a square wave of peak-to-peak amplitude $VD_1 + VD_2 + VD_3 + VD_4$ added. If one were now to look with an oscilloscope at point A , a positive half-sinusoid with an additional dc voltage of VD_1 would be observed when the signal at H swings positive. If the signal at point B were now subtracted from the signal at point A , a full-wave rectified signal riding on an added dc voltage equal to a diode voltage drop would be observed.

It is this added diode voltage drop which now permits compensation for the voltage drop of D_5 , which we will now reinsert.

C is the storage capacitor ($4.22 \mu\text{F}$) mentioned in Section V. If diodes D_1 and D_2 are conducting, then C is charged through the R_1 , across

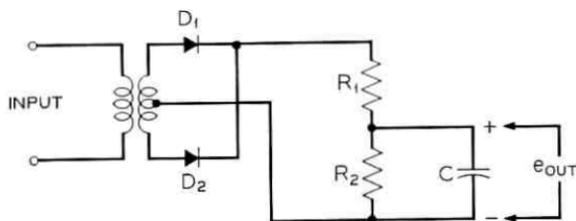


Fig. 9—Simple quasi-rms detector.

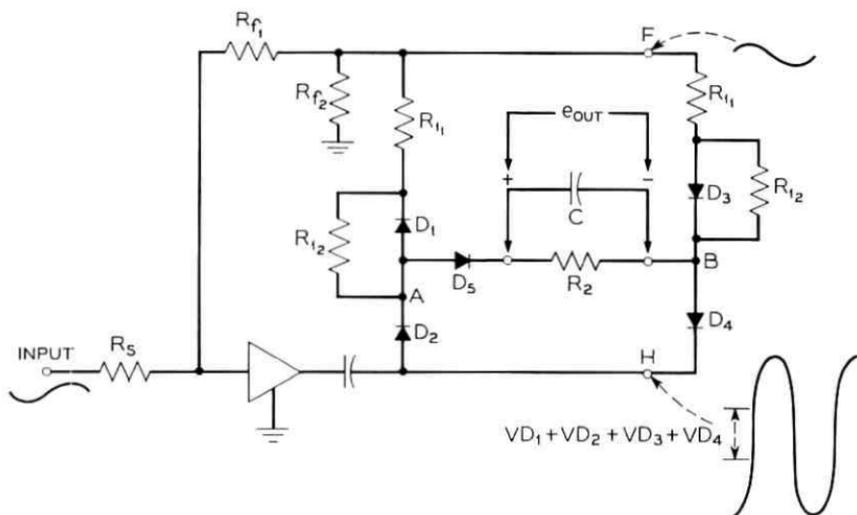


Fig. 10 — ATMS quasi-rms detector.

D_3 and the two R_{11} 's in series. The relationship between the resistors is as follows:

$$\frac{R_2}{R_{11} + R_{11} + R_{11} + R_2} = 0.796.$$

When the signal is first applied to this circuit, diode D_5 will be conducting most of the time. As C charges, D_5 will conduct less of the time as determined by the input signal waveform.

Diode D_5 is chosen so that its forward voltage drop is the same as the forward voltage drop across D_1 or D_3 , which carry higher currents than D_5 . The size of the R_1 's, R_2 and C are chosen as described below.

6.4 Noise Detector Transient Response

The ATMS quasi-rms measurement circuit, as in the 3A set, is designed to match the transient response of the human ear. This response was determined⁶ during tests in which listeners were asked to match the loudness of bursts of 1000-hertz tone to that of a steady 1000-hertz tone.

The response of the ear could not be exactly matched with a quasi-rms charging characteristic so a compromise was made to ensure a close match in the 150 to 250 millisecond range ($R_2 = 42.2\text{k}\Omega$ in Fig. 10).

The selection of a time constant must take into account the fact

that the discharge time constant of the quasi-rms circuit is more than four times the initial charging time constant. Thus, for noise which falls off during the measurement interval it is desirable to have R_2 as small as possible.

6.5 *Noise Measuring Interval*

The ATMS was designed to measure background noise rather than impulse noise. The most common types of background noise occurring on trunks are single-frequency tones, combinations of tones or white noise. The single-frequency tones can arise from such sources as power line harmonics and modulation products on carrier systems. The white noise arises from thermal and shot noise effects. The modulation products falling in a carrier channel from a large number of talkers in other channels also behave like white noise.

How long a period is necessary to measure white noise? It has been shown⁷ that the error resulting from a noise measurement made over a short interval of time decreases with increasing bandwidth and increasing measurement interval. It was desired that successive ATMS noise readings of a stable white noise source exhibit a standard deviation no larger than 0.25 dB. With the quasi-rms detector time constant as determined previously, a measurement interval of 0.375 second was found to meet this requirement.

Because of the characteristics of the quasi-rms detector, the measurement of a sine wave over this interval produces less than .05 dB error.

6.6 *ATMS vs 3A Noise Measuring Set Observers*

The ATMS noise measuring system performance was checked against 15 observers using a 3A Noise Measuring set on a series of noise tapes selected at random from a survey of 1069 intertoll trunks covering the whole Bell System. Fifteen-second noise samples from 15 different trunks were selected from each of two trunk length ranges: 250 to 500 miles and over 2000 miles.

Each individual noise segment occurred twice at random positions on the tape. The ATMS made three measurements during each 15-second segment for a total of six ATMS readings per segment.

The results of these tests are shown in Table II.

This data shows that the ATMS readings are consistent with the design requirements. Even greater reliability can be obtained by using the various repeat measurement modes. The possibility of rejecting

TABLE II—COMPARISON OF ATMS AND 3A OBSERVERS

		250 to 500 mile trunks	Over 2000 mile trunks
3A observers	δ	0.31–0.33 dB	0.33–0.35 dB
	Mean	31.7 dBrnC	38.8 dBrnC
ATMS	δ	0.82–0.84 dB	0.60–0.61 dB
	Mean	32.0 dBrnC	39.0 dBrnC

a good trunk for high noise readings on two successive measurements is remote.

VII. OVERALL ATMS OPERATION

7.1 General

So far, the basic measurement technique employed by ATMS and some special considerations for the measurement of noise have been discussed. In order to operate satisfactorily as a system, a number of other functions must be considered. These relate back to the nine simple steps described in the general description (Section 4.1) and include: accepting priming information, coordinating simultaneous operation of director and responder, making measurements, and displaying all results at the director location. All ATMS operations may be related to specific circuit functions or subsystems as follows:

- (i) Measurement circuits.
- (ii) Computational (counting) circuits.
- (iii) Storage (or registration) circuits.
- (iv) Signaling system.
- (v) Data transmission system.
- (vi) Control circuits.
- (vii) Timing circuits.
- (viii) Logic circuits.

Before describing how these circuits and subsystems function together, however, it is necessary to say a few more words about the measurement procedure.

7.1.1 Measurement Procedure

As described in Section V, the ATMS amplitude-to-pulse-width converter generates a pulse whose length is proportional to the

logarithm of the input signal level. A complete measurement requires that this pulse length be converted into a digital output which can be used to drive a Teletype machine or other display device. This is shown in Fig. 11, a block diagram of the ATMS measuring circuits. Thus, the pulse length is converted into a number of pulses which are then fed into a binary-coded decade counting circuit.

The counting circuit, functioning as described in Section V, determines the difference between the measured and expected values. Thus, upon completion of a measurement, the result, regardless of whether the actual measurement was made by the director or the responder, is stored in the director counting circuits.

7.2 Measurement Sequence

Using this information on the ATMS measurement procedure, the overall operation of the system may be described by a relatively simple sequence of events. In Fig. 12, a complete functional block diagram of the ATMS is shown. The procedure involved in gaining access to a trunk, making a measurement and advancing to the next trunk, is described in the General System Description, Section 4.3, which also discusses the functions of the automatic trunk test frame and the test line. The circuit blocks within the ATMS director and responder will be discussed as they occur in the description.

Before the measurement sequence begins, the test frame supplies the director with all necessary priming information. This permits the logic and control circuits to preset counters and limit circuits to

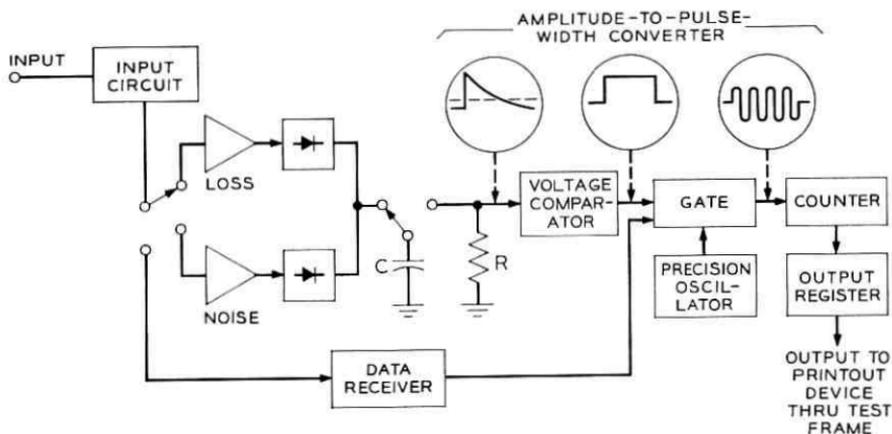


Fig. 11 — Block diagram ATMS measuring circuits.

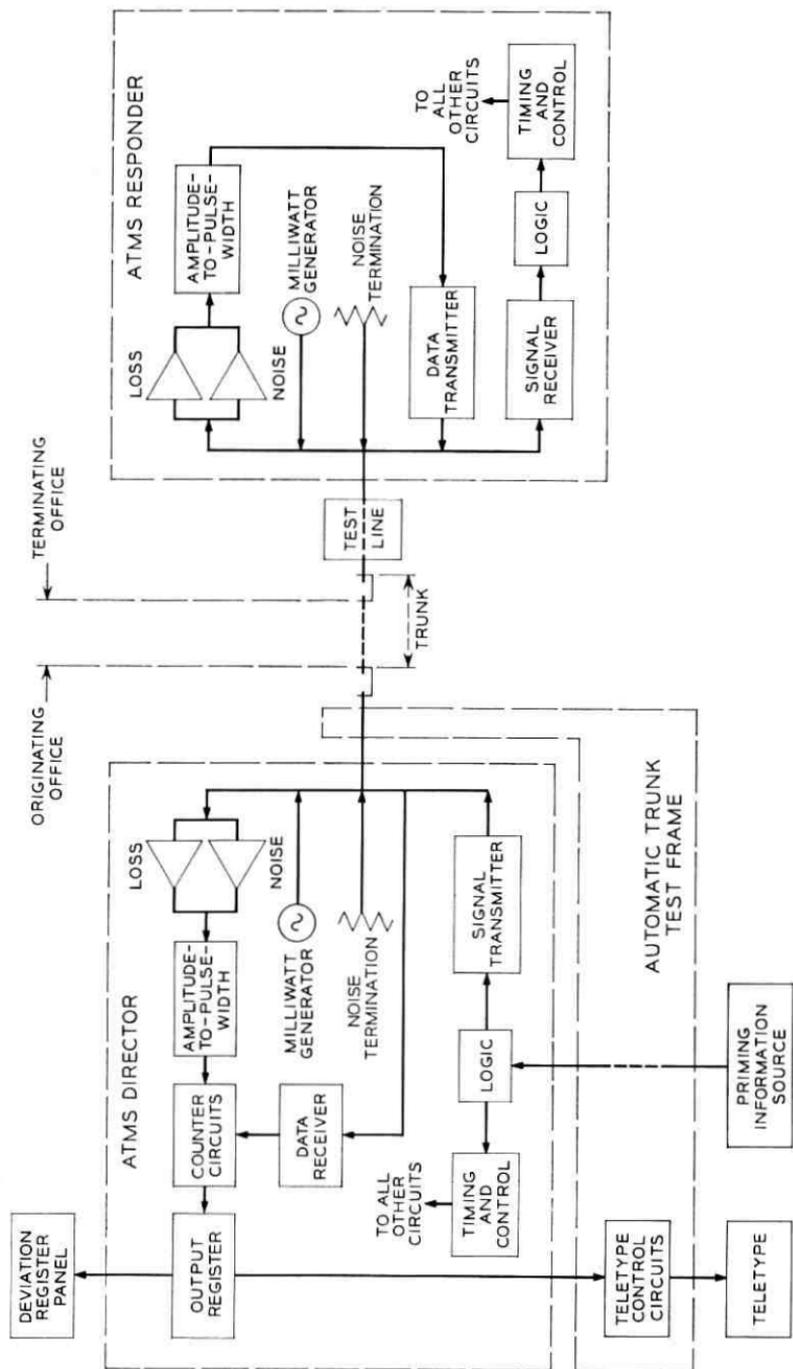
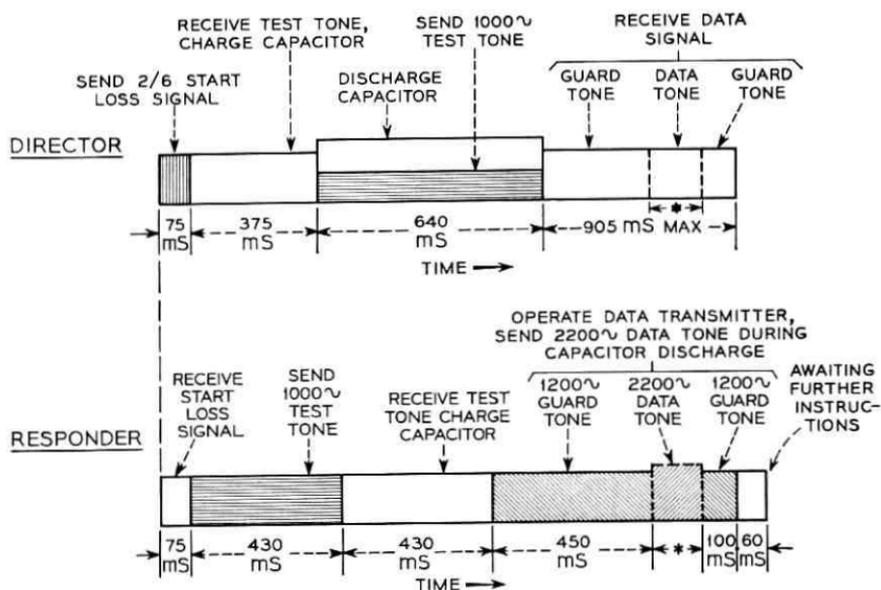


Fig. 12 — ATMS functional block diagram.

their correct values, set up the correct measurement mode (discussed in Section VIII), set the input circuit for the correct impedance and level, set the measuring circuits for self-check or trunk test, and other similar operations. Once the responder has been connected to the trunk, the test frame instructs the director to start the measurement sequence. The measurement sequence can be described using as an illustration a test sequence which includes loss in both directions and noise at both ends.

7.2.1 Loss Measurements

The simultaneous activities carried on by the director and the responder during loss measurements and the time allotted for each of these activities are shown in Fig. 13. Loss measurements are initiated when the director sends a 2-out-of-6 (2/6) multi-frequency (MF) signal to the responder commanding it to begin a loss measurement sequence. The timing circuit of the responder is triggered by the receipt of the signal, and the next three steps occur automatically. The responder sends a 1-kHz milliwatt test tone over the trunk to the director. The loss measurement circuit of the director amplifies and



* THIS TIME WILL VARY DEPENDING UPON THE REFERENCE LEVEL AND THE CHARGE ON THE STORAGE CAPACITOR

Fig. 13 — Loss measurement timing.

filters the signal. The amplitude-to-pulse-width converter rectifies the signal and uses the resulting dc potential to charge a storage capacitor.

The director then performs two simultaneous activities. It discharges the storage capacitor of the amplitude-to-pulse-width converter. This pulse gates on the precision oscillator, and the decade counting circuits count these pulses to rate the trunk on its far-to-near loss. At the same time, the director transmits a 1-kHz test tone to the responder as the first part of a near-to-far loss measurement. The loss measurement and the amplitude-to-pulse-width converter circuits of the responder are identical to those of the director. These circuits charge the storage capacitor in the responder to a level which is dependent upon the near-to-far loss characteristic of the trunk. The loss measurements are completed as the responder sends data back to the director to indicate the near-to-far loss characteristic. When the storage capacitor of the responder is discharged, the pulse is used to control a data transmitter which starts by sending guard tone (1200 Hz). It then shifts to data tone (2200 Hz) for the duration of the pulse, then returns to guard tone for a short period. The data receiver of the director converts this data signal into the dc pulse which gates on the precision oscillator so that the near-to-far loss deviation may be counted. Two-way loss measurements are accomplished in less than two seconds.

7.2.2 *Noise Measurements*

Noise measurements begin when the director samples the noise present on the trunk under test. The noise is amplified, weighted with C-message weighting and rectified in the noise measuring circuit, then used to charge the storage capacitor. The responder at this time functions only as a quiet termination for the trunk at the far end. Next, the director commands the responder to make a noise measurement. It does this by transmitting a 2/6 MF signal to the signaling receiver of the responder. After sending the "start noise" measurement command, the director provides a near-end termination for the trunk under test. The responder uses its own measuring circuit and amplitude-to-pulse-width converter to charge the responder storage capacitor from noise present at the far-end of the trunk. Simultaneously, the director discharges its storage capacitor which had been charged from the near-end noise. The resulting pulse gates the precision oscillator and the near-end deviation from the reference noise level is counted. Then, the responder discharges its storage capacitor and sends a data signal

indicating the results of the far-end noise measurement to the director. The data receiver of the director converts the data signal into a gate pulse for the counting of the far-end noise deviation from reference level.

7.2.3 *Nonmeasurement Functions*

At the appropriate time in the testing sequence, the measurement is transferred from the counter to the output register, a relay circuit which translates the results from a binary to a decimal code and stores them. The automatic test frame "reads" the ATMS output and causes the results to be printed. Once the results of a measurement are stored in the output register, the counting circuits may be preset and the sequence advanced to the next measurement. Note that this provides the director with the capacity for the simultaneous storage of two answers; one in the output register and one in the counter. This feature is used to advantage to decrease the measurement time.

In addition to the functions described above, counters are preset, results are compared with limits to determine cues (indications that a measurement has exceeded a limit), and a determination is made as to whether the trunk should be retested.

VIII. MEASUREMENT MODES

The ATMS provides its users with a choice in the amount of printout information that may be obtained. In all the preceding discussions, the operation of the ATMS was described with all of the measurement results printed out. This would include self-check results and both initial and repeat results when a trunk test is repeated, and is the maximum printout available. There are occasions, however, where such complete results are not necessary (and indeed, may even make it more difficult to utilize the results), and when desired, the ATMS may be instructed to print out the results of only those trunks whose measurement results have exceeded some limit. The advantages of such operation include increased testing speed and a printed record of only those trunks exceeding certain maintenance limit.

8.1 *Measurement Limits and Cues*

The ATMS director may be set for measurement limits that correspond to two different degrees of urgency: maintenance limits and immediate action limits. Any of ten maintenance limits and seven immediate action limits may be selected. During the measurement

sequence the director provides special indications called cues, along with the measurement results, whenever one or more measurements exceed one of these limits. A cue of "1" (Q_1) is provided when a maintenance limit is exceeded and a cue of "2" (Q_2) is provided when an immediate action limit is exceeded. In addition, self-check limits are built into the director. A cue of "0" (Q_0) indicates a satisfactory self-check and a cue of "9" (Q_9) indicates a self-check limit has been exceeded (± 0.1 dB for loss and ± 1 dB for noise).

8.2 Measurement and Printout Modes

Four different measurement and printout modes may be set into the director by switch selection at the test frame. These modes are as follows:

(i) Full Printout—No Repeat: All measurements are printed out and no repeat tests are made regardless of the cue.

(ii) Full Printout—Repeat on Q_2 : All initial measurements are printed out and if Q_2 , which is the highest limit, is exceeded, the measurements are repeated and printed out.

(iii) Full Printout—Repeat Q_1 or Q_2 : All initial measurements are printed out and the measurements are repeated and printed out if either Q_1 or Q_2 is exceeded.

(iv) Abbreviated Printout—Repeat on Q_1 or Q_2 : Initial measurements are not printed out. If no limit is exceeded, no record is made. If either Q_1 or Q_2 is exceeded, the measurements are repeated and the results of the repeat test are printed out.

All self-check results are printed out, both initial and repeat test, regardless of the print mode selected.

IX. MECHANICAL FEATURES

ATMS directors and responders each consist of a group of modules called circuit packs which plug into horizontal mounting shelves. The shelves, in turn, are fastened to the framework of 23-inch relay racks. A typical circuit pack is shown in Fig. 14. Each circuit pack is 8-3/8 inches high, 8 inches deep, and either 1 or 2 inches wide. Most electrical parts are mounted on epoxy glass printed wiring boards. A few components, such as keys and jacks, are mounted in the face panel of the cast metal frame. A multiple plug at the rear of the circuit pack provides interconnection to other units through a mating connector and the shelf wiring. On the director, installer wiring termi-

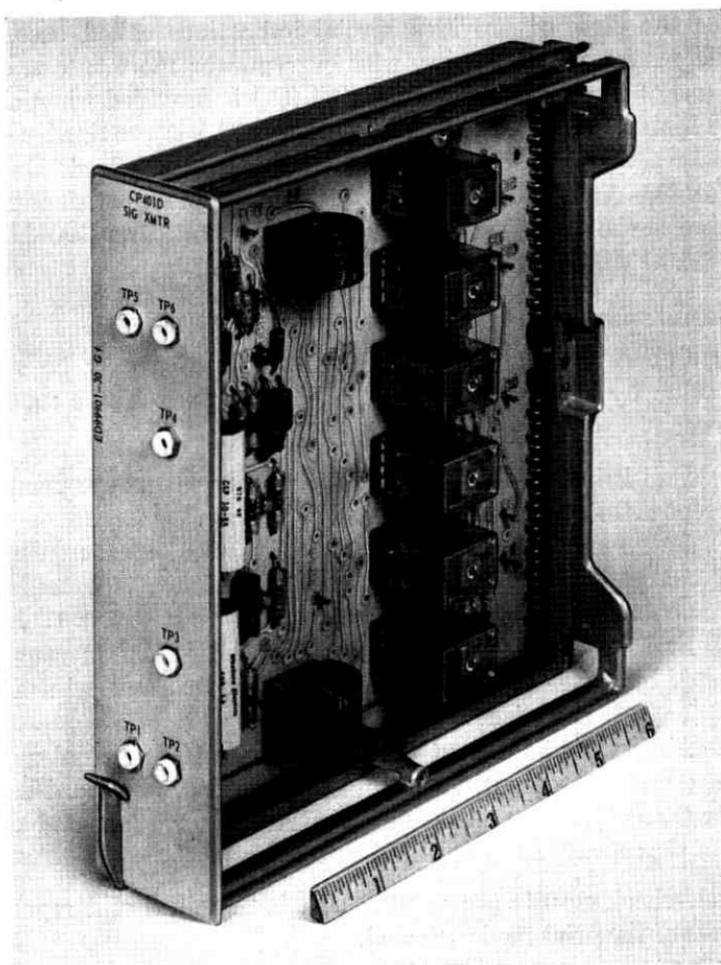


Fig. 14 — ATMS circuit pack.

nates in several multiple plugs which engage mating connectors wired to appropriate circuit pack connectors.

Circuit packs of the director mount in four horizontal shelves. The shelves are each 10 inches high. The overall assembly therefore occupies 40 inches in a 23-inch bay. The director is shown in Fig. 15.

The responder occupies only three horizontal shelves of a bay. The complete assembly, shown in Fig. 16 is 30 inches high and 23 inches wide. As shown in Fig. 15 and 16, both the director and responder contain circuit packs with no designations on the front. These are

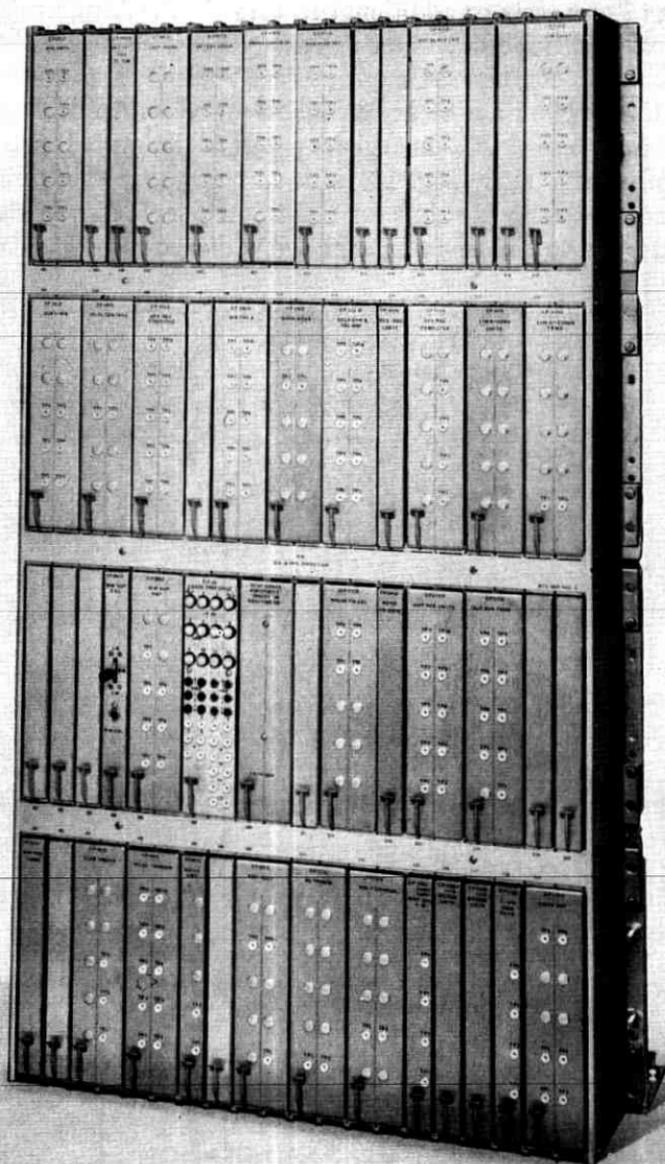


Fig. 15— ATMS director.

blank circuit pack frames and represent the expansion space for the addition of new tests or additional features in the future.

The loss and noise deviation register panel is 6 inches high and 23 inches wide. It contains 34 message registers which provide information on the distribution of deviations in an office. An early version of the loss and noise deviation register panel is illustrated in Fig. 17.

The alignment unit (Fig. 18) is a carrying case containing circuit packs used in testing the director and responder. It also holds a circuit pack extender to aid in making maintenance measurements and ad-

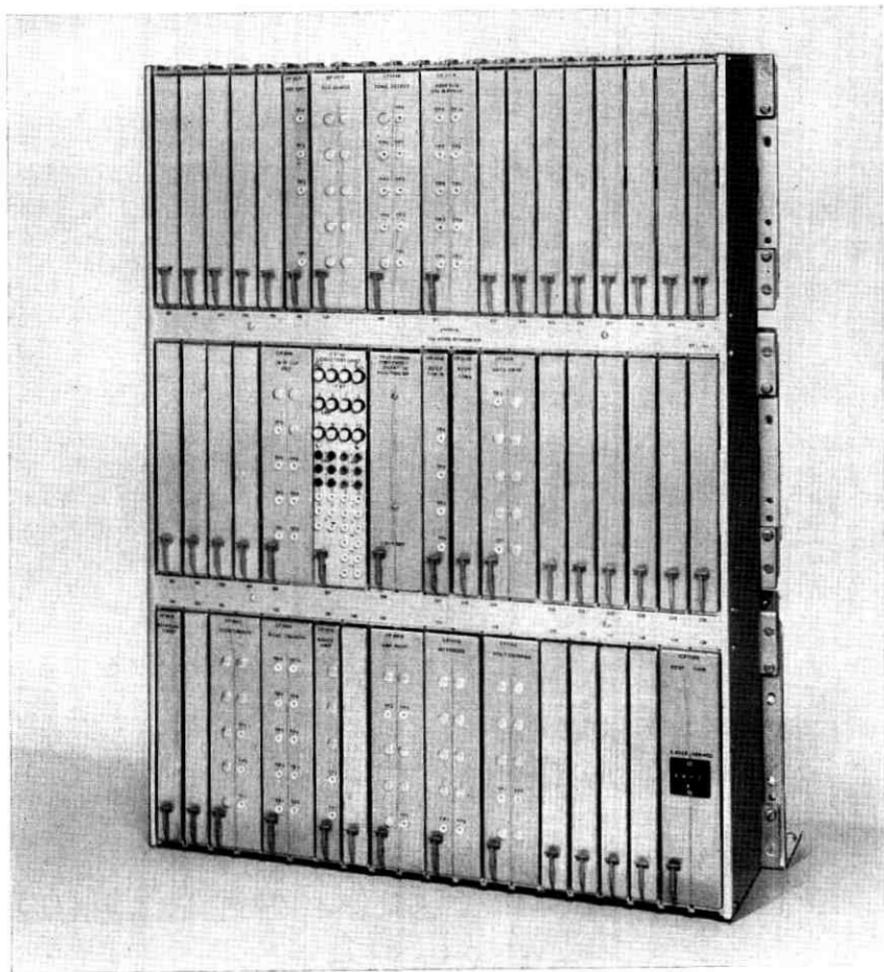


Fig. 16 — ATMS responder.

justments. The unit is 10 inches high, 15 inches wide, and 11 inches deep. It can be mounted by brackets in a 23-inch bay. In this case, the whole assembly is 10 inches high and 23 inches wide.

X. MAINTENANCE

10.1 Alignment

Alignment of the director or the responder is accomplished by use of test circuit packs which normally are stored in the alignment unit. Alignment is necessary upon installation, when a critical circuit pack is replaced, and on a routine basis. Routine alignment is not expected to be necessary more often than every six months.

10.2 Trouble Location

Maintenance is facilitated by use of test points located on the face panels of the circuit packs. The test points provide access to particularly important points in the circuits. The circuit packs can be placed on an extender (included in the alignment unit) to make internal measurements or adjustments. No maintenance or repair of individual circuit packs is required by the user. Instead, the faulty circuit pack is simply located and replaced. The faulty circuit pack is then sent to a repair center. Special test procedures are provided for rapidly identifying faulty circuit packs.

XI. ATMS FIELD TRIAL

An extensive field trial was undertaken to assure that the ATMS and the associated switching equipment would function properly in the actual telephone offices.

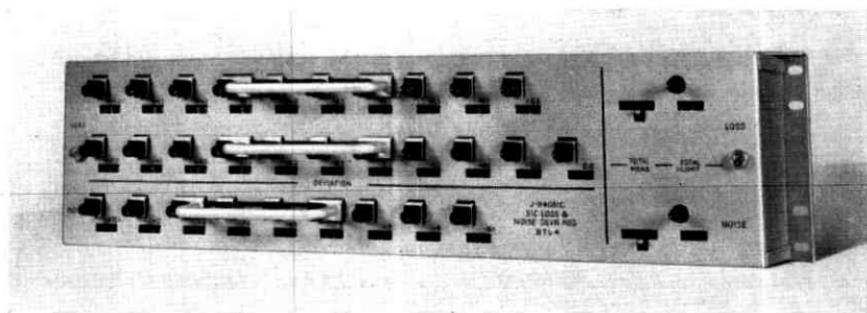


Fig. 17 — ATMS loss and noise deviation register.

11.1 *Equipment Location*

The ATMS was on field trial in the Norristown, Pennsylvania, area between January, 1965 and January 1967. An ATMS director was associated with a No. 5 Crossbar Automatic Progression Trunk Test Frame (APTT) in the Norristown, Pennsylvania, central office. Five responders were located as indicated in Table III. Other far-end offices with 102-type and 104-type test lines were included in the trial.

11.2 *Field Trial Results*

11.2.1 *General*

The ATMS was found to meet all its design requirements. Trunks were tested more frequently and precisely than would have been possible with manual trunk testing by the telephone office personnel.

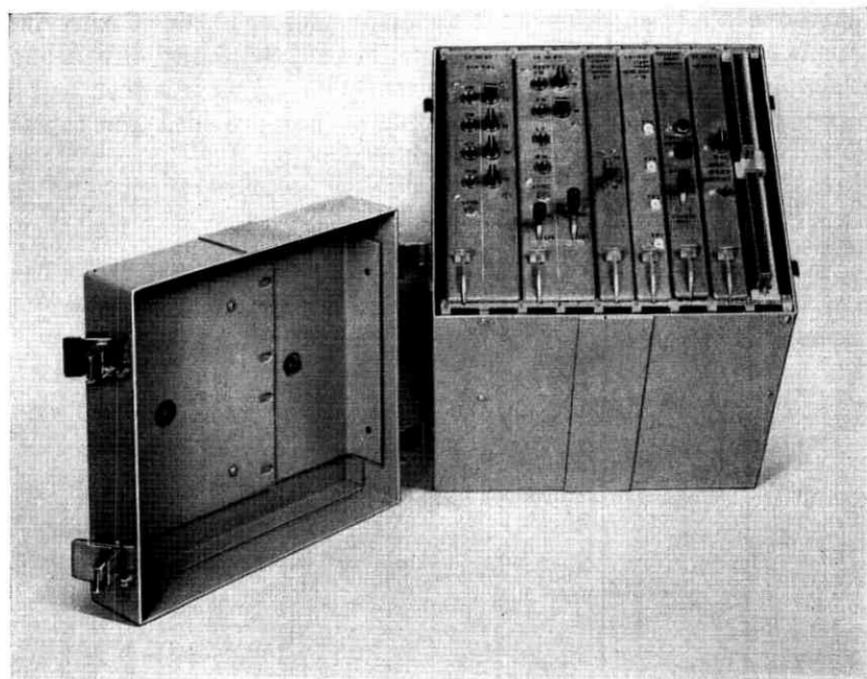


Fig. 18 — ATMS alignment unit.

TABLE III—RESPONDERS FOR ATMS FIELD TRIAL

Location of responder	Airline distance to Norristown, Pa. in miles	Type of access to responder through 105-type test line
Wayne, Pa.	6	4A toll
Lansdale, Pa.	9	SXS toll, SXS local
Philadelphia, Pa.	15	XB tandem, 4M toll IXB local, panel local
Pottstown, Pa.	19	5XB toll, 5XB local
Newark, N. J.	80	4A toll

11.2.2 Trunk Testing Time

It is desirable not only to reduce the time per trunk tested, but to reduce the holding time of the trunk so that it will be available to customers. The typical trunk was held for approximately eight seconds on a no-repeat loss and noise measurement to an ATMS responder. About half of this time was measurement time. The remaining time was necessary to complete the printing of the measurement results.

The speed at which trunks can be tested varies considerably depending on any or all of the factors below.

- (i) Number of trunks in the trunk group.
- (ii) Number of self-checks requested.
- (iii) Printout mode—no repeat or repeat.
- (iv) Number of trunks requiring a repeat measurement.
- (v) Busy trunks—this is a function of the time of day.
- (vi) Trunk seizure time.

The trunk noise readings are usually highest during the hours of peak office activity. Unfortunately, busy hour testing implies a maximum number of busy trunks (80 percent during one extended test) as well as competition with the customer for the few available trunks. The "busy hour" in Norristown extends almost all day, necessitating night-time testing.

During the field trial, measurements were made in the Repeat-on- Q_1 -or- Q_2 printout mode during the hours from midnight to 8 a.m. Dividing the total time by the number of trunks tested results in an average time of 50 to 60 seconds per trunk tested.

Assuming 10 hours of usage per day, a seven-day week and an

average trunk test time of 60 seconds per trunk, one may then test 4,200 trunks per week.

XII. REMOTE-OFFICE TESTING

The classes of ATMS testing previously discussed permitted testing of trunks between an office containing an ATMS director and other offices with 100-, 102-, 104- or 105-Type Test Lines. See Table I.

Trunks between offices too small to justify an ATMS director and its associated test frame could not be tested with the ATMS until the advent of the Remote Office Test Line (ROTL). The ROTL (to be available soon) permits the director at Office A to obtain the results of measurements on trunks between Office B, equipped with a ROTL, and Office C equipped with a Code 100-, 102- or 105-Type Test Line.

12.1 General

The office containing the director and its associated test frame will be referred to as the near-end office, the office with the ROTL as the remote office, and the office containing the test line as the far-end office. The trunk between the near-end office and the remote office will be called the access trunk.

Two kinds of remote office testing have been developed.

(i) Remote-Office-Responder Testing—full accuracy, measurements made at the remote office using a modified responder—see Fig. 19.

(ii) Remote-Office Through Testing—reduced accuracy, no measurements made at the remote office, lower cost—see Fig. 20.

The remote office concept may be used to measure trunks if the far-end office is equipped with a 100-, 102- or 105-Type Test Line.

Under control of the test frame the ROTL can seize an outgoing

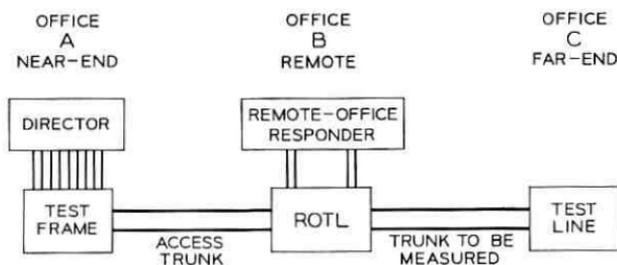


Fig. 19 — Remote-office-responder testing.

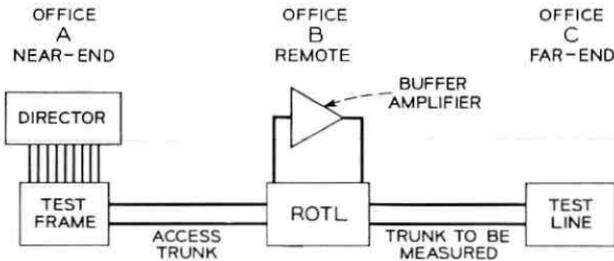


Fig. 20 — Remote-office through testing.

trunk to the far-end office, pulse forward a test line code and assist in making transmission tests on the outgoing trunk.

The responder and test line can be in the same office or building as the director and test frame to permit testing of incoming trunks. An incoming trunk is defined as one which can be seized only at a distant office. An access trunk and ROTL are used to gain access to the originating end of these incoming trunks.

12.2 Systems with Similarities to Remote-Office Testing

12.2.1 *L. M. Ericsson Remote-Controlled Measurement²*

The Swedish Ericsson system mentioned in Section II can operate in a mode in which the controlling set in office A can control test sets in offices B and C. The results of the measurement of the trunk between B and C is then relayed to A by means of a multifrequency code. This system employs a slower measurement method.

12.2.2 *Loop-Around Test Line*

At the present time manual, one-man, two-way loss measurements are possible without 104-Type Test Line or ATMS equipment if the remote office is equipped for loop-around testing.

All trunks in a group to the remote office are first measured in the far-to-near direction by seizing the Milliwatt Test Line in the remote office. One of these trunks is then selected as the reference trunk and is connected in turn through the loop-around test line to each of the other trunks in succession. Test tone is then sent from the originating office through the trunk to be tested and back through the reference trunk. Subtraction is then necessary to obtain the near-to-far loss of each trunk.

Loop-around testing necessarily requires that a means be available

at the test location in the originating office for originating and holding two connections simultaneously. Only outgoing trunks may be tested by this method.

There are many disadvantages to this method in addition to the subtraction required. If the loss of the reference trunk varies with time or level (see Section II), then the computed near-to-far losses for the other trunks will be in error. Mismatch errors may occur when the reference trunk is connected to the trunk to be measured. As with all manual measurements, the procedure is slow.

12.3 *Remote-Office-Responder Testing*

All measurements in this mode are made by either the far-end responder (for the case of a 105-Type Test Line) or a modified responder at the remote office. All measurement results are sent back to the director in the form of frequency-shift data signals. The loss and noise of the access trunk therefore do not degrade the accuracy from that of a director-to-responder measurement.

12.3.1 *Responder Modification*

A responder is modified to a remote-office responder by the addition of three circuit packs. These provide for modification of the timing cycles and independent output circuitry toward the access trunk. The remote-office responder is capable of transmitting different signals simultaneously, one toward the director and another toward the far-end equipment. This responder may still be utilized as a standard responder, if desired.

By using the isolation amplifier contained in the remote-office responder one can pass a signal *through* a remote-office responder in either direction. 2/6 commands may be passed through the remote-office responder to the far-end responder or the data from the far-end responder may be passed through the remote-office responder to the director.

12.3.2 *Remote-Office-Responder Testing Sequence*

Fig. 21 shows a two-way loss and noise measurement between a remote-office responder and a far-end responder. One new 2/6 multi-frequency command is necessary to make the near-end noise measurement. The other 2/6 commands are the same as those required for a normal director-to-responder measurement. It should be noted that a

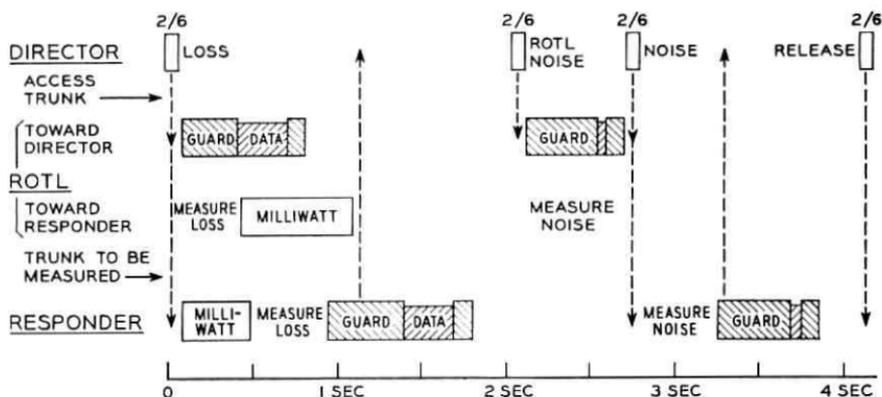


Fig. 21 — Remote-office-responder testing to a far-end responder.

normal responder will not reply to any other 2/6 commands than those mentioned in Section VII.

Both the remote-office responder and the far-end responder will reply to the 2/6 loss signal. The remote-office responder will measure the far-end responder test tone and at the same time transmit the 1200-Hz guard tone to the director. The remote-office responder will now complete the data signal to the director which it is transmitting test tone toward the far-end responder. The director then receives the loss data signal from the far-end responder through the remote-office responder.

The director then commands the remote-office responder to measure noise. After the director has received this data signal it commands the far-end responder to measure noise, and with the same signal, the remote-office responder to pass a data signal through to the director. When this final noise data signal has been completed, both responders return to a state where they can receive 2/6 commands.

Loss and noise self-checks of the remote-office responder and the far-end responder may be completed in a somewhat similar manner.

It is now clear that the loss and noise of the access trunk will have no effect on the accuracy or range of measurements.

12.4 Remote-Office Through Testing

This mode of testing necessitates connecting the access trunk to the trunk to be measured. If both of these trunks employed negative impedance repeaters, then a possibility of a singing condition exists due

to mistermination. This can be eliminated by the use of a buffer amplifier.

12.4.1 *Buffer Amplifier*

The buffer amplifier eliminates the effect of interaction between the impedances of the two trunks and provides terminations of nominal impedance during measurements. Because of its unilateral transmission, however, the buffer amplifier necessitates more control functions in the ROTL.

Since no measuring equipment is present in the remote-office equipment, the director must make the measurement for noise at the remote-office end of the trunk to be measured. Noise on the access trunk of the same level as that on the trunk to be measured can have a large effect on the measurement accuracy. For this reason a buffer amplifier gain of 20 dB was chosen for the period of this noise measurement. At all other times a buffer amplifier gain of 0 dB has the advantage of preserving signal levels.

In the sequences which follow, the buffer amplifier is used in such a manner that its actual gain is relatively unimportant as long as the amplifier is linear.

12.4.2 *Remote-Office Through Measurement Sequence*

Fig. 22 shows the actual sequence for a measurement to a far-end responder. Fig. 23 is a simplified diagram of the amplifier and the transmission portion of the ROTL. Table IV describes the sequence followed by the circuit shown in Fig. 23.

The far-to-near loss measurement is made by first measuring the loss of the access trunk (interval t_1) from the remote-office to the director and then subtracting this from the loss of the trunk to be measured and the access trunk in tandem (interval t_3). During interval t_4 the far-end responder is measuring the test power from the remote office.

The director makes a measurement of the noise at the ROTL end of the trunk to be measured during interval t_6 . During interval t_8 the responder can make a valid noise measurement because the buffer amplifier is pointed toward the director. Not only does it block any noise on the access trunk, but it terminates the trunk to be measured in the correct impedance independent of the impedance of the access trunk.

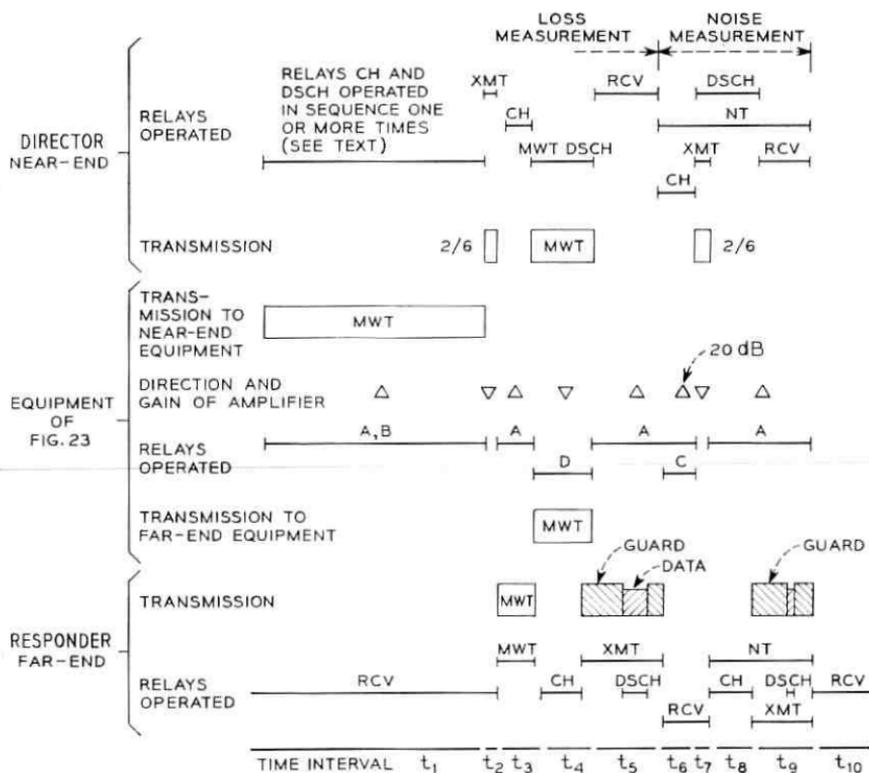


Fig. 22 — Remote-office through testing to a far-end responder.

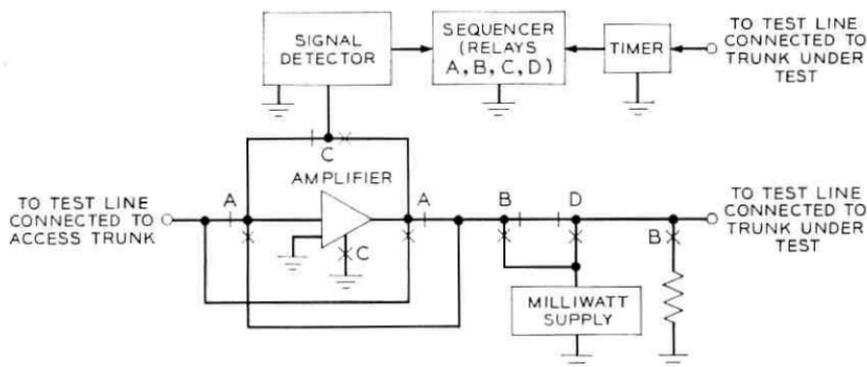


Fig. 23 — Simplified ROTL for remote-office through testing.

TABLE IV—REMOTE-OFFICE THROUGH TESTING SEQUENCE

Time interval from Fig. 22	Approximate time in seconds	Relays operated				Function being performed during the time relays are operated and released as shown
		A	B	C	D	
t_1	3.00	X	X			3 sec. MWT from Fig. 23 to near end
t_2	0.08					2/6 freq. signal from near to far end
t_3	0.43	X				MWT from far end to near end
t_4	0.64				X	MWT from Fig. 23 to far end
t_5	0.65	X				Loss data from far to near end
t_6	0.38	X		X		Near end measuring noise
t_7	0.15					2/6 freq. signal from near to far end
t_8	0.43	X				Far end measures noise
t_9	0.54	X				Noise data from near to far end
t_{10}	0.08					2/6 freq. signal from near to far end

It should be noted that the far-end responder receives commands from the director and acts on these commands in exactly the same manner as it would in a measurement without a ROTL. Thus, a responder does not have to know whether its commands come through a ROTL.

Testing to a 100- or 102-Type Test Line is accomplished by employing parts of the 105-Type Test Line sequence shown in Fig. 22.

12.4.3 Remote-Office Through Testing Limitation

In the loss measurement portion of the sequence the director must make a measurement of the access trunk loss which it will subsequently subtract from the loss of the trunk to be measured and the access trunk loss in tandem. The access trunk loss measurement is made during an initial transmission of the remote-office test tone through the remote-office buffer amplifier. Since the director cannot measure a received

level below -15 dBm the loss of the two trunks in tandem cannot be greater than 15 dB.

Raising the gain of the buffer amplifier at the remote-office would increase this range but would also imply initial transmission of the remote office milliwatt at the level about 0 dBm. Such a transmission could cause overload and crosstalk problems on an access trunk over a carrier system. The most practical plan, therefore, is to use access trunks with as low a loss as possible.

The subtraction process to obtain the loss of the trunk from the far-end office with the test line to the remote office involves two separate measurements made at different times and at different levels. With some carrier access trunks, beating of pilot frequencies can easily result in time-varying trunk loss variation of 0.2 dB. Compandor tracking errors can add another 0.1 dB or more of error.

If the far-end test (100- or 105-type) line permits noise measurement at the remote-office then the 20 dB buffer amplifier gain mode is employed. The loss of the access trunk now affects the noise measurement accuracy, for the noise level at the director must be reduced by 20 dB minus the loss of the access trunk. This access trunk loss is stored in pads in the noise measurement path in the director. This loss is stored to the nearest 1 dB—thereby introducing a noise error of ± 0.5 dB. Compandor tracking errors are greater at noise measurement levels. These errors add to the carrier beating problem already mentioned.

The necessity for the initial test tone transmission from the remote-office limits the printout to one mode—full printout—no repeat.

No remote-office through automatic testing is attempted to a 104-Type Test Line because its automatic mode (as opposed to manual mode) cannot be utilized. The half-minute cycle time for the 104 circuit in its manual mode restricts the number of trunks which could be tested.

XIII. SUMMARY

The Automatic Transmission Measuring System (ATMS) permits accurate and more rapid measurement of telephone trunks than was previously possible. The 1000-Hz loss of a trunk may be measured in both directions to an accuracy of ± 0.1 dB. Noise measurements at each end of the trunk are accurate to ± 1 dB. The results of these measurements are printed on page copy, or perforated on punched tape or cards.

An ATMS director is in one office and an ATMS responder is in

the other office. The responder may be commanded to make any one of several measurements in conjunction with the director.

The ATMS director works in conjunction with one of several automatic test frames or ESS central control which provides an interface for the director to the particular switching system. Interface for the ATMS responder is provided by a 105-Type Test Line. For most switching systems, the total ATMS measurement time for two loss and two noise readings is less than the overall time to read the trunk information from the input tape and seize the trunk.

Personnel are needed only for loading the input, reading and interpreting the output, periodic alignments and occasional maintenance. When an ATMS director and its associated test frame cannot be provided in a particular office, measurements may be made on trunks between central offices by using one of two Remote-Office Test Line (ROTL) concepts.

The ATMS director can make measurements to four different far-end test lines—

- (i) 100-Type 5 seconds of milliwatt followed by a quiet termination (to be available soon).
- (ii) 102-Type Milliwatt, interrupted at 10-second intervals.
- (iii) 104-Type Transmission Measuring and Noise Checking Circuit (TMANC).
- (iv) 105-Type ATMS responder.

A director may make measurements through a remote-office to a 100-, 102-, or 105-Type far-end test line.

Flexibility and ease of maintenance result from the use of transistor circuits on plug-in circuit packs. The director requires 40 inches of a 23-inch relay rack and the responder requires 30 inches of a 23-inch relay rack.

XIV. ACKNOWLEDGMENTS

Contributions to the development of the ATMS have been made by the transmission measurements department, switching maintenance engineering department and the various switching development departments of Bell Telephone Laboratories. Particularly valuable have been the work of R. W. Hatch, R. F. Rollman, and P. J. Dugal (formerly) of Bell Telephone Laboratories, D. T. Osgood of the American Telephone and Telegraph Company and the continuing efforts of F. J. Danik and R. L. Hanson of Bell Telephone Laboratories.

REFERENCES

1. Felder, H. H., Pascarella, A. J., and Shoffstall, H. F., Automatic Testing of Transmission and Operational Functions of Intertoll Trunks, *B.S.T.J.*, *35*, July, 1956, pp. 927-972.
2. Carlstrom, P., Automatic Transmission Measuring Equipment for Telephone Circuits 1. Equipment Design, *Ericsson Review*, *40*, #2, 1963, pp. 62-68.
3. Carlstrom, P., Automatic Transmission Measuring Equipment for Telephone Circuits 2. Measurement of Attenuation and Noise, *Ericsson Review*, *40*, #3, 1963, pp. 78-86.
4. Cochran, W. T. and Lewinski, D. A., A New Measuring Set for Message Circuit Noise, *B.S.T.J.*, *39*, July, 1960, pp. 911-932.
5. Ingle, J. F., U. S. Patent 3287651, issued November 22, 1966.
6. Munson, W. A., The Growth of Auditory Sensation, *J. Acoust. Soc. Am.*, *19*, 1947, pp. 584-589.
7. Rice, S. O., Filtered Thermal Noise—Fluctuation of Energy as a Function of Interval Length, *J. Acoust. Soc. Am.*, *14*, 1943, p. 223.

A Solid-State Regenerative Repeater for Guided Millimeter-Wave Communication Systems

By W. M. HUBBARD, J. E. GOELL, W. D. WARTERS,
R. D. STANDLEY, G. D. MANDEVILLE, T. P. LEE,
R. C. SHAW, and P. L. CLOUSER

(Manuscript received July 20, 1967)

Recent advances in solid-state device technology for generating millimeter waves as well as advances in component design for IF and baseband portions of repeaters have renewed interest in millimeter-wave guided-wave communication systems. This paper describes a 306 Mb/s, all solid-state repeater which has been built using a 1.3-GHz IF and a form of differentially-coherent phase modulation. A signal-to-noise ratio of 13.6 dB is required for an error probability of 10^{-9} (compared with a theoretical value of 13.0 dB for an ideal differentially-coherent phase-modulated system). Sufficient gain for 15-mile repeater spacings (using two-inch circular waveguide) has been obtained with an LSA diode, an IMPATT diode, and a varactor multiplier as the millimeter-wave power source.

I. INTRODUCTION

1.1 Guided Millimeter-Wave Communication Systems

High-speed, long-haul communication by means of millimeter waves transmitted in the circular-electric mode in a multimode circular waveguide was described by S. E. Miller¹ in 1954. Recent advances in solid-state devices for generating millimeter waves as well as advances in circuit design for the IF and baseband portions of the repeaters have renewed interest in such a system.

The purpose of this paper is to describe the design and performance of an experimental all solid-state millimeter-wave repeater which has recently been built and tested. It operates at a carrier frequency of 51.7 GHz and transmits binary PCM at a 306 Mb/s rate. The experimental repeater includes all of the active circuitry for one channel of

a complete repeater and contains channel filters representative of those needed to separate and combine the many channels of an actual system. It was built to demonstrate certain principles and no attempt is made here to describe or design a complete system. Certain system considerations are discussed in Section IV in order to give the reader some perspective concerning those factors which influenced the design of the repeater.

In section II, we discuss a modulation scheme which was conceived to satisfy the requirement imposed by the nature of the system. The circuitry used in the repeater is discussed in Section III. Particular emphasis is placed on those portions of the circuit which the authors feel represent a significant advance in the state of the art. The performance of the repeater is described in Section V. Finally, the conclusions which are to be drawn from the experimental performance of the repeater are summarized in Section VI.

For a given repeater gain and spacing the communication capacity of such a system is set by the attenuation characteristics of the waveguide. The system under consideration would use TE_{01} mode transmission in 2-inch helix or dielectric-lined circular waveguides.

The characteristics of these kinds of waveguides have been studied theoretically by H. E. Rowe and W. D. Warters,² S. P. Morgan and J. A. Young,³ and H. G. Unger;⁴ and studied experimentally by A. P. King and G. D. Mandeville⁵ and W. H. Steier⁶ for a straight waveguide. More recently a study of typical route loss has been undertaken by W. G. Nutt and others.⁷ Their results, shown in Fig. 1 along with the results of the other measurements cited above, are used for the calculations in Section 4.1. From these measurements, one finds that

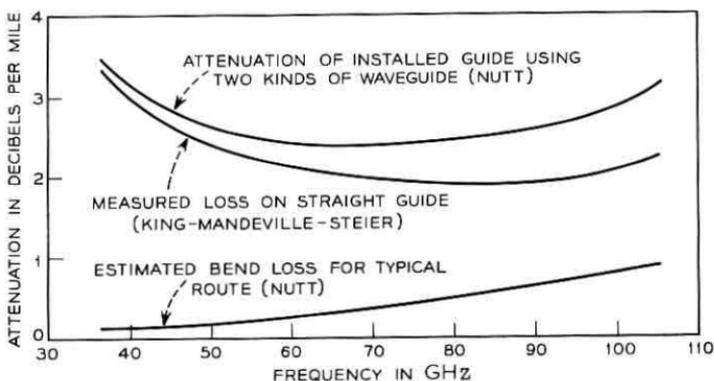


Fig. 1— TE_{01} mode attenuation characteristic of 2-inch circular waveguide.

the attenuation is less than 3 dB per mile over a band of frequencies extending from approximately 40 GHz to 100 GHz. This 60-GHz band of frequencies is considered the "usable bandwidth" of the waveguide. In Section IV, it is shown that approximately 200,000 two-way voice channels can be accommodated by the waveguide if 9 digit binary PCM transmission is used.

The purpose of this experiment was to demonstrate the feasibility of building repeaters for a millimeter-wave communication system. When this experiment was begun early in 1966, the band-splitting filters and the IF amplifiers had already been developed and no problems were expected in these areas—and in fact, none arose. Our initial efforts were concerned with building filters for dropping the individual channels and for injection of local oscillator power into the up- and down-converters, providing a source of millimeter-wave power, building up- and down-converters with attractive conversion loss, building FM deviators, and building baseband and timing recovery circuitry which would operate at the 306 Mb/s rate. Soon after this work began, the LSA (Limited Space-charge Accumulation) oscillator was developed by J. A. Copeland.⁸ It provides what seems to be a suitable millimeter-wave power source. In addition, a 12.6-GHz IMPATT (IMPact Ionization Avalanche Transit Time) diode driving a quadrupler, provides a suitable power source. More recently, a 50.4-GHz IMPATT diode has been successfully tested as a millimeter-wave power source. The other components were developed during the course of the experiment. Thus, it has been demonstrated that such a system is within the present state of the art.

One significant component, a delay distortion equalizer, which will be required in an actual system, was not considered in this experiment. Several possible equalizers have been proposed in the past and while considerable study is still required before a choice can be made from among these alternatives, there do not seem to be any problems associated with equalization that would affect the feasibility of the system. A review of some work which has been done on equalization of delay distortion is presented briefly in Appendix A.

II. MODULATION

2.1 Modulation Requirements

It was felt that the modulation scheme used in this repeater should satisfy four important requirements. First, in order to make efficient

use of the limited power available from solid-state devices—especially at millimeter-wave frequencies—it is important to use a type of modulation which gives good noise immunity, that is, one which will provide an acceptable error-rate with relatively small signal-to-noise ratio.

Second, because the repeaters are to be regenerative, timing information must be provided at each repeater. While this can be accomplished in several ways, a very attractive way is to use a type of modulation which allows the repeater to extract timing directly from the signal regardless of message statistics. This eliminates the necessity of sending timing information on a separate channel or of including pulses into the bit stream to insure a timing signal even in the event the message causes a particularly unfavorable pulse pattern.

Third, since the system is to operate at very high bit rates, it is important that the modulation scheme be one which can be implemented with a minimum of circuitry.

Finally, the modulation scheme must not be excessive in its bandwidth requirement even though, due to the large bandwidth capability of the waveguide, one is willing to make a reasonable trade of bandwidth for noise immunity.

The optimum noise immunity (consideration 1) would be achieved with binary coherent phase-shift-keyed modulation.⁹ However, this type of modulation requires that a reference phase be provided at each repeater. The need for a reference signal is eliminated by using a differentially-coherent signal at a cost of less than 0.5 dB in noise immunity at acceptable error rates (the order of one error in 10^9 bits), as can be calculated from the equations in a review paper by J. G. Lawton.¹⁰

2.2 Description of FM-DCPSK

A modulation scheme which we have designated FM-DCPSK (Frequency-Modulated Differentially-Coherent Phase-Shift-Keyed) modulation was conceived as a reasonable compromise among the four considerations. FM-DCPSK is a hybrid of frequency modulation and differentially-coherent phase-shift-keyed modulation. The signal has constant amplitude and is angle modulated in such a way that the information is carried in the relative phase, i.e., the phase shift between adjacent sampling instants. Optimum noise immunity occurs when the two possible signal states in a given time slot differ in phase by π . This can be achieved by shifting the phase by an amount $+\pi/2$ or

$-\pi/2$ between successive time slots (the choice of the sign depending on the message). A signal-space diagram is shown in Fig. 2. Fig. 3 shows the variation of phase and frequency resulting from modulation with the binary train indicated at the top of the figure. It will become apparent when we discuss repeater circuits in Section III that the simplicity consideration is satisfied by the FM-DCPSK signal.

Modulation from polar binary baseband to carrier IF is performed directly with an FM deviator. No flip-flop or other binary to differential-binary translator is required because of the differential relationship between frequency and phase. The deviator linearity is unimportant since only the area under the frequency-versus-time curve is important. The constant-amplitude continuous-phase nature of the signal allows phase-locked oscillators to be used for gain and limiting.

Because there is a phase change (hence, a frequency swing) in each time slot, regardless of the signal statistics, timing information is available in the signal itself. This can be readily extracted by means of a frequency discriminator and a narrow bandpass filter as described in Section 3.9. Finally, the bandwidth which gives optimum results is found experimentally to be slightly larger than the bit rate, which is in agreement with theoretical calculations for frequency modulated systems by R. R. Anderson and J. Salz.¹¹

III. CIRCUITRY FOR THE REPEATER

3.1 Introduction

The repeater circuit is shown in block diagram form in Fig. 4. Fig. 5 is a photograph of the repeater. This subsection will give a brief introductory discussion of the layout and operation of the repeater. Detailed discussion of the operation of various components is deferred to the following subsections.

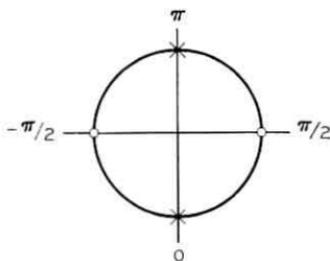


Fig. 2 — Signal space diagram for binary FM-DCPSK $\phi_n = 0$ or π for $n = n_0, n_0 + 2, n_0 + 4, \dots, \phi_n = +\pi/2$ or $-\pi/2$ for $n = n_0 + 1, n_0 + 3, n_0 + 5, \dots$

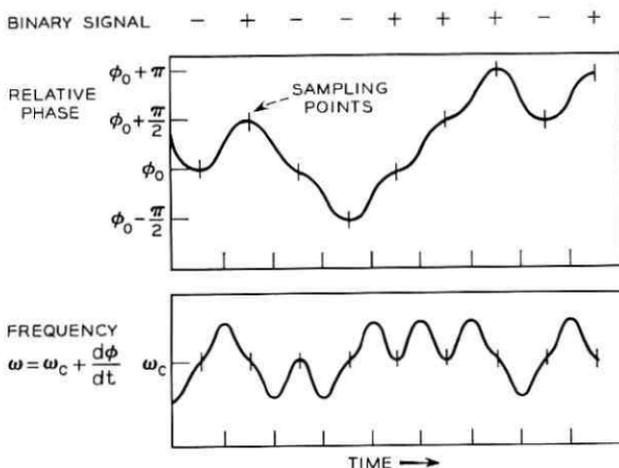


Fig. 3 — Phase and frequency variations of a FM-DCPSK signal.

The signal enters in the TE_{01} mode in 2-inch circular waveguide and first encounters a band-splitting filter which divides the 60-GHz band of the waveguide into two sub-bands. The signal in each of these sub-bands next encounters a channel-dropping filter which drops the individual channel for the individual repeater. In an actual system, several (perhaps as many as six) band-splitting filters would be used and a string of several (perhaps as many as 30) channel-dropping filters would follow each band-splitting filter. The first component which the individual signal encounters after the channel-dropping filter is a down-converter which translates the millimeter-wave signal frequency to the 1.3-GHz IF frequency of the repeater. The down-converter is followed by a low-noise transistor amplifier which provides approximately 52 dB of gain. This amplified signal is then used to lock an oscillator which serves as a limiter. The output of this limiter is amplified by a second transistor amplifier having 27 dB of gain. The next component is a combination differential-phase detector and timing recovery circuit. This component provides both a timing signal which consists of a sine wave at the bit frequency recovered from the transmitted signal and a baseband information signal whose polarity depends on the binary information transmitted. This polar baseband signal is then applied to a regenerator along with the timing signal and the regenerator makes a decision as to which of the binary states was transmitted in each time slot. The output of the regenerator drives an FM deviator which provides an

angle-modulated signal at IF. This signal is amplified by a third transistor IF amplifier and up-converted to the original millimeter-wave frequency. This millimeter-wave signal is now combined with the signals in other channels by means of a series of channel-adding filters and band-combining filters which are identical to the channel-dropping and band-splitting filters used at the input. The output is again in the TE_{01} mode in 2-inch circular waveguide.

3.2 Band-Splitting Filters

The band-splitting filters perform the function of splitting the 40-GHz to 100-GHz band into relatively wide sub-bands. The devices used for this purpose have been described in detail by Marcatili and Bisbee.¹² For completeness, their scheme will be reviewed briefly. Fig. 6 shows a constant resistance filter made up of two hybrids connected together by two identical high-pass filters.

Power entering port 1 is equally split by the hybrid H_1 with each

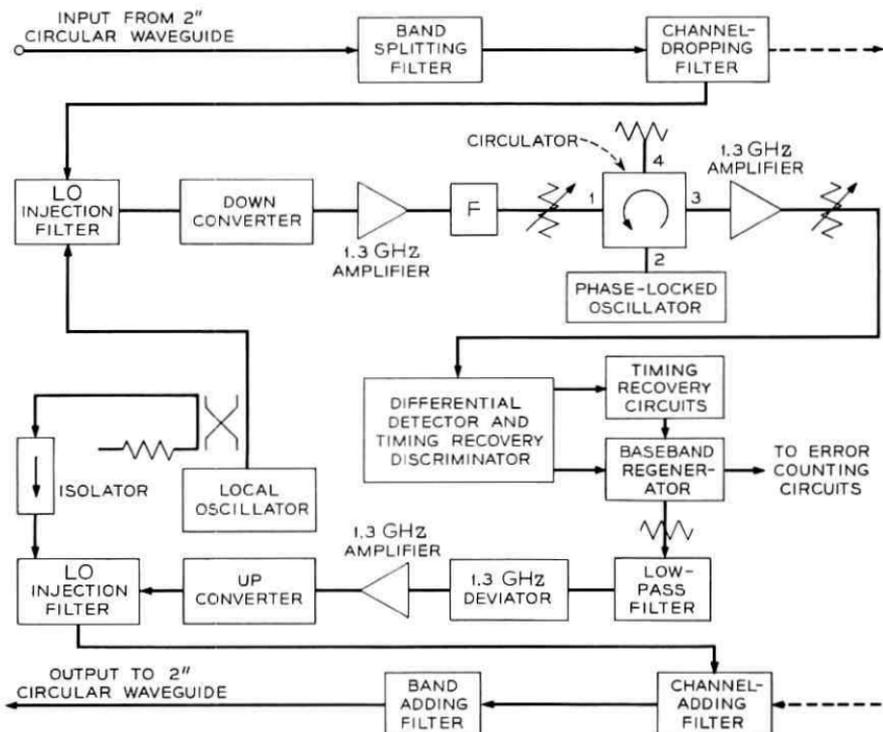


Fig. 4 — Repeater circuitry.

half propagating through equal line lengths toward the high-pass filters. Frequencies above the cutoff frequency of the high-pass filters pass through the filters unattenuated, are recombined in the second hybrid and emerge at port 4. Frequencies below the cutoff frequency of the high-pass filters are reflected back towards the first hybrid where they recombine and emerge at port 2. Marcetili and Bisbee realized this structure in low-loss TE_{01} circular electric mode components.

The hybrids developed consist of a right angle tee junction of two, 2-inch i.d. round waveguides with a thin sheet of dielectric material placed diagonally across the junction. The system can be analyzed on a quasi-optical basis with the result that proper selection of the dielectric material produces hybrid performance.

The high-pass filters used were TE_{01} mode guides with cutoff frequency equal to the splitting frequency. They were coupled to the 2-inch helix guides by means of helix waveguide tapers. Experimental results on a composite filter show that the maximum loss in either sub-band is 1.5 dB and that the transition region takes up only 160 MHz of the spectrum.

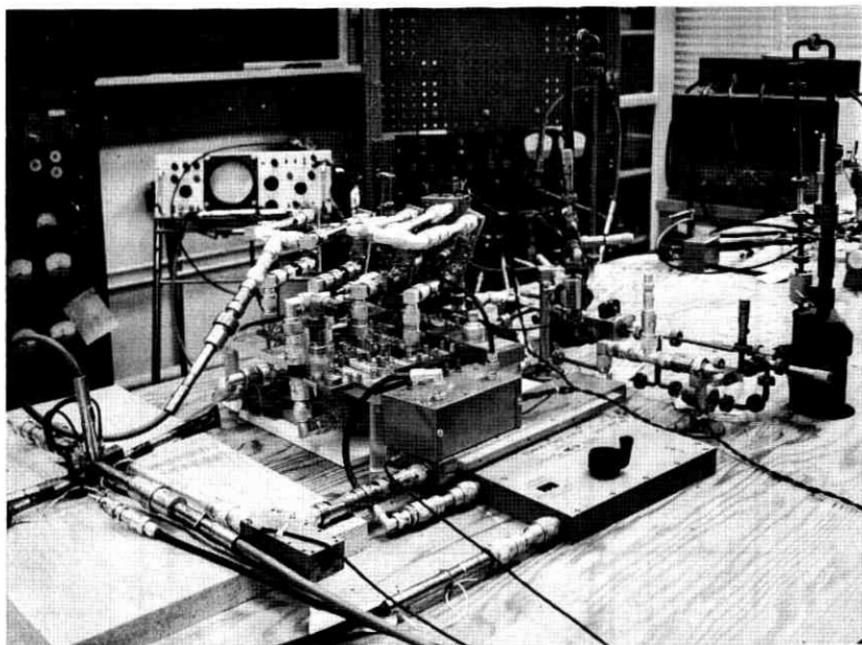


Fig. 5 — Experimental model of millimeter-wave repeater.

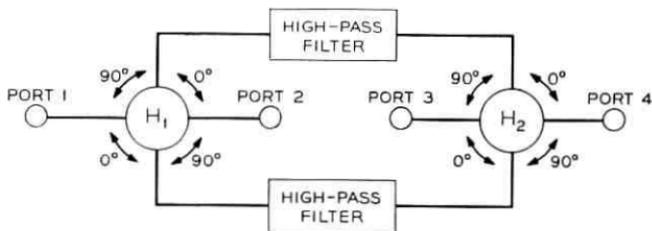


Fig. 6—Band-splitting filter. H_1 and H_2 are hybrids (taken from Ref. 12).

3.3 Channel-Dropping Filters

The requirements for the channel-dropping filters are determined by such factors as tolerable insertion loss, intersymbol interference, and interchannel interference. Explicit analysis of intersymbol interference and inter-channel interference problems associated with the type modulation used is as yet incomplete. Hence, a procedure for optimizing channel to channel spacing is not available.

For the experimental repeater, attention was directed to two-pole, wideband channel-dropping filters because they afford a significant reduction in required channel spacing relative to that for single pole filters. A bandwidth of 1 GHz was chosen to prove the flexibility of the design procedure. The theory developed¹³ employed narrow bandwidth approximations; thus, the design of filters having smaller bandwidth would be no problem. As stated in Section 2.2 an overall channel bandwidth slightly greater than the bit rate yields optimum error rate. Based on this fact, consideration of all of the band-limiting elements in a given channel indicates that channel-dropping filter bandwidths of less than twice the bit rate will be adequate.

Fig. 7 shows the physical structure and identifies the resonant elements of the channel-dropping filters. Ports 1 and 3 are the circular mode input and output ports, respectively. Port 2 is the dropped (or added) channel output (or input) port. The input and output guides are above cutoff for the TE_{01} mode and just below cutoff for the TE_{02} mode. The large guide sections are just above cutoff for the TE_{02} mode. The rectangular waveguide output is coupled to the mode-conversion resonator nearest the input by means of a wrapped resonator of rectangular cross section.

A qualitative description of the behavior of the structure is as follows. First, consider an individual rejection resonator. A signal incident in the TE_{01} mode is coupled to the TE_{02} mode by means of a symmetrical diameter discontinuity. Since the input and output

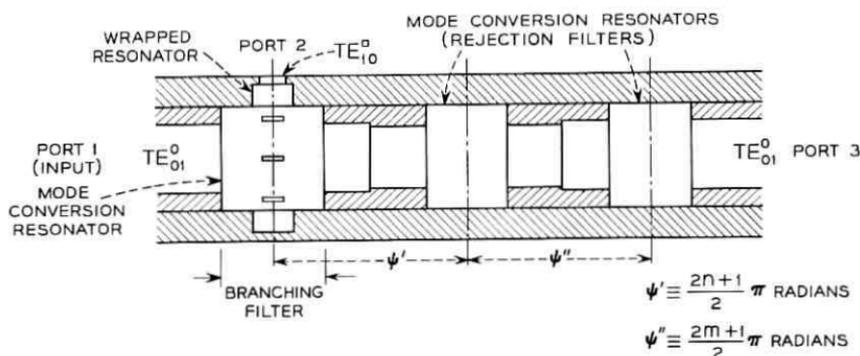


Fig. 7 — Cross section view of channel-dropping filter.

guides are below cutoff for the TE_{02} mode, the power in that mode is trapped in the large diameter region. Marcattili's analysis of the structure shows that at resonance the transverse mid-plane of the resonator is effectively a short circuit.¹⁴ The center frequency and bandwidth are dependent on the length of the resonator and the ratio of the input guide diameter to the resonator diameter. The details of the relationship are given by Marcattili.¹⁴

In the structure of Fig. 7, the mid-planes of adjacent mode-conversion resonators are electrically separated by odd multiples of $\pi/2$ radians. Hence, at resonance, the rejection-resonator pair presents an open circuit at the mid-plane of the input mode-conversion resonator. All of the incident TE_{01} mode power appears at the rectangular waveguide output when the various coupling coefficients are properly chosen.

Further insight into the electrical behavior of the structure is obtained by considering the prototype network shown in Fig. 8. The prototype network consists of complimentary admittances connected in shunt. The elements of the network have been chosen to yield a two-pole, maximally flat insertion loss response between ports 1 and 2 while maintaining a constant input admittance as a function of frequency. Total power transfer occurs at zero frequency, and half-power transfer occurs at an input angular frequency of one radian per second. The prototype network is converted to a network having total power transfer at some frequency ω_0 through use of the angular frequency mapping function

$$\omega' = Q_L \left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega} \right), \quad (1)$$

where

- ω' = angular frequency variable of the prototype network
 ω = angular frequency variable of the desired network
 Q_L = $\omega_0/(\omega_1 - \omega_2)$
 ω_1, ω_2 = half power angular frequencies of the desired network.

For the purpose of obtaining a qualitative understanding of electrical behavior it is sufficient to state that the performance of the microwave structure will be identical to that of the frequency-mapped prototype network subject only to the approximations involved in relating their respective parameters.

Four filters were constructed for use in the repeater system. The results were consistent from filter to filter. Figs. 9 and 10 show a set of typical characteristics. The insertion loss to the dropped channel is about 0.5 dB. The theory predicted an overall bandwidth of 1.13 GHz. The agreement between measured and theoretical values is good.

3.4 Solid-State Millimeter-Wave Power Sources

Three different solid-state millimeter-wave power sources were used. They were an LSA diode oscillator, an harmonic generator and an IMPATT diode oscillator.

The first solid-state device used successfully in the repeater was an LSA oscillator.⁸ The diode used in this experiment required dc power of 0.4 amps at 3.5 volts and delivered 4 mW of power at 50.4 GHz. (Similar units which deliver 20 mW at various frequencies in the 40- to 100-GHz band have been built by Copeland.) The LSA oscillator was used in all of the error-rate and gain experiments described in Section V.

The varactor quadrupler uses a zinc diffused gallium arsenide diode.¹⁵ The diode was a planar array structure similar to the "honeycomb"

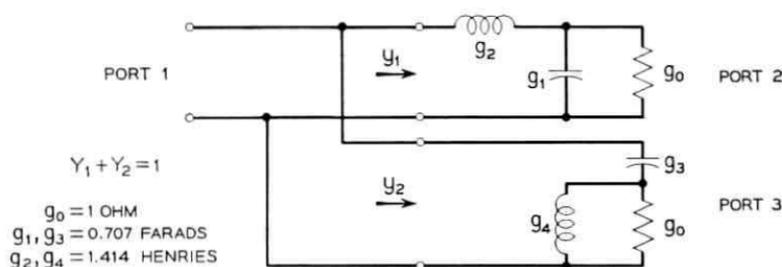


Fig. 8 — Prototype network for a two-pole diplexer.

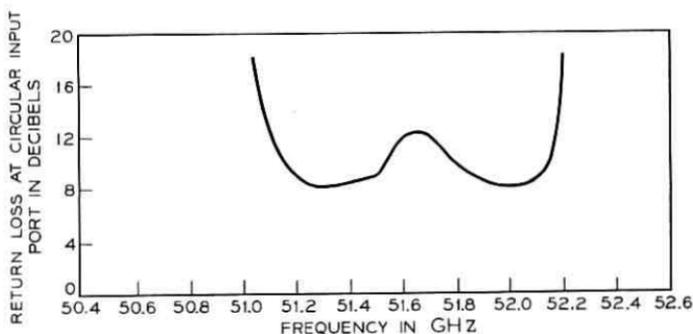


Fig. 9—Return loss at port 1.

type described by Young and Irvin.¹⁶ It was mounted in a Sharpless wafer as shown in Fig. 11. The input signal frequency was 12.6 GHz. The power output and overall efficiency of a typical unit are shown in Fig. 12. The maximum power output obtained was 10 mW at an efficiency of 6.5 percent. The input VSWR was less than 2 to 1 and the output VSWR was about 7 to 1. The power source for the quadrupler was an IMPATT oscillator which provided a 12.6-GHz signal.

The millimeter-wave IMPATT diode delivers approximately 50 mW at 50.4 GHz. (Diodes of this type which deliver 130 mW at 70 GHz have been built by T. Misawa.¹⁷)

Each of these power sources gave an overall performance as good as that obtained from an Oki Klystron.

3.5 Local Oscillator (LO) Injection Filters

The local oscillator power is coupled to the up- and down-converters by means of a three port diplexer as shown in Fig. 13. The local oscil-

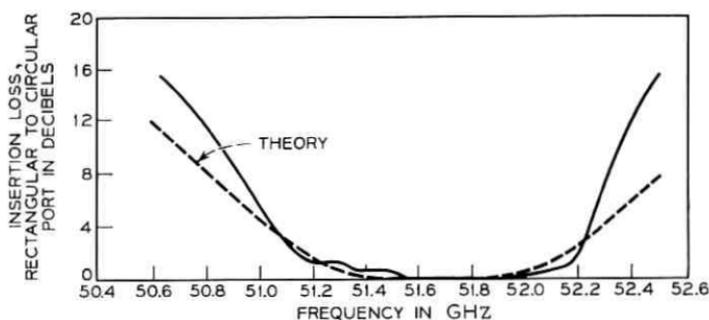


Fig. 10—Insertion loss from port 1 to port 2.

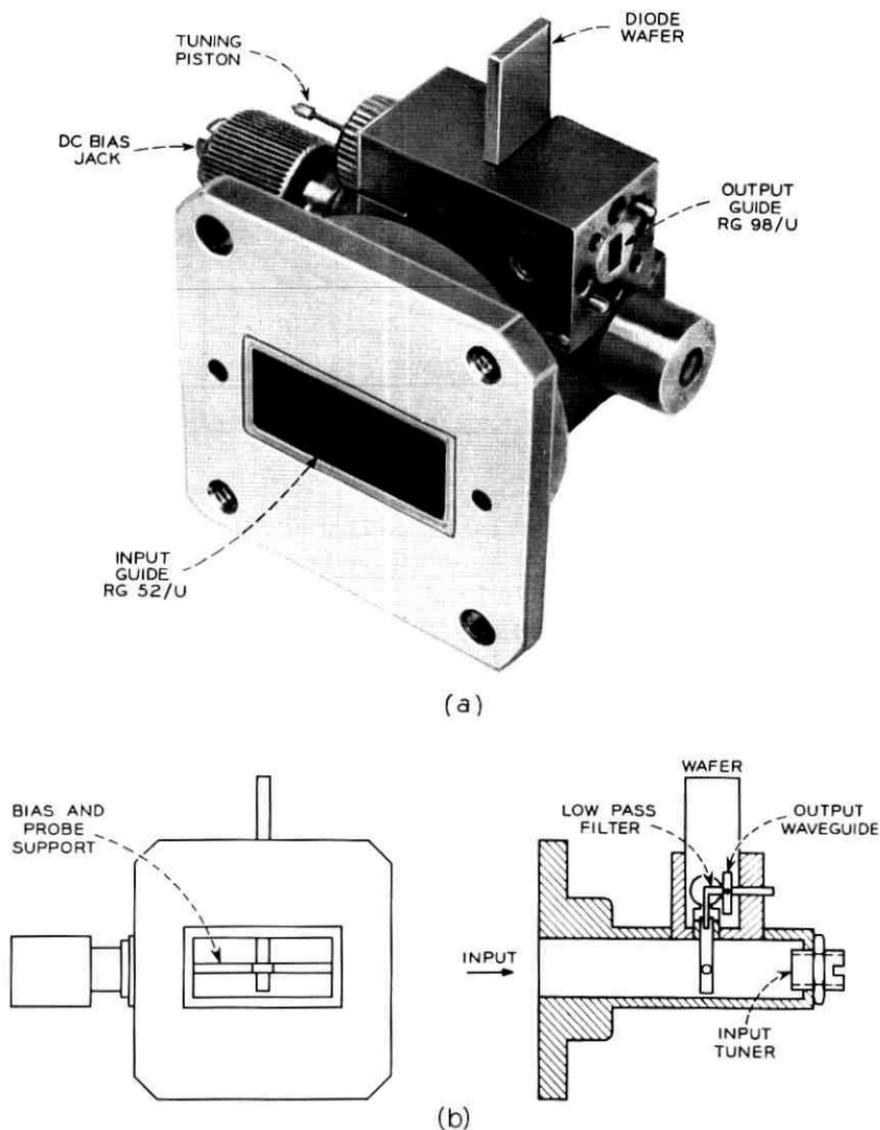


Fig. 11—The X-band to millimeter-wave band quadrupler. (a) Photograph. (b) Cross-section view.

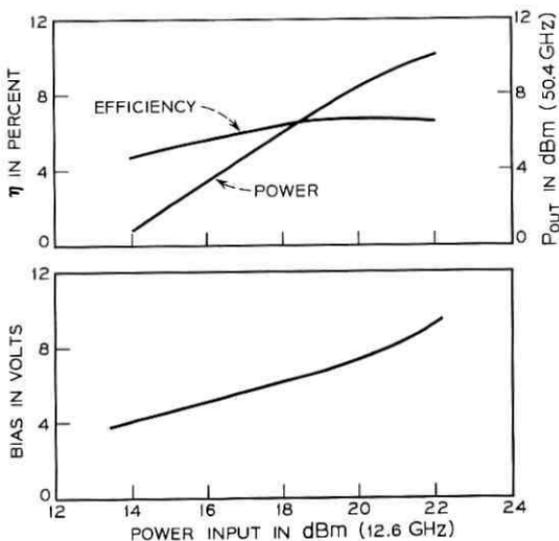


Fig. 12 — Power output and efficiency of quadrupler.

lator power is injected at the bandpass port (port 1) and the signal at the band rejection port (port 3). The up (or down) converter is connected to the constant resistance port (port 2).

The construction of an efficient millimeter wave diplexer was accomplished by utilizing two low-loss TE_{011} circular cylindrical cavity mode resonators as shown in Fig. 14. The device operates as follows. Both resonators are tuned to the local oscillator frequency. At resonance, the rejection cavity effectively open circuits the waveguide. The bandpass resonator (dropping filter) is located an odd number of quarter wavelengths from the transverse symmetry plane of the rejection resonator. Hence, at resonance, a short circuit appears to exist at plane A of Fig. 14.

Ideally, proper adjustment of the coupling apertures yields coupling

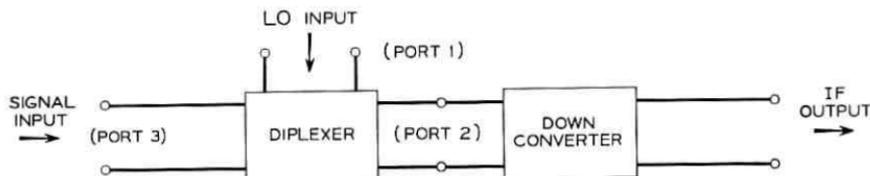


Fig. 13 — Schematic of the local oscillator injection arrangement.

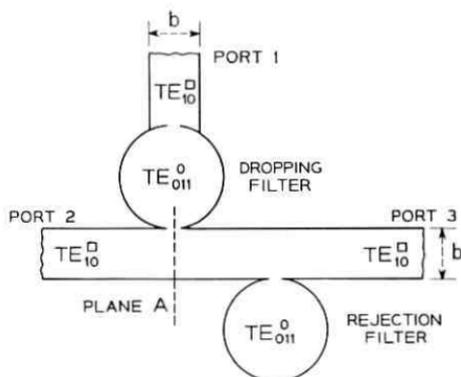


Fig. 14 — Physical realization of local oscillator injector filter.

of 100 percent of the LO power to port 1 in the absence of dissipation. The details of the design procedure are given in Ref. 18.

The requirements on the bandwidth of the diplexer were established by considering the tolerable dissipation of LO power at resonance, and the transmission loss through the diplexer over the signal band. The tolerable LO loss was set at 2 dB maximum based on the millimeter-wave power available from the solid-state source and the LO power requirements established for the up- and down-converters. Minimum attenuation to the signal band is achieved when the bandwidth is at a minimum consistent with the LO loss requirement. Experimental work indicated that a 50-MHz bandwidth in the power transfer from the LO input port to the converter port was about optimum.

Four diplexers were constructed with consistent results. The insertion loss to the LO averaged 1.4 dB. The return loss at the signal input port was better than 25 dB over the signal band. The return loss looking into the converter port was better than 15 dB at all frequencies of interest. Figs. 15 and 16 show typical frequency responses at the various ports.

3.6 Down-Converters

The down-converters developed for the system had the following characteristics:

- (i) IF frequency band from 1.0 to 1.6 GHz
- (ii) LO frequency of 50.4 GHz at a power level of -3 dBm
- (iii) Input signal from 51.4 to 52.0 GHz
- (iv) Conversion loss of 6.0 ± 0.5 dB over the above band
- (v) Converter noise temperature ratio of nearly unity.

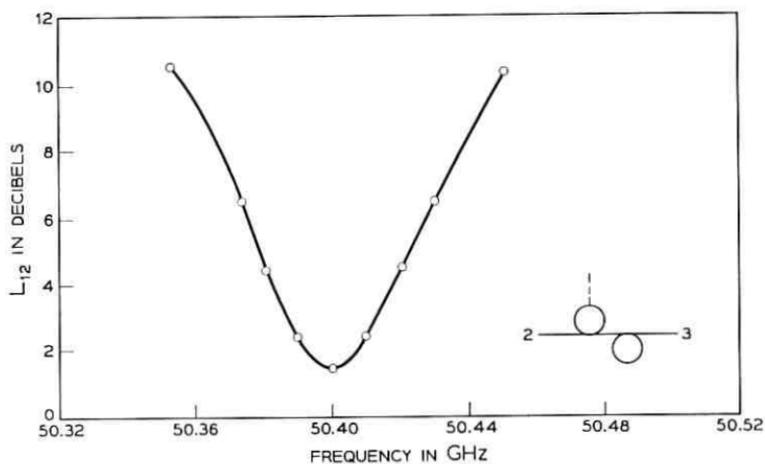


Fig. 15 — Pass-band response for local oscillator injection filter (ports 1 to 2).

This performance was achieved using Schottky barrier diodes at a fixed dc bias.

The basis for the design was the converter mount described by Sharpless.¹⁹ The only modification required was the addition of an IF impedance matching network. The millimeter-wave portion of the structure was not changed. Fig. 17 shows the structure. The following paragraphs describe the equivalent circuit and give a brief discussion of the design procedure.

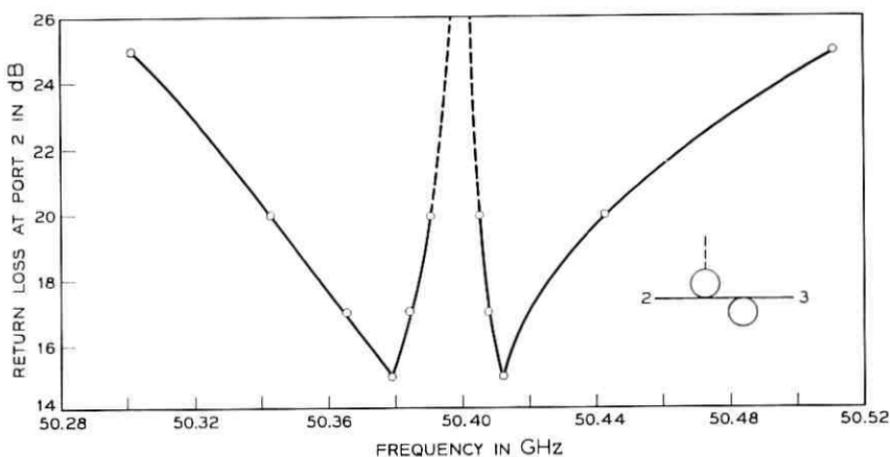


Fig. 16 — Return loss of local oscillator injection filter at port 2.

The equivalent circuit for the structure is shown in Fig. 18. The IF input admittance Y_L was measured over the band with the LO on and the bias fixed. It was found that Y_L could be closely approximated by a constant conductance shunted by a capacitive susceptance. At mid-band

$$Y_L = (59)^{-1} + j(150)^{-1} \text{ mhos.}$$

The admittance Y_L was matched to 50 ohms at mid-band by a short length of transmission line having a characteristic impedance of 83 ohms. This was followed by a biasing tap consisting of a quarter wavelength 50-ohm stub by-passed to ground. The excellent broadband behavior of the completed circuit is indicated by the small variation of conversion loss over the band. The latter is shown in Fig. 19.

3.7 IF Amplifiers

Wideband transistor amplifiers (with a center frequency of 1.3 GHz) of the balanced integrated circuit type originally developed by Engel-

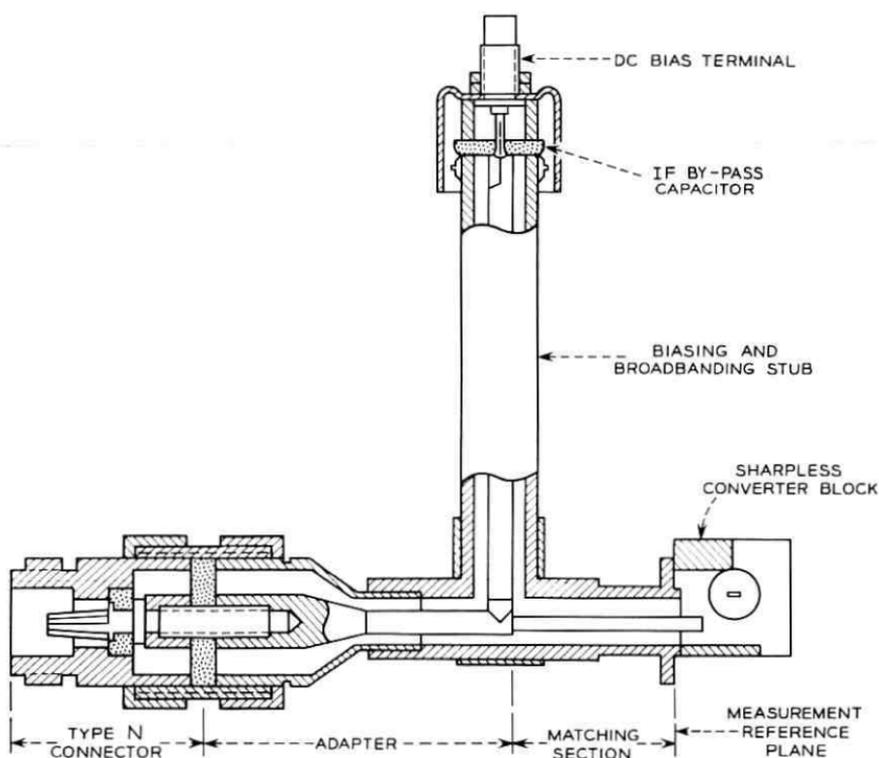


Fig. 17 — Down-converter structure.

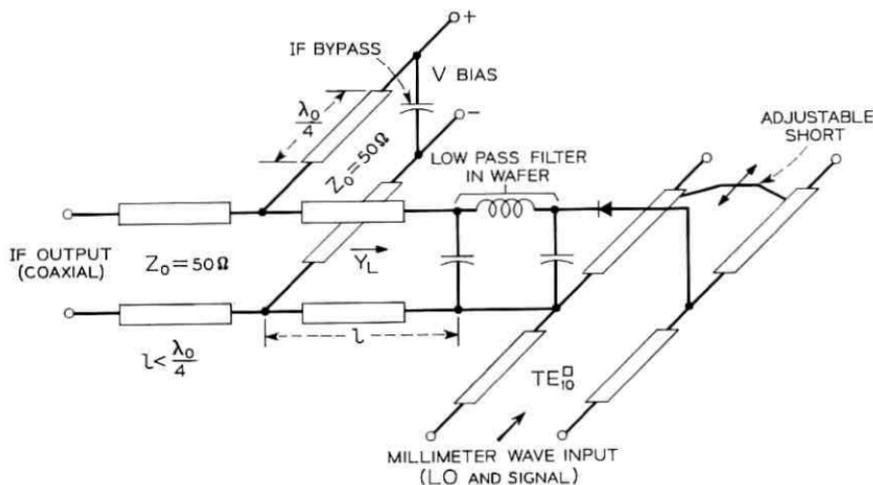


Fig. 18 — Down-converter equivalent circuit.

brecht and Kurokawa²⁰ were used in this repeater. Fig. 20, which is reproduced here from Ref. 20, shows the basic amplifier circuit. These amplifiers can be built with excellent noise figures (less than 4 dB) and, because of the excellent match between sections, can be cascaded to achieve high gain. These amplifiers exceeded the required specifications in all respects and have proved entirely satisfactory for this repeater.

3.8 Limiter

Because of the nature of the regenerator which is used in this repeater, an improvement in error-rate performance for a given signal-to-noise ratio is expected from the inclusion of a limiter ahead of the differential-phase detector.²¹ A simple but effective limiter was

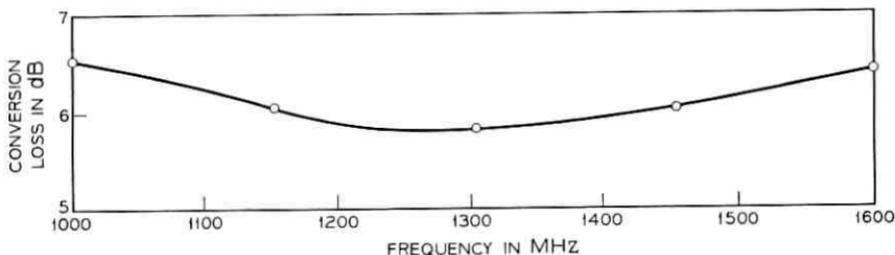


Fig. 19 — Conversion loss of down-converter.

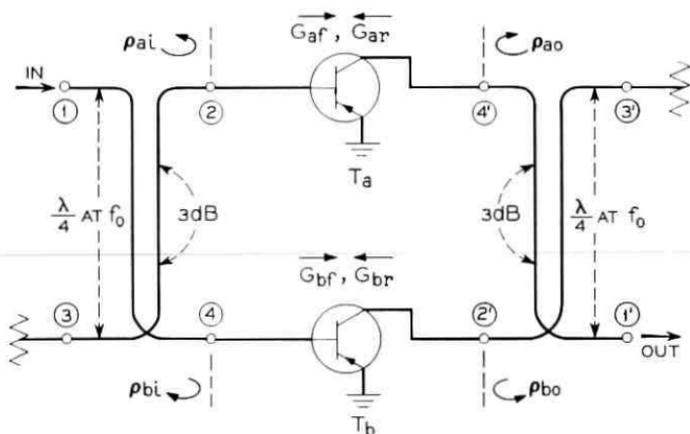


Fig. 20 — Schematic representation of single-stage balanced amplifier. (Taken from Ref. 20).

achieved by the use of a tunnel-diode oscillator built in 50-ohm coaxial transmission line and phase-locked to the IF signal (see Fig. 21). The tuning of the diode was inductive and was accomplished by means of a shorted 75-ohm transmission line stub. The tunnel diode was of germanium point-contact construction and had a peak current of 2 mA. Fig. 22 shows the output power of the oscillator versus gain (ratio of output power to input power) at the center frequency of the oscillator. Best error performance was found experimentally to occur at a gain of 8 dB. The change in output power with frequency for several values of gain is shown in Fig. 23.

3.9 Differential Phase Detector and Timing Recovery

The baseband and timing circuits are shown in Fig. 24. The couplers, delay lines, and diode mounts are microwave printed circuits; the filters and combining Tee are coaxial. The differential phase detector and the timing recovery circuit are combined (share a common delay line) in order to save space and cost.

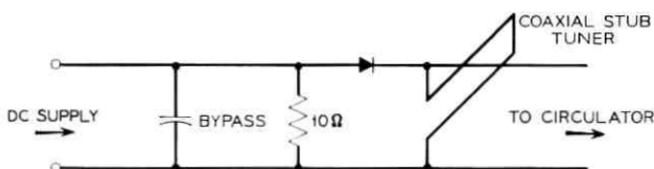


Fig. 21 — Limiter circuit.

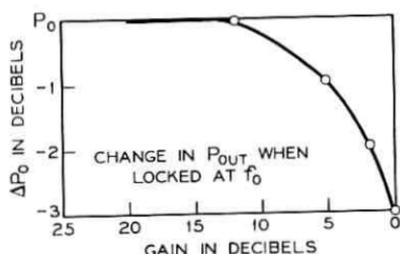


Fig. 22—Output power of limiter vs gain at the center frequency.

Fig. 25 shows the basic differential-phase detector or timing-recovery circuit. A straightforward analysis (see the Appendix of Ref. 22) shows that the output voltage of this circuit is given by

$$V(t) = \cos \left\{ \omega_0 t + \int_{t-\tau}^t \omega(t') dt' \right\} \quad (2)$$

for an input FM-DCPSK signal given by

$$S(t) = \sqrt{2} \cos \left\{ \omega_0 t + \int_0^t \omega(t') dt' \right\}, \quad (3)$$

with

$$|\omega(t)| = |\omega(t + nT)|, \quad \int_{(n-\frac{1}{2})T}^{(n+\frac{1}{2})T} \omega(t') dt' = \pm \pi/2.$$

One can readily see that if $\omega_0 \tau$ is chosen to be a multiple of π , $V(t)$ is independent, to first order, of the sign of $\omega(t)$; hence, the output is periodic in period T , where T is the reciprocal of the bit rate. By proper choice of $\omega_0 \tau$ a signal is obtained with a strong frequency component

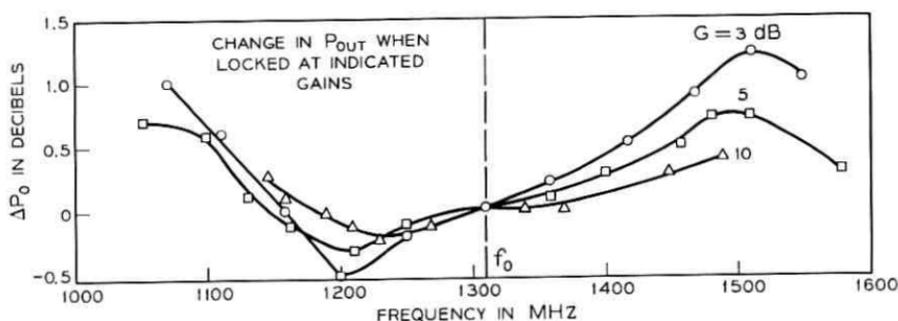


Fig. 23—Change in output of the limiter vs frequency for several values of gain.

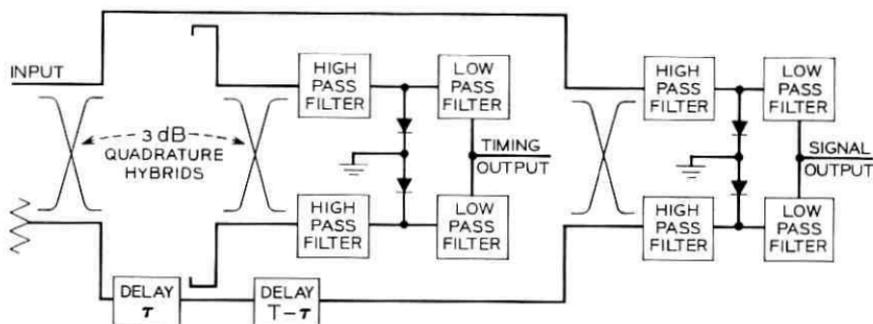


Fig. 24 — Differential phase detector and timing recovery circuit.

at the bit rate. This signal is used to lock an oscillator at the bit rate. The oscillator, in turn, provides the timing signal to the regenerator.

If $\omega_0\tau$ is chosen to be an odd multiple of $\pi/2$ and τ is chosen equal to T , the reciprocal of the bit rate, one sees from (2) and (3) that at the sampling instants, $[t = (n + \frac{1}{2})T]$, the output is given by

$$V(t) = \cos \{(m + \frac{1}{2})\pi \pm \pi/2\} = \pm 1.$$

Thus, under these conditions the device is the desired differential-phase detector for this signal.

3.10 Regenerator

The regenerator consists of a balanced-line logic element²³ which is a modification of the standard Goto-pair circuit. The input signal is applied at the midpoint between the two tunnel diodes and a timing signal is applied across the pair of tunnel diodes as shown in Fig. 26. Ideally, the timing signal causes one and only one of the diodes to switch once each time-slot. The input signal determines which of the two diodes switches. When one of the diodes switches, the resultant voltage drop across the other diode inhibits its switching and this voltage drop occurs across the output of the re-

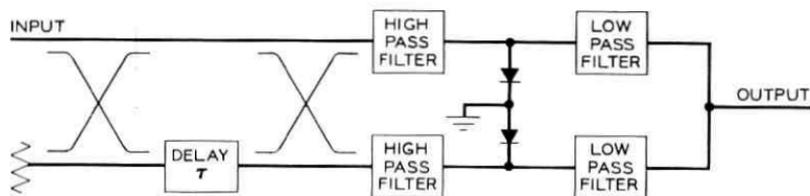


Fig. 25 — Basic differential phase detector or timing recovery circuit.

generator. If the diode labeled D_1 in Fig. 26 switches, the voltage pulse at the output of the regenerator is negative and, correspondingly, if D_2 switches the voltage pulse is positive. The transients initiated by the switching of the diode travel down the delay lines and are reflected with inverted polarity back to the diodes by the low-impedance termination. These reflected signals reset the diode to its original condition. Thus, the information content of the signal is translated into a sequence of polar baseband pulses at the regenerator.

E. G. Herzog²⁴ has discussed limitations on the speed of operation of the Goto-pair. His conclusions also apply to the balanced-line logic element. He showed that for bias voltages above a certain critical value the Goto-pair has a stable zero output state in addition to stable positive and negative output states. Due to the junction capacitance and the series inductance of the diode, it takes a finite time (talking time) for one diode to indicate to the other that it has switched. Thus, with a small input signal each diode can go to its second positive resistance region and if the synchronizing voltage passes the critical value too soon they will be left there and the zero output state will result. Also, we have observed that if the voltage does not pass the critical values soon enough both diodes will return to their first positive resistance region resulting in an intermediate amplitude output. In order to minimize the probability of occurrence of these undesirable operations, the following properties are desirable for the diodes: First, the diode must have adequate peak current (if the peak current is too

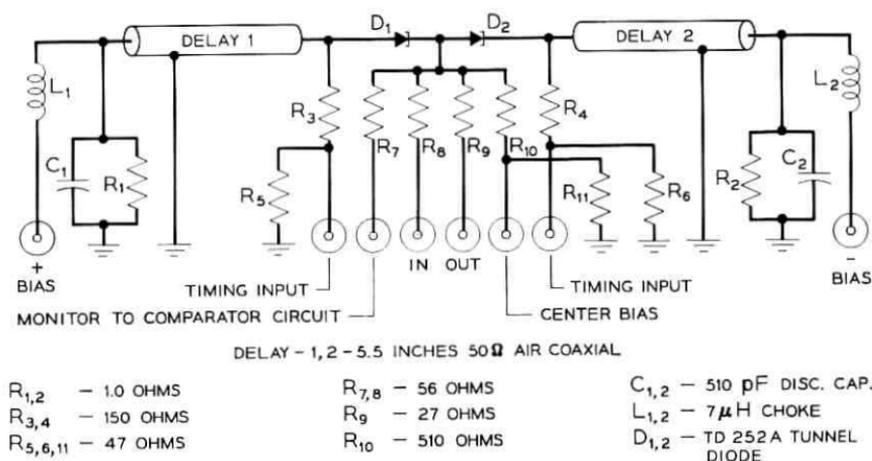


Fig. 26 — Regenerator circuit.

low, the load will prevent bistable operation); second, it must have low junction capacitance (this decreases the talking time); third, the magnitude of the product of the negative resistance and the junction capacitance should be low enough to make the switching time short compared with a time-slot; and finally, the diode should have low series inductance (this decreases the talking time). For the regenerator, the third condition must be strengthened to make the switching time short compared to the round-trip time on the delay line. Since several round trips are required for the pulse to die out, the switching time of the diodes in the balanced-line logic element must be several times faster than for a Goto-pair.

TD-252A germanium tunnel diodes have been found to meet the above requirements. They have a series inductance of 1.5 nH, a peak current of 4.7 mA and a junction capacitance less than 1.0 pF. The diodes are mounted in the circuit of Fig. 26 in the manner shown in Fig. 27. This circuit has been built in such a way that the diodes are placed as close together as possible in order to eliminate lead inductance. By building the diodes into the transmission line, connector mismatches have been eliminated. The inductance of the leads of the input and output resistors has very little effect on the output pulse and it is believed that it steers the current from one diode to the other during the short time required for switching. Fig. 28(a) shows an eye diagram of a low signal-to-noise ratio input signal to the regenerator and Fig. 28(b) the resulting output signal of the regenerator. This figure illustrates the ability of the regenerator to remove noise from the signal.

3.11 *FM Deviator*

The FM deviator is basically a tunnel-diode relaxation oscillator. The frequency of oscillation of the tunnel diode in this type of circuit is extremely sensitive to bias voltage, allowing it to be driven by the balanced-line logic element. Fig. 29 shows the circuit. Tests on a low-frequency prototype circuit showed that the oscillator could be tuned over a bandwidth of more than half an octave in a time interval corresponding to less than 1 RF cycle. The total tuning range of the L-band deviator was greater than an octave, as shown in Fig. 30.

The circuit of Fig. 29 was built for use in L-band using conventional (as opposed to printed) circuits with the circuit dimensions kept as small as possible. Fig. 28(c) shows the differentially detected output eye of the deviator.

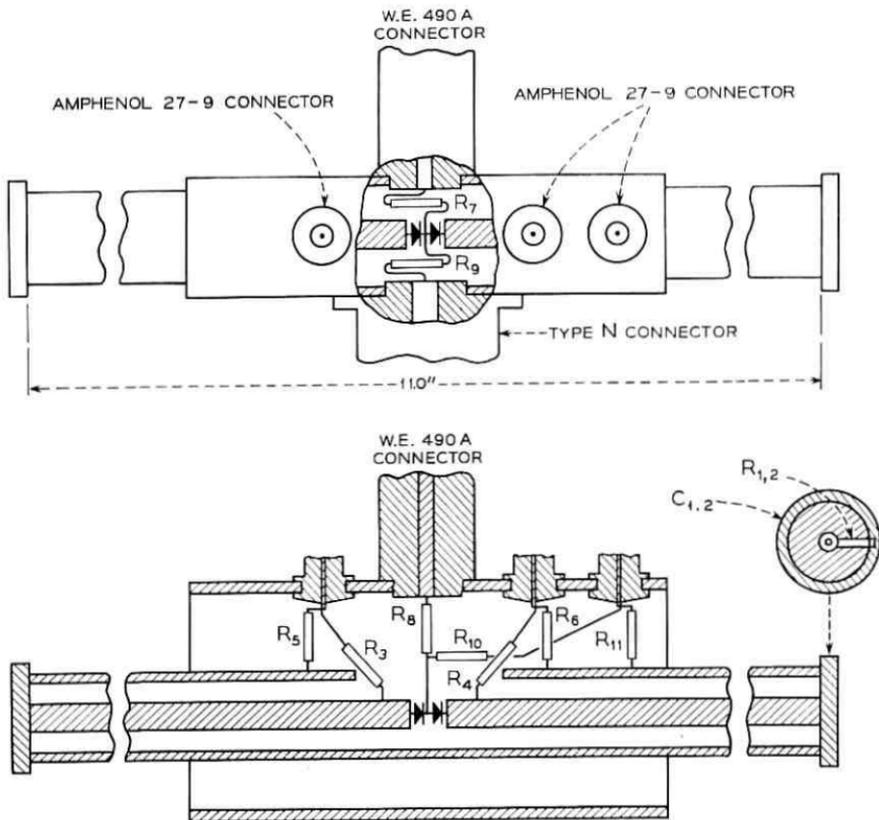


Fig. 27 — Mechanical layout of the regenerator.

3.12 Up-Converters

The chief goal in designing the up-converters was the maximization of output power over the band from 51.4 to 52.0 GHz when used with a local oscillator supplying + 3 dBm of power at 50.4 GHz. Typical units exhibited 6-dB LO to RF conversion loss across the band. Fig. 31 shows the frequency response of a typical unit. Both GaAs Schottky barrier diodes and planar diffused gallium arsenide varactor diodes were used with similar results—the latter exhibiting slightly lower conversion loss.

The physical structure for the units was similar in form to that of the down-converter described in Section 3.6. An *E-H* tuner preceded the converter block on the millimeter-wave side and was used to match the input impedance at the LO and signal band frequencies. All of the

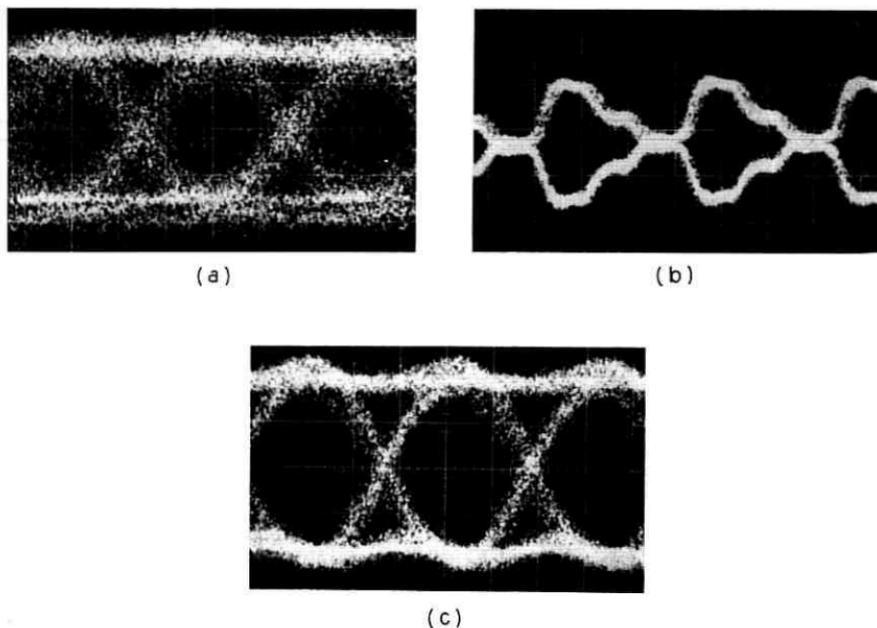


Fig. 28—Eye diagrams. (a) Degraded regenerator input. (b) Regenerator output. (c) Regenerated differentially detected IF.

diodes were operated at zero bias voltage. A more detailed description of planar diffused diode performance is given in Ref. 15.

3.13 Power and Space Requirements

The baseband circuitry of the repeater requires approximately 0.3 watts and the IF circuitry requires approximately 1.5 watts. Thus, the power requirement per channel per repeater is approximately 1.8 watts exclusive of the power required for the millimeter-wave power source.

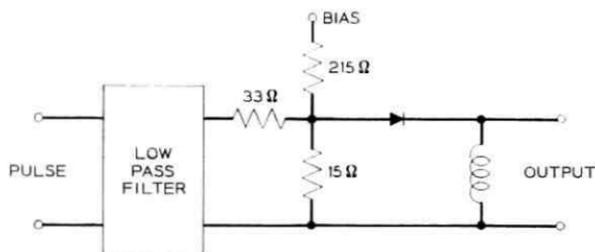


Fig. 29—Deviator circuit.

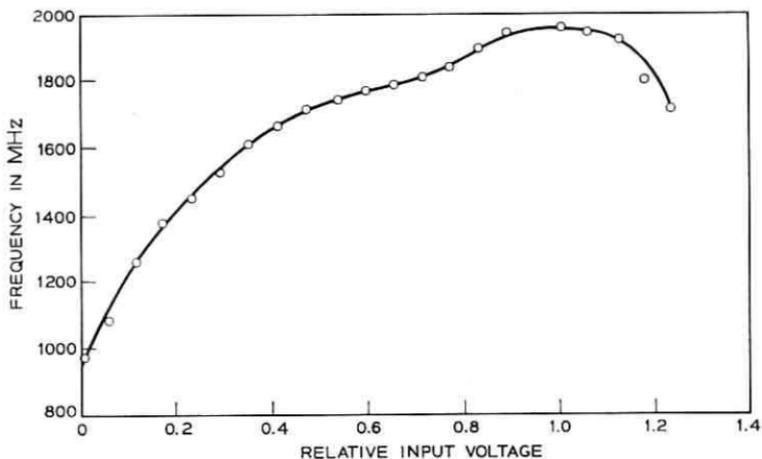


Fig. 30 — Deviator frequency vs relative input voltage.

The total power required per channel per repeater can thus be expressed by

$$\text{Total Power Required} = 1.8$$

$$+ \frac{\text{Millimeter-Wave Power Required}}{\text{Efficiency of Millimeter-Wave Source}} \text{ Watts}$$

The experimental repeater included many commercial components and no serious thought was given to miniaturization. Even so, it occupies only a volume of the order of 2 cubic feet (exclusive of band-splitting filters). With printed circuit techniques, the total volume per channel per repeater can be of the order of 0.5 cubic feet or less.

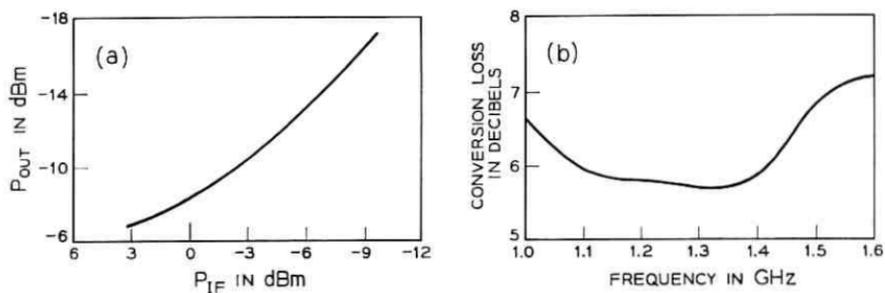


Fig. 31 — Varistor up-converter data. (a) Output versus input data. (b) Frequency response ($P_{IF} = +6$ dBm).

IV. SYSTEM CONSIDERATIONS

4.1 Error-Rate as a Function of Signal-to-Noise Ratio

The error-rate versus signal-to-noise ratio has been calculated in a manner which includes the effects of non-ideal regeneration of the signal and of intersymbol interference.²² Some results of this calculation are shown in Fig. 32 for an ideal regenerator and for a regenerator which has a threshold of operation, T , 9 dB below the expected signal level, S . The term threshold of operation is defined as follows. Suppose that the expected value of the signal at the input to the regenerator is V_1 or $-V_1$. The regenerator will then regenerate a positive or negative output pulse according to whether the input is positive or negative. However, if magnitude of the signal is too small the regenerator will not function properly. The minimum voltage at which the regenerator will function properly is the threshold of operation.

It is impossible to consider quantitatively the effects of intersymbol interference unless the waveform of the signal is known accurately. However, it is plausible to assume that the intersymbol interference contributes phase shifts of the order of a few degrees in the sense described in Ref. 22. For that reason, Fig. 32 shows the results of error-rate versus signal-to-noise ratio for the case where there is no inter-

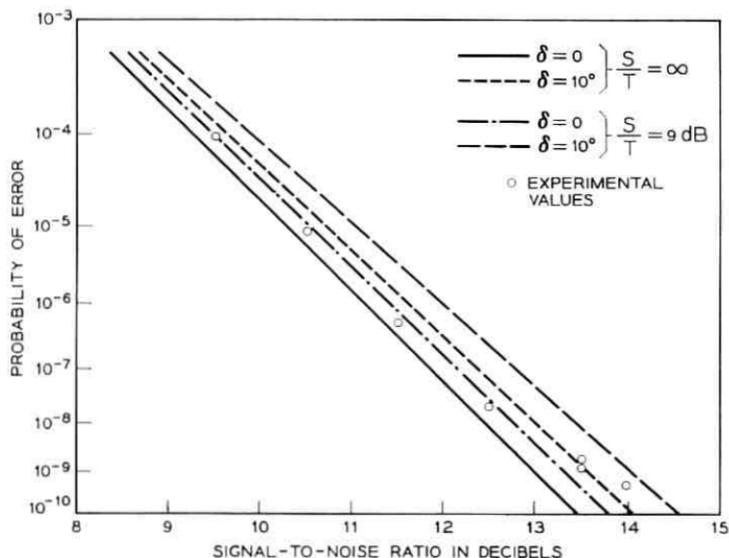


Fig. 32 — Error-rate vs Signal-to-noise ratio for the repeater.

symbol interference and for the case where the intersymbol interference corresponds to a phase shift, δ , of 10 degrees. These values should constitute the bounds on the expected error-rate. From Fig. 32 one observes that the expected value of S/N for 10^{-9} error rate lies between 13 and 14 dB.

4.2 Model of a System

Fig. 33 shows a model of a system which was used as an aid in the design of the repeater. This model is not an attempt to describe or design a complete system, it is intended only to give some perspective and insight into those factors which influenced the design of the repeater.

An actual system would use both frequency division and time division multiplex to separate individual voice channels. One possible arrangement of filters to separate the individual frequency multiplexed channels in the system is shown in Fig. 34. The skew arrangement of the filters is chosen to offset in part the variation of loss with frequency in the waveguide bandwidth. That is, since certain channels experience greater loss in the medium, the filters are arranged so as to give less loss to these channels at the expense of channels which have suffered less loss in the medium. Since the shape of the loss-versus-frequency curve is a function of repeater spacing (the relative loss in dB at two frequencies depends on the repeater spacing), the *details* of the arrangement of the filters are a function of repeater spacing. As an illustrative example, we assume a repeater spacing of 15 miles and attenuation curves for the medium given by Fig. 1. In addition, we

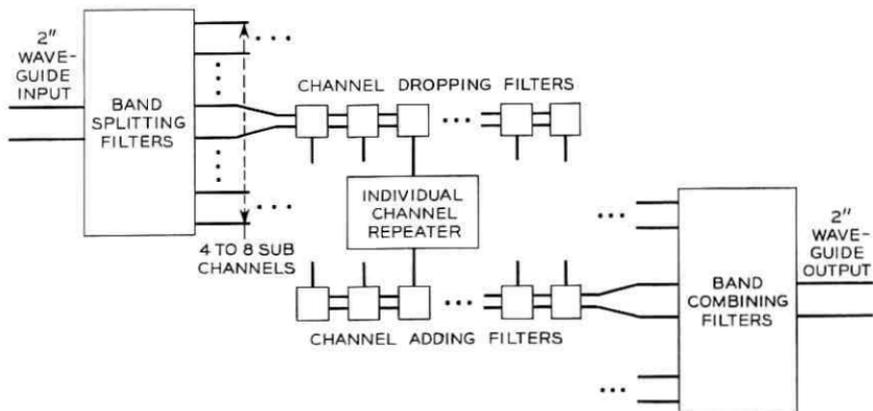


Fig. 33— Illustrative model of a system.

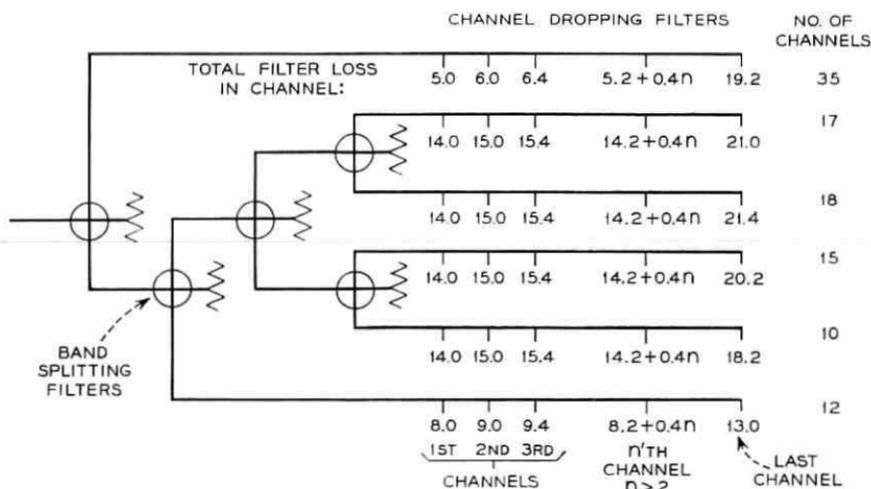


Fig. 34 — Channel-dropping filter array for the illustrative example.

make the assumption that the power available from realizable solid-state sources falls off at a rate of 3 dB per octave in frequency.²⁵ The loss of each band-splitting filter is taken to be 1.5 dB.¹² Based on the data of Figs. 9 and 10, conservative estimates of channel-dropping filter losses are 1.0 dB to the dropped channel, 0.5 dB to the adjacent channel and 0.2 dB to all other channels which pass through them.

Fig. 35 shows the waveguide loss as a function of frequency for a 15 mile repeater spacing. It also shows the power at the input to a repeater relative to the power at the output of the up-converter of the 50-GHz channel (based on the curve of Nutt in Fig. 1 and the assumed 3 dB per octave fall off in available power). The points in the figure then show the total relative signal power at the output of the channel-dropping filter for each channel (based on the filter arrangement shown in Fig. 34). The term "relative power" here means the power relative to that available at the output of the up-converter in a 50-GHz channel. From Fig. 35 one observes that in the worst case the relative signal level is -58 dB for this model. Thus, the repeater gain (defined as the ratio of the output power of the up-converter to the signal power at the input to the down-converter which gives an error rate of 10^{-9}) must be 58 dB. Since this is the value which gives the *maximum acceptable* error-rate, it seems expedient to include a 6-dB margin in the design of the repeater and thus the design goal for a 15-mile repeater spacing is a gain of 64 dB.

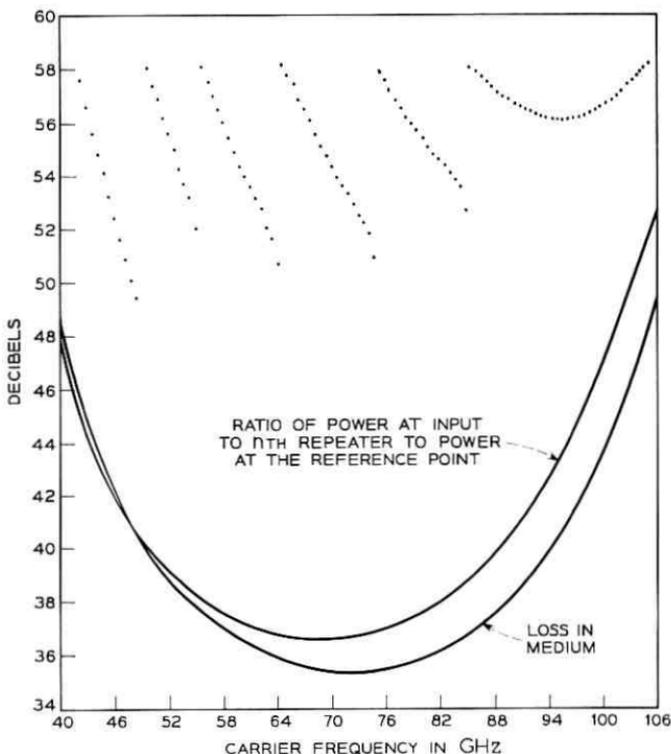


Fig. 35—Ratio of power at the output of the channel-dropping filters to power at reference point. The power reference point is the output of the up-converter of the 50-GHz channel of the preceding repeater.

In this model, there are 100 channels spaced at 600-MHz intervals across the band. The experimental repeater described here uses a 500-MHz bandwidth set by an inexpensive commercially available five-section Tschebycheff filter. (Only a slight degradation is experienced by using a 400-MHz filter of the same type.) Even smaller bandwidths might well be practical if suitable attention is given to the phase characteristic of the filters. Thus, the 600-MHz spacing assumed in the model is a conservative estimate. The capacity of the waveguide based on this model is 30,000 Mb/s. Since 72 Kb/s are required for each voice-grade circuit, the capacity of the system would be 416,000 voice-grade circuits or 208,000 two-way voice channels.

4.3 Theoretical Gain

The gain of the repeater in the sense in which it is used here can be expressed as

$$G = (P_{Lo} - L_{uc} - L_I) - (L_{DC} + F + KTB + S/N), \quad (4)$$

where

P_{Lo} is the local oscillator power,

L_{uc} and L_{DC} are the conversion losses of the up- and down-converters, respectively,

L_I is the loss in the isolator, the waveguide between the LO and the up-converter, and the injection filter

F is the noise figure of the first IF amplifier (since one finds experimentally that aside from conversion loss, the noise figure of the down-converter is negligible),

KTB is the thermal noise in the pass band of the IF section, and

S/N is the signal-to-noise ratio required for the acceptable error-rate.

As stated in Sections 3.12 and 3.16, one finds experimentally that

$$L_{uc} = 6 \text{ dB}, \quad L_{DC} = 6 \text{ dB} \quad \text{and} \quad F = 3.7 \text{ dB}.$$

Using a 500-MHz bandwidth, the thermal noise power is -87 dBm. Therefore, the required local oscillator power for 15-mile repeater spacing is

$$P_{Lo} = S/N + L_I - 7 \text{ dBm}.$$

If one assumes a value of 14 dB for S/N (from the discussion in Section 4.1) and 4 dB for L_I , he obtains 11 dBm as the required local oscillator power at the up-converter for a 15-mile repeater spacing. Since 0.5 mW of LO power is required for the down-converter, the total millimeter-wave power requirement for a 15-mile spacing is approximately 12 dBm.

V. EXPERIMENT AND RESULTS

5.1 Description of the Apparatus

The experimental apparatus used in the experiments to be described in Sections 5.2 and 5.3 consists of a transmitter, a receiver, and the repeater described in Section III as well as the necessary equipment and circuits for counting the errors made by the repeater, the receiver, or both. The transmitter is shown (in block diagram) in Fig. 36. The random-word generator consists of a regenerator of the type described in Section 3.10 driven by a similar regenerator which is, in turn, driven by differentially-detected wideband noise generated in a pair of X -band traveling-wave tubes. The random-word generator drives an FM deviator of the type described in Section 3.11. The remainder of

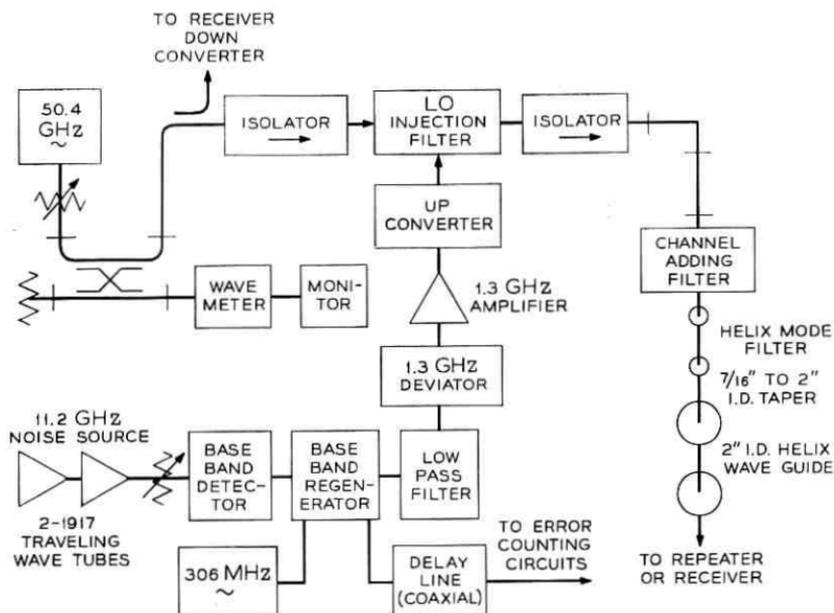


Fig. 36—Transmitter.

the transmitter is identical to the portion of the repeater which follows the FM deviator.

Just as the transmitter is a duplicate of the second half of the repeater, the receiver is a duplicate of the first half, beginning with the down-converter and ending with the regenerator. It is shown in Fig. 37.

All of the regenerators (including the random-word generator of the transmitter) are built with two outputs—one to drive the next following component in the circuit, the other to serve as a monitor port or as a source of pulses for error counting. The three pieces of apparatus, the transmitter, the repeater and the receiver are built so that they can be interconnected in either of two ways; the transmitter can be connected to the repeater which is in turn connected to the receiver, or the transmitter can be connected directly to the receiver. This affords an *A-B* test of the performance of the repeater which is the heart of the gain experiment to be described in Section 5.3.

5.2 Error-rate vs *S/N* Experiment

One of the experiments performed with this apparatus measured the error-rate as a function of signal-to-noise ratio. This experiment

was performed for the four possible cases, namely, errors introduced by the transmitter to repeater hop, those introduced by the repeater to receiver hop, those introduced by the transmitter to receiver hop, and those introduced by the complete transmitter to repeater to receiver hops. Allowing for the differences in the noise figures of the actual devices used in each of these components, the results were quite consistent. Therefore, the experiment will be described for one case only, transmitter to repeater.

The signal from the extra output of the random word generator was delayed in a transmission line for a time interval equal to the time for the transmitted signal to be regenerated. The outputs from the random word generator and from the regenerator of the repeater were then combined in a "baseband hybrid" as shown in Fig. 38. The output of this "hybrid" is 0 if the two input pulses are of the same polarity and is some amount $\pm v$ if the input pulses differ in polarity indicating that an error was made. These output pulses drive a pulse-height discriminator which has two output channels. This device delivers a pulse to one of its two outputs if the magnitude of the input pulse exceeds a certain threshold value. If the input pulse is positive the output occurs in one channel; if the input is negative the output occurs in the other channel. These "error-pulses" are counted on a dual-channel counter.

The experimental procedure is quite similar to that used in performing similar experiments on a prototype model of the IF portion of this repeater. It is described in some detail in a previous paper²⁶ and need not be repeated here. Certain differences should, however, be pointed out. First, the error-counting technique has been improved by the use of the dual-channel counter as described above. Second, in this experiment the signal-to-noise ratio was adjusted by changing the signal level and using the actual amplifier noise instead of injecting additional noise into the input of the repeater as was done in the experiment of Ref. 26. Finally, in addition to the checks on signal statistics listed in Ref. 26, the IF signal was observed on a spectrum analyzer.

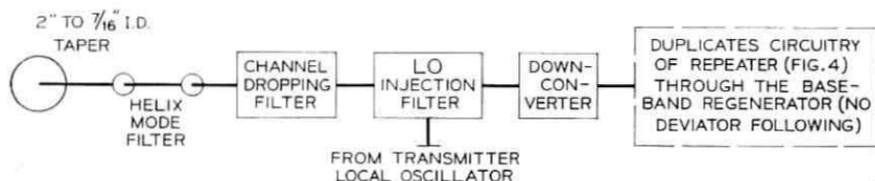


Fig. 37 — Receiver.

The biases on the random-word generator were adjusted until the spectrum was symmetric and free of "horns" or spikes (which are indicative of periodicities and hence nonrandomness in the signal).

The results of this experiment are shown by the points plotted in Fig. 32. Comparison between theory and experiment can readily be made from this figure.

5.3 Repeater Gain Experiment

The second experiment consisted of setting up two arrangements of components mentioned in Section 5.1 and setting the attenuators between these components to the value which gave an error rate of one error in 10^9 pulses (the assumed acceptable error-rate). This experiment is illustrated in block diagram form in Fig. 39. The gain of the repeater (in the sense of Section 4.3) is then given by

$$\begin{aligned} & (\text{Loss from Trans. to Rep.}) + (\text{Loss from Rep. To Rec.}) \\ & \quad - (\text{Loss from Trans. to Rec.}) \end{aligned}$$

after allowance is made for loss in the passive millimeter-wave circuitry of the repeater. Experimentally, the loss between the transmitter and the receiver was found to be an amount A_0 , the loss between transmitter and repeater plus the loss between repeater and receiver was found to be $A_0 + 43$ dB for 10^{-9} error probability at each regenerator. The loss in the passive millimeter-wave circuitry of the repeater was found to be 14 dB. Thus, the experimentally determined gain of the repeater is 57 dB. The measured local oscillator power is 6.0 dBm. From this one concludes that an additional 7 dB of LO power or a total of 13 dBm is necessary to achieve the 64-dB gain required for 15-mile repeater spacing with a 6-dB margin (from Section 4.2). This is in good agreement with the 12 dBm predicted by the argument of Section 4.3.

VI. CONCLUSIONS

A solid-state millimeter-wave repeater has been built which operates at a 306-Mb/s rate with an error-rate performance within 0.5 dB of

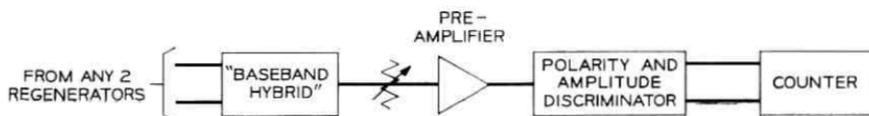


Fig. 38 — Error-counting circuitry.

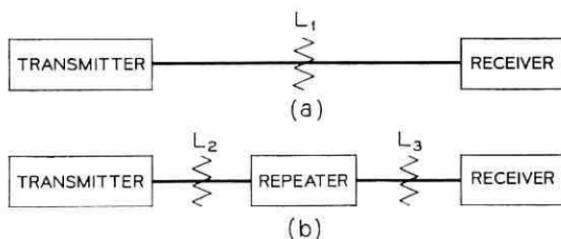


Fig. 39—Experimental arrangement for gain test. L_1 = loss from transmitter to receiver, L_2 = loss from transmitter to repeater, L_3 = loss from repeater to receiver.

the theoretical value. It gives a measured gain of 57 dB with a local oscillator power of 6 dBm. Since the repeater gain is proportional to LO power it is concluded that a 13-dBm local oscillator would give the 64-dB gain necessary for a 15-mile repeater spacing with a 6-dB margin and a suitable channel-dropping filter array for over one hundred 300-Mb/s channels.

VII. ACKNOWLEDGMENTS

The authors are particularly grateful to Mr. R. S. Engelbrecht and Mr. J. A. Copeland for providing LSA diodes immediately upon their first successful operation, and to Mr. A. Bakanowski for supplying the IF amplifiers used in this experiment. We wish to thank Mr. C. A. Burrus and Mr. J. C. Irvin for providing the millimeter-wave diodes used in this experiment and Mr. B. C. DeLoach and Mr. T. Misawa for the X-band and millimeter-wave IMPATT diodes. We also thank Mr. C. N. Dunn for supplying the tunnel diodes for the limiters.

We wish to thank Mr. J. H. Johnson for developing the limiter and gratefully acknowledge his assistance in the construction of the IF and baseband circuits. We are grateful to Mr. H. M. James for constructing the clock and timing recovery circuits.

APPENDIX

A.1 Introduction

Several types of delay distortion equalizers have been proposed during the past several years. Five of these equalizers will be discussed in the following paragraphs. Any of the five could, in principle, be used to equalize the delay distortion of the waveguide; economic considerations will dictate which is the most practical. It might, for eco-

conomic reasons, be desirable to use more than one type of equalization. For example, frequency frogging (Paragraph A.3) might be used to give partial equalization with a transversal equalizer used to complete the equalization. Other possible combinations will suggest themselves as the advantages and disadvantages of each type of equalizer are discussed.

Delay distortion is inversely proportional to the cube of the frequency. The delay distortion introduced across a 500-MHz band by 15 miles of waveguide varies from 34 nsec at 40 GHz to 2.2 nsec at 100 GHz. Therefore, considerable equalization is necessary in the lower bands and some equalization is desirable (although not required) in the upper bands.

The pertinent characteristics of the delay distortion are summarized as follows. The time delay, T , in a length, l , of waveguide can be written as

$$T = \frac{l}{v_g} = l \frac{\partial \beta}{\partial \omega},$$

where v_g is the group velocity in the medium and β , the waveguide propagation constant, is given by

$$\beta = \frac{\omega}{c} \sqrt{1 - (\omega_c/\omega)^2}.$$

Expanding β in a Taylor's series about ω_0 , the center angular frequency of the channel, gives

$$T = l[\beta_1 + 2\beta_2(\omega - \omega_0) + 3\beta_3(\omega - \omega_0)^2 + \dots], \quad (5)$$

where β_1, β_2, \dots are the expansion coefficients of the Taylor's series for β .

For frequencies and bandwidths of interest the terms of order β_3 and higher are negligible for the 2-inch waveguide. Since the β_1 term is a constant time delay, the only source of distortion is the β_2 term.

A.2 Reflection Equalizer

The reflection equalizer, proposed by J. R. Pierce and W. S. Alberheim,²⁷ is illustrated in Fig. 40. Since the higher-frequency components of the signal penetrate deeper into the taper before being reflected than the lower frequency components, their round trip transit time is longer. By properly designing the shape of the taper the distortion of the guide can, in principle, be exactly equalized. Equalizers of

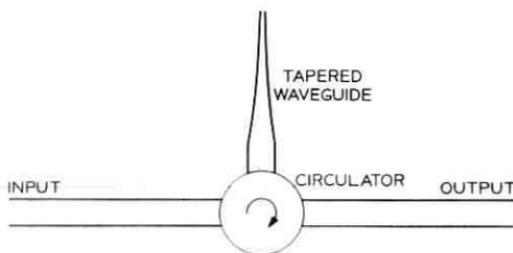


Fig. 40 — Reflection equalizer.

this type (operating at X-band) have been built by K. Woo²⁸ and also by C. C. H. Tang.²⁹ This type of equalizer can be built with an adjustable delay characteristic; as an alternative, one might build a small number of "stock" tapers which give approximate equalization and use a different type of equalizer to "trim" the equalization.

A.3 Transmission Equalizer

A second type of delay distortion equalizer is the transmission equalizer shown in Fig. 41. If the frequency spectrum of the signal in the channel is inverted and this signal is then passed through a short piece of waveguide near cutoff the delay distortion can be equalized since the frequency inversion causes a sign reversal in the β_2 term. Writing the transit time in the medium as

$$T_m = l_m[\beta_{1_m} + 2\beta_{2_m}(\omega - \omega_0)]$$

and the transit time in the equalizer as

$$T_e = l_e[\beta_{1_e} + 2\beta_{2_e}(\omega_0 - \omega) + 3\beta_{3_e}(\omega_0 - \omega)^2 + \dots],$$

one obtains for the total transit time

$$T_m + T_e = [l_m\beta_{1_m} + l_e\beta_{1_e}] \\ + 2[\beta_{2_m}l_m - \beta_{2_e}l_e](\omega - \omega_0) + 3\beta_{3_e}l_e(\omega_0 - \omega)^2 + \dots$$

The cutoff frequency of the equalizer can be chosen such that

$$\beta_{2_m}l_m = \beta_{2_e}l_e.$$

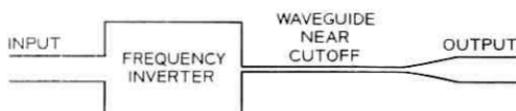


Fig. 41 — Transmission equalizer.

The delay distortion of the medium is thereby removed. If, however, l_e is short, the equalizer must be operated quite near cutoff in order to satisfy this equation. If l_e is too short, the terms of order β_{3e} and higher may contribute significant distortion to the signal. This, in fact, sets the lower limit on the length of the transmission equalizer. The minimum length of the equalizer depends on how much of this distortion is tolerable (which is not precisely known) and on the carrier frequency used in the equalizer. However, lengths of the order of ten feet or less are probably adequate for carrier frequencies above 50 GHz. For frequencies between 40 and 50 GHz, the length of the equalizer would probably be prohibitive for channel bandwidths of 500 MHz. However, transmission equalizers might be attractive as "trimming equalizers" for use with "stock" tapers or for use with the frequency-frogging scheme discussed in the next section.

Recently, J. H. Johnson³⁰ has built a transmission equalizer. His tests indicate that at the frequencies of interest, the phase characteristic is in agreement with the lossless theory and that the attenuation will not be prohibitive.

A.4 Frequency Frogging

The third approach to delay distortion equalization is due to D. H. Ring.³¹ It is known as "frequency frogging" and consists of replacing every other regenerative repeater in the system with nonregenerative repeaters that invert the frequency spectrum of each channel and provide linear gain. This scheme is illustrated in Fig. 42. The medium itself in span 2 (see Fig. 42) then acts as a long transmission equalizer for span 1. If the spans are of equal length the equalization is exact except for the contribution from the $\beta_3, \beta_5, \text{etc.}$, terms in (5) which is negligible for reasonable channel bandwidth and repeater spacings. If the spans are of unequal length, say x_1 and x_2 , respectively, only the distortion $2\beta_2(\omega - \omega_0)(x_1 - x_2)$ must be equalized. Since one would have

$$|2\beta_2(\omega - \omega_0)(x_1 - x_2)| \ll |2\beta_2(\omega - \omega_0)x_1|$$

the "trimming equalizer" required in such a system could be a comparatively simple transmission or transversal equalizer.

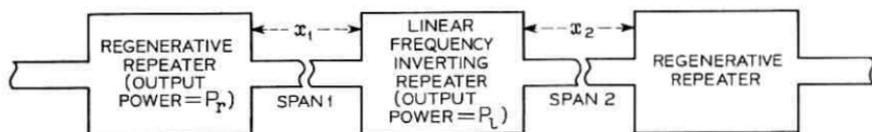


Fig. 42—Frequency frogging.

The chief disadvantage of frequency frogging is the requirement that the nonregenerative repeater be linear.* This will probably result in lower available power at these repeaters than is attainable at the regenerative repeaters. The effect on repeater spacing can be calculated. If we define P_r and P_l to be the average power available at the output of regenerative and linear repeaters, respectively, x_0 to be the maximum allowable spacing between repeaters in a system having all regenerative repeaters and ideal delay distortion equalizers, and x_1 and x_2 to be the lengths of the spans in the frequency-frogging system, the values of x_1 and x_2 which maximize the quantity $x_1 + x_2$ are given by

$$x_1 = x_0 - 1 \text{ miles}$$

$$x_2 = x_0 - 1 - \frac{10}{3} \text{Log} \frac{P_r}{P_l} \text{ miles,}$$

(A loss of 3 dB per mile has been assumed in the above equations.) The fractional decrease in repeaters spacing using this scheme is thus,

$$\frac{2x_0 - (x_1 + x_2)}{2x_0} = \frac{1 + \frac{5}{3} \text{Log} \frac{P_r}{P_l}}{x_0} = \frac{1 + \frac{(P_r/P_l) \text{ dB}}{6}}{x_0}$$

which, for example, amounts to only about 12 percent for $P_r/P_l = 8$ dB and $x_0 = 20$ miles. The amount of delay distortion which must be made up by a "trimming equalizer" is equivalent to Δ miles of guide where

$$\Delta = x_1 - x_2 = \frac{10}{3} \text{Log} \frac{P_r}{P_l}.$$

For the example cited above, $\Delta = 2.67$ miles.

A.5 Transversal Equalizer

Baseband transversal equalization can be used in a linear system to improve the pulse response.³² This type of equalizer functions by adding time shifted images of the input pulse to itself in such a manner that the pulse response of the system is set to zero at a finite number of instants an integral number of time slots from the pulse center. The addition of the time shifted images of the pulse is usually carried out by means of a tapped delay line and a summing network.

* Since the delay distorted signal at the nonregenerative repeater may possess amplitude modulation, this repeater may have to be linear up to power levels higher than the average signal power.

Since the conversion to and from baseband in our system is non-linear, baseband equalization cannot be used. However, an IF transversal equalizer can be used. A possible configuration of the circuit is shown in Fig. 43. It can be shown that any realizable transfer function can be approximated over a finite band using this type of circuit.³³ Thus, this circuit can be used to compensate for the waveguide delay distortion.

An alternate approach to transversal equalization can be used which is similar to baseband transversal equalization. It can be shown by taking quadrature components that the binary FM-DCPSK signal is equivalent to two polar pulse trains in phase quadrature. By proper choice of tap gains and phase shift the response of the system can be set to zero at instants that are multiples of a bit interval from the pulse center.

Computations made by one of the authors, JEG, which will be published at a later date, show that transversal equalizers with about 11 taps can be built to equalize the channels at 50 GHz and that above 70 GHz extremely good equalization can be achieved with 5 or fewer taps. Also, under certain circumstances, the phase shifters can be eliminated.

A.6 Equalization by Quasi-Periodic Structures

In a recent paper,³⁴ H. S. Hewitt has described a tapered meander line filter, shown in Fig. 44, which produced 300 ns of nearly linear delay distortion over a frequency band from 1.1 to 1.7 GHz. This device, which had a total length of less than 18 inches, demonstrates the feasibility of using a filter of this type. It is believed that this type of structure deserves careful consideration.

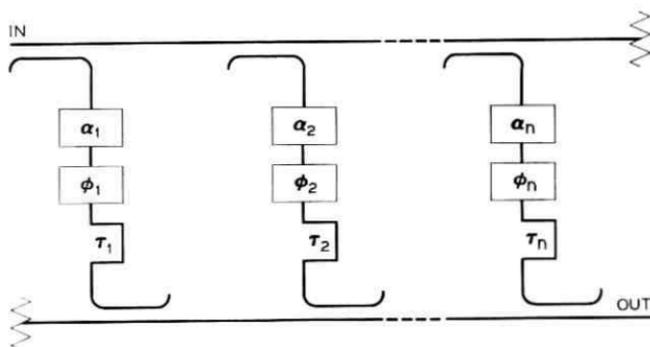


Fig. 43 — A microwave realization of the transversal equalizer.

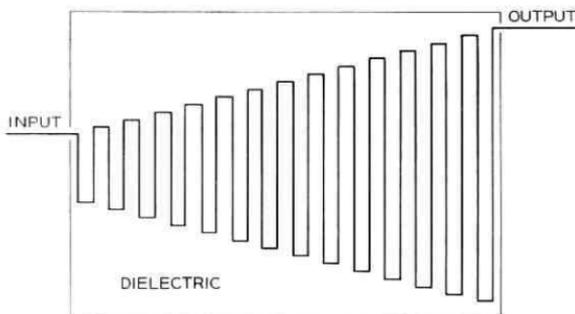


Fig. 44 — Tapered meander line.

REFERENCES

1. Miller, S. E., *Waveguide as a Communication Medium*, B.S.T.J., 33, November, 1954, pp. 1209-1265.
2. Rowe, H. E. and Warters, W. D., *Transmission in Multimode Waveguide with Random Imperfections*, B.S.T.J., 41, May, 1962, pp. 1031-1170.
3. Morgan, S. P. and Young, J. A., *Helix Waveguide*, B.S.T.J., 35, November, 1956, pp. 1347-1384.
4. Unger, H. G., *Round Waveguide with Lossy Lining*, Proc. Symp. Millimeter Waves, Polytechnic Inst. of Brooklyn, 1959, pp. 535-541.
5. King, A. P. and Mandeville G. D., *Observed 33 to 90 KMc Attenuation of Two-Inch Improved Waveguide*, B.S.T.J., 40, September, 1961, pp. 1323-1330.
6. Steier, W. H., *The Attenuation of the Holmdel Helix Waveguide in the 100-125 KMc Band*, B.S.T.J., 44, May, 1965, pp. 899-906.
7. Nutt, W. G., private communication.
8. Copeland, J. A., *CW Operation of LSA Oscillator Diodes— 44 to 88 GHz*, B.S.T.J., 56, January, 1967, pp. 284-287.
9. Kotel'nikov, V. A., *Theory of Optimum Noise Immunity*, McGraw-Hill Book Co. Inc., New York, 1959.
10. Lawton, J. G., *Comparison of Binary Data Transmission*, Proc. 1958 Conf. Mil. Elec.
11. Anderson, R. R. and Salz, J., *Spectra of Digital FM*, B.S.T.J., 44, July 1965, pp. 1165-1190.
12. Marcatili, E. A. and Bisbee, D. A., *Band Splitting Filter*, B.S.T.J., 40, January, 1961, pp. 197-212.
13. Standley, R. D., *A Millimeter-Wave, Two-Pole, Circular-Electric Mode, Channel-Dropping Filter Structure*, to be published, B.S.T.J., December, 1967.
14. Marcatili, E. A., *Mode Conversion Filters*, B.S.T.J. 40, January, 1961, pp. 149-184.
15. Lee, T. P. and Burrus, C. A., *A Millimeter-Wave Quadrupler and an Up-Converter Using Planar Diffused Gallium Arsenide Varactor Diodes*; Conf. High Freq. Gen. Ampl., Cornell University, Ithaca, N. Y., August, 1967.
16. Young, D. T. and Irvin, J. C., *Millimeter Frequency Conversion Using Au-n-Type GaAs Schottky Barrier Epitaxial Diodes with a Novel Contacting Technique*, Proc. IEEE, 53, December, 1965, pp. 2130-2131.
17. Misawa, T., *CW Millimeter-Wave IMPATT Diodes with Nearly Abrupt Type Junctions*, Solid-State Device Research Conf., Santa Barbara, Calif., June, 1967.
18. Standley, R. D., *Millimeter Wavelength Diplexing Filters Utilizing Circular TE₀₁₁ Mode Resonators*, PGMTT, to be published, January, 1968.
19. Sharpless, W. M., *Wafer-Type Millimeter Wave Rectifiers*, B.S.T.J., 35, November, 1956, pp. 1385-1420.

20. Engelbrecht, R. S., and Kurokawa, K., A. Wideband Low Noise L-Band Balanced Transistor Amplifier, Proc. IEEE, 53, March, 1965, pp. 237-247.
21. Hubbard, W. M., The Effect of a Finite-Width Decision Threshold on Binary Differentially Coherent PSK Systems, B.S.T.J., 45, February, 1966, pp. 307-320.
22. Hubbard, W. M., The Effect of Intersymbol Interference on Error-Rate in Binary Differentially Coherent Phase Shifted Keyed Systems, B.S.T.J., 46, July-August, 1967, pp. 1149-1172.
23. Axelrod, M. S., Farber, A. S., and Rosenheim, D. E., Some New High Speed Tunnel-Diode Logic Circuits, IBM J., April, 1962.
24. Herzog, G. B., Tunnel-Diode Balanced-Pair Switching Analysis, RCA Rev., June, 1962.
25. Copeland, J. A., private communication.
26. Hubbard, W. M. and Mandeville, G. D., Experimental Verification of the Error-Rate Performance of Two Types of Regenerative Repeaters for Differentially Coherent Phase Shift Keyed Signals, B.S.T.J., 46, July-August, 1967, pp. 1173-1202.
27. Pierce, J. R., Tapered Waveguide Delay Equalizer, U. S. Patent No. 2,863,126, issued December, 1958.
28. Woo, K., An Adjustable Microwave Delay Equalizer, IEEE Trans., MTT 13, March, 1965, pp. 224-232.
29. Tang, C. C. H., Delay Equalization by Tapered Cutoff Waveguides, IEEE Trans. Microwave Theory Tech., MTT-12, November, 1964, pp. 608-615.
30. Johnson, J. H., private communication.
31. Ring, D. H., U. S. Patent No. 2,629,782, issued.
32. Wheeler, H. A., The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, Proc. IRE, 27, June, 1939, pp. 359-385.
33. Burrows, C. R., Discussion of H. A. Wheeler paper (see Ref. 32) pp. 384-385.
34. Hewitt, H. S., A Computer Designed, 720 to 1 Microwave Compression Filter, Group Microwave Tech. Int. Microwave Symp. Digest, 1967, pp. 51-53.

A Quantitative Theory of $1/f$ Type Noise Due to Interface States in Thermally Oxidized Silicon

By E. H. NICOLLIAN and H. MELCHIOR

(Manuscript received June 22, 1967)

A quantitative theory of $1/f$ type noise is derived from the distribution of trapping times for charges in interface states. The distribution of trapping times has been recently explained quantitatively by means of a random distribution of surface potential caused by a random distribution over the plane of the interface of fixed charges located in the oxide. This model, which agrees with the interface state time constant dispersion measured by the MIS conductance technique, leads to a noise spectrum which is independent of frequency at very low frequencies, tends towards a $1/f^2$ dependence at high frequencies, and has an extended $1/f$ frequency dependence at intermediate frequencies. The mechanism for time constant dispersion is independent of temperature and silicon resistivity; it depends only on the majority carrier density at the silicon surface, the interface state density, and the density of fixed oxide charges. The dependence of open circuit mean square noise voltage on these parameters and frequency are illustrated for an MOS capacitor.

I. INTRODUCTION

It has long been recognized that states at the Si-SiO₂ interface which exchange charge with the silicon can give rise to $1/f$ type noise. Recently, Sah and Hielscher¹ have shown by experiment that the $1/f$ noise of a metal -SiO₂-silicon (MOS) capacitor is directly related to interface state density and capture conductance over the energy gap. Random capture and emission of carriers by interface states results in fluctuations of trapped charge. In an MOS capacitor, these charge fluctuations cause random changes in admittance constituting noise. These charge fluctuations can be calculated from the dispersion of interface state time constants. A major obstacle to a quantitative theory of $1/f$ type noise arising from interface states has been the lack

of an experimentally established mechanism for interface state time constant dispersion. This obstacle has recently been removed. With the MIS conductance technique,^{2, 3, 4} accurate small-signal measurements have been made of interface state density and capture conductance over the middle half of the energy gap in the Si-SiO₂ system. A large interface state time constant dispersion was observed in the depletion and accumulation regions. An explanation which quantitatively fits these measurements essentially without any arbitrary adjustable parameters is that the dispersion arises from a random distribution of surface potential over the plane of the interface. The random surface potential distribution is in turn caused primarily by a random distribution of built-in oxide charges and charged interface states over the plane of the interface. The noise measurements of Ref. 1 and the small signal conductance measurements of Ref. 2 through 4 suggest that a quantitative explanation of $1/f$ type noise of an MOS capacitor can be given in terms of the interface state time constant dispersion caused by the random distribution of surface potential and the resulting capture conductance.

It has been reported that low-frequency noise generated at semiconductor surfaces shows a $1/f^n$ spectrum with $n \approx 1$ over many decades of frequency.^{5, 6} Various mechanisms have been proposed to explain this, such as slow states in the oxide or at the oxide-air interface or slow time dependent changes in the density of states at the semiconductor surface.^{5, 7, 8} Atalla, et al,⁶ have shown that surface generated $1/f$ noise extending over many decades of frequency is considerably reduced in magnitude when silicon is thermally oxidized. The noise theory presented here is based on conductance measurements made on thermally oxidized silicon samples prepared as described in Ref. 4. In these samples, oxide thickness is greater than 500 Å. These samples have stable electrical characteristics at room temperature under bias. Also, losses in the oxide layer and bulk silicon are found to be negligible.^{2, 4} Thus, they should be free of noise mechanisms other than random emission and capture of carriers by interface states having a time invariant density. The case where interface state transitions dominate the loss as in the measurements of Ref. 4 will be the only case considered here.

This work clearly shows that in thermally oxidized silicon, surface-generated noise arising from random emission and capture of carriers by interface states does not explain a $1/f$ noise spectrum over many decades of frequency. Measurements in which a $1/f$ noise spectrum is found over several decades must involve additional mechanisms as mentioned.

The noise spectrum of a single level state, as is well known,^{9, 10} is independent of frequency at low frequencies and has a $1/f^2$ frequency dependence at high frequencies. We shall show that the time constant dispersion found by conductance measurements introduces an intermediate range in the noise spectrum with a $1/f$ type frequency dependence. The resulting open circuit noise voltage appearing across the terminals of an MOS capacitor has been calculated and found to have a large $1/f$ type range and the same dependence on interface state density and capture conductance as in Ref. 1.

The MOS capacitor is the simplest case of interface state $1/f$ type noise to treat quantitatively because there is no dc current flow. The theory for the MOS capacitor will be worked out here in detail. This theory can be extended to explain $1/f$ noise arising from interface states in MOS field effect transistors and oxide passivated bipolar transistors because in these devices, time constant dispersion also will be caused by the random surface potential distribution. This extension will not be made here.

II. THEORY

We shall use the Nyquist formula for the calculation of noise. This is justified by the fact that in the MOS capacitor it is reasonable to assume that the interface states and the silicon are in thermal equilibrium with each other at each bias when no dc leakage current flows through the oxide layer. We shall consider the case where the applied voltage biases the silicon into accumulation or depletion up to within a few kT/q of mid gap. In these regions, majority carrier density is several orders of magnitude greater than minority carrier density at the silicon surface. Neglecting minority transitions will cause little or no error at the frequencies considered in this paper (0.1 Hz to 10^8 Hz) because in these regions of bias there is virtually no recombination-generation through interface states or states in the silicon bulk.⁴ Diffusion from the bulk is also negligible. In the MOS capacitor, recombination-generation and diffusion are the only ways the minority carrier band can communicate with an external circuit. Thus, the noise we shall calculate arises primarily by the random capture and emission of majority carriers by the interface states. Fig. 1 shows the noise equivalent circuit for the MOS capacitor at a given bias and angular frequency ω . Using the Nyquist formula, the mean square noise current per cm^2 generated in $G_p(\omega)$ is

$$\langle i_p^2 \rangle = 4kTBG_p(\omega), \quad (1)$$

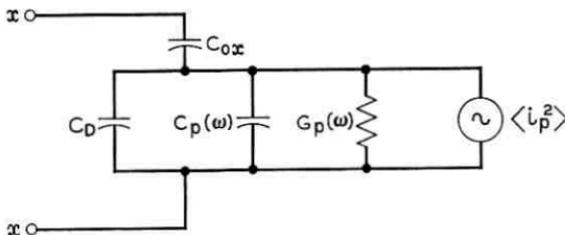


Fig. 1—Noise equivalent circuit of MOS capacitor. C_{ox} is oxide layer capacitance in farads/cm², C_D is depletion layer capacitance in farads/cm², $C_p(\omega)$ is the interface state capacitance given by (9) in farads/cm², $G_p(\omega)$ is the interface state equivalent parallel conductance given by (8) mhos/cm², and $\langle i_p^2 \rangle$ is the mean square short circuit noise current in amp²/cm².

where $G_p(\omega)$ is the interface state equivalent parallel conductance in mhos/cm², k is Boltzman's constant in Joules/°K, T is the absolute temperature in °K, and B is the bandwidth in Hz. The mean square open circuit noise voltage \times cm² appearing across the terminals $x-x$ in Fig. 1 is then

$$\langle v_o^2 \rangle = \frac{4kTBG_p(\omega)}{G_p^2(\omega) + \omega^2[C_D + C_p(\omega)]^2}, \quad (2)$$

where $C_p(\omega)$ is interface state capacitance in farads/cm², and C_D is the depletion layer capacitance in farads/cm².

The problem in evaluating (2) is essentially to find the interface state admittance as a function of bias and frequency. This has been done previously as described in Refs. 3 and 4. This derivation will be briefly outlined here. It is based on a model in which the interface state time constant dispersion required for a $1/f$ type noise spectrum is caused by a random distribution of surface potential. A detailed analysis of this mechanism complete with experimental documentation can be found in Ref. 4.

2.1 Depletion

With the silicon surface in depletion or accumulation, it has been shown experimentally (see Ref. 4) that the ohmic loss in the oxide layer and the silicon space-charge region is negligible compared to the ohmic loss arising from transitions between interface states and the majority carrier band. Bulk silicon series resistance and contact resistance can be made negligible in practice⁴ or calculated separately. Because this paper is restricted to a discussion of noise due to interface states, these two resistances will be ignored.

A single level interface state is not observed experimentally. Rather, the interface states are observed to be comprised of energy levels so closely spaced in energy that they cannot be distinguished as separate levels. They appear as a continuum over the bandgap of the silicon. The time constant dispersion observed is larger than expected for a continuum. A random distribution of surface potential caused by a random distribution of fixed oxide charges over the plane of the interface is found to quantitatively explain the time constant dispersion measured. To analyze this mechanism, we proceed as follows. Dividing the plane of the interface into a number of squares of equal area, the largest area within which surface potential is uniform is called the characteristic area of the random fixed oxide charge distribution. The admittance of the continuum of levels located in a characteristic area can be obtained by integrating the admittance of a single level over all the levels distributed in energy from the valence band to the conduction band. The resulting total interface state admittance in a characteristic area is⁴

$$Y_{ss} = j\omega \frac{q^2}{kT} \int_{E_v}^{E_c} \frac{N_{ss} f_0 (1 - f_0) d\psi}{1 + j\omega f_0 / c_p p_{so}}, \quad (3)$$

where q is the electronic charge in coulombs, $j = \sqrt{-1}$, N_{ss} is the density of interface states $\text{cm}^{-2} \times \text{eV}^{-1}$, f_0 is the Fermi function at a given bias, c_p is the majority carrier capture probability in cm^3/sec , p_{so} is the majority carrier density at the silicon surface in cm^{-3} , and $d\psi$ is a small energy interval in the bandgap in eV . The integrand of (3) is sharply peaked about the Fermi level with a width of about kT/q . Thus, (3) can be easily integrated because both N_{ss} and c_p are experimentally observed to vary only slightly over several kT/q in a range of bandgap energy of about half the gap centered about mid-gap. Making the substitution $f_0(1 - f_0) = (kT/q)(df_0/d\psi)$ transforms (3) into an integral over f_0 . Integrating from zero to unity yields

$$Y_{ss} = \frac{qN_{ss}}{2\tau_m} \ln(1 + \omega^2\tau_m^2) + jq \frac{N_{ss}}{\tau_m} \text{arc tan}(\omega\tau_m), \quad (4)$$

where $\tau_m = 1/c_p p_{so}$. Equation (4) was first derived by Lehevec.¹¹

Typically N_{ss} is in the range $10^{10} \text{cm}^{-2} \times \text{eV}^{-1}$ to $10^{11} \text{cm}^{-2} \times \text{eV}^{-1}$. This means that the interface states are spacially separated too far apart in the plane of the interface for the wave function of an electron in one center to overlap a neighboring center. Transitions from one center to another, even though the centers are closely spaced in energy, are therefore, highly improbable. Thus, transitions between the ma-

majority carrier band and a particular level in the continuum located in energy near the Fermi level are not correlated to transitions between the majority carrier band and other levels nearby in energy.

The total admittance Y_T is obtained by multiplying the admittance contributed by each characteristic area Y_{ss} from (4) by the number of characteristic areas in which the surface potential is between u_s and $u_s + du_s$ and integrating over all the characteristic areas under the field plate. The result is

$$Y_T = \int_{-\infty}^{\infty} Y_{ss} P(u_s) du_s, \quad (5)$$

where $P(u_s) du_s$ is the number of characteristic areas in which the surface potential (in units of kT/q) is between u_s and $u_s + du_s$ and Y_{ss} is given by (4). $P(u_s)$, the probability that the surface potential in a characteristic area is u_s , is obtained from the random distribution of fixed oxide charges.⁴ When the mean number of charges in a characteristic area is large, the probability of finding N charges in a characteristic area $P(N)$ is given by the Gaussian approximation of a Poisson distribution. Transforming $P(N)$ to $P(u_s)$ for the case of small fluctuations (see Refs. 3 and 4), we get

$$P(u_s) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp [-(u_s - \bar{u}_s)^2/2\sigma^2], \quad (6)$$

where σ is the standard deviation of surface potential and \bar{u}_s is the mean surface potential in units of kT/q at a given bias. The standard deviation of surface potential is

$$\sigma = \frac{(q/kT)W(q\bar{Q}_s/\alpha)^{\frac{1}{2}}}{WC_{ox} + \epsilon_{si}}, \quad (7)$$

where W is space-charge width in cm, \bar{Q}_s is the fixed oxide charge density in coul/cm², α is the characteristic area in cm², C_{ox} is the oxide layer capacitance in farads/cm², and ϵ_{si} is the permittivity of the silicon in farads/cm.

Substituting (4) and (6) into (5), we get

$$G_p(\omega) = \frac{1}{2} q N_{ss} (2\pi\sigma^2)^{-\frac{1}{2}} \cdot \int_{-\infty}^{\infty} \exp [-(u_s - \bar{u}_s)^2/2\sigma^2] \tau_m^{-1} \ln(1 + \omega^2 \tau_o^2) du_s \quad (8)$$

and

$$C_p(\omega) = q N_{ss} (2\pi\sigma^2)^{-\frac{1}{2}} \cdot \int_{-\infty}^{\infty} \exp [-(u_s - \bar{u}_s)^2/2\sigma^2] (\omega\tau_m)^{-1} \arctan(\omega\tau_m) du_s. \quad (9)$$

For p-type, $\tau_m = 1/c_p p_{s0} = (1/c_p N_A) \exp u_s$, where N_A is the acceptor density in the silicon in cm^{-3} and c_p is the hole capture probability.

These integrals can be evaluated numerically using the experimental observation that the density of states and the majority carrier capture probability vary very slowly over several kT of bandgap energy.

The values of $G_p(\omega)$ and $C_p(\omega)$ calculated from (8) and (9) can be used in (2) to obtain the open circuit mean square noise voltage of the MOS capacitor.

To illustrate the noise properties predicted by the statistical model, the spectrum of the trapped charge fluctuations in the interface states is derived from this model. First, the spectrum of charge fluctuations for a single time constant is^{9, 10}

$$S_{ss}(\omega) = \frac{4BN_s \tau f_0 (1 - f_0)}{1 + \omega^2 \tau^2}, \quad (10)$$

where $\tau = f_0 \tau_m$ and N_s is the density of states cm^{-2} .

The noise spectrum for the continuum of states located in a characteristic area is obtained by integrating (10) over bandgap energy in a manner identical to (3). The result is

$$S_{sc}(\omega) = (2kT/q)BN_{ss}(\omega^2 \tau_m^2)^{-1} \ln(1 + \omega^2 \tau_m^2). \quad (11)$$

Integrating (11) over all characteristic areas similarly to (5), (8), and (9) yields for the actual spectral distribution

$$S(\omega) = (2kT/q)BN_{ss}(2\pi\sigma^2)^{-1} \int_{-\infty}^{\infty} \exp[-(u_s - \bar{u}_s)^2/2\sigma^2](\omega^2 \tau_m^2)^{-1} \ln(1 + \omega^2 \tau_m^2) du_s. \quad (12)$$

Equations (8), (9), and (12) have been numerically integrated on an IBM 7094 computer using the trapezoidal rule.

III. DISCUSSION

3.1 Depletion

Curve (a) of Fig. 2 shows the noise spectrum for a single level state calculated from (10) with the Fermi level at the trap level. Curve (b) of Fig. 2 shows the noise spectrum for the continuum of levels located in a characteristic area calculated from (11). Both of these curves are normalized to their low-frequency values. Comparing curve (a) to curve (b), it is seen that integration over the continuum of levels results only in minor modifications of the shape of the spectrum. Curve

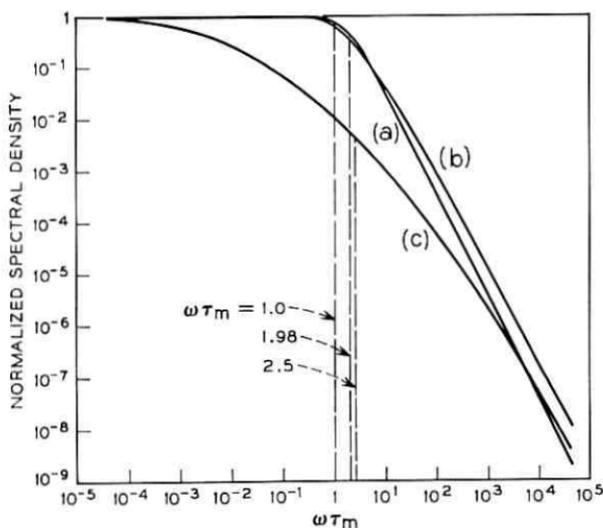


Fig. 2—Log-log plots of normalized spectral density vs $\omega\tau_m$ for the mean square fluctuations of the number of electrons trapped at the interface states. Curve (a) is for a single time constant calculated from (10). Curve (b) is for a continuum of states calculated from (11). Curve (c) is a plot of (12) using a standard deviation of surface potential of 2.6. All three curves are calculated using a hole density at the silicon surface of $6.4 \times 10^{12} \text{ cm}^{-3}$ and a hole capture probability of $2.2 \times 10^{-9} \text{ cm}^3/\text{sec}$. The conditions: $\omega\tau_m = 1.0, 1.98,$ and 2.5 correspond to the values of $\omega\tau_m$ at which the $G_p(\omega)/\omega$ curve peaks for each case.

(c) of Fig. 2 shows the noise spectrum calculated from (12) using a standard deviation of surface potential of 2.6. This curve is also normalized to its low-frequency value. Curve (c) is seen to be significantly different from curve (a) and curve (b). Fig. 2 shows that the random distribution of surface potential for an experimentally observed standard deviation of 2.6 is the dominant influence on the shape of the noise spectrum. In fact, the random distribution of surface potential will be the dominant influence over the range of standard deviation between 1.8 and 2.6. This is the range found by conductance measurements on several [111] and [100] crystals both n and p type.

In Fig. 3, curve (a) is the noise spectrum calculated from (12) and curve (b) the corresponding $G_p(\omega)/\omega$ vs frequency calculated from (8). For the parameters given in the caption under Fig. 3, $G_p(\omega)/\omega$ goes through a peak of 6 kHz. Fig. 3 shows that:

(i) The noise spectrum becomes independent of frequency at low frequencies. For the case considered, this occurs at frequencies much lower than 6 kHz.

(ii) The noise spectrum tends towards a $1/f^2$ frequency dependence at high frequencies. This will occur at frequencies much higher than 6 kHz for the case considered.

(iii) In the intermediate frequency range where $G_p(\omega)/\omega$ has its highest values, the noise has a $1/f$ type frequency dependence. For the case considered here, this occurs around 6 kHz.

A $1/f$ spectrum is drawn through curve (a) in Fig. 3. To see that the standard deviation determines the frequency range over which a $1/f$ spectrum fits our theory, we transform $P(u_s)$ to $P(\tau_m)$.

$$P(\tau_m) = P(u_s) du_s/d\tau_m, \quad (13)$$

where $P(u_s)$ is the probability that the time constant in a characteristic area is τ_m . From the relation $\tau_m = (c_p N_A)^{-1} \exp u_s$ given previously, (13) becomes

$$P(\tau_m) = P(u_s) \tau_m^{-1}. \quad (14)$$

We expand $P(u_s)$ given in (6) in a power series. As long as the condition $(u_s - \bar{u}_s)^2/2\sigma^2 \ll 1$ holds, all terms in the series except the first

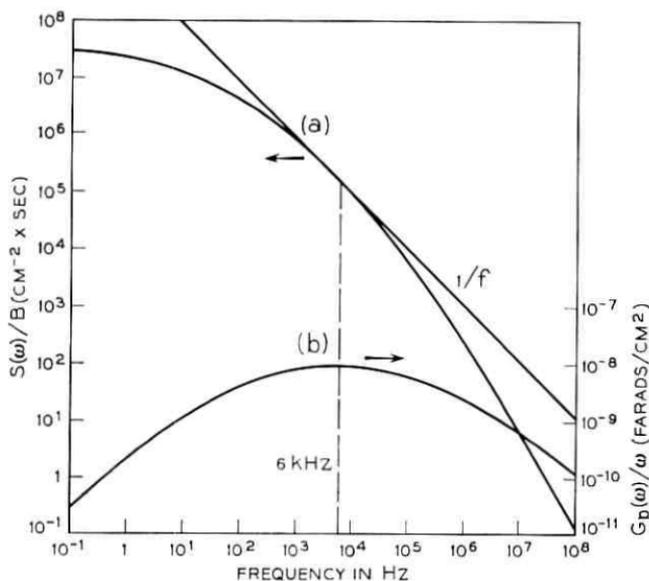


Fig. 3— Curve (a) is a log-log plot of (12) vs frequency and curve (b) a log-log plot of $G_p(\omega)/\omega$ from (8) vs frequency. Both curves are calculated for a standard deviation of 2.6, a hole density at the surface of $6.5 \times 10^{12} \text{ cm}^{-3}$, a hole capture probability of $2.2 \times 10^{-9} \text{ cm}^3/\text{sec}$, an interface state density of $3 \times 10^{11} \text{ cm}^{-2} \times \text{eV}^{-1}$, and a temperature of 300°K .

can be dropped. Then, from (14), $P(\tau_m) = (2\pi\sigma^2)^{-1}\tau_m^{-1}$. Superposing spectra of the type $\tau(1 + \omega^2\tau^2)^{-1}$ with a time constant distribution proportional to $1/\tau$ gives a $1/f$ noise spectrum.¹² The integration over the trap levels in one characteristic area results only in a minor change of shape of a $\tau(1 + \omega^2\tau^2)^{-1}$ spectrum as shown in Fig. 2. Essentially only the frequency for $G_p(\omega)/\omega$ maximum shifts to a higher value. Thus, a $1/f$ spectrum will fit our theory over a frequency range determined by the condition $(u_s - \bar{u}_s)^2/2\sigma^2 \ll 1$. The width of the frequency range given by this condition is determined by σ . This can be clarified by an illustrative example. Let us replace $P(u_s)$ with a rectangular distribution of height $(2\pi\sigma^2)^{-\frac{1}{2}}$ and width $(2\pi\sigma^2)^{\frac{1}{2}}$. This distribution gives a $1/f$ spectrum which is within a factor of 2 of curve (a) in Fig. 3 over four decades of frequency. The highest and lowest frequencies in this range are given by

$$f_{H,L} = f_p \exp [\pm(\pi/2)^{\frac{1}{2}}\sigma], \quad (15)$$

where f_p is the center frequency of the range and is the frequency at which the corresponding $G_p(\omega)/\omega$ vs $\log \omega$ curve peaks. This center frequency is proportional to the majority carrier density p_{s0} at the silicon surface and is almost independent of σ . For the statistical model with σ between 1.8 and 2.6, $f_p = (2.5/2\pi) c_p p_{s0}$. Equation (15) shows that the frequency range over which the $1/f$ spectrum is observed depends exponentially on σ .

$C_p(\omega)$ and $G_p(\omega)$ given by (8) and (9) are independent of temperature and silicon resistivity. For a given σ , c_p , and ω , these equations depend only on p_{s0} through the variable τ_m . The relation between τ_m and p_{s0} is $\tau_m = 1/c_p p_{s0}$. Measurements reported in Ref. 4 show that capture probability is independent of temperature. For a wide range of temperature and silicon resistivity, the same value of p_{s0} in depletion or accumulation can be obtained just by adjusting field plate bias. Thus, our mechanism for time constant dispersion is independent of temperature and silicon resistivity. Silicon conductivity type is important only because the capture probability for electrons is found to be about ten times larger than for holes.

Fig. 4 shows open circuit mean square noise voltage for two different values of p_{s0} or bias calculated from (2) using (8) and (9). Curve (a) is for $p_{s0} = 6.4 \times 10^{12} \text{ cm}^{-3}$ and curve (b) for $p_{s0} = 3.5 \times 10^{14} \text{ cm}^{-3}$ both for $\sigma = 2.6$. Fig. 4 illustrates the bias dependence of the noise voltage vs frequency. The curves in Fig. 4 will be a function of temperature and silicon resistivity as seen from (2).

Fig. 5 shows the influence of standard deviation of surface potential

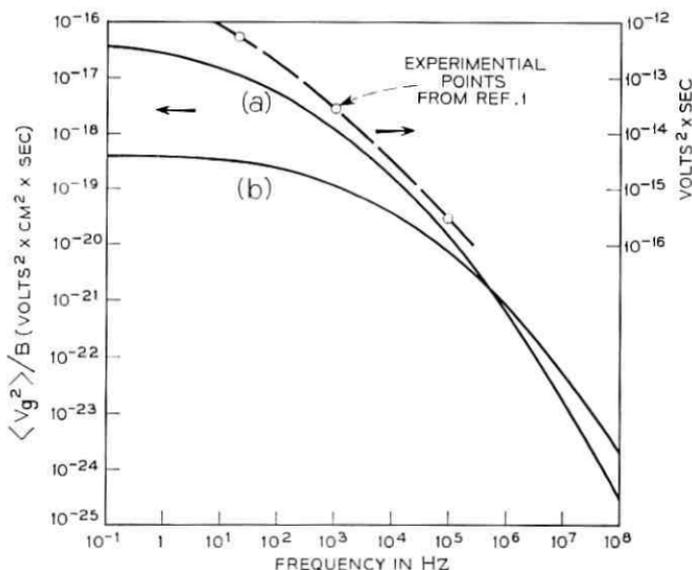


Fig. 4—Log-log plot of open circuit mean square noise voltage of MOS capacitor vs frequency calculated from (2), (8), and (9). Curve (a) and curve (b) are for hole densities at the silicon surface of $6.4 \times 10^{12} \text{ cm}^{-3}$ and $3.5 \times 10^{14} \text{ cm}^{-3}$ respectively. For both curves, standard deviation is 2.6, hole capture probability is $2.2 \times 10^{-9} \text{ cm}^3$, acceptor density is $2.1 \times 10^{16} \text{ cm}^{-3}$, interface state density is $3 \times 10^{11} \text{ cm}^{-2} \times \text{eV}^{-1}$, and temperature 300°K . Mean surface potential is 8.1 in curve (a) and 4.1 in curve (b). Experimental points are taken from Fig. 1 of Ref. 1 at a gate voltage of -2 volts. Notice similarity of shape to curves (a) and (b).

on the mean square noise voltage vs frequency for a majority carrier density at the silicon surface of $6.4 \times 10^{12} \text{ cm}^{-3}$. Curve (a) is for a standard deviation of 2.6 and curve (b) for a standard deviation of 1.8. These are the largest and smallest values found by conductance measurements on several MOS capacitors.

Fig. 5 shows that:

- (i) The standard deviation has the greatest influence on the magnitude of the mean square noise voltage at low frequencies.
- (ii) The range of frequencies over which the mean square noise voltage has a $1/f$ frequency dependence increases with increasing standard deviation of surface potential.

Standard deviation is experimentally observed to be independent of bias over most of the depletion range. It is shown in Refs. 3 and 4 that the relation between characteristic area and space-charge width is

$$\alpha^{\frac{1}{2}} \approx 2W. \quad (16)$$

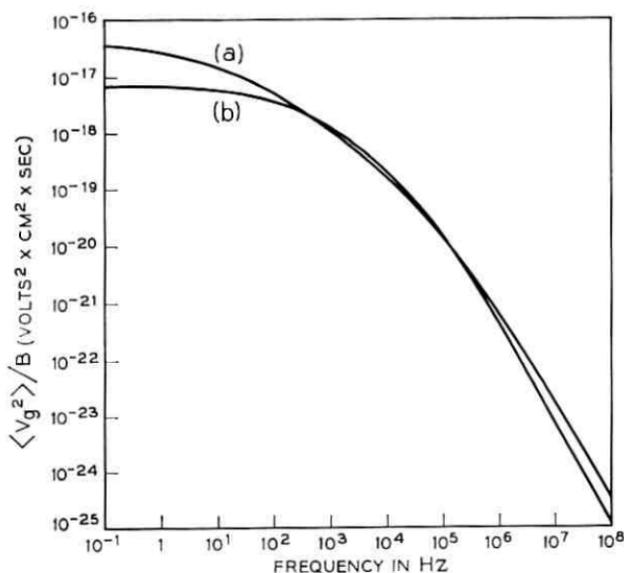


Fig. 5—Log-log plot of open circuit mean square noise voltage of MOS capacitor vs frequency calculated from (2), (8), and (9). Curve (a) is for a standard deviation of 2.6 and curve (b) for a standard deviation of 1.8. For both curves, hole density at the silicon surface is $6.4 \times 10^{12} \text{ cm}^{-3}$. Hole capture probability, acceptor density, interface state density, and temperature are the same as in Fig. 4. Mean surface potential is 8.1.

Substituting (16) into (7), the fixed charge density causing the random distribution of surface potential can be calculated from the standard deviation. For a standard deviation of 2.6, fixed oxide charge density will be $1 \times 10^{12} \text{ cm}^{-2}$ and for a standard deviation of 1.8, fixed oxide charge density will be $5 \times 10^{11} \text{ cm}^{-2}$. A doping density of $2.1 \times 10^{16} \text{ cm}^{-3}$ and a mean surface potential of 8.1 have been used in calculating these values of charge density.

It is found experimentally that (8) and (9) are valid over the frequency range from 50 Hz to 500 kHz. The curves in Figs. 3, 4, and 5 cover the frequency range from 10^{-1} to 10^8 Hz. In extending these curves over a wider frequency range than covered by the conductance measurements, it is assumed that no new important ohmic loss mechanisms arise at the lower and higher frequencies.

Fig. 6 shows open circuit mean square noise voltage calculated from (2) using (8) and (9) as a function of mean surface potential. A frequency of 10 kHz and a standard deviation of 2.6 have been used in this calculation.

In the practical case, the noise voltage curve peaks at the same value of mean surface potential and has the same shape as the equivalent parallel conductance vs \bar{u}_s which would be measured across terminals $x-x$ in Fig. 1.

At a given frequency, Fig. 6 shows that mean square noise voltage decreases at values of mean surface potential near flat bands and saturates in accumulation. A constant density of states with energy has been used in calculating the curve in Fig. 6. Actually, the density of states increases rapidly toward the band edges as shown by Gray and Brown.¹³ This means that mean square noise voltage would be greater near flat bands than indicated in Fig. 6. The noise spectrum in this region, however, would have a shape similar to the curves in Fig. 4.

The mean square noise voltage decreases at values of mean surface potential near mid-gap. Because the theory developed here considers only majority carrier transitions, it does not apply without error when the Fermi level is within a few kT/q of mid-gap where both majority and minority carrier transitions become important. For this reason, the curve in Fig. 6 is shown as a dotted extrapolation in this region.

In the region of weak inversion where the Fermi level has moved past mid-gap a few kT/q toward the minority carrier band, the time constant dispersion disappears.^{3, 4} In this region, the noise spectrum is expected to be similar to curve (a) in Fig. 2 for a single time constant.

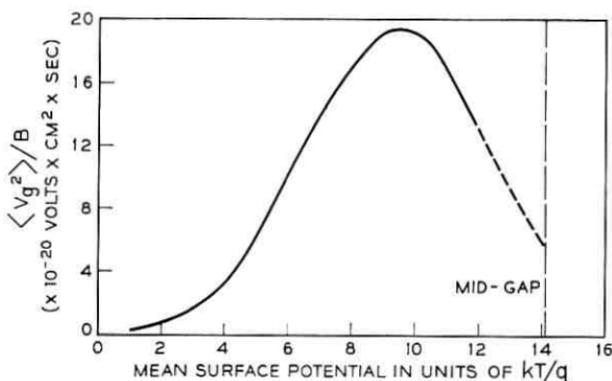


Fig. 6—Open circuit mean square noise voltage of MOS capacitor vs mean surface potential in units of kT/q . The curve is calculated from (2), (8), and (9) using a frequency of 10 kHz and a standard deviation, acceptor density, hole capture probability, interface state density, and temperature the same as in Fig. 4. Mid-gap is at $\bar{u}_s = 14.1$. Notice similarity in shape to experimental curves in Figs. 1 and 2 of Ref. 1.

IV. SUMMARY AND CONCLUSIONS

A theory of $1/f$ type noise has been presented based on a model for the interface state time constant distribution which quantitatively fits MIS conductance measurements. The noise spectra have been obtained from measured capture conductances $G_p(\omega)$ through the relation $s(\omega) = 4kTB(qw)^{-2} G_p(\omega)$ which is independent of the particular model used. Thus, the noise spectra presented in this paper are based solely on measured time constant dispersion.

The mean square noise voltage vs frequency in the model discussed in this paper depends on three quantities for a given temperature and silicon resistivity

(i) It depends upon the majority carrier density at the silicon surface which in turn is determined by field plate bias.

(ii) It depends upon the standard deviation of surface potential which is related to fixed oxide charge density.

(iii) It depends upon the density of interface states as seen from (8) and (9) combined with (2).

This theory predicts that $1/f$ type noise due to interface states can be reduced by decreasing the density of states which can exchange charge with the silicon and the density of fixed oxide charge. One consequence is that $1/f$ type noise can be calculated from the electrical properties of the interface obtained by measuring the admittance of an MOS capacitor.

V. ACKNOWLEDGMENTS

We wish to thank J. A. Morton for pointing out to us the possible connection between the results obtained from MIS admittance measurements and $1/f$ type noise. We also thank A. Goetzberger, H. K. Gummel, and R. M. Ryder for many useful discussions and their critical reading of the manuscript. A stimulating discussion with A. S. Grove is also acknowledged.

REFERENCES

1. Sah, C. T. and Hielscher, F. H., Evidence of the Surface Origin of the $1/f$ Noise, *Phys. Rev. Letters*, 17, October, 1966, pp. 956-958.
2. Nicollian, E. H. and Goetzberger, A., MOS Conductance Technique for Measuring Surface State Parameters, *Appl. Phys. Letters*, 7, October 15, 1965, pp. 216-219.
3. Nicollian, E. H. and Goetzberger, A., MOS Study of Interface-State Time

- Constant Dispersion, *Appl. Phys. Letters*, *10*, 15 January, 1967, pp. 60-62.
4. Nicollian, E. H. and Goetzberger, A., The Si-SiO₂ Interface-Electrical Properties as Determined by the Metal-Insulator-Semiconductor Conductance Technique, *B.S.T.J.*, *46*, July-August, 1967, pp. 1055-1133.
 5. McWhorter, A. L., 1/f Noise and Germanium Surface Properties, in R. H. Kingston (ed.), *Semiconductor Surface Physics*, University of Pennsylvania Press, Philadelphia, 1957, pp. 207-228.
 6. Atalla, M. M., Tannenbaum, E., and Scheibner, E. J., Stabilization of Silicon Surfaces by Thermally Grown Oxides, *B.S.T.J.*, *38*, May, 1959, pp. 749-783.
 7. Bess, L. A., A Possible Mechanism for 1/f Noise Generation in Semiconductor Filaments, *Phys. Rev.*, *91*, September, 1953, p. 1569.
 8. Jäntschi, O., On the Theory of 1/f Noise at Semiconductor Surfaces, (in German), *Verhandlungen der Deutschen Physikalischen Gesellschaft*, *6*, 1967, p. 35.
 9. Sah, C. T., Theory of Low-Frequency Generation Noise in Junction-Gate Field-Effect Transistors, *Proc. IEEE*, *52*, July, 1964, pp. 795-814.
 10. Lauritzen, P. O., Low-Frequency Generation Noise in Junction Field Effect Transistors, *Solid-State Elec.*, *8*, 1965, pp. 41-58.
 11. Lehovec, K., Frequency Dependence of the Impedance of Distributed Surface States in MOS Structures, *Appl. Phys. Letters*, *8*, 1966, pp. 48-50.
 12. Van der Ziel, A., On the Noise Spectra of Semiconductor Noise and of Flicker Effect, *Physica*, *16*, 1950, pp. 352-359.
 13. Gray, P. V. and Brown, D. M., Density of SiO₂-Si Interface States, *Appl. Phys. Letters*, *8*, 1966, p. 31.

Stability Considerations in Lossless Varactor Frequency Multipliers

By V. K. PRABHU

(Manuscript received May 31, 1967)

A general analysis of stability conditions of pumped nonlinear systems is presented in this paper. The type of instability investigated for these systems is that which causes spurious tones to appear at any point in the system in the vicinity of an appropriate harmonic carrier. A set of stability criteria that assure stability for the system has been given in terms of scattering parameters of the system. These criteria have then been applied to investigate the stability of lossless varactor harmonic generators that have been shown in this paper to be potentially unstable systems. It is then investigated for these multipliers how instability arises, and how it can be avoided by proper terminations. For some simple terminations, which are usually used in practice, sufficient conditions, that assure total stability of the multipliers, are explicitly given.

I. INTRODUCTION

One of the principal limitations to efficient wideband harmonic generation with varactor diodes is the generation of spurious signals.^{1, 2, 3} The origin of these signals is usually thought¹ to be due to a parametric "pumping up" of some signal in the multiplier passband, or to a parametric up-conversion process,¹ or a variation in the average capacitance of the diode at input frequency.³ A multiplier which contains these spurious signals is considered to be unstable,⁴ and it is this type of instability that is investigated in this paper.

At the present time, much is not known about the stability of harmonic generators, even though it is a widely-known experimental fact that this is a serious problem in high-efficiency varactor multipliers.^{2, 4} Very little is also known about the conditions imposed by stability on the available circuit configurations. Consequently, present design procedures leave the problem to be solved experimentally, and this is often done at the expense of efficiency. Very often isolators are used

to connect a chain of multipliers which are individually stable in order to guarantee stability of the chain.⁴ The isolators used in the chain always lower the overall efficiency.

A start on this problem of stability in multipliers has been made by Ref. 4 which considers the stability conditions of multipliers of order 2^n with minimum number of idlers. Some simple conditions on the terminations have been obtained⁴ in order to ensure stability of the multipliers. This paper extends this analysis to harmonic generators of arbitrary order and also obtains refinements to the conditions obtained in Ref. 4.

Varactor harmonic generators come under the general class of pumped nonlinear systems, which are systems driven periodically by a pump or a local oscillator at a frequency ω_0 .⁵ For such systems, a general method can be used⁵ to obtain the scattering parameters which relate the small-signal fluctuations present at various points in the system. In particular, Ref. 5 obtains these scattering relations for lossless abrupt-junction varactor multipliers of order 2^n , 3^s , and $2^n 3^s$, n and s integers, with the least number of idlers.

These scattering relations for pumped systems have been obtained in Ref. 5 when the difference frequency ω is real and small. The concept of analytic continuation has been used to obtain these scattering parameters when this difference frequency is complex, and is still small in magnitude.

Stability conditions for pumped systems are then expressed in terms of the scattering matrix of the system and a certain characteristic equation is obtained which determines the stability of the system. For the system to be stable it is necessary and sufficient that the roots of this characteristic equation must lie external to a region R of the complex frequency plane. Proper terminations that guarantee stability of the system can be determined for the pumped system from this equation.

We then discuss AM-to-PM and PM-to-AM conversion properties of a set of lossless interstage networks usually used with multipliers.

Stability conditions of lossless abrupt-junction varactor multipliers, most frequently encountered in practice, are then considered. It has been shown that if the bias circuit is properly designed⁶ so that there are no currents flowing in the vicinity of dc the characteristic equation[‡] of the multiplier can be expressed as a product of an AM characteristic

[‡]This condition can be achieved in practice by having a bias source with infinite internal impedance.

equation and a PM characteristic equation. If any root of the AM characteristic equation lies in the closed right-half of the complex plane there will not be a finite upper bound to the AM fluctuations originating at some point in the system. Such a system is defined to be unstable with respect to its AM fluctuations. Similarly, the PM fluctuations will be finite if and only if all zeros of the PM characteristic equation lie in the open left-half plane. For total stability of the multiplier no zero of its AM and PM characteristic equations should lie in the closed right-half plane.¶

It has been shown for multipliers of order 2^n that all roots of the AM characteristic equation always lie in the left-half plane for arbitrary values of input, output, and idler terminations.|| It has also been proved for these multipliers that PM stability is not achievable with arbitrary terminal impedances.

We then specifically consider PM stability of a 1-2 doubler, 1-2-4 quadrupler, and 1-2-4-8 octupler when their terminations are single-tuned series circuits.‡ Simple restrictions to be satisfied by these terminations are obtained to guarantee PM stability of the multipliers.

Stability of a 1-2-3 tripler for an arbitrary passive idler termination is the subject of discussion of the next section. We show that a tripler is potentially unstable for arbitrary input and output terminations. It has also been proven that a tripler is stable with respect to both AM and PM fluctuations if its terminations are single-tuned series circuits.

We next assume that the bias source impedance Z_0 can be a finite number. We then show that the stability characterization of a multiplier having finite bias source impedance is the same as that of a multiplier having infinite bias source impedance.

For a multiplier of any order, a general method of obtaining the conditions on available circuit configurations imposed by the condition of stability has also been presented.

§ The closed right-half of the complex plane is the region of λ -plane where $\text{Re } \lambda \geq 0$. The open left-half plane contains all the points of the λ -plane for which $\text{Re } \lambda < 0$.

¶ For total stability of systems whose characteristic equation $F(\lambda)$ cannot be expressed as a product of AM and PM characteristic equations, it is necessary and sufficient that no zero of $F(\lambda)$ lies in the closed right-half plane.

|| All terminations considered in this paper are assumed to be linear and passive.

‡ It can be shown that a single-tuned series circuit is a first-order approximation to any circuit usually used in practice, since the average elastance S_0 of the varactor diode is almost always nonzero.⁷

scribed⁵ by an equation of the form

$$\mathbf{V} = \mathbf{Z}_{\alpha-\beta} \mathbf{I} \quad (4)$$

where \mathbf{V} and \mathbf{I} are the terminal voltage and current column matrices and \mathbf{Z} is an impedance matrix. We shall now utilize the principle of analytic continuation⁸ to obtain \mathbf{Z} (and other parameters) of the pumped nonlinear system when the difference frequency is complex. This can be done by the simple expedient of replacing the variable $j\omega$ by the complex variable $\lambda = \sigma + j\omega$ wherever it occurs⁸ in (4).§ The truth of this statement, expressing a property of functions known as their permanence of form, follows directly from the identity theorem, since \mathbf{Z} and its continuation obviously coincide on the $j\omega$ -axis.⁸

We can, therefore, obtain scattering parameters of all pumped nonlinear systems (including those of lossless abrupt-junction varactor multipliers) when the difference frequency λ is complex.

III. STABILITY OF PUMPED NONLINEAR SYSTEMS

We shall first begin with a discussion of stability of pumped nonlinear systems in which small-signal fluctuations may be present at various points in the system. Since lossless varactor harmonic generators are specific pumped nonlinear systems all these results and remarks also apply to these harmonic multipliers.

A small-signal fluctuation originating at some point in the system is propagated, in general, throughout the system. We shall define a pumped nonlinear system to be stable if and only if the amplitude of small-signal fluctuations at any point in the system is finite for a finite small-signal fluctuation originating at some point in the system.

We shall make use of some of the results obtained in the study of stability of linear n -port systems.^{9,10,11,12,13,14} The stability of a linear n -port system is usually described by the statement that the roots of a certain characteristic equation $F(\lambda)$ of the system must be external to a region R of the complex frequency plane, that is, $F(\lambda) \neq 0$ in region R , where $\lambda = \sigma + j\omega$ is the complex frequency variable. Some set of stability criteria can also be obtained^{9,10,11,12,13} for a general class of linear reciprocal and nonreciprocal n -ports. For a reciprocal twoport, a well-known result by Gewertz¹⁰ states that it is stable under all passive terminations if and only if it is passive. This theorem has been generalized by Youla¹² to the reciprocal n -port. Very little, how-

§ In order that $\delta v_k(t)$ is small compared to the carrier at frequency $k\omega_0$ for all time t , it is required that $\sigma \leq 0$.

ever, is known^{13,14} about the stability of linear nonreciprocal n -ports, when $n \geq 3$.

It is shown in Section II that the terminal small-signal behavior of noise-free pumped nonlinear system can be described by[†]

$$\mathbf{V} = \mathcal{Z}_{\alpha-\beta} \mathbf{I} \quad (4)$$

where $\mathcal{Z}_{\alpha-\beta}$ is a function of $j\omega_0$ and $\lambda = \sigma + j\omega$.

We shall restrict ourselves in this paper to the consideration of stability of pumped nonlinear systems having only two (physical) accessible ports. It can be noted, however, that most of the concepts developed for the system having two accessible ports can be extended in a straightforward manner if the system possesses more than two accessible terminal pairs. This will be evident to the reader when we discuss stability of a tripler elsewhere in this paper.

If ω_0 and ω_0 are the input and out put carrier frequencies, it can be shown⁵ that the AM and PM fluctuations at different points in the system can be related through a scattering matrix $\underline{\mathcal{S}}$:

$$\begin{bmatrix} (m_r)_l \\ (m_r)_s \\ (\theta_r)_l \\ (\theta_r)_s \end{bmatrix} = \begin{bmatrix} \underline{\mathcal{S}}_{aa} & \underline{\mathcal{S}}_{ap} \\ \underline{\mathcal{S}}_{pa} & \underline{\mathcal{S}}_{pp} \end{bmatrix} \begin{bmatrix} (m_i)_l \\ (m_i)_s \\ (\theta_i)_l \\ (\theta_i)_s \end{bmatrix}, \quad (5)$$

where m and θ are the AM and PM indexes of the system, $\underline{\mathcal{S}}_{aa}$ is the AM scattering matrix, etc. We shall write (5) as

$$\mathbf{b} = \underline{\mathcal{S}}\mathbf{a}. \quad (6)$$

Let the system be terminated in linear passive impedances (see Fig. 1) z_1, z_2, z_3 , and z_4 with reflection coefficients ρ_1, ρ_2, ρ_3 , and ρ_4 . § Let us define a matrix ρ where

$$\rho = \text{dia.} [\rho_1, \rho_2, \rho_3, \rho_4]. \quad (7)$$

Since z_i 's are assumed passive, we have

† Let \underline{A} be an arbitrary matrix. Then $\underline{A}^t, \underline{A}^*, \underline{A}^\dagger$, and $\Delta \underline{A}$ stand for the transpose, the complex conjugate, the complex conjugate transpose, and the determinant of \underline{A} , respectively. Column vectors are denoted by \mathbf{V}, \mathbf{I} , etc. A diagonal matrix $[\mu_i \delta_{ij}]$ $\{\delta_{ij} = 1, i = j; \delta_{ij} = 0, i \neq j\}$, is denoted as $\text{dia.} [\mu_1, \mu_2, \dots, \mu_n]$. $\underline{1}_n$ is the unit matrix of order n .

§ The linear impedances z_1, z_2, z_3 , and z_4 are normalized with respect to "characteristic impedances" at corresponding carrier frequencies. Characteristic impedance at input port is the "input impedance" and that at the output port is the "load impedance".⁷

$$|\rho_i| \leq 1, \quad 1 \leq i \leq 4, \quad (8)$$

for $\text{Re } \lambda \geq 0$.

From (6), we can show that the system is stable if and only if§

$$\Delta\{\underline{1}_4 - \underline{S}\rho\} \neq 0, \quad \text{for } \text{Re } \lambda \geq 0. \quad (9)$$

We can, therefore, state that the characteristic equation of the system is given by

$$F(\lambda) = \Delta\{\underline{1}_4 - \underline{S}\rho\} = 0; \quad (10)$$

and for stability of the system it is necessary and sufficient that no root of $F(\lambda)$ lies in the closed right-half plane.¶

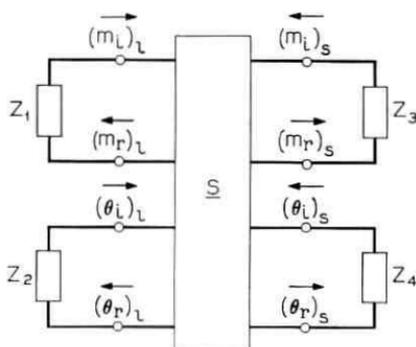


Fig. 1—Pumped nonlinear system, in amplitude-phase representation, terminated in linear passive impedances.

Theorem 1: We shall now show¹³ that two systems described by scattering matrices \underline{S}_1 and \underline{S}_2 possess identical stability characterizations if \underline{S}_1 and \underline{S}_2 possess identical principal minors¹⁵ of all order.

The characteristic equation $F(\lambda)$ of a system described by scattering matrix \underline{S} for a certain termination described by matrix ρ is given by (10). If \underline{S} is nonsingular, we can write (10) as

$$\Delta\{\underline{S}^{-1} - \rho\} = 0. \quad (11)$$

§ The constraints imposed on \underline{S} for a twoport system may be found in Ref. 14. These constraints, if satisfied, guarantee stability of the system independent of the terminations.

¶ The reader will recognize that $F(\lambda) = 0$ gives the natural frequencies of the system. For stability of a system, simple zeros of $F(\lambda)$ on the $j\omega$ -axis are usually allowed, since this just leads to sustained response of finite amplitude. However, multiple order zeros on the $j\omega$ -axis lead to instability of the system.

Now $\Delta\{\underline{S}^{-1} - \rho\}$ can be expanded in terms of the elements of ρ as follows:

$$\Delta\{\underline{S}^{-1} - \rho\} = \Delta\underline{S}^{-1} - \sum_{k=1}^4 \rho_k B_k + \sum_{k < r}^4 \rho_k \rho_r B_{k,r}, \quad (12)$$

where B_k is the principal minor of \underline{S}^{-1} obtained by striking out the k th row and column, $B_{k,r}$ is the principal minor obtained by deleting the k th and r th rows and the k th and r th columns. It, therefore, follows that two systems described by scattering matrices \underline{S}_1 and \underline{S}_2 have identical stability characterizations if \underline{S}_1^{-1} and \underline{S}_2^{-1} have identical principal minors of all order. We know that \underline{S}_1^{-1} and \underline{S}_2^{-1} possess identical principal minors of all order if and only if \underline{S}_1 and \underline{S}_2 possess identical principal minors of all order. This proves the theorem.

If $F(\lambda) \neq 0$ for $\text{Re } \lambda \geq 0$ for all allowable values of ρ , we shall say that the pumped system is absolutely stable. If there is only a set of ρ which meets this requirement the system will be considered to be conditionally (or potentially) stable. It can be observed that if one port of the system is terminated in a linear passive impedance z_i , and if the real part of the impedance across any other pair of terminals is negative for $\text{Re } \lambda \geq 0$, the system cannot be absolutely stable. This is one of the methods to investigate absolute stability of a system.

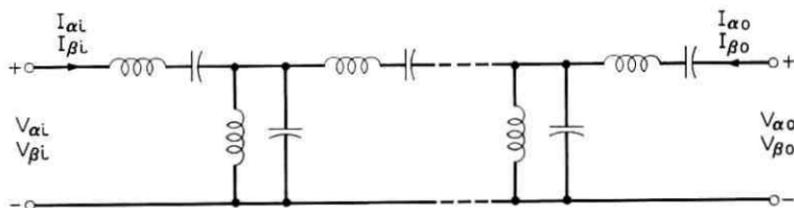


Fig. 2—Typical interstage network used in a multiplier. All series and shunt arms are resonant at frequency $k\omega_0$.

IV. SOME PROPERTIES OF A CLASS OF LOSSLESS INTERSTAGE NETWORKS

Frequency separation is obtained in harmonic generators by using linear bandpass[‡] filters. A typical example of a class of filters most commonly used in harmonic generators is shown in Fig. 2. This filter has a passband centered around carrier frequency $\pm k\omega_0$. Such filters with proper terminations are connected at accessible ports of a multiplier so as to obtain the desired frequency separations and proper impedance

[‡] This can be a low-pass filter at the lowest carrier frequency present in the multiplier and a high-pass filter at the highest carrier frequency.⁴

terminations at different carrier frequencies present in the multiplier. § A multiplier with input frequency ω_0 , output frequency $n\omega_0$, and interstage networks $N_1, N_2, \dots, N_k, \dots, N_n$ is shown in Fig. 3. ¶

For such interstage networks it will be shown that the scattering parameters || are given by

$$S = \begin{bmatrix} \underline{S}_{aa} & \underline{0} \\ \underline{0} & \underline{S}_{pp} \end{bmatrix} \quad (13)$$

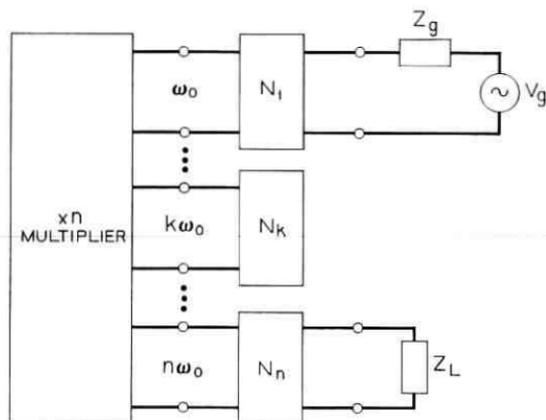


Fig. 3—Lossless interstage networks as used in a frequency multiplier.

so that these networks do not produce AM-to-PM or PM-to-AM conversion.

Since the series arms are resonant at frequency $k\omega_0$, and the antiresonant frequency of the shunt arms is also $k\omega_0$, if $\omega/\omega_0 \ll 1$, we can write

$$\begin{bmatrix} V_{\alpha i} \\ V_{\beta i} \\ V_{\alpha 0} \\ V_{\beta 0} \end{bmatrix} = \begin{bmatrix} z_{ii} & 0 & z_{i0} & 0 \\ 0 & z_{ii} & 0 & z_{i0} \\ z_{0i} & 0 & z_{00} & 0 \\ 0 & z_{0i} & 0 & z_{00} \end{bmatrix} \begin{bmatrix} I_{\alpha i} \\ I_{\beta i} \\ I_{\alpha 0} \\ I_{\beta 0} \end{bmatrix}. \quad (14)$$

§ For example, this filter should also act as a matching filter at the input carrier frequency ω_0 .

¶ It is assumed that all idler terminations are lossless.

|| Even though N_k is a two-port network we must obtain 4x4 scattering matrix of this network since amplitude and phase transmission characteristics of the pumped nonlinear system with which N_k may be used are not necessarily the same.⁶ See Ref. 5 for the definitions of amplitude and phase transmission characteristics as used in this paper.

We shall now assume that large signal voltage at carrier frequency $k\omega_0$ is in phase with the large signal current.‡

We can now write

$$\begin{bmatrix} V_{ai} \\ V_{pi} \\ V_{a0} \\ V_{p0} \end{bmatrix} = \begin{bmatrix} z_{ii} & 0 & z_{i0} & 0 \\ 0 & z_{ii} & 0 & z_{i0} \\ z_{0i} & 0 & z_{00} & 0 \\ 0 & z_{0i} & 0 & z_{00} \end{bmatrix} \begin{bmatrix} I_{ai} \\ I_{pi} \\ I_{a0} \\ I_{p0} \end{bmatrix}. \quad (15)$$

Equations (14) and (15) show that the scattering parameters of a lossless interstage network are given by (13). This shows that if such interstage networks are used in multipliers which are characterized by uncoupled§ scattering matrices the resultant scattering matrix is also uncoupled.

V. STABILITY OF LOSSLESS ABRUPT-JUNCTION VARACTOR MULTIPLIERS

The general analysis of the stability conditions presented in the earlier sections will be applied to investigate stability of frequency multipliers of order $2^n 3^s$, n and s integers, when lossless interstage networks of the form discussed in Section IV are used with these multipliers. It will be shown that these multipliers are potentially unstable and we shall obtain some circuit configurations which guarantee their conditional stability.

It has been shown⁵ that a multiplier of order $2^n 3^s$ with any input, output, and idler terminations can be considered as a chain of n doublers, s triplers, and $n + 2s + 1$ interstage networks (see Fig. 4). All these interstage networks¶ will be assumed to be of the form presented in Section IV. A lossless abrupt-junction varactor tripler with an arbitrary lossless idler termination is shown in Fig. 5. It is assumed that the tripler is tuned at the idler frequency, $Z_2(2\omega_0) = 0$, and that $\omega/\omega_0 \ll 1$. By the techniques of Ref. 5 we can show that the scattering parameters of a tripler can be represented as

‡ This condition usually leads to optimum efficiency of multipliers and is usually satisfied in practice.⁷

§ The scattering matrix is defined by us to be an uncoupled scattering matrix if $S_{ap} = S_{pa} = 0$.

¶ The average elastance S_0 of the varactor diode is considered as a part of the interstage networks used in the multipliers.

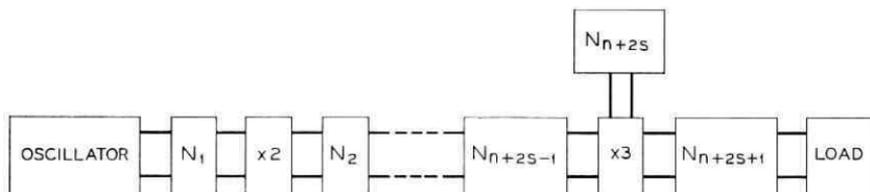


Fig. 4—Lossless abrupt-junction varactor multiplier of order $2^n 3^r$. N_i is an interstage network of the form shown in Fig. 2.

$$S = \left[\begin{array}{cc|cc} 0 & \frac{\mu - 1/2}{\mu + 3/2} & & 0 \\ 1 & \frac{-1}{\mu + 3/2} & & \\ \hline & & \frac{-1}{\mu + 1/2} & \frac{1}{3} \frac{\mu - 3/2}{\mu + 1/2} \\ 0 & & 3 & 0 \end{array} \right], \quad (16)$$

where

$$\mu = \frac{R_{02}}{Z_2}, \quad (17)$$

$$R_{02} = \frac{3 |S_1|^2}{8 |S_2| \omega_0}. \quad (18)$$

For a tripler, we can hence write

$$S_{aa} = \begin{bmatrix} 0 & \frac{\mu - 1/2}{\mu + 3/2} \\ 1 & \frac{-1}{\mu + 3/2} \end{bmatrix} \quad (19)$$

$$S_{pp} = \begin{bmatrix} \frac{-1}{\mu + 1/2} & \frac{1}{3} \frac{\mu - 3/2}{\mu + 1/2} \\ 3 & 0 \end{bmatrix} \quad (20)$$

and

$$S_{ap} = S_{pa} = 0. \quad (21)$$

Since a doubler,⁵ a tripler, and all interstage networks used in the multiplier have uncoupled scattering matrices it follows that general

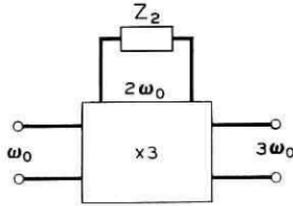


Fig. 5—Lossless abrupt-junction varactor tripler with an arbitrary lossless idler termination Z_2 .

scattering parameters of multipliers of order $2^n 3^s$ are given by the following equation:

$$\underline{S} = \begin{bmatrix} \underline{S}_{aa} & \underline{0} \\ \underline{0} & \underline{S}_{pp} \end{bmatrix}. \tag{22}$$

If such a multiplier is terminated in passive impedances as shown in Fig. 6, the characteristic equation of the system according to (10) can be written as

$$F(\lambda) = \Delta\{\underline{1}_A - \underline{S}\rho\} = 0, \tag{23}$$

where

$$\begin{aligned} \rho &= \text{dia.} [\rho_{m1}, \rho_{m2}, \rho_{\theta1}, \rho_{\theta2}] \\ &= \begin{bmatrix} \underline{\rho}_m & \underline{0} \\ \underline{0} & \underline{\rho}_\theta \end{bmatrix}. \end{aligned} \tag{24}$$

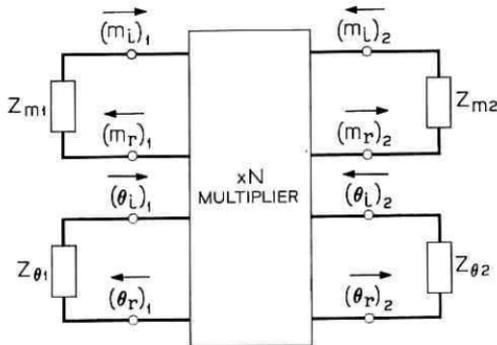


Fig. 6—Multiplier of order N . AM and PM ports of the multiplier are terminated in linear passive impedances.

From (22) through (24), we can write

$$F(\lambda) = \Delta\{\underline{1}_2 - \underline{S}_{aa}\rho_m\}\Delta\{\underline{1}_2 - \underline{S}_{pp}\rho_\theta\} \quad (25)$$

$$= F_a(\lambda)F_p(\lambda), \quad (26)$$

where

$$F_a(\lambda) = \Delta\{\underline{1}_2 - \underline{S}_{aa}\rho_m\} \quad (27)$$

and

$$F_p(\lambda) = \Delta\{\underline{1}_2 - \underline{S}_{pp}\rho_\theta\}. \quad (28)$$

For stability of the multiplier it is necessary and sufficient that the zeros of $F_a(\lambda)$ and $F_p(\lambda)$ lie in a region external to the closed right-half plane. $F_a(\lambda)$ and $F_p(\lambda)$ will be called the AM and PM characteristic equations of the multiplier respectively. It must be borne in mind that the uncoupled nature of the scattering matrix of the multiplier with a properly designed bias circuit enables us to express $F(\lambda)$ as a product of $F_a(\lambda)$ and $F_p(\lambda)$. If this cannot be done we will not be able to investigate the nature of roots of $F(\lambda)$ by studying only the roots of $F_a(\lambda)$ and $F_p(\lambda)$.

For multipliers for which we can express $F(\lambda)$ as the product of $F_a(\lambda)$ and $F_p(\lambda)$ we can define AM and PM stability independently. If no zeros of $F_a(\lambda)$ lie in the closed right-half plane we shall say that the multiplier is AM stable. A multiplier is PM stable if all roots of $F_p(\lambda)$ lie in the open left-half plane. For total stability of the multiplier it must be both AM and PM stable.

5.1 AM Stability of Multipliers of Order 2^n

The AM stability of lossless abrupt-junction varactor multipliers of order 2^n with minimum number of idlers will be considered in this section. It has been shown⁵ that a multiplier of order 2^n is equivalent to a cascade of n doublers as shown in Fig. 7. It will be assumed that interstage networks are passive, do not produce AM to PM or PM to AM conversion, and that the load z_n is a linear passive impedance. Since

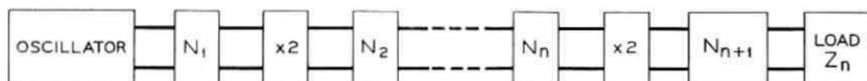


Fig. 7—Lossless abrupt-junction varactor multiplier of order 2^n . Only AM (or PM) ports of the doubler and interstage networks are shown in the figure.

N_{n+1} is a passive interstage network it follows that the amplitude terminal impedance for the n th doubler is also passive.

Let us now assume that the terminal impedance of the j th doubler is z_j where z_j is passive. We shall now show that the input impedance $(z_{in})_j$ of the j th doubler (see Fig. 8) is passive, $1 \leq j \leq n$. Since the generator impedance is assumed to be passive, no AM instability can arise in the multiplier.

The AM scattering matrix of a doubler is given by

$$\underline{S}_{aa} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ 1 & 0 \end{bmatrix}. \quad (29)$$

Let the reflection coefficient of z_j normalized to some convenient number be ρ_j . It can be shown¹⁶ that

$$|\rho_j| \leq 1, \quad \text{for } \text{Re } \lambda \geq 0. \quad (30)$$

From (29), we have,¹⁶

$$(\rho_{in})_j = \frac{1}{2}\{1 - \rho_j\}. \quad (31)$$

From (30) and (31), it follows that

$$|(\rho_{in})_j| \leq 1, \quad \text{for } \text{Re } \lambda \geq 0. \quad (32)$$

Equation (32) proves the desired result that if z_j is passive, $(z_{in})_j$ is also passive.

This shows that if input, output, and all idler terminations of a multiplier of order 2^n are passive, the impedance measured at any accessible pair of terminals is also passive. This result leads to the conclusion¹³ that a multiplier of order 2^n is absolutely stable with respect to its AM fluctuations.

5.2 PM Stability of Multipliers of Order 2^n

The phase terminal behavior of a multiplier of order 2^n has also been shown⁵ to be equivalent to a chain of n doublers as shown in Fig. 7.

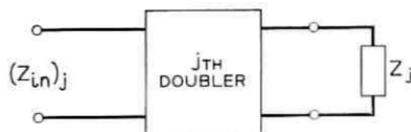


Fig. 8— j th doubler.

The PM scattering matrix of a doubler is given by

$$S_{pp} = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}. \quad (33)$$

If the phase terminal impedance of j th doubler has a reflection coefficient $(\rho_p)_j$, we have

$$\{(\rho_p)_{in}\}_j = \frac{-2(\rho_p)_j}{1 - (\rho_p)_j}. \quad (34)$$

For $(\rho_p)_j = \frac{1}{2}$, $\{(\rho_p)_{in}\}_j = -2$. This shows that the phase input impedance of j th doubler is not necessarily passive if its phase terminal impedance is passive. A doubler is, therefore, potentially unstable with respect to its PM fluctuations if its phase port is terminated in an arbitrary passive impedance. For this reason, we conclude that a multiplier of order 2^n , $n \geq 1$, can become unstable with respect to its PM fluctuations for some values of its input, output, and idler terminations.

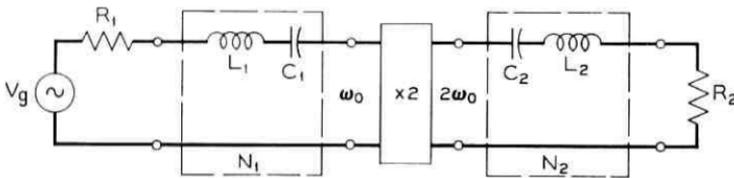


Fig. 9—Lossless abrupt-junction varactor doubler. Interstage networks N_1 and N_2 are assumed to be single-tuned series circuits.

The PM stability of a doubler, a quadrupler, and an octupler when interstage networks are single-tuned series circuits is studied next. Since the average elastance of a varactor diode is always nonzero, these circuits are always a first-order approximation to any circuits usually used in practice. For any other set of interstage networks used in the multiplier recourse can be had to Section V to obtain the constraints imposed by the condition of PM stability.

5.3 PM Stability of a Doubler

A lossless abrupt-junction varactor doubler with single-tuned series circuits for its generator and load impedances is shown in Fig. 9. R_1 and R_2 are the real parts of generator and load impedances of the multiplier.‡ These are given⁵ by

‡ It is assumed that the generator is matched to the varactor diode at carrier frequency ω_c .

$$R_1 = \frac{|S_2|}{\omega_0}, \quad (35)$$

and

$$R_2 = \frac{|S_1|^2}{4 |S_2| \omega_0}. \quad (36)$$

The bandwidths B_i 's for the single-tuned series circuits are defined as

$$B_i = \frac{R_{0i}}{L_i}, \quad 1 \leq i \leq 2, \quad (37)$$

where R_{0i} is the normalizing number for the i th termination. It is assumed for the doubler that

$$R_{0i} = R_i, \quad 1 \leq i \leq 2. \quad (38)$$

From (28), (33), and (37), we can show that the PM characteristic equation $F_p(\lambda)$ of the doubler can be represented as

$$F_p(\lambda) = 2\lambda^2 + B_2\lambda + B_1B_2 = 0. \quad (39)$$

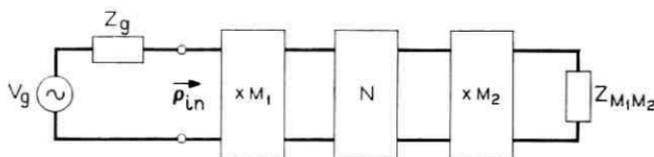
We can observe from (39) that a doubler is PM stable for any finite nonzero values of B_1 and B_2 . Therefore, it follows that a doubler is conditionally stable with respect to its AM and PM fluctuations if single-tuned series circuits are used for its input and output terminations.

5.4 PM Stability of a quadrupler

Before we discuss PM stability of a quadrupler we shall present in this section a systematic method to obtain the characteristic equation of a multiplier of any order which is equivalent to a chain of multipliers.⁵ Let us say that a multiplier of order $M_1 \times M_2$ is equivalent[‡] to a multiplier of order M_1 cascaded with a multiplier of order M_2 as shown in Fig. 10. It is assumed that the 2×2 scattering matrices of M_1 , M_2 , and the linear interstage network N are known. The impedance $Z_{M_1M_2}$ is assumed to be normalized with respect to its port number.¹⁶ The reflection coefficient $\rho_{M_1M_2}$ of the load termination $Z_{M_1M_2}$ is given by

$$\rho_{M_1M_2} = \frac{Z_{M_1M_2} - 1}{Z_{M_1M_2} + 1}. \quad (40)$$

[‡] The conditions under which this is true are given in Ref. 5.

Fig. 10 — Multiplier of order $M_1 \times M_2$.

Since the scattering matrices of M_1 , M_2 , and N are known, reflection coefficient ρ_{in} can be calculated. If the generator reflection coefficient ρ_g is given by

$$\rho_g = \frac{Z_g - 1}{Z_g + 1}, \quad (41)$$

the characteristic equation of the multiplier is given by

$$1 - \rho_g \rho_{in} = 0. \quad (42)$$

Let us now consider PM stability of a quadrupler. A lossless abrupt-junction varactor quadrupler is equivalent to a cascade of two doublers. We shall now investigate its PM stability when its input, output, and idler terminations are single-tuned series circuits as shown in Fig. 11. The normalizing impedance for the idler port is assumed to be

$$R_{02} = \frac{|S_1|^2}{4 |S_2| \omega_0}. \quad (43)$$

It can be noted that R_{02} is the "input impedance" of the second doubler. The bandwidths B_i 's are defined as in the earlier section.

We can now show that the PM characteristic equation of a quadrupler can be written as

$$F_p(\lambda) = 4\lambda^3 + 2\lambda^2(B_4 - B_2) + \lambda(2B_1B_2 + B_2B_4) + B_1B_2B_4 = 0. \quad (44)$$

In order that a quadrupler is PM stable it is necessary and sufficient that no zero of (44) lies in the closed right-half plane. The Routh-

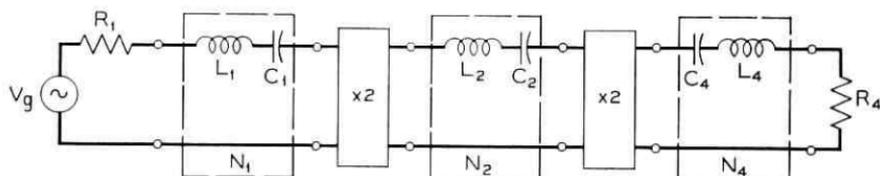


Fig. 11 — Lossless abrupt-junction varactor quadrupler. Interstage networks N_1 , N_2 , and N_4 are single-tuned series circuits.

Hurwitz¹⁷ criteria can be used to obtain the constraints on the coefficients so that the quadrupler is PM stable. It can be shown from this criterion that if

$$\frac{B_4}{B_2} > 2 \frac{B_1}{B_4} + 1 \quad (45)$$

all the zeros of (44) lie in the open left-half plane and the quadrupler is PM stable. Hence, we conclude that a quadrupler can be made conditionally stable[‡] if (45) is satisfied.

Let us now assume that

$$\frac{B_4}{B_2} = \frac{B_2}{B_1} = \gamma. \quad (46)$$

The minimum value of γ which guarantees PM stability of the multiplier can be obtained from (45). We can show that (45) is satisfied if and only if

$$\gamma > 1.629. \quad (47)$$

Specifically, we would like to note here that a quadrupler becomes unstable with respect to its PM fluctuations if $B_2 \rightarrow \infty$.

Also, we note that it is PM stable if simple bandwidth restrictions given by (45) or (47) are satisfied.

5.5 PM Stability of an Octupler

The AM stability of an octupler has been proved earlier in this section. The PM characteristic equation of an octupler with single-tuned series circuits for its input, output, and idler terminations can be shown to be given by the following equation:

$$\begin{aligned} F_p(\lambda) = & 8\lambda^4 + 4\lambda^3(B_8 - B_4 - B_2) \\ & + 2\lambda^2(2B_1B_2 + 3B_2B_4 - B_2B_8 + B_1B_8) \\ & + \lambda(2B_1B_2B_8 + B_2B_4B_8 - 2B_1B_2B_4) + B_1B_2B_4B_8 = 0. \end{aligned} \quad (48)$$

B_i is the bandwidth of the multiplier at carrier frequency $i\omega_0$.

The Routh-Hurwitz criterion can again be used to get the constraints on B_i 's so that the octupler is PM stable. These constraints can be shown to be

[‡] We have shown earlier in this section that a quadrupler is AM stable for all passive terminations.

$$\frac{B_8}{B_4} > \frac{B_2}{B_4} + 1 \quad (49)$$

$$2 \frac{B_1}{B_8} + 3 \frac{B_4}{B_8} + \frac{B_4}{B_2} > 1 \quad (50)$$

and

$$\begin{aligned} & 10 \left(\frac{B_8}{B_1} \right) + 2 \left(\frac{B_4}{B_1} \right) \left(\frac{B_8}{B_1} \right) - 14 \left(\frac{B_4}{B_1} \right) - 2 \left(\frac{B_8}{B_4} \right) \left(\frac{B_8}{B_1} \right) \\ & - \left(\frac{B_8}{B_1} \right)^2 + \left(\frac{B_4}{B_1} \right) \left(\frac{B_8}{B_1} \right) \left(\frac{B_8}{B_2} \right) - 3 \left(\frac{B_4}{B_1} \right)^2 + 6 \left(\frac{B_4}{B_8} \right) \left(\frac{B_4}{B_1} \right) \\ & - \left(\frac{B_4}{B_1} \right)^2 \left(\frac{B_8}{B_2} \right) - 4 \left(\frac{B_2}{B_4} \right) - 12 \left(\frac{B_2}{B_1} \right) + 4 \left(\frac{B_2}{B_8} \right) \\ & - 3 \left(\frac{B_2}{B_1} \right) \left(\frac{B_4}{B_1} \right) + 6 \left(\frac{B_2}{B_1} \right) \left(\frac{B_4}{B_8} \right) + 2 \left(\frac{B_2}{B_1} \right) \left(\frac{B_8}{B_4} \right) \\ & + \left(\frac{B_2}{B_1} \right) \left(\frac{B_8}{B_1} \right) > 0. \end{aligned} \quad (51)$$

If we can choose B_i 's so that we can satisfy (49) through (51), the multiplier will be PM stable. Let us now choose

$$\frac{B_8}{B_4} = \frac{B_4}{B_2} = \frac{B_2}{B_1} = x; \quad (52)$$

and see whether there exists a value of x which satisfies (49) through (51) simultaneously. The answer is in the affirmative and we can prove that if

$$x > 1.992 \quad (53)$$

the multiplier is PM stable. This shows that an octupler can be made conditionally stable by using single-tuned series circuits which satisfy certain bandwidth restrictions.

5.6 PM Stability of Multipliers of Order 2^n

Methods presented in earlier sections can be used to investigate PM stability of multipliers of order 2^n , $n \geq 4$. It is our conjecture based on earlier discussions and results that a multiplier of order 2^n with single-tuned series circuits as interstage networks is PM stable if bandwidths B_{2^i} 's, $0 \leq i \leq n$ satisfy the following equation:

$$\frac{B_{2^i}}{B_{2^{i+1}}} \ll 1. \quad (54)$$

VI. STABILITY OF A TRIPLER

The scattering relations for a tripler are given in (16). Even if the idler termination for the tripler is lossless it is evident from examining (19) and (20) that a tripler is not AM or PM stable[‡] for arbitrary input, and output terminations.

Hence, we shall assume that single-tuned series circuits are used for input, output, and idler terminations of the tripler as shown in Fig. 12. Bandwidths B_1 and B_3 are defined as usual. B_2 is defined as

$$B_2 = \frac{R_{02}}{L_2}, \quad (55)$$

where R_{02} is given in (18).

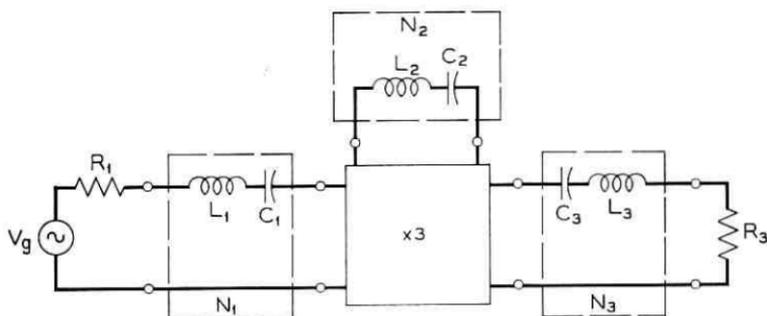


Fig. 12—Lossless abrupt-junction varactor tripler. N_1 , N_2 , and N_3 are single-tuned series circuits.

We can now obtain $F_a(\lambda)$ and $F_p(\lambda)$ for the tripler from (19) and (20). These can be shown to be given by

$$\begin{aligned} F_a(\lambda) &= 6\lambda^3 + \lambda^2(5B_1 + 3B_3) \\ &\quad + \lambda(B_1B_2 + B_2B_3 + 3B_3B_1) + B_1B_2B_3 \\ &= 0 \end{aligned} \quad (56)$$

and

$$\begin{aligned} F_p(\lambda) &= 6\lambda^3 + \lambda^2(B_1 + 3B_3) + \lambda(B_1B_2 + B_2B_3 + B_3B_1) + B_1B_2B_3 \\ &= 0. \end{aligned} \quad (57)$$

[‡] One of the reflection coefficients in \underline{S}_{pp} can be made in magnitude larger than unity by arbitrarily choosing μ . Also \underline{S}_{aa} does not satisfy the criterion given in Ref. 14 for the absolute AM stability of the system.

By Routh-Hurwitz criterion, it is necessary and sufficient that

$$5B_1(B_1B_2 + 3B_3B_1) + 3B_3(B_2B_3 + 3B_3B_1) + 2B_1B_2B_3 > 0 \quad (58)$$

so that no zero of $F_a(\lambda)$ lies in the closed right-half plane.

Similarly, for PM stability of the tripler, it is necessary and sufficient that

$$2\frac{B_3}{B_1} + \frac{B_1}{B_2} + 3\frac{B_3}{B_2} + \left\{ \frac{B_1}{B_3} + \frac{B_3}{B_1} - 2 \right\} > 0. \quad (59)$$

Since $(B_1/B_3) + (B_3/B_1) - 2 \geq 0$ for all positive values of B_1 and B_3 , it follows that a tripler is both AM and PM stable when single-tuned series circuits are used for its terminations. There are no bandwidth restrictions imposed by the condition of stability.

This does not mean that a tripler can be connected with another circuit (for example a stable doubler) without affecting the total stability of the system. We can indeed show that a 1-2-4-6 multiplier which is equivalent to a cascade of a doubler and a tripler imposes certain bandwidth restrictions on its external circuits so as to be assured of its stability.

VII. BIAS CIRCUIT AND ITS INFLUENCE ON THE STABILITY OF HARMONIC GENERATORS

It was assumed all along that the bias circuit in lossless abrupt-junction varactor multipliers is designed properly so that there are no currents flowing at sideband frequencies $\pm\omega$. We shall now assume that the varactor harmonic generator has a finite impedance at frequencies $\pm\omega$ so that there are currents flowing at those sideband frequencies. It will be our purpose in this section to investigate how this assumption affects the stability of the multiplier. The study of the influence of the bias circuit on the output signal-to-noise ratio of harmonic generators and other related results are reserved for a future publication in which we shall discuss noise performance of harmonic generators.

We shall also restrict ourselves in this section to the consideration of lossless abrupt-junction varactor harmonic generators which satisfy the following condition. If we choose the time origin so that carrier current I_1 is real and positive, all carrier currents I_k 's, $2 \leq k \leq n$, of the n th order harmonic generator are all real. We shall also assume that the multiplier is tuned at all carrier frequencies so that carrier voltages are in phase or out of phase with the respective carrier currents.

There are a large number of multipliers which by design satisfy

these conditions.^{7, 18} We know that the multipliers of order $2^n 3^m$ discussed in this paper come under this category. We can also show⁷ that the 1-2-4-5 quintupler can be designed to satisfy this condition.

Tuning circuits[†] for the multiplier are considered part of the terminations as shown in Fig. 13. We shall also assume that all idler terminations are lossless. The small-signal voltages $V_{\alpha k}$ and $V_{\beta k}$ at sideband frequencies $\pm k\omega_0 + \omega$ can be written as

$$V_{\alpha k} = \sum \frac{S_{k-l}}{j(l\omega_0 + \omega)} I_{\alpha l} + \sum \frac{S_{k+m}}{j(-m\omega_0 + \omega)} I_{\beta m} + \frac{S_k}{j\omega} I_{\alpha 0} \quad (60)$$

$$V_{\beta k} = \sum \frac{S_{-k+l}}{j(-l\omega_0 + \omega)} I_{\beta l} + \sum \frac{S_{-k-m}}{j(m\omega_0 + \omega)} I_{\alpha m} + \frac{S_{-k}}{j\omega} I_{\alpha 0} \quad (61)$$

$$V_{\alpha 0} = \sum \frac{S_{-l}}{j(l\omega_0 + \omega)} I_{\alpha l} + \sum \frac{S_m}{j(-m\omega_0 + \omega)} I_{\beta m} \quad (62)$$

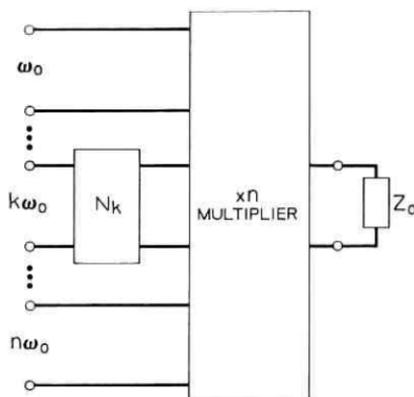


Fig. 13—Lossless abrupt-junction varactor harmonic generator of order n .

With the assumption that $\omega/\omega_0 \ll 1$, and using amplitude-phase representation, we can write (60) through (62) as[§]

$$V_{\alpha k} = \sum \pm \left| \frac{S_{k-l}}{l\omega_0} \right| I_{\alpha l} + \sum \pm \left| \frac{S_{k+m}}{m\omega_0} \right| I_{\beta m} \quad (63)$$

$$V_{\beta k} = \sum \pm \left| \frac{S_{k-l}}{l\omega_0} \right| I_{\beta l} + \sum \pm \left| \frac{S_{k+m}}{m\omega_0} \right| I_{\alpha m} \pm \left[\frac{S_k - S_k^*}{2\omega} \right] I_{\alpha 0} \quad (64)$$

[†] Average elastance S_0 of the varactor diode is included in these terminations.

[§] Note that S_k 's are all pure imaginary because of our assumptions about I_k 's.

and

$$V_{a0} = \sum \pm 2 \left| \frac{S_l}{l\omega_0} \right| I_{al} . \quad (65)$$

Let us now assume that all idler and bias terminations are such that[†]

$$V_{a0} = -Z_0 I_{a0} \quad (66)$$

$$V_{ak} = -Z_{ak} I_{ak} , \quad 2 \leq k \leq n-1 \quad (67)$$

and

$$V_{pk} = -Z_{pk} I_{pk} , \quad 2 \leq k \leq n-1 . \quad (68)$$

From (63) through (68), we can write

$$\begin{bmatrix} V_{a1} \\ V_{an} \\ V_{p1} \\ V_{pn} \end{bmatrix} = \begin{bmatrix} z_{n1a1} & z_{a1an} & 0 & 0 \\ z_{nna1} & z_{anna} & 0 & 0 \\ z_{p1a1} & z_{p1an} & z_{p1p1} & z_{p1pn} \\ z_{pna1} & z_{pnana} & z_{pnp1} & z_{pnpn} \end{bmatrix} . \quad (69)$$

The scattering parameters of a lossless abrupt-junction varactor harmonic generator hence can be described by

$$S = \begin{bmatrix} S_{aa} & 0 \\ S_{pa} & S_{pp} \end{bmatrix} . \quad (70)$$

It follows from (62) through (68) that S_{aa} and S_{pp} in (69) are the same as those that can be obtained by assuming $Z_0 = \infty$. For example, the scattering matrix of a doubler with finite bias source impedance Z_0 is given by⁶

$$S = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 1 & 0 & 0 \\ S_{pa} & 0 & -1 \\ & 2 & 1 \end{bmatrix} . \quad (71)$$

The characteristic equation of a harmonic generator with finite bias source impedance Z_0 can, according to (10), be represented as

$$F(\lambda) = \Delta\{\underline{1}_4 - S_{\rho}\} = 0, \quad (10)$$

[†] See Section IV.

where ρ is defined in Section III. From (10), (24), (25), and (70), we can write

$$F(\lambda) = \Delta\{\underline{1}_2 - \underline{S}_{aa}\rho_m\}\Delta\{\underline{1}_2 - \underline{S}_{pp}\rho_\theta\} \quad (72)$$

$$= F_a(\lambda)F_p(\lambda). \quad (73)$$

Equations (70) and (72) show that stability of a harmonic generator is not affected by the finite bias source impedance present in the multiplier even though it increases the output fluctuations of a harmonic generator.⁶ If a harmonic generator is stable for certain generator and load impedances for $Z_0 = \infty$, it is also stable when Z_0 is finite. This is one of the important results of this paper.

The conclusions arrived at in this section are applicable to harmonic generators of order $2^n 3^s$ discussed earlier in this section.

VIII. REMARKS AND CONCLUSIONS

A general method has been presented in this paper to investigate the stability of pumped nonlinear systems, and to obtain the conditions imposed thereby on the available circuit configurations. The type of instability investigated is that which causes spurious tones to appear at any point in the system in the vicinity of a carrier.

It has been shown that the roots of a certain characteristic equation

$$F(\lambda) = \Delta\{\underline{1}_4 - \underline{S}\rho\} = 0 \quad (10)$$

should lie in the open left-half plane for the system to be stable.

For lossless abrupt-junction varactor multipliers of order $2^n 3^s$ in which a certain set of interstage networks are used it has been shown that there is no AM-to-PM and PM-to-AM conversion and the characteristic equation can be expressed as

$$F(\lambda) = \Delta\{\underline{1}_2 - \underline{S}_{aa}\rho_m\}\Delta\{\underline{1}_2 - \underline{S}_{pp}\rho_\theta\} \quad (25)$$

$$= F_a(\lambda)F_p(\lambda), \quad (26)$$

and that we can treat separately AM and PM stabilities of the system.

A multiplier of order 2^n has been shown to be AM stable for all passive terminations. However, it is not absolutely stable with respect to PM fluctuations.

The conditional stability of a 1-2 doubler, 1-2-4 quadrupler, and 1-2-4-8 octupler is investigated next. All these multipliers are shown

to be PM stable if single-tuned series circuits are used as their terminations, and bandwidths B_i 's of these terminations satisfy certain conditions.

The PM characteristic equation of a doubler is given by

$$F_p(\lambda) = 2\lambda^2 + B_2\lambda + B_1B_2 = 0. \quad (39)$$

It is PM stable for any finite B_1 and B_2 .

A quadrupler has the following PM characteristic equation:

$$F_p(\lambda) = 4\lambda^3 + 2\lambda^2(B_4 - B_2) + \lambda(2B_1B_2 + B_2B_4) + B_1B_2B_4 = 0. \quad (44)$$

The quadrupler is PM stable if

$$\gamma > 1.629, \quad (47)$$

where

$$\frac{B_4}{B_2} = \frac{B_2}{B_1} = \gamma. \quad (46)$$

An octupler has also been shown to be PM stable if

$$x > 1.992, \quad (53)$$

where

$$\frac{B_8}{B_4} = \frac{B_4}{B_2} = \frac{B_2}{B_1} = x. \quad (52)$$

The scattering relations for a tripler when its idler termination is a passive impedance Z_2 are obtained. It has been shown that a tripler is not absolutely stable both with respect to its AM and PM fluctuations. However, it is stable when the interstage networks used in the tripler are single-tuned series circuits. The condition of stability does not impose any bandwidth restrictions.

Finally, it has been shown that the scattering matrix \underline{S} of a lossless abrupt-junction varactor harmonic generator with a finite bias source impedance Z_0 can be expressed as

$$\underline{S} = \left[\begin{array}{c|c} \underline{S}_{aa} & \underline{0} \\ \hline \underline{S}_{pa} & \underline{S}_{pp} \end{array} \right], \quad (70)$$

where \underline{S}_{aa} and \underline{S}_{pp} are the same as those obtained by assuming $Z_0 = \infty$. It is then shown that stability characterization of a lossless varactor harmonic generator is not affected by finite bias source impedance.

The noise analysis of harmonic generators and other related results will be discussed in a future publication.

REFERENCES

1. Buck, D. C., Origin of Spurious Signals in a Varactor Tripler, Proc. IEEE, *53*, No. 10, October, 1965, p. 1677.
2. Hines, M. F., Bloidsell, A. A., Collins, F., and Priest, W., Special Problems in Microwave Harmonic Generator Chain, Digest of Technical Papers, 1962 International Solid-State Circuits Conference, Philadelphia, Pa., 1962.
3. Burckhardt, C. B., Spurious Signals in Varactor Multipliers, Proc. IEEE, *53*, No. 4, April, 1965, pp. 389-390.
4. Dragone, C., Phase and Amplitude Modulation in High Efficiency Varactor Frequency Multipliers of Order $N = 2^n$ —Stability and Noise, B.S.T.J., *46*, No. 4, April, 1967, pp. 797-834.
5. Dragone, C. and Prabhu, V. K., Scattering Relations in Lossless Varactor Frequency Multipliers, B.S.T.J., *46*, October, 1967, pp. 1699-1731.
6. Prabhu, V. K., Noise Performance of Abrupt-Junction Varactor Frequency Multipliers, Proc. IEEE, *54*, No. 2, February, 1966, pp. 285-287.
7. Penfield, P., Jr. and Rafuse, R. P., *Varactor Applications*, The M.I.T. Press, Cambridge, Mass., 1962.
8. Guillemin, E. A., *The Mathematics of Circuit Analysis*, The Technology Press, Cambridge, Mass., and John Wiley and Sons, Inc., New York, N. Y., 1959.
9. Llewellyn, F. B., Some Fundamental Properties of Transmission Systems, Proc. IRE, *40*, No. 3, March, 1952, pp. 271-283.
10. Gewertz, C. M., Synthesis of a Finite, Four-Terminal Network from its Prescribed Driving-Point Functions and Transfer Function, J. Math. Phys., *12*, 1933, pp. 1-257.
11. Folke Bolinder, E., Survey of Some Properties of Linear Networks, IRE Trans. Circuit Theor., *CT-4*, 1957, pp. 70-78.
12. Youla, D. C., A Stability Characterization of the Reciprocal Linear Passive N -port, Proc. IRE, *47*, 1959, pp. 1150-1151.
13. Youla, D. C., A Note on the Stability of Linear, Nonreciprocal N -ports, Proc. IRE, *48*, 1960, pp. 121-122.
14. Ku, W. H., Unilateral Gain and Stability Criterion of Active Two-Ports in Terms of Scattering Parameters, Proc. IEEE, *54*, No. 11, November, 1966, pp. 1617-1618.
15. Hohn, F. E., *Elementary Matrix Algebra*, The MacMillan Company, New York, N. Y., 1963.
16. Carlin, H. J. and Giordano, A. B., *Network Theory*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1964.
17. Truxal, J. G., *Control Engineers' Handbook*, McGraw-Hill Book Company, Inc., New York, N. Y., 1958.
18. Burckhardt, C. B., Analysis of Varactor Frequency Multipliers for Arbitrary Capacitance Variation and Drive Level, B.S.T.J., *44*, No. 4, April, 1965, pp. 675-692.

Some Properties of a Classic Numerical Integration Formula

By I. W. SANDBERG

(Manuscript received May 19, 1967)

The numerical integration formula

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} + h \sum_{k=-1}^p b_k y'_{n-k}, \quad n \geq p \quad (1)$$

can be used to obtain a numerical solution of the system of nonlinear differential equations

$$\dot{x} + f(x, t) = 0, \quad t \geq 0 [x(0) = x_0]. \quad (2)$$

In many instances, it is known beforehand that the solution of (2) possesses a particular property such as boundedness or asymptotic periodicity with a given period, and it is then of interest to analytically determine the range of values of the step size h such that the sequence $\{y_n\}$ defined by (1) exhibits (at least) that property. In this paper, we consider problems of this type [but do not actually use assumptions concerning the character of the solution of (2)], and we study also the overall effect of solving instead of (1) the equation

$$z_{n+1} = \sum_{k=0}^p a_k z_{n-k} + h \sum_{k=-1}^p b_k z'_{n-k} + R_n, \quad n \geq p$$

which takes into account the effect of local roundoff errors and errors in the starting values. We consider explicitly only the case in which $x(t)$ is scalar valued.

I. INTRODUCTION

In this paper, we present some theorems concerning properties of the classic numerical integration formula¹

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} + h \sum_{k=-1}^p b_k y'_{n-k}, \quad n \geq p \quad (1)$$

a formula which can be used to obtain a numerical solution of the set of first-order nonlinear differential equations

$$\dot{x} + f(x, t) = 0, \quad t \geq 0 [x(0) = x_0]. \quad (2)$$

In (1) the y_n are approximations to the $x_n \triangleq x(nh)$, where h , a positive number, is the step-size parameter; y_0, y_1, \dots, y_p are starting vectors, the last p of which are obtained by an independent method; and

$$y'_n \triangleq -f(y_n, nh).$$

Specializations of (1) include, for example, Euler's method:

$$y_{n+1} = y_n + hy'_n, \quad (3)$$

and the more useful formula

$$y_{n+1} = y_n + \frac{1}{2}h(y'_n + y'_{n+1}). \quad (4)$$

In many instances it is known beforehand that the solution of (2) possesses a particular property such as boundedness or asymptotic periodicity with a given period, and it is then of interest to analytically determine the range (or ranges) of step sizes that will lead to a sequence $\{y_n\}$ which exhibits (at least) that property. This is one type of problem that we consider. For related material concerned with the overall effect of local truncation errors, see Ref. 2. Our results dealing with questions of asymptotic periodicity of the y_n are restricted to cases in which the basic period is a multiple of the step size h . However, it is often reasonable to choose h in this way to reduce programming complexity.

In addition to the fact that the solution of (1) differs from the samples of the solution of (2) due to truncation effects,^{1,3} the problem of solving (2) is further complicated by the fact that the numbers obtained from the computer differ from the y_n of (1) as a result of round-off errors. The local roundoff error R_n introduced in calculating y_{n+1} can be taken into account¹ by replacing (1) by

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} + h \sum_{k=-1}^p b_k y'_{n-k} + R_n, \quad n \geq p. \quad (5)$$

If $b_{-1} \neq 0$, the error in solving (1) for y_{n+1} , caused typically by truncating an iteration procedure^{1,3} after a finite number of steps, can be accounted for by redefining R_n . The second type of problem that we treat is to bound (from below as well as from above) a measure of the overall error in solving (5) instead of (1). The problem of estimating

the R_n before the calculations are performed is by no means trivial, and is not considered here. On the other hand, since there exist methods for bounding R_n given y_k for $(n-p) \leq k \leq n$ (see, for example, Wilkinson⁴, for bounds on the effect of roundoff in forming sums, products, etc.), our results suggest the feasibility of programming the computer to evaluate overall error bounds as the calculation of the successive y_{n+1} proceeds.

We shall explicitly consider only the case in which $x(t)$ and the y_n are scalars. Without much difficulty, each of the theorems can be extended to cover the vector case. In this extension, requirements on, for example, the derivative $\partial f(x, t)/\partial x$ are replaced by conditions on the Jacobian matrix of $f(x, t)$ (see Ref. 2).

For reasons that will become clear to the reader, our theorems are quite naturally characterized as "frequency-domain" results. Some of these theorems are close relatives of earlier results concerned with the input-output stability of nonlinear feedback systems* (see Ref. 5 and the difference-equation theorems stated without proof of Ref. 6). To the writer's knowledge, the only even remotely related material concerning (1) in the numerical-analysis literature, with the exception of Ref. 2, is Hamming's transfer-function approach.³

II. RESULTS†

We begin by introducing some definitions and assumptions. We assume throughout this section that y_n and $f(y_n, nh)$ are real-valued scalars.

Let α and β be two real constants, let $a_{-1} \triangleq 0$, and let

$$F(z) \triangleq 1 - \sum_{k=-1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)} \quad (6)$$

for all complex $z \neq 0$.

Assumption 1: It is assumed throughout that $1 + \frac{1}{2}(\alpha + \beta)hb_{-1} \neq 0$, and that $F(z) \neq 0$ for all $|z| \geq 1$.

This assumption implies that the sequence of approximations defined by (1) is bounded and approaches zero as $n \rightarrow \infty$ for all sets of starting values when $f(x, t) = \frac{1}{2}(\alpha + \beta)x$.

*The usual frequency-domain nonlinear system stability results such as Popov's criterion⁷ are not directly related because they do not deal with systems subjected to external inputs.

†The proofs of the theorems stated here are given in Section III.

Definitions

- (i) $\rho \triangleq \frac{1}{2}(\beta - \alpha)h \max_{0 \leq \omega \leq 2\pi} \left| \frac{\sum_{k=-1}^p b_k \exp[-i(k+1)\omega]}{F(e^{i\omega})} \right|$
- (ii)* $l_2 \triangleq \left\{ \{s_n\} \mid \sum_{n=0}^{\infty} |s_n|^2 < \infty \right\}$
 $l_{\infty} \triangleq \left\{ \{s_n\} \mid \sup_{n \geq 0} |s_n| < \infty \right\}$
- (iii)* Let K be a positive integer, and let
 $\mathcal{K} \triangleq \left\{ \{s_n\} \mid s_n = s_{n+K+1} \text{ for } n = 0, \pm 1, \pm 2, \dots \right\}$
- (iv) $\rho_K \triangleq \frac{1}{2}(\beta - \alpha)h \max_{\omega \in \mathcal{O}} \left| \frac{\sum_{k=-1}^p b_k \exp\left[-\frac{i(k+1)2\pi q}{K+1}\right]}{F\left[\exp\left(\frac{i2\pi q}{K+1}\right)\right]} \right|$
 in which $\mathcal{O} \triangleq \{0, 1, 2, \dots, K\}$.

2.1 Properties of (1)

Theorem 1: If

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=-1}^p b_k f[y_{n-k}, (n-k)h], \quad n \geq p$$

if $\rho < 1$, and if

$$\alpha \leq \frac{f(u, nh) - f(0, nh)}{u} \leq \beta, \quad n \geq 0$$

for all real $u \neq 0$, then

- (i) $\{f(0, nh)\} \in l_2$ implies that $\{y_n\} \in l_2$
- (ii) $\{f(0, nh)\} \in l_{\infty}$ implies that $\{y_n\} \in l_{\infty}$.

Remarks:

The condition that $\rho < 1$ is satisfied if and only if the locus of

$$\Theta(\omega) \triangleq \frac{\sum_{k=0}^p a_k \exp(ik\omega) - \exp(-i\omega)}{\sum_{k=-1}^p b_k \exp(ik\omega)} \tag{7}$$

* We consider only real sequences.

for $0 \leq \omega \leq 2\pi$ lies outside the "critical circle" C of radius $\frac{1}{2}(\beta - \alpha)h$ centered in the complex plane at $[\frac{1}{2}(\alpha + \beta)h, 0]$ (see Fig. 1).

For Euler's formula (3), we have $F(z) = 1 - [1 - \frac{1}{2}(\alpha + \beta)h]z^{-1}$, so that $F(z) \neq 0$ for $|z| \geq 1$ if and only if $0 < \frac{1}{2}(\alpha + \beta)h < 2$. For this formula the locus of Θ is the circle shown in Fig. 2, since $\Theta(\omega) = 1 - e^{-i\omega}$. If $\alpha h > 0$ and $\beta h < 2$, then the critical disk (Fig. 2) is not intersected by the locus of Θ , the condition that $0 < \frac{1}{2}(\alpha + \beta)h < 2$ is satisfied, and $\rho < 1$. Concerning the necessity of the condition $\rho < 1$, we note that if $\alpha h > 0$, but $\beta h > 2$, then for even the special case in which $f(x, t) = \beta x$, we have y_0, y_1, y_2, \dots unbounded (assuming merely that $y_0 \neq 0$).

For the formula (4):

$$F(z) = 1 + \frac{1}{4}(\alpha + \beta)h - [1 - \frac{1}{4}(\alpha + \beta)h]z^{-1}, \text{ and}$$

$$\Theta(\omega) = \frac{1 - e^{-i\omega}}{\frac{1}{2}(1 + e^{-i\omega})} = 2i \tan\left(\frac{\omega}{2}\right).$$

We have $1 + \frac{1}{4}(\alpha + \beta)h \neq 0$ and $F(z) \neq 0$ for $|z| \geq 1$ if and only if $(\alpha + \beta)h > 0$. The locus of Θ lies entirely on the imaginary axis of the complex plane,

$$\rho = \frac{\beta - \alpha}{\beta + \alpha},$$

and obviously $\rho < 1$ if $\alpha > 0$. On the other hand, if $\alpha < 0$, then for even the special case $f(x, t) = \alpha x : y_0, y_1, \dots$ is unbounded provided that $y_0 \neq 0$.

The following theorem is concerned with conditions under which

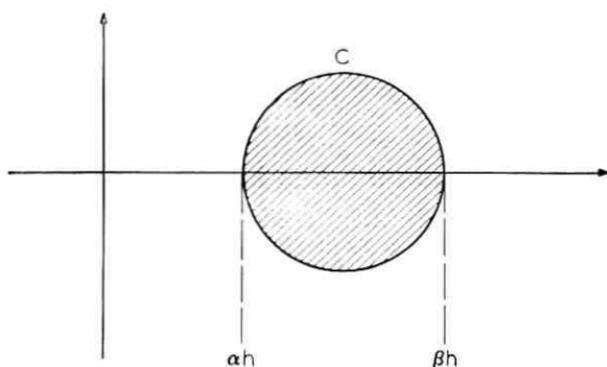


Fig. 1 — Location of the critical circle C .

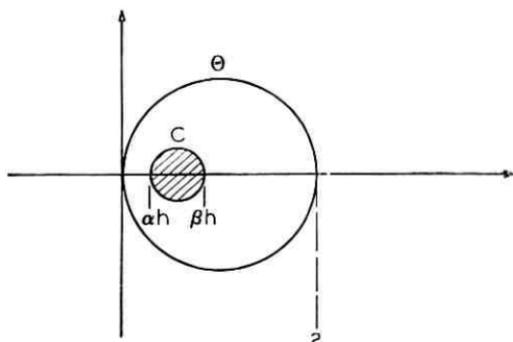


Fig. 2 — The locus of $\Theta(\omega)$ For Euler's method, and the critical circle C .

asymptotically periodic $f(0, nh)$ in (1) implies that $\{y_n\}$ is asymptotically periodic with the same period as that of $f(0, nh)$.

Theorem 2: If

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=1}^p b_k f[y_{n-k}, (n-k)h], \quad n \geq p$$

if $\rho < 1$, if $[f(u, nh) - f(0, nh)] = [f(u, (n + K + 1)h) - f(0, (n + K + 1)h)]$ for all real u and $n \geq 0$, if

$$\alpha \leq \frac{\partial f(u, nh)}{\partial u} \leq \beta, \quad n \geq 0$$

for all real u , and if there exists a $y_a^* \in \mathcal{K}$ such that $[f(0, nh) - y_a^*] \in l_2$, then there exists a $y_b^* \in \mathcal{K}$ such that

- (i) $(y - y_b^*) \in l_2$
- (ii) y_b^* is independent of $[f(0, nh) - y_a^*]$.

Remarks:

In many cases of interest $[f(u, nh) - f(0, nh)]$ is independent of n , and hence certainly satisfies the periodicity requirement.

Theorem 3, below, provides a condition under which the sequence $\{y_n\}$ of (1) cannot approach a "self sustained" limit cycle with period $(K + 1)$.

Theorem 3: If

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=1}^p f[y_{n-k}, (n-k)h], \quad n \geq p$$

if $[f(u, nh) - f(0, nh)] = [f(u, (n + K + 1)h) - f(0, (n + K + 1)h)]$ for all real u and $n \geq 0$, if

$$\alpha \leq \frac{\partial f(u, nh)}{\partial u} \leq \beta, \quad n \geq 0$$

for all real u , if $f(0, nh) \rightarrow 0$ as $n \rightarrow \infty$, and if there exists a $y^* \in \mathcal{K}$ different from the zero element of \mathcal{K} such that $(y_n - y_n^*) \rightarrow 0$ as $n \rightarrow \infty$, then $\rho_K \geq 1$.

Remark:

For $\rho_K \geq 1$, at least one of the complex numbers

$$\Theta\left(\frac{2\pi q}{K+1}\right) \quad q = 0, 1, 2, \dots, K$$

must lie on or within the circle C of Fig. 1.

2.2 Results Concerning the Effect of R_n and Errors in the Starting Values

Theorem 4, below, is essentially the same as a result concerning the effect of local roundoff and truncation errors proved in Ref. 2. The proof of Theorem 4 given in Section III is considerably more direct than the corresponding argument of Ref. 2.

Definition:

$$\langle s \rangle_N \triangleq \left(\frac{1}{N+1} \sum_{n=0}^N |s_n|^2 \right)^{\frac{1}{2}}$$

for all $N \geq 0$ and every sequence $\{s_n\}$.

Theorem 4: If

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=1}^p b_k f[y_{n-k}, (n-k)h], \quad n \geq p$$

$$z_{n+1} = \sum_{k=0}^p a_k z_{n-k} - h \sum_{k=1}^p b_k f[z_{n-k}, (n-k)h] + R_n, \quad n \geq p$$

if

$$\alpha \leq \frac{\partial f(u, nh)}{\partial u} \leq \beta, \quad n \geq 0$$

for all real u , then for all $N \geq 0$:

$$(i) \quad \langle y - z \rangle_N \geq (1 + \rho)^{-1} \min_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle \psi \rangle_N,$$

and

(ii) if $\rho < 1$,

$$\langle y - z \rangle_N \leq (1 - \rho)^{-1} \max_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle \psi \rangle_N$$

in which

$$\begin{aligned} \psi_n &= -R_{n-1}, \quad n \geq (p+1) \\ &= (y_n - z_n) - \sum_{k=0}^p a_k (y_{n-k-1} - z_{n-k-1}) \\ &\quad + h \sum_{k=-1}^p b_k \{f[y_{n-k-1}, (n-k-1)h] - f[z_{n-k-1}, (n-k-1)h]\}, \\ &\qquad\qquad\qquad n = 0, 1, 2, \dots, p \end{aligned}$$

with $y_n = f(y_n, nh) = z_n = f(z_n, nh) = 0$ for $n < 0$.

Remarks:

Ref. 2 considers two simple examples concerning the evaluation of the numbers

$$(1 + \rho)^{-1} \min_{\omega} |F(e^{i\omega})|^{-1} \quad \text{and} \quad (1 - \rho)^{-1} \max_{\omega} |F(e^{i\omega})|^{-1}.$$

Since

$$\rho = \frac{1}{2}(\beta - \alpha)h \left\{ \min_{\omega} |\Theta(\omega) - \frac{1}{2}(\alpha + \beta)h| \right\}^{-1},$$

we see that ρ is the ratio of the radius of the circle C of Fig. 1 to the distance between c and θ , where c is the center of C and θ is a point nearest c on the locus of $\Theta(\omega)$.

The following corollary provides asymptotic bounds on the difference between the solutions of (1) and (5) when the solution $\{y_n\}$ of (1) is, for example, asymptotically periodic.

Corollary to Theorem 4: If

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=-1}^p b_k f[y_{n-k}, (n-k)h], \quad n \geq p$$

with

$$\alpha \leq \frac{\partial f(u, nh)}{\partial u} \leq \beta$$

for all real u and $n \geq 0$, if there exists a sequence \tilde{y} such that $(y_n - \tilde{y}_n) \rightarrow 0$ as $n \rightarrow \infty$, and if

$$z_{n+1} = \sum_{k=0}^p a_k z_{n-k} - h \sum_{k=-1}^p b_k f[z_{n-k}, (n-k)h] + R_n, \quad n \geq p.$$

Then

(i)

$$\langle z - \tilde{y} \rangle_N \geq (1 + \rho)^{-1} \min_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle \psi \rangle_N - |q_N|$$

with $q_N \rightarrow 0$ as $N \rightarrow \infty$,

and

(ii) if $\rho < 1$,

$$\langle z - \tilde{y} \rangle_N \leq (1 - \rho)^{-1} \max_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle \psi \rangle_N + |r_N|$$

with $r_N \rightarrow 0$ as $N \rightarrow \infty$

in which

$$\begin{aligned} \psi_n &= R_{n-1}, & n &\geq (p+1) \\ &= 0, & n &= 0, 1, 2, \dots, p. \end{aligned}$$

Remark:

Note that the lower bound is valid under quite weak assumptions.

III. PROOFS

We first prove the following lemma which plays a role in the proofs of all of the theorems

Lemma 1: If

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=-1}^p b_k f[y_{n-k}, (n-k)h] + R_n, \quad n \geq p$$

then

$$y_n = \sum_{k=0}^n w_{n-k} g(y_k, kh) + \sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k, \quad n \geq 0$$

in which $\{w_n\}$ and $\{v_n\}$ are the inverse z -transforms of

$$W(z) \triangleq \frac{-h \sum_{k=-1}^p b_k z^{-(k+1)}}{1 - \sum_{k=-1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}}$$

and

$$V(z) \triangleq \frac{1}{1 - \sum_{k=1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}},$$

respectively;

$$\sum_{n=0}^{\infty} |w_n| < \infty, \quad \sum_{n=0}^{\infty} |v_n| < \infty,$$

$$g(y_k, kh) \triangleq f(y_k, kh) - f(0, kh) - \frac{1}{2}(\alpha + \beta)y_k,$$

and

$$\varphi_n = R_{n-1}, \quad n \geq (p+1)$$

$$= y_n - \sum_{k=0}^p a_k y_{n-k-1} + h \sum_{k=1}^p b_k f[y_{n-k-1}, (n-k-1)h],$$

$$n = 0, 1, 2, \dots, p$$

with $y_n = f(y_n, nh) \triangleq 0$ for $n < 0$.

Proof of Lemma 1:

From

$$y_{n+1} = \sum_{k=0}^p a_k y_{n-k} - h \sum_{k=1}^p b_k f[y_{n-k}, (n-k)h] + R_n, \quad n \geq p$$

we have

$$y_n = \sum_{k=0}^p a_k y_{n-k-1}$$

$$- h \sum_{k=1}^p b_k f[y_{n-k-1}, (n-k-1)h] + R_{n-1}, \quad n \geq (p+1)$$

and, with the φ_n as defined in the lemma,

$$y_n = \sum_{k=1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]y_{n-k-1} - h \sum_{k=1}^p b_k \delta_{n-k-1} + \varphi_n, \quad n \geq 0$$

where

$$\delta_k = f(y_k, kh) - \frac{1}{2}(\alpha + \beta)y_k.$$

Let $M > 0$. Then $y_n = \hat{y}_n$ for $n = 0, 1, \dots, M$, in which

$$\hat{y}_n = \sum_{k=1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]\hat{y}_{n-k-1} - h \sum_{k=1}^p b_k \hat{\delta}_{n-k-1} + \hat{\varphi}_n, \quad n \geq 0,$$

where

$$\begin{aligned} \hat{\delta}_n &= \delta_n \quad \text{for } n \leq M \\ &= 0 \quad \text{for } n > M, \\ \hat{\varphi}_n &= \varphi_n \quad \text{for } n \leq M \\ &= 0 \quad \text{for } n > M, \end{aligned}$$

and

$$\hat{y}_n = f(\hat{y}_n, nh) = 0 \quad \text{for } n < 0.$$

It is clear that $\{\hat{\varphi}_n\}$, $\{\hat{\delta}_n\}$, and $\{\hat{y}_n\}$ are z -transformable. Let

$$\psi(z) \triangleq \sum_{n=0}^{\infty} \hat{\varphi}_n z^{-n}, \quad \Delta(z) \triangleq \sum_{n=0}^{\infty} \hat{\delta}_n z^{-n},$$

and

$$Y(z) \triangleq \sum_{n=0}^{\infty} \hat{y}_n z^{-n}.$$

Then

$$\begin{aligned} \left[1 - \sum_{k=-1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)} \right] Y(z) \\ = -h \sum_{k=-1}^p b_k z^{-(k+1)} \Delta(z) + \psi(z). \end{aligned}$$

Therefore,

$$\begin{aligned} Y(z) &= \frac{-h \sum_{k=-1}^p b_k z^{-(k+1)}}{1 - \sum_{k=-1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}} \Delta(z) \\ &\quad + \frac{\psi(z)}{1 - \sum_{k=-1}^p [a_k - \frac{1}{2}(\alpha + \beta)hb_k]z^{-(k+1)}} \end{aligned}$$

and, with $\{w_n\}$ and $\{v_n\}$ the inverse z -transform of $W(z)$ and $V(z)$, respectively,* we have

$$\hat{y}_n = \sum_{k=0}^n w_{n-k} \hat{\delta}_k + \sum_{k=0}^n v_{n-k} \hat{\varphi}_k, \quad n \geq 0$$

* Recall that $W(z)$ and $V(z)$ are defined in Lemma 1.

with (in view of Assumption 1)

$$\sum_{n=0}^{\infty} |w_n| < \infty, \quad \text{and} \quad \sum_{n=0}^{\infty} |v_n| < \infty. \quad (8)$$

Thus,

$$y_n = \sum_{k=0}^n w_{n-k} \delta_k + \sum_{k=0}^n v_{n-k} \varphi_k \quad (9)$$

for $n = 0, 1, 2, \dots, M$. Since M is arbitrary, (9) is satisfied for all $n \geq 0$. Finally, with

$$g(y_k, kh) \triangleq f(y_k, kh) - f(0, kh) - \frac{1}{2}(\alpha + \beta)y_k$$

$$y_n = \sum_{k=0}^n w_{n-k} g(y_k, kh) + \sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k, \quad n \geq 0.$$

We now prove a lemma which is used in the proofs of most of the theorems. We repeat the

Definition:

$$\langle s \rangle_N \triangleq \left(\frac{1}{N+1} \sum_{n=0}^N |s_n|^2 \right)^{\frac{1}{2}}$$

for all $N \geq 0$ and every sequence $\{s_n\}$.

Lemma 2: If

$$y_n = \sum_{k=0}^n w_{n-k} a(k) y_k + b_n, \quad n \geq 0$$

and if $-\frac{1}{2}(\beta - \alpha) \leq a(k) \leq \frac{1}{2}(\beta - \alpha)$ for all $k \geq 0$, then

$$(i) \langle y \rangle_N \geq (1 + \rho)^{-1} \langle b \rangle_N \text{ for } N \geq 0,$$

and

$$(ii) \text{ if } \rho < 1, \text{ then } \langle y \rangle_N \geq (1 - \rho)^{-1} \langle b \rangle_N \text{ for } N \geq 0.$$

Proof of Lemma 2:

Let

$$q_n \triangleq \sum_{k=0}^n w_{n-k} a(k) y_k, \quad n \geq 0.$$

By Minkowski's inequality,

$$\langle y \rangle_N \leq \langle q \rangle_N + \langle b \rangle_N \quad (10)$$

and

$$\langle b \rangle_N \leq \langle y \rangle_N + \langle q \rangle_N . \tag{11}$$

Lemma 2 follows from (10), (11), and the inequality²

$$\langle q \rangle_N \leq \rho \langle y \rangle_N .$$

3.1 Proof of Theorem 1:

By Lemma 1, we have

$$y_n = \sum_{k=0}^n w_{n-k} g(y_k, kh) + \sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k , \quad n \geq 0$$

with (because $R_n = 0$ for all $n \geq p$) $\varphi_n = 0$ for all $n \geq (p + 1)$.

Let

$$b_n = \sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k , \quad n \geq 0 .$$

Since both $\{w_n\}$ and $\{v_n\}$ belong to l_1 [i.e., since (8) is satisfied], $b \in l_2$ if $\{f(0, kh)\} \in l_2$ and $b \in l_\infty$ if $\{f(0, kh)\} \in l_\infty$.

Suppose that $b \in l_2$, and let

$$\begin{aligned} a(k) &= \frac{g(y_k, kh)}{y_k} , \quad \text{for } y_k \neq 0 \\ &= 0, \quad \text{for } y_k = 0 . \end{aligned}$$

The function $a(k)$ satisfies the bounds of Lemma 2, and

$$y_n = \sum_{k=0}^n w_{n-k} a(k) y_k + b_n , \quad n \geq 0 \tag{12}$$

Therefore, by Lemma 2,

$$\sum_{n=0}^N |y_n|^2 \leq (1 - \rho)^{-2} \sum_{n=0}^N |b_n|^2 \leq (1 - \rho)^{-2} \sum_{n=0}^{\infty} |b_n|^2$$

for all $N \geq 0$, from which it is clear that $y \in l_2$.

If $b \in l_\infty$, then $\{y_n\}$ satisfies (12) with $b \in l_\infty$. According to the first conclusion of the following lemma, this implies that $y \in l_\infty$.

Lemma 3: If

$$y_n = \sum_{k=0}^n w_{n-k} a(k) y_k + b_n , \quad n \geq 0$$

with $b \in l_\infty$, if $\rho < 1$, and if $-\frac{1}{2}(\beta - \alpha) \leq a(k) \leq \frac{1}{2}(\beta - \alpha)$ for all $k \geq 0$, then

- (i) $y \in l_\infty$
 (ii) there exists a constant c_∞ , which depends on only the a_k , the b_k , α , and β such that

$$\sup_{n \geq 0} |y_n| \leq c_\infty \sup_{n \geq 0} |b_n|.$$

Proof of Lemma 3:

The proof is essentially the same as that of the second part of Theorem 2 of Ref. 2. The details are omitted.*

3.2 Proof of Theorem 2

Definitions: Let \mathcal{K} denote the set of all real sequences $\{s_n\}$ such that $s_n = s_{n+K+1}$ for all $n = 0, \pm 1, \pm 2, \dots$, and let $\mathcal{R} \triangleq \{0, 1, 2, \dots, K\}$.

Lemma 4: Let $g^*(x, nh)$ be defined for all real x and all $n = 0, \pm 1, \pm 2, \dots$, such that: $g^*(x, nh) = g^*[x, (n + K + 1)h]$ for all x and n , and

$$-\frac{1}{2}(\beta - \alpha) \leq \frac{\partial g^*(x, nh)}{\partial x} \leq \frac{1}{2}(\beta - \alpha)$$

for all x and n . If $p \in \mathcal{K}$ and if $\rho_K < 1$, then \mathcal{K} contains exactly one element y^* such that

$$y_n^* = \sum_{k=-\infty}^n w_{n-k} g^*(y_k, kh) + p_n$$

for $n = 0, \pm 1, \pm 2, \dots$.

Proof of Lemma 4:

With the norm

$$\|s\| \triangleq \left(\sum_{k=0}^K |s_k|^2 \right)^{\frac{1}{2}},$$

the set \mathcal{K} is a Banach space. The operator WG defined on \mathcal{K} by

$$(WGs)_n = \sum_{k=-\infty}^n w_{n-k} g^*(s_k, kh), \quad s \in \mathcal{K}$$

maps \mathcal{K} into itself. By the contraction-mapping fixed-point theorem, it suffices to show that WG is a contraction when $\rho_K < 1$. It is clear that

$$\begin{aligned} \|WGs_a - WGs_b\| &\leq \|W\| \cdot \|Gs_a - Gs_b\| \\ &\leq \frac{1}{2}(\beta - \alpha) \|W\| \cdot \|s_a - s_b\| \end{aligned}$$

for all $s_a \in \mathcal{K}$ and all $s_b \in \mathcal{K}$.

* See also Ref. 6.

If $s \in \mathfrak{K}$, then

$$s_k = \sum_{l=0}^K \hat{s}_l \exp\left(\frac{i2\pi lk}{K+1}\right) \text{ for } k = 0, \pm 1, \pm 2, \dots$$

in which

$$\hat{s}_l = (K+1)^{-1} \sum_{n=0}^K s_n \exp\left(-\frac{i2\pi ln}{K+1}\right)$$

and

$$\sum_{n=0}^K |s_n|^2 = (K+1) \sum_{n=0}^K |\hat{s}_n|^2.$$

Thus, if

$$u_n = \sum_{k=-\infty}^n w_{n-k} s_k \text{ for } n = 0, \pm 1, \pm 2, \dots$$

with $s \in \mathfrak{K}$, we find that

$$\begin{aligned} u_n &= \sum_{k=-\infty}^n w_{n-k} \sum_{l=0}^K \hat{s}_l \exp\left(\frac{i2\pi lk}{K+1}\right) \\ &= \sum_{l=0}^K \hat{s}_l \sum_{k=-\infty}^{\infty} w_{n-k} \exp\left(\frac{i2\pi lk}{K+1}\right) \quad (w_n = 0, n < 0) \\ &= \sum_{l=0}^K \hat{s}_l \exp\left(\frac{i2\pi ln}{K+1}\right) \sum_{n=0}^{\infty} w_n \exp\left(-\frac{i2\pi ln}{K+1}\right) \\ &= \sum_{l=0}^K W\left[\exp\left(\frac{i2\pi l}{K+1}\right)\right] \hat{s}_l \exp\left(\frac{i2\pi ln}{K+1}\right). \end{aligned}$$

Therefore, since

$$\|u\| = \|Ws\| \leq \max_{q \in \mathfrak{R}} \left| W\left[\exp\left(\frac{i2\pi q}{K+1}\right)\right] \right| \|s\|,$$

we have

$$\|W\| \leq \max_{q \in \mathfrak{R}} \left| W\left[\exp\left(\frac{i2\pi q}{K+1}\right)\right] \right|$$

and $\|WGs_a - WGs_b\| \leq \rho_K \|s_a - s_b\|$ for all $s_a \in \mathfrak{K}$ and all $s_b \in \mathfrak{K}$. This completes the proof of Lemma 4.

By Lemma 1,

$$y_n = \sum_{k=0}^n w_{n-k} g(y_k, kh) + \sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k, \quad n \geq 0$$

with $\varphi_k = 0$ for $k \geq (p + 1)$. Here, since both $\{w_n\}$ and $\{v_n\}$ belong to l_1 , we have

$$\sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k = p_n + c_n, \quad n \geq 0$$

with $p \in \mathcal{K}$ and $c \in l_2$. In fact, with y_n^* as defined in Theorem 2,

$$p_n = \sum_{k=-\infty}^n w_{n-k} y_{ak}^*, \quad n = 0, \pm 1, \pm 2, \dots$$

Let $g^*(x, nh)$ be defined by the conditions: $g^*(x, nh) = g^*[x, (n + K + 1)h]$ for all x and $n = 0, \pm 1, \pm 2, \dots$, and $g^*(x, nh) = g(x, nh)$ for all x and $n = 0, 1, \dots, K$. Then, since $\rho_K \leq \rho < 1$, by Lemma 4 there exists a $y_b^* \in \mathcal{K}$ such that

$$y_{bn}^* = \sum_{k=-\infty}^n w_{n-k} g^*(y_{bk}^*, kh) + p_n$$

for $n \geq 0$. Therefore,

$$y_n - y_{bn}^* = \sum_{k=0}^n w_{n-k} [g^*(y_k, kh) - g^*(y_{bk}^*, kh)] + d_n, \quad n \geq 0$$

in which

$$d_n = c_n - \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_{bk}^*, kh), \quad n \geq 0.$$

But

$$\left| \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_{bk}^*, kh) \right| \leq \sup_{n \geq 0} |g^*(y_{bn}^*, nh)| \sum_{k=-\infty}^{-1} |w_{n-k}|$$

and, using the fact that there exist constants $\eta > 0$ and $\zeta > 0$ such that $|w_n| \leq \eta \exp(-\zeta n)$ for $n \geq 0$,

$$\sum_{k=-\infty}^{-1} |w_{n-k}| = \sum_{m=(n+1)}^{\infty} |w_m| \leq \eta \exp[-\zeta(n+1)] \sum_{m=0}^{\infty} \exp(-\zeta m).$$

We see that

$$\sum_{k=-\infty}^{-1} w_{n-k} g^*(y_{bk}^*, kh) \in l_2,$$

and consequently $d \in l_2$.

Let

$$a(k) = \frac{g^*(y_k, kh) - g^*(y_{bk}^*, kh)}{y_k - y_{bk}^*}, \quad y_k \neq y_{bk}^* \\ = 0, \quad y_k = y_{bk}^*.$$

Then $-\frac{1}{2}(\beta - \alpha) \leq a(k) \leq \frac{1}{2}(\beta - \alpha)$, and

$$y_n - y_{bn}^* = \sum_{k=0}^n w_{n-k} a(k) (y_k - y_{bk}^*) + d_n, \quad n \geq 0.$$

By Lemma 2, we have $(y - y_b^*) \in l_2$, and since it is clear that y_b^* depends on y_a^* , but not on $[f(0, nh) - y_a^*]$, this completes the proof of Theorem 2.

3.3 Proof of Theorem 3

We need the following lemma.

Lemma 5: If $y_n = y_n^ + \eta_n$ with $y^* \in \mathcal{K}$ and $\eta_n \rightarrow 0$ as $n \rightarrow \infty$, if $g(x, kh) = g[x, (k + K + 1)h]$ for all $k \geq 0$ and all x , if there exists a positive constant c such that $|g(u_1, kh) - g(u_2, kh)| \leq c |u_1 - u_2|$ for all real u_1 and u_2 and all $k \geq 0$, and if*

$$y_n = \sum_{k=0}^n w_{n-k} g(y_k, kh) + p_n + \delta_n, \quad n \geq 0$$

with $p \in \mathcal{K}$ and $\delta_n \rightarrow 0$ as $n \rightarrow \infty$, then

$$y_n^* = \sum_{k=-\infty}^n w_{n-k} g^*(y_k^*, kh) + p_n$$

for all $n = 0, \pm 1, \pm 2, \dots$, in which $g^*(x, kh)$ is defined by the conditions:

$$g^*(x, kh) = g^*[x, (k + K + 1)h]$$

for all k and all x , and

$$g^*(x, kh) = g(x, kh)$$

for all x and $k = 0, 1, 2, \dots, K$.

Proof of Lemma 5:

For $n \geq 0$:

$$\begin{aligned} y_n^* + \eta_n &= \sum_{k=0}^n w_{n-k} g[y_k^* + \eta_k, kh] + p_n + \delta_n \\ &= \sum_{k=0}^n w_{n-k} g(y_k^*, kh) + \sum_{k=0}^n w_{n-k} [g(y_k^* + \eta_k, kh) - g(y_k^*, kh)] \\ &\quad + p_n + \delta_n \\ &= \sum_{k=-\infty}^n w_{n-k} g^*(y_k^*, kh) + \sum_{k=0}^n w_{n-k} [g(y_k^* + \eta_k, kh) - g(y_k^*, kh)] \\ &\quad - \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_k^*, kh) + p_n + \delta_n. \end{aligned}$$

Therefore,

$$\begin{aligned}
 y_n^* - \sum_{k=-\infty}^n w_{n-k} g^*(y_k^*, kh) - p_n = -\eta_n \\
 + \sum_{k=0}^n w_{n-k} [g(y_k^* + \eta_k, kh) - g(y_k^*, kh)] \\
 - \sum_{k=-\infty}^{-1} w_{n-k} g^*(y_k^*, kh) + \delta_n, \quad n \geq 0.
 \end{aligned}$$

Since $\{w_n\} \in l_1$, both sums on the right-side approach zero as $n \rightarrow \infty$. Thus, the left side also approaches zero as $n \rightarrow \infty$. But the values of the left side are periodic. Therefore,

$$y_n^* - \sum_{k=-\infty}^n w_{n-k} g^*(y_k^*, kh) - p_n = 0 \tag{13}$$

for all $n \geq 0$, and since $y^* \in \mathcal{K}$ and $p \in \mathcal{K}$, (13) holds for all n . This proves Lemma 5.

By Lemma 1,

$$y_n = \sum_{k=0}^n w_{n-k} g(y_k, kh) + \sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k, \quad n \geq 0$$

in which $g(y_k, kh)$ is defined in Lemma 1, and $\varphi_k = 0$ for $k \geq (p + 1)$. Since $\{w_n\}$ and $\{v_n\} \in l_1$, and $f(0, kh) \rightarrow 0$ as $k \rightarrow \infty$, we have

$$\sum_{k=0}^n w_{n-k} f(0, kh) + \sum_{k=0}^n v_{n-k} \varphi_k \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

By Lemma 5 and the hypotheses of Theorem 3,

$$y_n^* = \sum_{k=-\infty}^n w_{n-k} g^*(y_k^*, kh)$$

for $n = 0, \pm 1, \pm 2, \dots$, with $y^* \in \mathcal{K}$. If ρ_K were less than unity, it would follow from Lemma 4 (in particular the uniqueness property of y^* of Lemma 4) that $y_n^* = 0$ for all n , since $g^*(0, kh) = 0$ for all $k \geq 0$. Therefore, $\rho_K \geq 1$, which completes the proof of Theorem 3.

3.4 Proof of Theorem 4:

According to Lemma 1,

$$y_n - z_n = \sum_{k=0}^n w_{n-k} [g(y_k, kh) - g(z_k, kh)] + \sum_{k=0}^n v_{n-k} \psi_k, \quad n \geq 0.$$

Therefore, with

$$b_n = \sum_{k=0}^n v_{n-k} \psi_k, \quad n \geq 0$$

we have, by Lemma 2,

$$\langle y - z \rangle_N \geq (1 + \rho)^{-1} \langle b \rangle_N$$

and if $\rho < 1$,

$$\langle y - z \rangle_N \leq (1 - \rho)^{-1} \langle b \rangle_N.$$

Since²

$$\langle b \rangle_N \leq \max_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle \psi \rangle_N,$$

it remains only to prove the following lemma.†

Lemma 6: If

$$d_n = \sum_{k=0}^n v_{n-k} c_k, \quad n \geq 0$$

then

$$\langle d \rangle_N \geq \min_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle c \rangle_N.$$

Proof:

Let $\{e_k\}$ be the inverse z -transform of $V^{-1}(z)$. Clearly, $\{e_k\} \in l_1$. We have

$$\sum_{m=0}^n e_{n-m} d_m = \sum_{m=0}^n e_{n-m} \sum_{k=0}^m v_{m-k} c_k = c_n \quad \text{for } n \geq 0.$$

Thus,²

$$\langle c \rangle_N \leq \max_{0 \leq \omega \leq 2\pi} |V^{-1}(e^{i\omega})| \langle d \rangle_N$$

and, since $F(z) = V^{-1}(z)$,

$$\begin{aligned} \langle d \rangle_N &\geq \left(\max_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})| \right)^{-1} \langle c \rangle_N \\ &\geq \min_{0 \leq \omega \leq 2\pi} |F(e^{i\omega})|^{-1} \langle c \rangle_N \end{aligned}$$

which proves Lemma 6, and completes the proof of Theorem 4.

3.5 Proof of the Corollary to Theorem 4

Minkowski's inequality.

† Lemma 6 is proved in Ref. 2. The proof given here is simpler

IV. ACKNOWLEDGMENT

The writer is pleased to acknowledge the discussions held with his colleagues H. Schichman and J. F. Traub on various aspects of Numerical Analysis.

REFERENCES

1. Ralston, A., *First Course in Numerical Analysis*, McGraw-Hill Book Company, Inc., New York, 1965.
2. Sandberg, I. W., Two Theorems on the Accuracy of Numerical Solutions of Systems of Ordinary Differential Equations, *B.S.T.J.*, 46, July-August, 1967, pp. 1243-1266.
3. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill Book Company, Inc., New York, 1962.
4. Wilkinson, J. H., *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, New Jersey, 1963.
5. Sandberg, I. W., On the Theory of Physical Systems Governed by Nonlinear Functional Equations, *B.S.T.J.*, 44, May-June, 1965, p. 871.
6. Sandberg, I. W., On the Boundedness of Solutions of Nonlinear Integral Equations, *B.S.T.J.*, 44, March 1965, p. 439-453.
7. Aizerman, M. A. and Gantmacher, F. R., *Absolute Stability of Regulator Systems*, Holden-Day, San Francisco, 1964, p. 51.

A Normal Limit Theorem for Power Sums of Independent Random Variables

By N. A. MARLOW

(Manuscript received June 8, 1967)

Suppose that

$$P_n = 10 \log_{10} [10^{X_1/10} + \cdots + 10^{X_n/10}],$$

where $\{X_n\}$ is a sequence of independent random variables. The main result of this paper shows that under very general conditions on the sequence $\{X_n\}$, the power sums P_n will be asymptotically normally distributed. This result supports a commonly used normal approximation, and shows why many physical quantities obtained by power addition of random variables tend to be normally distributed in dB.

I. INTRODUCTION

In many areas of transmission engineering, logarithms of sums of powers are considered in the form

$$P_n = 10 \log_{10} [10^{X_1/10} + \cdots + 10^{X_n/10}],$$

where X_1, \dots, X_n are random variables. Specifically, if X_1, \dots, X_n are power levels in dB such that

$$X_j = 10 \log_{10} (w_j/w_0) \quad j = 1, 2, \dots, n,$$

where w_0, w_1, \dots, w_n are powers (e.g., expressed in watts), then the power level in dB of the sum $w \equiv w_1 + \cdots + w_n$ is given by the so-called "power sum,"

$$P_n = 10 \log_{10} (w/w_0) = 10 \log_{10} [10^{X_1/10} + \cdots + 10^{X_n/10}].$$

Quite often X_1, \dots, X_n are taken to be mutually independent random variables with specified distributions, and it is of interest to determine properties of their power sum P_n .

A major difficulty encountered in working with power sums is that the distribution and moments of such a sum usually cannot be ex-

pressed in simple closed form. This includes, for example, the important case when X_1, \dots, X_n are mutually independent and each has a truncated normal distribution. Even in the simpler case when X_1, \dots, X_n are mutually independent, identically distributed, and X_1 is normal, the problem is intractable. The difficulty and importance of the general problem, in turn, has led to a number of methods for approximating the distribution of a power sum.^{1, 2, 3, 4, 5, 6, 7, 8, 9}

In the present paper, the asymptotic distribution of a power sum is studied. The main result is a limit theorem which shows that under very general conditions on the components X_1, X_2, \dots , the corresponding power sums P_n will be asymptotically normal as $n \rightarrow \infty$. The particular *form* of the result is as follows: Given a sequence $\{X_n\}$ of mutually independent random variables satisfying certain conditions, there exist sequences of constants $\{c_n\}$ and $\{d_n\}$ such that

$$\lim_{n \rightarrow \infty} P\{[(P_n - c_n)/d_n] \leq x\} = [1/\sqrt{2\pi}] \int_{-\infty}^x \exp[-t^2/2] dt. \quad (1)$$

The conditions for (1) to hold are the central concern of this paper, but the implications of the results are equally important. In particular, one of the oldest and most useful approximations to the distribution of a power sum is a normal approximation. This approximation was first used at Bell Telephone Laboratories in 1934 by R. I. Wilkinson,² and is based on the fact that many observed power sum distributions are "nearly normal." This includes power sum distributions obtained by numerical convolution, and empirical distributions of physical quantities such as noise levels on trunks and connections where the resultant noise (on a dB scale) can be viewed as an approximate power sum.^{10, 11} The limit theorem proved in this paper thus provides mathematical support for a normal approximation, and substantially explains why many physical quantities obtained by power addition of random variables tend to be normally distributed in dB.

II. A NORMAL LIMIT THEOREM FOR POWER SUMS

2.1 Discussion

Before stating the main results, it is instructive to show informally why one would expect power sums to be asymptotically normal. To take a simple case, suppose that $\{X_n\}$ is a sequence of mutually independent, identically distributed random variables such that

$$\tau^2 \equiv \text{Var} [10^{X_n/10}]$$

is finite. Let $\theta = E10^{X_1/10}$ and put

$$S_n = 10^{X_1/10} + \dots + 10^{X_n/10}.$$

Then by the law of large numbers, one expects that for large n ,

$$\frac{S_n}{n\theta} \approx 1.$$

Next, note that if $x \approx 1$, then $\log_e x \approx x - 1$ so for large n

$$\log_e \frac{S_n}{n\theta} \approx \frac{S_n - n\theta}{n\theta}.$$

Multiplication by $(\theta \sqrt{n})/\tau$ then gives

$$\frac{\theta \sqrt{n}}{\tau} \log_e \frac{S_n}{n\theta} \approx \frac{S_n - n\theta}{\tau \sqrt{n}}. \tag{2}$$

But, by the central limit theorem, the right-hand side of (2) is asymptotically normal with mean 0 and variance 1. Thus, it is strongly suggested that

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left\{ \frac{\theta \sqrt{n}}{\tau} [\log_e S_n - \log_e (n\theta)] \leq x \right\} \\ = [1/\sqrt{2\pi}] \int_{-\infty}^x \exp [-t^2/2] dt. \end{aligned}$$

This, and more, is indeed true as will be shown.

2.2 The Main Result

The normal limit theorem for power sums is a consequence of the following result which will first be proved:

Lemma 1: Let $\{S_n\}$ be a sequence of positive random variables. Suppose there exist sequences of positive real numbers $\{a_n\}$ and $\{b_n\}$, and a distribution F such that

(i) *At each point of continuity of F ,*

$$\lim_{n \rightarrow \infty} P \left\{ \frac{S_n - a_n}{b_n} \leq x \right\} = F(x)$$

(ii) $\lim_{n \rightarrow \infty} (b_n/a_n) = 0.$

Then at each point of continuity of F ,

$$\lim_{n \rightarrow \infty} P \{ (a_n/b_n) \log_e (S_n/a_n) \leq x \} = F(x).$$

Proof: Let x be a continuity point of F , and let $\epsilon > 0$ be given. Because F has at most a countable number of discontinuities, there is a $\delta > 0$ such that F is continuous at $x + \delta$ and

$$F(x + \delta) - F(x) < \epsilon. \quad (3)$$

Next, define

$$U_n = (S_n - a_n)/b_n \quad \text{and} \quad V_n = (a_n/b_n) \log_e (S_n/a_n).$$

Then

$$\begin{aligned} |P\{V_n \leq x\} - F(x)| \\ \leq |P\{V_n \leq x\} - P\{U_n \leq x\}| + |P\{U_n \leq x\} - F(x)|. \end{aligned}$$

By assumption (i) therefore,

$$\overline{\lim}_{n \rightarrow \infty} |P\{V_n \leq x\} - F(x)| \leq \overline{\lim}_{n \rightarrow \infty} |P\{V_n \leq x\} - P\{U_n \leq x\}|.$$

Let

$$\Delta_n(x) = |P\{V_n \leq x\} - P\{U_n \leq x\}|.$$

To complete the proof it suffices to show that

$$\overline{\lim}_{n \rightarrow \infty} \Delta_n(x) = 0.$$

To prove this note first from the inequality $\log_e x \leq x - 1$, $x > 0$, that $V_n \leq U_n$ for all n . Thus,

$$\begin{aligned} \Delta_n(x) &= P[\{V_n \leq x\} \cap \{U_n > x\}] \\ &= P\{x < U_n \leq (a_n/b_n)[\exp(b_n x/a_n) - 1]\}. \end{aligned}$$

Using the inequality $e^y - 1 \leq ye^y$, $-\infty < y < \infty$, it follows that

$$0 \leq \Delta_n(x) \leq P\{x < U_n \leq x \exp(b_n x/a_n)\}.$$

By assumption, $(b_n/a_n) > 0$ for all n and $\lim_{n \rightarrow \infty} (b_n/a_n) = 0$. Thus, there exists a natural number N such that $n \geq N$ implies

$$x < x \exp(b_n x/a_n) \leq x + \delta.$$

So if $n \geq N$,

$$0 \leq \Delta_n(x) \leq P\{x < U_n \leq x + \delta\}.$$

Because x and $x + \delta$ are continuity points of F , it follows by assumption (i) and inequality (3) that

$$0 \leq \overline{\lim}_{n \rightarrow \infty} \Delta_n(x) \leq F(x + \delta) - F(x) < \epsilon.$$

Since $\epsilon > 0$ was arbitrary, the proof is complete.

The importance of Lemma 1 is that it gives a sufficient condition to go from limit theorems for sums of random variables to limit theorems for logarithms of sums. In the important case of power sums of independent random variables, general conditions for asymptotic normality can thus be obtained from classical central limit theory as shown in the next result.

Theorem 1: Let $\{X_n\}$ be a sequence of mutually independent random variables and suppose that

$$\tau_j^2 \equiv \text{Var} [10^{X_j/10}]$$

is finite for every j . Let $\theta_j = E10^{X_j/10}$ and put .

$$M_n = \sum_{j=1}^n \theta_j, \quad s_n^2 = \sum_{j=1}^n \tau_j^2.$$

Denote the distribution of $10^{X_j/10}$ by $H_j(x)$, and let

$$P_n = 10 \log_{10} [10^{X_1/10} + \dots + 10^{X_n/10}].$$

If the following conditions are satisfied:

(i) *The Lindeberg Condition: For every $\epsilon > 0$,*

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^2} \sum_{j=1}^n \int_{A_{jn}} (x - \theta_j)^2 dH_j(x) = 0,$$

where

$$A_{jn} = \{x: |x - \theta_j| \geq \epsilon s_n\}$$

(ii) $\lim_{n \rightarrow \infty} (s_n/M_n) = 0$

it will follow that

$$\lim_{n \rightarrow \infty} P\{(\lambda M_n/s_n)[P_n - 10 \log_{10} M_n] \leq x\} = \Phi(x) \tag{4}$$

where $\lambda = (\log_e 10)/10$ and

$$\Phi(x) = [1/\sqrt{2\pi}] \int_{-\infty}^x \exp [-t^2/2] dt.$$

Proof: Let

$$S_n = 10^{X_1/10} + \dots + 10^{X_n/10}.$$

Then condition (i) implies that

$$\lim_{n \rightarrow \infty} P \left\{ \frac{S_n - M_n}{s_n} \leq x \right\} = \Phi(x)$$

(cf. Feller,¹² p. 256). With the identifications $a_n = M_n$ and $b_n = s_n$ it follows from condition (ii) and Lemma 1 that

$$\lim_{n \rightarrow \infty} P \{ (M_n/s_n) \log_e (S_n/M_n) \leq x \} = \Phi(x).$$

The assertion of the theorem then follows by changing to logarithms with base 10.

An interesting thing to note is that if the conditions of Theorem 1 are satisfied then the sum of powers

$$S_n = 10^{X_1/10} + \dots + 10^{X_n/10}$$

and the power sum in dB, $P_n = 10 \log_{10} S_n$, will both be asymptotically normal. Thus, not only will normality be observed on a "power scale" but on a "dB scale" as well.

2.3 Identically Distributed Components

The preceding result implies the asymptotic normality of P_n when the components are identically distributed. To show this, suppose that $\{X_n\}$ is a sequence of mutually independent, identically distributed random variables with $H(x) = P\{10^{X_1/10} \leq x\}$. Let

$$\tau^2 = \text{Var} [10^{X_1/10}]$$

and $\theta = E10^{X_1/10}$. If τ^2 is finite, condition (ii) of Theorem 1 is clearly satisfied since

$$\frac{s_n}{M_n} = \frac{\tau}{\theta \sqrt{n}}.$$

Condition (i) is also satisfied because if $\epsilon > 0$,

$$\frac{1}{s_n^2} \sum_{i=1}^n \int_{A_{i,n}} (x - \theta_i)^2 dH_i(x) = \frac{1}{\tau^2} \int_{A_n} (x - \theta)^2 dH(x) \rightarrow 0 \text{ as } n \rightarrow \infty,$$

where $A_n = \{x : |x - \theta| \geq \epsilon \tau \sqrt{n}\}$. It thus follows that

$$\lim_{n \rightarrow \infty} P \left\{ \lambda \frac{\theta \sqrt{n}}{\tau} [P_n - 10 \log_{10} (n\theta)] \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp[-t^2/2] dt,$$

hence, P_n is asymptotically normal with mean $10 \log_{10} (n\theta)$ and variance $\tau^2/(n\lambda^2\theta^2)$.

2.4 *Bounded Components*

Suppose next that $\{X_n\}$ is a sequence of mutually independent random variables and that the following conditions are satisfied:

(i) There exist constants b and B such that

$$0 < b \leq 10^{x_i/10} \leq B \text{ for all } j$$

(ii) $s_n^2 \rightarrow \infty$ as $n \rightarrow \infty$.

The conditions of Theorem 1 are easily shown to be satisfied in this case, and it follows that P_n will be asymptotically normal. Note that condition (i) will be satisfied whenever $10^{x_i/10}$ represents power from a physical source. Condition (ii), on the other hand, will be satisfied if $\tau_j^2 \geq c > 0$ for some fixed c and an infinite number of indices j .

III. THE NORMAL LIMIT THEOREM AND WILKINSON'S NORMAL APPROXIMATION

One of the most useful approximations to the distribution and moments of a power sum is based on a normal approximation as mentioned in the introduction. The method consists of approximating the distribution of P_n by a normal distribution so that

$$P\{P_n \leq x\} \approx P\{\alpha\xi + \beta \leq x\},$$

where ξ is normal with mean 0 and variance 1. Writing as before,

$$M_n = E10^{P_n/10} \quad \text{and} \quad s_n^2 = \text{Var} [10^{P_n/10}],$$

the parameters α and β are chosen so that

$$M_n = E[10^{(\alpha\xi + \beta)/10}]$$

and

$$s_n^2 = \text{Var} [10^{(\alpha\xi + \beta)/10}]$$

which is equivalent to equating means and variances on a "power scale." If ξ is normal with mean 0 and variance 1 then

$$E[10^{(\alpha\xi + \beta)/10}] = e^{\lambda\beta} e^{\frac{1}{2}(\lambda\alpha)^2}$$

and

$$\text{Var} [10^{(\alpha\xi + \beta)/10}] = e^{2\lambda\beta} [e^{2\lambda^2\alpha^2} - e^{\lambda^2\alpha^2}],$$

where

$$\lambda = (\log_e 10)/10.$$

Solving the above equations for α and β , the approximation then asserts that P_n is normal with

$$E(P_n) = \beta = 10 \log_{10} M_n - 5 \log_{10} [1 + (s_n/M_n)^2] \quad (5)$$

and

$$\text{Var}(P_n) = \alpha^2 = \frac{10}{\lambda} \log_{10} [1 + (s_n/M_n)^2]. \quad (6)$$

In light of the normal limit theorem, it is quite natural to assume that P_n is approximately normal, provided the conditions of the theorem are satisfied, and n is large. On the other hand, the estimates given by (5) and (6) are different from those based on (4):

$$E(P_n) \doteq 10 \log_{10} M_n \quad (7)$$

$$\text{Var}(P_n) \doteq s_n^2/(\lambda M_n)^2. \quad (8)$$

The difference, however, is easily resolved once it is realized that if condition (ii) of Theorem 1 is satisfied then (5) and (6) are asymptotically equivalent to (7) and (8). In fact, it is a simple matter to show (cf. Feller,¹² p. 246) that if the conditions of Theorem 1 are satisfied then

$$\lim_{n \rightarrow \infty} P\{[(P_n - u_n)/\sqrt{v_n}] \leq x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp[-t^2/2] dt, \quad (9)$$

where

$$u_n = 10 \log_{10} M_n - 5 \log_{10} [1 + (s_n/M_n)^2]$$

and

$$v_n = \frac{10}{\lambda} \log_{10} [1 + (s_n/M_n)^2].$$

In numerical applications, the normal approximation based on (9) is to be favored over that based on (4). In the first place, when X_1, \dots, X_n are mutually independent, identically distributed, and X_1 has a truncated normal distribution, Monte Carlo studies by I. Nâsell⁹ have shown that the mean and variance estimates given by (5) and (6) are better than those given by (7) and (8) (although for large n and small variance of X_1 there is hardly any difference). Secondly, the normalizing factors in (9) were obtained quite naturally by equating moments on a power scale. This is analogous to the situation in classical central limit theory when the sequence $(S_n - M_n)/s_n$ converges in dis-

tribution to the standard normal. The normalizing factors M_n and s_n are not the only ones that give this result, but they are chosen in a natural way to insure that for every n , the mean and variance of $(S_n - M_n)/s_n$ agrees with its asymptotic distribution.

IV. ACKNOWLEDGMENTS

I would like to thank D. A. Lewinski and I. Näsell for their helpful comments and suggestions.

REFERENCES

1. Dixon, J. T., unpublished work, 1932.
2. Wilkinson, R. I., unpublished work, 1934.
3. Holbrook, B. D. and Dixon, J. T., Load Rating Theory for Multichannel Amplifiers, B.S.T.J., 18, October, 1939, p. 624.
4. Curtis, H. E., Probability Distribution of Noise Due to Fading on Multisection FM Microwave Systems, IRE Trans. Commun. Syst., September, 1959, p. 161.
5. Fenton, L. F., The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems, IRE Trans. Commun. Syst., March 1960, p. 57.
6. Roberts, J. H., Sums of Probability Distributions Expressed in Decibel Steps, Proc. IEE, 110, No. 4, April, 1963, p. 692.
7. Cyr, M. H. and Thuswaldner, A., Multichannel Load Calculation Using the Monte Carlo Method, IEEE Trans. Commun. Tech., COM-14, No. 2, April, 1966, p. 177.
8. Derzai, M., Power Addition of Independent Random Variables Normally Distributed on a dB Scale, IEEE Int. Conv. Record, Part I, March, 1967, p. 40.
9. Näsell, I., Some Properties of Power Sums of Truncated Normal Random Variables, B.S.T.J., this issue, p. 2091.
10. Näsell, I., The 1962 Survey of Noise and Loss on Toll Connections, B.S.T.J., 43, March, 1964, p. 697-718.
11. Lewinski, D. A., A New Objective for Message Circuit Noise, B.S.T.J., 43, March, 1964, p. 719-740.
12. Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. II, John Wiley and Sons, Incorporated, New York, 1966.

Some Properties of Power Sums of Truncated Normal Random Variables

By INGEMAR NÅSELL

(Manuscript received June 15, 1967)

The power sum of P_n n components X_1, X_2, \dots, X_n is defined by the relation

$$P_n = 10 \log_{10} [10^{X_1/10} + \dots + 10^{X_n/10}].$$

The distributions of such power sums are studied both analytically and by Monte Carlo simulation techniques for the case where the components are independent, identically distributed, truncated normal random variables. Results are given in terms of distributions and moments of P_n . The number of components varies from 2 to 256, and the standard deviation of the component variables before truncation ranges from 1 to 10 dB. The dependence of the results on the choice of truncation point is also investigated.

I. INTRODUCTION

It is common practice in communications engineering to express signal and noise powers on a logarithmic scale. As is well known, such a scale serves both to narrow the numerical range between large and small powers and to simplify some computations by replacing multiplication by addition. The decibel scale is most commonly used. Employing this scale, the power level x of a power w is defined by

$$x = 10 \log_{10} \frac{w}{w_0}, \quad (1)$$

where w_0 is a reference power, and x is expressed in decibels (dB) over the reference power w_0 . Note from (1) that $w/w_0 = 10^{x/10}$.

In the situation where a number of uncorrelated signal sources feed into the same load, the power level p_n of a sum of powers w_1, \dots, w_n is given by

$$p_n = 10 \log_{10} [10^{x_1/10} + \dots + 10^{x_n/10}], \quad (2)$$

where x_i is the power level of w_i . Examples of such sums arise in cross-talk computations, overload theory for multichannel amplifiers, noise calculations on carrier systems and multihop radio systems, and in the evaluations of noise distributions on built-up connections between telephone subscribers. Here, however, the power levels are in many situations random rather than deterministic variables. Thus, in analogy with (2), one is faced with the random variable

$$P_n = 10 \log_{10} [10^{X_1/10} + \dots + 10^{X_n/10}], \quad (3)$$

where each X_i is a random variable with known distribution. The classical power sum problem consists of finding the distribution function and the moments of the power sum P_n defined in (3). This problem does not, however, possess a simple closed-form mathematical solution. As a result, the task of finding approximate solutions has received extensive attention, beginning at least 35 years ago and persisting till this date.

Among earlier contributions to the problem, we can distinguish those that give specific methods for numerical evaluation of the power sum distribution without introducing any other approximations than those that are directly related to the numerical technique that is being used.^{1, 3, 4, 5, 6} Another approach is based on approximating the power sum with a normally distributed random variable.^{2, 7} This approach, due to R. I. Wilkinson,² is quite appealing, since it leads to simple evaluation formulas. Moreover, it has now been put on a firm mathematical foundation with the development of a limit theorem by N. A. Marlow. In a companion paper,⁸ he proves that power sums are asymptotically normally distributed, provided some mild conditions on the component variables are satisfied.

The present paper considers power sums of independent, identically distributed, truncated normal random variables, since this is a situation of considerable practical importance in transmission engineering work. Two approaches are being used. In the first one, asymptotic expressions are developed for the mean and variance of P_n . The second approach is based on Monte Carlo simulation.⁹ This method has a number of distinct advantages over other numerical methods in that

- (i) it can accept any number of component variables with arbitrarily specified distribution functions,
- (ii) independence among the component variables is not required,
- (iii) computation errors do not cumulate as more than two variables are added, and

(iv) accuracy can be determined through the evaluation of confidence limits.

Our main results are numerical estimates of moments of P_n and selected graphs of its distribution function. A wide range of component distributions is covered with n ranging from 2 to 256. Most of the results are based on a nominal symmetric truncation of the component variables at ± 3.5 standard deviations from the mean. In addition, the effect on P_n of choosing other truncation points is discussed, and some general trends are developed.

II. ANALYTICAL RESULTS

Consider first the case where the X_i are independent, identically distributed random variables. Assume that the expectation

$$\theta = E[10^{X_i/10}]$$

and the central moments

$$\tau_j = E[10^{X_i/10} - \theta]^j,$$

exist and are finite for a sufficiently large range of j . We require $-1 \leq j \leq 8$ to derive the results for the mean of P_n , $-2 \leq j \leq 12$ for the variance and wider ranges for higher-order moments.

Rewrite the power sum P_n of X_1, X_2, \dots, X_n , as

$$P_n = 10 \log_{10} S_n, \quad (4)$$

where

$$S_n = 10^{X_1/10} + \dots + 10^{X_n/10}.$$

Now expand (4) in a finite Taylor series about the mean, $n\theta$, of S_n . This gives

$$P_n = \frac{1}{\lambda} \left[\log(n\theta) + \frac{S_n - n\theta}{n\theta} - \frac{1}{2} \left(\frac{S_n - n\theta}{n\theta} \right)^2 + \dots + \frac{(-1)^{m+1}}{m} \left(\frac{S_n - n\theta}{n\theta} \right)^m + R_m \left(\frac{S_n - n\theta}{n\theta} \right) \right], \quad (5)$$

where

$$\lambda = \frac{1}{10 \log_{10} e} \approx 0.23026$$

and \log stands for \log_e . The remainder term in (5) can be expressed in integral form as

$$R_m(x) = (-1)^m x^{m+1} \int_0^1 \frac{t^m dt}{1+xt}, \quad x > -1, \quad (6)$$

or, alternatively, as

$$R_m(x) = \frac{(-1)^m}{m+1} \left(\frac{x}{1+\delta x} \right)^{m+1}, \quad x > -1, \quad (7)$$

where $0 < \delta < 1$.

With $R_m(x)$ given by (7), one obtains

$$R_m\left(\frac{S_n - n\theta}{n\theta}\right) \leq 0 \quad \text{for } m \text{ odd,}$$

so that from (5) we get our first result

$$E(P_n) \leq LAP_n, \quad (8)$$

where

$$LAP_n = 10 \log_{10}(n\theta) = \frac{1}{\lambda} \log(n\theta) \quad (9)$$

is the level of average power.

To derive asymptotic expressions for the moments of P_n , we apply the Lemma in Appendix A and (6) to get

$$E\left[\left(\frac{S_n - n\theta}{n\theta}\right)^\alpha \left(R_m\left(\frac{S_n - n\theta}{n\theta}\right)\right)^\beta\right] = O(n^{-\frac{1}{2}(\alpha+\beta(m+1))}) \quad (10)$$

Next, to derive an asymptotic expression for $E(P_n)$, we take the expected value of both sides of (5) with $m = 3$. An application of (10) then gives

$$E(P_n) = LAP_n - \frac{\tau_2}{2\lambda\theta^2} \frac{1}{n} + O(1/n^2) \quad \text{as } n \rightarrow \infty. \quad (11)$$

Here the independence of the component variables has been used to express the variance of S_n as $n\tau_2$, and the third central moment of S_n as $n\tau_3$. The term containing τ_3 is of order $1/n^2$.

To arrive at an asymptotic expression for the variance $\sigma^2(P_n)$, we use (5) with $m = 2$ and (11) to get

$$\begin{aligned} P_n - E(P_n) &= \frac{1}{\lambda} \frac{S_n - n\theta}{n\theta} - \frac{1}{2\lambda} \left(\frac{S_n - n\theta}{n\theta} \right)^2 \\ &\quad + \frac{1}{\lambda} R_2\left(\frac{S_n - n\theta}{n\theta}\right) + O(1/n) \quad \text{as } n \rightarrow \infty. \end{aligned} \quad (12)$$

Squaring (12), taking the expected value of both sides, and applying (10) to four of the resulting terms gives

$$\sigma^2(P_n) = \frac{\tau_2}{\lambda^2 \theta^2} \frac{1}{n} + O(1/n^2) \quad \text{as } n \rightarrow \infty. \quad (13)$$

A similar approach can be used to derive asymptotic expressions for higher-order moments. The measures of skewness and excess, denoted by $\gamma_1(P_n)$ and $\gamma_2(P_n)$, respectively, are defined by

$$\gamma_1(P_n) = \frac{E(P_n - EP_n)^3}{\sigma^3(P_n)}$$

and

$$\gamma_2(P_n) = \frac{E(P_n - EP_n)^4}{\sigma^4(P_n)} - 3.$$

They are found to satisfy the expressions

$$\gamma_1(P_n) = \left[\frac{\tau_3}{\tau_2^3} - \frac{3\tau_2^3}{\theta} \right] \frac{1}{n^3} + O(1/n^3) \quad \text{as } n \rightarrow \infty \quad (14)$$

and

$$\gamma_2(P_n) = \left[\frac{\tau_4}{\tau_2^4} - \frac{12\tau_3}{\theta\tau_2} + \frac{20\tau_2}{\theta^2} - 3 \right] \frac{1}{n} + O(1/n^2) \quad \text{as } n \rightarrow \infty. \quad (15)$$

The asymptotic results given in (11), (13), (14), and (15) are all consistent with Marlow's normal limit theorem.⁸ The main virtue of the asymptotic results above is that they indicate the rate of convergence of the four quantities considered. This is of practical interest since engineering applications often involve a finite and fairly small number of component variables.

In the particular case where the X_i are truncated normal with mean 0 dB, standard deviation before truncation of σ dB, and symmetric truncation at $\pm c\sigma$ dB, the results contained in Appendix B can be used to express (11), (13), (14), and (15) in terms of σ , c , and n . For the mean and the variance we get, respectively,

$$\mu(P_n) = LAP_n - \frac{\exp(\lambda^2 \sigma^2) U_c(\sigma) - 1}{2\lambda n} + O(1/n^2) \quad (16)$$

and

$$\sigma^2(P_n) = \frac{\exp(\lambda^2 \sigma^2) U_c(\sigma) - 1}{\lambda^2 n} + O(1/n^2), \quad (17)$$

where the truncation factor $U_c(\sigma)$ is defined by

$$U_c(\sigma) = \frac{T_c(2\sigma)}{T_c^2(\sigma)},$$

with

$$T_c(\sigma) = \frac{\Phi(c - \lambda\sigma) - \Phi(-c - \lambda\sigma)}{\Phi(c) - \Phi(-c)}$$

and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt.$$

Derivation of the Wilkinson estimates for the mean and variance of the power sum P_n is given in Appendix B. This derivation uses the same ideas employed by R. I. Wilkinson in 1934.² Thus, P_n is approximated by a normally distributed random variable P_{nw} . As above, the components are independent, identically distributed truncated normal with mean 0 dB, standard deviation before truncation of σ dB, and truncation at $\pm c\sigma$ dB. From Appendix B we then have

$$\mu(P_{nw}) = LAP_n - 5 \log_{10} \left[1 + \frac{\exp(\lambda^2 \sigma^2) U_c(\sigma) - 1}{n} \right] \quad (18)$$

$$\sigma^2(P_{nw}) = \frac{10}{\lambda} \log_{10} \left[1 + \frac{\exp(\lambda^2 \sigma^2) U_c(\sigma) - 1}{n} \right]. \quad (19)$$

The first terms in the asymptotic expansion of (18) and (19), respectively, agree exactly with the results in (16) and (17). This agreement establishes the important result that expressions (18) and (19) are asymptotically correct to the order of n included in (16) and (17). Finally, we note that the actual result due to Wilkinson is contained in (18) and (19); the case with nontruncated component variables is obtained by putting the truncation factor $U_c(\sigma) = 1$.

III. MONTE CARLO RESULTS FOR $C = 3.5$

Having established analytical estimates for the mean and variance of power sums of truncated normal random variables, let us now turn to estimation using the Monte Carlo technique. The power sum problem is basically solved by estimating the distribution function of P_n . Using the Monte Carlo method, one obtains an estimate of this function by random sampling. Each sample of the power sum is obtained by selecting

n independent samples, one from each of the component distributions on the dB scale. The corresponding sample value of the power sum is then directly computed from (2). For the results presented here, the component samples have been selected via computer generation of so-called pseudo-random numbers. These have approximately a uniform distribution over the unit interval. Using the inverse error function together with nominal truncation at $\pm 3.5\sigma$ gave a random variable with truncated normal distribution. Because of requirements of computing speed, this transformation has been achieved via a table look-up scheme with values of the transformation stored in the computer memory.

Table I summarizes Monte Carlo results in terms of estimates of the mean $\mu(P_n)$, the standard deviation $\sigma(P_n)$, and the measures of skewness and excess $\gamma_1(P_n)$ and $\gamma_2(P_n)$. Monte Carlo estimates of these quantities are denoted by the corresponding latin letters $m(P_n)$, $s(P_n)$, $g_1(P_n)$, and $g_2(P_n)$. The standard deviation and the measures of skewness and excess are estimated directly by the corresponding characteristics of the sample distribution. The mean is estimated through the formula

$$m(P_n) = LAP_n - (LAP_{MC} - m_{MC}). \quad (20)$$

The value of LAP_n is computed exactly from relation (9), while LAP_{MC} and m_{MC} are the LAP and the mean, respectively, of the sample distribution. The mean $\mu(P_n)$ could also be estimated by m_{MC} . However, $m(P_n)$ from (20) is preferred over m_{MC} because the Monte Carlo results show that it has a smaller sampling variance.

An indication of the accuracy of the results in Table I is given by the number of decimals included. The half-width of the 99 percent confidence interval that represents the sampling uncertainty is between one and five times the unit in the least significant digit. For the mean, the confidence interval width has, however, been computed for m_{MC} instead of for $m(P_n)$. The computation of these confidence intervals has been based on the asymptotic normality of the corresponding statistics.

Table I shows that the mean of the power sum increases by somewhat more than 3 dB when the number of component variables is doubled for a fixed σ . This effect is illustrated in Fig. 1, where the mean is plotted as a function of the number of components n . This figure shows that the increase in the mean is substantially more than 3 dB for a doubling of the number of components n in case n is small and σ is large. On the other hand, Fig. 1 indicates that the slope of the

TABLE I—MONTE CARLO ESTIMATES $m(P_n)$, $s(P_n)$, $g_1(P_n)$, $g_2(P_n)$ OF MEAN, STANDARD DEVIATION, MEASURE OF SKEWNESS, AND MEASURE OF EXCESS OF P_n . THE COMPONENTS ARE TRUNCATED NORMAL WITH MEAN $\mu = 0$, STANDARD DEVIATION BEFORE TRUNCATION σ , TRUNCATION AT $\pm 3.5\sigma$.

	$\sigma:$	1	2	3	4	5	6	7	8	9	10
$n = 2$	m	3.07	3.23	3.5	3.8	4.1	4.5	5.1	5.5	6.1	6.6
	s	0.70	1.43	2.20	3.0	3.8	4.6	5.4	6.2	7.1	7.9
	g_1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1
$n = 4$	g_2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	m	6.11	6.36	6.75	7.3	7.9	8.6	9.4	10.2	11.1	12.1
	s	0.50	1.03	1.60	2.25	2.9	3.5	4.2	4.9	5.6	6.3
$n = 8$	g_1	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.2	0.2
	g_2	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0
	m	9.13	9.43	9.90	10.54	11.3	12.2	13.2	14.3	15.4	16.7
$n = 16$	s	0.36	0.73	1.15	1.64	2.14	2.7	3.3	3.9	5.4	5.0
	g_1	0.0	0.0	0.0	0.1	0.2	0.2	0.2	0.3	0.3	0.3
	g_2	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.2	0.1
$n = 32$	m	12.15	12.47	12.99	13.70	14.59	15.6	16.8	18.0	19.4	20.9
	s	0.252	0.52	0.82	1.19	1.56	2.01	2.51	3.1	3.6	4.0
	g_1	0.0	0.0	0.0	0.1	0.2	0.2	0.3	0.3	0.4	0.4
$n = 64$	g_2	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.2
	m	15.162	15.49	16.04	16.79	17.74	18.86	20.12	21.5	23.0	24.7
	s	0.178	0.37	0.59	0.84	1.14	1.48	1.91	2.36	2.8	3.3
$n = 128$	g_1	0.0	0.0	0.0	0.1	0.2	0.2	0.3	0.3	0.4	0.4
	g_2	0.0	0.0	0.0	0.1	0.2	0.1	0.0	0.0	0.1	0.2
	m	18.174	18.51	19.07	19.84	20.82	21.99	23.34	24.85	26.5	28.2
$n = 256$	s	0.126	0.260	0.42	0.61	0.82	1.10	1.42	1.76	2.18	2.6
	g_1	0.0	0.0	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.4
	g_2	0.0	0.0	0.1	0.1	0.2	0.0	0.0	-0.1	0.0	0.0
$n = 512$	m	21.185	21.525	22.09	22.87	23.87	25.07	26.46	28.02	29.73	31.6
	s	0.089	0.184	0.29	0.43	0.59	0.79	1.04	1.30	1.63	1.98
	g_1	0.0	0.0	0.0	0.1	0.1	0.1	0.2	0.2	0.3	0.3
$n = 1024$	g_2	0.0	0.0	0.0	0.0	0.1	0.0	-0.1	-0.2	-0.1	-0.1
	m	24.196	24.537	25.10	25.89	26.90	28.12	29.53	31.13	32.89	34.81
	s	0.063	0.132	0.208	0.30	0.42	0.55	0.74	0.95	1.20	1.48
$n = 2048$	g_1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2
	g_2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	m	27.200	27.537	28.07	28.86	29.87	31.18	32.69	34.41	36.24	38.18

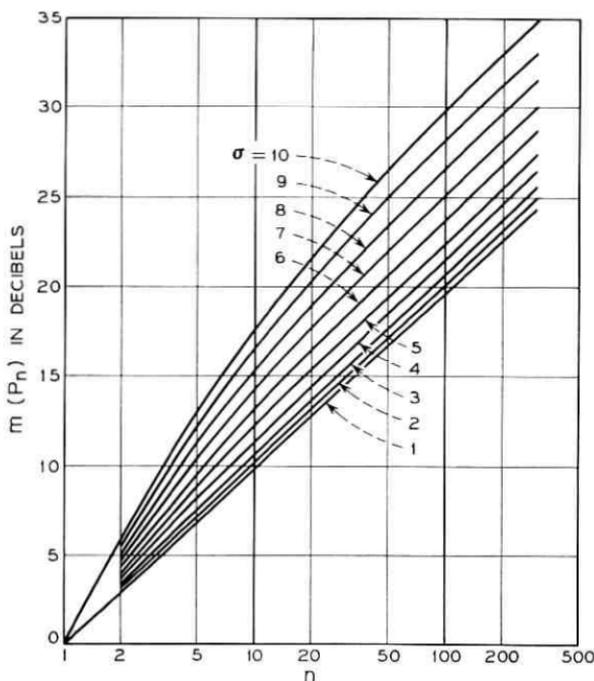


Fig. 1—Monte Carlo estimates of $\mu(P_n)$. The components are truncated normal; $\mu = 0$, truncation at $\pm 3.5\sigma$.

graph of the mean levels off at approximately 3 dB for each doubling of the number of components at all values of σ for n large enough.

It is illuminating to compare these properties of the mean with the properties of LAP_n . According to relation (9), LAP_n increases by $10 \log_{10} 2 \approx 3$ dB for each doubling of the number of components n , similar to the increase of the mean noted above. Furthermore, relations (8) and (11) imply that $LAP_n - \mu(P_n)$ is nonnegative and approaches 0 as n increases toward infinity. The rate of decrease of $LAP_n - \mu(P_n)$ is illustrated by the Monte Carlo results plotted in Fig. 2.

Table I also shows that the standard deviation of the power sum decreases as the number of component variables is increased for fixed σ . This is illustrated in Fig. 3, where Monte Carlo estimates of $\sigma(P_n)$ are plotted as a function of the number of components n .

The measures of skewness and excess in Table I can be taken as an indication of the deviation from normality of the distribution of the power sum. These measures are zero for the normal distribution and they have low values for distributions that deviate only slightly from

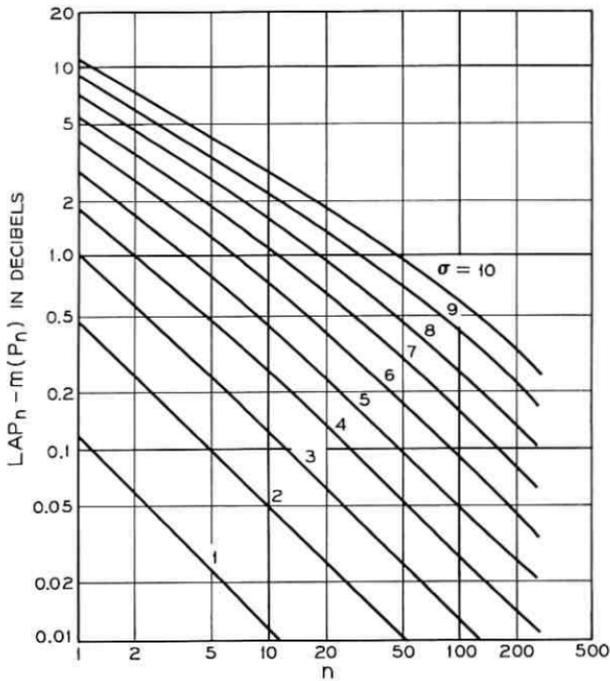


Fig. 2—Monte Carlo estimates of $LAP_n - \mu(P_n)$. The components are truncated normal; $\mu = 0$, truncation at $\pm 3.5\sigma$.

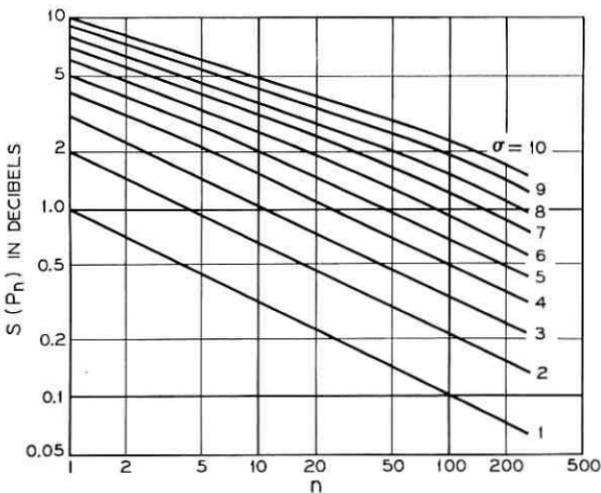


Fig. 3—Monte Carlo estimates of $\sigma(P_n)$. The components are truncated normal; $\mu = 0$, truncation at $\pm 3.5\sigma$.

normality. The table shows that g_1 and g_2 are very small for σ -values up to four over the range of n -values considered. The table also shows that g_1 is, in general, positive. This indicates that the power sum distribution is positively skewed. Moreover, g_1 considered as a function of the number of components n has definite maxima around $n = 32$ for all sufficiently large values of σ . In particular, this means that the magnitude of g_1 decreases as n becomes large enough. This behavior is consistent with the asymptotic behavior of the measure of skewness as expressed by relation (14).

The results of the previous section show that both $LAP_n - \mu(P_n)$ and $\sigma(P_n)$ converge to 0 as n becomes infinite. From these two facts it follows that the distribution of $P_n - LAP_n$ converges to a distribution degenerate at 0. Fig. 4 illustrates this convergence by plots of the Monte Carlo estimates of the distribution function of P_n for $n = 1, 4, 16, 64,$ and 256 . This convergence is also illustrated in Fig. 5 where the 1 percent and 99 percent points of the distribution function of P_n are plotted in addition to the mean $m(P_n)$ and the level of average power LAP_n , for $\sigma = 10$. It is seen that the slope of the 1 percent point with a doubling of the number of components can be considerably larger than 3 dB, while the 99 percent point changes by somewhat less than 3 dB whenever the number of components is doubled. LAP_n does not represent a fixed percentage point on the distribution function as n is changing. It is, therefore, seen that the plots in Fig. 5 of some

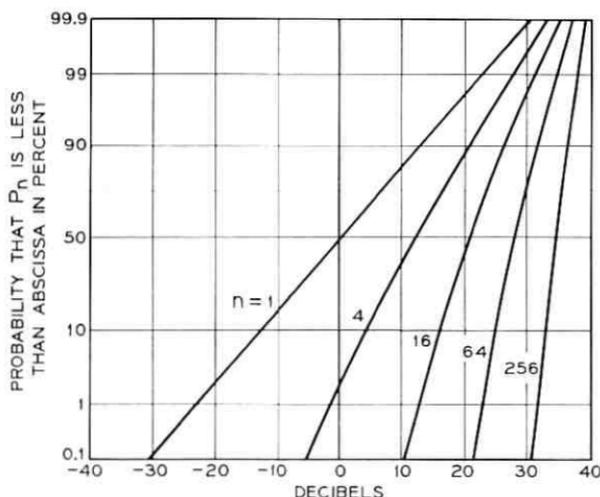


Fig. 4— Monte Carlo estimates of distribution function of P_n . The components are truncated normal; $\mu = 0, \sigma = 10$, truncation at $\pm 3.5\sigma$.

percentage points would actually cross their asymptote LAP_n from below before approaching it asymptotically from above. This is true for all percentage points of the component distribution that lie between 0 and LAP_1 . In other words, the fact that all percentage points approach LAP_n asymptotically does not imply that the approach is monotone.

IV. COMPARISON BETWEEN ANALYTICAL AND MONTE CARLO RESULTS

At this point it is natural to examine the relative agreement between the various analytical approximations and the Monte Carlo estimates. Figs. 6 and 7 contain plots of the asymptote (16), the Wilkinson approximation (18), and the Monte Carlo estimates of $LAP_n - \mu(P_n)$ for $\sigma = 6$ and 10, respectively. Both figures show the asymptote as an upper bound for $LAP_n - \mu(P_n)$. The plots also indicate that the Wilkinson expression gives a better agreement with the Monte Carlo results than the asymptote, and they illustrate the degree of agreement between the Monte Carlo results and the analytical expressions for various values of n . Finally, a comparison between the two figures shows that the analytical approximations are better for low values of σ than for high values. Figs. 8 and 9 present similar comparisons between Monte Carlo results and analytical approximations for $\sigma(N_n)$. The figures contain plots of the asymptotic expression (17), the Wilkinson expression (19), and the Monte Carlo estimate

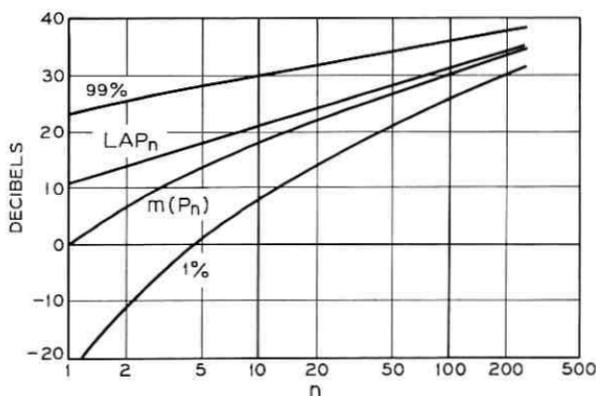


Fig. 5— LAP_n and Monte Carlo estimates of $\mu(P_n)$ and of two points on the distribution function of P_n . The components are truncated normal; $\mu = 0$, $\sigma = 10$, truncation at $\pm 3.5\sigma$.

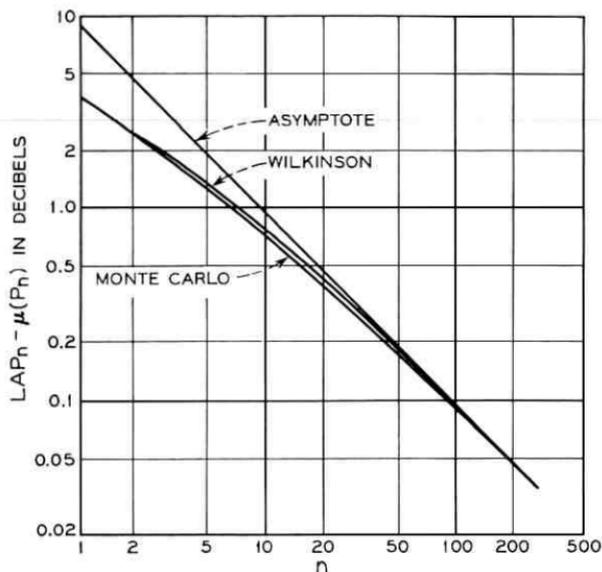


Fig. 6—Comparison between three estimates for $L_{AP_n-\mu}(P_n)$. The components are truncated normal; $\mu = 0$, $\sigma = 6$, truncation at $\pm 3.5\sigma$.

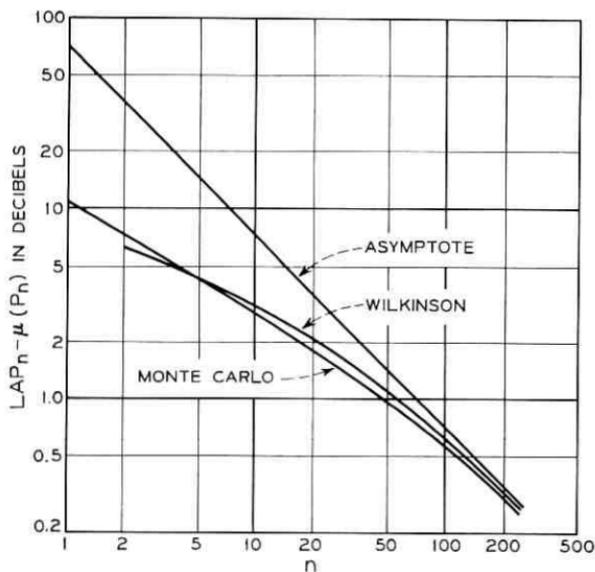


Fig. 7—Comparison between three estimates for $L_{AP_n-\mu}(P_n)$. The components are truncated normal; $\mu = 0$, $\sigma = 10$, truncation at $\pm 3.5\sigma$.

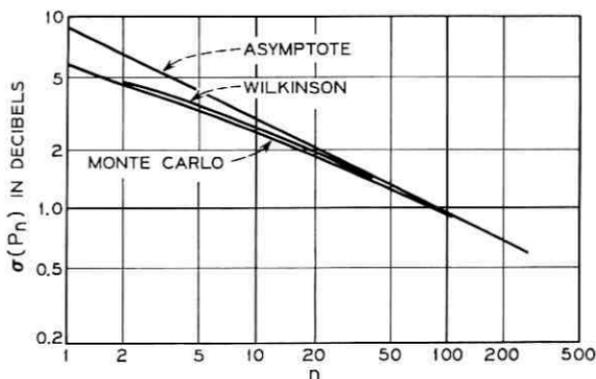


Fig. 8—Comparison between three estimates for $\sigma(P_n)$. The components are truncated normal; $\mu = 0$, $\sigma = 6$, truncation at $\pm 3.5\sigma$.

of $\sigma(P_n)$ for $\sigma = 6$ and 10, respectively. The figures serve as a basis for conjecturing that the asymptote provides an upper bound for $\sigma(P_n)$. Furthermore, the figures indicate as above the degree of agreement between the analytic approximations and the Monte Carlo results, and they show that the analytical approximations are better for low than for high values of σ .

V. INFLUENCE OF TAILS

The results discussed thus far are all based on a truncation of the component distributions at $\pm 3.5\sigma$. Truncations at other points can easily be studied with the tools used. Thus, Table II summarizes results

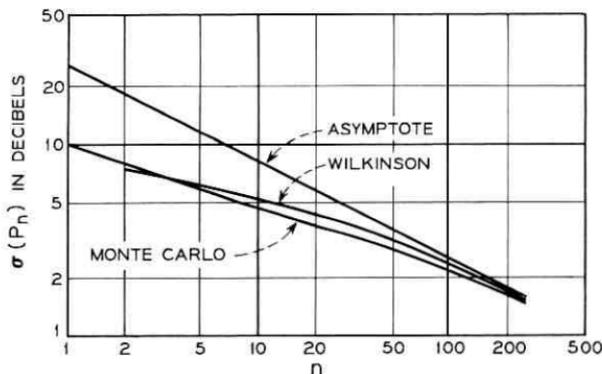


Fig. 9—Comparison between three estimates for $\sigma(P_n)$. The components are truncated normal; $\mu = 0$, $\sigma = 10$, truncation at $\pm 3.5\sigma$.

MEAN $\mu = 0$, STANDARD DEVIATION BEFORE TRUNCATION σ , TRUNCATION AT $\pm c\sigma$.

	$c:$ $\sigma:$	2.0 1	2.5 1	3.0 1	2.0 6	2.5 6	3.0 6	2.0 10	2.5 10	3.0 10
$n = 2$	m	3.06	3.06	3.07	4.3	4.4	4.6	5.9	6.3	6.4
	s	0.62	0.68	0.70	4.0	4.4	4.5	6.9	7.6	7.8
	g_1	-0.1	0.0	-0.1	-0.1	0.0	0.0	-0.2	-0.1	0.1
	g_2	-0.4	-0.3	-0.2	-0.5	-0.3	-0.1	-0.5	-0.3	-0.1
$n = 4$	m	6.09	6.10	6.10	8.1	8.4	8.6	10.9	11.5	11.8
	s	0.44	0.48	0.50	2.9	3.3	3.4	5.2	5.9	6.2
	g_1	-0.1	0.0	0.0	-0.3	-0.1	0.1	-0.3	-0.1	0.1
	g_2	-0.2	-0.1	0.0	-0.3	-0.3	-0.1	-0.4	-0.3	-0.1
$n = 8$	m	9.11	9.12	9.13	11.6	12.0	12.2	15.1	16.0	16.5
	s	0.31	0.34	0.35	2.06	2.39	2.60	3.7	4.5	4.9
	g_1	0.0	0.0	0.0	-0.3	-0.2	0.1	-0.3	-0.1	0.2
	g_2	-0.1	0.0	0.0	-0.1	-0.2	-0.1	-0.3	-0.4	-0.2
$n = 16$	m	12.12	12.14	12.15	14.80	15.30	15.53	18.9	20.0	20.6
	s	0.222	0.241	0.248	1.44	1.70	1.92	2.61	3.3	3.8
	g_1	0.0	0.0	0.0	-0.3	-0.1	0.1	-0.5	-0.1	0.2
	g_2	-0.1	0.0	0.0	0.1	-0.3	-0.2	0.0	-0.4	-0.3
$n = 32$	m	15.137	15.153	15.160	17.93	18.47	18.75	22.22	23.6	24.4
	s	0.157	0.170	0.174	1.00	1.21	1.40	1.81	2.39	2.96
	g_1	0.0	0.0	0.0	-0.3	-0.1	0.1	-0.4	-0.2	0.1
	g_2	-0.1	0.1	-0.1	0.2	-0.2	-0.2	0.1	-0.2	-0.4
$n = 64$	m	18.149	18.165	18.172	20.99	21.56	21.87	25.41	26.93	27.8
	s	0.110	0.121	0.125	0.70	0.86	1.00	1.25	1.70	2.20
	g_1	0.0	0.0	0.0	-0.2	-0.1	0.0	-0.3	-0.3	0.0
	g_2	0.0	0.0	-0.1	0.1	-0.1	-0.2	0.1	-0.1	-0.4
$n = 128$	m	21.160	21.176	21.183	24.03	24.61	24.93	28.51	30.10	31.10
	s	0.078	0.086	0.088	0.50	0.61	0.71	0.87	1.19	1.58
	g_1	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.2	-0.2	0.0
	g_2	0.0	0.0	0.0	-0.1	-0.2	-0.1	0.1	-0.1	-0.2
$n = 256$	m	24.170	24.186	24.194	27.06	27.65	27.97	31.56	33.19	34.25
	s	0.055	0.060	0.062	0.35	0.43	0.50	0.61	0.84	1.13
	g_1	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.2	-0.2	-0.1
	g_2	0.0	0.0	0.0	0.0	-0.1	0.0	0.1	0.0	-0.1

of Monte Carlo evaluations for symmetric truncations at $\pm 2\sigma$, $\pm 2.5\sigma$, and $\pm 3\sigma$ for $\sigma = 1, 6$, and 10 dB, respectively, and with the same range of n -values as considered previously. A study of the table reveals that the truncation point can have a considerable influence on the distribution of the resulting power sum. To exemplify this, Fig. 10 shows plots of the standard deviation of the power sum of 256 components as a function of the truncation point c . The plots cover a wider range of c -values and σ -values than found in Tables I and II. The extensions are based on the Wilkinson approximation.

The plots in Fig. 10 exhibit the important trend that the influence of the truncation point increases with an increase of the component standard deviation σ . The same conclusion can be drawn from a study of the c -dependence of the mean $\mu(P_n)$ or of the quantity $LAP_n - \mu(P_n)$.

Table II contains several cases of negative skewness of P_n . Hence, the earlier observation that P_n is in general positively skewed does not apply for c -values below 3.5.

VI. CONCLUDING REMARKS

The extension of the results given here to an even larger number of components ($n > 256$) is straightforward, but the computer time needed can easily become excessive. The agreement between asymptotic expressions and Monte Carlo results for large enough n does, however, indicate that the Monte Carlo technique is not necessary for

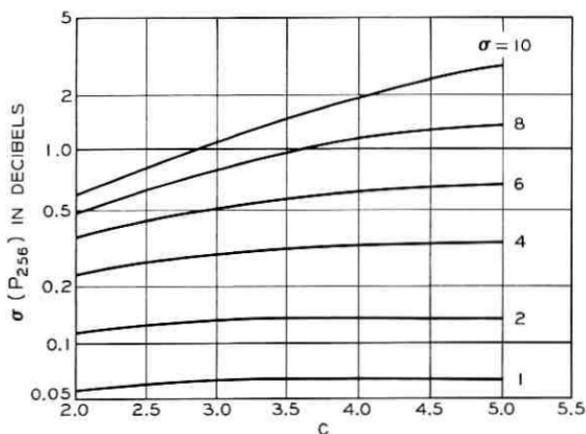


Fig. 10—Estimates of $\sigma(P_{256})$. The components are truncated normal; $\mu = 0$, truncation at $\pm c\sigma$.

power sum evaluations beyond a certain n -value, namely, the one where the asymptotic expressions become sufficiently accurate.

Finally, we note that the problem of evaluating the distribution of the power sum of nontruncated normal components has not been brought closer to its solution by the results presented here. This problem is certainly of mathematical interest even though it represents a physically unrealistic situation. Some Monte Carlo studies with larger values for the truncation points have indicated that the convergence of the power sum to normality is much less rapid in this case, and that considerably larger values of the measures of skewness and excess can occur than those contained in Table I.

VII. ACKNOWLEDGMENTS

Miss M. L. Chubb and F. P. Duffy wrote the computer programs leading to the numerical results presented here. R. J. Christie suggested the method for converting the random numbers from uniform to truncated normal distribution. The proof of the Lemma in Appendix A is due to N. A. Marlow with whom I had several interesting discussions. D. A. Lewinski made several valuable comments. I thank them all for their contributions.

APPENDIX A

Let X_1, X_2, \dots, X_n be independent identically distributed random variables. Put

$$S_n = 10^{X_1/10} + \dots + 10^{X_n/10}$$

and let $\theta = E[10^{X_1/10}]$. In order to prove the asymptotic results in the main body of the paper, we need the following.

Lemma: Suppose

$$Q(x) = x^l \left[\int_0^1 \frac{t^m dt}{1 + xt} \right]^j, \quad x > -1$$

where l, j, m are nonnegative integers. If $E10^{2iX_1/10}$ and $E10^{-iX_1/10}$ are bounded, then

$$E \left[Q \left(\frac{S_n - n\theta}{n\theta} \right) \right] = O(n^{-1/2}) \quad \text{as } n \rightarrow \infty.$$

Proof: Let

$$I_m(x) = \int_0^1 \frac{t^m}{1 + xt} dt, \quad x > -1.$$

Then $Q(x) = x^l [I_m(x)]^j$, and it follows from the Cauchy-Schwarz inequality that

$$\left| E \left[Q \left(\frac{S_n - n\theta}{n\theta} \right) \right] \right|^2 \leq E \left(\frac{S_n - n\theta}{n\theta} \right)^{2l} E \left[I_m \left(\frac{S_n - n\theta}{n\theta} \right) \right]^{2j}.$$

The asymptotic behavior of the central moment of S_n of order $2l$ is found from Cramér.¹⁰ Hence,

$$\left| E \left[Q \left(\frac{S_n - n\theta}{n\theta} \right) \right] \right|^2 = O(n^{-l}) E \left[I_m \left(\frac{S_n - n\theta}{n\theta} \right) \right]^{2j} \quad \text{as } n \rightarrow \infty.$$

To complete the proof, it suffices to show that

$$E \left[I_m \left(\frac{S_n - n\theta}{n\theta} \right) \right]^{2j} = O(1) \quad \text{as } n \rightarrow \infty.$$

To show this, we again apply the Cauchy-Schwarz inequality. Thus,

$$I_m^2(x) \leq \int_0^1 t^{2m} dt \int_0^1 \frac{dt}{(1+xt)^2} = \frac{1}{2m+1} \frac{1}{x+1}.$$

Hence,

$$E \left[I_m \left(\frac{S_n - n\theta}{n\theta} \right) \right]^{2j} \leq \left(\frac{\theta}{2m+1} \right)^j E \left(\frac{n}{S_n} \right)^j.$$

Consider now the function $u(x) = 1/x^j$, which is convex on $(0, \infty)$ for $j \geq 0$. By Jensen's inequality it follows that if $\alpha_1, \dots, \alpha_n, y_1, \dots, y_n$ are non-negative real numbers such that $\alpha_1 + \dots + \alpha_n = 1$, then

$$u(\alpha_1 y_1 + \dots + \alpha_n y_n) \leq \alpha_1 u(y_1) + \dots + \alpha_n u(y_n).$$

In particular,

$$\begin{aligned} (n/S_n)^j &= u(S_n/n) \leq (1/n)[u(10^{X_1/10}) + \dots + u(10^{X_n/10})] \\ &= (1/n)[10^{-jX_1/10} + \dots + 10^{-jX_n/10}]. \end{aligned}$$

Hence,

$$E(n/S_n)^j \leq E[10^{-jX_1/10}].$$

The right-hand side of this inequality is finite by assumption, so the proof is complete.

APPENDIX B

Derivation of the Wilkinson Results for Truncated Normal Components

As in Appendix A, let X_1, X_2, \dots, X_n be independent, identically distributed random variables, and assume further that they all have a

truncated normal distribution. The density function of X_1 is then

$$g(x) = \begin{cases} 0, & x < \mu - c\sigma, \quad x > \mu + c\sigma \\ \frac{1}{\Phi(c) - \Phi(-c)} \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), & \mu - c\sigma \leq x \leq \mu + c\sigma \end{cases}$$

where Φ stands for the standardized normal distribution function.

Now let W_i be the nonnegative random variable that expresses the power corresponding to X_i , i.e.,

$$W_i = 10^{X_i/10}.$$

The density function of W_1 is

$$f(w) = \begin{cases} 0, & w < 10^{(\mu - c\sigma)/10}, \quad w > 10^{(\mu + c\sigma)/10} \\ \frac{1}{\Phi(c) - \Phi(-c)} \frac{1}{\sqrt{2\pi} \sigma \lambda w} \exp\left(-\frac{(\log w - \lambda\mu)^2}{2\lambda^2 \sigma^2}\right), & 10^{(\mu - c\sigma)/10} \leq w \leq 10^{(\mu + c\sigma)/10} \end{cases}$$

The moments of W_1 are therefore,

$$EW_1^k = \int_0^\infty w^k f(w) dw = \exp(k\lambda\mu + \frac{1}{2}k^2\lambda^2\sigma^2)T_c(k\sigma),$$

where

$$T_c(\sigma) = \frac{\Phi(c - \lambda\sigma) - \Phi(-c - \lambda\sigma)}{\Phi(c) - \Phi(-c)}$$

accounts for the effect of the truncation. We note that $T_c(\sigma) \rightarrow 1$ as $c \rightarrow \infty$.

The mean and variance of W_1 are found to be

$$\theta = EW_1 = \exp(\lambda\mu + \frac{1}{2}\lambda^2\sigma^2)T_c(\sigma) \quad (21)$$

and

$$\tau_2 = \text{Var}(W_1) = \exp(2\lambda\mu + \lambda^2\sigma^2)T_c^2(\sigma)[\exp(\lambda^2\sigma^2)U_c(\sigma) - 1], \quad (22)$$

where

$$U_c(\sigma) = \frac{T_c(2\sigma)}{T_c^2(\sigma)}.$$

Now let P_n be the power sum of X_1, X_2, \dots, X_n and take $\mu = 0$. Furthermore, let P_n be approximated by a normally distributed random

variable P_{nw} . The independence of the X_i 's then allows us to establish two equations by adding the means and variances of the W_i 's to get the mean and variance, respectively, of

$$S_{nw} = 10^{P_{nw}/10} = W_1 + \dots + W_n.$$

Relations (21) and (22) allow mean and variance of S_{nw} to be expressed in terms of mean and variance of P_{nw} . Hence, we get

$$n \exp(\frac{1}{2}\lambda^2\sigma^2)T_c(\sigma) = \exp[\lambda\mu(P_{nw}) + \frac{1}{2}\lambda^2\sigma^2(P_{nw})]$$

and

$$\begin{aligned} n \exp(\lambda^2\sigma^2)T_c^2(\sigma)[\exp(\lambda^2\sigma^2)U_c(\sigma) - 1] \\ = \exp[2\lambda\mu(P_{nw}) + \lambda^2\sigma^2(P_{nw})][\exp(\lambda^2\sigma^2(P_{nw})) - 1]. \end{aligned}$$

Solving these two equations for $\mu(P_{nw})$ and $\sigma^2(P_{nw})$ we find

$$\mu(P_{nw}) = LAP_n - \frac{1}{2\lambda} \log \left[1 + \frac{\exp(\lambda^2\sigma^2)U_c(\sigma) - 1}{n} \right]$$

and

$$\sigma^2(P_{nw}) = \frac{1}{\lambda^2} \log \left[1 + \frac{\exp(\lambda^2\sigma^2)U_c(\sigma) - 1}{n} \right],$$

where

$$LAP_n = \frac{1}{\lambda} \log n + \frac{1}{2}\lambda\sigma^2 + \frac{1}{\lambda} \log T_c(\sigma).$$

REFERENCES

1. Dixon, J. T., unpublished work, 1932.
2. Wilkinson, R. L., unpublished work, 1934.
3. Holbrook, B. D. and Dixon, J. T., Load Rating Theory for Multichannel Amplifiers, B.S.T.J., 18, October, 1939, p. 624.
4. Curtis, H. E., Probability Distribution of Noise Due to Fading on Multisection FM Microwave Systems, IRE Trans. Commun. Syst., September, 1959, p. 161.
5. Roberts, J. H., Sums of Probability Distributions Expressed in Decibel Steps, Proc. IEE, 110, No. 4, April, 1963, p. 692.
6. Derzai, M., Power Addition of Independent Random Variables Normally Distributed on a dB Scale, 1967 IEEE International Convention Record, I, p. 40, March, 1967.
7. Fenton, L. F., The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems, IRE Trans. Commun. Syst., March, 1960, p. 57.
8. Marlow, N. A., A Normal Limit Theorem for Power Sums of Independent Random Variables, B.S.T.J., this issue, p. 2081.
9. Cyr, M. H. and Thuswaldner, A., Multichannel Load Calculation Using the Monte Carlo Method, IEEE Trans. Commun. Tech., COM-14, No. 2, April, 1966, p. 177.
10. Cramér, H., *Mathematical Methods of Statistics*, Princeton, 1945, p. 346.

Random Packings and Coverings of the Unit n -Sphere

By A. D. WYNER

(Manuscript received July 13, 1967)

It is well known that the quantity $M_p(n, \theta)$, the maximum number of nonoverlapping spherical caps of half angle θ (a "packing") which can be placed on the surface of a unit sphere in Euclidean n -space is not less than $\exp[-n \log \sin 2\theta + o(n)]$ ($\theta < \pi/4$). In this paper we give a new proof of this fact by a "random coding" argument, the central part of which is a theorem which asserts that if a set of roughly $\exp(-n \log \sin 2\theta)$ caps is chosen at random, that on the average only a very small fraction of the caps will overlap (when n is large).

A related problem is the determination of $M_c(n, \theta)$, the minimum number of caps of half angle θ required to cover the unit Euclidean n -sphere. We show that $M_c(n, \theta) = \exp[-n \log \sin \theta + o(n)]$. The central part of the proof is also a random coding argument which asserts that if a set roughly $\exp(-n \log \sin \theta)$ caps is chosen at random, that on the average only a very small fraction of the surface of the n -sphere will remain uncovered (when n is large).

I. INTRODUCTION

A problem in coding theory for the Gaussian channel is the determination of $M_p(n, \theta)$, the maximum number of points which may be placed on the surface of a unit n -sphere such that the spherical caps with centers at these points and half angle θ are disjoint (the "packing" problem). This quantity, though unknown, has been estimated by upper and lower bounds.⁵ In this paper, we give a proof of the known lower bound by a "random coding" argument. It is felt that this new method is of interest in itself.

A related problem is the "covering" problem, the determination of $M_c(n, \theta)$, the minimum number of caps of half angle θ required to cover the surface of a unit n -sphere. This problem is of interest when one wants to quantize an n -dimensional Gaussian vector with inde-

pendent components (which with very high probability lies near the surface of an n -sphere). In this paper, $M_c(n, \theta)$ is estimated with upper and lower bounds which are "exponentially" tight. The upper bound is also proved by a "random coding" argument.

The random coding arguments owe much to Shannon.^{3, 4} The random covering theorem in particular is similar to his approximation theorem in the latter reference. R. Graham has called my attention to the work of Rogers,^{1, 2} who has considered the problem of covering a large n -dimensional cube with spheres of a unit radius. Rogers' methods and result parallel those given here.

Let x, y with and without subscripts denote points on S_n , the surface of a unit sphere in n -dimensional Euclidean space. Let $\alpha(x, y)$ be the angle* between x and y , and note that $\alpha(x, y)$ satisfies the axioms of a metric. For $0 \leq \theta \leq \pi$, let $\mathcal{C}(x, \theta) = \{y : \alpha(x, y) < \theta\}$, the open spherical cap of half angle θ centered at x . A set $S \subseteq S_n$ is said to be a θ -covering ($0 \leq \theta \leq \pi$) if $\bigcup_{x \in S} \mathcal{C}(x, \theta)$ covers S_n , and $S \subseteq S_n$ is said to be a θ -packing if $\mathcal{C}(x, \theta) \cap \mathcal{C}(y, \theta)$ is empty for $x, y \in S, x \neq y$. Let $M_c(n, \theta)$ be the minimum number of points which can constitute a θ -covering of S_n and let $M_p(n, \theta)$ be the maximum number of points which can constitute a θ -packing. These quantities are related by

Lemma 1: $M_c(n, 2\theta) \leq M_p(n, \theta)$.

Proof: We say that $S \subseteq S_n$ is a *maximal θ -packing* if S is a θ -packing, and for all $y \notin S$, the union $\{y\} \cup S$ is not a θ -packing. We establish Lemma 1 by showing that every maximal θ -packing is a 2θ -covering. Let S be a maximal θ -packing. If S is not a 2θ -covering then there exists a y such that $\alpha(x, y) \geq 2\theta$ for all $x \in S$. Thus, from the triangle inequality for α , $\mathcal{C}(x, \theta) \cap \mathcal{C}(y, \theta) = \Phi$ for all $x \in S$, and $\{y\} \cup S$ is a θ -packing contradicting the maximality of S . Hence, the lemma.†

The quantity $M_p(n, \theta)$ is well studied.⁵ In particular, it is known that (for $\theta < \pi/4$)

$$\exp [nP_L(\theta)(1 + \beta_n(\theta))] \leq M_p(n, \theta) \leq \exp [nP_U(\theta)(1 + \gamma_n(\theta))], \quad (1a)$$

where $\beta_n, \gamma_n \rightarrow 0$ as $n \rightarrow \infty$ and

$$P_L(\theta) = -\log \sin 2\theta, \quad (1b)$$

* The angle is defined as follows. Say that the center of the unit sphere is the origin of coordinates in n -space. Then x and y may be thought of a unit vectors. The angle $\alpha(x, y)$ between them is defined by $\cos \alpha = \text{inner product of } x \text{ and } y$, where $0 \leq \alpha \leq \pi$.

† The fact that it does not seem possible to obtain a reverse inequality relating M_c and M_p may lead one to suspect that covering and packing are, in fact, not dual problems. This may account for the fact that random coding appears "better" for covering than for packing.

and

$$P_U(\theta) = -\log \sqrt{2} \sin \theta. \quad (1c)$$

Thus, roughly speaking $M_p(n, \theta)$ increases exponentially in n (as $n \rightarrow \infty$) with exponent between P_L and P_U .

In Section III we give another proof of the lower bound in (1). The central part of this proof is a theorem that asserts that if a packing with roughly $\exp [nP_L(\theta)]$ points is chosen at random, that on the average only a very small fraction of the caps will overlap (Theorem 1). The lower bound of (1) is a corollary to this theorem. It is felt that Theorem 1 is of interest in itself.

Now consider $M_c(n, \theta)$. We will show that it too increases roughly exponentially in n (as $n \rightarrow \infty$). But here we can find the exponent exactly, viz., (for $\theta < \pi/2$)

$$M_c(n, \theta) = \exp [nR_c(\theta)(1 + \epsilon_n(\theta))], \quad (2a)$$

where $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$ and

$$R_c(\theta) = -\log \sin \theta. \quad (2b)$$

The central part of the proof of the existence of a covering satisfying (2) is a theorem which asserts that if a covering with roughly $\exp [nR_c(\theta)]$ points is chosen at random, that on the average only a very small fraction of S_n will remain uncovered.

II. THEOREMS

In this section we give precise statements of our theorems, leaving the proofs for Section III. We begin with some definitions.

Assign the usual "area" measure to S_n . If $A \subseteq S_n$ is measurable, let $\mu(A)$ be its measure. In particular, let

$$C_n(\alpha) = \mu(\mathcal{C}(x, \alpha)) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma[(n+1)/2]} \int_0^\alpha \sin^{(n-2)} \varphi \, d\varphi \quad (3a)$$

be the area (measure) of a cap of half-angle α , and let

$$C_n(\pi) = \frac{n\pi^{n/2}}{\Gamma[(n+2)/2]} \quad (3b)$$

be the area of S_n . It is easy to show that (for $\alpha < \pi/2$)

$$\frac{C_n(\pi)}{C_n(\alpha)} = \exp \left\{ n \log \left(\frac{1}{\sin \alpha} \right) + o(n) \right\}. \quad (4)$$

as $n \rightarrow \infty$.

In connection with the packing problem, let $S = \{x_i\}_{i=1}^M \subseteq S_n$, and consider $\{C(x_i, \theta)\}_{i=1}^M$ the corresponding caps of half-angle θ . Define

$$F_p(S, \theta) = \frac{1}{M} \sum_{i=1}^M g_i(S, \theta), \quad (5a)$$

where g_i ($i = 1, 2, \dots, M$) is defined by

$$g_i(S, \theta) = \begin{cases} 1, & C(x_i, \theta) \cap C(x_j, \theta) = \Phi \text{ all } j \neq i, \\ 0, & \text{otherwise.} \end{cases} \quad (5b)$$

Thus, $F_p(S, \theta)$ is the fraction of the caps which do not overlap. Notice that S is a θ -packing if and only if $F_p(S, \theta) = 1$. We now state

Theorem 1: (Random Packing) Consider a random experiment in which the M members of S are chosen independently with uniform distribution on S_n . $F_p(S, \theta)$ is then a random variable. Let θ be fixed and let M increase as $n \rightarrow \infty$, then

$$\text{if } M \frac{C_n(2\theta)}{C(\pi)} \rightarrow \infty, \quad EF_p(S, \theta) \rightarrow 0 \quad (6a)$$

and

$$\text{if } M \frac{C_n(2\theta)}{C(\pi)} \rightarrow 0, \quad EF_p(S, \theta) \rightarrow 1, \quad (6b)$$

where E denotes expectation.

Thus, in particular, if $M = e^{\rho n}$ (ρ fixed), we have from (4) that $EF_p(S, \theta) \rightarrow 1$ or 0 according as $\rho < -\log \sin 2\theta = P_L(\theta)$ or $\rho > P_L(\theta)$. Further, since there must be a set S such that $F_p(S, \theta) \geq EF_p$, we conclude that for any $\rho < P_L(\theta)$ and any $\epsilon > 0$ there exists an n sufficiently large and a set $S \subseteq S_n$ with $M = e^{\rho n}$ members such that

$$F_p(S, \theta) \geq 1 - \epsilon. \quad (7)$$

If we delete the (ϵM) members of S with overlapping caps we obtain a θ -packing with $M = e^{\rho n}(1 - \epsilon)^n$ points. This is equivalent to the lower bound of (1).

Let us now turn to the covering problem. We can easily establish a lower bound on $M_c(n, \theta)$ as follows. Let $S = \{x_i\}_{i=1}^M \subseteq S_n$ be a θ -covering, so that $\bigcup_{i=1}^M C_n(x_i, \theta)$ covers S_n . Hence,

$$C_n(\pi) = \mu(S_n) = \mu \bigcup_{i=1}^M C(x_i, \theta) \leq \sum_{i=1}^M \mu(C(x_i, \theta)) = MC_n(\theta). \quad (8)$$

Thus, we have proved

Lemma 2: $M_c(n, \theta) \geq C_n(\pi)/C_n(\theta)$.

In the light of (4), Lemma 2 implies that M_c is not less than the right member of (2a) for $\theta < \pi/2$.

Let $\beta > 0$ and $S \subseteq S_n$ be given. Define the set

$$B(S, \beta) = \{y \in S_n : y \notin \mathcal{C}(x, \beta) \text{ for all } x \in S\}. \quad (9a)$$

Then

$$F_c(S, \beta) = \mu(B(S, \beta))/C_n(\pi) \quad (9b)$$

represents that fraction of S_n not covered by the caps $\mathcal{C}(x, \beta)$, $x \in S$. We now state

Theorem 2: (Random Covering) Consider a random experiment in which the M members of a set S are chosen independently with uniform distribution on S_n . Then $F_c(S, \beta)$ is a random variable. Let $\beta < \pi$ be fixed and let M increase as $n \rightarrow \infty$, then

$$\text{if } M \frac{C_n(\beta)}{C_n(\pi)} \rightarrow \infty, \quad E(F_c) \rightarrow 0, \quad (10a)$$

and

$$\text{if } M \frac{C_n(\beta)}{C_n(\pi)} \rightarrow 0, \quad E(F_c) \rightarrow 1. \quad (10b)$$

Further,

$$E(F_c) \leq \exp \left\{ -M \frac{C_n(\beta)}{C_n(\pi)} \right\}. \quad (11)$$

In particular, if $M = e^{\rho n}$ (ρ fixed) and $\beta < \pi/2$, we have from (10) and (4) that $E(F_c) \rightarrow 0$ or 1 according as $\rho > -\log \sin \beta = R_c(\beta)$ or $\rho < R_c(\beta)$. Further, since there must be at least one set S for which $F_c(S, \beta) \leq EF_c$, we conclude from (11) and (4) that for any $\beta < \pi/2$ and any $\rho > R_c(\beta)$ there exists for each $n = 1, 2, \dots$ a set $S \subseteq S_n$ with $M = e^{\rho n}$ members such that

$$\frac{\mu(B(S, \beta))}{C_n(\pi)} \leq \exp \{ -\exp [(\rho - R_c(\beta))n(1 + \lambda(\beta))] \}, \quad (12)$$

where $\lambda(\beta) \rightarrow 0$ as $n \rightarrow \infty$. The following corollary (also proved in Section III) follows from (12).

Corollary: Let $\theta(0 < \theta < \pi/2)$ be arbitrary and let $\rho > R_c(\theta)$. Then for n sufficiently large there exists a θ -covering of S_n with $M = e^{\rho n}$ points.

It remains to show that M_c is not more than the right member of (2a). For $\theta < \pi/2$ let

$$\rho^*(\theta) = \limsup_{n \rightarrow \infty} \frac{1}{n} \log M_c(n, \theta).$$

Say $\rho^* > R_c(\theta)$. Let $\rho' = (R(\theta) + \rho^*(\theta))/2 < \rho'$. We conclude that there is an infinite sequence of n 's such that any set of $e^{\rho'n}$ points in S_n cannot be a θ -covering. But since $\rho' > R_c(\theta)$, application of the above corollary yields a contradiction. Thus, $\rho^* \leq R_c(\theta)$. This taken together with Lemma 2 gives

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M(n, \theta) = R_c(\theta),$$

from which (2) follows.

III. PROOFS

Proof of Theorem 1: Let the points $x_1, x_2, \dots, x_M \in S_n$ be chosen independently with a uniform distribution on S_n . The random variables g_i ($i = 1, 2, \dots, M$) defined in (5b) may be rewritten

$$g_i(x_1, x_2, \dots, x_M, \theta) = \begin{cases} 1, & \alpha(x_i, x_j) \geq 2\theta, \quad j \neq i, \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

Thus, the random variable F_p of (5a) has expectation

$$EF_p = \frac{1}{M} \sum_{i=1}^M Eg_i = \frac{1}{M} \sum_{i=1}^M \Pr \{g_i = 1\}. \quad (14)$$

Let i be fixed. If $x_i = x$ then $g_i = 1$ if and only if the $(M - 1)$ independent choices of $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_M$ do not belong to $\mathcal{C}(x, 2\theta)$. Since the x_i are uniformly distributed on S_n we have

$$\Pr \{g_i = 1 \mid x_i = x\} = \left(1 - \frac{C_n(2\theta)}{C_n(\pi)}\right)^{M-1},$$

independent of x . Thus, from (14)

$$E(F_p) = \left(1 - \frac{C_n(2\theta)}{C_n(\pi)}\right)^{M-1} = \left(1 - \frac{1}{\mu_n}\right)^{\mu_n[(M-1)/\mu_n]}, \quad (15)$$

where $\mu_n = C_n(\pi)/C_n(2\theta)$. Our result follows on noting that as $n \rightarrow \infty$,

$$(1 - 1/\mu_n)^{\mu_n} \rightarrow e^{-1} \quad \text{and} \quad (M - 1)/\mu_n \approx MC_n(2\theta)/C_n(\pi).$$

Proof of Theorem 2: Let the points $x_1, x_2, \dots, x_M \in S_n$ be chosen independently with a uniform distribution on S_n . The random variable F_c may be written

$$F_c = \frac{1}{C_n(\pi)} \int_{S_n} h(y, x_1, x_2, \dots, x_M) d\mu(y), \quad (16)$$

where

$$h(y, x_1, \dots, x_M) = \begin{cases} 1, & \text{if } \alpha(x_i, y) \geq \theta, \quad 1 \leq i \leq M \\ 0, & \text{otherwise.} \end{cases}$$

Since $h \geq 0$ we may interchange the expectation and integration operations and obtain

$$EF_c = \frac{1}{C_n(\pi)} \int_{S_n} d\mu(y) Eh(y, x_1, \dots, x_M),$$

where as indicated Eh is computed with y held fixed. Now

$$\begin{aligned} Eh(y, x_1, \dots, x_M) &= \Pr \{h = 1\} = \Pr \bigcap_{i=1}^M \{\alpha(x_i, y) \geq \theta\} \\ &= \left(1 - \frac{C_n(\theta)}{C_n(\pi)}\right)^M \leq \exp \left[-M C_n(\theta)/C_n(\pi)\right], \end{aligned}$$

from which (10) and (11) follow.

Proof of Corollary to Theorem 2: Let $\rho > R_c(\theta)$ be given. Let γ be defined by $R_c(\gamma) = \rho$. Since $\rho > R_c(\theta)$ a decreasing function, we have $\gamma < \theta$. We will apply Theorem 2 with $\beta = (\theta + \gamma)/2$, so that $\rho > R(\beta)$. Let S_n ($n = 1, 2, \dots$) be the sets which satisfy (12). By (4) and (12), $C_n[(\theta - \gamma)/2]/C_n(\pi)$ decreases much more slowly (as $n \rightarrow \infty$) than $[\mu(B(S_n, \beta))]/C_n(\pi) \triangleq \delta_n$, so that we can find an N sufficiently large such that for $n \geq N$,

$$\delta_n < \frac{C_n[(\theta - \gamma)/2]}{C_n(\pi)}.$$

We claim that for $n > N$, the sets S_n are θ -coverings of S_n . To show this observe that if $y \notin \bigcup_{x_i \in S_n} \mathcal{C}(x_i, \theta)$, then $\alpha(x_i, y) < \theta$, all $x_i \in S_n$. Thus,

$$\mathcal{C}\left(y, \frac{\theta - \gamma}{2}\right) \cap \mathcal{C}\left(x_i, \frac{\theta + \gamma}{2}\right) = \Phi \text{ for all } x_i \in S_n,$$

which in turn implies

$$e\left(y, \frac{\theta - \gamma}{2}\right) \subseteq B\left(S_n, \frac{\theta + \gamma}{2}\right).$$

Thus,

$$\delta_n = \frac{\mu\left\{B\left(S_n, \frac{\theta + \gamma}{2}\right)\right\}}{C_n(\pi)} \geq \frac{\mu\left\{e\left(y, \frac{\theta - \gamma}{2}\right)\right\}}{C_n(\pi)} = \frac{C_n\left(\frac{\theta - \gamma}{2}\right)}{C_n(\pi)},$$

a contradiction. Thus, there is no such y and the corollary follows.

REFERENCES

1. Rogers, C. A., A Note on Coverings, *Mathematica*, 4, 1957, pp. 1-6.
2. Rogers, C. A., *Packing and Covering*, Cambridge University Press, Cambridge, 1964.
3. Shannon, C. E., A Mathematical Theory of Communication, *B.S.T.J.*, 27, 1948, pp. 379-423 and pp. 623-656.
4. Shannon, C. E., Coding Theorems for a Discrete Source with a Fidelity Criteria, 1959 IRE Conv. Record, Part 4, pp. 142-163.
5. Wyner, A. D., Capabilities of Bounded Discrepancy Decoding, *B.S.T.J.*, 44, 1965, pp. 1061-1122. (The tightest known bounds are summarized on pp. 1071-1072, Eq. 24.)

Slope Overload Noise in Differential Pulse Code Modulation Systems

By E. N. PROTONOTARIOS

(Manuscript received June 12, 1967)

In differential pulse code modulation (DPCM) systems, often referred to as predictive quantizing systems, the quantizing noise manifests itself in two forms, granular noise and slope overload noise. The study of overload noise in DPCM may be abstracted to the following stochastic processes problem. Let the input to the system be a Gaussian stochastic process $\{x(t)\}$ with a bandlimited $(0, f_0)$ spectrum $F(f)$. Denote the output of the system by $y(t)$. Most of the time $y(t)$ is equal to $x(t)$. During time intervals of this kind, the absolute value of the derivative $x'(t) = dx(t)/dt$ is less than a given positive constant x'_0 . (In a DPCM system, $x'_0 = kf_s$, where k is the maximum level of the quantizer and f_s is the sampling frequency.) There are time intervals, $I_i(t_0^{(i)}, t_1^{(i)})$ ($i = 0, \pm 1, \pm 2, \dots$), for which $y(t) \neq x(t)$. These time intervals begin at time instants $t_0^{(i)}$ such that $|x'(t_0^{(i)})|$ increases through the value x'_0 . For $t \in I_i$, $y(t) = x(t_0^{(i)}) + (t - t_0^{(i)})x'_0$. The interval ends at $t_1^{(i)}$, when $x(t)$ and $y(t)$ become equal again. The overload noise in the DPCM system is defined to be $n(t) = x(t) - y(t)$. The problem is to study the random process $\{n(t)\}$. In the present paper, we will give an upper bound to the average noise power $\langle n^2(t) \rangle_{av}$, which at the same time is a very good approximation to the noise power itself.

Two previous attempts have been made to find $\langle n^2(t) \rangle_{av}$. One, due to Rice and O'Neal, involves an approximation valid only for very large x'_0 . Another approach to the problem, due to Zetterberg, includes an ingenious way of avoiding the determination of $t_1^{(i)}$. A new approach is given here that combines the best features of the two methods. The present result is a better approximation for slope overload noise than has been previously obtained. The result differs from previous results but is asymptotically equal to that given by Rice and O'Neal for $x'_0 \rightarrow \infty$. In the region where overload noise is important, the present result is in very good agreement

with computer simulation and experiment. The technique used could be applied for the determination of other statistical characteristics of the error random process.

I. INTRODUCTION

This paper is concerned with the slope overload noise in Differential Pulse Code Modulation (DPCM) systems, often referred to as predictive quantizing systems. Delta Modulation (ΔM), the simplest member of the DPCM family, is a European invention of the mid-forties.¹ DPCM was first revealed in a Phillips Company patent² in 1951 and as a predictive quantizing system in a patent by C. C. Cutler³ of the Bell Telephone Laboratories in 1952. ΔM and DPCM are receiving renewed attention due to the present trend toward digital communications and general efforts aimed at redundancy reduction⁴ in picture transmission. The present work was motivated, to a large extent, by the application of DPCM to Picturephone[®] signal transmission.

Work on ΔM and DPCM was reported in the early and mid-fifties. Most representative are the papers by (i) DeJager⁵ on ΔM , mainly of introductory and descriptive nature, (ii) Van de Weg⁶ on uniform DPCM—we will refer to it in the sequel, and (iii) Zetterberg⁷ whose long paper on ΔM is the most detailed study of the subject to date. Recent publications note the beginning of a "renaissance" period for ΔM and DPCM.^{8,9,10,4}

In DPCM systems the quantization noise manifests itself in two forms, the granular noise and the slope overload noise. The granular noise is essentially uncorrelated with the input signal and has a more or less flat power spectrum and an approximately uniform amplitude probability distribution, resembling the granular noise in standard PCM. The granular noise for single integration DPCM systems with a uniform quantizer has been studied by Van de Weg.⁶

In contrast with a straight PCM system, which overloads in amplitude, a differential PCM system overloads in slope. Consider a DPCM system (Fig. 1) with a single integrator in the feedback path and a symmetric quantizer which is not necessarily uniform. Practical DPCM systems have leaky integrators. For simplicity, we are considering only perfect integrators here. Let k be the maximum level of the quantizer and f_s the sampling frequency. Then the maximum slope that the system can follow is $x'_0 = kf_s$, corresponding to the emission of a string of impulses of strength k by the quantizer of Fig. 1. For a fixed value of $x'_0 = kf_s$, and for $k \rightarrow 0$ the granular noise tends to zero, and the total

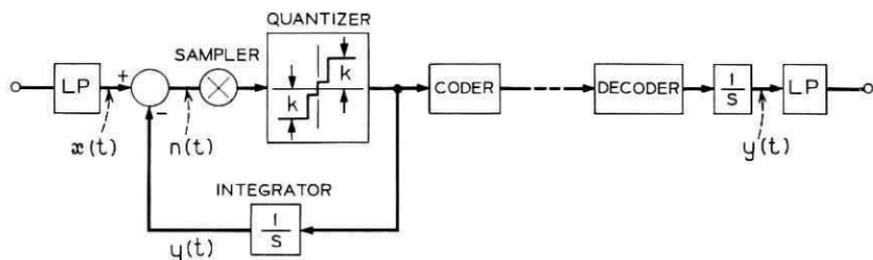


Fig. 1 — Single integration DPCM with a symmetric quantizer.

noise is due to slope overload alone. In this paper, we concentrate on certain statistics of the overload noise defined precisely in Section III.

II. SUMMARY OF RESULTS AND COMPARISON WITH PREVIOUS WORK

There exist two previous papers concerning overload noise in DPCM systems. Approximate results are given for the slope overload noise N_0 in terms of the slope capability x'_0 of the DPCM system and the power spectrum of the input signal, assumed to be Gaussian. The result due to Zetterberg⁷ (with some corrections) is as follows

$$N_{0,z} = \frac{4\sqrt{2}}{35\pi^{\frac{3}{2}}} \left(\frac{b_1^2}{b_2}\right) \left(\frac{3b_1^{\frac{3}{2}}}{x'_0}\right)^5 A(\lambda) \exp\left(-\frac{x_0'^2}{2b_1}\right),$$

where b_1 and b_2 are the variances of the first and second derivatives of the input signal, respectively, and they are given in terms of the spectrum in (1) of the following section. The quantity λ and the function $A(\lambda)$ are defined in (31) and (32), respectively. The second result is due to Rice and O'Neal.⁸ Their basic approximations are: (i) a truncation of the Taylor series for $x(t)$, around a transition point, including terms through the third derivative; and (ii) the assumption that the third derivative of $x(t)$ at the transition points has, as a random variable, a very small variance compared to its mean value. Therefore, the third derivative is taken to be a deterministic constant with value equal to its mean. With these assumptions, (22) of Ref. 8 results in

$$N_{0,r} = \frac{1}{4\sqrt{2\pi}} \left(\frac{b_1^2}{b_2}\right) \left(\frac{3b_1^{\frac{3}{2}}}{x'_0}\right)^5 \exp\left(-\frac{x_0'^2}{2b_1}\right).$$

There are two points that we want to make here:

(i) When the formula above together with an expression for the granular noise given in Ref. 8 are used to compute S/N we see that the

agreement with computer simulation is not very satisfactory in the region of severe slope overload. This formula does, however, identify the peak of the S/N ratio quite successfully (see Fig. 11).

(ii) When we compare Zetterberg's and Rice's results by considering the ratio $N_{0,Z}/N_{0,R}$ we get

$$\frac{N_{0,Z}}{N_{0,R}} = \frac{32}{35\pi} A(\lambda) = \frac{1}{3.44} A(\lambda)$$

$$10 \log_{10} \frac{N_{0,Z}}{N_{0,R}} = -5.36 + 10 \log_{10} A(\lambda) \text{ dB.}$$

Thus, we see that the two results differ substantially.

Hence, the question of the average slope overload noise power cannot be considered settled since the two results above are different and they both differ from computer simulation and experiment. The present paper sheds further light on the question of the slope overload noise. Our principle result is the approximation

$$N_0 = \frac{1}{4\sqrt{2\pi}} \left(\frac{b_1^2}{x_0'} \right) \left(\frac{3b_1^4}{x_0'} \right)^5 \exp \left(-\frac{x_0'^2}{2b_1} \right) A(\chi),$$

where the quantity χ and the function $A(\chi)$ are defined in (64) and (66), respectively. This expression, like the previous ones, is a function of only two things—the maximum slope capability x_0' of the DPCM system and the power spectrum of the input signal. Indeed all the variables appearing in this formula are calculated directly from these two quantities only [see (1) and (64)]. The present formula gives better agreement with computer simulation than the one by Rice and O'Neal, when used to compute S/N (see Fig. 11).

We might also point out here that the present work applies to any system which is slope limited, not just to DPCM or digital encoding systems.

III. PROBLEM DEFINITION

With reference to Fig. 1 let the input $\{x(t)\}$ be a stationary band-limited Gaussian random process. Let $\psi(\tau)$ be the autocorrelation function of $x(t)$ and $F(f)$ the one-sided power spectrum. Let f_0 be the bandwidth of $x(t)$ and $F_s = f_s/f_0$ the normalized sampling frequency. The random process $\{x(t)\}$ is assumed to be zero mean. Let b_n be the variance of the n th derivative of $x(t)$ ($n = 1, 2, \dots$). These numbers (b_n) will be extensively used in the sequel. They are given by the relation

$$b_n = \int_0^{f_0} (2\pi f)^{2n} F(f) df. \quad (1)$$

The output signal $y(t)$ follows the input signal $x(t)$ during certain time intervals. Within these time intervals

$$\left| \frac{dx(t)}{dt} \right| < x'_0.$$

The rest of the time $y(t)$ follows segments of straight lines having slope x'_0 or $-x'_0$. If t_0 is a time instant at which a transition from the input signal to the straight line segment takes place, we have

$$x'(t_0) = \frac{dx(t_0)}{dt} = x'_0, \quad x''(t_0) > 0$$

or

$$x'(t_0) = -x'_0, \quad x''(t_0) < 0. \quad (2)$$

For

$$x'(t_0) = x'_0 \quad (3)$$

$$y(t) = x(t_0) + (t - t_0)x'_0 \quad t \in (t_0, t_1)$$

and for

$$x'(t_0) = -x'_0$$

$$y(t) = x(t_0) - (t - t_0)x'_0 \quad t \in (t_0, t_1), \quad (4)$$

where t_1 is the smallest time $t_1 > t_0$ for which

$$x(t_1) = y(t_1) = x(t_0) + (t_1 - t_0)x'(t_0). \quad (5)$$

Since the overload noise is defined to be

$$n(t) = x(t) - y(t), \quad (6)$$

the problem boils down to the study of the random process $\{n(t)\}$. We will concentrate on the derivation of an upper bound to the average noise power $\langle n^2(t) \rangle_{av}$, which at the same time is a very good approximation to the noise power itself. Other statistical properties of $n(t)$ can be obtained, but we will only mention them at the conclusion of the paper.

In contrast with straight PCM the evaluation of the overload noise in DPCM systems is not easy. The beginning of a slope overload burst can be defined statistically in a clear manner. Difficulties arise in defining a valid tractable procedure for determining the duration of the burst and its end point (t_1).

As pointed out before two previous attempts have been made to find $\langle n^2(t) \rangle_{av}$.^{7,8} One, due to Rice and O'Neal,⁸ involves a Taylor series approximation for determining the end point t_1 of the burst valid only for very large x'_0 , i.e., in a region where slope overload noise is not dominant since it is over-shadowed by the granular part of the quantization error. Another approach to the problem is due to Zetterberg.⁷ His approach includes an ingenious way of avoiding the determination of t_1 . Unfortunately, his work contains a conceptual error in the averaging procedure. The error resides in his interpretation of continuous conditional probability density functions in the vertical window sense.

A new approach is given here that combines the best features of the two methods. The result is asymptotically equal to that given by Rice and O'Neal for $x'_0 \rightarrow \infty$. In the region where overload noise is important, the present result is in very good agreement with computer simulation and experiment. As noted above, the technique can also be applied to the determination of other statistical properties of the error random process.

In Section IV, we give a critique of Zetterberg's work. It must be emphasized that Zetterberg's valuable work contains concepts and techniques on which our improved results are based. The wedding of the best in the methods of Rice and Zetterberg is accomplished in our Section V. Theoretical results are compared with computer simulation in Section VI and agreement is seen to be excellent.* Finally, in Section VII we indicate how other statistical properties of $n(t)$ may be obtained by utilizing some of the approaches developed herein.

IV. CRITIQUE OF ZETTERBERG'S APPROACH

Using an argument based on the ergodicity of the random process $\{x(t)\}$ Zetterberg⁷ states that

$$\langle n^2(t) \rangle = \langle n^2(t) \rangle_{av} = S_{x_0} \left\langle \int_0^{s_1} n^2(t_0 + s) ds \right\rangle^\dagger, \quad (7)$$

where

$$\begin{aligned} s &= t - t_0 \\ s_1 &= t_1 - t_0 \end{aligned} \quad (8)$$

and S_{x_0} is the average number of points of transition per second. In what follows, we summarize his procedure deviating slightly from his notation and arguments to clarify a few points. Consider the ensemble

* Comparison with experiments will be given in another paper.¹¹

† $\langle \rangle$ denotes ensemble average and $\langle \rangle_{av}$ time average.

of the sequences $\{t_i(\zeta)\}$, $\dagger i = 0, \pm 1, \pm 2, \dots$, of time instants such that $x'(t_i(\zeta), \zeta) = x'_0$ and $x''(t_i(\zeta), \zeta) > 0$ or $x'(t_i(\zeta), \zeta) = -x'_0$ and $x''(t_i(\zeta), \zeta) \pm 0$ for $i = 0, \pm 1, \pm 2, \dots$. Zetterberg avoids the definition of the end point of the burst by defining a sequence of random processes $\{m_i(s, \zeta)\}$ (see Fig. 2), with index corresponding to the above time instants, in the following way:

$$m_i(s, \zeta) = [x(t_i(\zeta) + s, \zeta) - x(t_i(\zeta), \zeta) - x'_0 s] \cdot \mu(s) \\ \cdot \mu(x(t_i(\zeta) + s, \zeta) - x(t_i(\zeta), \zeta) - x'_0 s) \quad (9)$$

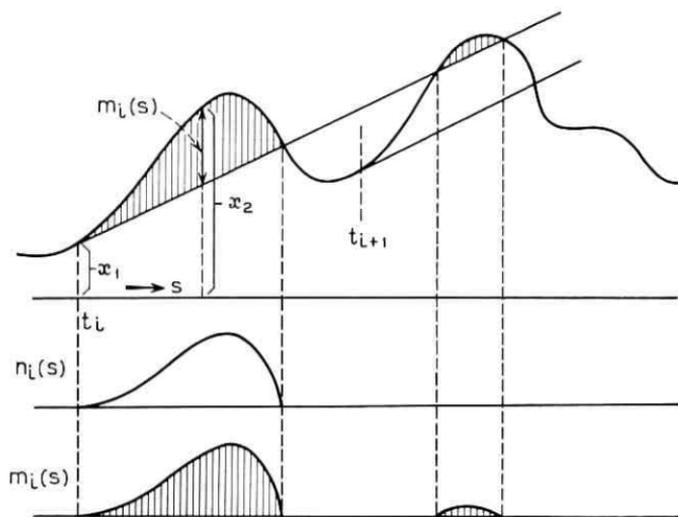


Fig. 2 — An overload noise "burst" $n_i(s)$ and the approximating function $m_i(s)$.

for

$$x'(t_i(\zeta), \zeta) = x'_0 \quad \text{and} \quad x''(t_i(\zeta), \zeta) > 0,$$

and

$$m_i(s, \zeta) = [x(t_i + s, \zeta) - x(t_i, \zeta) + x'_0 s] \mu(s) \\ \cdot \mu(-x(t_i + s, \zeta) + x(t_i, \zeta) - x'_0 s) \quad (10)$$

for

$$x'(t_i, \zeta) = -x'_0, \quad x''(t_i, \zeta) < 0.$$

($\mu(s)$ is the unit step.)

[†] For clarity we show in this paragraph the input random process as generated by an experiment with outcome ζ .

For brevity, we drop the index i and the argument ζ . We denote, as before, the beginning of a burst by t_0 , the end by t_1 and by s_1 its duration, such that for a "positive burst," i.e., $x'(t_0) > 0$ we have

$$m(s) = [x(t_0 + s) - x(t_0) - x'_0 s] \mu(s) \mu[x(t_0 + s) - x(t_0) - x'_0 s]. \quad (11)$$

In general, as shown in Fig. 2, $m(s)$ contains not only noise burst corresponding to the transition point t_0 but also some additional "bursts" cut from the function $x(t)$ by the straight line starting at the point $(t_0, x(t_0))$ and having slope x'_0 . This makes

$$\int_0^\infty m^2(s) ds \geq \int_0^{s_1} n^2(t_0 + s) ds \quad (12)$$

and

$$\left\langle \int_0^\infty m^2(s) ds \right\rangle \geq \left\langle \int_0^{s_1} n^2(t_0 + s) ds \right\rangle. \quad (13)$$

For sufficiently large values of $x'_0/\sqrt{b_1}$, (b_1 is the variance of $x'(t)$), however, the probability is small that the situation depicted in Fig. 2 will occur. Also, generally the additional sections in $m(s)$ occur in reduced amplitude and the squaring reduces the introduced error still further. Denote by $R_{x'_0}$ the average number of points for which

$$x'(t_i) = x'_0, \quad x''(t_i) > 0$$

or

$$x'(t_i) = -x'_0, \quad x''(t_i) < 0.$$

It is seen from Fig. 3 that $R_{x'_0} \geq S_{x'_0}$, since a burst cannot start when another is taking place even if the conditions on the first and second derivative are satisfied. But again, for sufficiently large $x'_0/\sqrt{b_1}$, $R_{x'_0}$ is a good estimate of $S_{x'_0}$. It follows from the discussion above that the quantity

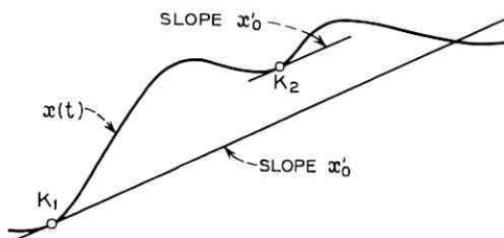


Fig. 3— $K_1, K_2 \subset R_{x'_0}$, whereas $K_1 \subset S_{x'_0}$ but $K_2 \not\subset S_{x'_0}$.

$$R_{x'_0} \left\langle \int_0^\infty m^2(s) ds \right\rangle \quad (14)$$

is an upper bound and actually under certain conditions a good estimate of $\langle n^2(t) \rangle_{av}$.

When one defines

$$Q_{x'_0} \equiv \left\langle \int_0^\infty m^2(s) ds \right\rangle = \int_0^\infty \langle m^2(s) \rangle ds, \quad (15)$$

where the equality above holds provided that the integrals exist, then

$$N_0 = R_{x'_0} Q_{x'_0} \quad (16)$$

is Zetterberg's upper bound to the overload noise.

At this point Zetterberg takes the ensemble average $\langle m^2(s) \rangle$ in the following way:

$$\begin{aligned} \langle m^2(s) \rangle &= \int_{-\infty}^{\infty} \int_{x_1+x'_0s}^{\infty} (x_2 - x_1 - x'_0s)^2 p(x_1, x_2 | \dot{x}_1 = x'_0; s) dx_2 dx_1 \\ &+ \int_{-\infty}^{\infty} \int_{-\infty}^{x_1-x'_0s} (x_2 - x_1 + x'_0s)^2 p(x_1, x_2 | \dot{x}_1 = -x'_0; s) dx_2 dx_1, \quad (17) \end{aligned}$$

where $p(x_1, x_2 | \dot{x}_1; s)$ is the conditional joint probability density function of the random variables $X_1 = x(t_0)$, $X_2 = x(t_0 + s)$ given the value of the random variable $\dot{X}_1 = dx(t_0)/dt$, understood in the vectorial window sense. It turns out that the averaging procedure as described by (17) is wrong for two reasons:

(i) The joint probability density of X_1 and X_2 should be subject not only to the condition $\dot{X}_1 = dx(t_0)/dt = \pm x'_0$, but also to the condition $\ddot{X}_1 = d^2x(t_0)/dt^2 \geq 0$. If we do not impose the above condition on the second derivative at the beginning of the burst, then an $m(s)$ of the form depicted in Fig. 4 would erroneously add to the approximation of the average slope overload noise power per burst.

(ii) It is known¹² that conditional probability densities must be treated with great caution. M. Kac and D. Slepian in Ref. 12 have illustrated with examples how different the expression for conditional probability densities might be, depending on the way we understand them. From the ensemble viewpoint quantities like the *conditional joint probability density for the rv $X_1 = x(t_0)$ and $X_2 = x(t_0 + s)$ given that $\dot{X}_1 = dx(t_0)/dt = x'_0$* are not clearly defined since the set of sample functions with $dx(t_0)/dt = x'_0$ has probability zero. We can of course give meaning to the conditional densities by means of limiting proce-

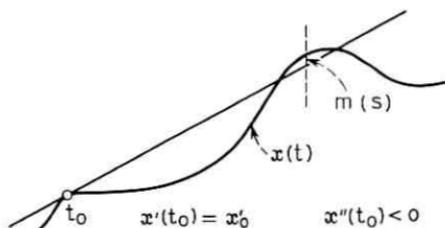


Fig. 4—Consequences of not requiring $x''(t_0)$ to be positive.

dures. As Kac and Slepian point out, a condition like $dx(t_0)/dt = x'_0$ would be replaced by a condition, A , with nonzero probability, depending on parameters, such that when these parameters tend to limiting values A becomes the condition $dx(t_0)/dt = x'_0$. It turns out that, in general, the resulting conditional probability density function depends on the manner in which A approaches the condition $dx(t_0)/dt = x'_0$. Two window conditions are considered below.

(i) A *vertical window* condition is a condition of the form

$$x'_0 < \frac{dx(t_0)}{dt} < x'_0 + \delta. \quad (18)$$

Then, with reference to Fig. 5(a),

$$\begin{aligned} p(x_1, x_2 | \dot{x}(t_0) = x'_0; s)_{vw} \\ = \lim_{\delta \rightarrow 0} \frac{\int_{x'_0}^{x'_0 + \delta} p(x_1, x_2, \dot{x}_1; s) d\dot{x}_1}{\int_{x'_0}^{x'_0 + \delta} p(\dot{x}_1) d\dot{x}_1} = \frac{p(x_1, x_2, x'_0; s)}{p(x'_0)}, \end{aligned} \quad (19)$$

where $p(x_1, x_2, \dot{x}_1; s)$ is the joint probability density function of the random variables $X_1 = x(t_0)$, $X_2 = x(t_0 + s)$ and $\dot{X}_1 = dx_1(t_0)/dt$ and $p(\dot{x}_1)$ is the probability density function of the derivative $\dot{X}_1 = dx(t_0)/dt$. Note that the time argument of the density functions above are written taking into account the stationarity of the input process $\{x(t)\}$.

(ii) A *horizontal window* condition is a condition of the form $dx(t)/dt = x'_0$ for some t such that

$$t_0 \leq t \leq t_0 + \delta.$$

Then,

$$p(x_1, x_2 | x'(t_0) = x'_0; s)_{hw}$$

$$\begin{aligned}
&= \lim_{\delta \rightarrow 0} \left\{ \int_0^{\infty} dx_1'' \int_{x_1' - x_1'' \delta}^{x_1'} p(x_1, x_2, x_1', x_1''; s) dx_1' \right. \\
&\quad + \int_{-\infty}^0 dx_1'' \int_{x_1' - x_1'' \delta}^{x_1'} p(x_1, x_2, x_1', x_1''; s) dx_1' \left. \right\} \\
&\quad \cdot \left\{ \int_0^{\infty} dx_1'' \int_{x_1' - x_1'' \delta}^{x_1'} p(x_1', x_1'') dx_1' \right. \\
&\quad + \left. \int_{-\infty}^0 dx_1'' \int_{x_1' - x_1'' \delta}^{x_1'} p(x_1', x_1'') dx_1' \right\}^{-1} \\
&= \frac{\int_{-\infty}^{\infty} |x_1''| p(x_1, x_2, x_1', x_1''; s) dx_1''}{\int_{-\infty}^{\infty} |x_1''| p(x_1', x_1'') dx_1''}, \tag{20}
\end{aligned}$$

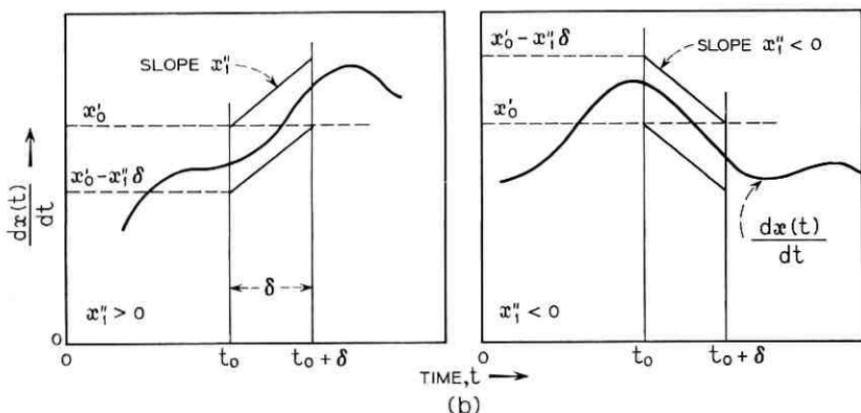
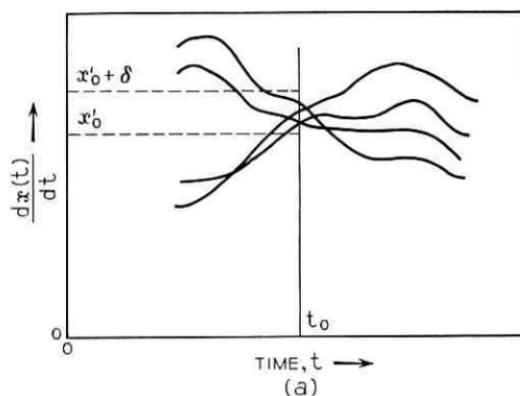


Fig. 5—(a) A “vertical window” condition. (b) A “horizontal window” condition.

where $p(x_1, x_2, x'_1, x''_1; s)$ is the joint probability density function of the random variables $X_1 = x(t_0)$, $X_2 = x(t_0 + s)$, $X'_1 = dx(t_0)/dt$ and $X''_1 = d^2x(t_0)/dt^2$ and $p(x'_1, x''_1)$ the joint probability density of the random variables X'_1 and X''_1 . Equation (20) follows from the fact that the "horizontal window" condition is equivalent (within first order in small quantities and for a given second derivative, say $x''_1 > 0$) to $x'_0 - x''_1 \delta \leq dx(t_0)/dt \leq x'_0$. For $x''_1 < 0$ condition A is satisfied only if $x'_0 \leq dx(t_0)/dt \leq x'_0 - x''_1 \delta$ [see Fig. 5(b)].

Consider now according to Kac and Slepian an "empirical or time derived joint probability density for x_1 and x_2 given that $x'(t_0) = x'_0$ " resulting from taking one sample function of the process and observing the values of $x(t)$ and $x(t + s)$ at each value of t for which $dx(t)/dt = x'_0$ (s is of course a given number). It turns out that the empirical or time derived density thus obtained is equal to the conditional density defined in the horizontal window sense.

Note that if we impose the additional condition $x''(t_0) > 0$ we have

$$p(x_1, x_2 | x'(t_0) = x'_0, x''(t_0) > 0; s)_{hw} = \frac{\int_0^\infty x''_1 p(x_1, x_2, x'_0, x''_1; s) dx''_1}{\int_0^\infty x''_1 p(x'_0, x''_1) dx''_1}. \quad (20a)$$

It will become clearer in a later section where the averaging is done carefully that one should interpret the conditional probability densities in the horizontal window sense.

Zetterberg defines the conditional densities in the integrals of (17) in the vertical window sense; this follows from the way that he computes them.

But let us overlook for a moment these shortcomings of Ref. 7 and continue with the approach presented there. For a Gaussian input process $\{x(t)\}$ Zetterberg derives the following expression for $Q_{x'_0}$.

$$Q_{x'_0} = \sqrt{\frac{2}{\pi}} \int_0^\infty \int_0^\infty k(s) u^2 \exp \left\{ -\frac{1}{2}(u + g(s))^2 \right\} du ds, \quad (21)$$

where

$$k(s) = 2(\psi_0 - \psi(s)) - \frac{1}{b_1} \left\{ \frac{d\psi(s)}{ds} \right\}^2 \quad (22)$$

$$g(s) = \frac{x'_0 s}{\sqrt{k(s)}} \left\{ 1 + \frac{1}{b_1 s} \frac{d\psi(s)}{ds} \right\}. \quad (23)$$

$b_n (n = 1, 2, \dots)$ is defined in (1) and

$$\psi_0 = \psi(0).$$

The following asymptotic expressions are valid (noted in Ref. 7).
For

$$s \rightarrow 0$$

$$k(s) \approx \frac{b_2 s^4}{4} \quad (24)$$

$$g(s) \approx x'_0 s \frac{\sqrt{b_2}}{3b_1}. \quad (25)$$

For

$$s \rightarrow \infty$$

$$k(s) \approx 2\psi_0 \quad (26)$$

$$g(s) \approx \frac{x'_0 s}{\sqrt{2\psi_0}}. \quad (27)$$

(Note the meaning of the symbol \approx as used here:

$$x(s) \approx y(s) \quad \text{for } s \rightarrow s_0$$

if

$$\lim_{s \rightarrow s_0} \frac{x(s)}{y(s)} = 1.)$$

An approximate calculation of the integral for Q_x , as given by (21) is based by Zetterberg on the following simplifications. He uses the asymptotic formula for $g(s)$ for small s . This is a justifiable approximation since the smaller values of $g(s)$ are more important in the evaluation of the integral (21) and in any case the slopes of $g(s)$ for $s \rightarrow 0$ and $s \rightarrow \infty$ do not differ drastically.

For $k(s)$ he sets

$$k(s) = \begin{cases} \frac{b_2 s^4}{4}, & \text{for } s < s_1 \\ 2\psi_0, & \text{for } s > s_1, \end{cases} \quad (28)$$

where s_1 is determined such that

$$\frac{b_2 s_1^4}{4} = 2\psi_0,$$

i.e.,

$$s_1 = \sqrt[4]{\frac{8\psi_0}{b_2}}. \quad (29)$$

The evaluation of the integrals (21) for $Q_{x'}$ are not correct as reported in Ref. 7.* In Appendix A the evaluation of the integral is made and the result is [see (90)]

$$Q_{x'} = \sqrt{\frac{2}{\pi}} \cdot \frac{4}{35} \cdot \frac{b_1^{5/2}}{b_2^{3/2}} \left(\frac{3b_1^{1/2}}{x'_0} \right)^5 A(\lambda), \quad (30)$$

where

$$\lambda = \frac{2}{3} \frac{x'_0}{b_1} \sqrt[4]{\frac{b_2\psi_0}{2}} \quad (31)$$

and

$$A(\lambda) = 1 + P(\lambda)e^{-\lambda^{3/2}} - Q(\lambda)\Phi(\lambda) \quad (32)$$

with

$$P(\lambda) = \frac{17}{24} \lambda^6 + \frac{4}{3} \lambda^4 - \frac{\lambda^2}{2} - 1 \quad (33)$$

$$Q(\lambda) = \frac{17}{24} \lambda^7 + \frac{35}{16} \lambda^5 \quad (34)$$

$$\Phi(\lambda) = \int_{\lambda}^{\infty} e^{-v^{3/2}} dv. \quad (35)$$

For the number $R_{x'}$ both Rice and Zetterberg agree since the formula comes from one of Rice's classic papers;¹³ namely

$$R_{x'} = \frac{1}{\pi} \left(\frac{b_2}{b_1} \right)^{3/2} \exp \left(-\frac{x'^2_0}{2b_1} \right). \quad (36)$$

Therefore, the overload noise according to Zetterberg is

$$N_{0,z} = R_{x'} Q_{x'} = \frac{4\sqrt{2}}{35\pi^{3/2}} \left(\frac{b_1^2}{b_2} \right) \left(\frac{3b_1^{1/2}}{x'_0} \right)^5 A(\lambda) \exp \left(-\frac{x'^2_0}{2b_1} \right). \quad (37)$$

V. OVERLOAD NOISE—THE NEW APPROACH†

In this section we will determine the overload noise using an approach which combines the more accurate model of Zetterberg with the correct averaging procedure given by Rice.

* Zetterberg's expression corresponding to the $A(\lambda)$ given in (32) was not positive for all values of λ — clearly a nonphysical situation.

† In the present section we assume, without loss of generality, $\psi(0) = \sigma^2 = 1$.

This formulation proceeds as follows:

(i) The average noise energy per burst is approximated by $\text{ave} \{ \int_0^\infty m^2(s) ds \}$, as per Zetterberg. This approach avoids Rice's approximation of $n(t)$ during a burst with a third-order polynomial and does not refer to the end point of the burst. On the other hand it yields clearly an upper bound on the overload noise, whereas in Rice's approach the sense of approximation is not clear.

(ii) The averaging process is done the "correct" physical way in the following paragraph. This paragraph is a paraphrasing of the lucid lecture given to us by Rice.

Consider a very long record of the input signal (Fig. 6) of time duration NT , where N is a very large positive integer and T is an extremely long time interval compared with the time unit. Mark on this time record of the input signal all points for which a positive burst begins—all points for which the derivative $dx(t)/dt$ increases through x'_0 . Mark on the record of the signal all time instants s time units following the beginnings of the bursts and measure the value of $m(s)$. Let K be the average number of "positive" bursts per unit time. Then the total number of "positive" bursts in the time interval NT will be: NTK . The average value of $m^2(s)$ over all these positive bursts will be

$$\text{ave} \{ m^2(s) \} = \frac{\sum_{i=1}^{NTK} \{ m_i^2(s) \}}{NTK} \quad (38)$$

Now break up the total signal record into N equal records of duration T and imagine them placed one below the other such that their beginnings lie on the same vertical line as shown in Fig. 6. Divide the time interval into $T/\Delta t$ equal small time intervals of length Δt and imagine vertical lines drawn at the dividing points. Consider a vertical strip of width Δt around time t and sum up the values of $m^2(s)$ over all members of the ensemble that have a "positive" burst which began in the time interval of duration Δt and around the time point $t - s = t_0$, i.e., s time units before t . This sum is independent of the vertical strip we consider and it is denoted by $\sum_{\Delta t} m^2(s)$.

It follows that

$$\sum_{i=1}^{NTK} m_i^2(s) = \frac{T}{\Delta t} \sum_{\Delta t} m^2(s) \quad (39)$$

When a member $x(t)$ is picked at random from the ensemble of the N $x(t)$'s we denote by p the chance that the following three things happen:

(i) A "positive" burst begins in the interval $(t - s, t - s + \Delta t)$ or equivalently the derivative $dx(t)/dt$ increases through x'_0 during $t - s, t - s + \Delta t$.

(ii) The slope of $dx(t)/dt$ at $t_0 = t - s$ lies between x'_1 and $x'_1 + dx'_1$.

(iii) In the time interval $(t, t + \Delta t)$, $m(s)$ lies between $m(s)$ and $m(s) + d(m(s))$. Since $m(s) = x(t) - x(t - s) - x'_0 s$, this is equivalent to asking that $X_1 = x(t - s)$ lie between x_1 and $x_1 + dx_1$ where x_1 is any real number and $X_2 = x(t)$ lie between x_2 and $x_2 + dx_2$ where $x_2 \in (x_1 + x'_0 s, \infty)$.

Then we have

$$p = x'_1 p(x_1, x_2, x'_0, x'_1; s) dx_1 dx_2 dx'_1 \Delta t,$$

where

$$p(x_1, x_2, x'_0, x'_1; s)$$

is the joint probability density function of the random variables

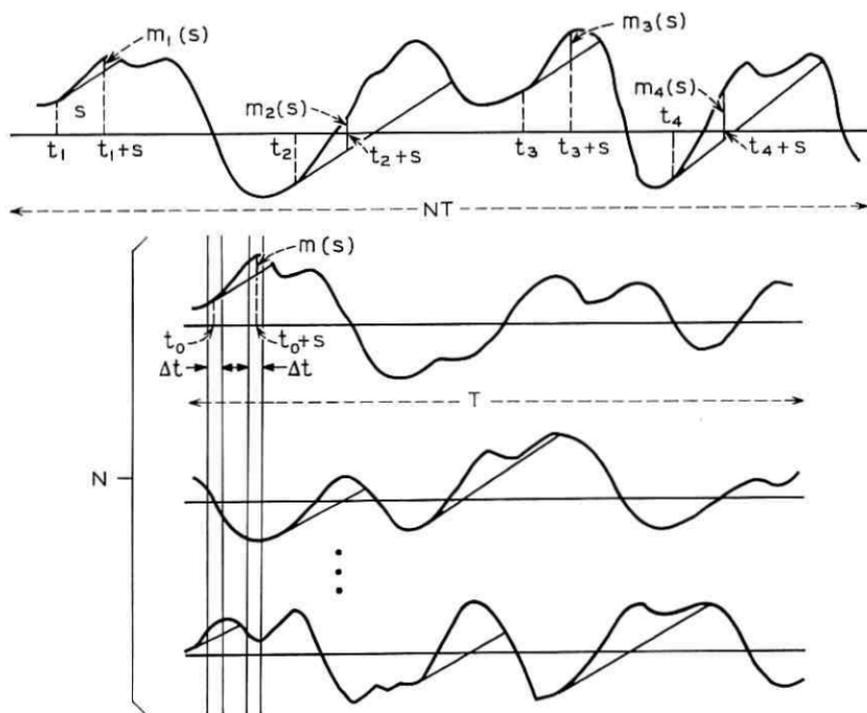


Fig. 6— Illustration of the averaging procedure.

$$\begin{aligned}
 X_1 &= x(t-s) = x(t_0) \\
 X_2 &= x(t) \\
 X_1' &= \left. \frac{dx(t)}{dt} \right|_{t=t_0} \\
 X_1'' &= \left. \frac{d^2x(t)}{dt^2} \right|_{t=t_0} .
 \end{aligned} \tag{40}$$

For an extremely large number N of members of the ensemble of $x(t)$'s the number of members satisfying the three conditions above will be

$$pN = (N\Delta t)x_1''p(x_1, x_2, x_0', x_1''; s) dx_2 dx_1 dx_1'' \tag{41}$$

and therefore,

$$\begin{aligned}
 \sum_{\Delta t} m^2(s) &= \int_0^\infty \int_{-\infty}^\infty \int_{x_1+x_0's}^\infty (x_2 - x_1 - x_0's)^2 (N\Delta t)x_1'' \\
 &\quad \cdot p(x_1, x_2, x_0', x_1''; s) dx_2 dx_1 dx_1'' .
 \end{aligned} \tag{42}$$

Consequently,

$$\begin{aligned}
 \text{ave } \{m^2(s)\} &= \frac{T}{\Delta t} \frac{\sum_{\Delta t} m^2(s)}{NTK} \\
 &= \frac{1}{K} \int_0^\infty \int_{-\infty}^\infty \int_{x_1+x_0's}^\infty (x_2 - x_1 - x_0's)^2 x_1'' \\
 &\quad \cdot p(x_1, x_2, x_0', x_1''; s) dx_2 dx_1 dx_1'' .
 \end{aligned} \tag{43}$$

Make the change of variables

$$x_2 = x_1 + x_0's + u.$$

Then

$$\begin{aligned}
 \text{ave } \{m^2(s)\} &= \frac{1}{K} \int_0^\infty \int_0^\infty \int_{-\infty}^\infty x_1'' u^2 p(x_1, x_1 + x_0's + u, x_0', x_1''; s) \\
 &\quad \cdot dx_1 dx_1'' du.
 \end{aligned} \tag{44}$$

Remark:

Note that¹³

$$K = \int_0^\infty x'' p(x_0', x'') dx'', \tag{45}$$

where $p(x', x'')$ is the joint probability density function of the random variables $X' = dx(t)/dt$ and $X'' = d^2x(t)/dt^2$. Using (20a), therefore, and substituting into (44) we can write

$$\begin{aligned} \text{ave} \{m^2(s)\} &= \int_0^\infty \int_{-\infty}^\infty u^2 p_{hw}(x_1, x_1 + x'_0 s + u \mid x'(t_0) = x'_0, x''(t_0) > 0; s) \\ &\cdot dx_1 du. \end{aligned} \quad (46)$$

Hence, the present "physical" averaging procedure amounts to taking conditional densities in the horizontal window sense.

$$K = \frac{R_{x'_0}}{2} = \frac{1}{2\pi} \left(\frac{b_2}{b_1}\right)^{\frac{1}{2}} \exp\left\{-\frac{x'_0{}^2}{2b_1}\right\}. \quad (47)$$

On the other hand,

$$p(x_1, x_1 + x'_0 s + u, x'_0, x'_1{}'; s) = \frac{1}{(2\pi)^2 |M|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\mathbf{x}'M^{-1}\mathbf{x}\right\}, \quad (48)$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_1 + x'_0 s + u \\ x'_0 \\ x'_1{}' \end{bmatrix}$$

and \mathbf{x}' is the transposed vector. M is the 4×4 cross-correlation matrix: $\{\mu_{ij}\}$, $i, j = 1, 2, 3, 4$ and it is given in Appendix B. $|M|$ is the determinant of M .

After some very lengthy algebraic manipulations which are summarized in Appendix B we find [see (137)]

$$\text{ave} \{m^2(s)\} = \frac{k_1(s)}{\sqrt{2\pi}} \int_0^\infty z^2 \exp\left[-\frac{1}{2}(z + g_1(s))^2\right] \frac{\sqrt{1 - \lambda^2(s)}}{\lambda(s)} \varphi(\xi) dz \quad (49)$$

where $k_1(s)$, $g_1(s)$, and $\lambda(s)$ are complicated functions of s , expressed in terms of the signal autocorrelation function and its derivatives. They are given in Appendix B, (138), (135), and (128), respectively. Note that they do not coincide with Zetterberg's $k(s)$ and $g(s)$ as given by (22) and (23). Other symbols in (49) are defined below.

$$\xi = \frac{\lambda(s)}{\sqrt{1 - \lambda^2(s)}} (z + g_1(s)) \quad (50)$$

$$\varphi(\xi) = e^{-\xi^2/2} + \xi\Phi(-\xi) \quad (51)$$

with

$$\Phi(x) = \int_x^\infty e^{-z^2/2} dz. \quad (52)$$

Consequently,

$$\begin{aligned} Q_{z'} &= \text{ave} \left\{ \int_0^\infty m^2(s) ds \right\} \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty k_1(s) \int_0^\infty z^2 \exp \left[-\frac{1}{2}(z+g_1(s))^2 \right] \frac{\sqrt{1-\lambda^2(s)}}{\lambda(s)} \varphi(\xi) dz ds. \end{aligned} \quad (53)$$

Up to this point we have made no approximations beyond those inherent in the initial model. In the following, additional approximations are required to evaluate (53). In Appendix C it is seen that at $s = 0$, $z = 0$

$$\xi = \xi_0 = \frac{b_2}{\sqrt{b_1 b_3 - b_2^2}} \cdot \frac{x'_0}{\sqrt{b_1}} \quad (54)$$

and for s and z large

$$\xi \cong \gamma_0 z + \delta_0 s,$$

where γ_0 and δ_0 are positive constants defined in Appendix C. The function $\varphi(\xi)$ is plotted in Fig. 7. It is easily seen that, for $\xi > 0$

$$\frac{\varphi(\xi)}{\sqrt{2\pi} \xi} = 1 + \frac{e^{-\xi^2/2}}{\xi \sqrt{2\pi}} \{1 - \xi e^{\xi^2/2} \Phi(\xi)\}.$$

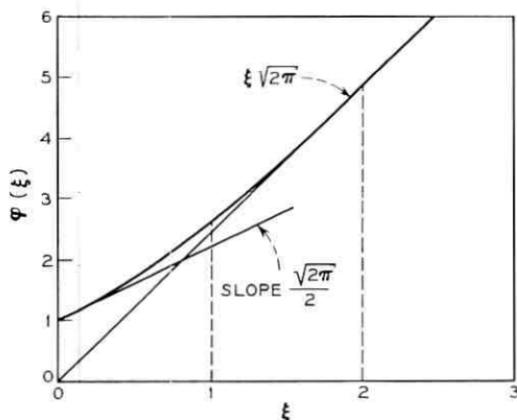


Fig. 7 — The function $\varphi(\xi)$.

For $\xi > 0$ we have

$$\xi e^{\xi^{3/2}} \Phi(\xi) < 1$$

and for ξ large

$$1 - \xi e^{\xi^{3/2}} \Phi(\xi) = \frac{1}{\xi^2} \left(1 - \frac{1}{\xi^2} + \dots \right).$$

Hence, for ξ large

$$\frac{\varphi(\xi)}{\sqrt{2\pi} \xi} = 1 + \frac{e^{-\xi^{3/2}}}{\xi^3 \sqrt{2\pi}} \left(1 - \frac{1}{\xi^2} + \dots \right).$$

The derivative of $\varphi(\xi)$ is very close to $\sqrt{2\pi}$ for large ξ . Namely,

$$\frac{\varphi'(\xi)}{\sqrt{2\pi}} = 1 - \frac{\Phi(\xi)}{\sqrt{2\pi}}.$$

Note also that

$$\frac{\varphi(1)}{\sqrt{2\pi}} \cong 1.08, \quad \frac{\varphi(2)}{2\sqrt{2\pi}} = 1.004, \quad \frac{\varphi(3)}{3\sqrt{2\pi}} \cong 1.0002.$$

Hence, for $\xi \geq 2$ the approximation

$$\varphi(\xi) \cong \xi \sqrt{2\pi} \quad (55)$$

is very good (error less than 1 percent). The approximations hold good even for ξ somewhat larger than 1, as seen in the calculations above. So that if $\xi_0 > 1$, as given by (54), it is justifiable, for the sake of simplicity, to substitute $\xi \sqrt{2\pi}$ for $\varphi(\xi)$ in the integral (53).

Another interesting comment here is that ξ_0 , as given in (54), is equal to the ratio of the absolute value of the mean of the third derivative of the input process $x(t)$ over its standard deviation. Indeed, the mean of $x''(t_0)$, where t_0 is the beginning of a positive burst, is $-b_2 x'_0/b_1$ and the standard deviation $\sqrt{\mathfrak{B}/b_1}$, where

$$\mathfrak{B} = \sqrt{b_1 b_3 - b_2^2}$$

[see Rice's comments above (18) of Ref. 8]. Rice assumed that this ratio is large compared to unity. Here the approximation is good even with ξ_0 close to 1. With the approximation introduced in (55) and using (50) we get

$$\frac{\sqrt{1 - \lambda^2(s)}}{\lambda(s)} \varphi(\xi) \cong (z + g_1(s)) \sqrt{2\pi}$$

and consequently,

$$Q_{x',s} = \int_0^\infty k_1(s) \int_0^\infty z^2(z + g_1(s)) \exp[-\frac{1}{2}(z + g_1(s))^2] dz ds. \quad (56)$$

Integrating in the inside integral by parts we get the simplified expression

$$Q_{x',s} = 2 \int_0^\infty k_1(s) \int_0^\infty z \exp[-\frac{1}{2}(z + g_1(s))^2] dz ds. \quad (57)$$

5.1 Approximate Evaluation of the Noise Energy per Burst

The following asymptotic expressions for $k_1(s)$ and $g_1(s)$ are found in Appendix C, for s small

$$k_1(s) \cong \frac{b_2 s^4}{4} \quad (58)$$

$$g_1(s) \cong \frac{\sqrt{b_2}}{3b_1} x'_0 s \quad (59)$$

and for large s

$$k_1(s) \approx k_\infty = \frac{b_1 \sqrt{2}}{\sqrt{b_2}} \quad (60)$$

$$g_1(s) \approx \frac{x'_0 s}{\sqrt{2}}. \quad (61)$$

The function $g(s)$ has an approximately linear variation for small and large values of s .

To calculate $Q_{x',s}$ according to (57) we will use essentially the same approach used by Zetterberg; namely, use the asymptotic expression for $g_1(s)$ near 0 [see (59)] and for $k(s)$ the expression (58) when $s \leq s_2$ and (60) when $s \geq s_2$. Here, s_2 is the value of s for which the two expressions are equal; namely,

$$s_2 = \sqrt[4]{\frac{4k_\infty}{b_2}} = \frac{2^{5/8} b_1^{1/4}}{b_2^{3/8}} \quad (62)$$

and

$$g_1(s) = \alpha s,$$

where

$$\alpha = \frac{\sqrt{b_2}}{3b_1} x'_0 \quad (63)$$

set

$$\chi = \alpha s_2 = \frac{\sqrt{2}}{3} \left(\frac{2b_2}{b_1^2} \right)^{1/8} \frac{x'_0}{\sqrt{b_1}}. \quad (64)$$

The nature of approximation of the functions $k_1(s)$ and $g_1(s)$ by their values for large and small s is indicated in the Figs. 8 and 9.

The evaluation of the integral (57) for $Q_{x'_0}$ is done in Appendix D and the result is

$$Q_{x'_0} = \frac{\sqrt{2\pi}}{8} \left(\frac{3b_1}{x'_0} \right)^5 b_2^{-3/2} A(\chi), \quad (65)$$

where

$$A(\chi) = 1 - \frac{e^{-\chi^2/2}}{\sqrt{2\pi}} P(\chi) + \frac{1}{\sqrt{2\pi}} \Phi(\chi) Q(\chi) \quad (66)$$

$$P(\chi) = 2 \left(\frac{16}{15} \chi^5 + \frac{1}{3} \chi^3 + \chi \right)$$

$$Q(\chi) = 2 \left(\frac{16}{15} \chi^6 + \chi^4 - 1 \right) \quad (67)$$

$$\Phi(\chi) = \int_{\chi}^{\infty} e^{-z^2/2} dz.$$

The average overload noise power is obtained by multiplying $Q_{x'_0}$ by the average number of bursts per unit time, given approximately in (36).

The average overload noise is, therefore,

$$N_0 = \frac{1}{4\sqrt{2\pi}} \left(\frac{b_1^2}{b_2} \right) \left(\frac{3b_1^3}{x'_0} \right)^5 \exp \left(-\frac{x'_0{}^2}{2b_1} \right) A(\chi), \quad (68)$$

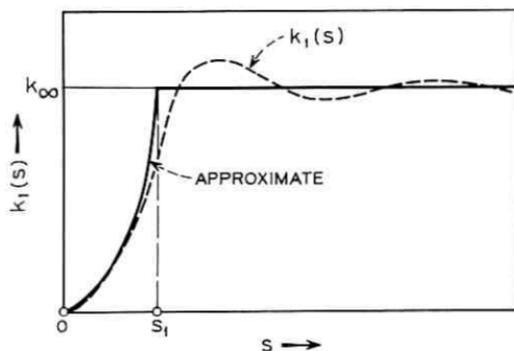


Fig. 8— $k_1(s)$ and the approximation used for the evaluation of $Q_{x'_0}$.

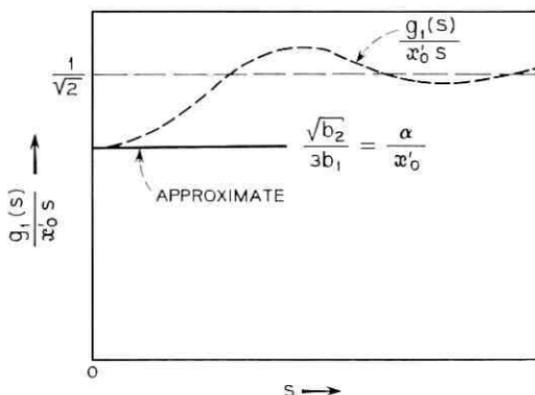


Fig. 9—Approximation to $g_1(s)$ used for the evaluation of $Q_{x'_0}$.

where χ and $A(\chi)$ are given in (64) and (66), respectively. This result is equal to Rice's result [see (22), Ref. 8] times $A(\chi)$. For χ large compared to unity $A(\chi)$ is very close to 1 and thus, in this case (equivalent to x'_0 being large compared to $\sqrt{b_1}$) the two results are identical. This is very interesting when we note that the route taken in the two approaches differ markedly.

The factor $A(\chi)$, for $\chi > 0$ is a positive monotonically increasing function of χ varying between 0 and 1. The function $A(\chi)$ is studied in Appendix E and $-10 \log_{10} A(\chi)$ is plotted in Fig. 10.

VI. COMPARISON WITH COMPUTER SIMULATION AND EXPERIMENTS

The new formula for the average slope overload noise power gives results for both, flat low-pass Gaussian, and band-limited RC Gaussian input signals, that agree in a very satisfactory manner with O'Neal's⁸ computer simulation. For flat low-pass Gaussian input signals we have

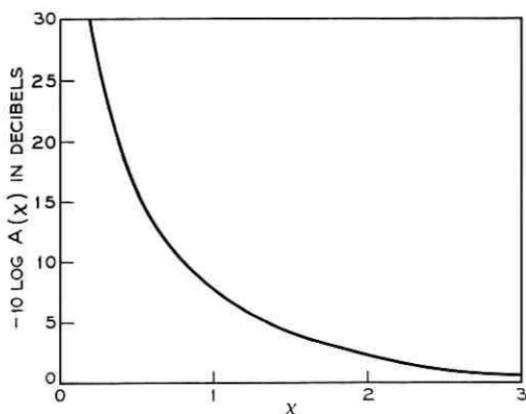
$$b_1 = \frac{(2\pi f_n)^2}{3}$$

$$b_2 = \frac{(2\pi f_n)^4}{5}$$

Using (64) we get, in this case,

$$\chi = \frac{(3.6)^{\frac{1}{2}}}{\pi \sqrt{6}} (kF_s) \cong 0.153(kF_s)$$

so that for $kF_s = 2, 4,$ and 8 we have, respectively,

Fig. 10—The function $-10 \log_{10} A(x)$.

$$\begin{array}{lll} \chi_1 = 0.306 & \chi_2 = 0.612 & \chi_3 = 1.224 \\ A(\chi_1) \cong 5.3 \times 10^{-3} & A(\chi_2) = 4.95 \times 10^{-2} & A(\chi_3) \cong 0.270 \end{array}$$

and the corresponding corrections in O'Neal's curves (Fig. 4 of Ref. 8) would be

$$\begin{array}{l} -10 \log_{10} A(\chi_1) \cong 23 \text{ dB} \\ -10 \log_{10} A(\chi_2) \cong 13 \text{ dB} \\ -10 \log_{10} A(\chi_3) \cong 5.7 \text{ dB.} \end{array}$$

With these significant corrections the present analytical points pass through the computer simulation points, as seen in Fig. 11. Note that the slope overload noise as defined depends only on (kF_s) and not F_s .

Excellent agreement with computer simulation also occurs for RC shaped bandlimited input signals. For RC-shaped signals with spectrum given by (6) of Ref. 8 we have

$$\begin{aligned} b_1 &= \frac{2\pi f_0 \alpha}{\tan^{-1} \left(\frac{2\pi f_0}{\alpha} \right)} - \alpha^2 \\ b_2 &= \frac{(2\pi f_0)^3 \alpha - 6\pi f_0 \alpha^3}{3 \tan^{-1} \left(\frac{2\pi f_0}{\alpha} \right)} + \alpha^4 \end{aligned}$$

so that for $\alpha = 0.25f_0 (= 1/RC)$

$$b_1 \cong 0.94f_0^2$$

$$b_2 \cong 13.2f_0^4.$$

And from (64)

$$\chi \cong 0.744(kF_s)$$

so that for $kF_s = 1$ and 2 we have, respectively, $\chi_1 = 0.744$ and $\chi_2 = 1.488$ yielding a correction to Rice's result of about 10.6 and 4.2 dB, respectively. A comparison with Fig. 5 of Ref. 8, reveals the agreement with computer simulation.

For RC-shaped signals (Gaussian and bandlimited) with $\alpha = 0.068$ [corresponding roughly to the envelope of a black and white entertainment TV signal (FCC standard)]

$$b_1 = 0.267f_0^2$$

$$b_2 = 3.57f_0^4$$

$$\chi = 1.62(kF_s)$$

so that for $kF_s = \frac{1}{4}, \frac{1}{2},$ and 1 the corrections are, respectively, 18.6, 9.7, and 3.4 dB. Good agreement with computer simulation in this case may be noted by applying these corrections to Fig. 6 of Ref. 8.

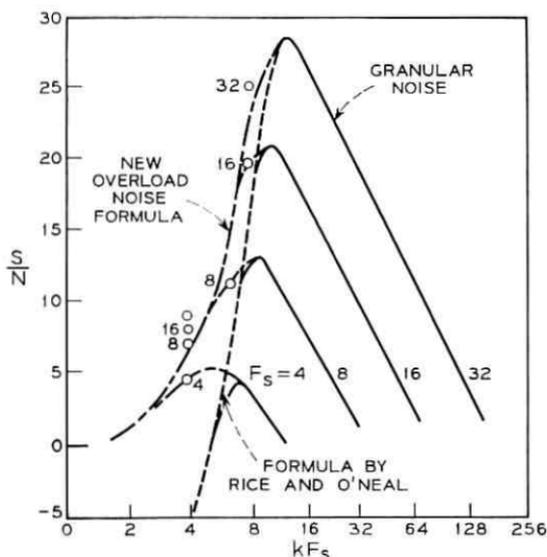


Fig. 11—Flat bandlimited Gaussian signals—comparison of the new results with previous analytic results and computer simulation.

Comparison of the new analytical result with experiment will be covered elsewhere.

VII. OTHER STATISTICAL CHARACTERISTICS OF THE OVERLOAD NOISE

7.1 Probability Density

The technique used in the present paper, i.e., the substitution of $m(s)$ for $n(t)$ and the application of the averaging procedure presented in Section V, can be used for the determination of other statistical characteristics of the slope overload noise.

For example, let $q(m, s, x'_0)$ be the probability density of $m(s)$, where s is a given number, i.e., a parameter, taking on nonnegative values. Let us define the following auxiliary probability functions $q^\pm(m, s, x'_0) ds$, the conditional probability that $x(t) - x(t - s) \mp x'_0 s$ lies between m and $m + dm$ given that the derivative of $x(\cdot)$ increases (decreases) through x'_0 between $t - s$ and $t - s + ds$, where $m \geq 0$ and $s > 0$. Clearly,

$$q^-(m, s, x'_0) = q^+(-m, s, x'_0) \quad \text{for } m < 0.$$

Also using the same averaging procedure as in Section V and the definition of conditional probability densities in the horizontal window sense, we find that [see (20a) and (46)]

$$\begin{aligned} q^+(m, s, x'_0) &= \int_{-\infty}^{\infty} p_{hw}(x_1, x_2 = x_1 + x'_0 s + m \mid x' = x'_0, x'' > 0) dx_1 \\ &= \frac{\int_{-\infty}^{\infty} \int_0^{\infty} x'_1 p(x_1, x_2 = x_1 + x'_0 s + m, x'_0, x'_1; s) dx'_1 dx_1}{\int_0^{\infty} x'_1 p(x'_0, x'_1) dx'_1} \\ &= \frac{1}{K} \int_{-\infty}^{\infty} \int_0^{\infty} x'_1 p(x_1, x_2 = x_1 + x'_0 s + m, x'_0, x'_1; s) dx'_1 dx_1. \end{aligned}$$

From (100) and (101) we see that

$$q^+(m, s, x'_0) = \frac{1}{(2\pi)^{\frac{3}{2}} K \sqrt{a(s)}} P(m, s).$$

$P(\cdot, \cdot)$ is defined in (101) and is determined in (116), Appendix B.

It is easy to verify that

$$q(m, s, x'_0) = \begin{cases} q^+(m, s, x'_0), & \text{for } m > 0 \\ q^-(m, s, x'_0), & \text{for } m < 0. \end{cases}$$

Note also that there is a finite probability that $m(s) = 0$. Hence, the density $q(m, s, x'_0)$ contains an impulse at $m = 0$ with strength $p(s)$

$$p(s) = 1 - 2 \int_0^\infty q^+(m, s, x'_0) dm.$$

The probability density of m , i.e., without specified s , is clearly

$$P_M(m, x'_0) = \int_0^\infty q(m, s, x'_0) ds.$$

Clearly, $P_M(m, x'_0)$ contains an impulse at $m = 0$ of strength

$$\int_0^\infty p(s) ds.$$

7.2 Other Statistical Characteristics

Another useful attribute of the noise is its covariance $\langle x(t)n(t) \rangle_{av}$ with the input random process. This quantity is of interest in comparing results obtained by a particular measured procedure with those obtained analytically. This will be discussed further in the paper referred to previously. The evaluation of $\langle xn \rangle_{av}$ has been performed applying the method presented in Section V. The calculations are even more complicated than the ones employed in the evaluation of $\langle n^2(t) \rangle_{av}$ and we will not consider them here. Moments of any order could be worked out. The expected value of $|n(t)|$ has also been determined. There are many statistical problems that may be generated by the study of slope overload noise in DPCM. These problems have their counter-part in the theory of level-crossings of random processes, but they are even more complicated.

VIII. ACKNOWLEDGMENT

I wish to thank M. R. Aaron, S. O. Rice, and Miss E. G. Cheatham for the assistance provided by them during the development of the present work. M. R. Aaron introduced me to the subject and suggested the problem. In addition, he contributed substantially with insight and ideas and checked the manuscript in all stages of the development of the present work. S. O. Rice gave us a lucid lecture on the averaging procedure. Miss E. G. Cheatham wrote the computer program for the evaluation of $A(\chi)$.

APPENDIX A

Correction of Zetterberg's $Q_{x'}$.

The evaluation of $Q_{x'}$ (Q_{γ} with the notation in equation 4.26 of Ref. 7) is not done correctly in Ref. 7 since $A(x)$ in Equation (4.32) attains negative values. The integral to be evaluated is

$$Q_{x'} = \sqrt{\frac{2}{\pi}} \int_0^{\infty} \int_0^{\infty} k(s)u^2 \exp[-\frac{1}{2}(u + g(s))^2] du ds \quad (69)$$

with

$$(i) \quad g(s) = as, \quad (70)$$

where

$$a = \frac{x'_0 \sqrt{b_2}}{3b_1}$$

[see (25) and the following comments] and

$$(ii) \quad k(s) = \begin{cases} \frac{b_2 s^4}{4}, & \text{for } s \leq s_1 \\ 2\psi_0, & \text{for } s > s_1, \end{cases} \quad (71)$$

where

$$s_1 = \sqrt[4]{\frac{8\psi_0}{b_2}}. \quad (72)$$

Make the change of variable

$$u + as = v.$$

Then

$$Q_{x'} = \frac{1}{a} \sqrt{\frac{2}{\pi}} \int_0^{\infty} \int_0^v k\left(\frac{v-u}{a}\right)u^2 e^{-v^2/2} du dv. \quad (73)$$

Set

$$X(v) = \int_0^v k\left(\frac{v-u}{a}\right)u^2 du = \int_0^v k\left(\frac{z}{a}\right)(v-z)^2 dz \text{ for } v \leq as_1 = \lambda^* \quad (74)$$

$$\frac{X(v)}{2\psi_0/\lambda^4} = Y_1 = \int_0^v (v-z)^2 z^4 dz = \frac{v^7}{105} \quad (75)$$

* λ here corresponds to Zetterberg's x . (λ is introduced to avoid confusion with the input $x(t)$.)

For $v \geq \lambda$

$$\begin{aligned} \frac{X(v)}{2\psi_0/\lambda^4} = Y_2 &= \int_0^\lambda (v-z)^2 z^4 dz + \lambda^4 \int_\lambda^v (v-z)^2 dz \\ &= -\frac{34}{105} \lambda^7 + \frac{\lambda^4 v^3}{3} - \lambda^5 v^2 + \lambda^6 v. \end{aligned} \quad (76)$$

Hence,

$$\begin{aligned} Q_{z'} &= \sqrt{\frac{2}{\pi}} \frac{1}{a} \frac{2\psi_0}{\lambda^4} \left[\int_0^\lambda e^{-v^2/2} \frac{v^7}{105} dv \right. \\ &\quad \left. + \int_\lambda^\infty e^{-v^2/2} \left(\frac{\lambda^4 v^3}{3} - \lambda^5 v^2 + \lambda^6 v - \frac{34}{105} \lambda^7 \right) dv \right]. \end{aligned}$$

Consequently,

$$\begin{aligned} Q_{z'} &= \sqrt{\frac{2}{\pi}} \frac{1}{a} \frac{2\psi_0}{35\lambda^4} \left[\frac{1}{3} J_7(\lambda) + \frac{35}{3} \lambda^4 I_3(\lambda) \right. \\ &\quad \left. - 35\lambda^5 I_2(\lambda) + 35\lambda^6 I_1(\lambda) - \frac{34}{3} \lambda^7 \Phi(\lambda) \right], \end{aligned} \quad (77)$$

where

$$\Phi(\lambda) = \int_\lambda^\infty e^{-z^2/2} dz \quad (78)$$

and

$$I_n(\lambda) = \int_\lambda^\infty z^n e^{-z^2/2} dz \quad (79)$$

$$J_n(\lambda) = \int_0^\lambda z^n e^{-z^2/2} dz. \quad (80)$$

Integrating by parts we find the following recursive relations for I_n and J_n , respectively,

$$I_n(\lambda) = \lambda^{n-1} e^{-\lambda^2/2} + (n-1) I_{n-2}(\lambda). \quad (81)$$

Clearly,

$$I_0(\lambda) = \Phi(\lambda) \quad (82)$$

and

$$I_1(\lambda) = e^{-\lambda^2/2} \quad (83)$$

$$J_n(\lambda) = -\lambda^{n-1} e^{-\lambda^2/2} + (n-1) J_{n-2}(\lambda) \quad (84)$$

with

$$J_0(\lambda) = \int_0^\lambda e^{-z^{2/2}} dz = \frac{\sqrt{2\pi}}{2} - \Phi(\lambda) \quad (85)$$

$$J_1(\lambda) = 1 - e^{-\lambda^2/2}. \quad (86)$$

Applying these recursive relations we find

$$I_2(\lambda) = \lambda e^{-\lambda^2/2} + \Phi(\lambda), \quad (87)$$

$$I_3 = \lambda^2 e^{-\lambda^2/2} + 2e^{-\lambda^2/2}, \quad (88)$$

and

$$J_7 = 48 \left[1 - e^{-\lambda^2/2} \left(\frac{\lambda^6}{48} + \frac{\lambda^4}{8} + \frac{\lambda^2}{2} + 1 \right) \right]. \quad (89)$$

Substituting in (77) we get

$$Q_{x_{\nu'}} = \sqrt{\frac{2}{\pi}} \frac{1}{a} \frac{32}{35} \frac{\psi_0}{\lambda^4} [1 + P_1(\lambda)e^{-\lambda^2/2} - Q_1(\lambda)\Phi(\lambda)], \quad (90)$$

where

$$P_1(\lambda) = \frac{1}{2} \frac{7}{4} \lambda^6 + \frac{4}{3} \lambda^4 - \frac{1}{2} \lambda^2 - 1$$

$$Q_1(\lambda) = \frac{1}{2} \frac{7}{4} \lambda^7 + \frac{3}{16} \lambda^5.$$

APPENDIX B

Algebraic Manipulations with the Statistical Parameters

Denote by

$$M = \{\mu_{ij}\} (i, j = 1, \dots, 4)$$

the cross-correlation matrix of the random variables

$$X_1 = x(t_0)$$

$$X_2 = x(t_0 + s)$$

$$X_1' = \frac{dx(t_0)}{dt}$$

$$X_1'' = \frac{d^2x(t_0)}{dt^2}.$$

Then we have

$$\mu_{11} = E(X_1^2) = 1$$

$$\begin{aligned}
\mu_{12} &= \mu_{21} = E(X_1 X_2) = E(x(t_0 - s)x(t_0)) = \psi(s) \\
\mu_{13} &= \mu_{31} = E(X_1 X_1') = 0 \\
\mu_{14} &= \mu_{41} = E(X_1 X_1'') = (-1)^2 \frac{d^2 \psi(\tau)}{d\tau^2} \Big|_{\tau=0} = -b_1 \\
\mu_{22} &= E(X_2^2) = 1 \\
\mu_{23} &= \mu_{32} = E(X_2 X_1') = -\frac{d\psi(s)}{ds} = -\dot{\psi}(s) \\
\mu_{24} &= \mu_{42} = E(X_2 X_1'') = (-1)^2 \frac{d^2 \psi(s)}{ds^2} = \ddot{\psi}(s) \\
\mu_{33} &= E(X_1'^2) = b_1 \\
\mu_{34} &= \mu_{43} = E(X_1' X_1'') = 0 \\
\mu_{44} &= E(X_1''^2) = b_2.
\end{aligned} \tag{91}$$

Therefore,

$$M = \begin{bmatrix} 1 & \psi(s) & 0 & -b_1 \\ \psi(s) & 1 & -\dot{\psi}(s) & \ddot{\psi}(s) \\ 0 & -\dot{\psi}(s) & b_1 & 0 \\ -b_1 & \ddot{\psi}(s) & 0 & b_2 \end{bmatrix}. \tag{92}$$

Call $|M|$ the determinant of M .

It turns out that

$$|M| = (b_2 - b_1^2) \{b_1(1 - \psi^2(s)) - \dot{\psi}^2(s)\} - b_1 \{\psi(s) + b_1 \ddot{\psi}(s)\}^2. \tag{93}$$

Denote by M_{ij} ($i, j = 1, \dots, 4$) the co-factors of the matrix M . Since M is a symmetric matrix, M^{-1} is also symmetric and $M_{ij} = M_{ji}$ and

$$M^{-1} = \frac{1}{|M|} \{M_{ij}\}. \tag{94}$$

These co-factors are given in terms of the statistics of the input process as follows:

$$\begin{aligned}
M_{11} &= b_1 b_2 - b_1 \dot{\psi}^2(s) - b_2 \dot{\psi}^2(s) \\
M_{12} &= -b_1 (b_2 \psi(s) + b_1 \ddot{\psi}(s)) \\
M_{13} &= -\dot{\psi}(s) (b_2 \psi(s) + b_1 \ddot{\psi}(s)) \\
M_{14} &= b_1 (b_1 + \psi(s) \ddot{\psi}(s) - \dot{\psi}^2(s))
\end{aligned}$$

$$\begin{aligned}
 M_{22} &= b_1(b_2 - b_1^2) \\
 M_{23} &= \psi(s)(b_2 - b_1^2) \\
 M_{24} &= -b_1(\dot{\psi}(s) + b_1\psi(s)) \\
 M_{33} &= (1 - \psi^2(s))(b_2 - b_1^2) - (\dot{\psi}(s) + b_1\psi(s))^2 \\
 M_{34} &= -\psi(s)(\dot{\psi}(s) + b_1\psi(s)) \\
 M_{44} &= b_1(1 - \psi^2(s)) - \dot{\psi}^2(s).
 \end{aligned} \tag{95}$$

It is easily seen that

$$\mathbf{x}' M^{-1} \mathbf{x} = \frac{1}{|M|} (ax_1^2 + 2bx_1 + c), \tag{96}$$

where

$$a = a(s) = M_{11} + 2M_{12} + M_{22}, \tag{97}$$

a function of s only

$$b = (M_{14} + M_{24})x_1' + (M_{12} + M_{22})(u + x_0's) + (M_{13} + M_{23})x_0', \tag{98}$$

a linear function in x_1' and $(u + x_0's)$

$$\begin{aligned}
 c &= M_{44}x_1'^2 + 2[M_{24}(u + x_0's) + M_{34}x_0'] + M_{22}(u + x_0's)^2 \\
 &\quad + 2M_{23}x_0'(u + x_0's) + M_{33}x_0'^2
 \end{aligned} \tag{99}$$

quadratic in x_1' and $(u + x_0's)$.

Integrating with respect to x_1 in (44) we get

$$\begin{aligned}
 &K \text{ ave } \{m^2(s)\} \\
 &= \frac{1}{(2\pi)^{\frac{1}{2}} \sqrt{a(s)}} \int_0^\infty u^2 \int_0^\infty x_1' \exp \left\{ -\frac{1}{2|M|} \left(c - \frac{b^2}{a} \right) \right\} dx_1' du \\
 &= \frac{1}{(2\pi)^{\frac{1}{2}} \sqrt{a(s)}} \int_0^\infty u^2 P(u, s) du,
 \end{aligned} \tag{100}$$

where

$$P(u, s) = \int_0^\infty x_1' \exp \left\{ -\frac{1}{2|M|} \left(c - \frac{b^2}{a} \right) \right\} dx_1'. \tag{101}$$

It is seen that

$$\frac{1}{|M|} \left(c - \frac{b^2}{a} \right) = A(s)x_1'^2 + 2B(s, u, x_0')x_1' + C(s, u, x_0'), \tag{102}$$

where

$$A(s) = \frac{1}{|M|} \left\{ M_{44} - \frac{(M_{14} + M_{24})^2}{a(s)} \right\} \quad (103)$$

is a function of s only and

$$\begin{aligned} B(s) &= B_1(s)(u + x'_0s) + B_2(s)x'_0 \\ &= B_1(s)u + (sB_1(s) + B_2(s))x'_0 \end{aligned} \quad (104)$$

is a linear function in u , with

$$B_1(s) = \frac{1}{|M|} \left\{ M_{24} - \frac{(M_{14} + M_{24})(M_{12} + M_{22})}{a(s)} \right\} \quad (105)$$

$$B_2(s) = \frac{1}{|M|} \left\{ M_{34} - \frac{(M_{14} + M_{24})(M_{13} + M_{23})}{a(s)} \right\}. \quad (106)$$

Further,

$$C(s, u, x'_0) = C_1(s)(u + x'_0s)^2 + 2C_2(s)x'_0(u + x'_0s) + C_3(s)x'_0{}^2 \quad (107)$$

is quadratic in u , where

$$C_1(s) = \frac{1}{|M|} \left\{ M_{22} - \frac{(M_{12} + M_{22})^2}{a(s)} \right\} \quad (108)$$

$$C_2(s) = \frac{1}{|M|} \left\{ M_{23} - \frac{(M_{12} + M_{22})(M_{13} + M_{23})}{a(s)} \right\} \quad (109)$$

$$C_3(s) = \frac{1}{|M|} \left\{ M_{33} - \frac{(M_{13} + M_{23})^2}{a(s)} \right\}. \quad (110)$$

Substituting in (101) we get

$$\begin{aligned} P(u, s) &= \exp \left\{ -\frac{1}{2} \left(C - \frac{B^2}{A} \right) \right\} \\ &\quad \cdot \int_0^\infty x_1'' \exp \left\{ -\frac{A}{2} \left(x_1'' + \frac{B}{A} \right)^2 \right\} dx_1''. \end{aligned} \quad (111)$$

Make the change of variables

$$\sqrt{A} \left(x_1'' + \frac{B}{A} \right) = \eta. \quad (112)$$

Then

$$\begin{aligned}
 P(u, s) &= \frac{1}{A(s)} \exp \left\{ -\frac{1}{2} \left(C - \frac{B^2}{A} \right) \right\} \int_{B/\sqrt{A}}^{\infty} \left(\eta - \frac{B}{\sqrt{A}} \right) e^{-\eta^2/2} d\eta \\
 &= \frac{1}{A(s)} \exp \left\{ -\frac{1}{2} \left(C - \frac{B^2}{A} \right) \right\} \left[e^{-B^2/2A} - \frac{1}{\sqrt{A}} \int_{B/\sqrt{A}}^{\infty} e^{-\eta^2/2} d\eta \right].
 \end{aligned}
 \tag{113}$$

Set

$$\xi = -\frac{B(u, s)}{\sqrt{A(s)}} = -\frac{B_1(s)}{\sqrt{A(s)}} u - \frac{sB_1(s) + B_2(s)}{\sqrt{A(s)}} x'_0
 \tag{114}$$

and

$$\varphi(\xi) = e^{-\xi^2/2} + \xi \Phi(-\xi),
 \tag{115}$$

where

$$\Phi(x) = \int_x^{\infty} e^{-z^2/2} dz.$$

Hence,

$$P(u, s) = \frac{1}{A(s)} \exp \left\{ -\frac{1}{2} \left(C - \frac{B^2}{A} \right) \right\} \varphi(\xi).
 \tag{116}$$

Clearly,

$$\begin{aligned}
 C(u, s) &= C_1(s)(u + x'_0 s)^2 + 2C_2(s)(u + x'_0 s) + C_3(s)x_0'^2 \\
 &= (u\sqrt{C_1(s)} + g^*(s))^2 + x_0'^2 \left[C_3(s) - \frac{C_2^2(s)}{C_1(s)} \right],
 \end{aligned}
 \tag{117}$$

where

$$g^*(s) = x'_0 \frac{sC_1 + C_2}{\sqrt{C_1}}
 \tag{118}$$

and C_1 , C_2 , and C_3 are given in (108), (109), and (110), respectively. Using these equations and the definition of $a(s)$ in (97) we find

$$\begin{aligned}
 |M| \left[C_3(s) - \frac{C_2^2(s)}{C_1(s)} \right] &= M_{33} - \frac{(M_{13} + M_{23})^2}{a(s)} \\
 &\quad - \frac{\{M_{23}(M_{11} + M_{12}) - M_{13}(M_{12} + M_{22})\}^2}{(M_{11}M_{22} - M_{12}^2)a(s)}.
 \end{aligned}
 \tag{119}$$

We also note that

$$\frac{M_{13}}{M_{12}} = \frac{M_{23}}{M_{22}} = \frac{\psi(s)}{b_1} \quad (120)$$

[use (95)].

From (119) and (120) it follows easily that

$$|M| \left[C_3(s) - \frac{C_2^2(s)}{C_1(s)} \right] = M_{33} - \frac{M_{22}\psi^2}{b_1^2}.$$

Substituting the expressions for M_{22} and M_{33} from (95) we find

$$M_{33} - \frac{M_{22}\psi^2}{b_1^2} = \frac{|M|}{b_1}.$$

Hence,

$$C_3(s) - \frac{C_2^2(s)}{C_1(s)} = \frac{1}{b_1}. \quad (121)$$

Set

$$v = u\sqrt{C_1(s)}. \quad (122)$$

Then we have

$$C = (v + g^*(s))^2 + \frac{x_0'^2}{b_1} \quad (123)$$

and from (114)

$$\xi = -\frac{B_1(s)}{\sqrt{A(s)C_1(s)}} \left\{ v + \sqrt{C_1(s)} \left(s + \frac{B_2(s)}{B_1(s)} \right) x_0' \right\}. \quad (124)$$

Using (120), the definitions of $B_1(s)$, $B_2(s)$, $C_1(s)$, and $C_2(s)$ in (105), (106), (108), and (109), respectively, and the relation

$$\frac{M_{34}}{M_{24}} = \frac{\psi(s)}{b_1} \quad (125)$$

resulting from (95) we find

$$\frac{B_2(s)}{B_1(s)} = \frac{C_2(s)}{C_1(s)} = \frac{\psi(s)}{b_1}. \quad (126)$$

Hence, (124) becomes

$$\xi = -\frac{B}{\sqrt{A}} = \lambda(s)(v + g^*(s)), \quad (127)$$

where

$$\lambda(s) = -\frac{B_1(s)}{\sqrt{A(s)C_1(s)}} \quad (128)$$

and g^* is defined in (118). Note also that

$$g^*(s) = \sqrt{C_1(s)} \left(s + \frac{\psi(s)}{b_1} \right) x'_0. \quad (129)$$

From (123) and (125) we get

$$C - \frac{B^2}{A} = (1 - \lambda^2(s))(v + g^*(s))^2 + \frac{x_0'^2}{b_1}. \quad (130)$$

Using the value of K given in (47) we find for the quantity P as given in (116) that

$$\frac{P}{K} = \frac{2\pi}{A(s)} \left(\frac{b_1}{b_2} \right)^{\frac{1}{2}} \exp \left[-\frac{1}{2}(1 - \lambda^2(s))(v + g^*(s))^2 \right] \varphi(\xi). \quad (131)$$

Substituting in (100) and using the change of variable as given in (122) we get

$$\text{ave} \{ m^2(s) \} = \frac{(b_1/b_2)^{\frac{1}{2}}}{\sqrt{2\pi} A(s) \sqrt{a(s)} (\sqrt{C_1(s)})^3} \int_0^\infty v^2 \exp \left[-\frac{1}{2}(1 - \lambda^2(s))(v + g^*(s))^2 \right] \varphi(\xi) dv, \quad (132)$$

where ξ and $\varphi(\xi)$ are given in (127) and (115), respectively. Now make the following change of variables:

$$v \sqrt{1 - \lambda^2(s)} = z \quad (133)$$

and set

$$g^*(s) \sqrt{1 - \lambda^2(s)} = g_1(s) \quad (134)$$

so that from (129)

$$g_1(s) = \frac{x'_0}{b_1} (b_1 s + \psi(s)) \frac{\sqrt{A(s)C_1(s) - B_1^2(s)}}{\sqrt{A(s)}}. \quad (135)$$

Then

$$\xi = \frac{\lambda(s)}{\sqrt{1 - \lambda^2(s)}} (z + g_1(s)) \quad (136)$$

and

$$\text{ave} \{ m^2(s) \} = \frac{k_1(s)}{\sqrt{2\pi}} \int_0^\infty z^2 \exp \left[-\frac{1}{2}(z + g_1(s))^2 \right] \frac{\sqrt{1 - \lambda^2(s)}}{\lambda(s)} \varphi(\xi) dz, \quad (137)$$

where

$$k_1(s) = \left(\frac{b_1}{b_2}\right)^{\frac{1}{2}} \frac{-B_1(s)\sqrt{A(s)}}{\sqrt{A(s)}\{A(s)C_1(s) - B_1^2(s)\}^{\frac{1}{2}}}. \quad (138)$$

Finally, the average energy per burst becomes

$$Q_{r.o.} = \frac{1}{\sqrt{2\pi}} \int_0^\infty k_1(s) \int_0^\infty z^2 \exp[-\frac{1}{2}(z + g_1(s))^2] \frac{\sqrt{1 - \lambda^2(s)}}{\lambda(s)} \varphi(\xi) dz ds. \quad (139)$$

APPENDIX C

Asymptotic Behavior of Several Functions of s for $s \rightarrow 0$ and $s \rightarrow \infty$

Assume that $f^n F(f)$ is integrable for n less than or equal to 8. For bandlimited signals, the usual case in practice, this requirement is automatically satisfied.

For small s the following Taylor expansions hold:

$$\psi(s) = 1 - b_1 \frac{s^2}{2!} + b_2 \frac{s^4}{4!} - b_3 \frac{s^6}{6!} + \frac{s^8}{8!} \psi^{(8)}(\theta s),$$

where θ is a number such that $0 < \theta < 1$ and $\psi^{(8)}(\theta s)$ is the 8th derivative of $\psi(s)$ evaluated at θs .

For a signal bandlimited to the band $(0, f_0)$ we have

$$|\psi^{(8)}(\theta s)| \leq b_4 \leq (2\pi f_0)^2 b_3.$$

Therefore, the absolute value of the remainder term satisfies the following inequality:

$$\left| \frac{s^8}{8!} \psi^{(8)}(\theta s) \right| \leq b_3 \frac{s^6}{6!} \frac{(2\pi f_0 s)^2}{56}$$

so that this term will be negligible if $(2\pi f_0 s)^2 \ll 56$, i.e., if $f_0 s \ll 1.2$. In the expansion for the first and second derivatives of $\psi(s)$ the first three terms are included and the remainder terms may be disregarded for the same values of s .

Consequently, we have

$$\dot{\psi}(s) = -b_1 s + b_2 \frac{s^3}{3!} - b_3 \frac{s^5}{5!} + \frac{s^7}{7!} \psi^{(8)}(\theta_1 s)$$

$$\ddot{\psi}(s) = -b_1 + b_2 \frac{s^2}{2!} - b_3 \frac{s^4}{4!} + \frac{s^6}{6!} \psi^{(8)}(\theta_2 s)$$

with

$$0 \leq \theta_1, \theta_2 \leq 1.$$

We, now, obtain the asymptotic behavior of some expressions ψ which appear in the functions of s involved in the integral for Q_x .

Note that

$$\begin{aligned} (i) \quad 1 - \psi(s) &= b_1 \frac{s^2}{2!} - b_2 \frac{s^4}{4!} + b_3 \frac{s^6}{6!} + 0(s^8) \\ (ii) \quad 1 - \psi^2(s) &= b_1 s^2 - \left(\frac{b_2}{12} + \frac{b_1^2}{4} \right) s^4 + \left(\frac{b_1 b_2}{4!} + \frac{2b_3}{6!} \right) s^6 + 0(s^8) \\ (iii) \quad \psi^2(s) &= b_1^2 s^2 - \frac{b_1 b_2 s^4}{3} + \left(\frac{2b_1 b_3}{5!} + \frac{b_2^2}{(3!)^2} \right) s^6 + 0(s^8) \\ (iv) \quad \check{\psi}(s) + b_1 &= \frac{b_2 s^2}{2} - \frac{b_3 s^4}{24} + 0(s^6) \\ (\check{\psi}(s) + b_1)^2 &= \frac{b_2 s^4}{4} - \frac{b_2 b_3 s^6}{24} + 0(s^8) \\ (v) \quad \check{\psi}(s) + b_1 \psi(s) &= \frac{b_2 - b_1^2}{2} s^2 + \frac{b_1 b_2 - b_3}{4!} s^4 + 0(s^6) \\ (vi) \quad (\check{\psi}(s) + b_1 \psi(s))^2 &= \frac{(b_2 - b_1^2)^2}{4} s^4 + \frac{(b_2 - b_1^2)(b_1 b_2 - b_3)}{4!} s^6 \\ &+ 0(s^8). \end{aligned}$$

Using these formulas and the formulas of definition of the different functions of s we find after a considerable amount of algebraic manipulations, the following asymptotic expressions for $s \rightarrow 0$. Set

$$\mathfrak{B} = b_1 b_3 - b_2^2.$$

$$\begin{aligned} (i) \quad a(s) &\approx \frac{b_2 \mathfrak{B}}{36} s^6 \\ (ii) \quad |M| &\approx \frac{(b_2 - b_1^2) \mathfrak{B}}{36} s^6 \\ (iii) \quad A(s) &\approx \frac{9b_1}{\mathfrak{B} s^2} \\ (iv) \quad B_1(s) &\approx -\frac{18b_1}{\mathfrak{B} s^4} \end{aligned}$$

$$(v) \quad B_2(s) = \frac{\psi(s)}{b_1} B_1(s) \approx \frac{18b_1}{3s^3}$$

$$(vi) \quad C_1(s) \approx \frac{36b_1}{3s^4}$$

$$(vii) \quad C_2(s) = \frac{\psi(s)}{b_1} C_1(s) \approx -\frac{36b_1}{3s^5}$$

$$(viii) \quad A(s)C_1(s) - B_1^2(s) \approx \frac{36b_1}{b_2 3} \frac{1}{s^6}$$

$$(ix) \quad \lambda(s) = -\frac{B_1(s)}{\sqrt{A(s)C_1(s)}} \approx 1$$

$$(x) \quad \frac{\sqrt{1 - \lambda^2(s)}}{\lambda(s)} \approx \frac{1}{3} \sqrt{\frac{3}{b_1 b_2}} \cdot s.$$

Using the formulas above we find that

$$k_1(s) = -\sqrt{\frac{b_1}{b_2}} \frac{B_1(s) \sqrt{A(s)}}{\sqrt{a(s)} \{A(s)C_1(s) - B_1^2(s)\}^2} \approx \frac{b_2 s^4}{4} \quad \text{for } s \rightarrow 0$$

and

$$g_1(s) = x'_0 \left(s + \frac{C_2(s)}{C_1(s)} \right) \frac{\sqrt{A(s)C_1(s) - B_1^2(s)}}{\sqrt{A(s)}} \approx x'_0 \frac{\sqrt{b_2}}{3b_1} s \quad \text{for } s \rightarrow 0,$$

i.e.,

$$g_1(s) \cong \alpha s \quad \text{for } s \text{ small}$$

with

$$\alpha = x'_0 \frac{b_2^{\frac{1}{2}}}{3b_1}.$$

For $s \rightarrow \infty$, $\psi(s)$, $\psi(s)$, and $\check{\psi}(s)$ approach zero.

The following asymptotic expressions are easily derived for $s \rightarrow \infty$:

$$(i) \quad a(\infty) = b_1(2b_2 - b_1^2)$$

$$(ii) \quad |M| = b_1(b_2 - b_1^2)$$

$$(iii) \quad A(\infty) = \frac{2}{2b_2 - b_1^2}$$

$$(iv) \quad B_1(\infty) = -\frac{b_1}{2b_2 - b_1^2}$$

$$\begin{aligned}
 (v) \quad & B_2(\infty) = 0 \\
 (vi) \quad & C_1(\infty) = \frac{b_2}{2b_2 - b_1^2} \\
 (vii) \quad & C_2(\infty) = 0 \\
 (viii) \quad & A(\infty)C_1(\infty) - B_1^2(\infty) = \frac{1}{2b_2 - b_1^2} \\
 (ix) \quad & \lambda(\infty) = \frac{b_1}{\sqrt{2b_2}} \\
 (x) \quad & \frac{\sqrt{1 - \lambda^2(\infty)}}{\lambda(\infty)} = \frac{\sqrt{2b_2 - b_1^2}}{b_1}.
 \end{aligned}$$

Using these expressions we find

$$k_\infty = \lim_{s \rightarrow \infty} k(s) = b_1 \sqrt{\frac{2}{b_2}}$$

and

$$g_1(s) \approx \frac{x'_0 s}{\sqrt{2}} \quad \text{for } s \rightarrow \infty.$$

Note also that for ξ as defined in (50) we have

(i) For $z = 0$ and s small

$$\xi = \xi_0 = \frac{\lambda(s)}{\sqrt{1 - \lambda^2(s)}} g_1(s) \cong 3\sqrt{\frac{b_1 b_2}{\mathfrak{B}}} \frac{1}{s} x'_0 \frac{\sqrt{b_2}}{3b_1} s.$$

Hence,

$$\xi_0 = \frac{b_2 x'_0}{\sqrt{\mathfrak{B} b_1}}.$$

(ii) For s large

$$\xi \cong \frac{\sqrt{2b_2 - b_1^2}}{b_1} z + \frac{\sqrt{2b_2 - b_1^2}}{b_1 \sqrt{2}} x'_0 s \triangleq \gamma_0 z + \delta_0 s.$$

APPENDIX D

Approximate Evaluation of $Q_{z'}$.

From (57) we have

$$Q_{z'}. = 2 \int_0^\infty k_1(s) \int_0^\infty z \exp[-\frac{1}{2}(z + g_1(s))^2] dz ds, \quad (140)$$

where

$$g_1(s) = \alpha s \quad 0 < s < \infty$$

$$k_1(s) = \begin{cases} k_\infty \left(\frac{s}{s_2}\right)^4, & \text{for } s \in (0, s_2) \\ k_\infty, & \text{for } s \in (s_2, \infty). \end{cases} \quad (141)$$

The symbols α , s_2 , and k_∞ are defined in the relations (63), (62), and (60), respectively.

Make the change of variables

$$y = z + \alpha s.$$

We then have

$$Q_{x'} = \frac{2}{\alpha} \int_0^\infty e^{-v^2/2} \int_0^v k_1\left(\frac{y-z}{\alpha}\right) z \, dz \, dy.$$

Set

$$X(y) = \int_0^v z k_1\left(\frac{y-z}{\alpha}\right) dz = \int_0^v (y-\eta) k_1\left(\frac{\eta}{\alpha}\right) d\eta.$$

Then

$$Q_{x'} = \frac{2}{\alpha} \int_0^\infty X(y) e^{-v^2/2} dy.$$

For $y \leq \chi = \alpha s_1$

$$\frac{X(y)}{k_\infty/\chi^4} = Y_1 = \int_0^v \eta^4 (y-\eta) d\eta = \frac{1}{30} y^5.$$

For $y \geq \chi$

$$\frac{X(y)}{k_\infty/\chi^4} = Y_2 = \frac{\chi^6}{30} + \chi^4 \int_\chi^v (y-\eta) d\eta = \frac{8\chi^6}{15} - \chi^5 y + \frac{\chi^4 y^2}{2}.$$

Consequently,

$$\frac{Q_{x'}}{2k_\infty/\alpha\chi^4} = \frac{1}{30} J_6 + \frac{1}{2} \chi^4 I_2 - \chi^5 I_1 + \frac{8}{15} \chi^6 \Phi(\chi), \quad (142)$$

where $\Phi(\chi)$, $I_n(\chi)$, and $J_n(\chi)$ are defined in (78), (79), and (80), respectively.

Applying the recursive relations (81) and (84), we find

$$I_1(\chi) = e^{-\chi^2/2}$$

$$I_2(\chi) = \chi e^{-\chi^2/2} + \Phi(\chi)$$

$$J_6 = \frac{15\sqrt{2\pi}}{2} - e^{-\chi^2/2}(\chi^5 + 5\chi^3 + 15\chi) - 15\Phi(\chi).$$

Substituting in (142) the values of J_6 , I_2 , I_1 , we get

$$\frac{Q_{x'0}}{2k_\infty/\alpha\chi^4} = \frac{\sqrt{2\pi}}{4} \left\{ 1 - \frac{e^{-x^{2/2}}}{\sqrt{2\pi}} P(x) + \frac{1}{\sqrt{2\pi}} \Phi(x)Q(x) \right\}, \quad (143)$$

where

$$P(x) = 2\left(\frac{1}{15}\chi^5 + \frac{1}{3}\chi^3 + x\right) \quad (144)$$

and

$$Q(x) = 2\left(\frac{1}{15}\chi^6 + x^4 - 1\right), \quad (145)$$

i.e.,

$$Q_{x'0} = \frac{\sqrt{2\pi}}{2} \frac{k_\infty}{\chi^4\alpha} A(x),$$

where

$$A(x) = 1 - \frac{e^{-x^{2/2}}}{\sqrt{2\pi}} P(x) + \frac{1}{\sqrt{2\pi}} \Phi(x)Q(x). \quad (146)$$

Using the expressions (63), (64), and (60) for α , χ , and k_∞ , respectively, we get

$$Q_{x'0} = \frac{\sqrt{2\pi}}{8} \left(\frac{3b_1}{x'_0}\right)^5 b_2^{-1} A(x) = (\text{Rice's Results}) \cdot A(x). \quad (147)$$

APPENDIX E

The Function $A(x)$

The function $A(x)$ as defined in (66) is a monotonically increasing function of x in the interval $(0, \infty)$ with $A(0) = 0$ and $A(\infty) = 1$.

The computation of $A(x)$ for different values of x was performed using the computer and $10 \log_{10} 1/A(x)$, the correcting factors of Rice's result, is shown in Fig. 10.

Expanding into Taylor series we can find that for x small

$$A(x) \approx x^4 - 2\sqrt{\frac{2}{\pi}}x^5 + \left(\frac{113}{120\sqrt{2\pi}} + \frac{16}{15}\right)x^6 \cong x^4(1 - 1.6x + 1.44x^2);$$

whereas, for x large, using the asymptotic expansion for

$$\frac{1}{\sqrt{2\pi}} \Phi(x) = \frac{1}{2} \operatorname{erfc} \left(\frac{x}{\sqrt{2}} \right),$$

we get

$$A(x) \approx 1 - \frac{4x^3 e^{-x^2/2}}{5\sqrt{2\pi}} \left(1 - \frac{3}{x^2}\right).$$

REFERENCES

1. Deloraine, E. M., VanMerlo, S. and Derjavitch, B., French Patent No. 932-140, August 10, 1946, p. 140.
2. Phillips, N. V., Gloeilampenfabrieken of Holland, French Patent No. 987,238, applied for May 23, 1949; issued August 10, 1951.
3. Cutler, C. C., Differential Quantization of Communications Signals, U. S. Patent No. 2,605,361, issued July 29, 1952.
4. Proc. IEEE, Special Issue on Redundancy Reduction, 55, No. 3, March, 1967.
5. deJager, F., Delta Modulation, A Method of PCM Transmission Using a 1-Unit Code, Philips Res. Rep. 7, 1952, pp. 442-466.
6. Van de Weg, H., Quantizing Noise of a Single Integration Delta Modulation System with an N-Digit Code, Philips Res. Rep. 8, 1953, pp. 367-385.
7. Zetterberg, L. H., A Comparison Between Delta and Pulse Code Modulation, Ericsson Technics, 11, No. 1, 1955, pp. 95-154.
8. O'Neal, Jr., J. B., Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Input Signals, B.S.T.J., 45, January, 1966, pp. 117-141.
9. O'Neal, Jr., J. B., Predictive Quantizing Systems (Differential Pulse Code Modulation) for the Transmission of Television Signals, B.S.T.J., 45, May-June 1966, pp. 689-721.
10. McDonald, R. A., Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems—Particular Application to Voice Signals, B.S.T.J., 45, September, 1966, pp. 1123-1151.
11. Aaron, M. R., Fleischman, J. S., McDonald, R. A. and Protonotarios, E. N., Delta Modulation Response to Gaussian Inputs—Analytical, Computer, and Experimental Results, to be published.
12. Kac, M. and Slepian, D., Large Excursions of Gaussian Process, Annals Math. Stat., 30, No. 4, December, 1959, pp. 1215-1228.
13. Rice, S. O., Mathematical Analysis of Random Noise, B.S.T.J., 23, 1944, pp. 282-332, and 24, 1945, pp. 46-156.

A Generalized Nyquist Criterion and an Optimum Linear Receiver for a Pulse Modulation System

By D. A. SHNIDMAN

(Manuscript received June 27, 1967)

A pulse modulation system is modeled with M waveforms $\{s_m(t)\}_1^M$, each of which is amplitude scaled and simultaneously transmitted over a single physical channel. An infinite pulse train is assumed with signal interval T , which is determined by bandwidth consideration of the channel. We restrict the receiver to be linear with M outputs, one for each signal waveform.

At a high signal-to-noise ratio the main sources of interference at the input to the receiver are the intersymbol interference and crosstalk; by crosstalk we mean the interference between the different waveforms. It is desirable, therefore, for the receiver to eliminate both types of interference and to minimize the remaining error due to additive noise in the channel. This constraint on the intersymbol interference and crosstalk is defined as the generalized Nyquist criterion.

The receiver which accomplishes the above is determined for a mean square error criterion. Finally, some examples are presented which demonstrate the ease with which the generalized Nyquist criterion can be used to design waveforms without intersymbol interference or crosstalk.

I. THE MATHEMATICAL MODEL

The mathematical model for a pulse modulation system is shown in Fig. 1. The M waveforms $\{s_m(t)\}_1^M$, which are assumed linearly independent and of equal energy, are simultaneously transmitted over a single physical channel. Information is carried on each waveform by amplitude scaling the waveforms $s_m(t)$ by the real numbers $\{a_m\}_1^M$ which are random variables. An infinite pulse train is assumed with signal interval T so that the resulting transmitted waveform is

$$\sum_{n=-\infty}^{\infty} \sum_{m=1}^M a_{nm} s_m(t - nT). \quad (1)$$

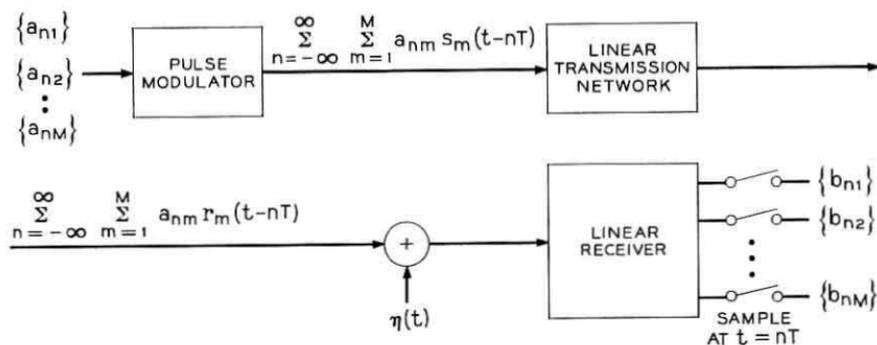


Fig. 1 — Model of the pulse modulation transmission system.

Characterizing the linear time invariant channel by its impulse response, $h(t)$, we define $r_m(t)$ as the convolution of $s_m(t)$ with $h(t)$ so that the received signal waveform is

$$\sum_{n=-\infty}^{\infty} \sum_{m=1}^M a_{nm} r_m(t - nT). \quad (2)$$

To this the channel adds stationary zero mean noise, $\eta(t)$, with correlation function $n(\tau)$ and spectral density $N(f)$. The received waveform is processed by a bank of receivers $\{w_k\}_1^M$ whose M outputs are sampled at times $t = nT$, $n = 0, \pm 1, \pm 2, \dots$ to give b_{nm} which are the estimates of the a_{nm} .

If we consider the set $\{s_m(t)\}_1^M$ with our one physical channel as comprising M different channels then we can refer to the interference of the waveform due to $s_k(t)$ with that of $s_m(t)$ ($m \neq k$) as crosstalk.

Restricting our attention to linear time-invariant receivers then we can characterize the receivers $\{w_k\}_1^M$ by impulse response $\{w_k(t)\}_1^M$ so that the output of the receivers can be expressed as

$$b_k(t) = \sum_{p=-\infty}^{\infty} \sum_{m=1}^M a_{pm} v_{mk}(t - pT) + \int_{-\infty}^{\infty} \eta(x) w_k(t - x) dx, \quad (3)$$

where

$$v_{mk}(t) = \int_{-\infty}^{\infty} w_k(t - x) r_m(x) dx. \quad (4)$$

The sampled outputs are designated by b_{nk} ,

$$b_{nk} = b_k(nT). \quad (5)$$

At high signal-to-noise ratio (S/N) where

$$S/N = \frac{\sum_{m=1}^M v_{mm}^2(0)}{\sum_{k=1}^M \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} n(y-x) w_k(x) w_k(y) dx dy}, \quad (6)$$

the main sources of interference at the input to the receivers are intersymbol interference and crosstalk. It is desirable, therefore, for the receiver to process its input so that the output eliminates intersymbol interference and crosstalk; i.e., that

$$\sum_{p=-\infty}^{\infty} \sum_{m=1}^M a_{pm} v_{mk}(nT) = a_{nk} \quad (7)$$

for all possible sequences of the a_{nm} . This is equivalent to requiring that

$$v_{mk}(nT) = \delta_{mk} \delta_{n0} \quad \begin{array}{l} m, k = 1, \dots, M \\ n = 0, \pm 1, \pm 2, \dots \end{array} \quad (8)$$

where the δ_{ij} are Kronecker delta functions. Further justification for imposing this constraint at high S/N is provided in the Appendix.

We use as our error criterion the mean square error averaged over the receiver outputs

$$J_n = \frac{1}{M} \sum_{m=1}^M E\{(a_{nm} - b_{nm})^2\}, \quad (9)$$

where the expectation is with respect to the random variables a_{nk} and the noise.

We are now in a position to specify the problem concisely: to determine the linear receiver which minimizes the mean square error under the constraint that there be no intersymbol interference or crosstalk.

II. A GENERALIZED NYQUIST CRITERION

A waveform $v(t)$ is said to satisfy the Nyquist criterion¹ for the signal interval T , if

$$v(nT) = \delta_{n0} \quad n = 0, \pm 1, \pm 2, \dots \quad (10)$$

Denoting the Fourier transform of $v(t)$ by $V(f)$ (upper case letters will be used throughout to denote the Fourier transforms of the func-

tions represented by lower case letters), we can state that (10) is true, if and only if,

$$\frac{1}{T} \sum_{\alpha=-\infty}^{\infty} V\left(f - \frac{\alpha}{T}\right) = 1. \quad (11)$$

This is easily shown using Poisson's sum formula (Papoulis)³

$$\frac{1}{T} \sum_{\alpha=-\infty}^{\infty} \Phi\left(f - \frac{\alpha}{T}\right) = \sum_{\alpha=-\infty}^{\infty} e^{-i\alpha 2\pi f T} \phi(\alpha T). \quad (12)$$

If we associate $\phi(t)$ with $v(t)$ then (10) implies and is implied by (11).

Our constraint that the $v_{mk}(t)$ satisfy (8) requires not only that the $v_{mm}(t)$ satisfy (10) but also that the $M(M-1)$ waveforms $v_{mk}(t)$ ($m \neq k$) be zero at $t = nT$. We refer to (8) as the generalized Nyquist criterion. The equation analogous to (11) is

$$\frac{1}{T} \sum_{\alpha=-\infty}^{\infty} V_{mk}\left(f - \frac{\alpha}{T}\right) = \delta_{mk}. \quad (13)$$

This will be used interchangeably with (8) in solving the optimization problem. Since the $\{V_{mk}(f)\}$ can be checked almost by inspection to see if they satisfy (13), the equation is very simple to use.

III. THE CONSTRAINED OPTIMUM RECEIVER

The object of this section is to determine the linear receiver which, subject to the constraints of (8), minimizes the error expression (9). Because of the constraint of (8) we have

$$b_{nk} - a_{nk} = \int_{-\infty}^{\infty} \eta(x) w_k(t-x) dx \quad (14)$$

so that the error becomes

$$J = \frac{1}{M} \sum_{m=1}^M \int_{-\infty}^{\infty} W_m(f) W_m^*(f) N(f) df \quad (15)$$

which is independent of n .

We are now left with the interesting variational problem of minimizing J with respect to all linear receivers $W_k(f)$ such that

$$\frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_m\left(f - \frac{\alpha}{T}\right) W_k\left(f - \frac{\alpha}{T}\right) = \delta_{mk}; \quad (16)$$

i.e., which satisfy the generalized Nyquist constraint. In order to do this, we vary each $W_k(f)$ by an amount $\epsilon \Gamma_k(f)$, where the $\Gamma_k(f)$ must

be such that (16) is still valid. We require

$$\begin{aligned} & \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_m \left(f - \frac{\alpha}{T} \right) \left[W_k \left(f - \frac{\alpha}{T} \right) + \epsilon \Gamma_k \left(f - \frac{\alpha}{T} \right) \right] \\ &= \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_m \left(f - \frac{\alpha}{T} \right) W_k \left(f - \frac{\alpha}{T} \right) + \epsilon \sum_{\alpha=-\infty}^{\infty} R_m \left(f - \frac{\alpha}{T} \right) \Gamma_k \left(f - \frac{\alpha}{T} \right) = \delta_{mk} \\ & \qquad \qquad \qquad k, m = 1, 2, \dots, M \end{aligned} \quad (17)$$

so that $\Gamma_k(f)$ must satisfy the condition

$$\sum_{\alpha=-\infty}^{\infty} R_m \left(f - \frac{\alpha}{T} \right) \Gamma_k \left(f - \frac{\alpha}{T} \right) = 0 \quad m, k = 1, 2, \dots, M. \quad (18)$$

The error with variations becomes

$$\begin{aligned} J(\epsilon) &= \frac{1}{M} \sum_{k=1}^M \int_{-\infty}^{\infty} [W_k(f) + \epsilon \Gamma_k(f)][W_k^*(f) + \epsilon \Gamma_k^*(f)] N(f) df \\ &= \frac{1}{M} \sum_{k=1}^M \int_{-\infty}^{\infty} W_k(f) W_k^*(f) N(f) df \\ & \quad + \frac{\epsilon}{M} \sum_{k=1}^M \int_{-\infty}^{\infty} [\Gamma_k(f) W_k^*(f) + \Gamma_k^*(f) W_k(f)] N(f) df \\ & \quad + \frac{\epsilon^2}{M} \sum_{k=1}^M \int_{-\infty}^{\infty} \Gamma_k(f) \Gamma_k^*(f) N(f) df. \end{aligned} \quad (19)$$

$J(0)$ is minimum if (the $w_k(t)$ are constrained to be real)

$$\int_{-\infty}^{\infty} \Gamma_k(f) W_k^*(f) N(f) df = 0 \quad (\text{for each } k = 1 \dots M), \quad (20)$$

where $\Gamma_k(f)$ must satisfy (18).

In order to solve for $W_k(f)$ we manipulate (20) as follows:

$$\begin{aligned} & \int_{-\infty}^{\infty} \Gamma_k(f) W_k^*(f) N(f) df \\ &= \sum_{\alpha=-\infty}^{\infty} \int_{-1/2T+\alpha/T}^{1/2T+\alpha/T} \Gamma_k(f) W_k^*(f) N(f) df \\ &= \sum_{\alpha=-\infty}^{\infty} \int_{-1/2T}^{1/2T} \Gamma_k \left(f - \frac{\alpha}{T} \right) W_k^* \left(f - \frac{\alpha}{T} \right) N \left(f - \frac{\alpha}{T} \right) df \\ &= \int_{-1/2T}^{1/2T} \sum_{\alpha=-\infty}^{\infty} \left[W_k^* \left(f - \frac{\alpha}{T} \right) N \left(f - \frac{\alpha}{T} \right) \right] \Gamma_k \left(f - \frac{\alpha}{T} \right) df = 0. \end{aligned} \quad (21)$$

Comparing (21) with (18) it can be recognized that (21) is satisfied by a $W_k(f)$ such that

$$W_k(f) = \sum_{c=1}^M \frac{R_c^*(f)}{N(f)} Z_{ck}(f) \quad k = 1, \dots, M, \quad (22)$$

where the $Z_{ck}(f)$ are arbitrary periodic functions of f with period $1/T$.

In order to completely specify $W_k(f)$ we must determine the $Z_{ck}(f)$. Substituting (22) into (16), we obtain

$$\begin{aligned} \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_m \left(f - \frac{\alpha}{T} \right) \sum_{c=1}^M \frac{R_c^* \left(f - \frac{\alpha}{T} \right)}{N \left(f - \frac{\alpha}{T} \right)} Z_{ck} \left(f - \frac{\alpha}{T} \right) \\ = \frac{1}{T} \sum_{c=1}^M Z_{ck}(f) \sum_{\alpha=-\infty}^{\infty} \frac{R_m \left(f - \frac{\alpha}{T} \right) R_c^* \left(f - \frac{\alpha}{T} \right)}{N \left(f - \frac{\alpha}{T} \right)} = \delta_{mk} \end{aligned} \quad (23)$$

$$m, k = 1, 2, \dots, m$$

since $Z_{ck}(f)$ is periodic with period $1/T$. Let

$$L_{mc}(f) = \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_m \left(f - \frac{\alpha}{T} \right) R_c^* \left(f - \frac{\alpha}{T} \right) / N \left(f - \frac{\alpha}{T} \right) \quad (24)$$

then (23) becomes

$$\frac{1}{T} \sum_{c=1}^M L_{mc}(f) Z_{ck}(f) = \delta_{mk} \quad m, k = 1, 2 \dots m \quad (25)$$

or in matrix form

$$L(f)Z(f) = I, \quad (26)$$

where

$$L(f) = [L_{ij}(f)]$$

and

$$Z(f) = [Z_{ij}(f)]$$

are M by M matrices.

Thus, we have, if L is nonsingular for all f , that

$$Z(f) = [L(f)]^{-1} \quad (27)$$

so that $|L| \neq 0$ is a necessary and sufficient condition for a solution to exist. $W_k(f)$ is now completely specified and a realization of the optimum constrained receiver is shown in Fig. 2.

A simple expression for the resulting mean square error is obtained from a manipulation similar to that of (21):

$$J_{\text{opt}} = \frac{T}{M} \sum_{m=1}^M \int_{-1/2T}^{1/2T} Z_{mm}(f) df. \quad (28)$$

IV. EXAMPLES

In this section examples are presented which demonstrate the ease with which the generalized Nyquist criterion can be used to design waveforms without intersymbol interference or crosstalk.

4.1 Example

We start out by making the simplifying assumption that $N(f) = 1$. In addition, if the transmitted waveforms $\{S_m(f)\}_1^M$ are chosen such that $R_m(f)$, where $R_m(f) = S_m(f)H(f)$ satisfy the equation

$$\sum_{\alpha=-\infty}^{\infty} R_m\left(f - \frac{\alpha}{T}\right) R_k^*\left(f - \frac{\alpha}{T}\right) = d_m \delta_{mk}, \quad (29)$$

then a solution exists since the L matrix becomes a diagonal matrix

$$L = Id; \quad d = \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix} \quad (30)$$

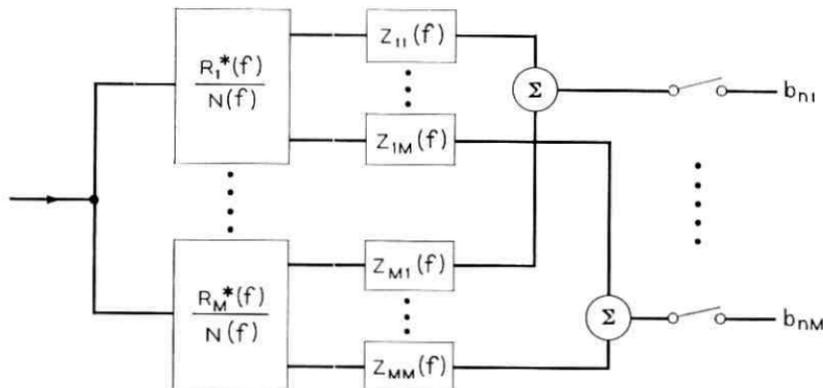


Fig. 2 — A realization of the optimum constrained receiver.

and

$$Z = L^{-1} = \begin{bmatrix} 1/d_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & 1/d_m \end{bmatrix} \quad (31)$$

with the resulting error

$$J = \frac{1}{M} \sum_{m=1}^M \frac{1}{d_m}. \quad (32)$$

Under these conditions the outputs of the matched filters satisfy the generalized Nyquist constraint except for scale factors and the $Z_{cm}(f)$ functions need only perform the appropriate scaling. We consider next two cases where (29) is satisfied.

4.2 Case I

Only the case of $M = 2$ is presented here in detail although other values of M can similarly be handled.

First, note that since matched filters are used the actual phases of the $R_i(f)$ are not important since the output depends only on phase difference between $R_i(f)$ and $R_j(f)$. We use the phase of $R_1(f)$ as a reference phase.

$$R_1(f) = \begin{cases} ce^{j\phi_1(f)}, & |f| \leq 1/T \\ 0, & |f| > 1/T \end{cases}$$

$$R_2(f) = \begin{cases} ce^{j[\phi_1(f) + \Delta\phi(f)]}, & |f| \leq 1/T \\ 0, & |f| > 1/T \end{cases},$$

where $\Delta\phi(f) = \Delta\phi(-f) \pm \pi$ for $|f| \leq 1/T$. The sign is chosen so that $|\Delta\phi(f)| \leq \pi$.

To simplify matters we can choose

$$\Delta\phi(f) = \begin{cases} \pi/2, & f > 0 \\ -\pi/2, & f < 0 \end{cases}$$

so

$$R_1(f)R_2^*(f) = \begin{cases} c^2 e^{-j\Delta\phi(f)}, & |f| \leq 1/T \\ 0, & |f| > 1/T \end{cases}$$

$$= \begin{cases} -jc^2, & 0 \leq f \leq 1/T \\ jc^2, & -1/T \leq f < 0 \\ 0, & \text{elsewhere.} \end{cases}$$

Therefore,

$$\sum_{\alpha=-\infty}^{\infty} R_1\left(f - \frac{\alpha}{T}\right)R_2^*\left(f - \frac{\alpha}{T}\right) = jc^2 - jc^2 = 0 \quad (\text{Fig. 3}).$$

Similarly,

$$\sum_{\alpha=-\infty}^{\infty} R_2\left(f - \frac{\alpha}{T}\right)R_1^*\left(f - \frac{\alpha}{T}\right) = 0$$

and

$$\sum_{\alpha=-\infty}^{\infty} \left| R_m\left(f - \frac{\alpha}{T}\right) \right|^2 = 2c^2 \quad m = 1, 2$$

so

$$L = \begin{bmatrix} 2c^2 & 0 \\ 0 & 2c^2 \end{bmatrix} = 2c^2 I$$

and

$$Z = \frac{1}{2c^2} I$$

$$J_0 = \frac{1}{2c^2}.$$

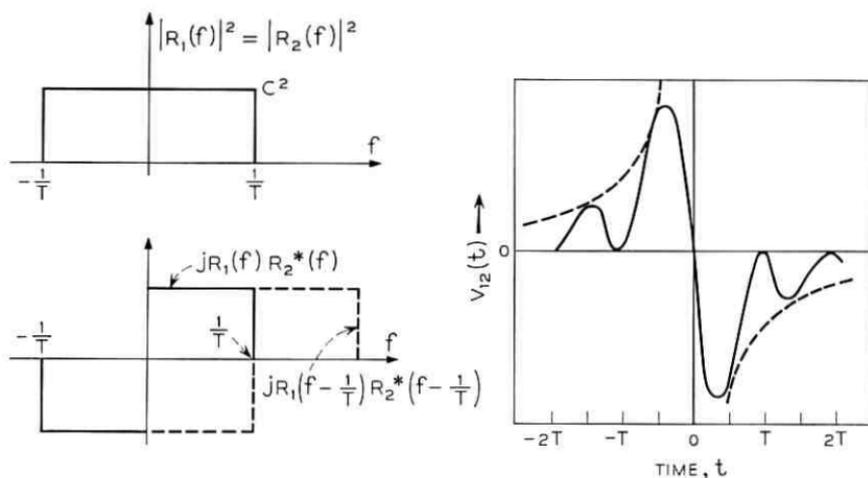


Fig. 3 — Case I transforms at the output of the matched filter.

In the time domain we have

$$v_{12}(t) = c^2 \left(\cos \frac{2\pi t}{T} - 1 \right) / \pi t$$

and it is easily seen that

$$v_{12}(nT) = 0 \quad \text{for } n = 0, \pm 1, \pm 2, \dots$$

4.3 Case II

Consider a set of band-limited frequency multiplexed signals $\{R_m(f)\}_1^M$. The bandwidths are $(1 + \gamma_m + \gamma_{m-1})/T$ where the $\gamma_m (0 \leq \gamma_m \leq 1)$ are parameters associated with the excess rolloff bandwidth, and the signals are separated in frequency by $1/T$ hertz so that the waveforms overlap the adjacent signals only. As in Case I, the actual phases are unimportant because of the matched filters so only phase differences $\phi_m(f)$ from a reference phase $\phi(f)$ will be important.

$$R_m(f) = |R_m(f)| \exp j[\phi(f) + \phi_m(f)]$$

$$R_m(f)R_{m+1}^*(f) = |R_m(f)R_{m+1}(f)| \exp j[\phi_m(f) - \phi_{m+1}(f)].$$

We define roll-off characteristics as a real function $Q_m(f)$ such that

$$Q_m(f) = 0 \quad \text{for } |f| > \frac{\gamma_m}{2T}$$

and

$$Q_m(f) = -Q_m(-f); \quad \text{for } |f| \leq \frac{\gamma_m}{2T}.$$

We can specify the $R_m(f)$ as follows:

$$\begin{aligned} |R_m(f)| &= c_m \sqrt{\text{rect} \left(\frac{|f| - m}{T} \right) + Q_m \left(|f| + \frac{1}{2T} - \frac{m}{T} \right) + Q_{m+1} \left(-|f| - \frac{1}{2T} - \frac{m}{T} \right)} \\ \Delta_m(f) &= \phi_m(f) - \phi_{m-1}(f) \end{aligned}$$

and

$$\Delta_m(f) = \Delta_m(-f) \pm \pi.$$

With the $R_m(f)$ specified it is easily checked that

$$\sum_{\alpha=-\infty}^{\infty} \left| R_m \left(f - \frac{\alpha}{T} \right) \right|^2 = c_m^2$$

$$\begin{aligned}
 R_m(f)R_{m-1}(f) &= |R_m(f)| |R_{m-1}(f)| \exp [-j\Delta_m(f)] \\
 &= c_m c_{m-1} \left\{ B_m \left(f - \frac{m - \frac{1}{2}}{T} \right) \exp [j\Delta_m(f)] \right. \\
 &\quad \left. + B_m \left(f + \frac{m - \frac{1}{2}}{T} \right) \exp [-j\Delta_m(f)] \right\},
 \end{aligned}$$

where $B_m(f)$ is an even real function with bandwidth $2\gamma_{m-1}/T$.

$$B_m(f) = \sqrt{Q_m(|f|)[1 - Q_m(-|f|)]}$$

We can specify $\Delta_m(f)$ as

$$\Delta_m(f) = \begin{cases} \frac{\pi}{2}, & f > 0 \\ -\frac{\pi}{2}, & f < 0 \end{cases}$$

without really restricting ourselves. The resulting $R_m R_{m-1}^*$ is shown in Fig. 4. Looking at Fig. 4, we see by inspection that the $\{R_m\}_1^M$ satisfy (29).

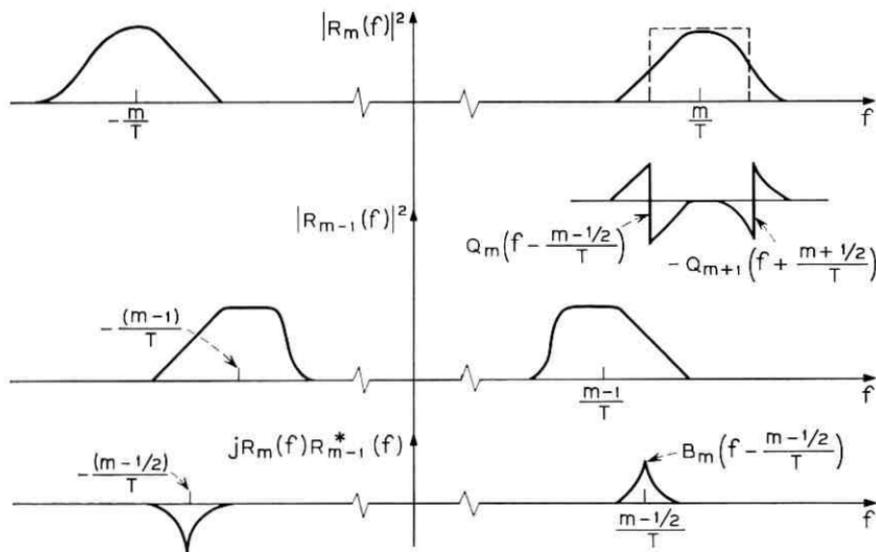


Fig. 4 — Case II transforms at the output of the matched filter.

V. ACKNOWLEDGMENT

The author is indebted to D. W. Tufts for introducing him to the Nyquist problem, and to L. E. Franks for pointing out some interesting aspects of the problem.

APPENDIX

The Optimum Mean Square Error Receiver

In this appendix the optimum mean square error receiver is obtained and it is shown that as $S/N \rightarrow \infty$ this receiver and the optimum constrained receiver of Section III converge to the same receiver when the $\{a_{nk}\}$ are stationary.

The general expression for the mean square error is

$$\begin{aligned}
 J_n &= \frac{1}{M} \sum_{m=1}^M E \left\{ a_{nm}^2 - 2 \sum_{p=-\infty}^{\infty} \sum_{k=1}^M a_{nm} a_{pk} v_{km}(nT - pT) \right. \\
 &\quad + \sum_{p=-\infty}^{\infty} \sum_{k=1}^M \sum_{r=-\infty}^{\infty} \sum_{i=1}^M a_{pk} a_{ri} v_{km}(nT - pT) v_{im}(nT - rT) \\
 &\quad - 2a_{nm} \int_{-\infty}^{\infty} \eta(x) w_m(nT - x) dx \\
 &\quad + 2 \sum_{p=-\infty}^{\infty} \sum_{b=1}^M a_{pk} v_{km}(nT - pT) \int_{-\infty}^{\infty} \eta(x) w_m(nT - x) dx \\
 &\quad \left. + \iint_{-\infty}^{\infty} \eta(x) \eta(y) w_m(nT - x) w_m(nT - y) dx dy \right\} \quad (33) \\
 &= \frac{1}{M} \sum_{m=1}^M \left[\rho_{mm}^{nn} - 2 \sum_{p=-\infty}^{\infty} \sum_{k=1}^M \rho_{mk}^{np} v_{km}(nT - pT) \right. \\
 &\quad + \sum_{p=-\infty}^{\infty} \sum_{k=1}^M \sum_{r=-\infty}^{\infty} \sum_{i=1}^M \rho_{ki}^{pr} v_{km}(nT - pT) v_{im}(nT - rT) \\
 &\quad \left. + \int_{-\infty}^{\infty} N(f) W_m(f) W_m^*(f) df \right],
 \end{aligned}$$

where

$$\rho_{mk}^{np} = E\{a_{nm} a_{pk}\}. \quad (34)$$

Since the $\{a_{nk}\}$ are stationary we can write

$$\rho_{mk}^{n,p} \equiv \rho_{mk}^{(n-p)}. \quad (35)$$

Defining

$$M_{mk}(f) = \sum_{n=-\infty}^{\infty} \rho_{mk}^{(n)} \exp(-j2\pi fnT), \quad (36)$$

then J_n can be written as

$$\begin{aligned} J = \frac{1}{M} \sum_{m=1}^M \left[\rho_{mm}^{(0)} - 2 \sum_{k=1}^M \int_{-\infty}^{\infty} M_{mk}(f) V_{km}^*(f) df \right. \\ \left. + \sum_{k=1}^M \sum_{i=1}^M \int_{-\infty}^{\infty} M_{ki}(f) V_{km}^*(f) \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} V_{im}\left(f - \frac{\alpha}{T}\right) df \right. \\ \left. + \int_{-\infty}^{\infty} W_m(f) W_m^*(f) N(f) df \right] \quad (37) \end{aligned}$$

which is independent of n so the index has been dropped.

Using variational calculus we obtain as a necessary condition on the optimum $\{W_m(f)\}_1^M$ that they satisfy the equations

$$\begin{aligned} \sum_{m=1}^M R_m^*(f) \left[\sum_{i=1}^M M_{mi}(f) \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_i\left(f - \frac{\alpha}{T}\right) W_k\left(f - \frac{\alpha}{T}\right) - M_{mk}(f) \right] \\ + W_k(f) N(f) = 0 \quad k = 1, 2, \dots, M. \quad (38) \end{aligned}$$

The solutions for the $\{W_k(f)\}_1^M$ are

$$W_k(f) = \sum_{c=1}^M \frac{R_c^*(f)}{N(f)} Y_{ck}(f) \quad k = 1, 2, \dots, M, \quad (39)$$

where the $Y_{ck}(f)$ are periodic functions of f with period $1/T$. In order to see that the $\{W_k(f)\}_1^M$ of (39) satisfy (38) for the appropriate determination of the $\{Y_{ck}(f)\}$, substitute (39) for $W_k(f)$ in (38) to obtain

$$\begin{aligned} \sum_{m=1}^M R_m^*(f) \left[\sum_{i=1}^M M_{mi}(f) \frac{1}{T} \sum_{\alpha=-\infty}^{\infty} R_i[f - (\alpha/T)] \right. \\ \left. \cdot \sum_{c=1}^M \frac{R_c^*[f - (\alpha/T)]}{N[f - (\alpha/T)]} Y_{cn}[f - (\alpha/T)] - M_{mk}(f) \right] \\ + \sum_{c=1}^M R_c^*(f) Y_{ck}(f) = 0 \quad k = 1, 2, \dots, M \quad (40) \end{aligned}$$

or since the $Y_{ck}(f)$ are periodic

$$\begin{aligned} \sum_{m=1}^M R_m^*(f) \left[Y_{mk}(f) - M_{mn}(f) + \sum_{i=1}^M M_{mi}(f) \sum_{c=1}^M L_{ic}(f) Y_{ck}(f) \right] = 0 \\ k = 1, 2, \dots, M, \quad (41) \end{aligned}$$

where $L_{ic}(f)$ is as defined in (24).

Defining $M(f)$ and $Y(f)$ as matrices whose elements are, respectively, $M_{ij}(f)$ and $Y_{ij}(f)$ and a column vector $R(f)$ whose elements are $R_m(f)$ (41) can be written as

$$R(f)^T(Y(f) - M(f) + M(f)L(f)Y(f)) = 0, \quad (42)$$

where $L(f)$ is as previously defined. Unless $R(f) = 0$ we require that

$$(I + ML)Y = M \quad (43)$$

$$Y = (I + ML)^{-1}M. \quad (44)$$

With Y so specified (39) satisfies (38) and the resulting mean square error is

$$J_{\text{opt}} = \frac{1}{M} \sum_{m=1}^M \left[\rho_{nm}^{(0)} - \sum_{k=1}^M \int_{-\infty}^{\infty} M_{mk}(f) R_k^*(f) \sum_{c=1}^M \frac{R_c(f)}{N(f)} Y_{cm}^*(f) df \right]. \quad (45)$$

Manipulating as in (21) and using the periodicity of $M(f)$ and $Y(f)$ we then obtain

$$J_{\text{opt}} = \frac{T}{M} \sum_{m=1}^M \int_{-1/2T}^{1/2T} \left[M_{mm}(f) - \sum_{k=1}^M \sum_{c=1}^M M_{mk}(f) L_{kc}^*(f) Y_{cm}^*(f) \right] df. \quad (46)$$

Lastly, recognizing that the integrand is $Y_{mm}^*(f)$ we get

$$J_{\text{opt}} = \frac{T}{M} \sum_{m=1}^M \int_{-1/2T}^{1/2T} Y_{mm}^*(f) df. \quad (47)$$

Finally, we wish to show that the optimum and constrained optimum receivers approach the same limit as $S/N \rightarrow \infty$.

We define U to be the resulting L matrix when the S/N is unity, and we write for any other S/N

$$L = aU, \quad (48)$$

where a is proportional to the signal energy. Since both receivers are of the same form, we need only show that $Y \rightarrow Z$ as $a \rightarrow \infty$.

$$\begin{aligned} Y &= (ML + I)^{-1}M \\ &= (aMU + I)^{-1}M \\ &= [(1/a)U^{-1}M^{-1} - (1/a^2)(U^{-1}M^{-1})^2 + (1/a^3)(U^{-1}M^{-1})^3 - \dots]M \\ &= (1/a)U^{-1} + O(1/a^2), \end{aligned} \quad (49)$$

where $O(1/a^2)$ indicates terms dropping off at least as fast as $1/a^2$. As $a \rightarrow \infty$ the terms of order $1/a^2$ become negligible with respect to the

$1/a$ term. Using the fact that $L^{-1} = 1/a U^{-1}$, we obtain the result

$$\lim_{a \rightarrow \infty} Y = L^{-1} = Z, \quad (50)$$

and the two receivers converge and the constrained optimum is optimum.

REFERENCES

1. Bennett, W. R. and Davey, J. R., *Data Transmission*, McGraw-Hill Book Company, Inc., New York, 1965, pp. 61-63.
2. Tufts, D. W., Nyquist's Problem—The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation, Proc. IEEE, March, 1965.
3. Papoulis, A., *The Fourier Integral and Its Applications*, McGraw-Hill Book Company, Inc., New York, 1962, p. 47.

An Automatic Equalizer for General-Purpose Communication Channels

By R. W. LUCKY and H. R. RUDIN

(Manuscript received June 19, 1967).

The restriction imposed by linear distortion on the flow of information in a communication channel is well known. In the past, the effects of this distortion have been alleviated through the use of manually adjusted equalizing or compensating networks. The adjustment of these networks is too cumbersome a process for the user of a switched communication service to perform each time a new connection is established. Therefore, in present switched networks, control of linear distortion is imposed only on the individual links. Variation between links and variation of the number of links in tandem result in channels with distributed performance. Lower distortion can be achieved by equalizing the overall connection.

Recent developments have made automatic linear distortion removal (equalization) practical for synchronous data communication systems. Here an implementation is described wherein these techniques have been generalized so that automatic equalization can be provided for a communication channel independent of the signal format used in that channel. For a number of applications the speed of automatic equalization makes efficient end-to-end equalization practical in a switched network.

The implementation described affords automatic minimization of the discrepancy between a specified response and the actual response of a linear transmission medium. Thus, on the one hand, it permits the automatic reduction of transmission defects such as signal dispersion and echoes, and, on the other hand, it permits the mechanized synthesis of filters with specified transfer functions.

This paper reviews the general aspects of automatic equalization, describes an implementation of a general purpose automatic equalizer, discusses the theoretical performance of such an equalizer as determined from computer simulations, and lastly presents results for the equalization of real channels using the implementation described.

I. INTRODUCTION

Recent years have witnessed an increasingly intensive investigation of automatic equalization techniques.¹ *Equalization*, itself, is necessary because of the increased demand for efficient use of communication channels. Fixed compromise equalizers have been used in terminal equipment but they cannot remove all of the distortion because of variation between connections in a switched service. Two factors contribute to the distribution of distortion on different connections—differences in the characteristics of the individual links that may be switched together and differences in the number of links in a connection. Better equalization and, therefore, greater transmission efficiency can be achieved by individually equalizing each connection after it has been established. *Automatic* equalization provides a practical means for rapidly and efficiently equalizing each connection.

Several automatic equalization schemes have been published which provide equalization for specific, usually synchronous, communication systems. Some of the techniques for synchronous data transmission systems are those of Coll and George,^{2,3} DiToro,⁴ Funk et al,⁵ and Lucky and Becker et al.^{6,7,8} These techniques are very powerful for the synchronous data transmission systems for which they are intended. Furthermore, the implementations of these equalization strategies possess considerable economy of design because they rely upon the peculiarities of the particular synchronous transmission systems for which they are intended. But, their use is restricted to such systems.

The present paper is concerned with an equalization technique which is essentially independent of the transmission format to be used on the channel. The inclusion of such an equalizer in a communication channel is shown in Fig. 1 in the simplest form. A test signal is transmitted through the channel and the equalizer controller adjusts the equalizer until optimum equalization has been attained. The equalizer adjust-

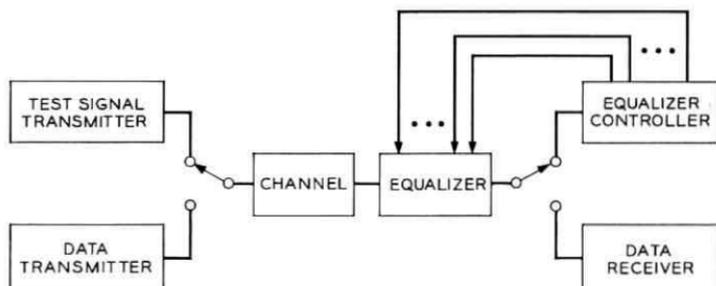


Fig. 1 — Preset mean-square channel equalizer.

ments are then locked and the equalized channel used for communication.

"Optimum" equalization is here defined as the minimization of the mean-squared difference between a specified channel response and the actual equalized channel response. Much work has been done on the problem of optimization under a mean-squared error criterion. The most famous of these is Wiener's classic paper.⁹ A paper by Narendra and McBride¹⁰ also relates to the present work.

In the equalization schemes for synchronous data transmission,²⁻⁸ the data receiver is inside the equalizer control loop and the actual data transmitter is used to generate the test signal. Here, the equalizer can correct for distortion introduced by imperfections in the transmitter and receiver as well as in the channel, resulting in very effective equalization. A general purpose equalizer of the type described does not have this capability (by intent) but instead has the advantage that it is not tied to a single communication system. The equalized channel can be used by arbitrary information transmission systems. The equalization is generally carried out at passband frequencies and the control circuitry could be shared by a number of communication channels. Thus, the technique described may be an attractive one when it is necessary to provide equalization for a variety of customers whose communication channels terminate at a common location.

The equalizer described here uses a transversal filter to operate on the channel response so that the equalized channel response approximates the desired response in an optimum fashion. Again, the criterion used to determine this optimum fashion is the minimization of the mean-squared error.

In summary, this paper reviews some of the general aspects of automatic equalization, describes an implementation of a general purpose automatic equalizer, discusses the theoretical performance of such an equalizer, and presents results for the equalization of real channels using the implementation described. Some laboratory results are also presented for the application of these techniques to a network synthesis problem.

The present paper expands on two previous brief disclosures in the literature.^{11, 12}

II. THE TECHNIQUE

2.1 *The Basic Mathematics*

The notion of the mean-square equalizer starts in the frequency domain. Here, the channel transmission characteristic is equalized so

that it best resembles the ideal transmission characteristic. This "best" fit is made using a mean-square error criterion. Thus, the distortion to be minimized is

$$E_1 = \int_{-\infty}^{\infty} |H(\omega) - G(\omega)|^2 d\omega, \quad (1)$$

where $H(\omega)$ is the equalized channel characteristic and $G(\omega)$ is the ideal channel characteristic. Notice that this error criterion includes both phase (and consequently delay) and amplitude information in the goodness of fit.

The error criterion given in (1) can be made more general by adding information concerning the relative importance of errors at various frequencies. For example, in most information transmission schemes the major portion of the signal energy is placed near the center of the band, so that the equalization should be most perfect there. Since relatively little signal energy is put near the band edges, the quality of equalization is not of as great concern in this region. Therefore, a real, nonnegative weighting function $|W(\omega)|^2$, which assigns a relative weight $|W(\omega)|^2$ to the equalization error at each frequency ω , is included in the criterion. The resultant criterion is

$$E = \int_{-\infty}^{\infty} |H(\omega) - G(\omega)|^2 |W(\omega)|^2 d\omega. \quad (2)$$

Usually the ideal characteristic $G(\omega)$ would have flat amplitude and linear phase within the band of interest, while the spectral weighting function $|W(\omega)|^2$ would resemble the spectral density of the signal likely to be transmitted, if this spectral density is known beforehand. The system is shown in block diagram form in Fig. 2.

The equalized channel characteristic is the product of the unequalized channel characteristic $X(\omega)$ and the equalizer characteristic $C(\omega)$.

$$H(\omega) = X(\omega)C(\omega). \quad (3)$$

The frequency characteristic function of a $(2N+1)$ - tap transversal equalizer with tap gains c_n , $n = -N, \dots, N$ spaced at τ second intervals is

$$C(\omega) = \sum_{n=-N}^N c_n e^{-in\omega\tau}. \quad (4)$$

Notice that this response is periodic with period $2\pi/\tau$, the real part of the response being even about frequencies $2n\pi/\tau$ and the imaginary

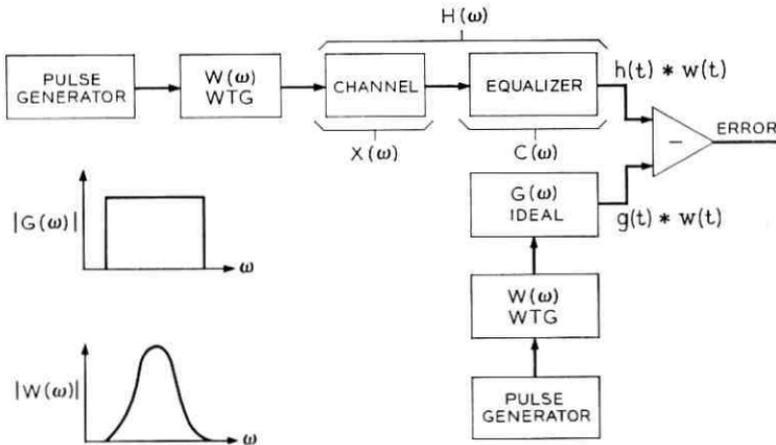


Fig. 2 — Mean-square equalizer.

part odd about these frequencies. Thus, the transversal equalizer offers independent control of the overall frequency response of the system over only one of the frequency intervals $n\pi/\tau \leq \omega \leq (n + 1)\pi/\tau$. The value of the tap spacing τ must be picked such that the desired equalization frequency range is included in one of these intervals. A frequent case is that where the channel is essentially low-pass in nature. Here the tap separation τ will be the Nyquist period $1/2W$, where W is the highest frequency of interest. For bandpass channels, τ will generally have to be less than a Nyquist period.

The objective is the minimization of the distortion E as a function of the $(2N + 1)$ variables c_n in automatic fashion. Because this minimization is more easily carried out in the time domain, Parseval's theorem is used to obtain an equivalent form for (2):

$$E = \int_{-\infty}^{\infty} \{|h(t) - g(t)| * w(t)\}^2 dt. \tag{5}$$

In (5), $h(t)$, $g(t)$, and $w(t)$ are the impulse responses corresponding to the frequency responses $H(\omega)$, $G(\omega)$, and $W(\omega)$, respectively, and the * symbol is used to represent convolution.

If $x(t)$ is the impulse response of the unequalized channel, the equalizer output response is

$$h(t) = \sum_{n=-N}^N c_n x(t - n\tau) \tag{6}$$

and (5) can be written

$$E = \int_{-\infty}^{\infty} \left\{ \sum_{n=-N}^N c_n x(t - n\tau) * w(t) - g(t) * w(t) \right\}^2 dt. \quad (7)$$

It can easily be demonstrated that E is a convex function of the tap gains c_n ; $n = -N, N$. Thus, there is a single minimum of E and this occurs when the $(2N+1)$ derivatives $\partial E / \partial c_n$ are zero. Setting these derivatives to zero gives $(2N+1)$ simultaneous linear equations which can be solved to effect a minimization of E . If the partial differentiation is carried out with respect to a particular tap setting (say c_j), the following relation is obtained:

$$\frac{\partial E}{\partial c_j} = 2 \int_{-\infty}^{\infty} \{h(t) * w(t) - g(t) * w(t)\} \{x(t - j\tau) * w(t)\} dt, \quad -N \leq j \leq N. \quad (8)$$

The set of (8) contains all the information required for automatic optimization. First, if these equations are set equal to zero and solved for the c_n 's, the desired tap coefficients are obtained. Second, if arbitrary values are chosen for the c_n 's, the set of (8) dictates the direction in which the coefficients must be changed to reduce the error E . Further, a comparison of the set of (8) with Fig. 3 yields a technique which facilitates the calculation of the partial derivatives which, in

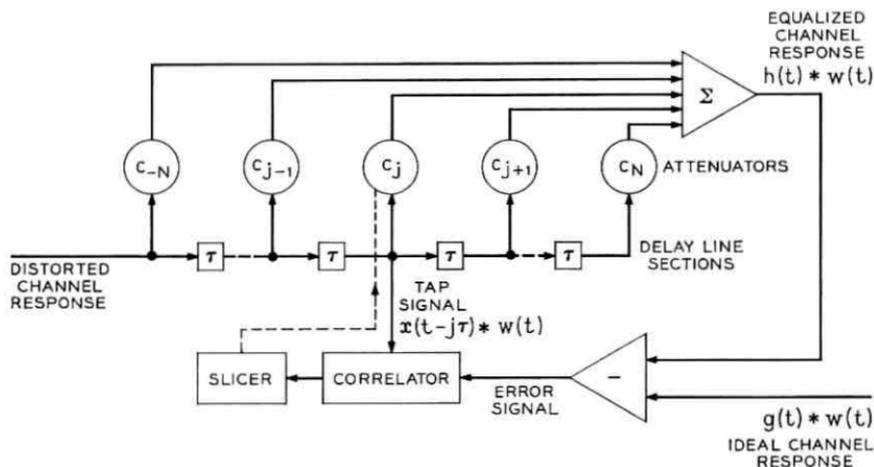


Fig. 3—Mean-square transversal equalizer.

turn, provides the basis for an algorithm for automatic equalization under the mean-square error criterion.

2.2 Basic Implementation

The tapped-delay line structure which forms the basis of a transversal filter is shown in Fig. 3. The "attenuators" have the capability of supplying both positive and negative weights. The first term in the set of (8) is simply the error signal, or the difference between the equalized channel response and the ideal channel response. The second term, which multiplies the first, is the signal at the j th tap when (8) is written for $\partial E/\partial c_j$. Thus, the partial derivative of the distortion with respect to a particular attenuator setting is given by the time-integral of the product of the error signal and the signal at the particular tap being considered. In other words, the partial derivative is given by the cross-correlation of the error signal with the tap signal.

Coincident with the start of the equalization process, the various cross-correlation coefficients for all of the delay-line taps are calculated by the correlators. The polarity of a particular cross-correlation coefficient indicates the polarity of the partial derivative of the distortion with respect to the corresponding tap weighting coefficient. Because of the convexity of the criterion this polarity information indicates the direction in which the tap weight must be changed to reduce the distortion. When all cross-correlation coefficients become zero, no further adjustment of the weights can lower the distortion and the desired equalization is achieved.

Some feeling for the algorithm can be obtained from the following argument. The signals at the various taps contribute to the error signal in linear fashion. The best that the equalizer can expect to achieve is the elimination of any systematic contribution between the tap signals and the error signal. Under a mean-squared error criterion the measurement of such a systematic contribution is cross-correlation. When all the cross-correlation coefficients are zero, nothing further can be done to reduce the error.

2.3 Related Applications

In the course of equalization, an automatic equalizer is called upon to perform a network synthesis. Specifically, it synthesizes that network within its repertoire which results in the minimum mean-squared error. It is possible to use the automatic equalizer simply as an automatic network synthesizer.

The distinction between these two cases (channel equalization and network synthesis) is made in Fig. 4. Fig. 4(a) shows the conventional application of the equalizer wherein the equalizer strives to first determine and then synthesize the function

$$C(\omega) \cong 1/X(\omega), \quad (9)$$

where $X(\omega)$ is the frequency response function of the distorting channel. If the equalizer could perfectly synthesize $1/X(\omega)$ (plus an arbitrary flat time delay) the distortion would be completely removed. The use of the equalizer for network synthesis is shown in Fig. 4(b). Here, the transversal filter with complex frequency response $C(\omega)$ strives to approximate $A(\omega)$ directly so that the quantity

$$E = \int_{-\infty}^{\infty} |A(\omega) - C(\omega)|^2 d\omega \quad (10)$$

is minimized. As in the case of channel equalization the error can be given a frequency sensitive weighting, $W(\omega)$.

So far, the discussion has centered upon an equalizer of the transversal filter type (as in Fig. 3). This is by no means the only possibility, and a more general equalizer/synthesizer is shown in Fig. 5. The common ground shared by the schemes (as shown in Fig. 3 and 5) is that both rely upon the sum of weighted responses. The parallel net-

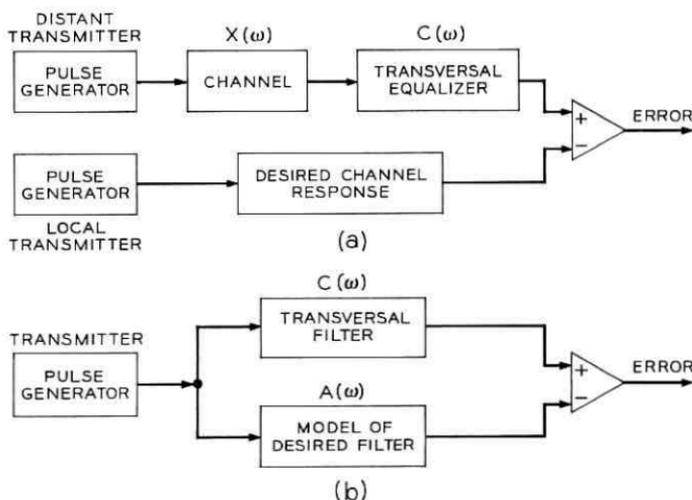


Fig. 4 — (a) Channel equalization. (b) Filter synthesis.

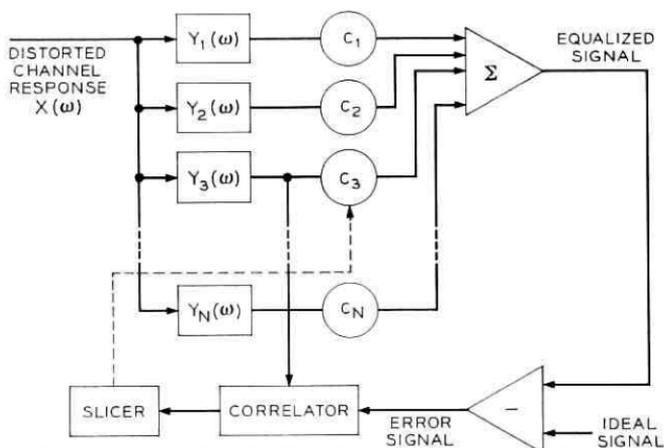


Fig. 5—Generalized mean-square equalizer/synthesizer.

work layout of Fig. 5 has greater flexibility than the series network layout of Fig. 3. The series layout has the advantage that much of the filtering necessary for one particular response is performed by the preceding networks.

For practical reasons only it is required that the responses of the various networks shown in Fig. 5 be linearly independent; it is desirable (but by no means necessary) to have network responses orthogonal to each other so as to minimize the interaction between the setting of the various weighting coefficients. The desirable orthogonality results when (11) is satisfied.

$$0 = \int_{-\infty}^{\infty} |X(\omega)|^2 Y_i(\omega) Y_j(\omega) d\omega, \quad i \neq j. \quad (11)$$

In (11) the $Y_i(\omega)$ are the transfer functions of the various networks and $X(\omega)$ the Fourier transform of their common input. A discussion of various sets of such orthogonal networks may be found in Lee.¹³

If $X(\omega)$ is constant from dc to f_1 Hz and if the taps on a delay line are spaced at $1/2f_1$ second intervals, the desired (but again not necessary) orthogonality is obtained. In the case of the equalization of a communication channel, orthogonality can not usually be obtained. Here $X(\omega)$ is affected by the amplitude response of the distorting channel and this of course is unknown, *a priori*.

An application closely related to network synthesis is that of a

relatively new technique for echo suppression: echo cancellation. This problem most commonly occurs in long-haul voice communication. Here the possibility of an improperly terminated hybrid makes undesirable returned echoes probable. These echoes are generally dispersed in time by the transmission medium. Previous techniques have introduced attenuation into the echo path. The recently developed technique uses, instead, principles identical to those developed here to generate a replica of the echo. The actual echo and its replica are then added together in such a fashion that they cancel. This can be achieved automatically and adaptively as discussed in Refs. 14 through 17.

2.4 *Performance in the Presence of Noise*

In the network synthesis problem, the environment is largely under the control of the designer and as a result noise represents a negligible problem. This is not the case for equalization, where noise is definitely to be reckoned with. Noise effectively alters the equalized channel's frequency characteristic. It will be shown in what follows that the change in the frequency characteristic is a desirable one, i.e., the total mean-square error is minimized.

Noise also increases the settling time in a very complicated fashion. However, in the implementation discussed, this increase is very small and for that reason will not be further discussed here.

2.4.1 *The Mean-Square Criterion in a Noisy Environment*

In the process of equalizing a communication channel to approach the desired flat amplitude and linear phase-frequency responses, care must be taken that the noise in the channel is not increased to harmful levels. Ideally, when noise is present the equalizer should minimize the average total error consisting of both the component resulting from the imperfect channel frequency characteristic and the component resulting from noise. If the spectral weighting function $W(\omega)$ is chosen properly, the equalizer described here attains this objective. The noise in the channel is assumed to be the same during and after equalization. It will be shown that the proper choice (in the sense above) is a $W(\omega)$ function which makes the equalizer test signal's power spectrum duplicate the information signal's power spectrum.

Consistent with the notation used previously, let the channel [with impulse response $x(t)$] be used to transmit information $w(t)$. (The square of the amplitude frequency response of the error weighting

filter is thus picked to be identical with the power spectral density of the transmitted signal.) The received noise $\eta(t)$ will be taken as a sample from a stationary random process. Thus, the received signal, $y(t)$, is given by

$$y(t) = w(t) * x(t) + \eta(t). \quad (12)$$

The error criterion E_n is again taken as the average mean-square error between the equalized received signal $h(t)$ and the transmitted signal passed through the ideal, noiseless channel $G(\omega)$.

$$E_n = \langle [h(t) - w(t) * g(t)]^2 \rangle. \quad (13)$$

The brackets $\langle \rangle$ denote a time average. The equalized signal $h(t)$ is given by

$$h(t) = \sum_{n=-N}^N c_n [\eta(t - n\tau) + w(t) * x(t - n\tau)] \quad (14)$$

using the transversal filter equalizer of Fig. 3. As before, the partial derivatives of the distortion are computed with respect to the various tap gains c_j .

$$\frac{\partial E_n}{\partial c_j} = 2 \langle [h(t) - w(t) * g(t)] [\eta(t - j\tau) + w(t) * x(t - j\tau)] \rangle. \quad (15)$$

When this relation is compared with Fig. 3, it is seen that the expected value of the output signal of the cross-correlator is given precisely by (15). Thus, the equalizer does minimize the total expected mean-squared error in the presence of noise. Again, this is true provided that the test signal used for purposes of equalization has a spectral density identical to that of the signal to be transmitted over the equalized channel.

Often the power spectrum of the information transmission signal is not known beforehand. In this instance a flat weighting can be used. Examples of the effect of various weighting functions are given in Section IV.

III. IMPLEMENTATION

This section is devoted to the description of an implementation of a general-purpose automatic equalizer. The discussion of the implementation will be broken down into three parts: The automatic

transversal filter itself, carrier recovery, and timing recovery. The first part is of general interest and pertinent to both the problems of automatic equalization and network synthesis; the second and third parts are peculiar to the equalization process. The need for carrier recovery arises because of the incidental modulation which can occur in some transmission channels (notably those involving carrier facilities). Timing is needed to ensure the proper synchronization of the desired signal generated at the equalizer with the signal received from the distant transmitter.

Throughout this discussion, reference will be made to Fig. 6.

3.1 Implementation of the Automatic Transversal Filter

Many sets of functions could have been used as the basis functions for the equalizer. Only one set, the set of functions generated at regu-

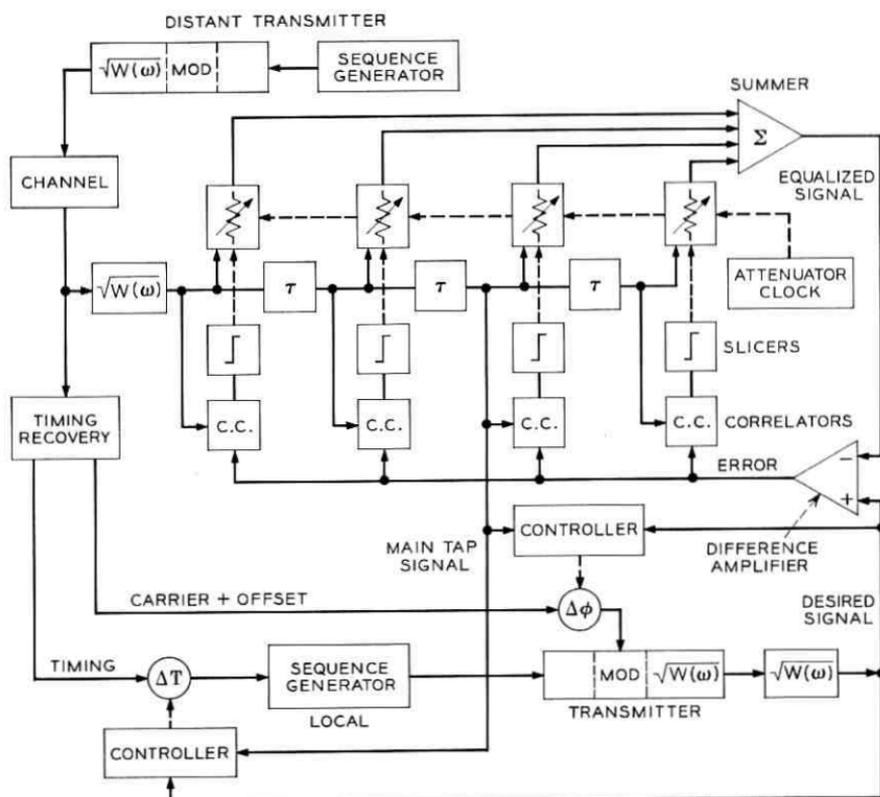


Fig. 6 — Equalizer block diagram.

larly spaced taps on a delay line, will be discussed here. There are two reasons for this, the first being that the transversal filter has been found to be a reasonably efficient means for the removal of distortion and the second being that considerable experience with the use of tapped delay lines is available.¹

In the selection of an appropriate delay line, three parameters must be established: bandwidth, tap spacing, and number of taps. Because the equalization is carried out at passband, it is clear that the usable bandwidth of the delay line must be at least coincident with the channel's bandwidth. Often, as in a telephone voice channel, the passband extends sufficiently close to dc that it is reasonable to use a line which provides delay from dc to the upper frequency limit of the passband.

The tap spacing has already been touched on in Section 2.1. For the case just mentioned, the tap spacing τ was chosen equal to the reciprocal of twice the upper band-edge frequency, thus making the tap spacing slightly smaller than the Nyquist interval. The alternative in this case would be separating the taps by the Nyquist interval and providing additional, frequency-independent phase shifting networks at the various taps. This is equivalent to translating the passband into a comparable low-pass channel, equalizing, and retranslating to passband.

The number of taps necessary depends on the nature and degree of the dispersion (or distortion) likely to be found in the channel and on the precision of the equalization desired. A very rough approximation can be obtained from paired-echo theory.¹⁸ This estimate equates the necessary number of taps to four times the number of cycles in the highest frequency Fourier series component needed to represent the distorting frequency characteristic function. The accurate determination of the necessary number of taps can be made only by case-by-case calculation. Examples showing the effect of a varying number of taps will be given later.

The attenuators associated with each tap on the delay line are capable of providing both positive and negative weights to the tap signals. The attenuators are controlled by digital counters composed of a number of binary memory elements. These are connected in such a fashion that the total count can be increased or decreased by one at any time and are therefore given the name up-down counters. All the attenuators are changed at the same time by a common clock. The outputs of the binary elements control the solid-state switching of constant-resistance ladder networks. A full count corresponds to a

normalized tap weight of +1, a zero count to -1, and a half-full count to 0. Each attenuator is thus a kind of granular potentiometer with constant increments. The number of increments is determined by the number of binary elements and for K elements is equal to 2^K .

The number of steps in the attenuator and the relative ranges of the attenuator determine an upper bound on the accuracy of the equalizer. Taking into account the required polarity information and assuming the attenuator settings to be off by half an increment, the accuracy to which an attenuator may be set is $1/(2)^K$. If there are $(2N+1)$ taps, then the maximum signal-to-noise ratio (considering the residual distortion as noise) attainable is

$$(S/N)_{RES} = 10 \log_{10} \frac{(2)^{2K}}{(2N+1)} \text{ dB} \quad (16)$$

or

$$(S/N)_{RES} \cong 6K - 10 \log_{10} (2N+1) \text{ dB.} \quad (17)$$

In the implemented equalizer of 19 taps and 10-bit attenuator-counters this residual signal-to-noise ratio is about 54 dB. The relations above assume the ranges of all attenuators to be the same. Often the characteristics of the channels to be equalized permit the ranges of the various attenuators to be tapered as one moves from the center towards the ends of the delay line. This would make the above estimate somewhat pessimistic.

The settings of the attenuators are controlled by the cross-correlators whose inputs are the error and delay line tap signals. The multiplying function necessary in measuring cross-correlation is accomplished through the use of a switched modulator driven by a pulse-width modulated signal. The output so obtained is directly proportional to the normalized cross-correlation coefficient and the magnitudes of the two input signals. This particular scheme was selected from the many available because first, it is capable of handling the very large dynamic ranges of the two input signals and second, it determines the true cross-correlation, thereby guaranteeing convergence for all reasonable input signals.

The measurement of cross-correlation also requires integration in time, in fact, integration over the infinite interval. This is, of course, simply too long to wait. A simple resistor-capacitor low-pass filter provides a suitable approximation to real integration.

The outputs of the low-pass filters in the correlators are sliced about

the zero level. The polarity of the output signals from the slicers determines whether the corresponding counter is incremented or decremented when a repetitive clock pulse occurs.* (The repetition rate of this clock pulse will be discussed later.) When the equalization reaches equilibrium, the clock pulses are removed and the attenuator weightings are retained permanently by the binary memory elements.

An equalizer consisting of the elements just described is shown in Fig. 7. The tapped delay lines are shown clustered in the top left-hand corner. The remaining cards in the top row are the resistive-ladder attenuators. There are 20 attenuators, 19 associated with the 19 delay-line taps and the remaining unit serving as part of the automatic gain control loop which regulates the signal level on the delay line. The two rows below the attenuators serve only as lamp indicators for the attenuator settings. The two rows below the lamps contain the binary memory elements and associated logic. The bottom row of cards consists of the cross-correlators. This equalizer was constructed for voiceband use; the delay line has a usable bandwidth in excess of 3,000 Hz, and the tap spacing is 150 microseconds. Examples of its performance will be given in a subsequent section.

3.1.1 *Settling Time*

The settling time (the time required for the equalizer to reach equilibrium) is determined in large measure by the time-constant of the low-pass filter in the correlators and the frequency of the clock which controls the counters.

Nothing has been said to this point about the nature of the test signal used to determine the equalizer settings. The test signal is a passband signal obtained by modulating a smoothed pseudo-random sequence† into the passband frequency range. The pseudo-random sequence¹⁹ was used because this facilitates the generation of identical signals at the transmitter and receiver. The smoothing is done in accordance with the error spectrum weighting filter $W(\omega)$; the modulation is necessary because of the likelihood of frequency offset on carrier transmission facilities. These subjects will be treated in greater detail later.

The pseudo-random sequence has a periodic auto-correlation func-

* The magnitude of the cross-correlation coefficients can be used to control the rate of change of the attenuator settings as in Refs. 10, 15, and 16.

† The pseudo-random sequence is a repetitive sequence of binary digits chosen in a random manner. The sequence can be generated by a binary shift register.

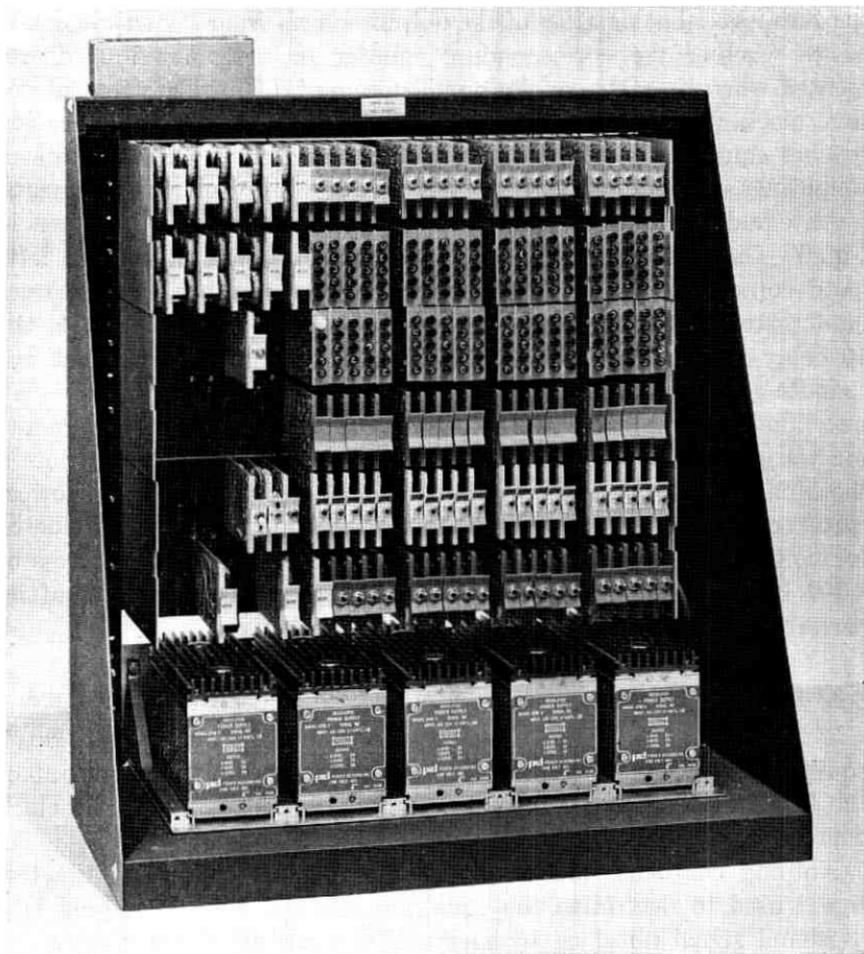


Fig. 7 — Photograph of 19-tap equalizer.

tion. The sequence generator must be designed in such a manner that the period of this auto-correlation function is larger than the length of the expected dispersion in the channel. If this were not the case, the correlator would react to properties of the test signal, rather than to properties of the channel. It is then necessary to integrate, in the correlator, for a length of time corresponding to several periods of the pseudo-random sequence. In the implemented equalizer, the integration is performed in a simple resistor-capacitor low-pass network; the RC-product was established at about four times the pseudo-random

sequence period. The polarity of the output of the correlator must be sampled at a still slower rate so that time is provided for the integrators to reach the steady-state after a change in the attenuator settings. Again, several time constants must be allowed for this to take place. The ratio of the repetition rate of the pseudo-random sequence to the sampling rate of the correlators is about 20 for this particular implementation.

Thus, in general, the length of the dispersion of the channel in time (i.e., the length of the significantly nonzero portion of the channel's impulse response) determines the repetition rate of the pseudo-random sequence and, in turn, this repetition rate determines the rate at which the attenuators are adjusted. The settling time for the equalizer can then be calculated by dividing the number of steps the attenuator must change by the rate of change.

As an example consider a voice channel wherein most of the dispersion is confined to a five millisecond interval. If the repetition rate of the pseudo-random sequence is established at 10 milliseconds, then in accordance with the above comments the clock rate for the pulse controlling the attenuators should be 20 times slower or about 5 Hz. In an equalizer using 10-bit attenuators and starting from the reset condition of zero attenuator weights, the longest travel of an attenuator would be some 500 increments. It would take 100 seconds to traverse the full range. This is a rather long time to wait, even for an equalizer used in such a manner that it is divorced from the communicating modems. There is, fortunately, an easy remedy and this involves letting the attenuators run rapidly to their approximate values and then slowly to their exact values. This dual-mode operation of the attenuator clock can decrease the settling time by a considerable factor.

The settling time for this particular implementation is 10 seconds. This is achieved by running the attenuator clock at a high rate for a fixed initial period and then by continuing operation at a slower rate.

3.2 *Carrier Recovery*

The implementation was designed for use on all voice-frequency channels, including carrier channels. The nature of carrier channels is such that the channel may introduce a slight frequency shift. If such a frequency shift were not compensated, the output of the correlators would be modulated by the shift frequency, ruling out the possibility of satisfactory operation. There are two equivalent means of dealing with this frequency shift. The first is to remove the frequency shift

from the received channel signal and the second is to alter the modulating frequency of the generator of the comparison or desired signal as indicated ("carrier + offset") in Fig. 6. In either case, the frequency offset generated by the channel must be detected; the latter scheme was selected here.

A technique suggested by F. K. Becker²⁰ was used to recover both modulating frequency and the frequency necessary to drive the random sequence generator. In this approach, two pilot tones are added to the transmitted signal, one each at the upper and lower edges of the band of interest. These tones can then be combined in such a fashion that the transmitted carrier plus frequency offset can be recovered. At the same time, by combining the two pilot tones in another manner, the sequence generator clock can be recovered. In the case of the equalizer shown in Fig. 7, the modulating carrier frequency (2400 Hz) plus carrier offset (δ Hz) is obtained from the two pilot tones at 600 Hz and 3000 Hz as indicated by (18).

$$(2400 + \delta) = (3000 + \delta) - \frac{(3000 + \delta) - (600 + \delta)}{4}. \quad (18)$$

Once the proper modulating frequency is obtained, it remains to establish the proper phase. This is achieved by transmitting energy at the carrier frequency. The phase of the carrier generated at the equalizer is adjusted until it agrees with the received carrier phase as it appears at the output of the "main" equalizer tap. This is achieved through the use of a cross-correlator. After the proper phase has been established, the variable phase shift element is locked.

3.3 Timing Recovery

In conjunction with the discussion of settling time, it was stated that the test signal is derived from a pseudo-random sequence generator. It is necessary to synchronize the remote and local generators (which are identical) so that near-optimum use is made of the transversal filter.

Like the modulating carrier, the clock frequency required to drive the sequence generator at the equalizer is derived from the two pilot tones. In the implemented equalizer shown in Fig. 7, the clock frequency of 2400 Hz is obtained via the relation

$$(2400) = (3000 + \delta) - (600 + \delta). \quad (19)$$

In addition to obtaining the proper phase for this clock, it is necessary to synchronize the random 63-bit sequences. These ends are attained in a sequence of two steps.

It is known that the autocorrelation function $R_{pp}(\nu)$ of a pseudo-random sequence of the variety used here has a shape like that of Fig. 8(a). (In the equalizer, the $W(\omega)$ weighting function causes a smoother function to be generated for the autocorrelation function of the desired signal.) The timing recovery circuitry is built upon this fact.

What in essence is needed is an estimate of the arrival time T of the received signal $x(t)$. Knowing this, the desired signal $g(t)$ can be properly synchronized. A maximum likelihood estimate of T is developed using a correlation detector.²¹ Under the assumption that the noise is Gaussian, white, and additive, it can be shown²¹ that the maximum likelihood estimate of T can be found by adjusting T so that

$$q(T) \equiv \int_0^{t_1} g(t+T)x(t) dt \quad (20)$$

is a maximum. Because of the noise component in $x(t)$ there will be some ambiguity in deciding exactly where the maximum of $q(T)$ is, but this ambiguity can be reduced by increasing the length of the observation time t_1 . In fact, when t_1 is very large $q(T)$ approaches the $R_{pp}(T)$ shown in Fig. 8(a), assuming no spectral weighting, band limiting, or channel distortion.

It is known that the effect of linear distortion in a bandlimited channel can be represented in terms of pairs of echoes of the impulse response in the time domain.¹⁸ An estimate of T is obtained for the distorted $x_d(t)$ just as it was in the distortion-free case but because of

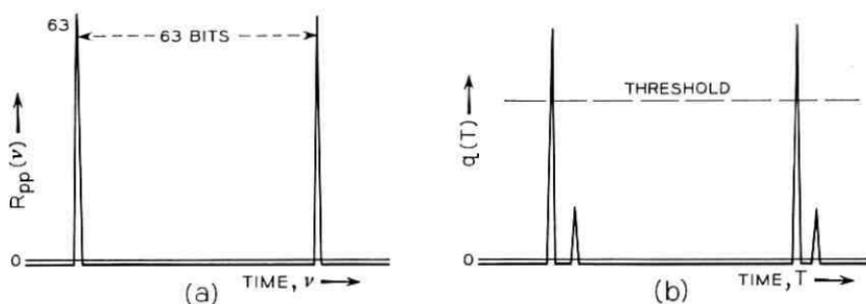


Fig. 8—Synchronization waveforms. (a) Autocorrelation function of pseudo-random word generator (63-bit length). (b) $q(T)$ function for pseudo-random word generator in presence of linear distortion.

the distortion only an approximation to the maximum likelihood estimate is obtained. Again

$$q(T) = \int_0^{t_1} g(t+T)x_d(t) dt \quad (21)$$

is maximized. As in the distortion-free case the ambiguity in q resulting from noise can be made vanishingly small by making t_1 very large. However, the distortion echoes do contribute systematically to $q(T)$ and an increase in the observation time does not diminish their contribution. If the distortion were such that a single echo were introduced by the channel, the $q(T)$ function might have the appearance of Fig. 8(b), again ignoring the effects of band-limiting and smoothing by the filter $W(\omega)$. It can be seen, then, that linear distortion makes the search for the absolute maximum of $q(T)$ more difficult by introducing greater undulations in the $q(T)$ function. Because of the complexity of the $q(T)$ function, the search for its absolute maximum is made in two successive modes.

In the first mode, gross synchronization is attained. This means that the timing of the desired waveform sequence is shifted until it is roughly lined up with the received signal as it appears at the "main" tap. This coarse alignment is obtained by cross-correlating the two signals just mentioned and comparing the result with a fixed threshold. Until the output of the cross-correlator, $q(T)$, reaches the threshold, the phase of the timing signal is continuously increased (over an interval which may be as large as 63 symbol periods in the case of the 63-bit sequence). When the threshold is reached the phase is locked. The threshold is determined empirically so that only the one large spike (corresponding to the undistorted pulse) penetrates the threshold. Thus, in the first mode the proper "spike" of $q(T)$ is found; in mode 2 the maximum of this spike is found.

The maximum of $q(T)$ can be found by partial differentiation of (21) with respect to T and setting the result equal to zero.

$$\frac{\partial q(T)}{\partial T} = 0 = \int_0^{t_1} g'(t+T)x_d(t) dt, \quad (22)$$

where $g'(t)$ is the time derivative of $g(t)$. This approach could not be used from the start because $q(T)$ can be assumed to be a convex function only over a small region. The operation indicated in (22) is achieved through yet another cross-correlator.

A few words are in order about what has been called the "main"

tap—that tap which is used for both carrier and timing reference. The main tap would normally be the center tap on the delay line. It turns out empirically, however, that most distorting echoes lag the undistorted impulse. Hence, lower residual distortion is obtained by shifting the main or reference tap to a position about two-thirds down the delay line.

IV. PERFORMANCE

4.1 Computer Simulations

In order to determine the theoretical performance of this equalization technique, a computer simulation was made. Fig. 9 shows results for a voiceband channel. The unequalized channel characteristic was taken as a typical Direct-Distance-Dialed connection as given in Alexander, Gryb, and Nast.²² The amplitude characteristic has a 15 dB/octave falloff starting at 240 Hz, is flat from 240 to 1100 Hz, has a linear logarithmic slope to 7.6-dB loss at 3000 Hz and an 80 dB/octave loss commencing at 3000 Hz. The delay characteristic is parabolic, centered at 1500 Hz, with a maximum delay of 1 millisecond at 0 and 3000 Hz. In the simulation, the error spectral weighting function $W(\omega)$ used was of raised cosine shape, symmetric about a peak at 1650 Hz and zero at 300 and 3000 Hz. The tap spacing was established at 150 microseconds. In Fig. 9 the amplitude and delay frequency-response curves for both unequalized and equalized channels are shown. Three cases are shown, those of 9, 13, and 25 taps.

A simulation was also made for a baseband channel with group bandwidth.‡ Only the amplitude frequency responses are shown because the delay distortion was not significant in this particular case and remained essentially invariant throughout the equalization process. Both uniform and nonuniform spectral weightings were investigated. In the cases where a nonuniform spectral weighting $W(\omega)$ was used, $W(\omega)$ was selected as a half-cosine rolloff shape, essentially flat to 12.5 kHz, and then falling to zero at 37.5 kHz as a cosine. Energy at very low frequencies was given small weighting by a simple high-pass filter with 2-kHz corner frequency. Fig. 10 displays the amplitude characteristics on both linear and logarithmic frequency scales as the number of taps is increased from 13 to 51, all with the half-cosine rolloff weighting. Performance improves with the number of taps but

‡ A "group" is twelve voice channels with a bandwidth of about 12×4 or 48 kHz.

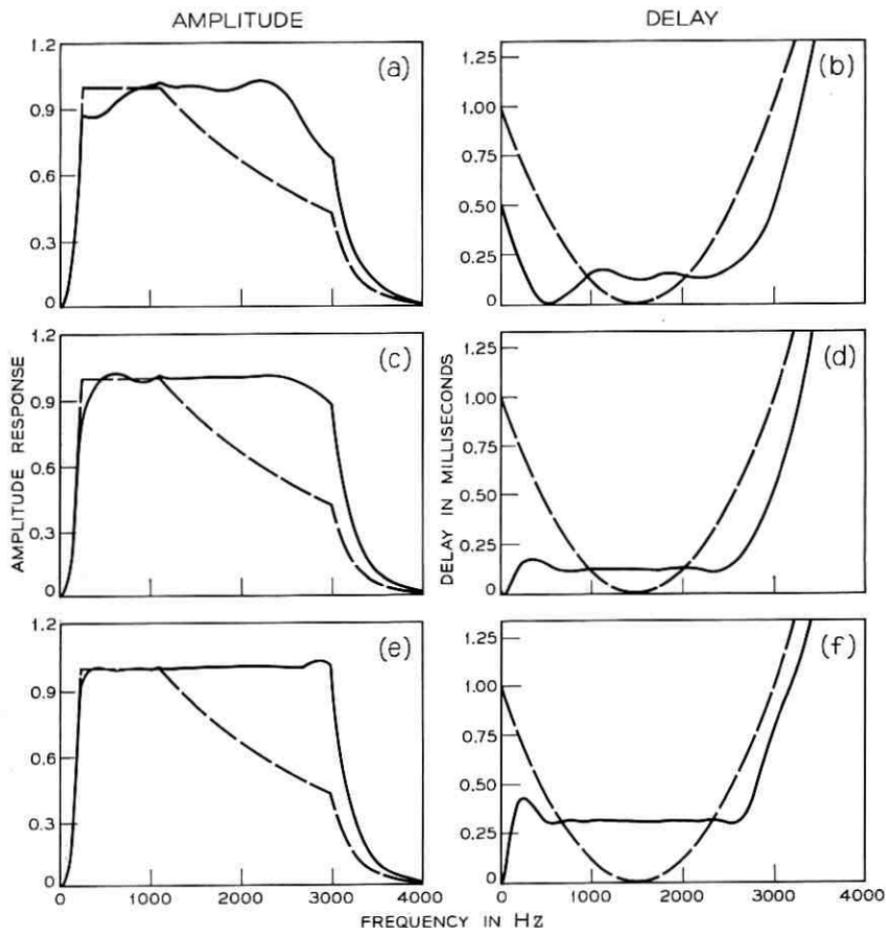


Fig. 9—Simulated voiceband performance (a) Amplitude characteristics, 9 taps, raised cosine weighting. (b) Delay characteristics, 9 taps, raised cosine weighting (c) Amplitude characteristics, 13 taps, raised cosine weighting. (d) Delay characteristics, 13 taps, raised cosine weighting. (e) Amplitude characteristics, 25 taps, raised cosine weighting. (f) Delay characteristics, 25 taps, raised cosine weighting.

the change is rather subtle compared with the voiceband case. Fig. 11 illustrates the effects of error weighting (weighted and unweighted) and the effect of signal-to-noise ratio for the case of white noise. Note that in the case of very small noise, the equalized channel characteristic may behave erratically in the region where the error has very little weight (i.e., near 37.5 kHz). The addition of noise or the use

of a more uniform weighting prohibits this behavior. Note also that the use of weighting forces the best equalization to occur in the region of highest weight, i.e., 2 to 25 kHz.

It may also be seen that in the case of abnormally high noise the equalizer minimizes the total error energy consisting of both distortion and noise as predicted in Section 2.3.

4.2 Measurements on Real Facilities

Measurements have also been made with the experimental equipment operating over real facilities. Fig. 12 shows the equalization of an actual L-carrier looped facility from Holmdel, N. J. to Chicago and return. Both the unequalized and equalized delay and amplitude frequency responses are shown. The results are those for thirteen taps with a raised cosine error spectral weighting function with zero weight at 600 and 3000 Hz. Fig. 13 shows the binary eyes resulting from

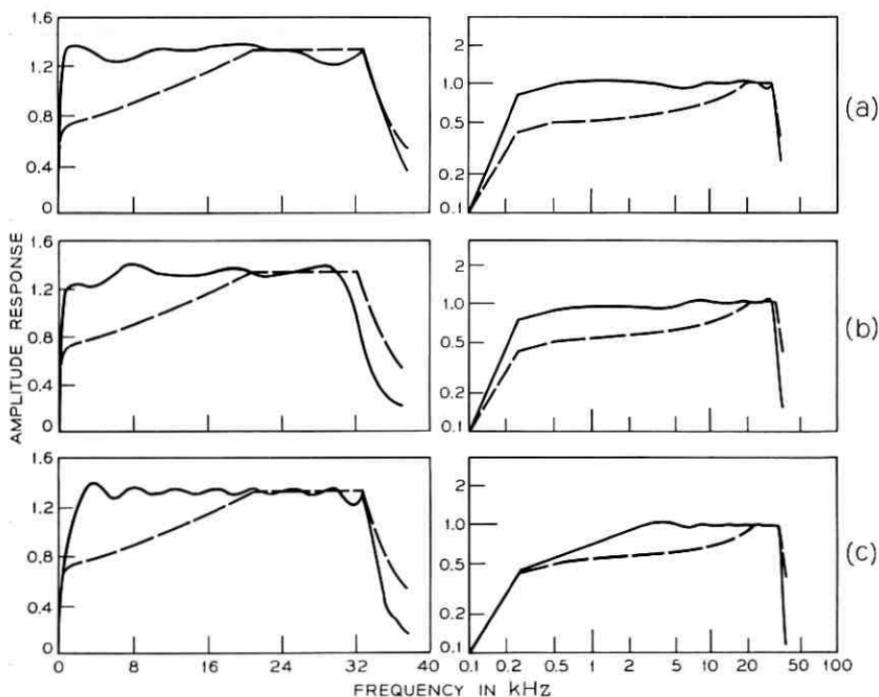


Fig. 10—Effect of number of taps.—group band, weighted error, $S/N=25\text{dB}$. (a) 13 taps. (b) 25 taps. (c) 51 taps.

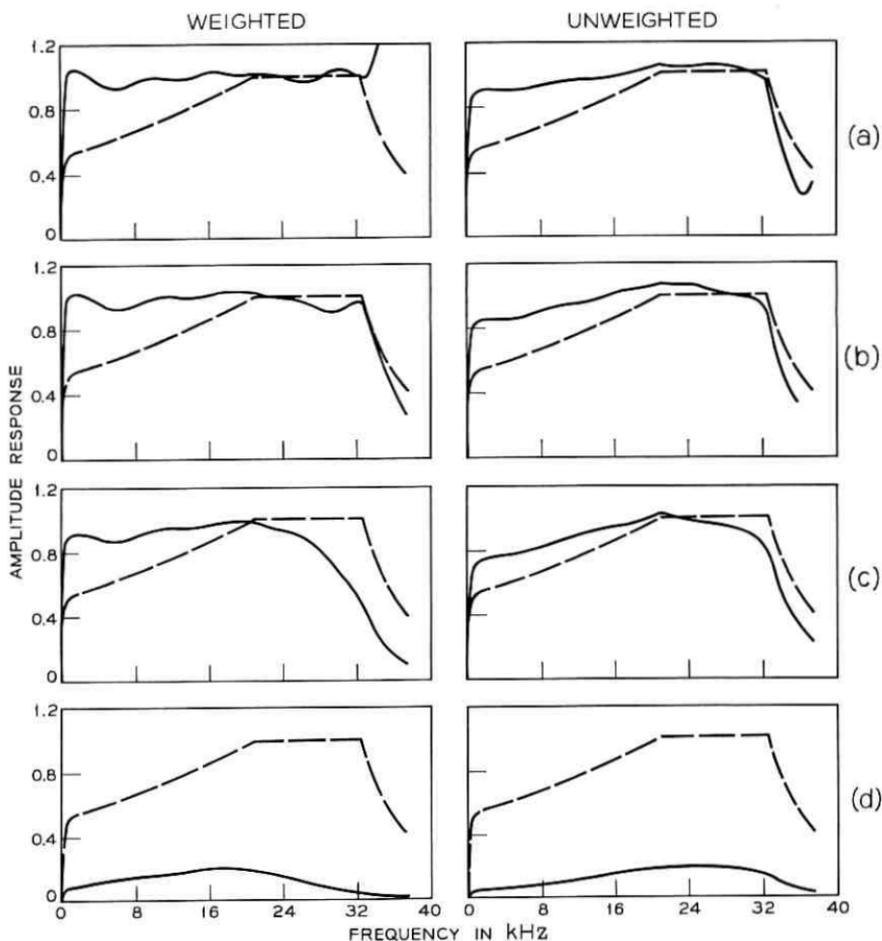


Fig. 11—Effect of signal-to-noise ratio and error weighting. Group band, 15 taps (a) $S/N=95\text{dB}$. (b) $S/N=25\text{dB}$. (c) $S/N=10\text{dB}$. (d) $S/N=-10\text{dB}$.

transmission of an FM data signal on a DDD looped facility from Holmdel to Denver and return. A 19-tap equalizer was used for the rms-equalized case. As an example of asynchronous transmission over similar facilities, Fig. 14 shows facsimile transmission, equalized and unequalized. Fig. 14 was obtained using the Bell System DATA-PHONE* Data Set 602A which contains an FM modem.

The nineteen-tap equalizer shown in Fig. 7 was used to equalize a

* Service mark of the Bell System.

looped facility from Holmdel, N. J. to Omaha and return. The results of this test are shown in Fig. 15 with the requirements for a schedule 4B line. The weighting function is identical to that used for the equalization obtained in Fig. 12. As can be seen, the 4B requirements are met by the equalized channel except at the band edges.

4.3 Filter Synthesis

As an example of automatic filter synthesis, the curves in Fig. 16 were obtained. The system configuration is that of Fig. 4(b). The desired or model amplitude response is shown for comparison with 5-, 9-, and 19-tap approximations to it.

V. CONCLUSION

Automatic Equalization is a powerful tool for increasing the efficiency of communication channels. The implementation described is of general utility and need not be married to a particular modem. It functions conveniently in the passband and is especially suited to the equalization of a large number of communication channels terminating at a common location where the adjustment circuitry can be shared. In addition, the principles of the techniques seem applicable to a wide class of problems.

Much work remains to be done before generalized automatic equali-

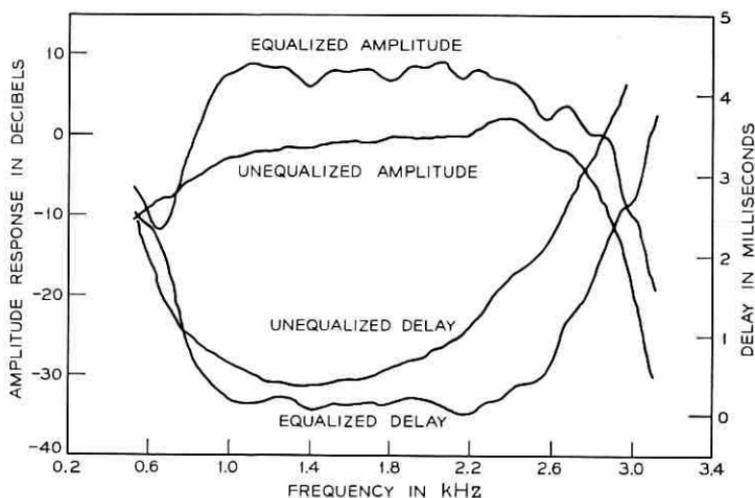
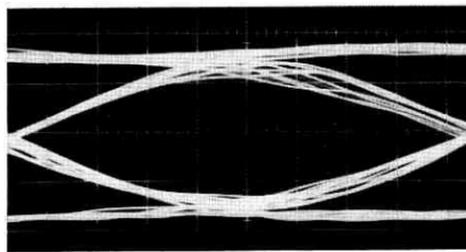
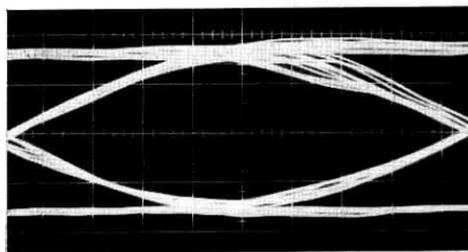


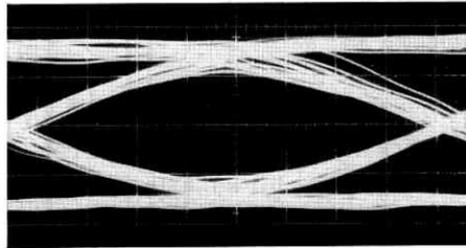
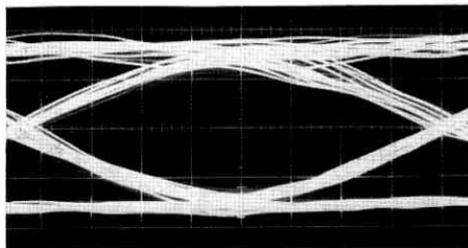
Fig. 12 — Actual performance—frequency response.

1200 BPS

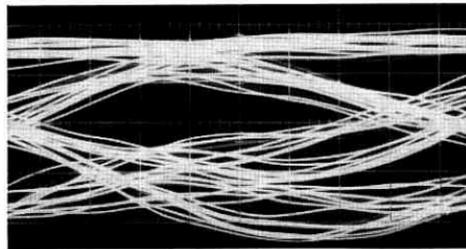
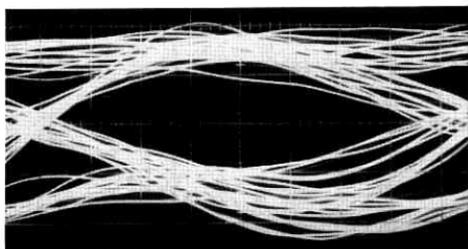
1400 BPS



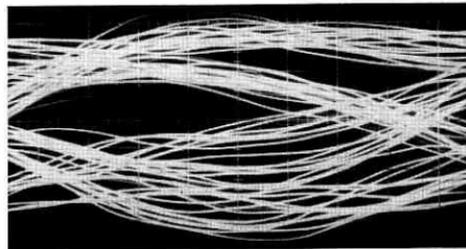
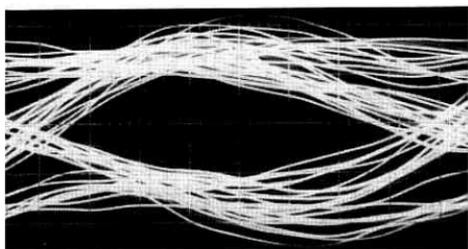
BACK TO BACK



RMS EQUALIZED



354A EQUALIZED

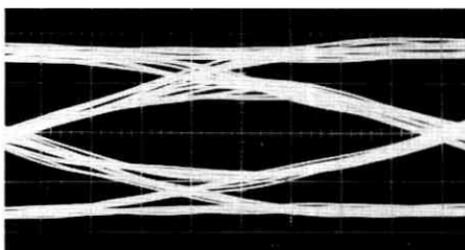
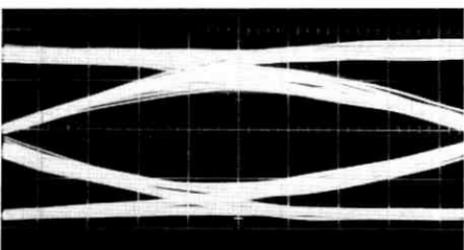


UNEQUALIZED

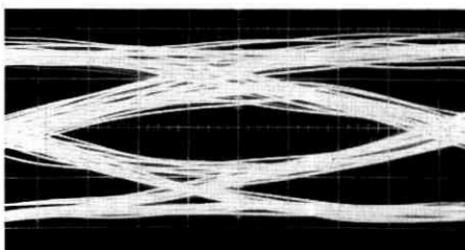
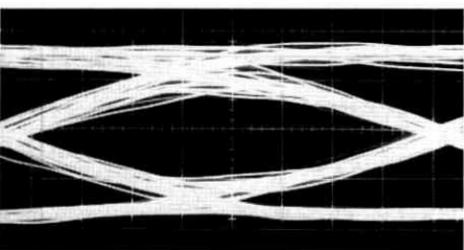
Fig. 13 — Actual performance—eye patterns for the Bell System Data Set 202D. —unequalized, partially equalized using a fixed compromise equalizer, automatically equalized, back-to-back operation.

1600 BPS

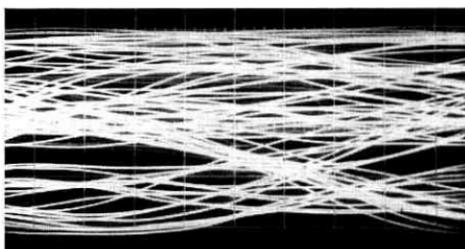
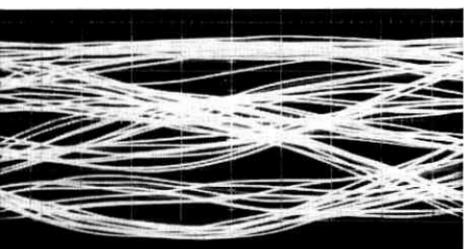
1800 BPS



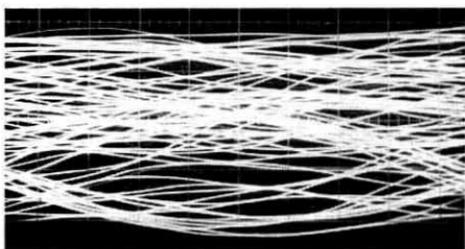
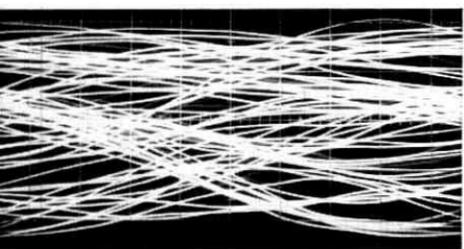
BACK TO BACK



RMS EQUALIZED



354A EQUALIZED



UNEQUALIZED

zation becomes practical. In addition, the equalizer described here does nothing for the problem of nonlinear distortion or for the time-varying channel. However, the results obtained thus far are encouraging.

The authors gratefully acknowledge the encouragement and contribution of their colleagues, especially F. K. Becker, L. N. Holzman, and E. Port.

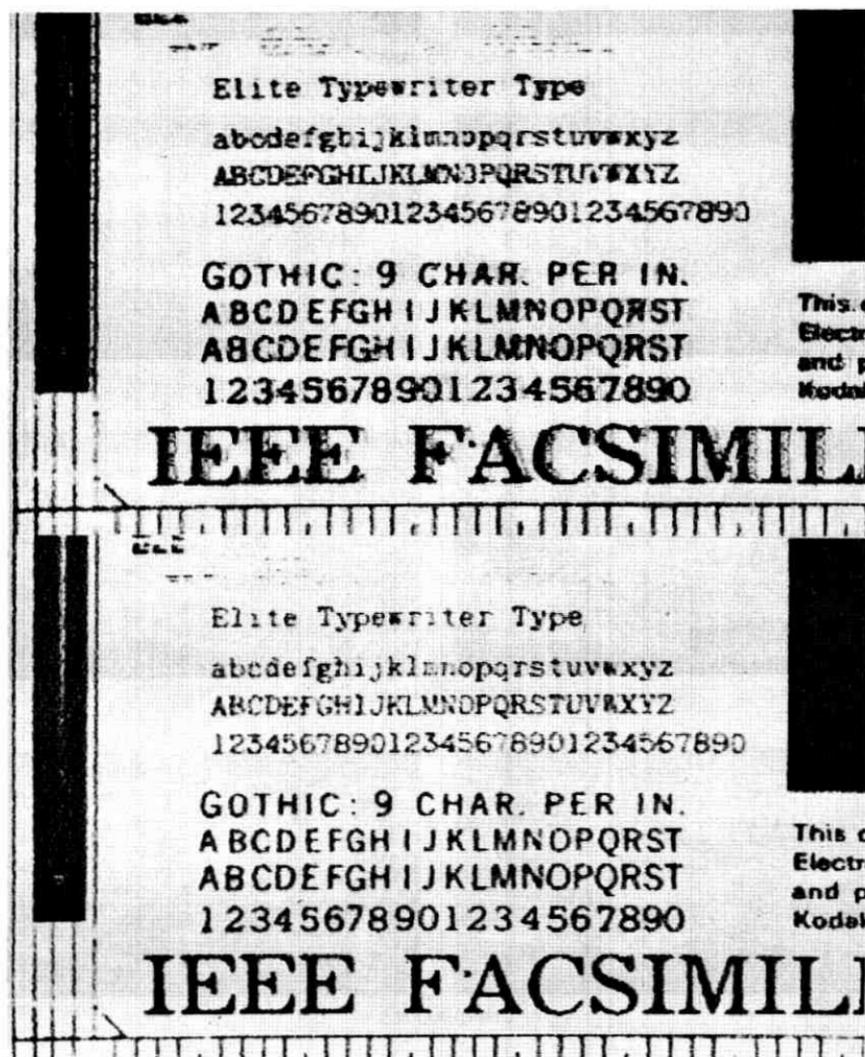


Fig. 14 — Actual performance—facsimile transmission.

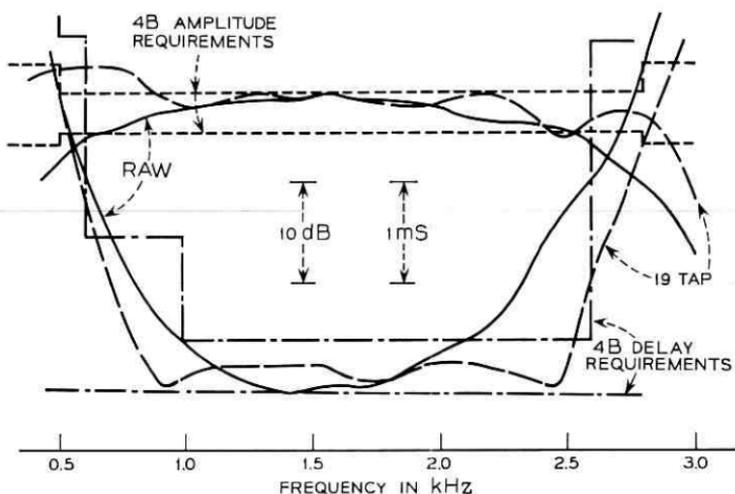


Fig. 15 — Actual performance—frequency response—Omaha loop.

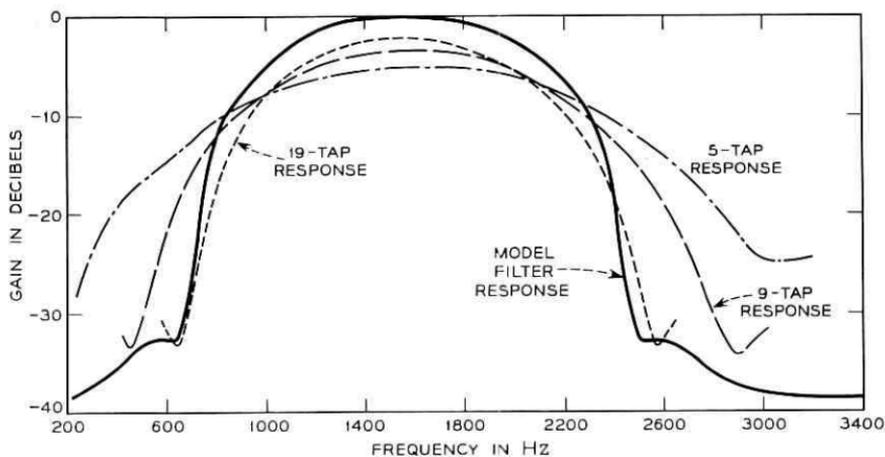


Fig. 16 — Actual performance—filter synthesis.

REFERENCES

1. Rudin, H. R., Automatic Equalization Using Transversal Filters, *IEEE Spectrum*, January, 1967, pp. 53-59.
2. Coll, D. C. and George, D. A., A Receiver for Time-Dispersed Pulses, *Conf. Record 1965 IEEE Ann. Commun. Conv.*, pp. 753-757.
3. Coll, D. C. and George, D. A., The Reception of Time-Dispersed Pulses, *Conf. Record 1965 IEEE Ann. Commun. Conv.*, pp. 749-752.
4. DiToro, M. J., A New Method of High-Speed Adaptive Serial Communica-

- tion Through Any Time-Variable and Dispersive Transmission Medium, Conf. Record 1965 IEEE Ann. Commun. Conv., pp. 763-767.
5. Funk, H. L., Hopner, E., and Schreiner, K. E., Automatic Distortion Correction for Efficient Pulse Transmission, IBM J., January, 1965, pp. 20-30.
 6. Lucky, R. W., Automatic Equalization for Digital Communication, B.S.T.J., 44, April, 1965, pp. 547-588.
 7. Lucky, R. W., "Techniques for Adaptive Equalization of Digital Communication, B.S.T.J., 45, February, 1966, pp. 255-286.
 8. Becker, F. K., Holzman, L. N., Lucky, R. W., and Port, E., Automatic Equalization for Digital Communication, Proc. IEEE, (Letters) 53, January, 1965, pp. 96-97.
 9. Wiener, N., *Extrapolation, Interpolation, And Smoothing of Stationary Time Series*, John Wiley & Sons, New York, 1957.
 10. Narendra, K. S. and McBride, L. E., Multiparameter Self-Optimizing Systems Using Correlation Techniques, IEEE Trans. Auto. Cont., January, 1964, pp. 31-38.
 11. Lucky, R. W. and Rudin, H. R., Generalized Automatic Equalization for Communication Channels, Proc. IEEE (Letters), 54, March, 1966, pp. 439-440.
 12. Lucky, R. W. and Rudin, H. R., Generalized Automatic Equalization for Communication Channels, Digest of Tech. Papers, 1966 IEEE Int. Commun. Conf., June, 1966, pp. 22-23.
 13. Lee, Y. W., *Statistical Theory of Communication*, John Wiley & Sons, Inc., New York, 1960.
 14. Becker, F. K. and Rudin, H. R., Application of Transversal Filters to the Problem of Echo Suppression, B.S.T.J., 45, December, 1966, pp. 1847-1850.
 15. Sondhi, M. M. and Presti, A. J., A Self-Adaptive Echo Canceller, B.S.T.J., 45, December, 1966, pp. 1851-1854.
 16. Sondhi, M. M., An Adaptive Echo Canceller, B.S.T.J., 46, March, 1967, pp. 497-511.
 17. Zmood, R. B., Some Aspects of the Development of a Self Adaptive Hybrid, unpublished report of the Postmaster General's Department, Melbourne, Australia, August, 1966.
 18. Wheeler, H. A., The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, Proc. IRE, 27, June, 1939, pp. 359-385.
 19. Golomb, S. W., *et al*, *Digital Communication with Space Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.
 20. Becker, F. K., An Exploratory, Multilevel, Vestigial Sideband Data Terminal for Use on High Grade Voice Facilities, IEEE Ann. Commun. Conv. Conf. Record, June, 1965, pp. 481-484.

Minimum Cost Communication Networks

By E. N. GILBERT

(Manuscript received July 21, 1967)

Cities A_1, \dots, A_n in the plane are to be interconnected by two-way communication channels. $N(i, j)$ channels are to go between A_i and A_j . One could install the $N(i, j)$ channels along a straight line, for every pair i, j . However it is usually possible to save money by rerouting channels over longer paths in order to group channels together. In this way, large numbers of channels share such preliminary expenses as real estate, surveying, and trench digging.

The geometry of the least expensive network will depend on the numbers of channels $N(i, j)$ and on the function $f(N)$ which represents the cost per mile of installing N channels along a common route. If the preliminary expenses are the only expenses then $f(N)$ is a constant, independent of N . In that case the best network is obtained by routing channels along lines of the "Steiner minimal tree", a graph which has been studied extensively and which can be constructed by ruler and compass. In part, this paper generalizes Steiner minimal trees for the case of an arbitrary function $f(N)$. One again obtains a ruler and compass construction for a minimizing tree, which is likely to provide a best or good solution when preliminary costs are a significant part of the total cost. However the minimizing tree need not be the best solution in general because further cost reductions may now be possible by using graphs which have cycles. Other properties of Steiner minimal trees generalize only part way, and some examples illustrate the new complications.

The remainder of the paper considers functions $f(N)$ with special properties. A convexity property

$$f(N + 2) - 2f(N + 1) + f(N) \leq 0, N = 1, 2, \dots$$

ensures that there is a minimizing solution in which all $N(i, j)$ channels between A_i and A_j take the same path (no split routing). If $f(N)$ is a linear function ($f(N) = a + bN$), one can obtain simple bounds on the minimum cost. The lower bound is fairly accurate.

I. INTRODUCTION

Let points A_1, \dots, A_n in the plane represent n cities which require a communications network. Let $N(i, j)$ denote the number of channels which the network must supply between A_i and A_j . The network sought must provide these channels at minimum cost. In calculating costs suppose that a monotone function $f(N)$ represents the cost in dollars per mile to install N channels together along a common route.

One possible network just connects each pair A_i, A_j by $N(i, j)$ channels installed along a straight line path. This network will be called the *complete network* because the routes used form a complete graph. Fig. 1(a) is the complete network for a case with $n = 4$; the numbers on the lines are the $N(i, j)$.

The complete network makes each channel as short as possible; it is the cheapest network if $f(N) = N$. However, most situations have more complicated functions $f(N)$. In particular, there are usually some *preliminary costs* for surveying, obtaining the right-of-way, digging a trench, etc. These items have a non-zero cost $f(0)$ dollars per mile regardless of how many channels are to be installed.

In some cases preliminary costs may be so high that a network which merely minimizes preliminary costs is a reasonable choice. Such a network must minimize the total number of miles of right-of-way. For the example in Fig. 1(a), the network which minimizes preliminary costs is Fig. 1(b) [or, more simply, Fig. 1(c)]. Such networks can be drawn with a ruler and compass in a finite (possibly large) number of steps (see Ref. 1, 2).

When $f(N)$ is not constant, the cheapest network is harder to find. Still the methods which minimize only the preliminary costs generalize far enough to be useful. Sections III and IV develop these generalizations. In particular, if preliminary costs are a large fraction of the

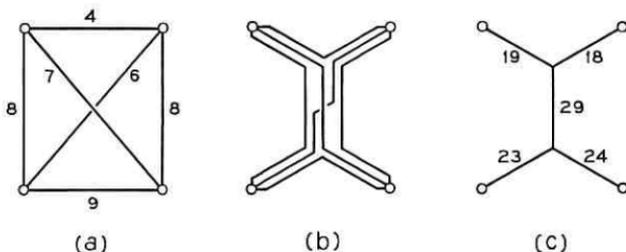


Fig. 1 — Networks.

total cost one has a good chance of constructing the cheapest network by these methods.

In many problems the cost function is linear, $f(N) = a + bN$. A linear cost function is obtained if the *incremental cost* $f(N) - f(N-1)$ of adding an N th channel to a group of $N-1$ channels is a fixed amount b dollars per mile, independent of N . The cost of additional copper wires, channel filters, or repeaters usually does not depend on N . By contrast, consider waveguide systems. Each guide can supply thousands of channels. The incremental cost is small for most values of N but is large when adding channel N requires adding another guide; $f(N)$ is a staircase function. Section VI obtains some bounds on the cost of the cheapest network when $f(N)$ is linear. Section V finds a property of the minimal cost network when $f(N)$ is merely convex.

II. STEINER MINIMAL TREES

A network may be represented, as in Fig. 1(c), as a set of lines (the routes or right-of-ways) connecting A_1, \dots, A_n and perhaps some other points where lines join. This representation will be called the *graph* of the network. Figs. 1(b) and 1(c) illustrate the distinction between a network and its graph. A *Steiner point* is a junction point of the graph which is not one of A_1, \dots, A_n . Fig. 1(c) has two Steiner points. The *minimal graph* is the graph of the cheapest network. A graph is *relatively minimal* if its Steiner points are located so that no small displacement of the Steiner points reduces the cost. If a graph is relatively minimal there is no guarantee that a more violent perturbation, altering the topology of the graph, may not secure a reduction; i.e., relatively minimal graphs need not be minimal.

The procedure to be described here finds relatively minimal graphs which are trees having exactly three lines incident at each Steiner point. The cheapest of these relatively minimal trees will be called the *Steiner minimal tree* for A_1, \dots, A_n . The procedure in question is a modification of one which applies when the cost function is simply $f(N) = 1$. In order to have an easy terminology by which one may compare a given problem against the corresponding problem with $f(N) = 1$, I use the adjective *ordinary* freely to mean "having $f(N) = 1$ ". Thus, "ordinary minimal graph, ordinary relatively minimal graph, ordinary Steiner minimal tree, \dots " mean "minimal graph, relatively minimal graph, Steiner minimal tree, \dots in the case $f(N) = 1$ ".

The ordinary case is a simpler one than the general case because the

ordinary minimal graph is the ordinary Steiner minimal tree. In general, the minimal graph need not be a tree (recall that the complete graph is minimal if $f(N) = N$). Moreover, even the cheapest tree need not be a Steiner minimal tree. For example, consider four cities A_1, \dots, A_4 at the corners of a unit square as shown in Fig. 2. For the demand matrix $N(i, j)$ take

$$\| N(i, j) \| = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 10 \\ 1 & 1 & 0 & 1 \\ 1 & 10 & 1 & 0 \end{pmatrix}$$

and let N channels cost $f(N) = 1 + N$ dollars per mile. Fig. 2(a) shows the cheapest tree. It is not a Steiner minimal tree because four lines meet at its Steiner point. Fig. 2(b) shows a typical tree in which three lines meet at each Steiner point. However, Fig. 2(b) is not relatively minimal; its cost decreases when the two Steiner points are displaced toward the center of the square. If one continues to displace these Steiner points, in hopes of finding a relatively minimal tree, they finally merge together as in Fig. 2(a).

III. GENERALIZATIONS FROM THE ORDINARY CASE

In Ref. 1 we gave some simple properties of ordinary relatively minimal trees and ordinary Steiner minimal trees. Some of these properties generalize directly while others do not. This section will discuss the simplest generalizations. In some cases the proofs are omitted because the arguments of Ref. 1 apply with only trivial changes.

3.1 *Mechanics*

A graph of a network may be interpreted as a mechanical system of elastic bands (the lines). A_1, \dots, A_n are fixed supports for the

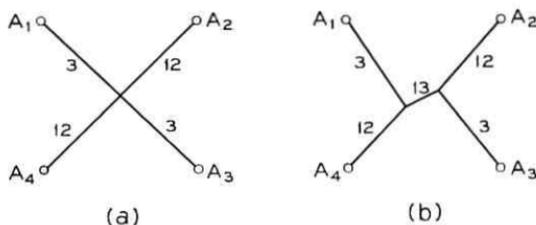


Fig. 2—Four cities problem.

bands incident there but the bands at a Steiner point are merely joined together and left free to move. Let each band have a tension equal to the cost per mile of the channels in the corresponding line. Then the mechanical system has a potential function equal to the cost of the graph; the system is in stable equilibrium if and only if the graph is relatively minimal.

3.2 *Angles at a Steiner point*

At a Steiner point S let vectors v, v', v'', \dots denote the forces (tensions) exerted by the elastic bands. The condition for mechanical equilibrium (relatively minimal graph) is $v + v' + v'' + \dots = 0$. The magnitudes $|v|, |v'|, |v''|, \dots$ are the costs per mile of the lines at S . When S has only three lines, the law of cosines determines the angles between the lines. For instance,

$$\cos(v', v'') = (|v|^2 - |v'|^2 - |v''|^2) / (2|v'| |v''|). \quad (1)$$

The analogous condition on ordinary relatively minimal trees, which stated that three lines meet at 120° at S , is an instance of (1) with $|v| = |v'| = |v''|$. When four or more lines meet at S the equilibrium condition does not determine the angles at S .

3.3 *Number of Steiner points*

Consider any tree joining A_1, \dots, A_n and let s be the number of Steiner points. It is no restriction to assume that no Steiner point has less than three lines; for clearly such Steiner points can save no cost. Then (see Ref. 1, Section 3.4)

$$s \leq n - 2$$

with equality holding if and only if each Steiner point has three lines and each A_i has one line.

3.4 *Uniqueness*

Suppose a graph, not necessarily a tree, is given for a network connecting A_1, \dots, A_n . The numbers of channels are also supposed prescribed for each line of the graph. Now perturb the positions of the Steiner points trying to reach a relative minimum cost for graphs with the same topology. As illustrated by Fig. 2, it can happen that a relative minimum may be only approached but not attained. In the ordinary case, when one does find a relatively minimal graph one can conclude that there are no others with the same topology.

In the general case, there is no such uniqueness. For example, sup-

pose A_1, A_2, A_3 are at the vertices of an equilateral triangle and suppose $N(i, j) = 1$ for all pairs (i, j) . Let $f(N) = 1 + (3^N - 1)(N - 1)$. Fig. 3 shows a possible graph and gives the angles, obtained from (1), which suffice for a relative minimum. These angles do not determine the locations of the Steiner points. It suffices to put each Steiner point S_i at the same distance from the center O of the triangle and on the line OA_i .

Fig. 4(a) shows that one may encounter non-uniqueness even when searching for a relatively minimal tree. To perturb S into a position of minimum cost, place S anywhere on the line segment A_2A_3 . The individual channels $[N(i, j) = 1$ for all $i, j]$ appear in Fig. 4(b). Steiner points, such as S in Fig. 4(b), at which all incident lines meet at either 180° or 360° have no real interest. Any channel which makes a 180° turn at S can be rerouted away from S over a shorter path using only existing right-of-ways. After the shortening [Fig. 4(c)] the Steiner point is gone.

In spite of examples like Figs. 3 and 4, a weak kind of uniqueness holds even in the general case. Any relatively minimal tree is either the unique relatively minimal tree with the given topology or else it has a Steiner point, like S in Fig. 4(b), at which all lines meet at angles of either 180° or 360° . An outline of the proof follows. As in Ref. 1 the argument uses an "averaging" operation for graphs. If G and G' are two graphs with the same topology, the *averaged graph* $pG + qG'$ (where $p \geq 0, q \geq 0$, and $p + q = 1$) has vertices of the form $pV + qV'$ where V, V' are corresponding vertices, $V \in G$ and $V' \in G'$. For each line V_1V_2 of G (and correspondingly, $V'_1V'_2$ of G') $pG + qG'$ has the line joining $pV_1 + qV'_1$ to $pV_2 + qV'_2$. If L is a line V_1V_2 of G and L' the corresponding line of G' , let $pL + qL'$ denote the corresponding line of $pG + qG'$. The lengths $|L|, |L'|, |pL + qL'|$ of these lines satisfy

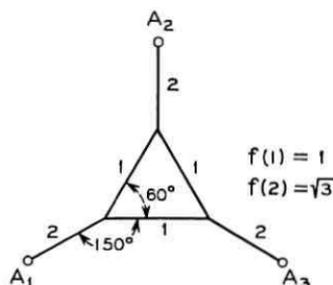


Fig. 3 — Example of non-uniqueness.

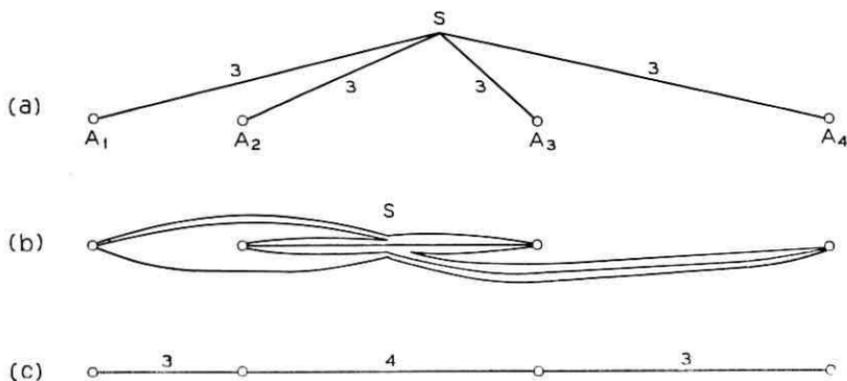


Fig. 4 — Non-uniqueness for trees.

$$|pL + qL'| \leq p|L| + q|L'| \tag{2}$$

with equality holding only if the directions of the line segments V_1V_2 and $V'_1V'_2$ are the same.

One can now prove that all relatively minimal graphs with the same topology have the same cost. For suppose, on the contrary, that graphs G, G' have the same topology and have costs c, c' with $c < c'$. Because of (2) the cost of $pG + qG'$ is no greater than $pc + qc'$. Then, taking p to be small, $pG + qG'$ is a slight perturbation of G' and costs less than c' . Then G' cannot have been relatively minimal, a contradiction.

If G and G' are two different graphs which both attain the relative minimum cost, then (2) shows that an average graph $pG + qG'$ will cost even less (a contradiction) unless every line of G' is parallel to its corresponding line in G . Note that the graphs obtained from Figs. 3 and 4(b) all had that property. Now suppose G and G' are relatively minimal trees. If G and G' differ some Steiner point S in G is connected to vertices V_1 and V_2 such that $V'_1 = V_1, V'_2 = V_2$, but $S' \neq S$. For instance, V_1 and V_2 might be two of A_1, \dots, A_n . But, to avoid the contradiction noted above, SV_1 and $S'V_1$ must be parallel, as must SV_2 and $S'V_2$. That can be true when $S \neq S'$ only if S, S', V_1, V_2 are colinear, whence V_1S makes a 360° angle with V_2S .

3.5 Number of choices

In Ref. 1 the solution to the ordinary case is found by constructing a relatively minimal tree, if one exists, for each of the topologically distinct ways of interconnecting A_1, \dots, A_n . Because of 3.3 there are only a finite number of cases to consider. For $s = 0, 1, \dots, n - 2$,

the number of cases with s Steiner points turns out to be

$$2^{-s} \binom{n}{s+2} (n+s-2)!/s!$$

In the general problem, each of these cases is again a candidate for the Steiner minimal tree. The total number of cases for $n = 3, 4, 5, 6, 7, \dots$ are 4, 27, 270, 3645, 62370, \dots . Of course the minimal graph may not be one of these trees; in general, there will be many more cases.

Fortunately, it seems to be easy to guess topologies which, if not actually best, cost only slightly more than the minimal cost. In Ref. 1, for example, we were unable to invent a problem in which the minimum cost was less than 86.6 percent of the cost of the (easily constructed) best tree having no Steiner points. The four cities in the unit square of Fig. 2 illustrate the same thing. Again let $f(N) = 1 + N$ and let $||N(i, j)||$ be the same as in Section II. Table I compares the cost of the cheapest graph, Fig. 2(a), with some other simple ones.

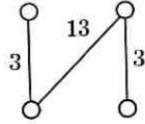
These comparisons suggest that one should be willing to accept a good network (perhaps the best relatively minimal network obtained for several reasonable topologies) even though it is not proved to have absolutely minimum cost. There are usually too many cases to find the best network by exhaustion; also the saving in cost is apt to be slight.

IV. CONSTRUCTION ALGORITHMS

The ruler and compass construction of relatively minimal trees is similar to the construction in the ordinary case.

Consider first the case $n = 3$. Fig. 5(a) shows a typical case with given points A_1, A_2, A_3 to be joined to a Steiner point S . The costs c_i per mile of the three lines SA_i are supposed known. The angles α_1, α_2 ,

TABLE I

Graph	Cost (in dollars)
Fig. 2(a)	24.04
complete graph	26.38
ordinary Steiner min. tree	25.55
	27.80

α_3 at which lines meet at S are now determined from the equilibrium condition,* e.g., by (1). A ruler and compass construction for $\alpha_1, \alpha_2, \alpha_3$ is easy because these angles are the exterior angles to a triangle with sides c_1, c_2, c_3 [Fig. 5(b)].

In general, c_1, c_2, c_3 might have any values, including some which are not constructable by ruler and compass (e.g., perhaps $c_1 = 2^{1/3}, c_2 = \pi, c_3 = e$). Then Fig. 5(b) is itself not constructable without first using the ruler as a "scale" to lay off segments of lengths c_1, c_2, c_3 . I assume that these segments have already been drawn. Then all other

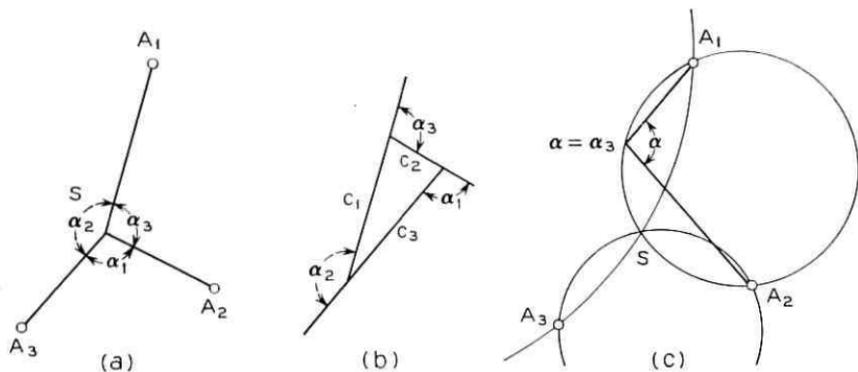


Fig. 5 — First construction with $n = 3$.

constructions, such as the one for $\alpha_1, \alpha_2, \alpha_3$, can use the ruler and compass in the manner intended by Euclid.

Since angle $A_1SA_2 = \alpha_3$, S lies on a circular arc of angle $2\pi - 2\alpha_3$ through A_1 and A_2 . By constructing this arc, and a similar arc for A_2A_3 or A_3A_1 , one constructs S as an intersection of circular arcs [Fig. 5(c)]. The same construction appears in Ref. 4.

In Fig. 5(c) consider the line A_3S extended to meet the circle through A_1 and A_2 again. Let $A_{1,2}$ denote this new point of intersection [Fig. 6(a)]. The point $A_{1,2}$ has interesting properties which are needed for solving cases with $n \geq 4$.

First note [Fig. 6(b)] that the exterior angles of the triangle $A_1A_2A_{1,2}$ are $\alpha_1, \alpha_2, \alpha_3$. Then this triangle is similar to the triangle of Fig. 5(b) and so can be constructed by ruler and compass (if $|A_1A_2|$

* If one of the c_i exceeds the sum of the other two, say $c_1 + c_2 < c_3$, no choice of angles satisfy the equilibrium condition. The minimal tree consists just of two lines (A_1A_3 and A_2A_3 in the case cited). In many cases the function $f(N)$ is convex, as defined in (3), and then $c_1 + c_2 < c_3$ cannot happen.

$=d$, then $|A_1A_{1,2}| = dc_2/c_3$. The important fact used later on is that this construction will produce $A_{1,2}$ from c_1, c_2, c_3, A_1 , and A_2 , without using A_3 .

Another construction for the case $n = 3$ proceeds as follows. With A_1A_2 as a base erect a triangle* with sides $|A_1A_2| = d, dc_2/c_3$, and dc_1/c_3 to construct $A_{1,2}$. Circumscribe this triangle in a circle C_{12} . If A_3 lies inside C_{12} there is no Steiner point (the cheapest solution consists of two lines A_3A_1 and A_3A_2). If A_3 lies outside C_{12} draw the line segment $A_{1,2}A_3$. Observe whether this segment crosses the arc A_1A_2 of C_{12} which does not contain $A_{1,2}$. If there is a crossing point

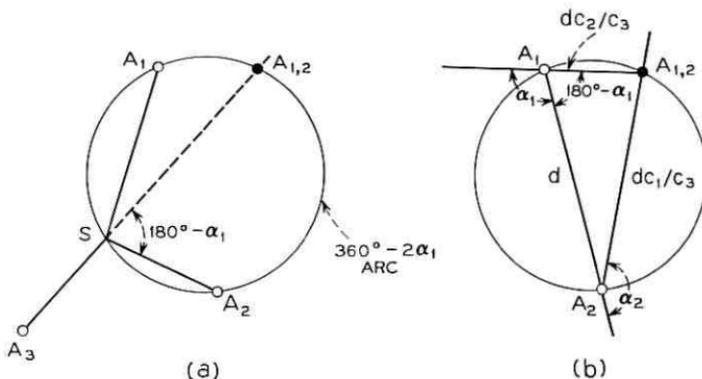


Fig. 6 — Construction of $A_{1,2}$.

S , then S is the desired Steiner point. If not, then there is no relatively minimal tree with the given topology. The best solution consists of A_1A_2 and A_1A_3 if A_2 and A_3 are on opposite sides of the line $A_1A_{1,2}$; use A_1A_2 and A_2A_3 if $A_2A_{1,2}$ separates A_1 from A_3 . Fig. 7 shows how the cheapest tree depends on the location of A_3 .

When the construction produces a legitimate Steiner point [Fig. 7(d)], Ref. 1 showed, in the ordinary case, that the length $|SA_1| + |SA_2| + |SA_3|$ of the tree is just $|A_3A_{1,2}|$. The appropriate generalization here is that the cost of the tree is the same as that of $|A_3A_{1,2}|$ miles of circuit costing c_3 dollars per mile, i.e.,

$$c_1 |SA_1| + c_2 |SA_2| + c_3 |SA_3| = c_3 |A_3A_{1,2}|. \quad (3)$$

The proof of (3) will follow from a theorem in Ptolemy's *Μεγάλῃ Σύνταξις* stating that the product of the diagonals of a quadrilateral

* In general, there are two such triangles. Construct the one which places $A_{1,2}$ and A_3 on opposite sides of the line A_1A_2 .

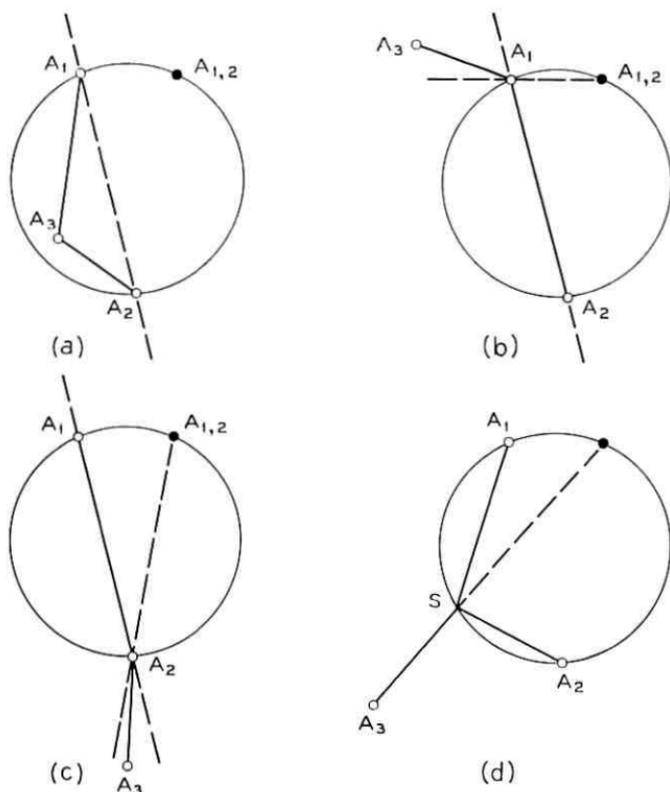


Fig. 7 — Second construction with $n = 3$.

equals the sum of the products of opposite sides.³ When applied to the quadrilateral $A_1SA_2A_{1,2}$ in Fig. 6 the theorem becomes

$$|SA_{1,2}| \cdot d = |SA_1| dc_1/c_3 + |SA_2| dc_2/c_3$$

or

$$c_1 |SA_1| + c_2 |SA_2| = c_3 |SA_{1,2}|.$$

Add $c_3 |SA_3|$ to both sides to get (3).

The construction of Fig. 7 may be used iteratively to find relatively minimal trees with $n \geq 3$ when each Steiner point is restricted to have only three incident lines.

The details are similar to the ordinary case¹ and so it suffices here to give an illustrative example. Fig. 8 shows cities A_1, \dots, A_5 to be interconnected by a graph having Steiner points S_1, S_2, S_3 . To locate

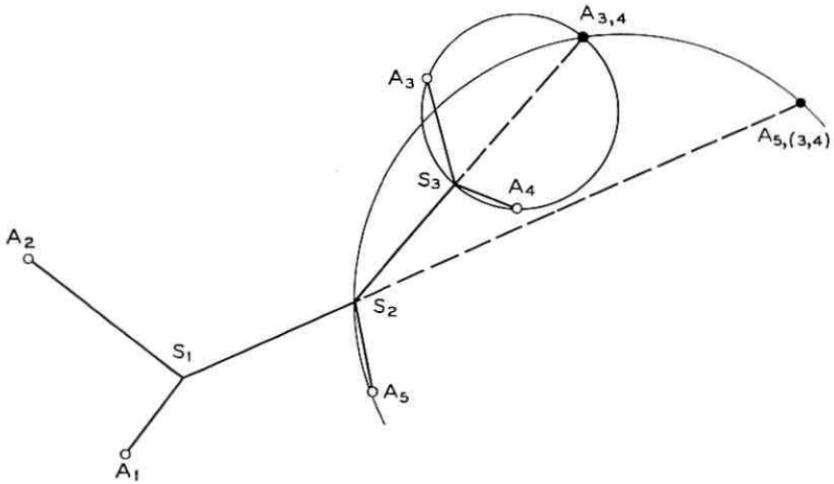


Fig. 8—A construction with $n = 5$.

the S_i one begins by finding a pair of cities which are to be connected to a common Steiner point; A_3 and A_4 will serve in Fig. 8. Construct $A_{3,4}$ as in Fig. 6(b), and draw the circle circumscribing A_3, A_4 and $A_{3,4}$. S_3 will be obtained ultimately by intersecting this circle with the line $S_2A_{3,4}$ [compare Fig. 7(d)] but at the moment the position of S_2 is unknown. Nevertheless, the problem is now reduced to drawing a new tree for $A_1, A_2, A_{3,4}$, and A_5 with Steiner points S_1 and S_2 (the cost per mile for the new line $S_2A_{3,4}$ is taken to be the same as the original cost per mile of S_2S_3). Again pick a pair of cities with a common Steiner point, say A_5 and $A_{3,4}$; draw the triangle with base $A_5A_{3,4}$ to construct a new point $A_{5,(3,4)}$. Now the problem reduces to drawing a tree for A_1, A_2 , and $A_{5,(3,4)}$. This is a case with $n = 3$ which is solved as described above. The solution locates S_1 . One can then locate S_2 on the line $S_1A_{(3,4),5}$ and finally, S_3 on the line $S_2A_{(3,4)}$.

In general, one has n cities A_1, \dots, A_n and at most $n-2$ Steiner points. By iterating the construction of Fig. 6(b) at most $n-2$ times one ultimately reduces the problem to a solvable case. There are three cautions to observe.

First of all, there are two triangles having a given base A_iA_j and given sides. The correct choice of triangle, and hence the correct $A_{i,j}$, is clear if one knows the location of the third point which shares the Steiner point of A_i and A_j . If this third point is itself a Steiner point and not yet located, one may have to try both possibilities for $A_{i,j}$.

However, if one can guess the correct choice of $A_{i,j}$ and then find a relatively minimal tree, the uniqueness result of Section III shows that one need not try the other choice.

Secondly, at some stage in the construction, one may find the situation shown in Fig. 7(a), (b), or (c) and so be unable to locate a Steiner point. This can happen either because no relatively minimal tree exists with the topology sought or because one of the $A_{i,j}$ was chosen wrong.

Thirdly, the construction described here produces only trees which have three lines at each Steiner point. A tree having Steiner points with four or more lines or a graph which is not a tree may be cheaper than the Steiner minimal tree in some cases.

V. SPLIT ROUTING

Unlike trees, which provide just one path between each pair of points, graphs with cycles offer a choice of paths. Then the $N(i, j)$ channels from i to j may be distributed over two or more paths (*split routing*). The example in Fig. 9 shows that split routing is sometimes economical. The three cities are at the corners of a unit equilateral triangle and the demands are $N(1, 2) = 13$, $N(1, 3) = N(2, 3) = 1$. The cost per mile for N channels is

$$f(N) = [(N + 2)/3].$$

Such a cost function might be encountered if channels are available only in cables containing 3 channels each; then $f(1) = f(2) = f(3)$, $f(4) = f(5) = f(6)$, etc. In Fig. 9(a) all channels follow direct paths in the complete graph. In Fig. 9(b) one of the channels from A_1 to A_2 has been rerouted through A_3 . This reduces the cost of the line A_1A_2 ;

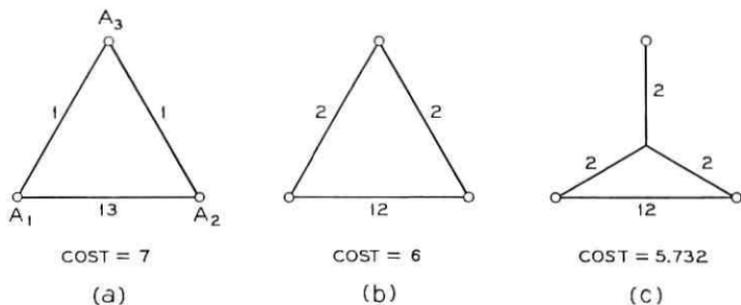


Fig. 9—Split routing.

it increases the number of channels in the other lines but does not increase their cost. Fig. 9(c) shows the minimal graph, which also uses split routing.

The remainder of this section will show that split routing gains nothing if $f(N)$ is a convex function, i.e., if

$$f(N+2) - 2f(N+1) + f(N) \leq 0 \quad (4)$$

for all N . Suppose $f(N)$ is convex and consider a network which uses split routing. Then one can find two channels, say α and β , which join cities A_i, A_j by different routes. To make cost comparisons easy, suppose that all other channels of the network have been installed and that the two channels for α and β have been installed on those lines of the graph which belong to both α and β . Now for $n = 0, 1, 2$ let $I_\alpha(n)$ be the incremental cost of installing n channels in each of the remaining lines of α and let $I_\beta(n)$ be a similar incremental cost for β . The cost to finish constructing the network is

$$\text{cost} = I_\alpha(1) + I_\beta(1). \quad (5)$$

However, $I_\alpha(n)$ is the sum of incremental costs of adding n channels to certain existing lines. If the k th line has $N_k[\alpha]$ channels

$$I_\alpha(n) = \sum_k \{f(N_k[\alpha] + n) - f(N_k[\alpha])\}.$$

Then (4) shows $f(N+2) - f(N) \leq 2\{f(N+1) - f(N)\}$, so

$$I_\alpha(2) \leq 2I_\alpha(1),$$

and similarly,

$$I_\beta(2) \leq 2I_\beta(1).$$

Now (5) shows

$$\begin{aligned} \text{cost} &\geq \frac{1}{2}\{I_\alpha(2) + I_\beta(2)\} \\ &\geq \text{Min}\{I_\alpha(2), I_\beta(2)\}. \end{aligned}$$

The last inequality shows that it would be as cheap to complete two copies of one of the channels α or β as to complete one of each.

VI. LINEAR COST FUNCTIONS

Suppose $f(N)$ is linear, $f(N) = a + bN$. Consider any graph. Let L_i denote the length of the i th line of the graph and N_i the number of

channels along that line. The cost of the network is

$$\text{cost} = aL + \sum_i N_i L_i b, \quad (6)$$

where $L = L_1 + L_2 + \dots$ is the total length of the graph.

A simple lower bound on the cost of networks which satisfy a given demand for channels may be obtained by bounding the two terms in (6) separately. The preliminary cost term aL is at least as large as aL_0 , where L_0 is the total length of the ordinary Steiner minimal tree connecting the given cities. The remaining cost in (6) would have been the cost of building the network if $f(N)$ had been bN . This cost is minimized by the complete network. Then

$$\text{cost} \geq aL_0 + b \sum_{i < j} |A_i A_j| N(i, j). \quad (7)$$

Another way of writing (7) uses two new quantities,

$$L_c = \sum_{i < j} |A_i A_j|,$$

(the length of the complete graph) and

$$\nu = L_c^{-1} \sum_{i < j} |A_i A_j| N(i, j)$$

(the average of the numbers of channels required between pairs of cities with the distance between cities as a weighting factor). Then (7), combined with the observation that the cost of the complete graph is an upper bound, becomes

$$aL_0 + b\nu L_c \leq \text{cost} \leq aL_c + b\nu L_c. \quad (8)$$

The form (8) of (7) is useful when numbers of channels which will be required between cities can be predicted only relatively but not absolutely. Then ν is a convenient measure of "traffic level".

The lower bound (8) is an instance of a more general inequality expressing a convexity property of the minimum cost function $c(\nu)$:

$$c(\nu) \geq \{(\nu_2 - \nu)c(\nu_1) + (\nu - \nu_1)c(\nu_2)\} / (\nu_2 - \nu_1) \quad (9)$$

for $\nu_1 \leq \nu \leq \nu_2$. According to (9) linear interpolation between known values $c(\nu_1)$, $c(\nu_2)$ gives a lower bound on $c(\nu)$. In particular, (9) becomes the left half of (8) in the limiting case $\nu_1 = 0$, $\nu_2 \rightarrow \infty$.

In the proof of (9) which follows it is convenient to extend the definition of $c(\nu)$ from a discrete set of ν values [at which all $N(i, j)$ are integers] to all positive real values. Although a line may require

a nonintegral number N of channels to satisfy traffic level ν exactly, its cost will be computed still at $a + bN$ dollars per mile. Now let $c(G, \nu)$ be the cost of providing channels for traffic level ν using graph G . In specifying G I intend that the location of any Steiner points be specified and not to depend on ν . Then $c(G, \nu)$ is a linear function of ν . Since

$$c(\nu) = \underset{G}{\text{Min}} c(G, \nu), \quad (10)$$

the region below the curve $c = c(\nu)$ is an intersection of the half-spaces lying below the lines $c = c(G, \nu)$. Then the region in question is convex and (9) follows.

The lower bound (8) is asymptotic to the minimum cost both for small ν and large ν . Even at intermediate values of ν the lower bound is reasonably accurate. For example, when there are three cities at the vertices of an equilateral triangle and ν channels are required between each pair of cities, the lower bound stays within 11.3 percent of the true minimum for all ν . The worst disagreement occurs when $\nu = (1+3^{1/2}) a/b$.

For a more realistic illustration, I took the four cities New York, Chicago, Houston, and Los Angeles and the numbers of channels given in Table II.

TABLE II—NUMBER OF CHANNELS BETWEEN CITIES

Cities	Separation (miles)	Number of channels
Houst.—L.A.	1374	x
Houst.—Chi.	940	$2x$
Houst.—N.Y.	1420	$4x$
L.A.—Chi.	1745	$5x$
L.A.—N.Y.	2451	$10x$
Chi.—N.Y.	713	$20x$

Here x is another parameter to specify traffic level; the average number of channels per pair of cities turns out to be $\nu = 6.52x$. The number of channels listed is nearly proportional to the product of the populations of the cities.* The cost function was $f(N) = 17,000 + 7N$ dollars per mile. The complete graph and ordinary Steiner minimal tree have lengths

$$L_c = 8,643 \text{ miles}$$

$$L_0 = 2,980 \text{ miles}$$

* N. Y. population includes Philadelphia; Chicago population includes Detroit.

so the lower bound is

$$50,660,000 + 394,400x$$

dollars. Table III compares this bound with the true minimum cost. Fig. 10 shows some of the minimum graphs. The upper bound in (7) differs from the lower bound by

TABLE III—COST OF MINIMUM GRAPHS (MILLIONS OF DOLLARS)

x	ν	Minimum cost	Lower bound	Discrepancy (percent)
30	195.6	63.2	62.5	1.1
50	326	72.0	70.0	2.2
100	652	93.2	90.1	3.4
200	1,304	135.2	129.5	4.2
500	3,260	260.4	247.9	4.8
1000	6,520	466.0	445.0	4.5
5000	32,600	2096.0	2022.7	3.5

$a(L_C - L_0)$, which in this example is about 240 million dollars. Then, for values of x larger than those shown in Table III the two bounds will agree to better than 4.6 percent.

Suppose one kind of technology, say coaxial cable, provides channels with a linear cost function

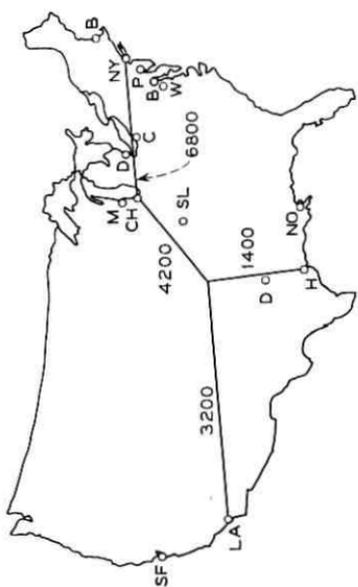
$$f(N) = a + bN$$

and suppose that a competing technology, say waveguide or microwave relay, has another linear cost function

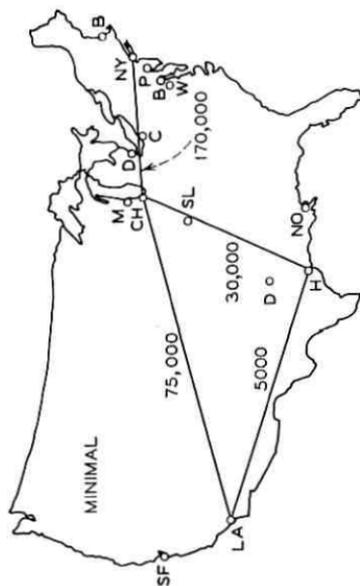
$$F(N) = A + BN.$$

Suppose that $a < A$ but $B < b$ so that the first technology is the more economical one to use if ν is small but the second is the more economical if ν is large. It is interesting to compare the two costs at various traffic levels and to find a value $\nu = \nu_0$ at which the two technologies are equally expensive.

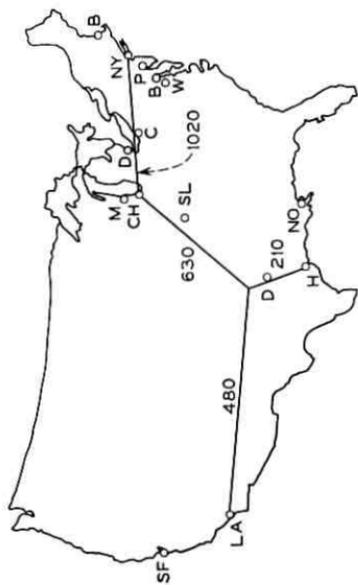
Suppose one computes minimal graphs and minimal costs $c(\nu)$, as in Table III, using the function $f(N)$. The corresponding minimal graphs and costs $C(\nu)$ for $F(N)$ may be obtained immediately by the following "scaling" argument. First, note that if $F(N)$ were just a multiple $\lambda f(N)$ of $f(N)$, the minimal networks in the two technologies would be identical and the costs would satisfy $C(\nu) = \lambda c(\nu)$. Secondly, note that if $F(N) = f(\mu N)$ for some multiplier μ , then the minimal



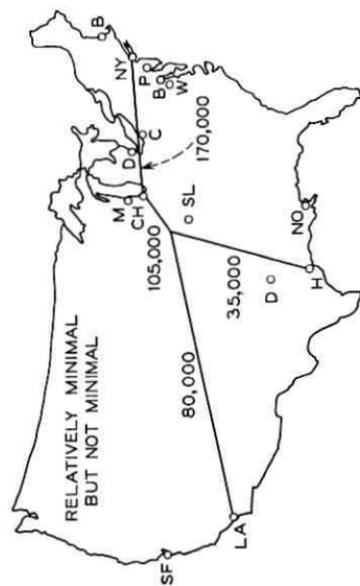
$x = 200$, COST = \$ 135,200,000



$x = 5000$, COST = \$ 2,096,000,000



$x = 30$, COST = \$ 63,200,000



$x = 5000$, COST = \$ 2,116,000,000

Fig. 10—Minimal graphs for three traffic levels (cf Table III).

network in the second technology is the same as the one which the first technology had at the traffic level $\mu\nu$; also $C(\nu) = c(\mu\nu)$. Since, in general,

$$F(N) = \lambda f(\mu N),$$

where $\lambda = A/a$, and $\mu = aB/(Ab)$, the two observations above combine to show that

$$C(\nu) = (A/a)c(aB\nu/(Ab)).$$

Moreover, the minimal graph for the second technology is the one found for the first at traffic level $aB\nu/(Ab)$.

To get a very rough estimate of the traffic level ν_0 at which the two technologies are equally expensive one might use the lower bound in (8) as an approximation to the minimal cost. Doing this provides the estimate

$$\nu_0 = (A - a)L_0/\{(b - B)L_c\}.$$

REFERENCES

1. Gilbert, E. N. and Pollak, H. O., Steiner Minimal Trees, to appear in SIAM J. Appl. Math.
2. Melzak, Z. A., On the Problem of Steiner, Canadian Math. Bu., 4, 1961, pp. 143-148.
3. Todhunter, Isaac, editor, *The Elements of Euclid*, Everyman, 851, London 1939. Book VI, Prop. D.
4. van de Lindt, W. J., A Geometrical Solution of the Three Factory Problem, *Mathematics Magazine* 39, 1966, pp. 162-165.

Contributors to This Issue

P. L. CLOUSER, B.S., 1960, Drexel Institute of Technology; Ph.D. (Physics), 1964, Duke University; Bell Telephone Laboratories, 1963–1967. Mr. Clouser was engaged in the study of esaki diodes by direct microscopic observation, esaki diode oscillators as power sources, and esaki diode phase lock oscillators. Member, IEEE.

E. N. GILBERT, B.S. (physics), 1943, Queens College; Ph.D. (Mathematics), 1948, MIT; Bell Telephone Laboratories, 1948—. Mr. Gilbert has been working in probability, combinatorial analysis, and information theory. Member, Amer. Math. Soc., IEEE.

J. E. GOELL, B.E.E., 1962, M.S., 1963, and Ph.D. (electrical engineering), 1965, all from Cornell University; Bell Telephone Laboratories 1965—. While at Cornell Mr. Goell was a teaching assistant and held the Sloan Fellowship and the National Science Cooperative Fellowship. At Bell Telephone Laboratories, he has worked on solid state repeaters for millimeter wave communication systems. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Phi Kappa Phi.

W. M. HUBBARD, B.S., 1957, Georgia Institute of Technology; M.S., 1958, University of Illinois; Ph.D., 1963, Georgia Institute of Technology; Bell Telephone Laboratories, 1963—. Mr. Hubbard's work has included analyses related to the design of millimeter-wave solid-state repeaters for use in a waveguide transmission system and the construction of prototype high-speed repeaters for this type of system. Member, Sigma Xi, Tau Beta Pi, Phi Kappa Phi, American Physical Society.

JAMES F. INGLE, B.E.E., 1955, Rensselaer Polytechnic Institute; M.E.E., 1961, New York University; Bell Telephone Laboratories, 1955—. Mr. Ingle was involved in designing frequency domain test equipment in the television and audio frequency ranges. More recently, he has been designing voice frequency automatic transmission measuring equipment. Member, Tau Beta Pi, Eta Kappa Nu.

JOHN J. KOKINDA, B.S.E.E., 1960, Purdue University; M.E.E., 1962, New York University; Bell Telephone Laboratories, 1960—. Mr.

Kokinda has been designing and developing manual voice frequency transmission measuring equipment and applying this equipment to maintaining telephone transmission plant. He is Supervisor of the Voice Frequency Measurements Group. Member, Eta Kappa Nu, Tau Beta Pi.

TIEN PEI LEE, B.S.E.E., 1957, National Taiwan University, Taiwan, China; M.S.E.E., 1959, Ohio State University; Ph.D., 1963, Stanford University; Bell Telephone Laboratories, 1963—. Mr. Lee participated in the research and development of solid-state microwave diodes and photodiodes. He is working on millimeter wave devices. Member, Sigma Xi, IEEE.

ROBERT W. LUCKY, B.S.E.E. 1957, M.S.E.E. 1959, and Ph.D. 1961, all from Purdue University; Bell Telephone Laboratories, 1961—. Mr. Lucky has been concerned with various analytical problems in the transmission of digital information over voice telephone facilities. He is Head of the Data Theory Department. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

GEORGE E. McLAUGHLIN, B.S.E.E., 1957, University of Rhode Island; M.E.E., 1959, New York University; Bell Telephone Laboratories, 1957—. Mr. McLaughlin has been involved in designing transmission measuring equipment in the television frequency range. He is Supervisor of the Automatic Transmission Measurements Group.

G. D. MANDEVILLE, 1933-34, Monmouth Junior College; 1935-36, Rutgers University; Western Electric Co., 1939-49; Bell Telephone Laboratories, 1949—. With Western Electric, Mr. Mandeville was concerned with radar development and shop test equipment. He headed the shop test equipment prove-in section for three years. With Bell Laboratories he has been associated with guided-wave research in the areas of waveguide and repeaters.

NORMAN A. MARLOW, B.S.E.E., 1960, M.S.E.E., 1961, both from the University of Southern California; Bell Telephone Laboratories, 1961—. Mr. Marlow has studied the effect of noise on voice-band data signals. Now he is concerned with research in systems and economic modeling, specifically to relate cost, reliability, and performance of communications equipment. Member, Inst. Math. Stat., Amer. Statistical Assn., Eta Kappa Nu.

HANS MELCHIOR, Dipl. El. Ing., 1959, Dr. Sc. Techn., 1965, both from Swiss Federal Institute of Technology, Zurich; Bell Telephone Laboratories, 1965—. From 1960 to 1965, with Department of Advanced Electrical Engineering, Swiss Federal Institute of Technology, he worked on noise problems of p-n junctions at breakdown, on high injection effects, and on second breakdown in diodes and transistors, and on tunnel diode mixers and oscillators. At Bell Telephone Laboratories he is working on avalanche photodiodes and on noise problems in metal-SiO₂-silicon devices. Member, IEEE.

INGEMAR NÅSELL, Civilingenjör, 1955, Royal Institute of Technology, Stockholm, Sweden; M.E.E., 1962, and M.S. (mathematics), 1965, both from New York University; Research Institute of National Defense, Stockholm, Sweden, 1955-1960; Bell Telephone Laboratories, 1960—. At Bell Telephone Laboratories he first worked on new noise objectives. He is supervisor of a group which conducts surveys throughout the Bell System to gather information for building a mathematical model of the Bell System toll network. Member, Svenska Teknologföreningen, and Eta Kappa Nu.

E. H. NICOLLIAN, M.E., 1951, Stevens Institute of Technology; M.A. (Physics) 1956, Columbia University; Bell Telephone Laboratories, 1957—. Mr. Nicollian's work has been in semiconductor device physics. He is currently engaged in research on the electrical properties of semiconductor-insulator interfaces. Member, American Physical Society, Electrochemical Society, RESA, AAAS.

VASANT K. PRABHU, B.Sc., 1958, Karnatak College, India; B.E.E., 1962, Indian Institute of Science; S.M., 1963, and Sc.D., 1966, both from Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu has been concerned with network theory, solid-state microwave power devices, optical communication, and noise theory. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, IEEE, AAAS.

E. N. PROTONOTARIOS, Electrical Engineer, 1963, National Technical University of Athens, Greece; Eng. Sc.D., 1966, Columbia University; Bell Telephone Laboratories, 1966—. At Bell Laboratories Mr. Protonotarios has been engaged in analytical studies of digital communi-

cation systems such as pulse code modulation and differential PCM. Member, Sigma Xi, IEEE, AAAS.

HARRY RUDIN, JR., B.E. 1958, M.Eng. 1960, D.Eng. 1964, all from Yale University; Bell Telephone Laboratories, 1964—. Mr. Rudin was an instructor in electrical engineering at Yale University from 1961 until 1964. At Bell Telephone Laboratories he has worked in data communication. Recently he has concentrated on automatic equalization and generalized equalization techniques. He is a former executive of the IEEE Connecticut section and is a member of the Yale Engineering Association executive board.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, synthesis and analysis of active and time-varying networks, studies of properties of nonlinear systems, and with a few problems in communication theory. His current interests are in the area of numerical analysis. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

R. C. SHAW, B.S.E.E., 1926, Michigan University; graduate studies, Columbia University and Stevens Institute of Technology; Bell Telephone Laboratories 1927-1945, 1947-1967; Chairman, Antenna Coordinating Committee, Office of the U. S. Secretary of Research and Development, 1945. Mr. Shaw has worked on developing high power radio transmitters for overseas telephone service, developing radar transmitters, methods of measuring radio field strength, and studies of antennas and propagation. Before his retirement this year he was helping develop A-1 mobile radio systems. Member, IEEE.

DAVID A. SHNIDMAN, B.S. and M.S. (electrical engineering), 1959, from Massachusetts Institute of Technology; PH.D. (applied mathematics), 1965, Harvard University; Data Sciences Lab at Air Force Cambridge Research Laboratory, 1959-1965; Bell Telephone Laboratories 1965—. At Cambridge he worked on communication theory and contract monitoring. At Bell Telephone Laboratories he is concerned with the efficient use of high quality transmission systems. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, IEEE, AAAS.

R. D. STANDLEY, B.S., 1957, University of Illinois; M.S., 1960, Rutgers University; Ph.D., 1966, Illinois Institute of Technology; U.S. Army Research and Development Laboratory, Fort Monmouth, N. J., 1957-1960; IIT Research Institute, Chicago, 1960-1966; Bell Telephone Laboratories, 1966—. At Fort Monmouth, Mr. Standley was project engineer on various microwave component development programs. His work at IITRI included microwave and antenna research, and management of an electromagnetic compatibility group. At Bell Telephone Laboratories he has been concerned with millimeter-wave up-converters, local oscillator injection filters, and channel dropping filters. He is investigating millimeter-wave impact ionization avalanche transit time diode devices, integrated circuits, and time delay equalizers. Member, IEEE, Sigma Tau, Sigma Xi.

WILLIAM D. WARTERS, A.B., 1949, Harvard College; M.S., 1950 and Ph.D., 1953, both from California Institute of Technology; Bell Telephone Laboratories, 1953—. Mr. Warters has done research in millimeter waveguide transmission and worked on repeaters for millimeter waveguide systems. He is Director of the Transmission Systems Research Center. Senior member, IEEE; member, American Physical Society, Sigma Xi, Phi Beta Kappa.

AARON D. WYNER, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, all from Columbia University; Bell Telephone Laboratories, 1963—. Mr. Wyner has been engaged in research in various aspects of information theory. He is also adjunct associate professor of electrical engineering at Columbia University and Chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group. Member, IEEE, SIAM, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

