THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XLI JANUARY 1962 NUMBER 1

An Experimental Pulse Code Modulation System for Short-Haul Trunks — C. G. DAVIS 1

A Bipolar Repeater for Pulse Code Signals — J. S. MAYO 25

PCM Transmission in the Exchange Plant — M. R. AARON 99

Performance Limitations of a Practical PCM Terminal — R. H. SHENNUM AND J. R. GRAY 143

A Companded Coder System for an Experimental PCM Terminal — H. MANN, H. M. STRAUBE AND C. P. VILLARS 173

Variational Techniques Applied to Phase-Controlled Oscillators — R. D. BARNARD 227

Ultimately Periodic Solutions to a Nonlinear Integrodifferential Equation — V. E. BENEŠ 257

Single-Server Systems—I. Relations Between Some Averages — S. O. RICE 269

Single-Server Systems—II. Busy Periods — S. O. RICE 279

Delay Distributions for Simple Trunk Groups with Recurrent Input and Exponential Service Times — L. TAKÁCS 311

The Transistorized A5 Channel Bank for Broadband Systems — F. H. BLECHER AND F. J. HALLENBECK 321

On the Use of Passive Circuit Measurements for the Adjustment of Variable-Capacitance Amplifiers — K. KUROKAWA 361

Contributors to this Issue 383

COPYRIGHT 1962 AMERICAN TELEPHONE AND TELEGRAPH COMPANY

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

H. I. ROMNES, *President, Western Electric Company*

J. B. FISK, *President, Bell Telephone Laboratories*

J. E. DINGMAN, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

A. C. DICKIESON, *Chairman*

A. J. BUSCH
L. R. COOK
R. P. CROSS
R. L. DIETZOLD
J. H. FELKER
K. E. GOULD
G. GRISWOLD, JR.
J. R. PIERCE
M. SPARKS
W. O. TURNER

EDITORIAL STAFF

G. E. SCHINDLER, JR., *Editor*
L. M. COLE, JR., *Assistant Editor*
C. POLOGE, *Production Editor*
J. T. MYSAK, *Technical Illustrations*
T. N. POPE, *Circulation Manager*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y., E. J. McNeely, President; Allen G. Barry, Vice President and Secretary; L. Chester May, Treasurer. Subscriptions are accepted at $5.00 per year. Single Copies $1.25 each. Foreign postage is $1.08 per year or 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM

TECHNICAL JOURNAL

VOLUME XLI JANUARY 1962 NUMBER 1

Copyright 1962, American Telephone and Telegraph Company

An Experimental Pulse Code Modulation System for Short-Haul Trunks

By C. G. DAVIS

(Manuscript received August 30, 1961)

An experimental 24-channel pulse code modulation system employing solid-state devices is described. Economy of design and ability to operate over existing exchange cable were dominant factors in selection of system organization and circuit alternatives. System requirements are unique to this type of system. Design considerations in meeting these requirements are presented in brief here and in detail in companion articles.

I. INTRODUCTION

The Bell System has a wealth of experience in frequency division systems. Through the years, as new systems have been developed, our knowledge of the factors affecting the performance of these systems has steadily increased. There has been no comparable experience with time division systems. Therefore, when the decision was made to develop a pulse code modulation system for commercial use, it was felt advisable to build an early experimental model to prove the feasibility of the general system arrangement and the circuit approaches adopted. The experimental system also established the over-all level of transmission performance that could be expected.

This and companion articles in this issue describe the experimental system, analyze the results of measurements on the actual system, and compare performance with objectives.

II. SYSTEM APPLICATION AND FEATURES

Development of a pulse code modulation (PCM) system was undertaken to answer the need in the Bell System for a carrier system economical for short distances of less than ten to more than 25 miles, working on exchange cable pairs. A carrier system for short distances must, of necessity, have inexpensive terminals. As will be shown, the PCM system achieves economy in the terminals by dint of a high percentage of common equipment, the cost of this common equipment being spread among all the channels. The close spacing of repeaters, resulting from the wide band of frequencies occupied by the PCM signal, sets an upper limit to the distance for which the system will be attractive.

A summary of the features of this system is given below:

Terminal:

Number of speech channels = 24.

Seven-digit binary code expresses amplitude of speech samples.

Instantaneous compandor reduces noise and crosstalk by 26 db.

Built-in signaling system uses eighth digit assigned to each channel and seventh digit during on-hook period for revertive pulsing.

Only solid-state devices are used.

Three terminals may be mounted in a 23-inch by 11-foot, 6-inch bay.

Repeatered Line:

Designed for use with 19- or 22-gauge cable pairs.

Nominal repeater spacing = 6000 feet.

Regenerative repeaters receive timing information from the pulse pattern.

Pulse repetition rate = 1.544 megabits per second.

Power supplied over phantom circuit.

Bipolar pulse pattern used to reduce base line wander and to reduce timing frequency crosstalk between systems.

This article will give a general description of the system and the design choices. Accompanying articles go into greater detail about the design and performance of critical areas of the system, viz, the coding complex and regenerative repeaters.

III. SYSTEM ORGANIZATION

3.1 *Speech Processing*

A block diagram of the speech portion of the system is shown in Fig. 1. Incoming speech to a channel unit, after passing through the hybrid,

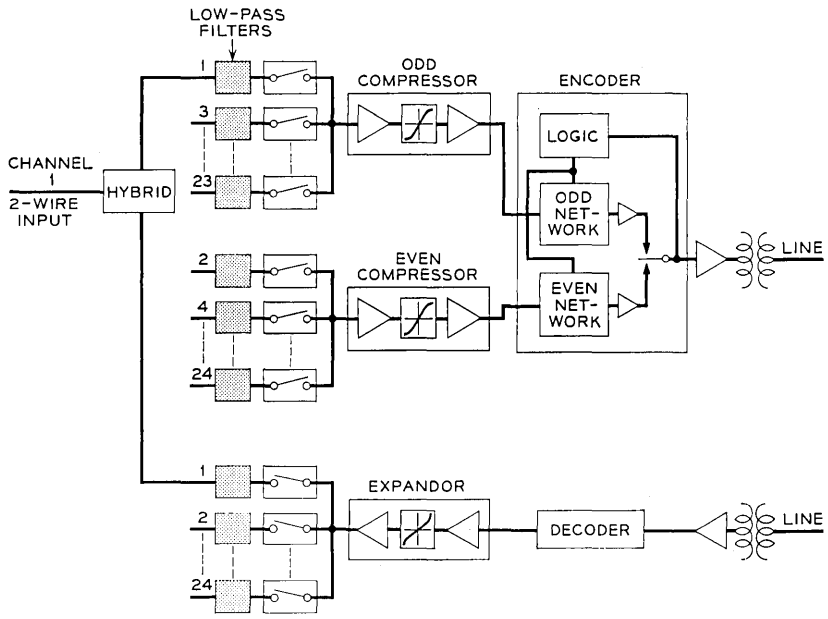


Fig. 1 — Experimental PCM system — speech portion of terminal.

is band-limited to reject all frequencies above 4 kc. This band-limited signal is sampled 8000 times per second by the sampling gate associated with this channel. The resultant sample, whose amplitude is proportional to the signal level at the instant of sampling, is passed through the compressor, which gives preferential gain to low-level signals, and presented to the coder. The coder expresses the sample amplitude as a seven-digit binary number or one of 128 different possible levels. The first digit has weight of 64; the last digit has weight of 1. The signal arrives at the coder on a pedestal 64 units high so that code 64 corresponds to zero signal amplitude. The seven-digit code goes onto the transmission line, followed by an eighth time slot which carries the supervisory signaling for that channel.

The channels are sampled in a recurring sequence, one sample from each channel or 24 samples being encoded and transmitted every 125 μsec. Since each sample requires eight time slots including the signaling, these 24 samples require a total of 192 time slots on the line. An additional or 193rd time slot is added to permit synchronizing or framing the two ends of the system. These 193 time slots comprise a framing period. This is illustrated in Fig. 2. There being 8000 such periods each second, the repetition rate of pulses on the line is 1.544 million pulses per second.

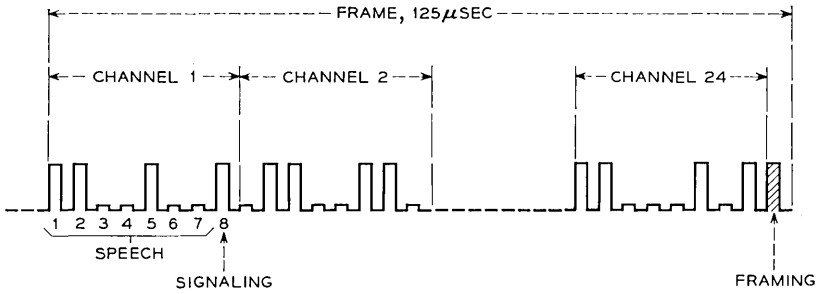


Fig. 2 — System time assignment.

The time assigned to one bit is about 0.65 μsec. The pulses have 50 per cent duty cycle and are therefore 0.325 μsec wide.

These pulses, containing speech, signaling and framing information, arrive at the receiving terminal after being reconstituted several times by the regenerative repeaters spaced at 6000-foot intervals on the line. At the receiver, the pulses are sorted out, the signaling pulses being directed to the individual channel signaling units and the speech-coded signals going into the decoder. The decoder output is a pulse amplitude modulated (PAM) signal whose amplitude is equal to the amplitude of the input to the coder, within one-half coder step. The decoder PAM pulse passes through the expandor, which has the inverse characteristic of the compressor, providing more gain for higher-level signals. The expandor is followed by a wide-band power amplifier which raises the signals to a level sufficient to require no further amplification after being switched to the channel units. A low-pass filter in the receiving section of the channel unit integrates the samples to yield the original signal.

With this general description in mind, let us proceed to a more detailed consideration of the circuits used to process the signal.

3.1.1 *Multiplex Gates*

Each of the 24 multiplex gates is operated 8000 times a second by a "channel pulse" from a control circuit named the "channel counter." The operation of the gate results in a pulse, whose amplitude is proportional to the signal level in that channel, being presented at the input of the compressor. This pulse must be held until coding is completed.

It is important that the signal stand still while it is being encoded, for if even small ripples synchronous with the digit rate are superimposed on the signal, the coder will spuriously encode one of the levels the sig-

nal reaches momentarily during the coding process. This can have the effect of enlarging the size of some coder steps by attenuating or actually eliminating others. The problem of crosstalk from control signals is a real one because of the very high level of the control pulses in the terminal. Since the effects of crosstalk and noise on the common bus are also a function of the signal level, it is advisable to keep the signal level as high as possible before encoding so that interference will have the smallest effect. For this reason voltage sampling was not considered satisfactory, and an energy-sampling approach was used. As is well known, voltage sampling results in signal attenuation equal to the percentage of time the gate is closed, whereas energy sampling[1] theoretically results in no attenuation. The energy sampling gate is shown in Fig. 3. Between sampling periods, the capacitor C1 is charged to the signal level. When the gate is closed, the complete charge of C1 is transferred through the inductor L to the common capacitor C_c. The gate is held closed for exactly one-half the resonant period of L with C1 and C_c, or 1.95 μsec (the width of the "channel pulse" driving the gate). After a signal has been encoded, the clamp is closed for 1.95 μsec to remove this signal from the common bus before the arrival of the next signal. The input impedance of the compressor is high to keep the voltage on C_c essentially constant through the coding period.

Use of the resonant transfer process results in an economy by making it unnecessary to use an input amplifier per channel. The choice of diode sampling gates is largely one of economics. The gates must, in all cases, be driven on hard enough that they can carry the maximum required signal current. If transistor gates are used, the control pulse does not need to be so powerful. If diode gates are used, the full gain requirement is placed on the drive pulse. Since it was possible to derive the required

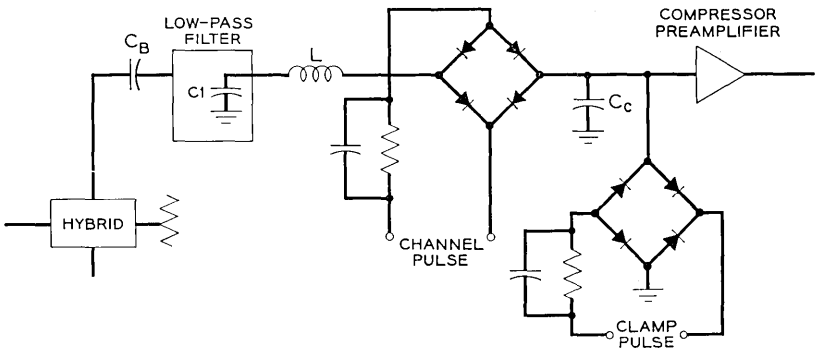


Fig. 3 — Multiplex gate and clamp.

amount of drive from the channel counter, and since this arrangement proved the most economical, diode gates were employed.

There is a balance requirement on the diode gates to minimize the pedestal associated with the operation of the gate. This is necessary to keep the signal centered on the coder characteristic. As R. H. Shennum and J. R. Gray[2] point out, any quiescent displacement from the center of the compressor curve results in increased idle circuit noise and cross-talk. The balance requirement is eased considerably by the capacitor C_B which builds up a dc potential equal, but opposite in polarity, to the average gate unbalance voltage.

3.1.2 *Compandor-Coding Circuits*

A companion article[3] describes the operation of the companding and coding portions of the system and the reasons for choosing the approach used. Suffice to say at this point that the purpose of the instantaneous compandor is to provide better definition (less quantizing* noise) for low-level signals at the expense of poorer definition for high-level signals, where quantizing noise is not so noticeable. The circuit is designed to provide 26 db of companding, which means that the coder steps near zero signal level, as reflected at the compressor input, are 20 times smaller than they would be without companding. This is important both from a quantizing noise standpoint and because the noise floor in such a system is equal to one coder step. The effect of the compandor is to make the seven-digit coder equivalent to an eleven-digit coder for small signals. The compandor achieves its nonlinear characteristic by the use of diodes whose forward voltage-current characteristic is logarithmic. The diodes shunt the signal path in the compressor and are in series with it in the expandor.

A network coder, shown schematically in Fig. 4, is used. This coder successively compares the input PAM signal to binary weighted currents and generates the PCM signal as the comparison is taking place. The switches 1 through 8 are swung from ground to battery in sequence under the control of leads from the digit generator (see Section 3.2). Each switch closure subtracts an amount of current from the current-summing point proportional to the conductance in series with the switch. Thus, closing switch No. 1 subtracts 64 units of current; switch No. 2 subtracts 32 units; etc. Whether a switch remains operated through the remainder of the coding operation or is released depends on whether current flows into or away from the summing amplifier as a result of the switch closure.

* Quantizing noise is the noise introduced by coding error which results from the finite size of coder steps.

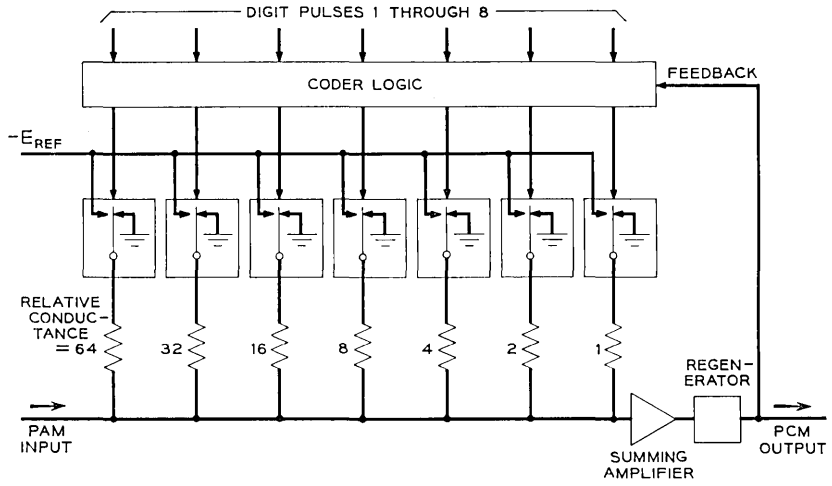


Fig. 4 — Simplified block diagram of network coder.

Operating the first switch, then, determines whether the signal is greater or less than 64. If it is greater, that switch is left closed, a space is put on the line, and the next comparison determines whether the signal exceeds 96. If the signal is less than 64, a pulse is generated, the first switch is released, and the signal is then compared to 32. This process is continued through the seven digits, a pulse being generated each time the PAM signal is less than the sum of the reference currents. As mentioned previously, code 64 corresponds to zero signal level.

The logic portion of the coder uses flip-flops to keep selected switches operated. All the flip-flops are reset by the eighth digit from the digit generator.

This type of coder demands that the input signal be present throughout the coding interval, which leads to the reason for having two compressors. A given PAM signal must be present throughout the coding interval of about 4.5 μsec, and then within 0.65 μsec that signal must be replaced by another. Rather than place this speed requirement on sampling gates and compressor, it was decided to switch the coder between two inputs and to sample on an odd-even channel basis, as shown in Fig. 1. To accomplish this switching at a digital rather than an analog position, the analog portions of the coder are duplicated and the logic section alternately put under the control of the two sections. This arrangement allows adequate time for removal of the previous signal from each compressor input and settling of the desired signal before encoding.

An alternative way to provide more guard space between channels is to use a coder which encodes at a 3-mc bit rate but reads out at a 1.5-mc rate. Such a coder, which would require a 50 per cent duty cycle of incoming PAM samples, is functionally quite simple and compares favorably in cost to the dual coder. The speed requirements for coder and compressor, however, would be doubled by such an approach.

The decoder is similar to the coder in its operation. In fact, the coder actually contains a decoder in its feedback path, as shown in Fig. 5. In response to a pulse in a given time slot on the PCM line, a reference voltage is applied to a binary-weighted resistor. This supplies into a summing point a current proportional to the weight of the pulse. Pulses from the PCM line are steered to the correct switches by logic gates using pulses from the receiving digit pulse generator (see Section 3.2).

Since all the weighted currents must be present at the time of demultiplexing, the decoder must contain storage or delay circuits to convert the serial PCM code to parallel. Capacitor storage is used for this purpose in the decoder, this approach yielding the most economical circuit. Other choices considered were blocking oscillators and flip-flops.

3.1.3 *Demultiplexing*

The pedestal applied to the PAM signal at the transmitter is removed following the decoder. This results in bipolar PAM at the receiver which has advantages in minimizing interference, as pointed out in Section IV.

Since it was not possible with available devices to derive all the re-

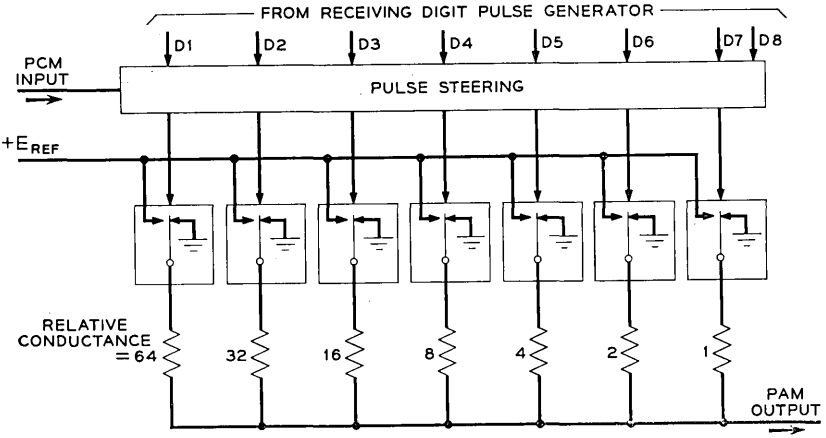


Fig. 5 — Simplified block diagram of decoder.

quired power at the receiver from the decoder and pass this power through the expandor, some signal gain must be supplied following the expandor. For economic reasons it was decided to provide this gain at a common point rather than furnish an amplifier per channel. This decision to amplify the decoded signals before distribution to the receiving channel filters results in a need for a broadband power amplifier and high-level demultiplex gates. These circuits are described in the following sections.

3.1.3.1 *Common Amplifier.* Only one expandor and common amplifier are required because the PAM pulses in the receiver do not have to occupy a full channel time slot as in the transmitter. In the receiver, the PAM pulses are 3.25 μsec wide with 1.95 μsec guard space between them.

The common amplifier is designed to provide a peak current of 300 ma at 13 volts. To obtain this high power output from a broadband amplifier, a diffused silicon transistor was used in class B operation in the manner shown in Fig. 6. Transistor Q2 with equal resistors R in emitter and collector circuits serves as a unity-gain phase shifter to drive the

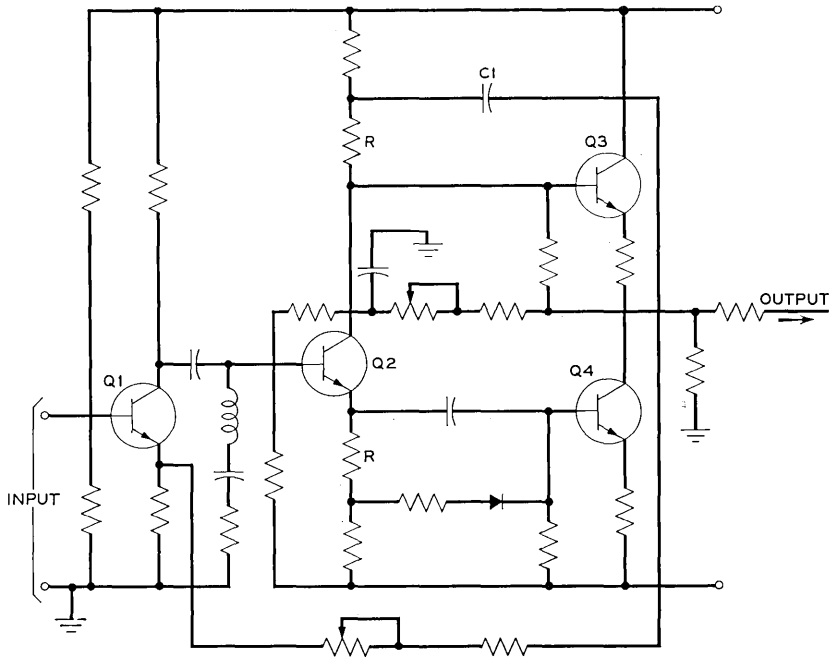


Fig. 6 — Common amplifier.

class B output stages. The capacitor C1 assures that the drive to Q3 is the same as that to Q4. Feedback is provided from the output to the emitter of the preamplifier transistor Q1.

The stabilization of this amplifier is complicated by the time variance of the load. During turn-on of the gate, the load is essentially resistive. Later in the cycle, the load becomes an RC combination. Between pulses, the load is an open circuit. To circumvent this problem, the amplifier was designed to have adequate gain and phase margin under all load conditions.

3.1.3.2 *Demultiplex Gate.* The demultiplex gate configuration is shown in Fig. 7. For large signal difference between common amplifier and filter input, two of the gate diodes are turned off, and the filter is charged by a constant current determined by the transistor drive and its current gain. As the signal levels become nearly equal, all diodes are forward biased by the drive pulse, and the circuit functions as a normal balanced gate. Constant current charging to near the signal level results in essentially full transfer of the signal voltage in the time available. This renders any amplitude modulation due to variations in width of the gate drive unnoticeable, and also provides good control of the net loss of the demultiplex gate.

Further advantages of the balanced demultiplex gate are mentioned in Section IV.

3.1.3.3 *Receiving Filter.* The decision to use a common amplifier resulted

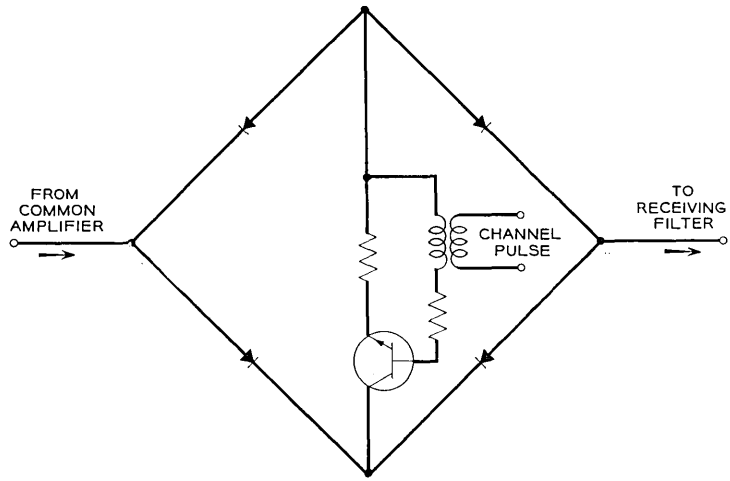


Fig. 7 — Demultiplex gate.

in loss of buffering between the demultiplex gates and hybrids. This results in the hybrid looking into a filter terminated by a time-varying impedance, as shown in Fig. 8. Balancing the hybrid to meet return loss requirements at all frequencies of interest requires that the impedance seen by the hybrid be one obtainable with passive elements.

M. R. Aaron found that the time-varying impedance could be represented by a simple series RL combination whose values were only slightly frequency dependent. Using average values of R and L, it was possible to match a constant K filter to this impedance so that a reasonable impedance was presented to the hybrid.

3.2 *Control Circuits*

The basic pulse repetition rate of the system is controlled by 1.544-mc clocks. These clocks take the form of a crystal oscillator in the transmitting terminal and slaved oscillator in the receiving terminal. The slaved oscillator is synchronized with the incoming pulse train from the repeatered line.

In addition to these clocks, control circuits are needed to define the eight time slots in a channel and the 24 channel slots in a frame. These control circuits, referred to in previous sections, are called digit pulse generators and channel counters. The digit pulse generators control the coder and decoder. The channel counters control multiplexing and demultiplexing of PAM signals. The two circuits together control the signaling circuits.

The following sections describe these two control circuits.

3.2.1 *Digit Pulse Generator*

Each terminal contains a transmitting and receiving digit pulse generator, synchronized by the transmitting or receiving clock. The digit pulse generator provides a pulse sequentially on each of eight outputs. It is of a self-starting ring counter design shown in Fig. 9.

The basic building block of this circuit is a blocking oscillator whose

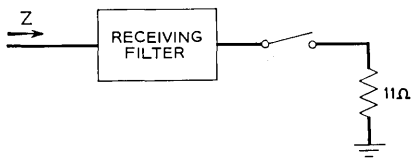


Fig. 8 — Time-varying impedance seen by hybrid.

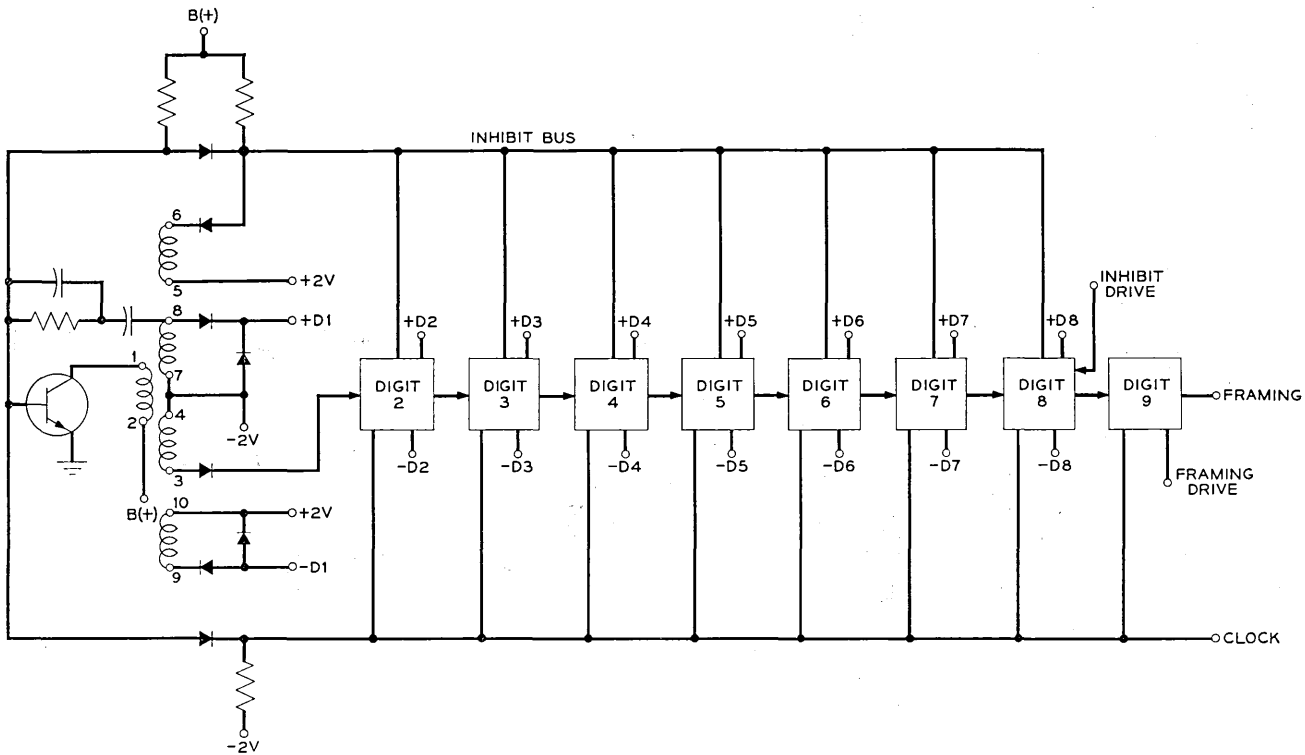


Fig. 9 — Transmitting digit pulse generator.

input is combined with the 1.544-mc clock in an AND circuit. In the presence of positive input signal, the blocking oscillator fires when the clock goes positive. The natural period of the blocking oscillator is greater than a half cycle of the clock, so that the turnoff of the blocking oscillator is initiated by the negative-going transition of the clock. The output pulse length is about 0.325 μsec. Both polarities of this pulse are available from the blocking oscillator.

Nine such blocks are used in the building of a digit pulse generator. As can be seen from Fig. 9, the output of each stage triggers the next stage by utilizing the positive overshoot of a negative-going pulse. This is one way of obtaining the required delay between stages.

A pulse is started through the ring counter by the firing of the first stage. Since we desire to have only one output active at a time, the first stage must be inhibited from firing again until the last stage has become active. This inhibiting action is accomplished by using the negative overshoot of the positive pulse from all but the eighth stage. The inhibit output from the eighth stage is used in channel 24 to make the digit pulse generator count to nine in that channel (see the inhibit drive lead), providing a time slot for framing. Under control of the framing drive lead, pulses are generated in this ninth time slot in alternate frames.

3.2.2 *Channel Counters*

Each terminal contains also a transmitting and receiving channel counter. The channel counters are similar to the digit pulse generators in that they use blocking oscillators as building blocks and pulse overshoots for delay. A channel counter, however, counts to 24, generating a pulse on each of 24 output leads in sequence. The pulse width is controlled not directly by the clock, but from a secondary "clock" derived from the digit pulse generator outputs. The pulse widths from the two channel counters are different, as pointed out in the multiplex and demultiplex gate sections. The transmitting channel counter generates 1.95-μsec pulses; the receiving channel counter generates 3.25-μsec pulses.

3.2.3 *General Design Considerations*

At the time this experimental system was built, the principal alternatives available for accomplishing these control functions were multivibrators, magnetic cores and blocking oscillators. Since the power and speed requirements were rather severe for both types of counter, and

since transformer coupling was desired, blocking oscillators were employed.

3.3 *Synchronization*

The pulse repetition rate in the receiving terminal must, of course, be exactly the same as that in the transmitting terminal. This synchronization is obtained by deriving the receiving terminal clock frequency from the incoming pulse train, just as is done at each regenerative repeater. Further synchronization between the two terminals is required to permit decoding of the speech and signal information. This synchronization must permit identification of each time slot in each channel. Since these time slots recur on a 125-μsec frame basis, this synchronization will be called "framing."

Basically, there are two types of framing systems — forward-acting and backward-acting. A forward-acting system transmits at the beginning of each frame a unique signal which cannot be encountered elsewhere in the pulse train. Due to the restrictions of the regenerative line, there can be no information in the pulse width or amplitude; therefore, this would have to be a unique code for the system herein described. The receiver control circuitry would start counting from this signal to steer the pulses to their correct destinations. A forward-acting framing circuit has the advantage of very fast action, since the system reframes each frame. Disadvantages are (1) expense, because the unique code must consist of many characters and therefore complex circuitry would be required at both ends, and (2) frequency of framing loss. Each time a pulse error is made during transmission of the framing pattern, the system is out of frame and each customer hears a transient noise.

A backward-acting system also requires a framing code to be transmitted each frame. Instead of reframing each new frame, however, the backward-acting system gets into frame and simply checks with each frame to ascertain whether it is still in frame. In this manner, the backward-acting framing system does not require an absolutely unique code and, furthermore, can be designed to ignore one or more transmission errors by insisting on a given number of errors in a specified time before deciding that the system is out of frame. Once having ascertained that framing is lost, the backward-acting system moves through the various pulse positions looking for the framing code. When it finds the framing code, it locks to it and normal operation is resumed.

The amount of time required for the backward-acting system to reframe depends on the probability of its being fooled by encountering the framing code by chance in the wrong portion of the pulse train. This probability decreases as the length of the framing code is increased. Thus,

even with the backward-acting system, reframing is accomplished more quickly with a more nearly unique framing code.

The objectives for the framing system for the PCM system were (1) the system should not go out of frame on single errors in transmission, and (2) the reframing should be rapid, consistent with low cost. The first objective leads to a backward-acting framing system. The ultimate choice of code for simplicity and therefore economy is a single digit. Since a pulse or space could exist for very long times in any position in the frame, however, neither of these can be used exclusively. Consideration of the coding procedure, though, reveals that an alternating pulse-space pattern cannot exist for long in any pulse position. This is true because the alternating pattern implies a 4-kc component in a signal and the input filters do not pass 4 kc. This alternating single pulse pattern was thus chosen. The 193rd time slot in each frame contains alternately a pulse and a space which the receiving framing circuit locks onto. The generation of this pattern is explained in Section 3.2.1. A time rate of errors in excess of a predetermined level is interpreted as loss of framing. In this event, the receiving framing circuit moves to the preceding time slot and looks for an alternating pattern. One violation of the alternating pattern is sufficient to drive it to the following time slot, etc., until a time slot is found which contains an alternating pattern for more than eight frames. To speed up the reframing process, the adjacent pulse position is looked at immediately upon violation of the alternating pattern, which means that it may be necesary to remain on a given position only 125 μsec to ascertain that it is not the 193rd position. The circuitry for accomplishing this motion from one pulse position to the next is not shown.

The framing circuit requires 0.4 to 6 milliseconds to detect an out-of-frame condition. The time required to reframe depends, of course, on how far away from the framing position the circuit is when it starts hunting. A simple calculation shows that even if it has to go through all 192 positions, only about 50 milliseconds will be required for the system to reframe.

The detection of the framing signal is illustrated in Fig. 10. A pulse on the SP lead once per frame changes the state of the binary cell when the system is in frame. The compare circuit looks for the alternating pulse pattern in the 193rd pulse position.

3.4 *Signaling*

Any transmission system must make provision for the signals necessary to establish a connection and to end the connection at the termina-

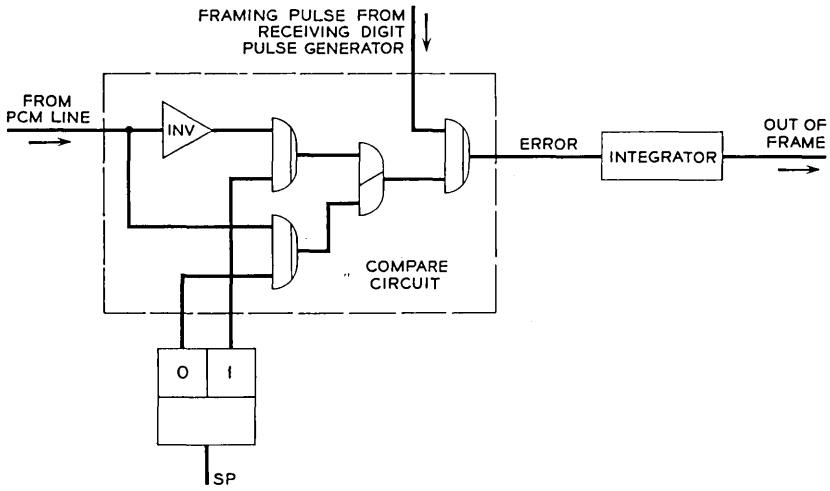


Fig. 10 — Framing detector.

tion of the call. The nature of the PCM signal makes it most economical to convert the dc signals containing this signaling information into digital form for transmission over the line and to convert the digital signals back into the desired dc conditions at the distant end.

As mentioned previously, the eighth time slot in each channel is assigned to signaling. To multiplex the signaling information from the various channels into their respective signaling time slots requires the use of simple scanning gates as shown in Fig. 11. The dc signaling information contained in the eighth time slot of each channel is demultiplexed at the receiving terminal, and the resulting pulse train is integrated to recapture the original dc signal.

If revertive pulse signaling is used between offices, two signal paths are required in one direction, one for supervision and one for revertive pulsing. In this case, both the seventh and eighth pulses of the channel are used — the eighth for supervision and the seventh, or least significant speech digit, for revertive pulsing. Except for special-purpose calls to operators, this has no effect on speech because revertive pulses are sent only before the called party answers. The seventh coder pulse is, of course, suppressed when the seventh pulse is being used for signaling.

3.5 *Power Supply*

The dc power for the terminal is furnished from a central regulated power supply. The required voltages are derived from a dc to dc con-

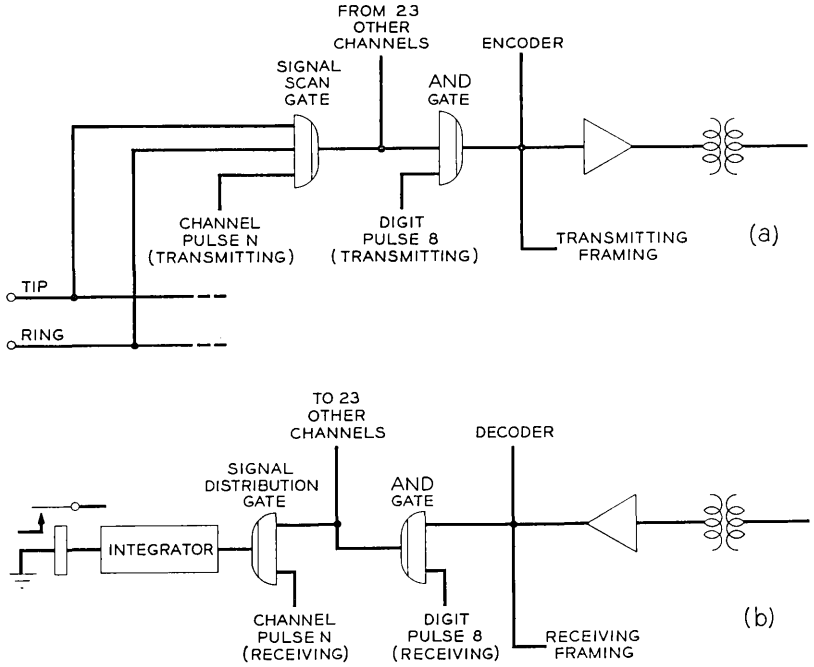


Fig. 11 — Experimental PCM system — signaling.

verter driven from the office 48-volt battery. Transistor regulators restrict absolute voltage level variations and ripple to permissible values.

3.6 *Repeatered Line*

The repeatered line is described in detail in a companion article.[4] It will therefore be treated very briefly here.

Repeaters were designed for 6000-foot spacing because this is a standard load-coil spacing. Build-out sections are provided for shorter spacing. The repeaters are fully regenerative and receive their timing information for the regenerative process from a resonant circuit driven by the incoming pulse train. Power for the repeaters is supplied over the phantom circuit.

While a pulse transmission system is relatively immune to noise, it is quite vulnerable to interference at the timing frequency. With unipolar pulses, crosstalk between timing waves was a problem if several systems were to be put into the same cable. Increased margin against this crosstalk was obtained by inverting alternate pulses on the line, thereby con-

centrating the bulk of the signal energy in the vicinity of half the pulse repetition rate. The timing frequency is derived by full-wave rectification of the bipolar pulse train. The advantage lies in the fact that the forward loss at the lower frequency is less, and the coupling loss greater, than for the 1.544-mc component. The bipolar signal also has minimum dc content, thereby restricting base line wander, which would otherwise reduce the signal-to-noise ratio.

Remote fault location and marginal checking of repeaters are accomplished by using a single fault-location pair for as many as 25 systems in a cable. Each repeater location is identified by the frequency to which a filter between its output and the fault-location line is tuned. The details and use of this system of remote testing are covered in Ref. 4.

The high bit rate of the PCM line provides obvious capability for data transmission. The entire line could be used for one broadband data channel, or time division methods could be used to subdivide the pulse train to accommodate several data channels or a mixture of data and voice channels.

IV. SYSTEM PERFORMANCE

The ultimate gauge of success of a speech communication system is determined by its subjective acceptability. Experience with frequency division systems has made it possible to spell out performance objectives which assure subjective acceptability. The signal impairments introduced by a PCM system are quite different from those introduced by a frequency division system. It is not realistic therefore, to set the objectives in the same way. A primary goal of the experimental system was to determine economically attainable system performance and to judge whether the speech quality of the system would be acceptable. Results of tests on our experimental system indicated satisfactory performance with the following design parameters.

i. Channel overload — The system presents an undistorted voltage range equal to the peak-to-peak excursion of a +3-dbm sinusoid at the zero-db system level.

ii. Idle circuit noise — Did not exceed 22 dba at the zero-db system level as measured with a 2B noise-measuring set with F1A weighting.* Performance was generally appreciably better.

iii. Idle channel crosstalk — Did not exceed −65 dbm on any channel when 15 dba of thermal noise power was applied at the input to that channel and a zero-dbm, 1-kc tone applied to any other channel.

* 28 dbrn (C message) in terms of the new 3A Noise Measuring Set.

iv. Transmission quality of quantized signals — The plot of measured signal-to-noise ratio for any channel did not fall more than 3 db below the theoretical value (shown in the article noted in Ref. 3) at any volume within 50 db of full sinusoidal excitation.

Unlike a frequency division system, essentially all the noise, distortion, and gain variation in a PCM system occurs in the terminals. The principal source of noise is quantizing error. The relation of quantizing noise to signal level depends on the number of coder steps used to describe the signal and the distribution of amplitudes of these coder steps. As described by H. Mann, et al,[3] the signal-to-quantizing noise ratio for the effective coder characteristic used can be calculated both for the ideal case and with random variation of step size.

The quantization of the signal has the further effect of placing a floor under crosstalk and idle circuit noise. This is true because, without the addition of complex circuitry, the zero-level signal voltage may be such that an infinitesimally small signal can trigger a coder step. The minimum guaranteed noise or crosstalk is thus that corresponding to one coder step. Since the coder range must accommodate up to the overload signal, this relates the crosstalk and idle circuit noise to overload level through the number of coding steps available. Such crosstalk would have the quality of infinitely clipped speech, which has been found to be remarkably intelligible. Interestingly, this same feature can eliminate all low-level crosstalk and noise from the transmitter if the zero-level signal sits at the middle of a coder step.

The principal source of crosstalk in a time division system is generally due to the carryover of the PAM sample on a common bus into the following time slot. To meet the crosstalk requirements, it is necessary that the time constant of all common points between samples be sufficiently small. This condition is met on the transmitting common bus by use of a low-impedance clamp between pulses, and on the receiving common bus by the low output impedance of the common amplifier.

Another crosstalk path is through the capacitance and reverse impedance of the back-biased diodes of the multiplex or demultiplex gates. This path can permit crosstalk from any one channel to all channels or from all channels to one channel.

The 15 dba of thermal noise specified in the discussion of the crosstalk objective is intended to scramble the zero crossings of the tone, thereby reducing the enhancement of weak crosstalk. Such noise may be expected to be present on any channel in service.

Subjective tests have shown characteristic PCM idle circuit noise to be more objectionable than equivalent thermal noise. The 22-dba quan-

tizing noise objective is the subjective equivalent of about 26-dba thermal noise.

Net loss variations can be attributed to any circuit in the signal path. The usual compandor enhancement of gain variations must be considered in assigning margins to compressor post-amplifier, expandor preamplifier, coder and decoder.

The noise, crosstalk, net loss, and distortion requirements are considered in the accompanying article by R. H. Shennum and J. R. Gray,[2] and the allocation of impairments among the various circuits of the terminal is described.

The repeatered line can introduce noise through digit errors and receiving terminal clock phase variations. The clock phase variations are attributable to dependence of the slave clock phase at repeaters and terminal on the received pulse pattern. This can result in audible noise or, strangely enough, intelligible crosstalk through amplitude, width, or position modulation of the PAM pulse.

Clock phase variation can modulate the signal amplitude only if the PAM signal amplitude is dependent on the width of the channel pulse driving the demultiplex gate. This can be true if the receiving filter is not completely charged in the time the demultiplex gate is closed. As mentioned previously, the demultiplex gate was designed to provide essentially complete charging of the receiving filter. Additionally, the demultiplex gate is a balanced configuration with no output in the absence of signal. This means that any interference due to amplitude, width, or position modulation of the received PAM will be proportional to the received signal and therefore unnoticeable.

Digit errors on the repeatered line may result in audible transients or clicks in the receiver. The magnitude of the transient depends on the weight of the digit in error. For reasonable error rates, only errors in the two most significant digits produce objectionable clicks. A given talker uses 16,000 of these digits per second or about 10^6 per minute. Thus, an error rate of one in 10^6 would result in one click per minute. An error rate of one in 10^7 would result in one click in 10 minutes. A permissible error rate, assuming strictly random error distribution, is between 10^{-6} and 10^{-7}, and the experimental system performed this well or generally appreciably better.

V. DEVICES

A PCM system makes prolific use of devices. The economy in such a design is predicated, therefore, on the availability of inexpensive devices

capable of realizing the speed, power, and other requirements. Previous unavailability of such devices has been the chief deterrent to the adoption of the PCM approach. To realize the advantages of the present device cost picture, one must take cognizance of the fact that devices are inexpensive only when produced in quantity. The experimental system design restricted itself to general-purpose transistors and diodes except where departure from this avenue could be economically justified.

Most transistor applications were filled with a diffused-base silicon NPN unit. There were two types available — one switch design and one amplifier design. A diffused-base germanium PNP was used in the compandor amplifiers and in a few logic applications where the complementtary aspect was of value. The common amplifier employs a large area diffused silicon design, chosen for its higher power capability.

Relatively low-performance alloyed germanium logic diodes, already in large-scale production, and therefore very low in cost, were used where possible. Small, high-speed diffused silicon units were used in critical positions where switching speed was important. Clamping diodes for the coder and decoder were selected from this design, the selection process being necessary to match forward voltage drops. Larger area diffused silicon diodes were used for the multiplex and demultiplex gates. Once more, a selection process was necessary for balance of forward voltage drops.

The compressor and expandor networks use small area diffused silicon diodes.

VI. EQUIPMENT DESIGN

A model of the two-way regenerative repeater is shown in Fig. 12. The design stressed efficiency of space usage to permit the maximum possible number to be mounted in a manhole. Watertight containers were designed to house the repeaters and fault-location circuits for the field experiment.

The experimental terminal, fully equipped, would contain about 300 transistors and 900 diodes. These and their accompanying components were mounted on printed wiring boards of a type shown in Fig. 13. The boards used two-sided wiring to achieve a high component density. Transistors were mounted in sockets for the experimental terminal, although in a production model they would be soldered to the board. Test points were brought out to the front panel made of formed aluminum.

The printed boards can be assembled to form a terminal, shown by the artist's conception in Fig. 14. Circuit functions are assigned to printed

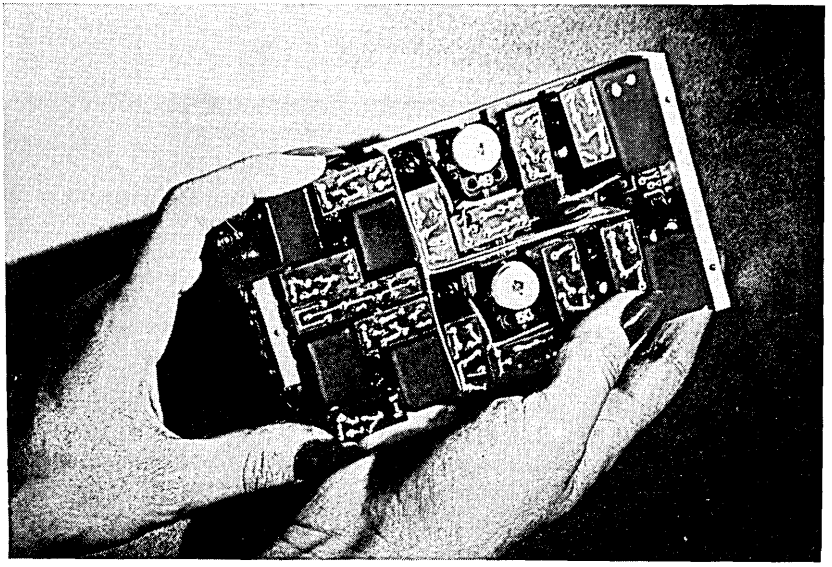


Fig. 12 — Experimental two-way repeater with cover removed.

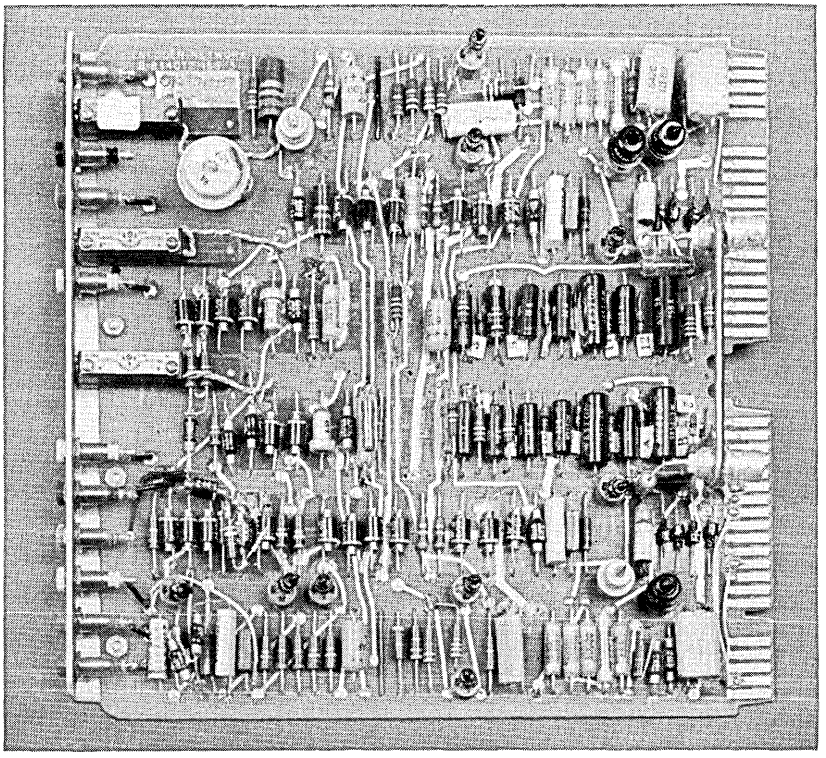


Fig. 13 — Plug-in type printed board used in experimental PCM system.

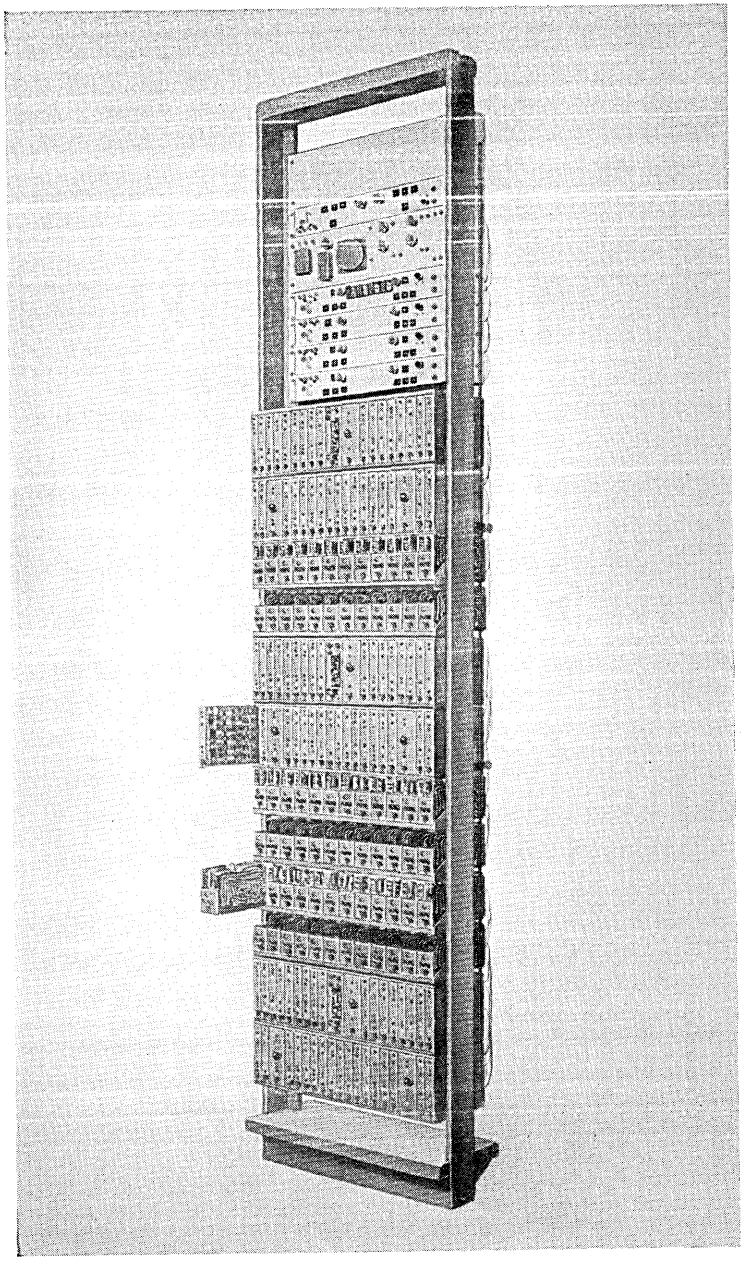


Fig. 14 — Artist's conception of a three-terminal PCM bay.

boards so that a terminal may be partially equipped in the most economical manner. One board per channel, containing hybrid and signaling circuits, is required in addition to the common circuitry. The type of channel unit selected depends on the mode of signaling used between the two offices.

One standard 23-inch by 11-foot, 6-inch bay will hold three PCM terminals. A power supply capable of supplying the three terminals would be mounted in an adjacent bay.

VII. FIELD EXPERIMENT

An early model of the experimental system was installed between Summit and South Orange, New Jersey. This field experiment permitted evaluation of the performance of the repeatered line, signaling, and some unpredicted phenomena in the speech portion of the terminal. While the offices were only about seven miles apart, the provision of several cable pairs permitted the testing of longer repeatered lines.

The Summit office having crossbar switching and the South Orange office having panel switching provided for test signaling from crossbar to panel and panel to crossbar. Loopback tests were used to test crossbar-to-crossbar and panel-to-panel signaling. The general results of the field experiment have been quoted in Section IV of this paper.

VIII. STATUS OF DEVELOPMENT

The experimental system has met design objectives both technically and economically. As described in this and companion papers, the principles of performance and design of such a PCM system are now quite well understood. Final design for manufacture is now nearly completed.

IX. ACKNOWLEDGMENTS

This article has summarized the work of several people. The development of this experimental system was carried out under the direction of E. E. Sumner at the Murray Hill Laboratory and O. L. Williams at the Merrimack Valley Laboratory. Engineering studies were supervised by K. E. Fultz. The author is indebted to members of these groups for their helpful consultation.

REFERENCES

1. Desoer, C. A., B.S.T.J., **36**, 1957, pp. 1403–1428.
2. Shennum, R. H. and Gray, J. R., this issue, p. 143.
3. Mann, H., Straube, H. M. and Villars, C. P., this issue, p. 173.
4. Mayo, J. S., this issue, p. 25.

A Bipolar Repeater for Pulse Code Modulation Signals

By J. S. MAYO

(Manuscript received July 12, 1961)

Designed for use on unloaded exchange cable, this repeater was developed to satisfy the transmission requirements of an experimental PCM system. It utilizes a pulse repetition frequency of 1.544 mc, and 6000-ft repeater spacing. Functionally, the repeatered line transmits the PCM signal without appreciable degradation over distances up to 25 miles, a feat accomplished by retiming and reshaping the signal at each repeater point. Retiming is accomplished by means of a clock extracted from the signal; reshaping is accomplished by regeneration with positive pulse width control.

Near-end crosstalk and pulse train jitter dominate the design parameters. Timing is made tolerant of near-end crosstalk by choice of bipolar transmission (where successive marks are of opposite polarity) with clock derived from the rectified and clipped signal. Tolerance in the decision circuit is obtained by automatic threshold control, spike sampling, and tight control of time and voltage parameters. Accumulated pulse train jitter is controlled to the extent dictated by the terminal equipment, principally through control of the bandwidth of the clock circuit.

Seven diffused-base transistors and ten logic diodes are used in the one-way repeater circuit. A two-way repeater consists of two such circuits with a common power unit, utilizes 135 components, and is packaged in a can of $1\frac{1}{16}$ x $3\frac{1}{8}$ x $5\frac{3}{4}$-inches outside dimensions. Accommodations are made for line-length and power options, as well as remote testing. Power for the repeater is transmitted over the signal pair. One watt is required for the two-way repeater.

I. INTRODUCTION

Companion papers[1,3,6] propose transmission of 24 voice channels by means of pulse code modulation over 22-gauge paper insulated cable pairs. Seven-digit coding, built-in signaling, and frame synchronization dictate a 1.544-mc pulse rate. Very simply, the repeater function is to

"look" at a received pulse train and emit a "new" pulse for each received pulse, a feat that becomes impossible when the transmission bandwidth is sufficiently limited, or the noise environment is sufficiently strong that realizable circuitry cannot distinguish pulse from interference. In practice these parameters are controlled by repeater spacing, and installation of repeaters is greatly convenienced if the nominal spacing is 6000 feet (a repeater then replaces a load coil). Such spacing produces a very difficult interference situation.

One of the many virtues of PCM is that the pulse train carries its own timing information, and this paper considers only the self-timed repeater. In such repeaters it is essential that the recovered clock be and remain phase locked to the incoming pulse train, even in the presence of severe interference.

Interference is mainly intra- or inter-system crosstalk via the near-end crosstalk path (NEXT). Such interference is strongest at high frequencies where timing information must be transmitted. A second important interference source is office impulse noise, which is confined to the vicinity of offices and requires shortened repeater spacing adjacent to offices. Interference of PCM into other carrier systems and vice-versa is considered briefly by Aaron.[1]

Manhole mounting of repeaters is anticipated and repeater size is of utmost importance. Extreme reliability is required, and excessive component miniaturization may not only jeopardize reliability but be costly besides. A reasonable compromise is the use of standard, well-known components with diminutive repeater size achieved by high packing density.

The repeater must withstand environmental conditions attendant to field mounted equipment, including lightning surge activity. A repeater should require no adjustments other than those made at time of manufacture.

Operating error rate per system should be of the order of one error in the most significant digit per channel per minute. This produces one audible click per channel per minute and corresponds to a pulse error rate of the order of one error per 10^6 time slots.

II. FEATURES

The following items summarize the design considerations.

2.1 *Bipolar Transmission*

At the pulse repetition rate, about one per cent of the NEXT paths have less loss than maximum line loss. Transmission of clock at this

rate leads to a very difficult near-end crosstalk situation. A solution is to transmit the clock at less than the bit rate. Aaron[1] presents the problem, considers many transmission schemes, and chooses 50 per cent duty-cycle bipolar transmission where successive marks are of opposite polarity because: (1) by obtaining a 1.5-mc clock from rectification of the bipolar train, most of the clock energy comes from midband (750 kc) where line loss is about 15 db less than at the bit rate, (2) by transmitting clock at half the bit rate, half bit-rate crosstalk loss is applicable, (3) by clipping the pulse train before clock extraction, baseline noise is removed, easing the adverse pulse density problem of sparse transmitted pattern-dense interfering pattern, and (4) by utilizing a transmission scheme with concentration of energy at midband, spectral nulls at zero frequency and the bit rate, physical realization is not so difficult.

2.2 *Threshold and Clipping Control*

Of the many methods of pulse recognition, comparison of the received pulse amplitude to a reference is by far the easiest to implement. With a fixed reference of half the nominal pulse height, a theoretical limit of 6-db reduction in received pulse height exists. A smaller reference accommodates a smaller pulse but also accommodates less noise, and indeed may be exceeded by the tail of a large pulse. Because (1) the received signal is subject to 6- to 12-db amplitude variations from location to location in the environment considered, (2) a pulse will interfere into an adjacent time slot, (3) field gain adjustments are undesirable, and (4) good noise performance is a decided goal, it is essential that the reference "threshold" be dependent upon the average received pulse amplitude. Also the clipping level in the clock path must be proportional to received signal amplitude. Otherwise the phase of the recovered clock would be dependent upon the amplitude of the received signal. It conveniently results that the same signal-dependent voltage (equal to approximately half the peak received signal) may be used in both the "threshold" and clock circuits.

2.3 *Spike Sampling*

The recovered clock zero-crossings tell where to "look" on the pulse train. To prevent partial pulsing, to optimize crosstalk performance, and to allow maximum phase instability in the recovered clock circuit, a narrow sampling pulse must be used to examine the received pulse train.

2.4 *Width Control*

Optimum line equalization results in rise-time limited transmission, the received pulse amplitude being proportional to both the height and width of the transmitted pulse. Good performance requires control of transmitted pulse width, particularly to make it independent of pattern or random variations. The use of the recovered clock for this purpose is most attractive, the positive-going clock zero-crossing "starts" the output pulse, the negative-going zero-crossing "stops" it.

2.5 *Powering*

Powering of repeaters over the cable pair is most attractive. For the exchange plant, interoffice distances up to 25 miles must be spanned. Half that distance must be powerable from each office, preferably with existing power supplies. Because of line resistance, maximum power available per two-way repeater is approximately one watt, and this requires both +130 and −130 volts at each end of the 25-mile system.

III. CONFIGURATION

The stage has been set for the configuration of Fig. 1. The received pulse train is acted upon by linear shaped gain to produce an optimum signal-to-noise condition at the output. The preamplifier gain characteristic is set by the characteristic of 6000 feet of cable. Lesser cable lengths are accommodated by selected quantized line-build-out (LBO) networks that make any cable length appear as approximately 6000 feet to the preamplifier. An average preamplifier output level is established by the automatic threshold and clipping circuit. Rectified and clipped pulses enter a tuned circuit of high Q to produce a sine-wave clock. The zero crossings of the sine wave are extracted and delivered to the regenerator as "sample" and "turn-off" pulses. The regenerator puts out a "new" pulse when the preamplifier output exceeds the reference during the sampling instant. For bipolar transmission the regenerator is a "balanced" circuit of identical halves in push-pull. Positive pulses are automatically routed through one half of the regenerator and negative pulses through the other half. The repeater is thus in a sense not a bipolar repeater, but a generalized pseudo-ternary repeater, theoretically capable of handling any 50 per cent duty cycle psuedo-ternary code that has no zero frequency component.

The arguments have led to a forward-acting repeater with complete retiming and pulse width control. Comparison with other repeater arrangements has been made by Aaron.[1]

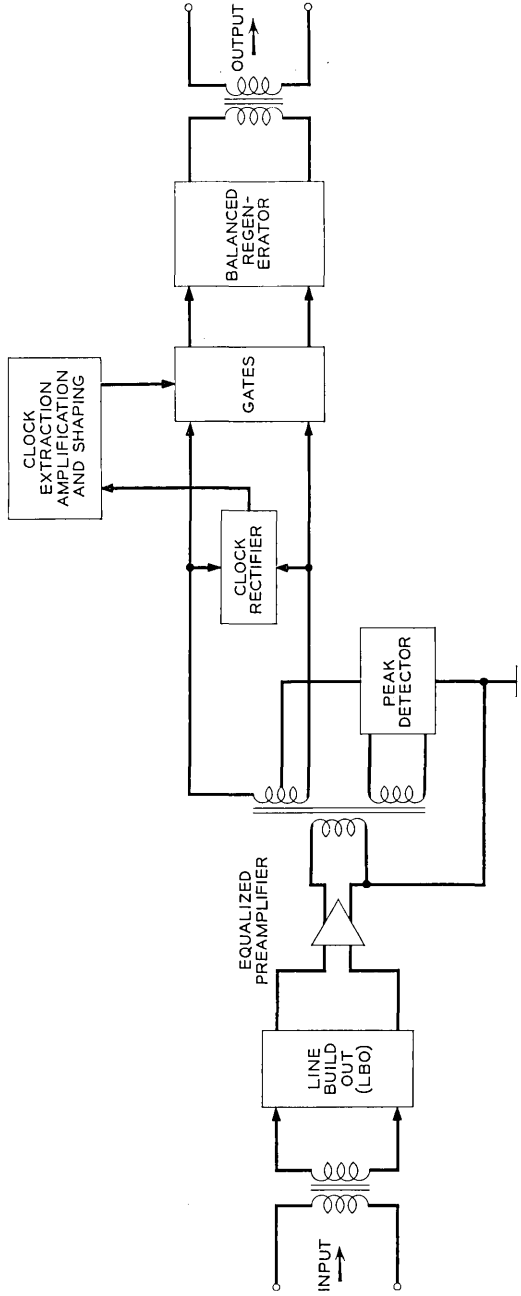


Fig. 1 — Repeater configuration.

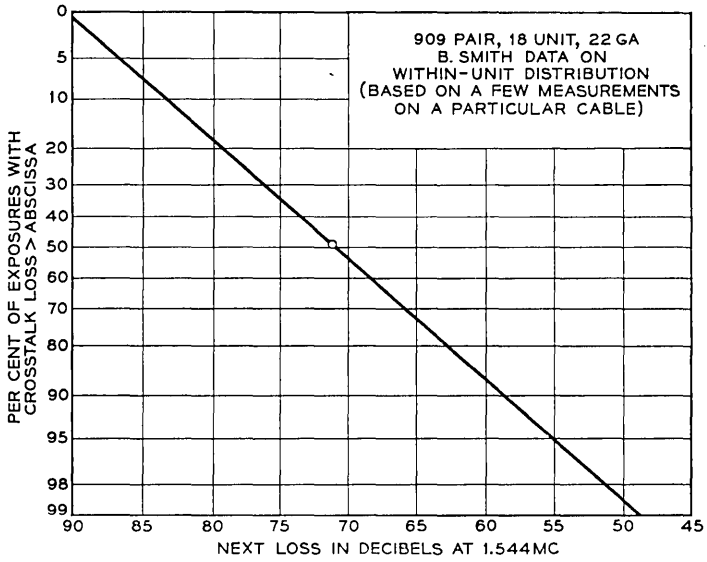


Fig. 2 — Distribution of near-end crosstalk loss.

IV. EQUALIZATION OPTIMIZATION

Define optimum equalization as that cable-LBO-preamplifier gain-frequency characteristic which allows maximum NEXT coupling for the condition of marginal pulse detection. This optimum will be dependent not only upon the cable and crosstalk path characteristics, but upon the "goodness" of the detection process — that is, how accurately the clock spike and threshold positions can be held. Good time crosshair positioning with poor voltage crosshair control dictates a characteristic quite different from the optimum for good voltage but poor time control. Optimization on a crosstalk basis applies primarily to the case of nominal line lengths, where near-end crosstalk is important. Short line lengths are likely to be near offices where optimization on the basis of impulse noise immunity is important. A reasonable approach is to shape the preamplifier for good crosstalk performance for 6000 feet of cable, then design the LBO to optimize impulse noise performance.

Fig. 2 shows the distribution of near-end crosstalk loss for 22CSA cable.* The distribution appears to be log normal with about 1 per cent of the couplings having less than 49 db loss at 1.5 mc. The coupling loss decreases with frequency, in the band of interest, at a rate of ap-

* Aaron[1] considers the cable loss and crosstalk characteristics in some detail. The curve presented is based on a small number of samples on a particular cable.

proximately 4.5 db per octave. No truncation of the distribution is indicated in Fig. 2, though common sense would indicate some lower limit must exist.

Approximate 6000-ft cable response is shown in Fig. 3. Superimposing the 1 per cent crosstalk limit vividly shows that the usable transmission band does not quite extend up to the 1.5-mc bit rate.

A random 50 per cent duty-cycle bipolar pulse train $p(t)$ starting at $t = 0$ with the first pulse positive has a Laplace transform given by

$$P(s) = \frac{1}{s}\left[1 - \exp\frac{-sT_0}{2}\right]\sum_0^\infty a_n e^{-nsT_0} \tag{1}$$

where T_0 = pulse repetition period and

$$a_n = 1, 0, -1 \quad \text{under the bipolar constraint} \quad \sum_0^k a_n = 0, 1.$$

Transmission of this train over a cable and an equalized amplifier of over-all transfer function $G(s)$ gives a signal

$$R(s) = \frac{G(s)}{s}\left[1 - \exp\frac{-sT_0}{2}\right]\sum_0^\infty a_n e^{-nsT_0} = L\{r(t)\}. \tag{2}$$

This signal must be compared to a reference voltage periodically to determine the presence or absence of a pulse.

Let $H(s)$ represent the over-all NEXT path and amplifier characteristic, so the crosstalk signal $N(s)$ at the decision point is

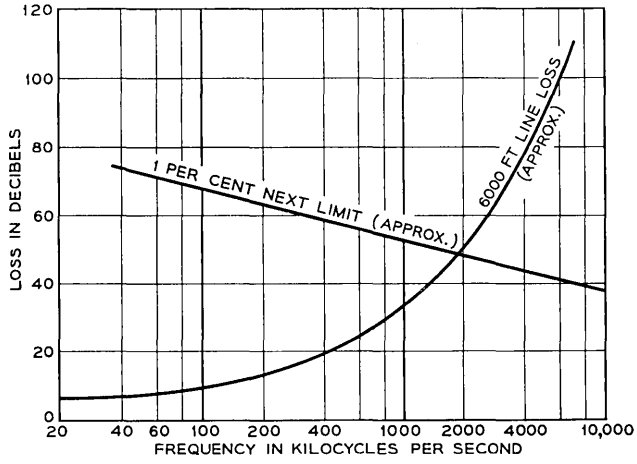


Fig. 3 — Line loss and NEXT loss as a function of frequency.

$$N(s) = \frac{H(s)}{s}\left[1 - \exp\frac{-sT_1}{2}\right]\sum_0^\infty b_n e^{-nsT_1} = L\{n(t)\} \tag{3}$$

where b_n describes the interfering pattern and $1/T_1$ the interfering frequency. With an amplifier characteristic $A(s)$ and a NEXT path described by ks (an approximation), and the cable characteristic by $C(s)$,

$$G(s) = C(s)A(s) \tag{4}$$

$$H(s) = ksA(s) \tag{5}$$

$$H(s) = ks\frac{G(s)}{C(s)} \tag{6}$$

$$N(s) = k\frac{G(s)}{C(s)}\left[1 - \exp\frac{-sT_1}{2}\right]\sum_0^\infty b_n e^{-nsT_1}. \tag{7}$$

A good approximation for the cable characteristic of Fig. 3 must consider a matching of time response. A fair approximation in the range of interest is

$$C(s) = \frac{0.4}{\left(\frac{s}{\omega_1} + 1\right)\left(\frac{s}{\omega_2} + 1\right)} \tag{8}$$

where $\omega_1 = 2\pi \times 80$ kc and $\omega_2 = 2\pi \times 800$ kc.

It is desired to solve (2), (7), and (8) for that $G(s)$ which maximizes k for a workable $r(t)/n(t)$ ratio at decision instants. With $r(t)$ and $n(t)$ asynchronous, for error rates of 1 in 10^6, it is necessary to consider only $n(t)_{\max}$ which adds or subtracts from $r(t)$ to form the composite signal upon which decisions must be made. Furthermore, $H(s)$ has accentuated high-frequency response, so each crosstalk pulse is pretty well confined to its own time slot and $n(t)_{\max}$ may be computed from (7) with $b_0 = 1$ and $b_j = 0, j \geqq 1$.

Such an assumption does not hold for the severely band-limited signal path. Generally, however, in a workable situation the received signal is confined to 3 time slots, a received pulse being principally affected only by the one preceding it and the one following it. From (2), $r(t)$ may be computed for three important cases:

$$\text{Case 1:} \quad a_{j-1} = 1, \quad a_k = 0, \quad k \neq j - 1 \tag{9}$$

$$\text{Case 2:} \quad a_{j+1} = 1, \quad a_k = 0, \quad k \neq j + 1 \tag{10}$$

$$\text{Case 3:} \quad a_{j-1} = -1, \quad a_j = 1, \quad a_{j+1} = -1,$$
$$a_k = 0, \quad k \neq j - 1, j, j + 1. \tag{11}$$

Case 1 gives the interference of a pulse in $j - 1$ time slot into the jth time slot. Case 2 gives the interference of a pulse in the $j + 1$ time slot into the jth time slot. The bipolar rule does not allow Case 1 and Case 2 to apply simultaneously. With a pulse in the jth time slot, only Case 3 is of real interest, where interference of adjacent pulses is most severe. For various $G(s)$, both signal and interference may be computed. Typical solutions for the signal are given in Fig. 4.

The area of Fig. 4 above Case 1 and Case 2 enclosed by Case 3 is the working area for detection. In the absence of any interference, the voltage and time crosshairs must intersect within this area. This area is referred to as an (worst case) "eye," and a chart of all eyes as the "eye diagram," which is just an oscilloscopic type display of a random pulse train. A typical eye diagram is shown in Fig. 5. It is seen to be symmetrical due to bipolar transmission.

Fig. 6 shows eye parameters of interest: the height, h, of the maximum opening which occurs at time T_m and the breadth, w, which occurs at amplitude E_t (generally taken so as to divide h into two equal parts). Let one half the difference between the height of an isolated pulse, E_m, at time T_m, and h define I, intersymbol interference,

$$I = \tfrac{1}{2}[E_m - h]. \tag{12}$$

Peak crosstalk interference hinders pulse recognition when, as seen in Fig. 4, it adds to Case 1 and Case 2 and when it subtracts from Case 3.

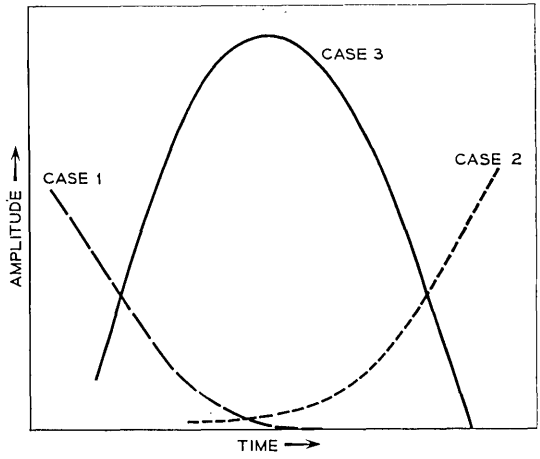


Fig. 4 — Typical received signal.

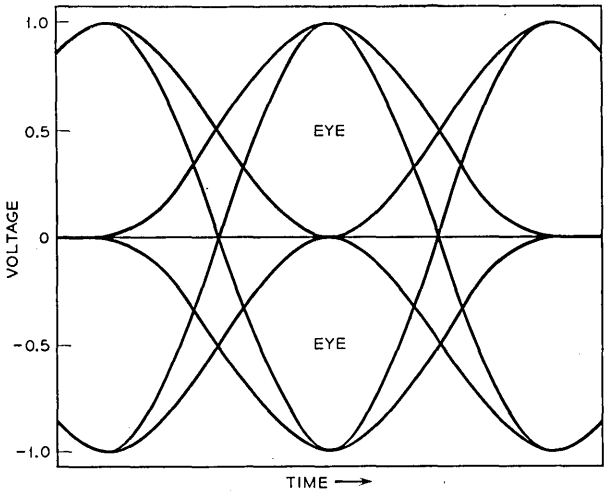


Fig. 5 — Eye diagram.

The maximum allowable interference is $h/2$, at which time the eye is closed.

Preliminaries aside, under the stated assumptions, for any $G(s)$ one may compute h and $n(t)_{\max}$. The value of k which makes $n(t)_{\max} = h/2$ defines the minimum crosstalk loss for which a repeater may work with a *single* interferer, and no other source of interference.

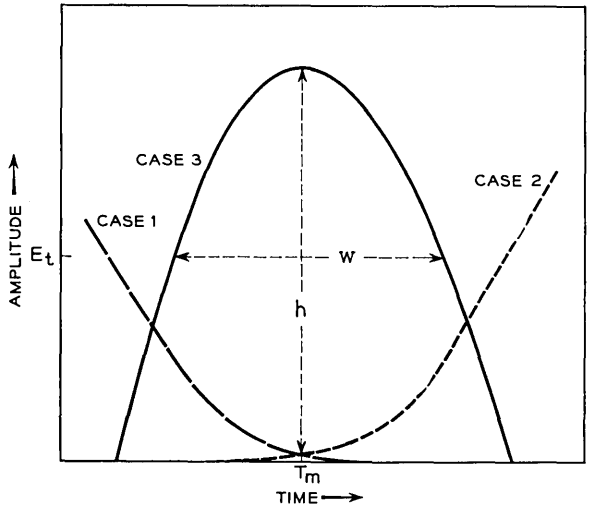


Fig. 6 — Eye parameters.

4.1 *Gross Features*

A convenient choice for $G(s)$ is*

$$G(s) = \frac{1}{\left(\frac{s}{2\pi f_i} + 1\right)^{\alpha}} \cdot \frac{s^2}{(s + \omega_3)(s + \omega_4)} \tag{13}$$

where f_i largely determines the bandwith and α the rate of rolloff. ω_3 and ω_4 describe the low-frequency cutoffs that may be expected in the input and output transformers.

TABLE I — VALUE of $h(f_i, \alpha)$

f_i \ α	3	4	5
1.5 mc	0.90	0.90	0.95
1.25 mc	0.92	0.94	1.0
1.0 mc	0.98	0.93	0.78
0.75 mc	0.80	0.63	0.53

Solving (2) for Cases 1, 2 and 3, and computing h, one gets Table I, for transformer cutoffs at 20 and 40 kc, and henceforth taking $E_m = 1$. Equation (7) yields $n(t)_{\max}$, giving Table II.

TABLE II — VALUES OF $n(t)_{\max} \times 10^{-9} \times k^{-1}$

f_i \ α	3	4	5
1.5 mc	0.55	0.30	0.20
1.25 mc	0.43	0.22	0.14
1.0 mc	0.31	0.15	0.09
0.75 mc	0.19	0.07	0.05

The margin of operation in pulse detection, M, is

$$M = \frac{h}{2} - n(t)_{\max} \tag{14}$$

and by (12)

$$M = 0.5 - I - n(t)_{\max}. \tag{15}$$

Near-end crosstalk loss at 1.544 mc is given by

$$x = -20 \log k - 140 \text{ db}. \tag{16}$$

* Aaron arrives at essentially the same result using the Gaussian cutoff. Equation (13) is not a bad representation of what one can do short of conditioning the equalization to the cable type, gauge or exact length.

Since

$$n(t)_{\max} = 10^9 qk \tag{17}$$

where q are the entries of Table II,

$$M = 0.5 - I - q[10^{(40-x)/20}]. \tag{18}$$

Equation (18) is plotted in Fig. 7 for various f_i and α.

A typical curve is shown in Fig. 8. As $x \to \infty$, $n(t) \to 0$; so the differ-

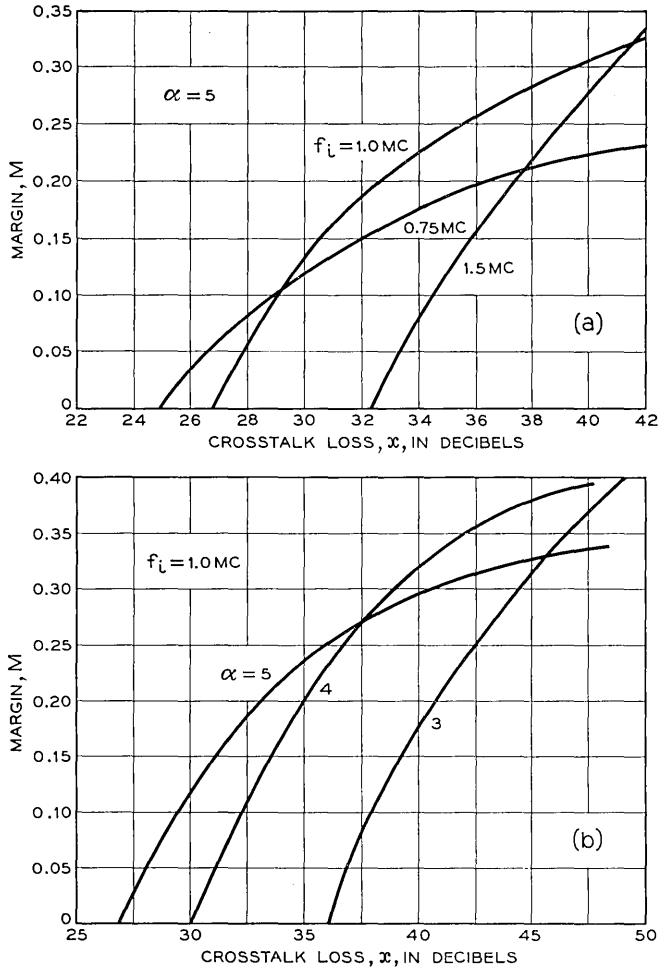


Fig. 7 — Margin as a function of x for (a) various f_i and (b) various α.

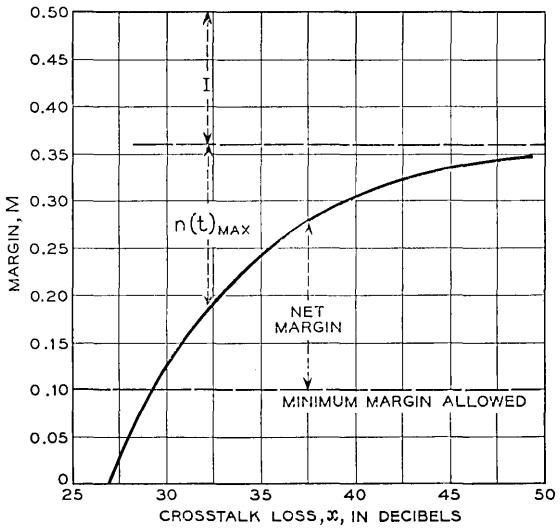


Fig. 8 — Typical margin-crosstalk curve.

ence between 0.5 and the asymptotic value of each curve of Fig. 7 gives intersymbol interference. For any x, the difference between the asymptotic value of each curve and the value of the curve gives the peak crosstalk signal. A narrow transmission band with sharp cutoff is indicated as optimum for the closed eye condition ($M = 0$). On the other hand, perfect threshold setting is unrealizable; the decision level is a region rather than a line; and some minimum M is required for a practical repeater. A reasonable figure is

$$M_{\min} = 0.1 \tag{19}$$

so the region below $M = 0.1$ on Fig. 7 is not usable. Very near optimum performance is obtained with a 1-mc cutoff and a 30 db per octave slope, giving,

$$G(jf) = \frac{1}{(1 + jf)^5}, \quad \text{with } f \text{ in mc and } f > 0.1. \tag{20}$$

Transmission through $G(jf)$ results in the pulse shape of Fig. 9. It is seen the pulse is not symmetrical, and some improvement should be obtainable via additional equalization.

4.2 *Fine Features*

Appropriately, one might now consider other forms for $G(jf)$ to achieve near optimum phase equalization. One may indeed proceed to choose a

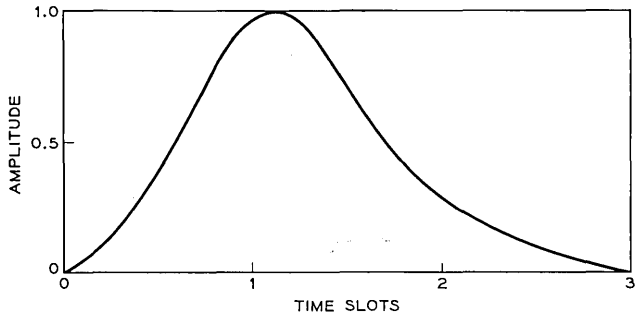


Fig. 9 — Pulse shape resulting from transmission through G(*jf*).

function of known performance, such as one of the Thomson functions,[10] and design an equalized amplifier-cable characteristic that closely approximates this function. Such was the course taken; but two difficulties arose. First, to achieve reasonably symmetrical pulses, higher-order Thomson functions are required. These lead to a very complicated equalizer (in terms of number of components). Secondly, statistical and length variation in the cable do not allow precise equalization short of a field equalizer adjustment, made for each repeater at time of installation.

Equalizer development, herein, follows a course of realization of a fairly simple characteristic of near optimum bandwidth and controlled cutoff, leading to a reasonable, but asymmetrical, time response. A nonlinear technique is then applied to reduce intersymbol interference in a manner not so critical of the exact cable characteristic.

4.2.1 *Preamplifier Design*

Several factors point to a feedback preamplifier with the cable equalizer characteristic built into the feedback path. Feedback is required to achieve gain stability. By placing the equalizer in the feedback loop, a large amount of feedback exists at lower frequencies, greatly reducing the output impedance — a very important consideration because the load driven by the preamplifier is decidedly nonlinear.

From a power gain point of view, a two-transistor amplifier is sufficient. The ac configuration chosen is shown in Fig. 10. The ac feedback network is shown in Fig. 11. Functionally, R_3 and C_2 allow the closed loop gain to rise at 6 db per octave, starting at 80 kc, to match the cable response falloff. At higher frequencies, the rising gain is checked and the rolloff established by R_1, L_1 and Z_f.

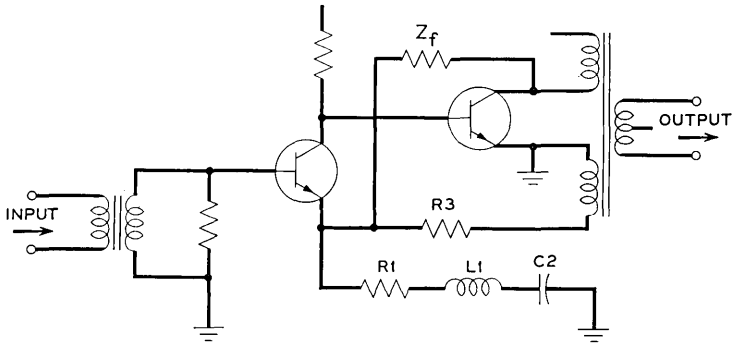


Fig. 10 — AC configuration of preamplifier.

From Fig. 11,

$$\frac{e_1}{e_0} = \frac{n}{1 + \dfrac{nR_3}{Z_f}}\left[1 + \frac{R_3}{Z_e} + \frac{R_3}{Z_f}\right] \tag{21}$$

which for large amounts of feedback is the closed-loop gain of the amplifier, K_v,

$$K_v = \frac{n}{Z_f + nR_3}\,\frac{Z_eZ_f + R_3Z_f + R_3Z_e}{Z_e}. \tag{22}$$

Or

$$K_v = 1 + \frac{\dfrac{Z_f(nR_3)}{Z_f + nR_3}}{\dfrac{Z_e\left(\dfrac{nR_3}{n-1}\right)}{Z_e + \dfrac{nR_3}{n-1}}} \tag{23}$$

and if $Z_F(s)$ is the parallel combination of Z_f and nR_3, and $Z_E(s)$ is the

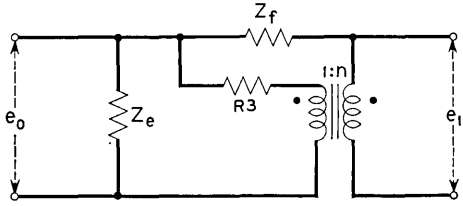


Fig. 11 — Feedback network configuration.

parallel combination of Z_e and $nR_3/(n-1)$,

$$K_v = 1 + \frac{Z_F(s)}{Z_E(s)}. \tag{24}$$

Neglecting the effects of transistor and transformer cutoffs, the voltage gain may be shaped according to (24) by proper selection of R_3, Z_e and Z_f. Z_e has been chosen as a simple series RLC circuit, and Z_f has been chosen as the parallel combination of two RLC circuits, both as shown in Fig. 12.

The parameter n is chosen to give the desired low-frequency gain. Considering the low-frequency cable loss, the voltage gain of the input transformer, and the loss of amplitude due to the bandlimited transmission, one readily establishes $n = 3$ if peak preamplifier output is to equal the transmitted pulse height (3 volts). The combination R_3, C_2 is chosen to give a zero of K_v (24) to cancel the effects of the 80-kc pole in the cable characteristic $C(s)$ (8). The rising gain is checked at 800 kc by C_3, and the rolloff is started by the pole of $C(s)$. The preamplifier gain starts to fall at 1.5 mc due to L_1 and has a decided null (for repeater timing considerations) at 2.25 mc due to L_3 and C_4. The response above 2.25 mc is suppressed by the L_2R_2 combination.

For the parameter values of Fig. 12, one may evaluate $Z_F(s)$ and $Z_E(s)$.

$$Z_E(s) = 0.705\,\frac{s^2 + 0.9091s + 0.9183}{s^2 + 22.27s + 0.9183} \tag{25}$$

$$Z_F(s) = 0.5198\,\frac{(s^2 + 0.1s + 1.961)(s^2 + 1.2195s + 4.545)}{s^4 + 3.429s^3 + 4.5396s^2 + 8.0376s + 3.2627} \tag{26}$$

where Z is in kilohms and s in units of 10^7 radians per second.

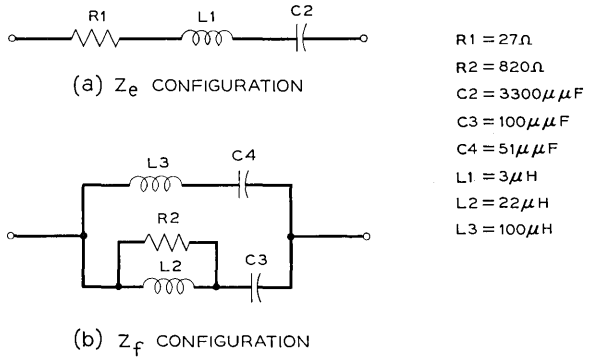


Fig. 12 — Network configurations; (a) Z_e configuration; (b) Z_f configuration.

By (24), the poles and zeros of K_v in megacycles per second are

Poles	*Zeros*
-0.800	-0.0928
$-0.724 \pm j1.342$	$-0.0965 \pm j2.22$
$-0.193 \pm j2.47$	$-0.876 \pm j3.12$
-4.28	$-17.88.$

The corresponding asymptotic response is shown in Fig. 13. The complete circuit is shown in Fig. 14.

The low-frequency response of the circuit of Fig. 14 is controlled by the input transformer T_1, the interstage transformer T_2, and by-pass capacitors C_4 and C_5. Very important also is the low-frequency performance of the regenerator output transformer. The components chosen produce a net effect of a 2-kc break due to T_1 cutoff, another 2-kc break due to T_2 and C_4 and C_5 (an approximation) and a 16-kc cutoff due to the regenerator output transformer. The low-frequency transfer function is then

$$L(s) = \frac{s^3}{(s + 0.01)(s + 0.0012)(s + 0.0012)} \tag{27}$$

with s again in units of 10^7 radians per second. The preamplifier transfer function is then (including effect of regenerator output transformer)

$$A(s) = K_v(s)L(s) \tag{28}$$

and is plotted (excluding effect of regenerator output transformer) in Fig. 15. It was planned that a filter ahead of the preamplifier would remove the higher lobe of response. However, the presence of the lobe is not sufficiently injurious to merit the cost of the filter.

Measured upper-frequency open loop gain and phase is shown in Fig.

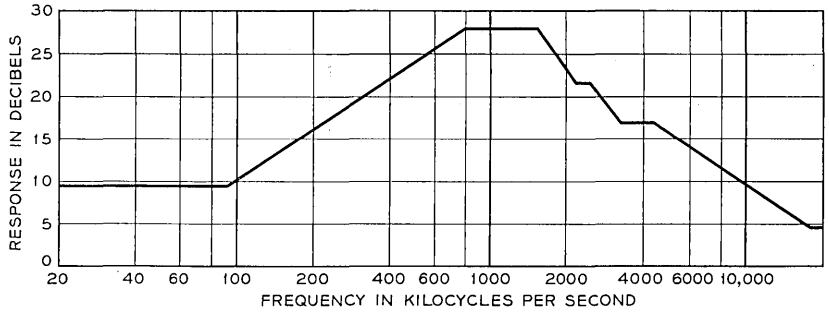


Fig. 13 — Asymptotic response of K_v.

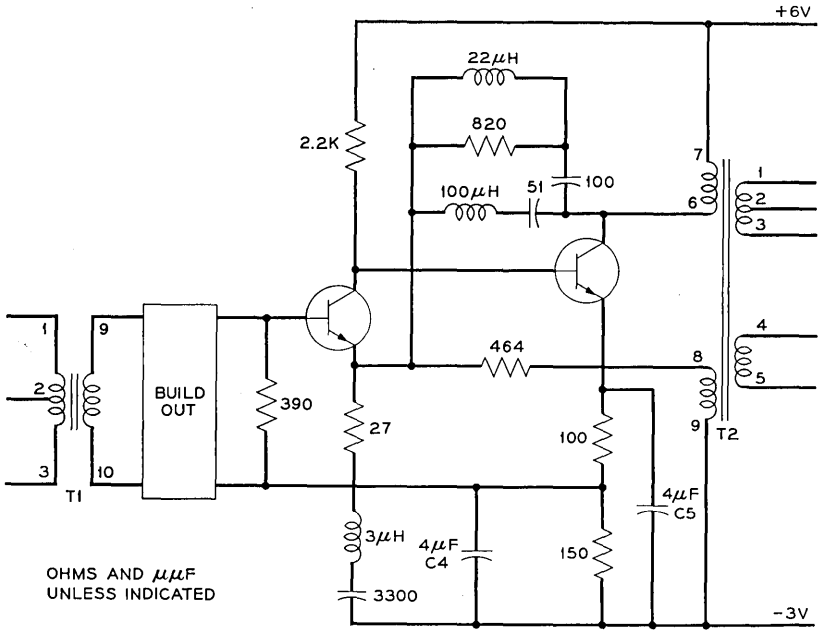


Fig. 14 — Repeater preamplifier circuit.

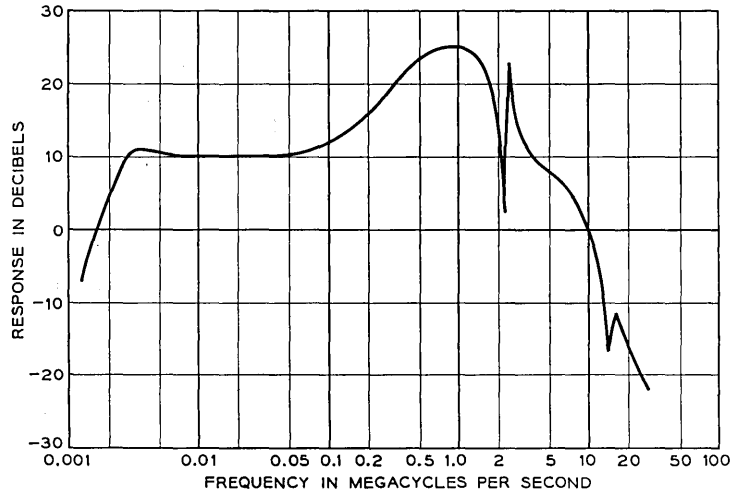


Fig. 15 — Measured preamplifier response.

16. The phase margin (36 mc) is 70°, and there is no phase crossover of 180° in the range of the measurement. For the by-pass capacitors used, the system is stable at low frequencies also, for all transistor gains from zero to infinity.

The equalized transmission medium from regenerator to decision point is

$$G(s) = C(s)K_v(s)L(s). \tag{29}$$

The measured response from line input (not including the effect of regenerator output transformer cutoff) to preamplifier output at pins 1 and 2 of T_2 is shown in Fig. 17.

4.2.2 *Time Response*

Response of $G(s)$ to a 50 per cent duty cycle, 1.5-mc rectangular pulse is shown in Fig. 18. The measured pulse is seen to be quite asymmetrical with a tail that is 20 per cent of the peak pulse height, one time slot after the peak. The difference between computed and measured response is attributed primarily to the poor analytical representation of the cable, (8).

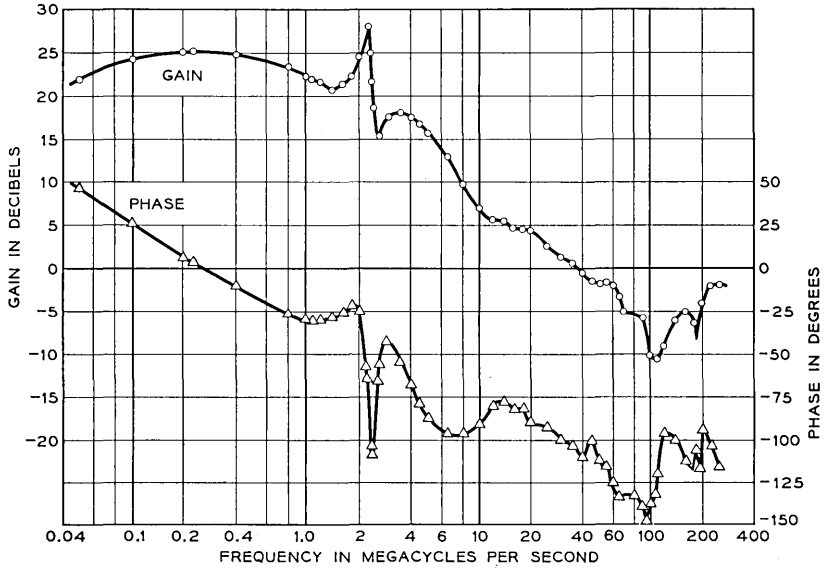


Fig. 16 — Measured open-loop transmission of repeater preamplifier.

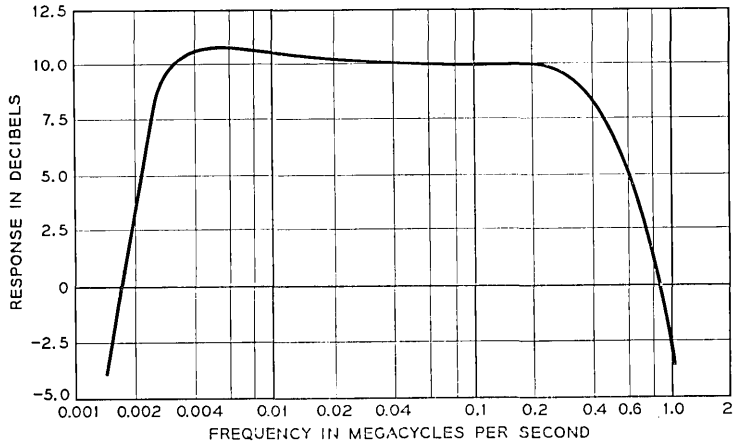


Fig. 17 — Approximate over-all response (not including regenerator output transformer).

Examining Fig. 18, one naturally asks: What can be done to further improve the signal-to-crosstalk situation? There appears to be as much as several db (additional allowed crosstalk coupling) available by improved equalization. Most of this may be obtained without field equalization adjustment by a very simple nonlinear equalizer.

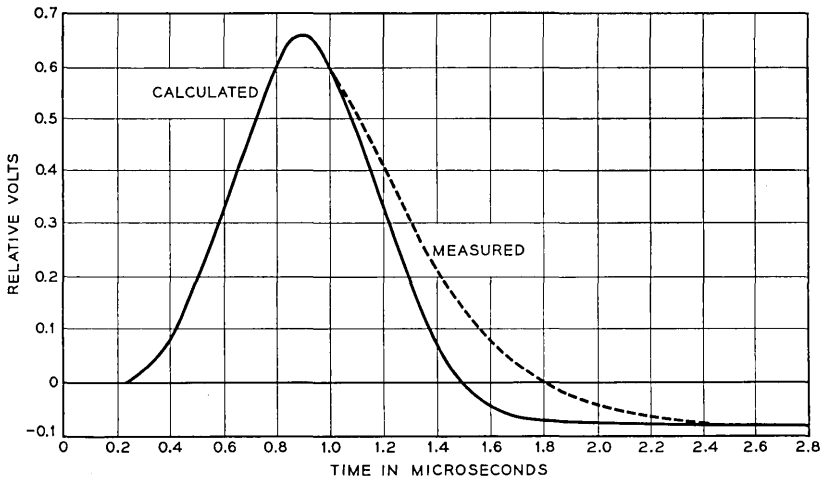


Fig. 18 — Response of line plus preamplifier to a 0.324-microsecond pulse.

4.2.3 *Nonlinear Equalization*

Special transmitted pulse shapes may be used to reduce intersymbol interference. The shape may be adjusted to optimize a decreasing intersymbol interference-increasing peak crosstalk interference situation. An example of such transmission is bipolar-dipulse,* which may be sent directly over the unequalized cable (or an ideal integrator) and which produces received pulse shapes that are almost completely resolved to one time slot. One can interpret the action of the two halves of the dipulse as a mark-erase operation that may not be so critical of the exact nature of the transmission medium. A similar but different preshaping will be considered and will be referred to as nonlinear equalization because of the way it is derived, and, as applied, has other nonlinear effects.

Assume a blocking oscillator is operating into an inductive load as shown in Fig. 19. The effect of the inductive load, in conjunction with the

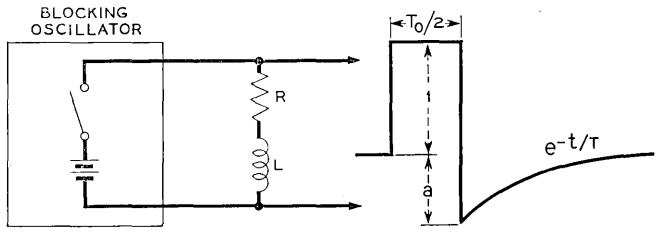


Fig. 19 — Simplified blocking oscillator operating into inductive load.

nonlinear blocking oscillator output impedance is equivalent to a "black box" transfer characteristic

$$Y_2(s) = 1 - \frac{ase^{-(sT_0/2)}}{\left(s + \frac{1}{T}\right)(1 - e^{-(sT_0/2)})} \tag{30}$$

where a is the amplitude of the afterkick and T is the decay time constant. It is convenient if the afterkick is assumed to be of rectangular form of duration $T_0/2$ with an area equal to the actual afterkick area, a fair approximation in a severely band-limited medium. The transfer function becomes

$$Y_3(s) = 1 - \bar{a}e^{-(sT_0/2)} \tag{31}$$

* Achieved by passing bipolar signals through $\left(1 - \exp \frac{-sT_0}{2}\right)$.

where

$$\bar{a} = \frac{2a}{T_0}\int_0^\infty e^{-(t/T)}\,dt = \frac{2aT}{T_0}. \tag{32}$$

Let $g(t)$ represent the actual received pulse of Fig. 18. The effect of (31) may be added in the time domain. If T_m is the time of the pulse peak, and if the afterkick does not materially shift the position of this peak, the received pulse amplitude becomes

$$A_{\max} = g(T_m) - \bar{a}g\left(T_m - \frac{T_0}{2}\right). \tag{33}$$

And the height of the tail of the new pulse, one time slot after the peak, is

$$I_T = g(T_m + T_0) - \bar{a}g\left(T_m + \frac{T_0}{2}\right). \tag{34}$$

The crosstalk path is highly differentiative, so the peak crosstalk interference is approximately proportional to the height of the largest voltage transition of the transmitted pulse. Therefore, letting $N(t)_{\max}$ represent the new peak interference,

$$N(t)_{\max} = n(t)_{\max}\ (1 + \bar{a}). \tag{35}$$

Reading from Fig. 18,

$$g(T_m) = 0.65 \tag{36}$$

$$g\left(T_m - \frac{t_0}{2}\right) = 0.30 \tag{37}$$

$$g\left(T_m + \frac{t_0}{2}\right) = 0.40 \tag{38}$$

$$g(T_m + t_0) = 0.10. \tag{39}$$

Thus

$$A_{\max} = 0.65 - 0.30\ \bar{a} \tag{40}$$

$$I_T = 0.10 - 0.40\ \bar{a} \tag{41}$$

$$N(t)_{\max} = n(t)_{\max}\ (1 + \bar{a}). \tag{42}$$

One observes that I_T decreases faster with $\bar{a}$ than $N(t)_{\max}$ increases with $\bar{a}$. The optimum $\bar{a}$ is then that which reduces I_T to zero

$$\bar{a}_{\text{opt}} = \frac{0.1}{0.4} = 0.25. \tag{43}$$

The half height of the eye, h, is approximately $\frac{A_{\max}}{2} - I_T$,

$$h = \frac{0.65}{2} - 0.1 = 0.225, \qquad \bar{a} = 0 \tag{44}$$

$$h = \frac{0.58}{2} - 0 = 0.290, \qquad \bar{a} = 0.25 \tag{45}$$

If 25 per cent of the received pulse height is allotted to other degradations, the allowed peak crosstalk,

$$n(t)_{\max} = 0.75\left(\frac{A_{\max}}{2}\right) - I_T \tag{46}$$

$$n(t)_{\max} = 0.75\left(\frac{0.65}{2}\right) - 0.1 = 0.142, \qquad \text{for } \bar{a} = 0 \tag{47}$$

$$n(t)_{\max} = 0.75\left(\frac{0.58}{2}\right) - 0 = 0.218, \qquad \text{for } \bar{a} = 0.25. \tag{48}$$

But the actual $N(t)_{\max}$ for $\bar{a} = 0.25$ is

$$N(t)_{\max} = 0.142(1 + 0.25) = 0.178. \tag{49}$$

The new equalization thus allows additional crosstalk coupling to the extent of a factor of 2.18/1.78 or about 2 db.

The measured received pulse for $a = 0.25$ and $T = 0.3$ μsec (values found experimentally to be optimum, considering all effects of the RL network) is shown in Fig. 20. The corresponding $\bar{a}$ is given by (32)

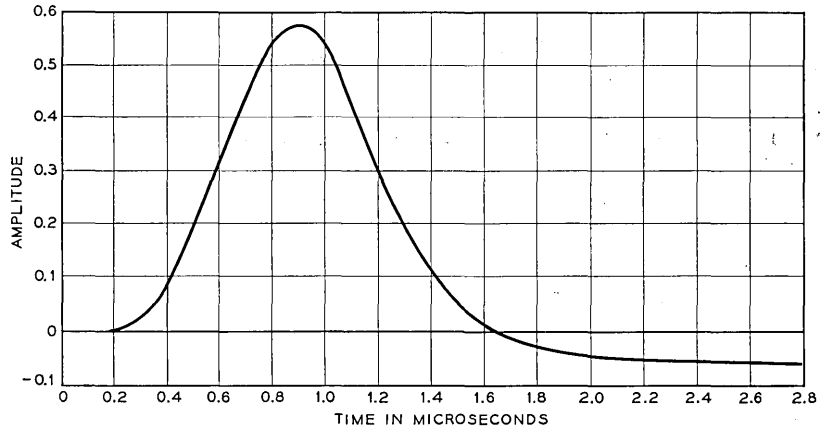


Fig. 20 — Preamplifier output including effect of RL network.

$$\bar{a} = \frac{2(0.25)(0.3)}{0.628} = 0.24 \tag{50}$$

a value almost equal to that expected considering the effect of the network on equalization alone.

4.3 *Line-Build-Out Networks*

The nature of PCM transmission is such that small variations in received pulse shape may not be consequential. A field equalization adjustment dependent upon the exact length of a particular section of cable does not appear justified. An attractive proposal is to build out any line length by means of one of a quantized set of networks to an equivalent 6000 feet plus or minus some quantization error.

The build-out network is a separable option on the repeater and, for many reasons, should be a simple network. A single-break-frequency network is desirable, and is a good approximation when the cable length

TABLE III — BUILD-OUT NETWORK PARAMETERS

Build-Out Network (feet)	Flat Loss (db)	Corner Freq. (kc)
5000	4.0	90
4000	3.0	120
3000	3.0	250
2000	1.5	340
1000	1.2	700

to be simulated is small. For shorter-spaced repeaters, a poor approximation is satisfactory, for little margin must be allotted to interference.

A bridged-T network has the desirable feature that flat-loss and break-frequency may be controlled independently. In the vicinity of central offices, the over-all transmission frequency band may be increased and flat loss adjusted to further open the eye, yet maintain nominal received pulse level. An equalization optimization problem exists for each build-out. The optimization, however, is relative to a statistical quantity (allowed impulse noise), and is not so "sharp" as was the case for "fixed amplitude" NEXT. Very nearly optimum LBO's result if the transmission band is limited to produce only slight intersymbol interference. A set of resulting loss and corner frequency values is given in Table III.

Using the $C(s)$ of (8) and the preamplifier characteristic of (28) produces the amplitude versus cable length curve of Fig. 21.* One thousand

* The actual computation was based on an amplifier characteristic slightly different from (28) and a better approximation to the cable characteristic than (8). The effect of the RL equalizer was not included. It is felt, even so, that the amplitude-line length function obtained is in close agreement with that available from (28), (30) and (8).

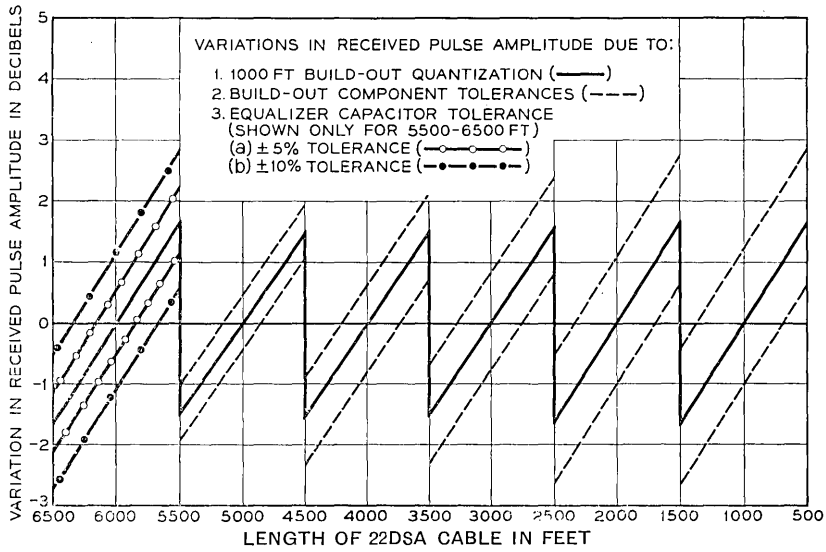


Fig. 21 — Variations in amplitude vs cable length.

foot build-out quantization is seen to contribute about 3 db to the expected variation in received pulse height. Five per cent component tolerances in the LBO network contribute a maximum gain variation of 0.5 to 1 db, depending upon length of cable represented. A 5 per cent variation in C_2 of the amplifier contributes another 0.5 db to variation in received pulse height.

V. VOLTAGE CROSSHAIR

In a severely band-limited system, the received pulse amplitude is dependent upon both width and height of the transmitted pulse. In addition, variations in line loss, line-build-out parameters, and amplifier characteristics affect the received pulse amplitude (and shape). A realizable set of variations is shown in Table IV.

A 3-to-1 variation in received pulse amplitude is thereby expected, aerial operation being accommodated by better LBO control (twice as many networks). To materially reduce this variation in received pulse amplitude requires components of extreme tolerances. Even a field gain adjustment to eliminate items 3, 5, 6, 7 would leave a total variation of 5 db for underground and 8 db for aerial operation. A field gain adjustment can also eliminate items 1 and 2, but this requires adjustments at three locations each time a two-way repeater is replaced. Item 4 alone accounts for 2 db and 5 db for the two types of plants. In a repeater

TABLE IV — FACTORS AFFECTING RECEIVED PULSE HEIGHT

Factor	Manhole 32–100°F	Aerial −40 to +140°F
1. Pulse height	±1.0 db	±1.0 db
2. Pulse width	±0.5 db	±0.5 db
3. Line loss (3σ limit)	±1.0 db	±1.0 db
4. Line loss (temp.)	±1.0 db	±2.5 db
5. LBO Quantization	±1.5 db	±0.7 db
6. LBO Parameters	±1.5 db	±0.8 db
7. Preamp characteristic	±0.5 db	±0.5 db
Total	±7.0 db	±7.0 db
Items 4 + 5 + r.s.s. of others	±4.7 db	±4.9 db

with a fixed threshold at one half the nominal received pulse height, a loss of 1 db in pulse height results in a 2-db penalty in crosstalk performance. A 2.5-db loss of amplitude produces a 5-db penalty.

How should level control be implemented? There are two important approaches, automatic gain control and automatic threshold control. Automatic gain control is most desirable for it maintains fixed preamplifier output, which in turn prevents line and gain variations from affecting recovered clock level. But automatic gain control is rather difficult to realize, requiring a voltage-sensitive device, making it generally expensive and difficult to maintain linearity. It is, however, rather easy to obtain a voltage proportional to received pulse height, and this voltage may be used as the detection voltage crosshair. This simplicity gives automatic threshold control a decided advantage over automatic gain control. Variation in clock level with transmission gain is not too consequential, for it varies violently with pattern anyway.

5.1 *Realization*

Fig. 22 shows a possible implementation of automatic threshold control. A balanced preamplifier output delivers bipolar pulses to two decision gates. Assuming the transistor base-emitter drop V_{BE} equals the diode forward drop, regenerator R_1 works when the output "a" exceeds ground, and R_2 works when output "b" exceeds ground. A negative voltage V_c applied at c allows R_1 to work when the preamplifier output $v_o > V_c$, and R_2 works when $v_o < -V_c$. Thus the threshold voltage at c is applied to both pulse polarities, and the bipolarity of the signal is automatically maintained through the regenerator.

The proper "threshold voltage," $|V_c|$, is achieved by setting the turns ratio of the preamplifier output transformer and selecting V_R such that it cancels the peak-detecting diode drop. Capacitor C_6 is large enough to make V_c a slow function of time, able to follow long-term variations but unable to vary between pulses.

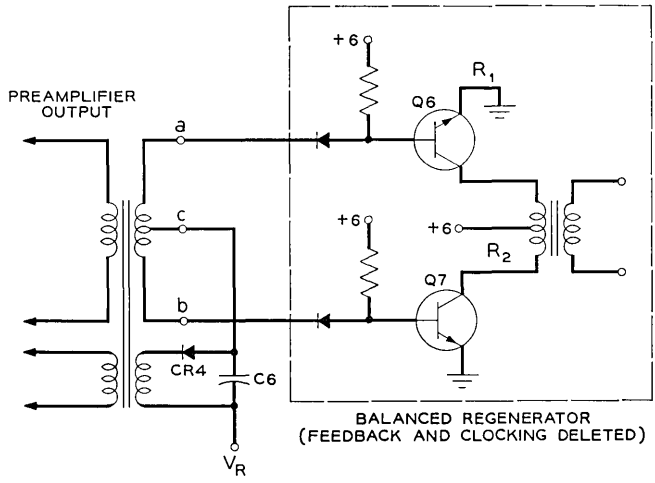


Fig. 22 — Conceptual threshold circuit.

Two changes are necessary to make the circuit workable. These are shown in Fig. 23. First the peak-detecting diode must be able to carry the average gate current. Peak diode current is approximately

$$I_P = \frac{I_{\text{gate}}}{\tau} \tag{51}$$

where τ is an equivalent fractional conduction time for the diode. If the drop across the diode is not to exceed 8 per cent of V_c, Fig. 20 shows the conduction time to be approximately 0.200 μsec. For minimum density pattern of 1 out of 8

$$I_P = \frac{I_{\text{gate}}}{\frac{200}{648} \times \left(\frac{1}{16}\right)} = 52\, I_{\text{gate}}. \tag{52}$$

The preamplifier is not likely to be able to deliver this peak current, so dc gain must separate the peak detector and gate. This is the function of Q_3 in Fig. 23.

Secondly, consideration must be given to the noise performance of the peak rectifier. In Fig. 22 the charging time constant for C_6 is much less than the discharge time constant. A short burst of noise may charge C_6 to an extreme value, producing a long tail of repeater errors after the disturbance. In Fig. 23, C_6 is chosen small and C_9 as large as economical and R_{15} as large as allowed. Selection of the proper time constants for the threshold circuit depends upon the nature of the interference, the over-

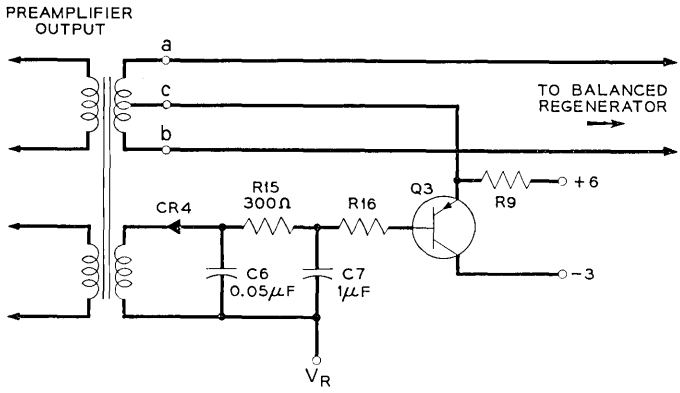


Fig. 23 — Actual threshold circuit.

load performance of the preamplifier, the low-frequency response of the preamplifier, the low-frequency signal handling capability of the preamplifier, and allowed performance degradation. A reasonable compromise design results in the parameters of Fig. 23.

For optimum performance the voltage crosshair must be set at the "center of the eye" [$V_c \approx \frac{1}{2}E_m$], so ideally $\pm h/2$ volts of peak interference are allowed. With a threshold misplaced Δ volts, $(h/2) - \Delta$ volts of peak interference is allowed. The reduction in allowed interference is

$$20 \log \frac{h/2}{(h/2) - \Delta} = 20 \log \frac{1}{1 - \frac{2\Delta}{h}} \text{ db} \tag{53}$$

which is plotted in Fig. 24 and illustrates the importance of precise threshold control.

Performance of an ideal automatic threshold is compared to a fixed threshold in Fig. 25. Ideally the margin against errors of commission (extra pulses) equals the margin against errors of omission (lost pulses) for all input amplitudes, while for a fixed threshold there is a constant margin against errors of commission.

Tracking of the automatic threshold circuit requires that the regenerator be "on the verge" for no input signal, that is, from Fig. 23 with the regenerator of Fig. 22

$$-V_{BEX} + V_f(I_g) + V_{BEY} + \frac{I_c}{\beta_0}(R_{15} + R_{16}) + V_f\left(\frac{I_c}{\beta_0 \tau}\right) + V_R = 0 \tag{54}$$

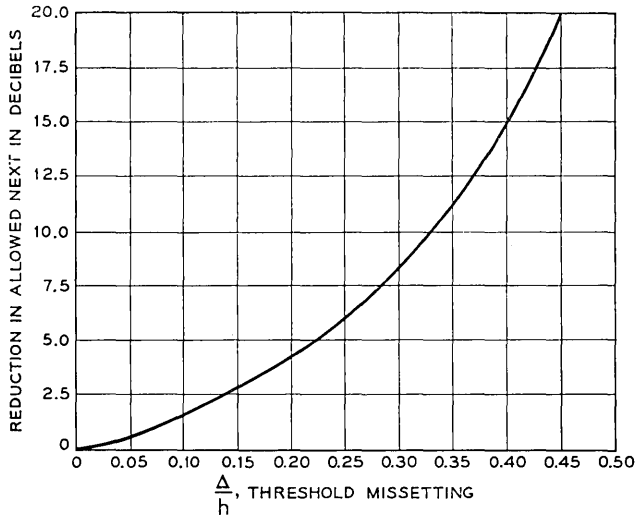


Fig. 24 — Reduction in allowed NEXT as a function of threshold mis-setting in fraction of height of eye.

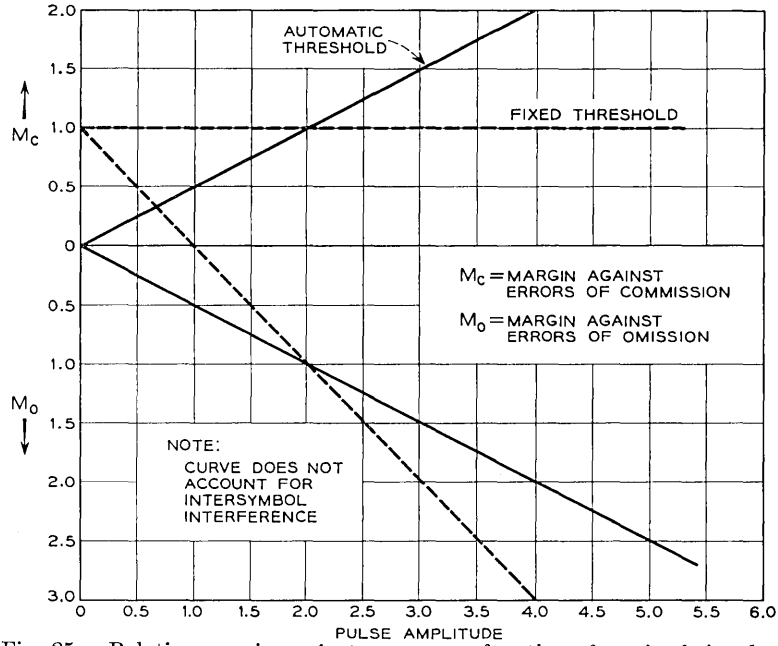


Fig. 25 — Relative margin against errors as a function of received signal amplitude.

where

V_{BEX} = "just conduct" base-emitter voltage of Q_6 or Q_7

$V_f(I_g)$ = diode forward drop

V_{BEY} = base-emitter drop of Q_3

I_c = collector current of Q_3

β_0 = dc common emitter gain of Q_3 .

But V_{BE} , V_f and β_0 are statistical quantities, and V_R can be selected only to satisfy (54) for nominal parameter values. With that V_R , however, reasonable transistor and diode specifications make the left-hand side of (54) as large as ± 375 millivolts, which for a 2-volt eye is a 4-db penalty in allowed peak interference (Fig. 24). A factory adjustment of each repeater is called for, but the expense, size, and unreliability of a potentiometer do not seem justified. A satisfactory solution is a strapping option made at time of factory checkout. V_R is selected for nominal device voltage drops. One strap option raises V_R by 0.25 volt, another strap lowers V_R by 0.25 volt. The resulting threshold placement is correct to within 0.125 volt. Furthermore, since the deviation is due to many device parameters, extreme deviations should seldom be encountered, so most repeaters should come out with option 1, no strap at all.

Temperature dependence of the threshold voltage is primarily through the dependence of β_0 of Q_3 . The diode and transistor drops around the threshold loop generally "track" with temperature. However, any change in Q_3 base current produces a voltage drop in R_{15} and R_{16} . The variation in this drop for $\beta_0 = 20$, $\beta_{\min} = 12$ and $\beta_{\max} = 50$, with $R_1 = 300$ ohms $R_2 = 470$ ohms and $I_c = 10$ ma, is approximately ± 0.25 volt, a worst case. For higher initial β_0's, the situation is not so severe.

5.2 *Effect of NEXT*

The effect of NEXT on the automatic threshold is interesting. There are many cases, but let us consider only the case of a single interferer which is only slightly asynchronous.

The threshold will then be set by the sum of the signal and crosstalk when both signal and interference peak fall in phase. When the crosstalk is out of phase with the signal, the sum signal has peaks no higher than the desired signal alone. There thus appears a beat-frequency component in the threshold circuit equal to the height of the interference. If the bit rates of the two signals differ considerably, the low-pass filter in the threshold loop will reduce the beat-frequency component.

Assuming the threshold interference voltage sinusoidal

$$V_c = r[E_m + \tfrac{1}{2}n(t)_{\max}(1 + \sin \Delta\omega t)] \tag{55}$$

where r is the fractional threshold setting.

Neglecting intersymbol interference, and letting M_c be the margin against errors of commission and M_0 be the margin against errors of omission, then

$$M_c = V_c - n(t)_{\max} \tag{56}$$

$$M_0 = E_m - n(t)_{\max} - V_c \tag{57}$$

when V_c is a minimum, M_c is a minimum, and by (55) and (56)

$$(M_c)_{\min} = rE_m - n(t)_{\max} \tag{58}$$

when V_c is a maximum, M_0 is a minimum, and

$$(M_0)_{\min} = (1 - r)E_m - (1 + r)n(t)_{\max}. \tag{59}$$

The optimum r allows M_c and M_0 to simultaneously go to zero as $n(t)_{\max}$ increases. When

$$(M_c)_{\min} = (M_0)_{\min} = 0 \tag{60}$$

$$n(t)_{\max} = rE_m = \frac{1 - r}{1 + r}E_m \tag{61}$$

$$r = 0.414. \tag{62}$$

For the case considered, the effect of $n(t)$ on the threshold voltage requires, for optimum performance, that the voltage crosshair be set at 41.4 per cent of the peak received signal. Optimum setting for a fixed threshold voltage (for the assumptions made) would be at 50 per cent of the received signal.

The automatic crosshair positioning thereby potentially introduces up to a 1.6-db penalty in crosstalk performance. All of the 1.6 db is not lost in a practical repeater because, for example, $n(t)_{\max}$ must be less than 0.5 E_m, especially due to intersymbol interference, and the special case considered is highly idealized in many other respects.

VI. TIME CROSSHAIR

The equalized and amplified pulse shape shown in Fig. 20 can be crudely matched by

$$e(t) = 1 + \cos \frac{\pi t}{T_0}. \tag{63}$$

Peak crosstalk interference is given by something between the (1.0,5) and (1.25,4) entries of Table II. The measured quantity is almost in the center of this range, 0.18,

$$n(t)_{\max}(10^{-9})(k^{-1}) = 0.18 \tag{64}$$

and by (17)

$$n(t)_{\max} = 0.18(10^{-(x-40)/20}). \tag{65}$$

Fig. 26 is a plot of (65) and shows the peak crosstalk at the decision point as a function of crosstalk loss. Contours of constant x may now be drawn on the eye diagram by adding the corresponding $n(t)_{\max}$ to the Case 1, Case 2 curves, and subtracting $n(t)_{\max}$ from the Case 3 curve. Fig. 27 results. The eye is seen to close at $x = 31.5$ db. At $x = 34$ db, the time crosshair must be held within ± 0.108 μsec minus the crosshair half-width.

6.1 *Effect of Clipping Level*

As previously stated, clock may be recovered from a bipolar pulse train by rectification of the received signal. For various reasons, a primary one being the elimination of adverse pattern conditions in near-end crosstalk, the received pulses are clipped to about half amplitude by the clock rectifier. This eliminates interference on the pulse train baseline and resolves the pulses to their own time slot. However, the clipping operation affects both the amplitude and phase of the recovered clock.

Empirically, the variation of phase of the recovered clock with clipping level is shown in Fig. 28(a), and the variation of clock amplitude with clipping level is shown in Fig. 28(b). As a compromise between phase variation with pattern and loss of amplitude, and because it is

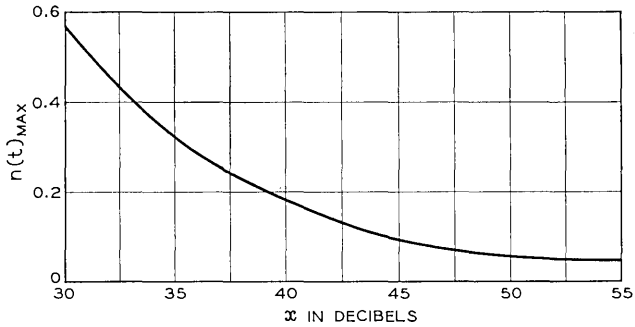


Fig. 26 — Peak crosstalk interference as a function of 1.544-mc crosstalk loss.

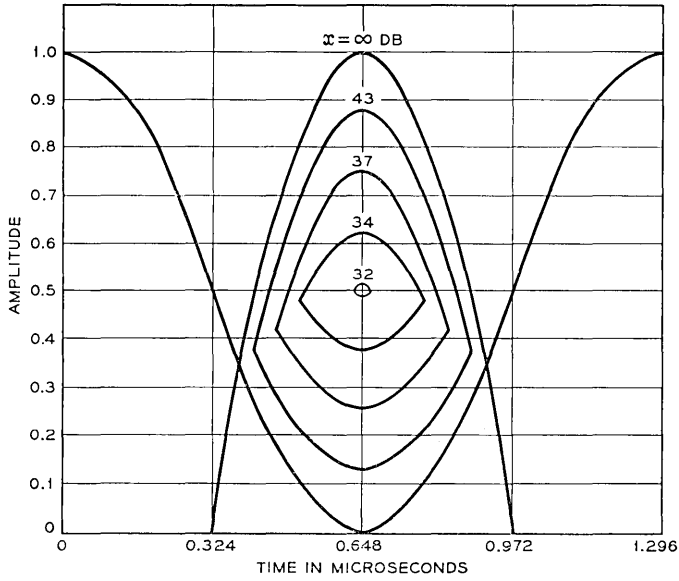


Fig. 27 — Eye size for various values of NEXT loss.

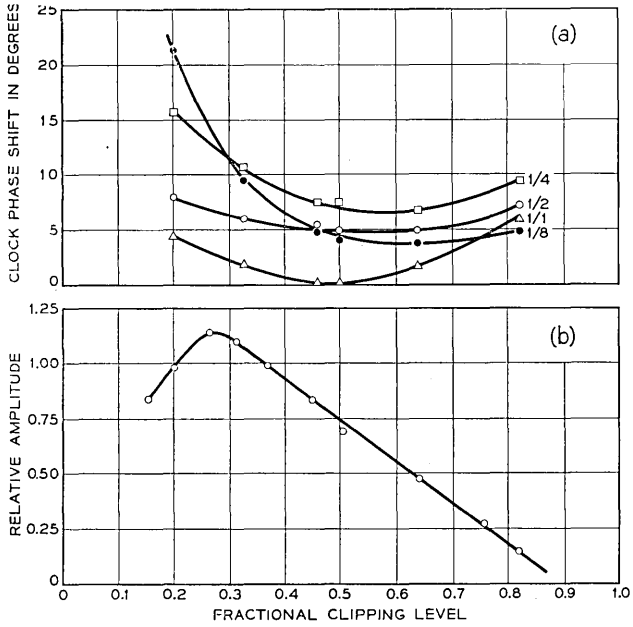


Fig. 28(a) — Measured clock phase as a function of clipping level for various patterns; (b) measured clock amplitude as a function of clipping level for 4/8 pattern.

very easy to implement, the clock clipping level may be set by the voltage crosshair signal to approximately half the peak received pulse. This produces an expected pattern jitter of 6°, and a nominal clock phase that varies 3° as the clipping level is varied ±10 per cent from half-peak amplitude (for an average pattern such as $\frac{4}{8}$*).

6.2 *Effect of Width*

The time crosshair is subject to jitter due to timing interference via near-end crosstalk. Assuming the timing signal is transmitted at 772 kc, (8) gives

$$20 \log | C(i2\pi \cdot 772 \text{ kc}) | = -30 \text{ db} \tag{66}$$

and by (5) and (16)

$$20 \log \frac{H(S)}{A(S)} = -(x + 6) \text{ db}. \tag{67}$$

The signal-to-noise in the timing channel is then, due to crossta k interference,

$$20 \log S/N = -30 + (x + 6) = x - 24 \text{ db}. \tag{68}$$

A word about the assumed 772-kc timing. The assumption holds almost exactly for $a_j = b_j = 1$ (full pattern); it fails considerably for some particular patterns. For random patterns, it is a fair approximation. The S/N ratio quoted is then, at best, a fair approximation. The maximum clock jitter for any $S/N < 1$

$$j_{\max} = \sin^{-1} \frac{N}{S} \tag{69}$$

$$j_{\max} = \sin^{-1}[10^{-(x-24)/20}], \qquad x > 24 \text{ db}. \tag{70}$$

Fig. 29(a) is a plot of (70). The eye closes when $j_{\max} \approx 25°$. The allowable jitter is

$$J(x) = \frac{W(x) - \delta - 2\epsilon}{2} \cdot \frac{360}{0.648} \tag{71}$$

where

$W(x)$ = eye width in presence of crosstalk in μsec
δ = crosshair width in μsec
ϵ = deviation of crosshair from optimum placement, in μsec.

* n/m will be used to indicate repetitive words of m time slots, each with n *consecutive* pulses in the word.

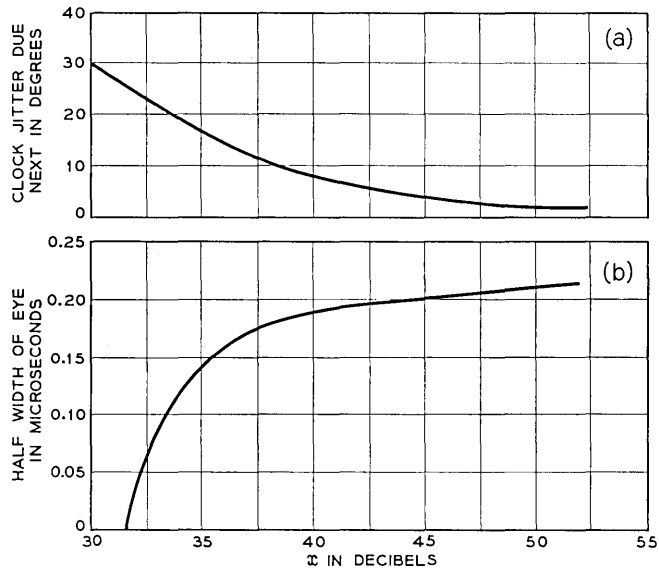


Fig. 29 — Clock jitter (a) and half width of eye (b) as a function of 1.544-mc NEXT loss.

$W(x)$ is available from Fig. 27 and is plotted in Fig. 29(b). Equating $J(x)$ to $j_{\max}$, one obtains the minimum allowable crosstalk loss as a function of sample width.

$$\frac{360}{0.648} \cdot \frac{W(x) - \delta - 2\epsilon}{2} = \sin^{-1}[10^{-(x-24)/20}]. \tag{72}$$

For each x one may read $W(x)$ from Fig. 29 and compute δ from (72). Fig. 30 results, which can be interpreted as the minimum crosstalk loss as a function of crosshair width. In reality, $\epsilon = 0$ cannot be maintained; $\epsilon = 0.054$ μs ($\pm 30°$) is realistic. One concludes that, for $\epsilon = 30°$, a crosshair width of 0.1 μs is reasonable. The penalty in crosstalk performance for this finite width is about a db.

6.3 *Timing Extraction*

Many circuits for clock extraction were considered, and three circuits were reduced to hardware. The preferred circuit (adequate performance at minimum cost) has a single LC tank for clock extraction. Automatic phase-locked oscillator and crystal tank circuits, while workable and possessing certain desirable features, require all the components of the

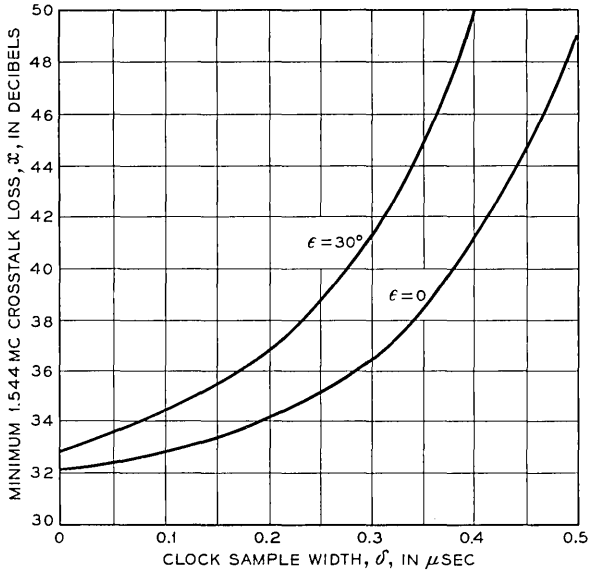


Fig. 30 — Minimum crosstalk loss as a function of crosshair width.

LC tank plus others. Stability requirements on the LC tank are lessened in the latter two approaches. (The phase-locked oscillator developed uses an LC tank in the oscillator; the damped crystal works into an LC tank.) The crystal tank is particularly difficult because of the poor phase stability of a crystal and the difficulties in compensating for crystal shunt capacitance.

The effective Q of the timing circuit must be large enough that (1) the timing circuit can bridge gaps between pulses in the received train and such that (2) the phase of the recovered clock varies with pulse pattern in an acceptable manner. The better resolved the received pulses, the smaller the pattern shift. Generally, bandwidth and gain limitations result in an unresolved pulse driving the timing circuit. Separate equalization for the clock path does not seem justified for, with a clock Q of 50 to 100, a maximum pattern shift of 10° per repeater appears tolerable and realizable.

The effective Q of the timing circuit must be small enough that over the temperature range (−40 to +140°F) and 20-year life, the phase of the clock is stable to ±30° (1 db crosstalk penalty). It will be seen that the time crosshair can be placed only to the nearest ±10°, so that only ±20° are available for temperature and aging. With minimum Q's of

75 and maximum Q's of 100, the required stability in resonant frequency is

$$\pm\tan 20^\circ \geqq 2Q \frac{\Delta f}{f} = 200 \frac{\Delta f}{f} \tag{73}$$

$$\frac{\Delta f}{f} \leqq \pm 0.0018 \tag{74}$$

or the total allowed component variation is 0.36 per cent.

A carbonyl iron inductor is expected to age $\leqq 0.1$ per cent over a 20-year life. Tests on tubular ceramic capacitors show an aging trend of <0.1 per cent over a 20-year life. The allowed temperature mistracking (-40 to ± 140) is then ± 0.16 per cent. Temperature coefficient of carbonyl iron is typically 0.6 per cent ± 0.3 per cent over the temperature range; a realizable capacitor (-0.6 per cent ± 0.3 per cent) produces an over-all coefficient of $+0.6$ per cent to -0.6 per cent over the range -40 to $+140$°F. Initial tuning is done at midrange temperature, so the expected over-all variation is $< \pm 0.3$ per cent, almost double the requirement of ± 0.16 per cent. It is not likely that both inductor and capacitor will age in the wrong direction by the maximum amount and individually have worst-case temperature coefficients that aid the aging effect. If this should happen, the clock phase will be off as much as

$$10^\circ + \tan^{-1}[200(0.25 \times 10^{-2})] = 36.6^\circ. \tag{75}$$

This additional shift increases the minimum crosstalk loss by another db. The situation as presented appears workable but is pushing the art of component stability.

6.4 *Implementation*

The clock circuit must accept pulses of heights from 1 to 3 volts, clip off the bottom half of the pulse, and from this pattern-varying train deliver a stable uniform clock to the regenerator.

For a full-pulse pattern, the clipped signal delivered to the tuned tank is very nearly a rectified sine wave of 0.5 volt minimum amplitude. The 1.544-mc component is $4/3\pi \times 0.5$ or 0.212 volt. The minimum density level (1 out of 8 pulses present) is about $(1/8)(0.212) = 0.0265$ volt. The maximum signal is $3 \times 0.212 = 0.636$ volt. The clock amplitude thus varies over a 28-db range, and the clock circuit must clip this back to a constant-level signal, yet introduce no appreciable amplitude-to-phase conversion.

Two LC clock circuits have been built. Both have a gated output stage.

One has a two-transistor feedback amplifier driving the output stage, with limiting in the feedback path. The other is forward-acting with diode clipping. The latter circuit is shown in Fig. 31.

The automatic threshold voltage serves as a clipping voltage for the clock. An emitter follower provides a low driving impedance for the tank (the dc gain of this transistor is used in the automatic threshold circuit). A phase shifting stage with current feedback drives a switched output stage.

6.4.1 *Effective Q*

The effective Q of the tank circuit, Q_C, is given by

$$\frac{1}{Q_C} = \frac{1}{Q_S} + \frac{1}{Q_T} + \frac{1}{Q_L} \tag{76}$$

where

$$Q_S = \text{source } Q$$
$$Q_T = \text{tank } Q$$
$$Q_L = \text{load } Q$$

The source Q is controlled by the output impedance of Q_3

$$Q_S \approx \frac{\omega_0 L}{R_S \left[\frac{C_8}{C_7 + C_8 + C_{15}}\right]^2} \tag{77}$$

where R_S is given by

$$\frac{1}{R_S} = \frac{1}{R_9} + \frac{1}{R_0} \tag{78}$$

where

$$R_0 = r_e + \frac{r_b' + R_p}{1 + \beta}, \qquad R_p \approx \frac{R_{16} \times R_{19}}{R_{16} + R_{19}}, \tag{79}$$

r_e, r_b', and β are conventional parameters of Q_3, and L is the full inductance of T_4. R_S is about 15 ohms for the minimum allowed $\beta(\approx 20)$, so for the parameters of Fig. 31

$$Q_S \approx 270. \tag{80}$$

The tank Q is controlled by many factors. Carbonyl iron possesses the required stability, and the resulting Q_T is quite dependent on inductor size. Three inductor sizes were developed, all 100-μhenry coils. The larger

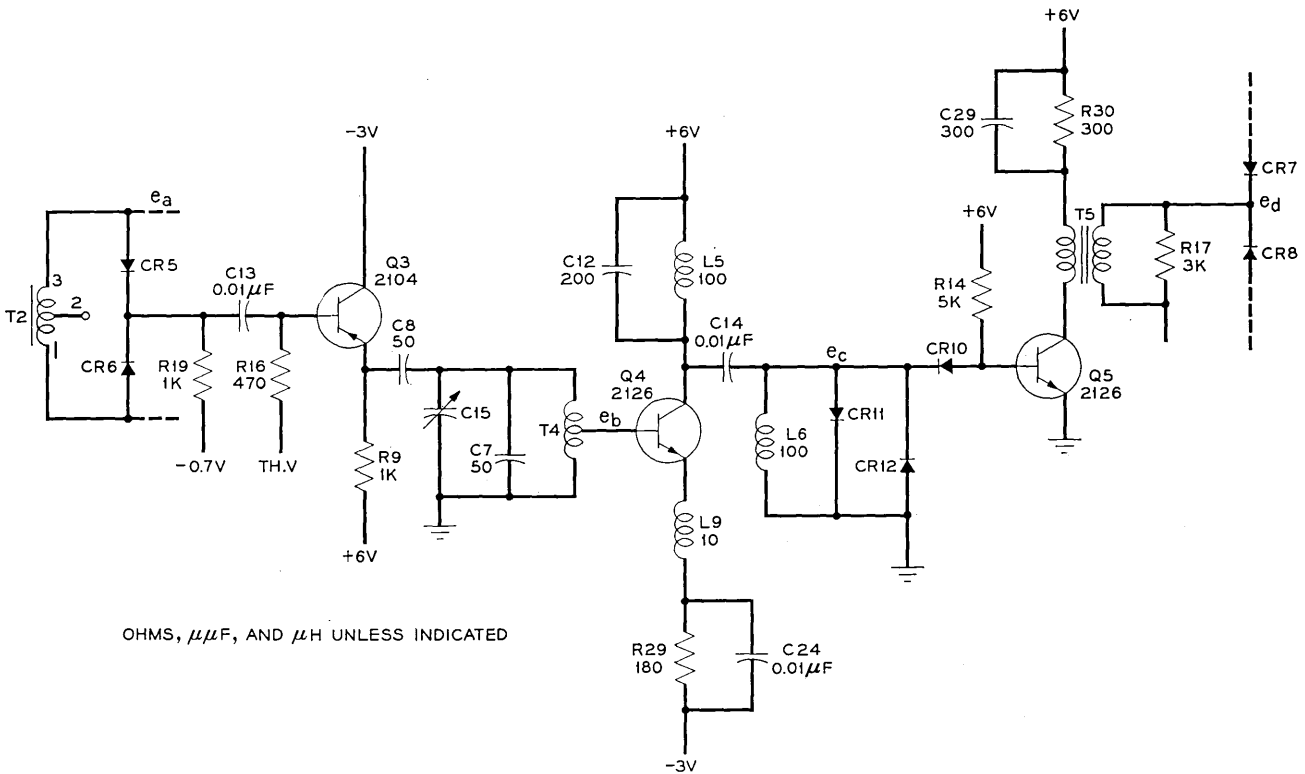


Fig. 31 — Repeater clock circuit.

inductor has $Q_T \approx 200$; the intermediate size inductor has $Q_T \approx 150$; and the smaller inductor has $Q_T \approx 100$. The intermediate size was selected to give an over-all circuit Q of 75–100.

The load Q is controlled by R_i, the input resistance to Q_4

$$\frac{1}{R_i} = \frac{1}{\omega_\beta \beta_0 L_9} + K\omega_0 C_c \tag{81}$$

where

ω_β = common emitter cutoff frequency of Q_4

β_0 = low-frequency common emitter gain of Q_4

K = voltage gain of Q_4

C_c = collector capacity of Q_4.

The first term represents a conductance of about 0.1 millimho and the second term a conductance that may vary with signal level from about 0.2 millimho to almost zero. The load Q becomes

$$Q_L = \frac{n^2 R_i}{\omega_0 L}. \tag{82}$$

With $n = 10$, for nominal signal levels $Q_L > 330$.

The resulting minimum effective Q is approximately

$$\frac{1}{Q_C} = \frac{1}{270} + \frac{1}{150} + \frac{1}{330} \tag{83}$$

$$Q_C = 75. \tag{84}$$

6.4.2 *Gain*

The voltage gain of the tank is

$$\frac{e_0}{e_i} \approx \frac{j}{n}\left(\frac{C_8}{C_7 + C_8 + C_{15}}\right) Q_C, \qquad \text{for } Q_C \gg 1. \tag{85}$$

Using the parameters stated,

$$\frac{e_0}{e_i} = j3.75. \tag{86}$$

The signal delivered to the base of Q_4 is then in the range 0.1 to 2.3 volts peak. For random patterns with average line loss, the average level delivered to the base of Q_4 is about 1 volt peak.

A voltage gain of five in Q_4 brings the signal level up to the point where the input gate to Q_5 can be precisely operated. Clipping diodes

CR11 and CR12 limit the collector swing of Q_4, which operates in its linear region for all signal levels.

One would expect a square wave of clock out of Q_5. This would be true, except that the primary inductance of T_5 is selected small (<60 μhenry) so that the transformer output overshoots measurably. This overshoot is used as the time crosshair for the regenerator input gate. The output of T_5 is then a three-level signal of shape shown in Fig. 32. The spike overshoot "initiates" the regenerator action and the negative excursion turns off the regenerator, thereby controlling regenerator output pulse width. T_5 must deliver enough current to rapidly shut the regenerator off; it must accept the gate current after the sample instant (if no received pulse is present), making the regenerator immune to any received signal changes that occur outside the sampling instant.

6.4.3 *Clock Placement*

The approximate phase relation of the clock signals is shown in Fig. 33. The clock spike is nominally centered in the "eye." The main sources of clock misplacement are:

i. Mistuning Effects

(a) LC temperature coefficient mistracking	$\pm 16°$ max
(b) LC aging	$\pm 11°$ max
(c) mistuning due source	$\pm 0.5°$ max
(d) mistuning due load	$\pm 6°$ max

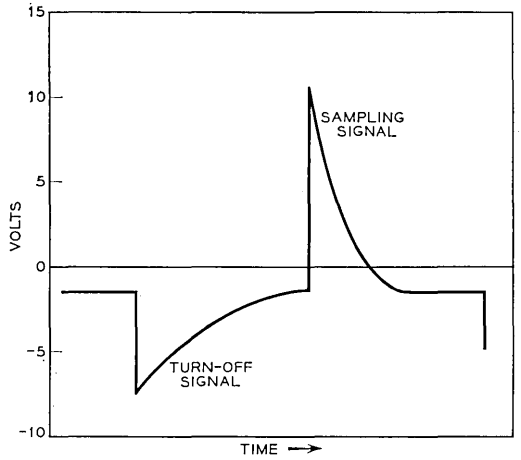


Fig. 32 — Idealized clock waveform.

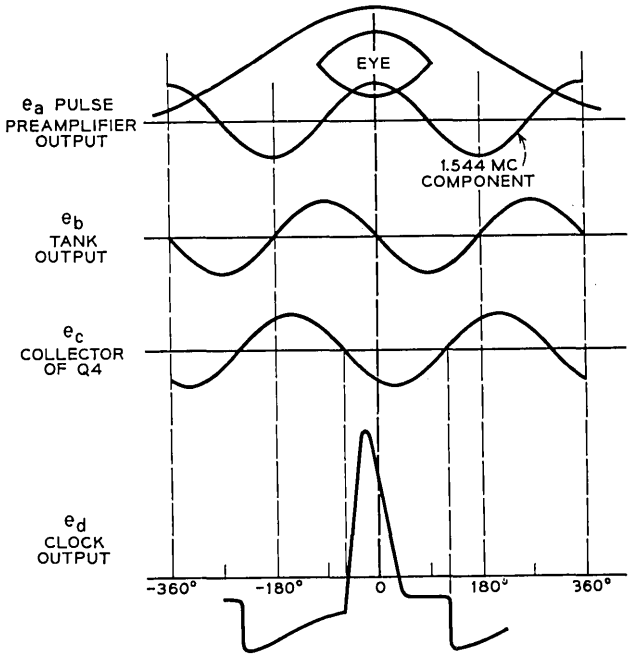


Fig. 33 — Clock phase at various points of Fig. 31.

ii. Placement Effects.

(a) tolerances on L_5, L_6, C_{12} (5%)	±5° max
(b) diode balance ($CR11$, $CR12$)	0.5° max
(c) V_{BE} variation in Q_5	±8° max
(d) storage time of Q_5	±5° max

The effects under *ii* are additive; the effects under *i* can be added as long as the sum is small (directly on a $\Delta f/f$ basis). Direct addition gives ±52°. An r.s.s. addition gives ±23°. Most of the items under *ii* occur only for worst-case transistors and maximum line loss with minimum received clock, and are reduced considerably for reasonably random pulse trains. Clock position variations from nominal of the order of 20° to 40° should be expected, producing about a 1-db penalty in NEXT performance in accordance with Fig. 30.

VII. REGENERATOR

The regenerator design for a bipolar repeater is particularly governed by economic factors. Indeed, the requirement of a balanced regenerator is one that must be pondered before a decision against unipolar trans-

mission is made. On the other hand, of the four basic configurations, there is one, the shunt-series configuration, that allows a rather simple circuit. This arrangement, shown in Fig. 34, requires only one transformer, and the action of the feedback is such that one set of gate diodes may be used both for spike sampling and turn-off. For this configuration spike sampling comes "free."

Although in many ways advantageous, the shunt-series circuit has one outstanding disadvantage. Since the feedback windings are directly coupled to the transmission line, any line reflections are coupled into the feedback winding and become "noise" on the signal. The problem is aggravated by the fact that the regenerator is a poor termination for any reflected signal.

If ρ is the voltage reflection coefficient due to a discontinuity in the line

$$V_{\text{reflected}} = \rho V_{\text{incident}}$$

and if midband loss (5 db per 1000 feet) is assumed applicable, then the voltage coupled into the feedback winding is

$$V_{fb} \approx \frac{\rho(1 + \Gamma)}{3^{l-1}} \text{ volts} \tag{87}$$

for a 3-volt transmitted pulse, where l is the distance from repeater output to the discontinuity in kilofeet, and Γ the reflection coefficient the

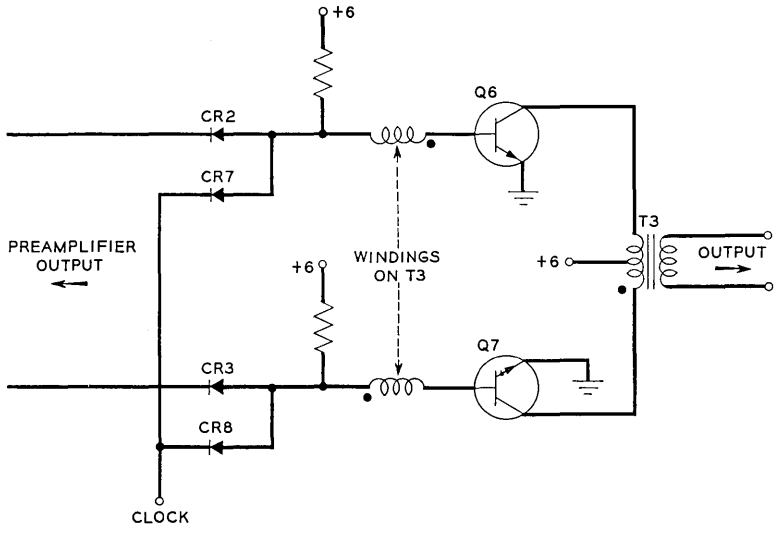


Fig. 34 — Regenerator circuit.

repeater output presents to the returned signal. It is planned that some matching will be provided by the coupling network to the fault-location system. Assuming a 1-volt eye half-height, V_{fb} reduces the allowable near-end crosstalk interference by

$$20 \log \frac{1}{1 - \frac{\rho(1 + \Gamma)}{3^{l-1}}} \tag{88}$$

a worse case occurring when the reflected signal peak occurs at the sampling instant. Equation (88) is plotted in Fig. 35. These curves vividly illustrate the gravity of line discontinuities close to a repeater output.

There are many ways to make the regenerator more tolerant or completely tolerant of line reflections, but all methods add to the cost. A small resistive pad between regenerator and line aids considerably. This can be inserted by either increasing output level, decreasing signal

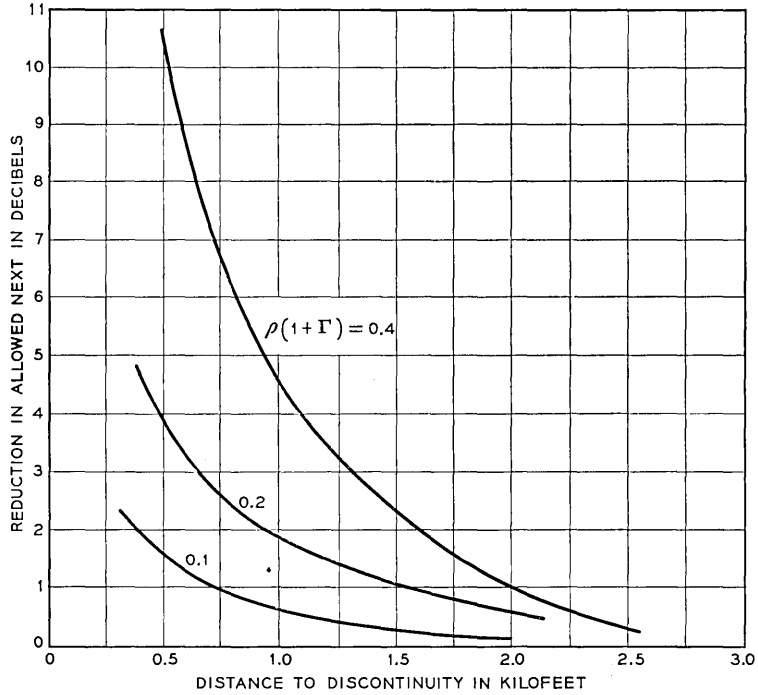


Fig. 35 — Effect of reflected signals on NEXT performance.

levels, and/or increasing preamplifier gain. Up to 6 db of padding is possible by a combination of these means without materially affecting the repeater design. It does not appear that further action is required, though a "bad environment" regenerator has been considered. A possible implementation is shown in Fig. 36; the feedback windings have been placed on a separate core.

7.1 *Requirements*

The following requirements have been set for the regenerator:

Height of output pulse	$3v \pm 0.3$ volt
Unbalance in height of positive and negative pulse	± 0.15 volt
Width of output pulse (half amplitude)	0.32 ± 0.03 μsec
Unbalance in width of positive and negative pulse	± 0.015 μsec
Maximum rise or fall time	0.090 μsec

The variation in the height of the output pulse is primarily controlled by the variation in supply voltage. The 6-volt supply of Fig. 34 is derived from a zener diode, ± 10 per cent voltage control being quite economical. The effect of the transistor "on" voltage is extremely small,

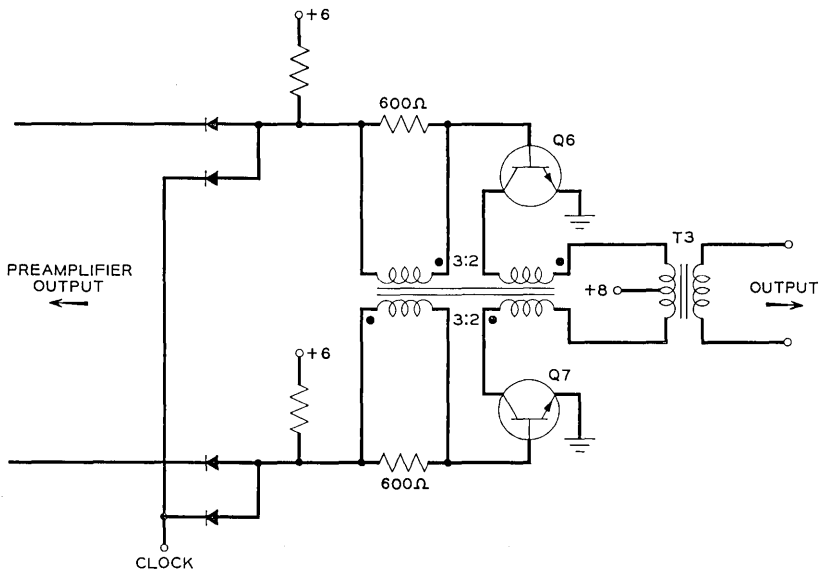


Fig. 36 — Modified regenerator.

$V_{CE_{SAT}}$ being typically 0.3 volt for the transistor used for the operating conditions designed. The allowed unbalance requirement is primarily a transformer balance requirement. The effect of $\Delta V_{CE_{SAT}}$ is slight, typically about 0.1 volt (compared to 6-volt supply).

The width of the output is controlled by the clock and transistor rise and fall times and the response of the output transformer. The transistor requirements derived allow a rise time of 0.06 ± 0.01 μsec and a fall time of 0.035 ± 0.015 μsec. For the diffused silicon transistors used, the turn-on time requirements set the gate current at 2 ma, giving a switching current gain of 7.5. The requirement on pulse width control greatly dominates any criterion for regeneration.

7.2 *Feedback Ratio*

The feedback voltage is bounded in magnitude by reverse emitter breakdown in the regenerator transistors on the one hand and by noise margin on the other. The peak negative signal delivered by the preamplifier to the gate is approximately $3/2\ E_m$. The maximum reverse voltage seen by the base-emitter junction of the blocking oscillator transistor is

$$V_{BER} = 3/2\ E_m + V_f - V_d \tag{89}$$

where V_f is the feedback voltage and V_d is the gate diode forward drop, typically 0.7 volt. The curves labelled "max" in Fig. 37 show the allowed V_f as a function of peak received pulse amplitude for various values of $V_{BER_{max}}$. For $V_{BER} = 7$ volts (minimum for the transistors used), and $E_m = 3$ volts, V_f must be less than 3.2 volts.

When one half of the regenerator is operating, its gate point falls V_f volts. Should the signal fall more than V_f volts below threshold, it may turn off the regenerator prematurely. The voltage at the preamplifier output is approximately

$$v = e_s + n(t) - \frac{E_m}{2} \tag{90}$$

where e_s is the received signal and may be represented by (worst case)

$$e_s = E_m \sin \omega_0 t \tag{91}$$

$$v = E_m(\sin \omega_0 t - \tfrac{1}{2}) + n(t). \tag{92}$$

If regenerator turn-on is initiated at the peak of the received pulse, the greatest change that can occur in the received signal (less noise) while

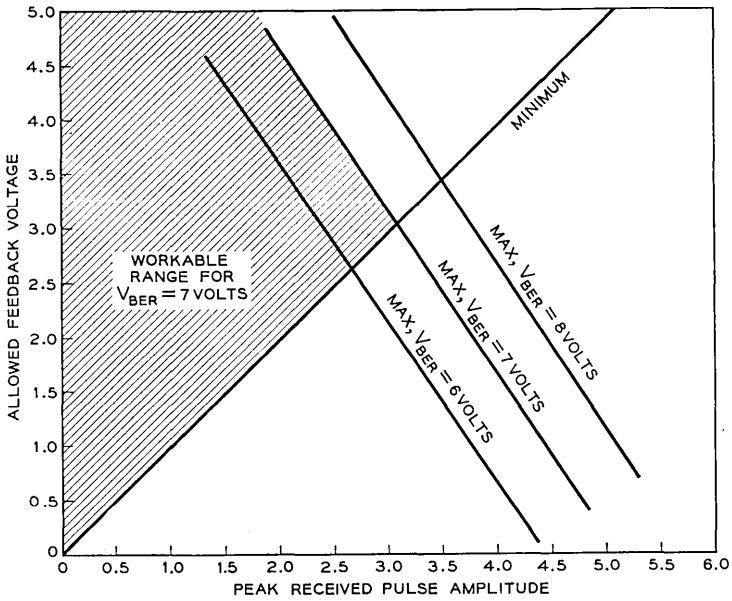


Fig. 37 — Maximum and minimum allowed feedback voltage.

the regenerator is on is $E_m/2$, so the most negative the gate can go during regeneration is

$$v_{\max} = \frac{E_m}{2} + n(t)_{\max}. \tag{93}$$

The $n(t)_{\max}$ cannot exceed $E_m/2$ in a workable repeater so

$$v_{\max} \leqq E_m. \tag{94}$$

It is then safe to feed back a voltage as small as E_m, as shown by the curve labelled "minimum" in Fig. 37. Thus, for $E_m = 3$ volts, V_f must be $\geqq 3$ volts to ensure uninterrupted regeneration and V_f must be $\leqq 3.2$ volts to prevent reverse emitter breakdown. The feedback voltage is thus set equal to the regenerator output level (3 volts), making the primary-to-output-to-feedback windings ratio for T_3 equal to $4CT$:-$1CT$:1:1.

Leakage inductance in T_3 affects the turn-on of the regenerator, particularly the "maybe" region of regeneration. If the received signal exceeds threshold by Δ volts at the sampling instant, this voltage is im-

pressed across the feedback winding and must produce sufficient current in the base of Q_6 or Q_7 to turn these units on within the width of the time crosshair, or perhaps more importantly, must not greatly affect the output rise time. It appears quite sufficient if a $\Delta = 0.1$ volt allows the base current to build up to 0.5 ma in 0.01 μsec, or the leakage inductance of T_3 is less than 2 μhenrys looking into the feedback winding. If the high-frequency cutoff of T_3 is due to leakage inductance, this cutoff must then exceed 7 mc.

Capacitance of the feedback windings-to-ground affect the delay and rise time of the regenerator. With the clock output returned to -1.5 volts and 2 ma gate current, the delay due to gate capacitance is 0.7 C nanoseconds with C in $\mu\mu$f. Capacitances up to 20 $\mu\mu$f can be tolerated provided the two gates are approximately balanced. Because the transformer used has unbalanced feedback winding capacitances (approximately 15 and 25 $\mu\mu$f), the gate currents have been made to differ slightly to produce the same pulse widths for both the positive and negative pulse output. Perhaps a better arrangement, at the expense of a component, is to build out the capacity of the low-capacity winding.

It is worth noting that the regenerator may be triggered into self-oscillation in the absence of a received signal (clock). This cannot be prevented so long as the same gate diode is used for spike sampling and turn-off. Actually, the amount of received clock required to dampen the self-oscillation is so small that the clock energy received via near-end crosstalk is generally sufficient to dampen the oscillation. The oscillation, when it occurs, is not detrimental to the repeater circuit.

The RL network shunted across the repeater output as a nonlinear equalizer (Fig. 19) has two important effects on the regenerator that will be treated only superficially. First, the afterkick of the network is coupled into the regenerator feedback windings. This has the effect of reducing intersymbol interference or, equivalently, of partially invoking the bipolar rule at each repeater. The feedback voltage is made equal to a few tenths of a volt at the next sample instant to discourage, to the extent desired, the output of successive pulses of common polarity.

A second effect of the network is that the afterkick coupled into the feedback winding affects the rise time of the output. The net effect is that when an output pulse is immediately preceded by a pulse (which must be of opposite polarity), it rises sooner than would be the case otherwise; this additional area on the front of the pulse aids in cancelling the tail of its predecessor, thereby further reducing intersymbol interference. Because the faster rise time is pulse-correlated, its effect on the position of the received pulse peak is straightforward, and properly

applied, the technique can even improve the pattern jitter performance of a repeater.

VIII. SECONDARY FEATURES

8.1 *Power Arrangement*

Repeaters are to be powered over the signal pair. Of all the series and shunt power arrangements, the simplex current loop of Fig. 38 is the most attractive. A constant current is fed over the phantom of the EW and WE circuits, and the voltage at each repeater is obtained across a zener reference diode.

For 22-gauge cable, the maximum line resistance (140°F) in the current path is approximately 116 ohms for 6000-ft span length. Optimum power transfer to the repeater circuit then occurs when the current demanded by the repeater, I_0, and the operating voltage for the repeater, E_0, satisfy

$$E_0 = 116\, I_0 . \tag{95}$$

Diffused-base transistors are noted for their alpha defect at low current, and an average current of about 5 ma/stage is required to maintain reasonable gain. This accounts for 35 ma per one-way repeater. Power considerations make it not always possible to design for minimum current, and considering base drives and other current demands, minimum current drain per one-way repeater for the signal levels desired and the transistors used is approximately 50 ma.

For reasons of economy, as well as ability to power maximum-length spans, it is convenient to design one power unit for a two-way repeater

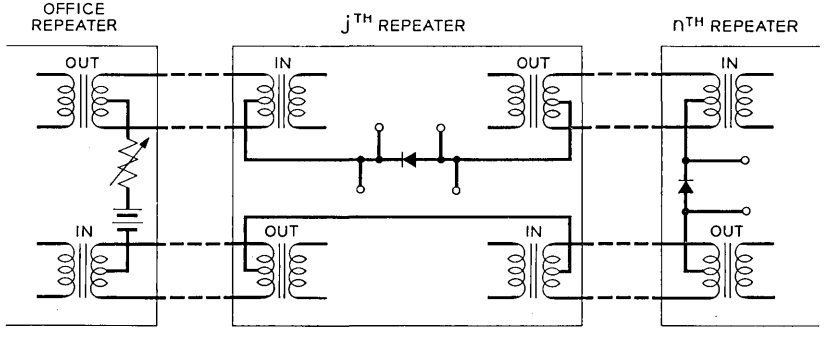


Fig. 38 — Simplex power loop.

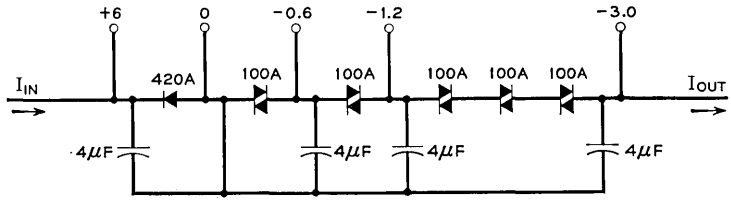
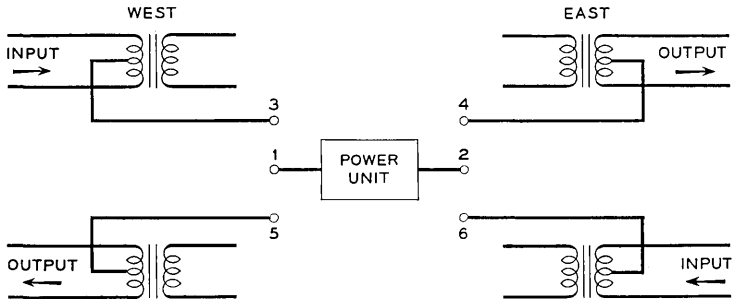


Fig. 39 – Twin repeater power unit.

(twin repeater circuits in one enclosure). The repeater designed is such a "twin" unit with common power supply. It operates with a minimum current of 110 ma and has a voltage drop of 9 volts (which is almost optimum for nominal temperature). This voltage is derived from a 6-volt zener diode and several click-reducer diodes as shown in Fig. 39. The voltage required for the power loop is then 22 volts per section. Using a +130 and −130 volts in the office, 11 sections may be powered, corresponding to 12.5 miles per office or 25 miles maximum interoffice spacing.

The power options on the repeater are shown in Fig. 40. Power may be routed through the repeater in either direction. It may be looped back readily, and under the loopback condition, the repeater at the loopback point may be powered from either end of the line. The option requires 6 terminals. By eliminating some of the minor options, the number of terminals may be reduced.



1. STRAIGHT THROUGH	1 TO 3,	2 TO 4,	5 TO 6
2. LOOPBACK, POWER FROM EAST	1 TO 6,	2 TO 4,	3 TO 5
3. LOOPBACK, POWER FROM WEST	1 TO 3,	2 TO 5,	4 TO 6

Fig. 40 — Repeater power options.

The two halves of the "twin" repeater, with midspan power loopback, serve 24 two-way voice channels or 48 one-way channels.

8.2 *Protection*

The transistors used are low-power devices, and any transistor circuit attached to exposed cable must consider the problem of lightning surge activity.

P. A. Gresh and D. W. Bodle of Bell Laboratories have divided the plant into three exposure classes: (1) low exposure typified by underground plant, (2) moderate exposure typified by aerial and buried cable in suburban and rural areas, and (3) high exposure typified by unshielded

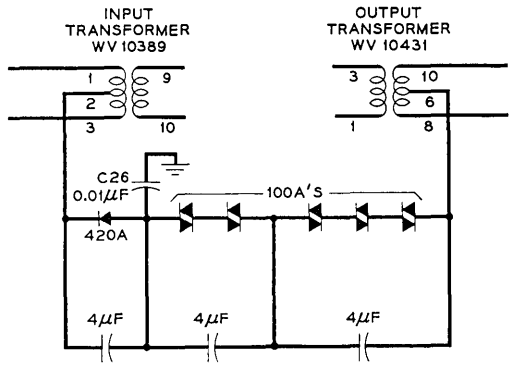


Fig. 41 — Repeater power circuit.

facilities such as open wire or distribution wire. Only the first two classes will be considered as potential PCM environments.

8.2.1 *Repeater Capability*

Repeaters may be subjected to either longitudinal or metallic voltage or current surges. Longitudinal surge currents flow through the power circuit (Fig. 41), producing currents at the secondary of the input transformer and the primary of the output transformer diminished by 40 db (the balance requirement placed on the transformers). The primary effect of longitudinal surge current is in the power zener. The unit used, WE420A, has a reverse surge current capability of the order of 10 amperes.*

* All surge currents will assume a near triangular wave of 10 μs rise and 600 μs fall times.

Longitudinal voltages are impressed across the by-pass capacitor C_{26}, and across the entire repeater circuit to ground. The circuit to can and C_{26} breakdown voltage must exceed expected surges, and has been set at 1000 volts. Longitudinal voltages may exist, particularly at power loopback points, between the repeater input and output terminals and the repeater circuit. There, transformer windings must be insulated against surge voltages, a difficult situation with small pulse transformers. However, 1000 volts surge breakdown has been achieved in the transformers used.

Metallic currents induce sizeable currents into the transistor circuits. These currents cannot readily damage the input transistors Q_1 and Q_2 because of the large emitter and collector resistors associated with these stages. The output transistors Q_6 and Q_7 are not so protected. Tests indicate these transistors will fail for metallic line surges of 5 to 10 amperes. Further protection is available by limiting the metallic voltage by shunting the line with semiconductor diodes, such as the 100A click-reducer diode, which is capable of surge currents up to 80 amperes. Six or more units must be used to prevent clipping of the repeater output signal. These diodes also protect the transformer windings which, however, have been found in limited tests to be able to withstand surges of 50–100 amperes.

8.2.2 *Class 1 Repeaters*

Gresh and Bodle found very slight lightning activity in their Class 1 environment. The 1000-volt longitudinal voltage capability and the 10-ampere longitudinal and 5-ampere metallic surge current capability should be sufficient for this class, except for repeaters mounted in central offices with carbon block line protection. Breakdown of the carbon block will convert longitudinal surges to metallic surges and may damage the output transistors. For this reason it is felt that with carbon blocks on a PCM line, the repeater output should be shunted by the 100A click-reducer diode string. At the office, the power unit of the repeater is generally returned through a current-limiting resistance to a power source. This resistance should not be less than 10 ohms so that carbon block breakdown does not damage the zener diode.

Then, for Class 1 exposure, line repeaters have no special protection. Office repeater outputs have 100A diode protection.

8.2.3 *Class 2 Repeaters*

For Class 2 exposure, the 1000-volt insulation and 10-ampere longitudinal current capability are inadequate. There are two major alterna-

tives: (1) better insulated input and output transformers and a heavy-duty zener diode, or (2) carbon block or gas tube protection for each repeater. The problem is still under study. The interim solution uses carbon block and 100A diode protection on both the input and output of each Class 2 repeater, and builds out the power supply impedance with a small resistance (10–20 ohms) to limit the surge current. A better solution would likely require a high-power zener diode, and gas tube protection for the transformers, all built within the repeater can.

8.3 *Maintenance Equipment*

8.3.1 *Error Meter*

One acid test of the performance of a repeatered line is a test of line error rate. Indeed, if the basic design assures adequate phase jitter performance, the error rate test is all important.

With pseudo-ternary transmission according to some basic law, such as bipolar, measurement of error rate is very simply accomplished by observing the number of pulses per unit time that violate the basic code. For bipolar transmission one looks for those pulses that fail to obey the alternate polarity rule. A circuit to accomplish this is shown in Fig. 42. Its operation is straightforward and will not be discussed. Suffice it to

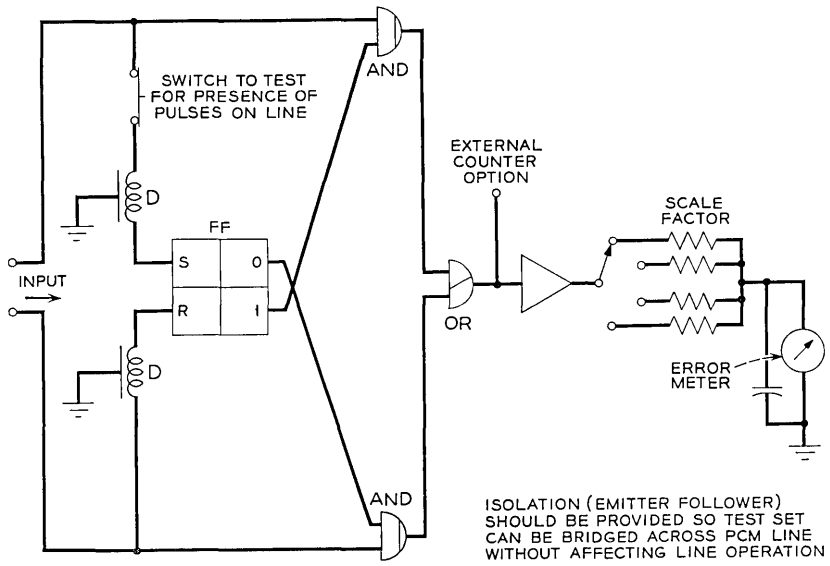


Fig. 42 — Proposed error rate meter for exchange carrier PCM.

say the circuit detects all single errors and most multiple errors. A comparison of bipolar violations per second against actual errors per second for various error rates has been established empirically and is shown in Fig. 43. There are, of course, certain types of errors that could conceivably occur in large numbers that would not be detected by the error meter.

The meter may be placed across an operating line to evaluate the performance of the line without affecting the operation of the line. Conceptually, it is as simple to use as a simple voltmeter.

8.3.2 *Marginal Checking*

The contemplated PCM system operates with large numbers of tandem repeaters. Provision for remote identification of a faulty repeater is a requisite, particularly in the exchange cable environment characterized by severe NEXT. Lacking a complete description of NEXT (particularly the tail of the distribution), it seems likely that an aging repeater will make excessive transmission errors before failing altogether. Indeed, a PCM route may work satisfactorily for years, then

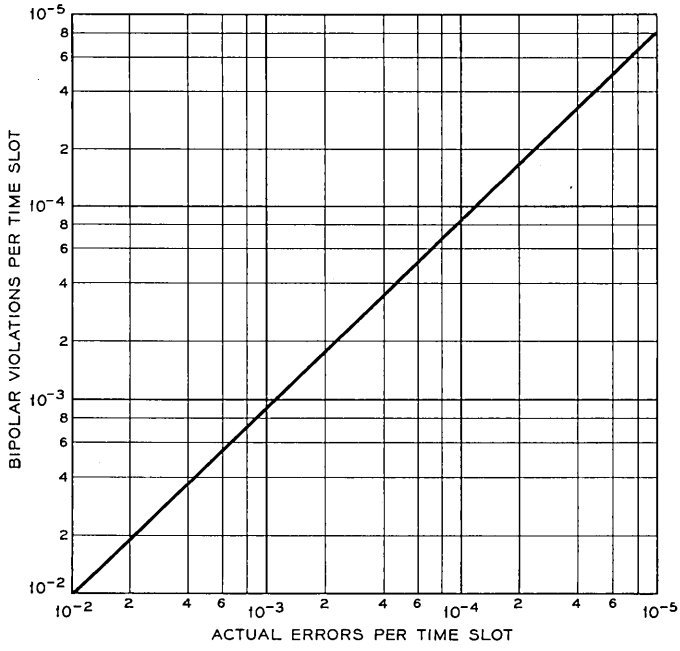


Fig. 43 — Comparison of error rates, actual and as indicated by error meter, for errors generated by exposing working repeater to random noise.

fail due, perhaps, to increased density of PCM in the cable. Remote marginal checking, as well as fault location, seems extremely desirable.

What meaningful parameter may be varied from a central office to produce failure of a line? The author has concluded that in view of the pulse-detection process, the most meaningful parameter available for remote testing is the balance of the transmitted bipolar pattern. This parameter gets to the very heart of the repeater — the decision-making — for the greater the unbalance in the pulse train, the smaller the opening of the eye. The marginal check parameter is then the number of unipolar pulses that may be superimposed on a bipolar wave before a repeater malfunctions.

Perhaps the simplest method for identification of one of many serially operated units associates an identification tone with each equipment location. It is then convenient to associate a particular marginal check frequency and filter with each manhole along a repeater route. The filter may serve many repeaters and is coupled to the outputs of the many repeaters at a common location. A special test PCM pattern may be transmitted to excite any selected repeater along a route. The output of the filter is returned to the office over a marginal checking cable pair. The general plan is shown in Fig. 44.

If the identification frequencies are in the audio band, a balanced bi-

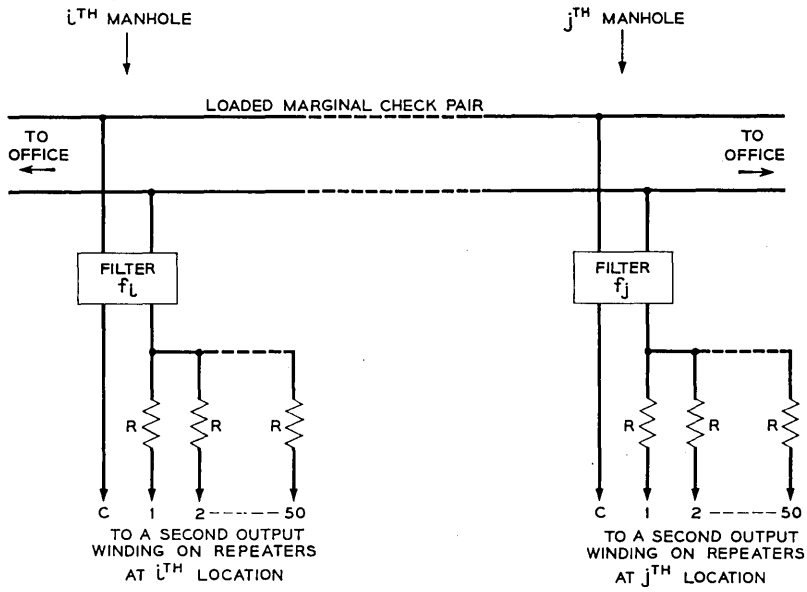


Fig. 44 — General plan for marginal checking 25 two-way systems.

polar pulse train contributes very little energy to the filter output. On the other hand, added unipolar pulses gated on-off at an identification frequency provide considerable excitation to the filter. Furthermore, the return signal should be proportional to the unipolar density up to the point of repeater failure. The margin of operation of a repeater is then given by the number of unipolar pulses that may be added to a bipolar train before the return signal fails to be proportional to unipolar density. In practice, a sparse unipolar density is transmitted and the return signal amplitude adjusted to a reference mark on a meter. As the unipolar density is increased, the meter must go to calibrated markings. When the meter fails to toe the mark, the unipolar density is a measure of the margin in the repeater under test or any repeater that precedes it. The indication that the meter gives under the marginal condition is an indication of the type of failure.

How many repeaters may a single filter serve? A random bipolar signal with probability of a pulse being present equal to 0.5 has a power density spectrum

$$W(\omega) = \frac{1}{16\pi^2}(1 - \cos \omega)\left(\frac{\sin \omega/4}{\omega/4}\right)^2 \tag{96}$$

where frequency has been normalized to the bit rate, and the signal is assumed to be one volt peak amplitude applied to a one-ohm load.

If $\omega \ll 1$

$$W(\omega) \approx \frac{\omega^2}{32\pi^2}. \tag{97}$$

The power resulting from passing N such signals through a band-pass filter of bandwidth Δf is

$$P_N = \frac{\omega^2 N}{16\pi}\Delta f. \tag{98}$$

A unipolar signal of pulse density d burst on-off at frequency ω_1 produces a signal of amplitude d/π, or power

$$P_S = \frac{d^2}{\pi^2}. \tag{99}$$

Thus, the signal returned consists of P_N due to the disturbance of N operating systems and P_S due to the desired return signal

$$\frac{P_S}{P_N} = \frac{16d^2}{\pi N \omega^2 \Delta f}. \tag{100}$$

Even for a 20-db signal-to-noise ratio, a sparse unipolar pattern ($d = 1/32$), with a fault-location frequency of 3 kc and a $2\pi\Delta f = 0.01\omega$ (typical conditions), N is extremely large. Consequently, when working in the audio range, one filter may readily serve 50 repeaters from an interference point of view. Equipment boxes will be designed for 25 two-way repeaters, and for various reasons it is felt that there should be one marginal checking pair per repeater box.

Perhaps the greatest problem in this area lies in absolute signal levels. If there is no power gain in the marginal check filter and 50 repeaters are resistively coupled to that filter, and if the output of the filter is transmitted up to 12.5 miles back to an office, the received power level is exceedingly small. Typical major losses are given in Table V.

To keep line transmission loss down to 0.9 db/ mile, audio frequencies and loaded cable have been employed. The signal available at the repeater output is about -27 dbm for $d = \frac{1}{32}$.

The situation is summarized in Fig. 45. For example, for $d = \frac{1}{16}$, transmission over 12.5 miles of loaded cable (requiring measurement from each end of a 25-mile system) produces a signal level at the office of -77 dbm. For $N = -70$ dbm (thought to be near worst-case office noise level across the 1.5 to 3-kc band if certain harmonics of 60 cps are avoided), the receiver bandwidth must be less than 75 cycles for a 6 db S/N ratio. Or a signal with $d = \frac{1}{32}$ over 12.5 miles requires a receiver bandwidth of about 20 cycles for a 6-db S/N ratio. The situation is somewhat different if peak detection is used (20 log bandwidth ratio applies).

It is felt that reasonable unipolar densities are $d = \frac{1}{32}, \frac{2}{32}, \frac{3}{32}, \frac{4}{32}$. The repeater will be only slightly affected by $d = \frac{1}{32}$, but will almost surely fail for $d = \frac{4}{32}$. The effect of the added unipolar pattern in reducing operating margin is shown in Fig. 46. The difference in performance for positive or negative added pulses results from the action of the half-wave rectifier that controls the automatic threshold circuit.

With $d = \frac{1}{32}$, there is a quantized set of tones available as fault-loca-

TABLE V — MAJOR SIGNAL LOSSES

Repeater Output Transformer (1.5 kc)	16.0 db
Filter Matching Loss	4.0 db
Filter Transmission Loss	3.0 db
Coupling Loss	16.5 db
Other Losses	5.5 db
	45.0 db

Line Transmission Loss 0.9 db/mi.

tion frequencies, corresponding to the number of unipolar pulses allowed per burst. There are 17 such frequencies in the band 1.005 to 3.017 kc. Such close-spaced frequencies put rather stringent requirements on the marginal check filter. The filter should offer 20-db rejection of adjacent frequency tones, provide maximum power transfer to the loaded pair, not load the audio pair at frequencies out of its passband, and have a pass-band, preferably, that is independent of the distance from the line-driving point to the nearest load coil. Suffice it to say that such a filter has been designed.

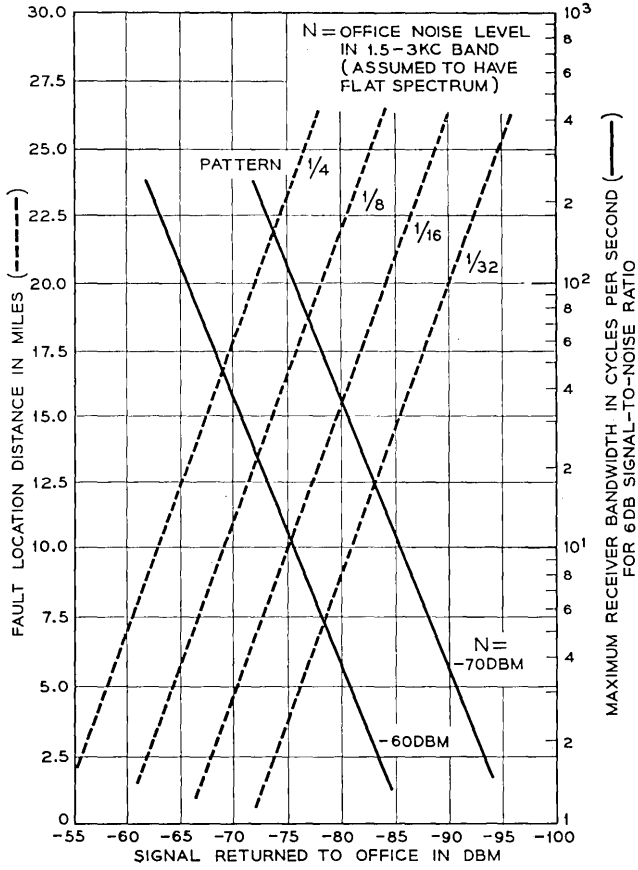


Fig. 45 — Received signal for various distances and patterns, and required receiver bandwidth for various received levels and office noise.

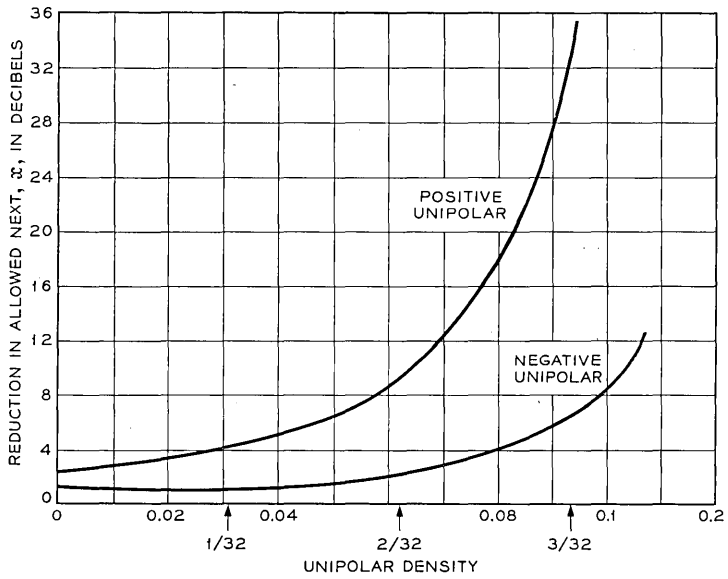


Fig. 46 — Effect of added unipolar pulses on NEXT performance.

8.4 *Equipment Design*

The guide words for the equipment design are size and reliability. Size is of utmost importance, for a repeater is expected to be manhole mounted, and manhole space is quite valuable. Manhole enlargement is generally sufficiently expensive that, if required, it represents a sizeable fraction of repeater cost, even when divided up among many repeaters. On the other hand, a manhole repeater must be extremely reliable. It would be a mistake to sacrifice reliability for size, for example, for the feasibility of operation of the proposed system demands a very high degree of reliability. For example, a 25-mile repeatered line will contain about 350 transistors, 675 diodes, and 2350 other components. With a reliability figure of 0.1 per cent per 1000 hours for transistors, the line could be expected to fail every 3000 hours (125 days) from this cause alone. At 0.01 per cent per 1000 hours, the line should fail every 1250 days (3.4 years). A complement of 25 repeatered lines along a 25-mile route would develop trouble every 5 days for the 0.1 per cent figure and every 50 days for the 0.01 per cent figure. It is expected that routes will be sectionalized and spare lines incorporated into the system.

So, miniaturization should not be achieved at a sacrifice to reliability.

For this reason, standard, well-known components are used throughout. Only the timing inductor represents a special component development. Diminutive size is achieved by a three-dimensional packaging of the components.

The three-dimensionality is achieved by assembling sub-blocks of the circuit into small component modules, as shown in Fig. 47. These modules are assembled (as if they were components) onto a master board as shown in Fig. 48. The 135 components are packaged into a can of overall dimensions $5\frac{3}{4} \times 3\frac{1}{8} \times 1\frac{1}{16}$ inches.

There are three options on the two-way assembly. As shown in Fig. 49, the center screw terminal is a power option that controls the power flow through the repeater. An LBO option for each half of the repeater is located on the corners of the master board. The LBO network is a single connection for nominal cable length. For other lengths, a network secured by the three screws replaces the simple strap shown. The three screws make the required connections from LBO to printed wiring board.

Fig. 50 shows the printed wiring side of the repeater. Inputs, outputs, ground, and the marginal checking output appear on the printed wiring connector.

Thirty repeaters have been built in the manner of Figs. 48–50.

8.5 *Terminal Repeaters*

In addition to the normal line repeaters, special repeaters are required at the transmitting and receiving terminals.

The transmitting repeater utilizes the terminal clock. It has no crosstalk problem, and is peculiar only in that it must convert the unipolar

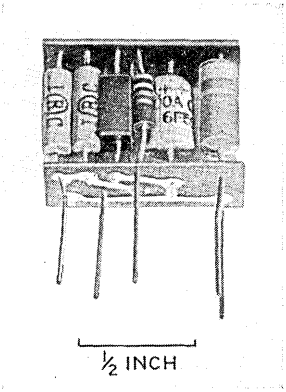


Fig. 47 — Small component module.

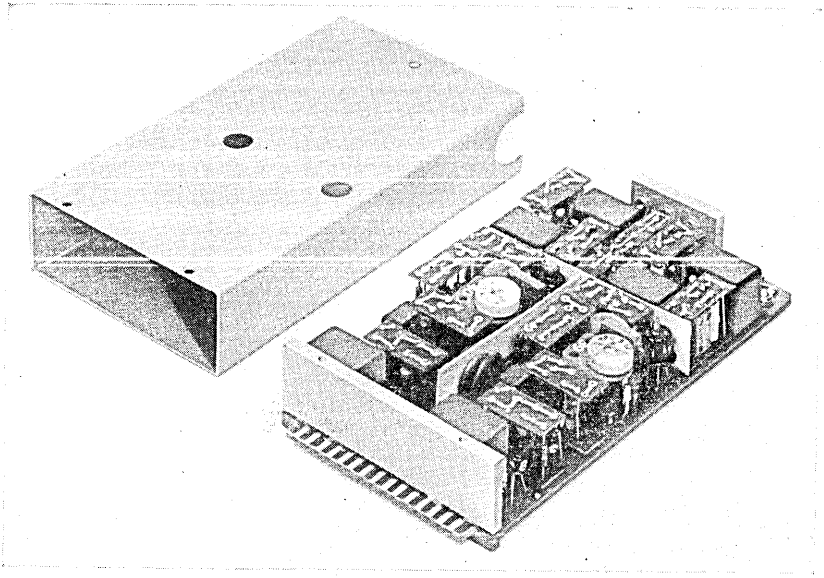


Fig. 48 — Experimental PCM repeater, cover removed.

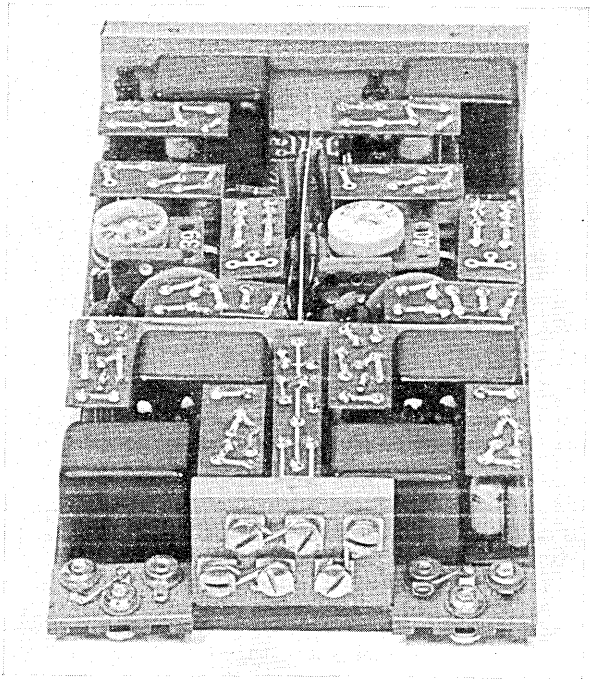


Fig. 49 – Another view of repeater circuit.

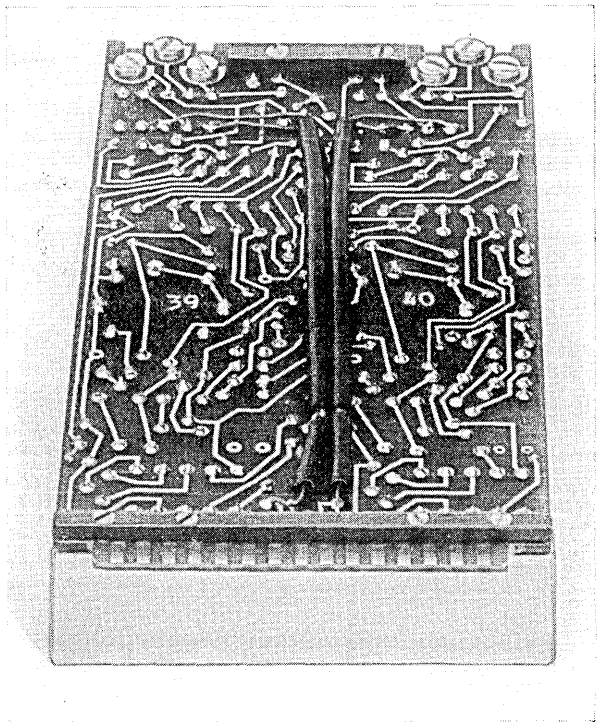


Fig. 50 — Printed wiring of repeater.

terminal signals to a bipolar pattern. This is done by use of a regular repeater regenerator with flip-flop control of the input gates to enforce the alternate polarity rule. The flip-flop is set and reset by the regenerator output. A block diagram is shown in Fig. 51.

The receiving repeater must deliver a unipolar signal and clock to the receiving terminal. The design is very close to a regular repeater with the bipolar regenerator replaced by an OR gate and unipolar regenerator. The output of the repeater clock circuit is used to drive a clock amplifier that provides a square wave of clock, properly positioned, for the receiving terminal. Fig. 52 shows a block diagram. In many ways a regular repeater circuit at the receiving terminal is desirable. In particular, in the design of Fig. 52, the error meter cannot be readily used at the receiving terminal.

These repeaters are physically a part of the experimental terminal, and the equipment design is consistent with the over-all terminal design.

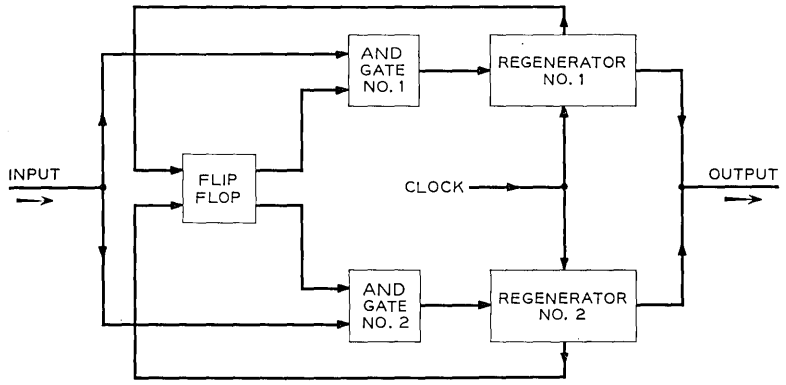


Fig. 51 — Block diagram of transmitting repeater.

IX. PERFORMANCE EVALUATION

Two parameters dominate the performance evaluation picture. First the repeater must maintain acceptable crosstalk performance over life and temperature. Secondly, the jitter due to random pattern changes and any other phenomena on the output pulse must be sufficiently controlled that a 25-mile system will deliver an acceptably stable signal to the receiving terminal. But first consider the simple operation of the repeater.

9.1 *Operating Waveforms*

Fig. 53(a) shows the repeater output signal which is applied to the transmission medium. The particular pattern displayed consists of three

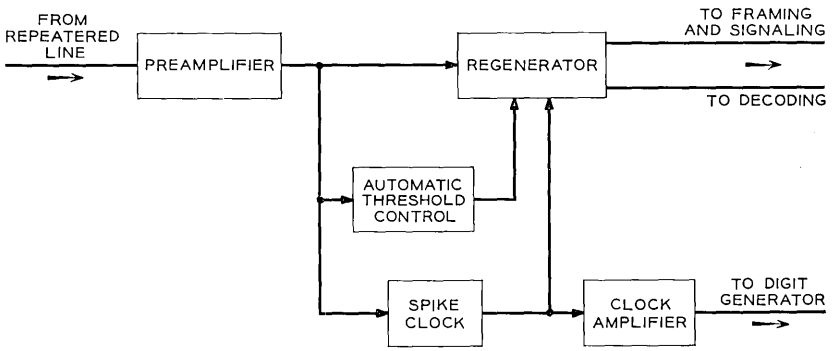


Fig. 52 — Block diagram of receiving repeater.

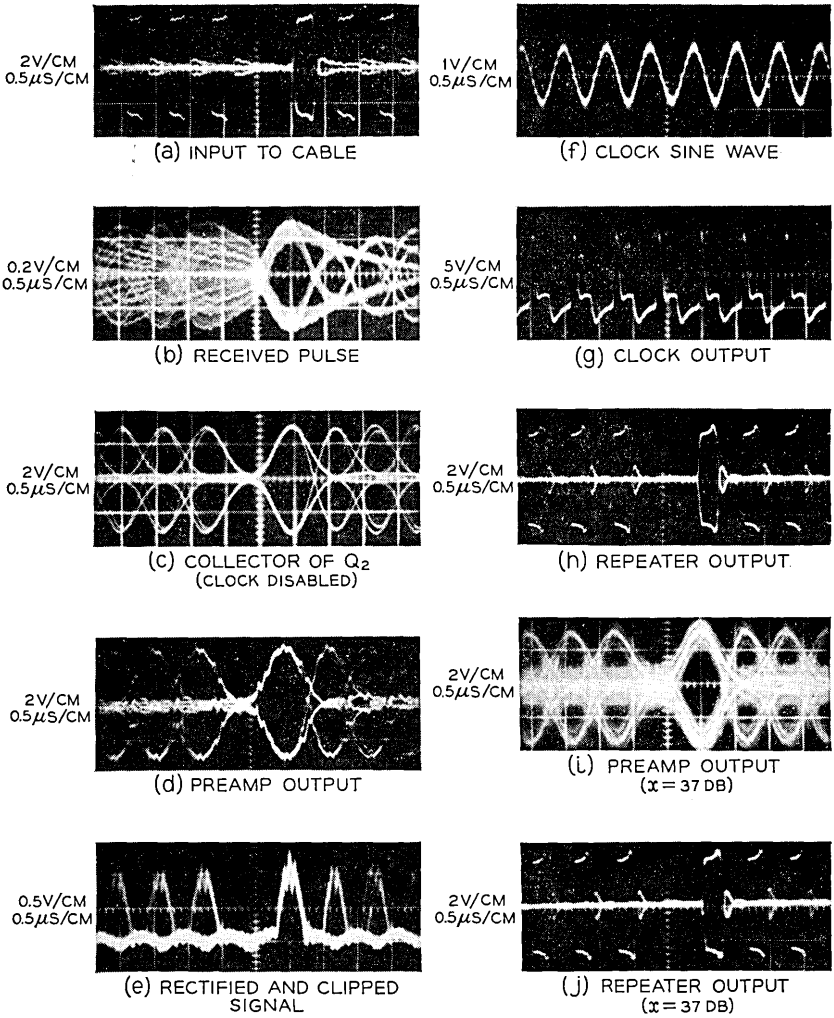


Fig. 53 — Operating waveforms. One large division in the illustration equals one cm.

random pulses followed by a forced space, followed by a forced pulse, followed by 3 random pulses. The pulses are 3 volts in amplitude.

Fig. 53(b) shows the signal after transmission through 6300 feet of cable as it appears at the secondary of the input transformer. The peak pulse height is about 0.25 volt, and a pulse is spread out over about 4 time slots.

Fig. 53(c) shows the amplified and equalized train as it appears on

the collector of the output stage of the preamplifier. The pulses are now about 3 volts in amplitude and essentially resolved to a single time slot. For this picture the repeater clock has been disabled.

Fig. 53(d) shows the signal at the secondary of the preamplifier output transformer. The signal has been shifted by the automatic threshold circuit to achieve the threshold and clock clipping function. The waveform has pronounced "nicks" that are a result of periodic gate current flow in the leakage inductance of the transformer.

Fig. 53(e) shows the rectified and clipped signal ready for delivery to the clock circuit.

Fig. 53(f) shows the sine wave at the tap point of the timing inductor.

Fig. 53(g) shows the clock output at the secondary of the clock output transformer. The positive-going spike samples the incoming wave; the negative-going spike shuts off the regenerator.

Fig. 53(h) shows the repeater output wave which is again applied to the transmission line.

Fig. 53(i) is the same as 53(d) with a crosstalk signal introduced through 37 db of loss (x = 37 db).

Fig. 53(j) is the output signal corresponding to Fig. 53(i). The output error rate is set to one error per 10^6 time slots.

9.2 *Crosstalk Performance*

Although work has been done with actual cable, it was convenient, for the most part, to use a simulated crosstalk path. An experimental crosstalk set-up is shown in Fig. 54. A small coupling capacitor is used to produce a crosstalk path-frequency characteristic that slopes at 6 db per octave in the frequency range of interest. For this set-up 1.544-mc

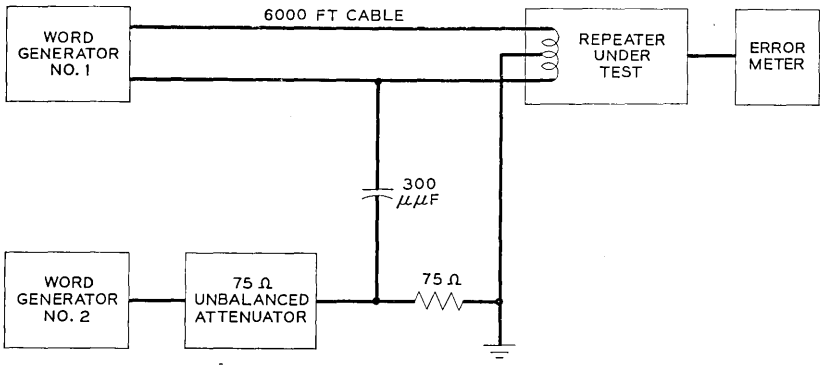


Fig. 54 — Experimental crosstalk simulator.

crosstalk coupling loss (x) is obtained by adding 24 db to the attenuator setting.

9.2.1 *Information Path*

Over-all crosstalk effect on repeater performance has been measured in terms of the allowed x versus the time and voltage crosshair positions, both varied forcibly without affecting other repeater conditions.

Typical effect of time crosshair position is shown in Fig. 55. This experimental curve should be compared to Fig. 30. The experimental clock sample width is 0.1 μsec. With perfect voltage crosshair positioning and performance, Fig. 30 predicts $x = 33$ db for $\epsilon = 0$; the measured value is 35 db. A shift of 30° ($\epsilon = 30°$) reduces the allowed crosstalk by 1.3 db ideally (Fig. 30). The measured value is 1.3 db for a 30° lead and slightly less than 2 db for a 30° phase lag. One concludes the allowed instability in repeater clock phase should not penalize crosstalk performance (x) by more than 1 to 2 db.

For similar conditions, the clock was unaltered and the threshold voltage varied. For each voltage crosshair position, the value of x that produces 10^{-6} error rate was recorded. The result is shown in Fig. 56. From Fig. 24, for a 2 db reduction in allowed NEXT, $\Delta = 0.10h_x$, and from

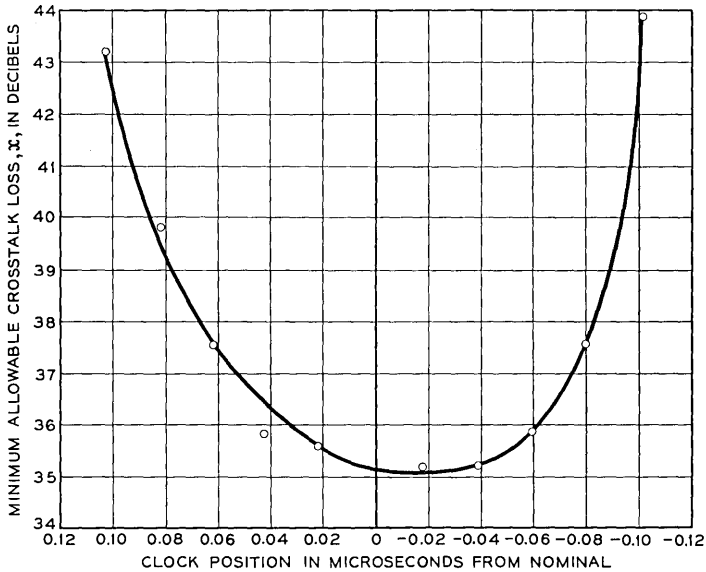


Fig. 55 — Typical measured effect of clock phase on crosstalk performance.

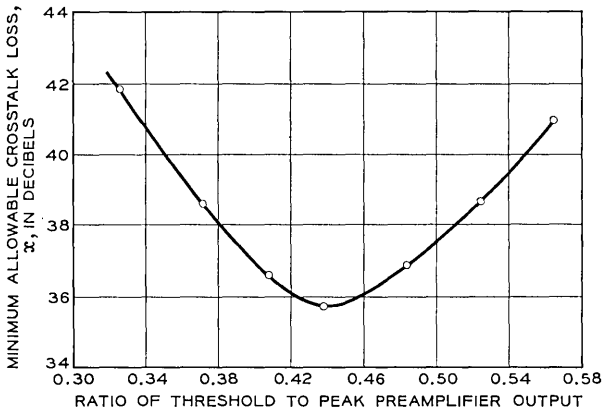


Fig. 56 — Typical measured effect of threshold on crosstalk performance.

Fig. 56, $\Delta = 0.06E_m$, suggesting $h_x = 0.6E_m$. It appears that only 60 per cent of the peak received pulse height exists as the eye height at the extreme of the jittered time crosshair position.* Fig. 29(a) predicts 30° of clock jitter for $x = 35$ db and Fig. 30 predicts a 1.3 db crosstalk penalty for the 30° shift, suggesting that the peak height of the eye is approximately 70 per cent of the received peak pulse height. Thirty per cent of the potential eye is thus lost to intersymbol interference, undershoot, bipolarity unbalance, noise pickup in the test cable, and deviation of reality from the analytical assumptions. Examination of Fig. 53(c) and 53(d) leads one to expect this result.

9.2.2 *Timing Path*

The simulated crosstalk path was used to examine the timing crosstalk performance. In this case, for various transmitted and received patterns, crosstalk loss was adjusted to produce ±30° maximum clock jitter. These values of x are recorded in Fig. 57. They are averaged, first by column, then completely. For each transmitted pattern, an "equivalent" timing frequency is recorded. This is the frequency where the cable loss is 5 db less than the crosstalk loss, and does not depart greatly from 772 kc. In some cases, the equivalent frequency falls below 772 kc, particularly due to dense transmitted pattern-sparse interfering pattern. It should be noted that odd patterns (odd numbers of pulses per word) produce harmonics of 96.5 kc, while even patterns produce harmonics of

* Bear in mind that changing the threshold voltage affects clock clipping and therefore alters the phase of the clock.

CROSSTALK PATTERN	TRANSMITTED PATTERN										
	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8	1/4	1/2	
1/8	42	27	32	<24	26	<24	<24	<24	27	<24	
2/8	34	39	27	34	<24	30	<24	27	39	34	
3/8	44	32	39	29	34	26	32	25	32	28	
4/8	36	42	31	39	29	35	27	27	40	30	
5/8	44	33	40	32	39	29	35	29	35	31	
6/8	38	41	33	39	31	38	30	29	40	35	
7/8	44	33	40	32	39	31	38	30	35	33	
8/8	35	41	33	39	32	39	31	37	35	34	
1/4	32	36	25	33	<24	26	<24	30	42	25	
1/2	33	38	29	29	26	29	25	24	31	42	GRAND AVG
AVG	38	36	34	33	31	31	29	28	35	31	33
f_e	900	820	760	740	660	660	630	600	800	660	740

LINE LOSS = 32DB AT 772 KC, 48 DB AT 1.5MC

f_e = EQUIVALENT TIMING FREQUENCY (WHERE LINE LOSS IS 5DB LESS THAN CROSSTALK LOSS BASED ON ABOVE 1.5MC FIGURES) IN KC.

n/8 MEANS AN 8-DIGIT WORD OF n SUCCESSIVE PULSES.

1/n MEANS ONE PULSE EVERY n TIME SLOTS.

Fig. 57 — Values of x that produce 30° of timing jitter.

193 kc. The result is that worst interference exists when both patterns are either odd or even.

9.3 *Pattern Shift*

R. C. Chapman of the Laboratories has studied the mechanism and result of phase jitter accumulation along a string of PCM repeaters. The result of his work (based on a simple model), which also considers the nature of the PCM terminals, indicates satisfactory terminal performance should result from PCM line lengths up to 25 miles if the repeater clock circuit Q is greater than 50, and the worst-case variation of delay

(due to pattern change) through a single repeater is less than 0.018 μsec (10° of clock shift). It is our purpose to show that these requirements have been met, and to investigate the various phenomena that may make the delay through a repeater pattern sensitive.

Fig. 58(a) shows the phase variation through eight repeaters excited by a long period of all pulses present, followed by a long period of one pulse out of eight present. It shows that the delay through a repeater varies about 0.009 μsec (5° of clock shift) as the pattern changes from the sparsest allowed to the densest possible. Measured pattern shift is shown in Fig. 58(b).

A simple model for phase accumulation (which treats the tank as a simple low-pass filter and neglects clock amplitude variation) along a string of repeaters predicts a maximum rate of change of clock phase, expressed in cycles per second, of $f_1 = \pi f_0(\theta_0/Q)$ where θ_0 is the shift per

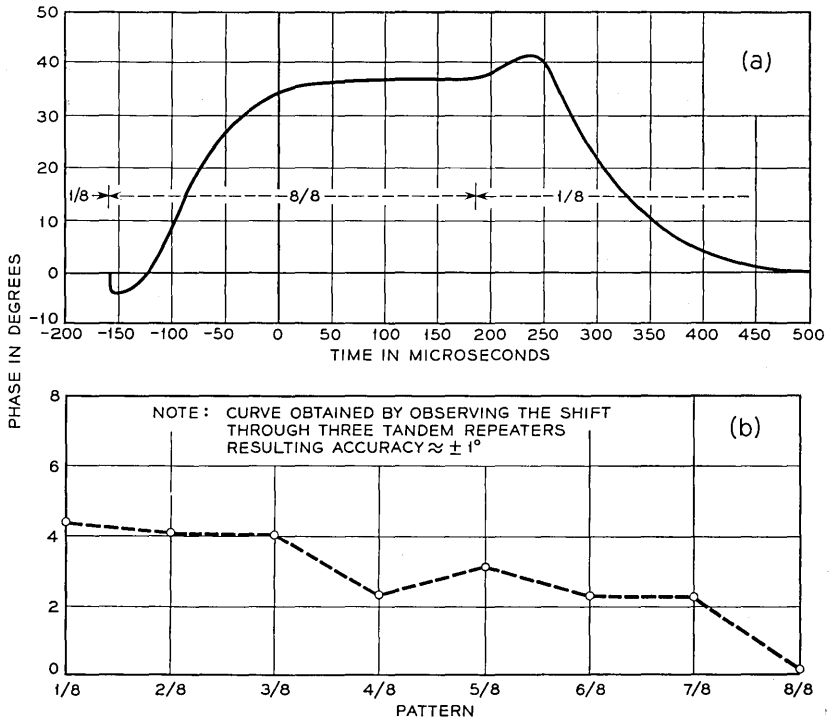


Fig. 58 — Measured clock pattern shift performance: (a) measured phase shift through eight repeaters due to abrupt change in pulse pattern; (b) measured phase shift per repeater due to change of pattern.

repeater in fraction of a time slot and Q is the quality factor of the repeater tank. From Fig. 58, f_1 is about 1.1×10^3. The corresponding average effective $Q(\theta_0 = 5/360)$ is 60. Direct measurement indicates the actual circuit Q is the upper design value ($\approx$100). There is reason to believe that the simple model, by neglecting the large clock amplitude variation with pattern, thus predicts a slower phase accumulation than actually is the case.

An attempt has been made to determine the contributors to the 5° of pattern shift per repeater. These measurements are very difficult, and the results are approximate only. There is indication that about half of the 5° comes from bottom-up clipping in the clock input stage, Q3 (see Fig. 31), and variations in the automatic clipping level with pattern. To separate the threshold and clock functions, C13 is used. An average voltage builds up across C13 , dependent on the received pulse density, producing additional clipping via CR5 and CR6. This effect can be eliminated by use of a separate threshold amplifier, or by transformer coupling of the clock into Q3. The transformer primary would be made constant resistance by use of the appropriate RC network, so the operation of the clipping diode is not affected by the primary inductance of the transformer. Both solutions add slightly to the cost of the repeater.

It appears that the remaining two to three degrees is that expected due to clock extraction. The blocking oscillator does not seem to contribute significantly to pattern jitter. With the inductive load, the blocking oscillator compensates slightly for some of the shift introduced in clock extraction.

9.4 *Temperature Performance*

Expected range of temperature variation is 32–100°F for manhole-mounted repeaters and −40 to +140°F for units mounted above ground. The design range was −40 to +140°F. The main effects of temperature are (1) effect on time crosshair position, and (2) effect on voltage crosshair position.

Fig. 59 shows the over-all effect of temperature on crosstalk performance for a repeater. This has been broken down into variation with clock position (by mistuning the tank) in Fig. 60 and variation with threshold in Fig. 61. It appears that the dominant effect is threshold variation with temperature, due primarily to the loss of dc gain in the threshold amplifier. The case shown is almost a worst case in that the threshold amplifier transistor was selected for approximately minimum

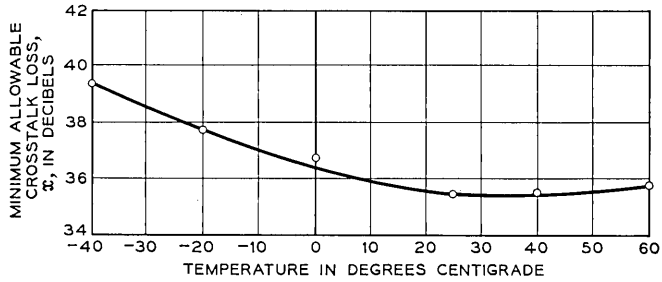


Fig. 59 — Typical measured over-all affect of temperature on crosstalk performance.

allowed β_0. Some threshold temperature compensation is possible but has not been thoroughly investigated.

9.5 *Field Experiment*

Approximately 15 repeaters were installed on two two-way lines between Summit and South Orange, New Jersey. Twenty cable pairs were leased, and crosstalk and line loss measurements were made at all re-

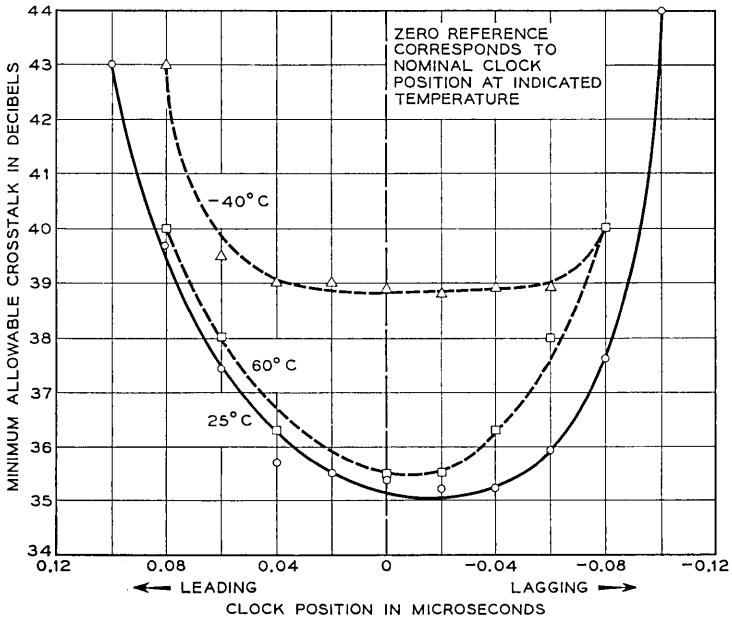


Fig. 60 — Typical crosstalk vs clock position.

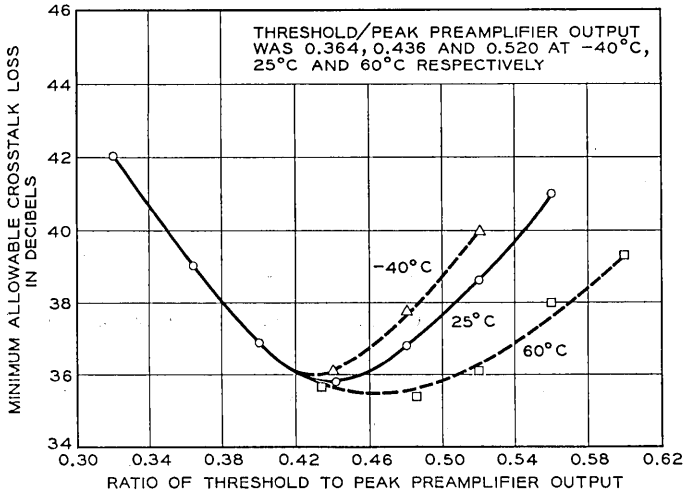


Fig. 61 — Typical crosstalk vs threshold.

peater points. One repeatered line was made up for maximum intra-system crosstalk (worse crosstalk pairs at each repeater point). The other line was of typical exposure. Placing the two lines in tandem gave one-way operation through 32 repeaters.

The resulting error rate was close to that which could be predicted from measured worst-case office impulse noise. Average error rates over a 10-minute interval were in the range 10^{-7} to 10^{-8}. About 50 per cent of the time the error rate over 10-minute intervals was less than 10^{-8}. In about 1 per cent of the 10-minute intervals the error rate exceeded 10^{-6}. These errors tended to come in short bursts and do not represent sustained error activity over the 10-minute interval.

X. CONCLUSIONS

A 1.544-mc experimental bipolar PCM repeater has been designed for one- or two-way use on unloaded exchange cable pairs. The circuit utilizes seven diffused-base transistors, and two repeaters with a common power unit are packaged in a can of $1\frac{1}{16} \times 3\frac{1}{8} \times 5\frac{3}{4}$ inches outside dimensions. The repeater includes power and line-length options, as well as provision for protection and remote testing, and is powered over the signal pair.

Timing jitter has been controlled to the extent required for 25-mile transmission, and the circuit has been optimized for near-end crosstalk performance.

Over sixty repeater circuits have been constructed, and measured performance is in reasonably good agreement with that predictable from design considerations.

Although many alternatives have been investigated for realization of each major function, one consistent design has been presented.

XI. ACKNOWLEDGMENT

The work reported herein has for the most part been carried out by various members of the Transmission Systems Development Department under the supervision of the author. At points the author has also drawn on the work of other departments. The guidance of E. E. Sumner and the prior work of F. T. Andrews are gratefully acknowledged.

REFERENCES AND BIBLIOGRAPHY

1. Aaron, M. R., this issue, p. 99.
2. Carbrey, R. L., Proc. I.R.E., **48,** p. 1546, Sept. 1960.
3. Davis, C. G., this issue, p. 1.
4. DeLange, O. E., B.S.T.J., **35,** p. 67, Jan., 1956.
5. DeLange, O. E., and Pustelnyk, M., B.S.T.J., **37,** p. 1487, Nov., 1958.
6. Mann, M., Straube, H. M., and Villars, C. P., this issue, p. 173.
7. Rowe, H. E., B.S.T.J., **37,** p. 1543, Nov. 1958.
8. Sunde, E. D., B.S.T.J., **36,** p. 891, July, 1957.
9. Thomson, W. E., Wireless Engineer, **29,** p. 256, Oct., 1952.
10. Wrathall, L. R., B.S.T.J., **35,** p. 1059. Sept. 1956.

PCM Transmission in the Exchange Plant

By M. R. AARON

(Manuscript received July 12, 1961)

Attention is focused on the choice of a code for transmitting a PCM signal in the exchange plant via paper-insulated cable pairs. Inter- and intra-system crosstalk via near-end coupling is the principal source of interference. A code and repeater structure for reconstructing this code to minimize the effect of crosstalk into the timing channel of a reconstructive repeater is emphasized. It is shown that the conventional self-timed unipolar repeater is not suited to this environment. Several pseudo-ternary codes are developed and surveyed for overcoming crosstalk interference. A bipolar code and a repeater for reconstructing this code are chosen for a variety of reasons. The repeater features nonlinear timing wave extraction and complete retiming and pulse width control, and its realization is discussed in detail in a companion paper by J. S. Mayo.[3]

I. INTRODUCTION

The choice of a modulation method or code for processing a signal preparatory to transmission over a communication channel depends on a multiplicity of conflicting parameters. All of the technical factors must be considered to arrive at a system that serves a need at a justifiable cost. Generally a few of the characteristics of the transmission medium stand out to limit performance. In the situation at hand — pulse transmission at high rates in the exchange plant — near-end crosstalk between cable pairs is the principal transmission deterrent. Examination of the factors that enter into the choice of a simple code and a repeater to reconstruct this code, to combat crosstalk interference, is the principal objective of this paper. Stated another way, we are seeking a transmission scheme that maximizes the number of pairs in a cable that can be used without pair selection.

1.1 *Road Map*

Before we attack this objective it is profitable for later comparisons to provide some system background, to review the functions involved in a PCM reconstructive repeater, and to define terminology. Though the bulk of this introductory material has been covered in the literature, we introduce some new approaches and concepts. It is essential to a clear definition of the problem and our chosen solution. Most of this material is covered in Sections 1.2 through 2.4. Some preliminary comparisons are made in Section 2.5 that contribute to the choice of complete retiming and regeneration. The crosstalk problem is brought into focus in Section III. In Section IV it is shown that the conventional pulse-absence of pulse method for transmitting binary PCM is not suited to the present application. Section V contains a survey and evaluation of various pseudo-ternary codes that minimize the problem of crosstalk into the timing channel of a repeater. Reasons for choosing a self-timing bipolar repeater are enumerated at the end of this section. Section VI is devoted to a few words on equalization and equalization optimization. The final section deals with some implications of the interaction of the bipolar repeater with the actual exchange plant.

1.2 *System Background*

1.2.1 *General*

Other papers [1,2] in this issue have shown that the signal to be transmitted over 22-gauge paper-insulated cable pairs is a 1.544-mc pulse train. Repeaters to reconstruct this pulse train are spaced throughout the medium with spacing dictated by economics, near-end crosstalk coupling and impulse noise. Since loading coils must be removed from voice-frequency pairs prior to pulse transmission, it is economically and administratively desirable simply to replace the loading coils with repeaters. The most common loading coil spacing is nominally 6000 feet, thereby dictating a repeater spacing equal to an integer multiple of 6000 feet. Twelve thousand foot spacing is precluded by near-end crosstalk considerations, as seen later. Therefore the nominal spacing between line repeaters is 6000 feet. Effective variation of electrical length about the nominal is constrained to be ± 500 feet by the use of line-build-out networks. The importance of this assumed constraint on repeater spacing cannot be overemphasized. In effect it removes repeater spacing as a system parameter to be used to combat crosstalk as far as this study is concerned. For example, with shorter repeater spacing the crosstalk problem could

be substantially reduced. The economic penalty could be countered by raising the channel capacity above 24 channels. This of course, would increase the bit rate and bring the crosstalk problem back into prominence. The choice of a transmission scheme most tolerant to crosstalk under this new situation might differ from that chosen for the 6000-ft repeater spacing.

In the vicinity of a central office another source of interference becomes important, namely impulse noise due to switching transients coupled from pair to pair via crosstalk coupling. To combat impulse noise, repeaters adjacent to an office are spaced nominally 3000 feet from the office.

It should be emphasized that with the above spacings and practical signal levels, thermal noise is not a problem.

1.2.2 *Reconstructive Repeaters*

It is well known that the salient feature of PCM transmission is the ability to reconstruct the transmitted pulse train after it has traveled through a dispersive, noisy medium. We will call such repeaters reconstructive repeaters, as opposed to the more commonly used name of regenerative repeaters, since regeneration in the circuit sense is only one facet in reconstructing pulse trains. There are basically three functions that must be performed by such a repeater, namely the three R's—*reshaping*, *retiming*, and *regeneration*.* This operational breakdown is depicted schematically in Fig. 1(a) where the pulse train is traced from repeater to repeater. For purposes of illustration, we assume that the transmitted pulse train at (b) consists of a series of pulses and spaces representing the binary 1 and 0 respectively. This is called a unipolar pulse train, and eventually we will show that this type of code is not suited to the exchange plant environment. However, at this point it suffices for our simple explanations. After transmission over 6000 feet of cable, the high-frequency content of each pulse is severely attenuated and the received pulse train is corrupted by additive interference. The spread-out, "noisy" train appears at (c) (cable output).

The primary function of the first repeater block, *reshaping*, is to shape the signal and raise its level to the point where a pulse vs no pulse decision can be made.

This process involves the inevitable compromise between interference

* *Pulse regeneration* as defined by the American Standards Association in ASA C-42, Group 65 (definition 65.02-102) includes all of the three R's. However, for our purposes the three-way split is more convenient.

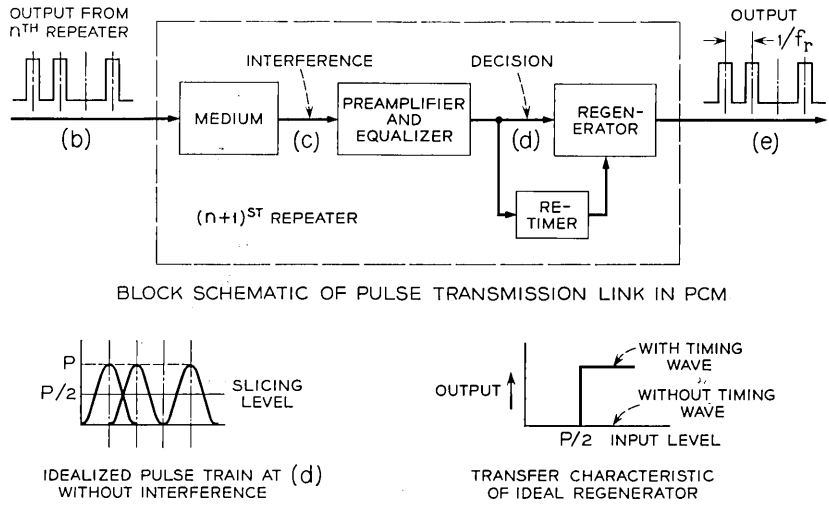


Fig. 1 — PCM transmission: (a) block schematic of pulse transmission link in PCM, (b) idealized pulse train at (d) without interference, (c) transfer characteristic of ideal regenerator.

reduction and pulse resolution.* We will examine reshaping in minor detail in this paper. A more extensive coverage is given in the paper by Mayo,[3] which goes into the details of the realization of the bipolar repeater.

Final reconstruction of the pulse train at d (point of decision) is accomplished by the simultaneous operation of *regeneration* and *retiming*. For purposes of this preliminary discussion we will assume that the regenerator takes the form of a simple threshold detector. The regenerator is enabled when the incoming signal plus interference exceeds the threshold *and* when the timing wave at the output of the retimer has the proper amplitude, or polarity. Other types of regenerators in which regeneration is only partially accomplished in a single repeater are known as partial regenerators.[4] Several partial regenerators in tandem approach complete regeneration. Partial regeneration will not be considered further since complete regeneration can be approached closely with available circuitry (i.e., blocking oscillators) at the frequencies of interest. This situation does not prevail at microwave frequencies.[4]

The purpose of the retimer is threefold: (1) to provide a signal to

* Readers schooled in the lore of digital transmission will recognize the reshaping network as that circuit which, in conjunction with the line and transmitted pulse shape, determines the "eye" picture, i.e., a snapshot of all possible pulse combinations that contribute to each time slot. See Ref. 3.

sample the pulse train where its peak is expected, i.e., where the signal-to-interference ratio is best, (2) to maintain the proper pulse spacing and reduce pulse jitter to minimize distortion in the receiving terminal, and (3) to allow the timing signal to be used to turn off the regenerator to maintain proper pulse width.

In the above manner, provided the signal-to-interference ratio is sufficiently high and the time jitter sufficiently low, the reconstructed pulse train at e is *almost* a replica of the original.

II. TIMING

Since timing plays an important role in PCM transmission, we will introduce and classify methods for *launching*, *extracting* and *using* the timing wave. In addition, we will summarize the sources of timing jitter. Finally, we will make a preliminary comparison of various timing techniques with respect to sources of jitter exclusive of crosstalk. After the initial comparison is made we will characterize the medium and concentrate on the crosstalk problem.

2.1 *Timing Generation*

Methods for launching and extracting timing information can be classified broadly according to whether or not an extra pair is used for transmitting the timing signal. These broad divisions are further subdivided in Table I.

The self-timing category in Table I has been explored in several papers. For this class, the timing wave is obtained by processing the information-bearing pulse train by either linear or nonlinear means or both. Linear extraction, as exemplified by the work of Wrathall,[5] Sunde,[6] and Bennett,[7] is used when the power spectral density of the transmitted pulse train has a discrete line spectrum with a component at the pulse repetition frequency. We will display other pulse trains in which a discrete line at submultiples of the bit rate is produced that can be used

TABLE I — TIMING CLASSIFICATION

1. Same-pair timing
   a. Timing from information-bearing pulses, generally called self timing.
      (1) Linear extraction — forward or backward
      (2) Nonlinear extraction — forward or backward
   b. Timing wave added
      (1) Without spectral null
      (2) With spectral null
   c. Hybrid or dual-mode timing
2. Separate-pair timing

for timing. In other words, the pulse train may have a periodic average value with a fundamental component at frequencies below the bit frequency. This timing component is obtained by exciting any one of several circuits or combinations thereof with the transmitted pulse train. Circuit approaches will be mentioned later.

Self timing can also be employed when the discrete line at the bit rate is absent. Bennett[7] discussed one method for achieving this, and we will consider other possibilities in a later section. Obviously, nonlinear techniques must be used here that take advantage of the underlying phase structure of the pulse train.

When the timing wave is obtained from the transmitted pulse train, it is known as forward-acting. Backward-acting timing indicates that the timing wave is obtained from the reconstructed pulse train at the repeater output and fed back to an internal point in the repeater. We will cover the relative merits of these two approaches when we discuss methods for using the timing wave.

Category (b) under same-pair timing in Table I has not been covered previously in the literature. In this approach a separate sinusoidal signal is added to the transmitted or reconstructed pulse train and extracted at the next repeater. The added timing wave may be used to augment the timing component inherent in the pulse train to insure adequate timing when the pulse train is sparse. Alternatively, the added timing wave can assume the entire timing burden. This may be achieved by any of several methods. First, the energy in the pulse train in the neighborhood of the bit rate can be eliminated by an appropriate filter prior to the addition of the timing wave. Another approach is to use a pulse transmission scheme in which the power spectral density has a null at the bit frequency. A timing wave can be inserted in the resulting slot. Spectral nulls at the bit rate may be obtained by pulse shaping or by converting the binary pulse train to a three-level code, as we shall demonstrate. With the latter method it is possible to produce spectral nulls at submultiples of the bit frequency.

The last same-pair timing category of Table I is really a transition class that leads into separate timing. Dual-mode timing involves the use of self timing in one direction of transmission and the simultaneous use of this timing wave for timing the other direction of transmission. The slave system can use a transmission scheme that has a spectral null at the timing frequency. This approach eliminates the near-end crosstalk (NEXT) interference at the timing frequency.

The final category in Table I is self-explanatory. Further comparison of all these schemes is deferred until we can bring some other factors to bear.

2.2 *Methods of Using Timing Wave*

Just as in the case of regeneration, retiming can be classified as complete or partial. In this section these and related terms, as well as some of those used previously, are given more concrete significance with the aid of the diagrams in Fig. 2.

The simplest scheme to instrument is the partial retiming approach used by Wrathall and analyzed by Sunde. In this approach, the peak of the recovered timing wave (clock) is pinned to ground or some other convenient reference and added to the incoming pulse train, as shown in Fig. 2(a). When the signal-plus-timing wave exceeds the threshold level the regenerator is fired. Obviously, as the timing wave amplitude becomes larger, sampling of the input pulse occurs closer and closer to the pulse peak where the signal-to-interference ratio is best. The timing wave

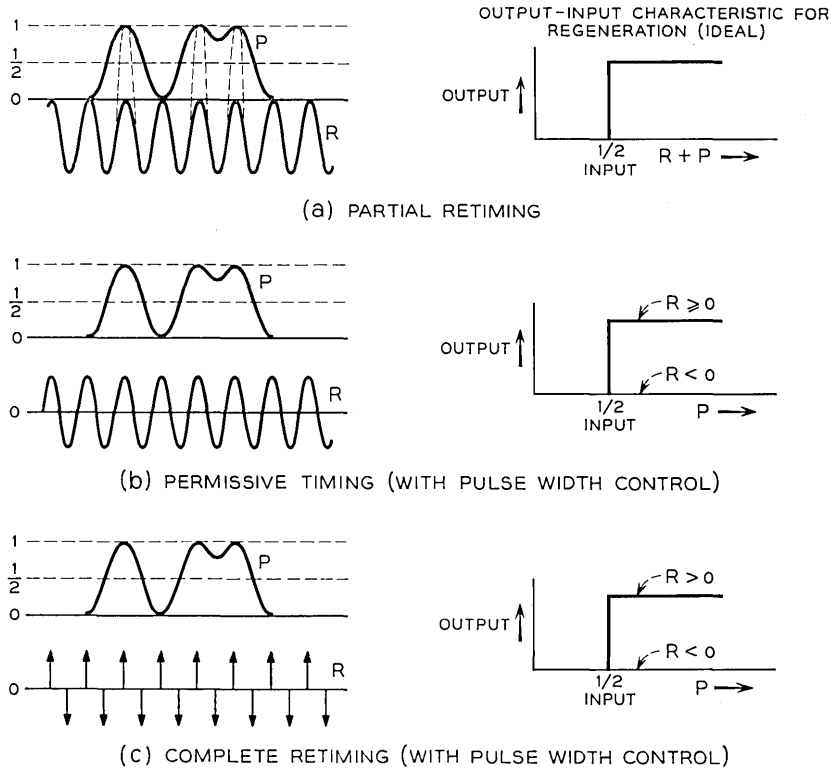


Fig. 2 — Retiming methods: (a) partial retiming, (b) permissive timing (with pulse width control), (c) complete retiming (with pulse width control).

may be obtained from either the incoming or regenerated pulse train or from a separate pair.

Another method for using the timing wave is illustrated in Fig. 2(b). We call this permissive timing with pulse width control. Both the incoming pulse train and the clock are put into an "and" gate. When the signal pulse exceeds the threshold and the clock is in its positive half cycle, standard current is fed to a regenerator (a blocking oscillator, for example) and the regeneration process is started. When the clock wave enters its negative half cycle, current is extracted from the regenerator to turn it off. In this manner the pulse width is constrained to be less than one-half cycle of the clock.

One final approach is depicted in Fig. 2(c). This approach is known as complete retiming with pulse width control. In this method, narrow pulses are generated at the positive-going or negative-going zero crossings of the timing wave and used for gating the incoming pulse train. Similarly, pulses generated at the negative-going or positive-going zero crossings are used to turn off the regenerator to control the width of the regenerated pulses. In passing we note that the narrower the sampling pulse, the smaller the fraction of time that interference is effective in obscuring a pulse decision. The importance of the width of the sampling pulse in combating crosstalk is discussed in the paper by Mayo.[3]

As discussed by Sunde[6] and Rowe,[10] complete retiming precludes timing from the reconstructed pulse train. Backward-acting timing is also taboo with permissive timing.

2.3 *Timing Wave Extractors*

In self-timed repeaters, several circuit configurations have been investigated for timing-wave extraction. Since the timing recovery problem is similar to that involved in synch recovery in TV, it is to be expected that circuit approaches investigated for that application should be pertinent here.[8] They include:

1. Tuned circuits
   a. single tuned — LC
   b. double tuned — LC and other narrow-band filters
   c. crystal filters
2. Controlled passive filter
3. Clutched (locked) oscillator
4. Phase-locked oscillator — APC loop
5. Phase- and frequency-locked oscillator — dual mode.

By far the simplest and cheapest circuit is the single tuned LC tank.

Most of the other possible implementations have been examined for this application.

Detailed discussion of the merits of each approach is beyond the scope of this paper. It suffices to point out that an automatic phase-control (APC) loop is the ideal timing wave extractor, and an APC loop for this application has been constructed. This relatively complicated circuit is available, if required, to mop-up timing deviations that propagate through a chain of repeaters with simple LC tanks. For the short repeater chains encountered in the exchange cable application, the LC tank with a loaded Q of 100 can be designed to meet jitter requirements, system economics, and space requirements. Therefore, in the remainder of the paper we will concentrate on this simple implementation.

2.4 *Sources of Timing Jitter*

In self-timed reconstructive repeaters using LC tanks for timing recovery, several sources of timing jitter and mistiming arise. They are (1) thermal and impulse noise, (2) mistuning, (3) finite pulse width, (4) amplitude-to-phase conversion in nonlinear devices, and (5) crosstalk.

2.4.1 *Thermal and Impulse Noise*

De Lange and Pustelnyk[9] have shown that thermal noise is not a serious contributor to timing jitter in the timing channel of a reconstructive repeater. Stated another way, if thermal noise is small enough such that the pulses can be recognized with low probability of error with no noise in the timing channel, then the addition of noise in the timing channel results in a negligible increase in the error probability. Physically this is to be expected since the narrow-band timing extractor accepts only a small fraction of the noise power for circuit Q's of the order of 100. This conclusion is equally valid for impulse noise for the same reason as above and also because the average rate of occurrence of impulses is considerably less than the bit rate or the minimum pulse density allowable in the system.

At this juncture it is appropriate to point out that the timing wave amplitude must exceed a certain minimum value to operate the timing gate and to limit some of the sources of mistiming to be discussed below. This presents a limitation on the minimum pulse density that can successfully be reconstructed in a repeater string for the transmission schemes of type 1a in Table I. It should be emphasized that this limita-

tion will prevail with any of the circuit implementations in a practical environment that includes timing recovery with a mistuned extractor excited by a baseband pulse train that permits spaces. Furthermore, this lower bound is a function of the statistics of the signal being transmitted, the allowable error rate, the kind of interference, the effective Q of the timing circuit, and the amplification following the timing tank. Once the timing wave has reached a sufficiently large amplitude, the higher the Q, the longer the gap between pulses that can be bridged. However, the higher the Q, the greater the phase slope of the tuned circuit and the smaller the allowable mistuning for a given signal-to-interference ratio. This inherent compromise between Q and mistuning will be discussed quantitatively below. In the exchange carrier application, it is necessary to limit the maximum number of spaces between pulses to about 15. This is achieved by eliminating the all-zeros code in the encoder. Attendant to this constraint is a small change in terminal distortion due to clipping.

2.4.2 *Mistuning*

When the tuned circuit is mistuned from the p.r.f. (pulse repetition frequency), the information-bearing pulses are sampled away from their peaks. This lowers the interference allowable for a specified error rate. As noted above, the higher the Q, the larger the shift in the zero-crossings of the timing wave at the output of the tuned circuit for a fixed mistuning. However, a high Q is desirable to bridge gaps and reduce timing jitter due to finite pulse width, as discussed below. Furthermore, the effect of mistuning is dependent upon whether forward- or backward-acting timing extraction is used, as well as whether partial or complete retiming is employed. We will shortly compare all of these techniques quantitatively.

2.4.3 *Finite Pulse Width and Pattern Effects*

As shown by Rowe,[10] when the pulses exciting the tuned circuit are not impulses or 50 per cent duty cycle rectangular pulses, the zero-crossings at the output of the tuned circuit are perturbed from their nominal positions. These deviations are dependent upon the pulse density (or pattern), the Q of the tuned circuit, and the shape of the pulses exciting the tank; and they differ for positive- and negative-going zero-crossings. This is a form of amplitude-to-phase conversion analogous to the amplitude-to-phase conversion in FM limiters.

Pattern jitter also occurs in partial retiming due to the above effects.

While backward-acting timing with ideal 50 per cent duty cycle rectangular pulses exciting the tuned circuit eliminates the finite pulse width effect, it does suffer from severe pattern jitter due to variation in timing wave amplitude with pulse density.

2.4.4 *Nonlinear Amplitude-to-Phase Conversion*

Any of the nonlinear circuits either preceding or following the tuned circuit will inevitably perturb the zero-crossings of the timing wave. In addition the regenerator, nominally a blocking oscillator, will not be an ideal zero-memory nonlinear device. Consequently, the spacing between reconstructed pulses will be altered by its inherent storage and will be dependent on the past history of the reconstructed pulse train.

2.4.5 *Crosstalk*

Intra- and inter-system crosstalk into the timing channel of a forward-acting repeater, either partially or completely retimed, results in a shift of the zero-crossings of the timing wave. We will devote the bulk of the paper to this problem.

2.5 *Preliminary Comparisons*

At this point, we will make preliminary quantitative comparisons of some of the methods of extracting and using the timing wave in a self-timed repeater. Our attention will be confined to the effects of mistuning and pattern jitter.

2.5.1 *Mistuning*

As noted previously, when the tuned circuit is mistuned from the pulse repetition frequency, tolerance to interference is reduced. To compare the various self-timing schemes quantitatively, we will make the following assumptions:

1. The information-bearing pulses are raised cosine pulses of base width, either T, $1.5T$, or $2T$* wide.

2. The output of the tuned circuit is a sinusoid whose peak-to-peak amplitude is equal to the height of the information-bearing pulse. For partial retiming, the positive peak is clamped to ground and added to the signal.

---

* In the $2T$ case, the discrete component at the bit rate disappears. Therefore, nonlinear methods must be used to create a discrete line at the bit frequency for purposes of timing.

Under these conditions, we can use the methods developed by Sunde[6] to compare backward and forward-acting partial retiming with complete retiming. In the latter case, we assume that the phase shift in the zero crossings of the timing wave is given by its average value, approximately $2Q(\Delta f/f)$; $\Delta f/f$ is the fractional mistuning of the tuned circuit. Fig. 3 shows the reduction in allowable interference in the presence of mistuning for the three self-timing methods of interest. As expected, forward-acting partial retiming and complete retiming are far superior to backward-acting partial retiming. This is particularly noticeable for the case where the pulse width is from 1.5 to 2 time slots wide at its base. It will be shown that a pulse shape in this region is required to minimize the effects of crosstalk interference in making the pulse-no pulse decision. Based on the fact that there are several sources of impairment that must share the allowable margin against error, that portion of the margin allocated to mistuning must be as small as possible, consistent with presently available components. In the present state of the art, including initial misplacement and aging, it appears that the maximum phase shift in the tuned circuit (LC tank) can be held to about $\pm 30°$ for a Q of about 100. This corresponds to $2Q(\Delta f/f) \doteq 0.6$, or $\Delta f/f = 0.003$. From Fig. 3 and the fact that equalization to minimize crosstalk interference must yield a pulse close to the 1.5 to $2T$ cases, it can be seen that backward-acting timing should not be used for this application.

It should be pointed out that there is nothing restrictive about the use of raised cosine pulses in making this point. It can be shown that the conclusion arrived at above remains valid for other similar pulse shapes, time limited or not, and for other practical ratios of timing wave amplitude to pulse peak. Indeed, for larger ratios of timing wave amplitude to pulse peak, the stability problem is further aggravated. This is due to the positive feedback nature of the timing loop in backward-acting timing. A straightforward extension of Sunde's work will verify this contention. In addition, experimental work by A. C. Norwine* confirms this expected behavior.

2.5.2 *Finite Pulse Width and Pattern Effects*

In this category, we will compare partial retiming (forward-acting) with complete retiming for both periodic and random pulse patterns. With periodic patterns, both raised cosine and Gaussian pulses will be considered. Several ratios of average timing wave amplitude to peak

* Private communication.

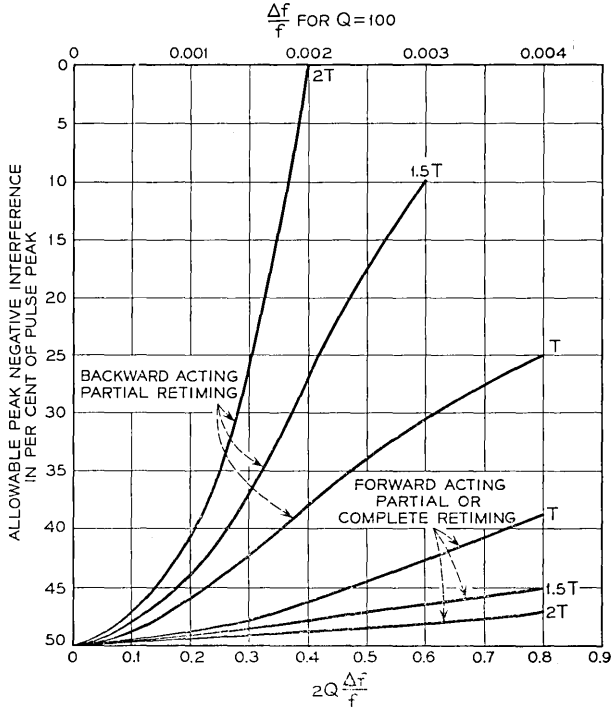


Fig. 3 — Effects of mistuning.

pulse height will be considered with partial retiming. We continue to consider binary (1,0) pulse trains.

2.5.2.1 *Periodic Pulse Patterns — Partial Retiming.* To divorce pattern effects from mistuning, we assume that the peak of the timing wave occurs at the pulse peak. Further, the positive peak of the timing wave is pinned to ground and the timing wave amplitude is given by

$$-\frac{an}{2M}\left(1 - \cos\frac{2\pi t}{T}\right) \tag{1}$$

where

a is the peak-to-peak amplitude of the wave when all pulses are present; i.e., $n = 1 = M$, and

n is the number of pulses that occur in an M-bit word.

When $n \neq 1$, we assume that the pulses are adjacent, and that the time slot examined is the one containing the last pulse in the group. In the following, we will specialize to $n = 1$, and M will vary from 1 to 8.

The assumptions underlying (1) are, first, that the pulses exciting the tuned circuit are very narrow pulses obtained by processing the incoming signal, and second, that the timing wave amplitude in a real repeater is dependent upon the density of pulses exciting the tuned circuit. While the first assumption does not correspond to the practical case, it does give an optimistic view for partial retiming. We will find that even with this idealization, partial retiming is considerably inferior to complete retiming in maintaining the proper pulse spacing under both steady-state and random conditions. It is possible to drop the first assumption here, as we do for complete retiming. The additional analysis simply lends further emphasis to the conclusion.

Two pulse shapes will be considered. First, the raised cosine with

$$P(t) = \frac{1}{2}\left(1 + \cos\frac{2\pi st}{T}\right) \quad \text{where} \quad |t| < \frac{T}{2s} \qquad (2)$$
$$= 0 \qquad \text{elsewhere.}$$

The pulse width is T/s. We consider $s = 1, \frac{2}{3}$. Secondly,

$$P(t) = e^{-(\pi^2/\ln 2)(f_6 t)^2}. \qquad (3)$$

In (3), f_6 is the frequency at which the transform of the Gaussian pulse is 6 db down from its low-frequency asymptote. Alternatively, we can write the Gaussian pulse as

$$P(t) = e^{-4\ln 10(t/T_w)^2}. \qquad (4)$$

In (4), T_w is the pulse width between points where the pulse amplitude is 0.1 of its peak.

In partial retiming, the incoming pulse is regenerated at the instant when the sum of the signal plus the timing wave exceeds the slicing level, which is assumed to be at $\frac{1}{2}$. The resulting problem can be solved graphically, iteratively, or, for some cases, analytically. Fig. 4 displays the results of these computations for raised cosine pulses.* The ordinate gives the steady-state phase shift in degrees, measured from the pulse peak, as a function of the pulse pattern. In all cases we have subtracted out the phase shift corresponding to all pulses present. The important point is the fact that the difference in phase shift between all pulses present and $\frac{1}{8}$ present is quite large: about 60° for the widest pulse and 50° for a pulse width of $1.5T$. Fig. 5 presents the same information for Gaussian pulses, and the results are not substantially different from those obtained for raised cosine pulses. It is worth noting that inter-

* Similar computations have been made in an unpublished memorandum by W. M. Goodall and O. E. De Lange.

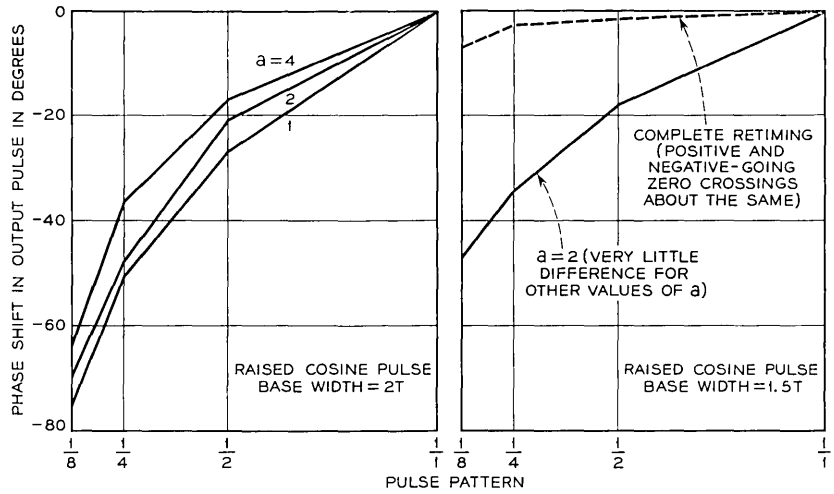


Fig. 4 — Pattern effects — raised cosine pulse: (a) base width = $2T$, (b) base width = $1.5T$.

ference riding on the leading edge of the pulse will result in additional timing jitter.

2.5.2.2 *Periodic Pulse Patterns — Complete Retiming.* In complete retiming, timing jitter results from the finite width of the pulses exciting the tuned circuit and the variation of pulse density. Using a minor mod-

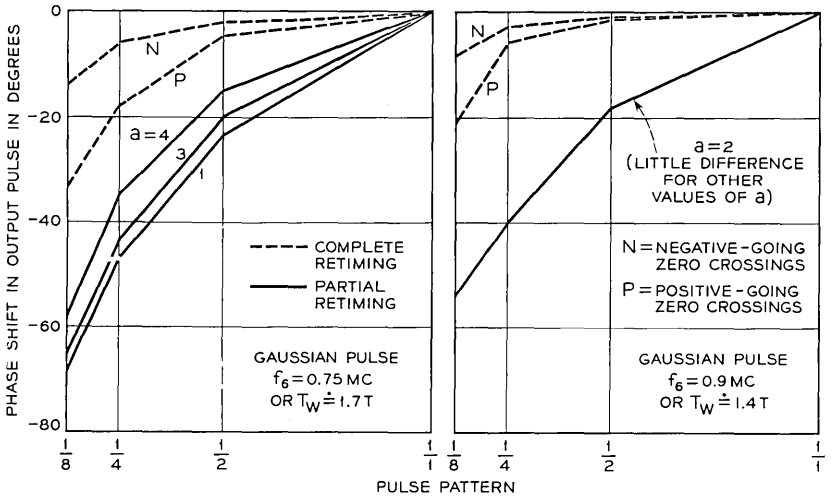


Fig. 5 — Pattern effects — Gaussian pulse: (a) $f_6 = 0.75$ mc, (b) $f_6 = 0.90$ mc.

ification of Rowe's method, we have displayed the phase shift in the zero crossings at the output of the tuned circuit in Figs. 4 and 5. A Q of 100 is assumed, and both positive- and negative-going zero-crossings are considered. It can be seen that the phase shift in this case is considerably smaller than in the case of partial retiming. Since these figures display static performance, we cannot immediately conclude that the resulting distortion in the reconstructed analog signal will be either tolerable or intolerable. Distortion at the output of the demultiplex filter in the terminal due to phase modulation on the PAM samples is dependent upon both the amplitude of the jitter and its rate of change. These quantities in turn depend upon the bandwidth (Q) of the timing extractor used in each repeater. Without going into detail, based on the dynamic effects of timing jitter propagation due to pattern shift in a string of repeaters, it can be concluded that partial retiming will be unsatisfactory for this application. Furthermore, the additional circuitry required in complete retiming over that needed for partial retiming (forward) is relatively simple and cheap. Finally, and most important, complete retiming makes pulse width control relatively easy, and the use of narrow sampling pulses yields a substantial crosstalk advantage.[3]

From Fig. 3, we see that wide pulses are desirable to minimize the effects of mistuning, while Fig. 5 shows that narrower pulses are desirable for minimizing jitter in the zero-crossings at the output of the timing wave extractor. Furthermore the positive-going zero-crossings are perturbed to a larger extent than negative-going zero-crossings for sparse patterns. This suggests separate equalization in the timing path to narrow the pulses prior to excitation of the tuned circuit. Slicing in the timing path is useful in this regard and also serves to eliminate low-level interference in the absence of a pulse. This will be emphasized later in connection with crosstalk.

2.5.2.3 *Random Pulse Patterns.* Another view of the dynamic effects of pattern jitter may be obtained by determining the probability distribution of phase jitter for random pulse patterns. This approach will be published in another paper[17] where we will include both mistuning and finite pulse width effects. This information can be used to further cement our choice of complete retiming for this application.

2.6 *Interim Summary*

At this point a short summary of our preliminary conclusions is in order. Based primarily on mistuning and pattern effects, with a peek ahead at crosstalk considerations, it is concluded that complete retiming with pulse width control should be used for a self-timed repeater.

This follows from the fact that our objective is to leave most of the margin against error to crosstalk interference in order to avoid pair selection. For this same reason, and the fact that it is readily approached, complete regeneration should be employed.

It should be apparent that pattern effects ideally can be eliminated with timing wave added or by transmitting the timing wave on a separate pair. There are important economic and technical reasons why these approaches are undesirable and we will cover them in Section V.

III. NATURE OF NEAR-END CROSSTALK

3.1 *General*

With the above preliminaries largely disposed of, we can concentrate on the crosstalk problem. This problem arises when we consider both directions of transmission for a single 24-channel system and is further compounded when many 24-channel systems are transmitted on separate pairs in the same cable bundle. In effect, such a system is a combination of time division and space division, and a typical repeater-to-repeater link can be depicted in block diagram form as shown in Fig. 6.

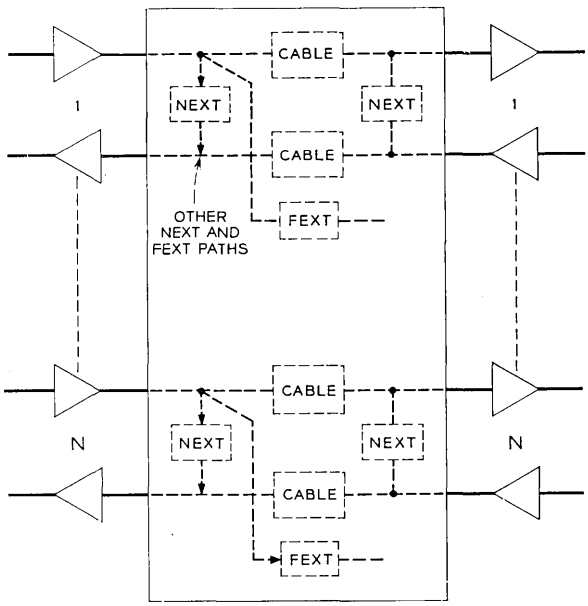


Fig. 6 — "Channel" for N PCM systems.

We could write a matrix relationship between the transforms of the voltages at the $2N$ output terminals and the $2N$ input terminals. The elements of the matrix would contain the transfer function of the direct cable paths on the main diagonal, and the transfer functions of the near-end and far-end crosstalk paths would reside in the off-diagonal terms. Such a representation, though elegant, would be of little use without a detailed characterization of the elements of the matrix (channel). At the present state of our knowledge, the above viewpoint is not too useful.

The most important couplings fall in the category of near-end crosstalk coupling (NEXT) between pairs operating in different directions of transmission. A comprehensive treatment of the physical mechanism from which crosstalk arises will not be attempted. A useful image of the situation results from considering the $2N$ pairs (for N systems) as $2N$ lossy coupled, tapped delay lines in which the "taps" (inductive and capacitive coupling due to unbalances) have both random amplitude and spacing. For our purposes, the macroscopic view obtained from looking at the $4N$ ports of this network will suffice. Furthermore, it will be convenient to make a further abstraction from reality and define an "equivalent crosstalker" and an "equivalent crosstalk path" to replace the complicated statistical model given above. This approach will permit us to attach a "crosstalk figure of merit" to each transmission scheme, which will serve as a valuable aid in sorting out those approaches most tolerant to near-end crosstalk (NEXT).

3.2 *NEXT Measurements*

3.2.1 *Single-Frequency Distributions*

Over the past twenty years or so, hundreds of thousands of single-frequency measurements have been made of crosstalk between pairs in trunk cables. Relatively few of these measurements have been made in the 100-kc to 10-mc region. A small number of measurements in this frequency range have been made at Bell Laboratories by B. Smith on a unit-constructed cable. We have used these preliminary data as a guidepost to choose a transmission scheme.* The distribution of these measurements at 1.5 mc (close to the bit rate of interest) is shown in Fig. 7. Two cases of interest are shown. The worst case prevails when the interfering pair is in the same unit as the pair into which it is crosstalking. When the two pairs are in adjacent units, crosstalk coupling is reduced.

* It is to be emphasized that these are preliminary data and are included to convey a feel for the magnitude of the problem; they do not necessarily typify the average cable plant.

The most favorable situation (not shown) occurs where opposite directions of transmission are located in diametrically opposed units of the cable. Where unit integrity is maintained throughout the length of the cable, it is obviously prudent to take advantage of the lowest crosstalk coupling in installation. However, unit integrity is not universally adhered to. Therefore, for purposes of discussion we will use the data on Fig. 7 corresponding to the worst case.

It should be noted that, as shown, the distribution of pair-to-pair crosstalk is distributed according to a log normal distribution. There is some theoretical justification for this shape over most of the crosstalk loss range. It is to be expected from physical considerations that the distribution should truncate in the neighborhood of the tails. Indeed, R. J. Herman in work at Bell Laboratories has shown that the log normal hypothesis is not supported by the data in this region.

3.2.2 *Loss vs Frequency*

Single-frequency crosstalk loss measurements are most useful for comparing some transmission schemes with respect to crosstalk in the

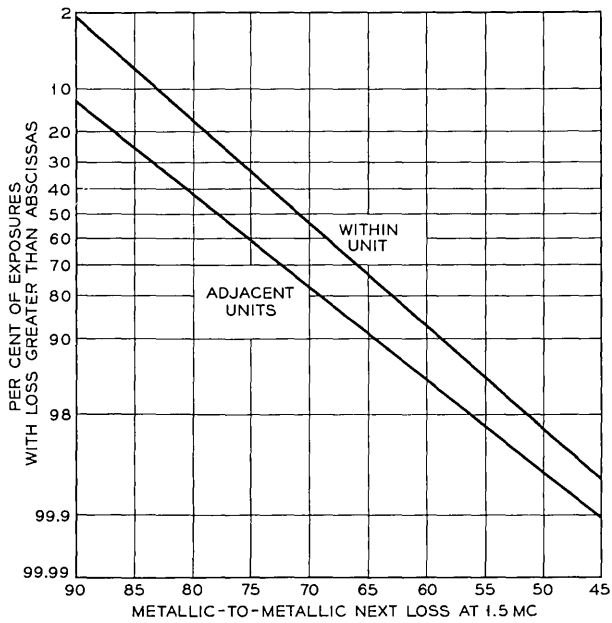


Fig. 7 — Preliminary near-end crosstalk loss distribution.

timing channel. However, these data are insufficient to define the effect of crosstalk interference on the desired information-bearing signals. An appraisal of the effect of crosstalk on the desired pulse train requires the determination of the loss and phase of the crosstalk path or, alternatively, the pulse response of this path. Since the interfering signal is made up of a combination of random pulse trains, and since the paths they traverse to the system being interfered with also come from a random family, a complete statistical description of the situation is extremely complicated and nonexistent. Measurements of the NEXT loss between pairs, as a function of frequency, display a broad structure in which the loss decreases with increasing frequency at about 4.5 db per octave (coupling proportional to $f^{3/4}$) for frequencies greater than about 200 kc. This is in good agreement with some unpublished theoretical work of D. K. Gannett of Bell Laboratories. In addition to this nominal smooth behavior as a function of frequency, measurements have revealed other more rapid variations with frequency. This fine structure may be attributed to relatively large localized capacitive and inductive unbalances. These results are in accord with the model specified previously as a visual aid and give a qualitative explanation of the classes of pulse responses observed with pulse excitation of the crosstalk path.

3.2.3 *Simplification*

In our comparison of various transmission schemes, we will find it convenient to neglect the fine structure associated with the crosstalk path. We take this bold step with the obvious realization that we preclude a completely definitive evaluation. It was mandatory to make this abstraction from reality early in the system development when detailed crosstalk data were not available and a development decision had to be made. The principal advantage of this simplification is that it permits us to isolate the classes of codes most tolerant to NEXT interference in the timing channel of the disturbed repeater.

It was convenient and expedient to go one step further and replace the multiplicity of crosstalk paths and interferers by a single capacitive coupling and a single interferer for purposes of repeater design and experimentation. An equivalence between this gross simplification and the real world can be determined empirically. Mapping of results obtained in the simple domain onto the real world by analytical means is an extremely difficult chore. We will have a little more to say about this correlation of models in a later section.

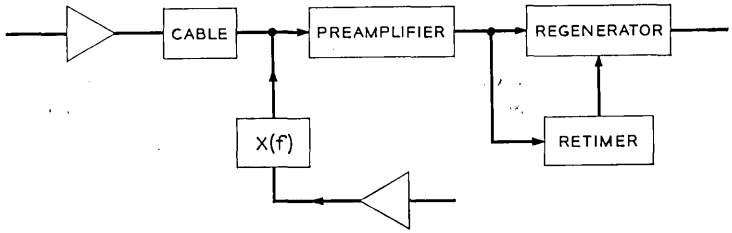


Fig. 8 — "Equivalent" crosstalk situation.

IV. CROSSTALK INTO THE TIMING CHANNEL

4.1 *Assumptions*

With the simple "smooth" crosstalk model we can begin to be more quantitative in our definition of the crosstalk problem, particularly with respect to timing. For purposes of definition, and later comparisons, we will consider the situation depicted in Fig. 8. The following assumptions are appropriate:

1. Nominal repeater spacing is 6000 feet on 22-gauge cable, and "within-unit" distribution of metallic-to-metallic NEXT is applicable.

The line loss corresponding to this length at 55°F is shown on Fig. 9.

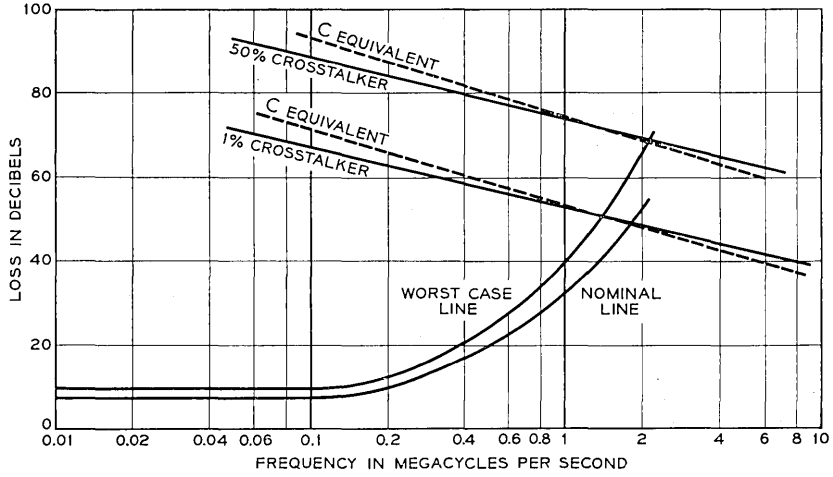


Fig. 9 — Line loss and crosstalk loss.

In addition, the line loss under extreme temperature conditions (100°F), under extreme manufacturing deviations (3σ limit), and line length of 6500 feet is also depicted. The mean crosstalk and the 1 per cent crosstalk loss are also shown using both 6 db and 4.5 db per octave extrapolations from the mean and 1 per cent values of Fig. 7.

2. A disparity of pulse densities of 7/1 on disturbing and disturbed pairs is quite likely.

This assumption is of course intimately tied up with the code pattern transmitted by the terminals and can arise when there are several idle channels in the disturbing system and the coder is misbiased. Under conditions of misbiasing, the coder may put out level 63 for the idle circuit or no speech condition instead of 64. Clearly this also depends upon noise riding on the reference. There are several techniques for minimizing the effects of this disparity in pulse densities on repeater timing, and they will be considered anon. An 8/1 disparity in pulse densities can occur for speech transmission but with a low probability. For transmission of high speed data over T-1 lines, the disparity in pulse densities can be 8/1 with significant probability.

3. Finally, we will determine the allowable crosstalk at the pulse repetition frequency such that the desired timing frequency component at the input to the tuned circuit is 6 db stronger than the timing frequency component that traverses the crosstalk path. This corresponds to a maximum of 30° phase shift in the timing signal relative to the position of the desired pulse peak. It turns out that this amount of jitter is tolerable as far as pattern jitter effects on the reconstructed PAM signal are concerned for the bandwidths ($Q = 100$) of the tuned circuits used in the repeaters.[3] We will not be concerned with these dynamic effects in any detail. Our picture will, by the large, be a static one, and the bandwidth of the tuned circuit will not enter. The allowable crosstalk loss as determined above is defined as the "crosstalk figure of merit" for a transmission scheme. The smaller this crosstalk figure, the larger the number of pairs that can be used for PCM transmission.

From the data on Fig. 9, we can anticipate the difficulty associated with transmitting timing information down the repeatered line at the bit frequency. For example, if we assume the worst case line loss and equal transmitted timing components on both disturbing and disturbed pairs, then $53 + 6 = 59$ db is the required crosstalk loss to meet the 30° phase shift requirement. With this crosstalk figure of merit, it is apparent from the distribution of Fig. 7 that pair selection cannot be ruled out even with a single interferer and no adverse disparity in pulse densities.

It should be quite evident from an extrapolation of Fig. 9 that 12,000-foot spacing, twice the normal load-coil spacing, is an unworkable situation.

4.2 *Pair Selection*

Now that we have raised the specter of pair selection, it will be instructive to bring the magnitude and scope of the associated measurement problem to the fore. Fig. 6 gives a picture of the crosstalk situation at the extremities of a repeater-to-repeater link. For N PCM systems in the cable, there are N^2 crosstalk exposures at each end of the link. In an L-mile route containing N PCM systems, approximately $2N^2L$ exposures exist. As an example, consider the installation of 20 PCM systems, each 25 miles long. Approximately 20,000 exposures exist, and the measurement, characterization and exclusion of pairs is an enormous and expensive task. To avoid this problem we must choose a transmission method with a high degree of tolerance to intra- and inter-system crosstalk. More precisely, a transmission scheme sufficiently insensitive to near-end crosstalk interference must be chosen such that the probability of failure of a single repeater (due to this cause), out of the approximately $2NL$ installed, is small. This calls for a low crosstalk figure of merit.

4.3 *Unipolar Pulse Train*

The simplest and most common form for transmitting binary PCM is to represent a one by a pulse and a zero by the absence of a pulse. Under these conditions, with pulses and spaces uncorrelated, it is well known that the power spectral density of the pulse train is

$$P_1(f) = \frac{|\, G(f)\,|^2 p(1-p)}{T} + |\, G(f)\,|^2 \frac{p^2}{T^2} \sum_{n=-\infty}^{\infty} \delta(f - nf_r). \quad (5)$$

In (5), p and $(1 - p)$ are the probabilities of pulse and no-pulse respectively, $G(f)$ is the Fourier transform of the pulse shape, and T is the reciprocal of the pulse repetition frequency f_r. The presence of a discrete line in the spectrum at the bit rate (assuming $|\, G(f_r)\,| \neq 0$) permits the extraction of timing information from the pulse train by means of a simple tuned circuit. With a random pulse train on both disturbing and disturbed pairs, the crosstalk figure of merit for worst case line loss is 59 db as before. Even with nominal line loss the figure of merit is 49 db. In either case, pair selection cannot be excluded for only one system. Unfavorable disparity in pulse densities of 7/1 raises the worst case figure

of merit by 17 db to 76 db. From Fig. 7 we note that only 40 per cent of the exposures have crosstalk loss exceeding 76 db. This shows clearly that this simple transmission scheme is not suited to the exchange plant environment. Obviously, the direction to proceed is to consider transmission schemes in which the timing information is transmitted at a frequency below the bit rate where the line loss is reduced and the crosstalk loss is increased. In the succeeding sections we will present some methods for achieving this end.

V. OTHER PULSE TRANSMISSION SCHEMES

5.1 *General*

All of the transmission schemes that we will consider have a common thread. They involve a conversion of the binary train to a three-level code (pseudo-ternary) for transmission over the line. Conversions are accomplished on a bit-by-bit basis. Translation from the binary code to a three-level code by operating on multiple bits or binary words will not be discussed. In addition, hybrid combinations (series or parallel or both) of the bit-by-bit converters are possible. However, the latter two approaches have been ruled out for this application, either on the basis of economics or the fact that they do not afford substantial technical advantages over the simpler methods. It should be understood, however, that the provision of a three-level reconstructive repeater permits the utlization of some of the more complex code translators at the terminal should unforeseen conditions dictate their choice.

Several of the pseudo-ternary pulse trains will be described prior to placing them in the crosstalk environment. In addition, other factors will be brought to bear in the comparison of the various three-level codes. Consideration of these other factors involves judgments of the degree of difficulty and the costs associated with the realization of the various approaches.

One of these additional factors involves low-frequency suppression. Transformer coupling is required to couple an unbalanced repeater to the balanced line and to provide a phantom path for remotely powering the repeaters. Transmission of a pulse through transformers results in a long transient undershoot that extends over several time slots and interferes with subsequent pulses in the train. This of course reduces tolerance to crosstalk. There are a host of techniques that have been used to combat this low-frequency wander. A summary of most of these methods is given in Ref. 11, and we do not intend to review them in this paper. We

have found that many of the circuit approaches for minimizing or theoretically eliminating the effects of low-frequency suppression that appear ideal on paper have been found wanting when actually implemented and placed in a real world environment. Pulse trains whose spectra contain discrete lines at dc suffer most from low-frequency suppression. This is another reason for discarding unipolar.

A second factor of interest involves compatability of PCM systems with AM carrier systems using pairs in the same cable. Compatability is a function of the number of PCM systems involved, the crosstalk loss in the frequency region occupied by the AM system, and other details and layout of the particular AM system. An indication of the *relative* compatability of the various PCM transmission approaches due to interference from PCM into the AM system may be obtained from a comparison of their power spectra. Since crosstalk loss decreases with increasing frequency, the AM system that extends to the highest frequency will be most affected. N carrier fits this picture. Therefore, we will use a frequency close to the top of the N carrier band as a bench mark for comparison of the relevant power spectra. For convenience we choose $f_r/6 \doteq 257$ kc as the representative frequency. Since N carrier employs frequency frogging, other higher frequencies are also of importance at a high-low N repeater. In particular, frequencies up to 440 kc are of importance. This is one reason why timing at $f_r/4$ does not seem attractive.

5.2 *Time Polarity Control* (TPC)

The first transmission scheme we consider is called time polarity control. It was suggested and demonstrated by L. C. Thomas. In this approach, time slots are labeled alternately positive and negative. If a unipolar pulse occurs in a positively labeled time slot, it is transmitted unaltered. On the other hand, if a unipolar pulse occurs in a time slot with a negative label, the pulse is transmitted with negative polarity. Zeros in the unipolar train are unaffected. The block diagram of a circuit for achieving this end is shown in Fig. 10, along with an example of an idealized pulse train before and after conversion. Intuitively, discrete lines in the power spectral density of the TPC train are expected at multiples of half the bit rate. This inherent periodicity may be demonstrated by computing the ensemble average of the TPC train. To do this, we assume that the original binary pulse train has independent pulses and spaces that occur with probability p and $1 - p$. Furthermore, we assume that the positive and negative pulse shapes in the TPC train are given by $g_1(t)$ and $-g_2(t)$, respectively. The latter assumption is introduced to

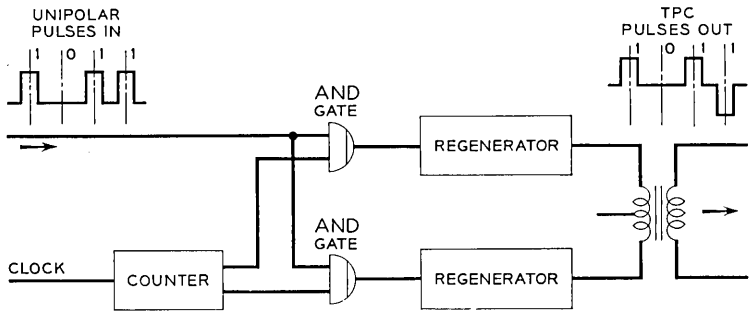


Fig. 10 — TPC converter.

account for practical differences in the two circuits that generate the positive and negative pulses. Under these conditions, the ensemble average of the pulse train is

$$\text{av } x(t) = \frac{p}{2} \sum_{n=-\infty}^{\infty} (g_1(t - 2nT) - g_2(t - (2n + 1)T)). \tag{6}$$

That (6) is periodic with period $2T$ can be seen from

$$\begin{aligned} \text{av } x(t + 2T) &= \frac{p}{2} \sum_{n=-\infty}^{\infty} (g_1(t - (2n - 2)T) - g_2(t - (2n - 1)T)) \\ &= \text{av } x(t) \end{aligned} \tag{7}$$

by making a change in the summation index from $n - 1$ to m.

The power spectral density for TPC may be computed by any one of several approaches† to give

$$P(f)_c = \frac{p(1 - p)}{2T} [| G_1 |^2 + | G_2 |^2] \tag{8}$$

for the continuous portion of the spectrum. The discrete spectrum is

$$P(f)_d = \frac{p^2}{4T^2} [| G_1 |^2 + | G_2 |^2 - 2ReG_1G_2{}^*e^{-j2\pi fT}] \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{nf_r}{2}\right) \tag{9}$$

In the ideal case when $G_1 = G_2$

$$P_2(f) = \frac{p(1 - p)}{T} | G_1 |^2 + \frac{p^2}{T^2} | G_1 |^2 \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{(2n - 1)f_r}{2}\right). \tag{10}$$

From (10) it can be seen that the continuous part of the spectrum is

† Signal flow graph techniques[12,13] were used in deriving the expressions for the power spectral densities.

identical with that of unipolar assuming the same pulse shape in both cases. In ideal TPC discrete lines occur at odd integer multiples of one-half the bit rate. This permits timing at half the bit rate. Reconstruction of the original binary train at the receiver is achieved simply by rectification of the TPC train consisting of non-overlapping pulses. Half bit rate timing in the presence of a random crosstalk interferer with a crosstalk loss of 39 db at half the bit rate yields a crosstalk figure of merit of $39 - 4.5 = 34.5$ db for worst case line loss.

It is advisable for other comparisons to catalogue certain features of the power spectrum of TPC. Unlike unipolar, TPC with balanced pulses has no discrete line at dc. This eases but does not eliminate the low-frequency suppression problem. It still is necessary to consider long runs of alternating positive (or negative) pulses and spaces. Lack of dc transmission will increase the susceptibility of the system to errors due to crosstalk. DC restoration operating on both the positive and negative peaks can be employed to reduce this effect. This is a relatively difficult circuit problem and results in an increased repeater cost over other methods to be considered.

There are several modifications of the basic TPC converter that can be made to achieve other codes. Increasing the number of stages in the counter produces higher-order TPC. For an M-stage counter; M time slots are labeled positive, the next M negative, and the cycle is repeated. If the positive and negative output pulses are identical (except of course for sign), the continuous spectrum of TPC-M, is identical with unipolar. Discrete lines appear at odd integer multiples of $f_r/2M$. For $M > 1$, the fundamental occurs at frequencies occupied by AM carrier systems and precludes compatability between PCM and the AM system. Furthermore, more instrumentation is required in the timing path of a repeater to process the signal preparatory to performing complete retiming with pulse width control.

5.3 *Bipolar*

Another and more useful modification of the converter of Fig. 10 results when the clock input to the counter stage is replaced by the incoming unipolar train, as in Fig. 11. The output pseudo-ternary code is thereby constrained such that two successive pulses, whenever they occur, must be of opposite sign. We assume the same conditions on the unipolar train used previously. Further, it is assumed that the transforms of the positive and negative pulses are G and $[-(1 + a)G + G_1]$ respectively. This brings out the differences in the two shapes more explicitly.

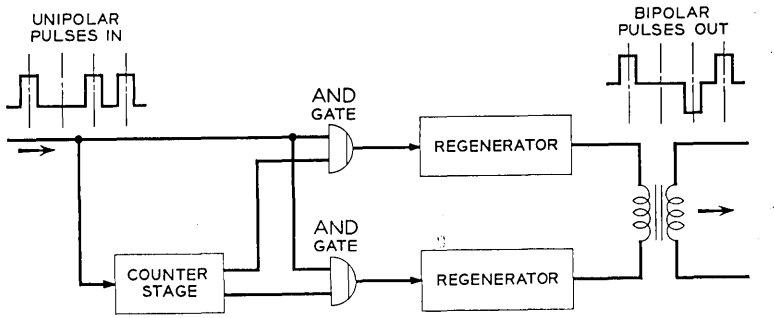


Fig. 11 — Bipolar converter.

In the ideal case (balanced pulses) $a = 0 = G_1$. With the general assumptions, the resulting expression for the continuous part of the power spectrum is long and will be covered elsewhere. The discrete spectrum is given in general by:

$$P(f)_d = \frac{p^2}{4T^2} | -aG + G_1 |^2 \sum_{n=-\infty}^{\infty} \delta(f - nf_r). \tag{11}$$

This obviously disappears under ideal conditions. Subject to this idealization, the spectrum contains only the continuous component given by:

$$P_3(f) = \frac{p(1-p)}{T} | G |^2 2\left[\frac{1 - \cos \omega T}{1 + 2(2p-1)\cos \omega T + (2p-1)^2}\right]. \tag{12}$$

For the same pulse shapes in unipolar and bipolar

$$P_3(f) = 2\left[\frac{1 - \cos \omega T}{1 + 2(2p-1)\cos \omega T + (2p-1)^2}\right] P_1(f)_c. \tag{13}$$

The subscript on P_1 indicates the continuous component of (5). With 50 per cent duty cycle rectangular pulses, the spectrum is shown in Fig. 12 (labeled $N = 1$) for $p = 1/2$. From this figure or (13) it is apparent that the spectrum has nulls at integer multiples of the bit rate (as well as nulls where $| G | = 0$). Absence of power at dc and the reduction of the spectrum in the neighborhood thereof eases requirements on the low-frequency performance of the transformers. This also follows from the fact that no two successive pulses, independent of their spacing, can be of the same sign.

In making this conversion we have suppressed all discrete components that might be useful for timing. As shown in the Appendix, rectification of the bipolar train produces discrete lines in the spectrum, and the com-

ponent at the bit rate can be used for purposes of timing.* The question arises as to where the energy for timing comes from. The analysis given in the Appendix (Section A.1) shows that the timing component arises *largely* from the region where the bipolar power spectrum is most concentrated. This is in the neighborhood of half the bit rate, and we will use this frequency to characterize the "timing frequency" for this method of transmission.† This result is intuitively satisfying since the rectifier acts as a frequency doubler. A square law rectifier has been assumed in the derivation given in the Appendix for several reasons. First, the square law device serves to reduce low-level interference in the absence of pulses. In this way, the adverse pulse density effects are reduced. Secondly, the square law assumption is a good approximation to a combination of slicer and rectifier that is actually used in the physical embodiment of this scheme. Finally, the square law characteristic is somewhat more convenient for analysis than the actual symmetrically biased rectifier.

The reader familiar with the literature will recognize that in the case $p = 1/2$, the spectrum of bipolar is identical with Meacham's twinned binary.[14] This follows from the fact that in Meacham's approach, twinned binary is generated by taking the unipolar train, delaying it by one time slot, and subtracting it from the original. This modifies the spectrum of unipolar by $| 1 - e^{-j\omega T} |^2 = 2(1 - \cos \omega T)$, which is identical to the modifying factor of (13) for $p = 1/2$. Despite the fact that the spectra are identical for equally likely unipolar pulses and spaces, the relationships between the original unipolar code and the transmitted code differ for the two methods. If the probability of error (both insertion and deletion) for the unipolar train is P_e for interference that is equally likely to be positive and negative, then it can be shown that the error probabilities for bipolar and twinned binary are $(3/2)P_e$ and $2P_e$ respectively. The above figures assumed that the repeaters to reconstruct either train do not have the bipolar constraint built in. In addition, due to the manner in which the twinned binary is converted to binary at the receiver, double errors per word are much more likely than in bipolar. For purposes of classification, bipolar belongs to the class of codes in which the conversion from unipolar is digital and the reconversion is analogue. In twinned binary the reverse is true.

There are two points of departure from the bipolar converter of Fig.

* A timing component could be added at the spectral null at the bit rate. This would not help the crosstalk timing problem except for the adverse pulse density effect.

† Obviously energy for timing does not come from a single frequency. However the characterization is convenient. Further clarification of this point is given in the Appendix and Section VI.

11. Increasing the number of counter stages to N results in a code in which N successive unipolar pulses are transmitted positive and the next N, wherever they occur, are transmitted negative. Spaces are unaffected. We have called such a pseudo-ternary code N-pulse. For $N = 2$, $p = 1/2$, and balanced pulses, the power spectrum has a continouus component only, given by:

$$P(f) = \frac{4(1 - \cos \omega T)(3 - 2 \cos \omega T)}{5 - 12 \cos \omega T + 8 \cos^2 \omega T} P_1(f)_c. \qquad (14)$$

The spectrum is shown in Fig. 12 for 50 per cent duty cycle rectangular pulses. The power spectrum under similar conditions for $N = 3$ is also shown in this figure. As before, rectification can be used to convert back to unipolar.

In N-pulse, for dense pulse patterns the low-frequency suppression problem is more severe than in straight bipolar. Furthermore, the spectrum peaks up at low frequencies, thereby increasing the interference sprayed into AM systems using pairs in the same cable. Therefore, for this application, N-pulse is inferior to straight bipolar.

5.4 *Higher-Order Bipolar*

Another possible direction, first suggested by M. Karnaugh, involves paralleling bipolar converters and routing the unipolar pulses to each of

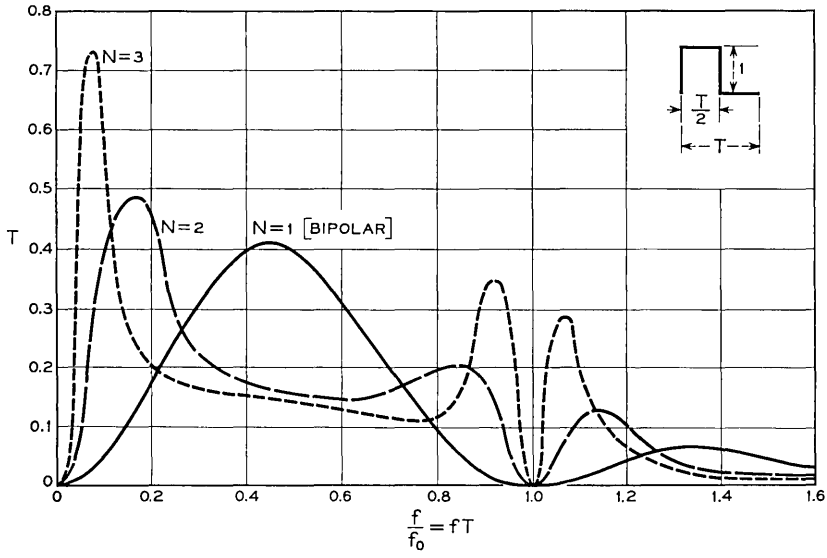


Fig. 12 — N pulse power spectra.

the converters in sequence. Such a composite converter is shown in Fig. 13. We will only give results for the power spectrum for balanced pulses and $p = 1/2$. With these conditions, for N converters in parallel the power spectral density is

$$P(f) = 2[1 - \cos N\omega T]\, P_1(f)_c. \tag{15}$$

Rectification suffices to reconstruct the original binary train.

The most interesting realization for our purposes is interleaved bipolar i.e., $N = 2$. It can be seen from (15) that the spectrum contains nulls at dc and integer multiples of half the bit rate. This permits the addition of a sinusoidal component at half the bit frequency to the pulse train for purposes of timing. In this manner, timing jitter due to finite pulse width effects is ideally removed, as is the adverse pulse density penalty. These are the principal features of this approach. As in 2-pulse, two pulses in a row may be of the same sign, thereby roughly doubling the low-frequency tail in the succeeding time slot and reducing tolerance to crosstalk. The addition and extraction of the sinusoidal component from the pulse train is not an easy nor an inexpensive circuit problem, and in practice results in an inevitable interaction between the pulse train and the sinusoidal signal to impair the decision-making process in the repeater. Specifically, balance requirements on the positive and negative pulses are particularly severe to suppress the discrete component in the pulse spectrum. The added timing component can be made sufficiently large to overcome the unbalances. However this increases the power-handling requirements of the repeater. Complete retiming and pulse width control are more difficult to implement with this approach than in bipolar for 50 per cent duty cycle pulses. The continuous spectrum at

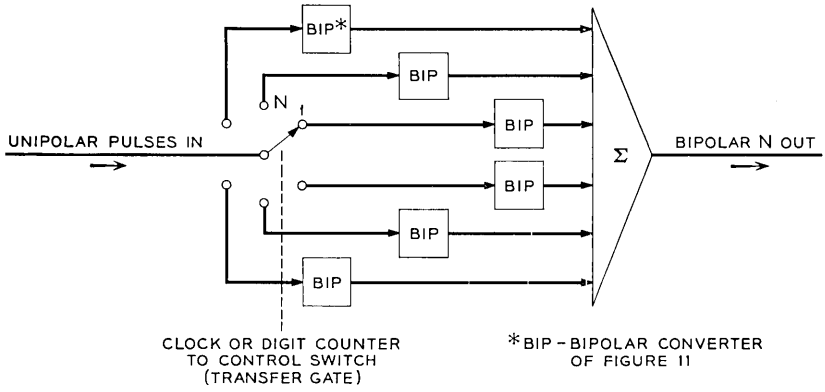


Fig. 13 — Bipolar N block diagram.

$f_r/6$ is $[1 - \cos(4\pi/6)]/[1 - \cos(2\pi/6)] = 3$ times that of bipolar, thereby making compatability between N carrier and PCM more difficult. For 100 per cent duty cycle pulses, complete retiming and width control are even more difficult, and the interference into N carrier is increased by another factor of 4* over 50 per cent duty cycle bipolar. All of the above factors point to bipolar as the preferable scheme provided that the timing jitter accumulation is satisfactory for the system lengths under consideration. This turns out to be the case.[3]

From the form of (15) it can be seen that Meacham's twinned binary can be extended to realize the same spectrum simply by delaying the unipolar train N time slots prior to subtraction. The reconversion at the receiver consists of N parallel converters, each similar to the twinned binary converter. This is an extension of the previous classification; i.e., bipolar-N involves N parallel digital converters at the transmitting terminal and a single analogue restorer at the receiver, while the extension of Meacham's twinned binary consists of an analogue operation at the transmitting terminal and N parallel digital restorers at the receiving terminal.

5.5 *Reduction of Adverse Pulse Density Penalty*

Before we go on to summarize the various pseudo-ternary codes given above, and their realization, we should pause to consider methods for reducing the adverse pulse density penalty. Several means are available, some of which have already been indicated. They include:

1. Use of a slicer to prevent interference from entering the tuned circuit in the absence of a desired pulse.
2. Addition of a timing wave at a spectral null.
3. Use of a very high Q timing circuit to integrate over a long time interval.
4. Addition of noise or a tone in an idle channel.
5. Homogenizing the code by using alternate interchange.[15]

The first three techniques given above relate to repeater design, while the last two methods involve terminal processing to minimize the problem at the source. Addition of noise or a tone to an idle channel or alternate interchange are believed to be too expensive for this application. Furthermore, these approaches will not be useful for data transmission. Item 3 above is included to point out that the adverse pulse patterns are transient effects that can be smoothed out by narrow timing filters. Too high a Q is precluded by previous considerations of stability of the tuned

* This assumes that the transmitted pulse peaks are identical in both cases.

circuit. The remaining items have been considered under bipolar and interleaved bipolar respectively.

5.6 *Other Timing Approaches*

Use of the dual-mode scheme (listed in Table I), with timing information transmitted in only one direction, can in principle avoid the crosstalk timing problem. However, delay differences between pairs requires delay build-out of the various links, which is akin to the magnitude of the pair selection problem. Similarly, we have devised separate pair timing schemes in which the NEXT problem, as far as timing is concerned, is ideally eliminated, and of course pattern jitter is removed. Again this implies the provision of several clock phases to account for differences in delay among the pairs in the cable. Variations in frequency and phase among the terminals and the transmitted clock will reduce tolerance to pulse crosstalk that obscures the decision-making process. In addition, an economic penalty results from utilizing a separate pair for timing. While it is possible to reduce the economic penalty by sharing a single clock among several pairs, this introduces a reliability problem that again is translated into increased cost. For these reasons a self-timed approach is more suitable.

5.7 *Choice*

Other self-timing approaches have been considered for this application. Some are variants of the schemes discussed, and all are inferior in most respects to those previously enumerated.

From the standpoint of timing alone, "idealized" interleaved bipolar with timing wave added at $f_r/2$ is the best approach. When we consider the problems of realization, economics, compatability with N carrier, and crosstalk into the information-bearing path this approach falls down the preference ladder below bipolar.

Bipolar is chosen for this application since it has the following advantages:

1. Energy for timing comes from a region of lower attenuation and higher crosstalk loss than in unipolar.
2. Low-frequency suppression problems are minimized, thereby relaxing transformer requirements and cost.
3. Slicing in the timing path reaps two rewards, namely (a) low-level interference in the absence of a pulse is removed; this reduces the adverse pulse density penalty; and (b) pulses exciting the tuned circuit are narrower; consequently finite pulse width effects are reduced.

4. Turnover problems associated with connecting the unbalanced repeater to the balanced line are eliminated.

5. A simple error meter can and has been devised to count bipolar violations and monitor the state of health of the line.

6. Relative compatability with N carrier is as good or better than in other schemes considered.

Most important, however, is the fact that all of these features plus complete retiming and pulse width control can be translated into a physical repeater with presently available components.

VI. PULSE SHAPING (EQUALIZATION)

Most of the foregoing material has dwelt on the problem of timing in self-timed repeaters subject to severe crosstalk interference. Even in the absence of timing interference, it should be clear that crosstalk interference will reduce the ability of a repeater to make a pulse-no pulse decision. Furthermore, the effect of crosstalk in causing errors will be dependent upon the equalization employed in the repeater. In reality, the effects of crosstalk on both timing and pulse recognition cannot be divorced. Ideally, we would like to be able to synthesize the complete repeater structure from input-output specifications. Unfortunately, the present state of the art is such that this is not possible. For this reason we address ourselves to a simpler question. What pulse shaping (both linear and nonlinear) should be employed such that a perfectly timed threshold detector can operate on the incoming signal to reconstruct the desired signal with a minimum probability of error? Since the statistics of the interference are not sufficiently well known, we cannot answer this question. Therefore, we lower our sights further and specialize to the case where the over-all equalized medium (cable plus equalization) is linear and the desired pulse presented to the threshold circuit belongs to a specified family. Several pulse shapes are suitable: i.e., linear phase low-pass Gaussian (LPG), raised cosine, or time response of maximally flat delay transfer function. We will use the LPG simply because the analysis is more tractable and because it can be approached with realizable circuitry. If we also use the lumped capacitor and single interferer to represent the crosstalk interference, we can (subject to certain simplifications) determine that member of the LPG family that minimizes the error probability for a specified coupling. Alternatively, we can consider the "worst-case eye" assumed by Mayo.[3] Under this condition, we can find the parameter of the LPG characteristic such that the sum of intersymbol interference plus peak crosstalk is a minimum for a preselected

crosstalk loss. An LPG characteristic which is 6 db down at half the pulse repetition frequency gives close to the optimum performance with an allowable crosstalk loss of about 30 db that just closes the eye. Details of this exercise in the application of the Fourier transform are not covered herein. The analysis is straightforward but messy. With the addition of low-frequency suppression, mistuning, finite width of the sampling pulse and threshold variations to intersymbol interference, it is not surprising that the allowable equivalent crosstalk loss must be raised to about 35 db to just cause errors in the real life repeater. This allowable coupling must be increased for lines longer than nominal and other factors discussed in Mayo's paper. In passing we note that the actual equalized pulse at the output of the repeater under nominal conditions agrees closely with the response of the optimum LPG characteristic to a 50 per cent duty cycle rectangular pulse.

The spectrum of the bipolar pulse train peaks up at odd integer multiples of $f_r/2$. Peaking in the neighborhood of $(3/2)f_r$ can be troublesome because energy in this region transmitted via the crosstalk path can beat with the desired timing component in the rectifier to produce additional interference at the timing frequency. Therefore, it is desirable to reduce the preamplifier gain in this region. This feature is included in the repeater.[3] Attendant to this modification is a slight change in intersymbol interference. The most pronounced advantage of this feature occurs for dense interfering pulse trains.

In principle, equalization in the timing path can profitably differ from that employed in the information path in the repeater. By the analysis outlined in the Appendix (Section A.2), it can be concluded that the pulses presented to the rectifier should be essentially the same as those giving optimum performance in the information path if the criterion is that the desired bit rate component at the rectifier output be twice that of the undesired component. This result is based on a random crosstalk interferer through a capacitive path and a desired pulse train that is random. In addition, this analysis permits us to conclude that under random conditions, crosstalk into the timing path should not be limiting, but crosstalk effects on the information bearing path are limiting. This improvement over the half bit rate figure used in Section V is due to the different spectral content of the desired pulse train and the crosstalking train that enter the square law rectifier assumed in the analysis. In this regard the bipolar repeater is superior to interleaved bipolar with timing wave added. Furthermore, it should be realized that this advantage is even greater for a dense desired pulse train and a sparse interferer. On the other hand, for adverse pulse

densities in periodic patterns the converse is true — timing becomes limiting before the pulses are obscured by crosstalk. These conclusions agree with experiment.

Before we bring this paper to its conclusion, it might be worthwhile to make a few remarks about our philosophy for choosing a transmission scheme for this application. In many respects we have focused our attention on worst cases. Attendant to this approach is the inherent danger of over-engineering. That we are free from this accusation follows from the fact that the resulting repeater satisfies economic objectives, size requirements, and permits considerable growth of the exchange plant. Furthermore, even if we neglect timing interference, other factors would dictate a repeater of virtually the same complexity. Indeed a carefully designed unipolar repeater is not significantly simpler than the bipolar repeater and suffers more loss in margin due to low-frequency suppression.

There has been a tendency in the literature on the design of PCM repeaters to consider only the mean square value of each source of interference. This is generally followed by the addition of the mean square values of all of the sources of interference and a statement that the resulting distribution of the sum is normal with mean square value given by the sum of the respective mean square values. In effect, the central limit theorem is invoked by incantation, not by proof. Furthermore even if the resulting distribution is normal in the *neighborhood* of the mean, this says nothing about the behavior in the neighborhood of the tails. This is the region of interest in most high-quality pulse systems. In point of fact it can be shown that most of the sources of interference considered herein have distributions that deviate considerably from the normal and are skewed adversely. We will demonstrate this for timing jitter due to mistuning and finite pulse width in another paper.[17] Therefore the treatment of the sum of the sources of interference as being normal involves a dangerous pitfall that we have studiously avoided.

VII. DISCUSSION

With the exception of some model building in the Appendix (Section A.2) we have dealt with essentially deterministic extrapolations rather than with true stochastic models of the real world. As noted previously, this simplification was essential in order to proceed with the development. Subject to these and other assumptions we have shown that the bipolar transmission scheme is well suited to the exchange plant environment. With a random crosstalker and a random desired signal, bipolar has a crosstalk figure of merit for worst case line loss about the same as

TPC, about 35 db. This figure is about 23 db better than that for unipolar under the same conditions. In addition, slicing in the timing path goes a long way toward eliminating the adverse pulse density penalty of 17 db.

The crosstalk figure of merit of 35 db is well below the 1 per cent loss point on the distribution given in Fig. 7. Despite the preliminary nature of these data, this crosstalk figure indicates that it certainly is possible to place a single PCM system within a unit of a unit-constructed cable with no pair selection. Indeed it should be possible to accommodate several systems. Neither of these statements can be made for unipolar. However, the question of how many systems can be installed cannot be answered quantitatively from the work reported in this paper. The multi-system performance is dependent upon how the crosstalking interferers add. This difficult statistical problem is presently being considered both analytically and experimentally by the Systems Engineering Department at Bell Telephone Laboratories. Fortunately, there are a host of techniques that can be employed to further improve the crosstalk situation. They all imply an economic penalty and involve special system layouts. For example, shorter repeater spacing and staggering repeater locations for different directions of transmission can appreciably reduce crosstalk interference. Of course, where it is possible to use separate units or separate cables for the two directions of transmission, this will significantly decrease the probability of crosstalk-induced failure.

We can add nothing to what we have already said about compatability with N carrier. This question is still under study. Here again special measures can be adopted to combat the problem.

With regard to impulse noise, extensive measurements made by T. V. Crater indicate that this should not be limiting provided that the repeater spacing from an office is about half of that of the normal line repeaters.

In short, further consideration of the interaction of the plant with the bipolar repeater is receiving considerable attention by the Systems Engineering Department at Bell Laboratories. Therefore a more quantitative picture will have to await the outcome of their work.

VIII. Conclusion

Based largely on timing considerations we have shown that the conventional unipolar PCM transmission scheme is unsuited to the Exchange Plant where crosstalk interference is the principal transmission deterrent. Several pulse transmission schemes have been examined for

replacement of unipolar. We have chosen a bipolar transmission scheme and a self-timed reconstructive repeater incorporating nonlinear timing wave extraction with complete retiming and pulse width control. Reasons for making this choice have been covered in Section V. For random pulse trains on the disturbed and disturbing pairs, bipolar virtually eliminates crosstalk timing interference as a contributor to failure of a repeater.

IX. ACKNOWLEDGMENTS

Many of the concepts and ideas put forth in the paper arose out of useful discussions with J. S. Mayo and E. E. Sumner. It is a pleasure to acknowledge their contributions to this effort.

APPENDIX

Bipolar Through a Square Law Device

A.1 *Random Signal Alone*

The signal transmitted in bipolar can be represented by

$$y(t) = \sum_{n=-\infty}^{\infty} a_n g(t - nT) \tag{16}$$

where a_n is a random variable taking on the values -1, 0, 1 with a priori probabilities $p/2$, $1 - p$, and $p/2$ respectively. It is readily shown that ave $y(t)$ is zero. When this pulse train is put through a square law device, we have at its output

$$\begin{aligned} y^2(t) &= \left[\sum_{n=-\infty}^{\infty} a_n g(t - nT) \right]^2 \\ &= \sum a_n{}^2 g^2(t - nT) + \sum_n \sum_{\substack{m \\ n \neq m}} a_n a_m g(t - nT) g(t - mT). \end{aligned} \tag{17}$$

The first term above is due to the squares of individual pulses, while the second term arises due to pulse overlaps. The average value of the squared pulse train is

$$\begin{aligned} \text{ave } y^2(t) &= \sum R(0) g^2(t - nT) \\ &\quad + \sum_n \sum_{\substack{n+k \\ k \neq 0}} R(k) g(t - nT) g(t - (n + k)T) \end{aligned} \tag{18}$$

with

$$R(k) = \text{ave } a_n a_{n+k}\,. \tag{19}$$

That the (average) output of the square law device is periodic may be shown by replacing t by $t + T$ in (18) to get

$$\begin{aligned}\text{ave } y^2(t+T) &= \sum_n R(0)g^2(t-(n-1)T) \\ &\quad + \sum_n \sum_{\substack{n+k \\ k\neq 0}} R(k)g(t-(n-1)T)g(t-(n-1+k)T) \\ &= \text{ave } y^2(t)\end{aligned} \tag{20}$$

with appropriate changes in summation indices.

Since the ensemble average is periodic, it can be expanded in a Fourier series. Following the same procedure used by Bennett[2] we get

$$\text{ave } y^2(t) = \sum_{n=-\infty}^{\infty} C_m \exp(j2m\pi f_r t) \tag{21}$$

with

$$\begin{aligned}C_m = f_r \Bigg[R(0) \int_{-\infty}^{\infty} g^2(u) e^{-j2\pi m f_r u}\, du \\ + \sum_k{}' R(k) \int_{-\infty}^{\infty} g(u)g(u+kT)e^{-j2\pi m f_r u}\, du \Bigg]\end{aligned} \tag{22}$$

The prime on the summation over k indicates that the $k = 0$ term is to be omitted.

Our main interest resides in the component at the bit frequency, namely C_1 or C_{-1}. By using the relationship between the Fourier transform of the product of two functions and the convolution of their transforms, we obtain

$$\begin{aligned}C_1 = f_r \Bigg[R(0) \int_{-\infty}^{\infty} G(-f)G(f_r+f)\, df \\ + \sum_k{}' R(k) \int_{-\infty}^{\infty} G(-f)G(f_r+f)e^{j2\pi k f T}\, df \Bigg].\end{aligned} \tag{23}$$

Since $R(k) = R(-k)$, (23) becomes

$$C_1 = f_r \int_{-\infty}^{\infty} G(-f)G(f_r+f) \Bigg[R(0) + 2\sum_{k=1}^{\infty} R(k) \cos 2\pi k f T \Bigg] df. \tag{24}$$

Multiplying numerator and denominator of the integral by $G(f)$ gives

$$C_1 = \int_{-\infty}^{\infty} \frac{G(f_r - f)}{G(-f)} P_3(f)\, df \tag{25}$$

where we have taken advantage of the fact that $P_3(f) = P_3(-f)$ and

$$P_3(f) = f_r \mid G(f) \mid^2 \left\{ R(0) + 2 \sum_{k=1}^{\infty} R(k) \cos 2\pi k f T \right\}. \quad (26)$$

It will be recalled from the text that $P_3(f)$ is the label attached to the bipolar spectrum. That the power spectrum of bipolar is given by (26) follows from the general relationship derived for the spectrum of a digital source given by Bennett[7] or from Wolds' theorem for the power spectrum of a stationary time series.[16] Alternatively it can be shown that

$$\begin{aligned} R(k) &= (-1)^{|k|} p^2 (2p - 1)^{|k|-1} \mid k \mid > 1 \\ R(1) &= -p^2 \\ R(0) &= p. \end{aligned} \quad (27)$$

Substitution of these expressions for $R(k)$ in (26) yields the expression for $P_3(f)$ given in (12).

To give a graphic picture of the frequency region that contributes most to the bit frequency component, we make the following assumptions:

1. The pulse shape is Gaussian with transform given by the low-pass Gaussian characteristic below

$$G(f) = \exp\left[-0.693 \left(\frac{f}{f_6} \right)^2 \right] \quad (28)$$

where f_6 is defined as the frequency at which the response is 6 db down from its low-frequency asymptote.

2. $p = 1/2$; therefore $R(k) = 0$ for $\mid k \mid > 1$. Under these assumptions (25) becomes

$$C_1 = \frac{f_r}{2} \int_{-\infty}^{\infty} \exp\left\{ \frac{-0.193}{f_6^2} [(f - f_r)^2 + f^2] \right\} (1 - \cos 2\pi f T)\, df. \quad (29)$$

The integrand of (29) (neglecting a multiplicative constant) is plotted in Fig. 14 for $f_6 = f_r/2$. From this figure it is apparent that the major contribution to the integral comes from the neighborhood of $f_r/2$.

A.2 *Random Signal Plus Random Crosstalkers*

If we consider the addition of several crosstalkers to the desired signal, the composite signal exciting the square law characteristic is

$$z(t) = y(t) + x(t) \quad (30)$$

where $y(t)$ is given by (16) and the crosstalk interference $x(t)$ is given by

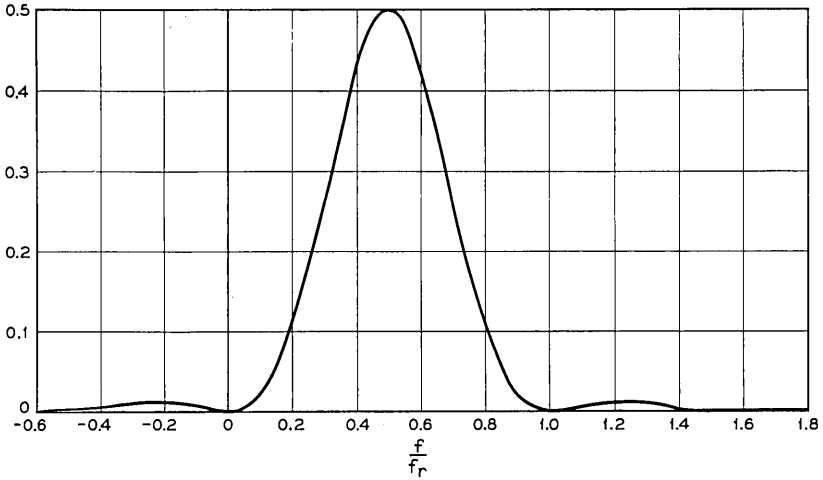


Fig. 14 — Integrand of (25) (except for a constant).

$$x(t) = \sum_{i=1}^{N} \sum_{n=-\infty}^{\infty} b_{in} h_i(t - nT_i - \tau_i) \tag{31}$$

for N interferers. It is assumed that the b_{in} are independent random variables with identical distributions, each distributed in the same manner as the a_n of the desired message. The function $h_i(t)$ is the response of the ith crosstalk path and the preamplifier of the disturbed repeater to a PCM pulse; τ_i is a measure of the phase difference between the ith crosstalk pulse train and the desired message; and T_i is the reciprocal of the pulse repetition frequency of the ith pulse train. If we let $T_i = T + \Delta_i$, then Δ_i is a measure of the difference in pulse repetition frequency between the ith crosstalker and the desired message. In general $h_i(t)$, τ_i, and Δ_i are random variables with the latter two being slowly time varying. We will assume the same frequency for all interferers; therefore $\Delta_i = 0$ for all i. When we average $z(t)$ over the a and b ensembles, we find that its average value is zero as expected. At the output of the square law device, the average value of $z^2(t)$ is

$$\text{ave } z^2(t) = \text{ave } y^2(t) + \text{ave } x^2(t) \tag{32}$$

since $x(t)$ and $y(t)$ are independent random variables with zero mean (with respect to averaging over the message ensembles). By the same argument used previously, ave $z^2(t)$ is periodic with period T and can

be expanded in a Fourier series. The component at the pulse repetition frequency, d_1, is

$$d_1 = C_1 + C_{1x} \tag{33}$$

where

$$C_{1x} = \sum_{i=1}^{N} e^{j2\pi f_r \tau_i} \int_{-\infty}^{\infty} \frac{H_i(f - f_r)}{H_i(-f)} P_{3i}(f)\, df \tag{34}$$

and C_1 is given by (24). $P_{3i}(f)$ is the power spectrum of the ith crosstalker at the input to the square law device, and is given by $P_3(f)$ when $H_i(f)$ is substituted for $G(f)$.

Let us assume that each $H_i = A_i H$. In this context, H is the transform of a "representative" crosstalk pulse and the A_i are random variables. With this relationship substituted in (33), we get

$$C_{1x} = \sum_{i=1}^{N} A_i^2 e^{j2\pi f_r \tau} f_r \cdot \int_{-\infty}^{\infty} H(f_r - f)H(f)[R(0) + 2 \sum R(k) \cos 2\pi k f T]\, df. \tag{35}$$

From (33) we can write the average timing wave component as

$$w(t) = 2 \mid C_1 \mid \cos (2\pi f_r t + \theta_1) + 2 \mid C_{1x} \mid \cos (2\pi f_r t + \theta_x) \tag{36}$$

where θ_1 and θ_x are the angles of C_1 and C_{1x} respectively. The representation of (34) enables us to define an "equivalent crosstalk interferer," namely

$$x(t)_e = \left| \sum_{i=1}^{N} A_i^2 e^{j2\pi f_r \tau_i} \right|^{1/2} \sum b_n h(t - nT - \tau). \tag{37}$$

In (37) τ is a random variable chosen to have the same distribution as the random component of $\theta_x/2\pi f_r$. Of major interest is the magnitude of the equivalent interferer. If we assume that $H(f)$ represents the frequency dependence of the "smooth" crosstalker as modified by the characteristic of the equalizer of the disturbed repeater, then we can assume that the A_i are random variables with the distribution of Fig. 7 that serve to change the level of the smooth characteristic. This model is useful primarily to give some idea as to how the crosstalkers combine to interfere with the desired timing component. We are still left with the difficult problem of determining the distribution of a complicated function of approximately log-normally distributed random variables in order to determine how the ensemble of average timing waves varies

with crosstalk interference. This of course does not indicate how the crosstalk interferers add to confuse the threshold circuit in the repeater.

Again we must lower our sights and revert back to the equivalent capacitive coupling in order to get some feel for how equalization affects the timing path. For an over-all LPG characteristic we can compute the allowable crosstalk coupling such that the ratio of desired to undesired timing component is 2 corresponding to a maximum of 30° phase shift in the tuned circuit; i.e. $| C_1 |/| C_{1x} | = 2$. The maximum allowable crosstalk will be a function of the equalization as characterized by the 6-db point on the LPG characteristic. Evaluation of the integrals for C_1 and C_{1x} (as modified for the deterministic capacitive coupling) shows that the optimum is rather broad and occurs in the neighborhood of $f_6 = f_r/2$. Furthermore the allowable crosstalk coupling is a few db larger than the allowable crosstalk coupling to close the eye in the absence of timing interference. Therefore under random pulse conditions, timing is not limiting.

Similar computations can be made for periodic pulse patterns. This has been done to check the experimental results given by Mayo[3] in his Fig. 57. Analytical results are within 2 db of measured results for the patterns that were spot checked.

REFERENCES

1. Davis, C. G., this issue, p. 1.
2. Mann, H., Straube, H. M., and Villars, C. P., this issue, p. 173.
3. Mayo, J. S., this issue, p. 25.
4. De Lange, O. E., B.S.T.J., **35,** Jan., 1956, p. 67.
5. Wrathall, L. R., B.S.T.J., **35,** Sept., 1956, p. 1059.
6. Sunde, E. D., B.S.T.J., **36,** July, 1957, p. 891.
7. Bennett, W. R., B.S.T.J., **37,** Nov., 1958, p. 1501.
8. Richman, D., Proc. I.R.E., **42,** Jan., 1954, p. 106.
9. De Lange, O. E., and Pustelnyk, M., B.S.T.J., **37,** Nov., 1958, p. 1487.
10. Rowe, H. E., B.S.T.J., **37,** Nov., 1958, p. 1543.
11. Bennett, W. R., *Proc. Symposium on Modern Network Synthesis,* **5,** 1956, p. 45.
12. Huggins, W. H., Proc. I.R.E., **45,** Jan., 1957, p. 74.
13. Zadeh, L. A., Proc. I.R.E., **45,** Oct., 1957, p. 1413.
14. Meacham, L. A., Twinned Binary Transmission, U.S. Patent No. 2,759,047.
15. Carbrey, R. L., Proc. I.R.E., **48,** Sept., 1960, p. 1546.
16. Wold, M., *A Study in the Analysis of Stationary Time Series,* Dissertation Upsala, 1938.
17. Aaron, M. R., and Gray, J. R., to be published.

Performance Limitations of a Practical PCM Terminal

By R. H. SHENNUM and J. R. GRAY

(Manuscript received July 12, 1961)

This paper discusses the performance of a practical PCM terminal for time-division speech transmission. Limitations which make performance different from more conventional frequency-division systems are considered. Particular emphasis is placed on limitations resulting from nonideal circuits and signal conditions. Both analytical and experimental results are presented with comparisons given where appropriate.

I. INTRODUCTION

The limitations of frequency-division systems have been discussed in the literature. This paper treats the limitations of a practical PCM terminal which make its performance different from more conventional systems. The areas to be discussed include (1) signal deterioration due to noise and distortion, (2) idle circuit noise and interchannel crosstalk, (3) net loss and (4) load capacity. Analytical discussions are presented and compared with experimental results on models assembled at Bell Telephone Laboratories.

II. SIGNAL DETERIORATION DUE TO NOISE AND DISTORTION

The deliberate error imparted to the signal by quantization is the significant source of signal impairment in a PCM terminal. Other effects arising in a practical terminal also affect signal quality. The sections to follow consider both the quantizing impairment found in an "ideal" terminal,* and the tolerances required in the laboratory version to meet system signal-to-noise objectives. Signal-to-noise measurements over 24 channels are also provided and substantiate calculations and assumptions.

* The word "ideal" is used because it is found in the literature. In some instances the experimental terminal was designed intentionally to depart from the so-called "ideal" performance as will become evident later.

2.1 *Quantizing Noise*

Quantization is the process of converting the exact sample values of the signal to their nearest equivalent in a discrete set of amplitudes to permit digital encoding and, therefore, essentially noise-free transmission in the medium. The error or noise produced by this "rounding-off" procedure is the major source of signal impairment. The choice of quantization or coding function is governed by the statistics of the signal to be processed. For time-division speech channels, nonuniform quantization yields the best over-all signal-to-noise performance for a given bandwidth. To obtain comparable performance with uniform quantization would require an increase in sampling frequency and/or an increase in the number of quantizing levels for the same signal range. Both of these would result in an increase in bandwidth over nonuniform quantization.

Nonuniform quantization can be achieved directly by nonlinear encoding, or by first compressing the input signal and then applying uniform quantization. In the case of the latter, an instantaneous compressor network attempts to improve performance for weak signals by giving them preferential amplification. After encoding, transmission and decoding, a device called the expandor employing the inverse characteristic restores the proper quantized amplitude distribution. The two networks together are referred to as a compandor, and the signal-to-noise advantage for weak signals over that obtained with the simpler uniform quantization is the companding improvement. In the experimental system, a logarithmic compandor and a separate uniform encoder-decoder combination are used as discussed in a companion paper.[1]* Analytic treatment of the signal-to-noise behavior of a PCM terminal with this type of quantization has been covered in a paper by B. Smith.[2] His results for an input speech signal, assuming 7-digit encoding with 26-db logarithmic companding are reproduced in Fig. 1. These latter are the broad design parameters of the system.

2.2 *Departures from "Ideal" Performance*

The performance of the laboratory terminal deviates from the signal-to-noise behavior of Fig. 1. This is attributed to several causes. One source is the difference between the logarithmic compression characteristic derived from a practical diode network and the $\mu = 100$ curve discussed by Smith. The departure is in part intentional as a result of the desire to improve the signal-to-noise ratio at low signal amplitudes.

* Discussions of certain nonlinear encoding schemes are also included therein.

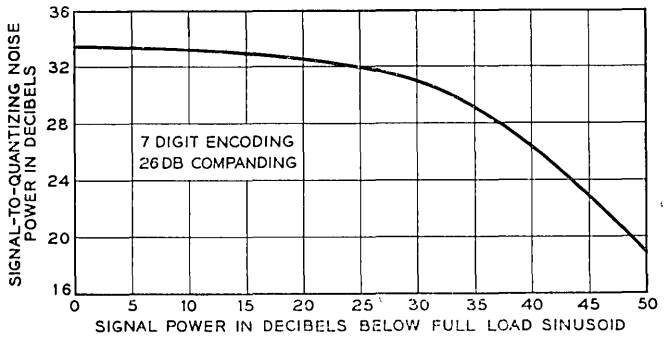


Fig. 1 — Plot of signal-to-quantizing noise as a function of signal level below full load sinusoid. Seven-digit encoding and 26-db companding are assumed.

At other values of the input signal, the diode network characteristic is constrained so that signal-to-noise ratios calculated with sinusoidal inputs using the experimental compression curve are a maximum of 3 db down at any point. The acceptance of a maximum penalty of 3 db is based on the fact that this maximum is confined to a relatively narrow signal range and on the existence of other ranges for which performance is better than shown in Fig. 1. This is shown elsewhere.[1]

Other major sources of imperfection are listed below. Each is assigned a margin such that the measured signal-to-noise performance is no more than 3 db lower than *calculated* performance as described above. These numbers assume that the contribution of each adds on a power basis to the theoretical quantizing impairment. They are also based on the fact that all do not degrade signal quality over the entire range of signal amplitudes:

(1)	Encoder-decoder fine structure	1 db
(2)	Pedestal variation prior to compression	2 db
(3)	Compandor mistracking	1 db
(4)	Encoder dc shift	1 db

In a practical encoder-decoder design, uniformity of step size and infinitely sharp transitions between adjacent quantized levels can only be realized approximately. An estimate of the effect of these imperfections on signal-to-noise behavior is calculated elsewhere.[1] Based on the margin assigned to code scale fine structure above, tolerances on the elements of the encoder and decoder networks are also given therein.

Pedestal variation is introduced at the multiplex gates as a result of sampling and causes misbiasing of the compressor network. The pre-

dominant effect in this case is a loss of companding improvement for weak signals as demonstrated by Smith.[2] Strong and midrange signals are relatively unaffected. The experimental system exhibits less of a degradation in this regard than predicted by Smith for $\mu = 100$ compandors. The reason for this is the greater range of linearity near the origin in the characteristics of the practical compandor networks. With a bias error at the compressor input of 1 per cent of the system overload voltage, approximately 2 db loss in companding improvement is found experimentally. This appears to be a realistic tolerance on pedestal variation, based on a desire for economy consistent with reasonable performance.

Compandor mistracking results when the transmission characteristics provided by the compressor and expandor networks are not exactly complementary. This introduces nonlinearities on through transmission. The primary effect is 3rd harmonic distortion as discussed by Mann, Straube, and Villars.[1] The 1-db margin allowed is based on the fact that only large signals are distorted. Weak signals are relatively unaffected since over-all linearity is maintained reasonably well in this range.

Besides 3rd harmonic distortion introduced by mistracking, 2nd harmonic distortion is also important. The outstanding contributor is unwanted dc reference drift at the coder, which in conjunction with companding results in a departure from linear transmission. To illustrate this effect, the transmission between compressor input and expandor output with an intermediate reference shift ϵ at the encoder is determined below.

Using the compression characteristic given by Smith, the compressor output is defined by*

$$\begin{aligned} y &= \frac{\log (1 + \mu x)}{\log (1 + \mu)} \qquad (0 \leqq x \leqq 1) \\ y &= \frac{-\log (1 - \mu x)}{\log (1 + \mu)} \qquad (-1 \leqq x \leqq 0) \end{aligned} \tag{1}$$

where μ is the compression parameter, and x is the input normalized to the compressor overload voltage. If quantization at the encoder is neglected, but its reference shifts by an amount ϵ, then the decoded output becomes $y + \epsilon$. Taking the expandor characteristic to be the inverse of the above, and assuming ϵ to be small and positive, then the expandor output as a function of the compressor input is approximately

* Unless otherwise specified, natural logarithms are implied.

$$F(x) \sim \alpha x \qquad (0 \leqq x \leqq 1)$$
$$F(x) \sim \frac{1}{\alpha} x \qquad (-1 \leqq x \leqq 0) \tag{2}$$

where $\alpha \sim 1 + \epsilon \log (1 + \mu)$. Equation (2) is illustrated by Fig. 2.

For a sinusoidal test signal of amplitude E at the compressor input, the expandor output would appear as a positive half-cycle sine wave with amplitude αE followed by a negative half-cycle of amplitude E/α. The 2nd harmonic content of such a wave is $-[(2/3\pi)E(\alpha^2 - 1/\alpha)]$ and the signal-to-2nd harmonic distortion becomes $| S/D | = 3\pi\alpha/2(\alpha^2 - 1)$. This result is plotted in Fig. 3, with ϵ given as a percentage of compressor overload voltage. In calculating this curve, μ is taken to be 100, corresponding approximately to the companding advantage obtained in the experimental system.

To meet the 1-db noise penalty assigned to this cause, coder reference drift must be restricted. The allowable drift is determined from the minimum acceptable signal-to-distortion ratio. It is assumed that only strong signals are to be penalized. Since the "ideal" signal-to-noise curve

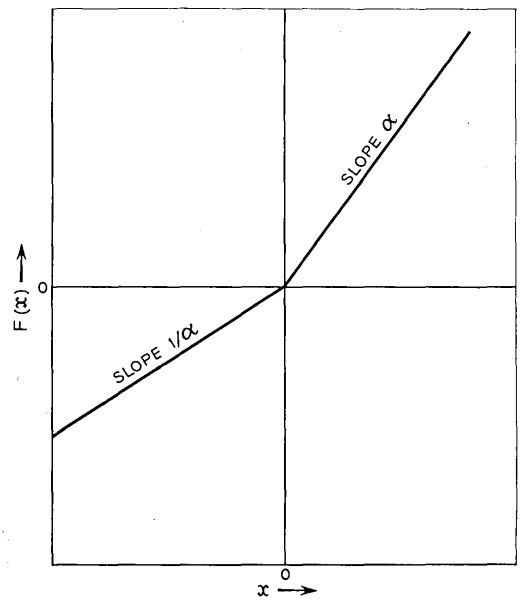


Fig. 2 — System transmission characteristic with a shift ϵ in the coder reference. Quantization is neglected.

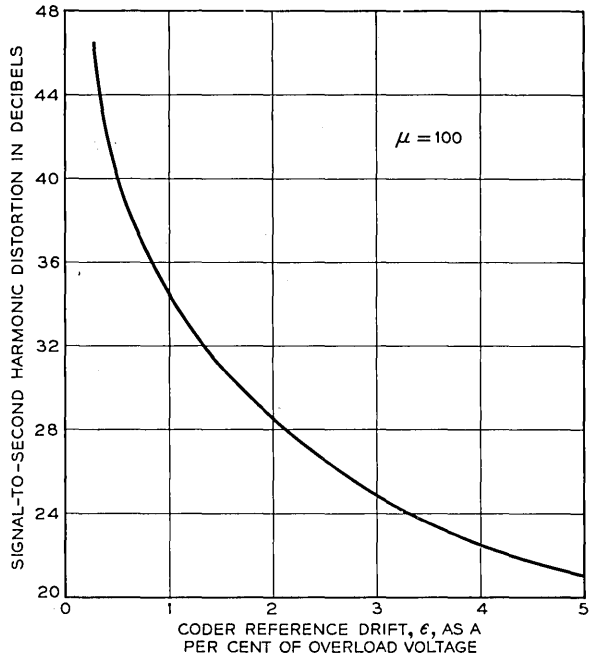


Fig. 3 — Plot of signal-to-second harmonic distortion as a function of coder reference shift expressed as a per cent of full load.

of Fig. 1 is approximately 34 db near full load, then using power law addition a $| S/D |$ of about 40 db is required to meet the 1-db margin. From Fig. 3 the 40-db figure implies a $\frac{1}{2}$ per cent tolerance on the coder dc shift.

2.3 *Signal-to-Noise Behavior of the Experimental Terminal*

A series of tests was performed to evaluate the degree to which the previous assumptions and calculations describe the laboratory model. In all cases, tests were made with sinusoids to permit probing the largest quantizing steps of the companded-coding system without the difficulty of peak clipping.

The signal-to-noise ratio was measured with an instrument which compares the average power of signal plus noise to the average power of the noise including distortion products. No measurements were made which allowed separation of the total noise into its components. The noise was determined by removing the signal with a narrow-band rejection filter. Both indications are calibrated to read rms for sinusoidal

signals. With the signal-to-noise ratios of interest, no special problem exists in establishing the output signal power in this manner. Careful checks indicate that errors of up to 1 db are possible for low-level signals. Fig. 4 shows the signal-to-noise ratio as a function of signal level for the extreme channels and the mean curve for 24 channels. The measurements were made with no attempt to optimize individual channels. Channel-to-channel variation is attributed primarily to pedestal variation prior to compression and dc shifts in the encoder reference. The curve corresponding to the previously stated objective is also shown.* Adequate performance is indicated.†

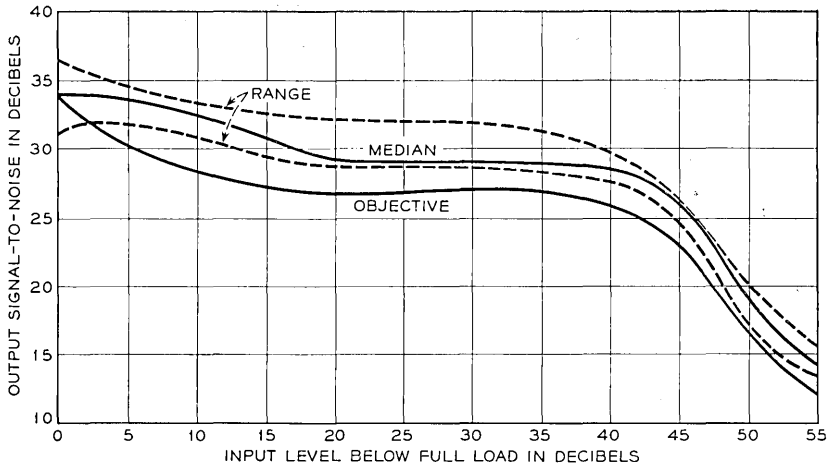


Fig. 4 — Plot of signal-to-noise measurements over 24 channels.

When sinusoidal modulating signals are used for system testing, variation in signal-to-noise as a function of the excitation frequency is to be expected.[3] When the frequency is small in relation to the sampling frequency, approximately the same performance is obtained for both sine waves and speech. As the frequency is increased, quantizing impairment is reduced. This is illustrated by the signal-to-noise measurements of Fig. 5. A constant input level is used for these measurements. The channel bandpass characteristics are given in Fig. 6 to explain the signal-to-noise falloff above 2600 cycles.

* This curve is obtained by reducing the signal-to-noise ratio by 3 db at all points on the curve for the actual compandor in Fig. 11 of Ref. 1.

† These results do not display common equipment variations contributing to system-to-system differences.

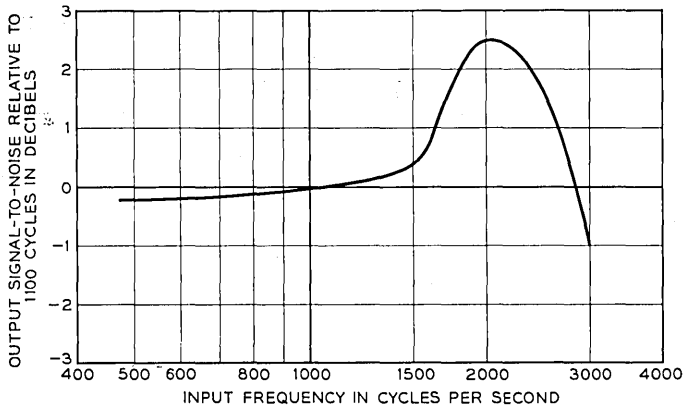


Fig. 5 — Plot of signal-to-noise ratio vs input frequency.

Submultiples of the sampling frequency were avoided in the measurements because at such points harmonics of the input signal produced by quantization beat with harmonics of the sampling frequency. For this reason, the measurements described previously were made with an 1100-cycle tone.

III. IDLE CIRCUIT NOISE AND INTERCHANNEL CROSSTALK

As a result of pedestal variation, PCM channels in the absence of speech modulation exhibit an important enhancement of weak interference. Both noise and interchannel crosstalk are affected. The en-

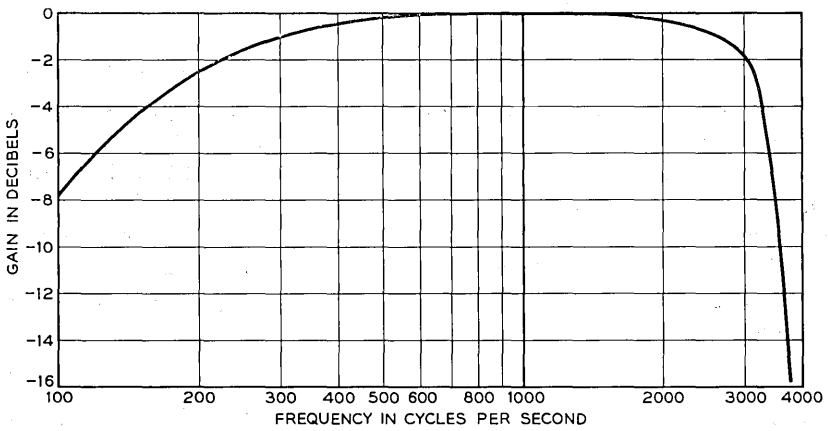


Fig. 6 — Typical bandpass characteristic of the channels.

hancement varies with pedestal drift and input interference, but is most pronounced when a quiescent channel is biased at the boundary between two adjacent quantizing steps or "mid-riser." Under this condition, any minute interference changes the quantized output and considerable enhancement is possible. In the following, a calculation of the interference power resulting at the system output under idle circuit conditions is presented. In addition, a series of noise and crosstalk measurements made on the experimental terminal are discussed.

3.1 *Analysis*

As a starting point for the analysis, it is assumed that noise and crosstalk are introduced into an idle channel prior to compression. The transmission characteristic between compressor and expandor is then derived, and the power spectrum of the interference at the expandor output is calculated. Finally, the noise and crosstalk spectrum components of interest are modified by the response of the system output filter and performance evaluated.

3.1.1 *Small-Signal Compressor-Expandor Transmission Characteristic*

With logarithmic companding and a perfect quantizer, the transmission characteristic between compressor input and expandor output is shown in Fig. 7. As indicated in the figure, the intermediate quantizer characteristic is decomposed into the sum of a linear term and a periodic sawtooth function for analytical convenience. This model for the staircase transducer was suggested by S. O. Rice and discussed in an earlier paper by W. R. Bennett.[4] The individual network characteristics can therefore be expressed as follows:

$$
\begin{aligned}
&\text{Compressor:}\begin{cases} Y = \dfrac{V_0 \log\left(1 + \mu \dfrac{X}{V_0}\right)}{\log(1+\mu)} & (0 \leqq X \leqq V_0) \\ Y = \dfrac{-V_0 \log\left(1 - \mu \dfrac{X}{V_0}\right)}{\log(1+\mu)} & (-V_0 \leqq X \leqq 0) \end{cases} \\
&\text{Quantizer:} \quad Z = Y + \frac{V}{\pi}\sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin\frac{2\pi n Y}{V} \quad (-V_0 \leqq Y \leqq V_0) \qquad (3) \\
&\text{Expandor:}\begin{cases} W = \dfrac{V_0}{\mu}\left(e^{Z/V_0 \log(1+\mu)} - 1\right) & (0 \leqq Z \leqq V_0) \\ W = \dfrac{-V_0}{\mu}\left(e^{-Z/V_0 \log(1+\mu)} - 1\right) & (-V_0 \leqq Z \leqq 0). \end{cases}
\end{aligned}
$$

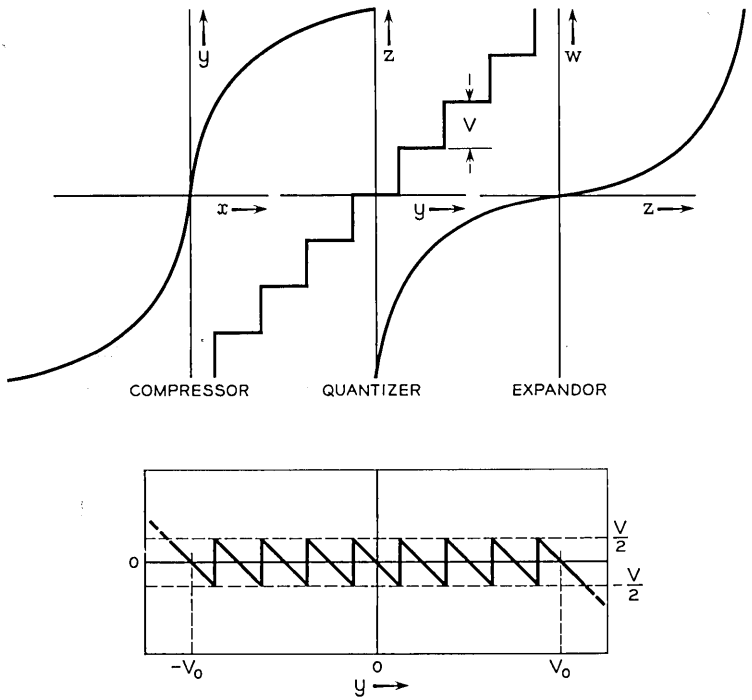


Fig. 7 — Transmission characteristics of the compressor, quantizer, and expandor. The quantizer transmission is decomposed into the sum of a linear term plus a periodic sawtooth function.

In these expressions V_0 is the system overload voltage, μ the compression parameter, and V the quantizer step size.

For a sufficiently small interference at the compressor input riding on a dc pedestal X_0, we can expand Y in a power series in the neighborhood of X_0, retaining only the first two terms. For positive X, therefore

$$Y \sim Y_0 + G(X - X_0) \tag{4}$$

where

$$Y_0 = \frac{V_0 \log \left(1 + \mu \frac{X_0}{V_0}\right)}{\log (1 + \mu)}$$

is the dc level at the quantizer input and

$$G = \frac{\mu}{\log (1 + \mu)} \cdot \frac{1}{1 + \mu \frac{X_0}{V_0}}$$

is the compressor small-signal gain.

Similarly, since the quantizer ac output will also be relatively small, and assuming that the operating point Z_0 at the expandor input is approximately Y_0,* then the expandor transmission can be written as

$$W \sim \frac{V_0}{\mu}\left(e^{Y_0/V_0\log(1+\mu)} - 1\right) + \frac{\log(1+\mu)}{\mu} e^{Y_0/V_0\log(1+\mu)}(Z - Y_0) \quad (5)$$

or substituting for Y_0 from above

$$W \sim X_0 + \frac{1}{G}(Z - Y_0).$$

Therefore, using the functional dependence of Z on Y as given by (3) and combining it with (4), W as a function of X becomes

$$W \sim X_0 + (X - X_0) + \frac{V}{G}\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin\left[\frac{2\pi n}{\frac{V}{G}}\left((X - X_0) + \frac{Y_0}{G}\right)\right]. \quad (6)$$

The input X consists of the idle channel interference (noise and crosstalk) plus the dc pedestal X_0. Therefore, $X' = X - X_0$ is the interference alone. If we denote Y_0/G by X_0' and V/G by V', then W may be written as

$$W \sim (X_0 - X_0') + (X' + X_0') + \frac{V'}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin\frac{2\pi n}{V'}(X' + X_0'). \quad (7)$$

Neglecting $X_0 - X_0'$, which only produces a dc component at the output, W as a function of $X' + X_0'$ is given by

$$W \sim (X' + X_0') + \frac{V'}{\pi}\sum_{n=1}^{\infty}\frac{(-1)}{n}\sin\frac{2\pi n}{V'}(X' + X_0'). \quad (8)$$

This result is shown in Fig. 8 and is the small-signal transmission characteristic. The quantity X_0' gives the shift in reference produced by the dc pedestal X_0, and V' is the original quantizer step size reduced by the gain factor G. Since G varies inversely with X_0, then V' becomes larger as X_0 and thus X_0' are increased.

3.1.2 *Power Spectrum of the Interference at the Expandor Output*

The expandor output spectrum can be obtained from the statistical properties of the train of interference samples obtained at this point.

* This neglects dc shift as a result of quantization.

Assuming unit impulse sampling, the output spectrum is given by[5]

$$P(f) = \frac{1}{T}\sum_{k=-\infty}^{k=\infty} \psi_W(kT)e^{i2\pi f\,kT}$$
$$= \frac{1}{T}\left[\psi_W(0) + 2\sum_{k=1}^{\infty}\psi_W(kT)\cos 2\pi f\,kT\right] \tag{9}$$

where T is the reciprocal of the sampling frequency f_s, and $\psi_W(kT)$ is the autocorrelation function at the expandor output evaluated at the sampling instants kT. This function can be determined directly from (8). Assuming $X'(t) = n(t) + E \sin 2\pi f_0 t$, where $n(t)$ is the input noise and the sinusoid represents crosstalk, then

$$W(t) = n(t) + E \sin 2\pi f_0 t + X_0'$$
$$+ \frac{V'}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin\left\{\frac{2\pi n}{V'}\,(n(t) + E\sin 2\pi f_0 t + X_0')\right\}. \tag{10}$$

Using this result, $\psi_W(kT)$ is calculated by definition:

$$\psi_W(kT) = \langle W(t)W(t + kT)\rangle$$

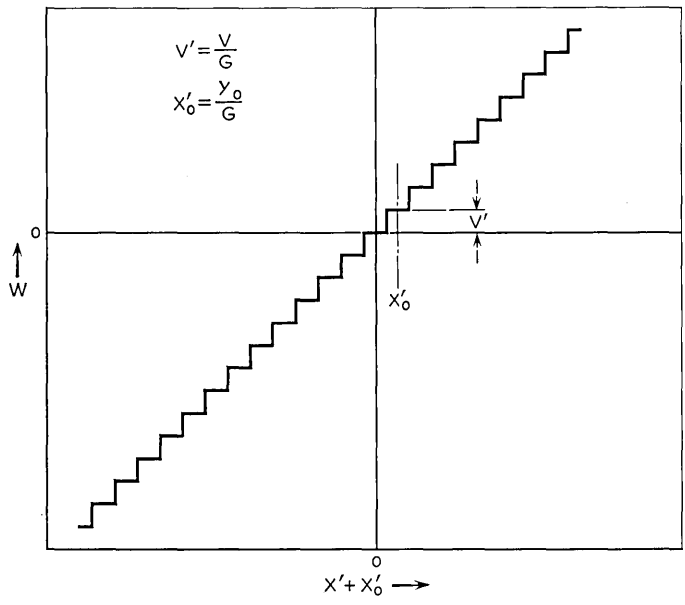


Fig. 8 — Small-signal transmission characteristic between compressor input and expandor output. The quantity X_0' gives the shift in reference produced by a dc pedestal at the compressor input.

where $\langle\ \rangle$ denotes expected value. If the input variables are all independent, the calculation of the autocorrelation function, although tedious, presents no formal difficulties. For the terms in the computation involving averages over the noise $n(t)$, a Gaussian amplitude distribution is assumed. This involves the autocorrelation of the input noise φ_{kT}, which is given by

$$\varphi_{kT} = \varphi_0 \frac{\sin \pi k}{\pi k} = \begin{cases} \varphi_0 & \text{for } k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

for a uniform noise spectrum over the voice band $(-f_s/2, f_s/2)$. The input noise power is φ_0. Under these conditions, (12) for $\psi_W(kT)$ is obtained. The J's in this expression are Bessel functions of the first kind. This result is used in (9) to determine the spectral components of noise and crosstalk.

$$\begin{aligned}
\psi_W(kT) &= \varphi_{kT}\left(1 + 4\sum_{n=1}^{\infty}(-1)^n J_0\left(\frac{2\pi nE}{V'}\right)\cos\frac{2\pi nX_0'}{V'}e^{-2(n\pi)^2\varphi_0/V'^2}\right) \\
&+ \left(\frac{E^2}{2} + 2EV'\sum_{n=1}^{\infty}\frac{(-1)^n}{n\pi}J_1\left(\frac{2\pi nE}{V'}\right)\cos\frac{2\pi nX_0'}{V'}\,e^{-2(n\pi)_2\varphi_0/V'^2}\right) \\
&\cdot\cos 2\pi f_0 kT + \left(X_0'^2 + 2X_0'V'\sum_{n=1}^{\infty}\frac{(-1)^n}{n\pi}J_0\left(\frac{2\pi nE}{V'}\right)\sin\frac{2\pi nX_0'}{V'}\right. \\
&\left.\cdot e^{-2(n\pi)^2\varphi_0/V'^2}\right) + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{V'^2}{\pi^2}\frac{(-1)^{m+n}}{mn}\,e^{-2(m^2+n^2)\pi^2\varphi_0/V'^2} \\
&\cdot\left[\cos\frac{2\pi(m-n)X_0'}{V'}\,e^{4mn\pi^2\varphi_{kT}/V'^2}\left(\tfrac{1}{2}J_0\left(\frac{2\pi nE}{V'}\right)J_0\left(\frac{2\pi mE}{V'}\right)\right.\right. \\
&+ \sum_{l=1}^{\infty}J_l\left(\frac{2\pi nE}{V'}\right)J_l\left(\frac{2\pi mE}{V'}\right)\cos 2\pi l f_0 kT\Big) + \cos\frac{2\pi(m+n)X_0'}{V'} \\
&\cdot e^{-4mn\pi^2\varphi_{kT}/V'^2}\left(\sum_{l=0}^{\infty}J_{2l+1}\left(\frac{2\pi nE}{V'}\right)J_{2l+1}\left(\frac{2\pi mE}{V'}\right)\cos 2\pi(2l+1)f_0 kT\right. \\
&- \tfrac{1}{2}J_0\left(\frac{2\pi nE}{V'}\right)J_0\left(\frac{2\pi mE}{V'}\right) - \sum_{l=1}^{\infty}J_{2l}\left(\frac{2\pi nE}{V'}\right)J_{2l}\left(\frac{2\pi mE}{V'}\right) \\
&\left.\left.\left.\cdot\cos 2\pi(2l)f_0 kT\right)\right)\right].
\end{aligned} \tag{12}$$

3.1.3 *System Output Power*

To determine the system output power, the expandor output spectrum, obtained from (9) and (12), is modified by the response of a channel filter. At the filter output we are interested in the following quantities.

(1) Total voice band noise power in the continuous spectrum in the absence of an input crosstalk signal.

(2) Total crosstalk power at the frequency f_0 with noise masking.*

Assuming that the output filter is flat over the voice band, has no dc transmission, and cuts off sharply at f_s, then the noise and crosstalk powers of interest become

$$N = \frac{1}{T^2}\left[\varphi_0\left(1 + 4\sum_{n=1}^{\infty}(-1)^n \cos\frac{2\pi n X_0'}{V'}\, e^{-2(n\pi)^2\varphi_0/V'^2}\right)\right]$$
$$+ \frac{1}{T^2}\left[\frac{V'^2}{2\pi^2}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{(-1)^{m+n}}{mn}\left(\cos\frac{2\pi(m-n)X_0'}{V'}\, e^{-2(m-n)^2\pi^2\varphi_0/V'^2}\right.\right.$$
$$\left.\left. - \cos\frac{2\pi(m+n)X_0'}{V'}\, e^{-2(m+n)^2\pi^2\varphi_0/V'^2}\right)\right] \qquad (13)$$
$$- \frac{1}{T^2}\left[\frac{V'^2}{\pi}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{(-1)^{m+n}}{mn}\, e^{-2(m^2+n^2)\pi^2\varphi_0/V'^2}\right.$$
$$\left.\cdot\sin\frac{2\pi n X_0'}{V'}\sin\frac{2\pi m X_0'}{V'}\right]$$

$$\overline{X} = \frac{1}{2T^2}\left[E + \frac{2V'}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}J_1\left(\frac{2\pi nE}{V'}\right)\cos\frac{2\pi n X_0'}{V'}\,\frac{e^{-2(n\pi)^2\varphi_0}}{V'^2}\right]^2. \qquad (14)$$

These expressions are evaluated in the following sections.

3.1.4 *Idle Circuit Noise*

Using the above result for N, curves of N vs X_0' with φ_0/V'^2 as a parameter can be plotted as shown in Fig. 9. The constant $1/T^2$ is omitted. In calculating these results, it is assumed that when X_0 varies such that X_0' moves between two adjacent decisions levels of Fig. 8, the step size V' remains constant.† That constant is determined by the value of X_0 which places the input X' at the midpoint of one of the input step intervals. Hence if $X_0' = \bar{m}V'$, for $\bar{m} = 0, 1, 2, 3, \cdots$ then

$$\bar{m} = \frac{V_0}{V}\,\frac{\log\left(1 + \mu\,\dfrac{X_0}{V_0}\right)}{\log(1+\mu)}$$

and solving the above for X_0 in terms of $\bar{m}$

* The output power at f_0 is a measure of the intelligible crosstalk interference.

† This approximation is valid for a small range of variation since the slope G of the compressor characteristic does not change materially for a small change in X_0.

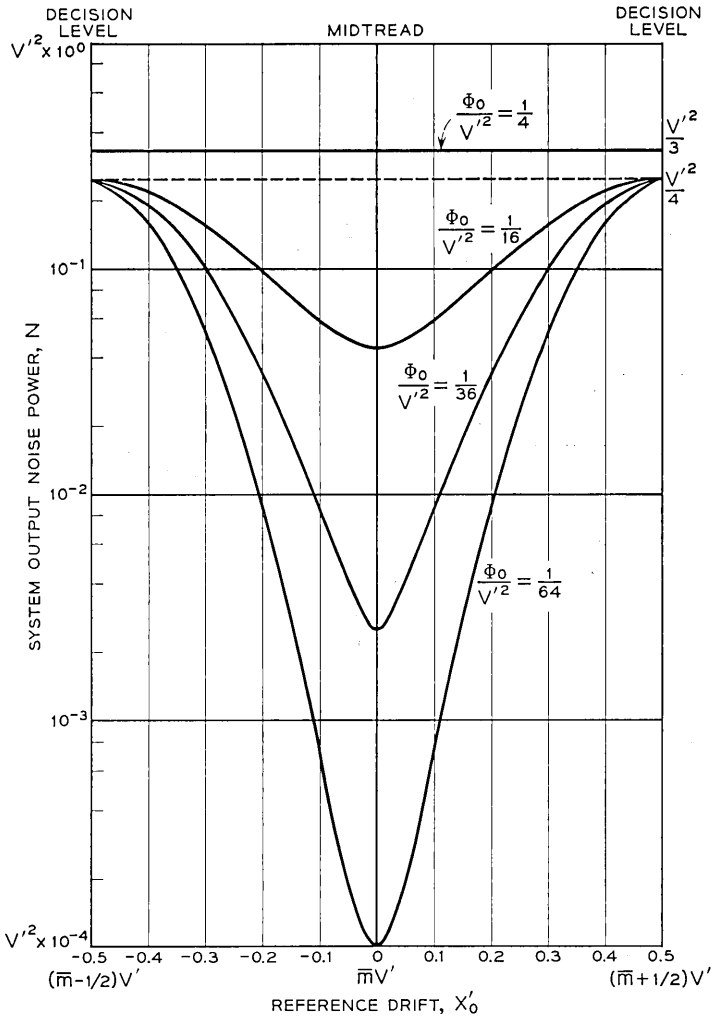


Fig. 9 — System output noise power vs reference drift X_0. The input noise power as a fraction of the step size V' is taken as a parameter. The reference drift is shown to vary between any two adjacent decision levels.

$$V' = \frac{V}{G} = V \frac{\log (1 + \mu)}{\mu} \exp \left\{ \bar{m} \frac{V}{V_0} \log (1 + \mu) \right\}. \qquad (15)$$

This result determines the value of V' to be used in Fig. 9 for any dc pedestal X_0 and thus $\bar{m}$.

With the above conditions in mind, the results of Fig. 9 are interpreted. First for φ_0/V'^2 sufficiently large, N becomes independent of

X_0'. The curve with $\varphi_0/V'^2 = 1/4$ is representative of this case and follows directly from (13). For example, the contribution of the summation appearing in the first bracketed term of this equation approaches zero under this condition, and the only contribution of any importance in the second and third terms is that for $n = m$. Therefore N reduces to the approximate expression

$$N \sim \frac{1}{T^2}\left[\varphi_0 + \frac{V'^2}{2\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\right] = \frac{V'^2}{T^2}\left[\frac{\varphi_0}{V'^2} + \frac{1}{12}\right] = \frac{V'^2}{3T^2} \qquad \left(\frac{\varphi_0}{V'^2} = \frac{1}{4}\right). \tag{16}$$

Hence, in this case, the output noise consists of the input noise plus a quantizing noise term in accord with Bennett's analysis.[4]*

When the input noise becomes smaller, the largest output power occurs for $X_0' = (m \pm \frac{1}{2})V'$ or the reference at "mid-riser" (quantizer decision level). In this situation the output is independent of the input when the latter is sufficiently small. Under this condition, the quantizer of Fig. 8 puts out a square wave with ac swing V' and random zero crossings. This results in a lower limit beyond which the system output noise power cannot be reduced. This lower limit is $V'^2/4$ and is called the noise "floor." As the dc pedestal becomes larger, V' and the "floor" become higher. This is a worst-case situation and occurs only for "mid-riser" biasing. As indicated in Fig. 9, performance improves as the reference departs from this location.

3.1.5 *Interchannel Crosstalk*

In evaluating crosstalk performance from the expression for $\overline{\underline{X}}$ in (14), we consider only the "mid-riser" condition or $X_0' = (m \pm \frac{1}{2})V'$ for $m = 0, 1, 2, 3, \cdots$. This yields the worst possible crosstalk performance for weak input crosstalk signals. Hence $\overline{\underline{X}}$ becomes

$$\overline{\underline{X}} = \frac{1}{2T^2}\left[E + \frac{2V'}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}J_1\left(\frac{2\pi nE}{V'}\right)\exp - \left(\frac{2(n\pi)^2\varphi_0}{V'^2}\right)\right]^2. \tag{17}$$

This result is plotted in Fig. 10 as a function of input crosstalk power with φ_0/V'^2 as a parameter. As before, $1/T^2$ is omitted, and it is assumed that E/V' is <1 so that the input crosstalk signal is confined to one step. V' is related to V by (15).

The results of Fig. 10 parallel those of Fig. 9 very closely. For example

* Since $V' = V/G$, where $1/G$ is the expandor small signal gain, then the term $V'^2/12$ also illustrates companding improvement as a function of dc pedestal X_0 as discussed by Smith in Ref. 2.

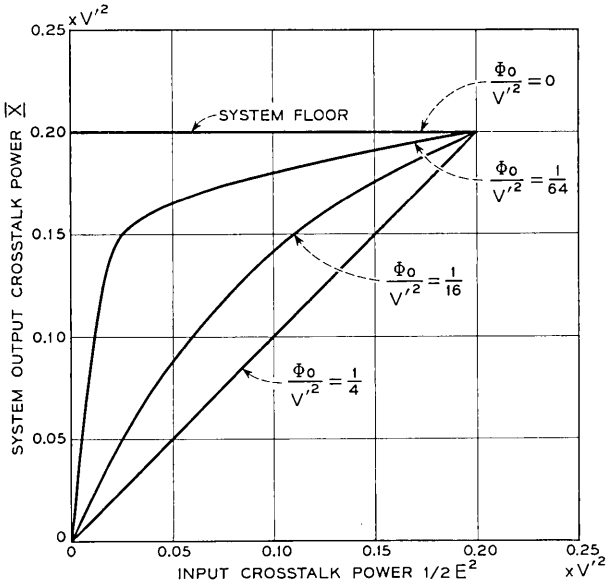


Fig. 10 — System output crosstalk power vs input crosstalk power with the input noise power as a fraction of the step size V' taken as a parameter. The quantizer reference is assumed to be at a decision level for purposes of determining this set of curves.

with φ_0/V'^2 sufficiently large ($\varphi_0/V'^2 = 1/4$), $\bar{X}$ becomes $E^2/2T^2$. In other words, the system transmits the input crosstalk signal without enhancement. Qualitatively, when the input crosstalk is sinusoidal, background noise tends to scramble its zero crossings, thus producing a width modulation in the output square wave. The net result is linear transmission at f_0, the frequency of the input, provided the noise masking is large enough.

For other values of input noise, a nonlinear input-output power relationship exists. In the special case $\varphi_0/V'^2 = 0$, $\bar{X}$ reduces to

$$1/2T^2(2V'/\pi)^2.$$

This corresponds to the power in the fundamental of an output square wave with unperturbed zero crossings. The system output crosstalk power that exists under these conditions is called the crosstalk "floor," and is the minimum power output when the bias is at "mid-riser." As in the case of idle circuit noise, the crosstalk "floor" depends directly on V', and is therefore higher as the dc pedestal becomes larger.

3.1.6 *System Performance*

In this section, performance is put in numerical terms by using the parameters of the experimental system. The parameters of interest are: (1) 7-digit encoding (128 quantizing levels), (2) 26-db companding ($\mu = 100$), (3) clipping level or overload point defined by a +3-dbm sine wave at the system input (0-db TL) and (4) 2-db net loss from system input to output. Using (1) through (4), the following relationships can be inferred

$$\frac{V_0}{V} = 64$$

$$G\Big|_{x_0 \to 0} = 20 \log_{10} \frac{\mu}{\log (1 + \mu)}\Big|_{\mu=100} = 26 \text{ db}$$

$$10 \log_{10} \frac{V_0^2}{2} = \begin{cases} +3 \text{ dbm at 0-db TL (system input)} \\ +1 \text{ dbm at } -2\text{-db TL (system output)}. \end{cases}$$

These numbers, used in conjunction with Figs. 9 and 10, are sufficient information to evaluate system performance with any dc pedestal X_0. A worst-case situation exists when the dc reference for the interference is at a decision level. The performance obtained under this condition will therefore be considered exclusively in what follows.

First, it is assumed that the dc pedestal is sufficiently small so that $X_0 = V'/2$ and $V' = V[\log (1 + \mu)/\mu]$ for $\bar{m} = 0$, as given by (15). In this case the system noise "floor" is 18 dba at the −2TL with F1A line weighting,* and the crosstalk "floor" at this same point is −65 dbm. The output noise power of 18 dba is the minimum obtainable when the reference shifts to the first decision level. This then is a fundamental system limitation. On the other hand, the crosstalk "floor" exists only without the masking effect of background noise. The improvement obtained with noise masking is related in a nonlinear fashion to the mean square noise at the compressor input as indicated in Fig. 10. With 20 dba of input noise referred to the 0-db TL ($\varphi_0/V'^2 = 1/4$), no crosstalk enhancement occurs, and the system output crosstalk power is linearly related by the system net loss to the input power at the 0-db TL.

For a larger dc pedestal, V' is increased and the system noise and crosstalk "floors" are higher. The degradation in system performance is determined directly from (15) and is given by $20 \log_{10} (1 + \mu X_0/V_0)$ db. It should be pointed out that this same factor is involved in deter-

* Using F1A line weighting, "white" noise in a 3-kc band at 0 dbm corresponds to 82 dba.

mining the amount of noise masking required to achieve a given improvement over the crosstalk "floor". For example, for a larger value of V' the input noise power φ_0 must also be larger in order to maintain φ_0/V'^2 constant and hence the same masking ratio.

The effect of a larger dc pedestal, as illustrated above, is essentially the same as the loss in companding improvement from this same cause, as discussed by Smith. Therefore, as mentioned in Section 2.2, the effect is not as pronounced in the experimental system as one would predict from "ideal" logarithmic companding when the factor $(1 + \mu X_0/V_0)$ applies.*

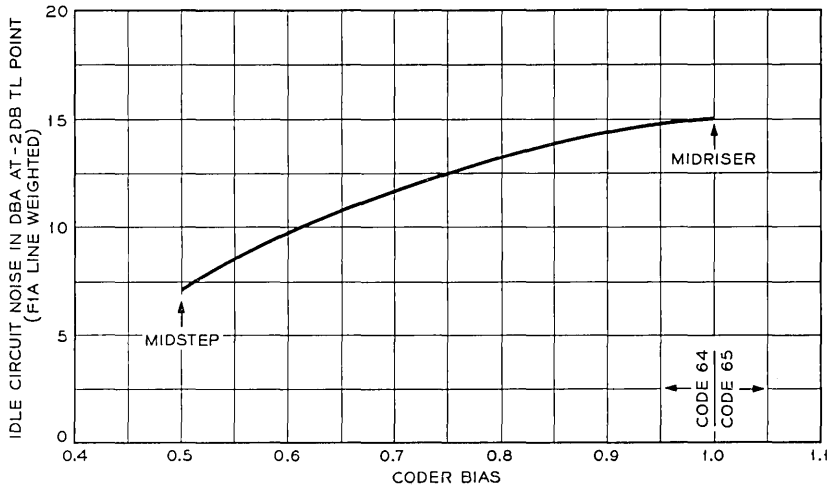


Fig. 11 — Measured output noise power vs coder reference drift in the range between codes 64 and 65. All measurements are in dba and are referred to the −2-db TL point.

3.2 *Measured Noise Performance*

All evaluations of the output idle circuit noise were made with a 2B noise meter using F1A line weighting. In all cases, the other end of each trunk was terminated in 900 ohms + 2μf, the nominal terminating impedance of the trunks. Variations of noise as a function of the bias position on the coder step are shown in Fig. 11 with variations from 7 to 15 dba. All measurements are at the −2-db TL point.

The general flattening of the experimental curve of Fig. 11 could be attributed to a substantial input noise at the compressor referring to

* For $X_0/V_0 = 1\%$, $(1 + \mu X_0/V_0)$ yields a 6-db degradation. The loss in companding improvement for a 1 per cent pedestal is only about 2 db in the experimental system so that the noise and crosstalk "floors" would only be 2-db higher.

TABLE I — IDLE CIRCUIT NOISE IN DBA AT −2-db TL

Channel Number	DBA	Channel Number	DBA
1	11.0	13	7.0
2	11.0	14	6.0
3	12.0	15	11.0
4	11.5	16	14.0
5	10.5	17	11.0
6	15.0	18	11.0
7	6.0	19	12.0
8	14.0	20	13.5
9	11.0	21	14.0
10	10.5	22	13.5
11	14.0	23	11.0
12	13.0	24	11.0

the curves of Fig. 9. Also, when the coder is adjusted for the decision-level bias position, some variation in bias occurs, so that the output power at this point tends to be an average over a range of bias. For this reason, the 15-dba measured noise floor is lower than the calculated value of 18 dba. Table I gives a typical distribution of noise measured on all 24 channels with no special care being taken to optimize each.

The total range 6 to 15 dba is explained by Fig. 11 with the distribution fairly uniform between these extremes. The mean value in dba is 11.4, which would be expected from a large number of channels uniformly distributed in bias if Fig. 11 were a linear plot.

3.3 *Measured Crosstalk Performance*

To obtain realistic results, all crosstalk evaluation was made with 15 dba (F1A weighted) of noise applied to the input of the trunks under test. The crosstalking channel was supplied with a 0-dbm sinusoid (at the 0-db TL point) at a frequency of 1100 cycles.

Fig. 12 shows the results of measurements of crosstalk of each channel into every other channel. The only pattern of significance in these data is that of a minimum of crosstalk among channels in the same compressor. Earlier descriptions of the dual compressor plan by Davis[6] have explained the grouping of all even-numbered channels in one compressor and odd in a second. The best crosstalk performance between channels in the same compressor is in the range of −72 to −75 dbm, whereas for channels in the other compressor the range is −73 to −80 dbm. This is a clear indication of transmitter terminal crosstalk.

Extensive measurements have not been made of crosstalk in the absence of intentionally introduced random noise in the channel being measured; however, sample tests show that crosstalk varied up to 6 db

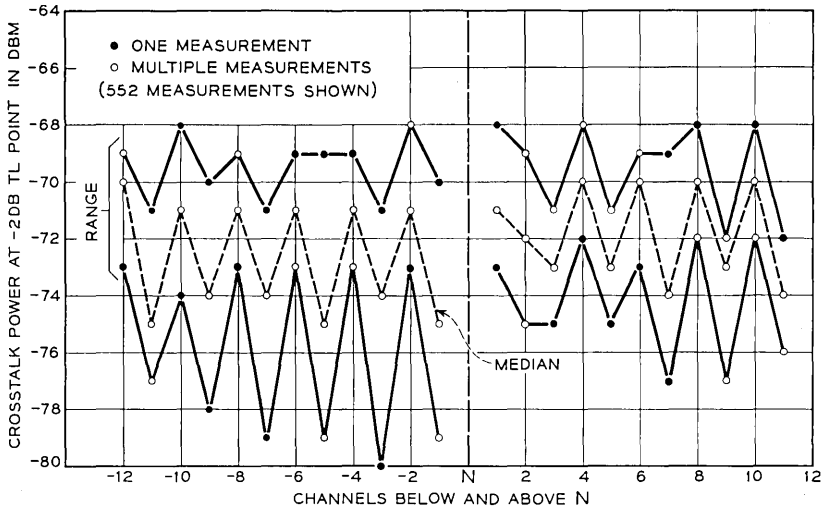


Fig. 12 — Measured output crosstalk power in the $(N + K)$th channel due to a tone in the Nth. All measurements are referred to the −2-db TL point.

with and without noise added. The crosstalk enhancement is in practice never infinite, even with the analog crosstalk small compared to one coder step, because the system is not noise-free. At the input to the compressor, the inband rms value of random noise was approximately $\frac{1}{6}$ of a code step during normal operation.

IV. NET LOSS INSTABILITY AND GRANULARITY

The control and measurement of net loss presents two major problems in the system. The first is the enhancement of variations in linear gain occurring between compressor input and expandor output. The second is a result of the encoding process which introduces granularity in the system response to a transmitted sine wave. The latter poses the problem of net loss measurement with a sinusoidal input, a standard testing procedure. Both of these effects are considered in the sections to follow. In addition, a gain variation margin assignment to the major blocks of the transmitting and receiving terminals is made. Finally, net loss measurements made on the terminals are presented.

4.1 *Linear Gain Variation Enhancement*

This effect can be analyzed by determining $G(x)$, the transmission characteristic provided by a back-to-back compressor-expandor com-

bination with an intermediate linear gain perturbation. Using the logarithmic compression curve given by (1) previously, multiplying the output y by $1 + \delta$, where δ is the linear gain variation, and feeding the result into the inverse expandor characteristic, the following expression is obtained for small δ.

$$\begin{aligned} G(x) &= x + \frac{\delta}{\mu}(1 + \mu x)\log(1 + \mu x) \qquad (0 \leqq x \leqq 1) \\ G(-x) &= -G(x). \end{aligned} \tag{18}$$

The first term gives the desired linear transmission and the second term the nonlinearity produced by δ.

The measurement of net loss is generally made with a sinusoidal input. Using the transmission characteristic given by $G(x)$, the recovered fundamental can be determined. This is accomplished by setting $x = E \sin wt$ and computing the fundamental component in the Fourier series at the output. In-band distortion produced by higher harmonics in the response is not considered. Under these conditions, the linear gain variation enhancement, i.e., the factor by which the gain is multiplied, is shown in Fig. 13. This curve gives the enhancement as a function of the amplitude E of the input sine wave normalized to the system overload voltage. Unity input corresponds to full load. The compression parameter μ is taken to be 100 in accordance with the laboratory design. The enhancement will be larger for a larger μ, however.

For signals considerably below full load the enhancement is small. This follows since compressor and expandor transmission are essentially linear in this range. At the other extreme, the enhancement takes on its maximum value and is approximately equal to the slope of the expandor characteristic at full load. This quantity is $\log(1 + \mu) = 4.6$ for $\mu = 100$ and indicates that a variation in linear gain is magnified 4.6 times for full-load signals. It is therefore apparent that considerably more control is required over deviations in linearity between compressor and expandor than for simple gain variations in other parts of the system.

4.2 *Net Loss Granularity*

Granularity of transmitted signal values for sinusoidal inputs is inherent in the PCM process. Measurement of system net loss with a sine wave, therefore, presents a problem independent of any circuit gain instability that may exist. The effect involved can be described by an

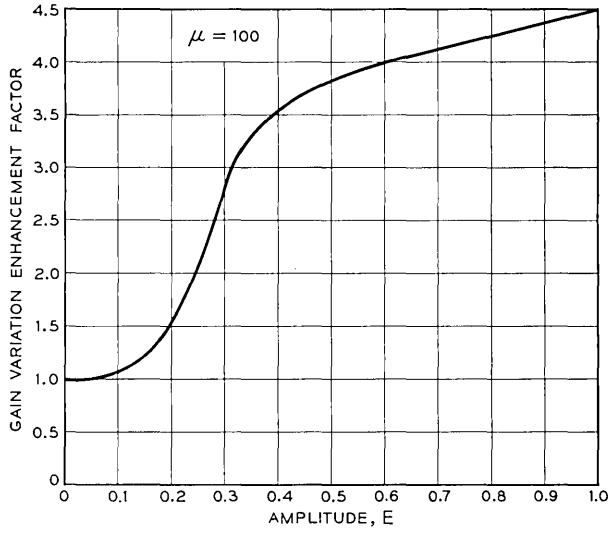


Fig. 13 — Gain variation enhancement factor as a function of the amplitude of the input sinusoid. Unity input corresponds to full load. The compression parameter μ is taken to be 100.

oscillatory gain curve as a function of the amplitude of the input sine wave. The extent of the gain variation is much greater for weak signals than for signals near full load due to quantization. However, values of the input signal may be found in any signal range for which the gain variation vanishes. To minimize the error in the testing procedure, therefore, test tones for net loss measurement should be confined to the vicinity of full load in the neighborhood of one of the nulls.

4.3 *Net Loss Objectives*

As noted earlier (Section 3.1.6), an objective of the experimental design was to provide a 2-db loss trunk. In this section a gain variation margin assignment is made to the major circuit blocks of the transmitting and receiving terminals. The allocation assumes that in the worst case an upper limit of 1.5 db variation from nominal is not to be exceeded. Therefore margins assigned to individual circuits are the maximum allowable assuming in-phase addition of effects from block-to-block.

The portions of the terminal equipment essential to this discussion are shown below with the margins assigned.

Transmitting Terminal	*Receiving Terminal*
Hybrid ± 0.05 db	Decoder ± 0.05 db
Filter and gate ± 0.1 db	Expandor network ± 0.45 db
Compressor preamp ± 0.1 db	Expandor postamp ± 0.1 db
Compressor network ± 0.45 db	Common amplifier ± 0.1 db
Compressor postamp ± 0.05 db	Gate and filter ± 0.1 db
Encoder ± 0.05 db	

The numbers in this table cannot be added directly to obtain the overall 1.5-db requirement. They have been determined so that ±0.55 db is allotted to those circuits which are physically outside of the path between compressor input and expandor output. The remaining ±0.95 db is proportioned among the remaining companding and encoding-decoding circuits. The latter allotment is higher to allow for enhancement effects introduced by the nonlinear transmission characteristics of the compressor and expandor networks.

As mentioned in Section 4.1, an enhancement factor of 4.6 for full-load signals applies when changes in linear gain between compressor and expandor networks occur. Therefore, the *linear* gain variation contributed by circuits located electrically between the compressor and expandor networks must be more accurately controlled. Included are the post-amplifier associated with the compressor network and the encoder and decoder. Allowing 0.5 db of the 0.95 db to this cause results in the 0.05 db allotment specified for each of these circuits. For the compressor and expandor networks themselves, we consider the gain variation produced by mistracking as discussed in Ref. 1. In view of these results, it would appear that the 0.45-db allocation is adequate.

4.4 *Net Loss Measurements*

Net loss measurements are made with an 1100-cycle 0-dbm tone at the 0-db TL. A full load sinusoid is +3 dbm at the 0-db TL. To illustrate the net loss as a function of signal level, the data shown in Fig. 14 were taken on the experimental terminal. The circuit transmission relative to that found for a 0-dbm tone is plotted. The relative measurement accuracy with the oscillator detector combination used for the measurements was approximately ± 0.05 db.

The large oscillations for weak signals show the granularity effect described previously. Fig. 14 shows the influence of a second phenomenon of approximate magnitude of 0.3 db. The broad slopes upon which the cyclic variations are superimposed are a result of compandor mistracking.

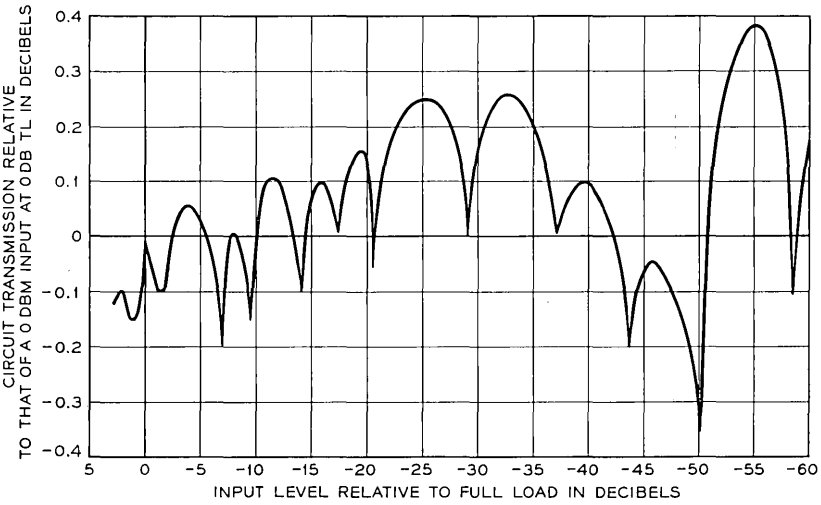


Fig. 14 — Net loss measurements — measured transmission relative to that for a 0-dbm input sinusoid.

V. LOAD CAPACITY

The overload performance of the system is determined primarily by clipping in the encoder-decoder combination. Neglecting quantization this can be represented ideally by the limiter characteristic of Fig. 15.* The following discussion is concerned with signal compression and harmonic distortion as a result of this limiting action. Measurements provided corroborate the analytical discussions.

5.1 *Sine Wave Response*

The system sine wave response is handled most conveniently using the integral representation given below for the characteristic of Fig. 15

$$H(x) = \frac{2}{\pi}\int_0^\infty \frac{\sin \mu \sin \mu x}{\mu^2}\, d\mu. \tag{19}$$

Denoting x by $E \sin wt$, the response becomes

$$H[E \sin wt] = \frac{4}{\pi}\sum_{n=0}^{\infty} I_{2n+1} \sin (2n+1)wt \tag{20}$$

* This neglects any shift in code scale introduced by the encoder. Since the latter is required to be small to reduce second-harmonic distortion as discussed earlier, its effect on overload performance is not considered.

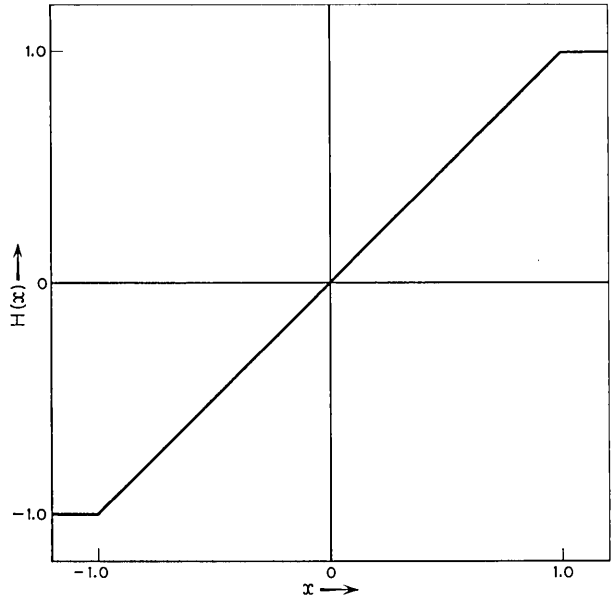


Fig. 15 — Limiting characteristic of system transmission in the absence of quantization.

where

$$I_{2n+1} = \int_0^\infty \frac{\sin \mu J_{2n+1}(\mu E)}{\mu^2}\, d\mu .$$

The J's in the preceding expression are Bessel functions of the first kind. Performing the required integration we get

$$I_1 = \frac{1}{2}\left[E \sin^{-1}\frac{1}{E} + \frac{\sqrt{E^2-1}}{E}\right] \qquad (n = 0)$$

$$I_{2n+1} = \frac{\sqrt{E^2-1}\,\sin\left[(2n+1)\sin^{-1}\frac{1}{E}\right]}{[(2n+1)^2-1]} - \frac{\cos\left[(2n+1)\sin^{-1}\frac{1}{E}\right]}{(2n+1)[(2n+1)^2-1]} \qquad (n > 0). \tag{21}$$

These results apply for signals above the clipping level or $E \geqq 1$, and give the harmonic content of the response. The first and third harmonics

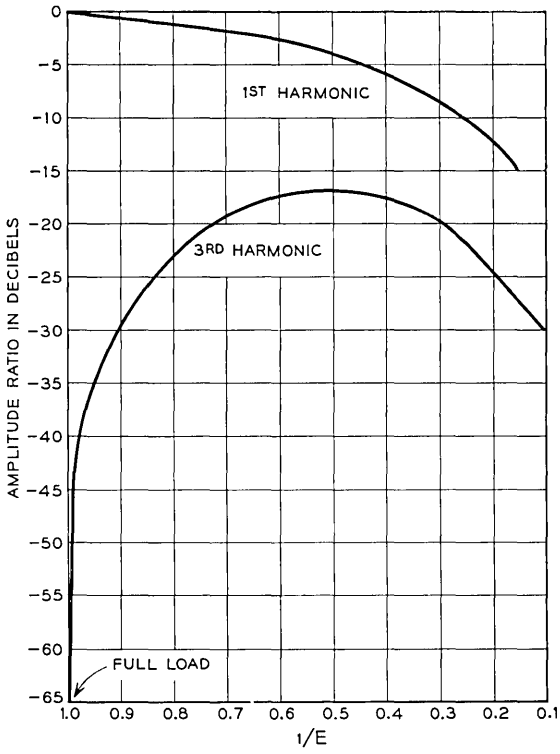


Fig. 16 — Plot of first and third harmonics of limiter response to a sinusoid.

relative to the input amplitude E are shown in Fig. 16 as a function of $1/E$.

5.2 *Measured Overload Performance*

The measured overload signal compression of the system is shown in Fig. 17. Calculated points extracted from the first harmonic response in Fig. 16 are also shown therein, indicating good agreement between calculation and experiment.

In addition to the response at the fundamental, harmonic distortion is also of interest. In this connection, the signal-to-noise ratio for signals above full load was measured. The measurement procedure was the same as indicated in Section 2.3. The results are shown in Fig. 18. Calculated results for the ratio of the first to third harmonic extracted from Fig. 2 are also shown at several points. Only points where clipping is

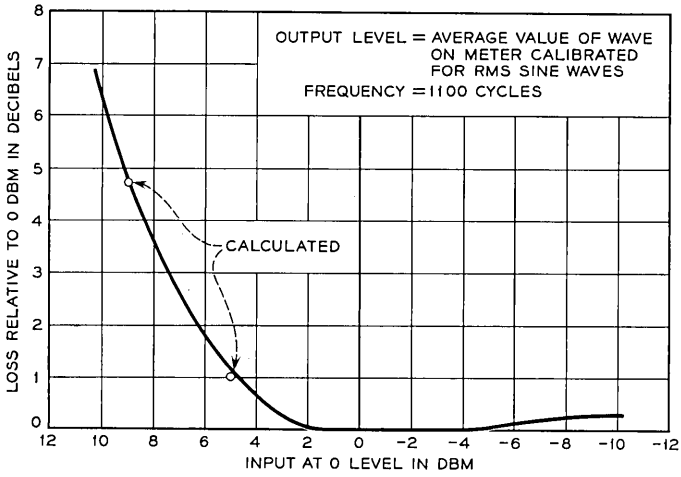


Fig. 17 — Measured signal compression vs input signal level.

appreciable are included, so that a valid comparison without quantizing impairment can be made. Agreement is good, thus indicating that only third harmonic distortion is appreciable.

VI. CONCLUSIONS

Four areas of system performance have been discussed. The measured performance has been compared with anticipated theoretical predictions. In all cases the system appears to be acceptable. No important or unex-

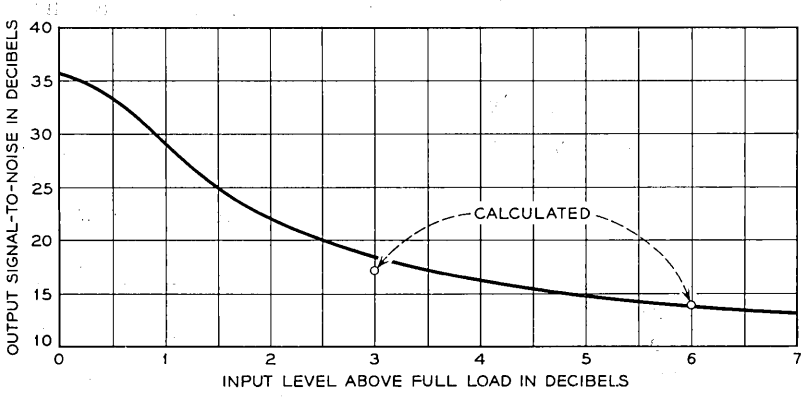


Fig. 18 — Measured signal-to-noise ratio vs input signal level for signals above full load.

plained differences between theoretical and actual performance have been discovered.

VII. ACKNOWLEDGMENTS

As in any corporate development, many members of Bell Laboratories have made significant contributions to the terminal design and evaluation programs. Particular mention should be made of the work of C. G. Davis on early phases of the system planning and functional arrangements, and of the work of D. J. Leonard in the design and execution of experiments for the system evaluation.

REFERENCES

1. Mann, H., Straube, H. M., and Villars, C. P., this issue p. 173.
2. Smith, B., B.S.T.J., **36,** May, 1957, p. 653.
3. Levin, G. A., and Golovichner, M. M., Radiotechnika SSRV V. **10,** No. 8, 1955, p. 3.
4. Bennett, W. R., B.S.T.J., **27,** July, 1948, p. 446.
5. Bennett, W. R., B.S.T.J., **37,** Nov., 1958, p. 1501.
6. Davis, C. G., this issue, p. 1.

A Companded Coder for an Experimental PCM Terminal

By H. MANN, H. M. STRAUBE, and C. P. VILLARS

(Manuscript received July 12, 1961)

The heart of the terminal for an experimental PCM system developed at Bell Laboratories is the "companded coder," which consists of a logarithmic instantaneous compandor plus a linear (equal-step) coder. The companded coder plays a major role in determining the over-all performance of the $1\frac{1}{2}$ megabit system. This paper includes a discussion of the fundamental design concepts, the practical realization and the performance of the all-semiconductor equipment that performs the analog-to-digital and digital-to-analog conversions for the PCM terminal. The instantaneous compandor employs matched semiconductor diodes to obtain a logarithmic gain characteristic. A 7-digit network type coder performs the conversion of the signal from analog to digital form and vice versa.

I. INTRODUCTION

An earlier paper[1] has described the framework of an experimental pulse code modulation (PCM) system for short-haul telephone trunks. The system consists of two 24-channel PCM terminals interconnected by two regenerative-repeatered lines. Within each terminal are compressor-encoder and decoder-expandor functional blocks that constitute a "companded coder system." The latter all-semiconductor system, which performs analog-to-digital and digital-to-analog conversions, is the subject of this paper.

The compressor-encoder block, in the transmitting portion of each terminal, accepts 7.4-microsecond bipolar sample pulses from the multiplex circuits. During a particular 4.5 microseconds of each pulse duration, the amplitude of the sample pulse is accurately compressed in accordance with a desired law, measured to the nearest half-quantum step out of 128, and converted to a pattern of seven $\frac{1}{3}$-microsecond on-off pulses in accordance with 7-digit binary notation. A succession of the latter pulse patterns, constituting a $1\frac{1}{2}$ megabit-per-second PCM signal, is finally transmitted to another terminal.

The decoder-expandor block, in the receiving portion of each terminal, accepts transmitted PCM, decodes it, and expands it by processes that are essentially inverse to those in the compressor-encoder block. It delivers a train of 3.2-microsecond bipolar pulses to the demultiplexing circuits, each pulse of which bears a linear amplitude relationship to its 7.4-microsecond mate at the sending end.

A discussion of fundamental design concepts is first presented, including such subjects as quantizing distortion, companded coding, coding techniques, companding characteristic, overload level, volume range and number of digits. The system realization is then described in quite some detail as to plan, circuits, and performance. A brief summary concludes the paper.

II. FUNDAMENTAL DESIGN CONCEPTS

2.1 *Quantizing Distortion*

It is well known that an analog signal will be inherently distorted when it is converted to PCM. Such distortion is the inevitable result of approximating sample amplitudes, which may take on an infinite number of values, by a finite number of codes. The integrated effect of such errors constitutes so-called "quantizing distortion" or "quantizing noise". Fortunately, such distortion can be made acceptably small if the chosen number and distribution of quantum steps are consistent with the volume range and statistics of the signal.

2.2 *Companded Coding*

If a moderate number of equal steps are used to quantize a given range of signals, the weakest signals experience the most serious quantizing distortion. As shown in Fig. 1(a), for a weak sample that traverses only a single quantum step, a half-step error amounts to about a 50 per cent error! This is indeed serious, but can be alleviated by resorting to one of several alternatives that, in effect, provide more steps for the weak signals.

Increasing the total number of equal steps will reduce the quantizing distortion, but requires more accurate coder decisions in less time, and increased bandwidth in the repeatered transmission line.

A second alternative is to taper the size of the steps (nonlinear coding) over the signal range in such a way that weak signals traverse their fair share of steps. However, this solution, although avoiding the

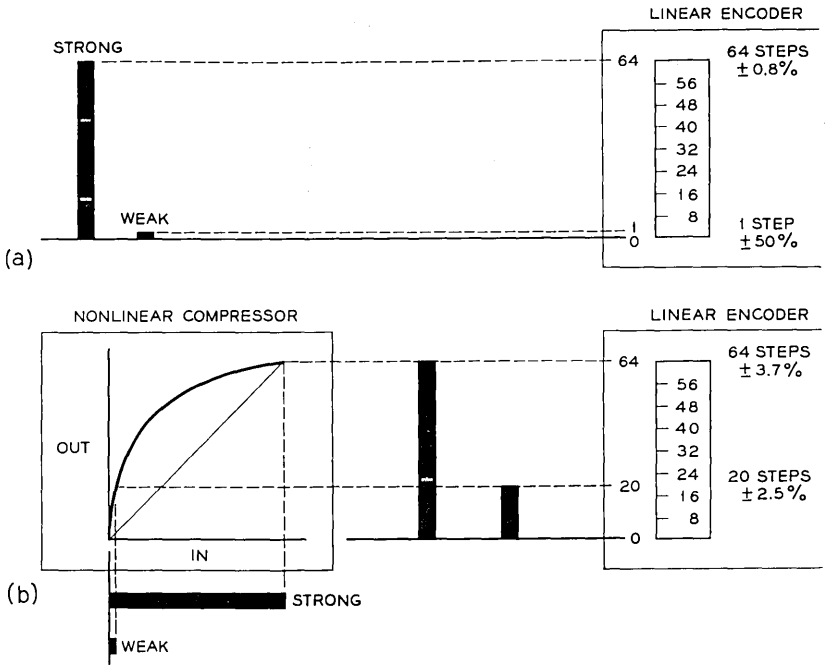


Fig. 1 — Reduction of quantizing distortion.

speed and bandwidth problems mentioned above, still poses a very serious problem in coding accuracy for the smallest steps.

The third alternative is to taper the signal in the manner shown in Fig. 1(b), and thereby spread weak signals over a considerable number of quantum steps. This allows the use of an encoder with moderate speed and accuracy, at the expense of a precise, but feasible, compressor circuit. Of course, it follows that a signal that has been compressed in transmission must be expanded in reception if the over-all transmission is to be linear. Accordingly, if a linear decoder is used in reception, it must be followed by an expandor circuit having an inverse characteristic to the compressor. This method, utilizing preferential amplification of weak signals prior to linear encoding and preferential attenuation of weak signals after linear decoding, is applied in the experimental system to be described.

It is significant that the latter compressor and expandor must respond to instantaneous values of very short pulses. They therefore constitute

an "instantaneous compandor", as opposed to the slow-acting "syllabic compandor" used per channel in some AM frequency multiplex systems.

2.3 *Companding Characteristic*

A modified logarithmic characteristic of the type defined in Fig. 2 has been found to be desirable when the message is speech.[2] Through its use, quantizing distortion may be reduced to an acceptable value for weak signals, with an acceptable impairment for strong signals.

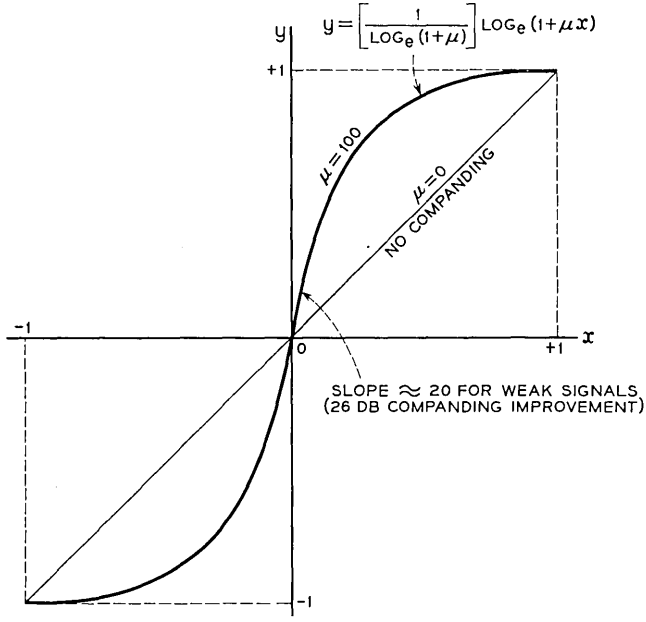


Fig. 2 — Desirable companding characteristic.

In the case of a compressor, x represents the input and y the output, while for an expandor, y represents the input and x the output. Parameter μ determines the degree of compression (or expansion) and hence the amount of companding improvement (the change in the signal-to-quantizing noise power ratio relative to a noncompanded quantizer) for weak signals. Fig. 3 shows how the signal-to-quantizing noise ratio varies with relative signal power when perfect $\mu = 0$, 50, 100, or 200 companding is combined with perfect seven-digit coding.*

The optimum value of μ depends on many considerations, as discussed below.

* This datum is calculated from (39) in Ref. 2.

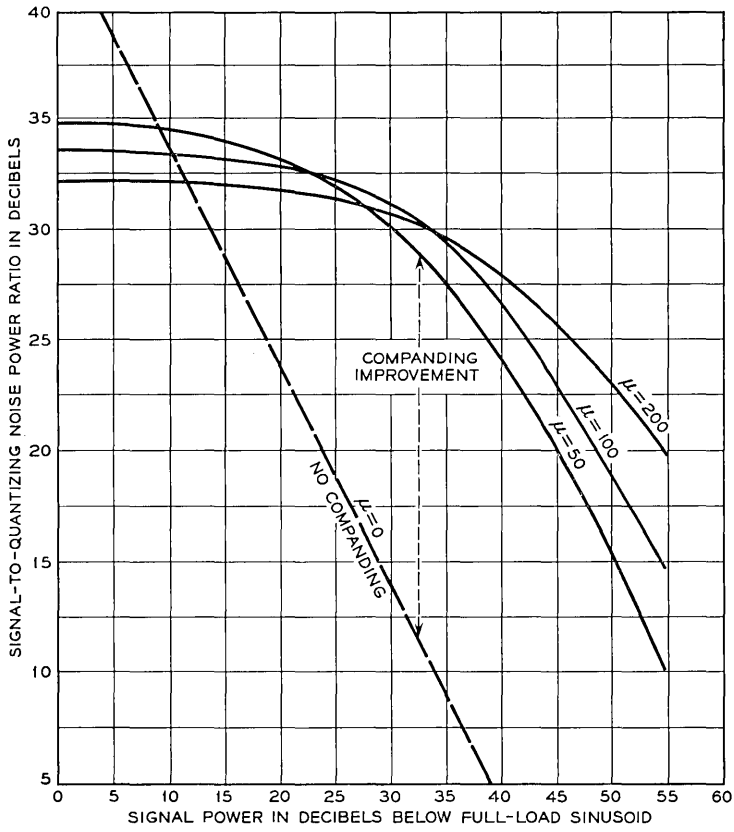


Fig. 3 — Quantizing noise performance of the logarithmic companding characteristic for $\mu = 50$, 100 and 200 (assuming perfect seven-digit linear encoding).

Considerations that encourage high values of μ are:

1. The desire to obtain large companding improvement for weak signals.

2. The desire to reduce idle circuit noise and interchannel crosstalk caused by irregular excitation of weak-signal quantum steps.[3] When an idle channel is biased at or near the boundary between weak-signal quantum steps, a fractional-step perturbation will produce a full-step output.

3. The desire to maintain a high system overload value relative to the weak talker volume so as to minimize clipping of large signals. If the overload amplitude is doubled, the signal volume range (and hence the size of each quantum step) is doubled. The adverse effect this has

on weak-signal quantizing distortion, as well as on idle circuit noise and interchannel crosstalk, can be offset by more extreme companding (higher μ).

Considerations that discourage high values of μ are:

4. The difficulty of achieving sufficient stability in system net loss for high-level signals. Although the expandor lends preferential attenuation to weak signals, it also lends preferential gain to strong signals. It therefore enhances any strong signal gain or loss changes that may occur in the compressor-encoder-decoder-expandor transmission path. The strong signal enhancement factor is approximately 4.6 for $\mu = 100$ and rises with increasing μ.

5. The difficulty of achieving and maintaining satisfactory "tracking" (true inversity) between compressor and expandor. At high signal levels mistracking is exaggerated by the enhancement factor mentioned in 4. At medium and low signal levels "mistracking" is aggravated by deficiences in reproducibility and stability that normally plague devices with greater non-linearity (higher μ).

6. The difficulty of achieving sufficient bandwidths in the compandor networks. Greater nonlinearity (higher μ) implies a greater ratio between the low-level and high-level impedances of the network. Difficulty is experienced in keeping the higher of these impedances low enough to yield sufficient bandwidth in the presence of normal stray capacitance.

7. The difficulty of holding the dc component of the multiplexed signals to a value low enough for full exploitation of high μ.[2]

Little additional companding improvement is realized for μ greater than about 100 unless the dc component of the multiplexed signals is well below 1 per cent of the overload amplitude. The latter is not easy to guarantee in view of practical deficiencies in sampling gates, encoder reference, etc.

A choice of $\mu = 100$, corresponding to a weak-signal companding improvement of almost 26 db, was made for the experimental compandor. This choice was largely dictated by the above practical considerations 4 through 7, and was only broadly influenced by the first three considerations that involve overload level, volume range, and number of digits.

2.4 *Coding Techniques*

2.4.1 *Nonlinear Coding*

It is possible to introduce the desired nonlinear characteristic of the compandor in the coder.[4] In such an implementation the coder var-

ies its step size as a function of the amplitude of the signal to be encoded or the code to be decoded.

The functional relationship between the analog signal and its associated code may take a variety of forms. The choice of the coding characteristic is a function of the statistics of the signal to be coded, the signal-to-quantizing noise power ratio desired, technical feasibility and economics.

The nonlinear characteristic required for the experimental PCM system is a logarithmic one, whereas early proposals for nonlinear coding led to a hyperbolic one. It is possible to generate the logarithmic coding characteristic with a network coder.[5]

Hyperbolic, logarithmic and other types of coding characteristics can also be approximated by a piecewise linear coding process.[5,6] A logarithmic coding characteristic, for example, is approximated by a number of straight lines of varied slopes, the slope of each line segment resulting from a different coder step size. The transition points at which the coder step size is altered are chosen to best approximate the desired logarithmic curve.

A comparison of the theoretical signal to quantizing noise power ratio versus signal power for a number of such characteristics is shown in Fig. 4. It is apparent that the piecewise linear coders will suffer a noise penalty when compared with a companded-coder system or a nonlinear coder with a logarithmic characteristic. The hyperbolic coder suffers a noise penalty only for the very loud talker.

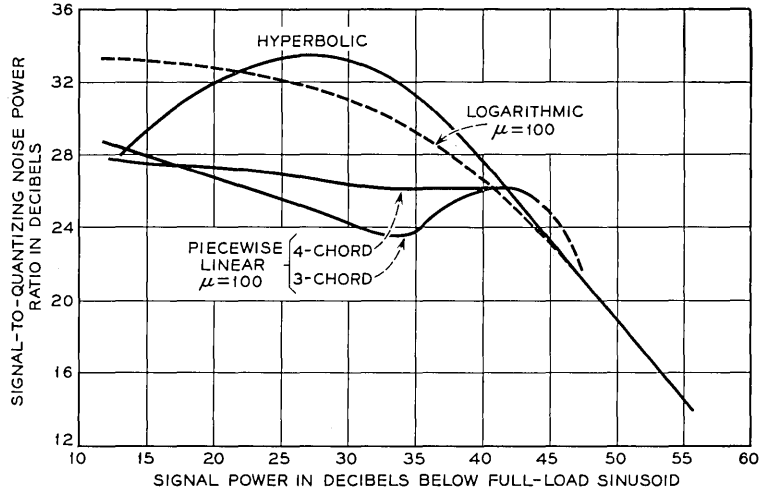


Fig. 4 — Quantizing noise performance of the hyperbolic, piecewise linear, and logarithmic companding characteristics.

The main problem encountered in nonlinear coding is the accuracy in the production of the steps. For linear coding with seven digits, the ratio of step size to the peak-to-peak signal is obviously 1 to 128. For nonlinear coding with the desired compression characteristic, this ratio is of the order of 1:2560 for the smallest steps. For these and other reasons relating to cost and to the performance characteristic of these coders, nonlinear coding was not deemed suitable for the experimental PCM system.

2.4.2 *Linear Encoding*

Three basic encoding techniques may be considered: (1) step-at-a-time,[7] (2) digit-at-a-time,[4,8] and (3) word-at-a-time.[9,10]

In the "step-at-a-time" encoder, each input sample is measured by determining which quantum steps, one by one, are contained in its value. In an n-digit system, this can require 2^n counts or decisions per input sample but storage of only one piece of information (single step value).

In the "digit-at-a-time" encoder, each input sample pulse is measured by determining which binary digits, one by one, are contained in its value. This requires n decisions per sample and storage of n pieces of information (digit values).

In the "word-at-a-time" encoder, each input sample pulse is measured by determining which word (code combination) represents its value. This requires only one decision per sample but storage of 2^n pieces of information (word values).

Which technique is least expensive depends on the relative costs of speed and storage. In the present application, wherein approximately 4.5 microseconds are available to encode each sample, the times allowed per decision may be shown to be about 0.035 microsecond for technique (1), 0.65 microsecond for technique (2), and 4.5 microseconds for technique (3). Although semiconductor devices may soon be available to implement technique (1), the speed and cost of present commercial semiconductor devices is more in keeping with technique (2). Technique (3) is discouraged by its storage expense, which far offsets the fact that it can tolerate slower devices. Accordingly, the "digit-at-a-time" technique was chosen for use in the present experimental system.

2.5 *Overload Level and Volume Range*

The quantized amplitude range was chosen to accommodate peak-to-peak excursions of a +3-dbm sinusoid at the two-wire input to the system.

The volume range, from the weakest signal to the overload level given above, was defined as the 50-db range between −47 dbm and +3 dbm at the system two-wire input. This range includes more than 99 per cent of the talker volumes experienced in exchange applications. Of course, talker volumes somewhat above and below these limits will be transmitted, but they will be expected to suffer more distortion.

2.6 *Number of Digits*

Once the degree of companding, overload level and volume range are defined, the number of digits required for an acceptable signal-to-quantizing noise ratio can be determined.

No absolute standards have been established for the minimum acceptable signal-to-quantizing noise ratio. The best available estimate of the least number of digits, consistent with the probable subjective acceptability (based on computer simulation of companded-coder systems) and the state of transistor circuit art, led to the choice of seven digits.[11]

III. SYSTEM PLAN

3.1 *System Block Diagram*

Fig. 5 shows a block diagram of the 24-channel exploratory PCM system. For simplicity, only those elements used for transmission in one direction are indicated, but it will be understood that two-way conversations may be provided by appropriate duplication of the equipment shown.* Transmission (from left to right) is seen to include operations of sample collection, compression, encoding, regenerative repeating, decoding, expansion, and sample distribution. The compressor-encoder and decoder-expandor blocks comprise the companded coding system with which this paper is primarily concerned.

Note that two multiplex busses are used in the transmitting portion of a terminal, one for odd-numbered channels and the other for even-numbered channels. The odd- and even-numbered signals are interleaved in time in the encoder by a transfer switch, which in effect dwells on each bus an appropriate half the time. This arrangement delivers full-length samples to the encoder (allowing maximum decision time) yet provides full-length guard spaces between samples (yielding low intersymbol interference with moderate PAM bandwidth).

* For two-way conversations, each terminal is provided with both transmitting and receiving equipment, and a second transmission line (with oppositely directed transmission) is included between terminals.

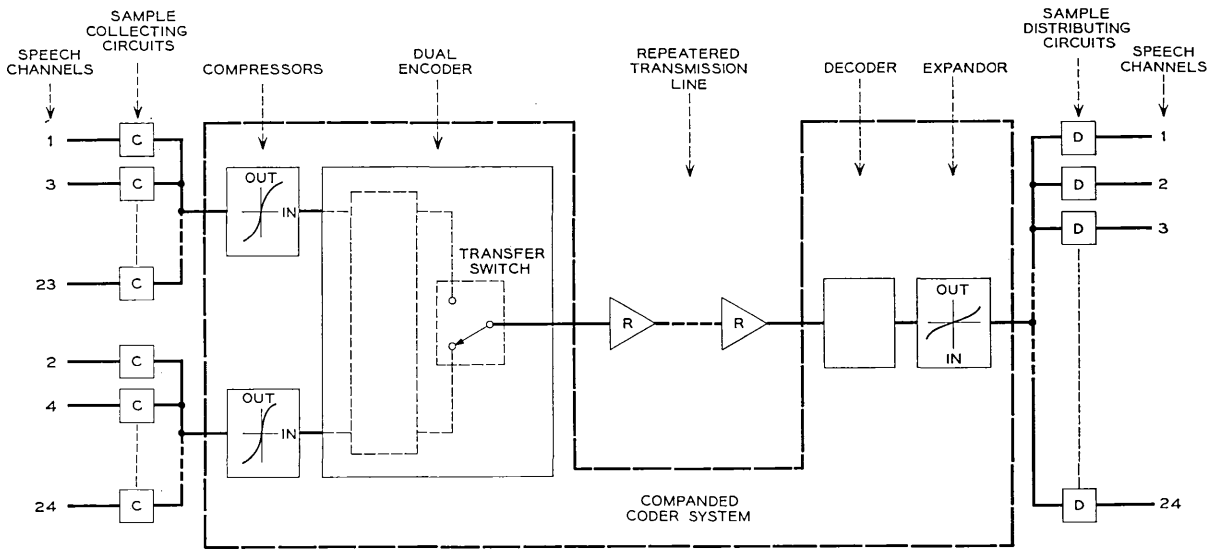


Fig. 5 — Companded coder system for the experimental PCM system.

In the receiving portion of a terminal, there is no particular need for full-length, flat-topped samples. A single receiving PAM channel of moderate bandwidth can therefore be realized by a judicious trade between sample length and guard space.

3.2 *System Time Intervals*

A better understanding of "odd-even" multiplexing, as well as other companded coding problems, can be gained through detailed consideration of system time intervals. Pertinent is the fact that signaling and framing information, as well as speech, is transmitted from terminal to terminal via on-off pulses. There are 8000 frames per second (125 microseconds per frame) and 193 digit-slots per frame (approximately 0.648 microsecond per digit-slot). This provides for information transmission at the rate of $8000 \times 193 = 1.544$ megabits per second. Each channel-slot (approximately 5.18 microseconds) contains eight digit-slots, one for signaling and seven for pulse-code-modulation speech. The 193rd digit-slot in each frame is reserved for framing information. The time slots (time intervals) reserved for each purpoes are defined in Fig. 6 where the timing pattern for a sample at the transmitter and receiver is shown.

The digit-at-a-time encoder, which will be described later, requires an input sample that is flat-topped over at least seven digit-slots. If a single multiplex bus is used, it is apparent from Fig. 6 that the time available to build up and decay such a sample will be only one digit-slot, that is, the last digit-slot of each channel-slot. However, if "odd-even" multiplexing is used, wherein the transients on one bus are allowed to dissipate while the encoder is acting on the other bus, a total of nine digit-slots (the eighth digit-slot plus one channel-slot) are ideally available for build-up and decay. Thus approximately a two-fold increase in equipment yields about a nine-fold decrease in bandwidth required per equipment. The present "odd-even" embodiment requires only about 0.6-megacycle bandwidth, as compared with the 5-megacycle or so bandwidth that would be required for single-bus multiplexing.

In the receiving portion of a terminal full-length, flat-topped samples are not a necessity. Accordingly, the width of a decoded sample (including its build-up transient) is intentionally reduced to $\frac{5}{8}$ channel-slot so as to allow $\frac{3}{8}$ channel-slot for decay. Under these conditions a single-bus demultiplexer requires only about 0.8-megacycle bandwidth and is therefore quite feasible.

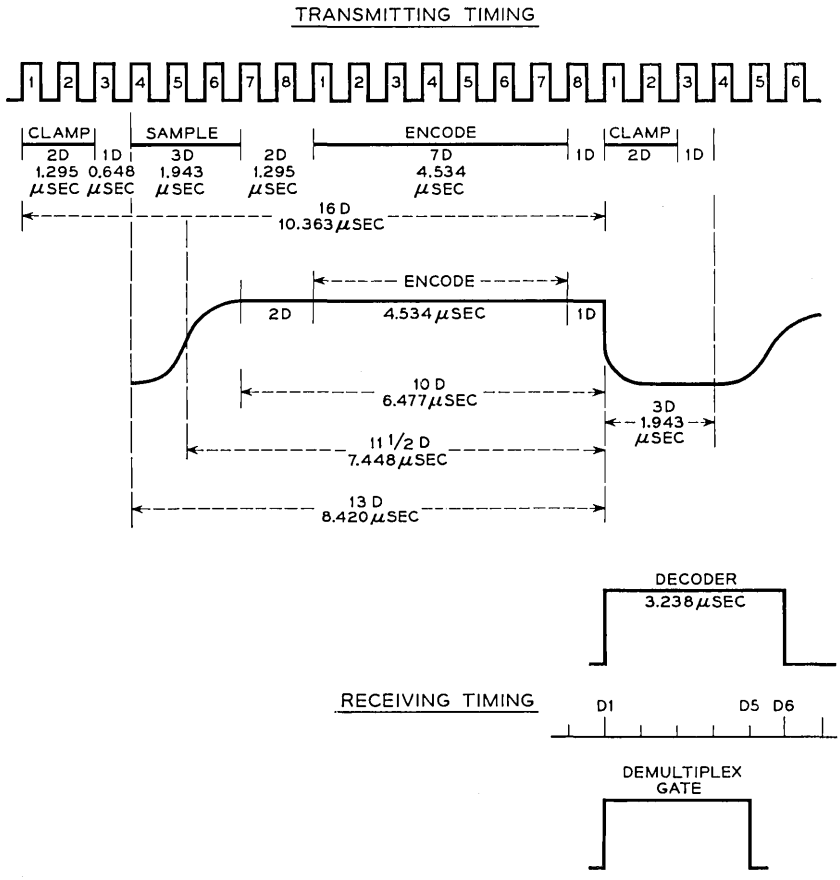


Fig. 6 — Timing pattern for a sample at the transmitter and receiver of the experimental PCM system.

IV. INSTANTANEOUS COMPANDOR

4.1 *Compandor Block Diagram*

The circuit blocks that constitute a compressor or expandor are shown in Fig. 7. In either case, a nonlinear network is connected in tandem with appropriate transistor amplifiers. The networks provide the desired compression or expansion characteristic, and the amplifiers simply adjust the power and impedance levels to appropriate values.

Note that in the compressor some 26 db more *gain* is provided for weak signals than for strong signals. In the expandor the inverse situa-

tion exists; that is, some 26 db more *loss* is provided for weak signals than for strong signals.

Not only must the proper gain be realized at each signal level, but also that gain must be held (regardless of ambient temperature and time) to a stability consistent with over-all system net loss and harmonic distortion requirements. As indicated in Fig. 7, the stability requirement is particularly severe for high-level signals. This is the result of the preferential gain provided by the expandor for strong signals. If the circuit blocks are assumed to have equal and correlated instabilities, not more than about ±0.05-db change per block may be tolerated. This encourages the use of negative-feedback amplifiers and temperature-controlled highly stable networks. Since readjustment of these circuits to the required accuracy is not within the scope of simple field procedures, at least a 20-year-stable design is desirable.

As shown in Fig. 6, the compressor and expandor input signals comprise pulses and guard spaces a few microseconds in duration. Considering the encoder's need for a flat-topped sample, and assuming a 74-db crosstalk loss requirement, one may show that it is necessary to have flat transmission over the frequency bands indicated in Fig. 7, that is,

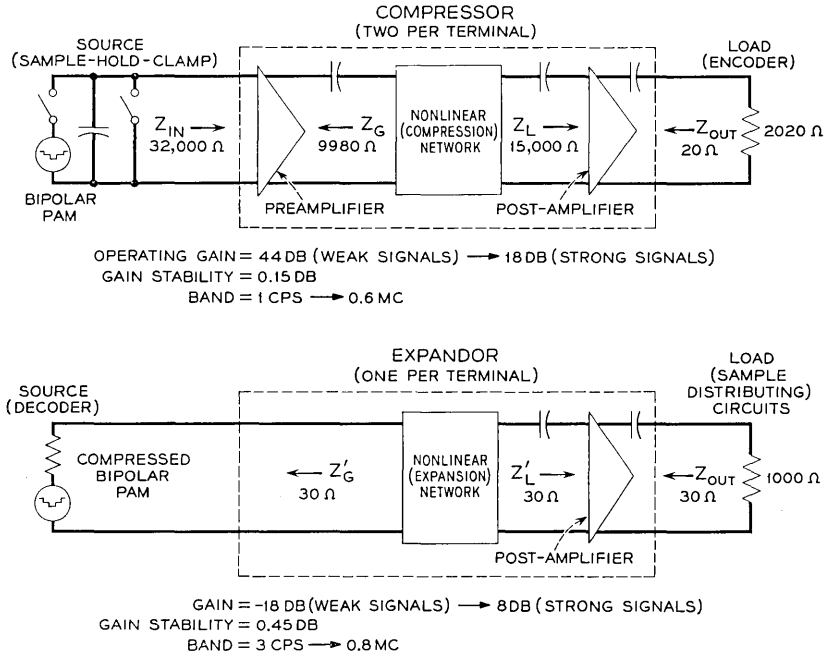


Fig. 7 — Block diagrams showing compressor and expandor.

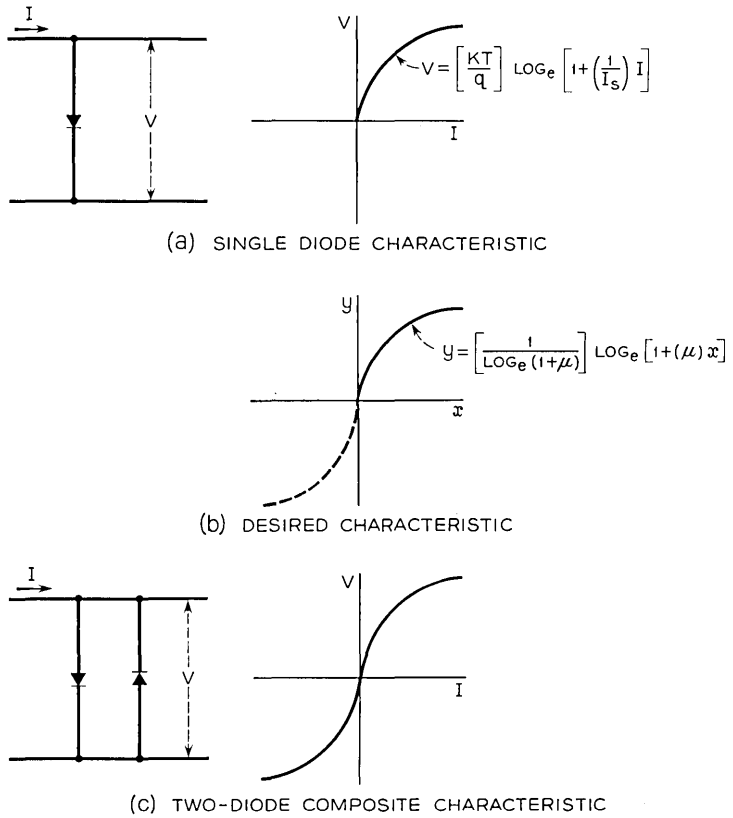


Fig. 8 — Smooth curve approximation to the logarithmic companding characteristic via the voltage-current characteristic of a semiconductor diode.

substantially from a cycle per second to nearly a megacycle per second.* In the compressor, the upper frequency cutoff is dictated by the rise time requirements of the encoder. The lower frequency cutoff is dictated by the crosstalk requirement.[12]

4.2 *Method of Achieving Nonlinear Characteristic*

The desired companding characteristic of Fig. 2 may be approached by exploiting the inherent voltage-current characteristics of semiconductor diodes. A single diode tends to provide a first quadrant approximation, as shown in Fig. 8(a). Note that the voltage-current (approximate) expression for the semiconductor junction in (a) has the same

* It is the "full-load" upper cutoff frequency that is shown in Fig. 7.

form as the equation of the desired theoretical curve in (b). Two diodes, paralleled with the polarity indicated in (c), tend to provide the desired approximation in both first and third quadrants. In a given quadrant, the composite characteristic is substantially the same as for a single diode, except at low currents where the composite characteristic becomes more nearly linear. More specifically, the voltage across the composite tends to become an inverse hyperbolic sine function of the current, rather than a logarithmic function. Although this represents a slight departure from the nominally desired characteristic, it is fortunately in a direction to reduce network sensitivity to small dc changes.

4.3 *Practical Companding Networks (Seven-Point Fit)*

Since the voltage-current characteristics of practical diodes are never identical from unit-to-unit, it is helpful to provide trimming resistors in series and parallel with the diodes, as shown in Fig. 9. Initial pairing of diodes at medium currents plus appropriate adjustments of R1, R2,

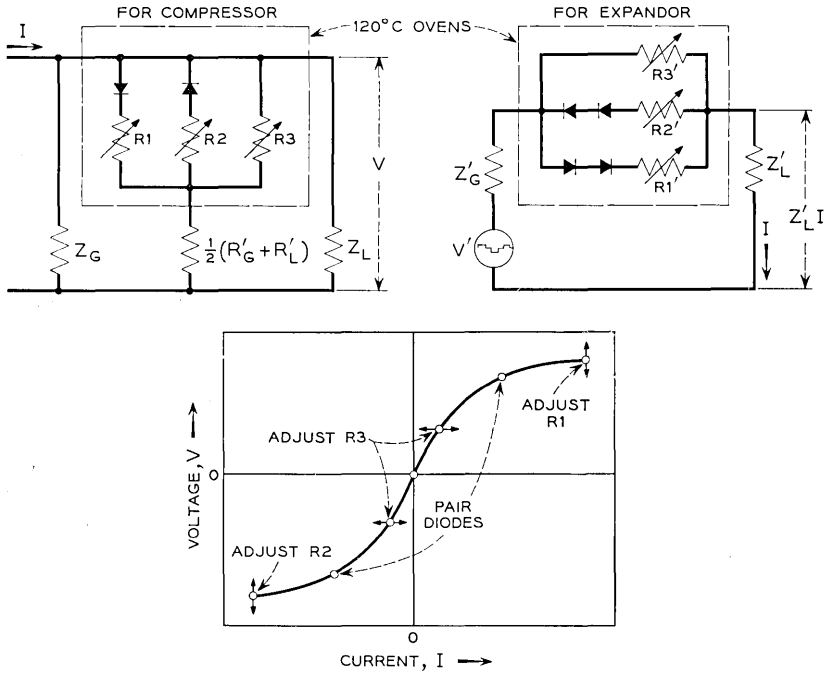


Fig. 9 — Practical companding network (seven-point fit).

and R3 (or R1′, R2′, and R3′) facilitate a seven-point fit to a "standard" curve, as suggested at the bottom of the illustration. A compression network results (at the left) when the network is operated in shunt with a high-impedance generator Z_G and a high-impedance load Z_L. The inverse or expansion network results (at the right) when the network is operated in series with a low-impedance generator Z'_G and a low-impedance load Z'_L.

It is clear that the addition of shunt and series resistances to the diode pair will tend to linearize the characteristic and so make the "standard" curve mentioned above depart from the nominally desired law. However, this addition yields a number of advantages that are of major importance in practice. Specifically:

1. It allows initial differences between diodes to be partially compensated by appropriate choice of the resistors (as already mentioned).

2. It allows the impedance of the generator (output impedance of a transistor preamplifier) and load (input impedance of a transistor postamplifier) to have reasonable and realizable values. Shunt resistors avoid the need of infinite-impedance generator and load for the compression network. Series resistors avoid the need of zero impedance generator and load for the expansion network.

3. It allows the direct and stray capacitances of the diodes and their interconnecting configuration to be reasonably high without sacrifice of the bandwidth required. This follows to the extent that the added resistors lower the shunt resistance more than they raise the shunt capacitance.

4. It provides masking of those changes that will inevitably occur in the diodes (due to temperature variation and aging). This effect is particularly significant in the case of the shunt resistors which in present circuitry provide masking factors of about 6 at low currents, thereby reducing a $1\frac{1}{2}$ per cent change in diode characteristic to $\frac{1}{4}$ per cent change in network characteristic.

5. It reduces network sensitivity to those dc components that will inevitably occur in the multiplexed signals. This follows from the fact that the resistance-masked characteristic is more nearly linear near the origin than the theoretical "$\mu = 100$" characteristic.

In the present circuits the latter advantages are obtained at the expense of about 3-db impairment in system signal-to-noise performance, due primarily to resistor-induced misfit of the nominally desired "$\mu = 100$" curve. Fortunately, as shown in Fig. 10, this impairment occurs only in the region of medium-strong signals where the "$\mu = 100$" signal-to-noise performance is well above the over-all system requirement. At low signal levels, where adequate companding is most vital, the result-

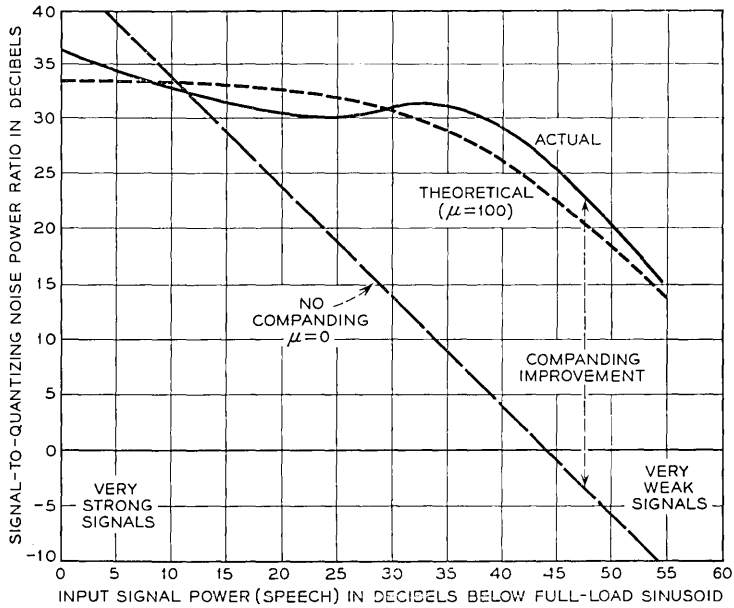


Fig. 10 — Quantizing noise performance of the companding characteristic (assuming perfect seven-digit encoding).

ant signal-to-noise performance is somewhat better than would obtain with a true "$\mu = 100$" characteristic.

4.4 *The Standard Curve*

Typical voltage-current plots for individual and paired diodes, as well as for a completed network, are shown in Fig. 11. The addition of shunt and series resistors, as described in the previous section, accounts for the difference between paired diodes and completed network. The standard curve so approached is defined by the fact that it passes through the origin and exhibits specified ratios between voltages that occur at high, medium and low currents. Specifically (for $\mu = 100$):

$$\text{When I} = 2000\mu\text{a, V} = 5.69\ \text{V}_0$$

$$\text{When I} = 200\mu\text{a, V} = 3.16\ \text{V}_0$$

$$\text{When I} = 20\mu\text{a, V} = \text{V}_0$$

It should be understood that the above equations apply to both the first and third quadrant plots of the network characteristic.

Note that the standard curve is defined in terms of voltage ratios

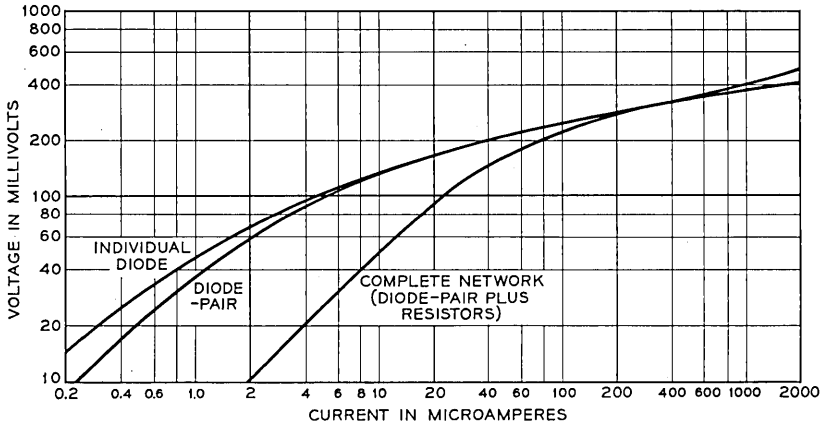


Fig. 11 — Compandor voltage vs current characteristics (diodes at 120°C, resistors at 25°C).

rather than specified voltages. On this basis, networks tend to follow the same law, rather than have identical voltages at specified currents. Absolute voltage differences that occur between networks are finally compensated by way of simple gain adjustments in the associated pre- and post-amplifiers. The over-all effect is to provide excellent tracking between each and every compressor and expandor.

4.5 *Mistracking Considerations*

The preceding sections have tended to concentrate on the mathematical law that a compressor or expandor should obey. It has been pointed out that a "standard" curve, which differs somewhat from the nominal "$\mu = 100$" curve, is quite adequate. However, it should be noted that once a standard curve has been chosen, great care must be taken to reproduce all points on it in every compressor and expandor. Clearly, if either the compressor or the expandor deviates from the chosen standard, the transfer relation between compressor input and expandor output will be nonlinear. Such "mistracking" will result in variation of system net loss with signal level, and in reduction of system signal-to-noise ratio because of increased quantizing and harmonic distortions. Obviously these effects must be controlled to an extent that renders acceptable system performance.

What causes mistracking? Within the networks, the following factors are significant (listed in order of decreasing magnitude):

1. Initial misfit to the standard curve in regions between the seven key points, due to differences between diodes.

2. Initial misfit to the standard curve at the seven key points, due to imperfect adjustment during manufacture.

3. Eventual misfit to the standard curve, due to diode changes with time (aging).

4. Initial misfit to the standard curve in regions between the seven key points, due to the fact that the practical compression and expansion networks are inherently not *exact* inverse configurations.

5. Eventual misfit to the standard curve due to diode changes with temperature variation.

6. Eventual misfit to the standard curve due to resistor changes with time (aging).

7. Eventual misfit to the standard curve due to resistor changes with temperature variation.

Outside the networks, the following effects are pertinent:

8. Misbiasing of the networks or the equivalent, due to a dc component induced in the signal by imperfect sampling, encoding, and decoding, and/or leaky coupling capacitors.

9. Nonlinearity in the transmission characteristics of the compandor amplifiers and the coder (nonuniform step size).

10. Changes in the gain or loss of the compandor amplifiers and the coder (nonconstant reference voltage).

4.6 *Network Tolerances*

Fig. 12 shows what are believed to be feasible tolerances for each of the seven network imperfections itemized above. These results are based on detailed studies of each factor, considering human and instrument errors, as well as diode and resistor statistics, temperature coefficients, and estimated twenty-year aging characteristics. Due account has been taken of the resistor masking effects mentioned in Section 4.3.

Note that "initial diode differences" is the major cause of initial misfit to the standard curve. If desired, this particular tolerance can be made comparable to the "initial adjustment" and "diode aging" tolerances by rejecting about 20 per cent of the manufactured diodes.

"Diode aging" and "diode temperature" are seen to be the most serious time-variant tolerances. Although both are well under control, special measures were necessary to achieve this result, as discussed below.

4.7 *Diode and Network Instability*

A rather extensive one-year measurement program yielded the estimates shown in Fig. 13 for the maximum per cent change in voltage

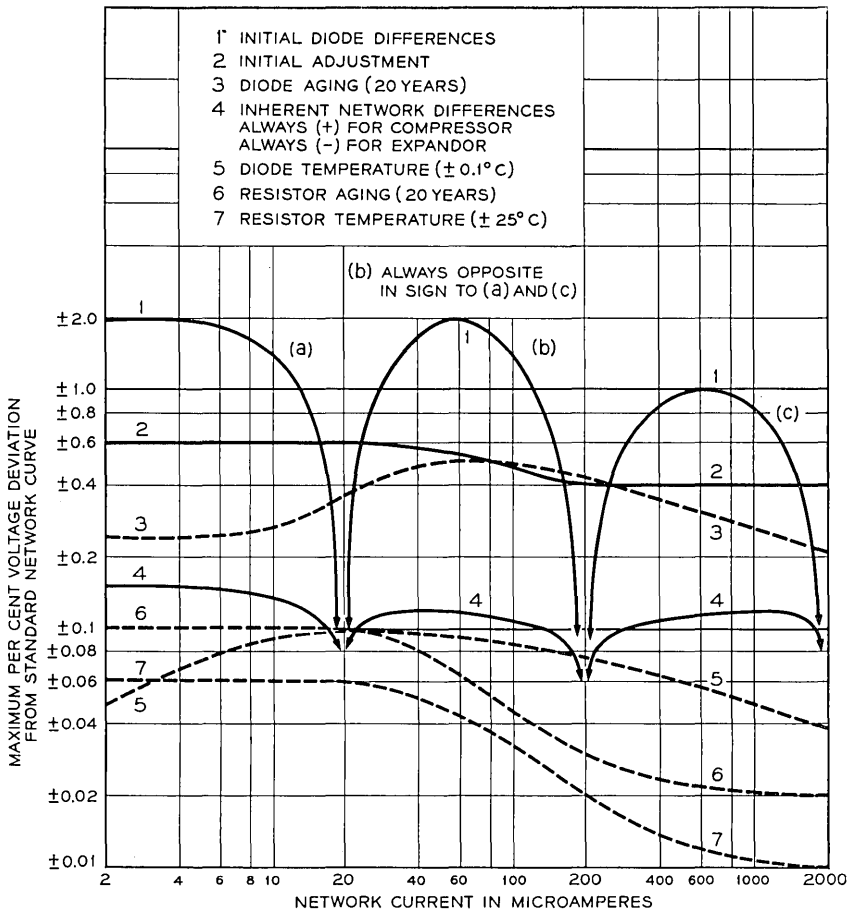


Fig. 12 — Compandor network tolerances that contribute to mistracking.

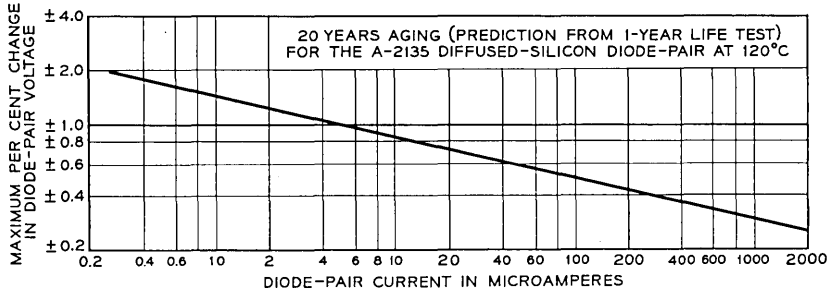


Fig. 13 — Compandor diode-pair aging characteristic.

across a diode pair that will occur during a twenty-year period. The per cent aging is seen to be an inverse function of current, at least within the range of currents shown. These results apply to a specific small-area, diffused-silicon diode operating at 120°C (Bell Laboratories Type A-2135). Several other types of diodes have been investigated, but most of them have yielded at least an order of magnitude poorer performance.*

Fig. 14 shows the per cent change in voltage across a diode pair per 0.1°C change in diode temperature. Here again, the per cent change is seen to be an inverse function of current.

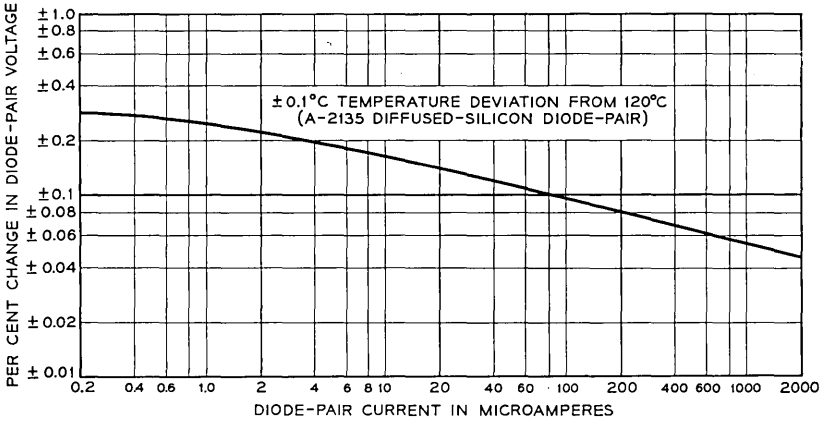


Fig. 14 — Compandor diode temperature characteristic.

Fortunately, the above variations are reduced to acceptable magnitudes by the masking effect of the network resistors. This assumes, of course, that the resistors are appreciably more stable than the diodes.† The factor by which diode-pair variations are reduced by masking is shown in Fig. 15, as derived from Fig. 11. Fortunately, masking is most effective at low currents where the per cent change in diode pairs is most severe. Application of the masking factor to Figs. 13 and 14 then yields the "diode aging" and "diode temperature" network tolerances shown in Fig. 12. Despite the advantages of resistor masking, good net-

* Certain point-contact or gold-bonded diodes will yield the desired curve with acceptable low capacitance, but they are unstable. Alloy junction units tend to have excessive capacitance. The A-2135 small-area, diffused-silicon diode operating at 120°C has proven satisfactory in all respects.

† Very stable resistors, such as Western Electric Type 106C or Weston Vamistor must be used. Consistently high stability must also be designed into those impedance elements contributed by amplifier inputs and outputs.

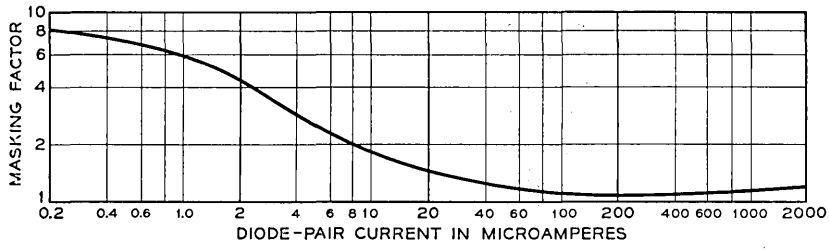


Fig. 15 — Factor by which per cent diode-pair voltage variations are reduced by resistance masking.

work stability is achieved only through the use of stable diodes housed in an oven with ±0.1°C temperature control.

4.8 *Network Temperature Control*

A photograph of the temperature-controlled oven used to house the network diodes is shown in Fig. 16. A combination of thermostat-controlled heater and Dewar-flask insulation provides the necessary 120 ± 0.1°C temperature control.

Rather than operate the associated thermostat contacts directly in the oven heater circuit, it is advantageous to include the contacts in the base circuit of a transistor switch, as shown in Fig. 17. The base-to-collector amplification of the transistor then makes possible low-voltage, low-current operation of the contacts, thereby prolonging contact life significantly.

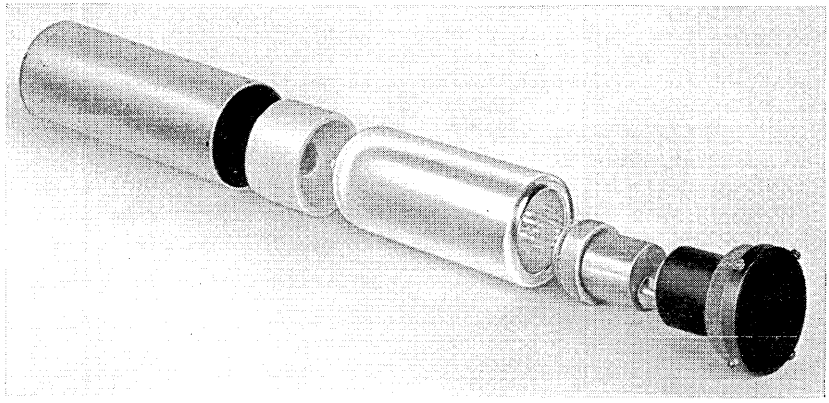


Fig. 16 — Temperature-controlled oven.

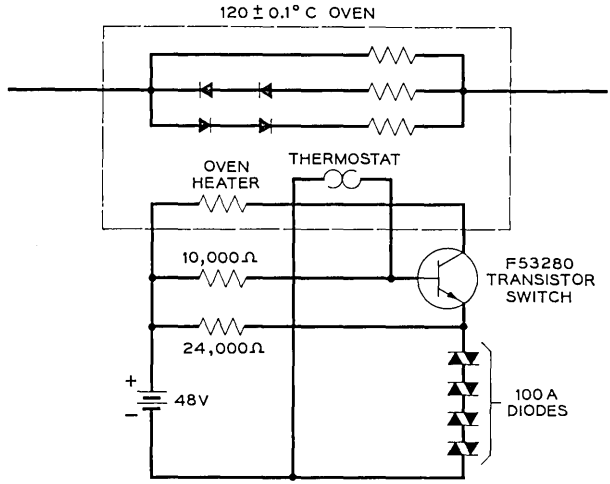


Fig. 17 — Increased thermostat life via transistor gain.

4.9 *Network Bandwidth*

In Item 3, of Section 4.3, it was pointed out that resistor-masked networks allow reasonable high values of stray capacitance without sacrifice of required bandwidth. Fig. 18 shows bandwidth as a function of level for the practical compression network achieved. Actually, the

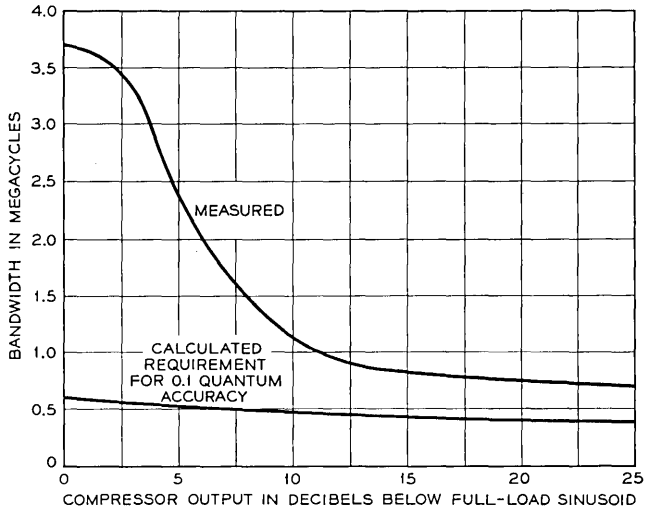


Fig. 18 — Compressor output signal vs compressor bandwidth.

bandwidth required is a function of signal level, as might be expected from the nonlinear nature of the compandor networks. At high levels, large bandwidths are required to guarantee (in the time allowed) that a given sample pulse will rise to within, say, 0.1 quantum step of its many-step final value. At low levels, less bandwidth is required, since the sample pulse need only rise to within 0.1 step of a few-step final value. The requirement shown in Fig. 18 is based on such 0.1-step accuracy. The measured result is seen to exceed the objective. In order to include all the stray capacitance associated with the network, these measurements included the transmission characteristics of the pre- and post-amplifiers associated with the network. However, since the amplifier bandwidths far exceed these values, the result is essentially one of network bandwidth.

The bandwidth needs of the expansion network are somewhat greater than those of the compression network. However, because of the expansion network configuration and its low impedance terminations, the required bandwidth is easily achieved.

4.10 *Compandor Performance*

4.10.1 *Quantizing Noise*

The signal-to-quantizing noise power performance of the compandor (for speech) has been calculated from experimental data on diode-pair characteristics under the condition of ideal 7-digit coding. The results are compared in Fig. 10 with the performance to be expected from a compressor having an ideal $\mu = 100$ characteristic. Note that for weak signals, the actual companding characteristic is somewhat better than the theoretical curve approximated. In the worst case, the actual result falls below the theoretical objective, but this occurs in a region where the signal-to-noise ratio is more than adequate. At all points the two curves differ by less than 3 db.

4.10.2 *Net Gain Stability and Third-Harmonic Distortion*

After twenty years, the net gain variation at full load and the ratio of the third harmonic to the fundamental, introduced by the compandor, should have values less than 0.45 db and -38 db, respectively.[3]

With the aid of the curves of Fig. 12 for the effects of initial and time-variant tolerance on the companding characteristic, it is possible to calculate the net gain stability and the third-harmonic distortion for a variety of conditions.

From Fig. 12 it is obvious that the principal contributors to network departures from the ideal are: (1) initial diode difference, (2) initial adjustment, and (3) diode aging.

If it is assumed that: (1) diodes having an initial difference greater than one standard deviation away from the average are rejected; (2) initial adjustment is uniformly distributed; and (3) diode aging and initial difference factors are statistically independent* and normally distributed; then 95 per cent of the compandors after twenty years will meet the gain variation and third-harmonic distortion objectives.

If, however, only diodes beyond the two-standard-deviations limit are rejected, then after twenty years 65 per cent of the compandors in service will meet the system objectives.

Table I indicates the net gain variations and third-harmonic distortion corresponding to various signal levels for the case in which those diodes within the one-standard-deviation limit are accepted.

TABLE I

Signal Level in db below Full-load Sinusoid	Net Gain Variations in db	3rd Harmonic in db below Fundamental
0	0.2	38.6
−6	0.43	48
−14	0.41	53

V. EQUAL-STEP SEVEN-DIGIT CODER

5.1 *Encoder*

5.1.1 *Encoding—General*

A digit-at-a-time or sequential comparison encoder[4] consists of three major functional blocks (see Fig. 19), namely: (1) the weighing network and switches, (2) the memory and logic circuits, and (3) the decision circuits or the summing amplifier and quantizer (consisting of a differential amplifier and flip-flop).

The code corresponding to a given compressor output is determined by comparing the compressed signal current with reference currents generated by the weighing network. The reference currents are proportional to $2^{n-1}, 2^{n-2}, \ldots 4, 2, 1$, units of current ($n$ = number of digits in

* This seems reasonable from measurements made at Bell Laboratories.

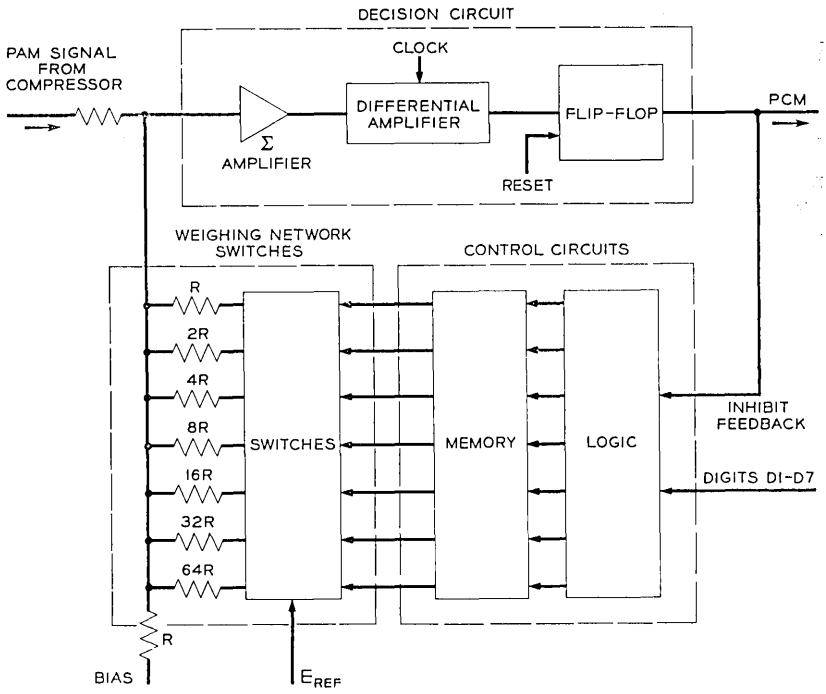


Fig. 19 — Sequential comparison encoder.

the code). In this particular implementation a code value of zero is generated whenever the decision circuits determine that a reference current is required to express the magnitude of the message. Otherwise a code value of one is generated. Thus the encoder produces the prime of the code corresponding to the signal amplitude. For example

$$11 = 0' \times 8 + 1' \times 4 + 0' \times 2 + 0' \times 1$$

$$= 1 \times 8 + 0 \times 4 + 1 \times 2 + 1 \times 1$$

where 11 is the magnitude of the signal and 8, 4, 2, and 1 are the magnitudes of the references. The transmitted code corresponding to the signal amplitude, for this example, is 0100.

Each reference current is produced by connecting a weighing resistor of value r, $2r$, $4r \ldots 64r$ either to ground or to a stable voltage. The switches providing these connections are controlled by memory elements consisting of flip-flops, which in turn are controlled by digit timing pulses D1 to D7. The network output is compared with the signal at the

summing amplifier. Because of the bipolar nature of the signal, a dc bias is introduced to the signal before comparison to insure that the zero signal level is encoded as 64, and the maximum negative signal level as 127.

During the first bit interval, the weighing resistor of value r (yielding a reference of one-half of the peak-to-peak excursion to be encoded) is connected to the negative reference voltage, while all other resistors are connected to ground. The polarity of the three-way sum of signal, bias, and negative reference is determined by the summing amplifier. If the resultant signal at the summing amplifier output is positive, indicating that the reference current exceeds in magnitude the sum of bias and signal currents, then a most significant PCM bit is transmitted. A feedback signal to the logic causes the memory element to reset; i.e., it will connect r to ground for each successive trial, since no further useful information can be obtained with the most significant reference current. However, if the resultant output of the summing amplifier is negative, thereby indicating that the signal plus the bias exceeds the reference current, the decision circuit transmits a zero for the most significant PCM bit. For this case, no feedback signal is generated, and thus r is connected to the reference voltage for the remainder of the encoding cycle. In the successive bit periods resistors of value $2r$, $4r \ldots 64r$ are tried until the final elements of the character are reached, and the resultant signal to the summing amplifier represents the quantizing error of the system. This process is illustrated in Fig. 20.

It will be recalled that crosstalk considerations lead to the use of an "odd and even" multiplexing arrangement. All the odd-numbered channels (1, 3, . . . 23) appear on one multiplexing bus, while the even-numbered ones (2, 4, . . . 24) appear on another. Each bus has its own compressor. This arrangement allows a full channel period as guard space between signal samples. To avoid the use of two encoders, the logical circuits of the encoder are made to serve in common for the encoding of the odd and the even channels* (see Fig. 21). There are two sets of weighing networks, summing and differential amplifiers, with each set attached to one compressor. The switching between the odd and even channels is performed, on the amplitude-limited analog signal, at the differential amplifier output.

5.1.2 *Weighing Resistor Network and Associated Semiconductor Switches*

A major factor in the accuracy of the coding process is the accuracy of the network outputs, that is, the spacing of the 2^n levels generated

* Suggested by C. G. Davis.

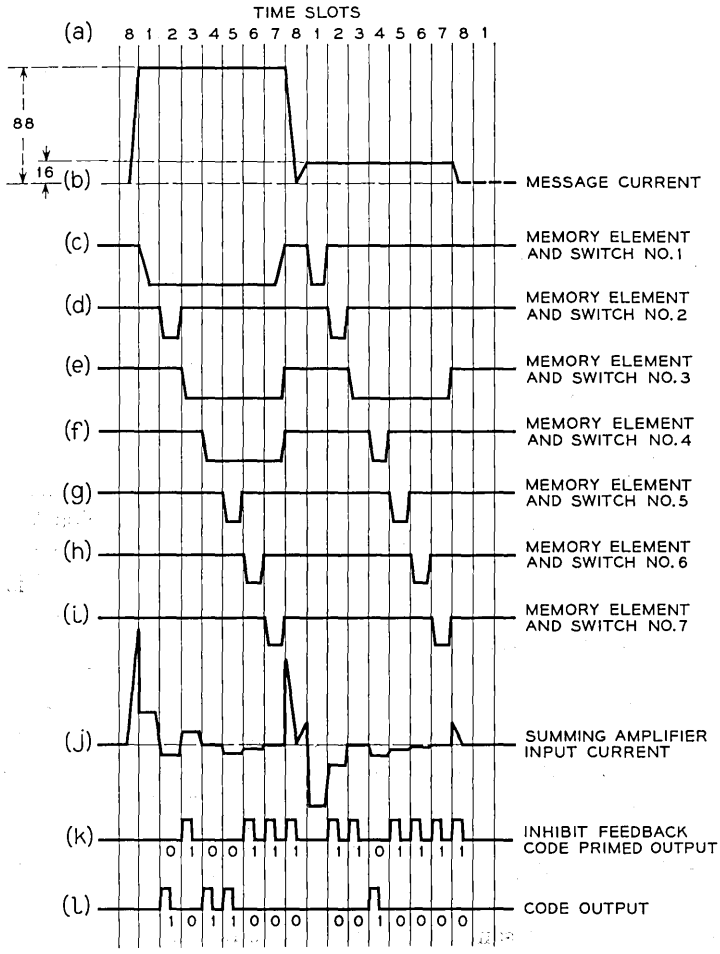


Fig. 20 — The encoding process.

by the 2^n combinations of the on-off condition of the n switches. In an ideal system, all the steps (the difference between adjacent levels) would be equal. In a physically realizable system, however, the steps will differ from one another.

There are two major sources of coder imperfections attributable to the network, namely (1) the deviations of the n weighing resistors from their nominal values, and (2) the unavoidable nonideal characteristic of the n semiconductor network switches. The effect of the imperfections

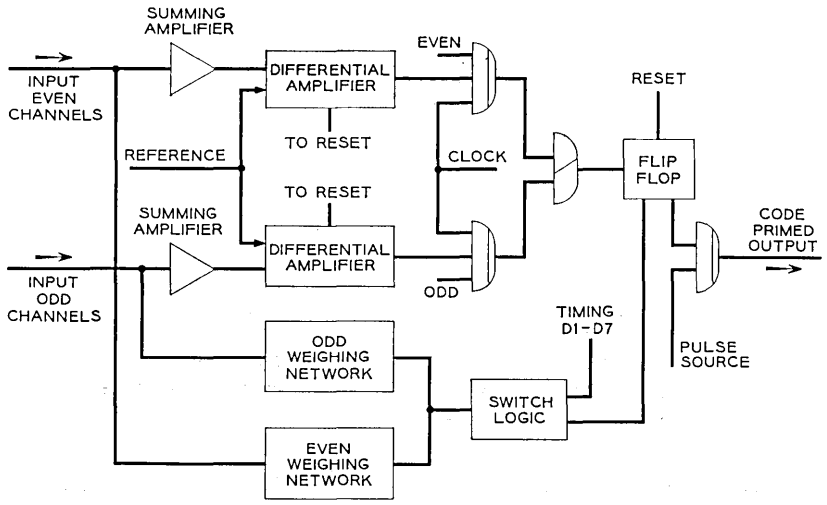


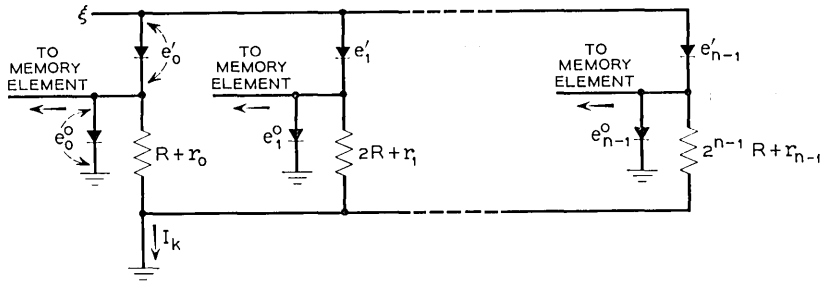
Fig. 21 — Dual encoder block diagram.

of the network is to decrease (on the average) the signal-to-quantizing noise power ratio below its theoretical value.

To obtain proper binary weighing, a network resistor must be within ± 0.13 per cent of its theoretical value. This tolerance should be met under all circumstances, such as aging, temperature variation, etc. Furthermore, the reactive components must be small enough so that transients are reduced to a reasonable level at the time a decision is made. Presently the most satisfactory resistor is of the metallic film type.

Each resistor of the network has attached to it two paths containing diodes; the diodes open and close in a complementary fashion (see Fig. 22). This arrangement permits the network to have a constant output impedance, thereby eliminating the effects of variations of the network load impedance on the binary relationship of its output. When a diode is closed for conductance, its forward voltage drop will be in series with a weighing resistor and will therefore affect the spacing of the levels of the coder. Because of the balanced diode arrangement, only the difference in the voltage drop of the switch in the open and closed position is important.* Diodes must match to within 15 mv to be satisfactory for this application. Western Electric Co. type 2030 diodes are used in the coder. A photograph of the network performance is shown in Fig. 23.

* See Section 5.3.2.



ξ: REFERENCE VOLTAGE, INCLUDING COMMON PART OF ALL SWITCH VOLTAGE DROPS

e_j^0: RESIDUAL SWITCH VOLTAGE—CONNECTION TO GROUND → CODE 0

e_j': RESIDUAL SWITCH VOLTAGE—CONNECTION TO REFERENCE → CODE 1

r_j: ERROR OF THE WEIGHING RESISTOR

Fig. 22 — Network model for the study of the fine structure effects.

5.1.3 *Decision Circuit*

The decision circuit functions to produce a binary output from the two possible network polarity states, as with a bistable regenerator. The input voltage uncertainty (region of indecision) of a regenerative element is of the order of 100 millivolts, which corresponds at the compandor output to an uncertainty of three quantum steps. To reduce the uncertainty to a fraction of a quantum step, it is necessary to introduce linear gain ahead of the regenerative element in the decision circuit.

The uncertainty of the regenerator, when referred through a summing amplifier to the summing node, is equivalent to a jitter at the

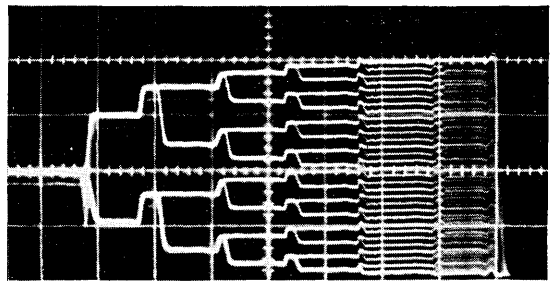


Fig. 23 — Performance of the feedback network. Each trace of the oscilloscope represents one of the 128 possible combinations of the network resistors. The lowest trace occurs when all of the network resistors are sequentially connected to ξ, the reference voltage. The highest trace occurs when all network resistors are sequentially connected to ground.

boundary of successive code outputs. The code boundary is the region in which the difference between the signal current and the sum of the network reference currents is small. As will be shown later, this jitter increases the quantizing noise associated with a coder. The summing amplifier provides enough gain to reduce this jitter to an acceptable fraction of a quantum step. To reduce the uncertainty from three steps to $\frac{1}{15}$ of a step, for example, requires a linear amplifier having a transresistance (ratio of output voltage to input current) of approximately 50,000 ohms.

In a sequential encoder, the difference between the signal current and the network reference current at the start of the comparison sequence can be as large as one half of the full amplitude range. Thus, to avoid saturation effects, an ideal transmission characteristic for the summing amplifier is one in which there is a large gain for very small currents and a low gain for large currents. Such a characteristic can be obtained by placing limiting diodes in the feedback path of a high-gain amplifier.

5.1.4 *Description of the Summing Amplifier*

The main characteristics of the summing amplifier are its transresistance, bandwidth, dc stability, and its ability to recover from heavy overload.

A schematic of the summing amplifier is as shown in Fig. 24. This configuration permits the first stage to work into a small load impedance and the last stage to provide a low output impedance, which is lowered further by shunt feedback. All three stages are protected against saturation by means of a nonlinear feedback loop which be-

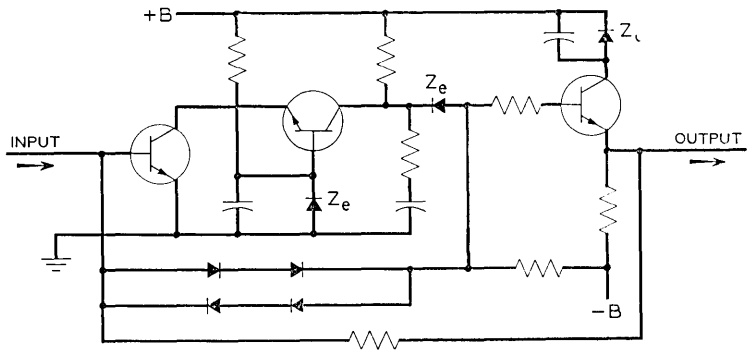


Fig. 24 — Summing amplifier schematic.

comes effective when the signal amplitude at the collector of the second stage exceeds ± 1 volt. The nonlinear feedback loop consists of a back-to-back connection of two sets of series diodes. Two diodes are used instead of one to widen the amplitude range of the high-gain region and to reduce the shunt capacitance across the diodes.

The transresistance of the summing amplifier for the high-gain region is approximately equal to the value of the feedback resistor of the linear feedback loop. A practical upper limit for this resistance is set by considering the shunting effect of stray capacitance across it, especially the capacitance of the low-gain feedback loop. The amplifier alone has a transresistance of 40,000 ohms, and it is followed by a differential amplifier which effectively increases the over-all transresistance to the required amount.

For good transient response, it is important that the gain of the open-loop response current (high-gain region) fall off with a slope not exceeding 8 or 9 db per octave. This is accomplished by having the high-frequency response controlled mainly by the cutoff of the last stage. The summing amplifier transresistance and bandwidth characteristics are as indicated in Fig. 25 and Fig. 26. The over-all bandwidth of 3 megacycles in the high-gain condition is adequate.

The dc stability of the amplifier is important because the center of the encoding characteristic must be accurately maintained to preserve tracking between the compressor and the expandor curve. Dc drift is caused mainly by variations of the base-to-emitter voltage and the base-to-collector current gain of the first stage of the amplifier with temperature changes. These variations are compensated by an appropriate complementary temperature variation of the bias current feeding the summing node.

5.2 *Decoder*

A network decoder consists of a weighing network, switches, and storage devices that permit serial-to-parallel conversion of the received

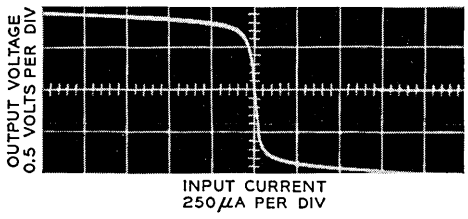


Fig. 25 — Transresistance of summing amplifier: high-gain region 40K ohms, low-gain region 200 ohms.

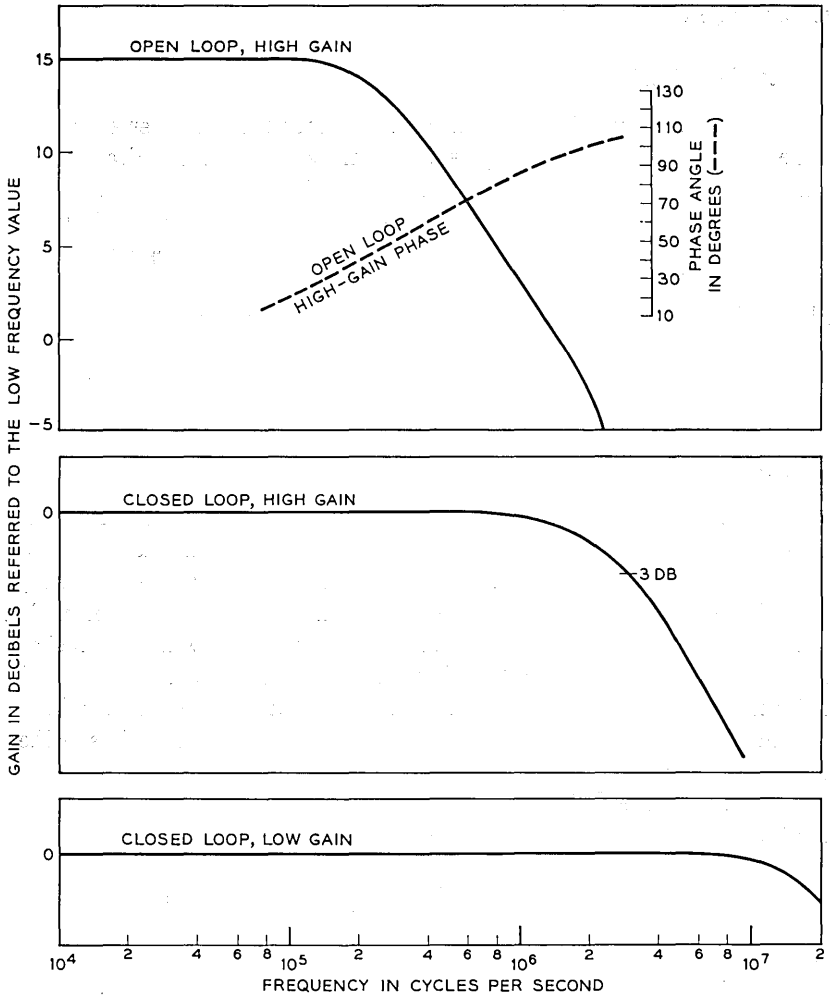


Fig. 26 — Summing amplifier performance.

code. When the code has been completely stored, bits are destructively read out in parallel to activate flip-flops, which in turn operate diode switches. These switches connect binary weighted resistors r, $2r$, ... $64r$ to either a negative reference voltage or to ground, according to the state of a flip-flop. The resultant decoded signal is then expanded and amplified. The flip-flops in the decoder are reset after a 5-bit interval.

5.3 *Coding Scale Imperfections*

5.3.1 *Introduction*

The usual computations on coder performance assume a perfect coding scale, i.e., one with perfectly equidistant steps and infinitely sharp transitions between adjacent codes. In practice, such a model can only be approximated, and the final system always shows some variation of the step size about a mean value and a finite transition region between codes. A model that simulates perturbations of the code scale is required along with a theory that relates the code scale imperfections to quantizing noise. In addition, the subjective effect of the given quantizing noise must be estimated. Specifications can then be written for the coder components to control these fine structure effects.

5.3.2 *The Fine Structure Model*

The weighing network of the coder produces at its output 2^n reference currents (encoder) or voltage steps (decoder) by combinations of the n weighing resistors. The magnitude of the reference currents or voltage steps is determined by the voltage feeding each resistor as shown in Fig. 22. This voltage is either the residual diode voltage e^0 when the resistor is connected to ground, or $\xi + e'$ when the resistor is connected to the reference potential. The error r of each resistor and the residual voltages e^0 and e' can be considered as random variables within their specified tolerance range.

Referring to the definitions indicated in Fig. 22, the network current* for the kth code is shown (in Appendix A) to be:

$$I_k'' = k + \sum_{j=0}^{n-1} \delta_{jk} \frac{2^{n-1}}{2^j} \left(-\frac{r_j}{2^j R} + \frac{e_j' - e_j^0}{\xi} \right) \tag{1}$$

where

$\delta_{jk} = 1$ if the current of weight 2^{-j} is required to express the kth reference,

$\delta_{jk} = 0$ if the current of weight 2^{-j} is not required to express the kth reference,

n = number of digits in the code, and

k = number of step in the code scale.

For simplicity of notation let us define

$$a_j = \frac{2^{n-1}}{2^j} \left(-\frac{r_j}{2^j R} + \frac{e_j' - e_j^0}{\xi} \right), \tag{2}$$

* The network currents are normalized with respect to the least significant reference current and are referred to I_0.

so that

$$I_k'' = k + \sum_{j=0}^{n-1} \delta_{jk} a_j \tag{3}$$

and a_j gives the deviation introduced by the jth arm in the coder network.

Equation (3) gives the value of the network current at which the coder output switches from the kth $-$ 1 state to the kth state, or the value of the network current in the decoder corresponding to a binary code number whose decimal equivalent is k. Table II compares the ideal and nonideal encoder decision levels or decoder output levels for a three-bit code.

Thus, the general formula for Y_k'' the kth encoder decision level or W_k'' the decoder kth output level is

$$\begin{aligned} Y_k'' &= k + \sum_{j=0}^{n-1} a_j \delta_{jk} \\ W_k'' &= k + \sum_{j=0}^{n-1} a_j' \delta_{jk} \end{aligned} \tag{4}$$

where a_j' is employed to emphasize that the decoder and encoder networks may have the same faults yet are physically different networks.

Translating the origin to $k = 2^{n-1}$ and normalizing

$$\begin{aligned} Y_k' &\cong -1 + \frac{k + \sum_{j=0}^{n-1} a_j \delta_{jk}}{2^{n-1}} \\ W_k' &\cong -1 + \frac{k + \sum_{j=0}^{n-1} a_j' \delta_{jk}}{2^{n-1}} \end{aligned} \tag{5}$$

TABLE II

Encoder Code Transition at Decision Level		Decision Levels or Output Levels		Decoder
From	To	Ideal	Nonideal	Input Code
000	000	0	0	000
000	001	1	$1 + a_2$	001
001	010	2	$2 + a_1$	010
010	011	3	$3 + a_1 + a_2$	011
011	100	4	$4 + a_0$	100
100	101	5	$5 + a_0 + a_2$	101
101	110	6	$6 + a_0 + a_1$	110
110	111	7	$7 + a_0 + a_1 + a_2$	111

where

$$\delta_{jk} \text{ for } j \neq 0 \text{ is as previously defined,}$$

but

$$\delta_{0k} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \text{ rather than } \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ as previously defined.}$$

Up to this point no attempt has been made to relate the encoder decision levels to the decoder output levels.

Generally speaking, there exists a signal amplitude difference between the encoder input and decoder output. The ratio of the signals may be considered to be the gain of the system. Hence the signals to be compared occur at different amplitudes. Furthermore, a nonlinear process separates the two signals. To be able to compare the coder input with the coder output requires a knowledge of the correspondence of input to output levels. In the absence of a sampled speech signal, the encoder is biased midway between the 64th and 65th decision levels. The decoder output rests at the 64th level. When the input signal is at its maximum amplitude, it is assumed to be one-half the average step size above the highest decision level. This amplitude therefore corresponds to the maximum decoder output. With the above in mind, it is possible to compare a normalized encoder input with a normalized decoder *quantized* output.

If the decoder output is defined as zero, then the 65th encoder decision level lies approximately one-half step above the zero signal level at the encoder input.

Therefore Y_k and W_k, the normalized relatable encoder decision levels and decoder output levels, are

$$Y_k \cong -1 + \frac{k + \sum_{j=0}^{n-1} a_j \delta_{jk}}{2^{n-1}} - \frac{1 + a_{n-1}}{2^n} \tag{6}$$

$$W_k \cong -1 + \frac{k + \sum_{j=0}^{n-1} a_j' \delta_{jk}}{2^{n-1}}. \tag{7}$$

The term $(1 + a_{n-1})/2^n$ can safely be ignored in practice.

It is reasonable to consider normalization of both signals as long as one is interested in the initial tolerance of components. It is always possible to make an initial adjustment when a coder is placed in service. However, once this system is in operation, a variation in gain will cause

compandor mistracking. In this paper, only the effects on the coder imperfections of a compandor with perfect tracking will be considered.

5.3.3 *The Relationship of the Component Tolerance to the Fine Structure*

It is assumed that the random variables r_j, $e_j' - e_j^0$ have rectangular distributions which are statistically independent with means equal to zero. Let us define

$$\sigma_j^2 \equiv E(a_j)^2 = E(a_j')^2 \tag{8}$$

and substituting (2) into (8) yields

$$\sigma_j^2 = \left(\frac{2^{n-1}}{2^j}\right)^2 \left[E\left(\frac{r_j}{2^j R}\right)^2 + E\left(\frac{e_j' - e_j^0}{\xi}\right)^2\right]. \tag{9}$$

Because of the factor $2^{n-1}/2^j$ in (9), it is preferable to allow identical tolerances for all the components of the same type. That is to allow

$$E\left(\frac{r_j}{2^j R}\right)^2 = \frac{\Delta r^2}{3R^2} \tag{10}$$

$$E\left(\frac{e_j' - e_j^0}{\xi}\right)^2 = \frac{\Delta e^2}{3\xi^2} \tag{11}$$

where $\pm\Delta r$ and $\pm\Delta e$ are the upper and lower bounds of their respective distribution.
Thus

$$\sigma_j^2 = \left(\frac{2^{n-1}}{2^j}\right)^2 \left[\left(\frac{\Delta r}{R}\right)^2 + \left(\frac{\Delta e}{\xi}\right)^2\right] \frac{1}{3} \tag{12}$$

and therefore

$$\sigma_j^2 = \left(\frac{2^{n-1}}{2^j}\right)^2 \sigma_{n-1}^2. \tag{13}$$

It turns out that a good compromise is to allow the resistor tolerance to equal one-half of the diode tolerance. Thus (12) reduces to

$$\sigma_j^2 = \left(\frac{2^{n-1}}{2^j}\right)^2 \left(\frac{\Delta r}{R}\right)^2. \tag{14}$$

Once the effect of the imperfections on the quantizing noise is known, it will be possible to specify σ_{n-1}^2 for a given noise penalty. With σ_{n-1}^2 specified, the tolerance for the components will be known.

5.3.4 *Effect of Fine Structure on the Quantizing Noise*

The mean square error* has been generally accepted as a significant measure of error introduced by ideal quantization.[2] Because of the statistical nature of the problem, the ratio of the mean value of the mean square error introduced by the nonideal coder to that generated by an ideal coder will be considered as a measure of the coder performance.

In Appendix B it is shown that the ratio of the expected value of the mean square error for a nonideal coder to the mean square error for an ideal coder in a system employing compandors is:

$$\frac{E(MSE')}{MSE} \cong C_I + 24\left(\frac{\ln(1+\mu)}{\mu}\right)^2\left[\int_0^1 (1+2e\mu+(e\mu)^2)f(e)\,de - \left[\int_0^1 (1+e\mu)f(e)\,de\right]^2\right]\sum_{j=0}^{n-1}(2^j\sigma_{n-1})^2 \tag{15}$$

where:

$E(MSE')$ is the expected value of the mean square error for imperfect coding taken with respect to the coder imperfections,
MSE is the mean square error for a perfect coder,
C_I is the companding improvement factor,[2]
$f(e)$ is the density distribution of the signal defined over the previously normalized signal range, and
σ_{n-1} is as defined in Section 5.3.3.

To use (15) it is necessary to specify $f(e)$.

5.3.5 *Representation of Speech*

It shall be assumed,[2] that the distribution of the amplitudes in speech at a constant volume can be represented by

$$f(e) = \frac{\lambda}{2}\exp(-\lambda\,|\,e\,|). \tag{16}$$

With this choice of $f(e)$ the solution of (15) becomes

$$\frac{E(MSE')}{MSE\,C_I} \cong 1 + \frac{6}{C_I}\left[\frac{\ln(1+\mu)}{\mu}\right]^2 \cdot\left[1+\frac{2\mu}{\lambda}+\frac{3\mu^2}{\lambda^2}\right]\left[\sum_{j=0}^{n-1}(2^j\sigma_{n-1})^2\right] \tag{17}\dagger$$

* See (42) in Appendix B.

† $C_I = \sqrt{2}\,\frac{\ln(1+\mu)}{\lambda}\left(1+\frac{\lambda^2}{2\mu^2}+\frac{\lambda}{\mu}\right)^{\frac{1}{2}}$ See ref. 2.

5.4 *Tolerance Specification for a Seven-Digit Coder*

Assuming a compandor having a $\mu = 100$ companding characteristic, it is possible to calculate the expected increase in quantizing noise power due to the coder imperfections. Since this increase will be a function of talker volume, calculations will be computed for a talker having a volume 13 db above that of an average power talker (−16.5 dbm). Since the system overload is at +3 dbm, the ratio of average power for this talker to peak sinusoid power is −6.5 db. It can then be shown that:

$$\lambda \cong 3. \tag{18}$$

Hence for a seven-digit coder:

$$\frac{E(MSE')}{(MSE)C_I} \cong 1 + 19.7 \sum_{j=0}^{n-1} (2^j \sigma_{n-1})^2. \tag{19}$$

Allowing $\frac{3}{4}$ db of the total allowance for this impairment

$$\sigma_{n-1}^2 \cong 1.75 \times 10^{-6}. \tag{20}$$

Using (14) yields

$$\frac{\Delta R}{R} = 1.32 \times 10^{-3} \tag{21}$$

$$\frac{\Delta e}{\xi} = 1.86 \times 10^{-3}. \tag{22}$$

If $\xi = 8$ volts, then the diode matching requirement $\Delta e = 14.88$ millivolts. The resistor tolerance is 0.13 per cent. One must bear in mind that these values were calculated for a specific talker. For a coder without companding, (17) would yield

$$\frac{E(MSE')}{(MSE)} \cong 1 + 6 \sum_{j=0}^{n-1} (2^j \sigma_{n-1})^2 \tag{23}$$

and the tolerances for an expected degradation of about $\frac{3}{4}$ db for a seven-digit coder would be:

$$\frac{\Delta R}{R} \cong 2.4 \times 10^{-3} \tag{24}$$

$$\frac{\Delta e}{e} \cong 3.4 \times 10^{-3}. \tag{25}$$

5.5 *Effect of Encoder Transition Uncertainty on the Quantizing Noise*

The boundary level between two adjacent code values will depart from its ideal value either because of noise inherently present at the

summing node or because of an uncertainty in the triggering level of the regenerative comparator. In other words, when the input signal is scanned over some amplitude range, the probability $p(x)$ of obtaining the correct code level is equal to one for a major portion of the amplitude step, but is not immediately reduced to zero in the neighborhood of the next code value. Accordingly (see Fig. 27), the next code value will start to appear with the complementary probability: $1 - p(x)$.

To study the effect of this gradual transition, it is reasonable to assume a uniform probability density $f(x_n)$ of the signal across the nth level.

For a perfect system the range of the error signal is $\pm E/2$, E being the width of a code step. In the nonideal system, with transition uncertainty, the range for the considered code is decreased to $\pm(E - \epsilon)/2$, and between $(E - \epsilon)/2$ and $(E + \epsilon)/2$ this code is produced with a probability $p(x)$.

Therefore the mean square error voltage for this region is:

$$e_n^2 = 2\left[\int_0^{(E-\epsilon)/2} x^2\,dx + \int_{(E-\epsilon)/2}^{(E+\epsilon)/2} p(x)x^2\,dx\right]f(x_n)E \qquad (26)$$

when $E \gg \epsilon$.
Let

$$p(x) = 1 - \frac{x - \dfrac{E-\epsilon}{2}}{\epsilon}$$

which implies that in the transition region the probability of obtaining a given code decreases linearly with respect to distance (see Fig. 27). Then e^2, the mean square error over all regions, is given by

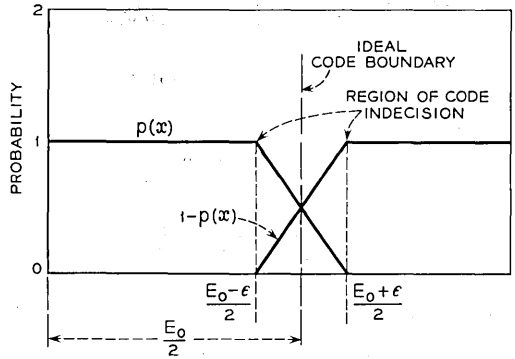


Fig. 27 — Model for the study of the effect of transition uncertainty on quantizing noise.

$$e^2 = \Sigma e_n{}^2 = \frac{E^2}{12}\left(1 + \frac{\epsilon^2}{E^2}\right). \tag{27}$$

The term $E^2/12$ is the noise power associated with a perfect step of size E. Therefore:

$$\frac{\epsilon^2}{E^2}$$

is the amount by which the noise power has been increased.

If $\epsilon/E = 1$, the noise is doubled; that is, extending the uncertainty region to the full width of a step is equivalent to losing $\frac{1}{2}$ binary digit.

The increase in the quantizing noise power in db as a function of relative width of the uncertainty region is shown in Table III.

5.6 *Discussion and Evaluation of Coder Performance*

5.6.1 *Scale Linearity*

A dc measurement of the encoding and decoding characteristics is shown on Fig. 28. The over-all transfer characteristic has excellent linearity, which is to be expected from the nature of the network. It is, of course, impossible to measure exactly the harmonic distortion introduced by the nonlinearity of the coder becaues it is masked to some extent by quantizing distortion. However, the ratio of fundamental to second- and third-harmonic distortion at full load (as measured with a slot filter) was found to be 55 db, indicating that generation of harmonic distortion by the coder is negligible.

5.6.2 *Transition Noise*

Ideally the transition region from a code word to an adjacent code word should be a small fraction of a code step. The effect of the transition region is to increase the quantizing noise associated with the signal. The average transition width measured for a coder was $\frac{1}{15}$ of a step, which corresponds to an input current change of approximately 2 microamperes to cover the transition region between adjacent codes.

From Table III it is seen that the impairment due to this effect is negligible, being less than 0.1 db.

TABLE III

ϵ/E	1/10	1/5	1/2	1/1
S/N Impairment in db	0.04	0.17	0.97	3.0

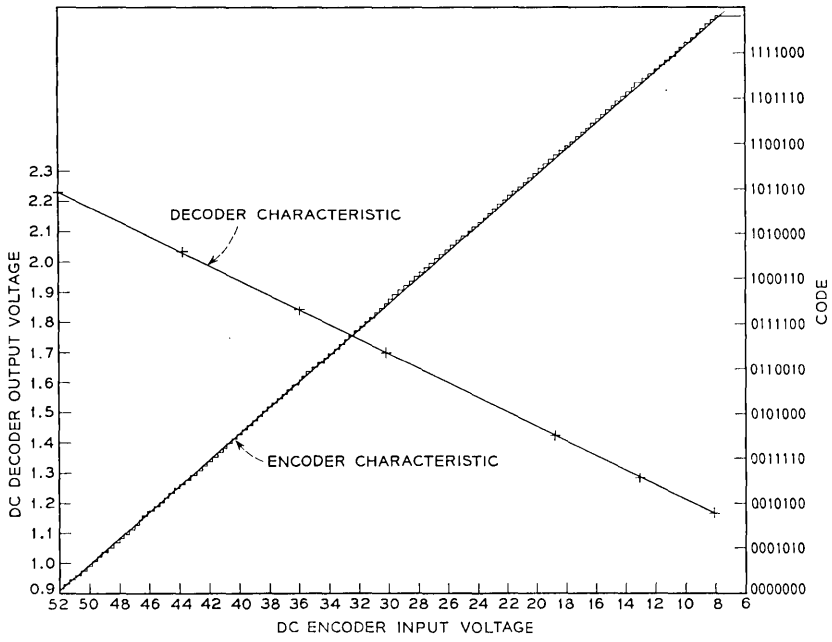


Fig. 28 — The encoding and decoding dc transfer characteristics.

5.6.3 *Quantizing Noise Performance*

While no direct noise measurements on the *coder* were performed, it was noted experimentally that a reduction of one digit (from seven- to six-digit coding) reduced the signal-to-noise ratio by 5 db at full load. Thus the seventh digit of the encoder contributes substantially all of the 6-db advantage it should yield, if perfect.

5.6.4 *The Stability of Origin of the Code*

The system specification for the stability of the origin of the code is ± 0.32 of a step ($\frac{1}{2}$ per cent relative to overload point).[3] Measurements over a short period of 35 days indicate a deviation of less than ± 0.1 of a step over this time interval. The major changes in the code center occurred because of day-to-day ambient temperature changes. In Fig. 29 is shown the deviation of the code origin as a function of temperature. In the range from 10° to 50°C, the system criterion is met.

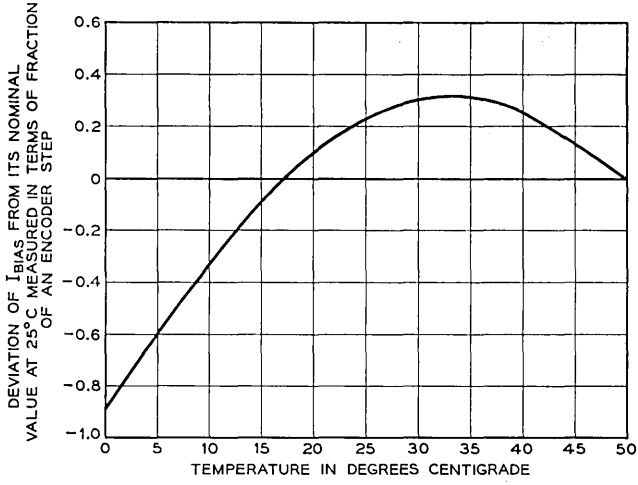


Fig. 29 — Stability of the origin of the code as a function of temperature. The nominal value of I_{bias} corresponds to a signal midway between the 64th and 65th encoder decision level.

VI. SYSTEM PERFORMANCE AND CONCLUSION

As the companded coder system comprises the major portion of the PCM system, only over-all system measurements were taken. The ratio of signal to total noise for a wide range of signals, as given by Gray and Shennum,[3] is shown in Fig. 30. It will be noted that almost all of the

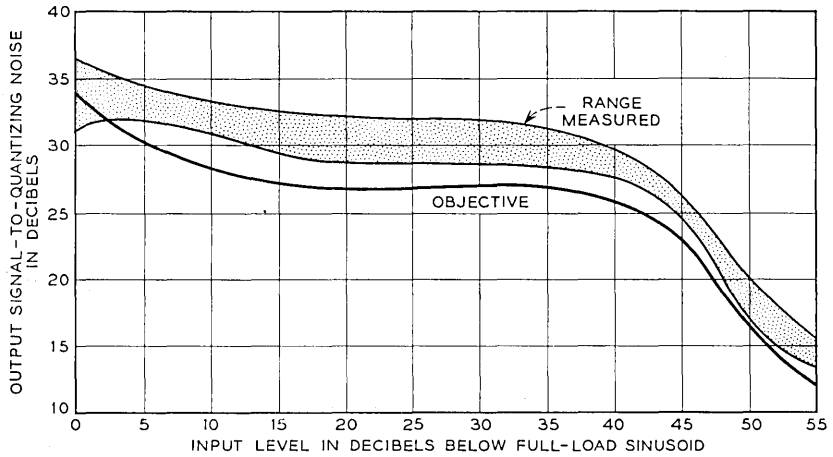


Fig. 30 — The range of the output signal-to-quantizing noise ratio for 24 channels.

channels of the system meet the specified requirement. Only at high levels in very few channels does the signal-to-noise ratio fail to meet the specification.

The measured performances of the combined and individual blocks of the companded-coder system are satisfactory and the over-all PCM system meets the desired design objectives.

VII. ACKNOWLEDGMENTS

As in any project of this magnitude, many individuals contributed to its success. In particular, the authors wish to express their appreciation for the able assistance of F. P. Rusin, who made many of the measurements associated with the compandor, F. D. Waldhauer, who consulted in the design of the compandor amplifiers, F. T. Andrews, who guided the work on the coder, and of J. R. Gray and R. C. Chapman, with whom discussions on coder imperfections were most illuminating.

APPENDIX A

The Fine Structure of the Coding Scale

The weighing network of the coder produces at its output 2^n reference currents (encoder) or voltage steps (decoder) by combinations of the n weighing resistors. The magnitude of the reference currents or voltage steps is determined by the voltage feeding each resistor, as shown in Fig. 22. This voltage is either the residual diode voltage e^0 when the resistor is connected to ground, or $\xi + e'$ when the resistor is connected to the reference potential. The error r of each resistor and the residual voltages e^0 and e' can be considered as random variables within their specified tolerance range.

Referring to the definitions indicated on Fig. 22, the network current is

Code 0

$$I_0 = \frac{e_0^{\,0}}{R + r_0} + \frac{e_1^{\,0}}{2R + r_1} + \cdots + \frac{e_{n-1}^{\,0}}{2^{n-1}R + r_{n-1}} = \sum_{j=0}^{n-1} \frac{e_j^{\,0}}{2^j R + r_j} \tag{28}$$

Code 1

$$\begin{aligned} I_1 &= \frac{e_0^{\,0}}{R + r_0} + \frac{e_1^{\,0}}{2R + r_1} + \cdots + \frac{\xi + e_{n-1}{}'}{2^{n-1}R + r_{n-1}} \\ &= \sum_{j=0}^{n-1} \frac{e_j^{\,0}\delta_{j1}{}' + (\xi + e_j{}')\delta_{j1}}{2^j R + r_j} \end{aligned} \tag{29}$$

Code k

$$I_k = \sum_{j=0}^{n-1} \frac{e_j^0 \delta_{jk}' + (\xi + e_j')\delta_{jk}}{2^j R + r_j} \tag{30}$$

where

$\delta_{jk} = 1$ if the current of weight 2^{-j} is required to express the kth reference.

$\delta_{jk} = 0$ if the current of weight 2^{-j} is not required to express the kth reference.

$\delta_{jk}'\delta_{jk} \equiv 0$.

For convenience, let I_0 be considered the reference for all other currents; then

$$\begin{aligned} I_k' = I_k - I_0 &= \sum_{j=0}^{n-1} \frac{e_j^0 \delta_{jk}' + (\xi + e_j')\delta_{jk}}{2^j R + r_j} - \sum_{j=0}^{n-1} \frac{e_j^0}{2^j R + r_j} \\ &= \sum_{j=0}^{n-1} \frac{e_j^0(\delta_{jk}' - 1) + (\xi + e_j')\delta_{jk}}{2^j R + r_j} \end{aligned} \tag{31}$$

but

$$\delta_{jk}' - 1 = -\delta_{jk} \tag{32}$$

so that

$$I_k' = \sum_{j=0}^{n-1} \frac{(\xi + e_j' - e_j^0)}{2^j R + r_j} \delta_{jk}. \tag{33}$$

If $r_j/2^j R \ll 1$, and ignoring higher-order products, then

$$I_k' = \sum_{j=0}^{n-1} \frac{\xi}{2^j R} \delta_{jk} \left(1 - \frac{r_j}{2^j R} + \frac{e_j' - e_j^0}{\xi}\right). \tag{34}$$

Setting $\xi/(2^{n-1}R) = 1$, which normalizes all the currents with respect to the nominal value of the current corresponding to the least significant digit, we have

$$\begin{aligned} I_k'' &= \sum_{j=0}^{n-1} \delta_{jk} \frac{2^{n-1}}{2^j} \left(1 - \frac{r_j}{2^j R} + \frac{e_j' - e_j^0}{\xi}\right) \\ &= \sum_{j=0}^{n-1} \delta_{jk} \frac{2^{n-1}}{2^j} + \sum_{j=0}^{n-1} \delta_{jk} \frac{2^{n-1}}{2^j} \left(-\frac{r_j}{2^j k} + \frac{e_j' - e_j^0}{\xi}\right) \end{aligned} \tag{35}$$

$$I_k'' = k + \sum_{j=0}^{n-1} \delta_{jk} \frac{2^{n-1}}{2^j} \left(-\frac{r_j}{2^j R} + \frac{e_j' - e_j^0}{\xi}\right) \tag{36}$$

where k of course ranges from 0 to $2^n - 1$. For simplicity of notation let us define

$$a_j = \frac{2^{n-1}}{2^j}\left(-\frac{r_j}{2^j R} + \frac{e_j' - e_j^0}{\xi}\right) \tag{37}$$

so that

$$I_k'' = k + \sum_{j=0}^{n-1} \delta_{jk} a_j$$

and a_j gives the deviation introduced by jth arm in the encoder network.

Thus, Y_k'', the kth decision point of the encoder, or W_k'', the kth output level of the decoder, is given by

$$\begin{aligned} Y_k'' &= k + \sum_{j=0}^{n-1} a_j \delta_{jk} \\ W_k'' &= k + \sum_{j=0}^{n-1} a_j' \delta_{jk} \end{aligned} \tag{38}$$

where a_j' is used to emphasize the fact that while the decoder and encoder networks have the same faults, they are still physically different networks.

If the zero signal level of the encoder or decoder corresponds to $k = 2^{n-1}$, then

$$\begin{aligned} Y_k''' &= k - 2^{n-1} + \sum_{j=0}^{n-1} a_j \delta_{jk} \\ W_k''' &= k - 2^{n-1} + \sum_{j=0}^{n-1} a_j' \delta_{jk} \end{aligned} \tag{39}$$

where

δ_{jk} for $j \neq 0$ is as previously defined, but

$\delta_{0k} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ rather than $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ as previously defined.

If the encoder input and decoder output are normalized to ± 1, then it is necessary to divide Y_k and W_k by the factors

$$\begin{aligned} \frac{2^n - 1 + \sum_{j=0}^{n-1} a_j}{2} &= \frac{2^n - 1}{2} \cong 2^{n-1} \\ \frac{2^n - 1 + \sum_{j=0}^{n-1} a_j'}{2} &\cong 2^{n-1} \end{aligned} \tag{40}$$

under the assumptions that

$$\text{a.} \quad \sum_{j=0}^{n-1} a_j \ll 2^{n-1}$$
$$\text{b.} \quad \tfrac{1}{2} \ll 2^n. \tag{41}$$

APPENDIX B*

The Functional Relation Between Mean Square Error of an Ideal Coder and the Expected Mean Square Error of a Practical Coder

The mean square error (MSE'') for a *particular* companded coded system is given by

$$MSE'' = \sum_{k=0}^{2^n-1} \int_{e_k+\epsilon_k}^{e_{k+1}+\epsilon_{k+1}} (e - d_k - \mu_k)^2 f(e)\, de \tag{42}$$

where

e_k = the compressor input amplitude at which the kth encoder transition should occur,
ϵ_k = the amount the kth encoder transition is displaced due to encoder imperfections,
d_k = the kth expandor output amplitude due to the decoder,
μ_k = the amount the kth expandor output amplitude has been displaced by decoder imperfections,
e = the signal at the compandor input, and
$f(e)$ = the probability density of the input signal.

It can be shown that

$$E(MSE'') \cong MSE''' + \sum_{k=0}^{2^n-1} \int_{e_k}^{e_{k+1}} E(\mu_k^2) f(e)\, de + \sum_{k=1}^{2^n-1} E(\epsilon_k^2) f(e) de \tag{43}$$

where MSE''' is the mean square error when $\epsilon_k = \mu_k = 0$ and $E(MSE'')$ implies the expected value of MSE'' over all coders, and where ϵ_k and μ_k are independent random variables (having mean zero) corresponding to the deviations in I_k presented in the main text.

* This problem has been treated in a private unpublished memorandum by W. L. Ross. His work has been modified and extended in this section.

† In this expression it is implied that the effect at the end points can be ignored. Such an assumption is reasonable since $f(e)$, for the average talker, is nearly zero at the extremes of the coding range.

The theoretical compression characteristic is given by

$$y = \frac{v}{\ln (1 + \mu)} \ln \left(1 + \mu \frac{x}{v}\right) \qquad (x > 0) \tag{44}$$

$$y = -\frac{v}{\ln (1 + \mu)} \ln \left(1 - \mu \frac{x}{v}\right) \qquad (x < 0)\,. \tag{45}$$

The above equation can be used to translate the imperfections at the coder terminals to imperfections at the compandor terminals.

Let:

Y_k = the amplitude at the input to the encoder at which the kth transition should occur,

γ_k = the amount the kth encoder transition is displaced due to encoder imperfections,

W_k = the kth output amplitude of the decoder, and

Ω_k = the amount the kth output amplitude has been displaced by decoder imperfections. Then the signal levels at the coder terminals corresponding to the signal levels at the compandor terminals are given by

$$Y_k + \gamma_k = \pm \frac{v}{\ln (1 + \mu)} \ln \left(1 \pm \frac{\mu(e_k + \epsilon_k)}{v}\right) \tag{46}$$

$$W_k + \Omega_k = \pm \frac{v}{\ln (1 + \mu)} \ln \left(1 \pm \frac{\mu(d_k + \mu_k)}{v}\right). \tag{47}$$

Now under the assumption of small perturbations

$$\epsilon_k = \pm \frac{\gamma_k \ln (1 + \mu)}{\mu} \left(1 \pm \frac{e_k \mu}{v}\right) \tag{48}$$

$$\mu_k = \pm \frac{\Omega_k \ln (1 + \mu)}{\mu} \left(1 \pm \frac{d_k \mu}{v}\right). \tag{49}$$

Thus from (43)

$$\begin{aligned} E(MSE'') = MSE''' + \sum_{k=1}^{2^n-1} \left[\frac{\ln (1 + \mu)}{\mu}\right]^2 \int_{e_k}^{e_{k+1}} E(\gamma_k^2) \\ \cdot \left(1 \pm \frac{e_k \mu}{v}\right)^2 f(e)\, de + \sum_{k=0}^{2^n-1} \left[\frac{\ln (1 + \mu)}{\mu}\right]^2 \\ \cdot \int_{e_k}^{e_{k+1}} E(\Omega_k^2) \left(1 \pm \frac{d_k \mu}{v}\right)^2 f(e)\, de \end{aligned} \tag{50}$$

where

$$e_k \gtrless 0 \qquad k \gtrless 2^{n-1}$$

$$d_k \gtrless 0 \qquad k \gtrless 2^{n-1}.$$

Referring to Section 5.3.2 on the fine structure of the code scale and recalling that the encoder is considered centered between 64 and 65, it is possible to relate γ_k and Ω_k to the a_j's. However, for simplicity in calculations and since for a large number of levels the $\frac{1}{2}$-step displacement can be ignored, it is assumed that the encoder zero is at 64 (mid-riser). If the compandor coder system is normalized to $v = \pm 1$, then

$$Y_k = \frac{k - 2^{n-1} - \frac{1}{2}}{2^{n-1}} \cong \frac{k - 2^{n-1}}{2^{n-1}} \tag{51}$$

$$W_k \cong \frac{k - 2^{n-1}}{2^{n-1}} \tag{52}$$

which implies $d_k \cong e_k$, and

$$\gamma_k = \sum_{j=0}^{n-1} \frac{a_j \delta_{jk}}{2^{n-1}} \tag{53}$$

$$\Omega_k = \sum_{j=0}^{n-1} \frac{a_j' \delta_{jk}}{2^{n-1}}. \tag{54}$$

Substituting in (50) and rearranging terms and recalling that

$$E[(a_j')^2] = E(a_j^2) \equiv \sigma_j^2$$

$$E(MSE'') = MSE''' + \sum_{j=0}^{n-1} \frac{\sigma_j^2 2}{(2^{n-1})^2} \left[\frac{\ln (1 + \mu)}{\mu}\right]^2 \cdot \left[\sum_{k=1}^{2^n-1} \delta_{jk}^2 (1 \pm e_k \mu)^2 \int_{e_k}^{e_{k+1}} f(e)\, de + \int_{e_0}^{e_1} \beta_0 f(e)\, de\right]. \tag{55}$$

An obvious solution for the $E(MSE'')$ can be obtained by replacing the discrete variable e_k by the variable e.*

Furthermore, due to the complementary nature† of the code, the symmetry of all the functions of e involved (see Fig. 31) and the fact that a

* For 2^n large (a dense code), the replacement of e_k by e is reasonable.

† We conveniently ignore the encoder half-step displacement and the $k = 0$ term of (55).

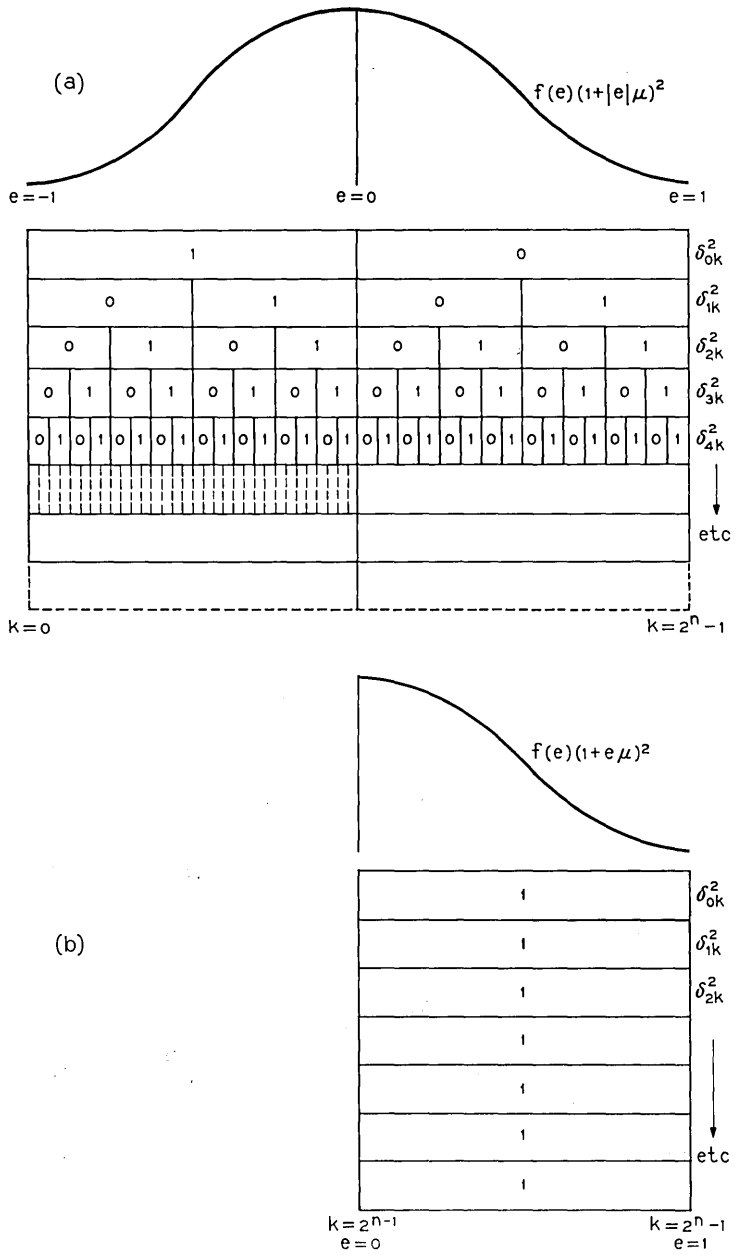


Fig. 31 — The complementary nature of the binary code gives δ_{jk} a complementary nature: for example, $\delta_{ok} = 0$ for all values of $k \geqq 64$. Because of the complementary nature of δ_{jk} and the symmetrical nature of $f(e)$ and $(1+ \mid e_k \mid \mu)^2$ about $k = 64$, the summation on k of (55) can be reduced to a summation over one-half of the range of interest which is independent of δ_{jk} .

finite sum of integrals can be replaced by a single integral (55) can be reduced to

$$E(MSE'') \cong MSE''' + 2\frac{\sum_{j=0}^{n-1}\sigma_j^2\left[\frac{\ln(1+\mu)}{\mu}\right]^2}{(2^{n-1})^2} \cdot \int_0^1 (1 + 2e\mu + e^2\mu^2) f(e)\, de. \tag{56}$$

Note that if $\mu = 0$, then by L'Hospital's rule (56) becomes

$$E(MSE'') = MSE + \frac{\sum_{j=0}^{n-1}\sigma_j^2}{(2^{n-1})^2} \tag{57}$$

since

$$\int_0^1 f(e)\, de = \frac{1}{2}.$$

The $E(MSE'')$ is independent of the statistics of the signal source.

For a coder without a compandor then

$$\frac{E(MSE'')}{MSE} = 1 + 12\sum_{j=0}^{n-1}\sigma_j^2. \tag{58}$$

One important point remains to be considered before (56) can be used; that is, what is the time average value of the additional error caused by imperfections of the coder? Since the system cannot pass dc, the average value of the error term can be ignored. Under the assumption of ergodicity, the time average and the ensemble average are equal and hence can be used interchangeably.

Here we are concerned with the signal ensemble and not the coder ensemble.

To avoid confusion, the symbol $\langle\ \rangle$ will signify the mean with respect to the signal distribution. Therefore

$$\text{Time average of error} = \langle e - d_k - \mu_k \rangle$$

$$\langle e - d_k - \mu_k \rangle = \sum_k \int_{e_k+\epsilon_k}^{e_{k+1}+\epsilon_{k+1}} (e - d_k - \mu_k) f(e)\, de \tag{59}$$

$$\langle e - d_k - \mu_k \rangle \cong \sum_k \int_{e_k}^{e_{k+1}} (e - d_k) f(e)\, de$$

$$- \sum_k \int_{e_k}^{e_{k+1}} \mu_k f(e)\, de \qquad (60^*)$$

$$+ \sum_k \int_{e_k}^{e_k + \epsilon_k} [-d_{k-1} + d_k - \mu_{k-1} + \mu_k] f(e)\, de.$$

But d_k is assumed to be the mean value of e in the interval $[e_k, e_{k+1}]$, hence

$$\langle e - d_k - \mu_k \rangle = -\sum_k \int_{e_k}^{e_{k+1}} \mu_k f(e)\, de$$

$$- \sum_k \int_{e_k}^{e_k + \epsilon_k} (-d_k + d_{k-1} - \mu_k + \mu_{k-1}) f(e)\, de. \qquad (61)$$

But by the mean value theorem:

$$-\sum_k \int_{e_k}^{e_k + \epsilon_k} (-d_k + d_{k-1} - \mu_k + \mu_{k-1}) f(e)\, de$$

$$\cong \sum_k - \epsilon_k f(e_k)(d_{k-1} - d_k) - \sum_k \epsilon_k f(e_k)(\mu_{k-1} - \mu_k).$$

Assuming $d_k - d_{k-1} = e_{k+1} - e_k$, then the above expression becomes

$$\cong \sum \epsilon_k \int_{e_k}^{e_{k+1}} f(e)\, de - \sum_k \epsilon_k f(e_k)(\mu_{k-1} - \mu_k). \qquad (62)$$

Hence

$$\langle e - d_k - \mu_k \rangle$$

$$\cong \sum_k \left[\int_{e_k}^{e_{k+1}} (\epsilon_k - \mu_k) f(e)\, de - \epsilon_k (\mu_{k-1} - \mu_k) f(e_k) \right]. \qquad (63)$$

Ignoring the second-order term $\epsilon_k(\mu_{k-1} - \mu_k) f(e_k)$

$$E[\langle e - d_k - \mu_k \rangle^2] = E\left[\left[\sum_k \int_{e_k}^{e_{k+1}} (\epsilon_k - \mu_k) f(e)\, de\right]^2\right]. \qquad (64)$$

* In this expression it is implied that the effect of the end points can be ignored. Such an assumption is reasonable since $f(e)$, for the average talker, is nearly zero at the extremes of the coding scale.

Substituting in (64) equations (48) and (49) for ϵ_k and μ_k respectively, assuming that $e_k = d_k$, and recalling the steps involved in (56)

$$E[\langle e - d_k - \mu_k \rangle^2] \cong 2 \sum_{j=0}^{n-1} \left(\frac{\sigma_j}{2^{n-1}}\right)^2 \left[\int_0^1 (1 + e\mu) f(e)\, de\right]^2 \left[\frac{\ln (1 + \mu)}{\mu}\right]^2. \quad (65)$$

Subtracting (65) from (56) gives the ac component of the $E(MSE')$

$$E(MSE') \cong MSE''' + \frac{2}{(2^{n-1})^2} \left(\frac{\ln (1 + \mu)}{\mu}\right)^2 \cdot \left[\left[\int_0^1 (1 + 2e\mu + (e\mu)^2) f(e)\, de\right] - \left[\int_0^1 (1 + e\mu) f(e)\, de\right]^2\right] \sum_{j=0}^{n-1} \sigma_j^2. \quad (66)$$

$$\text{Now } MSE''' = \frac{C_I}{(2^{n-1})^2 12} \quad (67)$$

where C_I is the companding improvement, which equals 1 if $\mu = 0$. Thus we obtain

$$\frac{E(MSE')}{MSE} \cong C_I + 24 \left[\frac{\ln (1 + \mu)}{\mu}\right]^2 \left[\int_0^1 (1 + 2e\mu + (e\mu)^2) f(e)\, de - \left[\int_0^1 (1 + e\mu) f(e)\, de\right]^2\right] \sum_{j=0}^{n-1} \sigma_j^2. \quad (68)$$

If there is to be identical tolerance among all the branches of the network, then,

$$\sigma_j^2 = \left(\frac{2^{n-1}}{2^j}\right)^2 \sigma_{n-1}^2. \quad (69)$$

Substituting in (68)

$$\frac{E(MSE')}{MSE} \cong C_I + 24 \left[\frac{\ln (1 + \mu)}{\mu}\right]^2 \left[\int_0^1 (1 + 2e\mu + (e\mu)^2) f(e)\, de - \left[\int_0^1 (1 + e\mu) f(e)\, de\right]^2\right] \sum_{j=0}^{n-1} (2^j \sigma_{n-1})^2 \quad (70)$$

or if $\mu = 0$, then

$$\frac{E(MSE')}{MSE} \cong 1 + 6 \sum_{j=0}^{n-1} (2^j \sigma_{n-1})^2. \quad (71)$$

REFERENCES

1. Davis, C. G., this issue, p. 1.
2. Smith, B., B.S.T.J., **36,** May, 1957, p. 653.
3. Shennum, R. H., and Gray, J. R., this issue p. 143.
4. Smith, B. D., Proc. I.R.E., **41,** Aug., 1953, p. 1053.
5. Villars, C. P., unpublished memorandum.
6. Mann, H., unpublished memorandum.
7. Black, H. S., and Edson, J. O., Trans. A.I.E.E., **66,** 1947, p. 895.
8. Goodall, W. M., B.S.T.J., **26,** July, 1947, p. 395.
9. Goodall, W. M., B.S.T.J., **30,** Jan., 1951, p. 33.
10. Meacham, L. A., and Peterson, E., B.S.T.J., **27,** Jan., 1948, p. 1.
11. Smith, B., Fultz, K. E., and Glaser, J. L., unpublished memorandum.
12. Straube, H. M., Lecture Session 36/3, Western Electronic Show and Convention, San Francisco, Cal., Aug. 25, 1961.

Variational Techniques Applied to Capture in Phase-Controlled Oscillators

By R. D. BARNARD

(Manuscript received April 4, 1961)

A variational formulation for deriving bounds on the capture (pull-out) range of phase-controlled oscillators is developed. This technique is applied to the more common types of systems, viz., those involving symmetric comparators and simple lag-RC filters. Exact asymptotic capture expressions relating to the simple RC filters with small bandwidth are obtained.

I. INTRODUCTION

The synchronization behavior of phase-controlled oscillators has been for many years the subject of extensive analytic studies. Although yielding significant and experimentally consistent results, these studies have been based primarily on laborious graphical procedures and linearizing approximations. In this paper we discuss and apply a variational formulation with which to analyze the synchronization phenomena of phase-controlled systems more directly and somewhat more generally.

The phase-controlled oscillator is represented by the block diagram of Fig. 1(a). The basic function of such a system is to establish and maintain, within certain limits, an approximately zero difference in frequency between the input and local carriers; i.e., the local oscillator is constrained so as to follow variations in the frequency of the input carrier. As indicated, this control is effected by applying the two carriers to a phase comparator and linear filter in cascade, and utilizing the resultant output to appropriately control the frequency of the local oscillator. The comparator output in most cases is a simple function of the phase difference $\varphi_i - \varphi_o$ between the carriers, the type of function depending to a large extent on the particular physical operation by which the phase difference is measured. For example, if the frequency spectrum of $\varphi_i - \varphi_o$ is baseband, then multiplying the two carriers and filtering out the resultant high-frequency component yield an output which approximates that of an ideal "cosine" comparator, viz., $f(\varphi_i - \varphi_o) \sim \cos(\varphi_i - \varphi_o)$. In general,

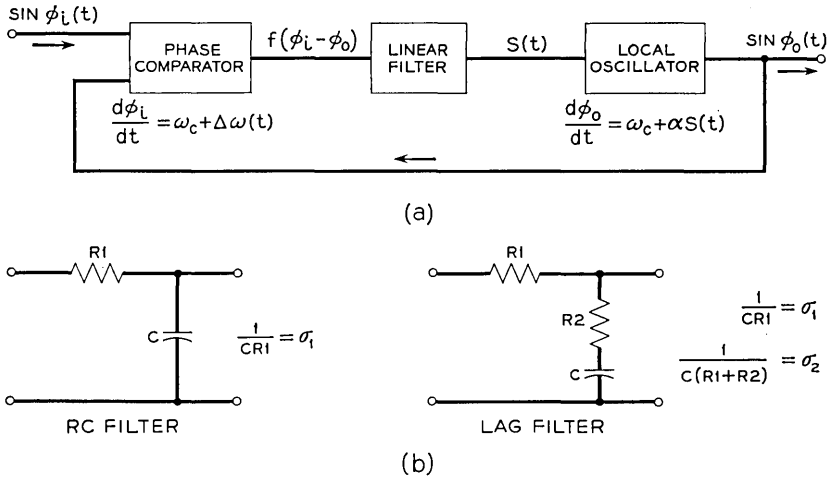


Fig. 1 — (a) Block diagram of the phase-controlled oscillator; (b) RC and lag filters.

$f(\varphi_i - \varphi_o)$ is not a baseband function; however, the filter depicted in Fig. 1(a) is usually a low-pass lag or RC type, serving to eliminate any high-frequency deviations in the control branch.

As applied here, the term synchronization relates to the difference in frequency between the input and output carriers; specifically, we consider the system to be in a synchronous state provided this frequency difference is either identically or asymptotically zero, i.e.,

$$(d\varphi_i/dt) - (d\varphi_o/dt) = 0 \quad \text{or} \quad \to 0 \ (t \to \infty).$$

Despite the fact that these conditions are realized in special cases only, near-synchronous states are readily established by most practical systems. Whether an exact, near, or nonsynchronous state prevails in a given situation depends largely on the particular form of input $d\varphi_i/dt$ and the related initial conditions. If completely general, the analysis of such phenomena necessarily involves both arbitrary inputs and arbitrary initial conditions. Owing to the intractable nonlinear formulation encountered, however, a broad treatment of synchronization does not appear possible at present except for trivial systems.

To establish effective criteria for evaluating system performance yet obviate the mathematical difficulties associated with a more detailed analysis, we consider special classes of input functions that in cases of interest characterize actual inputs and lead to exact or approximate solu-

tions. The most acceptable as well as convenient input is perhaps the step function, a single "jump" in frequency of magnitude $C_1(t \geqq 0)$, normally applied with the phase difference initially zero and the system initially in a synchronous state. Such functions represent the more extreme or sudden type of input deviation, and yield in the simple lag-RC filter cases (cf. Fig. 1(b)) autonomous formulations to which phase plane descriptions and techniques are applicable. We often find it possible to consider inputs with less restrictive initial conditions, namely those satisfying the asymptotic relation

$$(d\varphi_i/dt) - (\omega_c) \to C_2 = \text{const } (t \to \infty).$$

(Here, ω_c represents the "free-running" frequency of the local oscillator and C_2, an asymptotic frequency jump.*) The maximum positive values of C_1 and C_2 for which the related functions restore the system to a synchronous state are defined as the positive "pull-out" (capture) and "pull-in" ranges, respectively. As measures of performance, these definitions have been used extensively, especially in cases involving sine comparators and lag-RC filters.

One essential purpose of this paper is to formulate a variational procedure whereby one can obtain analytically bounds on the capture range. The method applies specifically to the second order, autonomous, nonlinear equation,

$$\frac{d^2x}{dt^2} = F\left(x, \frac{dx}{dt}\right)$$

which for $x \sim \varphi_i - \varphi_o$ describes phase-controlled systems with lag-RC filters and step function inputs. Relative to the phase plane representation given in Fig. 2, solution trajectories of the above equation for lag-RC systems are of two principal types: a nonsynchronous form running on indefinitely, and a synchronous or capture form satisfying the con-

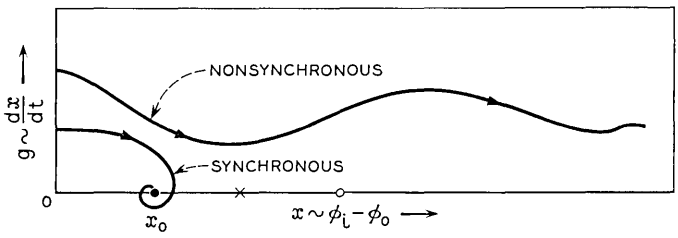


Fig. 2 — Phase plane representation of solution trajectories.

* See list of symbols, Appendix D.

dition $(d\varphi_i/dt) - (d\varphi_o/dt) \to 0$. As discussed previously, the basic problem is determining values of C_1 for which only the latter form results. To find these values by phase plane methods requires fairly accurate graphical plots of a large number of solution trajectories. Of main concern, however, is the relationship between C_1 and the resultant solution type, not detailed knowledge of the trajectory. Utilizing this principle, we show below that subintervals of C_1 can be identified as corresponding to either capture or noncapture trajectories by means of "test" trajectories, prescribed functions exhibiting the essential features of the actual solutions. The more closely the test contours fit those of the solutions, the more closely the derived subintervals bound the capture and noncapture ranges. For obtaining optimal bounds, we normally construct a class of appropriate test functions and vary over that class, the procedure simulating to some extent that of linear vector-space theory where variational formulations based on trial eigenfunctions furnish bounds on eigenvalues. Although the development is restricted to the second-order, autonomous equation given above, straightforward generalizations enable one to deal with higher-order, nonautonomous equations.

In the treatment that follows, we consider first the mathematical formulation of phase-controlled systems involving lag-RC filters, then the variational formulation, and finally the application of the latter to symmetric comparator functions, i. e.,

$$f(\epsilon) = -f(-\epsilon) \quad (\epsilon = \varphi_i - \varphi_o)$$
$$f(\epsilon) = f(\epsilon + 2\pi)$$
$$f(\epsilon) \geqq 0 \quad (0 \leqq \epsilon \leqq \pi).$$

Regarding the symmetric comparator and RC filter case, several results are derived: (*i*) bounds on the capture range under general bandwidth conditions; (*ii*) exact asymptotic formulae for the capture range under small and large bandwidth conditions; (*iii*) the fact that for capture solutions, the steady-state phase error lies in the interval

$$-\pi \leqq \epsilon \leqq \pi.$$

With respect to sawtooth comparator functions, i. e.,

$$f(\epsilon) = \sum_{n=-\infty}^{\infty} f_0(\epsilon - 2\pi n)$$
$$f_0(\epsilon) = \begin{cases} \epsilon & (|\epsilon| < \pi) \\ 0 & (|\epsilon| \geqq \pi) \end{cases}$$

we determine exact capture relations for RC filters and capture bounds for lag filters.

II. GENERAL FORMULATION

As in previous investigations,[1,2] the mathematical representation of the phase-controlled system shown in Fig. 1 is based primarily on three physical constraints:

(*i*) With regard to the control of the local oscillator, the instantaneous frequency $\omega_0(t)$ of the local carrier consists of a constant, free-running component ω_c which is independent of the control signal $S(t)$ applied to the local oscillator, and a component which is directly proportional to $S(t)$; i. e.,

$$\omega_0(t) = \frac{d\varphi_0}{dt} = \omega_c + \alpha S(t). \tag{1}$$

In general, ω_c is chosen so as to lie within the frequency range of the input signal. Accordingly, we represent the input as the sum of two frequency components, ω_c and a time-dependent deviation $\Delta\omega(t)$; viz.,

$$\omega_i(t) = \frac{d\varphi_i}{dt} = \omega_c + \Delta\omega(t). \tag{2}$$

(*ii*) Regarding synchronism, it is required that the system exhibit a "natural" synchronous state for which $\omega_i(t) = \omega_0(t) = \omega_c (t \geqq 0)$ provided only $\epsilon(0) = \varphi_i(0) - \varphi_o(0) = 0$ and $\Delta\omega(t) = 0$ $(t \geqq 0)$. If, after being initially established, this state is to persist indefinitely, then $S(t) = 0$ $(t \geqq 0)$. The latter condition requires the comparator output to vanish for zero error input and the linear filter to be initially inert. More explicitly, it is stipulated that

$$S(t) = 0 \quad (\epsilon(t) = 0, \qquad t \geqq 0) \tag{3}$$

$$f(0) = 0 \tag{4}$$

$$\left(\frac{df}{d\epsilon}\right)_{\epsilon=0} \leqq m = \text{const} < \infty. \tag{5}$$

In addition, the usual supposition of physical realizability implies the requisite

$$|f(\epsilon)| \leqq f_m = \text{const} < \infty \quad (-\infty < \epsilon < \infty). \tag{6}$$

(*iii*) Relative to boundary conditions, we assume that with the system initially in the synchronous state outlined in (*ii*), there is applied at

$t = 0$ an arbitrary $\Delta\omega(t)$ depending only on the behavior to be analyzed:

$$\epsilon(0^-) = 0, \qquad \Delta\omega(t) = 0 \quad (t < 0). \tag{7}$$

The constitutive conditions given by (1) through (7) yield

$$\frac{d\epsilon}{dt} = \Delta\omega(t) - \alpha \int_0^t H(t - \tau) f[\epsilon(\tau)]\, d\tau \quad (t \geqq 0) \tag{8}$$

where the filter impulse response $H(t)$, $\Delta\omega(t)$, and $f(\epsilon)$ are regarded as generalized functions.[3] For physically realizable filters and comparators, the convolution in the last term of this expression is bounded in a neighborhood of the point $t = 0$; hence, provided $\Delta\omega(t)$ is bounded, $d\epsilon/dt$ behaves similarly in this neighborhood and

$$\epsilon(0^-) = \epsilon(0) = \epsilon(0^+) = 0. \tag{9}$$

Considering specifically the lag filter for which

$$H(t) = \mathfrak{L}^{-1}\left[\frac{\sigma_1}{\sigma_2}\left(\frac{s + \sigma_2}{s + \sigma_1}\right)\right] = \frac{\sigma_1}{\sigma_2}[\delta(t) + (\sigma_2 - \sigma_1)e^{-\sigma_1 t}] \quad (t \geqq 0) \tag{10}$$

and differentiating (8) with respect to t, one obtains

$$\begin{aligned} \frac{k}{\sigma_1}\frac{dg}{dt} = \frac{1}{\alpha f_m}\left(\Delta\omega + \frac{1}{\sigma_1}\frac{d\Delta\omega}{dt}\right) \\ - C(x) - kg + (1 - \beta)kg\frac{dC}{dx} \quad (t > 0) \end{aligned} \tag{11}$$

$$\frac{dx}{dt} = \frac{\sigma_1}{k} g \tag{12}$$

where

$$x(t) = \frac{1}{\pi}\epsilon(t), \qquad C(x) = \frac{1}{f_m} f(\pi x)$$

$$\beta = 1 + \frac{\alpha f_m}{\pi\sigma_2}, \qquad k = \left(\frac{\pi\sigma_1}{\alpha f_m}\right)^{\frac{1}{2}}.$$

One notes from (1), (8), and (10) that

$$|\omega_0 - \omega_c| = \alpha |S(t)| = \alpha |H^* f| \leqq \alpha f_m \int_0^\infty H(\tau)\, d\tau = \alpha f_m.$$

Thus, the local oscillator cannot establish synchronism, i.e., $\omega_i = \omega_0$, if $\Delta\omega(t) = \omega_i - \omega_c > \alpha f_m$ $(t \geqq 0)$. Anticipating that $\Delta\omega/\alpha f_m \leqq 1$ in cases

which involve synchronism, we define the relative deviation $\Omega(t)$ by the relation

$$\Omega(t) = \frac{\Delta\omega(t)}{\alpha f_m}. \tag{13}$$

Equations (11), (12) and (13) can be combined to give

$$g\frac{dg}{dx} = \Omega(t) + \frac{1}{\sigma_1}\frac{d\Omega}{dt} - C(x) - kg + (1-\beta)kg\frac{dC}{dx}. \tag{14}$$

Also, by (4), (8), (9), (11), and (12),

$$g[x(0^+)] = \frac{k}{\pi\sigma_1}\left(\frac{d\epsilon}{dt}\right)_{0^+} = \frac{\Omega(0)}{k} \tag{15}$$

$$x(0^-) = x(0) = x(0^+) = 0. \tag{16}$$

To render the analysis of (14) to (16) tractable, we consider only specific classes of $\Omega(t)$. These deviations, although restricted, are chosen so that the associated synchronization behavior is representative of the more general phenomena. Accordingly, we discuss the following two classes:

(*i*) As a characterization of sudden input changes, $\Omega(t)$ is assumed to be a step function of magnitude ξ, causing either a temporary or permanent loss of synchronism. The maximum value of ξ for which the system returns to a synchronous state is denoted by ξ_+; namely,

$$\xi_+ = \sup\{\xi \mid \xi \geqq 0; \quad g(x) \to 0\ (t \to \infty)\} \tag{17}$$

where

$$\Omega(t) = \begin{cases} \xi & (t \geqq 0) \\ 0 & (t < 0) \end{cases}.$$

Similarly, for negative values of ξ,

$$\xi_- = \inf\{\xi \mid \xi < 0; \qquad g(x) \to 0\ (t \to \infty)\}. \tag{18}$$

We define the "relative" capture range ξ_c by the relation

$$\xi_c = \xi_+ - \xi_-. \tag{19}$$

(*ii*) As a generalization of the deviation type above, $\Omega(t)$ is assumed initially arbitrary but asymptotically such that

$$\Omega(t) \to \gamma = \text{const}\ (t \to \infty).$$

In a manner similar to that of (i), we define the relative "pull-in" range γ_p by the relation

$$\gamma_p = \gamma_+ - \gamma_- \tag{20}$$

where

$$\gamma_+ = \sup \{\gamma \mid \gamma \geqq 0; \quad g(x) \to 0, \Omega(t) \to \gamma \ (t \to \infty)\}$$
$$\gamma_- = \inf \{\gamma \mid \gamma < 0; \quad g(x) \to 0, \Omega(t) \to \gamma \ (t \to \infty)\}.$$

Since this latter class contains the class of (i), γ_p serves as a lower bound on ξ_c.

The capture and pull-in range as well as the steady-state phase error are significant as measures of the effectiveness of the system to correct for input deviations. This synchronizing capability is of prime importance; however, in the treatment here solutions $x(t)$ are not of direct interest.

III. PHASE PLANE REPRESENTATION

The phase plane is perhaps the most natural means of describing the behavior of the phase-controlled system. As shown in Fig. 2, the solution trajectories in the (g,x) plane assume two essential forms: Those approaching singular or critical points x_0 on the abscissa represent the capture phenomenon; those running on indefinitely, the noncapture or nonsynchronous state. The values of x_0 are determined directly from (11) and (12).[4,5] For $\Omega(t) = \xi \ (t \geqq 0)$,

$$C(x_0) = \xi \quad g(x_0) = 0. \tag{21}$$

Similarly, for $\Omega(t) \to \gamma \ (t \to \infty)$,

$$C(x_0) = \gamma, \qquad g(x_0) = 0 \quad (t \to \infty). \tag{22}$$

Since in both of these instances $\mid C(x_0) \mid \leqq 1$, a necessary condition for the presence of critical points and, consequently, capture and pull-in is that

$$\mid \xi \mid \leqq 1, \qquad \mid \gamma \mid \leqq 1. \tag{23}$$

As regards the nature of these points, it is shown in Appendix A that x_0 is either a stable or saddle point according as $(dC/dx)_{x_0}$ is positive or negative. Borderline cases, e.g., those involving the condition

$$(dC/dx)_{x_0} = 0$$

and the coalescence of stable and saddle points, are not of physical

interest as infinitesimal variations of the system parameters eliminate the borderline properties. In the event that $C(x)$ is discontinuous, the above criteria are applied to a regular sequence, i.e., a sequence $\{C_n(x)\}_0^\infty$ of continuous and differentiable functions for which

$$C_n(x) \rightarrow C(x) \ (n \rightarrow \infty),$$

the desired conditions obtained in the limit.

IV. TEST TRAJECTORIES — A VARIATIONAL APPROACH

At this point we describe a method with which to obtain upper and lower bounds on the relative capture range. With reference to the generalized phase plane portrait of Fig. 3, the capture and noncapture trajectories are regarded as terminating on two different sets of points M and N, respectively, and emanating from various regions of the initial state domain D, e.g., D_M and D_N. Terminal points in N represent points at infinity; those in M, critical points. Saddles are included in the latter group because a trajectory which approaches such a point involves an infinite amount of time and therefore satisfies capture conditions (17) to (19). Stated in the most general manner, the basic problem to be treated here is the determination of contours Γ_c, namely, boundaries separating regions in D from which either capture or noncapture trajectories originate. For $\Omega(t) = \xi$, domain D becomes the g axis and contours Γ_c degenerate to points.

As a technique for locating Γ_c, we first construct "test trajectories," i.e., functions $h(x)$, which run from D to a position somewhere between sets M and N. The test paths are selected so that if the solution paths are known to lie entirely on either side of the former, then the latter must terminate on points in the terminal set lying on the same side. It is then

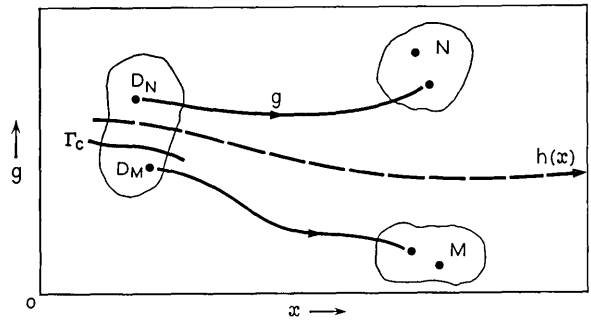


Fig. 3 — Generalized phase plane portrait.

possible to show by means of comparison theorems that for certain regions of the initial state domain positioned on one side of a test trajectory, the corresponding solution trajectories are similarly positioned. Consequently, the regions in D so determined are identified as either capture or noncapture domains forming bounds on Γ_c.

To make the above procedure precise, we utilize the following comparison theorem: Let $g(x(t))$ be a trajectory branch satisfying the autonomous equation

$$\frac{dg}{dx} = F(x,g) \quad (x \varepsilon I, t \varepsilon J)$$

where I denotes a closed interval of $x(x_a \leqq x \leqq x_b)$ and J, a branch interval of t defined by the inverse function $t = t(x)$ $(x \varepsilon I)$. If for a test trajectory $h(x)$,

$$\begin{aligned} \frac{dh}{dx} &< (>)\ F(x,h) \quad (x \varepsilon I) \\ h(x_a) &\leqq (\geqq)\ g(x_a) \\ h(x), g(x), t(x) &\varepsilon C \quad (x \varepsilon I, t \varepsilon J) \end{aligned} \tag{24}$$

then

$$h(x) < (>)\ g(x), \qquad x_a < x \leqq x_b . \tag{25}$$

A brief proof of (24) and (25) is outlined in Appendix B. Although normally restricted to continuous functions, this theorem can be applied to generalized functions through the related regular sequences.

With $g(x_a)$ positioned in the initial state domain and the test trajectory constructed as a family of functions involving parameters $\lambda_1, \cdots, \lambda_n$, i.e., $h(x,\lambda)$, the inequalities of (24) can usually be reduced to the form

$$g(x_a) > (<)\ G(x,\lambda) \tag{26}$$

where the inequalities are sometimes reversed. Since the G function is known, there are obtained from (26) regions of domain D for which the resulting trajectories lie either above or below those of the test functions; if in the sense described above, sets M and N are isolated by $h(x,\lambda)$, then these regions constitute portions of the capture and noncapture domains to be determined. The optimal bounds on $g(x_a)$ for the given family $h(x,\lambda)$ are derived by extremalizing G with respect to x and λ:

$$g(x_a) \quad \begin{cases} > g_U = \inf\limits_{\lambda} \cdot \sup\limits_{x} G(x,\lambda) \\ < g_L = \sup\limits_{\lambda} \cdot \inf\limits_{x} G(x,\lambda) \end{cases} . \tag{27}$$

The quality of these bounds depends on that of the test function in that the closer $g(x)$ is approximated by $h(x)$, the better the contours Γ_c are delineated. Unlike the functions associated with Liapounoff's second method,[4,5] $h(x)$ is related directly to whatever properties of the actual trajectory are known and is constructed so as to conform as closely as necessary to these properties.

V. GENERAL SYMMETRIC COMPARATOR — RC FILTER

In this section we apply the method of test trajectories to phase-controlled systems employing RC filters and symmetric comparators, viz., those defined by the relations

$$C(x) = C(x + 2n) \quad (n = 0, \pm 1, \cdots,) \tag{28}$$

$$C(x) = -C(-x) \tag{29}$$

$$C(x) \geqq 0 \quad (0 \leqq x \leqq 1) \tag{30}$$

where in accordance with (3) to (5) and (6)

$$C(0) = 0 \tag{31}$$

$$C(x) \leqq 1 \quad (0 \leqq x \leqq 1). \tag{32}$$

For the RC filter $\sigma_2 \to \infty$ ($\beta = 1$) in (14); therefore,

$$g \frac{dg}{dx} = \Omega + \frac{1}{\sigma_1} \frac{d\Omega}{dt} - C(x) - kg. \tag{33}$$

5.1 *Relative Capture Range*

With $\Omega(t) = \xi$, (33) reduces to

$$\frac{dg}{dx} = \frac{1}{g} [\xi - kg - C(x)] = F(x,g). \tag{34}$$

It is noted in this expression that trajectories corresponding to inputs ξ and $-\xi$ are symmetric about the origin of the phase plane in that the respective solutions $g_\xi(x)$ and $g_{-\xi}(x)$ are related by

$$g_\xi(x) = -g_{-\xi}(-x).$$

Consequently, both solutions exhibit the same capture characteristics, and

$$\xi_+ = -\xi_- \tag{35}$$

or

$$\xi_c = 2\xi_+ = -2\xi_- . \tag{36}$$

As a result, only positive input deviations need be considered. Referring to (15), one obtains for the initial state line

$$g(0) = \frac{\xi}{k} \geqq 0. \tag{37}$$

In order to establish crude noncapture intervals on this line, we employ the following test function:

$$h_1(x) = \sum_{n=0}^{\infty} \mu_1(x - 2n) \quad (0 \leqq x < \infty) \tag{38}$$

where

$$\mu_1(x) = \begin{cases} \frac{\lambda}{k}\left[1 - \frac{1}{\rho}\int_0^x C(z)\,dz\right]^{\frac{1}{2}} & (-1 \leqq x \leqq 1) \\ 0 & (|\,x\,| \geqq 1) \end{cases}$$

$$\rho = \int_0^1 C(z)\,dz.$$

The functional form chosen for $h_1(x)$, as well as most of those to follow, serves two basic aims:

(*i*) In addition to having the general appearance of a noncapture trajectory, $h_1(x)$ isolates the set of critical points on the x axis from the region above the test trajectory; i.e., solution trajectories which lie above $h_1(x)$ are bounded away from the x axis (cf. Section IV).

(*ii*) With respect to factor $\mu_1(x)$, $h_1(x)$ relates closely to the algebraic structure of (34).

We first verify property (*i*). By (28) to (32),

$$\rho = \sup_x \int_0^x C(z)\,dz\,.$$

Thus

$$\frac{1}{\rho}\int_0^x C(z)\,dz \leqq 1$$

and

$$\mu_1(x) \geqq 0$$

whence

$$h_1(x) \geqq 0 \quad (0 \leqq x \leqq \infty). \tag{39}$$

The comparison theorem applied to (34), (38) and (39) insures that

$$0 \leqq h_1(x) < g(x) \quad (x > 0)$$

for all values of ξ satisfying the inequalities

$$\xi > C(x) + kh_1 + h_1 \frac{dh_1}{dx} = \left(1 - \frac{\lambda^2}{2\rho k^2}\right) C(x) + kh_1(x) = G(x,h_1) \quad (x \geqq 0) \tag{40}$$

$$h_1(0) \leqq g(0) = \frac{\xi}{k}. \tag{41}$$

Since $C(0) = 0$, condition (41) is contained implicitly in (40), the latter expression corresponding to that of (26).

Functions $g(x)$ as trajectories bounded away from the x axis constitute noncapture solutions; hence, any ξ interval consistent with (40) is a noncapture interval. For obtaining rough limits, we set $\lambda^2 = 2\rho k^2$. Inequality (40) then becomes

$$\xi > kh_1(x) \quad (x \geqq 0).$$

Therefore, as indicated by (27) and (35), all values of ξ which satisfy the relation

$$\xi > \xi_1 = \sup_{x \geqq 0} kh_1(x) = \sup_{|x| \leqq 1} k\mu_1(x) = k(2\rho)^{\frac{1}{2}} \tag{42}$$

lie in the noncapture interval with the extremes $+\xi_1$ and $-\xi_1$ representing exterior bounds on the capture range. Although directly significant, these limits are derived primarily to provide essential information in more general calculations.

For a refined treatment we construct a more complex test function $h_2(x)$: Let $g_0(x)$ denote the separatrix solution which runs from the initial point $(g_0(0), 0)$ through the last saddle in the interval nearest to the point $x = 1$, and let ξ_0 denote the related input deviation. We then define

$$h_2(x) = \sum_{n=0}^{\infty} \mu_2(x - 2n) \quad (0 \leqq x < \infty) \tag{43}$$

where

$$\mu_2(x) = \begin{cases} \mu_1(x) & (\lambda = \xi_0, -1 \leqq x < 0) \\ g_0(x) & (0 \leqq x \leqq x_0) \\ 0 & (x_0 \leqq x \leqq 1) \end{cases}.$$

In addition to assuming the existence and uniqueness of $g_0(x)$, we suppose further that $g_0(x) \geqq 0$ $(0 \leqq x \leqq x_0)$, for if $g_0(x) = k/\sigma_1 \, dx/dt < 0$,

then $x(t)$ decreases as time increases, and the trajectory is depicted receding from the point x_0 in opposition to the definition of $g_0(x)$; thus,

$$\mu_2(x) \geqq 0 \quad (-1 \leqq x \leqq 1)$$

and

$$h_2(x) \geqq 0 \quad (0 \leqq x < \infty). \tag{44}$$

Also,

$$h_2(0) = g_0(0) = \frac{\xi}{k}. \tag{45}$$

Therefore, we have the same sufficient requirements for noncapture values of ξ as those of (40) and (41):

$$\xi < C(x) + kh_2 + h_2 \frac{dh_2}{dx} = G(x,h_2) \tag{46}$$

$$h_2(0) = \frac{\xi_0}{k} \leqq g(0) = \frac{\xi}{k}. \tag{47}$$

To determine $\sup_{x \geqq 0} G(x,h_2)$, we first derive several conditions relative to x_0, ξ_0, and g_0. With the selection of g_0 as a proper solution, (34), (46) and (47) combine to yield

$$G(x,g_0) = \xi_0 \quad (0 \leqq x \leqq x_0). \tag{48}$$

We next let x_m represent the greatest critical point value in the interval $x_0 < x \leqq 1$; viz.,

$$x_m = \sup_i \{x_i \mid C(x_i) = \xi_0 ; \qquad x_0 < x_i \leqq 1\}.$$

Inasmuch as x_0 is required by definition to be the only saddle in the interval $x_0 \leqq x \leqq 1$, x_m must be a stable point; then, $(dC/dx)_{x_m} > 0$, as discussed in Section III, with the result that $C(x) > C(x_m)$ for $x_m \leqq x \leqq 1$. However, since $C(x)$ as an odd function satisfying

$$C(1) = \tfrac{1}{2}\{C(1^+) + C(1^-)\} = C(0) = 0$$

must either approach zero or jump discontinuously to zero as $x \to 1$. x_m cannot exist; hence,

$$C(x) \neq \xi_0 (x_0 < x \leqq 1), \quad C(x_0) = \xi_0$$

and

$$\left(\frac{dC}{dx}\right)_{x_0} < 0.$$

These conditions imply that

$$C(x) \leqq C(x_0) \quad (x_0 \leqq x \leqq 1)$$

whence

$$\sup_{x_0 \leqq x \leqq 1} G(x,0) = \sup C(x) = C(x_0) = \xi_0 . \tag{49}$$

In addition, it is noted that because of the infinite amount of time for the g_0 trajectory to reach the saddle point, ξ_0 lies within the capture range and outside that defined by inequality (42); i. e.,

$$\xi_0 \leqq \xi_1 = k(2\rho)^{\frac{1}{2}}.$$

Consequently,

$$\begin{aligned} \sup_{-1 \leqq x < 0} G(x,\mu_1) &= \sup \left[\left(1 - \frac{\xi_0^2}{2\rho k^2} \right) C(x) + k\mu_1(x) \right] \\ &= \sup [k\mu_1(x)] \\ &= \xi_0 . \end{aligned} \tag{50}$$

As a result of (48), (49) and (50), the noncapture range is given by

$$\xi > \sup_{x \geqq 0} G(x,h_2) = \sup_{|x| \leqq 1} G(x,\mu_2) = \xi_0 . \tag{51}$$

That values of ξ less than ξ_0 are in the capture range is shown by reversing the inequalities in (46) and (47) and invoking topological properties of the phase plane portrait: Values of ξ for which

$$\xi < G(x,g_0) = \xi_0 \quad (0 \leqq x \leqq x_0) \tag{52}$$

correspond to solutions $g(x)$ which satisfy the relation

$$g(x) < g_0(x) \tag{53}$$

a condition indicating that for such ξ the associated solution trajectories either terminate on stable nodes or cross the x axis at a point $x = \hat{x} < x_0$ and proceed in the $-x$ direction. As shown by the phase portrait of Fig. 4, the crossover point $\hat{C}$ in the latter case must be located so that a saddle S lies just to the right and a focus F just to the left, for the trajectory would otherwise intersect either itself or a separatrix of S. Using the same argument proves that $g(x)$ must in addition spiral in toward F. Consequently, (52) is a capture condition and $\pm\xi_0$ are exact exterior bounds on the capture range. By (23) and (36)

$$\xi_c = 2\xi_0 \leqq 2 \tag{54}$$

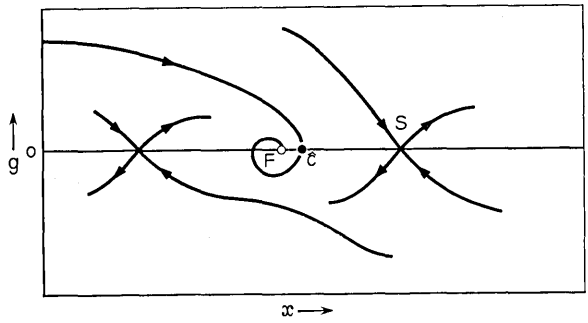


Fig. 4 — Phase portrait of trajectories.

or

$$\xi_+ = \xi_0 \leqq 1. \tag{55}$$

Two significant results have so far been obtained:

(*i*) The possibility of the system's reaching a synchronous state in a region outside the first domain does not exist; i.e., the steady state phase error x_{ss} must be such that

$$x_{ss} \epsilon \left\{ x_i \mid \xi = C(x_i); \qquad \left(\frac{dC}{dx}\right)_{x_i} > 0; \qquad 0 \leqq x_i \leqq 1 \right\}. \tag{56}$$

(*ii*) The determination of the capture range depends solely on that of the separatrix solution $g_0(x)$.

Among the set of stable points in (56) the appropriate x_i, namely, the exact value of x_{ss}, can be found by the application of the above procedure to each separatrix spanning the intervals $0 \leqq \xi/k \leqq \xi_0/k$ and $0 \leqq x \leqq x_0$ on the initial state line and x axis, respectively. For example, the first stable point $x_1 = \inf \{x_i\}$ is isolated by the separatrix which connects the points $\hat{\xi}/k$ and

$$\hat{x} = \inf \left\{ x_j \mid \hat{\xi} = C(x_j); \qquad \left(\frac{dC}{dx}\right)_{x_j} < 0; \qquad 0 \leqq x_j \leqq x_0 \right\},$$

in that the solutions for all $\xi < \hat{\xi}$ ($\xi > 0$) lie below this separatrix and $x_{ss} = x_1$. Each x_i can be handled in this manner.

In general, g_0 cannot be determined exactly; however, as regards specific comparators, graphical methods are found to yield results as accurate as one desires because of the relatively small distance traversed by the associated trajectory. The use of such graphical approximations for obtaining the capture range is prevalent in the literature in advance of any formal justification. It is important to point out that even for

the RC filter case, interior and exterior bounds are sometimes more convenient and useful than the graphical solutions.

5.2 *Asymptotic Results for Small and Large k*

It is possible in the case of vanishing k to derive explicit relations for both the capture range and $g_0(x)$. We first consider an exterior bound formulation based on the test function

$$h_3(x) = \sum_{n=0}^{\infty} \mu_3(x - 2n) \quad (x \geqq 0) \tag{57}$$

where

$$\mu_3(x) = \begin{cases} \left[2\rho - 2\int_0^x C(z)\,dz\right]^{\frac{1}{2}} & (-1 \leqq x \leqq x) \\ 0 \quad (|x| \geqq 0). \end{cases}$$

Note that

$$\begin{aligned} \mu_3(1) &= 0 \\ \mu_3(x) &> 0 \quad (-1 < x < 1) \\ h_3(x) &\geqq 0 \quad (x \geqq 0) \\ h_3(0) &= (2\rho)^{\frac{1}{2}} \end{aligned}$$

in accordance with noncapture criteria (cf. (40) et seq.). Hence, as in (40), (41) and (42), exterior bounds $\pm\xi_2$ are given by the expression

$$\begin{aligned} \xi > \xi_2 &= \sup_{x \geqq 0} G(x,h_3) = \sup_{|x| \leqq 1} G(x,\mu_3) \\ &= \sup_{|x| \leqq 1} \left[2\rho k^2 - 2k^2 \int_0^x C(z)\,dz\right]^{\frac{1}{2}} \\ &= k(2\rho)^{\frac{1}{2}} \end{aligned} \tag{58}$$

which is consistent with the initial condition of (41); viz.,

$$h_3(0) = (2\rho)^{\frac{1}{2}} \leqq g(0) = \frac{\xi}{k}.$$

To treat the interior bound calculation, we use

$$h_4(x) = \left[\frac{\lambda^2}{k^2} + 2\lambda x - 2\int_0^x C(z)\,dz\right]^{\frac{1}{2}} \quad (0 \leqq x \leqq \bar{x} \leqq 1) \tag{59}$$

where

$$h_4(\bar{x}) = 0$$

$$h_4^2(x) > 0 \quad (0 \leqq x < \bar{x} \leqq 1)$$

and λ equals the largest positive value for which the quantity in the above brackets has one zero in x in the interval $0 \leqq x \leqq 1$. Such a value must exist since $\int_0^x C(z)dz$ $(x \leqq x \leqq 1)$ is monotonic increasing. In addition, for this zero condition to hold independently of k, $\lambda \to k(2\rho)^{\frac{1}{2}}$ and $\bar{x} \to 1$ as $k \to 0$; thus,

$$h_4(x) \underset{k\to 0}{\to} \left[2\rho - 2\int_0^x C(z)\,dz\right]^{\frac{1}{2}} \quad (0 \leqq x \leqq \bar{x}). \tag{60}$$

Test function $h_4(x)$ is seen to satisfy the same requirements as those discussed in connection with (52) through (55). Accordingly, interior bounds $\pm\xi_3$ are given by the relation

$$\begin{aligned} \xi < \xi_3 &= \inf_{x<\bar{x}} G(x,h_4) = \inf\left\{\lambda + k\left[\frac{\lambda^2}{k^2} + 2\lambda x - 2\int_0^x C(z)\,dz\right]^{\frac{1}{2}}\right\} \\ &= \lambda \underset{k\to 0}{\to} k(2\rho)^{\frac{1}{2}} = \xi_2 \end{aligned} \tag{61}$$

which is also consistent with the initial condition

$$h_4(0) = \frac{\lambda}{k} \geqq g(0) = \frac{\xi}{k}.$$

Inasmuch as $\xi_3 \leqq \xi_+ \leqq \xi_2$ and $\xi_3 \to \xi_2 (k \to 0)$, the results of (58), (60), and (61) yield

$$\begin{aligned} \xi_+ = \frac{\xi_c}{2} &\sim k(2\rho)^{\frac{1}{2}} \qquad (k \to 0) \\ g_0(x) &\sim \left[2\rho - 2\int_0^x C(z)\,dz\right]^{\frac{1}{2}} \quad (k \to 0) \end{aligned} \tag{62}$$

where

$$\rho = \int_0^1 C(z)\,dz.$$

Therefore, the choice of asymptotically equal test functions results in coalescent bounds.

In considering large values of k, we take

$$h_5(x) = \hat{x} + \lambda - x \quad (0 \leqq x \leqq \hat{x} + \lambda < 1) \tag{63}$$

where

$$0 \leqq C(x) < 1 \quad (0 \leqq x < \hat{x} \leqq 1)$$
$$C(\hat{x} + \lambda) \underset{\lambda \to 0}{\to} 1$$
$$\hat{x} > 0 \qquad \hat{x} + \lambda > 0$$

and $\lambda \gtreqless 0$ according as $(dC/dx)_{\hat{x}} \gtreqless 0$. Clearly, λ is employed only in the event that $C(x)$ has a discontinuity at $x = \hat{x}$. Since the values of ξ specified by

$$\xi < G(x,h_5) = h_5 \frac{dh_5}{dx} + 2kh_5 + C(x)$$

also satisfy the basic initial criterion

$$kg(0) = \xi < G(0,h_5(0)) = -(\hat{x} + \lambda) + kh_5(0) \leqq kh_5(0)$$

interior bounds $\pm\xi_4$ are given by

$$\xi < \xi_4 = \inf_{x < \hat{x}+\lambda} G(x,h_5) = \inf [(k-1)(\hat{x} + \lambda - x) + C(x)]. \tag{64}$$

Employing the limits

$$G(x,h_5) \underset{k \to \infty}{\to} \infty \quad (x < \hat{x} + \lambda)$$
$$G[(\hat{x} + \lambda), h_5(\hat{x} + \lambda)] = C(\hat{x} + \lambda)$$

yields

$$\operatorname*{Lim}_{\lambda \to 0} \operatorname*{Lim}_{k \to \infty} \xi_4 = \operatorname*{Lim}_{\lambda \to 0} C(\hat{x} + \lambda) = C(\hat{x}) = 1 \leqq \xi_+ .$$

However, according to (23), $\xi_+ \leqq 1$; therefore,

$$\xi_+ = \frac{\xi_c}{2} \underset{k \to \infty}{\to} 1. \tag{65}$$

One can show in a similar manner that if the system captures for a given ξ and k, a capture condition results for all smaller values of k. Consequently, the relative capture as a function of k^{-2} assumes the general form depicted in Fig. 5 with the asymptotic behavior given by (62) and (65).

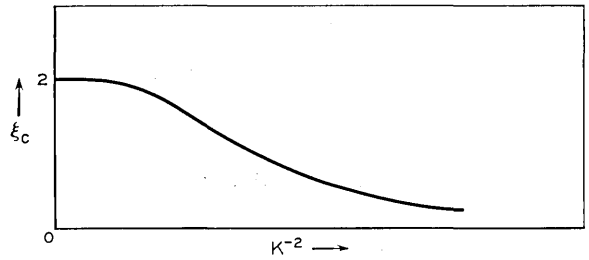


Fig. 5 — General form of relative capture as a function of k^{-2}.

VI. SAWTOOTH COMPARATOR — RC FILTER

Directly related to the formulation and results of Section V is the RC filter-sawtooth comparator system in which

$$C(x) = \sum_{n=-\infty}^{\infty} C_0(x - 2n) \quad (-\infty < x < \infty)$$

$$C_0(x) = \begin{cases} x \ (-1 < x < 1) \\ 0 \ | \ x \ | \geqq 1 \end{cases} \tag{66}$$

$$f_m = \pi, \qquad k = \left(\frac{\sigma_1}{\alpha}\right)^{\frac{1}{2}}, \qquad \beta = 1.$$

The relatively simple form assumed by $C(x)$ in the first domain enables one to derive exact expressions for both $g_0(x)$ and ξ_0. It is shown in Appendix C (cf. (99)) that

$$\xi_0 = \xi_+ = \frac{\xi_c}{2} = \begin{cases} \dfrac{k}{k + \exp\left[\dfrac{k}{\tau_0}\left(\tan^{-1}\dfrac{\tau_0}{k} - \pi\right)\right]} & (k < 2) \\ 1 \quad (k \geqq 2) \end{cases} \tag{67}$$

where $\tau_0^2 = (4 - k^2)$ and $0 < \tan^{-1}(\cdot) < \pi/2$. For small k, (67) can be expressed as

$$\xi_0 \sim 2k = 2\sqrt{\frac{\sigma_1}{\alpha}} \quad \left(k, \frac{\sigma_1}{\alpha} \to 0\right) \tag{68}$$

which checks with (62). In addition, from (20 et seq.)

$$\gamma_p \leqq \xi_c \,. \tag{69}$$

An exact derivation of γ_p valid for the sawtooth and general filter case has been obtained by A. J. Goldstein.[6]

As regards steady state phase error, relations (21) and (56) give

$$x_{ss} = C^{-1}(\xi_0)\,|_{|x|\leqq 1} = \frac{\xi_c}{2}. \tag{70}$$

VII. SAWTOOTH COMPARATOR — LAG FILTER

The capture results as developed in Section V do not apply completely to the lag filter; however, the calculation of exterior and interior bounds in this case follows closely that of (40), (41) and (52). For $\Omega(t) = \xi$, $\beta > 1$, and $C(x)$ specified by (66), (14) reduces to

$$\begin{aligned}\xi &= C(x) + \beta kg + g\frac{dg}{dx} - 2(\beta - 1)kg \sum_n \delta\{x - (2n+1)\} \\ &= G_1(x,g)\end{aligned} \tag{71}$$

where

$$k = \left(\frac{\sigma_1}{\alpha}\right)^{\frac{1}{2}}$$

and

$$\beta = 1 + \frac{\alpha}{\sigma_2}.$$

Using test function $h_1(x)$ of (38) yields the following noncapture values of ξ (cf. (40)):

$$\xi > G_1(x,h_1)(x \geqq 0)$$

where, as required, $kg(0) = \xi > k\beta h_1(0) > kh_1(0)$. Hence, exterior bounds $\pm\xi_U$ are given by

$$\begin{aligned}\xi_U &= \sup_{x\geqq 0,\lambda\geqq 0} G_1(x,h_1) = \sup_{|x|\leqq 1,\lambda\geqq 0} G_1(x,\mu_1) \\ &= \begin{cases} \dfrac{\beta k}{2}(4 - \beta^2k^2)^{\frac{1}{2}} & (\beta k < \sqrt{2}) \\ 1 & (\beta k \geqq \sqrt{2}). \end{cases}\end{aligned} \tag{72}$$

Using separatrix solution $g_L(x)$ of Appendix C as a test function yields the following capture values of ξ (cf. (52) and (98) et seq.):

$$\xi < G_1(x,g_L) = \xi_L \quad (0 \leqq x \leqq 1) \tag{73}$$

where

$$\xi_L = \begin{cases} \dfrac{k}{k + [1 - k^2(\beta - 1)]^{\frac{1}{2}} \exp\left\{\dfrac{\beta k}{\tau_1}\left[\tan^{-1}\dfrac{\tau_1}{(2-\beta)k} - \pi\right]\right\}} & (\beta k < 2) \\ 1 & (\beta k \geqq 2) \end{cases}$$

$$\tau_1^2 = (4 - \beta^2 k^2)$$

and

$$0 < \tan^{-1}(\cdot) < \pi.$$

Therefore,

$$\xi_L \leqq \frac{\xi_c}{2} \leqq \xi_U . \tag{74}$$

With regard to small k, the relations for ξ_U and ξ_L become

$$\xi_U \sim \beta k \quad (k \to 0) \tag{75}$$

$$\xi_L \sim k \quad (k \to 0). \tag{76}$$

As reflected by parameter β, the spread between ξ_U and ξ_L for small k appears significant, but in most cases involving small k, $\sigma_2 \gg \alpha$ and $\beta \cong 1$. Under more general conditions these bounds can be refined by constructing more complex test functions.

According to the discussion in Section II (cf. (20) et seq.), ξ_U serves also as an upper bound on the pull-in range; viz.,

$$\gamma_p = 2\gamma_+ \leqq 2\xi_U . \tag{77}$$

VIII. SINE COMPARATOR — RC FILTER

The sine-RC system is perhaps the most important practically, having received a great deal of attention in the literature. Several investigators have obtained capture results based on phase plane constructions of $g_0(x)$ for large and medium values of k.[1,7] For small k, approximate results have been derived by both analytic and experimental means, although these approximations are found to differ somewhat.[1,2] In this section we discuss the exact asymptotic form of the relative capture range for small k.

The normalized sine comparator is represented by

$$C(x) = \sin(\pi x) \quad (-\infty < x < \infty) \tag{78}$$

whence

$$\rho = \int_0^1 C(x)\, dx = \frac{2}{\pi}$$

$$f_m = 1$$

and

$$k = \left(\frac{\pi\sigma_1}{\alpha}\right)^{\frac{1}{2}}.$$

Consequently, the first relation of (62) can be written as

$$\xi_+ = \frac{\xi_c}{2} \sim \frac{2}{\sqrt{\pi}}\, k = 2\sqrt{\frac{\sigma_1}{\alpha}} \quad \left(k, \frac{\sigma_1}{\alpha} \to 0\right) \tag{79}$$

or

$$\xi_+ \sqrt{\frac{\alpha}{\sigma_1}} \to 2 \quad \left(\frac{\sigma_1}{\alpha} \to 0\right). \tag{80}$$

The numerical limit in (80) has been estimated by McAleer[2] and Jelonek[1] to be 1.00 and 1.27, respectively. As lower bounds, these approximations differ appreciably.

IX. SUMMARY

The more significant results of the above development can be outlined as follows:

(*i*) The variational formulation given by (27), a generalization of which allows one to treat similar problems in n-dimensional phase space.

(*ii*) Bounds given by (42) and (61), yielding respectively noncapture and capture ranges for the general symmetric comparator and the simple RC filter.

(*iii*) The first domain phenomenon associated with the symmetric-RC case of (*ii*) (cf. (56) et seq.).

(*iv*) The exact asymptotic capture relations given by (62) for the symmetric-RC case.

(*v*) The pull-in criteria given by (20 et seq.), (42), (61), and (62) for the symmetric-lag and RC cases.

(*vi*) The specific capture relations given by Sections VI through VIII for the sawtooth-sine and lag-RC cases.

(*vii*) The asymptotic relation given by (80) for small k, simple RC filters, and sine comparators.

APPENDIX A

Critical Point Type

To determine the behavior of trajectories near a critical point

$$(x = x_0, \qquad g = 0),$$

we linearize (11) and (12) about this point and investigate the natural modes of the corresponding linear solutions: As shown by Liapounoff,[5] these modes identify the type of critical point in that the presence of at least one exponentially increasing component indicates a saddle point, and the presence of only decreasing components, a stable (nodal or focal) point. Equations (11) and (12) are linearized by expanding the right-hand sides in a double Taylor series in x and g about the point

$$(x = x_0, \qquad g = 0),$$

and dropping all terms beyond the first order. Accordingly, we obtain from (11), (12), (21) and (22)

$$\begin{aligned} \frac{dx}{dt} &= ag \\ \frac{dg}{dt} &= b_1 x + b_2 g \end{aligned} \tag{81}$$

where

$$\begin{aligned} a &= \frac{\sigma_1}{k} \\ b_1 &= -\frac{\sigma_1}{k}\left(\frac{dC}{dx}\right)_{x_0} \\ b_2 &= \sigma_1\left[(1-\beta)\left(\frac{dC}{dx}\right)_{x_0} - 1\right]. \end{aligned}$$

Equations (81) combine to yield

$$\frac{d^2x}{dt^2} - b_2\frac{dx}{dt} - ab_1 x \tag{82}$$

the exponential components of the related solutions increasing or decreasing according as the roots of the secular equation

$$\lambda^2 - b_2\lambda - ab_1 = 0 \tag{83}$$

are positive or negative. Moreover, if $(dC/dx)_{x_0} > 0$, then $b_1 < 0$, $b_2 < 0$, and

$$\lambda = \frac{b_2}{2} \pm \left[\frac{b_2^2}{4} + ab_1\right]^{\frac{1}{2}} < 0. \tag{84}$$

Also, if $dC/dx_{x_0} < 0$, then $b_1 > 0$ and at least one value of λ given by (84) is positive; i. e.,

$$\lambda = \frac{b_2}{2} + \left[\frac{b_2^2}{4} + ab_1\right]^{\frac{1}{2}} > 0.$$

Hence, the critical point in question is stable provided $(dC/dx)_{x_0} > 0$ and a saddle provided $(dC/dx)_{x_0} < 0$.

APPENDIX B

Comparison Theorem

Theorem

If for a branch trajectory $g(x(t))$ and a test trajectory $h(x)$,

$$\begin{aligned} \frac{dg}{dx} &= F(x,g) \quad (x \in I, t \in J) \\ \frac{dh}{dx} &< (>) F(x,h) \quad (x \in I) \\ h(x_a) &\leqq (\geqq) g(x_a) \\ h(x), \quad g(x), \quad t(x) &\in C \quad (x \in I, t \in J), \end{aligned} \tag{85}$$

where I denotes a closed interval of $x(x_a \leqq x \leqq x_b)$ and J, a branch interval of t defined by the inverse function $t = t(x)$ $(x \in I)$, then

$$h(x) < (>) g(x) \quad (x_a < x \leqq x_b). \tag{86}$$

Proof

Assume first that $h(x_a) < g(x_a)$ and $dh/dx < F(x,h)$ $(x \in I)$. With h and g continuous, either $h(x) < g(x) (x \in I)$ or the two trajectories intersect at some point $\bar{x}$ in the interval, i.e., $h(\bar{x}) = g(\bar{x}) (x_a < \bar{x} \leqq x_b)$; however, an intersection implies that $(dh/dx)_{\bar{x}} \geqq (dg/dx)_{\bar{x}}$, a condition which contradicts the combined relation

$$\left(\frac{dh}{dx}\right)_{\bar{x}} < F(\bar{x},h(\bar{x})) = F(\bar{x},g(\bar{x})) = \left(\frac{dg}{dx}\right)_{\bar{x}}. \tag{87}$$

Therefore, in this case,

$$h(x) < g(x) \quad (x \varepsilon I).$$

If

$$h(x_a) = g(x_a)$$

and

$$\frac{dh}{dx} < F(x,h) \quad (x \varepsilon I)$$

$$\left(\frac{dh}{dx}\right)_{x_a} < F(x_a, h(x_a)) = F(x_a, g(x_a)) = \left(\frac{dg}{dx}\right)_{x_a}.$$

Consequently, $h(x) < g(x)$ over some portion of I. Ruling out additional intersections by (87), we obtain the general result

$$h(x) < g(x) \quad (x_a < x \leqq x_b).$$

The alternative set of inequalities is shown to be valid in a similar manner.

APPENDIX C

First Domain Trajectories in the Sawtooth Case

For treating the sawtooth comparator-lag filter case, we seek here solution trajectories of (71) which span the initial state line and saddle points in the first domain, viz., $0 \leqq x \leqq 1$. As defined in (66),

$$C(x) = \begin{cases} x & (0 \leqq x < 1) \\ 0 & (x = 1). \end{cases} \tag{88}$$

Hence, by (21 et seq.) the only saddle present in the first domain is that at the point $(x = 1, g = 0)$. We intend, therefore, to determine a solution $g_L(x)$ spanning the two points $(0,\xi/k)$ and $(1,0)$. It is noted that in the simple RC case, $g_L(x)$ is identical to $g_0(x)$ and ξ_L, to ξ_0 (cf. (43) and preceding).

Restricted to the interval $(x \leqq x \leqq 1)$ and solution $g_L(x)$, (71) becomes

$$\xi_L = x + \beta k g_L + g_L \frac{dg_L}{dx} - 2(\beta - 1) k g_L \delta(x - 1) \quad (0 \leqq x \leqq 1). \tag{89}$$

However, since $g_L(x)$ is required to vanish as $x \to 1$, $g_L\delta(x-1) = 0$; consequently,

$$\xi_L = x + \beta k g_L + g_L \frac{dg_L}{dx}. \tag{90}$$

Setting $z = x - \xi_L$ and $v = g_L/z$ reduces (90) to the form

$$\begin{aligned} 2\frac{dz}{z} &= -\frac{2v\,dv}{v^2 + \beta k v + 1} \\ &= \frac{\beta k}{\left(v + \frac{\beta k}{2}\right)^2 + \frac{\tau_1^2}{4}} - \frac{2v + \beta k}{v^2 + \beta k v + 1} \end{aligned} \tag{91}$$

where

$$\tau_1^2 = (4 - \beta^2 k^2).$$

Integrating, we have

$$\ln\left[g_L^2 + \beta k(x - \xi_L)g_L + (x - \xi_L)^2\right] = \frac{2\beta k}{\tau_1}\tan^{-1}\left[\frac{2g_L + \beta k(x - \xi_L)}{\tau_1(x - \xi_L)}\right] + C_3 \tag{92}$$

where $C_3 = \text{const}$. At the saddle point ($g_L = 0$, $x = 1$), (92) becomes

$$\ln(1 - \xi_L)^2 = \frac{2\beta k}{\tau_1}\tan^{-1}\left(\frac{\beta k}{\tau_1}\right) + C_3. \tag{93}$$

Therefore, g_L is given implicitly by the relation

$$\frac{g_L^2 + \beta k(x - \xi_L)g_L + (x - \xi_L)^2}{(1 - \xi_L)^2} = \exp\left\{\frac{2\beta k}{\tau_1}\left[\tan^{-1}\left(\frac{2g_L + \beta k(x - \xi_L)}{\tau_1(x - \xi_L)}\right) - \tan^{-1}\left(\frac{\beta k}{\tau_1}\right)\right]\right\}. \tag{94}$$

An explicit expression for ξ_L is derived by inserting the initial condition $g_L(0) = \xi_L/k$ in (94); i. e.,

$$\frac{\xi_L^2}{k^2}\left[\frac{1 - k^2(\beta - 1)}{(1 - \xi_L)^2}\right] = \exp\left\{\frac{2\beta k}{\tau_1}\left[\tan^{-1}\left(\frac{\beta k^2 - 2}{\tau_1 k}\right) - \tan^{-1}\left(\frac{\beta k}{\tau_1}\right)\right]\right\}. \tag{95}$$

For $\xi_L \leqq 1$ (cf. 23) and $\beta k < 2$ (cf. 91), the argument of the first arc-tan function in (94) takes on an infinite value at only one point in the interval $(0 \leqq x \leqq 1)$; i. e.,

$$\frac{2g_L + \beta k(x - \xi_L)}{\tau_1(x - \xi_L)} \rightarrow \begin{cases} \dfrac{\beta k}{\tau_1} \geqq 0 & (x \rightarrow 1) \\ +\infty & (x \rightarrow \xi_L^{+}) \\ -\infty & (x \rightarrow \xi_L^{-}) \\ \left(\dfrac{\beta k^2 - 2}{\tau_1 k}\right) \gtreqless 0 & (x \rightarrow 0). \end{cases}$$

Thus, the arc-tan difference in (95) must be such that

$$0 < \tan^{-1}\left(\frac{\beta k^2 - 2}{\tau_1 k}\right) - \tan^{-1}\left(\frac{\beta k}{\tau_1}\right) = \tan^{-1}\left[\frac{-\tau_1}{(2 - \beta)k}\right] < \pi. \quad (96)$$

Equation (95) can then be written as

$$\frac{\xi_L[1 - k^2(\beta - 1)]^{\frac{1}{2}}}{k(1 - \xi_L)} = \exp\left\{\frac{\beta k}{\tau_1}\left[\tan^{-1}\left(\frac{-\tau_1}{(2 - \beta)k}\right)\right]\right\} \quad (97)$$

or

$$\xi_L = \frac{k}{k + [1 - k^2(\beta - 1)]^{\frac{1}{2}} \exp\left\{\dfrac{\beta k}{\tau_1}\left[\tan^{-1}\left(\dfrac{\tau_1}{(2 - \beta)k}\right) - \pi\right]\right\}} \quad (98)$$

$$(\beta k < 2)$$

where $0 < \tan^{-1}(\cdot) < \pi$. That $\xi_L = 1$ for $\beta k \geqq 2$ is shown by considering two values of βk, say $\beta k_1 \rightarrow 2$ and $\beta k_2 > 2$, and two trajectories with respective initial conditions $\xi_1/k_1 = 1/k_1$ and $\xi_2/k_2 = 1/k_2$. Since from (98), $\xi_L \rightarrow 1 = \xi_1 = \xi_2$ as $\beta k_1 \rightarrow 2$, the former trajectory is a capture type. However, $1/k_1 > 1/k_2$; therefore, the latter initial point and trajectory lie below those of the former. This implies that relations $\xi_2 = 1$ and $\beta k_2 > 2$ correspond to capture conditions. On the other hand, if $\xi_2 > 1$, noncapture trajectories result. Hence, $\xi_L = \xi_2 = 1$ for $\beta k > 2$.

Finally, capture criteria in the simple RC case are given by

$$\xi_L \xrightarrow{\beta \rightarrow 1} \xi_0 = \begin{cases} \dfrac{k}{k + \exp\left[\dfrac{k}{\tau_0}\left(\tan^{-1}\dfrac{\tau_0}{k} - \pi\right)\right]} & (k < 2) \\ 1 & (k \geqq 2) \end{cases} \quad (99)$$

where

$$\tau_0^2 = (4 - k^2)$$

and

$$0 < \tan^{-1}(\cdot) < \pi/2.$$

APPENDIX D

List of Symbols

φ_i — input phase

φ_0 — output phase

$\epsilon = \varphi_i - \varphi_0$ — phase error

$x = \frac{\epsilon}{\pi}$ — normalized phase error

x_{ss} — normalized steady-state phase error in the synchronous state

$f(\epsilon)$ — comparator output

f_m — $\sup_{\epsilon} |f(\epsilon)|$

$C(x) = \frac{f(\pi x)}{f_m}$ — normalized comparator function

$$\rho = \int_0^1 C(z)\, dz$$

ω_c — free-running frequency of the local oscillator

$\Delta\omega(t)$ — input frequency deviation

$\Omega(t) = \frac{\Delta\omega(t)}{\alpha f_m}$ — normalized input frequency deviation

$C_1 = \text{const.}$ — magnitude of input frequency step

$\xi = \frac{C_1}{\alpha f_m}$ — relative input frequency step

ξ_+, ξ_- — relative positive and negative capture ranges, respectively

$\xi_c = \xi_+ - \xi_-$ — relative capture range

ξ_0 — relative input step corresponding to separatrix $g_0(x)$

$C_2 = \text{const.}$ — magnitude of asymptotic input frequency step

$\gamma = \dfrac{C_2}{\alpha f_m}$ — relative asymptotic input frequency step

γ_+ , γ_- — relative positive and negative pull-in ranges, respectively

$\gamma_\nu = \gamma_+ - \gamma_-$ — relative pull-in range

$H(t)$ — filter impulse response

$S(t)$ — filter output

$g = \dfrac{k}{\sigma_1}\dfrac{dx}{dt}$

$g_0(x)$ — primary first domain separatrix, as defined by the description prior to (43)

$\delta(t)$ — Dirac delta function

α — gain constant; viz., $\dfrac{d\varphi_o}{dt} = \omega_c + \alpha S(t)$

φ_1 , σ_2 — filter parameters (cf. Fig. 1(b))

$k = \left(\dfrac{\pi\sigma_1}{\alpha f_m}\right)^{\frac{1}{2}}$

$\beta = 1 + \dfrac{\alpha f_m}{\pi\sigma_2}$

— normalized system parameters

$*$ — convolution

s — Laplace transform variable; i.e., $\mathfrak{L}\cdot f(t) = F(s)$

REFERENCES

1. Jelonek, Z., Celinski, O., and Syski, R., *Pulling Effect in Synchronized Systems*, Monograph No. 79, The Institution of Electrical Engineers, 1953.
2. Mc Aleer, H. T., Proc. I.R.E., **47**, 1959, p. 1137.
3. Temple, G., Proc. Roy. Soc. (London), **A228**, 1955, p. 175.
4. Stoker, J. J., *Nonlinear Vibrations*, Interscience Publishers, Inc., New York, 1950.
5. Minorsky, N., "The Theory of Oscillations," in *Dynamics and Nonlinear Mechanics*, Vol. 2, John Wiley & Sons, Inc., New York, 1958.
6. Goldstein, A. J., to be published.
7. Celinski, O., *Investigation of Synchronization in Phase Lock Oscillators*, unpublished Diploma Thesis, Polish University College, London, 1950.

Ultimately Periodic Solutions to a Non-Linear Integrodifferential Equation

By V. E. BENEŠ

(Manuscript received July 6, 1961)

Tychonov's fixed point theorem is used to study the existence of ultimately periodic solutions of an integrodifferential equation that arises in the theory of the phase-controlled oscillator. The principal result describes conditions under which solutions of the equation exist which have a given ultimate period T, not necessarily the minimal period.

I. INTRODUCTION

Let $h(\cdot)$ be an integrable function of integral unity, vanishing for negative argument; let α, ω, and $x(0)$ be constants; and let $f(\cdot)$ be a periodic function of period (say) 2π. It is of interest to know what choices of $h(\cdot)$, α, ω, $x(0)$, and $f(\cdot)$ give rise to asymptotically periodic solutions of the equation

$$\dot{x} = \omega - \alpha \int_0^t h(t - u)f(x(u))\, du \quad (t \geqq 0). \tag{1}$$

This equation arises in the theory of various synchronization phenomena. (Cf. Refs. 1, 2, and 3 and references therein.) For example, the synchronous motor and the phase-controlled oscillator are devices often described by (1). For a specific physical application, we consider the phase-controlled loop depicted in Fig. 1 and described in detail by Goldstein,[3] q.v.

The system is described by the equations, in $t \geqq 0$,

$$\begin{aligned} \dot{\varphi}_o(t) &= \omega_c + \alpha v(t) \\ \dot{\varphi}_i(t) &= \omega_c + \omega(t) \\ x(t) &= \varphi_i(t) - \varphi_o(t) \\ v(t) &= \int_0^t h(t - u)f(x(u))\, du \end{aligned}$$

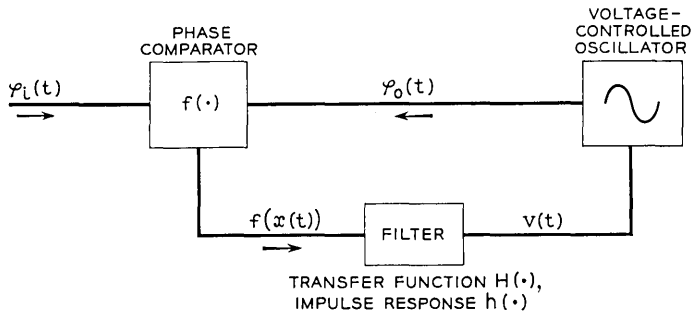


Fig. 1 — Phase-controlled loop.

where ω_c is the center or "free-running" frequency of the oscillator, $\varphi_o(\cdot)$ is the instantaneous phase of the output of a voltage-controlled oscillator, $\varphi_i(\cdot)$ is the instantaneous phase of an input signal (driving function, $h(\cdot)$ is the impulse response of a filter with dc gain unity, $f(\cdot)$ is a periodic phase comparator characteristic, and α is a gain constant. If the initial phase difference $x(0)$ is given, $\omega(t) = 0$ for $t < 0$, and $\omega(t) = \omega$ for $t \geqq 0$, we obtain (1); these conditions describe a sudden step of size ω in the frequency of the input signal.

Since the system described by (1) is autonomous, one may conjecture that a solution $x(\cdot)$ of (1) is always ultimately periodic, if we count identically constant functions as periodic. The present paper attempts to shed light on the question: what choices of ω give rise to an $x(\cdot)$ satisfying (1) with a *given* ultimate period T? We shall investigate solutions of (1) that are *ultimately periodic* in the following sense: a function $y(\cdot)$ is ultimately periodic with period T if there is a periodic function $p_y(\cdot)$ of period T such that

$$\lim_{n\to\infty} y(nT + t) = p_y(t) \quad (t \in [0,T]).$$

In this case we say that $y(\cdot)$ is u.p. $[T]$, and write $p_y(\cdot)$ for the periodic function approached by $y(\cdot)$. The number T is not necessarily the *minimal* period.

If $h(\cdot)$ is the Green's function of a differential operator of low order, the nature of solutions of (1) can be studied by the classical phase-plane method, as by Barnard.[2] To retain maximum generality and to exhibit (to some extent) the core of the problem, however, we shall use Tychonov's fixed point theorem. The possible novelty of our approach lies in using Tychonov's theorem to obtain specific asymptotic information about solutions $x(\cdot)$ of (1) by finding a fixed point (corresponding to a solution) in a relatively small region of a function space. This is achieved

by verifying some local properties of the operator whose fixed point is sought, and ensuring that a particular set is mapped into itself. A similar method has been used on (1) by the author in a previous paper[4] discussing the question whether $\dot{x}(t)$ approaches zero for large t.

II. PRELIMINARY CONSIDERATIONS

We rewrite (1) as the functional equation

$$g(t) = f\left(x(0) + \omega t - \alpha \int_0^t \eta(t-u)g(u)\,du\right) \quad (t \geqq 0) \qquad (2)$$

where

$$\eta(t) = \int_0^t h(u)\,du \quad (t \geqq 0)$$

$$\eta(\infty) = 1$$

$$g(t) = f(x(t)) \qquad (t \geqq 0)$$

and we seek an ultimately periodic solution $g(\cdot)$ of (2) in the space $B \frown C$ of bounded continuous functions. However, what periods $T > 0$ should be considered? If we choose a period T arbitrarily and define an operator J by

$$Jg(t) = \begin{cases} f\left(x(0) + \omega t - \alpha \int_0^t \eta(t-u)g(u)\,du\right) & (t \geqq 0) \\ f(x(0)) & (t \leqq 0) \end{cases}$$

then even if $g(\cdot)$ is ultimately periodic with period T, we have no guarantee that the image function $Jg(\cdot)$ is u.p. $[T]$, or that it is ultimately periodic at all. This circumstance is due to the presence of the constant ω in the definition of J, which has no immediate relation to $g(\cdot)$ or to a specific period T of interest.

We next observe heuristically that if $g(\cdot)$ is u.p. $[T]$, then $Jg(\cdot)$ can be u.p. $[T]$ only if ω bears a suitable relation to both the period 2π of $f(\cdot)$ and the desired period T. Roughly speaking, one effect of the integration in (2) is to subtract a linear term $c_0 t$ from the linear term ωt already present; the coefficient c_0 will be proportional to the mean (over a period T) of the periodic function $p_g(\cdot)$ to which $g(\cdot)$ is asymptotic; the remainder of the contribution of the integral will be u.p. $[T]$. It is intuitively clear that c_0 should have the form

$$c_0 = \frac{\alpha}{T} \int_0^T p_g(u)\,du.$$

Let $a(\cdot)$ be a function of period T. Then certainly the function defined by

$$f(\omega t - c_0 t + a(t))$$

will have period T if

$$\omega = c_0 + \frac{2n\pi}{T} \quad (n \text{ an integer}). \tag{3}$$

For in such a case

$$f[(\omega - c_0)t + (\omega - c_0)T + a(t + T)] = f[(\omega - c_0)t + a(t)].$$

In view of this, we shall allow the "constant" ω in (2) to depend on the function $g(\cdot)$ being mapped according to (3), for a fixed choice of n. That is, we define a transformation $A(\cdot)$ of $g(\cdot)$'s that are u.p. $[T]$ by

$$Ag(t) = \begin{cases} f\left(x(0) + \omega_g t - \alpha \int_0^t \eta(t-u)g(u)\,du\right) & (t \geqq 0) \\ f(x(0)) & (t \leqq 0) \end{cases} \tag{4}$$

$$\omega_g = \frac{2n\pi}{T} + \frac{\alpha}{T}\int_0^T p_g(u)\,du \qquad (n \text{ fixed}).$$

By this device we shall be able to consider an arbitrary period T.

III. SUMMARY OF HYPOTHESES AND RESULTS

If $h(\cdot)$ is the impulse response of a physically realizable network then $\eta(t) = \int_0^t h(u)\,du$ is its response to a unit step-function, and $\eta(\infty)$ is its dc gain, here taken to be unity. The function $\psi(t) = \eta(\infty) - \eta(t)$, $t \geqq 0$, is basic to much of our discussion, and is assumed to be absolutely integrable. The integral

$$\int_0^\infty [\eta(\infty) - \eta(u)]\,du = \int_0^\infty \psi(u)\,du \tag{5}$$

can be invested with physical meaning as follows: a partial integration,

$$\int_0^t uh(u)\,du = t\eta(t) - \int_0^t \eta(u)\,du$$

$$= t[\eta(t) - \eta(\infty)] + \int_0^t \psi(u)\,du$$

and the observation that $\eta(\infty) - \eta(t) = o(t^{-1})$ as $t \to \infty$, show that (5) is the "mean" of $h(\cdot)$, i.e.,

$$\int_0^\infty \psi(u)\, du = \int_0^\infty uh(u)\, du.$$

The integrable function $h(\cdot)$ has a Fourier transform $H(\cdot)$ defined by

$$H(s) = (2\pi)^{-\frac{1}{2}} \int_0^\infty e^{isu} h(u)\, du \quad (s \text{ real})$$

and the absolute convergence of (5) implies that

$$H'(0) = i(2\pi)^{-\frac{1}{2}} \int_0^\infty \psi(u)\, du.$$

The derivative of $H(\cdot)$ at $s = 0$ is closely related to the phase characteristic of the network whose impulse response is $h(\cdot)$. For upon representing $H(\cdot)$ as

$$H(s) = A(s)e^{i\varphi(s)}$$

we find

$$H'(s) = A'(s)e^{i\varphi(s)} + i\varphi'(s)H(s).$$

The amplitude characteristic $A(\cdot)$ is a real, even function, so $A'(0) = 0$. The phase characteristic is representable as an arctangent, so $\varphi'(0) \geqq 0$ in general. Since $A(0) = \eta(\infty)(2\pi)^{-\frac{1}{2}} = (\text{dc-gain})(2\pi)^{-\frac{1}{2}}$, we have

$$H'(0) = \frac{i}{(2\pi)^{\frac{1}{2}}} \times \text{dc-gain} \times \text{coefficient of } s \text{ in Taylor's expansion of } \varphi(\cdot) \text{ around zero}$$

$$= i(2\pi)^{-\frac{1}{2}} \int_0^\infty \psi(u)\, du.$$

We shall call $\varphi'(0)$, the derivative of the phase characteristic of $H(\cdot)$ at the origin, the delay of the network, so that

$$\int_0^\infty \psi(u)\, du = (\text{dc gain})(\text{delay}).$$

A condition on the function $\psi(\cdot)$, stronger than integrability, will be used. This is that

$$h_j = \sup_{0 \leqq t \leqq T} \int_{t+jT}^{t+(j+1)T} |\psi(u)|\, du \quad (j \geqq 0) \tag{6}$$

be an absolutely summable sequence. The magnitude of $\psi(\cdot)$ measures to a certain extent the speed of the response of the network whose impulse response is $h(\cdot)$; our condition requires that this speed be sufficient to make $\sum_{i \geqq 0} h_j$ finite.

We shall assume that the non-linear function $f(\cdot)$ is Lipschitz of order 1 with a constant β,

$$| f(x) - f(y) | \leqq \beta | x - y | .$$

To obtain an estimate of the rate of convergence of a solution $x(\cdot)$ of (1) to a periodic function, and to verify the compactness condition needed for use of Tychonov's theorem, we shall assume that a positive, absolutely summable sequence $\{k_j , j \geqq 0\}$ exists, satisfying the integral inequality

$$\alpha[T \sum_{i \geqq j} k_i + \sup_u | f(u) | \sum_{i>j} h_i + \sum_{i=0}^{j} k_{j-i} h_i] \leqq \beta^{-1} k_j \quad (j \geqq 0). \quad (7)$$

Under these hypotheses we shall prove that for each n, there exist a value of ω, a corresponding solution $x(\cdot)$ of (1), and a function $g(\cdot)$ that is ultimately periodic of period T, such that

$$f(x(t)) = g(t)$$

$$| f(x(jT + t)) - p_g(t) | \leqq k_j \quad (t \in [0,T]) \quad (j \geqq 0)$$

$$\omega = \omega_g = \frac{2n\pi}{T} + \frac{\alpha}{T} \int_0^T p_g(u)\, du .$$

In a corollary we give a condition under which the constants ω for various n are all distinct and lie roughly on a lattice.

IV. TOPOLOGY

In the linear space $B \frown C$ of bounded continuous functions we introduce a topology by means of the metric (distance function):

$$d(g_1 , g_2) = \sum_{n=1}^{\infty} 2^{-n} \max_{-n \leqq x \leqq n} | g_1(x) - g_2(x) |$$

$$+ \sup_{0 \leqq t \leqq T} \limsup_{n \to \infty} | \{ g_1(nT + t) - g_2(nT + t) \} | .$$

The sum term defines a metric for the topology of uniform convergence on compact sets, and the other term (so to speak) "strengthens" the topology at infinity. The number $T > 0$ occurring in the metric is a

parameter, the period of interest. The d-topology so defined is convenient for studying solutions of (1) that are u.p. $[T]$.

Since the metric $d(\cdot,\cdot)$ depends only on the difference function $g_1 - g_2$, it is invariant under translation to zero

$$d(g_1, g_2) = d(g_1 - g_2, 0).$$

Also it can be verified that for $a > 0$,

$$d(ag_1, ag_2) = ad(g_1, g_2).$$

Let $S_\epsilon(g)$ denote an open sphere of radius ϵ about an element g in $B \frown C$,

$$S_\epsilon(g) = \{g_1 \mid d(g_1, g) < \epsilon\}.$$

Let g_1 and g_2 be elements of $S_\epsilon(g)$ and consider a convex combination

$$a_1g_1 + a_2g_2 \quad (a_1, a_2 \geqq 0, \qquad a_1 + a_2 = 1).$$

Then

$$\begin{aligned} d(a_1g_1 + a_2g_2, g) &= d(a_1g_1 - a_1g + a_2g_2 - a_2g, 0) \\ &\leqq d(a_1g_1 - a_1g, 0) + d(a_2g_2 - a_2g, 0) \\ &\leqq a_1d(g_1, g) + a_2d(g_2, g) \\ &< \epsilon. \end{aligned}$$

Hence $S_\epsilon(g)$ is convex. The family of such spheres is a base for the d-topology consisting entirely of convex sets. Hence with the d-topology. $B \frown C$ is a locally convex, linear, topological space (Cf. Ref. 5).

V. PRELIMINARY RESULTS

We define a modulus $m(\cdot)$ of continuity by the equation

$$m(|\epsilon|) = \beta|\epsilon| \left\{\frac{2n\pi}{T} + 2\alpha \sup_u |f(u)| \sup_u |\eta(u)|\right\} + \beta \sup_{|\delta| \leqq |\epsilon|} \int_0^\infty |\psi(t + \delta) - \psi(t)|\, dt.$$

Lemma 1: If $g(\cdot)$ is u.p. $[T]$, and

$$\sup_u |g(u)| \leqq \sup_u |f(u)|$$

then $Ag(\cdot)$ has modulus of continuity $m(\cdot)$.

Proof: The mean value (first Fourier coefficient) of the limit function $p_g(\cdot)$ is at most $\sup_u |f(u)|$ in magnitude. Hence

$$\begin{aligned}
|Ag(t+\epsilon) - Ag(t)| &\leqq \beta|\epsilon|\omega_g + \beta\left|\int_t^{t+\epsilon} \eta(t+\epsilon-u)g(u)\,du\right| \\
&\quad + \beta\left|\int_0^t [\eta(t+\epsilon-u) - \eta(t-u)]g(u)\,du\right| \\
&\leqq \beta|\epsilon|\left\{\frac{2n\pi}{T} + 2\alpha \sup_u |f(u)| \sup_u |\eta(u)|\right\} \\
&\quad + \beta\left|\int_0^t [\psi(t+\epsilon-u) - \psi(t-u)]g(u)\,du\right| \\
&\leqq m(|\epsilon|)
\end{aligned}$$

by a known result of Lebesgue (Ref. 6, p. 14).

Let S be the subset of $B \frown C$ consisting of the functions $g(\cdot)$ with the properties

(i) There is a continuous function $p_g(\cdot)$ of period T such that

$$|g(jT+t) - p_g(t)| \leqq k_j \quad (t\ \varepsilon\ [0,T] \qquad (j \geqq 0).$$

(ii) $\sup_u |g(u)| \leqq \sup_u |f(u)|$

(iii) $g(\cdot)$ has modulus of continuity $m(\cdot)$.

Lemma 2: S is compact.

Proof: To show S is closed, let $\{x_m\} \subset S$ be a sequence converging to x. The second term of the $d(\cdot,\cdot)$ metric ensures that $p_{x_m}(t)$ converges as $m \to \infty$, uniformly for $t\ \varepsilon\ [0,T]$. Denote the continuous limit function by $p_x(\cdot)$. Then

$$\begin{aligned}
|x(jT+t) - p_x(t)| &\leqq |x(jT+t) - x_m(jT+t)| \\
&\quad + |x_m(jT+t) - p_{x_m}(t)| \\
&\quad + |p_{x_m}(t) - p_x(t)|.
\end{aligned}$$

With j and $t\ \varepsilon\ [0,T]$ fixed, let $m \to \infty$; the first and third terms on the right go to zero; the second is at most k_j, for all m. Hence

$$|x(jT+t) - p_x(t)| \leqq k_j \quad (t\ \varepsilon\ [0,T]) \quad (j \geqq 0).$$

Also

$$\begin{aligned}
|x(t+\epsilon) - x(t)| &\leqq |x(t+\epsilon) - x_m(t+\epsilon)| \\
&\quad + |x(t) - x_m(t)| \\
&\quad + |x_m(t+\epsilon) - x_m(t)|.
\end{aligned}$$

Letting $m \to \infty$ for fixed t, the first two terms on the right vanish; the last is at most $m(\epsilon)$ for all $t \geqq 0$. Thus $x(\cdot)$ has modulus of continuity $m(\cdot)$, and so belongs to S. Hence S is closed.

Also, for $y \in S$ and $t \in [0,T]$,

$$\begin{aligned} | p_y(t+\epsilon) - p_y(t) | \leqq & | p_y(t+\epsilon) - y(jT+t+\epsilon) | \\ & + | p_y(t) - y(jT+t) | \\ & + | y(jT+t+\epsilon) - y(jT+t) |. \end{aligned}$$

With t and ϵ fixed, let $j \to \infty$; the first two terms on the right vanish, and the last is at most $m(\epsilon)$ for all $j \geqq 0$. Hence, for $y \in S$, the limiting period function has modulus of continuity $m(\cdot)$.

Now let x_m be an arbitrary sequence of S. From the associated sequence p_{x_m} of periodic functions we can pick a subsequence converging uniformly on $[0,T]$ to a function $p(\cdot)$. From the x_m's associated with this subsequence we can pick, by a standard diagonal argument, a further subsequence $x_{k(i)}$ $i = 1, 2, \cdots$ such that for some $x \in S$

$$x_{k(i)} \to x \text{ uniformly on any compact set}$$

$$p_{x_{k(i)}} \to p \text{ uniformly on } [0,T].$$

Then $x_{k(i)}$ converges to x in the d-topology, so that S, being closed and sequentially compact, is compact.

Lemma 3: $AS \subset S$

Proof: In view of Lemma 1 and the form of A, it sufficies to show that A preserves the defining property (i) of the set S. Accordingly, let $g(\cdot) \in S$, and define the periodic function $p_{Ag}(\cdot)$ of period T by

$$\begin{aligned} p_{Ag}(t) = f\Bigg(x(0) + \frac{2\pi n}{T} t + \alpha \int_0^\infty [p_g(u) - g(u)]\, du \\ + \alpha \int_0^t [Mp_g - p_g(u)]\, du + \int_0^\infty p_g(t-u)\psi(u)\, du\Bigg) \end{aligned}$$

where Mp_g is the mean value of $p_g(\cdot)$. We shall show that

$$| Ag(jT+t) - p_{Ag}(t) | \leqq k_j \quad (t \in [0,T]) \quad (j \geqq 0).$$

Let us rewrite

$$\omega_g t - \alpha \int_0^t \eta(t-u) g(u)\, du$$

in the form, for $t \geqq 0$,

$$\frac{2n\pi}{T}t + \alpha\int_0^t \psi(t-u)p_g(u)\,du + \alpha\int_0^t \psi(t-u)[g(u)-p_g(u)]\,du + \alpha\int_0^t [p_g(u)-g(u)]\,du + \alpha\int_0^t [Mp_g - p_g(u)]\,du. \tag{8}$$

From the Lipschitz condition satisfied by $f(\cdot)$, we obtain

$$|Ag(jT+t) - p_{Ag}(t)| \leqq \beta\left\{\int_{jT+t}^{\infty} |p_g(u)-g(u)|\,du + \alpha\int_{jT+t}^{\infty} |p_g(t-u)|\cdot|\psi(u)|\,du + \alpha\int_0^{jT+t} |g(jT+t-u)-p_g(t-u)|\cdot|\psi(u)|\,du\right\}.$$

Since $g(\cdot)$ belongs to S, it is true that for $t \in [0,T]$

$$\int_{jT+t}^{\infty} |p_g(u)-g(u)|\,du \leqq T\sum_{i\geqq j} k_i,$$

$$\int_{jT+t}^{\infty} |p_g(t-u)|\cdot|\psi(u)|\,du \leqq \sup_u |f(u)|\cdot\sum_{i\geqq j} h_i,$$

$$\int_0^{jT+t} |g(jT+t-u)-p_g(t-u)|\cdot|\psi(u)|\,du \leqq \int_0^t |g(jT+t-u)-p_g(t-u)|\cdot|\psi(u)|\,du + \sum_{i=1}^{j-1}\int_{iT+t}^{(i+1)T+t} |g(jT+t-u)-p_g(t-u)|\cdot|\psi(u)|\,du \leqq \sum_{i=0}^{j} k_{j-i}h_i.$$

Lemma 3 now follows from the integral inequality (7).

VI. PRINCIPAL RESULTS

Theorem 1: If $\sum_{i\geqq 0} h_i < \infty$ and $\sum_{i\geqq 0} k_i < \infty$, where $\{h_i\}$ is given by (6) and $\{k_i\}$ satisfies (7), then for each integer n there exist a value of ω, a (corresponding) solution $x(\cdot)$ of (1), and a function $g(\cdot)$ that is u.p. $[T]$, such that

i. $Ag = g$, i.e., $g(\cdot)$ satisfies (4),
ii. $g(t) = f(x(t))$, for all t,
iii. $|f(x(jT + t)) - p_g(t)| \leqq k_j \quad (t \in [0,T]) \quad (j \geqq 0)$,
iv. the periodic function $p_g(\cdot)$ is a solution on $[0,T]$ of the equation

$$p_g(t) = f\left(x(0) + \frac{2\pi n}{T}t + \alpha\int_0^\infty [p_g(u) - g(u)]\,du + \alpha\int_0^t [Mp_g - p_g(u)]\,du + \alpha\int_0^\infty p_g(t-u)\psi(u)\,du\right),$$

v.

$$\omega = \frac{2\pi n}{T} + \frac{\alpha}{T}\int_0^T p_g(u)\,du = \frac{2\pi n}{T} + \alpha M p_g\,.$$

Proof: We first show that A is a continuous transformation on the set S. Let $g_m \to g$ with $g_m \in S$; then $g \in S$, because S is closed. Let L be a compact set of the line and set $z = \sup_{t\in L} |t|$. For $t \in L$ we have

$$|Ag_m(t) - Ag(t)| \leqq \beta\left|t(\omega_{g_m} - \omega_g) + \alpha\int_0^t \eta(t-u)[g(u) - g_m(u)]\,du\right|$$

$$\leqq \alpha\beta\left(z\,|Mp_{g_m} - Mp_g| + \sup_u |\eta(u)|\cdot\int_0^z |g(u) - g_m(u)|\,du\right).$$

Since $p_{g_m} \to p_g$ uniformly, and $g_m \to g$ uniformly on compact sets, it follows that $Ag_m \to Ag$ uniformly on compact sets. It remains to show that $p_{Ag_m} \to p_{Ag}$ uniformly. Now for $t \in [0,T]$,

$$|p_{Ag_m}(t) - p_{Ag}(t)| \leqq \alpha\beta\left|(Mp_{g_m} - Mp_g)t + \int_0^\infty [p_{g_m}(u) - g_m(u)] - \int_0^\infty [p_g(u) - g(u)]\,du + \int_0^\infty [p_{g_m}(t-u) - p_g(t-u)]\psi(u)\,du\right|. \tag{9}$$

The first and fourth terms on the right of (9) obviously converge to zero uniformly in t, by the uniform convergence of p_{g_m} to p_g. The middle two terms on the right are together at most

$$\int_0^{jT} |p_{g_m}(u) - p_g(u)|\,du + \int_0^{jT} |g(u) - g_m(u)|\,du + 2T\sum_{i\geqq j} k_i$$

in magnitude; this bound can be made arbitrarily small by first choosing j large, and then m large.

The set S is compact and convex, and A is a continuous map of S into

itself. Hence by Tychonov's fixed point theorem (Ref. 5, p. 456) there is a fixed point $g \varepsilon S$ with $Ag = g$. Define $x(\cdot)$ by

$$x(t) = \begin{cases} x(0) + \omega_g t - \alpha \int_0^t \eta(t-u)g(u)\,du & (t \geqq 0) \\ x(0) & (t \leqq 0). \end{cases}$$

Then $x(\cdot)$ satisfies (1) almost everywhere in $t \leqq 0$, and $g(t) = f(x(t))$. Since $g \varepsilon S$, there is a function $p_g(\cdot)$ of period T such that

$$| g(jT + t) - p_g(t) | \leqq k_j \quad (t \varepsilon [0,T]) \qquad (j \geqq 0).$$

The limit equation for $p_g(\cdot)$ is obtained by taking the limit as $j \to \infty$ in the equation

$$g(jT + t) = Ag(jT + t) \quad (t \varepsilon [0,T])$$

and using (8) to expand the right-hand side. The second term of (8) approaches

$$\int_0^\infty p_g(t-u)\psi(u)\,du$$

while the third term of (8) goes to zero, by an elementary Abelian result. Finally the value of ω associated with the solution $x(\cdot)$ of (1) is given by (v) because $f[x(\cdot)]$ is a fixed point of A.

Corollary: If $h(\cdot)$ is bounded, and

$$\frac{\pi}{T} > \alpha \sup_u | f(u) | = \gamma$$

then each interval $(2n\pi/T) \pm \gamma$, n an integer, contains a value of ω for which the (unique) solution of (1) is u.p. $[T]$.

Proof: The condition stated guarantees that the values of (v), Theorem 1, for various integers n, are all distinct. Uniqueness of the solution $x(\cdot)$ of (1) can be proved from the boundedness of $h(\cdot)$ and the Lipschitz condition on $f(\cdot)$ by standard methods.

REFERENCES

1. Barnard, R. D., this issue, p. 227.
2. Byrne, C. J., to be published.
3. Goldstein, A. J., to be published.
4. Beneš, V. E., Journal of Math. and Phys. **40**, (1961) pp. 55–65.
5. Dunford, N., and Schwarz, J. T., *Linear Operators*, Interscience Publishers, Inc., New York, 1958.
6. Wiener, N., *The Fourier Integral and Certain of its Applications*, Dover Publications, New York, 1951.

Single Server Systems — I. Relations Between Some Averages

By S. O. RICE

(Manuscript received April 13, 1961)

This is the first of two papers dealing with single server systems. Two subjects are discussed in the present paper, (i) relations between such items as the probability of loss, probability of no delay, and the average number of customers served in a busy period, and (ii) the statistical behavior of a single server system in which no waiting for service is allowed.

I. INTRODUCTION

A typical but rather homely example of the systems considered here is a barber shop in which there is only one barber, the "single server." Let $N - 1$ be the number of chairs provided for customers waiting for service so that the "capacity" of the system is N. When the shop is full, one customer is being served and $N - 1$ are waiting. A prospective customer (a "demand" for service) arriving when the shop is full is turned away and is said to be "lost." If the shop is not full, he waits and is eventually served.

The demands arrive at an average rate of a per unit time. The server would serve b customers per unit time (on the average) if he were to work steadily. It follows that the average interval between arrivals is $1/a$ and the average service time is $1/b$. It is assumed that the rates a and b do not change with time.

The first part of the paper is concerned with several quantities of interest, including the fraction L of demands lost and the average length of the busy periods, i.e., the periods during which the server is continuously busy. The values of these quantities are expressed in terms of a, b, and two other quantities p_0 and τ. Here p_0 is the probability that the server is idle (at an instant selected at random) and τ is the average duration of an idle period. Both p_0 and τ depend upon N and upon the probability laws governing the arrivals and service times. However, only the simplest cases of this dependence are mentioned in the first part of the paper. The results are summarized in Table I.

The second part of the paper is concerned with the single server "loss system" in which no waiting is allowed ($N = 1$). The input is now assumed to be "recurrent," i.e., the distances between arrivals are independent and have the general distribution function $A(t)$. The service times have the distribution function $B(t)$ and are independent of each other and of the arrivals. The functions $A(t)$ and $B(t)$ are such that the interarrival distances and service lengths have the respective average values

$$a^{-1} = \int_0^\infty [1 - A(t)]\, dt$$
$$b^{-1} = \int_0^\infty [1 - B(t)]\, dt. \tag{1}$$

Further conditions are imposed on $A(t)$ and $B(t)$ in the course of the derivations. The principal items of interest are (*i*) the loss L, (*ii*) the probabilities p_0 and p_1 that the server is idle or busy, respectively, at a time selected at random, and (*iii*) the distribution of the lengths of the idle periods. The expression for the loss is equivalent to one obtained in a different manner by F. Pollaczek.[1] A discussion of how the loss increases with the variability of arrival has been given by P. M. Morse.[2]

Results for a Type I counter, i.e., one in which the registration of an arrival is followed by a "dead time," may be applied to the single server loss system by identifying the dead time (supposed variable) with the intervals the server is busy. For example, results obtained here are closely related to some for counters given by R. Pyke[3] and earlier workers to whom he makes reference.

The infinite capacity system ($N = \infty$) is discussed in a companion paper.[4] In this case there is no loss, and attention is focused on the distributions of the waiting times and busy period lengths. Some of the results of Section II of the present paper find application there.

I am indebted to John Riordan for many helpful discussions on the subject matter of these two papers and for numerous improvements in presentation.

II. AVERAGES OBTAINED FROM FIRST PRINCIPLES, GENERAL INPUT AND LIMITED CAPACITY

Several important averages associated with a single server system may be readily obtained by considering its behavior over a long period of time T, statistical equilibrium being assumed, and then letting $T \to \infty$. The input and service are assumed to be general with respective arrival

and service rates a and b. The capacity of the system is N, so that demands arriving when $N - 1$ are waiting are lost.

The probability p_0 that the server is idle (at an instant selected at random) and the average duration τ of an idle period are supposed given. For Poisson input, i.e., one in which the probability that a demand will arrive in the interval t, $t + dt$ is $adt + 0(dt^2)$ (irrespective of the arrival times of the other demands), the value of τ is $1/a$.

Let $\nu(T)$ be the number of arrivals in T and let ϵ be a given, arbitrarily small, positive number. The input and service are assumed to be such that the probability of $1 - \epsilon < \nu(T)/aT < 1 + \epsilon$ may be made as close to unity as desired by choosing T large enough. For recurrent input this restriction is satisfied, by virtue of the law of large numbers, when the first integral in (1) gives a finite value for a^{-1}. We shall refer to the above inequality by saying that the number of arrivals in the long interval T is "equal" to aT. Similar statements made below regarding total idle and busy time, number served, etc., in the interval T are to be interpreted in a similar way.

The total idle time of the server is Tp_0, the total busy time is $T - Tp_0$, and the total number served is $b(T - Tp_0)$. The number lost is the number of arrivals less the number served, and the fraction of arrivals lost is

$$L = [aT - b(T - Tp_0)]/aT = 1 - (1 - p_0)\rho^{-1} \tag{2}$$

where $\rho = a/b$. L is the "probability of loss." Loss occurs only when the waiting room is filled, i.e., when the system is in state N (probability p_N and average duration τ_N). The number of times state N occurs is $p_N T/\tau_N$, and the average number of demands lost during such a period is $\tau_N[a - b(1 - p_0)]/p_N$. For exponential service the service time lengths have the distribution function $B(t) = 1 - e^{-bt}$, and the value of τ_N is $1/b$.

The number served in T without delay is equal, to within one, to the number of idle periods. The number of idle periods is $p_0 T/\tau$; hence $p_0/a\tau$ is the probability $W(0)$ that a demand will be served upon arrival. Thus,

$$W(0) = p_0/a\tau. \tag{3}$$

Because a demand is either served without delay, delayed, or lost, the probability of delay is

$$1 - L - (p_0/a\tau) = (1 - p_0)\rho^{-1} - (p_0/a\tau).$$

Since busy and idle periods alternate, the number of busy periods is, to within one, the number of idle periods, namely $p_0 T/\tau$. Dividing this

into the total busy time $T - Tp_0$ gives $\tau(1 - p_0)/p_0$ for the average length of a busy period. Similarly, the total number served gives $b\tau(1 - p_0)/p_0$ for the average number served in a busy period.

When N is infinite, statistical equilibrium requires $\rho < 1$, and all demands are eventually served. Equating the number of arrivals aT to the number served $b(T - Tp_0)$ gives $p_0 = 1 - \rho$.

These results and the forms they assume for N infinite and $\tau = 1/a$ (Poisson input) are summarized in Table I. The results for Poisson input are well known.

When the capacity of the waiting room is infinite, the average number $\bar{n}$ of demands served in a busy period is

$$\bar{n} = 1/W(0). \tag{4}$$

In this case L is zero, demands are either delayed or not delayed, and $W(0)$ becomes the probability of no delay. This follows from (3), Table I, and the fact that both sides of (4) are equal to $a\tau/(1 - \rho)$. For limited capacity

$$\bar{n} = (1 - L)/W(0) \tag{5}$$

which follows from Table I. Here $W(0)$ is the chance that a demand chosen at random will be served upon arrival.

III. SINGLE SERVER LOSS SYSTEM

In a single server loss system any demand arriving when the server is busy is lost. The system capacity N is 1, there is no waiting line, and every busy period consists of a single service interval. As mentioned in the introduction, the input is assumed to be recurrent, the service is

TABLE I — STATIONARY AVERAGES FOR A GENERAL SYSTEM

Average	Limited Capacity General Input	Infinite Capacity, $p_0 = 1 - \rho$	
		General Input	Poisson Input, $\tau = a^{-1}$
Rel. number lost	$1 - (1 - p_0)\rho^{-1}$	0	0
Rel. number not delayed	$p_0(a\tau)^{-1}$	$(1 - \rho)(a\tau)^{-1}$	$1 - \rho$
Rel. number delayed	$(1 - p_0)\rho^{-1} - p_0(a\tau)^{-1}$	$1 - (1 - \rho)(a\tau)^{-1}$	ρ
Length of busy period	$\tau(1 - p_0)p_0^{-1}$	$\tau\rho(1 - \rho)^{-1}$	$(b - a)^{-1}$
Number served in busy period	$b\tau(1 - p_0)p_0^{-1}$	$a\tau(1 - \rho)^{-1}$	$(1 - \rho)^{-1}$

Notation: a = arrival rate, b = service rate, $\rho = a/b$, τ = average length of idle period, p_0 = fraction of time server is idle.

general, and the principal items of interest are (*i*) the loss L, (*ii*) the probabilities p_0 and p_1 that the server is idle or busy, respectively, at a time selected at random, and (*iii*) the probability $q(u)\,du$ that the length of an idle period will lie between u and $u + du$.

3.1 *Values of L, p_0 and p_1*

For general input and service, the results of Section II for $N = 1$ and the relation $1 - p_0 = p_1$ show that

$$L = 1 - p_1\rho^{-1}, \qquad p_0 = b\tau p_1, \qquad \rho = a/b \tag{6}$$

where the second relation is obtained by equating the average busy period length $\tau(1 - p_0)/p_0$ to the average service length $1/b$, τ being the average idle period length.

For Poisson input, $\tau = 1/a$ and (6) gives

$$p_1 = \rho p_0, \qquad p_0 = (1 + \rho)^{-1}, \qquad L = p_1 = \rho/(1 + \rho). \tag{7}$$

In this case, the loss is independent of the service distribution, a property also possessed by the many-server loss system for Poisson input.

For recurrent input and general service, the ratio $L/(1 - L)$ of the number of demands lost to the number served is equal to the expected number of demands arriving while the server is busy serving one demand, i.e., during one service interval. Thus

$$\frac{L}{1 - L} = \int_0^\infty \overline{n(t)}\, dB(t) \tag{8}$$

where $\overline{n(t)}$ is the expected number of arrivals during a service of length t (not counting the one starting the service) and $B(t)$ is the service time distribution function. As t becomes large, $\overline{n(t)}$ is $0(at)$ and the integral converges because the average service time is finite.

It will be shown that

$$\overline{n(t)} = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{\alpha(s)e^{st}\,ds}{[1 - \alpha(s)]s} \qquad (c > 0) \tag{9}$$

where $\alpha(s)$ is the Laplace-Stieltjes transform of the distribution function $A(t)$ for the separation between arrivals:

$$\alpha(s) = \int_0^\infty e^{-st}\, dA(t). \tag{10}$$

It is assumed that $\alpha(s)$ and t are such that the integral in (9) converges. Equation (9) is but one of a number of similar results; see, for example, D. R. Cox and W. L. Smith.[5]

To obtain (9), note that the service starts with an arrival, and the probability that n or more additional arrivals will occur in the ensuing interval of length t is the probability that

$$S_n \equiv X_1 + X_2 + \cdots + X_n \leqq t. \tag{11}$$

Here X_i is the separation between arrivals $i - 1$ and i. Since the X_i's are independent and have the distribution function $A(t)$, the rules for determining the distribution of the sum of n random variables may be applied to find the chance that $S_n \leqq t$. In particular, the Laplace transform of the probability density for S_n is $[\alpha(s)]^n$, and the Laplace transform of Prob $[S_n \leqq t]$ is $[\alpha(s)]^n/s$.

The probability of exactly n arrivals in an interval of length t which starts just after an arrival is

$$\begin{aligned} P_n(t) &= \text{Prob}\,[S_n \leqq t) - \text{Prob}\,[S_{n+1} \leqq t] \\ &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} [1 - \alpha(s)][\alpha(s)]^n s^{-1} e^{st}\, ds. \end{aligned} \tag{12}$$

Multiplying $P_n(t)$ by n, noting that $|\alpha(s)| < 1$ on the path of integration, and summing from $n = 0$ to $n = \infty$ then gives (9).

When $\overline{n(t)}$ is known as a function of t, either from (9) or otherwise, and is used in (8), the result is an equation which may be solved for the loss L. When L is known, (6) gives

$$p_1 = \rho(1 - L), \qquad p_0 = 1 - p_1. \tag{13}$$

A few examples follow.

i. Poisson input. Here $1 - A(t) = e^{-at}$ and $\alpha(s) = a/(a + s)$. Then

$$\overline{n(t)} = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} as^{-2} e^{st}\, ds = at$$

$$\frac{L}{1 - L} = \int_0^\infty at\, dB(t) = a/b = \rho$$

and the results given in (7) are again obtained.

ii. Arrivals spaced $1/a$ apart. By inspection

$$\overline{n(t)} = n, \qquad na^{-1} < t < (n + 1)a^{-1}$$

and (8) becomes

$$\frac{L}{1 - L} = \sum_0^\infty n \int_{n/a}^{(n+1)/a} dB(t) = \sum_1^\infty n\left[B\left(\frac{n+1}{a}\right) - B\left(\frac{n}{a}\right)\right]. \tag{14}$$

iii. When $\alpha(s)$ is $0(1/s)$ as $s \to c \pm i\infty$, as it is when the probability density $A'(t) = dA(t)/dt$ exists and is of bounded variation, and when

$B'(t)$ exists and is $0(e^{-\epsilon t})$, $\epsilon > 0$, as $t \to \infty$, substitution of (9) in (8) gives

$$\begin{aligned} \frac{L}{1-L} &= \int_0^\infty \frac{dB(t)}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{\alpha(s)e^{st}\,ds}{[1-\alpha(s)]s} \\ &= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{\alpha(s)\beta(-s)\,ds}{[1-\alpha(s)]s} \qquad (0 < c < \epsilon) \end{aligned} \tag{15}$$

where

$$\beta(s) = \int_0^\infty e^{-st}\,dB(t). \tag{16}$$

In (15) the singularities of $\beta(-s)$ lie to the right of the path of integration and those of $\alpha(s)/s[1-\alpha(s)]$ to the left. The conditions imposed on $\alpha(s)$ and $B'(t)$ are sufficient to ensure the absolute convergence of the double integral and hence to justify the inversion of the order of integration. The result (15) is equivalent to one obtained by F. Pollaczek[1] by a different method.

iv. Exponential service. Here, $1 - B(t) = e^{-bt}$ and $\beta(-s) = b/(b-s)$. Substituting this value of $\beta(-s)$ in (15), closing the path of integration by an infinite semicircle on the right, and evaluating the residue at the pole $s = b$ gives

$$\frac{L}{1-L} = \frac{\alpha(b)}{1-\alpha(b)}, \qquad L = \alpha(b), \qquad p_1 = \rho[1-\alpha(b)].$$

This expression for p_1 is a special case of the stationary state probabilities (determined by both F. Pollaczek[6] and L. Takács[7]) for the many server system with recurrent input and exponential service.

v. Let $\alpha(s)$ satisfy the same condition as in example *iii* and in addition, suppose that (9) may be written as

$$\overline{n(t)} = R(t) + \frac{1}{2\pi i} \int_{-c-i\infty}^{-c+i\infty} \frac{\alpha(s)e^{st}\,ds}{[1-\alpha(s)]s} \qquad (c > 0)$$

where the only singularity of $e^{st}\alpha(s)/s[1-\alpha(s)]$ to the right of $\mathrm{Re}(s) = -c$ (s finite) is a double pole at $s = 0$ with residue $R(t)$. Then (8) gives

$$\frac{L}{1-L} = \int_0^\infty R(t)\,dB(t) + \frac{1}{2\pi i} \int_{-c-i\infty}^{-c+i\infty} \frac{\alpha(s)\beta(-s)\,ds}{[1-\alpha(s)]s}. \tag{17}$$

It should be noticed that not all cases can be handled by (15) and (17). An example to the contrary is furnished by

$$\begin{aligned} A(t) &= 1 - (1+at)^{-2} \\ B(t) &= 1 - (1+bt)^{-2} \end{aligned} \tag{18}$$

where the averages a^{-1} and b^{-1} exist but the corresponding variances are infinite. In this case

$$\alpha(s) = 1 - x + x^2 e^x \int_x^\infty e^{-y} y^{-1}\, dy \qquad (x = s/a).$$

As $|s| \to \infty$ in $\mathrm{Re}(s) \geqq 0$, $\alpha(s)$ is $0(s^{-1})$. Near $s = 0$,

$$\alpha(s) = 1 - sa^{-1} - s^2 a^{-2} \ln s + 0(s^2).$$

Since $\beta(s)$ is similar to $\alpha(s)$, both $\beta(-s)$ and $\alpha(s)/s[1 - \alpha(s)]$ have branch points at $s = 0$. Equation (17) fails in this case because $s = 0$ is not a double pole. Equation (15) fails because it is impossible to draw a path of integration which separates the singularities of $\beta(-s)$ from those of $\alpha(s)/s[1 - \alpha(s)]$. When $A(t)$and $B(t)$ are given by (18) it seems necessary to work directly with (8) and (9), or else use some sort of limit.

3.2 *Lengths of Idle Periods*

Now turn to the distribution of l_b and l_i, the lengths of a busy period and the following idle period, recurrent input and general service being assumed. The probability that $l_b \leqq t$ is simply $B(t)$, the service length distribution function. The distribution of the idle period length l_i is more complicated. Its average length is, from (6),

$$\tau = p_0/bp_1 = a^{-1}(1 - L)^{-1} - b^{-1} \tag{19}$$

where L is given by (8). The first step towards obtaining the probability $q(u)\, du$ that $u < l_i < u + du$ is to determine the conditional probability $q(u; t)\, du$ that $u < l_i < u + du$ given $l_b = t$.

Consideration of the arrival patterns which give no arrivals in $(t, t + u)$ followed by one in $(t + u, t + u + du)$ leads to

$$q(u; t)\, du = \sum_{n=0}^{\infty} \Pr\,[S_n < t; t + u < S_{n+1} < t + u + du] \tag{20}$$

where S_0 is 0 and S_n for $n > 0$ is the sum (see (11)) of n interarrival intervals X_i. An expression for the joint probability density of S_n and S_{n+1} may be obtained by inverting its double Laplace transform

$$\begin{aligned}\text{ave exp}\,[-rS_n - sS_{n+1}] &= \text{ave exp}\,[-(r + s)(X_1 + \cdots + X_n) - sX_{n+1}] \\ &= [\alpha(r + s)]^n \alpha(s).\end{aligned}$$

Integrating the density over the region $0 < S_n < t$, $u + t < S_{n+1} < u + t + du$ shows that the nth term, $n > 0$, in (20) is

$$\frac{du}{(2\pi i)^2} \int_{c-i\infty}^{c+i\infty} ds e^{s(u+t)} \alpha(s) \int_{c-i\infty}^{c+i\infty} [\alpha(r+s)]^n (e^{rt} - 1) r^{-1} dr \quad (21)$$

where $c > 0$.

Expression (21) holds only for $n > 0$. However, replacing the factor $(e^{rt} - 1)$ by e^{rt} gives an expression which holds for $n \geqq 0$. Indeed, closing the path of r-integration on the right shows that the integral of $[\alpha(r+s)]^n r^{-1}$ is zero for $n > 0$. Closing it on the left shows that the integral of $e^{rt} r^{-1}$ is $2\pi i$ for $t > 0$ and leads to the correct value for $n = 0$.

Setting the modified form of expression (21) in the series (20) and performing the summation shows that $q(u; t)\, du$ is equal to an expression obtained by replacing $[\alpha(r+s)]^n (e^{rt} - 1)$ in (21) by $[1 - \alpha(r+s)]^{-1} e^{rt}$. The joint probability density of l_b and l_i is $q(u; t)B'(t)$ where $B'(t) = dB(t)/dt$. The probability density $q(u)$ is the integral of $q(u; t)B'(t)$ taken from $t = 0$ to $t = \infty$. Assuming $B'(t)$ to be $0(e^{-\epsilon t})$ as $t \to \infty$ and choosing the paths of integration $c \pm i\infty$ so that $0 < 2c < \epsilon$ makes the integral of $B'(t) \exp [t(s+r)]$ converge and have the value $\beta(-s-r)$. Changing the variable of integration from r to $z = r + s$ and, for convenience in writing (22), taking $\alpha(s)$ to be such that the path of integration for s may be shifted to $-\eta \pm i\infty$ (this implies that $A'(t)$ is $0(e^{-\eta t})$ as $t \to \infty$) gives finally

$$q(u) = \left(\frac{1}{2\pi i}\right)^2 \int_{-\eta-i\infty}^{-\eta+i\infty} ds e^{su} \alpha(s) \int_{\eta-i\infty}^{\eta+i\infty} \frac{\beta(-z)\, dz}{[1 - \alpha(z)](z - s)} \quad (22)$$

where $u \geqq 0$ and η is an arbitrarily small positive number.

It may be shown that (22) reduces to a e^{-au} for Poisson input, as it should, and to

$$q(u) = b[1 - \alpha(b)]^{-1} \int_0^\infty e^{-bv} A'(u+v)\, dv \quad (23)$$

for recurrent input and exponential service. Multiplying $q(u)$ by $\exp(-s'u)$, integrating u from 0 to ∞, closing the path of integration for s on the right, and dropping the prime from s' shows that the Laplace transform of $q(u)$ is

$$\text{ave } e^{-su} = \text{ave } e^{-sl_i} = \frac{1}{2\pi i} \int_{\eta-i\infty}^{\eta+i\infty} \frac{\beta(-z)}{z - s} \left[\frac{\alpha(s) - \alpha(z)}{1 - \alpha(z)}\right] dz. \quad (24)$$

We also have

$$\text{ave}\, e^{-rl_b - sl_i} = \frac{1}{2\pi i} \int_{\eta - i\infty}^{\eta + i\infty} \frac{\beta(r - z)}{z - s} \left[\frac{\alpha(s) - \alpha(z)}{1 - \alpha(z)} \right] dz \qquad (25)$$

which may be regarded as a double Laplace transform. In (24) and (25) the singularities of $\beta(-z)$ and $\beta(r - z)$ are supposed to lie to the right of the path of integration and the remaining singularities of the integrands to the left.

REFERENCES

1. Pollaczek, F., "Problems stochastiques," in *Memorial des Sciences Mathematiques*, Fasc. No. **136,** Gauthier Villars, Paris, 1957, p. 113, Eq. (9.40).
2. Morse, P. M., *Queues, Inventories and Maintenance,* John Wiley and Sons, New York, 1958, Chapter 5.
3. Pyke, R., Annals of Math. Stat., **29,** 1958, p. 737.
4. Rice, S. O., this issue, p. 279.
5. Cox, D. R., and Smith, W. L., Biometrika, **41,** 1954, p. 91.
6. Pollaczek, F., C. R. Acad. Sci., Paris, **236,** 1953, p. 1469.
7. Takács, L., Acta Math. Acad. Sci. Hungar., **7,** 1956, p. 419.

Single Server Systems — II. Busy Periods

By S. O. RICE

(Manuscript received July 10, 1961)

This is the second of two papers dealing with single server systems. Statistical problems associated with the busy periods, i.e., the periods during which the server is continuously busy, are considered in the present paper. The input and the service time distributions may be quite general, and the length of the waiting line is unrestricted. Among the new results is an asymptotic expression for the probability density of the lengths of the busy periods. This expression holds when the arrival rate is almost equal to the service rate and when the busy periods tend to be long. It is hoped that the methods used here will throw additional light on known results.

I. INTRODUCTION

This is the second of two papers dealing with single server systems. In the first paper,[1] two subjects were discussed, (*i*) a method which led to the average values of several quantities of interest and (*ii*) the statistical behavior of a single server loss system. The present paper is concerned with the busy periods (the periods during which the server is continuously busy) in a single server system when no restrictions are placed on the queue length.

The distribution of the busy period lengths has been studied by a number of investigators, among them E. Borel,[2] D. G. Kendall,[3] F. Pollaczek,[4,5] L. Takács,[6,7] and B. W. Connolly.[8,9]

A closely related problem is that of the storage of water behind a dam. An interesting survey of this subject has been given by J. Gani.[10] The moving server problem treated by McMillan and Riordan[11] and by Karlin, Miller and Prabhu[12] is also related to the busy period problem. Again, for Poisson input, the distribution of the busy period lengths is related to the distribution of the delays in "last come, first served" type of service (see Riordan[20] and the references to earlier work given there).

The most general results are those due to Pollaczek. His work leads to the joint distribution function of n,S,y where n is the number of services

comprising the busy period, S the length of the period, and y the length of the following idle period.

The object of the present paper is to obtain, in another way, results equivalent to those derived by Pollaczek. Integral equations similar to Lindley's[13] equation for the waiting time distribution are first set up. These equations are then solved by methods similar to the Wiener-Hopf technique used by W. L. Smith[14] to solve Lindley's equation. The aim of the presentation is to illustrate the methods; the discussion is heuristic; and no claims are made for completeness or rigor. On the other hand, it is hoped that the different point of view will add something to the understanding of the earlier results.

In Section II a brief review of some of the earlier results is given. In Section III the length of the busy period, as measured by the number n of customers served, is considered. The results of Section III are generalized in Section IV to apply to the joint distribution of n, S, and y. Special attention is paid to the distribution of S. The transient behavior of the queue length is considered in Section V. Since the method used here to study the busy times is closely related to Smith's method of dealing with waiting times, a review of the waiting time problem is given in the Appendix. Some changes are made in Smith's method in order to make it fit in better with the busy problem.

Throughout the paper, except possibly for (2), the intervals between arrivals are considered to be independent of each other and of the service times (recurrent input). The service times are also taken to be independent of each other and of the input.

As mentioned in the companion paper, I am indebted to John Riordan for much help in the preparation of this paper. I am also indebted to V. Beneš and L. Takács for helpful discussions and a number of references.

II. PRELIMINARY REMARKS

In this paper, $A(t)$ will denote the distribution function for the intervals between successive arrivals, $B(t)$ the service time distribution, and $A'(t)$,$B'(t)$ the respective probability densities. The arrival rate is a, the service rate is b, the average interval between arrivals is

$$a^{-1} = \int_0^\infty [1 - A(t)]\, dt$$

and the average service time is

$$b^{-1} = \int_0^\infty [1 - B(t)]\, dt.$$

The Laplace-Stieltjes transforms defined by

$$\alpha(s) = \int_0^\infty e^{-st}\, dA(t)$$
$$\beta(s) = \int_0^\infty e^{-st}\, dB(t) \tag{1}$$

play an important role.

Some information on the average length of a busy period (chosen at random from the universe of busy periods) may be obtained from the general results of the companion paper.[1] Thus, from Table I of that paper, when the service rate b exceeds the arrival rate a so that statistical equilibrium exists, the average length $\bar{S}$ of a busy period and the average number $\bar{n}$ of customers served (during a busy period) are given by

$$\bar{S} = \frac{a\tau}{b-a} = \frac{1}{bW(0)}, \qquad \bar{n} = \frac{1}{W(0)}. \tag{2}$$

Here the system capacity N is infinite, τ is the average length of an idle period, and $W(0)$ is the chance that a customer is served immediately upon arrival (zero waiting time). For Poisson input τ is $1/a$ and

$$\bar{S} = \frac{1}{b-a}, \qquad \bar{n} = \frac{1}{1-\rho}, \qquad \rho = \frac{a}{b} \tag{3}$$

irrespective of the service time distribution. For inputs other than Poisson, one must know either τ or $W(0)$ in order to compute $\bar{S}$ from (2).

For the sake of orientation we state the following known results.

i. For Poisson input, general service and $a < b$, Kendall[3] has shown that the transform

$$\gamma(z) = \int_0^\infty e^{-zS}\, dG(S) \tag{4}$$

of the distribution function $G(S)$ of the busy period lengths satisfies the functional equation

$$\gamma(z) = \beta[z + a - a\gamma(z)]. \tag{5}$$

Here $\beta(s)$ is defined by (1). For Poisson input and exponential service, this leads to the probability density

$$G'(S) = \frac{dG(S)}{dS} = b(S\sqrt{ab})^{-1} I_1(2S\sqrt{ab})\, e^{-(b+a)S} \tag{6}$$

where $I_1(z)$ denotes a Bessel function of the first kind for imaginary argument.

ii. Suppose that we are given a busy period which has just begun. Let f_n be the chance that it will end after exactly n services. For Poisson input, Takács[6] has shown that the generating function

$$f(x) = \sum_1^\infty x^n f_n \tag{7}$$

must satisfy the functional equation

$$f(x) = x\beta[a - af(x)]. \tag{8}$$

For exponential service this leads to

$$f_n = \frac{(2n-2)!}{n!(n-1)!} \frac{\rho^{n-1}}{(1+\rho)^{2n-1}} \tag{9}$$

and for constant service time (Borel[2]) to

$$f_n = (\rho n)^{n-1} \frac{e^{-n\rho}}{n!}. \tag{10}$$

When $a < b$, all busy periods end eventually and $\sum_1^\infty f_n = 1$. When $a > b$, the customers arrive faster than the server can handle them, and sooner or later a busy period will start and never end. Given a busy period which has just begun, the probability that it will never end is $1 - \sum_1^\infty f_n$, where now $\sum_1^\infty f_n$ is less than 1. For Poisson input, exponential service, and $a > b$, the probability that the busy period will not end is $1 - (b/a)$.

iii. As mentioned in the Introduction, the results most closely related to the work of the present paper are those due to Pollaczek.[4,5] In particular he has shown the following. Given a busy period which has just begun, let $G_n'(y,S)dy\,dS$ be the chance that it will consist of exactly n services, have a length between S and $S + dS$, and be followed by an idle period whose length lies between y and $y + dy$. Let

$$\gamma_n(s,z) = \int_0^\infty dS \int_0^\infty dy\, e^{sy - zS} G_n'(y,S) \tag{11}$$

where e^{sy} appears instead of e^{-sy} for later convenience. Then, for rather general input and service distributions,

$$\sum_{n=1}^\infty x^n \gamma_n(s,z) = 1 - \exp\left\{\frac{1}{2\pi i} \int_c \frac{\ln[1 - x\beta(z+\zeta)\alpha(-\zeta)]\,d\zeta}{\zeta - s}\right\} \tag{12}$$

where $0 \leqq x \leqq 1$, $\mathrm{Re}(s) < \mathrm{Re}(\zeta) < 0 < \mathrm{Re}(z)$, $\mathrm{Re}(z+\zeta) > 0$, and

the path of integration for ζ runs from $-i\infty$ to $+i\infty$ in the strip specified by the foregoing inequalities. In writing (11) and (12) it has been convenient to change from Pollaczek's notation to a notation resembling that used by Smith.[14]

III. NUMBER SERVED IN A BUSY PERIOD

3.1 *Derivation of Integral Equation*

Consider the busy period to start with the arrival of customer numnumber 1 at time 0, and let the service time of the rth arrival be s_r and t_r the interval between the arrivals of customers r and $r+1$. The busy period consists of one service if $s_1 < t_1$, i.e., if $u_1 = s_1 - t_1 < 0$. It consists of two services if $s_1 \geqq t_1$ and $s_1 + s_2 < t_1 + t_2$, i.e., if $u_1 \geqq 0$ and $u_1 + u_2 < 0$. In general, n demands (customers) are served in a busy period if $U_1, U_2, \cdots, U_{n-1} \geqq 0$ and $U_n < 0$ where

$$\begin{aligned} U_n &= u_1 + u_2 + \cdots + u_n \\ u_r &= s_r - t_r . \end{aligned} \tag{13}$$

Since the u_r's are independent random variables whose distributions are known (indeed, ave $\exp(-su_r) = \beta(s)\alpha(-s)$) one may, in principle, find the joint distribution of $U_1, U_2, \cdots, U_n$. Then the probability f_n that n demands are served may be obtained by integration over the proper region in $U_1, \cdots, U_n$ space. However, it is more convenient to use an indirect method which depends upon the solution of an integral equation similar to that for the waiting time distribution.

Let $p_1(V)dV$ be the probability that $V < U_1 < V + dV$, and $p_n(V)dV$ the probability that $U_1, \cdots, U_{n-1} \geqq 0$ and $V < U_n < V + dV$, where V may be either positive or negative. Then

$$\begin{aligned} p_1(V) &= C'(V) \\ p_{n+1}(V) &= \int_0^\infty p_n(v)C'(V - v)\, dv \end{aligned} \tag{14}$$

where $C'(t) = dC(t)/dt$ is the probability density of $u_r = s_r - t_r$. Equations (14) lead to

$$J(x,V) = xC'(V) + x\int_0^\infty J(x,v)C'(V - v)\, dv \tag{15}$$

$$J(x,V) = \sum_{n=1}^{\infty} x^n p_n(V) \tag{16}$$

where the series and the integral are assumed to converge for $0 \leqq x \leqq 1$. Equation (15) is the integral equation which must be solved to obtain $J(x,V)$.

The integral of $p_n(V)$ from $V = -\infty$ to 0 gives the probability f_n that n demands are served and hence

$$f(x) = \sum_{n=1}^{\infty} x^n f_n = \int_{-\infty}^{0} J(x,V)\, dV. \tag{17}$$

When $a < b$, we expect all busy periods to end. Therefore

$$\sum_{1}^{\infty} f_n = 1 = \int_{-\infty}^{0} J(1,V)\, dV. \tag{18}$$

On the other hand, when the arrival rate a exceeds the service rate b we expect the queue length to increase with time, on the average. At the beginning of operations there may be a few busy periods, but eventually a busy period starts which does not end. This state of affairs is indicated by the fact that $\sum_1^\infty f_n$ is less than 1 when $a > b$. Thus, given a busy period which is just beginning, the probability that it will never end is

$$1 - \sum_{1}^{\infty} f_n = 1 - \int_{-\infty}^{0} J(1,V)\, dV. \tag{19}$$

The length of the idle period following a busy period of length n (services) is $-U_n$. The chance that a busy period will be of length n and will be followed by an idle period whose length lies between y and $y + dy$ is $p_n(-y)dy$. This is true for all values of a/b. When $a < b$, the chance that the length of an idle period picked at random from the universe of idle periods lies between y and $y + dy$ is

$$\sum_{1}^{\infty} p_n(-y)dy = J(1,-y)dy. \tag{20}$$

3.2 *Solution of Integral Equation*

The procedure used in the Appendix to solve Lindley's integral equation (95) suggests writing $J(x,V)$ as the sum of $J_-(x,V)$ and $J_+(x,V)$ where J_- is 0 for $V \geqq 0$ and J_+ is 0 for $V < 0$. Multiplying (15) by $\exp(-sV)$ and integrating with respect to V from $-\infty$ to $+\infty$ gives

$$\Phi_-(x,s) + \Phi_+(x,s) = x\beta(s)\alpha(-s)[1 + \Phi_+(x,s)] \tag{21}$$

where $\mathrm{Re}(s)$ is confined to some suitable range and Φ_+ is the Laplace transform of J_+ and Φ_- is similarly related to J_- [cf. (97)].

Assume functions $\Psi_+(x,s)$, $\Psi_-(x,s)$ and positive numbers D_1, D_2 (which may depend on x) may be found such that

$$x\beta(s)\alpha(-s) - 1 = \frac{\Psi_+(x,s)}{\Psi_-(x,s)} \tag{22}$$

where

(*i*) $\Psi_+(x,s)$ is analytic in s and free from zeros in $\mathrm{Re}(s) > D_1 \geqq 0$,

(*ii*) $\Psi_-(x,s)$ is analytic and free from zeros in $\mathrm{Re}(s) < D_2$ with $D_2 > D_1$ when $0 < x \leqq 1$ and

(*iii*)

$$\Psi_+(x,s) \to s \quad \text{as} \quad |s| \to \infty \quad \text{in} \quad \mathrm{Re}(s) > D_1$$

$$\Psi_-(x,s) \to -s \quad \text{as} \quad |s| \to \infty \quad \text{in} \quad \mathrm{Re}(s) < D_2.$$

It is seen that $\Psi_+(x,s)$ and $\Psi_-(x,s)$ reduce to $\psi_+(s)$ and $\psi_-(s)$ of the Appendix when $x = 1$.

When $x\beta(s)\alpha(-s)$ is eliminated from (21) and (22) one obtains

$$\Phi_-\Psi_- - \Psi_- - s = \Phi_+\Psi_+ + \Psi_+ - s \tag{23}$$

where s has been subtracted from both sides in order to keep them finite at infinity. Considerations of analytic continuation and the analogy with (101) suggest that both sides are equal to some quantity $K(x)$ independent of s. When this turns out to be the case

$$\Phi_+(x,s) = \frac{[K(x) - \Psi_+(x,s) + s]}{\Psi_+(x,s)}.$$

Suppose for the moment that $b > a$. To determine $K(x)$ note (1) that $\Phi_+(x,s)$ is analytic in $\mathrm{Re}(s) \geqq 0$, and (2) that $\Psi_+(x,s)$ has a zero at $s = s_0 \equiv s_0(x)$ where $s_0 = 0$ when $x = 1$ and $s_0 \approx (1 - x)/(a^{-1} - b^{-1})$ when x is near 1. Also note that s_0 is positive when $b > a$ and x is slightly less than 1. Statement (1) is true since $\Phi_+(x,s)$ is the Laplace transform of $J_+(x,V)$ and the integral

$$\int_0^\infty J_+(x,V)\,dV = x\,\Pr(U_1 \geqq 0) + x^2\,\Pr(U_1, U_2 \geqq 0) + \cdots < x(1-x)^{-1}$$

converges. Statement (2) follows from relation (22) and the fact that the expansion of $x\beta(s)\alpha(-s) - 1$ in powers of s shows that it is zero at $s = s_0$. Hence either $\Psi_+(x,s_0)$ is zero or $\Psi_-(x,s_0)$ is infinite. The second possibility is ruled out if $1 - x$ is small enough to make $s_0 < D_2$ since $\Psi_-(x,s)$ is analytic in $\mathrm{Re}(s) < D_2$. Thus $K(x)$ must be taken to be

$-s_0$ in order to keep $\Phi_+(x,s)$ from having a pole in the region $\mathrm{Re}(s) \geqq 0$ at $s = s_0$.

Equating the left-hand side of (23) to $-s_0$ and solving for Φ_- gives

$$\int_{-\infty}^{0} e^{-sV} J_-(x,V)\, dV = \Phi_-(x,s) = 1 + \frac{s - s_0}{\Psi_-(x,s)} \tag{24}$$

$$J_-(x,V) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{sV} \Phi_-(x,s)\, ds \tag{25}$$

where c is chosen so that the path of integration in the s-plane passes to the left of the singularities of $\Phi_-(x,s)$. Comparison of (24) and (17) leads to

$$\sum_{n=1} x^n f_n = \Phi_-(x,0) = 1 - \frac{s_0}{\Psi_-(x,0)} = 1 + \frac{(1-x)s_0}{\Psi_+(x,0)}. \tag{26}$$

In this result, $\Psi_-(x,s)$ and $\Psi_+(x,s)$ are obtained from (22); and s_0, a function of x, is that zero of $x\beta(s)\alpha(-s) - 1$ which approaches $s = 0$ as $x \to 1$. In the foregoing definition of s_0, b is supposed to exceed a. When a and b are in any ratio, the statement is amended to read, "$s = s_0$ is the one and only zero of $\Psi_+(x,s)$ in Re $s > 0$ when $0 \leqq x < 1$." When $a \leqq b$, s_0 tends to 0 as $x \to 1$, and when $a > b$, s_0 tends to a positive number as $x \to 1$. These statements about s_0 are made on the strength of the examples given below in Section 3.3 and hence cannot be regarded as proved in general.

When $a < b$, (26) gives $\Phi_-(1,0) = 1$ in agreement with (18). When $a > b$, s_0 is not zero for $x = 1$, $\Phi_-(1,0)$ is less than 1, and (19) shows that the probability that a busy period will not end is

$$1 - \Phi_-(1,0) = \frac{s_0}{\Psi_-(1,0)} \qquad (a > b) \tag{27}$$

The relations corresponding to (24) and (25) for $J_+(v,V)$ are

$$\int_{0}^{\infty} e^{-sV} J_+(x,V)\, dV = \Phi_+(x,s) = -1 + \frac{s - s_0}{\Psi_+(x,s)} \tag{28}$$

$$J_+(x,V) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{sV} \Phi_+(x,s)\, ds \tag{29}$$

where c is chosen so that the path of integration passes to the right of the singularities of $\Phi_+(x,s)$.

3.3 *Examples*

The special cases used as examples in the Appendix to illustrate the waiting time distribution will be used here for the busy period length. It has been pointed out earlier that $\Psi_{\pm}(x,s)$ reduces to $\psi_{\pm}(s)$ for $x = 1$.

Example a. Poisson Input, Exponential Service. Equation (22) to determine the Ψ's becomes

$$\frac{\Psi_+(x,s)}{\Psi_-(x,s)} = \frac{xab - (a - s)(b + s)}{(a - s)(b + s)} = \frac{s^2 + (b - a)s - (1 - x)ab}{(a - s)(b + s)} = \frac{(s - s_0)(s - s_1)}{(a - s)(b + s)} \tag{30}$$

where

$$\left.\begin{matrix} s_0 \\ s_1 \end{matrix}\right\} = \frac{a - b}{2} \pm \tfrac{1}{2}\sqrt{(b + a)^2 - 4abx}. \tag{31}$$

As x runs from 0 to 1 the roots s_0 and s_1 move along the real axis in the s-plane as shown in Fig. 1.

The conditions for (22) lead us to take, as in (109),

$$\Psi_+(x,s) = \frac{(s - s_0)(s - s_1)}{(s + b)}, \qquad \Psi_-(x,s) = a - s \tag{32}$$

with $D_1 = s_0$ and $D_2 = a$. Expressions (26) give the generating function for the probability f_n that, given a busy period which has just begun, it will consist of exactly n services. Setting $s = 0$ in (32) to obtain $\Psi_-(x,0)$ leads to

$$\sum_{n=1} x^n f_n = 1 - \frac{s_0}{\Psi_-(x,0)} = 1 - \frac{s_0}{a} = \frac{a + b}{2a} - \frac{1}{2a}\sqrt{(b + a)^2 - 4abx}. \tag{33}$$

The coefficient of x^n in the power series expansion of the last expression

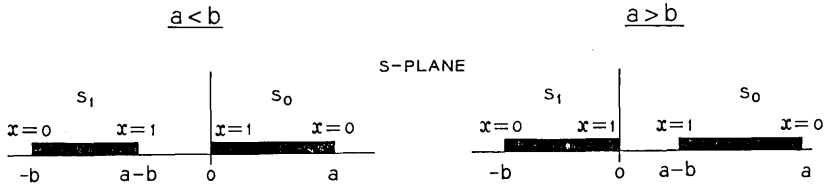


Fig. 1 — Ranges for s_0 and s_1 as x runs from 0 to 1.

gives the value (9) for f_n. When $x = 1$, the sum $\sum_1^\infty f_n$ is equal to 1 when $a < b$ and to b/a when $a > b$. Thus, from (19) or (27) the probability that the busy period will not end is $1 - (b/a)$, with $b < a$.

Inserting the expression

$$\Phi_-(x,s) = 1 + \frac{s - s_0}{a - s} = \frac{a - s_0}{a - s} \tag{34}$$

obtained from (24) in the integral (25), and evaluating the integral by closing the path of integration on the right when $V < 0$, shows that

$$J_-(x,V) = \begin{cases} (a - s_0)e^{aV} & (V < 0) \\ 0 & (V \geqq 0) \end{cases}. \tag{35}$$

From (20), the probability density for the lengths of the idle periods when $a < b$ is

$$J_-(1,-y) = a\, e^{-ay}$$

as expected for Poisson arrivals. When $a > b$, the idle period probability density is still given by the expression on the right, but now $J_-(1,-y)$ has to be divided by the normalizing factor $1 - (s_0/a)$ in which $x = 1$, i.e., by b/a, the probability that the busy period will end.

Although $\Phi_+(x,s)$, $J_+(x,V)$ are not needed to calculate the probabilities given above, their values as obtained from (28) and (29) will be stated for the sake of completeness

$$\Phi_+(x,s) = \frac{b + s_1}{s - s_1}$$

$$J_+(x,V) = \begin{cases} (b + s_1)e^{s_1 V} & (V \geqq 0) \\ 0 & (V < 0) \end{cases}.$$

Note that $b + s_1 = a - s_0$.

Example b. Poisson Input, General Service (Takács[6]). In this case, (22) and the related conditions are satisfied by the analogue of (110)

$$\Psi_+(x,s) = s - a + ax\beta(s), \qquad \Psi_-(x,s) = a - s \tag{36}$$

with $D_1 = s_0$ and $D_2 = a$. With the help of Rouche's theorem it may be shown that $\Psi_+(x,s)$ has only one zero, $s = s_0$, in Re $s > 0$ when $0 \leqq x < 1$. Furthermore, since $\Psi_+(x,0)$ is negative and $\Psi_+(x,a)$ is nonnegative, $0 < s_0 \leqq a$. Equations (24) and (25) show that $\Phi_-(x,s)$ and $J_-(x,V)$ are given by (34) and (35), just as for Poisson input and exponential service, and we still have

$$\sum x^n f_n = 1 - \frac{s_0}{a}. \tag{37}$$

However, s_0 no longer has the simple form (31), but it still tends to 0 if $a \leqq b$ and to a positive number if $a > b$ as $x \to 1$.

When Lagrange's expansion theorem is applied to $1 - (s_0/a)$ it is found that (see Pollaczek,[5] p. 102, Eq. (8.37))

$$f_n = \frac{(-a)^{n-1}}{n!}\left[\frac{d^{n-1}}{ds^{n-1}}\{\beta(s)\}^n\right]_{s=a} \tag{38}$$

which gives (9) and (10) as special cases. It should be recalled that from (3) the average number served in a busy period, for Poisson input and $a > b$, is $\bar{n} = 1/(1 - \rho)$.

Example c. Recurrent Input, Exponential Service. When one takes the steps leading to (111) (which pertains to the waiting time distribution for this case) as a guide, he is led to

$$\Psi_+(x,s) = \frac{(s - s_0)(s - s_1)}{b + s}, \qquad \Psi_-(x,s) = \frac{(s - s_0)(s - s_1)}{xb\alpha(-s) - b - s} \tag{39}$$

where s_1 is the only zero of

$$h(s) = s + b - xb\alpha(-s) \qquad (0 < x < 1) \tag{40}$$

which lies in $\mathrm{Re}(s) > 0$ and s_0 is the left-most zero of $h(s)$ in $\mathrm{Re}(s) > 0$ when x is close to 1. The existence of s_1 may be established with the help of Rouche's theorem. Then $h(0) = (1 - x)b > 0$ and $h(-b) = -x\alpha(b) > 0$ show that $-b > s_1 > 0$. To make the existence of s_0 plausible, consider the case when a is nearly equal to b and x is close to 1. When s is small, the series (114) for $\alpha(-s)$ gives

$$h(s) = (1 - x)b + s(1 - xba^{-1}) - \frac{xba_2s^2}{2} + \cdots. \tag{41}$$

It is seen that $h(s)$ has a double zero at $s = 0$ when $x = 1$ and $a = b$. When $x = 1 - \epsilon$, with ϵ small and positive, and $1 - ba^{-1} = \eta$ is small, the double zero splits into two simple zeros given approximately by

$$0 = -\epsilon b - s\eta + \frac{ba_2s^2}{2}.$$

The two roots of this equation are small, real, and of opposite sign. The positive root corresponds to s_0 and the negative one to s_1.

From (26)

$$\sum_1^\infty x^n f_n = 1 + \frac{(1 - x)s_0}{\Psi_+(x,0)} = 1 + \frac{(1 - x)b}{s_1} \tag{42}$$

and Lagrange's expansion theorem gives (see Pollaczek,[5] p. 103, Eq. (8.42))

$$f_n = c_{n-1} - c_n \qquad (c_0 = 1)$$
$$c_n = \frac{b^{n+1}}{n!}\left[\frac{d^{n-1}}{ds^{n-1}}\{s^{-2}[\alpha(-s)]^n\}\right]_{s=-b}. \tag{43}$$

When $a < b$, the average number served in a busy period may be obtained either by differentiating (42) and setting $x = 1$, or using the value of $W(0)$ obtained just below (111). Both methods give

$$\bar{n} = \frac{1}{W(0)} = \left[-\frac{b}{s_1}\right]_{x=1}. \tag{44}$$

Example d. Erlangian Input, Erlangian Service. Let $f(s)$ denote the same polynomial in s as in Example d of the Appendix. Then

$$x\beta(s)\alpha(-s) - 1 = x\left(1 + \frac{s}{bk}\right)^{-k}\left(1 - \frac{s}{al}\right)^{-l} - 1 = \frac{x - f(s)}{f(s)}.$$

When $x = 0$ the polynomial $x - f(s)$ has a zero of order k at $s = -bk$ and a zero of order l at $s = al$. Now let $0 < x < 1$. From Rouche's theorem and the fact that $|f(s)| > x$ on the imaginary s axis, it may be shown that $x - f(s)$ has k zeros in $\mathrm{Re}(s) < 0$ and l zeros in $\mathrm{Re}(s) > 0$. Denote the zeros in $\mathrm{Re}(s) < 0$ by $s_1, \cdots, s_k$, the left-most zero (when x is close to 1) in $\mathrm{Re}(s) > 0$ by s_0, and the remaining zeros in $\mathrm{Re}(s) > 0$ by $s_{k+1}, \cdots, s_{k+l-1}$. Then (22) takes the form

$$\frac{x - f(s)}{f(s)} = -\frac{(s - s_0)(s - s_1)\cdots(s - s_{k+l-1})}{(s - al)^l(s + bk)^k} = \frac{\Psi_+(x,s)}{\Psi_-(x,s)} \tag{45}$$

and the conditions on the Ψ's are satisfied by

$$\Psi_+(x,s) = \frac{(s - s_0)(s - s_1)\cdots(s - s_k)}{(s + bk)^k}$$
$$\Psi_-(x,s) = -\frac{(s - al)^l}{(s - s_{k+1})\cdots(s - s_{k+l-1})}. \tag{46}$$

From (26) the generating function for the number served in a busy period is

$$\sum_1^\infty x^n f_n = 1 - (al)^{-l}s_0s_{k+1}s_{k+2}\cdots s_{k+l-1} = 1 - \frac{(1 - x)(-bk)^k}{s_1\cdots s_k}$$

where $s_0, s_1, \cdots, s_{k+l-1}$ are functions of x.

IV. DURATION OF BUSY PERIOD

The distribution of the busy period lengths for recurrent input and general service may be obtained by an extension of the foregoing analysis. A number of the steps will be omitted here because of the similarity with the work of Section III. For the sake of simplicity, throughout this section it will be assumed that $a < b$ and hence that all busy periods eventually end.

4.1 *Derivation of Integral Equation*

Let $p_n(V,S)dV\,dS$ be the probability that $U_1, \cdots, U_{n-1} \geqq 0$, $V < U_n < V + dV$ and $S < S_n < S + dS$ where U_n is defined by (13) and S_n is the sum $s_1 + s_2 + \cdots + s_n$ of the first n service times. It may be shown that $p_n(V,S)$ is zero for $V > S$ and

$$p_1(V,S) = B'(S)A'(S - V)$$

$$p_{n+1}(V,S) = \int_0^S d\sigma \int_0^\sigma dv p_n(v,\sigma)B'(S - \sigma)A'(S - V - \sigma + v).$$

These may be combined to give the integral equation

$$\begin{aligned} J(x,V,S) &= xB'(S)A'(S - V) \\ &\quad + x\int_0^S d\sigma \int_0^\sigma dv J(x,v,\sigma)B'(S - \sigma)A'(S - V - \sigma + v) \end{aligned} \tag{47}$$

where

$$J(x,V,S) = \sum_{n=1}^{\infty} x^n p_n(V,S). \tag{48}$$

The probability that a busy period will consist of exactly n services, have a length between S and $S + dS$, and be followed by an idle period of length between y and $y + dy$ is

$$p_n(-y,S)dS\,dy = G_n'(y,S)dS\,dy \tag{49}$$

where $G_n'(y,S)$ is the density function introduced in Section II. The probability that a busy period will consist of exactly n services and have a length between S and $S + dS$ is dS times

$$G_n'(S) = \int_{-\infty}^{0} p_n(V,S)\,dV. \tag{50}$$

Summing with respect to n shows that the probability of a busy period length between S and $S + dS$ is dS times

$$G'(S) = \int_{-\infty}^{0} J(1,V,S)\, dV. \tag{51}$$

Furthermore, the probability the busy period consists of exactly n services is

$$f_n = \int_0^{\infty} dS \int_{-\infty}^{0} p_n(V,S)\, dV.$$

4.2 *Solution of Integral Equations*

Multiplying (47) by $\exp[-zS - sV]$, and integrating V from $-\infty$ to S and S from 0 to ∞ gives

$$\begin{aligned} \Phi_-(x,s,z) + \Phi_+(x,s,z) &= x\beta(s+z)\alpha(-s)[1 + \Phi_+(x,s,z)] \\ \Phi_+(x,s,z) &= \int_0^{\infty} dS \int_0^{S} dV\, e^{-zS-sV} J(x,V,S) \\ \Phi_-(x,s,z) &= \int_0^{\infty} dS \int_{-\infty}^{0} dV\, e^{-zS-sV} J(x,V,S). \end{aligned} \tag{52}$$

From the last equation

$$\int_{-\infty}^{0} e^{-sV} J(x,V,S)\, dV = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{zS} \Phi_-(x,s,z)\, dz \tag{53}$$

where c is such that the singularities of the integrand lie to the left of the path of integration.

Assume the factorization

$$x\beta(s+z)\alpha(-s) - 1 = \frac{\Psi_+(x,s,z)}{\Psi_-(x,s,z)} \tag{54}$$

where the Ψ functions satisfy conditions (*i*), (*ii*), (*iii*) set forth in connection with (22), z being regarded as an imaginary constant (or is at least on the path of integration $\mathrm{Re}(z) = c$ in (53)). When $z = 0$, $\Psi_\pm(x,s,z)$ reduces to the $\Psi_\pm(x,s)$ of (22) and when $z = 0$ and $x = 1$, to the $\psi_\pm(s)$ of (99). The analogue of (24) turns out to be

$$\Phi_-(x,s,z) = 1 + \frac{s - s_0(x,z)}{\Psi_-(x,s,z)} \tag{55}$$

where $s = s_0(x,z)$ is that root of

$$x\beta(s + z)\alpha(-s) - 1 = 0 \tag{56}$$

which tends to $s = 0$ as $x \to 1$ and $z \to 0$.

The various probabilities of interest may be computed, at least in theory, from $\Phi_-(x,s,z)$. The relations are indicated in Table I. The Laplace transforms $\gamma_n(s,z)$, $\gamma_n(z)$, $\gamma(z)$ are the same as those introduced in Section II.

From (55) and the last two entries in the table it follows that the Laplace transforms of $\sum x^n G_n'(S)$ and $G'(S)$ are respectively

$$\Phi_-(x,0,z) = 1 - \frac{s_0(x,z)}{\Psi_-(x,0,z)} = 1 + \frac{[1 - x\beta(z)]s_0(x,z)}{\Psi_+(x,0,z)} \tag{57}$$

$$\gamma(z) = \Phi_-(1,0,z) = 1 - \frac{s_0(1,z)}{\Psi_-(1,0,z)} = 1 + \frac{[1 - \beta(z)]s_0(1,z)}{\Psi_+(1,0,z)}. \tag{58}$$

Furthermore, the generating function (26) for f_n may also be written as $\Phi_-(x,0,0)$.

4.3 *Examples*

The special cases used earlier will again serve as examples. The results of Examples *b* and *c* are equivalent to results given by Pollaczek.

TABLE I

Probability Function	Laplace Transform
$G_n'(y,S) = p_n(-y,S)$	$\gamma_n(s,z) =$ Expression (11) $= \int_0^\infty dS \int_{-\infty}^0 dV e^{-zS-sV} p_n(V,S)$
$\sum_1^\infty x^n G_n'(y,S) = J(x,-y,S)$	$\sum_1^\infty x^n \gamma_n(s,z) = \Phi_-(x,s,z)$
$G_n'(S)$	$\int_0^\infty e^{-zS} G_n'(S)\, dS = \gamma_n(0,z) = \gamma_n(z)$
$\sum_1^\infty x^n G_n'(S) = \int_{-\infty}^0 J(x,V,S)\, dV$	$\sum_1^n x^n \gamma_n(z) = \Phi_-(x,0,z)$
$G'(S) = \int_{-\infty}^0 J(1,V,S)\, dV$	$\int_0^\infty e^{-zS} G'(S)\, dS = \gamma(z) = \Phi_-(1,0,z)$

Example a. Poisson Input, Exponential Service. In this case (54) becomes

$$\frac{\Psi_+(x,s,z)}{\Psi_-(x,s,z)} = \frac{s^2 + s(z + b - a) - a(z + b - xb)}{(s + b + z)(a - s)} = \frac{(s - s_0)(s - s_1)}{(s + b + z)(a - s)} \tag{59}$$

where $s_j \equiv s_j(x,z)$ is given by

$$\left.\begin{matrix} s_0 \\ s_1 \end{matrix}\right\} = \frac{a - b - z}{2} \pm \frac{1}{2}\sqrt{(b + a + z)^2 - 4xab}. \tag{60}$$

When x is fixed and z runs from $-i\infty$ to $+i\infty$, s_0 traverses an oval-shaped path in the s-plane. The oval lies in $\mathrm{Re}(s) \geqq 0$, it starts and ends at $s = a$ (corresponding to $z = \pm i\infty$), and when $z = 0$ it crosses the real s-axis between $s = 0$ and $s = a$ as indicated by Fig. 1, case $a < b$. At the same time, s_1 traverses a path roughly parallel to the imaginary axis. The path lies in the region $\mathrm{Re}(s) < 0$, is asymptotic to the line $\mathrm{Re}(s) = -b$ at $z = \pm i\infty$, and crosses the real s-axis between $s = -b$ and $s = a - b$. As $x \to 0$, the oval shrinks to the point $s = a$ and the path of s_1 tends to the straight line $\mathrm{Re}(s) = -b$.

This behavior of s_0 and s_1 and the similarity with (32) lead us to take

$$\Psi_+(x,s,z) = \frac{(s - s_0)(s - s_1)}{s + b + z}, \qquad \Psi_-(x,s,z) = a - s. \tag{61}$$

From (57)

$$\Phi_-(x,0,z) = 1 - \frac{s_0(x,z)}{a} = \frac{1}{2a}[b + a + z - \sqrt{(b + a + z)^2 - 4xab}]. \tag{62}$$

Inverting this Laplace transform (see, for instance, Pair 556.1, Campbell and Foster[16]) gives (Takács[6])

$$\sum_1^\infty x^n G_n'(S) = S^{-1}\sqrt{\frac{xb}{a}}\, e^{-(b+a)S} I_1(S\sqrt{4xab}) \tag{63}$$

where $I_1(z)$ is a Bessel function of the first kind for imaginary argument. Putting $x = 1$ in (63) gives Kendall's[3] expression (6) for $G'(S)$.

Example b. Poisson Input, General Service (Prabhu[17]). In analogy with (36)

$$\Psi_+(x,s,z) = s - a + xa\beta(s + z), \qquad \Psi_-(x,s,z) = a - s \tag{64}$$

and (57) gives

$$\Phi_-(x,0,z) = 1 - a^{-1}s_0(x,z) \tag{65}$$

where $s = s_0(x,z)$ is that root of

$$s - a + xa\beta(s + z) = 0 \tag{66}$$

which tends to zero as $x \to 1$ and $z \to 0$. Rouche's theorem may be used to show that if $0 \leqq x < 1$ and $\mathrm{Re}\, z \geqq 0$, then (66) has one and only one root (namely, $s_0(x,z)$) in $\mathrm{Re}(z + s) > 0$.

When s in (66) is replaced by $s_0(x,z) = a - a\Phi_-$ (from (65) with $\Phi_- \equiv \Phi_-(x,0,z)$) one obtains

$$\Phi_- = x\beta(z + a - a\Phi_-). \tag{67}$$

This is a functional equation satisfied by the generating function

$$\Phi_- = \sum x^n \gamma_n(z). \tag{68}$$

Setting $x = 1$ gives Kendall's equation (5)[3]

$$\gamma(z) = \beta(z + a - a\gamma(z))$$

for the Laplace transform of $G'(S)$.

Example c. Recurrent Input, Exponential Service (Connolly,[8] Takács[7]). The analogue of (39) is

$$\Psi_+(x,s,z) = \frac{(s - s_0)(s - s_1)}{s + b + z}, \qquad \Psi_-(x,s,z) = \frac{(s - s_0)(s - s_1)}{xb\alpha(-s) - s - b - z} \tag{69}$$

where $s = s_1$ is the only root of

$$xb\alpha(-s) - s - b - z = 0 \tag{70}$$

which has a negative real part when $\mathrm{Re}(z) \geqq 0$ and $0 \leqq x \leqq 1$, and $s = s_0$ is the root which tends to $s = 0$ when $z \to 0$ and $x \to 1$.

Equations (57) and (58) give

$$\Phi_-(x,0,z) = 1 + \frac{z + (1 - x)b}{s_1(x,z)}, \qquad \gamma(z) = 1 + \frac{z}{s_1(1,z)} \tag{71}$$

which may be used in Table I. The coefficients of $-z$ and $z^2/2$ in the power series expansion of $\gamma(z)$ give the first and second moments of the

distribution of the busy period length S. With the help of (70) it is found that

$$\bar{S} = -\frac{1}{s_{10}}$$
$$\frac{\text{variance } S}{[\bar{S}]^2} = \left[\frac{1 - b\alpha'(-s_{10})}{1 + b\alpha'(-s_{10})}\right] \tag{72}$$

where s_{10} denotes $s_1(1,z)$ for $z = 0$ and $\alpha'(s) = d\alpha'(s)/ds$. This value of $\bar{S}$ is equal to $1/bW(0)$ where $W(0)$ may be obtained from Example c in the Appendix.

Example d. Erlangian Input, Erlangian Service (special case of Connolly[9]). In this case the value of $\Psi_-(x,s,z)$ is given by the $\Psi_-(x,s)$ of (46) where now $s_0, \cdots, s_{k+l-1}$ are the roots of

$$x - \left(1 - \frac{s}{al}\right)^l \left(1 + \frac{s+z}{bk}\right)^k = 0.$$

Replacing $(s + bk)$ by $(s + z + bk)$ in the denominator of expression (46) for $\Psi_+(x,s)$ give $\Psi_+(x,s,z)$. As $x \to 0$, the k zeros $s_1, \cdots, s_k$ cluster around $s = -z - bk$ while the l zeros $s_0, s_{k+1}, \cdots, s_{k+l-1}$ cluster around $s = al$.

Example e. Arrival Rate Almost as Large as Service Rate. For this case some of the busy periods tend to be long. An idea of the behavior of the probability density $G'(S)$ for large S may be obtained by the following heuristic procedure.

Inverting the first of expressions (58) for the Laplace transform $\gamma(z)$, taking $S > 0$, and deforming the path of integration gives

$$G'(S) = \frac{1}{2\pi i}\int_C e^{zS}\left[1 - \frac{s_0(1,z)}{\Psi_-(1,0,z)}\right]dz \tag{73}$$

where C is a large loop starting and ending at $z = -\infty$ and enclosing the singularities of the integrand. The right-most of these singularities determine the form of $G'(S)$ when S is large. For the special case of Poisson input and exponential service, (62) with $x = 1$ shows that the singularities of the integrand are branch points at $z = -(\sqrt{b} - \sqrt{a})^2$ and $-(\sqrt{b} + \sqrt{a})^2$. Since the right-most branch point approaches $z = 0$ as $a \to b$, we assume that the principal contribution to $G'(S)$ in (73) comes from the region around $z = 0$ when S is large. This rules out the case of constant service time. For constant service time, the busy period lengths are multiples of b^{-1}, and $G'(S)$ is given by $\sum f_n\delta(S - nb^{-1})$ where $\delta(t)$ is the unit impulse function. A special case is illustrated below by (76).

When $x = 1$ and $z \to 0$, $s_0(1,z)$ tends to zero. To determine $s_0(1,z)$ when a is nearly equal to b, consider the expression obtained from the first few terms of the power series (114). As in Example e in the Appendix

$$\beta(s + z)\alpha(-s) - 1 \approx -b_1 z + (a_1 - b_1)s + 2^{-1}s^2(v_A + v_B) \quad (74)$$

where $a_1 = a^{-1}$, $b_1 = b^{-1}$ and v_A, v_B denote variances. All three terms in (74) are assumed to be of the same order of magnitude. Taking $v_A + v_B$ to be $0(b^{-2})$, it follows that s is $0(b(1 - \rho))$ and z is $0(b(1 - \rho)^2)$ where $\rho = a/b$ and $(1 - \rho) \ll 1$.

The expression on the right side of (74) vanishes for two values of s. The value which approaches $s = 0$ as $z \to 0$ is

$$s_0(1,z) = \frac{-(a_1 - b_1) + \sqrt{(a_1 - b_1)^2 + 2b_1(v_A + v_B)z}}{v_A + v_B}.$$

When this is substituted in (73), only the radical contributes to the value of the integral. Upon replacing $\Psi_-(1,0,z)$ by $\Psi_-(1,0,0)$ and using

$$\frac{1}{2\pi i}\int_C e^{zS}\sqrt{A + Bz}\, dz = -\frac{1}{2}\, S^{-3/2}\sqrt{\frac{B}{\pi}}\, e^{-AS/B},$$

$$\Psi_-(1,0,0) = \psi_-(0) = (a^{-1} - b^{-1})^{-1}W(0)$$

where $W(0)$ is the probability that a customer does not have to wait, it is found that

$$G'(S) \approx \frac{S^{-3/2}}{bW(0)}\left(\frac{D}{\pi}\right)^{\frac{1}{2}} e^{-DS}. \quad (75)$$

Here S is supposed to be large, $1 - \rho \ll 1$, and

$$D = \frac{(a_1 - b_1)^2}{2b_1(v_A + v_B)} = \frac{(1 - \rho)^2}{2a\rho(v_A + v_B)} = \frac{1 - \rho}{4\rho\bar{w}}$$

where $\bar{w}$ is the approximate average waiting time $-1/s_1$ given by (113). From (2) it is seen that $1/bW(0)$ may be replaced by $\bar{S}$, the average length of a busy period.

For Poisson input and arbitrary service $W(0) = 1 - \rho$ and for recurrent input and exponential service $W(0)$ is $-s_1/b$ where s_1 is the negative root of $s + b - b\alpha(-s) = 0$. However, as Example e in the Appendix shows, there appears to be no simple expression for $W(0)$ in the general case as $a \to b$.

It is interesting to note that for Poisson input and constant service time (10) gives

$$G'(S) = \sum_1^\infty \delta(S - nb^{-1})(\rho n)^{n-1}e^{-n\rho}/n! \quad (76)$$

where $\delta(t)$ is a unit impulse. Since the right-hand side of (75) is a continuous function of S, it cannot hold in this case. However, formal calculation of (75) gives

$$G'(S) \approx S^{-3/2}(2\pi b)^{-\frac{1}{2}}e^{-bS(1-\rho)^2/2}. \tag{77}$$

By using $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$ and $\rho \approx 1$ it may be shown that (77) becomes a smoothed version of (76) when S is replaced by nb^{-1}. A better approximation to (76) may be obtained by noting that the Laplace transform $\gamma(z)$ of (76) is of the form $\sum f_n \exp(-znb^{-1})$. Hence $\gamma(z)$ is a periodic function of z with period $i2\pi b$, and contributions to (73) may be expected not only from the region around $z = 0$ but also from regions near $z = i2\pi kb$ where $k = \cdots, -1, 0, 1, 2, \cdots$. It may be shown that the sum of these contributions does indeed approximate (76) when n is large and $p \approx 1$.

The approximation (75) holds for large values of S. One may obtain an idea of the behavior of $G'(S)$ for small values of S by noting that now the busy period contains only a small number of services. In particular, the assumption that the very short busy periods consist of a single service leads to the approximation $G'(S) \approx G_1'(S) = B'(S)[1 - A(S)]$. Higher-order approximations may be obtained by computing $p_1(V,S)$, $p_2(V,S)$, $\cdots$ step by step and using (50). The first step gives the result cited, namely

$$G_1'(S) = \int_{-\infty}^{0} B'(S)A'(S - V)\, dV = B'(S)[1 - A(S)]. \tag{78}$$

V. GROWTH OF QUEUE

In Section III the functions $\Psi_{\pm}(x,s)$ were introduced to obtain the chance f_n that a busy period will consist of n services. It is interesting to note that these functions may also be used to determine $W_r(t)$, the waiting time distribution for the rth customer. Only a sketch of the procedure is given here. More complete information on $W_r(t)$ is given by Pollaczek,[5] Spitzer,[18] and Lindley.[13]

5.1 *Solution of Integral Equation*

Upon setting

$$L_+(x,t) = \sum_0^\infty x^r W_r(t) \tag{79}$$

so that $L_+(x,t) = 0$ for $t < 0$, it is found from (94) that

$$L_+(x,t) = W_0(t) + x \int_{-\infty}^{\infty} L_+(x,t-\tau)\, dC(\tau)\,. \tag{80}$$

This holds only for $t \geqq 0$. To solve the integral equation, let $L_-(x,t)$ denote the value of the right-hand side for $t < 0$ and set $L_-(x,t) = 0$ for $t \geqq 0$. It is found that

$$\varphi_-(x,s) + \varphi_+(x,s) = s^{-1} + x\varphi_+(x,s)\beta(s)\alpha(-s) \tag{81}$$

where φ_+ is the Laplace transform of L_+ and φ_- is similarly related to L_- (cf. (97)). Equation (81) is similar to (21). When the functions $\Psi_\pm(x,s)$ appearing in (22) are introduced, the analogue of (23) turns out to be

$$\begin{aligned}\varphi_-(x,s)\Psi_-(x,s) - s^{-1}\Psi_-(x,s) + s^{-1}\Psi_-(x,0)& \\ = \varphi_+(x,s)\Psi_+(x,s) + s^{-1}\Psi_-(x,0)&\end{aligned} \tag{82}$$

where $s^{-1}\Psi_-(x,0)$ is added to both sides in order to cancel the pole of $s^{-1}\Psi_-(x,s)$ at $s = 0$. Setting both sides of (82) equal to a quantity $K(x)$ independent of s leads to

$$\varphi_+(x,s) = \frac{[s - s_0(x)]\Psi_-(x,0)}{s\Psi_+(x,s)s_0(x)}$$

where $s = s_0(x)$ is the one and only zero of $\Psi_+(x,s)$ in $\mathrm{Re}(s) > 0$.

Inversion gives the required result:

$$\sum_0^\infty x^r W_r(t) = \frac{\Psi_-(x,0)}{s_0(x)} \frac{1}{2\pi i} \int_{\epsilon-i\infty}^{\epsilon+i\infty} \frac{[s - s_0(x)]e^{st}}{s\Psi_+(x,s)}\, ds \qquad (\epsilon > 0)\,. \tag{83}$$

5.2 *Poisson Input and Exponential Service*

For the special case of Poisson input and exponential service, the values of $\Psi_\pm(x,s)$ are given by (32). Insertion in (83) gives

$$\begin{aligned}\sum_0^\infty x^r W_r(t) &= \frac{1}{2\pi i} \int_{\epsilon-i\infty}^{\epsilon+i\infty} \frac{a(s+b)e^{st}}{s_0 s(s - s_1)}\, ds \\ &= \frac{1}{1-x} - \frac{(a+b)e^{-bt}}{2b(1-x)} (1 - \sqrt{1-cx})e^{\tau - \tau\sqrt{1-cx}}\end{aligned} \tag{84}$$

where $\epsilon > 0$, $t > 0$ and

$$c = \frac{4ab}{(a+b)^2}, \qquad \tau = \frac{t(b+a)}{2}\,. \tag{85}$$

Upon using

$$(1 - \sqrt{1 - cx})^{k+1} = \sum_{n=k}^{\infty} \frac{(cx)^{n+1}(2n - k)!(k + 1)2^{k-1}}{(n + 1)!(n - k)!4^n}$$
$$(1 - \sqrt{1 - cx})\exp[\tau - \tau\sqrt{1 - cx}] = \frac{1}{2}\sum_{n=0}^{\infty} \frac{(cx)^{n+1}}{(n + 1)!4^n} P(n,\tau) \tag{86}$$

it is found that

$$W_r(t) = 1 - \frac{ae^{-bt}}{a + b}\sum_{n=0}^{r-1} \frac{c^n P(n,\tau)}{(n + 1)!4^n} \tag{87}$$

where $P(n,\tau)$ is the polynomial

$$\begin{aligned} P(n,\tau) &= \sum_{k=0}^{n} \frac{(2n - k)!(k + 1)(2\tau)^k}{k!(n - k)!} \\ &= (2\tau)^{n+\frac{1}{2}} e^{\tau} \pi^{-\frac{1}{2}}[(1 + \tau)K_{n+\frac{1}{2}}(\tau) - \tau K_{n-\frac{1}{2}}(\tau)] \end{aligned} \tag{88}$$

and K denotes a Bessel function of the second kind for imaginary argument.[19] It may be shown that $P(n,\tau)$ is $0(4^n n!/\sqrt{n})$ as $n \to \infty$ with τ finite.

Setting $x = 1$ in (86) and using the result to transform (87) leads to

$$W_r(t) = 1 - \rho\, e^{-(b-a)t} + a\, e^{-bt} F_r(t) \qquad (a \leqq b)$$
$$W_r(t) = 0 + a\, e^{-bt} F_r(t) \qquad (a \geqq b)$$

where

$$F_r(t) = (a + b)^{-1} \sum_{n=r}^{\infty} \frac{c^n P(n,\tau)}{(n + 1)!4^n}.$$

The value of $F_r(t)$ is unchanged when the values of a and b are interchanged.

When $a = b = 1$, differentiation of the generating function (84) leads to expressions for the probability density $W_r'(t)$, $r > 0$:

$$\begin{aligned} W_r'(t) &= \frac{t^r}{2^r r!}\left(\frac{2t}{\pi}\right)^{\frac{1}{2}} [K_{r+\frac{1}{2}}(t) - K_{r-\frac{1}{2}}(t)] \\ &= \frac{e^{-t}}{2^{2r-1}} \sum_{n=0}^{r-1} \frac{(2r - n - 1)!(2t)^n}{r!n!(r - n - 1)!} \\ &= \frac{2}{\pi}\int_0^{\infty} \frac{\cos tu\, du}{(u^2 + 1)^r}\left[\frac{1}{u^2 + 1} - \frac{t}{2r}\right]. \end{aligned} \tag{89}$$

5.3 *Queue Behavior as* $r \to \infty$

Return now to the case of recurrent input and general service. When r becomes large $W_r(t)$ approaches the limit $W(t)$ discussed in the Appendix if $a < b$. Questions related to this approach have been studied by Lindley, Pollaczek, Spitzer and others.

When $a > b$, the customers arrive faster than the server can handle them and the waiting line tends to grow steadily. An idea of the behavior of $W_r(t)$ in this case may be obtained by using the notation of the Appendix. The rth customer arrives at time $T_r = t_0 + t_1 + \cdots + t_{r-1}$ and his service begins at $S_r + I_r$, where $S_r = s_0 + s_1 + \cdots + s_{r-1}$ and I_r is the total amount of time the server is idle in the interval $(0,T_r)$. The waiting time of the rth customer is $w_r = S_r + I_r - T_r$. Since the server may be expected to be continuously busy after a few initial idle periods, we expect I_r to approach some constant value as $r \to \infty$. Thus, I_r becomes small in comparison with $S_r - T_r = u_0 + u_1 + \cdots + u_{r-1}$. Upon making the approximation $w_r \approx S_r - T_r$ and using the central limit theorem, it is found that, for $a > b$ and $r \gg 1$

$$W_r'(t) = \frac{dW_r(t)}{dt} \approx \frac{1}{\sigma\sqrt{2\pi r}} \exp\left[-\frac{(t-\mu r)^2}{2r\sigma^2}\right]. \tag{90}$$

Here $\mu = b^{-1} - a^{-1}$ is the average value of $u_r = s_r - t_r$ and $\sigma^2 = v_B + v_A$ is its variance, v_B and v_A being the respective variances of s_r and t_r.

When $a = b$, the approximation (90) no longer holds. However, Pollaczek[5] has shown that, at least for Poisson input, in place of (90) we have

$$W_r'(t) \approx \frac{2}{\sigma\sqrt{2\pi r}} \exp\left[-\frac{t^2}{2r\sigma^2}\right] \qquad (t > 0). \tag{91}$$

This agrees with the asymptotic form of the integral in (89).

Appendix

*Waiting Time Distribution**

In Sections III and IV the busy period problem has been investigated by a method similar to the Wiener-Hopf technique used by Smith[14] to deal with the waiting time distribution. The application of the method to waiting time problems is reviewed in this Appendix. Smith's approach has been changed slightly in order to make it fit in better with the busy period problem.

* Here the "waiting time" of a customer is the interval between his arrival and the instant his service begins.

A.1 *Derivation of Integral Equation*

The basic integral equation is due to Lindley.[13] Suppose the single server system (with unrestricted queue length, recurrent input, and general service) starts operations when the 0th customer arrives at time 0. Following Lindley, write w_r for the waiting time of the rth customer, s_r for his service time, and t_r for the interval between the arrivals of customers r and $r + 1$. The rth customer stays in the system for an interval of length $w_r + s_r$, and the $(r + 1)$th customer arrives t_r units of time after this interval begins. If $w_r + s_r \leqq t_r$, the $(r + 1)$th customer does not have to wait and w_{r+1} is zero. When $w_r + s_r > t_r$, the $(r + 1)$th customer has to wait $w_r + s_r - t_r$ units of time. Thus

$$\begin{aligned} w_{r+1} &= 0 && (w_r + s_r \leqq t_r) \\ w_{r+1} &= w_r + s_r - t_r && (w_r + s_r > t_r). \end{aligned} \tag{92}$$

It is assumed that the s_r are independent random variables with the common distribution function $B(t)$, and also that the t_r are independent (of each other and of the s_r) with common distribution function $A(t)$. The rule for combining probability distributions shows that the distribution function $C(t)$ of the variable $u_r = s_r - t_r$ is given by

$$C(t) = \int_0^\infty B(t + \tau)\, dA(\tau). \tag{93}$$

Since w_r and u_r are independent random variables, it follows upon rewriting (92) as

$$\begin{aligned} w_{r+1} &= 0 && (w_r + u_r \leqq 0) \\ w_{r+1} &= w_r + u_r && (w_r + u_r > 0) \end{aligned}$$

that the distribution function for w_{r+1} is

$$W_{r+1}(t) = \int_{-\infty}^{t} W_r(t - \tau)\, dC(\tau) \qquad (t \geqq 0). \tag{94}$$

By starting with $W_0(t) = 1$ for $t \geqq 0$, which states that the 0th customer is served immediately, one may compute $W_1(t)$, $W_2(t)$, $\cdots$, in succession from (94). When the service rate b exceeds the arrival rate a, i.e., when $\rho = a/b < 1$, $W_r(t)$ tends to $W(t)$ as $r \to \infty$ where $W(t)$ satisfies Lindley's integral equation

$$W(t) = \int_{-\infty}^{t} W(t - \tau)\, dC(\tau) \qquad (t \geqq 0). \tag{95}$$

$W(t)$ is the distribution function for the waiting time when statistical

equilibrium prevails. Note that (95) holds only for nonnegative t. When t is negative, $W(t)$ is zero but the integral on the right does not vanish.

A.2 *Solution of Integral Equation*

Let $W_-(t)$ be the value of the integral for $t < 0$ and take $W_-(t) = 0$ for $t \geqq 0$. Since $W(t - \tau)$ is zero for $\tau > t$, (95) may be written as

$$W_-(t) + W(t) = \int_{-\infty}^{\infty} W(t - \tau)C'(\tau)\, d\tau \tag{96}$$

which holds for all real values of t. The derivative $C'(\tau) = dC(\tau)/d\tau$ is the probability density of the random variable $u_r = s_r - t_r$.

Multiply both sides of (96) by $\exp(-st)$, where $0 < \mathrm{Re}(s) < D$ with D such that the following integrals converge. The existence of D is ensured by assumptions made below. Integrate from $t = -\infty$ to $t = \infty$ and introduce the transforms

$$\begin{aligned} \varphi_+(s) &= \int_{-\infty}^{\infty} e^{-st}W(t)\, dt = \int_0^{\infty} e^{-st}W(t)\, dt \\ \varphi_-(s) &= \int_{-\infty}^{\infty} e^{-st}W_-(t)\, dt = \int_{-\infty}^{0} e^{-st}W_-(t)\, dt \end{aligned} \tag{97}$$

$$\int_{-\infty}^{\infty} e^{-st}C'(\tau)\, d\tau = \text{ave exp}\,[-ss_r + st_r] = \beta(s)\alpha(-s)$$

where $\beta(s)$ and $\alpha(s)$ are the respective Laplace-Stieltjes transforms of the service and interarrival distribution functions $B(t)$ and $A(t)$. This carries (96) into

$$\begin{aligned} \varphi_-(s) + \varphi_+(s) &= \varphi_+(s)\beta(s)\alpha(-s) \\ \varphi_-(s) &= \varphi_+(s)[\beta(s)\alpha(-s) - 1]. \end{aligned} \tag{98}$$

Since $W(t)$ and $B(t)$ are distribution functions, both $\varphi_+(s)$ and $\beta(s)$ are analytic in the region $\mathrm{Re}(s) > 0$. To ensure convergence of the integrals in (97) involving $W_-(t)$ and $C'(\tau)$, assume that the probability density $A'(t) = dA(t)/dt$ exists and is $0[\exp(-Dt)]$ as $t \to \infty$, where D is positive but may be arbitrarily small. It may then be shown from (93) that $C(t)$ is $0[\exp(Dt)]$ as $t \to -\infty$ and, using this in (95), that $W_-(t)$ is also $0[\exp(Dt)]$ as $t \to -\infty$. It follows that both $\varphi_-(s)$ and $\alpha(-s) = \text{ave}\exp(st_r)$ are analytic in the region $\mathrm{Re}(s) < D$.

Now suppose that functions $\psi_+(s)$ and $\psi_-(s)$ may be found such that

$$\beta(s)\alpha(-s) - 1 = \frac{\psi_+(s)}{\psi_-(s)} \tag{99}$$

where (i) $\psi_+(s)$ is analytic and free from zeros in the half-plane $\mathrm{Re}(s) > 0$, and (ii) $\psi_-(s)$ is analytic and free from zeros in $\mathrm{Re}(s) < D$. Although these functions may be expressed as integrals when suitable conditions are satisfied (see, for example, Smith[14]) their expression in tractable form is usually the most difficult step in obtaining $W(t)$. For future convenience assume that $\psi_+(s)$ and $\psi_-(s)$ may be chosen so that

$$\begin{aligned} \psi_+(s) &\to s \quad \text{as} \quad |s| \to \infty \quad \text{in} \quad \mathrm{Re}(s) > 0 \\ \psi_-(s) &\to -s \quad \text{as} \quad |s| \to \infty \quad \text{in} \quad \mathrm{Re}(s) < D. \end{aligned} \tag{100}$$

The difference in sign is required by the fact that the left-hand side of (99) tends to -1 as $s \to \pm i\infty$ (unless both $A(t)$ and $B(t)$ have discontinuous jumps, a case we shall rule out).

When the resolution (99) is possible, (98) becomes

$$\varphi_-(s)\psi_-(s) = \varphi_+(s)\psi_+(s), \qquad 0 < \mathrm{Re}(s) < D. \tag{101}$$

The right-hand side is analytic for $\mathrm{Re}(s) > 0$ and the left hand for $\mathrm{Re}(s) < D$. Equality in the strip implies that each is the analytic continuation of a function which has no singularities in the finite part of the s-plane. It turns out that when conditions (100) are satisfied, this function may be taken to be a constant K. Indeed, conditions (100) were imposed to make this so. Then the Laplace transform of $W(t)$ is

$$\varphi_+(s) = \frac{K}{\psi_+(s)} \tag{102}$$

which is analytic in $\mathrm{Re}(s) > 0$ by virtue of the requirements on $\psi_+(s)$. Since

$$\lim_{s\to 0} s\varphi_+(s) = \lim_{s\to 0} \int_{-0}^{\infty} e^{-st}\, dW(t) = 1$$

it follows that $\psi_+(0)$ is 0 and K is given by

$$K = \lim_{s\to 0} \frac{\psi_+(s)}{s} = \left[\frac{d\psi_+(s)}{ds}\right]_{s=0} = \psi_+'(0). \tag{103}$$

The constant K is also equal to $W(0)$, the probability that a customer will not have to wait for service. One way to see this is to note that from (102) and (100)

$$\begin{aligned} K = \varphi_+(s)\psi_+(s) &= \lim_{s\to\infty} \varphi_+(s)s = \lim_{s\to\infty} s \int_0^{\infty} e^{-st} W(t)\, dt \\ &= \lim_{s\to\infty} \int_0^{\infty} e^{-u} W\left(\frac{u}{s}\right) du = W(0). \end{aligned} \tag{104}$$

The limit exists since $0 < W(t) \leqq 1$ and $W(t)$ decreases monotonically as t decreases to 0. Although $W(t)$ is discontinuous on the left at $t = 0$ and may have discontinuities for positive values of t (as it does when both $A(t)$ and $B(t)$ have discontinuities), we take it to be continuous on the right at $t = 0$.

Thus when $\psi_+(s)$ is known, $\varphi_+(s)$ is determined and $W(t)$ may be obtained by inversion,

$$W(t) = \frac{K}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{e^{st}\,ds}{\psi_+(s)} \qquad (c > 0). \tag{105}$$

The Laplace-Stieltjes transform of $W(t)$ is

$$w(s) = \int_0^\infty e^{-st}\,dW(t) = s\varphi_+(s) = \frac{sK}{\psi_+(s)}. \tag{106}$$

As $s \to \infty$, $w(s)$ tends to $W(0) = K$. The value of $-dw(s)/ds$ at $s = 0$ is equal to the average waiting time.

In some cases $\psi_-(s)$ is simpler than $\psi_+(s)$ and in place of (102) one may use

$$\varphi_+(s) = \frac{K}{[\beta(s)\alpha(-s) - 1]\psi_-(s)} \tag{107}$$

where differentiation of (99) gives

$$K = \psi_+'(0) = (a^{-1} - b^{-1})\psi_-(0). \tag{108}$$

A.3 *Examples*

The results just obtained will be illustrated by several special cases, all of which have appeared in the literature.

Example a. Poisson Input, Exponential Service. Here $\alpha(s) = a/(a + s)$, $\beta(s) = b/(b + s)$, and $\psi_+(s)$, $\psi_-(s)$ are to be determined from (99), which now takes the form

$$\frac{ba}{(b + s)(a - s)} - 1 = \frac{s(s + b - a)}{(b + s)(a - s)} = \frac{\psi_+(s)}{\psi_-(s)}.$$

Inspection shows that when $D = a$

$$\psi_+(s) = \frac{s(s + b - a)}{b + s}, \qquad \psi_-(s) = a - s \tag{109}$$

satisfy the requirements set forth in connection with (99). Then (103),

(102) and (105) give

$$K = \lim_{s\to 0} \frac{\psi_+(s)}{s} = \frac{(b-a)}{b} = 1 - \rho$$

$$\varphi_+(s) = \frac{K}{\psi_+(s)} = \frac{(1-\rho)(b+s)}{s(b-a+s)} = \frac{1}{s} - \frac{\rho}{b-a+s}$$

$$W(t) = 1 - \rho e^{-(b-a)t}.$$

Incidentally, for this case the probability density of $u_r = s_r - t_r$ is

$$C'(u) = \begin{cases} ab(a+b)^{-1}e^{au} & (u < 0) \\ ab(a+b)^{-1}e^{-bu} & (u \geqq 0) \end{cases}.$$

Example b. Poisson Input, General Service. In this case (99) becomes

$$\frac{\beta(s)a}{a-s} - 1 = \frac{s - a + a\beta(s)}{a-s} = \frac{\psi_+(s)}{\psi_-(s)}.$$

With the help of Rouche's theorem (Titchmarsh,[15] p. 116) and the fact that $[1-\beta(s)]/s$ is the Laplace transform of $1 - B(t)$, it may be shown that $s - a + a\beta(s)$ has no zero in $\mathrm{Re}(s) > 0$ when $b > a$.

Then, with $D = a$,

$$\psi_+(s) = s - a + a\beta(s), \qquad \psi_-(s) = a - s,$$

$$K = (a^{-1} - b^{-1})\psi_-(0) = 1 - \rho = W(0), \tag{110}$$

$$w(s) = s\varphi_+(s) = \frac{sK}{\psi_+(s)} = \frac{s(1-\rho)}{[s - a + a\beta(s)]}.$$

The expression for $w(s)$ is sometimes called the Pollaczek-Khinchin formula. The coefficient of $-s$ in the power series expansion of $w(s)$ is equal to the average waiting time.

Example c. Recurrent Input, Exponential Service. Replacing $\beta(s)$ by $b/(b+s)$ in (99) gives

$$\frac{b\alpha(-s) - b - s}{b+s} = \left[\frac{b\alpha(-s) - b - s}{s(s-s_1)}\right]\left[\frac{s(s-s_1)}{b+s}\right] = \frac{\psi_+(s)}{\psi_-(s)}$$

where s_1 is the one and only zero of $s + b - b\alpha(-s)$ which lies in $\mathrm{Re}(s) < 0$ when $b > a$. The existence of s_1 may be established with the help of Rouche's theorem.[15] If $s = 0$ is the only new zero of $s + b - b\alpha(-s)$ introduced when the region $\mathrm{Re}(s) < 0$ is extended to $\mathrm{Re}(s) < D$, one may take

$$\psi_+(s) = \frac{s(s-s_1)}{b+s}, \qquad \psi_-(s) = \frac{s(s-s_1)}{b\alpha(-s) - b - s}. \tag{111}$$

These functions satisfy the conditions stated just below (99). Then

$$K = \lim_{s\to 0} \frac{\psi_+(s)}{s} = -\frac{s_1}{b} = W(0)$$

$$\varphi_+(s) = \frac{K}{\psi_+(s)} = -\frac{s_1(b+s)}{bs(s-s_1)} = \frac{1}{s} - \frac{1 + s_1 b^{-1}}{s - s_1}$$

$$W(t) = 1 - (1 + s_1 b^{-1}) e^{s_1 t}.$$

Example d. Erlangian Input, Erlangian Service. In this case the input and service time probability densities are

$$A'(t) = \frac{al(alt)^{l-1}}{(l-1)!} e^{-alt}, \qquad B'(t) = \frac{bk(bkt)^{k-1}}{(k-1)!} e^{-bkt}$$

and have the Laplace transforms

$$\alpha(s) = \left(1 + \frac{s}{al}\right)^{-l}, \qquad \beta(s) = \left(1 + \frac{s}{bk}\right)^{-k}.$$

It is found that (99) becomes

$$\frac{1 - f(s)}{f(s)} = \frac{\psi_+(s)}{\psi_-(s)}, \qquad f(s) = \left(1 - \frac{s}{al}\right)^{l} \left(1 + \frac{s}{bk}\right)^{k}.$$

For $a < b$ (the only case considered here), the polynomial $F(s) \equiv 1 - f(s)$ has a zero at the origin, zeros $s_1, \cdots, s_k$ in $\mathrm{Re}(s) < 0$ and zeros $s_{k+1}, \cdots, s_{k+l-1}$ in $\mathrm{Re}(s) > 0$. This may be shown with the help of Rouche's theorem. It turns out that $s_1, \cdots, s_k$ lie inside a circle of radius bk centered on $s = -bk$, and $s_{k+1}, \cdots, s_{k+l-1}$ lie inside a circle of radius al centered on $s = al$. Hence

$$\psi_+(s) = \frac{s(s - s_1) \cdots (s - s_k)}{(s + bk)^k},$$

$$\psi_-(s) = -\frac{(s - al)^l}{(s - s_{k+1}) \cdots (s - s_{k+l-1})} \tag{112}$$

$$W(0) = K = \frac{s_1 \cdots s_k}{(-bk)^k} = \frac{(a^{-1} - b^{-1})(al)^l}{s_{k+1} \cdots s_{k+l-1}}.$$

For the case of regular arrivals and constant service time, inspection shows that $W(t) = 1$. It appears that in this case letting $k = l \to \infty$ should give

$$\psi_+(s) = s, \qquad \psi_-(s) = s[e^{sa^{-1} - sb^{-1}} - 1]^{-1}.$$

Example e. Arrival Rate Almost as Large as Service Rate. It often happens that when the arrival rate a is almost as large as the service rate b, most of the customers have to wait a long time for service. In such cases one may obtain an approximation for the waiting time distribution function $W(t)$.

Let a_1, a_2, b_1, b_2 be the first and second moments, and v_A, v_B be the variances (which are assumed to exist in the following discussion) associated with $A(t)$ and $B(t)$. Then $a_1 = 1/a$ is the average spacing between arrivals and $b_1 = 1/b$ is the average service time. From the integral (105) for $W(t)$ it is seen that the behavior of $W(t)$ for large values of t is determined by the right-most singularities of $K/\psi_+(s)$. In Example a (Poisson input, exponential service), these are poles at the zeros of $\psi_+(s)$ which occur at $s = 0$ and $s = s_1 = a - b$. Note that s_1 is negative and $s_1 \to 0$ as a tends to b. Furthermore, the value of $\psi_-(s)$ near the origin does not change markedly as $a \to b$; and the same is true for the remaining factor $(b + s)^{-1}$ appearing in $\psi_+(s)$.

Many other cases show the same type of behavior as $a \to b$. In general, the function (99)

$$\beta(s)\alpha(-s) - 1 = \frac{\psi_+(s)}{\psi_-(s)}$$

has a double zero at $s = 0$ when $a = b$. When a becomes slightly less than b, and a_1 slightly greater than b_1, one of these zeros remains at $s = 0$ and the other moves to $s \approx s_1$ where

$$s_1 = -\frac{2(a_1 - b_1)}{v_A + v_B} < 0. \tag{113}$$

This may be seen upon using

$$\begin{aligned} \beta(s) &= 1 - b_1 s + \frac{b_2 s^2}{2} + o(s^2) \\ \alpha(-s) &= 1 + a_1 s + \frac{a_2 s^2}{2} + o(s^2) \end{aligned} \tag{114}$$

and assuming that $(a_1 - b_1)^2$ is negligible in comparison with $v_A + v_B$. The approximation (113) for s_1 is given by Smith[14] who points out its importance in the present case.

An examination of the earlier examples leads us to take

$$\psi_+(s) \approx s(s - s_1)C \tag{115}$$

when s is near the origin. Here C is a constant equal to the value of

the remaining portion of $\psi_+(s)$ at $s = 0$. Equations (103) and (105) then give

$$K = \lim_{s \to 0} \frac{\psi_+(s)}{s} = -s_1 C \tag{116}$$

$$\frac{K}{\psi_+(s)} \approx \frac{-s_1}{s(s - s_1)} = \frac{1}{s} - \frac{1}{s - s_1} \tag{117}$$

$$W(t) \approx 1 - e^{s_1 t}.$$

Thus, when $a \to b$ and $v_A + v_B$ does not tend to zero, $W(t)$ is given by the approximation (117) where s_1 is given by (113). The average waiting time is $-1/s_1$. It should be noted that this approximation gives $W(0) \approx 0$ instead of the true (small) value $W(0) = K = -s_1 C$. Unfortunately, there appears to be no simple expression for C corresponding to (113) for s_1.

A.4 *Conclusion*

Finally, it will be mentioned that when a customer departs after being served, the chance p_n that n customers remain in the system is equal to the chance that an arriving customer will find n in the system. Furthermore, for $0 \leqq x < 1$

$$\sum_{n=0}^{\infty} x^n p_n = \frac{(1 - x)}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{\alpha(-s)\beta(s)}{[1 - x\alpha(-s)]} \frac{K\, ds}{\psi_+(s)}$$

where $c > 0$ and is such that the singularities of $\beta(s)/\psi_+(s)$ and $\alpha(-s)/[1 - x\alpha(-s)]$ lie on opposite sides of the path of integration (this restricts $A(t)$ and $B(t)$ somewhat). The right-hand side may be replaced by one of several integrals which differ slightly from the one shown.

REFERENCES

1. Rice, S. O., this issue, p. 269.
2. Borel, E., C. R. Acad. Sci., Paris, **214,** 1942, p. 452.
3. Kendall, D. G., Jour. Roy. Stat. Soc., Series B, **13,** 1951, p. 151.
4. Pollaczek, F., C. R. Acad. Sci., Paris, **234,** 1952, p. 2042.
5. Pollaczek, F., "Problems stochastiques," in *Memorial des Sciences Mathematiques*, Fasc. No. **136,** Gauthier Villars, Paris, 1957.
6. Takács, L., Acta Math. Acad. Sci. Hungar., **6,** 1955, p. 101.
7. Takács, L., Operations Research, **8,** 1960, p. 231.
8. Connolly, B. W., Biometrika, **46,** 1959, p. 246.
9. Connolly, B. W., Jour. Roy. Stat. Soc., Series B, **22,** 1960, p. 89.
10. Gani, J., Jour. Roy. Stat. Soc., Series B, **19,** 1957, p. 181.
11. McMillan, B., and Riordan, J., Annals Math. Stat., **28,** 1957, p. 471.
12. Karlin, S., Miller, R. G., Jr., and Prabhu, N. U., Annals. Math. Stat., **30,** 1959, p. 243.
13. Lindley, D. V., Proc. Cambr. Phil. Soc., **48,** 1952, p. 277.

14. Smith, W. L., Proc. Cambr. Phil. Soc., **49,** 1953, p. 449.
15. Titchmarsh, E. C., *The Theory of Functions*, Oxford University Press, Oxford, 1932.
16. Campbell, G. A., and Foster, R. M., *Fourier Integrals for Practical Applications*, D. Van Nostrand and Co., New York, 1948.
17. Prabhu, N. U., Jour. Roy. Stat. Soc., Series B, **22,** 1960, p. 104.
18. Spitzer, F., Duke Math. Jour., **24,** 1957, p. 327.
19. Watson, G. N., *Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1944.
20. Riordan, J., B.S.T.J., **40,** May, 1961, p. 785.

Delay Distributions for Simple Trunk Groups with Recurrent Input and Exponential Service Times†

By LAJOS TAKÁCS

(Manuscript received June 16, 1961)

At a telephone exchange, calls appear before a simple trunk group of m lines in accordance with a recurrent process. If every line is busy, calls are delayed. The call holding times are mutually independent random variables with common exponential distribution. In this paper, methods are given for the determination of the distribution of the delay for a stationary process and various orders of service. Three orders of service are considered: (1) order of arrival, (2) random order, and (3) inverse order of arrival.

1. INTRODUCTION

In the theory of telephone traffic, the following process is of considerable interest. In the time interval $0 \leqq t < \infty$, calls appear before a simple trunk group with m lines at instants $\tau_1, \tau_2, \cdots, \tau_n, \cdots$ where the interarrival times $\tau_{n+1} - \tau_n (n = 1, 2, \cdots)$ are identically distributed, mutually independent, positive random variables with distribution function

$$\mathbf{P}\{\tau_{n+1} - \tau_n \leqq x\} = F(x) \qquad (n = 1, 2, \cdots). \tag{1}$$

We say that the call input is a recurrent process. If an incoming call finds a free line, a connection is realized instantaneously. If every line is busy, the incoming call is delayed and waits for service as long as necessary (no defections). Denote by χ_n the holding time of the nth call. It is supposed that $\{\chi_n\}$ is a sequence of identically distributed, mutually independent, positive random variables with distribution function

$$\mathbf{P}\{\chi_n \leqq x\} = H(x) \qquad (n = 1, 2, \cdots) \tag{2}$$

† An address presented on September 14, 1961, at the Troisième Congrès International de Télétrafic, Paris.

and independent of $\{\tau_n\}$. We shall consider only those systems of service which satisfy the requirements that there is no free line if there are calls waiting and that the same principle of service applies to every call (no priorities). Such a service system can be characterized by the symbol $[F(x),H(x),m]$ provided that the order of service is specified.

The ideal order of service, "*order of arrival*" or "*first come — first served*," is not always realizable, particularly at times of heavy traffic; therefore it is important to consider other orders of service also. One of these is "*service in random order*" which often describes the practical situation with high accuracy. In this case, waiting calls are chosen for service at random. Every call, independently of the others, and of its past delay, has the same probability of being chosen. Further, it is of great informative value to consider the extreme case, "*inverse order of arrival*," or "*last come — first served.*" (At present we are not concerned with "priority systems" in which "last come — first served" service is the natural order, e.g., the last information to be received may be the most important in the process of the arrival of messages.)

In what follows we shall consider the system $[F(x),H(x),m]$ in the particular case when call holding times have the exponential distribution

$$H(x) = \begin{cases} 1 - e^{-\mu x} & (x \geqq 0) \\ 0 & (x < 0) \end{cases} \tag{3}$$

and the process is stationary. We shall give methods for finding delay distributions for the three service orders mentioned.

We introduce the notation

$$\varphi(s) = \int_0^\infty e^{-sx}\, dF(x) \tag{4}$$

for the Laplace-Stieltjes transform of the distribution function of interarrival times,

$$\alpha = \int_0^\infty x\, dF(x) \tag{5}$$

for the average interarrival time, $W(x)$ for the delay distribution function, i.e., $W(x)$ is the probability that the delay is $\leqq x$, and

$$\Omega(s) = \int_0^\infty e^{-sx}\, dW(x) \tag{6}$$

for the Laplace-Stieltjes transform of $W(x)$.

If $\Re(s) \geqq 0$ then denote by $z = \gamma(s)$ the root with the smallest absolute value of the equation

$$z = \varphi(s + m\mu(1 - z)).$$

We have $|\gamma(s)| \leqq 1$ and for $r = 1, 2, \cdots$

$$[\gamma(s)]^r = r \sum_{n=r}^{\infty} \frac{(m\mu)^{n-r}}{n(n-r)!} \int_0^{\infty} e^{-(s+m\mu)x} x^{n-r} \, dF_n(x) \tag{7}$$

where $F_n(x)$ denotes the nth iterated convolution of $F(x)$ with itself. Let $\omega = \gamma(0)$; then

$$\omega = \sum_{n=1}^{\infty} \frac{(m\mu)^{n-1}}{n!} \int_0^{\infty} e^{-m\mu x} x^{n-1} \, dF_n(x). \tag{8}$$

If $m\alpha\mu \leqq 1$, then $\omega = 1$ while if $m\alpha\mu > 1$, then ω is real and $0 < \omega < 1$.

II. GENERAL THEORY

A. K. Erlang[1] was the first to consider the process $[F(x),H(x),m]$ in the particular case $F(x) = 1 - e^{-\lambda x} (x \geqq 0)$, $H(x) = 1 - e^{-\mu x} (x \geqq 0)$. The case of general $F(x)$ has been treated earlier by D. G. Kendall,[2] F. Pollaczek,[3] and the author.[4]

Denote by $\xi(t)$ the number of calls in the system at the instant t; i.e., $\xi(t)$ is the total number of calls either waiting or being served. Denote by $\chi(t)$ the time difference between t and the arrival of the next call after t. Let $\xi_n = \xi(\tau_n - 0)$, i.e., the nth call finds ξ_n calls in the system, and denote by η_n the delay of the nth call. The initial state is given by $\xi(0)$ and $\chi(0)$.

The vector process $\{\xi(t),\chi(t); 0 \leqq t < \infty\}$ is a Markov process and has the same stochastic behavior for each order of service provided that there is no free line if there are calls waiting. In Ref. 4 it is proved that if $m\alpha\mu > 1$, then there exists a unique stationary process. By choosing the suitable distribution for $\{\xi(0),\chi(0)\}$ we arrive at the stationary process. For the stationary process, $\{\xi(t),\chi(t)\}$ has the same distribution for all t, and the distribution of ξ_n is independent of n. Let $\mathbf{P}\{\xi_n = k\} = P_k (k = 0, 1, \cdots)$. As shown in Ref. 4

$$P_k = \begin{cases} \sum_{r=k}^{m-1} (-1)^{r-k} \binom{r}{k} U_r & (k = 0, 1, \cdots, m-1) \\ A\omega^{k-m} & (k = m, m+1, \cdots) \end{cases} \tag{9}$$

where

$$U_r = AC_r \sum_{j=r+1}^{m} \binom{m}{j} \frac{[m(1-\varphi(j\mu)) - j]}{C_j[1-\varphi(j\mu)][m(1-\omega)-j]} \tag{10}$$

$$A = \left\{ \frac{1}{1-w} + \sum_{j=1}^{m} \binom{m}{j} \frac{[m(1-\varphi(j\mu)) - j]}{C_j[1-\varphi(j\mu)][m(1-\omega)-j]} \right\}^{-1} \tag{11}$$

$$C_j = \prod_{i=1}^{j} \left(\frac{\varphi(i\mu)}{1-\varphi(i\mu)} \right) \tag{12}$$

and ω is defined by (8).

Remark 1 — In the particular case when $\{\tau_n\}$ is a Poisson process of density λ, i.e., $F(x) = 1 - e^{-\lambda x}$ if $x \geqq 0$, we have $\varphi(s) = \lambda/(\lambda + s)$ and thus

$$\omega = \frac{\lambda}{m\mu}$$

and

$$A = \frac{\frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m}{\sum_{j=0}^{m-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j + \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \left(1 - \frac{\lambda}{m\mu}\right)^{-1}}.$$

Also for the stationary process the distribution function of η_n is independent of n but depends on the order of service. We shall use the notation $\mathbf{P}\{\eta_n \leqq x\} = W(x)$ for all cases. It is to be noted that the expectation $\mathbf{E}\{\eta_n\}$ is independent of the order of service if the same principle of service applies to every call. We shall see later that in each case, the mean waiting time is given by

$$\int_0^\infty x \, dW(x) = \frac{A}{m\mu(1-\omega)^2}. \tag{13}$$

III. SERVICE IN ORDER OF ARRIVAL

The following theorem has been proved earlier by D. G. Kendall,[2] F. Pollaczek[3] and the author.[4]

Theorem 1 — The delay distribution function for order of arrival service is given by

$$W(x) = 1 - \frac{A}{1-\omega} e^{-m\mu(1-\omega)x} \qquad (x \geqq 0) \tag{14}$$

where ω is defined by (8) *and A by* (11).

Proof — We have for $x \geqq 0$ that

$$W(x) = \sum_{k=0}^{m-1} P_k + \sum_{k=0}^{\infty} P_{m+k} \int_0^x e^{-m\mu u} \frac{(m\mu u)^k}{k!} m\mu \, du \qquad (15)$$

where $\{P_k\}$ is defined by (9). For, if an arriving call finds a free line, which has probability

$$\sum_{k=0}^{m-1} P_k = 1 - \sum_{k=0}^{\infty} P_{m+k} = 1 - \frac{A}{1-\omega}$$

then its service starts without delay; if it finds every line busy and k ($k = 0, 1, \cdots$) calls waiting, which has probability $P_{m+k} = A\omega^k$, then its service starts at the $(k+1)$st departure after the arrival. Since the departures follow a Poisson process of density $m\mu$, under this condition, the probability that the delay $\leqq x$ is

$$\int_0^x e^{-m\mu u} \frac{(m\mu u)^k}{k!} m\mu \, du.$$

Thus (15) follows from these and (14) agrees with (15).

It follows from (14) that the average delay is

$$\int_0^{\infty} x \, dW(x) = \frac{A}{m\mu(1-\omega)^2} \qquad (16)$$

and the second moment of the delay is

$$\int_0^{\infty} x^2 \, dW(x) = \frac{2A}{(m\mu)^2(1-\omega)^3}. \qquad (17)$$

IV. SERVICE IN RANDOM ORDER

In the particular case of Poisson input this process has been investiated by S. D. Mellor,[5] E. Vaulot,[6] C. Palm,[7,8]† F. Pollaczek,[9] J. Riordan,[10] R. I. Wilkinson,[11] and J. LeRoy.[12] Now we shall consider the case of recurrent input.

Denote by $W_j(x)$ ($j = 0, 1, \cdots$) the probability that the delay of a call is $\leqq x$, given that on its arrival all lines are busy and j other calls are waiting. For the stationary process the probability that an arriving call finds all m lines busy and j calls waiting is $P_{j+m} = A\omega^j$. If there is a free line when a call arrives, which has probability

$$\sum_{j=0}^{m-1} P_j = 1 - \frac{A}{1-\omega},$$

then there is no delay.

† Ref. 8 is an English version of the material in Ref. 7.

Thus we have

$$W(x) = \sum_{j=0}^{m-1} P_j + \sum_{j=0}^{\infty} P_{m+j} W_j(x), \tag{18}$$

that is,

$$W(x) = 1 - \frac{A}{1-\omega} + A \sum_{j=0}^{\infty} W_j(x)\omega^j. \tag{19}$$

Let us introduce the notation

$$\Omega_j(s) = \int_0^{\infty} e^{-sx}\, dW_j(x) \tag{20}$$

and

$$\Phi(s,z) = \sum_{j=0}^{\infty} \Omega_j(s) z^j \tag{21}$$

which is convergent if $\Re(s) \geqq 0$ and $|z| < 1$.

Theorem 2 — The Laplace-Stieltjes transform of the delay distribution function for random service is given by

$$\Omega(s) = 1 - \frac{A}{1-\omega} + A\Phi(s,\omega) \tag{22}$$

where $\omega = \gamma(0)$ *is defined by* (8), A *by* (11), *and*

$$\begin{aligned}\Phi(s,z) = {} & \frac{m\mu}{s + m\mu[1-\gamma(s)]} \exp\left\{\int_{\gamma(s)}^{z} \frac{du}{\varphi(s+m\mu(1-u)) - u}\right\} \\ & + m\mu \int_{\gamma(s)}^{z} \frac{[1-\varphi(s+m\mu(1-u))]}{(1-u)(s+m\mu(1-u))[u-\varphi(s+m\mu(1-u))]} \\ & \cdot \exp\left\{\int_u^z \frac{dv}{\varphi(s+m\mu(1-v)) - v}\right\} du\end{aligned} \tag{23}$$

where $\gamma(s)$ *is the only root in* z *of the equation*

$$z = \varphi(s + m\mu(1-z)) \tag{24}$$

in the unit circle $|z| < 1$. *The explicit form of* $\gamma(s)$ *is given by* (7) *with* $r = 1$.

Proof — For $j = 0, 1, \cdots$ we can write that

$$\begin{aligned}W_j(x) = {} & \sum_{k=0}^{j} \frac{(j+1-k)}{(j+1)} \left[\int_0^x e^{-m\mu u} \frac{(m\mu u)^k}{k!}\, dF(u)\right] * W_{j+1-k}(x) \\ & + \sum_{k=0}^{j} \frac{1}{j+1} \int_0^x e^{-m\mu u} \frac{(m\mu u)^k}{k!} [1 - F(u)] m\mu\, du\end{aligned} \tag{25}$$

where $*$ denotes convolution. If an arriving call finds every line busy and j ($j = 0, 1, \cdots$) calls waiting, then the event that the delay is $\leqq x$ can occur in the following mutually exclusive ways: either in the subsequent interarrival interval k ($k = 0, 1, \cdots, j$) services terminate and the service of the given call does not start during this time interval or, in the subsequent interarrival interval at least $k + 1$ ($k = 0, 1, \cdots, j$) services end and the service of the given call starts at the termination point of the $(k + 1)$st service.

Forming the Laplace-Stieltjes transform of (25) we get

$$(j+1)\Omega_j(s) = \sum_{k=0}^{j} (j+1-k)\Omega_{j+1-k}(s) \int_0^\infty e^{-(s+m\mu)x} \frac{(m\mu x)^k}{k!} dF(x) + \sum_{k=0}^{j} \int_0^\infty e^{-(s+m\mu)x} \frac{(m\mu x)^k}{k!} [1 - F(x)] m\mu \, dx \tag{26}$$

whence

$$z \sum_{j=0}^{\infty} (j+1)\Omega_j(s) z^j = \varphi(s + m\mu(1-z)) \sum_{j=0}^{\infty} j\Omega_j(s) z^j + \frac{m\mu z}{(1-z)} \frac{[1 - \varphi(s + m\mu(1-z))]}{(s + m\mu(1-z))},$$

that is,

$$[z - \varphi(s + m\mu(1-z))] \frac{\partial \Phi(s,z)}{\partial z} + \Phi(s,z) = \frac{m\mu}{(1-z)} \frac{[1 - \varphi(s + m\mu(1-z))]}{(s + m\mu(1-z))}. \tag{27}$$

If $m\mu\alpha > 1$, then $|\varphi(s + m\mu(1-z))| \leqq \varphi(m\mu\epsilon) < 1 - \epsilon$ when $|z| = 1 - \epsilon$ and ϵ is a sufficiently small positive number. Consequently by Rouché's theorem it follows that

$$z = \varphi(s + m\mu(1-z))$$

has one and only one root $z = \gamma(s)$ in the circle $|z| < 1 - \epsilon$, where ϵ is a sufficiently small positive number. The explicit form (7) for $[\gamma(s)]^r$ can be obtained by Lagrange expansion. By definition $\Phi(s,z)$ is a regular function of z if $|z| < 1$ and $\Re(s) \geqq 0$. If we put $z = \gamma(s)$ in (27), then we get

$$\Phi(s,\gamma(s)) = \frac{m\mu}{s + m\mu[1 - \gamma(s)]}. \tag{28}$$

The solution of the differential equation (27) which satisfies (28) can be written in the form (23). Finally, (22) follows from (19).

Remark 2 — Let us introduce the notation

$$A(s,z) = z - \varphi(s + m\mu(1 - z)) \tag{29}$$

and

$$B(s,z) = \frac{m\mu}{(1 - z)} \frac{[1 - \varphi(s + m\mu(1 - z))]}{(s + m\mu(1 - z))}. \tag{30}$$

Then (23) can be written in the following equivalent form

$$\Phi(s,z) = B(s,z) - \int_{\gamma(s)}^{z} \exp\left\{-\int_{u}^{z} \frac{dv}{A(s,v)}\right\} \frac{\partial B(s,u)}{\partial u}\, du. \tag{31}$$

The function $\Phi(s,z)$ can also be expressed in the form of an infinite series as follows

$$\Phi(s,z) = \sum_{j=0}^{\infty} \Phi_j(s)[z - \gamma(s)]^j \tag{32}$$

which is convergent if $| z - \gamma(s) |$ is small enough. If

$$A(s,z) = \sum_{j=1}^{\infty} A_j(s)[z - \gamma(s)]^j \tag{33}$$

[note that $A(s,\gamma(s)) = 0$ by definition of $\gamma(s)$] and

$$B(s,z) = \sum_{j=0}^{\infty} B_j(s)[z - \gamma(s)]^j \tag{34}$$

then $\Phi_j(s)$ $(j = 0, 1, \cdots)$ can be obtained by the following recurrence formula

$$\sum_{k=0}^{j} k\Phi_k(s)A_{j+1-k}(s) + \Phi_j(s) = B_j(s) \qquad (j = 0, 1, \cdots). \tag{35}$$

This follows from (27). In particular by (35) we obtain

$$\Phi_0(s) = B_0(s), \qquad \Phi_1(s) = \frac{B_1(s)}{1 + A_1(s)},$$

$$\Phi_2(s) = \frac{B_2(s)}{1 + 2A_1(s)} - \frac{B_1(s)A_2(s)}{[1 + A_1(s)][1 + 2A_2(s)]}.$$

Formula (32) can conveniently be used to determine the moments of the distribution function $W(x)$. The rth moment

$$\int_0^{\infty} x^r dW(x)$$

can be calculated by the aid of the derivatives

$$\left(\frac{d^i\Phi_j(s)}{ds^i}\right)_{s=0} \qquad (i + j \leqq r).$$

By using the relation $\gamma(s) = \varphi(s + m\mu(1 - \gamma(s))$ we can write that

$$\gamma(s) = \omega + \frac{s\varphi'(m\mu(1-\omega))}{[1 + m\mu\varphi'(m\mu(1-\omega))]} + \frac{s^2\varphi''(m\mu(1-\omega))}{2[1 + m\mu\varphi'(m\mu(1-\omega))]^3} + \cdots. \tag{36}$$

Now in particular we have $\int_0^\infty x\, dW(x) = \frac{A}{m\mu(1-\omega)^2}$ (37)

and $\int_0^\infty x^2\, dW(x) = \frac{2A}{(m\mu)^2(1-\omega)^3}\left[\frac{2}{2 + m\mu\varphi'(m\mu(1-\omega))}\right]$. (38)

V. SERVICE IN INVERSE ORDER OF ARRIVAL

The particular case when $\{\tau_n\}$ is a Poisson process was investigated earlier by E. Vaulot,[13] J. Riordan,[14] and D. M. G. Wishart.[15] The case of recurrent input can be treated in a similar way. As noted by J. Riordan[14] the problem can be reduced to finding the distribution of the length of the busy period for the process of type $[F(x),H(x),1]$ where $F(x)$ is defined by (1),

$$H(x) = \begin{cases} 1 - e^{-m\mu x} & (x \geqq 0) \\ 0 & (x < 0) \end{cases} \tag{39}$$

and there is only one server. In this case denote by $G(x)$ the probability that the length of the busy period is $\leqq x$. The busy period is defined as the time interval during which the server is continuously busy. Evidently $G(x)$ is independent of the order of service, provided that the server is idle if and only if there is no waiting customer in the system.

If $m\mu\alpha > 1$, then there is a unique stationary process, and for the stationary process $W(x)$ is given by

Theorem 4 — The delay distribution function for last-come, first-served service is

$$W(x) = 1 - \frac{A}{1-\omega} + \frac{A}{1-\omega}G(x) \tag{40}$$

where ω is defined by (8), *A by* (11), *and for $x \geqq 0$*

$$G(x) = m\mu \sum_{n=1}^{\infty} e^{-m\mu x}\frac{(m\mu x)^{n-1}}{n!}\int_0^x [1 - F_n(u)]\, du \tag{41}$$

with $F_n(u)$ the nth iterated convolution of $F(u)$ with itself.

Proof — If a call arrives and finds a free line, which has probability $1 - A(1-\omega)^{-1}$, then its service starts without delay; if on its arrival every line is busy, then we can remove all the calls waiting without

affecting the distribution of the delay of the call in question. The service of this call starts when the queue size decreases to m for the first time. The waiting time of this call evidently has the same distribution as the length of the busy period for the queueing process of type $[F(x),H(x),1]$ with $H(x) = 1 - e^{-m\mu x}$ $(x \geqq 0)$. For, in both cases the arrivals have identical stochastic law and the departures follow a Poisson process of density $m\mu$. Thus we get (40). In Ref. 16 it is proved that

$$\int_0^\infty e^{-sx}\, dG(x) = \frac{m\mu[1 - \gamma(s)]}{s + m\mu[1 - \gamma(s)]} \tag{42}$$

where $\gamma(s)$ is the root with smallest absolute value in z of the equation

$$z = \varphi(s + m\mu(1 - z)). \tag{43}$$

$\gamma(s)$ is given by (7) with $r = 1$. By Lagrange expansion we find that

$$\int_0^\infty e^{-sx}\, dG(x) = \frac{m\mu}{s + m\mu} + s \sum_{n=1}^{\infty} \frac{(-1)^n (m\mu)^n}{n!} \cdot \frac{d^{n-1}}{ds^{n-1}} \left(\frac{[\varphi(s + m\mu)]^n}{(s + m\mu)^2} \right) \tag{44}$$

whence (41) follows by inversion.

By using the expansion (36) we get from (42) that

$$\int_0^\infty x\, dW(x) = \frac{A}{m\mu(1 - \omega)^2} \tag{45}$$

and

$$\int_0^\infty x^2\, dW(x) = \frac{2A}{(m\mu)^2(1 - \omega)^3[1 + m\mu\varphi'(m\mu(1 - \omega))]}. \tag{46}$$

REFERENCES

1. Erlang, A. K., Post Office Elect. Engrs. Jour., **10,** 1917–18, p. 189.
2. Kendall, D. G., Annals of Math. Statistics, **24,** 1953, p. 338.
3. Pollaczek, F., C. R. Acad. Sci., Paris, **236,** 1953, p. 578.
4. Takács, L., Acta Math. Acad. Sci. Hungar., **8,** 1957, p. 325.
5. Mellor, S. D., Post Office Elect. Engrs. Jour., **35,** 1942, p. 53.
6. Vaulot, E., C. R. Acad. Sci., Paris, **22,** 1946, p. 268.
7. Palm, C., Tek. Meddelanden Från. Kungl. Telegrafstyrelsen, Stockholm, **110,** 1946, p. 70.
8. Palm, C., Tele. No. 1, 1957, p. 68.
9. Pollaczek, F., C. R. Acad. Sci., Paris, **222,** 1946, p. 353.
10. Riordan, J., B.S.T.J., **32,** 1953, p. 100.
11. Wilkinson, R. I., B.S.T.J., **32,** 1953, p. 360.
12. LeRoy, J., Ann. des Télécom., **12,** No. 1–2, 1957, p. 2.
13. Vaulot, E., C. R. Acad. Sci., Paris, **238,** 1954, p. 1188.
14. Riordan, J., B.S.T.J., **40,** 1961, p. 785.
15. Wishart, D. M. G., Operations Research, **8,** 1960, p. 591.
16. Takács, L., Operations Research, **8,** 1960, p. 231.

The Transistorized A5 Channel Bank for Broadband Systems

By F. H. BLECHER and F. J. HALLENBECK

(Manuscript received April 27, 1961)

This article presents a brief historical background to the A-series of terminal units, used extensively in long-haul and short-haul transmission facilities to provide the first step of modulation from voice to carrier and the final step of demodulation from carrier to voice. Most of the paper is devoted to a description of the latest unit of this series — the A5 channel bank, which through the use of transistors and other modern components achieves significant improvements in size, power requirements and operating characteristics.

I. INTRODUCTION

All of the long-haul, broadband transmission facilities of the Bell System, and many of the short-haul microwave radio systems, employ a common unit in their terminal multiplexes. This is the A type channel bank which provides the first step of modulation from voice to carrier spectrum and the reverse function of demodulation. From its inception over twenty-five years ago, the channel bank has undergone a number of size reductions primarily utilizing new types of crystal filters. This paper describes a radically new version, the A5 Channel Bank, with a detailed discussion of the new circuit feature — the transistorized voice-frequency amplifier.

II. HISTORICAL BACKGROUND

2.1 *General*

Over many years, the threads of the "channel bank" story have been woven into papers of much broader scope. It is the aim of this section to combine these pieces into one background picture.

In the middle 1930's, Bell System research and development effort in long-distance communication was focusing on the expansion of frequency-

division, single-sideband carrier systems.[1] The invention of the negative-feedback amplifier by H. S. Black had made feasible the development of transmission media capable of carrying multi-channel, high-quality systems across the country.[2] Other important advances in both the electronic art and the network art opened the way to the realization of practical multiplexes to translate many voice or other information channels to spectra suitable for line transmission.

At this time, many important decisions were made on multiplexing methods. Since that time, they have set the broad pattern for the terminals of long-haul carrier systems, both for wire and microwave radio. Inclusion in the recommendations of such bodies as the CCITT* made the general pattern worldwide. The decision of greatest interest to this paper involves the first step of modulation from voice frequency to a carrier spectrum and its counterpart of demodulation at the receiving end.

Actually, not one, but many considerations were involved. The fundamental ones concerned carrier spacing, the number of voice channels to be handled in the modulation process, and the frequency spectrum to be used. The answers did not evolve without much thought directed toward the future long-distance plant and the carrier systems which would make it a reality. The development trend was toward the utilization of much higher frequencies for line transmission than had previously been attempted.[3] Three high-frequency broadband systems were being conceived for application to:

1. Available 19-gauge toll cable, both underground and aerial (Type K).[4,5]
2. Open-wire lines then carrying the three-channel Type C system. The new system (Type J) would lie above Type C in frequency and require extensive retranspositions of the line.[6]
3. A new medium — coaxial cable — with much greater potential for high-frequency transmission. This would be the transmission medium for the L1 and L3 systems of the future.[7,8,9,10]

As the study progressed, it became more and more obvious that, if possible, the same terminal arrangements should serve all of these systems.

2.2 *Carrier Spacing*

Of primary importance in the pattern for the new systems was the question of carrier spacing. Assuming that the new broadband transmis-

* Comité Consultatif International Téléphonique et Télégraphique.

sion media did not have sharp cut-offs, the penalty for wide carrier spacing was more repeaters per link. Of course, the open-wire and cable systems did have practical limitations such as crosstalk, which presented economic but not insurmountable restrictions. Contrariwise, the new coaxial line seemed restricted only by attenuation due to thermal noise and by repeater spacing. Under these circumstances, the cost penalty per voice circuit due to wide spacing would not be controlling. For short systems where terminal costs predominate, the cost was insignificant; even for very long systems, it was quite small.[11]

The standardization of 4000-cycle spacing for broadband systems was based primarily on other considerations. The Bell System was striving to improve the over-all quality of the service it was offering. Subscriber sets, telephone instruments, and end links were all being developed to provide better quality. It seemed logical to make the backbone carrier circuits good enough not to be limiting in the over-all effective net loss. The 4000-cycle spacing made this possible. With advances in the filter art, the width of the frequency band could be extended at both low and high ends. The former helped to maintain the naturalness of the voice; the latter helped to improve articulation. At this time it was also decided to make the carrier intervals exactly uniform — in other words, to relate all carriers harmonically to 4 kc. This was not a matter of quality of voice transmission, but one of terminal equipment.

A corollary of high-capacity broadband systems is heavy concentrations of equipment at major terminal offices. Under these circumstances, supplying of carrier power poses an economic problem. Needed are many frequencies of high precision both initially and with time and changing environment. Economic feasibility of producing many harmonically related frequencies of approximately equal amplitudes had been demonstrated.[12] For the new broadband systems the circuit took the form of a highly stable 4-kc source driving a pulse producer. Odd harmonics were directly generated; even harmonics were provided by rectification. Sufficient power was derived to feed many systems from common bus-bar arrangements.

2.3 *The Modulation Process*

Of prime importance in devising a multiplexing plan is the decision as to single- or multi-stage modulation from voice to line frequency. Systems prior to the broadband development employed a single stage. For the new systems, several factors influenced the choice of multi-stage modulation.

It was known early that the line frequencies of the three systems — cable, open-wire and coaxial — would be quite different. Economic restrictions rather than technical feasibility imposed by crosstalk and line equalization set a practical limit for the long-distance, cable-pair system between 10 and 60 kc. Since the new open-wire system was to be placed above the standard Type "C", the lower limit was about 35 kc; the upper limit of about 150 kc was dictated by the costs of line transpositions. The coaxial cable system seemed to have a reasonable lower limit of about 60 kc, but an upper limit in the order of one megacycle. The latter was soon to become three megacycles, and finally, for L3, about eight megacycles. For the open-wire and cable-pair systems, single-stage modulation was possible but the terminals, except for an overlap region between 30 and 60 kc, required different channel filters and carrier supplies. Single-stage modulation was out of the question for the high channel capacity coaxial system. It would mean hundreds of different, closely spaced channel filters not yet feasible even with the rapidly advancing network art. Also, for every channel a different carrier frequency would be needed with resultant terminal complications.

The answer to these problems was the provision of a group of channels common to all three systems. With 4-kc spacing, both the cable-pair and open-wire systems could economically accommodate 12 channels. Of course, the coaxial system did not have this restriction, but 12 channels seemed a good common denominator.

This resolved the technical differences; in addition, the choice of a common group would mean (1) flexibility of interconnection of systems, (2) large-scale production of one major equipment unit, and (3) minimum development effort in system areas and in the supporting network and component areas as well.

2.4 *The Group Spectrum*

The foregoing discussion on the size of the group might seem to imply a completely free choice. Perhaps, at the present time, it could be made independently of the spectrum to be employed. This was not true in the 1930's. The limitations of available inductor-capacitor filters in the low-frequency ranges made a 12-channel group impracticable, even starting at a low frequency such as 12 kc. The first experimental cable system had demonstrated the difficulties above about 40 kc. It had provided nine channels with this top frequency.

Of course, as was done later in certain European and U. S. systems, another stage of modulation could have been introduced and the subgroup technique used, i.e., three or four channels for instance, operating

in the most efficient low-frequency range translated to a 12-channel group.

However, the expanding art had provided another answer. Intensive research and development in the Laboratories by Mason, Sykes, and Lane had culminated in successful application of the piezoelectric effects in quartz crystals to wave filters.[13,14,15] These were the same effects which had been studied by Langevin and applied to submarine detection in World War I and by Cady, Pierce, and others.

The decision was made to employ such crystal filters for the selection of the desired single sideband. They offered many advantages. The transmission bands could be positioned in the frequency spectrum with high precision. Steeply rising attenuation characteristics, low distortion and low flat loss could now be achieved rather easily.

The actual choice of 60 to 108 kc as the standard group frequency range of the channel bank was mainly based on economics. Lower frequency crystals were physically possible. They were large and expensive and, if high production was envisioned, might be difficult to obtain from the raw Brazilian quartz available. Crystals at frequencies higher than 108 kc were small and easily available, but their fabrication and frequency adjustment posed difficult problems in the 1930's. Thus, the final frequency choice of 60 to 108 kc as the group spectrum to provide 12 channels was an economic and technical compromise.

2.5 *Microwave Radio Usage*

As described, the channel bank was developed in the 1930's, long before the potentialities of the microwave radio band were realized. However, when radio systems at frequencies in the 4-kmc range (TD-2 Radio) were being developed, their high channel capacities could utilize the available coaxial system terminals based on the channel banks.[16,17] Interconnection with wire systems was thus made easy, and the high production needs of the new radio systems further proved the value of a standard and basic group of channels. Later microwave radio systems, TH and TJ, also employ the standard "A" type bank.[18,19]

III. OBJECTIVES OF A5 DESIGN

From these beginnings, the channel bank has progressed through several previous redesigns. Aimed specifically at size reduction, they were based mainly on advances in the crystal filter art. These had led to smaller filters as well as to much improved crystal arrangements.[20]

The A5 design had broader objectives. Modernization and miniaturiza-

tion were important, but equally so were transmission improvement and maintenance simplification. The introduction of new services such as Direct Distance Dialing and data transmission has imposed new requirements on the long-distance plant. Specifically, the circuits must maintain better net loss stability. This must be true in the face of office voltage variations and changes in the performance characteristics of the active devices employed.

With the ever-expanding, long-haul toll plant, the maintaining of performance within close limits becomes a problem of large proportions. Improved circuit designs and devices can in the long run avail little if proper maintenance is made difficult. The A5 bank design aimed at facilitating this important feature of Bell System service. Accessibility of adjustment and ease of interchanging parts were important factors in the design. Circuit compatibility with existing long-haul terminal gear was another important requirement.

IV. GENERAL CIRCUIT FEATURES

For the channel banks, the general circuit arrangements have remained unchanged through the several equipment redesigns, including the A5. The channel bank equipment does not include circuits for signaling or for the conversion from two-wire to four-wire operation. This permits greater flexibility in circuit arrangements. As shown in Fig. 1, the transmitting circuit starts at the voice-frequency input transformer, whose leakage reactance also offers high impedance to carrier frequencies. A simple shunt varistor bridge employs copper oxide as the modulating element. The modulator operates at a nominal carrier power level of 0 dbm. The high-pass filter following the modulator provides high impedance to voice frequencies, thus increasing the modulator efficiency. The desired lower sideband is then selected by a crystal band-pass filter operating in parallel with eleven others. A compensating network in parallel with the common output improves the transmission characteristic of the channel filters. The hybrid output coil provides an alternate output to facilitate switching a working bank to an alternate group facility without interruption. It also provides a means of inserting a program terminal occupying the frequency space of two or three regular voice channels.

The receiving circuit operates in much the same fashion as the transmitting circuit except, of course, in reverse. The high-frequency line terminates in a hybrid transformer, which provides for connection to a program receiving terminal. The twelve channel filters are paralleled at this hybrid. The individual sidebands are then demodulated in a bridge

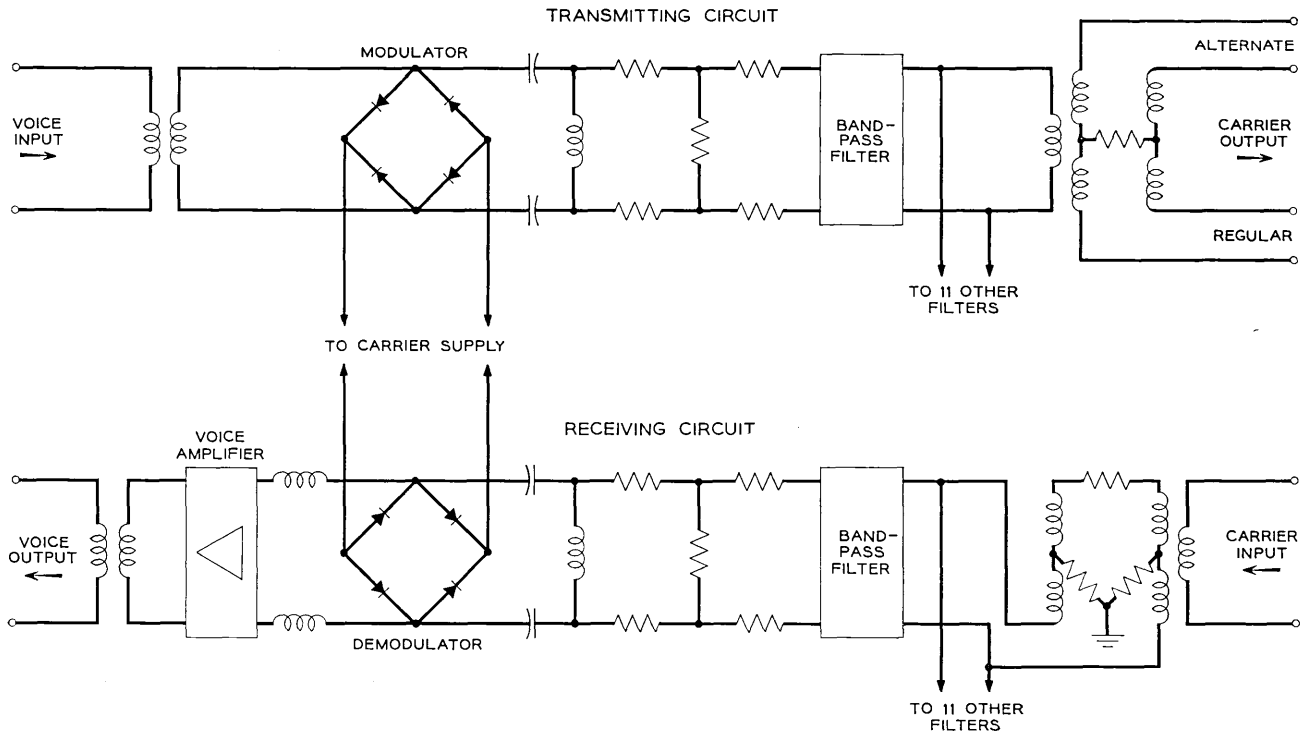


Fig. 1 — Basic circuit of the A5 channel bank.

similar to the modulator. The demodulator, however, is poled oppositely on the carrier supply to neutralize the dc components of the modulation process in the two units. High-pass and low-pass filters terminate the demodulator to improve its efficiency.

Following the low-pass filter is a voice amplifier with sufficient gain to establish the correct received voice level. The potentiometer for gain control is remotely mounted at the four-wire patching bay in the test area of the toll office.

From their inception, as described previously, the Bell System channel banks have been based on a single modulation process from voice to carrier output and upon the crystal filter as the sideband selecting medium. Other methods are feasible and have been used in this country as well as in Europe. Pregrouping both at low and high frequencies has been used, with the standard 60- to 108-kc group spectrum being obtained by a second stage of modulation. With the availability of high efficiency ferrite inductors, it is also possible to obtain good LC filter performance, usually with the addition of some band equalization. Economic studies indicated that, with Western Electric's highly developed capabilities in crystal filter production, the single modulation step using such filters was still the least expensive procedure.

Design choices had to be made regarding two other major circuit units — the modulator-demodulators and the only active unit for each channel, the voice amplifier at the output of the demodulator. In previous banks the modulating unit, a quad of copper-oxide varistors, had proved generally satisfactory. During the A5 development, newer devices such as silicon diodes were studied. However, problems of incompatibility with the carrier power available at existing offices and other considerations made it advisable to employ the standard copper-oxide varistors.

In previous banks, the demodulator amplifier employed a single electron tube and provided approximately 35 db of gain in each channel over the voice range. Only modest feedback was used with a maximum value of 7 db at minimum gain. This was reduced to zero at maximum gain. With the small amount of feedback, the tube amplifier was sensitive to tube aging, battery variations, and component changes. The resultant variations in net loss were tolerable for the voice plant with switching under operator control. The introduction of Direct Distance Dialing and the need to transmit various data signals impose more severe requirements. It is difficult to meet these consistently without frequent and very careful maintenance.

An important part of the A5 development was the improvement of this demodulator amplifier. The main focus, of course, was on the

provision of sufficient feedback to overcome the instabilities of the existing design. Two design approaches were available, a multi-tube or a transistor amplifier. With well-established techniques, the former offered no particular design problems; however, it did mean heavier power consumption, more heat to be dissipated, and a definite restriction on the degree of miniaturization possible.

On the other hand, a transistor amplifier offered a reduction in power, a cooler operating environment, and the possibility of a high degree of size reduction. In addition, the promise of very long life could mean a large reduction in maintenance cost. On the basis of these considerations, the decision was made to proceed with a transistor channel bank.

V. TRANSISTOR AMPLIFIER

The design specifications for the amplifier were:

1. Voltage gain between 600-ohm terminations should be adjustable between 30 to 40 db.
2. The voltage gain should have a 5-db rise at 200 cycles in order to compensate for the low-frequency rolloff of the crystal band-pass filters in the modulator and demodulator.
3. A particular gain adjustment should not change by more than ± 0.1 db over a period of years and for ± 10 per cent variations in the -24-volt central office battery.
4. Amplifier should provide 80 milliwatts of output power into a 600-ohm load with less than 1 per cent second and third harmonic distortion.
5. Output noise power should be less than 1.4×10^{-7} milliwatts (6.5 dba at zero-level point).
6. Return loss at input and output circuits should be greater than 40 db over the frequency band of 200 to 4000 cycles.

In addition to the above requirements, it is desirable that the amplifier be as simple as possible and employ the minimum number of components.

5.1 *Basic Design Considerations*

In order to satisfy the requirement that the gain of the amplifier be constant over a relatively long period of time, it is evident that a feedback amplifier is required. A hybrid feedback or a series feedback connection at the output of the amplifier appears to be most attractive. Hybrid feedback has the advantage over other types of feedback in permitting a 50 per cent reduction in dc power dissipated in the output stage. Series feedback, on the other hand, has the advantage that for a given

amount of negative feedback, it yields a somewhat better return loss than hybrid feedback. In addition, the series feedback circuit is simpler than the hybrid circuit and permits the use of a less expensive output transformer. For these reasons, it was decided to use a series feedback connection at the output of the amplifier rather than hybrid feedback. Similar considerations of return loss and circuit simplicity applied to the input of the amplifier indicate that series feedback is again the optimum connection.

It is demonstrated in Appendix A that the voltage gain of the series feedback amplifier shown in Fig. 2 is given by the expression

$$G_V = \frac{R_L}{Z_{12F}} \cdot \frac{A\beta}{1 - A\beta} \tag{1}$$

where R_L is the load resistance, Z_{12F} is the open-circuit transfer impedance of the feedback network ($R1$, $R2$ and $Z3$) and $A\beta$ is the loop current transmission. (In the case of transistor feedback amplifiers, it is convenient to define feedback as a loop current transmission.[21]) It is evident from (1) that if the magnitude of $A\beta$ is much greater than one, the voltage gain is determined by the load resistance and the feedback network. In order to achieve the desired voltage gain and distortion performance, R_L is of the order of several hundred ohms, and the magnitude of the transfer impedance $| Z_{12F} |$ is of the order of several ohms. Therefore, $R1$, $R2$ and $Z3$ are relatively small impedances, and $R2$ (a potentiometer) can be mounted at a considerable distance from the amplifier without introducing excessive electrostatic pickup.

It is shown in Appendix A that the input and output impedances of the series feedback amplifier are equal to

$$Z_{IN} = (Z_{IN}' + Z_{11F})\,[1 - A\beta(R_G = 0)] \tag{2}$$

$$Z_{OUT} = (Z_{OUT}' + Z_{22F})\,[1 - A\beta(R_L = 0)] \tag{3}$$

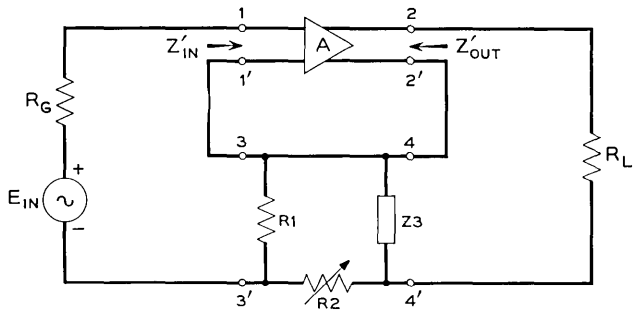


Fig. 2 — Equivalent circuit of a series feedback amplifier.

where Z_{IN}' and Z_{OUT}' are the input and output impedances respectively of the amplifier A (refer to Fig. 2) with the feedback loop opened, Z_{11F} and Z_{22F} are the open-circuit input and output impedances respectively of the feedback network, and $A\beta(R_G = 0)$ and $A\beta(R_L = 0)$ are the values of loop current transmission with R_G and R_L respectively set equal to zero. Since the magnitude of the loop current transmission is much greater than one and Z_{IN}' and Z_{OUT}' are of the order of 1000 ohms, a good return loss can be obtained by shunting the input and output circuits of the feedback amplifier with 600-ohm terminating resistors.

In order to satisfy electrical requirements 3, 4 and 6 in the above list, a minimum of 30 db of negative feedback is required. This amount of feedback and 40 db of external voltage gain can be obtained with three common-emitter connected junction transistors in the amplifier A. If two transistors are used in the amplifier, one of the transistors would have to be connected in the common-collector configuration and the other in the common-emitter configuration in order to provide the phase reversal necessary for negative feedback. Under these conditions the design would be marginal, and severe electrical requirements would have to be placed on the alphas of the transistors (alphas in excess of 0.995 would be required).

Fig. 3 shows a simplified circuit diagram of the amplifier. The dc biasing and the feedback shaping networks have been omitted. Strictly speaking, this structure is not a "pure" series feedback amplifier because the emitter of the transistor in the second stage and the interstage networks are returned to the common connection between the input and output transformers instead of to the emitters of the input and output

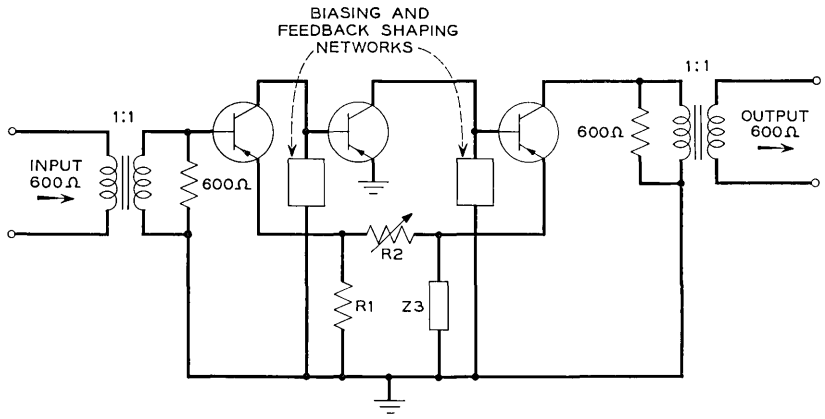


Fig. 3 — Simplified circuit of emitter feedback amplifier.

transistors. This type of feedback structure is defined as an emitter feedback amplifier and is analogous to the cathode feedback circuit for electron tubes.[22] In the case of the series feedback circuit, the network $R1$, $R2$, and $Z3$ provides feedback only around the main loop. In the case of emitter feedback, the network provides local feedback for the first and third stages of the amplifier in addition to feedback around the main loop.

The principal advantage of emitter feedback over series feedback, in a transistor amplifier, is that emitter feedback often makes possible a simpler biasing circuit. Emitter feedback, however, has two disadvantages. First, since it introduces local feedback into the first and third stages, it increases the alpha requirements on the transistors in order to obtain the necessary amount of loop feedback. This disadvantage is partially compensated for by the fact that the local feedback improves the return loss at the input of the amplifier and the distortion performance of the output stage. Secondly, emitter feedback stabilizes the emitter current of the output stage instead of the collector current (refer to Appendix A). As a result, the expression for voltage gain (1) must be multiplied by the alpha of the output transistor. This is not a serious limitation in the case of the demodulator amplifier since the gain of the amplifier is initially set by the use of a potentiometer. In addition, it is expected that alpha will not vary by more than ± 2 per cent over the life of a transistor.

5.2 *Magnitude of Feedback in the Useful Operating Band*

In order to calculate the feedback, it is convenient to redraw the circuit as a series feedback amplifier, as shown in Fig. 4. The local feedback

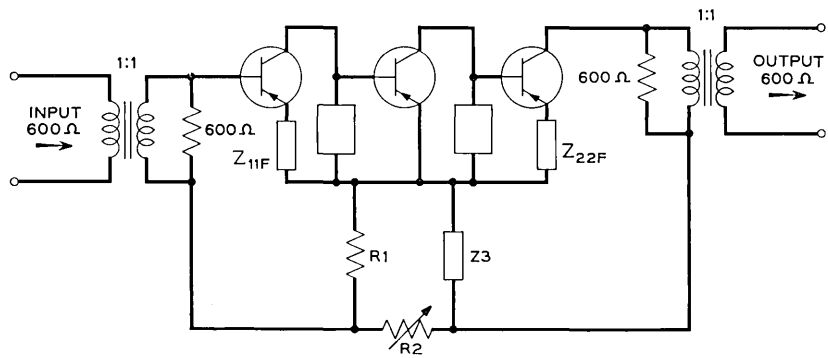


Fig. 4 — Series feedback amplifier equivalent to emitter feedback amplifier.

introduced by the feedback network is taken into account (to a good approximation) by the impedances Z_{11F} and Z_{22F} in the emitters of the first and third transistors respectively. It is evident from (2) and (3) that if the magnitude of $A\beta$ is sufficiently large, the input and output impedances of the amplifier will be equal to 600 ohms, and the load resistance, R_L, is equal to 300 ohms. Since the voltage gain of the amplifier must be adjustable from 30 to 40 db, and allowing for a 2-db loss in the input and output transformers, the transfer resistance of the feedback network must be adjustable between 7 and 2.2 ohms as determined by (1) for $R_L = 300$ ohms. The open-circuit transfer impedance of the feedback network is equal to

$$Z_{12F} = \frac{R1 \cdot Z3}{R1 + R2 + Z3}. \tag{4}$$

The resistance $R2$ is used to control the flat gain of the amplifier while the impedance $Z3$ controls the shape of the gain-frequency characteristic which is required to have a 5-db rise at 200 cycles. With reference to (1) and (4) it is evident that if $|Z3|$ is much less than $(R1 + R2)$, then the voltage gain of the amplifier is proportional to $1/|Z3|$. An additional requirement on $R1$ and $|Z3|$ is that they be as small as possible in order to minimize the amount of loop feedback lost through local feedback. $R1$ was chosen as 68.1 ohms while the flat-gain value of $|Z3|$ was chosen as 10 ohms. From (4), the maximum value of $R2$ (corresponding to maximum gain) is equal to 270 ohms.

In Appendix A it is shown that the loop current transmission of the amplifier shown in Fig. 4 is equal to

$$A\beta = \frac{-G_i Z_{\text{OUT}}' Z_{12F}}{(Z_{\text{IN}}' + Z_{11F} + R_G)(Z_{\text{OUT}}' + Z_{22F} + R_L) - Z_{12F}^2} \tag{5}$$

where G_i is the short-circuit current gain of the amplifier A. In the frequency range of 200 to 4000 cycles per second,

$$G_i = \frac{a_{01}}{1 - a_{01}} \cdot \frac{a_{02}}{1 - a_{02}} \cdot \frac{1}{1 - a_{03}} \tag{6}*$$

$$Z_{\text{IN}}' = r_{b1}' + \frac{r_{e1} + Z_{11F}}{1 - a_{01}} \tag{7}$$

$$Z_{\text{OUT}}' = r_{c3}(1 - a_{03}) \tag{8}$$

* With reference to Fig. 3, the feedback network is driven by the emitter current of the output stage, and as a result this transistor acts as a common collector stage as far as the feedback is concerned. Therefore, the current gain of the third stage is equal to $1/1 - a_{03}$ instead of $a_{03}/1 - a_{03}$ as determined from the equivalent series feedback amplifier shown in Fig. 4.

where a_{01}, a_{02}, and a_{03} are the low-frequency common-base current gains (alphas) of the first, second, and third transistor stages respectively, r_{e1} and r_{b1}' are the emitter and base resistances respectively of the transistor in the first stage and r_{c3} is the collector resistance of the output stage. Substituting (6) to (8) into (5) yields the expression for mid-band feedback:

$$A\beta_0 = \frac{-\dfrac{a_{01}a_{02}Z_{12F}}{1 - a_{02}}}{[(r_{b1}' + Z_{11F} + R_G)(1 - a_{01}) + r_{e1} + Z_{11F}]} \cdot \left[1 - a_{03} + \frac{R_L + Z_{22F}}{r_{c3}}\right] - \frac{Z_{12F}^{\,2}(1 - a_{01})}{r_{c3}} \quad (9)$$

It will be evident that for all practical designs, the second term in the denominator of (9) can be neglected. As previously discussed, a minimum of 30 db of negative feedback is required in order to satisfy the electrical requirements. The loop feedback is a minimum when the amplifier is adjusted for its maximum gain ($| Z_{12F} | = 2.2$ ohms) and the transistors used have the minimum allowed values of alpha.

$$a_{01} = a_{02} = a_{03} = 0.975$$

$$r_{b1}' = 100 \text{ ohms}^*$$

$$r_{e1} = 13 \text{ ohms}$$

$$r_{c3} = 50{,}000 \text{ ohms.}$$

If the above transistor parameters are substituted into (9), the resulting magnitude of $A\beta_0$ is 30 db. Consequently, the minimum magnitude of feedback is about 30 db over the frequency range of 400 to 4000 cycles per second. Due to the increase in voltage gain at frequencies between 200 and 400 cycles per second, the minimum magnitude of feedback in this frequency range can be less than 30 db.

5.3 *DC Biasing and Low-Frequency Shaping of the Negative Feedback Amplifier*

Fig. 5 shows a complete circuit diagram of the transistor amplifier. The first stage is biased at two milliamperes of collector current. This value of current is a compromise between low noise figure (requiring a low collector current) and a large magnitude of feedback [requiring a large

* The values stated for the transistor parameters r_{b1}', r_{e1} and r_{c3} are average values for the particular transistors used in the amplifier. Refer to Section 5.3 below for a discussion of the transistors used and the dc operating points.

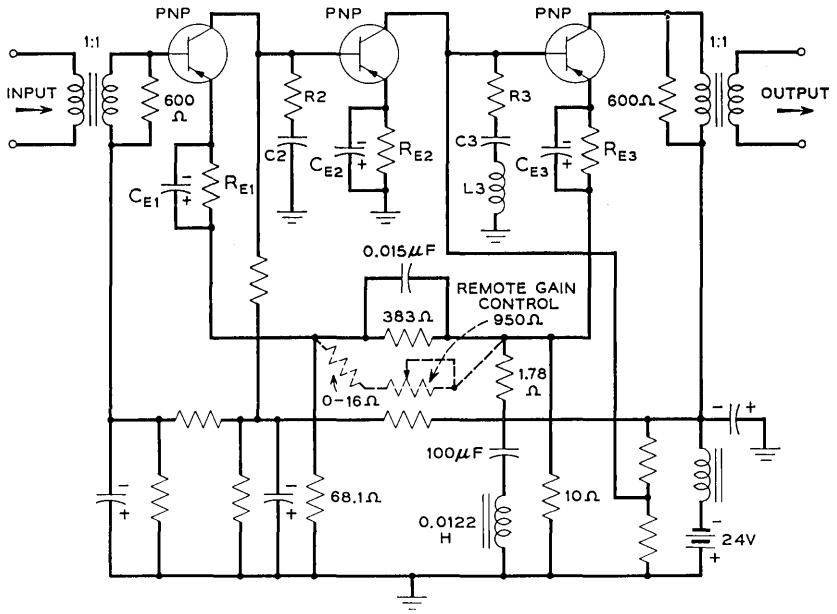


Fig. 5 — Complete circuit diagram of the transistor amplifier.

collector current for low emitter resistance — refer to (9)]. The second transistor stage is biased at 4.7 milliamperes of collector current while the third stage is biased at 45 milliamperes of collector current in order to satisfy the output power requirement of 80 milliwatts into a 600-ohm load. The first and second stages are biased at 2.5 volts collector-to-emitter voltage, while the third stage is biased at 14 volts collector-to-emitter voltage. The collector currents are stabilized by use of dc feedback provided by resistors in the emitter circuits of the transistors.

Type 12B alloy germanium PNP transistors are used in the first and second stages of the amplifier, and a type 9D alloy germanium PNP transistor is used in the output stage. The 12B is a low-power device limited to about 100 milliwatts of dc power dissipation at a 60°C ambient, while the 9D is a medium-power transistor capable of handling 1.5 watts of dc power with an adequate heat sink at a 60°C ambient. Since the 12B transistors in the first and second stages are biased at 5 milliwatts and 12 milliwatts of dc power respectively, and the 9D is biased at 630 milliwatts of dc power, good transistor reliability should be realized.

One of the most important considerations in the design of a feedback amplifier is shaping the negative feedback at high and low frequencies

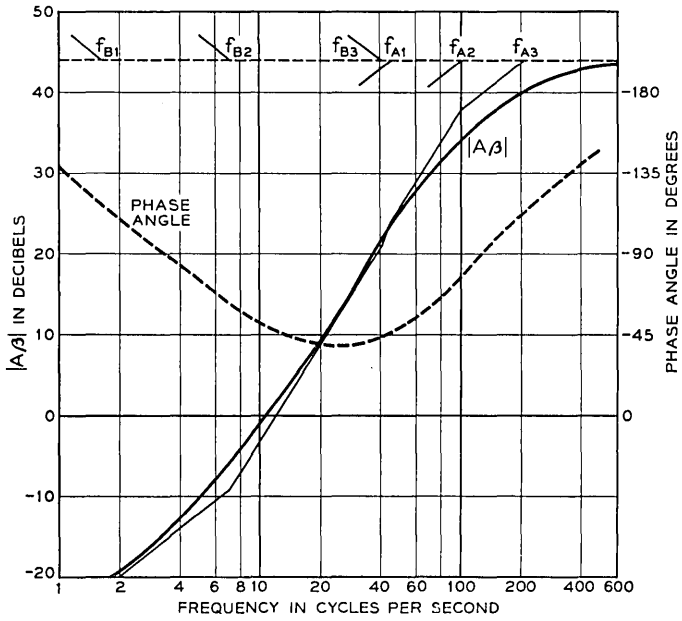


Fig. 6 — Plot of low-frequency loop current transmission with transistors as characterized in Table I.

in order to insure stability for all transistors that will be used in the amplifier. A detailed analysis of the low-frequency shaping of the negative feedback is presented in Appendix B. Fig. 6* shows a plot of the low-frequency loop current transmission calculated for transistors with the electrical parameters listed in Table I and for maximum negative feedback ($| Z_{12F} | = 7$ ohms). At the frequency 25 cycles, the phase of the feedback makes its closest approach to the critical phase angle of 0 degrees. It is evident that the amplifier has a low-frequency phase margin of 40 degrees. If transistors with alphas of 0.99 are used in the amplifier, the phase margin is reduced to 25 degrees. Since the phase of the feedback does not reach 0 degrees, it is not possible to define a gain margin.

5.4 *High Frequency Shaping of the Negative Feedback*

A detailed analysis of the high-frequency feedback shaping is presented in Appendix C. The results of that analysis will be used to show

* In this figure it is assumed that $Z3$ is equal to its flat gain value of 10 ohms. The variation of $Z3$ with frequency has negligible effect on the low-frequency stability.

TABLE I

$a_{01} = a_{02} = a_{03} = 0.975$
$f_{a1} = f_{a2} = f_{a3} = 3$ megacycles
$c_{c1} = c_{c2} = c_{c3} = 35\ \mu\mu f$
$r_{b1} = r_{b2} = r_{b3} = 100$ ohms
$r_{e1} = 13$ ohms
$r_{e2} = 5.5$ ohms
$r_{e3} = 0.6$ ohms
$r_{c3} = 50{,}000$ ohms
$m_1 = m_2 = m_3 = 0.2$ radian

that the amplifier is stable at high frequencies for all transistors for which

$$C_c \leqq 35\ \mu\mu f\ (|V_{cE}| = 2.5 \text{ volts}) \tag{10}$$

$$f_a \geqq 3 \text{ megacycles.} \tag{11}$$

Fig. 7 shows a plot of the high-frequency loop current transmission calculated for transistors with electrical parameters listed in Table I and for maximum feedback. At the frequency 133 kc, the magnitude of the

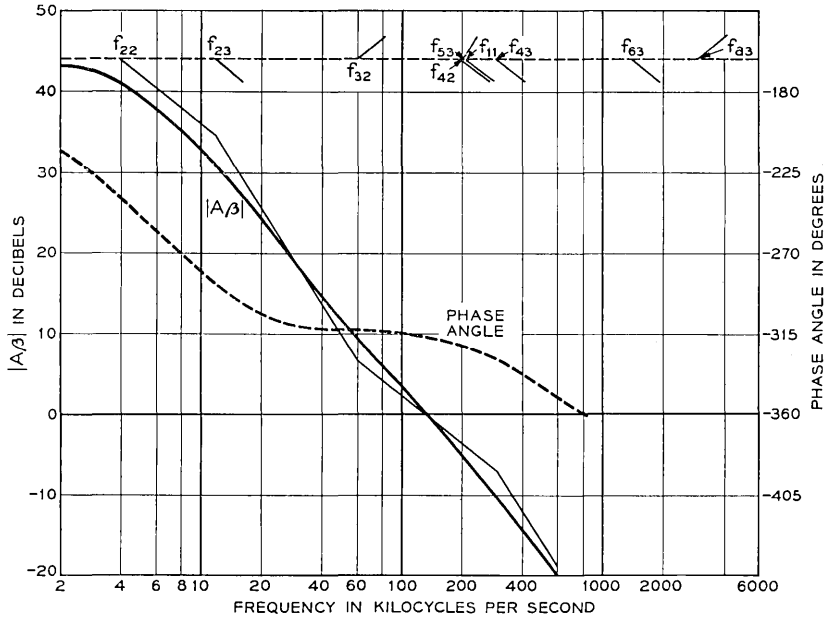


Fig. 7 — Plot of high-frequency loop current transmission with transistors as characterized in Table I.

feedback is 0 db while the phase of the feedback is −317 degrees. At the frequency 800 kc, the phase of the feedback is −360 degrees while the magnitude is −24 db. The phase and gain margins against instability are 43 degrees and 24 db respectively.

Except for the first two cutoff frequencies f_{22} and f_{23}, all of the other cutoff frequencies are essentially independent of the low-frequency common-emitter current gain $(a_0/(1 - a_0))$ and are determined by the transistor parameters f_a, r_b', and C_c and the circuit elements in the two interstage shaping networks, $R2$, $C2$, and $R3$, $C3$, $L3$. The cutoff frequency f_{22} is equal to the frequency at which the reactance of the capacitor $C2$ is equal to the input resistance of the second transistor stage, while the cutoff frequency f_{23} is equal to the frequency at which the reactance of $C3$ is equal to the input resistance of the third stage. To a good approximation, the input impedance of a common emitter stage is directly proportional to the current gain of the stage.* Since the current gain of a common-emitter stage may vary from 39 to 200 (corresponding to alpha variations of 0.975 to 0.995), the cutoff frequencies f_{22} and f_{23} may be as small as one-fifth the values shown in Fig. 7. Fortunately, this variation in cutoff frequencies f_{22} and f_{23} is almost exactly compensated for by the variation in current gain of the second and third stages respectively, and the asymptotic loop current gain is independent of the common emitter current gain at frequencies above f_{23}. The first stage of the amplifier acts as a common-base stage as far as the feedback is concerned. With reference to (9), it is evident that the magnitude of the feedback is essentially independent of the factor $(1 - a_{01})$ if $|(r_{e1} + Z_{11F})|$ is much greater than $|(r_{b1}' + R_G + Z_{11F})|(1 - a_{01})$. These results are important since it means that the high-frequency stability of the amplifier is essentially independent of the low-frequency common emitter current gain.

If transistors are used that have high-frequency parameters superior to those listed in Table I, then the high-frequency stability of the amplifier is improved. In particular, an increase in alpha cutoff frequency or a reduction in collector capacitance or base resistance will tend to increase cutoff frequencies f_{11}, f_{42}, and f_{43}, thus providing larger gain and phase margins against instability.

5.5 *Electrical Performance and Gain Control Circuit*

The amplifier satisfies all of the electrical requirements (1) to (6). Fig. 8 shows a plot of the closed loop voltage gain of the amplifier. To

* This approximation assumes that the component of input resistance due to emitter resistance is much larger than r_b'. This approximation is valid for the transistors in the amplifier.

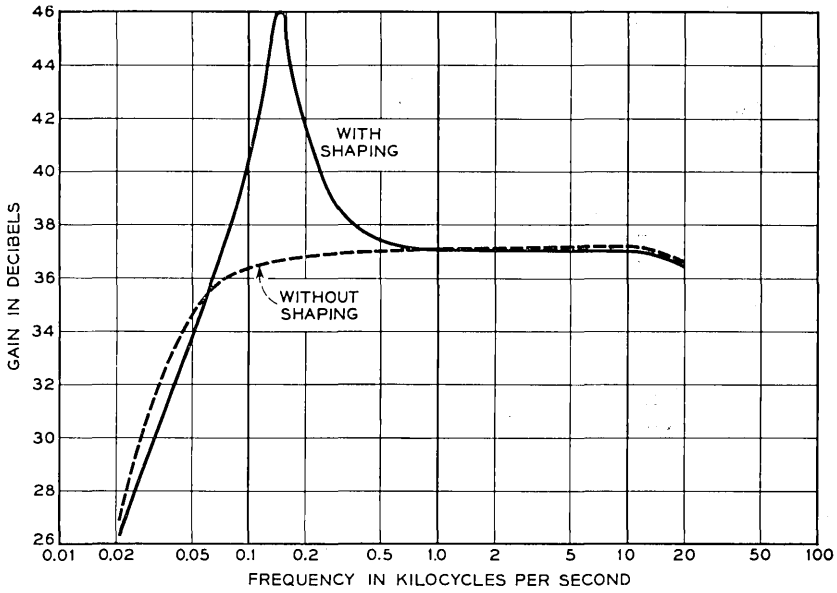


Fig. 8 — Closed loop voltage gain of the amplifier.

a good approximation, the shape of the gain-frequency characteristic is independent of the flat gain.

The gain control circuit consists of a 950-ohm potentiometer in series with the resistance of a 22-gauge wire with a length that may vary between 0 and 1000 feet. The resistance of the connecting wire has a maximum value of 16 ohms which can be neglected. In Section 5.2 it was pointed out that $R2$ had a maximum value of 270 ohms. This is obtained by placing a 383-ohm resistance across the remote gain control circuit as shown in Fig. 5. A 0.015-microfarad capacitor is also placed across the remote gain control circuit in order to minimize the effect of undesirable impedance variations introduced by the connecting wire acting as a transmission line at high frequencies.

VI. EQUIPMENT DESIGN

The equipment features of the A5 bank differ greatly from those of any previous long-haul carrier arrangements. In achieving the new format the chief objectives were: (1) miniaturization, (2) economical manufacture, (3) economical installation, and (4) simplified maintenance.

Taking advantage of the transistor amplifier and newly available passive components such as ferrite transformers, a substantial size re-

duction is achieved. The new bank, which is now a single equipment unit rather than a grouping of panels, is designed for 19-inch rack mounting. It occupies $12\frac{1}{2}$ inches of vertical height and has an over-all depth of 10 inches. Fig. 9 shows an A5 unit compared to an A4 bank. With presently available fuse arrangements and inter-bay cables, nine A5 banks can be provided on an 11-foot 6-inch bay. This represents a 3:1 improvement in space utilization over the A4 bank. It is expected that with development now in progress on new fuses and on smaller cables, ten A5 banks will be mounted in the 11-foot 6-inch bay. A picture of the size reductions from the first designs until now is given in Fig. 10. A close-up front view of the A5 bank is shown on Fig. 11. An A5 bank weighs about 70 pounds as compared to 270 pounds for the A4.

Consideration of the factors of repetitive manufacture and simplification of maintenance led to the particular arrangements of the A5 bank. Essentially, a channel bank provides twelve distinct circuits. These are basically the same in purpose, but each differs in one respect — operating frequency of the sideband produced. In the A5 design the circuit portions of each channel which are not frequency-sensitive are combined in identical units. Thus there are twelve exactly similar packages for each bank which contain the modulator and demodulator, the voice amplifier, and the level adjusting pads. Fig. 12 shows this "modem" unit. This twelve-fold increase in repetitive manufacture of a major portion of the bank leads to manufacturing simplification and economy. Also, for the first time it will be possible for the Operating Telephone Companies to stock a few spares of the active units which can be used in any channel.

As illustrated by Fig. 13, ease of installation and maintenance was carefully considered in the A5 design. Essentially the bank is comprised of two main sections. The main mounting frame contains the twelve channel filters and the common units such as the hybrids and the network which improves the operation of the end channels. All of these units are removable from the front of the assembly. Simple terminal strips are provided on both sides of this frame for the input and output cables. These are arranged for solderless wire-wrapping.

Attached to this frame is a hinged panel carrying the twelve identical "modem" units. Connection to these is by multi-pin connectors, and the units themselves are easily demountable. Thus, any "modem" unit can be removed from a channel without disturbing other working units. Seen on the rear of the door are small inductors which form part of the terminating filters for the modulators and demodulators. Since these differ for certain groups of channels they, too, have been isolated from

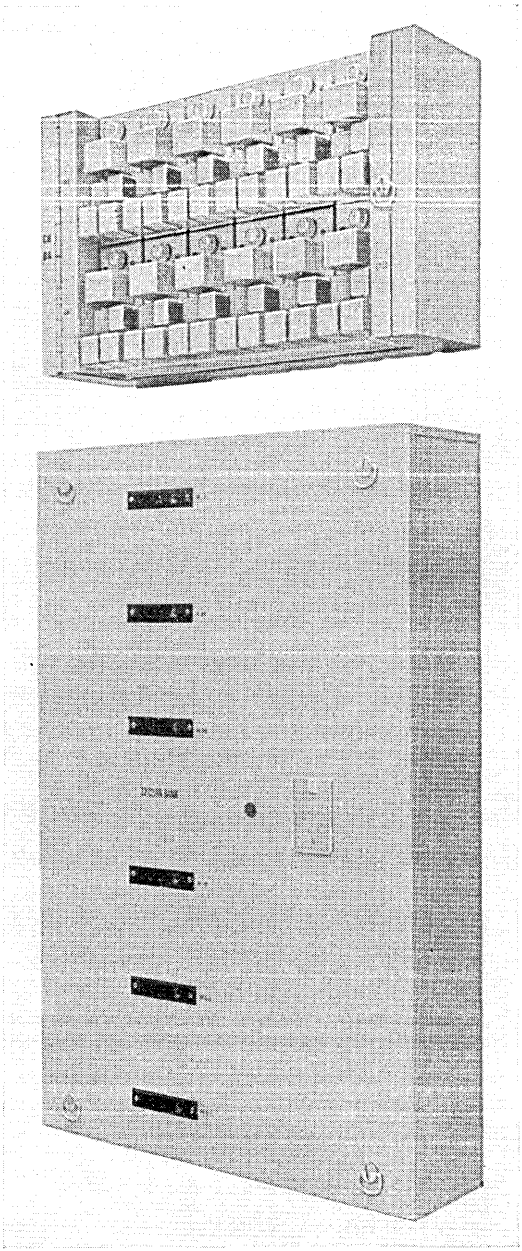


Fig. 9 — Comparison of A5 and A4 channel banks.

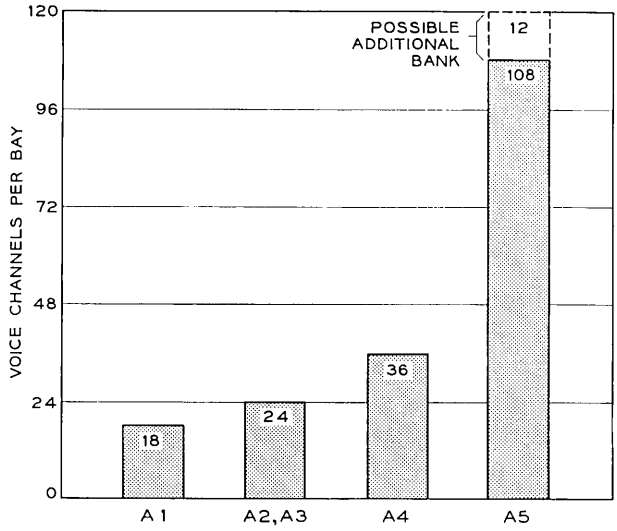


Fig. 10 — Size reductions in A-type channel banks since the A1.

the common "modem" units. The new equipment arrangements of the A5 bank provide for easy front-side maintenance and make even more attractive the common practice of mounting bays back to back.

Maintenance of the transistor amplifier has been made very simple. Pin jacks are provided on the front of each modem unit which permit

Fig. 11 — A5 bank with hinged panel closed.

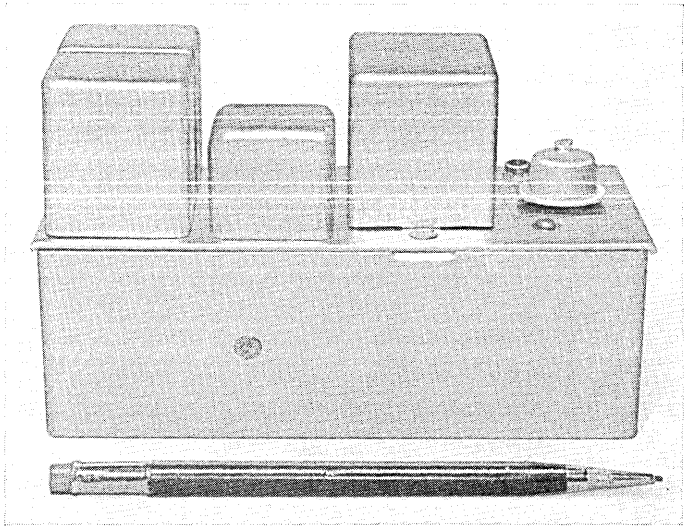


Fig. 12 — Modem unit of A5 bank.

measurements to indicate any degradation in the transistors. With the expected very long life of transistors, it is hoped that field experience will prove that no maintenance tests of this kind are needed. In this event it is likely that the pin jacks would no longer be provided.

From a power basis, the A5 bank offers substantial operational savings. It dissipates only 17 watts as compared to the 59 watts of an A4 bank, better than a 3:1 reduction. This means that even with miniaturization, there is no heat problem. A full complement of nine or ten A5

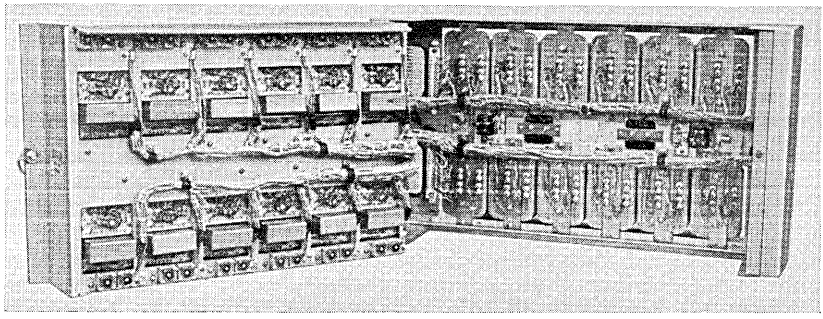


Fig. 13 — A5 bank with hinged door open showing rear of modem units and the filters mounted on rear frame.

units dissipates less than the present full bay of three A4 banks. Of course, an additional favorable factor is that a 130-volt power plant is not needed; the only voltage used is standard office 24 volts.

VII. CHANNEL FILTERS

From the very start of channel bank development, its filters have been a most important component. As previously described, they actually set the pattern which has since been followed. Each of the redesigns from A1 to A4 were also based mainly on filter changes, including reductions in their size. For the A1, the filters were quite large and expensive.[14] Later filters based on newly developed inductors and improvements in quartz crystal mounting and encasement were much smaller and led the way to the over-all equipment size reductions previously shown on Fig. 10.[20]

In the A5 development the filters represent mainly a packaging rearrangement of the A4 design. The main components, crystals and inductors, had already been very substantially reduced in size. To obtain minimum volume, the transmitting and receiving filters, each having the configuration shown in Fig. 14, are combined in one container. By careful placement of the two units and the wiring, and by individual shielding of the inductors, acceptable crosstalk coupling has been achieved. The filter mechanical arrangements are shown on Fig. 15.

Suppression characteristics of a typical filter are shown on Fig. 16 and the performance over the transmission band on Fig. 17.

VIII. OTHER COMPONENTS

The impressive reduction in size of the channel bank cannot be attributed solely to the transistor. True, the transistor itself, by virtue of its over-all size and low power consumption, contributed markedly to the miniaturization of the active circuitry. And the low impedances and voltages of the transistor amplifier circuit aided, as, of course, did efficient equipment design.

Full advantage could not have been taken of these factors, however, without the parallel advances that had been made in various passive components. These are effective not only in the active circuit but in other portions of the over-all transmission path. Outstanding in this field are the contributions of the ferrites. Their availability made possible extreme reductions in the size of transformers in all categories. Fig. 18 shows a comparison between some of the old and new units.

Similar reductions were achieved in the capacitor field. Small Mylar units are used in many instances. In addition, in the amplifier circuitry the very high capacitance values demanded by the low-impedance levels are furnished by miniature solid tantalum units. Without the latter the miniature equipment design would not have been possible.

IX. OVER-ALL PERFORMANCE

The over-all performance of the A5 channel bank meets its development objectives. For convenience in depicting the various characteristics they are discussed under two general headings: (1) those factors which were considered satisfactory in the earlier banks and are essentially unchanged, (2) those in which improvement was sought.

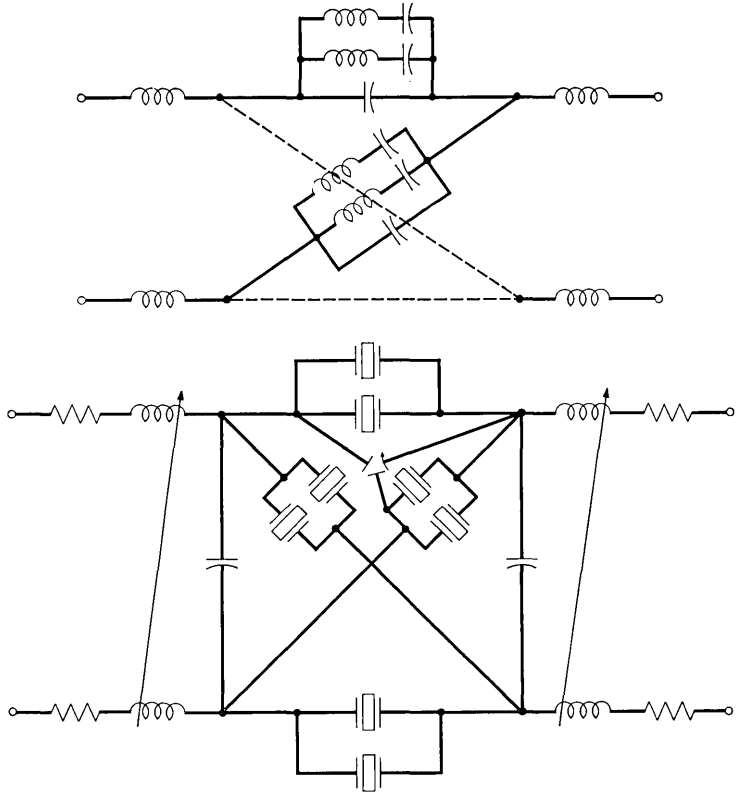


Fig. 14 — Configuration of a crystal channel filter.

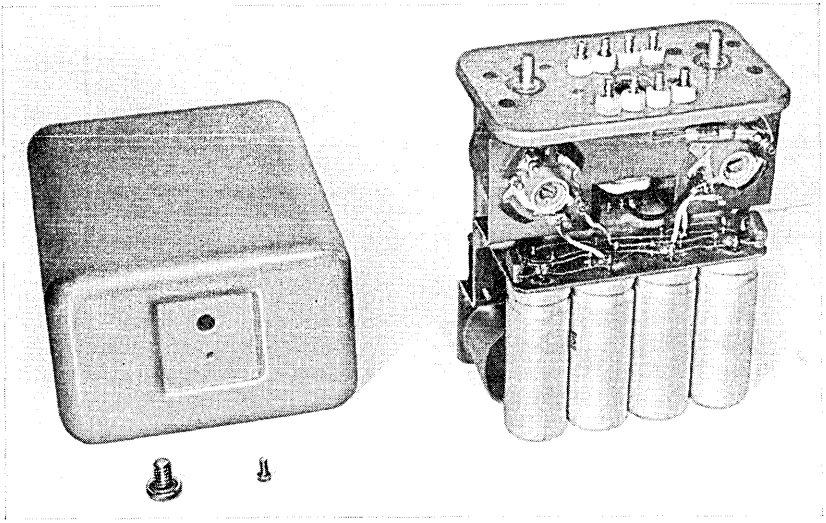


Fig. 15 — Crystal channel filter assembly comprising a transmitting and a receiving unit.

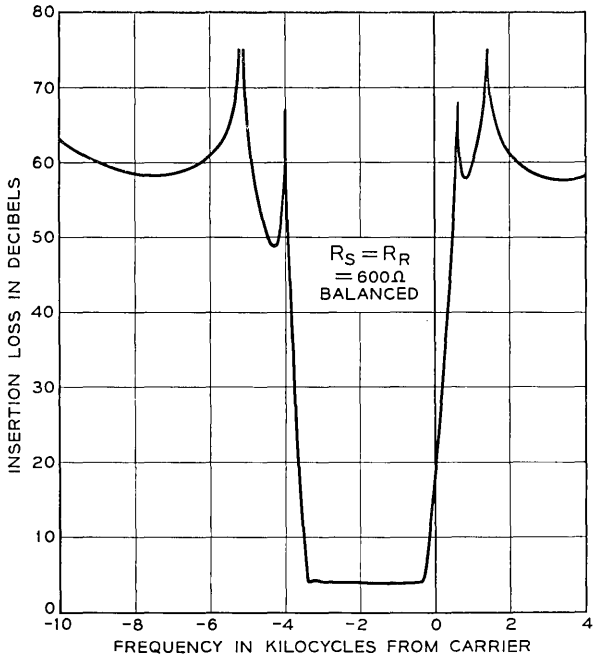


Fig. 16 — Typical loss-frequency characteristic of a channel filter.

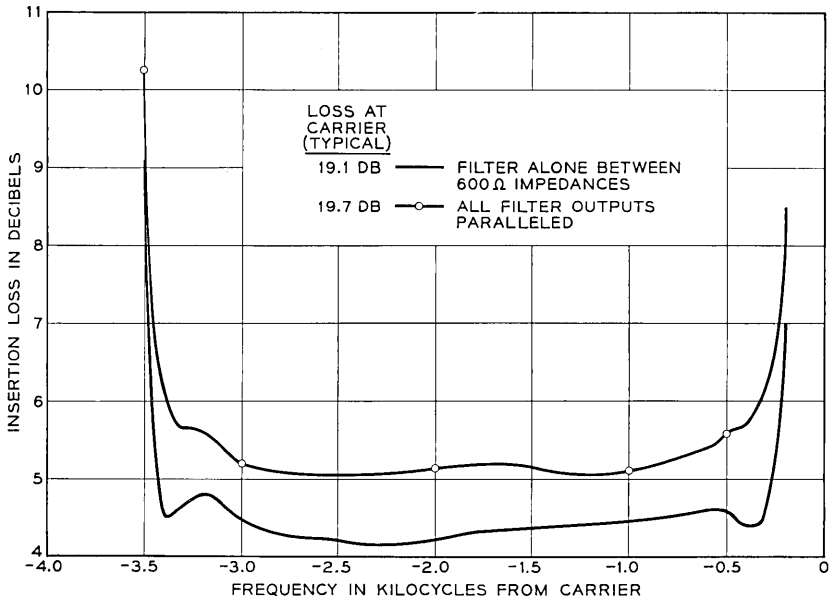


Fig. 17 — Typical pass-band characteristic of a channel filter.

In the first category are frequency response and modulator limiting. Since the chief determinants of the frequency response are the channel band filters, and since they are unchanged in electrical design, the same frequency performance is to be expected. A typical over-all channel characteristic is shown on Fig. 19. The gain-frequency behavior of a demodulator amplifier is shown on Fig. 8.

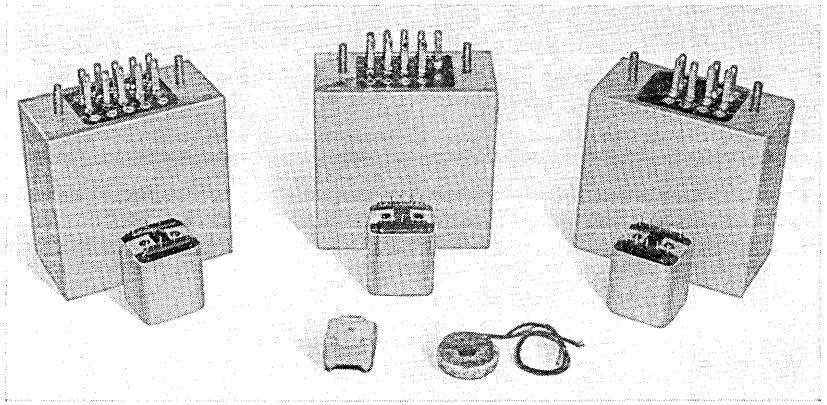


Fig. 18 — Comparison of transformers and inductors used in A5 and A4 banks.

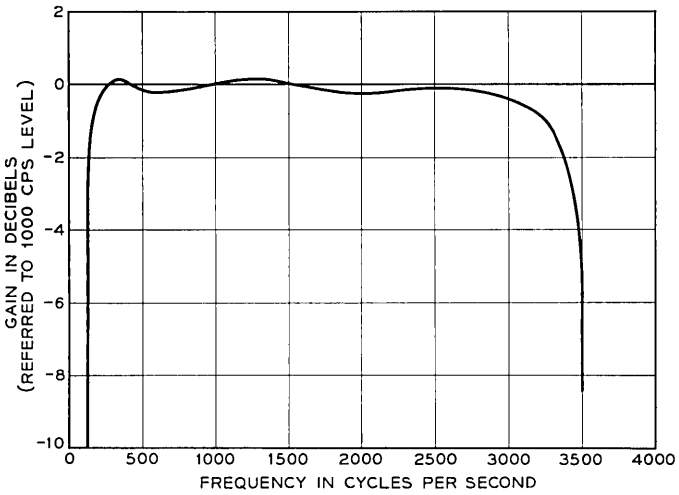


Fig. 19 — Typical channel gain-frequency characteristic of A5 bank.

The channel envelope delay is shown on Fig. 20. In terms of delay distortion, the following is derived from the curve:

Frequency Range (cycles per second)	Delay Distortion (microseconds)
1000–2500	100
850–2700	200
750–2900	300
600–3100	500

Modulator limiting is the same for both A5 and A4 and is shown on Fig. 21.

Temperature cycling tests indicate only very slight effect on band distortions; the largest effect is on channel net loss. From a nominal temperature of 80°F the bank was subjected to variations as great as ±60°F. The net loss variations are as follows:

Temperature Swing	Net Loss Change
±20°F	0.05 db
±40°F	0.25 db
±60°F	0.50 db

Earlier this paper outlined certain desired improvements in operating characteristics. These were particularly concerned with the behavior of the demodulator amplifier, primarily due to its lack of sufficient feedback. A very important aim was the stabilizing of net loss. In the A4 with normal battery variations, changes of two or three db were not uncommon. This situation was, of course, aggravated by aged tubes

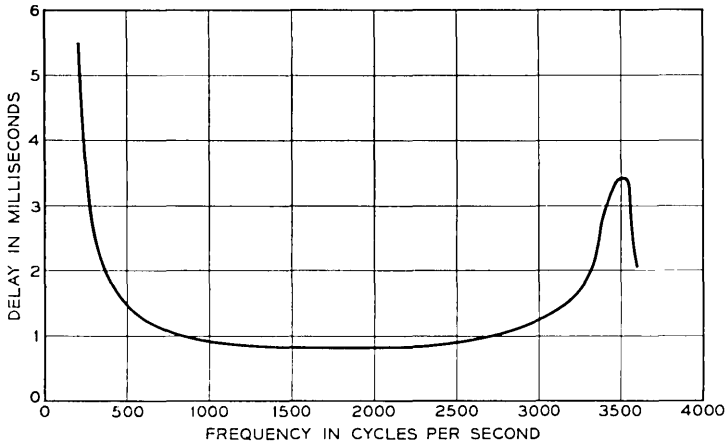


Fig. 20 — Typical channel delay-frequency characteristics of A5 bank.

and maximum gain settings. With 30-db minimum feedback, the aging of active elements should have negligible effect. Also, tests indicate that battery variations cause changes of only about 0.01 db per volt.

A4 channel banks require careful placing with regard to 60-cycle power sources in order to reduce noise pickup. The trouble arises in the amplifier input transformer which has a very high-impedance secondary and a large air gap. In the A5, the transistor circuitry requires a low-impedance transformer with a small air gap. Noise tests indicate that the pickup has been reduced to almost unmeasurable levels.

The intermodulation products created in previous banks make them

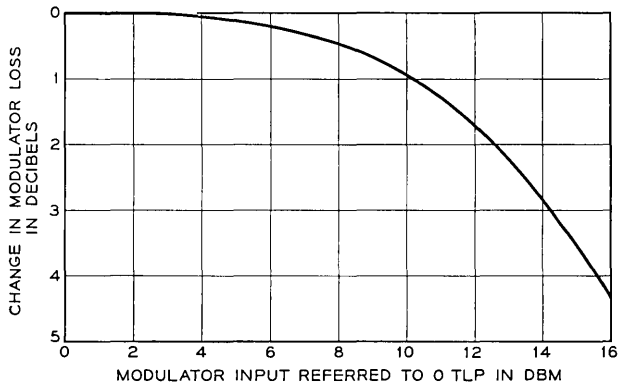


Fig. 21 — Typical channel overload characteristic showing modulator limiting.

unusable for certain special services, such as telephoto, without extreme lowering of levels and the addition of special amplifiers. The A5 amplifier has greatly improved this characteristic. Measurements show that for 0-dbm outputs of each of two frequencies, the $A - B$ product is -63 dbm and the $2A - B$ is -59 dbm. The requirement for telephoto, for example, is -48 dbm.

The amplifier, as designed to meet all of the requirements and employing the 9D power transistor, has a power-handling capacity superior to the older amplifier by about 6 db. The output characteristic as shown on Fig. 22 indicates a break point at about 22 to 23 dbm.

X. CONCLUSION

The A5 channel bank introduces the transistor into the long-haul wire and radio plant of the Bell System. Undoubtedly it is the forerunner of transistor circuitry in other portions of this equipment.

At about the same first cost of equipment, the A5 bank provides definitely improved service to meet the needs of today's communications. In the over-all it means savings to the Operating Telephone Companies through its reduced space and power requirements and its easier maintenance. With the expected high reliability and long life of transistors, the new bank should give many years of service without replacement of its active elements.

XI. ACKNOWLEDGEMENTS

A development of the magnitude of the A5 bank represents the contributions of many people. Certain ones played key roles, and the au-

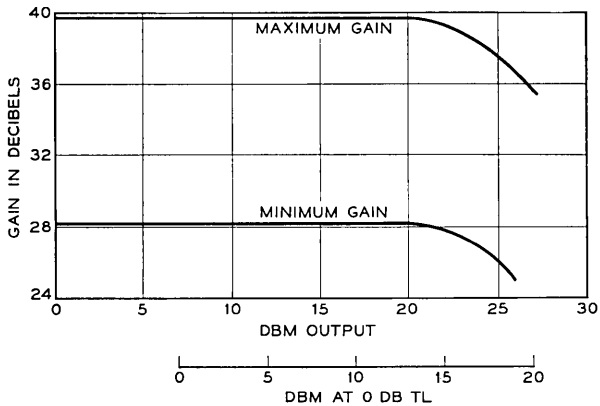


Fig. 22 — Overload characteristic of the transistor amplifier at 1 kc.

thors wish to acknowledge their very substantial part in the development. These are: W. G. Albert, F. R. Bies, J. L. Donoghue, J. B. Evans and J. J. Ginty of Systems Development, Merrimack Valley Bell Laboratories; W. E. Ballentine of Systems Development, Murray Hill Bell Laboratories; and H. G. Wells, now of the Western Electric Co., Merrimack Valley. Also, they wish to thank R. C. Boyd and J. J. Mahoney, Jr. of Systems Engineering for their co-operation and many helpful suggestions. To all the many others, unnamed, the authors also express deep appreciation.

APPENDIX A

Voltage Gain and Feedback Analysis for Transistor Amplifier

A.1 *Series Feedback Amplifier*

In this part of Appendix A, the expressions for voltage gain, feedback and input and output impedances for the series feedback amplifier shown in Fig. 2 are derived. It will be assumed that the amplifier A has no internal feedback and that the Z-parameter matrix for the amplifier is equal to

$$\begin{bmatrix} Z_{IN}' & 0 \\ Z_{21} & Z_{OUT}' \end{bmatrix} \tag{12}$$

where Z_{IN}' and Z_{OUT}' are the input and output impedances of the amplifier (without external feedback) respectively. The short-circuit current gain of the amplifier is related to the elements of this matrix by the expression

$$G_i = -\frac{Z_{21}}{Z_{OUT}'}. \tag{13}$$

The Z-parameter matrix for the feedback network is

$$\begin{bmatrix} Z_{11F} & Z_{12F} \\ Z_{12F} & Z_{22F} \end{bmatrix} \tag{14}$$

where

$$Z_{11F} = \frac{R1 \cdot (R2 + Z2)}{R1 + R2 + Z3}$$

$$Z_{12F} = \frac{R1 \cdot Z3}{R1 + R2 + Z3}$$

$$Z_{22F} = \frac{Z3 \cdot (R1 + R2)}{R1 + R2 + Z3}.$$

Since the amplifier A and the feedback network are connected in series, the over-all matrix for the series feedback amplifier is

$$\begin{bmatrix} Z_{IN}' + Z_{11F} & Z_{12F} \\ Z_{21} + Z_{12F} & Z_{OUT}' + Z_{22F} \end{bmatrix}. \tag{15}$$

The voltage gain of the series feedback amplifier is equal to

$$G_v = \frac{(Z_{21} + Z_{12F})R_L}{(Z_{IN}' + Z_{11F} + R_G)(Z_{OUT}' + Z_{22F} + R_L) - Z_{12F}(Z_{21} + Z_{12F})}. \tag{16}$$

At this point in the analysis it is convenient to introduce the loop current transmission, which is a convenient measure of feedback for a transistor amplifier. The current transmission is evaluated by opening the feedback circuit at terminals 1-1', terminating the left-hand pair of terminals in an impedance equal to Z_{IN}' and applying a unit input current to the right-hand pair of terminals. The loop current transmission is equal to the current in Z_{IN}'. The positive direction for this current is chosen so that if the original feedback circuit is restored, the current flows in the same direction as the unit input current. By straightforward calculation

$$A\beta = -\frac{G_i Z_{OUT}' Z_{12F}}{(Z_{IN}' + Z_{11F} + R_G)(Z_{OUT}' + Z_{22F} + R_L) - Z_{12F}^2} \tag{17}$$

$$\frac{A\beta}{1 - A\beta} = -\frac{G_i Z_{OUT}' Z_{12F}}{(Z_{IN}' + Z_{11F} + R_G)(Z_{OUT}' + Z_{22F} + R_L) - Z_{12F}^2 + G_i Z_{OUT}' Z_{12F}}. \tag{18}$$

If (13) and (18) are substituted into (16), then

$$G_v = \left[\frac{R_L}{Z_{12F}} - \frac{R_L}{G_i Z_{OUT}'}\right]\cdot\left[\frac{A\beta}{1 - A\beta}\right]. \tag{19}$$

In practice $| G_i | \gg 1$ and $| Z_{OUT}' | \gg | Z_{12F} |$. To a good approximation

$$G_v = \frac{R_L}{Z_{12F}}\cdot\frac{A\beta}{1 - A\beta} \tag{20}$$

The expressions for input and output impedances for the series feedback amplifier are readily obtained using Blackman's formula.[23]

$$Z = \frac{Z_0(1 - A\beta_0)}{(1 - A\beta_\infty)}. \tag{21}$$

In this formula, Z is the driving point impedance between two points in a circuit, Z_0 is the driving point impedance when one of the transistors in the amplifier A is in a reference condition so that $G_i = 0$ (feedback circuit is opened), $A\beta_0$ is the loop current transmission when the two points between which the impedance is measured are shorted, and $A\beta_\infty$ is the loop current transmission when the two points are open circuited.

In the case of input impedance, $Z_0 = (Z_{IN}' + Z_{11F})$, the input impedance of the amplifier A with the feedback loop opened plus Z_{11F} . $A\beta_0$ is equal to the loop current transmission with the input terminals to the series feedback amplifier, 1-3′, shorted. This value of loop current transmission can be calculated from (17) with R_G set equal to zero, and will be designated as $A\beta(R_G = 0)$. $A\beta_\infty$ is equal to the loop current transmission with the input terminals, 1-3′, open. Obviously $A\beta_\infty = 0$.

$$Z_{IN} = (Z_{IN}' + Z_{11F})[1 - A\beta(R_G = 0)]. \tag{22}$$

Similarly the expression for output impedance measured between terminals 2-4′ is

$$Z_{OUT} = (Z_{OUT}' + Z_{22F})[1 - A\beta(R_L = 0)] \tag{23}$$

where Z_{OUT}' is the output impedance of the amplifier A with the feedback loop opened and $A\beta(R_L = 0)$ is equal to (17) with R_L set equal to zero.

A.2 *Emitter Feedback Amplifier*

The emitter feedback amplifier shown in Fig. 3 differs from the series feedback amplifier in that the emitter of the second transistor and the interstage networks are returned to the common connection between the input and output transformers instead of to the emitters of the first and third transistors. As a result of this connection, the collector current of the first transistor must pass through the feedback network ($R1$, $R2$ and $Z3$) in returning to the emitter and, therefore, the network introduces local feedback to the first transistor stage. This feedback is present even when the main feedback loop is opened. Similarly, the base current of the third transistor stage must pass through the feedback network in returning to the emitter. The feedback network, consequently, introduces feedback to the third transistor stage.

The emitter feedback amplifier also differs from the series feedback amplifier in that the feedback network is driven by the emitter current of the output stage instead of by the collector current. Since the small signal collector current of a junction transistor is equal to the small

signal emitter current multiplied by the alpha of the transistor, expression (20) must be modified for the emitter feedback amplifier.

$$G_v = \frac{R_L a_3}{Z_{12F}} \cdot \frac{A\beta}{1 - A\beta}. \tag{24}$$

In order to calculate the feedback developed by an emitter feedback amplifier, it is convenient to use the circuit shown in Fig. 4. This circuit, to a good approximation, takes into account the local feedback introduced by the feedback network. Expressions (17), (22) and (23) are valid for the emitter feedback amplifier if Z_{IN}' and Z_{OUT}' are calculated for the circuit shown in Fig. 4.

APPENDIX B

Low-Frequency Stability Analysis of Transistor Amplifier

All of the low-frequency feedback shaping is introduced by the three capacitors in the emitter circuits of the transistors (refer to Fig. 5). Fig. 6 shows a plot of the low-frequency loop current transmission. The gain-cutoff at the frequency f_{A3} is introduced by the capacitor C_{E3}, in the emitter circuit of the last stage. At the cutoff frequency, the input impedance of the third stage is equal to the total shunt resistance, R_{S3}, between the base of the transistor and ground. To a good approximation

$$f_{A3} = \frac{1}{2\pi C_{E3} R_{S3}(1 - a_{03} + \delta)} \tag{25}*$$

where

$$\delta = \frac{R_L}{r_{c3}}.$$

The cutoff introduced by C_{E3} is terminated at the frequency f_{B3} at which the reactance of C_{E3} is equal to R_{E3}.

$$f_{B3} = \frac{1}{2\pi C_{E3} R_{E3}}. \tag{26}$$

The gain-cutoff at the frequency f_{A2} is introduced by C_{E2}. At this cutoff frequency, the input impedance of the second stage is equal to

* The first subscript for corner frequencies refers to the type of cutoff, while the second subscript refers to the transistor stage.

the total shunt resistance, R_{S2}, between the base of the transistor and ground. To a good approximation

$$f_{A2} = \frac{1}{2\pi C_{E2} R_{S2}(1 - a_{02})}. \tag{27}$$

The cutoff introduced by C_{E2} is terminated at the frequency f_{B2} at which the reactance of C_{E2} is equal to R_{E2}.

$$f_{B2} = \frac{1}{2\pi R_{E2} C_{E2}}. \tag{28}$$

With reference to (9), it is evident that C_{E1} will introduce a gain cutoff at the frequency where the reactance of the capacitor is equal to

$$f_{A1} = \frac{1}{2\pi C_{E1}[r_{b1}' + R_{11F} + R_G)(1 - a_{01}) + r_{e1} + R_{11F}]}. \tag{29}*$$

$$[(r_{b1}' + Z_{11F} + R_G)(1 - a_{01}) + r_{e1} + Z_{11F}]$$

This cutoff is terminated at the frequency f_{B1} at which the reactance of C_{E1} is equal to R_{E1}.

$$f_{B1} = \frac{1}{2\pi R_{E1} C_{E1}}. \tag{30}$$

APPENDIX C

High-Frequency Stability Analysis of Transistor Amplifier

In this Appendix we will calculate the high-frequency characteristic of the loop current transmission. Expression (17) is valid for the emitter feedback amplifier if Z_{IN}' and Z_{OUT}' are calculated for the circuit shown in Fig. 4.

$$G_i = \frac{a_1}{1 - a_1} \cdot \frac{a_2}{1 - a_2} \cdot \frac{1}{1 - a_3} \tag{31}$$

$$Z_{IN}' = r_{b1}' + \frac{r_{e1} + Z_{11F}}{1 - a_1} \tag{32}$$

$$Z_{OUT}' = Z_{c3}(1 - a_3) + Z_{22F}. \tag{33}$$

In practice, the term Z_{22F} in the expression for Z_{OUT}' can be neglected.

* At the frequency f_{A1}, Z_{11F} is real and has the value R_{11F}.

If (31), (32) and (33) are substituted into (17), then

$$A\beta = -\frac{\dfrac{a_1 a_2 Z_{12F}}{1-a_2}}{[r_{e1} + Z_{11F} + (r_b' + Z_{11F} + R_G)(1-a_1)] \cdot \left[(1-a_3) + \dfrac{R_L + Z_{22F}}{Z_{c3}}\right] - \dfrac{Z_{12F}^2(1-a_1)}{Z_{c3}}} \quad (34)$$

In all practical designs, the term $Z_{12F}^2(1-a_1)/Z_{c3}$ in the denominator of (34) can be neglected. In the following high-frequency analysis, it is assumed that the transistor parameters a and Z_c have the following frequency characteristic

$$a = \frac{a_0 \exp\left(-j\dfrac{fm}{f_a}\right)}{1 + j\dfrac{f}{f_a}} \quad (35)^{24}$$

$$Z_{c3} = \frac{r_{c3}}{1 + j2\pi f r_{c3} C_{c3}} \quad (36)$$

where a_0 is the dc value of the device parameter a, f_a is the frequency at which the magnitude of a is 3 db below its dc value, m is the number of radians by which the phase shift of a exceeds $\pi/4$ (45°) at f_a, C_{c3} is the collector capacitance and r_{c3} is the collector resistance of the third transistor. In practice, m is of the order of 0.2 for alloy types of transistors. If expressions (35) and (36) are substituted in (34), then to a good approximation

$$A\beta = \frac{A\beta_0\left(1 + j\dfrac{f}{f_{a3}}\right)}{\left(1 + j\dfrac{f}{f_{11}}\right)\left(1 + j\dfrac{f}{f_{12}}\right)\left(1 + j\dfrac{f}{f_{13}}\right)} \quad (37)^*$$

where $A\beta_0$ is equal to the mid-band value of $A\beta$ given by (9),

$$f_{11} = \frac{f_{a1}\left[1 - a_{01} + \dfrac{r_{e1} + R_{11F}}{r_{b1}' + R_{11F} + R_G}\right]}{1 + a_{01}m_1 + \dfrac{r_{e1} + R_{11F}}{r_{b1}' + R_{11F} + R_G}}$$

$$f_{12} = f_{a2}(1 - a_{02})$$

* At high frequencies, Z_{11F}, Z_{12F} and Z_{22F} are real with values R_{11F}, R_{12F} and R_{22F} respectively.

$$f_{13} = \frac{(1 - a_{03} + \delta)}{\frac{1 + a_{03}m_3 + \delta}{f_{a3}} + \frac{1}{f_{c3}}}$$

$$f_{c3} = \frac{1}{2\pi(R_L + R_{22F})C_{c3}}, \qquad \delta = \frac{R_L + R_{22F}}{r_{c3}}.$$

Expression (37) represents the high-frequency behavior of the loop current gain of the circuit without high-frequency shaping. In order to insure adequate margins against instability, two interstage high-frequency shaping networks are employed as shown in Fig. 5. These networks modify the high-frequency current gain of the second and third stages. In Ref. 21 it is shown that if a series RC circuit is placed in the base circuit of a transistor ($R2$, $C2$) the current gain of the stage can be represented by the following expression:

$$G_i = \frac{\frac{a_{02}}{1 - a_{02}}\left(1 + j\frac{f}{f_{32}}\right)}{\left(1 + j\frac{f}{f_{22}}\right)\left(1 + j\frac{f}{f_{42}}\right)} \tag{38}$$

where

$$f_{22} = \frac{1}{2\pi\left[r_{b2}' + R2 + \frac{r_{e2}}{1 - a_{02}}\right]C2}$$

$$f_{32} = \frac{1}{2\pi R2 \cdot C2}$$

$$f_{42} = \frac{f_{12}\left[r_{b2}' + R2 + \frac{r_{e2}}{1 - a_{02}}\right]}{r_{b2}' + R2}.$$

The corner frequency f_{22} corresponds to the frequency at which the reactance of $C2$ is equal to the input impedance of the second transistor stage. The corner frequency f_{32} is equal to the frequency at which the reactance of $C2$ is equal to $R2$.

In Ref. 21 it is shown that if a series RLC circuit is placed in the base circuit of a common emitter transistor ($R3$, $L3$ and $C3$), the current gain of the stage can be represented by the following expression:

$$G_i = \frac{\frac{a_{03}}{1 - a_{03} + \delta}\left(1 + j\frac{f}{f_{53}}\right)^2}{\left(1 + j\frac{f}{f_{23}}\right)\left(1 + j\frac{f}{f_{43}}\right)\left(1 + j\frac{f}{f_{63}}\right)} \tag{39}$$

where

$$f_{23} = \frac{1}{2\pi\left[r_{b3}' + R3 + \dfrac{(r_{e3} + R_{22F})(1 + \delta)}{1 - a_{03} + \delta}\right] C3}$$

$$f_{43} = \frac{f_{13}\left[r_{b3}' + R3 + \dfrac{(r_{e3} + R_{22})(1 + \delta)}{1 - a_{03} + \delta}\right]}{r_{b3}' + R3}$$

$$f_{53} = \frac{1}{2\pi\sqrt{L3C3}}$$

$$f_{63} = \frac{r_{b3}' + R3}{2\pi L3}.$$

The corner frequency f_{23} corresponds to the frequency at which the reactance of $C3$ is equal to the input impedance of the third transistor stage. The corner frequency f_{43} corresponds to the frequency at which the magnitude of the input impedance of the third stage plus $R3$ is within 3 db of $(r_{b3}' + R3)$. The corner frequency f_{53} corresponds to series resonance of the RLC circuit and f_{63} is the frequency at which the reactance of $L3$ is equal to $(r_{b3}' + R3)$. In order for (39) to be valid, $R3$ in the series resonant circuit must be chosen so that the circuit has a Q of one-half at f_{53}.

$$R3 = \frac{1}{\pi f_{53} C3}. \tag{40}$$

The complete expression for the high-frequency characteristic of $A\beta$ for the amplifier shown in Fig. 5, is

$$A\beta = \frac{A\beta_0\left(1 + j\frac{f}{f_{32}}\right)\left(1 + j\frac{f}{f_{53}}\right)^2\left(1 + j\frac{f}{f_{a3}}\right)}{\left(1 + j\frac{f}{f_{11}}\right)\left(1 + j\frac{f}{f_{22}}\right)\left(1 + j\frac{f}{f_{42}}\right)\cdot\left(1 + j\frac{f}{f_{23}}\right)\left(1 + j\frac{f}{f_{43}}\right)\left(1 + j\frac{f}{f_{63}}\right)}. \tag{41}$$

The magnitude and phase of $A\beta$ is plotted in Fig. 7 for the transistor parameter values listed in Table I and for $|Z_{12F}| = 7$ ohms.

REFERENCES

1. Colpitts, E. H., B.S.T.J., **16,** April, 1937, p. 119.
2. Black, H. S., Elect. Eng., **53,** Jan., 1934, p. 114.

3. Affel, H. A., Demarest, C. S., and Green, C. W., B.S.T.J., **7,** July, 1928, p. 564.
4. Green, C. W., and Green, E. I., B.S.T.J., **17,** Jan., 1938, p. 80.
5. Chestnut, R. W., Ilgenfritz, L. M., and Kenner, A., B.S.T.J., **17,** Jan., 1938, p. 106.
6. Kendall, B. W., and Affel, H. A., B.S.T.J., **28,** Jan., 1939, p. 119.
7. Espenschied, L., and Strieby, M. E., B.S.T.J., **13,** Oct., 1934, p. 654.
8. Strieby, M. E., B.S.T.J., **16,** Jan., 1937, p. 1.
9. Crane, R. E., Dixon, J. T., and Huber, G. H., Trans. A.I.E.E., **66,** 1947, p. 1451.
10. Elmendorf, C. H., Ehrbar, R. D., Klie, R. H., and Grossman, A. J., B.S.T.J. **32,** July, 1953, p. 781.
11. Affel, H. A., B.S.T.J., **16,** Oct., 1937, p. 487.
12. Peterson, E., Manley, J. M., and Wrathall, L. R., B.S.T.J., **16,** Oct., 1937, p. 437.
13. Buckley, O. E., J. Applied Phys., **8,** Jan., 1937, p. 40.
14. Mason, W. P., B.S.T.J., **13,** July, 1934, p. 405.
15. Lane, C. E., B.S.T.J., **17,** Jan., 1938, p. 125.
16. Grieser, T. J., and Peterson, A. C., Elect. Engg., **70,** Sept., 1951, p. 810.
17. Roetken, A. A., Smith, K. D., and Friis, R. W., B.S.T.J., **30,** Oct., 1951, p. 1041.
18. McDavitt, M. B., Trans. A.I.E.E., **76,** Part 1, Jan., 1958, p. 715.
19. Gammie, J., and Hathaway, S. D., B.S.T.J., **39,** July, 1960, p. 821.
20. Willis, E. S., Trans. A.I.E.E., **65,** March, 1946, p. 134.
21. Blecher, F. H., Trans. I.R.E., **CT-4,** Sept., 1957, p. 145.
22. Bode, H. W., *Network Analysis and Feedback Amplifier Design,* D. Van Nostrand and Co., New York, 1945, p. 39.
23. Ibid., p. 66.
24. Pritchard, R. L., Proc. I.R.E., **40,** Nov., 1952, p. 1476.

On the Use of Passive Circuit Measurements for the Adjustment of Variable Capacitance Amplifiers

By KANEYUKI KUROKAWA

(Manuscript received April 12, 1961)

In the field of microwave tubes, the cold test plays an important role. Hcwever, no attempt in this direction has been made for the variable capacitance amplifier. It is the purpose of this paper to present the theory of a cold test procedure for parametric amplifiers. The cold test is essentially the measurement of the impedance locus of the input, the output and the pump circuits under the no-pump condition, with the diode bias voltage as the variable parameter. From these impedance loci one can evaluate all the important circuit parameters of the equivalent circuit of the parametric amplifier, including the value of the dynamic quality factor of the diode. Using these data, it is relatively easy to design or adjust the circuit so as to give the best noise performance. Since the theory of the cold test procedure neglects the effects of (1) higher harmonics, (2) any parallel conductance of the diode, and (3) circuit losses, the validity of the theory can be established only by experiment. For this reason, this paper also presents some of the experimental results obtained with a 6-kmc degenerate amplifier. The correlation between theory and experiment has proved to be excellent.

I. INTRODUCTION

Although variable capacitance amplifiers have been built and operated successfully, their design has been an art practiced by the individual designer rather than a systematic construction. The designer generally provides various adjustable components in his amplifier and obtains the final optimum result on a trial and error basis. However, the designer can never be free from the fear that just one more adjustable component may considerably improve the noise figure. Furthermore, when the measured noise figure is not as good as expected from some previous

experience, there is no simple way to determine which circuit is the main cause of the poor noise figure, i.e., the signal, the idler or the pump circuit. The poor result might also be due to a poor diode. The cold test procedure goes a long way towards resolving this uncertainty.

In microwave tubes, such as the klystron and magnetron, the cold test plays an important role. However, no attempt in this direction has been made for the variable capacitance amplifiers. It is the purpose of this paper to present both the theory of a cold test procedure for parametric amplifiers and some experimental results on a 6-kmc degenerate amplifier showing the validity and the limit of applicability of the theory. It can be shown[1] that the minimum noise figures of lower-sideband idler-output and circulator type amplifiers, degenerate amplifier and upper-sideband up-converter are all basically determined by a dynamic quality factor of the diode and that, for optimum noise figure operation, each type of amplifier requires certain values of R_s/R_g and R_s/R_L, where R_s is the series resistance of the diode, R_g the internal resistance of the generator and R_L the load resistance. The cold test is essentially the measurement of the impedance locus of the input, the output, and the pump circuits, under the no-pump condition with the diode bias voltage as the variable parameter. From these loci all the important circuit parameters of the equivalent circuit of the parametric amplifier, including the diode dynamic quality factor, can easily be evaluated. Using these data, it is quite straightforward to design or adjust the circuit so as to give the best performance under the restriction of a three-frequency assumption. Further, should a poor noise figure be obtained, one can easily detect from the equivalent circuit where the main trouble lies.

The cold test procedure thus has a large advantage over the conventional trial and error method. It must be pointed out, however, that the cold test checks only the necessary conditions which must be satisfied to obtain a low-noise amplifier. Sufficient conditions are far beyond the scope of this paper, since the present theory of parametric amplifiers completely neglects the effect of higher harmonics. Also, it must be mentioned that if, in addition to the series resistance, there is an appreciable amount of parallel leakage conductance in the diode, the technique discussed here is not directly applicable. Such a diode is regarded as poor and would generally not be used; the theory of minimum noise figure for such a diode has, therefore, not yet been fully developed.

II. EQUIVALENT CIRCUIT OF PARAMETRIC AMPLIFIER

The parametric amplifier which we shall consider is a network with a variable capacitance, and an input (ω_1), an output (ω_2), and a pump

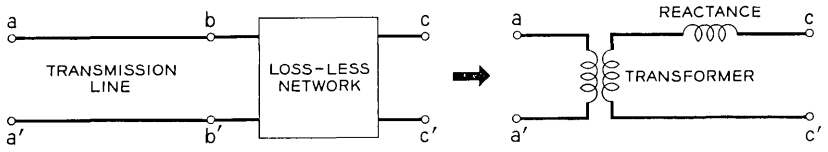


Fig. 1 — Equivalence of two networks.

(ω_p) circuit. Without any loss of generality, we can assume that the input is in the form of a transmission line with a matched generator, and the output a transmission line with a matched load. For a properly designed amplifier the network should have negligible losses except for the diode series resistance. Further, the input, the output and the pump circuits should be isolated from each other except through the parametric action. Therefore, the whole circuit connecting the input terminals to the diode is considered as a lossless two-terminal-pair network at ω_1, and the whole circuit connecting the output terminals to the diode as another lossless two-terminal-pair network at ω_2. By choosing a proper reference plane along the transmission line, each lossless two-terminal-pair network can be considered as a simple combination of an ideal transformer and a series reactance, as shown in Fig. 1 (see Appendix A). Therefore, the equivalent circuit of the amplifier becomes that shown in Fig. 2. For our purpose, however, the transformer can be eliminated as shown in Fig. 3, since only the ratios between the various impedances are important, and not their actual values. For example, the minimum noise figure condition requires a certain value of R_s/R_g, where R_g is the internal resistance of the generator looking back from the diode. R_s/R_g is the diode series resistance normalized to the generator resistance R_g in Fig. 3, and it corresponds to $R_s/(n_1^2 r_g)$ in Fig. 2, which is again the diode series resistance seen from and normalized to the generator resistance r_g, i.e., $(R_s/n_1^2)/r_g$. Thus, there is no difference between Fig. 2 and Fig. 3 as long as the impedances are all measured from the transmission lines.

The pump circuit is yet another lossless two-terminal-pair network, but this circuit does not have to satisfy any stringent requirements.

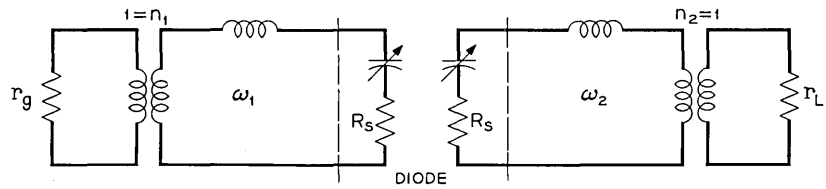


Fig. 2 — Equivalent circuit of the parametric amplifier.

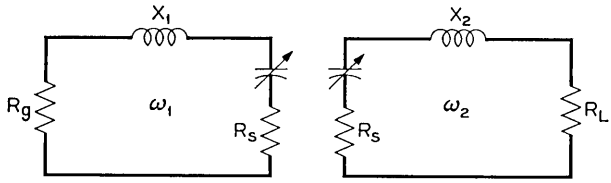


Fig. 3 — Elimination of the transformer in the equivalent circuit of Fig. 2.

The function of this circuit is to provide enough pump voltage across the diode junction without letting the signal and the idler energy escape into the pump generator. The coupling between the diode and the pump input is easily checked in a way similar to that used for the ω_1 and ω_2 circuits.

III. DETERMINATION OF CIRCUIT PARAMETERS

The impedance looking into the network from the input (or the output) is a series connection of a reactance, a variable capacitance and a resistance. If we change the capacitance value by applying a dc voltage to the diode, the impedance changes but the resistive part remains constant. Thus, the locus of the impedance should be a part of a constant resistance circle on the Smith chart. If we set the reference plane arbitrarily, the circle is generally not a constant resistance circle, but it is always possible (under the model assumed) to rotate the experimentally determined circle about the center of the Smith chart until it fits one of the constant resistance circles. This procedure corresponds to the proper choice of the reference plane in the previous section. The resistance value of the circle thus obtained gives R_s/R_g (or R_s/R_L). The mean value of the reactance gives the additional reactance in the input (or output) circuit, i.e., the diode average reactance plus the circuit reactance. The minimum noise figure condition requires that this additional reactance be zero. The locus should, therefore, be located symmetrically on the two sides of the zero reactance line when the bias voltage changes in the same manner as the pump voltage across the diode junction.

Making the open circuit assumption for the unwanted frequencies, the diode dynamic quality factor $\tilde{Q}$ is defined by

$$\tilde{Q} = \frac{1}{\omega 2 K_1 R_s} \tag{1}$$

$$\simeq \frac{\text{Total reactance variation}}{4R_s} = \frac{2\Delta X}{4R_s}. \tag{2}$$

The quantity K_1 is defined through the relationship

$$\frac{1}{C(t)} = \frac{1}{K_0} + \frac{1}{K_1} \cos \omega_p t + \cdots \tag{3}$$

where $C(t)$ is a junction capacitance which is a periodic function of time. The value of $\tilde{Q}$ is thus easily evaluated from the impedance locus using (2).

In certain cases, the circle of the impedance locus does not come close to the periphery of the Smith chart, and, thus no appropriate constant resistance circle can be obtained by a rotation of the experimentally determined circle. This indicates either that the network is lossy, or else that the diode has a parallel conductance in addition to the series resistance. In the former case the network should be redesigned or readjusted to eliminate this additional loss. In the latter case the present technique is not directly applicable.

IV. MINIMUM NOISE FIGURE ADJUSTMENT

In addition to the condition that the locus be symmetrical about the zero reactance line, each type of amplifier requires certain values of R_s/R_g and R_s/R_L for optimum noise figure operation. These values are detailed in a previous paper[1] and will be quoted in this section without further reference.

4.1 *Upper Sideband Up-converter*

For the upper sideband up-converter, the minimum noise figure condition is given by

$$\frac{R_s}{R_g} = \frac{1}{\sqrt{1 + \tilde{Q}_1^2}} \tag{4}$$

and

$$\frac{R_s}{R_L} = \frac{1}{1 + \dfrac{\tilde{Q}_1 \tilde{Q}_2}{1 + \sqrt{1 + \tilde{Q}_1^2}}}. \tag{5}$$

The corresponding noise figure and gain are

$$F = 1 + 2\frac{T_s}{T_g}\left(\frac{1}{\tilde{Q}_1^2} + \frac{1}{\tilde{Q}_1}\sqrt{1 + \frac{1}{\tilde{Q}_1^2}}\right) \tag{6}$$

where T_s and T_g (290°K) are the equivalent noise temperatures of the diode and the source respectively, and

$$G = \frac{\frac{\omega_2}{\omega_1}}{\left(1 + \frac{1}{\tilde{Q}_1\tilde{Q}_2} + \frac{1}{\tilde{Q}_1\tilde{Q}_2}\sqrt{1 + \tilde{Q}_1^2}\right)\left(1 + \frac{1}{\sqrt{1 + \tilde{Q}_1^2}}\right)} \tag{7}$$

From (2) and (4),

$$\frac{\Delta X_1}{R_g} = \left(\frac{R_s}{R_g}\right)\left(\frac{\Delta X_1}{R_s}\right) = \frac{2\tilde{Q}_1}{\sqrt{1 + \tilde{Q}_1^2}} \tag{8}$$

$$\simeq 2, \quad \text{if } \tilde{Q}_1 \text{ is large.}$$

Combining (4) and (8), a numerical calculation shows that the two ends of the locus should be found on the solid line shown on the Smith chart of Fig. 4. The dotted line is an example of the locus. It is worth noting

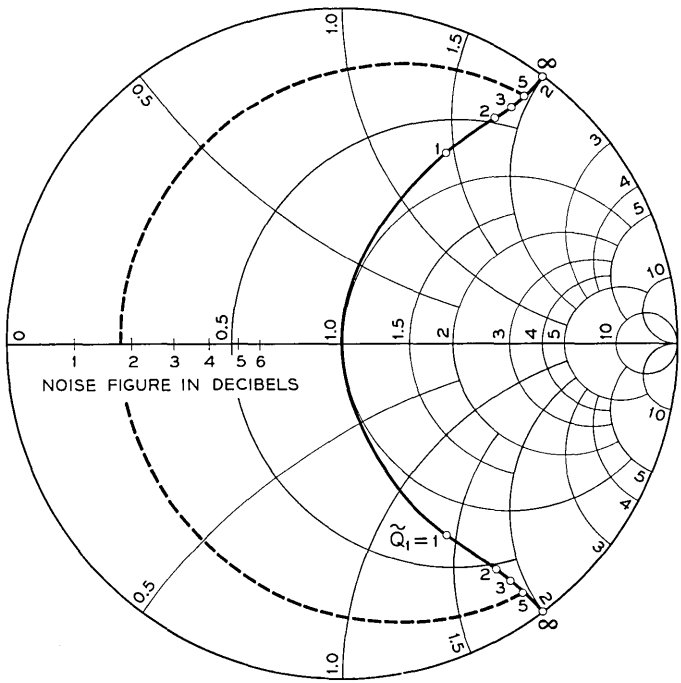


Fig. 4 — Impedance locus of input of up-converter for minimum noise figure. The dotted line is an example. The solid line is the limit of impedance swing.

that in most practical cases $\tilde{Q}_1$ is large, and, hence, the locus should extend nearly from $j2$ to $-j2$. Next let us consider the output circuit. From (5) we have

$$\frac{\Delta X_2}{R_L} = \frac{2\tilde{Q}_2}{1 + \dfrac{\tilde{Q}_1\tilde{Q}_2}{1 + \sqrt{1 + \tilde{Q}_1^2}}}. \tag{9}$$

If $\tilde{Q}_1$ and $\tilde{Q}_2$ are large, as is usually the case,

$$\frac{\Delta X_2}{R_L} \simeq 2. \tag{10}$$

This means that the output impedance locus also should extend from $j2$ to $-j2$.

Whenever the succeeding stage has a poor noise figure compared to the noise figure of the up-converter, the maximum gain condition of the up-converter gives a better over-all noise performance than the minimum noise figure condition. The maximum gain condition is given by

$$\frac{R_s}{R_g} = \frac{R_s}{R_L} = \frac{1}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2}}. \tag{11}$$

The corresponding noise figure and the maximum gain of the up-converter are given by

$$F = 1 + \frac{T_s}{T_g}\left(\frac{1}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2}} + \frac{\omega_1}{\omega_2}\frac{1}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2}}\frac{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2} + 1}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2} - 1}\right) \tag{12}$$

and

$$G = \frac{\omega_2}{\omega_1}\frac{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2} - 1}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2} + 1}. \tag{13}$$

Equation (11) shows that both the input and output impedance loci should lie on the same resistance circle of the Smith chart. The reactance variations are given by

$$\frac{\Delta X_1}{R_g} = \frac{2\tilde{Q}_1}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2}} \tag{14}$$

and

$$\frac{\Delta X_2}{R_L} = \frac{2\tilde{Q}_2}{\sqrt{1 + \tilde{Q}_1\tilde{Q}_2}}. \tag{15}$$

When $\tilde{Q}_1\tilde{Q}_2$ is large, they become

$$\frac{\Delta X_1}{R_g} \simeq 2\sqrt{\frac{\omega_2}{\omega_1}} \tag{16}$$

and

$$\frac{\Delta X_2}{R_L} \simeq 2\sqrt{\frac{\omega_1}{\omega_2}}. \tag{17}$$

In most practical cases, however, because of the moderate noise figure of the succeeding stage, neither the minimum noise figure condition nor the maximum gain condition gives the best over-all noise figure — this is always obtained somewhere between these two conditions. If the noise figure of the succeeding stage is given, the condition for the best over-all noise performance, and, hence, the required impedance loci, are easily calculated.[1] Thus the cold test procedure can also be used to adjust for the best over-all noise figure adjustment.

4.2 *Lower Sideband Nondegenerate Amplifiers*

For the lower sideband nondegenerate amplifiers (both idler-output and circulator types), the minimum noise figure condition for large gain is given by

$$\frac{R_s}{R_g} = \frac{1}{\tilde{Q}_1\tilde{Q}_2 - 1} \tag{18}$$

and

$$R_s/R_L \to \infty. \tag{19}$$

The minimum noise figure is

$$F = 1 + \frac{\frac{T_s}{T_g}}{\tilde{Q}_1\tilde{Q}_2 - 1}\left(1 + \frac{\omega_1}{\omega_2}\tilde{Q}_1\tilde{Q}_2\right). \tag{20}$$

Since $\tilde{Q}_1 = \Delta X_1/2R_s$ from (2), (18) may be rewritten as

$$\frac{\Delta X_1}{R_g} = \frac{2\tilde{Q}_1}{\tilde{Q}_1\tilde{Q}_2 - 1} \tag{21}$$

and so the input reactance should vary from

$$\frac{+j2\tilde{Q}_1}{\tilde{Q}_1\tilde{Q}_2 - 1} \quad \text{to} \quad \frac{-j2\tilde{Q}_1}{\tilde{Q}_1\tilde{Q}_2 - 1}.$$

From (19), the output locus should converge to the infinite resistance circle, keeping the center of the locus on the zero reactance line.

It sometimes happens that because of bandwidth or stability requirements, condition (19) can not be satisfied, i.e., the value of R_s/R_L has to remain finite. The minimum noise figure condition under this restriction, and the corresponding noise figure, are given by (18) and (20) respectively, provided that everywhere $\tilde{Q}_1\tilde{Q}_2$ is replaced by

$$\tilde{Q}_1\tilde{Q}_2\left(\frac{R_s}{R_s + R_L}\right)$$

and that the load and diode temperatures are equal. The reactance variations are given by

$$\frac{\Delta X_1}{R_g} = \frac{2\tilde{Q}_1}{\tilde{Q}_1\tilde{Q}_2\left(\frac{R_s}{R_s + R_L}\right) - 1} \tag{22}$$

and

$$\frac{\Delta X_2}{R_L} = \frac{2R_s}{R_L}\,\tilde{Q}_2. \tag{23}$$

When R_s/R_L is finite, $\tilde{Q}_1$ must be replaced by

$$\tilde{Q}_1\sqrt{\frac{R_s}{R_s + R_L}}$$

to obtain the corresponding noise figure from Fig. 4 of the referenced paper.[1]

4.3 *Degenerate Amplifier*

For the degenerate amplifier, the minimum noise figure for large gain is

$$F = 1 + \frac{T_s}{T_g}\frac{1}{\tilde{Q} - 1} \tag{24}$$

and is obtained when

$$\frac{R_s}{R_g} = \frac{1}{\tilde{Q} - 1} \tag{25}$$

Noting again that $\tilde{Q}_1 = \Delta X/2R_s$ from (2), we have

$$\frac{\Delta X}{R_g} = \frac{2\tilde{Q}}{\tilde{Q} - 1}. \tag{26}$$

A little calculation shows that the locus terminates on the straight lines through the $(\infty, j\infty)$ and $(0, \pm j2)$ points, as shown in Fig. 5. The dotted line is an example of the locus.

The above discussion is entirely based on two assumptions, (*i*) that the unwanted frequencies are open-circuited, and (*ii*) that the reactance variation of the diode junction is sinusoidal. In practice, however, it is difficult to satisfy these assumptions completely, and the actual reactance variation required for optimum performance may be slightly larger than that indicated above. Nevertheless, experiments on lower sideband amplifiers have shown that the effect of improperly terminating the higher sidebands is small as long as the circuit is well detuned at the frequency of the upper sideband. Experiments with an upper sideband up-converter also support the theory, even though in the actual experimental case the lower sideband was not perfectly open circuited. The experimental results on a degenerate amplifier will be discussed in detail in Section VI.

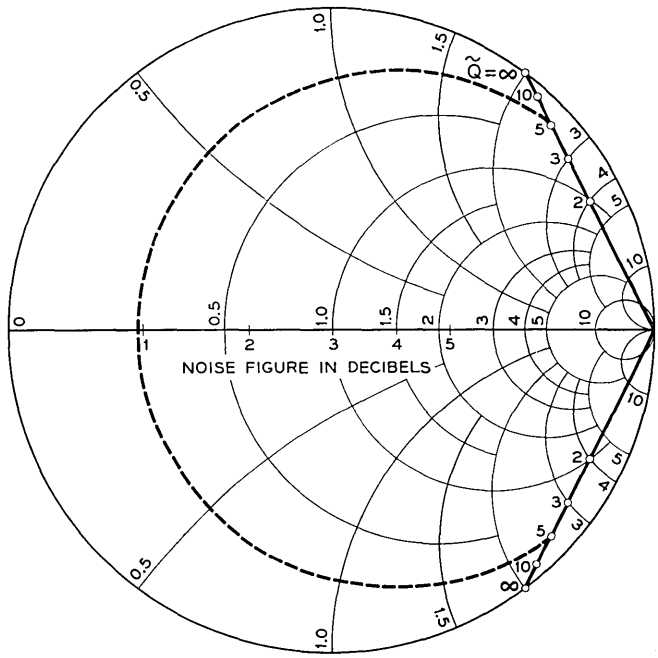


Fig. 5 — Impedance locus of degenerate amplifier with large gain and minimum noise figure. The dotted line is an example. The solid line is the limit of impedance swing.

The best starting point for adjusting an amplifier is to provide about 10 to 20 per cent larger reactance variation than that indicated above, and to test the amplification with reduced pump power. If the amplification is satisfactory, one can decrease R_s/R_g, improving the noise figure. The normalized reactance variation, however, decreases proportionally to R_s/R_g, and accordingly the pump power required for proper amplification increases. If the pump voltage across the diode junction increases too much, additional noise such as microplasma or shot noise becomes effective and the noise figure deteriorates again. Therefore, the decrease of R_s/R_g is stopped just before the noise figure begins to deteriorate.

V. BIAS SWEEPING METHOD

Since a large number of standing-wave measurements have to be made for just one impedance locus, it is a time-consuming job to get the proper locus by adjusting the actual circuit. However, the procedure is considerably simplified by modulating the bias and displaying the output from the standing-wave detector on an oscilloscope. A schematic diagram of the bias sweeping method is shown in Fig. 6. The diode bias voltage is modulated at 60 cps by means of an ac voltage applied through a transformer, and this voltage is also applied to the horizontal amplifier of the oscilloscope. The vertical axis of the oscilloscope shows the output of the standing-wave detector. Thus, if a square-law detector is used, the pattern on the oscilloscope represents the square of the length from a reference point on the periphery of the Smith chart to the locus versus the bias voltage. This relation is shown in Fig. 7 (see Appendix B). If one moves the probe position of the standing-wave detector k wave-

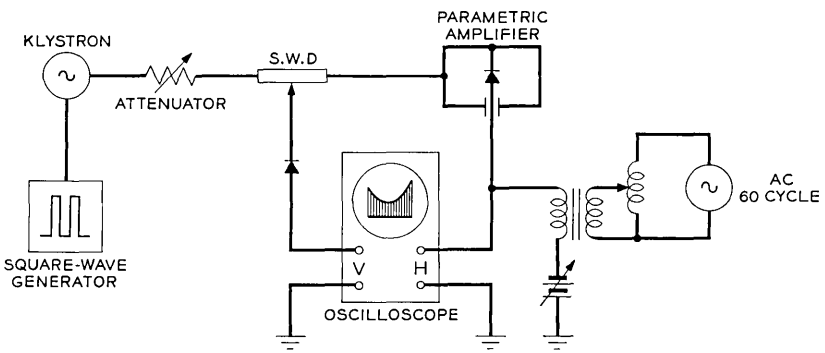


Fig. 6 — Bias sweeping method.

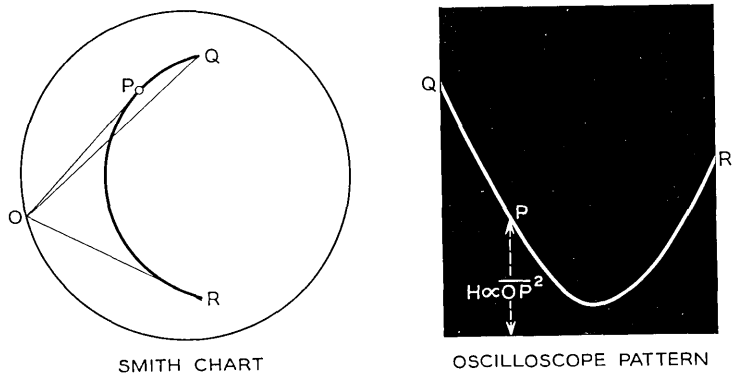


Fig. 7 — Relation between impedance locus and oscilloscope pattern.

lengths towards the load, the reference point moves k wavelengths clockwise along the periphery of the Smith chart (in the direction of the arrow indicating "wavelength toward generator"). Therefore, the pattern changes as shown in Fig. 8. Keeping these relations in mind, observing the pattern on the oscilloscope and moving the probe position of the standing-wave detector, one can visualize the shape of the locus on the Smith chart. After having adjusted the circuit for the proper pattern on the scope, the usual point-by-point measurement can be made, if desired.

The pump circuit can also be checked by the bias sweeping method. The proper impedance variation is an indication of good coupling between the pump input and the diode. No critical adjustment is required

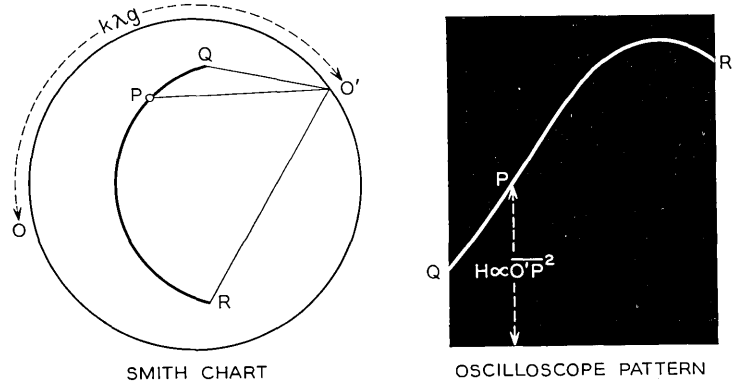


Fig. 8 — Shift of probe position and oscilloscope pattern.

for the pump circuit; however, it is desirable that the locus be reasonably symmetrical about the zero reactance line, and that the impedance swing be reasonably large.

VI. EXPERIMENTAL RESULTS

Since the theory neglects the effects of (*i*) higher harmonics, (*ii*) any parallel conductance of the diode, and (*iii*) circuit losses, the validity of the theory can be established only by experiment. For such a validity check, it is best to start with a degenerate amplifier, since this has the simplest circuit configuration. In this section we shall present some of the experimental results obtained with a 6-kmc degenerate amplifier.

The diode mount and also the noise figure measuring setup are similar to those described by M. Uenohara.[2] To make input impedance measurements, a standing-wave detector is inserted between the diode mount and the circulator. The circuit adjustment is done by the bias sweeping method. After obtaining the desired impedance locus, the locus is plotted on the Smith chart by the usual point-by-point measurement. Figs. 9

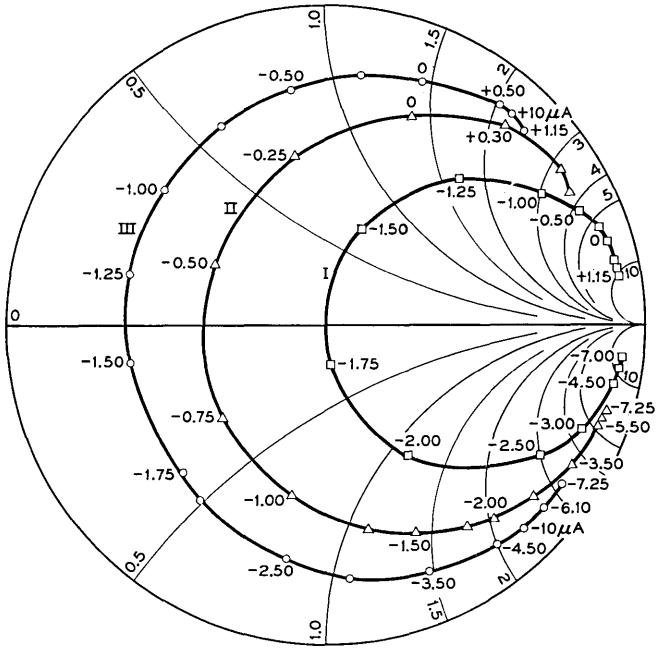


Fig. 9 — Impedance loci of a gallium-arsenide diode. DC bias voltage is applied through a 10,000-ohm series resistance.

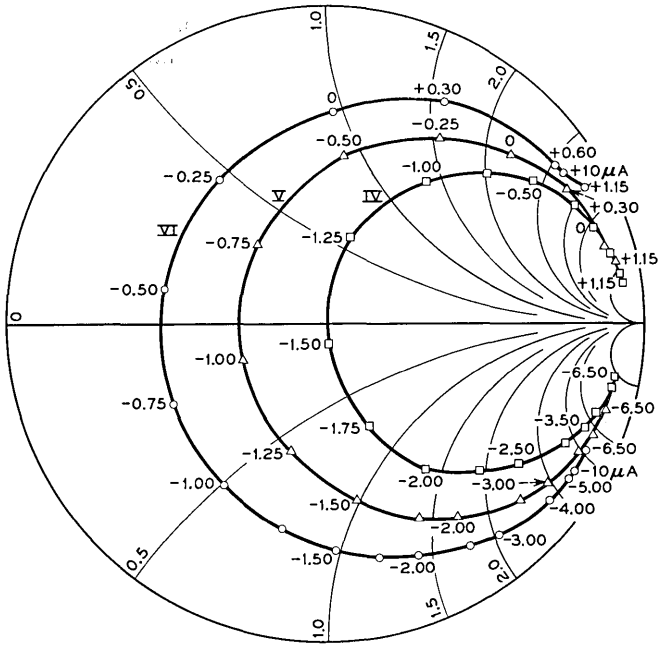


Fig. 10 — Impedance loci of a silicon mesa diode.

and 10 show two examples of such loci, one for a gallium arsenide point-contact diode and the other for a highly doped silicon mesa diode. The corresponding noise figures and the dc bias voltages for amplification are given in Table I. The accuracy of the noise figure measurements is believed to be ±0.2 db. From each impedance locus, a theoretical noise figure is calculated using (24) and (25) and is also given in the Table. The noise figure decreases as the locus moves in the direction of over-coupling. However, if it goes too far, large gain is no longer obtainable.

TABLE I — NOISE FIGURES AND OPERATION BIAS VOLTAGE

Diode Type	Locus.	Noise Figure (db)	Bias Voltage (volts)	Calculated N.F. (db)
Gallium Arsenide (※2499)	I	3.1	−1.75	3.0
	II	1.7	−1.00	1.6
	III	1.0	−2.45	0.9
Silicon (SI-R 73-62)	IV	3.1	−1.50	3.0
	V	2.1	−1.10	2.0
	VI	1.1	−0.95	1.2

This is shown in Fig. 11 for the same diode as that of Fig. 9. For comparison, the locus III of Fig. 9 is here given again. The locus VII gave a maximum gain of 15.4 db, and VIII a maximum of 5.4 db. An attempt to obtain larger gain by increasing the pump power failed. Apparently the loss of the diode increases due to the parallel conductance which comes in when the applied voltage exceeds either the contact potential in the forward bias region or the breakdown voltage in the reverse bias region. The theoretical limit on the length of the locus for the large gain condition is given by the straight lines drawn through the points $(\infty, j\infty)$ and $(0, \pm j2)$, provided that the locus is symmetrical about the zero reactance line. For a given coupling, if the locus cuts these two lines, the reactance swing can be decreased by decreasing the pump power until the locus terminates on these lines, thus giving the correct condition for large gain. If the locus does not extend to these lines, large gain cannot be achieved. The loci VII and VIII in Fig. 11 certainly correspond to this latter case.

The most interesting quantity for a diode is the best noise figure com-

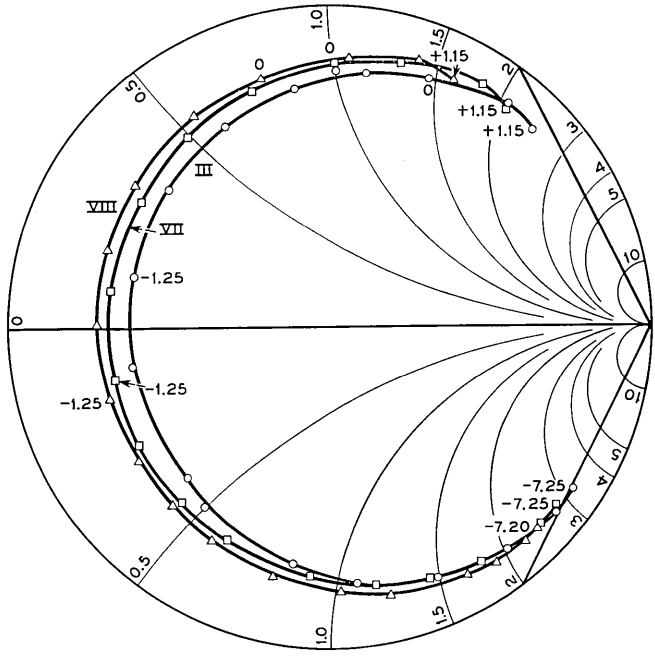


Fig. 11 — Departure from large gain condition. Locus III gives more than 20 db gain, VII 15.4 db and VIII 5.4 db.

patible with large gain. This can be found by changing the locus step-by-step towards overcoupling and measuring the noise figure for each adjustment (as has been done in Figs. 9 and 10) until one gets the optimum noise figure compatible with large gain: beyond this point either the gain becomes small (in our case less than 16 db) or the noise figure becomes worse because of microplasma and/or shot noise. Our final results for ten different diodes are plotted in Fig. 12 against the measured dynamic quality factor $\tilde{Q}$. The quoted values of $\tilde{Q}$ correspond to the reactance variation for a change of bias voltage up to the points where the dc current reaches $\pm 10\mu a$, and these $\tilde{Q}$ values are accordingly designated as $\tilde{Q}(\pm 10\mu a)$ in the Figure. This choice of $10\mu a$ for the limit is made mainly for convenience, and the question remains whether or not this choice is appropriate for all diodes. As diodes are improved, the noise contribution from the parasitic series resistance R_s becomes smaller: the microplasma and shot noise will then predominate, and one may therefore have to change this current limit to a smaller value. For the present diodes, however, the $10\mu a$ limit seems to give an appropriate guide.

VII. SOME REMARKS ON NEGLECTED FACTORS

The theory for the cold test procedure neglects the effects of higher harmonics, the parallel conductance of the diode, and circuit losses.

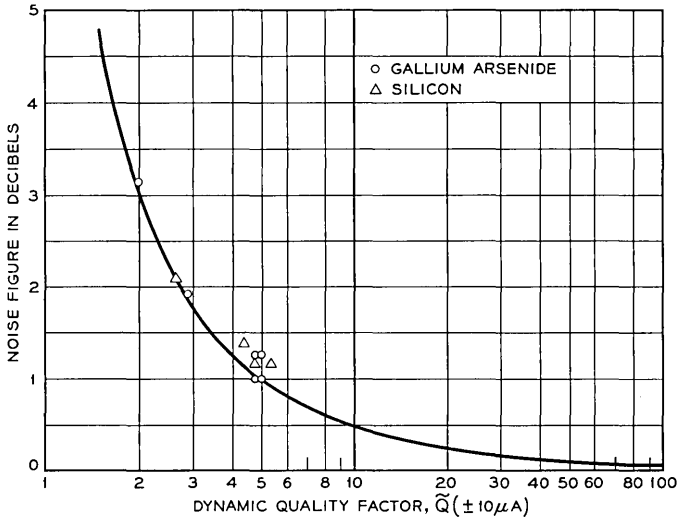


Fig. 12 — Measured noise figure versus measured dynamic quality factor Q $(\pm 10\ \mu a)$.

These effects, however, are actually noticeable in the experiments. Referring to Figs. 9 and 10, strictly speaking the impedance loci are not part of constant resistance circles: extrapolating the circles on which the loci lie, one finds that they do not touch the periphery of the Smith chart. This occurs partly because the diode has a small parallel conductance in addition to the series resistance, and partly because the circuit has finite losses. The series resistance R_s which is read directly off the Smith chart includes some contribution from the circuit itself. It is this R_s which is employed for the noise calculation in the previous section, and, hence, the noise contribution from the circuit losses is partly included in the calculated noise figure.

It will be noted that the operating bias voltage is not at the center of the locus. This is due to the effect of higher harmonics of the reactance variation. To see this clearly, the small-signal reactance of locus III is plotted against the dc bias voltage in Fig. 13. If the pump voltage across the junction is sinusoidal and the operating bias voltage is set at a zero reactance condition, the average reactance is inductive as shown schematically in Fig. 14. In actual practice, a distortion of the pump voltage takes place. As the bias is increased for the forward direction, the capacitance is increased, and, hence, the impedance is decreased. The pump voltage across the junction thus tends to decrease, making the waveform flat. Since it is possible to increase the pump power to produce the same maximum voltage as before without introducing shot noise or

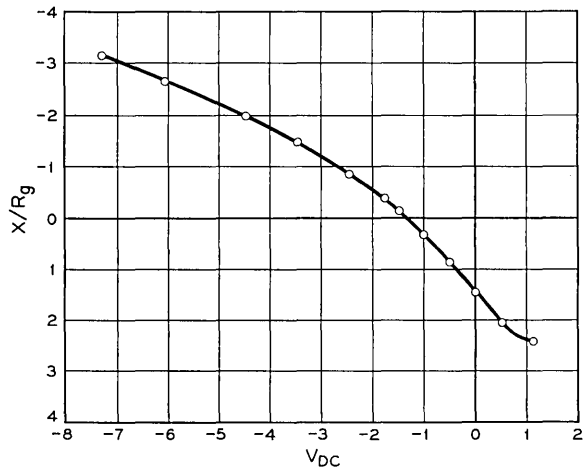


Fig. 13 — Small-signal reactance versus bias voltage (Locus III).

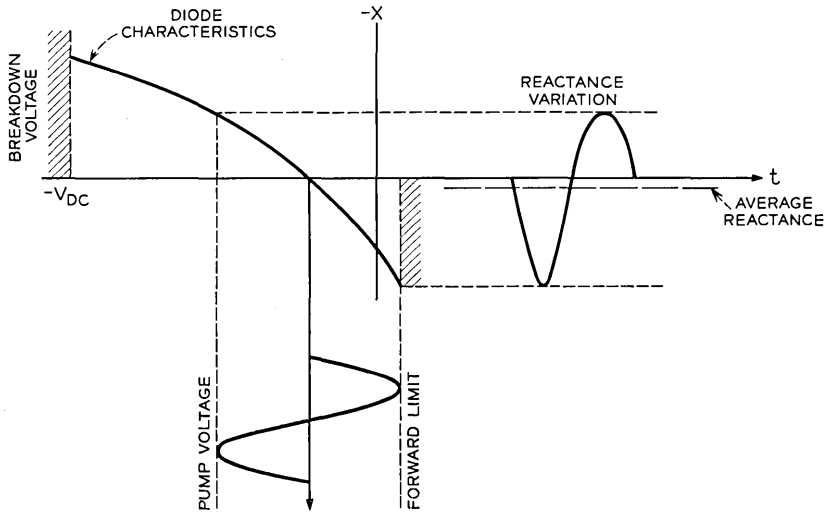


Fig. 14 — Distortion of reactance variation.

parallel conductance, the net result is an increase of the area under the reactance variation curve in the forward bias region. The opposite tendency appears in the reverse bias region, there increasing the area of the opposite reactance swing, but this is insufficient to compensate for the increase in the forward swing. The operating bias voltage must, therefore, be made further negative to cancel the additional inductance. This, then, may be the reason why the best bias voltage for amplification is always found on the capacitive side of the locus.

A rigorous theory for the minimum noise figure would necessarily include the effect of higher harmonics and of the parallel conductance. However, the fruitfulness of further efforts in this direction is doubtful, since the noise contribution which comes from the microplasma produced in avalanche breakdown is at least as great, and an analytical treatment of this is difficult. In our calculation, this is taken care of implicitly by choosing a current limit of $-10\mu a$ when evaluating $\tilde{Q}$.

In the circuit actually used in the experiments, the upper sideband impedance is not necessarily infinite. Whenever the upper sideband impedance becomes low, as indicated by the gain being smaller than that expected from the impedance locus, the shorted plunger or slide screw tuner in the pump circuit was shifted a half or one pump wavelength, thereby detuning the upper sideband without affecting the signal and

pump circuits. The upper sideband is thus expected to be adequately detuned.

VIII. CONCLUSIONS

We have discussed the principle of the cold test procedure for variable capacitance amplifiers, the adjustment of the circuit for minimum noise figure, a simplified method for visualizing the impedance locus on the Smith chart, and the experimental results on a 6-kmc degenerate amplifier. It has been shown that all the important circuit parameters of the equivalent circuit of the parametric amplifier can be evaluated from the cold test result, making the adjustment of the circuit for minimum noise figure relatively straightforward.

For the calculation of the dynamic quality factor, it is suggested that a limiting value for the dc bias current be chosen of $\pm 10\mu a$.

Despite the neglect in the theory of higher harmonics, parallel conductance, and circuit losses, the correlation between the measured noise figures and those calculated from the cold test results is very good, in fact, surprisingly so.

Brief consideration has also been given to certain factors neglected in the simple theory. The best operating bias voltage is always found on the negative side of the locus center, and this is explained qualitatively by the harmonics of the reactance variation.

Acknowledgments are due to H. Seidel, K. D. Bowers and M. Uenohara for their stimulating and helpful criticism and encouragement during this work. The author also wishes to thank H. H. Lehner and B. H. Johnson who performed the experiments.

APPENDIX A

Equivalent Circuit of Lossless Two-Terminal-Pair Network with a Transmission Line[3]

Consider a lossless two-terminal-pair network with a transmission line on the left-hand side of it. The reflection coefficient at an arbitrary plane on the transmission line is a bilinear function of the impedance Z connected to the right-hand side of the two-terminal-pair network. There are reference planes where the reflection coefficient becomes unity when Z approaches infinity, since in this case the circuit has no losses at all. Take one of such reference planes and consider the reflection coefficient at this plane with an arbitrary impedance Z. It can be expressed

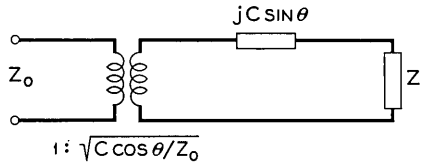


Fig. 15 — Equivalent circuit for a lossless two-terminal-pair network and a certain length of transmission line.

in the form

$$r = \frac{Z + C_1}{Z + C_2} \tag{27}$$

where C_1 and C_2 are complex quantities, because this is the most general bilinear expression for Z for which r becomes unity when Z is infinite. Suppose $Z = jX$ (pure imaginary), then $|r| = 1$, since the circuit is again lossless. Therefore, we have

$$\left| \frac{jX + C_1}{jX + C_2} \right| = 1 \tag{28}$$

for arbitrary real X's. From this relation, we obtain

$$\mathrm{Im}C_1 = \mathrm{Im}C_2 \tag{29}$$

$$|C_1|^2 = |C_2|^2. \tag{30}$$

Therefore, using real quantities C and θ, C_1 and C_2 can be expressed in the form

$$C_1 = -Ce^{-j\theta} \tag{31}$$

$$C_2 = Ce^{j\theta}. \tag{32}$$

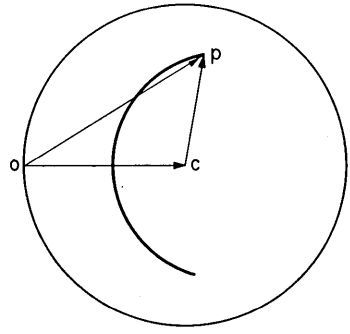


Fig. 16 — Relation between impedance and voltage on Smith chart.

Equation (27) then becomes

$$r = \frac{Z - Ce^{-j\theta}}{Z + Ce^{j\theta}} = \frac{Z + jC \sin \theta - C \cos \theta}{Z + jC \sin \theta + C \cos \theta}. \tag{33}$$

Comparison of this equation and the standard formula for the reflection coefficient in terms of impedance indicates that the same reflection would be given by the circuit shown in Fig. 15. Since Z is arbitrary, this means that a lossless two-terminal-pair network with a certain length of transmission line attached to one side of it is always equivalent to a combination of an ideal transformer and a reactance.

APPENDIX B

Relation Between Impedance and Voltage on Smith Chart

The Smith chart is a reflection coefficient plane in which the constant resistance curves and the constant reactance curves are mapped. Suppose an impedance at a reference plane is given by a point p in the Smith chart shown in Fig. 16, then the vector $\overrightarrow{cp}$ represents the corresponding reflection coefficient, provided that the radius of the Smith chart is unity. The voltage at the reference plane is therefore given by the length of the vector $\overrightarrow{op}$, i.e., a sum of the incident wave $\overrightarrow{oc}$ and the reflected wave $\overrightarrow{cp}$ provided that the incident wave voltage is unity. Keeping the incident wave voltage constant, the change in the length of the vector $\overrightarrow{op}$ is then equivalent to the change in the voltage at the reference plane. If one moves the reference plane k wavelengths towards the load, then the whole impedance locus rotates anticlockwise k wavelengths. However, as far as the length of the vector $\overrightarrow{op}$ is concerned, the same relative effect is obtained when the locus stands still and the point o rotates k wavelengths clockwise (in the direction of the arrow indicating "wavelength toward generator"). Therefore, depending on the probe position of the standing-wave detector, the oscilloscope pattern changes in the manner discussed in the text.

REFERENCES

1. Kurokawa, K., and Uenohara, M., B.S.T.J., **40**, May, 1961, p. 695.
2. Uenohara, M., Proc. I.R.E., **48**, Feb., 1960, p. 169.
3. Marcuvitz, N., *Waveguide Handbook*, Radiation Laboratories Series, **10**, p. 123.

Contributors to This Issue

M. R. AARON, B.S. in E.E., 1949 and M.S. in E.E., 1951, University of Pennsylvania; Bell Telephone Laboratories, 1951—. He first worked on analysis, design and synthesis of transmission networks for L3 and submarine cable systems. From 1954 to 1956 he supervised a group concerned with design of networks for the L3 system. Since 1956 he has been in charge of a group engaged in systems analysis of PCM. Member I.R.E.

R. D. BARNARD, B.E.E., 1952, and M.E.E., 1955, Polytechnic Institute of Brooklyn; Ph.D., 1959, Case Institute of Technology; Bell Telephone Laboratories, 1959–61; faculty, Wayne State University, 1961—. At the Laboratories he was primarily concerned with theoretical problems in communications and control. Member I.R.E., American Physical Society, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

VÁCLAV E. BENEŠ, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. Member American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, Mind Association, Phi Beta Kappa.

FRANKLIN H. BLECHER, B.E.E., 1949, M.E.E., 1950, and D.E.E., 1955, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1952—. His early work concerned the design of transistor circuits for application in analog and digital computers; design of wideband transistor feedback amplifiers for application in carrier systems; and development of active filters, IF amplifiers, and wideband video amplifiers. He later headed a group engaged in the development of solid-state and short-haul carrier circuits, and millimeter wave networks. This included the area of solid-state circuits for active communications satellites. Since May, 1961, he has been director of the Carrier Transmission Laboratory. Member I.R.E., Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

CLAUDE G. DAVIS, B.S. in E.E., 1950, Case Institute of Technology; M.S. in Mathematics, 1960, Stevens Institute of Technology; Bell Telephone Laboratories, 1950—. He has specialized in transmission systems development, including the development of armorless submarine cable for a transoceanic telephone system, a PCM system for exchange trunks, PCM repeaters for an experimental waveguide transmission system, and the time assignment speech interpolation (TASI) system. He is currently responsible for groups concerned with satellite repeater design and data analysis. Member Eta Kappa Nu.

JAMES R. GRAY, B.S. in E.E., 1954, and M.S.E., 1955, University of Florida; Bell Telephone Laboratories, 1955—. He first engaged in repeater design for pulse code modulation systems. Since 1958 he has concentrated on PCM transmission impairment studies.

F. J. HALLENBECK, E.E., 1936, Polytechnic Institute of Brooklyn; Western Electric Co., 1923–25; Bell Telephone Laboratories, 1925—. For many years he was involved in the development of transmission networks for Bell System and military communication facilities. He later supervised a group engaged in the development of broadband carrier systems. In 1958 he assumed responsibility for L-carrier terminal development. Senior member I.R.E.; member A.I.E.E., Tau Beta Pi, Eta Kappa Nu.

KANEYUKI KUROKAWA, B.S., 1951, and Dr. of Eng., 1958, University of Tokyo; Bell Telephone Laboratories, 1959–61; University of Tokyo, 1961—. Dr. Kurokawa was engaged in research on low-noise amplifiers at Bell Laboratories for two years, during which time he was on leave of absence from his position as assistant professor at the University of Tokyo. Member I.R.E., Institute of Electrical Engineers (Japan), Institute of Electrical Communication Engineers (Japan).

HENRY MANN, B.A., 1950, Brooklyn College; M.S. in E.E., 1955, Columbia University; Bell Telephone Laboratories, 1954—. His work has included the design of the synchronizing circuits, demultiplex gate, and portions of the encoder for the experimental pulse code modulation system. He also engaged in the development of a system for the transmission of two PCM groups over short-haul microwave carrier circuits. He is presently responsible for the design of a command decoder for an experimental active satellite communications system. Member I.R.E., Pi Mu Epsilon.

JOHN S. MAYO, B.S. in E.E., 1952; M.S. in E.E., 1953; Ph.D. in E.E., 1955; North Carolina State College; Bell Telephone Laboratories 1955—. He first engaged in computer research, including studies relating to the use of digital computers for measurement and automatic tracking of pulsed radar range information, and in military weapons control systems. His recent work has involved the development of line repeaters for an exchange carrier PCM system, and high-speed PCM terminals for an experimental waveguide transmission system. He has been in charge of the PCM Transmission Department since December, 1960. Member I.R.E., Sigma Xi.

STEPHEN O. RICE, B.S., 1929, Oregon State College; Graduate Studies, California Institute of Technology, 1929–30 and 1934–35; Bell Telephone Laboratories, 1930—. In his first years at the Laboratories, Mr. Rice was concerned with nonlinear circuit theory, especially with methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigation of telephone transmission theory, including noise theory, and applications of electromagnetic theory. He was a Gordon McKay Visiting Lecturer in applied physics at Harvard University for the Spring, 1958 term. Fellow I.R.E.

R. H. SHENNUM, B.S. in E.E., 1944, and M.S. in E.E., 1948, Montana State College; Ph.D., 1954, California Institute of Technology; Bell Telephone Laboratories, 1954—. He first worked on the design of microwave parts for the TJ microwave system. Later he was responsible for companding, signaling and voice-frequency circuit development, and field experiments, for an exchange carrier PCM system. Currently, as head of the Satellite Design Department, he is responsible for the development of the electronic system for the active satellite for an experimental satellite communications system. Member A.I.E.E., Sigma Xi, Tau Beta Pi, Phi Kappa Phi.

HAROLD M. STRAUBE, B.S. in E.E., 1939, University of Michigan; Northwestern University, 1939–41; Bell Telephone Laboratories, 1941–60; R.C.A., 1960—. His early work at the Laboratories concerned the design and development of panoramic receivers, underwater range devices, radar test equipment, and television test equipment. He later engaged in electronic-switching research, transmission research, and work on high-frequency transistor circuits, and subsequently was involved in various developmental aspects of an experimental PCM ex-

change carrier system, particularly the design and evaluation of a compandor. Mr. Straube is presently concerned with the development of digital communication systems at R.C.A. Senior member, I.R.E.; member Sigma Xi.

LAJOS F. TAKÁCS, Doctor's degree, 1948, University of Technical and Economical Sciences, Budapest; Doctor of Mathematical Sciences, 1957, Hungarian Academy of Sciences; Tungsram Research Laboratory (Telecommunications Research Institute), Budapest, 1945–55; Research Institute for Mathematics of the Hungarian Academy of Sciences, 1950–58; Roland Eötvös University, Budapest, 1953–58; Columbia University, 1959—; consultant, Bell Telephone Laboratories, 1959—. At present he is teaching probability theory and stochastic processes, and he is engaged in research in the mathematical theory of telephone traffic. Member American Mathematical Society, Mathematical Association of America, Society for Industrial and Applied Mathematics, Institute of Mathematical Statistics, Sigma Xi.

CLAUDE P. VILLARS, M.S. and Ph.D., Swiss Federal Institute of Technology, Zurich; Bell Telephone Laboratories, 1954–60. At the Laboratories, Mr. Villars was responsible for the design of the coder for the experimental exchange carrier pulse code modulation system. Member I.R.E., Institute of Electrical Engineers (Great Britain).