THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 54 February 1975 Number 2

Echo Performance of Toll Telephone Connections in the United States

By F. P. DUFFY, G. K. McNEES, I. NÅSELL,
and T. W. THATCHER, JR.

(Manuscript received June 18, 1974)

A field survey to characterize echo performance of toll telephone connections was conducted in 1972. Information on echo path loss and echo path delay for talker echoes was obtained from a sample of nearly 1700 connections in the continental United States. This paper discusses the survey data acquisition techniques, the sample design, and the statistical results. A major result of the survey was the determination that echo path delay is significantly less than previously estimated. For the longest connections (2700 miles or 4345 km), the median round-trip echo delay is 45 ms, 11 ms less than previously calculated from the sum of connection segments.

I. INTRODUCTION

Echo may be experienced by talkers on long telephone connections when conditions exist analogous to those producing acoustic echoes, i.e., a two-way transmission path, a point of reflection, a perceptible time delay between transmission and reception, and received energy of sufficient amplitude to be detected. In the presence of a loud, long-delayed echo, whether acoustic or telephonic, conversation is likely to be difficult. Figure 1 is a simplified representation of a long-distance telephone connection with two-wire loops, four-wire trunk, and the hybrids (H) and balancing networks used in joining two-wire and four-wire circuits. The hybrids are the principal points of reflection in the telephone network. When a hybrid is perfectly balanced, none of

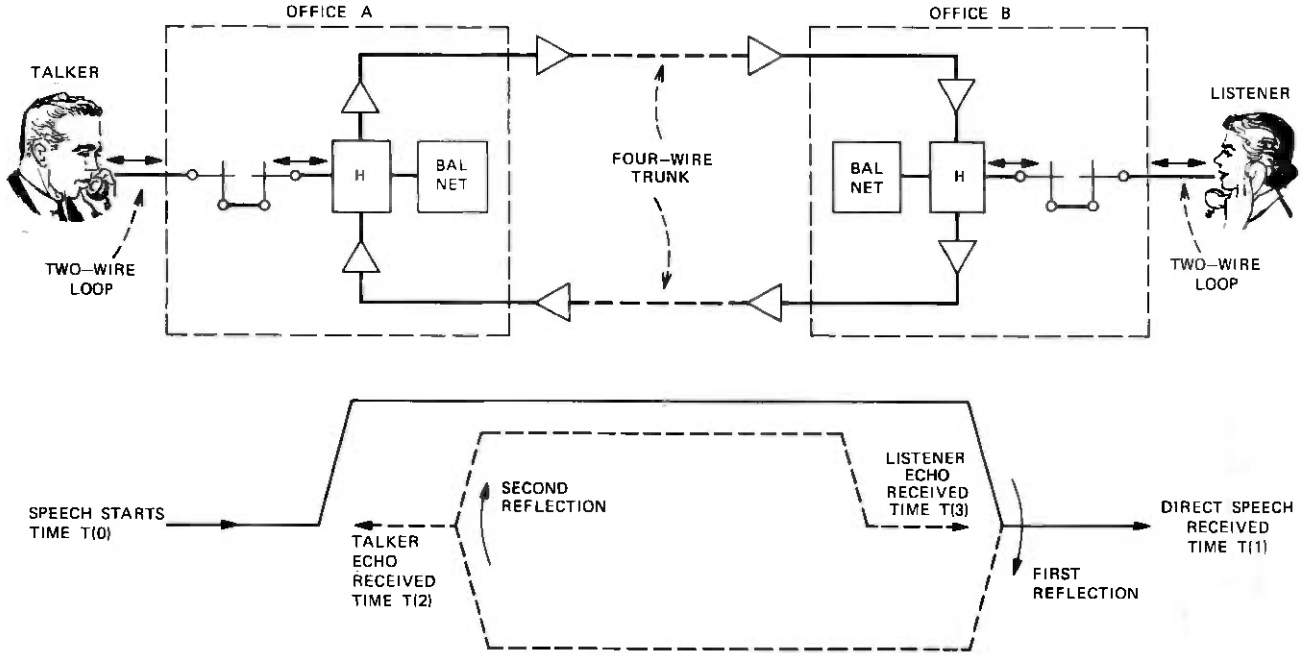


Fig. 1—Echo paths in a simple connection.

the energy from the receive pair of the four-wire path passes to the transmit pair. Since many different two-wire circuits, presenting a range of impedances, may be switched to a trunk while its associated balancing network remains fixed, some energy, which varies in amount from connection to connection, may be returned to the talker. Figure 1 also shows the direct and echo speech paths and their relative delay times. Only talker echoes are discussed here, since listener echoes are not significant when talker echoes are controlled to acceptable magnitudes.

Control of echo has been a concern of telephone engineers since long-distance telephone connections were made practicable by the introduction of low-distortion gain devices. The factors causing telephone echoes and most of the methods used since then to control echo were discussed by A. B. Clark in 1923.[1] These factors included the tolerance of talkers to echo as a function of echo amplitude and echo delay, the velocity of propagation of facilities, the degree of control of reflected signals at reflection points, and the choice of trunk losses to insure acceptable direct speech amplitudes while keeping echo amplitudes low. Another method of controlling echo that is presently used was soon added, the installation of echo suppressors on trunks having long echo delays to open the echo return path when speech is present in the direct path.[2] These measures, as appropriate, were applied to long toll trunks to control echo (from longer delayed reflections) and to short toll trunks to control singing and near-singing distortion (from shorter delayed reflections). The toll network trunking plan generally required only one or two short trunks and one long trunk to establish any long-distance connection. These trunk design methods for echo and singing control were continued from the 1920s through the 1940s, though with the passage of time knowledge of subscriber preferences was refined, impedance balancing was improved, new echo suppressors were developed, and carrier-type toll transmission facilities having propagation velocities approaching that of light were placed in service. These improvements permitted substantial reductions in the overall losses of long-distance connections.

A major change in echo design of toll trunks occurred in the late 1940s and early 1950s in conjunction with the change from operator to machine switching of toll calls, changes in the trunking plan allowing automatic alternate routing, and an increase to 7 in the maximum number of toll trunks in a long-distance connection. The Via Net Loss (VNL) plan was developed and implemented to assure acceptable echo performance on connections involving a few or many trunks, to provide low overall connection losses, and to avoid more than one echo suppressor on a long connection.[3,4]

In more recent years, attention has been given to new echo problems. The very long delayed echoes resulting from the transmission paths provided by communication satellites require changes in echo control measures.[6] The extension of speech transmission on digital facilities to greater and greater distances also will require new echo control measures. The T1 digital short-haul carrier system was introduced in the early 1960s, and its use continues to grow rapidly.[6] Long-haul digital trunk transmission systems are being developed.[7] The No. 4 ESS Toll Switching System will electronically switch digital bit streams to effect circuit switching.[8] Such digital arrangements will not easily permit adjustment of direct transmission loss on a trunk-by-trunk basis as specified by the VNL plan, and so require development of alternate methods of echo control.

Information was desired on the echo performance of the existing switched telephone network to provide an improved data base for echo control studies and planning, both for improvement of the present network and for evaluation of echo control measures proposed for digital networks. Information also was desired on whether there had been changes in subscriber reactions to echoes after some years of experience with low-loss long-distance connections.

This paper reports on the testing methods and results of a field survey to characterize the echo performance of the public switched-telephone network by making observations on a large sample of long-distance calls placed between many locations throughout the continental United States. The information obtained has been used to update mathematical models of echo performance of the telephone network. These models are being used in a variety of studies to evaluate results of changes proposed for the network. The echo survey disclosed that round-trip echo delays on the longer distance connections were shorter than had been predicted by older models. As a result of this and other information from the survey, the trunk lengths at which echo suppressors are installed have been increased, and significant cost savings are anticipated.

In this survey, observations were made on long-distance telephone connections extending from the local switching offices visited during the survey to distant called subscriber stations. The switching offices to be visited were selected using sampling techniques, copies of billing records were obtained, the billed calls were stratified by length into four mileage bands, and called numbers were randomly selected for the survey in each mileage band. Thus, the echo test calls repeated telephone calls previously made from the sampled offices. Figure 2 lists the sampled central office locations and shows the route traveled between locations.

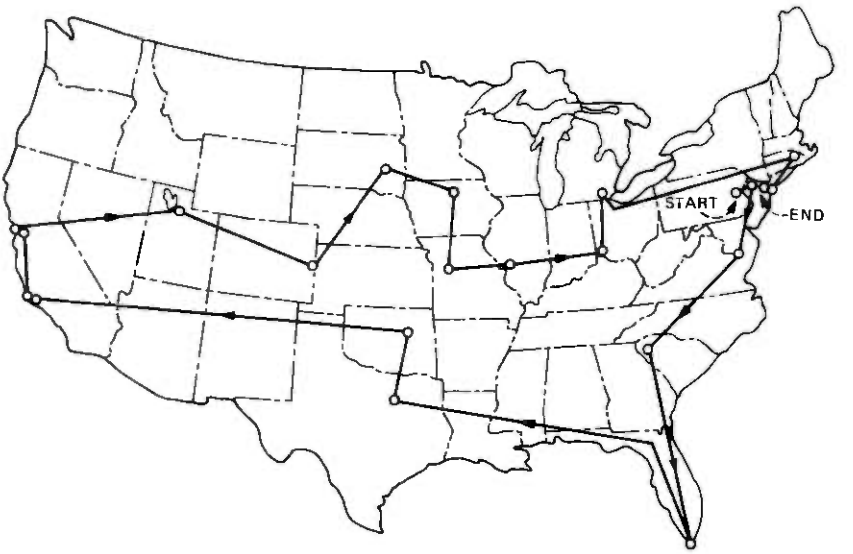
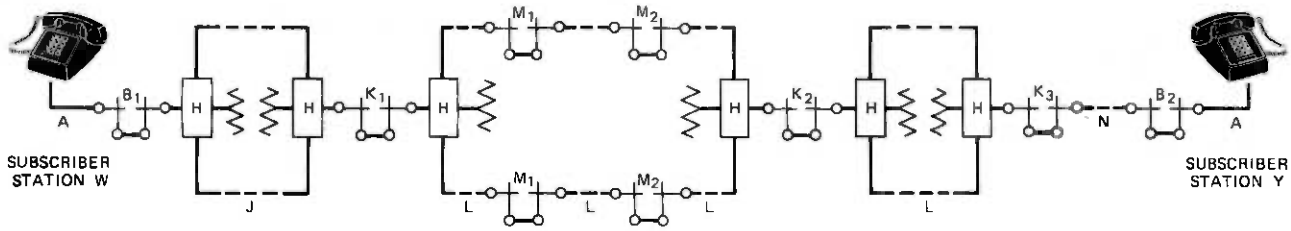


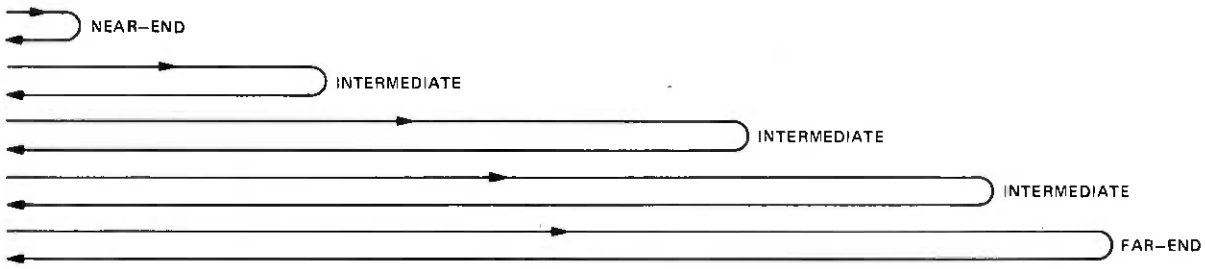
Fig. 2—Echo survey tour route and cities visited.

Prior echo measurements[9] had been made only on portions of connection echo paths, e.g., trunk transmission facilities, trunk terminating equipment, and the return loss between trunks and loops and between trunks and other trunks. Statistical modeling techniques were used to derive echo path loss and echo path delay distributions for overall connections. The availability of new measuring techniques, minicomputers, and processing software made possible this first survey in which actual echo path loss and delay were determined on calling-end office to called-subscriber connections, with only the station and loop from the calling subscriber to his serving local office being excluded. Information is available from a loop survey[10] to refer echo path loss to the originating station, if desired.

Figure 3 shows schematically the subscriber stations, loops, switching offices, and trunks comprising a possible long-distance telephone connection. The figure also indicates locations where signals may be reflected back into a path leading to the point of origin, causing talker echoes. The figure shows only those talker echoes heard by subscriber W. However, since telephone connections have the same general



ECHO PATHS AND REFLECTION LOCATIONS



CONNECTION ELEMENTS

A–LOOP
B–LOCAL SWITCHING OFFICE
H–TWO–WIRE TO FOUR–WIRE JUNCTION (HYBRID)
J–FOUR–WIRE TOLL CONNECTING TRUNK
K–TOLL SWITCHING OFFICE, TWO–WIRE
L–FOUR–WIRE INTERTOLL TRUNK
M–TOLL SWITCHING OFFICE, FOUR–WIRE
N–TWO–WIRE TOLL CONNECTING TRUNK

Fig. 3—A possible toll connection and subscriber W talker echo paths.

structure at both ends, similar echoes of his own speech from reflection points throughout the connection (talker echoes) may be heard by subscriber Y. The characteristics of echo paths extending from the local office (B) adjacent to subscriber W to the distant end and back, shown as the far-end echo in Fig. 3, are reported here. The specific characteristics determined for each connection were the round-trip echo path loss and echo path delay.

In the following sections, the survey sampling plan is presented, the measuring technique and instrumentation are described, and the survey results are presented and discussed.

II. SAMPLE DESIGN

The population about which information was desired was the set of toll calls originating in Bell System end offices and spanning an airline distance between originating and terminating local offices of 180 miles or more. Both originating and terminating ends of toll calls were confined to the continental United States. Toll calls shorter than 180 miles were excluded because the echo delay on these calls is short, and such short echoes are seldom perceived by subscribers.

The sampling plan used to select specific toll connections for field testing was a two-stage plan with primary stratification and substratification. The two-stage plan adopted has the advantage of limiting the number of locations to be visited for survey measurements. Bell System end-office buildings were identified as primary units of the sampling plan. Ten primary strata were formed. Each stratum was identified with the set of Bell System end-office buildings located in the area served by one of the ten regional centers in the DDD (direct-distance-dialing) network of the continental United States. The first-stage sample of primary units contained a total of 22 end-office buildings. Four primary units were selected from the White Plains region, and two were selected from each of the remaining nine regions. The first-stage sampling was made with probabilities proportional to estimates of size. The size of a primary unit was defined by the total number of outgoing toll calls based on billing records from the 1966 Message Minute Mile Study[11] and the 1964 Wire Center Study.[12]

The subjective effect of an echo is related to its delay, which is correlated with distance. This distance dependence influenced the structure of the sampling plan and resulted in the use of substratification. Four subclasses of toll calls were identified for data analysis purposes. They were defined on the basis of the airline distance between originating and terminating central offices. The four mileage bands were 180–360 miles (290–580 km), 360–725 miles (580–1167 km), 725–1450 miles (1167–2333 km), and 1450–2900 miles (2333–4667 km). The

purpose of the substratification was to give a sample of approximately equal size in each of these mileage bands. Four substrata were defined so that they approximately coincided with the four subclasses defined above. For convenience in the establishment of the second-stage sampling frames, the substrata were defined in terms of the numbering plan area (NPA) of the terminating end of a call rather than in terms of the exact distance between the end offices. Thus, if an NPA fell entirely inside one of the four mileage bands (measured from the originating end of the call), then all calls terminating in that NPA were referred to the corresponding substratum. If an NPA straddled the boundary between two mileage bands, then calls in that NPA were referred to the substratum that corresponds to the mileage band in which the majority of calls into the NPA were expected to terminate. This arbitrariness in the substratum definition does not affect the ability to analyze data with reference to each of the four subclasses defined. It has the advantage of avoiding the computation of the exact airline distance between originating and terminating central offices for the large number of calls listed in the second-stage frames.

Lists of outgoing toll traffic during one or more days were acquired from each of the 22 end office buildings comprising the first-stage sample. The lists covered one day's traffic for large offices and two or more days' traffic for small offices. The number of days was adjusted to give a sufficiently large listing of long toll calls. The substratification indicated above was imposed on each of the 22 lists. The second-stage sample of calls to be tested in the survey then was selected by simple random sampling. Independent selections were made in each substratum of each primary unit in the sample. Each second-stage sample element was identified with the telephone number of a called customer. The sample size was determined in such a way that the sample was approximately self-weighting in each of the four substrata (all observations contribute equally in calculating the statistical estimates within a mileage band). This self-weighting feature extended across all primary units within a specific substratum.

The sample size was determined on the basis of precision requirements and variance estimates. The precision requirements took the form of a maximum width of ± 1 dB for the 90-percent confidence interval of the mean echo path loss in each of the four mileage bands. Available data on variance components for the echo path loss were then used to derive the sample size. Successful transmission tests were completed on a sample of 1681 connections. Of these, 393 were in the first mileage band (180–360 miles), 470 in the second, 411 in the third, and 407 in the fourth mileage band.

All estimates given in Section V refer to the population defined above. The statistical estimation procedures used to derive the results were the appropriate ones for multistage-structured-sample surveys.[13]

III. SURVEY INSTRUMENTATION

In planning the instrumentation of this field survey, goals were: (*i*) the measurement results will accurately represent the field conditions, (*ii*) the test equipment will perform reliably during use and after repeated shut-downs, moves, and start-ups at new test sites, (*iii*) the equipment design will permit its operation and relocation by technical staff personnel without requiring excessive time for training and hands-on experience, (*iv*) the operation will require a minimum number of persons, (*v*) the overall equipment operation will be monitored by built-in self-checking and operator-checking features, (*vi*) the output data will be in a form that will simplify subsequent processing and use, and (*vii*) the field travel and expense will be minimized. These goals were generally met. The methods used are briefly described in the next five paragraphs; greater detail is given in subsequent paragraphs.

The heart of the echo test set was a minicomputer. Software programs directed the testing sequences of translating stored test signals from digital to analog form for application to the sampled connections, and translation of the applied and echo return signals to digital form for recording on magnetic tape. These digital/time domain results were processed by a fast Fourier transform (FFT) program in the computer and translated to the frequency domain. Further processing (division by the transform of the transmitted signal) gave the frequency response of the entire connection echo path. The FFT was then used to translate to the time domain, giving the impulse response of the connection with the echoes separated in time. This permitted identification of the echo of interest, which then was transformed back to the frequency domain, giving the wanted output parameters-echo path loss and echo path envelope delay versus frequency for the selected far-end echo. The accuracy of the test set is determined by the gain in its input path and by the A/D (analog-to-digital) converter step size. Referred to the central office loop input, the quantizing noise from the A/D converter digital sampling is -84.6 dBm, or 5.4 dBrn, well below the telephone line noise. Thus, echo path loss values are bounded by telephone line circuit noise, not by test equipment characteristics.

The echo path test set was installed in a small van that was driven from site to site and parked by the sampled central office buildings.

Alternating-current power and telephone line connections were made between the van and the central office buildings. Setup, checkout, and calibration of the test equipment required only an hour or two. Figure 4 shows the equipment lineup within the test van.

The test set was put in operation by mounting program and data magnetic tapes and loading four command words or conditions into the computer via the computer console switches. Subsequent operational commands were entered via pushbuttons that lighted to indicate test status or available choice options. The software programs were written to provide checks of the steps involving operator actions, to provide automatic multiple attempts when tape reading errors were encountered, and to permit returns to the start of sequences in case of operator error. These arrangements permitted single-operator operation after a short training period. However, two-man test teams operated in the field on overlapping two-week assignments to provide continuity of testing throughout the day, to take care of other than measurement details, and to provide guided hands-on experience.

In operation, the called telephone number was dialed by a repertory dialer, there was a short conversation with the answering party in which the test was explained, the telephone number verified, and cooperation obtained. The subscriber was asked to cover the telephone transmitter with the palm of his hand to reduce room noise interference,

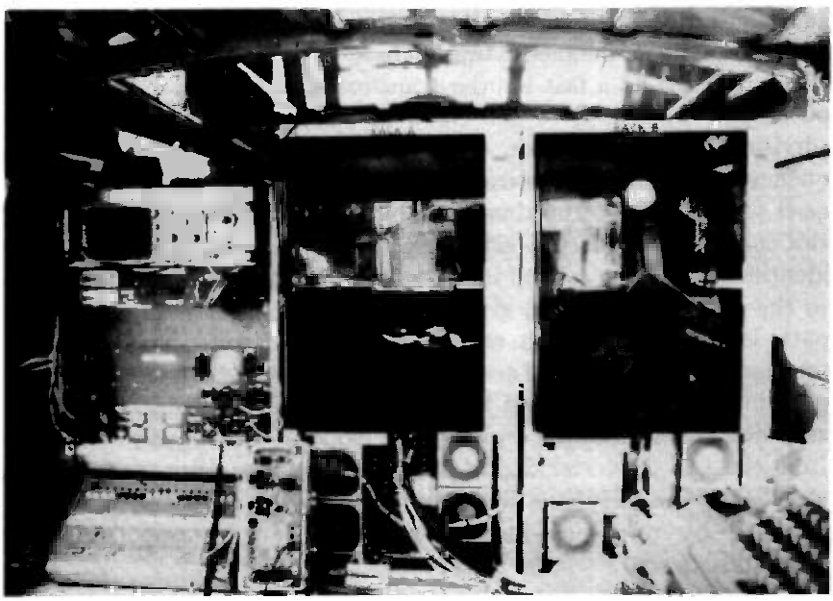


Fig. 4—Interior of test van.

the test tones were transmitted, the subscriber was thanked, and the call was terminated.

The echo path field data were recorded on magnetic tapes in standard data format to be compatible with large computers on which final processing was carried out after completion of the field portion of the survey.

3.1 Programmable test set

This section describes the hardware and software of the test set and its capabilities, which include analysis of observations and display of echo path loss and phase versus frequency. The principal information flow to and from the test set minicomputer is via digital-to-analog (D/A) and A/D converters. A block diagram of the test set is shown in Fig. 5. One magnetic tape unit is used to read stored programs and the other to record raw and processed data. The quantized interrogation signals stored in the computer memory as PCM binary words (part of the test program) are translated into a stair-step signal by the D/A converter and passed to the output low-pass filter (LPF), which reconstructs the original interrogation signals. Following the LPF is an attenuator used for setting the signal power transmitted to the local switching office. The talk-test switch connects the hybrid either to a dial/talk circuit for dialing a connection and talking to the subscriber or to the interrogation signal source. The hybrid is used to interconnect

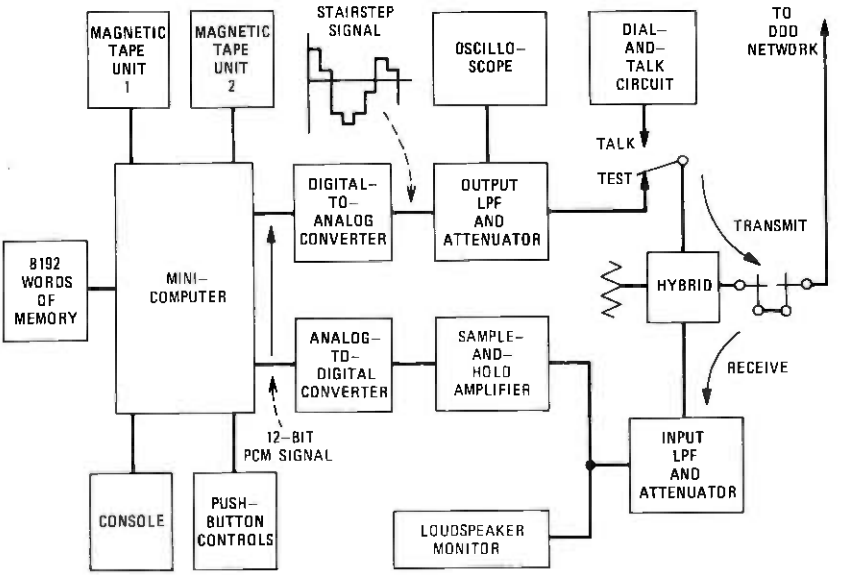


Fig. 5—Echo path test set.

the four-wire test set transmit-and-receive paths to the two-wire central office loop. In the receive path, the monitoring loudspeaker permits test personnel to verify transmission of the test signals. The LPF at the input to the sample-and-hold (S&H) amplifier is used to filter out frequencies higher than half the sampling rate. The S&H amplifier samples the incoming signal and holds the sample value constant while its A/D conversion is taking place.

The operator controls the test set by pushing buttons on the control panel to direct the test system to perform the various test or analysis operations. The control panel is also used to load numerical data in the test set and to display entered numbers and results of computations on a light-emitting diode (LED) display. The oscilloscope displays the transmitted signal and, following analyzation, the impulse, amplitude, or phase responses via the D/A converter.

As directed by the operator's pushbuttons, the computer's central processor unit executes various operational programs that are stored on magnetic tape and in core memory. These programs are overlaid in core memory from tape as they are needed by a monitor program[14] that always resides in the core memory.

3.2 Measurement of network impulse response by deterministic source interrogation signal

A number of methods were investigated for determining the impulse response of networks. The one chosen for the echo survey used a deterministic signal comprised of a finite number of evenly spaced frequency components spanning the frequency band of interest, whose phases were specified to obtain a minimum signal amplitude peak-to-average ratio.

The impulse response, $h(t)$, of a network can be found by first calculating its frequency response, $H(\omega)$, and then computing the inverse Fourier transform of $H(\omega)$.

$$h(t) = F^{-1}[H(\omega)].$$

A sample value of the frequency response can be calculated by applying a sine wave to the input of the network, determining the resulting output, and dividing the output by the input. Thus, if $X(\omega_i)$ is the input signal at frequency ω_i and $Y(\omega_i)$ is the output signal resulting from this input, then $H(\omega_i)$, the network response at ω_i, is

$$H(\omega_i) = Y(\omega_i)/X(\omega_i).$$

For a linear time-invariant network (essentially attained by the

testing method), the principle of superposition holds and the observations can be made simultaneously at all frequencies ω_i of interest.

The period of the lowest-frequency sine wave used was made to exceed the maximum expected delay of the network to avoid the ambiguity caused by one cycle of a sine wave being indistinguishable from the next. Since the maximum expected round-trip delay, based on models of the telephone network, was less than 60 ms, the test signal was designed to have a period of about 100 ms. This allows for a 40-ms duration of the impulse response of an echo delayed 60 ms. It follows, from the frequency sampling theorem,[15] that a signal essentially time-limited to 100 ms is completely specified by samples every 10 Hz. These samples in the frequency spectrum should cover the entire spectrum of interest, 200 to 3400 Hz, plus an additional upper band in which the energy can be reduced to zero using realizable filtering techniques. The transmitted interrogation signal, $x(t)$, was made up by summing 390 sine waves from about 10 Hz up to about 3800 Hz spaced approximately 10 Hz apart so that

$$x(t) = C \sum_{i=1}^{390} \cos [\omega_i t + \phi_i].$$

The amplitude distribution of the resulting interrogation signal $x(t)$ depends on the relative phases ϕ_i chosen for the component sine waves. This waveform can range from a very peaked impulse, when all the components are in phase, to a signal that has a relatively low peak-to-rms ratio for certain other phase relationships. Since there are many devices in the echo path that could overload and cause nonlinear distortion, such as amplifiers and syllabic compressors, a signal that has the least peak-to-rms ratio is desirable. When the phases of the component sine waves are proportional to the square of their frequencies, the peak-to-rms ratio is minimized.[16] The phases, ϕ_i, in the interrogation signal are given by

$$\phi_i = i^2/390.$$

The resulting sum of all the components is a good approximation to frequency modulation of a carrier with a sawtooth waveform. At the beginning of the approximately 100-ms sweep period, the energy is centered around 3800 Hz and linearly decreases in frequency with time to around 10 Hz at the end of a period. Figure 6 shows the waveform of one period of the interrogation signal.

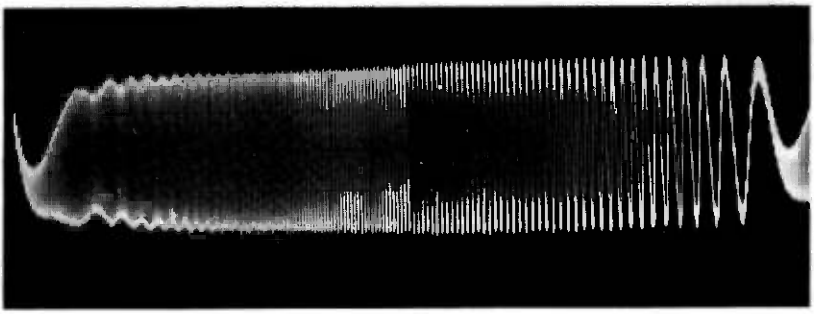


Fig. 6—One 102.4-ms period of the interrogation test signal.

IV. DATA ACQUISITION AND PROCESSING

4.1 Examples of echo path characteristics

The impulse response of an overall long-distance connection includes the impulse response of the near-end and intermediate paths as well as that of the desired far-end echo path. The occurrence of such echoes is depicted in Fig. 3. To obtain the wanted far-end echo path characteristics, the energy reflected from the far end must be separated from that reflected from near-end and intermediate discontinuities. This can be done only if the reflections are sufficiently separated in time.

Figure 7a shows the impulse response of an actual connection in which near-end and far-end echoes were the significant contributing

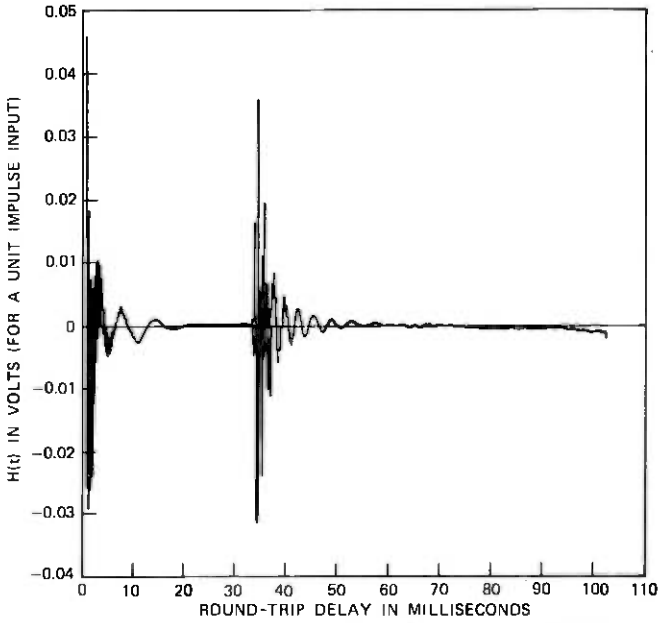


Fig. 7a—Impulse response of a telephone connection.

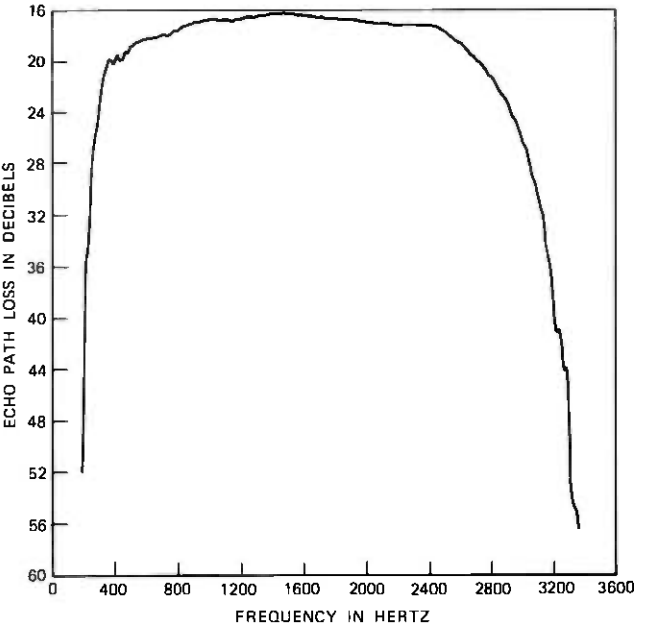


Fig. 7b—Far-end echo path loss of a telephone connection versus frequency;

elements. To obtain the echo path characteristics of the far-end echo alone, the amplitude values of the first impulse response (from the near-end discontinuity, 0 to 30 ms) are set equal to zero. The inverse transform is then taken, giving the amplitude and delay characteristics of the second or far-end impulse response that characterize the echo path. Figure 7b shows the far-end echo path amplitude response and Fig. 7c the envelope delay response (the derivative of the phase response) for the connection.

Figures 8a, 8b, and 8c show the impulse response, echo path amplitude, and envelope delay response of a connection with several reflections at the far end that could not be separated. The ripples versus frequency in the amplitude response result from the relative phasing of the components from two reflections. Large nulls occur when the two reflected components are nearly equal and 180 degrees out of phase. These correspond to absorption bands, and in these regions the actual delay is not equal to the envelope delay.[17] In such instances, delay values for the connection were taken from smooth curves that continued the trends adjacent to the absorption bands.

4.2 Data processing during and after acquisition

Basic to processing of the echo data is the discrete Fourier transform (DFT).[18] The fast Fourier transform algorithm for calculating

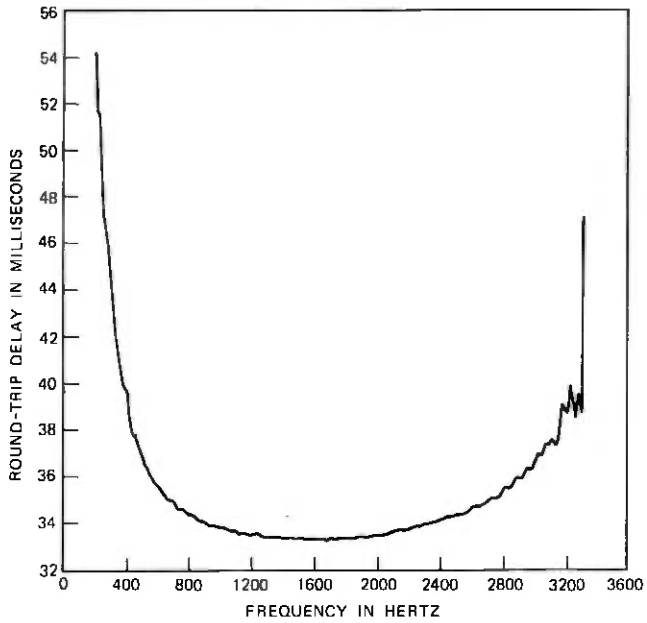


Fig. 7c—Far-end echo path envelope delay of a telephone connection versus frequency.

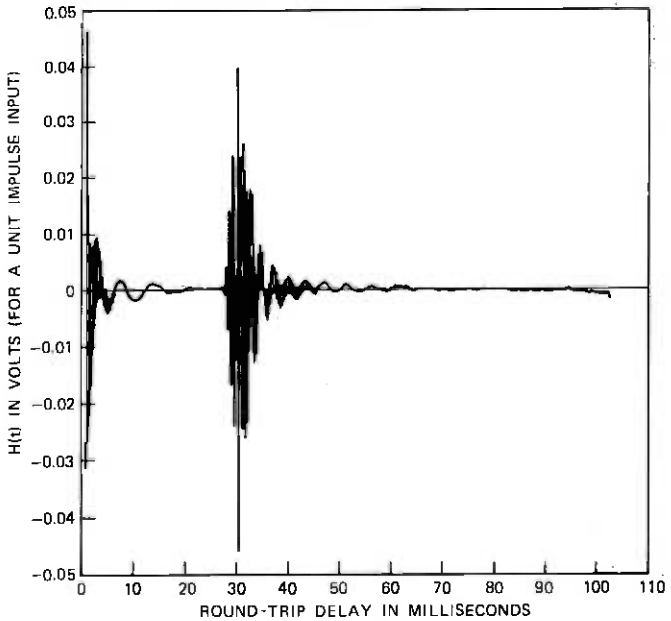


Fig. 8a—Impulse response of a telephone connection with multiple far-end echoes.

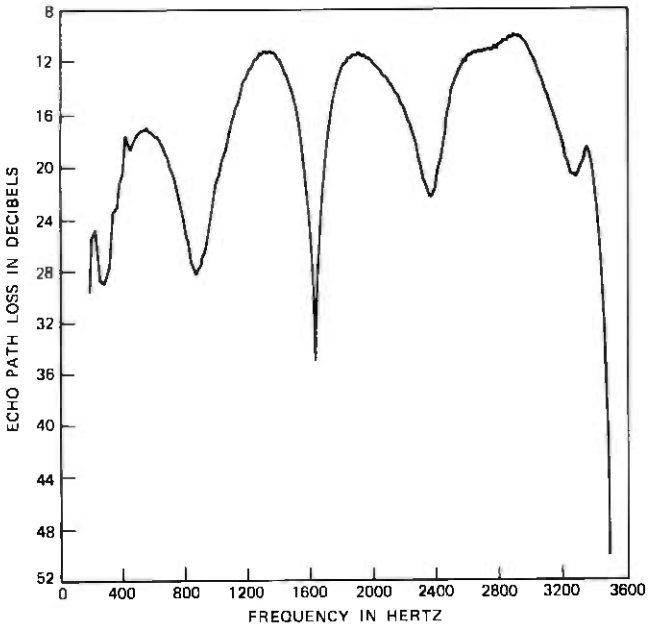


Fig. 8b—Far-end echo path loss of a telephone connection versus frequency; multiple echoes present.

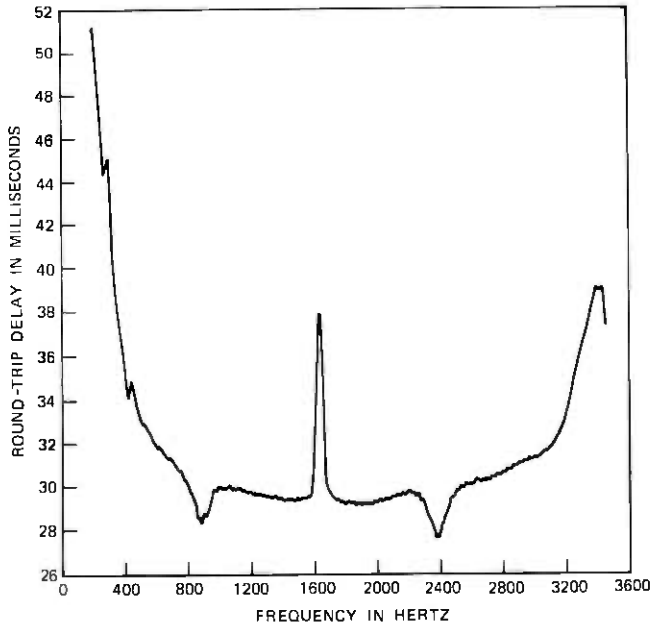


Fig. 8c—Far-end echo path envelope delay of a telephone connection versus frequency; multiple echoes present.

the DFT is most efficient if the number of data points to be transformed is an integral power of two. As previously mentioned, to measure delays of 100 ms the fundamental frequency should be about 10 Hz. The sampling rate for the converters was set to 10 kHz to allow some margin against aliasing of the voice-band measurements. For a 1024-point transform, a maximum period of 102.4 ms is possible, which meets our objective and gives a fundamental frequency $F_o = 10000/1024 = 9.76525$ Hz. The bandwidth of the echo paths is less than 3300 Hz[19] and, therefore, to cover the spectrum, the interrogation signal consisted of the sum of 390 sine waves spaced every 9.765625 Hz up to 3808.59375 Hz. The digital amplitude samples of the interrogation signal stored in the test set were preshaped to take into account the $(\sin x)/x$ weighting caused by using finite width samples instead of impulses to reconstruct the continuous wave.[20] Thirty-five periods of the interrogation signal were sent and, since each period lasts 102.4 ms, the test signal lasted 3.574 seconds. Received signals recorded simultaneously with transmission of the last 32 periods of the test signal were processed during data reduction. The time elapsing during transmission of the first three periods permitted any syllabic compandors present in the connection to reach equilibrium and transients to subside. The first step in signal processing was averaging of amplitude samples over the 32 signal periods, which improved the signal-to-noise ratio by 15 dB.

Although the desired echo path response is band-limited to less than 3300 Hz by filters in the facilities making up the trunks,[19] the reflected energy resulting from near-end discontinuities is not band-limited in this manner and normally will extend beyond 3800 Hz. The result of measuring a network whose bandwidth exceeds the bandwidth of the interrogation signal is equivalent to truncating the spectrum describing the wider bandwidth network, or measuring with these impulse response techniques an ideal low-pass filter in tandem with the desired network. This substantial discontinuity in the frequency spectrum causes Gibb's phenomenon[21] in the impulse response. To avoid this distortion, the returned signal was further digitally filtered by a 3400-Hz low-pass filter (DLPF) to assure that the response dropped off sufficiently at 3800 Hz. The DLPF loss characteristic is included in the response shown in Fig. 9a, where echo path loss versus frequency for the test set is shown. The high-frequency roll-off is determined entirely by the DLPF. This response is a calibration check for a 100-percent reflection (open circuit at the loop side of the hybrid) and thus includes all frequency weighting by the test set. Figure 9b is the envelope delay response of the calibration test.

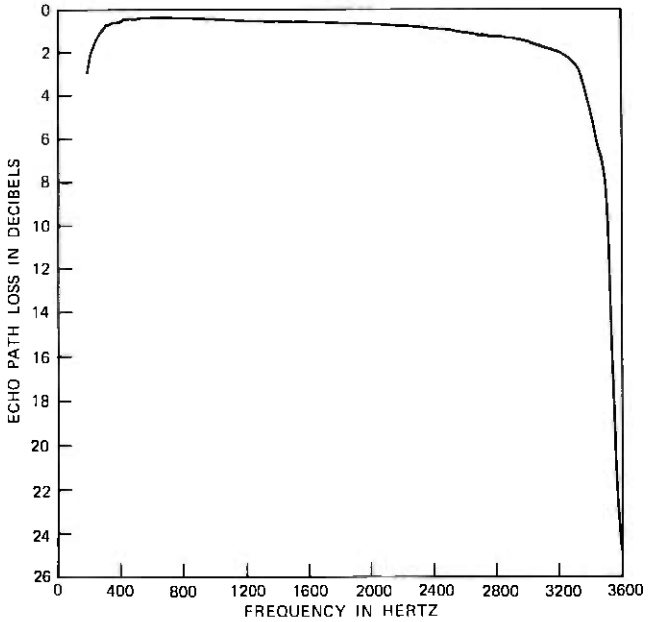


Fig. 9a—Frequency response of the echo path test set for a 100-percent reflection calibration test.

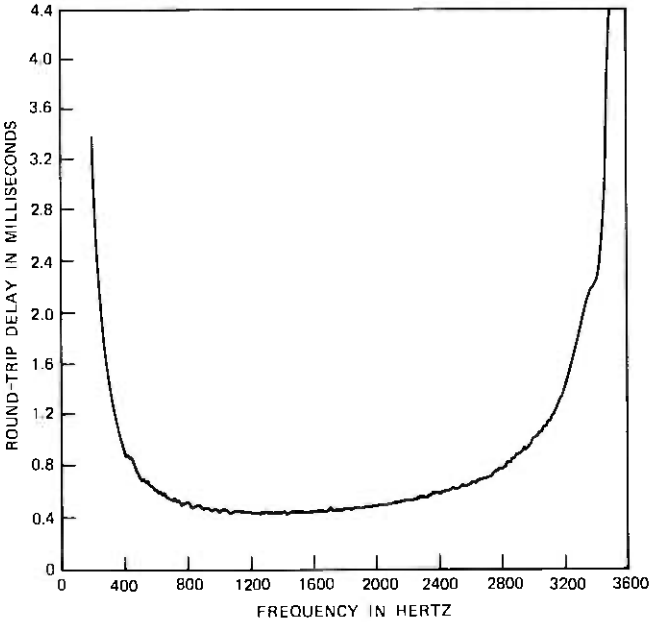


Fig. 9b—Envelope delay of the echo path test set for a 100-percent reflection calibration test.

Immediately following transmission of the 35-period interrogation signal, a 2109-Hz tone* was sent for 409.6 ms to disable echo suppressors that may have been present in the connection, and the interrogation signal was transmitted a second time. If an echo suppressor were present in the connection, it would have opened the return path and suppressed the echo from the far end during the first interrogation signal transmission. Sending the echo suppressor disabling tone made the echo suppressors inoperative, which kept the return path connected so that the echoes from all portions of the connection were recorded the second time the interrogation signal was sent.

The last 32 periods of the interrogation signal were averaged and then transformed by the FFT during processing of the recorded data. These transformed data were multiplied by the transform of the 3400-Hz DLPF mentioned above. The filtered data were then divided by the transform of the transmitted interrogation signal to obtain the system function of the connection modified by the filter. The inverse FFT was then computed, and the resulting impulse response was stored on magnetic tape. Both sets of data, with echo suppressors enabled and disabled, were processed. When this was completed, control was returned to the operator, and he could observe either impulse response on the oscilloscope. After viewing the impulse response, he could set to zero those portions of the response he desired to omit, and the test set would compute the spectrum of the echo path of interest. Upon completion of the transformation, either the amplitude spectrum or the phase response could be displayed on the oscilloscope. In addition, the minimum loss value and the average loss for the 500- to 2500-Hz band were displayed on the LED display. All calibration measurements, and periodically a test measurement, were processed in this manner to verify proper test set performance.

After the field survey was completed, processing was carried out on large-scale batch processing computers at the Holmdel location of Bell Laboratories. The echo path loss and phase were calculated, and test set characteristics were subtracted. In addition, the absolute envelope delay was calculated from the phase response, and microfilm graphs were created of all responses. The results were screened for errors, consistency checks were made, and the processed data were analyzed to obtain the results presented in Section V.

---

* In establishing a path through the DDD network for data transmission, a tone in the band 2010 to 2240 Hz is transmitted briefly just before application of the data signals to disable echo suppressors and permit simultaneous two-way transmission (Ref. 22).

V. ECHO PATH CHARACTERISTICS—DATA ANALYSIS RESULTS

The echo path loss and echo path delay discussed in the following sections are described in terms of means, standard deviations, and cumulative distribution functions. Each estimate of a population mean is accompanied by a 90-percent confidence interval to indicate the uncertainty because of sampling. Scatter diagrams and plots of cumulative distribution and probability density functions are used to illustrate data behavior in specific instances.

As stated in Section I, test connections originated from local switching offices and terminated at subscriber stations. All data have been adjusted to remove the influence of the lines used to connect the test equipment to the local switching offices and that of the testing equipment itself. Thus, results given in the following sections apply to connections having loops of zero length and 0-dB loss at the originating ends. Since test connections terminated at subscriber stations, customer loops were encountered at the far ends.

5.1 Loss characteristics of echo paths

Loss is intentionally introduced into the transmission path of a telephone connection to control echo performance, as previously noted. The total loss is allocated to various segments of the transmission path according to the Via Net Loss plan adopted by the Bell System in the early 1950s.[3,23] The goal of that design is to provide enough loss to control echo performance and simultaneously to insure adequate received levels for satisfactory direct transmission between subscribers.

5.1.1 Losses for far-end echoes

Three measures of echo path loss were extracted from each amplitude response characteristic. These are (*i*) the unweighted average echo path loss in the frequency band 500 to 2500 Hz, (*ii*) the echo path loss at 1000 Hz, and (*iii*) the minimum echo path loss. The average loss in the 500- to 2500-Hz band was calculated on the power scale over those test signal frequencies that fell within the indicated frequency band. This measure of echo path loss is used to evaluate subjective reaction to talker echoes in the telephone network. All three measures are discussed in this section.

Results of a statistical analysis of data for these three loss characteristics are tabulated in Table I. Echo suppressors were disabled when the information to calculate these loss characteristics was recorded.

The 500- to 2500-Hz echo path loss for far-end echoes is, on the average, 23.8 dB. Its distribution is approximately normal with a

Table I — Losses for far-end echo paths on toll telephone connections (echo suppressors disabled)

Connection Length (Airline Miles)	500- to 2500-Hz Echo Path Loss		1000-Hz Loss		Minimum Loss	
	Mean (dB)	Std. (dB)	Mean (dB)	Std. (dB)	Mean (dB)	Std. (dB)
180–2900	23.8 ± 1.9	6.3	26.2 ± 1.6	7.6	19.4 ± 1.8	5.8
180–360	23.1 ± 1.6	5.7	25.2 ± 1.5	7.0	18.7 ± 1.6	5.3
360–725	24.3 ± 2.2	6.8	26.8 ± 1.9	8.0	19.8 ± 2.0	6.3
725–1450	24.6 ± 2.2	6.3	27.2 ± 1.9	7.9	20.1 ± 2.0	5.7
1450–2900	23.3 ± 2.1	6.4	25.9 ± 1.8	7.6	18.9 ± 1.9	5.9

standard deviation of 6.3 dB. Table I shows that the estimated mean echo path losses increase slightly with increasing connection length in the first three mileage categories and the mean loss decreases slightly in the last one. This dependence upon connection length, while not statistically significant, probably results from application of the Via Net Loss plan to trunks used to establish connections. The trunk design loss under that plan is dependent upon the length of the trunk and is an increasing function of trunk length for trunks less than 1565 miles long. Echo suppressors were required on trunks longer than 1565 miles at the time the survey tests were made. Trunks containing an echo suppressor have a design loss of 0 dB. Although trunks are placed in tandem to establish connections, and airline distances instead of total trunk lengths are used to present the results in Table I, the overall influence of the Via Net Loss plan seems apparent in these estimates.

The estimated mean echo path loss at 1000 Hz exceeds the estimated mean 500- to 2500-Hz echo path loss by 2.4 dB, while the estimated standard deviation is larger by 1.3 dB. The difference in standard deviations is caused by ripples in the amplitude responses for some far-end echo paths (see Fig. 8b). These were caused by two or more reflections at the far end that could not be separated on some test connections. The 1000-Hz echo path loss is approximately normally distributed.

The estimated mean minimum echo path loss is 4.4 dB less than the estimated mean 500- to 2500-Hz echo path loss, and the estimated standard deviation is 0.5 dB smaller. The standard deviation is smaller because the minimum losses are less influenced by ripples in the amplitude responses. The distribution for minimum echo path loss is close to normal.

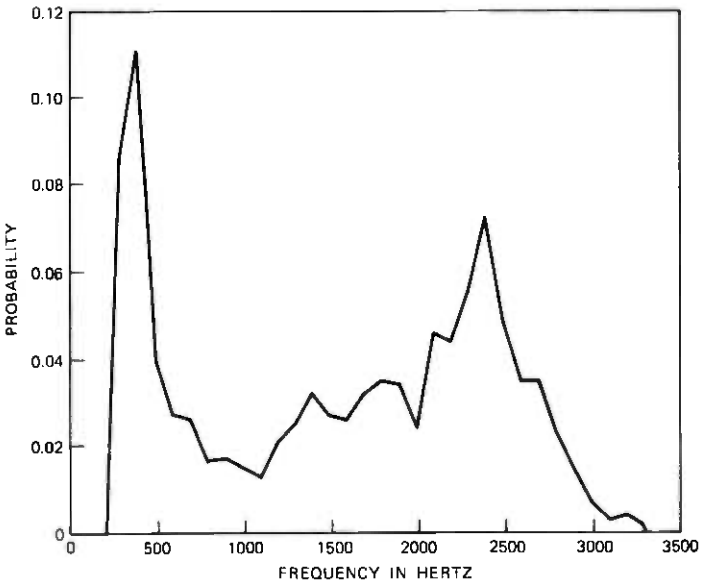


Fig. 10—Estimated probability density for the frequency at which minimum echo path loss occurs for far-end echoes.

The frequency at which the minimum echo path loss occurred was determined for each test connection. The estimated probability density for minimum echo path loss frequency is given in Fig. 10. This density function shows that the distribution is bimodal. The mode for low frequencies occurs around 400 Hz and the mode for high frequencies around 2400 Hz. When singing occurs because of excess gain on a connection, it usually is at frequencies between 200 and 500 Hz or 2500 and 3200 Hz. The bimodal behavior illustrated in Fig. 10 is in good agreement with that observed phenomenon.

5.1.2 Influence of echo suppressors

At the time field tests were conducted, echo suppressors were required on interregional intertoll trunks greater than 1565 miles long and on most intertoll trunks directly connecting regional-center toll-switching offices.[4]

Table II shows that 18 percent of toll connections longer than 180 airline miles contain an echo suppressor. When considered by mileage category, the table shows that essentially no echo suppressors are found on connections shorter than 725 airline miles. An estimated 25.4 percent of the connections belonging to the 725 to 1450 airline-mile category contain an echo suppressor. The airline distance between

Table II — Echo suppressor usage and operation on toll telephone connections

Connection Length (Airline Miles)	Percent Encountering Echo Suppressors	Far-end Echoes 500- to 2500-Hz Echo Path Loss			
		E. S. Disabled		E. S. Enabled	
		Mean (dB)	Std. (dB)	Mean (dB)	Std. (dB)
180–2900	18.0 ± 2.6	23.8 ± 1.9	6.3	28.9 ± 2.4	12.1
180–360	0	23.1 ± 1.6	5.7	23.2 ± 1.6	5.6
360–725	0.4 ± 0.3	24.3 ± 2.2	6.8	24.3 ± 2.2	6.8
725–1450	25.4 ± 5.1	24.6 ± 2.2	6.3	31.4 ± 3.5	11.9
1450–2900	90.1 ± 1.7	23.3 ± 2.1	6.4	49.2 ± 2.4	10.9

originating and terminating local switching offices is substantially less than the total trunk length for many connections. The locations of toll switching offices, alternate routing in the network, and the physical routes of transmission facilities between switching offices contribute to these differences between airline and actual connection lengths. Thus, some connections in this category may contain intertoll trunks greater than 1565 miles in physical length; others could contain intertoll trunks connecting two regional center switching offices. An estimated 90.1 percent of the connections in the longest mileage category contain an echo suppressor. The switching of intertoll trunks in tandem in a telephone connection accounts for some of these connections not having echo suppressors. Some may have had echo suppressors that did not suppress properly. These reasons account for the absence of echo suppressors on approximately 10 percent of the long connections. It was also estimated that 5.4 percent of echo suppressors did not respond to disabling tones.

Figure 11 is a scatter diagram of 500- to 2500-Hz echo path loss with echo suppressors disabled versus connection length in airline miles between originating and terminating local switching offices. Figure 12 is the corresponding scatter diagram of echo path loss with suppressors enabled. A comparison of these diagrams illustrates the influence of echo suppressors. Echo path loss is substantially increased on long connections when echo suppressors are enabled, while it remains unchanged for short connections that generally do not contain echo suppressors. Results for echo path loss with and without echo suppressors are listed in Table II. The presence of echo suppressors on connections in the last two mileage categories increases the mean echo path loss.

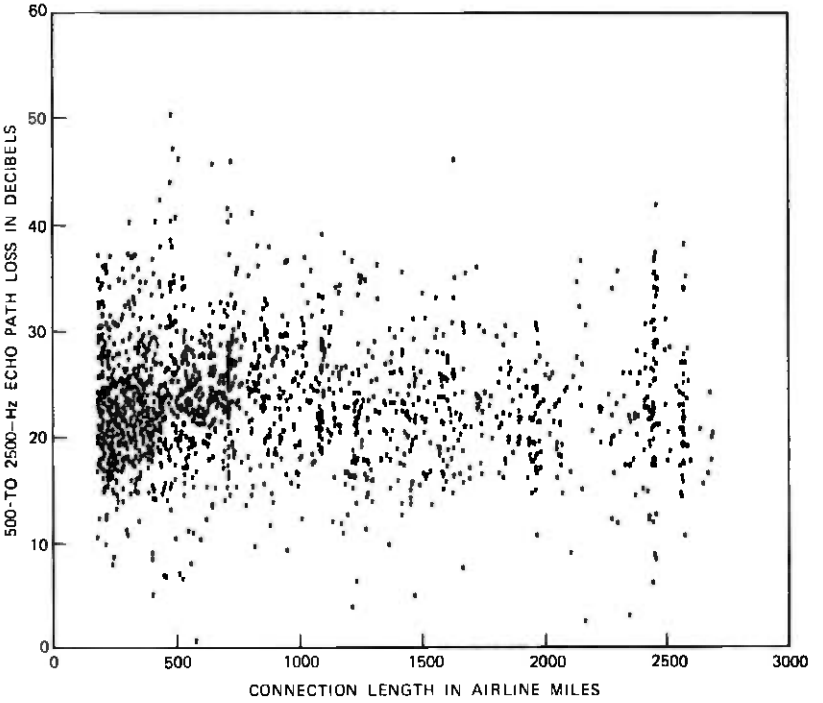


Fig. 11—Scatter diagram of far-end echo path loss on toll telephone connections with echo suppressors disabled versus connection length.

In many cases, the test signal was so highly attenuated by echo suppressors that only line noise was observed. The mixture of connections with and without echo suppressors accounts for the high standard deviations in the last two mileage categories. Figure 13 shows that the distribution of echo path loss with echo suppressors disabled is close to normal, while operation of echo suppressors causes positive skewness in the echo path loss distribution. Though not illustrated, the distribution for echo path loss with echo suppressors enabled is positively skewed in the third mileage category and negatively skewed in the fourth. The positive skewness in the third mileage category is caused by the 25 percent of the connections that contain echo suppressors. The negative skewness in the fourth category is caused by the 10 percent of the connections that do not contain echo suppressors.

Figure 14 graphically displays the suppression introduced by echo suppressors. The 500- to 2500-Hz echo path loss with echo suppressors disabled is plotted against the echo path loss with suppressors enabled. In cases where telephone connections did not contain echo suppressors,

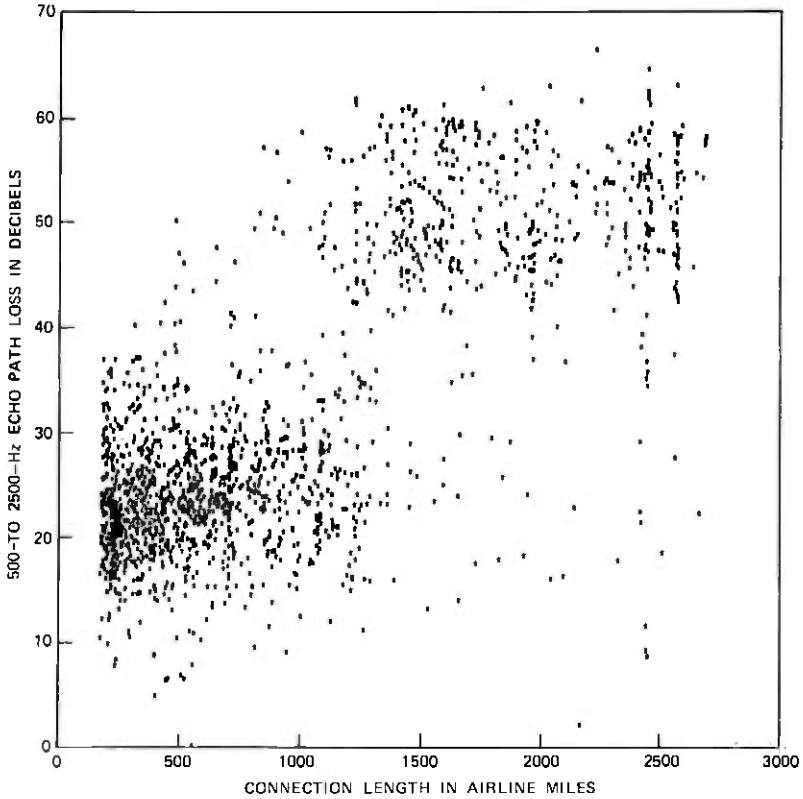


Fig. 12—Scatter diagram of far-end echo path loss on toll telephone connections with echo suppressors enabled versus connection length.

the two losses are close to being identical. These points in the figure lie about a line with a slope of +1 passing through the origin. In cases where properly functioning echo suppressors were encountered, the points lie well to the right of that line. An analysis of the data for connections that contained echo suppressors shows that the average additional loss inserted by the echo suppressors is greater than 28.4 dB and that this detectable additional loss is normally distributed with a standard deviation of 7.6 dB. In most cases, echo suppressors attenuated the reflected test signals to such an extent that they were below the noise present on the connections. In these cases, the actual losses with suppressors enabled were obscured by the circuit noise. Because of this, the estimated average additional loss inserted by echo suppressors is a lower bound on the average amount of suppression actually introduced.

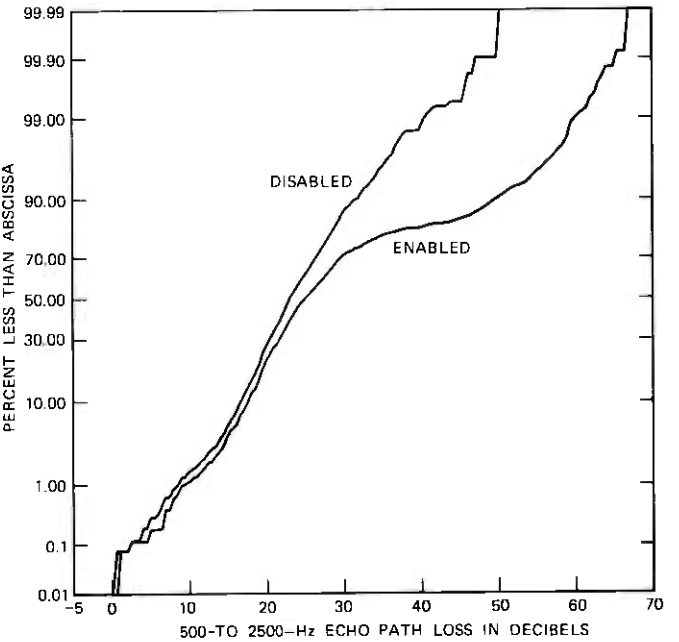


Fig. 13—Distributions of far-end echo path loss on toll telephone connections, with and without echo suppressors enabled.

5.2 Delay characteristics of echo paths

In voice transmission for a given echo path loss, the subjective disturbance caused by talker echo increases as receipt of the echo is increasingly delayed in time.[1–4] This delay is primarily determined by the length of the transmission path traversed, by the physical transmission media encountered, and by the number of modulation-demodulation steps associated with the individual transmission facilities encountered. Echo path delay on toll telephone connections is discussed in the following sections.

5.2.1 Delays experienced by far-end echoes

In Section I it was noted that echo path delay is the round-trip transmission delay experienced by an echo. This delay was computed from the impulse responses at approximately 10-Hz intervals across the voice frequency band. 1000-Hz and minimum echo path delays are discussed in this section. The minimum echo path delay may occur at different frequencies for different echo paths. The frequency at which the minimum occurs is also discussed.

Estimates of echo path delay are listed in Table III. The estimated average 1000-Hz delay is 19.5 ms for connections longer than 180 air-

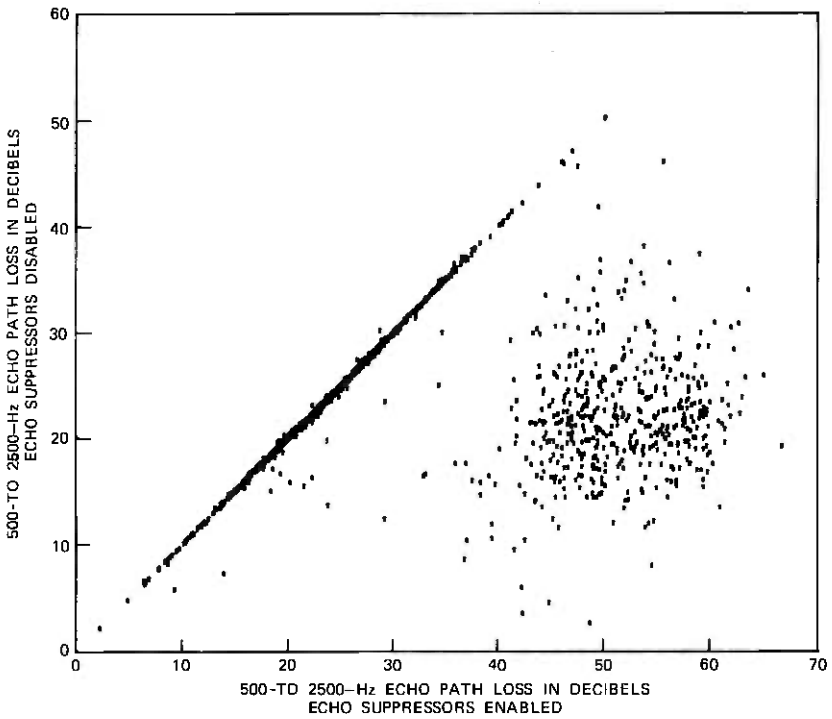


Fig. 14—Scatter diagram of far-end echo path loss on toll telephone connections with echo suppressors disabled versus echo suppressors enabled.

line miles. The distribution of 1000-Hz delay has an estimated standard deviation of 9.5 ms. Figure 15 shows that the distribution is truncated at about 5 ms in the lower tail, due to exclusion of toll connections shorter than 180 miles, and is highly skewed toward the positive direction.

Table III — Delays for far-end echo paths on toll telephone connections

Connection Length (Airline Miles)	1000-Hz Echo Path Delay		Minimum Echo Path Delay	
	Mean (ms)	Std. (ms)	Mean (ms)	Std. (ms)
180–2900	19.5 ± 0.9	9.5	18.5 ± 0.9	9.4
180–360	11.7 ± 0.8	3.4	10.6 ± 0.7	3.2
360–725	16.4 ± 0.5	3.8	15.4 ± 0.7	3.6
725–1450	24.8 ± 2.1	5.0	23.7 ± 2.0	4.8
1450–2900	37.3 ± 1.3	6.1	36.2 ± 1.4	6.0

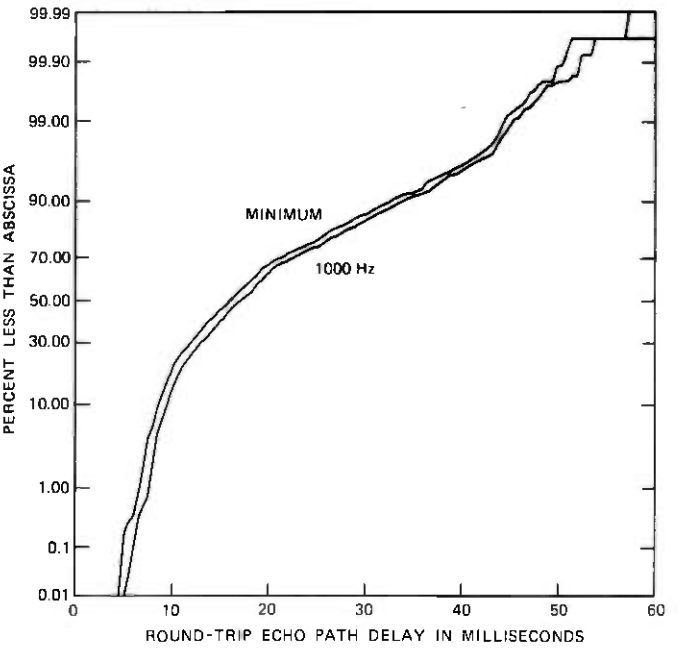


Fig. 15—Distributions of echo path delay on toll telephone connections for far-end echoes.

Average 1000-Hz delay increases monotonically with increasing connection length. This reflects the increased propagation delay required to travel greater distances and the increased likelihood of encountering more modulation-demodulation equipment on long connections. A previous study of intertoll trunks established that more channel bank pairs are found on long trunks than on short ones.[24] Alternate routing in the telephone network also accounts for an increased number of channel banks in long connections because more trunks are established in tandem to set up the connections. The standard deviation for 1000-Hz delay also increases with increasing connection length. The distribution of 1000-Hz delay in the shortest mileage subclass exhibits a high degree of positive skewness just as the overall distribution does. However, the distributions for the three remaining mileage subclasses are close to normal with slight deviations from normality found in the upper and lower tails.

Minimum delay closely follows the same trends discussed above for 1000-Hz delay. An analysis of the differences between the two delays calculated for each test connection estimates the average difference to be 1.1 ± 0.1 ms. These delay differences are close to being normally distributed with an estimated standard deviation of 0.6 ms. The cumulative distribution for minimum delay is also plotted in Fig. 15.

This figure clearly shows the similarity between 1000-Hz and minimum delay.

In many instances, the echo path delay versus frequency curves are rather flat in the middle of the voice frequency band. Since the minimum delay generally occurs in that area of the band, the frequency of minimum delay was arbitrarily defined to be 1700 Hz in these cases. This occurred in approximately one-third of the observations. The average frequency of minimum delay is estimated to be 1743 ± 33 Hz. The standard deviation is estimated to be 229 Hz.

5.2.2 Observed changes in echo path delay

Comparison of the current echo survey results with previously available information for echo path delay shows a noticeable decrease in the amount of transmission delay experienced in the telephone network. Figure 16 is a scatter diagram of the 1000-Hz echo path delay

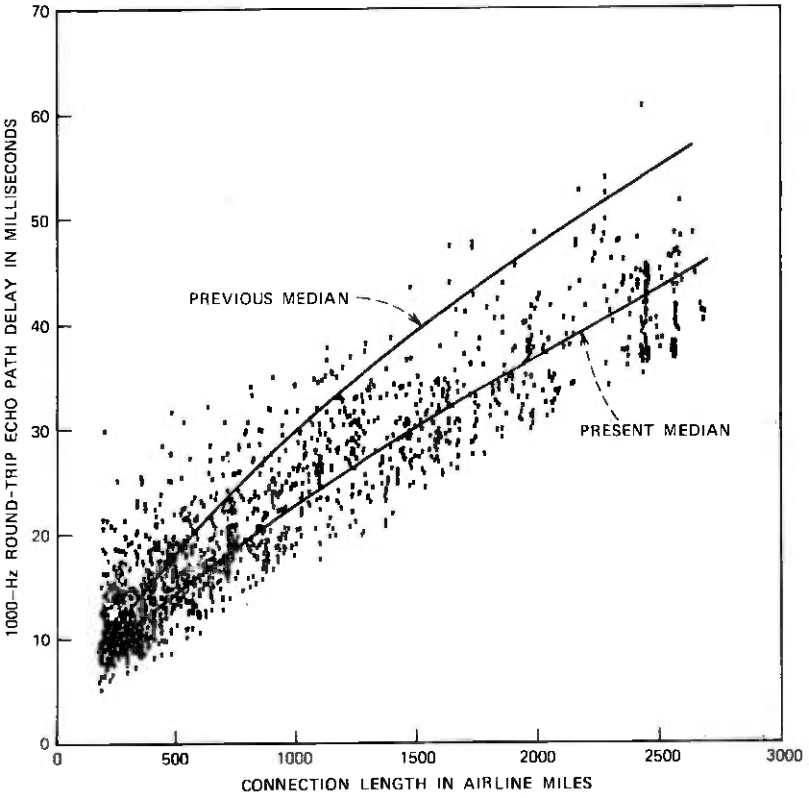


Fig. 16—Comparison of previous and present echo path delays on toll telephone connections.

observations for the echo survey versus the lengths of the connections on which the observations were made. Median 1000-Hz echo path delays are indicated by the two curves superimposed on the scatter diagram. The curve labeled "Previous Median" is based upon the echo path delay information available before this survey,[25] and the curve labeled "Present Median" is based upon the echo path delay data obtained in this survey. To generate the present median curve, the median delay was calculated in nine nonoverlapping, all-inclusive mileage bands. A linear, least-squares, curve-fitting routine was used to fit a quadratic equation to these nine points to obtain the curve. The standard error of this fit is 1.2 ms.

Examination of Fig. 16 shows a reduction in the estimated echo path delay between previous and present median echo path delays of about 11 ms for the longest connections. This improvement in median delay gradually decreases as the connection length gets shorter. For the shortest connections observed, the indicated improvement in median delay is very slight. This trend towards shorter echo path delays may have resulted from the following trends in the telephone plant over recent years: (*i*) provision of more direct high-usage trunk groups between cities, (*ii*) increasing use of carrier-type transmission facilities, and (*iii*) fewer voice-to-carrier frequency conversions in the longer trunks. These trends together produce the cumulative effects of reducing propagation delays attributable to physical transmission media and signal delays attributable to modulation-demodulation equipment.

5.3 Echo path loss versus delay

Echo path loss and delay have been discussed individually. In this section, echo path loss is described in terms of its observed relationship with 1000-Hz echo path delay. Connections are grouped into delay categories to analyze echo path loss. The interval of delay is 5 ms wide for each category. A particular delay category contains all connections having observed 1000-Hz echo path delays that fall within the specified time interval. Table IV lists the average 500- to 2500-Hz echo path losses estimated for each of the delay categories for echo suppressors disabled and enabled.

Results for echo path loss with echo suppressors disabled do not exhibit any trends related to 1000-Hz echo path delay. This is also evident in Fig. 17. Results listed in Table IV for echo path loss with echo suppressors in their normal operating conditions (enabled) show that the estimated average loss increases monotonically with increasing delay once echo suppressors begin to be encountered (around a 1000-Hz echo path delay of 15 ms). The standard deviation also starts changing at that point and continues to get larger until around 35 ms of delay,

Table IV — Losses versus delay for far-end echo paths on toll telephone connections

1000-Hz Echo Path Delay (ms)	500- to 2500-Hz Echo Path Loss			
	Echo Suppressors Disabled		Echo Suppressors Enabled	
	Mean (dB)	Std. (dB)	Mean (dB)	Std. (dB)
5–10	24.3 ± 1.2	5.9	24.4 ± 1.2	5.9
10–15	22.6 ± 1.6	5.9	22.6 ± 1.6	5.9
15–20	24.9 ± 2.3	6.2	25.0 ± 2.3	6.2
20–25	25.2 ± 2.4	6.5	29.3 ± 4.0	10.7
25–30	23.5 ± 1.6	6.1	35.3 ± 5.6	14.3
30–35	22.4 ± 1.8	6.8	38.6 ± 5.8	15.4
35–40	22.5 ± 1.6	5.9	48.5 ± 3.3	10.8
40–45	24.7 ± 3.0	7.4	49.3 ± 2.8	11.6
45–50	22.4 ± 0.9	4.2	52.8 ± 4.2	9.0

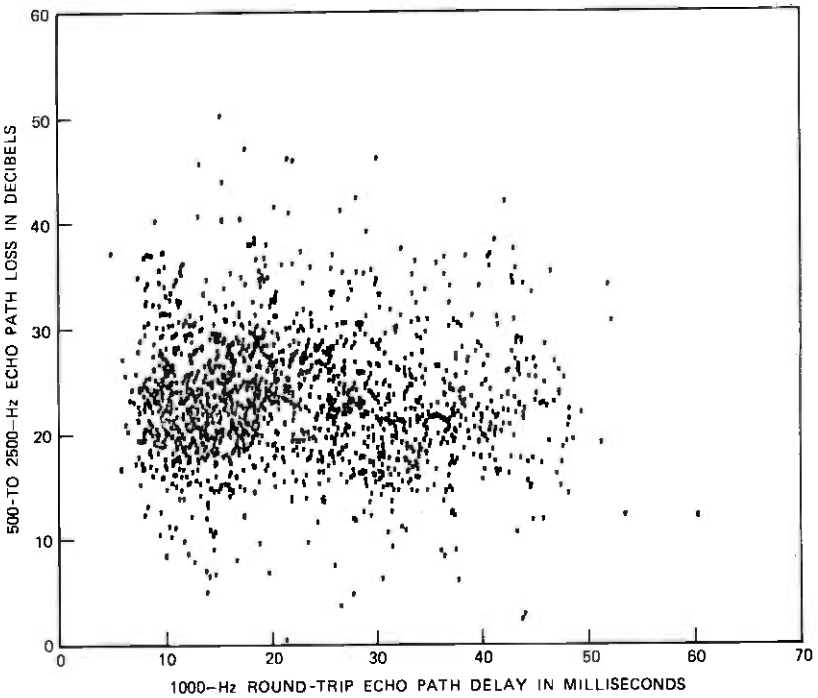


Fig. 17—Scatter diagram of far-end echo path loss with echo suppressors disabled versus echo path delay on toll telephone connections.

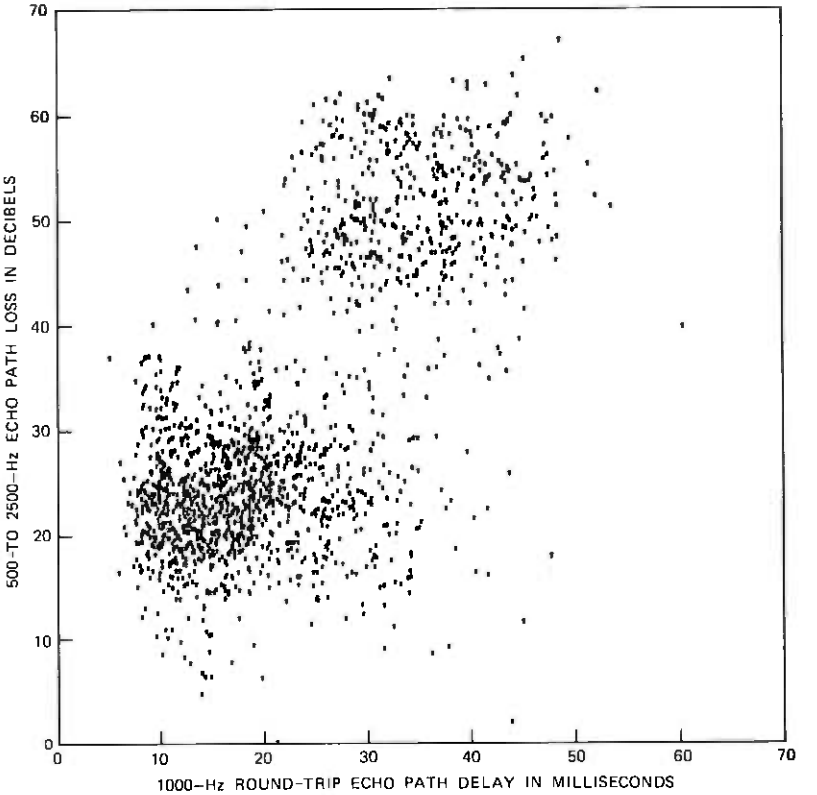


Fig. 18—Scatter diagram of far-end echo path loss with echo suppressors enabled versus echo path delay on toll telephone connections.

when it begins to decrease. This behavior reflects the process of encountering an increasing number of echo suppressors until, at around 35 ms of delay, only a relatively few connections remain that do not contain echo suppressors. This behavior is displayed in the scatter diagram presented as Fig. 18.

5.4 Intermediate echoes

Discussion of the data analysis results has been restricted to far-end echoes in the previous sections. In addition to far-end echoes, distinctly identifiable echoes occurring at intermediate toll switching offices were observed on approximately 28 percent of the connections. On those connections, the mean of the 500- to 2500-Hz echo path loss is 10.3 $\pm$ 0.7 dB higher for intermediate echoes than for far-end echoes. The estimated standard deviation of the loss difference is 7.2 dB, and the distribution deviates from normality in the lower tail. Approxi-

mately 5 percent of the intermediate echoes have lower loss than the corresponding far-end echoes on the same connections. In these few cases, the estimated average echo path losses are 30 dB for the far-end echoes and 27 dB for intermediate echoes. On two-thirds of these connections, the two losses are within 3 dB of each other.

The detection of intermediate echoes depended upon the relationship between the magnitude of the intermediate impulse and the peak amplitude of the total impulse response of the test connection. If the peak amplitude of the impulse response was high relative to the amplitude of intermediate echoes on the same connection, it is possible that the intermediates were not detected. Because of this peculiarity in the detection scheme, the estimates above are conservative, i.e., the echo path losses of intermediate echoes are, on the average, *at least* 10.3 dB greater than the echo path losses of far-end echoes.

VI. CONCLUSION

Acquisition of actual far-end talker echo path loss and echo path delay data on dialed-up long-distance telephone connections is now possible using digital computer techniques. By specifying such echo tests and analyzing the results according to sample survey procedures, echo performance of the continental United States switched telephone network has been characterized.

The echo survey illustrates the power inherent in modern sample survey methods. This is exemplified by the matching of the structure of the sampling plan to the structure of the population under study and by the flexibility in the sample design that allows analysis in subclasses that are not identical to the substrata of the sampling plan.

Results of the survey are being used to model the telephone network and evaluate network changes proposed to improve transmission performance. Significant changes in echo path delay were uncovered, and that information will be valuable in administering the United States DDD network from a transmission viewpoint.

VII. ACKNOWLEDGMENTS

Many individuals contributed to the echo survey. J. E. Unrue performed early theoretical work and measurement studies. J. R. Rosenberger developed software to record and process the echo data during field tests. Misses K. Haskell and J. Russo and Messrs. R. F. Cook, R. A. Gustafson, J. M. MacMaster, and E. J. Vlacich contributed to processing traffic data from the sampled locations, to processing data gathered during the field tests, and to assembling the data base for characterizing echo performance. Mrs. C. A. Hassol assisted in performing consistency checks and in screening the data. Seventeen

members of Bell Laboratories spent from one to three weeks in the field performing the tests. Finally, we would like to acknowledge the help of many Bell System craftspeople and central office foremen who assisted in the work, and R. Redilla, formerly of the AT&T Headquarters Engineering-Transmission Services Division, who acted as liaison between Bell Laboratories and the operating telephone companies.

REFERENCES

1. A. B. Clark, "Telephone Transmission over Long Cable Circuits," B.S.T.J., *2*, No. 1 (January 1923), pp. 67–94.
2. A. B. Clark and R. C. Mathes, "Echo Suppressors for Long Telephone Circuits," AIEE Trans., *44*, April 1925, pp. 481–489.
3. H. R. Huntley, "Transmission Design of Intertoll Telephone Trunks," AIEE Trans., *72*, part 1, November 1953, pp. 670–676.
4. *Notes on Distance Dialing*, New York: American Telephone and Telegraph Company, 1968, Section 6, part 3.
5. J. E. Unrue, Jr., "Controlling Echo in the Bell System," Bell Laboratories Record, *47*, No. 7 (August 1969), pp. 233–238.
6. C. G. Davis, "An Experimental Pulse Code Modulation System for Short Haul Trunks," B.S.T.J., *41*, No. 1 (January 1962), pp. 1–24.
7. K. L. Seastrand and L. L. Sheets, "Digital Transmission Over Analog Microwave Radio Systems," *International Switching Symposium Record*, Cambridge, Massachusetts, June 6–9, 1972, IEEE, New York, 1972.
8. H. E. Vaughan, "An Introduction to No. 4 ESS," *International Switching Symposium Record*, Cambridge, Massachusetts, June 6–9, 1972, IEEE, New York, 1972.
9. J. C. Davenport and D. T. Osgood, unpublished work.
10. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48*, No. 10 (December 1969), pp. 3337–3385.
11. I. Dolobowsky, unpublished work.
12. L. R. Pamm, unpublished work.
13. M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*, Volumes I and II, New York: John Wiley, 1953.
14. J. R. Rosenberger, unpublished work.
15. B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The Philosophy of PCM," Proc. IRE, *36*, No. 11 (November 1948), pp. 1324–1331.
16. W. R. Bennett, unpublished work.
17. Leon Brillouin, *Wave Propagation and Group Velocity*, New York: Academic Press, 1960.
18. W. T. Cochran, et al., "What is the Fast Fourier Transform?" Proc. IEEE, *55*, No. 10 (October 1967), pp. 1664–1674.
19. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1311–1347.
20. W. R. Bennett, M. Schwartz, and S. Stein, *Communication Systems and Techniques*, New York: McGraw-Hill, 1966.
21. A. Papoulis, *The Fourier Integral and Its Applications*, New York: McGraw-Hill, 1962.
22. "Data Communications Using the Switched Telecommunications Network," Technical Reference—PUB 41005, American Telephone and Telegraph Company, New York, 1971, p. 16.
23. F. T. Andrews, Jr. and R. W. Hatch, "National Telephone Network Transmission Planning in the American Telephone and Telegraph Company," IEEE Trans. on Commun. Technology, *COM-19*, No. 3 (June 1971), pp. 302–314.
24. I. Nåsell, C. R. Ellison, and R. Holmstrom, "The Transmission Performance of Bell System Intertoll Trunks," B.S.T.J., *47*, No. 8 (October 1968), pp. 1561–1614.
25. T. C. Spang, unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Optical-Fiber Packaging and Its Influence on Fiber Straightness and Loss

By D. GLOGE

(Manuscript received August 5, 1974)

Glass fibers are in general not thick enough to withstand external forces on their own without suffering axial distortion, mode coupling, and loss. Thus, plastic jackets must be carefully designed to provide effective protection. We evaluate jacket designs ranging from the mere use of soft materials to the application of multiple plastic coatings and graphite reinforcement. We compute the distortion loss as a function of dimensional variations and of lateral forces considered typical for cable packaging. The protective quality of a jacket is found to depend on a combination of stiffness and compressibility and on the fiber characteristics.

I. INTRODUCTION

Surprisingly small external forces can cause lateral deformations, mode coupling, and optical loss in clad fibers. For example, minute irregularities in the machined surface of a metal drum suffice to cause substantial loss in fibers wound on this drum with only a few grams of tension.[1] (An interesting and valuable study of this subject is described by W. B. Gardner.[2]) The pressure exerted on the individual fiber in a cable will almost certainly be considerably stronger and less uniform. The concern with this effect has heightened recently with the notion that lowest loss values are measured almost invariably in connection with extremely small mode coupling and after carefully eliminating external forces on the fiber.[3–5] Maintaining these loss values in a cable may require better fiber and, more importantly, effective jackets designed to optimally shield against external forces. This paper addresses the latter problem.

After gaining some insight into fiber deformation, we compute the excess transmission loss[6] resulting from statistical surface variations and lateral pressures affecting the fiber. The reader who is mainly interested in the results of this theory may wish to turn to Sections V or VI immediately, where practical examples and suggestions for an improved jacket design are discussed. We show that some care in this

respect substantially reduces the excess loss resulting from fiber distortion by outside forces.

II. ELASTIC DEFORMATIONS

We begin with the simple model of a fiber pressed against an elastic plane surface that is slightly rough (Fig. 1). The pressure from above is uniform, but as a result of the roughness, the contact forces between the fiber and the surface are not uniform along the fiber. Thus, the fiber bends slightly yielding to a force $f(z)$ per unit length.

According to the theory of the thin elastic beam, the lateral displacement $x(z)$ of the fiber axis is related to $f(z)$ by

$$\frac{d^4x}{dz^4} = \frac{f}{H}, \tag{1}$$

where

$$H = EI \tag{2}$$

is the flexural rigidity or stiffness; E is Young's modulus and I the moment of inertia. For the circular cross section of the fiber,

$$I = \frac{\pi}{4} a_1^4, \tag{3}$$

where a_1 is the radius of the fiber.

The force $f(z)$ not only causes a bending action, but also a deformation $u(z)$ of the surface. Provided that $f(z)$ does not change too drastically along z, $u(z)$ is a linear function of the force applied.[7] We introduce a factor of proportionality D, which we call the lateral rigidity, so that

$$u(z) = \frac{f(z)}{D}. \tag{4}$$

For the case of the elastic surface of Fig. 1, D is simply Young's modulus of the compressed surface material (we ignore a coefficient close to unity). To simplify the following steps, we assume temporarily that the surface is sufficiently compressible to conform to the fiber, producing a continuous line of contact. This imposes the relation

$$x + u - u_0 = v, \tag{5}$$

where $u_0 = \langle u \rangle$ is the average of $u(z)$ along z (see Fig. 1). Equations (1), (4), and (5) combined yield the differential equation

$$\frac{H}{D}\frac{d^4x}{dz^4} + x = v. \tag{6}$$

We now introduce the Fourier transforms $X(K)$ and $V(K)$ of $x(z)$ and $v(z)$. They are functions of a wave number K or a wavelength Λ

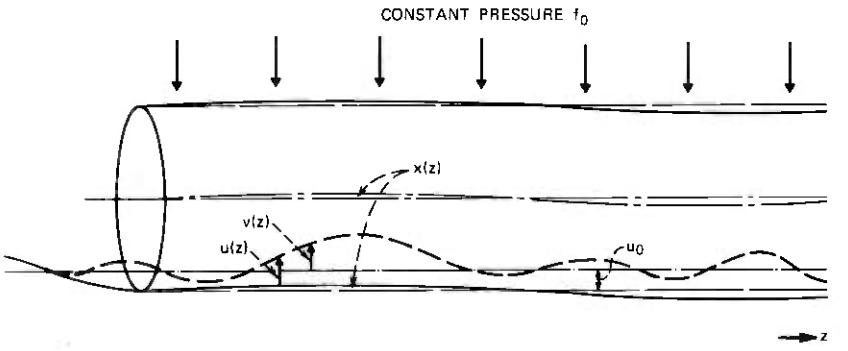


Fig. 1—Sketch of a fiber pressed against a rough surface by a uniform force (vertical dimensions strongly magnified).

related to K by

$$K = 2\pi/\Lambda. \tag{7}$$

In the Fourier domain, eq. (6) takes the form

$$X = \frac{V}{1 + K^4H/D}. \tag{8}$$

According to (8), the effect each Fourier component V has on the fiber displacement depends strongly on the wavelength of that component. Periodic disturbances having a wavelength smaller than

$$R = 2\pi(H/D)^{\frac{1}{4}} \tag{9}$$

hardly affect the fiber, while those with longer wavelengths than (9) are almost fully reproduced. The length R is called the retention length in the following, because it qualifies the usefulness of a given fiber package to keep the fiber in its natural straight condition.

III. INCOMPLETE CONTACT

Assume $v(z)$ to be a random variable measured from a suitable reference plane so that its mean is zero as in Fig. 1. Characterize the random process of which $v(z)$ is a sample function by the (power) spectral density $P_v(K)$. If a complete line of contact exists between the fiber and the surface, we can apply (8) to P_v so that the spectral density P_x of x becomes $P_x = P_v(1 + HK^4/D)^{-2}$.

If the contact is not complete, we have the situation of Fig. 2. Figure 2a depicts the case in which the fiber is very stiff and stays almost straight, while only the highest elevations of the rough surface are compressed. We assume a mean spacing t between the fiber periph-

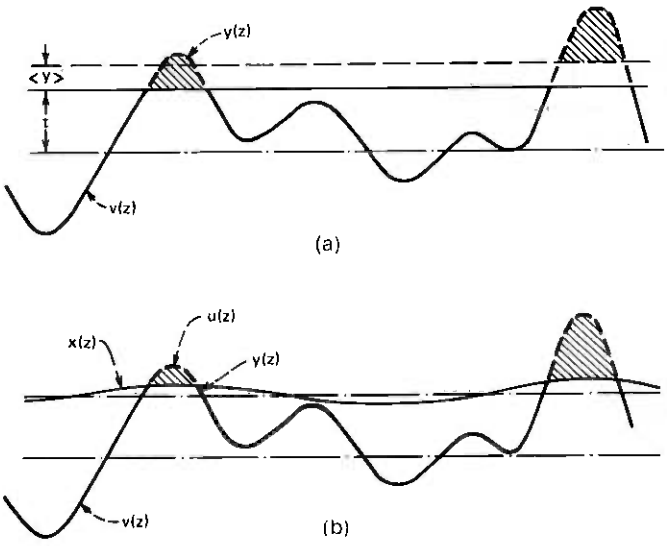


Fig. 2—Sketch of a fiber in incomplete contact with a rough surface. (a) The fiber is very stiff. (b) The fiber yields to bending. Cross-hatched areas indicate surface deformation (vertical dimensions strongly magnified).

ery and the surface. In this case the function causing the deformation is

$$y(z) = \begin{cases} v(z) - t & \text{for } v \geqq t \\ 0 & \text{for } v < t \end{cases} \tag{10}$$

rather than $v(z)$ itself. To obtain an approximate characterization of the random function y, we assume that $v(z)$ obeys a gaussian random process with standard deviation σ. The first two moments of y are, with this assumption,

$$\langle y \rangle = (2\pi)^{-\frac{1}{2}}\sigma^{-1} \int_t^\infty (v - t)e^{-v^2/2\sigma^2}dv \tag{11}$$

and

$$\langle y^2 \rangle = (2\pi)^{-\frac{1}{2}}\sigma^{-1} \int_t^\infty (v - t)^2 e^{-v^2/2\sigma^2}dv. \tag{12}$$

The variance of y is

$$s^2 = \langle y^2 \rangle - \langle y \rangle^2. \tag{13}$$

If one relates the spectral density P_y of $y - \langle y \rangle$ to that of v (for example, with the help of the Price method[7]), one finds the functional shape of both spectra to differ little for most cases of interest, so that the relation

$$\frac{P_y(K)}{P_v(K)} \approx \frac{s^2}{\sigma^2} \tag{14}$$

seems to be a useful approximation for all K. As y takes the place of v in (6) and (10), we can write

$$P_x = \frac{P_y}{(1 + HK^4/D)^2}, \tag{15}$$

which is

$$\approx \frac{P_v s^2/\sigma^2}{(1 + HK^4/D)^2}$$

because of (14).

It remains to find a relation between s^2/σ^2 and the (mean) lateral pressure which determines the extent of the contact. In the limit in which the fiber is stiff, as indicated in Fig. 2a, we have approximately $\langle f \rangle \approx D\langle y \rangle$, since $\langle u \rangle \approx \langle y \rangle$ in (4). The mean force $\langle f \rangle$ per unit length is, of course, the (linear) pressure we are interested in. To express (13) as a function of $\langle y \rangle$ only, we must eliminate t from (11) and (12). The result of this calculation is presented in approximate form:

$$\frac{s^2}{\sigma^2} = \left(1 + \frac{\pi^2}{4}\frac{\sigma^4}{\langle y \rangle^4}\right)^{-\frac{1}{2}}. \tag{16}$$

If the fiber cannot be assumed as stiff, the situation of Fig. 2b applies. We find that the surface deformation is more correctly given by the function

$$u(z) = \begin{cases} v(z) - x & v \geqq x + t \\ 0 & v < x + t \end{cases} \tag{17}$$

and that the mean of (17), rather than $\langle y \rangle$, determines the lateral pressure. The statistics of (17) are difficult to evaluate, since v and x are interrelated as a result of (15). According to (15), $x(z)$ essentially comprises all Fourier components of $v(z)$ having wave numbers $K < (D/H)^{\frac{1}{4}}$. As is evident from Fig. 2b, it is the remaining spectrum with $K > (D/H)^{\frac{1}{4}}$ that contributes to $u(z)$ of (17). This fact is the basis for the following estimate for the mean of (17):

$$\langle u \rangle^2 = \frac{\langle y \rangle^2}{s^2} \int_{(D/H)^{\frac{1}{4}}}^{\infty} P_v dK$$

$$\approx \frac{\langle y \rangle^2}{\sigma^2} \int_{(D/H)^{\frac{1}{4}}}^{\infty} P_v dK. \tag{18}$$

If the mean lateral pressure is $f_0 = \langle f \rangle$, we can write with (4), (16), and (18)

$$\frac{s^2}{\sigma^2} = \left[1 + \frac{\pi^2}{4}\frac{D^4}{f_0^4}\left(\int_{(D/H)^{\frac{1}{4}}}^{\infty} P_v dK\right)^2\right]^{-\frac{1}{2}}. \tag{19}$$

This relation together with (15) permits us to calculate the spectral

density P_x of the fiber deformation, if the spectrum of v and the mean pressure f_0 are known.

IV. DISTORTION LOSS

A significant exchange of power between two modes in a multimode fiber occurs when a periodic disturbance exists whose wave number K equals the phase lag between the two modes. This phase lag is in general a complicated function of the mode numbers involved and of the refractive index profile across the fiber core. Only if the index decreases as the square of the fiber radius (parabolic profile) is the phase lag the same for all modes coupled. This phase lag is

$$K_c = \frac{(2\Delta)^{\frac{1}{2}}}{a_c}, \tag{20}$$

when a_c is the core radius and Δ the relative index difference between core axis and cladding. If the index is uniform within the core and decreases abruptly to the cladding value (step-index profile), the coupled mode pairs have typically a smaller phase lag than K_c, although the phase lag approaches K_c for modes close to cutoff. For this profile, it is the spectral density $P_x(K)$ in the regime $0 < K \leqq K_c$, which determines coupling and coupling loss.

Equation (15) relates P_x to the spectral density P_v, which characterizes the original source of disturbance. We know little about its character; thus, to cover a broad variety of possibilities, we use the rather general functional description

$$P_v = \frac{P_v(0)}{(1 + l^2K^2)^\mu}, \tag{21}$$

with $\mu > 1$ and l large compared to $1/K_c$ and $(H/D)^{\frac{1}{2}}$. This stipulates a decrease of $P_v(K)$ in the vicinity of K_c in agreement with available experimental evidence.[2] The parameter l has the physical significance of a correlation distance. Integration of (21) yields a relation between $P_v(0)$ and the standard deviation σ introduced earlier:

$$P_v(0) = \frac{2\Gamma(\mu)\sigma^2 l}{\Gamma(\frac{1}{2})\Gamma(\mu - \frac{1}{2})}; \quad \mu > 1. \tag{22}$$

Coupling among neighboring modes dominates the power transfer inside the fiber. In the limit of very large mode numbers, the resulting power flow can be modeled by a diffusion process.[8] More specifically, if one defines a (continuous) mode variable r, one finds the power $\phi(r)$ in a mode group characterized by r from diffusion equations of

the form[8,9]

$$\frac{d}{dr}\left[\frac{K_c^4}{8\Delta} r^4 P_x(rK_c)\frac{d\phi}{dr}\right] + \gamma\phi = 0 \tag{23}$$

for the step profile and

$$\frac{d}{dr}\left[\frac{K_c^4}{4\Delta} r P_x(K_c)\frac{d\phi}{dr}\right] + \gamma\phi = 0 \tag{24}$$

for the parabolic profile. The term $\gamma\phi$ accounts for an overall decay of ϕ as a result of coupling loss. This excess loss is caused by radiation from modes at or beyond cutoff ($r \geqq 1$). The mathematical model considers the steeply rising loss at $r > 1$ to first approximation by the boundary condition $\phi(1) = 0$. In addition, we have $d\phi/dr = 0$ at $r = 0$, since no power can be lost at $r = 0$.

Equation (24) has an infinite set of eigensolutions[8] for arbitrary P_x; such solutions also exist for (23) at least if $\mu > 1$ and $l > 1/K_c$. In any of these cases, the lowest eigenvalue γ_0 is also the smallest and denotes the loss value approached asymptotically by long fibers once a "steady state" is reached. In the case of (24), γ_0 can be computed rigorously for arbitrary P_x; a way of finding a good upper limit for γ_0 of (23) is outlined in the appendix. The result is

$$\gamma_0 \leqq \frac{4\mu - 3}{16}\frac{K_c^4}{\Delta}\frac{P_y(K_c)}{1 + \dfrac{4\mu - 3}{4\mu + 13}(HK_c^4/D)^2} \tag{25}$$

for the step profile and

$$\gamma_0 = 0.36\frac{K_c^4}{\Delta}\frac{P_y(K_c)}{(1 + HK_c^4/D)^2} \tag{26}$$

for the parabolic profile, with

$$P_y(K_c) = \frac{2\Gamma(\mu)\sigma^2 l}{\Gamma(\frac{1}{2})\Gamma(\mu - \frac{1}{2})(lK_c)^{2\mu}} \times \left(1 + \frac{\pi\Gamma^2(\mu)\sigma^4 H^{\mu-\frac{1}{2}}}{(2\mu - 1)^2\Gamma^2(\mu - \frac{1}{2}) f_0^4 l^{4\mu-2} D^{\mu-9/2}}\right)^{-\frac{1}{4}} \tag{27}$$

from (19) and (22). Note that (25) is an upper limit and that these derivations are subject to the limitation $\mu > 1$ and that l must be large compared to $1/K_c$ and $(H/D)^{\frac{1}{4}}$.

In general, it will be necessary to determine the parameters in (21) from experimental evidence. For the numerical results following in the next sections, we have used $\mu = 3$ as a typical and realistic example.[2]

In this case, the use of (20) and (27) converts (25) and (26) into

$$\gamma_0 \leqq \frac{3}{2\pi} \frac{\sigma^2 a_c^2}{l^5 \Delta^2} \frac{1}{\left(1 + \frac{144\Delta^4 H^2}{25 a_c^8 D^2}\right)\left(1 + \frac{64\sigma^4 H^{\frac{1}{2}} D^{\frac{1}{2}}}{225 f_0^4 l^{10}}\right)^{\frac{1}{4}}} \tag{28}$$

for the step profile and

$$\gamma_0 = \frac{96}{25\pi} \frac{\sigma^2 a_c^2}{l^5 \Delta^2} \frac{1}{\left(1 + \frac{4\Delta^2 H}{a_c^4 D}\right)^2 \left(1 + \frac{64\sigma^4 H^{\frac{1}{2}} D^{\frac{1}{2}}}{225 f_0^4 l^{10}}\right)^{\frac{1}{4}}} \tag{29}$$

for the parabolic profile.

V. FIBER STORAGE DRUM

The main objective of this theory is, of course, the design of jackets that protect the fiber from distortion and the loss associated with it; but, to begin with a simple problem, let us first ask how the loss increases in a fiber when it is wound on a drum. Clearly, the drum surface properties and the winding force are important factors. We assume the radius ρ of the drum to be so large that the constant curvature of the fiber has no noticeable influence on the loss. If we apply a tensile force F, the fiber presses against the drum surface with a (linear) pressure

$$f_o = F/\rho. \tag{30}$$

With these definitions, the distortion loss as a result of the winding pressure can be directly computed from (25) and (26). The results are illustrated by the following representative example:

(*i*) Fiber characteristics:
Core radius $a_c = 40$ μm.
Outside radius $a_1 = 60$ μm.
Relative index difference $\Delta = 0.005$.
Young's modulus (silica) $E_1 = 7000$ kg/mm^2 (10^7 psi).

(*ii*) Drum surface statistics:
Standard deviation $\sigma = 1$ μm.
Correlation length $l = 1$ mm.
Spectral coefficient $\mu = 3$.

The evaluation of (25) and (26) for $\mu = 3$ is given in (28) and (29). We discuss only the step-index profile in the following. The results for the parabolic profile can be obtained from (29); they differ little from those of the step profile.

Figure 3 is an evaluation of (28) as a function of Young's modulus of the drum surface material with the pressure f_o as a parameter. The

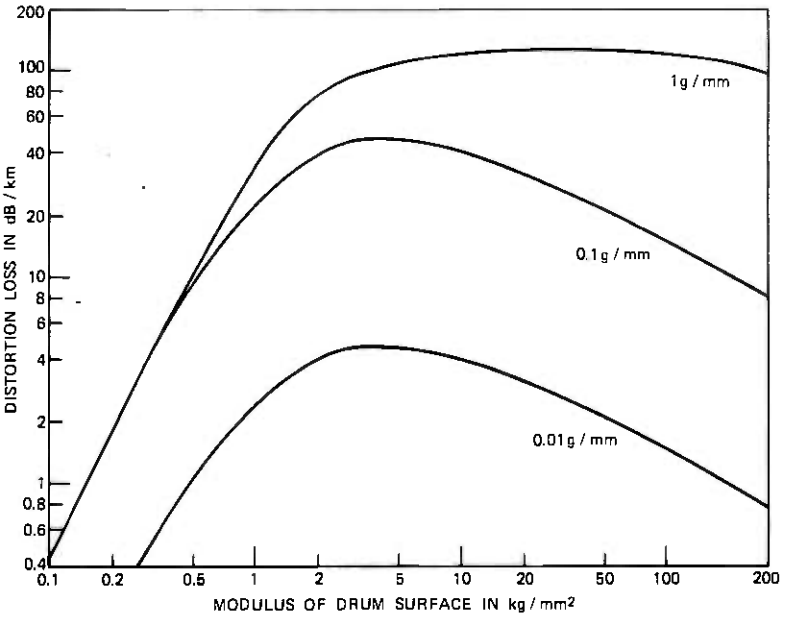


Fig. 3—Distortion loss versus drum surface modulus according to eq. (30); fiber diameter is 120 μm, core diameter is 80 μm, relative index difference is 0.5 percent, rms drum surface roughness is 1 μm, correlation distance is 1 mm. Mean lateral force per unit length is the parameter.

plot represents the loss for drums of different elasticity provided that the surface statistics are the same for all. If $F = 100$ g and $\rho = 10$ cm ($f_o = 1$ g/mm), the distortion loss can be as high as 130 dB/km. For low pressures, the loss decreases with increasing rigidity of the drum surface, as the fiber ceases to conform to the irregularities of the surface. If the drum is soft, the loss is reduced independently of the pressure, since the fiber sinks into the surface and smoothes the irregularities. Thus, both hard and soft surfaces have a tendency to decrease the excess loss for a given pressure. The effect of a hard surface, however, strongly depends on the pressure applied. A reduction of the loss to 0.5 dB/km independently of pressure requires an extremely soft surface (0.11 kg/mm^2 = 157 psi) for the kind of fibers characterized by this example. Typical winding forces which are caused by the pulling operation itself or applied in rewinding operations are in the range between 10 to 100 g. Thus, a loss increase of 100 dB/km or more as a result of drum storage is not surprising.

Equation (28) shows that γ_0 is proportional to Δ^{-6} in the case of "soft" surface conditions, i.e., when the first parenthesis in the denominator of (28) is much larger than unity. Hence, an increase in the

index difference by a factor of 1.5 would reduce the loss coefficient in this range by one order of magnitude. Of course, these results depend on the surface statistics assumed here. For arbitrary μ, the excess loss coefficient is proportional to $\Delta^{1-\mu}$ if the surface is hard and to $\Delta^{-3-\mu}$ if the surface is soft.

Next, let us consider a jacketed fiber wound onto a slightly rough drum. The lateral rigidity D_2 of the jacket is, in general, different from the rigidity D_1 of the drum surface. To account for the compressibility of both, one must use an effective rigidity

$$D_e = \frac{1}{1/D_1 + 1/D_2} \tag{31}$$

in (25) to (29). There will be statistical variations of the jacket thickness and these are likely to differ from those of the drum surface. If one or the other dominates and follows the characteristics (21) with $\mu = 3$, one can still use (28) or (29) or Fig. 4 to determine the distortion loss if one incorporates (31).

VI. PLASTIC JACKET DESIGN

In a cable, the fibers will be organized in bundles and pressed together by binding or sheathing forces, by cable deformations, and by pressure on the cable, once it has been placed.

Considering only one cross-sectional dimension, we assume a typical fiber of the bundle to be contacted by two others, one on either side. All fibers have plastic jackets, so that elastic surfaces of equal modulus press against each other. The situation is similar to that described by (31) except that now D_1 and D_2 of that equation are identical and equal to the modulus (or the rigidity D) of the jacket material. Hence, $D_e = D/2$. Other differences with respect to the previous model are the two lines of variable pressure and a total of four random variables involved in the deformation of the fiber. These variables are the jacket thickness variations v_1 and v_2 of the fiber in the middle and the variations v_3 and v_4 referring to the jackets on the outside. If we assume again complete and continuous contact, the resulting differential equation becomes

$$2\frac{H}{D}\frac{d^4x}{dz^4} + 2x = v_1 - v_2 - v_3 + v_4, \tag{32}$$

where the relation $D_e = D/2$ has been used. The variables v_1 to v_4 are statistically independent, but they are samples of the same ensemble. Therefore, they all have the same spectral density $P_v(K)$ and the spectral density of the sum on the right of (32) is 4 $P_v(K)$. After Fourier transformation and the insertion of spectral densities into (32),

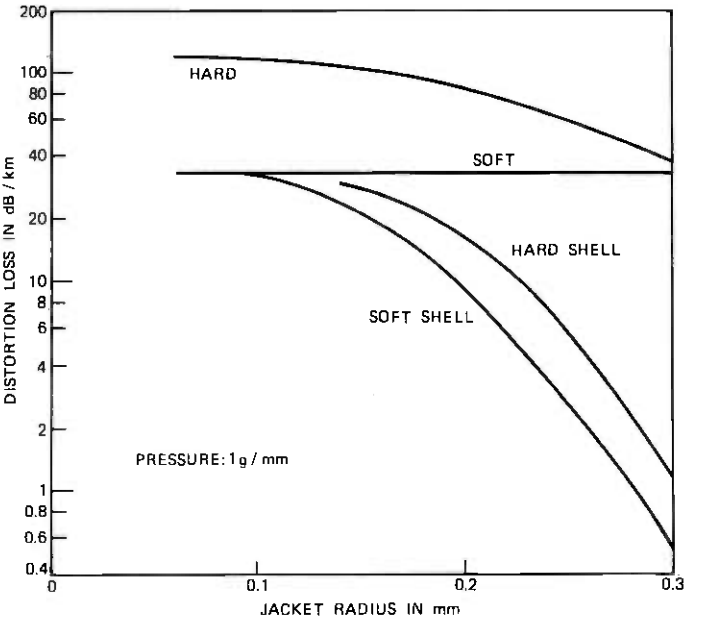


Fig. 4—Distortion loss versus outside jacket radius according to eq. (30); fiber characteristics as in Fig. 3, rms jacket thickness variation 1 μm, correlation distance 1 mm, mean lateral pressure 1 g/mm. Curves refer to four jacket configurations listed in Table I.

all numerical factors cancel, leaving us with the mathematical relationships derived earlier. As a result, (28) and (29) are also applicable to the problem of the jacketed fiber in a bundle, provided that the statistics of the jacket thickness variations can be described by (21) with $\mu = 3$. Now D stands for the modulus of the jacket and H for the combined stiffness of fiber and jacket. The stiffness of the latter is

$$H_2 = \frac{\pi}{4} E_2(a_2^4 - a_1^4) \tag{33}$$

with E_2 being Young's modulus of the jacket material and a_2 and a_1 its outer and inner radius, respectively. In the case of several jackets, H is generally the sum over all stiffnesses. If the outer jacket is the softer one and sufficiently thick that a deformation beyond its elastic limit is unlikely, D is simply the modulus of the outer jacket. If the outer jacket is harder than the inner one and has a thickness b small compared to its outer radius a_2, we have[10]

$$D \approx E_2 + E_3 \frac{b^3}{a_2^3}, \tag{34}$$

where E_2 and E_3 are the moduli of the inner and the outer jacket, re-

spectively. If the inner jacket is very soft and thick, we must consider the hard outer shell and the fiber as two independent systems, each undergoing deformations governed by differential equations similar to (32). The result are four instead of two expressions in the denominator of (28) and (29), one pair comprising the H and D parameters of a hard shell surrounding a soft material, and the other pair comprising the H and D parameters of a fiber imbedded in a soft material.

The following is a discussion of four alternative jacket configurations. As a realistic example, we consider the same fiber characteristics and the same statistical parameters listed in the previous section for the drum surface. Table I gives a description of the jackets. The first is made entirely from a soft plastic, the second from a hard plastic, and the third and fourth are hybrid structures. We assume a modulus of 1 kg/mm² (1400 psi) for a typical soft material and 100 kg/mm² for a typical hard material. In Figs. 4 and 5, the outer jacket radius a_2 is plotted versus the excess loss computed for each structure if the mean lateral pressure is either 1 g/mm (Fig. 4) or 0.1 g/mm (Fig. 5). The pressure obviously determines the choice between a soft or a hard material, if the jacket is to be made from one material alone. The decrease of the loss contribution with increasing jacket radius in case of the hard jacket comes about as a result of the increase in stiffness. The corresponding increase afforded by the soft jacket is negligible. The last two columns of Table I list the D and H parameters used in each case.

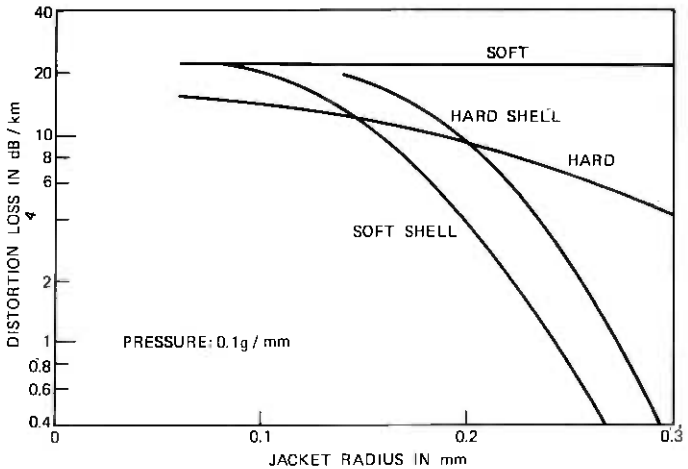


Fig. 5—Distortion loss versus outside-jacket radius according to eq. (30); fiber characteristics as in Figs. 3 and 4, jacket statistics as in Fig. 4, and mean lateral pressure 0.1 g/mm. Curves refer to four jacket configurations listed in Table I.

Table I — Characteristics of several types of protective jackets for optical fibers

	Modulus in kg/mm²	Inside Radius in mm	Outside Radius in mm	Rigidity in kg/mm²	Stiffness in kg mm²
Soft Jacket	1	0.06	a_2	1	0.0713
Hard Jacket	100	0.06	a_2	100	$0.0702 + 78.5a_2^4$
Inside Jacket	100	0.06	$a_2 - 0.02$	1	$0.0702 + 78.5(a_2 - 0.02)^4$
Outside Jacket	1	$a_2 - 0.02$	a_2		
Inside Jacket	1	0.06	$a_2 - 0.04$	1	0.0713
Outside Jacket	100	$a_2 - 0.04$	a_2	$1 + \frac{0.0064}{a_2^3}$	$78.5a_2^4 - 78.5(a_2 - 0.04)^4$

The third structure has a hard jacket padded with a soft outer layer. The layer thickness of 20 μm was chosen to avoid any deformation beyond its elastic limit. The fourth structure has a hard shell surrounding a soft material. The thickness of this shell should be approximately 0.02 a_2. This optimum is a result of an increase both in stiffness and lateral rigidity as the shell thickness is increased, so that the retention length R, which is the ratio of the two, passes through a maximum. To simplify matters, we have chosen a thickness of 40 μm independent of the shell radius. The two pairs of D-H values listed in the case of the fourth configuration refer to the two independently deforming structures (shell and fiber) which must be considered in this case, as was mentioned earlier. The slight advantage of the soft over the hard shell, evident in Figs. 4 and 5, is too small to be decisive. It may well be offset by weight and cost considerations. The substantially improved fiber protection afforded by the hybrid structures as compared to simple jackets, however, is well worth considering. A jacket diameter of 0.5 to 0.6 mm permits a virtual elimination of the distortion loss in case of the example considered here. A similar reduction by a single hard jacket requires at least twice this jacket diameter.

The excess loss computed for the structure with a hard shell vanishes when the modulus of the inner jacket is reduced to zero. This implies that the protection provided by a stiff shell that surrounds the fiber in a loose way without any material in between is perfect. Of course, this would indeed be true if the only forces present were lateral outside forces borne by the shell. In practice, there are other forces not considered here; forces that press the fiber against the inside jacket wall in a cable bend, for example. Such forces determine the distortion loss of the loosely jacketed fiber. Although this is an important problem to consider, it is beyond the scope of this work.

Properties similar to those of hybrid jackets can also be obtained with reinforced jackets. The reinforcement could, for example, consist of strong fine fibers running parallel or slightly stranded to the optical fiber imbedded in a relatively soft jacket material. The fiber material could be plastic, glass, or graphite, the latter being particularly suited because of its low weight, high tensile modulus, and high strength. Also, as graphite fiber is available with diameters down to 5 μm, its incorporation into the jacket should be manageable without causing permanent internal stresses resulting in distortion loss by itself. The advantage of the reinforced jacket is its anisotropy which combines stiffness with lateral compressibility. Although these properties are difficult to compute, an estimated loss reduction of two orders of magnitude for a jacket 0.4 mm in diameter seems achievable with the fiber characteristics listed in the previous section.

The effect of the reinforced jacket and of the configurations 1, 3, and 4 in Table I is a combination of stiffness and compressibility, while that of configuration 2 is based on stiffness alone, preventing its conformance to surface irregularities. Mathematically, these two effects are distinguished by the two parentheses in the denominator of (25) or (26). It is important to note that the first parenthesis depends strongly on the fiber characteristics. An effective jacket implies $HK_c^4/D \gg 1$ in (25), so that the loss reduction afforded by the first group of jackets is proportional to K_c^8 or, with (20), to Δ^4/a_c^8. As a result, a small increase in index difference substantially increases the effectiveness of these jackets. If, for example, $\Delta = 2$ percent instead of 0.5 percent as previously assumed, the soft jacket, the reinforced jacket, and the two hybrid structures reduce the excess loss coefficient by an additional factor of 256, while the effect of the hard jacket remains the same. This strongly emphasizes the importance of this first group of jackets and the need for fibers with large index difference.

Of course, the above dependence on Δ holds only as long as the predominant sources of loss are indeed those assumed here. If other sources of loss dominate, as, for example, the influence of a very lossy cladding material, typically only a fraction of all trapped modes propagates in the steady state. In this case, Δ in (20) and in the above arguments must be replaced by $N^2/2n$, where N is the effective numerical aperture characterizing the mode distribution of the steady state and n the refractive index of the core.

VII. CONCLUSIONS

Optical fibers need protection from lateral forces and this requires a careful design of the fiber jacket. The jacket should have a high flexural rigidity or stiffness in combination with a good lateral compressibility. These properties define a retention length within which the jacket essentially absorbs irregularities impressed from the outside. Longer irregularities deform the fiber and can lead to distortion loss if they comprise spectral components in the vicinity of the critical wave number of the fiber.

Although the forces to which a fiber is subjected in a cable are difficult to estimate, one gains a fair notion of the sensitivity of the fiber to such forces by winding it on a drum with minute surface irregularities. This can best be illustrated by way of a representative example. Consider a silica fiber, 120 μm in diameter, that has a relative index difference of $\Delta = 0.5$ percent and a core diameter of 80 μm. Assume a tensile force of between 10 and 100 g applied when winding the fiber on a drum, which has a diameter of 10 cm and an rms surface roughness of 1 μm. The estimated loss increase is between 50 and 130 dB/km

depending on the winding force applied. A winding force of 10 g corresponds to a mean pressure of 0.1 g/mm on the fiber.

Now consider five types of jackets:

(*i*) A soft plastic jacket having a modulus of 1 kg/mm^2.
(*ii*) A jacket of hard plastic with 100 kg/mm^2.
(*iii*) The same as (*ii*) padded with a thin layer of the material used in (*i*).
(*iv*) A shell of the material of (*ii*) on top of soft material as used in (*i*).
(*v*) A soft jacket reinforced by a filler of strong plastic, glass, or graphite fiber.

We find that, for equal jacket diameters, (*i*) is almost always better than (*ii*) except when Δ and the lateral forces are small. For the fiber of the previous example, the jacket (*i*) reduces the excess loss cofficient by a factor of 3. If optimized in thickness, the shell (*iv*) is about as useful as (*iii*). An overall thickness of 0.6 mm permits in both cases a reduction of the loss coefficient by two orders of magnitude. A graphite reinforced jacket of equal size should have at least the same effect.

The effectiveness of a jacket is a strong function of the fiber to be protected. For example, the factor by which the jacket reduces the loss coefficient is proportional to Δ^4. In addition, the distortion loss of the unprotected fiber is a function of Δ. Hence, the loss coefficient may typically scale as Δ^{-2} for a fiber without jacket, but as Δ^{-6} for the jacketed fiber. In other words, if the index difference in the previous example had been 1 percent instead of 0.5 percent, the excess loss would have been initially less than 35 dB/km on the drum and 0.5 dB/km after protection with a simple soft jacket. Cable forces are likely to be stronger and less uniform than those encountered on a storage drum and may necessitate a fiber protection by the more expensive hybrid jackets or even by reinforcement.

VIII. ACKNOWLEDGMENTS

Helpful discussions with E. A. J. Marcatili, W. B. Gardner, and L. L. Blyler are gratefully acknowledged.

APPENDIX

Rayleigh-Ritz Limit for Steady-State Loss

The Rayleigh-Ritz method[12] provides a surprisingly close upper limit for the lowest eigenvalue of differential equations of the type in (25) or (26) if a reasonable trial solution for the lowest eigenvalue can be constructed. We demonstrate this for an important subclass of (25).

Consider the case that l in (23) is very large and $(H/D)^{\frac{1}{2}}$ in (17) is very small compared to $1/K_c$, so that $P_x \approx P_y \approx 2\Gamma(\mu)\sigma^2 l/\Gamma(\frac{1}{2})\Gamma(\mu - \frac{1}{2})(lK)^{2\mu}$. After suitable normalization, (25) is then of the form

$$\frac{d}{dr} c(r) \frac{d\phi}{dr} + g\phi = 0 \quad \text{with } c = r^{-\sigma} \tag{35}$$

and $\sigma > -2$. Multiply (36) by ϕ, integrate over r from 0 to 1, and solve for g. With the boundary condition $\phi(1) = 0$, one arrives at

$$g = \frac{\int_0^1 c \left(\frac{d\phi}{dr}\right)^2 dr}{\int_0^1 \phi^2 dr}. \tag{36}$$

We choose the trial solution

$$\phi = 1 - r^\nu \quad \text{with } \nu > 1, \tag{37}$$

so that the boundary condition $d\phi/dr = 0$ at $r = 0$ is also satisfied. We insert (37) into (36) to obtain

$$g = \frac{1}{2} \frac{(2\nu + 1)(\nu + 1)}{2\nu - \sigma - 1}. \tag{38}$$

Since (38) is larger than the true eigenvalue γ_0 for all ν, we find the best approximation from $dg/d\nu = 0$. The result is

$$\nu = \frac{1}{2}[(\sigma + 1) + (\sigma^2 + 5\sigma + 6)^{\frac{1}{2}}] \approx \sigma + \frac{7}{4} \tag{39}$$

and

$$g \approx \sigma + \frac{10}{4}. \tag{40}$$

The quality of this result can be checked against the rigorous solution $\gamma_0 = \pi^2/4 = 2.467$ as compared to $g = 2.5$ for $\sigma = 0$. One can show that (40) converges on γ_0 for increasing σ. For $\sigma < 0$, (40) proves useful even beyond the regime of validity of the trial solution. For $\sigma = -1$, for example, the rigorous solution is[8] $\gamma_0 = 1.446$, while (40) yields $g = 1.5$. This case, by the way, is the solution of (26).

The trial solution (37) is useful also in the case that P_v of (24) has the more general form given by (17) and (22), but it becomes substantially more difficult to optimize ν. One can convince oneself that the final result (27) converges on the form derived in (40) in the limits $(D/H)^{\frac{1}{2}} \gg K_c$ and $(D/H)^{\frac{1}{2}} \ll K_c$.

REFERENCES

1. R. D. Maurer, private communication, 1972.
2. W. B. Gardner, "Microbending Loss in Optical Fibers," B.S.T.J., this issue, pp. 457.
3. W. A. Gambling, D. N. Payne, and H. Matsumura, "Gigahertz Bandwidths in Multimode, Liquid-Core, Optical Fiber Waveguide," Optics Communications *6*, No. 4 (December 1972), pp. 317–322.
4. D. B. Keck, "Observation of Externally Controlled Mode Coupling in Optical Waveguides," Proc. IEEE, *62*, No. 5 (May 1974), pp. 649–650.
5. L. G. Cohen and D. Gloge, "Mode Delay in Selfoc Fibers of the New Type," unpublished work.
6. D. Marcuse, "Pulse Propagation in Multimode Dielectric Waveguides," B.S.T.J., *51*, No. 7 (July–August 1972), pp. 1199–1232.
7. L. A. Galin, *Contact Problems in the Theory of Elasticity*. Edited by I. N. Sneddon. Translated by Mrs. H. Moss. Raleigh, N.C.: North Carolina State College, 1961.
8. R. Deutsch, *Nonlinear Transformations of Random Processes*, Englewood Cliffs, N.J.: Prentice-Hall, 1962, pp. 15–26.
9. D. Gloge, "Optical Power Flow in Multimode Fibers," B.S.T.J., *51*, No. 8 (October 1972), pp. 1767–1783.
10. D. Marcuse, "Losses and Impulse Response of a Parabolic Index Fiber with Random Bends," B.S.T.J., *52*, No. 8 (October 1973), pp. 1423–1437.
11. S. Timoshenko and S. Woinowsky-Krieger, *Theory of Plates and Shells*, New York: McGraw-Hill, 1959.
12. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, II*, New York: McGraw-Hill, 1953, p. 1115.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Resonant-Grid Quasi-Optical Diplexers

By J. A. ARNAUD and F. A. PELOW

(Manuscript received August 1, 1974)

Experimental results are reported concerning the transmission, reflection, and depolarization of metal grids that reflect an upper band, centered at 30 GHz, and transmit a lower band, centered at 20 GHz. With a single grid, the transmission loss is less than 0.1 dB, and the rejection exceeds 44 dB. Depolarization is measured under realistic conditions with direct dual-mode feed excitation. In a 10-percent band, depolarization is below 34 dB at all scanning angles for both the transmitted and reflected waves. Experiments with two parallel resonant grids are also reported.

I. INTRODUCTION

In many millimeter-wave systems associated with communication-satellite antennas or Hertzian cables,[1] quasi-optical filters and diplexers are attractive. Because of their large areas, quasi-optical devices have large power-handling capability and the multimoding problem is, in a sense, avoided. The ohmic losses can be small, and the grids are easy to manufacture by photographic techniques.

A simple type of diplexer is the plane-parallel Fabry-Perot resonator, incorporating parallel inductive grids and operating under oblique incidence.[2] The transmission of a plane-parallel Fabry-Perot resonator is essentially the same as that of a single-pole filter. This type of diplexer, however, suffers from the walk-off effects associated with the diffraction and lateral displacement of the incident beam.[3,4] This effect is aggravated if more than two grids are used to obtain a maximally flat response. The number of grids required, and therefore the walk-off effects, are minimized if the grids have resonant properties of their own. Narrow-band resonant crosses have been used in the far-infrared region.[5] No diplexing operation, however, was considered. The special features of the grid patterns considered here are their broadband characteristics and their capability of operating under oblique incidences. Preliminary results were reported in Ref. 6. In the present paper, new experimental results concerning the transmission, reflection, and depolarization of resonant-grid diplexers are reported. We give special attention to the depolarization of incident waves because, in

some applications, it is required that two orthogonally polarized channels be transmitted, and depolarization introduces crosstalk. An ideal grid does not cause depolarization of incident plane waves under normal incidence when the pattern is invariant under a 90-degree rotation. This useful property of square array grids does not hold, in general, for waves under oblique incidence. However, we have observed that the depolarization remains small for a particular orientation of the grid in its own plane.

Although the main features of the diplexer response are easy to understand, some details are not yet fully understood. This is the case for the sharp spurious dips in transmission observed at certain angles of incidence and for the coupling through evanescent fields. In most of our experiments we tried to avoid the evanescent coupling by misorienting the grids. This coupling mechanism would perhaps be useful if it were precisely understood.

II. TRANSMISSION AND REFLECTION

The transmission characteristics of resonant grids are described here in an essentially qualitative manner. We assume that the grid periods, $\mathbf{p}_1$, $\mathbf{p}_2$, have equal magnitude and are perpendicular to one another. If the wavefronts are plane, unlimited, and parallel to the plane of the grid, the electric field can be decomposed into two components, one parallel to $\mathbf{p}_1$ and one parallel to $\mathbf{p}_2$. If the grid pattern is invariant under a 90-degree rotation, the transmission is the same in amplitude and phase for these two components, and therefore no depolarization is suffered. The transmission curves in Fig. 1 are applicable to such grids. Because the periods are much smaller than the wavelength, the grids can be represented by lumped circuit elements. If the filter incorporates more than one grid, the fine field structure of one grid (space harmonics) is assumed to be negligible at the other grid location. When the grid spacing is small, it is advisable, as indicated before, to set the grids at a small angle to one another, of the order of 2 degrees, to prevent a spurious coupling from taking place.

Figure 1a shows a simple mesh. This grid can be represented by an inductance in parallel on a transmission line representing free space. The smaller the opening areas, the smaller the grid reactance. When two such grids are parallel to each other, with a spacing slightly less than $l\lambda_o/2$, where l denotes an integer, a resonance takes place that can be pictured as resulting from the wave being reflected back and forth between the two grids. At each reflection, a wave of small amplitude is transmitted. Because the round-trip path length is of the order of $l\lambda_o$, the waves transmitted at the successive passes are in phase and add up. If the system has a plane of symmetry and the losses can be

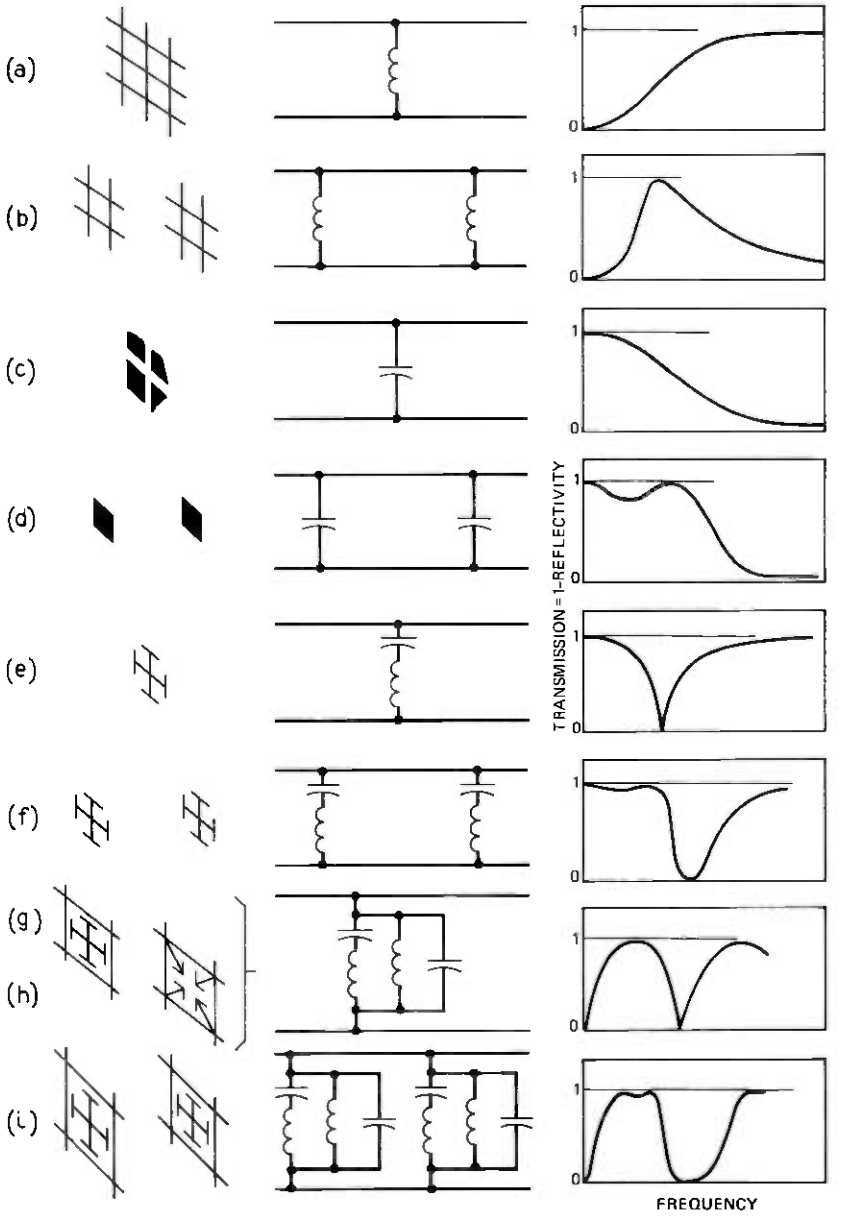


Fig. 1—Schematic representation of single-grid and double-grid diplexers. (a) Inductive grid. (b) Double-inductive grid. (c) Capacitive grid. (d) Double-capacitive grid. (e) Jerusalem cross (with rejection frequency). (f) Double Jerusalem cross. (g) Gridded Jerusalem cross. (h) Self-supported diplexer. (i) Double-gridded Jerusalem cross.

neglected, the transmission reaches 100 percent at one frequency at least. When the bandwidth is large, the response curve exhibits some dissymmetry, with a slower decay above resonance than below. If the metal and opened parts are exchanged, as shown in Figs. 1c and 1d, a capacitive grid is obtained, whose reflectivity is precisely equal to the inductive grid transmissivity. The square capacitive elements need, of course, to be supported by a dielectric sheet, perhaps in mylar. When the array period is small compared with the wavelength, it is very difficult to obtain even a moderately large reflectivity, such as $R = 0.9$. This is because the gap between squares required to provide such a reflectivity with a small period-to-wavelength ratio is of the order of a few micrometers. The power transmissivity is, in general,

$$T = [1 + (B/2)^2]^{-1} \tag{1}$$

for a susceptance, B, normalized to free space. For infinitely thin strips with a gap g and period p, we have ($p \ll \lambda$, thickness $\ll g$)

$$B = (2p/\lambda) \log_e [1/\sin (\pi g/2p)]. \tag{2}$$

It follows from eqs. (1) and (2) that if, for example, $\lambda = 10$ mm, $p = 3$ mm, a 10-percent transmissivity ($R = 0.9$, $B = 6$) requires a gap as small as 0.1 μm. This was one of our motivations for proposing a modified capacitive grid with an inductance in series with the capacitance, shown in Fig. 1e. These capacitive elements resemble "Jerusalem" crosses. At the resonance frequency, the Jerusalem-cross grid is perfectly reflecting and behaves as a plain sheet of copper. In a typical case, the measured transmission at the rejection frequency, 30 GHz, is at least 44 dB below the incident power.

Perfect transparency is obtained only at very low frequencies. To obtain a transmission band, two arrangements are considered. One consists in assembling two Jerusalem-cross grids parallel to each other. The response curve in Fig. 1f is obtained. A second possibility consists in introducing inductive elements in parallel with the crosses, as shown in Fig. 1g. The resulting grid is called a "gridded Jerusalem cross." An alternative configuration that does not require a mylar support but has essentially the same characteristics as the gridded Jerusalem cross is shown in Fig. 1h. These grids (type g or h) are very simple and attractive. Most of the experiments that we report were made on these types of grids. Finally, two Jerusalem-cross grids can be used, parallel to each other. Broad, uniform, transmission bands and broad, uniform, rejection bands are then obtained.

A grid, whatever its design, can be represented by circuit elements that are found empirically by fitting the measured response curve to the one calculated from the equivalent circuit. Sometimes, circuit ele-

ments that cannot be localized on the pattern need to be added to match the experimental curve. For instance, in the grid in Fig. 1g, a parallel capacitance needs to be added to account for a second, unexpected, transmission band above the rejection band. Even for a simple grid such as the one shown in Fig. 1a, a parallel capacitance needs to be added that tunes the inductance at the frequency c/p, p being the array period and c the speed of light in free space. Just below that frequency, the decay of the energy density of the space harmonics is very slow. This additional stored energy is represented by a capacitance. Above the frequency c/p, the transmission becomes a complicated function of frequency because of the excitation of grating lobes, and a circuit representation is of limited usefulness.

The transmission at resonance does not reach exactly 100 percent because of dissipation losses (ohmic losses in the metal and dielectric losses if mylar backings are used) and of scattering losses caused by the array not being perfectly periodic. The walk-off losses discussed in the introduction result from the incident beam being finite in cross section. For a single grid, the term "walk-off loss" is not applicable, but a similar physical effect exists. A beam with finite cross section has a finite angular spread, and a loss is suffered if the grid response depends significantly on the angle of incidence.

In a Fabry-Perot resonator incorporating conventional grids, the ohmic losses are small, particularly if the spacing between the grids is large (that is, if the axial mode number l is large). In contrast, high Q-factor resonant grids (e.g., narrow slits in a metal sheet) have rather high ohmic losses. Thus, resonant grids should be used only for broadband applications. Such grids are ideal, for example, for separating two channels widely separated in frequency, such as 20 and 30 GHz, or 4 and 6 GHz. Low-Q resonant grids are also useful in conjunction with conventional grids to eliminate side resonances in narrow-band filters. In any case, they provide greater flexibility in the filter design.

For convenient reference, let us give the expression for the susceptance, B, of the circuit shown in Fig. 1h (series L, C and parallel L', C')

$$B = (-L\omega + 1/C\omega)^{-1} + C'\omega - 1/L'\omega. \tag{3}$$

In terms of the reactances at the resonance angular frequency of the LC circuit, $\omega_o = (LC)^{-1}$; that is, with $X \equiv L\omega_o = 1/C\omega_o$, $X' = L'\omega_o$, $X'' = 1/C'\omega_o$, and with $f \equiv \omega/\omega_o$, B is

$$B(f) = X^{-1}(-f + f^{-1})^{-1} + f/X'' - 1/X'f. \tag{4}$$

This expression was used to generate the theoretical response curve in Fig. 7. For a single grid, this expression is to be substituted in eq. (1) to obtain the transmission T.

III. DEPOLARIZATION

The characterization of the depolarization introduced by a grid is discussed in this section. Let us first assume that the grid is an infinite plane and that the incident wave is plane. The normal to the wavefront of the incident wave and the normal to the grid plane define the angle of incidence i (= reflection angle). The orientation of the grid wires can be defined by the angle ν that they make with the normal to the incident plane in the clockwise direction for the direction of propagation. For a linear incident polarization, the polarization is defined by the angle Ω that the electric field makes with the normal to the incident plane, again in the clockwise direction. Arbitrary incident polarizations can be defined by their two components, along the normal to the incident plane and along the perpendicular direction. A wave is said to be E-polarized (or TM) if the magnetic field is linearly polarized and perpendicular to the incident plane ($\Omega = 90°$), and H-polarized (or TE) if the electric field is perpendicular to the plane of incidence ($\Omega = 0$). For given i, ν, the grid response is characterized by the complex transmission coefficients t_{EE}, t_{EH}, t_{HE}, and t_{HH}. These four parameters are, in general, functions of frequency. Because of linearity we have

$$\begin{aligned} e'_E &= t_{EE}e_E + t_{EH}e_H, \\ e'_H &= t_{HE}e_E + t_{HH}e_H. \end{aligned} \tag{5}$$

The parameters can be considered the elements of a complex 2-by-2 matrix t defined as

$$\mathbf{t} \equiv \begin{bmatrix} t_{EE} & t_{EH} \\ t_{HE} & t_{HH} \end{bmatrix}. \tag{6}$$

In terms of this matrix, (5) is written

$$\mathbf{e}' = \mathbf{te}. \tag{7}$$

Ideally, we would like to have $t_{EE} = t_{HH} \equiv t(\omega)$ and $t_{EH} = t_{HE} = 0$, in which case $\mathbf{e}' = t\mathbf{e}$, t being a scalar. We may require the less stringent condition that the state of polarization of the incident field be preserved to within an arbitrary, but fixed, rotation angle: $\mathbf{e}' = t\mathbf{Re}$, where $\mathbf{R}$ denotes a fixed rotation matrix and t a scalar function of ω. For this to happen, t must have the form $\mathbf{t} = t\mathbf{R}$. Reflection by an even number of plane mirrors, for example, preserves polarization in that sense. An even less stringent requirement is that orthogonal incident polarizations remain orthogonal. It can be shown[7] that this is the case if and only if t has the form $t\mathbf{U}$, with t a complex or real number and $\mathbf{U}$ a unitary matrix (that is, $\mathbf{U}^\dagger\mathbf{U} = \mathbf{1}$, where $\dagger$ denotes transposition and complex conjugation).

Symmetry considerations sometimes show that depolarization should not be present. Let us, for example, assume that the grid is a rectangular mesh, with a period vector perpendicular to the incident plane (that is, $\nu = 0°$, modulo 90°). By reason of symmetry, we must have $t_{EH} = t_{HE} = 0$. In general, however, $t_{EE} \neq t_{HH}$. Thus, for that case, E waves and H waves are not depolarized, but waves with any other polarization, for instance, a wave with linear polarization at 45 degrees, will be depolarized and acquire an elliptical polarization. To avoid depolarization, it is not sufficient that $|t_{EE}| = |t_{HH}|$. The phases of t_{EE} and t_{HH} must be equal, too. However, if the magnitudes of t_{EE} and t_{HH} are found equal over a large band of frequency, the phases of

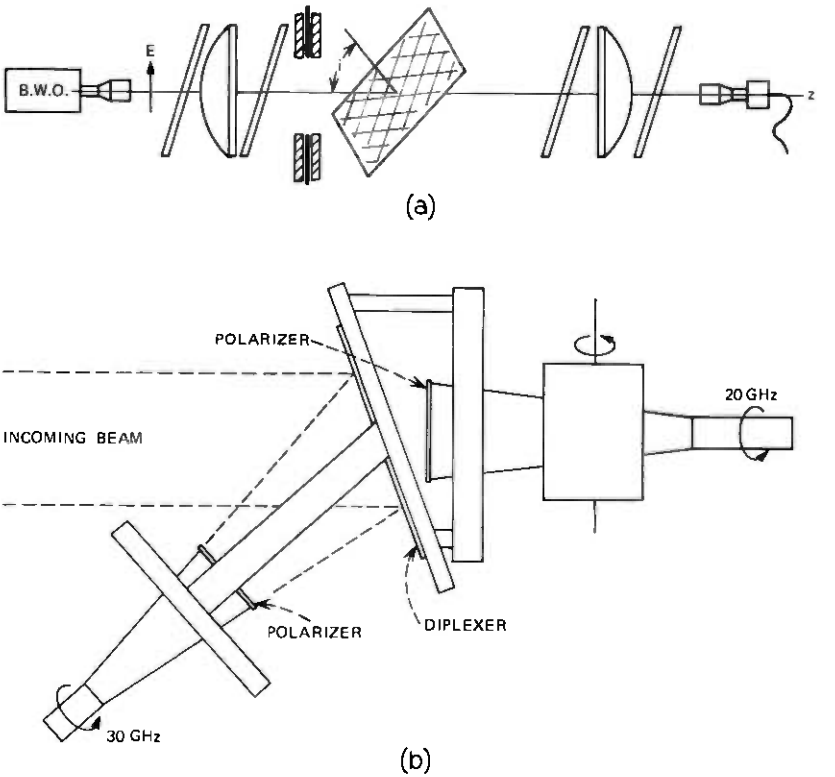


Fig. 2—Measurement system. (a) For near-plane wave excitation. The source is a backward wave oscillator feeding a dielectric lens (300-mm focal length and diameter). The system exhibits low depolarization (< -50 dB) when properly aligned. The incident electric field is vertical or horizontal. The grid under test can be rotated in its own plane. The angle ν refers to the angle between the grid wires and the normal to the incident plane. The angle of incidence (i) can be varied, and the grid assembly can be rotated by steps of 45 degrees about the z-axis (angle Ω) with the help of a 45-degree wedge. (b) Diplexer mounted with the feed.

t_{EE} and t_{HH} are most likely equal because of the integral relations existing between phase and amplitude for minimum phase circuits. Furthermore, if the grid can be considered very thin and lossless, there exists a simple relation between the phases and the moduli of the transmitted (t) and reflected (r) fields. The relation is

$$\begin{aligned} r &= -\cos\phi \exp(i\phi), \\ t &= -i \sin\phi \exp(i\phi). \end{aligned} \tag{8}$$

Thus, for thin grids, the phase of t can be obtained from the modulus of t at any frequency.

Another important result applicable to thin grids is that two-dimensional scaling applies. That is, nothing is changed if the wavelength and the grid dimensions in the plane are multiplied by the same factor. For thin grids, the Babinet principle also applies, which says that the reflectivity of a grid is equal to the transmissivity of its complement.

Let us now describe the experimental setup for plane wave measurement. The source (a backward wave oscillator) radiates through a dual-mode feed. The beam is collimated by a dielectric lens and sub-

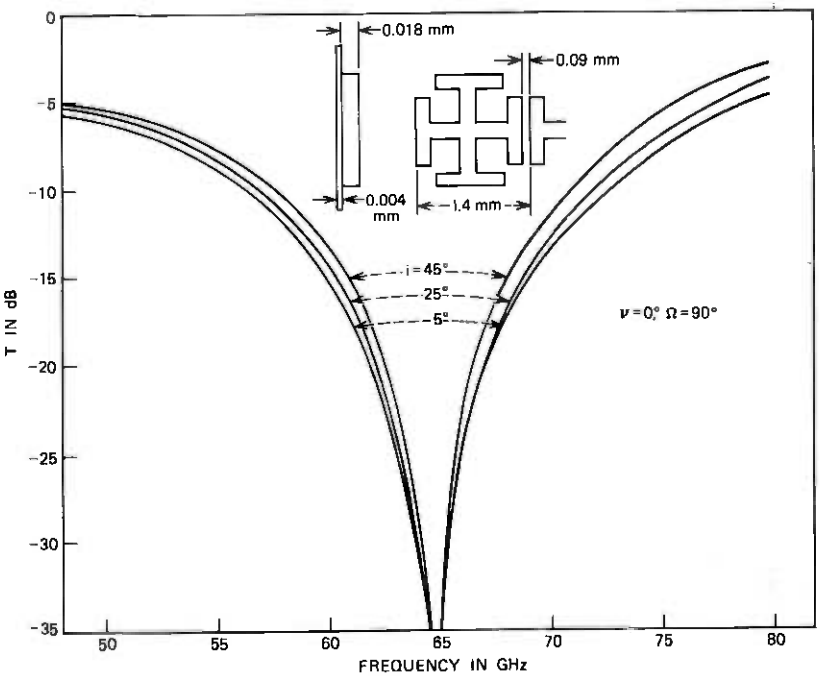


Fig. 3—Dimensions and response curve of a Jerusalem-cross grid. i denotes the angle of incidence ($\nu = 0°$, $\Omega = 90°$: horizontal electric field).

sequently focused by a second lens back on a collecting dual-mode feed. The size of the beam between the two lenses is sufficiently large that the field can be well approximated by that of a plane wave. To check the depolarization of the system, the collecting feed is rotated about its axis until a null is obtained. Because the system has a plane of symmetry, the cross-polarized components should be equal to zero. These components are found to be less than −50 dB. To damp the reflections that take place from the lenses and the feeds and between lenses, attenuating glass plates are introduced. These glass plates are tilted, but symmetry in the vertical plane is preserved. In this arrangement, the transmitted electric field must remain either vertical or horizontal. Otherwise, the tilted glass plates would depolarize the beam because of differential loss. Thus, the filter under test, rather than the source, needs to be rotated about the system axis (z) to measure its response for various polarization angles. In practice, it is sufficient to measure the depolarization for three angles, $\Omega = 0$, 45°, and 90°, and pick up the worst number. This maximum depolarization (X) is a function of the angle of incidence (i), of the angle that the grid wires

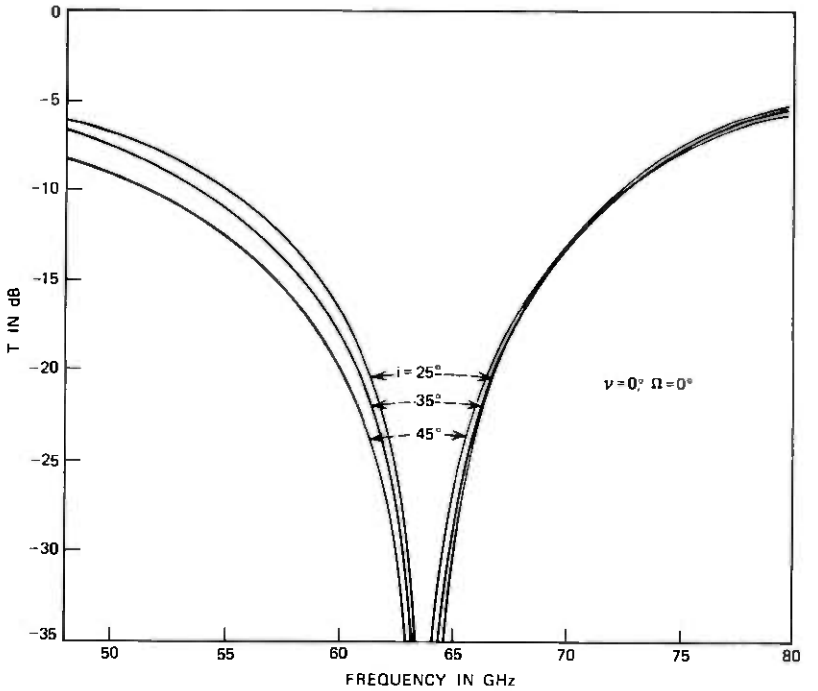


Fig. 4—Same as Fig. 3 with $\Omega = 0°$. For $\Omega = 45°$, the response (not shown) is found to be almost independent of i, for $i = 25°$ to 45°.

make with the normal to the incident plane (ν), and, of course, of frequency. Thus, the plane-wave depolarization of a diplexer in transmission is characterized by a function

$$X_t = X_t(i, \nu, \omega) \qquad \text{(dB)}.$$

The depolarization in reflection is similarly defined, but the mechanical arrangement is more complicated.

In some applications, the incident wave is diverging rather than plane. This is the case when the diplexer is used to separate beams just before the feed of a primary feed antenna, as shown in Fig. 2b. The definition of what constitutes a perfect feed from the point of view of polarization is not obvious. It has been observed that, if the feed is intended to be used at the focal point of a parabolic dish, its polarization pattern should be the same as that of an Huygens source, a combination of electric and magnetic dipoles.[8] The feeds used in this paper have relatively narrow beam patterns (about ±4 degrees at the 3-dB points), and it seems that the ambiguities associated with the definition

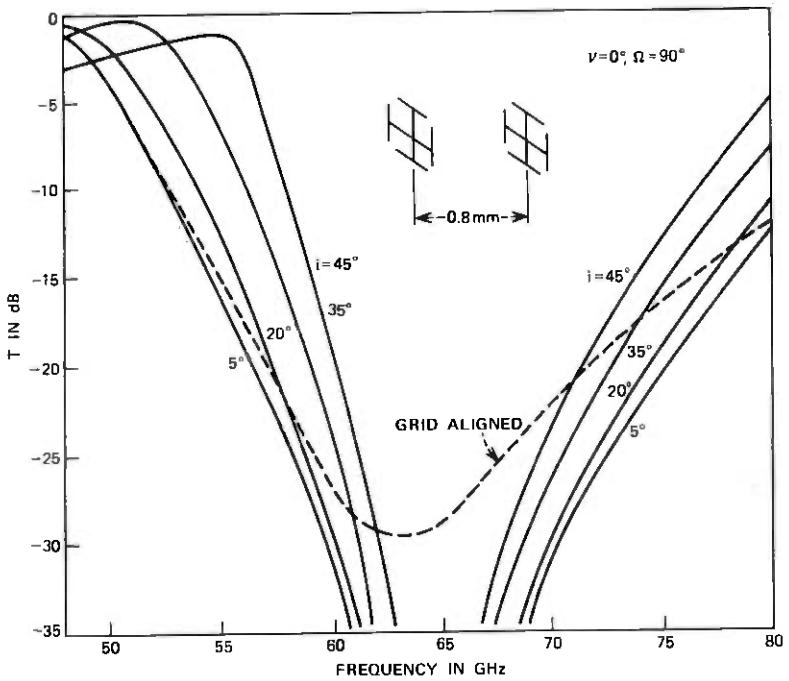


Fig. 5—Response curve of a filter incorporating two Jerusalem-cross grids as in Fig. 3, with a 0.8-mm spacing. The grids are at a small angle (~3°) to one another to avoid evanescent wave coupling. When the two grids are aligned, the response shown by a dashed line is obtained. For all curves, $\Omega = 90°$ (horizontal electric field).

of what constitutes a "good" feed can be ignored. The test thus simply consists of scanning the diplexer assembly in azimuth for various incident polarizations (0, 45, and 90 degrees) and measuring the cross-polarized component.

IV. GRID FABRICATION

The grids were made either in beryllium copper (75-μm thick) or from a sheet of copper bound on a mylar film. The fabrication consists of generating a mask from a computer program and using conventional photoetching techniques.

To generate the mask, three plotting techniques were used, all computer-driven. For simple grid patterns, the most convenient system seems to be the "Litehead" plot, which is a scanned focused beam of light. It was used for fabricating polarizers. For the complicated patterns considered in this paper, we used the "Rubylith" plot, which uses mechanical cuts. Because of computer limitations, only moderate array sizes can be obtained. To obtain large array sizes, it is necessary to use a "step-and-repeat" camera technique. This photographic technique has size limitations, and further mask joining is required.

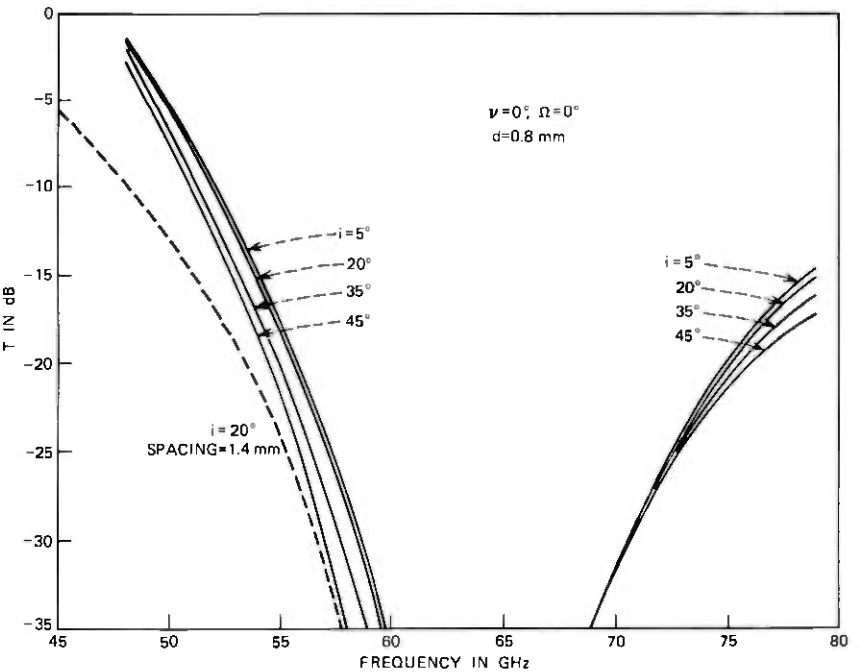


Fig. 6—Continuation of Fig. 5 with $\Omega = 0°$ (vertical electric field). The dashed line is for a 1.4-mm spacing.

An alternative technique is the PPG (primary pattern generator) system. This system is a raster-scan-type device that generates 190-mm by 240-mm arrays with a 7-μm by 7-μm resolution. This is the technique used to generate the grids whose depolarization is shown in Figs. 9 to 12, 15, and 16, except for the dashed line in Fig. 9.

V. THE JERUSALEM-CROSS GRID

The dimensions of the crosses of a typical Jerusalem-cross grid and the response curve are shown in Figs. 3 and 4 (E and H polarizations) for various angles of incidence. The resonance frequency of the series

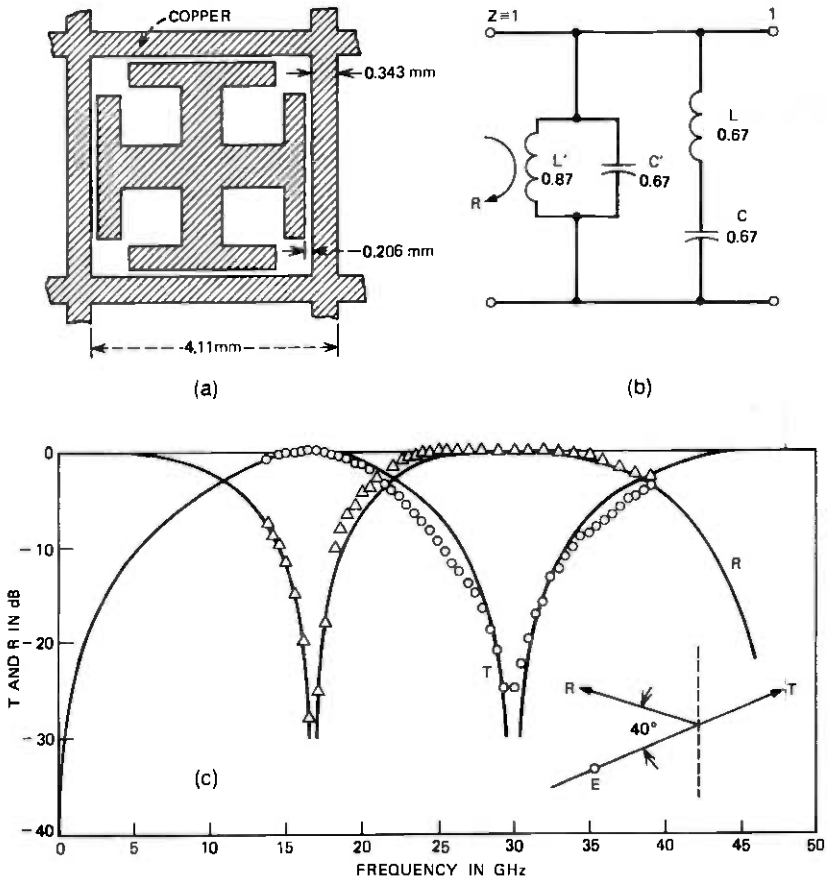


Fig. 7—Gridded Jerusalem-cross diplexer. The critical dimensions are shown in (a) and the equivalent circuit obtained by fitting the measured transmission and reflection curves in (c) is shown in (b). The peak transmission (loss < 0.1 dB) is at 16 GHz, and the peak rejection (> 44 dB) is at 30 GHz. The plain curve in (c) is theoretical from the equivalent circuit.

LC circuit is at 65 GHz. We note that the resonance frequency is almost independent of the angle of incidence.

The transmission characteristic of a pair of such grids, with a 0.8-mm spacing, at various angles of incidence, is shown in Figs. 5 and 6. At an angle of incidence of 35 degrees, for example, the transmission reaches its maximum at a frequency of 51 GHz. The rejection exceeds 30 dB from 61 to 69 GHz. In this experiment, the two grids are slightly rotated with respect to one another (about 3 degrees). The dashed curve in Fig. 5 shows that, when the grids have exactly the same orientation, the rejection is smaller. This probably results from a coupling through evanescent fields.

VI. THE GRIDDED JERUSALEM CROSS

A single grid exhibits a transmission band if parallel inductances are added to the series circuit. The grid dimensions and the measured response, both in transmission and in reflection, are shown in Fig. 7 (see also Ref. 6). The equivalent circuit has been selected to match the experimental response curve. The curve in Fig. 8 shows the effect of

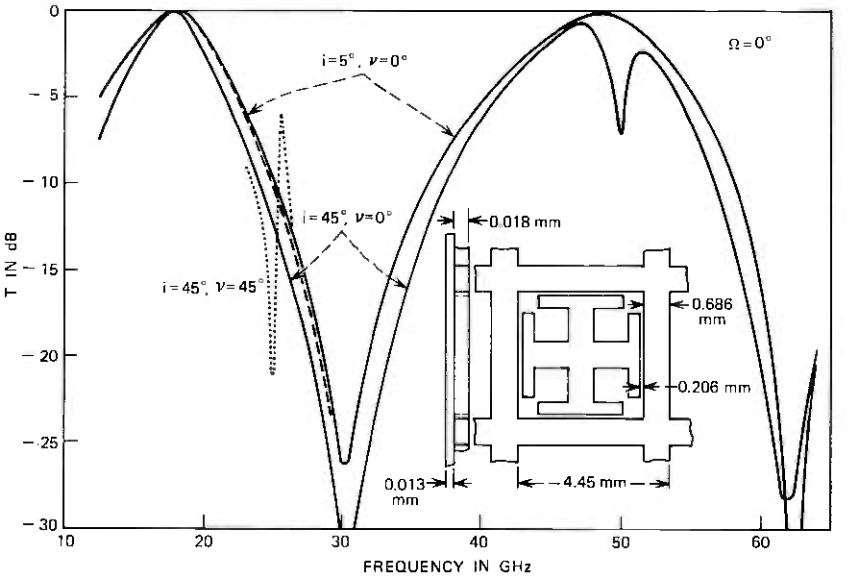


Fig. 8—Gridded Jerusalem-cross diplexer. This curve shows that the transmitted frequency can be raised from 16 to 18 GHz by increasing the width of the parallel wires. The transmitted curves are for angles of incidence i of 5 and 45 degrees, $\nu = 0°$. The small-dash curve is for the grid located between the feed (two wavelengths across) and the lens, that is, under strongly diverging wave excitation. The transmission curve is essentially unaffected. The curve for an angle of incidence of 45 degrees and the grid rotated in its own plane at $\nu = 45°$ exhibits a spurious dip.

increasing the width of the parallel strips from 0.343 to 0.686 mm. This brings the passband closer to the rejection band. By increasing the width of the parallel grids further to 0.840 mm and reducing the capacitive gap to 0.126 mm and the series inductance strip width to 0.336 mm, we found it possible to obtain a ratio of rejection frequency to transmission frequency as low as 1.5:1. An approximate theoretical analysis of the operation of the gridded Jerusalem cross under normal incidence will be reported shortly.[9] At a large angle of incidence and for some orientation of the grid, a sharp dip is observed. This dip is observed when the electric field is at 45 degrees to the series inductance wires. (The same rule applies to the "self-supported grid" discussed in Section VII.) A possible mechanism is the following. Consider as a simpler model an array of parallel strips with series capacitances. Because the phase velocity along the strips exceeds the velocity of light in free space, this system can, in principle, radiate. However, if the symmetry is perfect, the relevant space harmonic has zero amplitude. Thus, a very small lack of symmetry is needed to have radiation. The Q-factor of this resonance can be very high because the coupling to

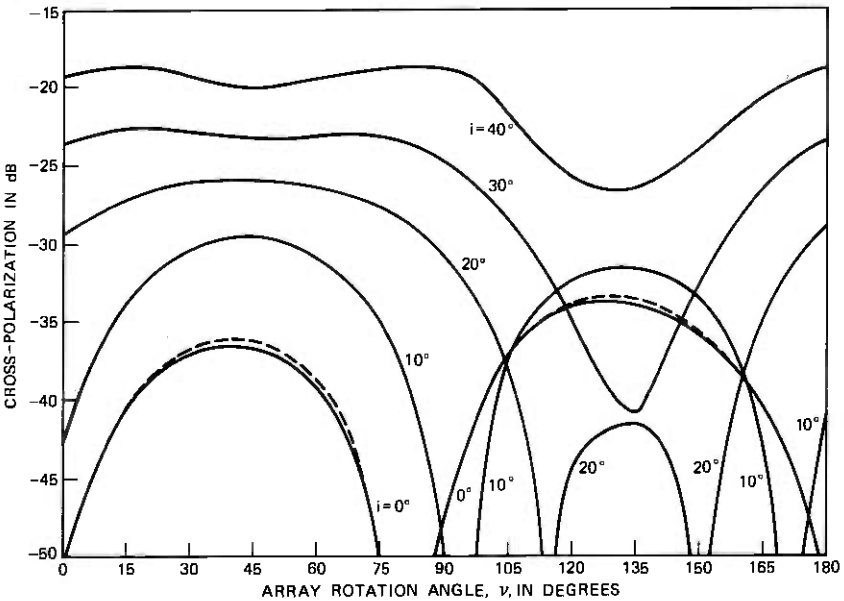


Fig. 9—Depolarization introduced by the gridded Jerusalem-cross diplexer (diplexer obtained from a PPG mask) in Fig. 8 under near-plane wave excitation for different angles of incidence (i), as a function of the rotation of the grid in its own plane (angle ν). The incident polarization is at 45 degrees to the incidence plane ($\Omega = 45°$). The dashed line is for joint masks (Rubylith plot).

incident plane waves is very small. Whether this model is applicable to the grids investigated remains to be seen.

The depolarization introduced by a grid of this type has been measured, both under plane wave excitation and under diverging wave excitation. The depolarization under plane wave excitation is shown in Fig. 9 as a function of the orientation of the grid in its own plane. It should be noted that, even under normal incidence, the depolarization is not zero and, under oblique incidence, the curve does not have a 90-degree period, contrary to our expectations based on the nominal symmetry of the grid. The first grid that we tested (whose response curve is shown as a dashed line in Fig. 9) was obtained from a composite mask made of two smaller masks. A slight discontinuity between the two masks was noted, which was thought to be responsible for the

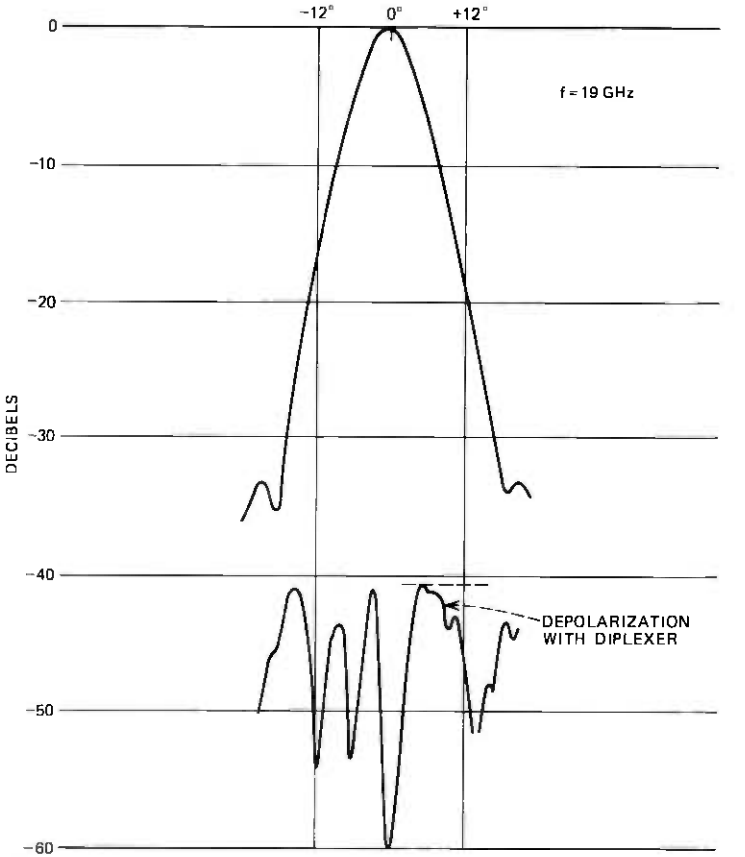


Fig. 10—Typical azimuth scan showing the transmission and depolarization of the diplexer in Fig. 8, at 19 GHz.

depolarization under normal incidence (34 dB maximum). However, the depolarization for a grid obtained from a single mask (PPG generator) is almost the same. The effect of diffraction at the edges of the filter has been eliminated as a possible cause for depolarization. No systematic lack of symmetry of the grid under a 90-degree rotation is noticeable under microscopic observation. The residual depolarization that we observe under normal incidence for the PGG grid, therefore, remains unexplained. A second observation is that excellent cross-polarization properties are obtained for angles of incidence less than 20 degrees for some orientations of the grid in its own plane (e.g., cross-polarization < -50 dB for $\nu = 110°$, $i = 20°$).

More complete tests were made in an anechoic chamber with dual-mode feed horns at 20 GHz (transmitted beam) and 30 GHz (reflected

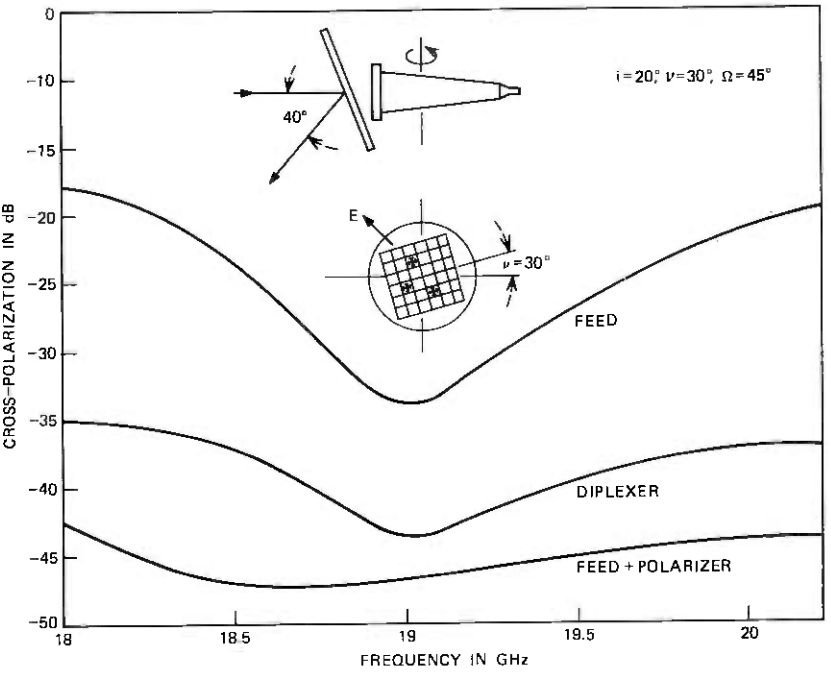


Fig. 11—Depolarization introduced by the gridded Jerusalem-cross diplexer in Fig. 8 under diverging wave excitation as a function of frequency. The feed diameter is 130 mm (3 dB points of the far-field pattern at ±5° at 19 GHz). The angle of incidence is 20 degrees. At each frequency, the diplexer (grid and feed together) is scanned in azimuth. The points shown correspond to the worst depolarization within a ±12° angle, corresponding to the −18-dB points of the far-field radiation pattern of the feed. The incident polarization is at 45 degrees to the incident plane. The orientation of the grid in its own plane ($\nu = 30°$) is shown in the figure. The upper curve is for the dual-mode feed alone. The lower curve is for the feed mounted with a grid polarizer. The central curve gives the depolarization introduced by the diplexer and feed combination.

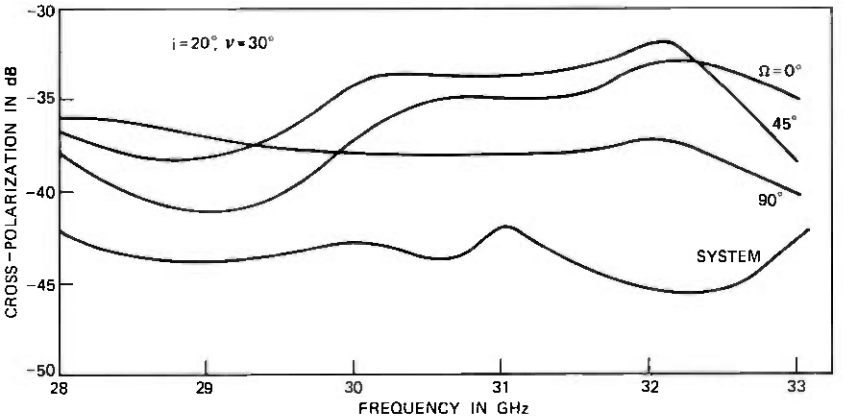


Fig. 12—Depolarization introduced on reflection in the 30-GHz band by the gridded Jerusalem-cross diplexer in Fig. 8. Measurements were made for the same grid orientation as in Fig. 11 for incident polarizations corresponding to angles $\Omega = 0$, 45°, and 90°. The worst depolarization within a $\pm 12°$ scanning angle, corresponding to the −22-dB point in the feed response at 30 GHz, is selected. The curve labeled "system" is for the dual-mode feed-mounted with a polarizer.

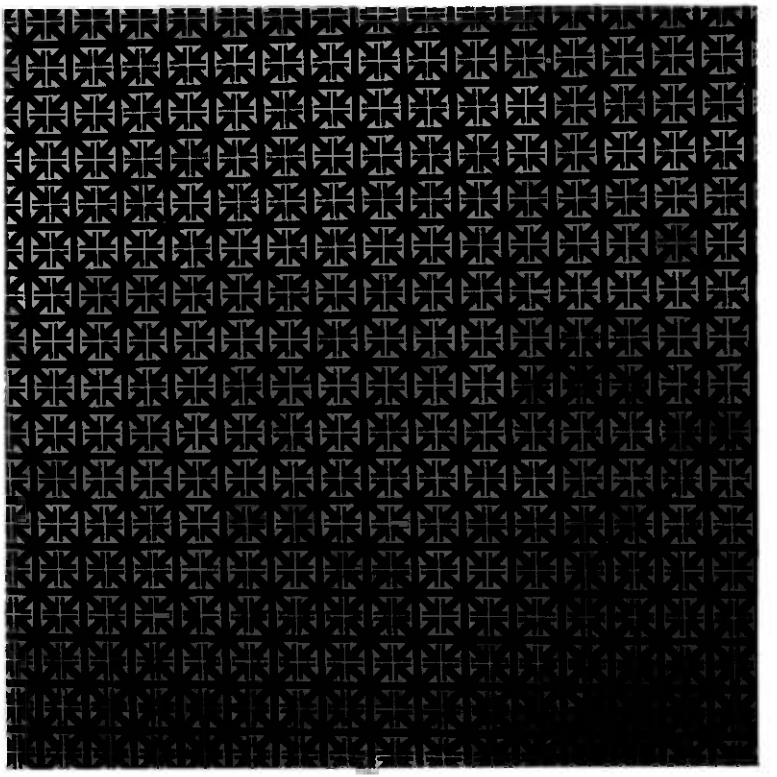


Fig. 13—Photograph of a self-supported diplexer in beryllium copper.

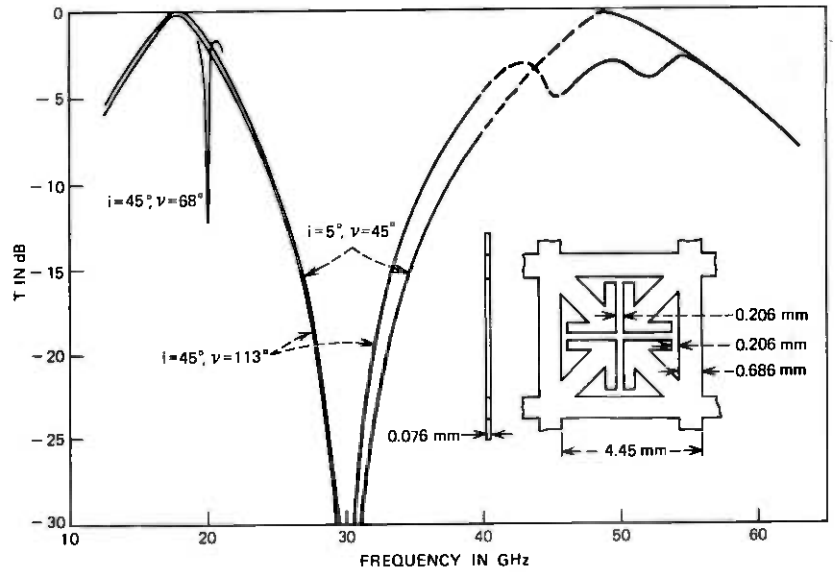


Fig. 14—Transmission of a self-supported grid in beryllium copper whose dimensions are shown on the figure. At large angles of incidence (i = 45°) and for some orientations of the grid, a large dip appears in the response.

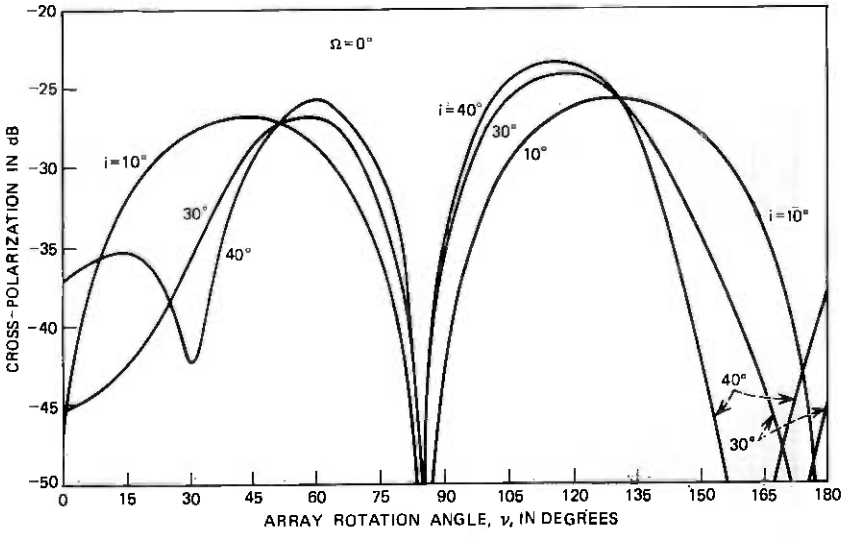


Fig. 15—Depolarization introduced by the self-supported grid in Fig. 13 as a function of the orientation of the grid in its own plane (angle ν) for various angles of incidence. The frequency is 19.5 GHz, the beam diameter is 130 mm, and the incident polarization is perpendicular to the incidence plane (Ω = 0°).

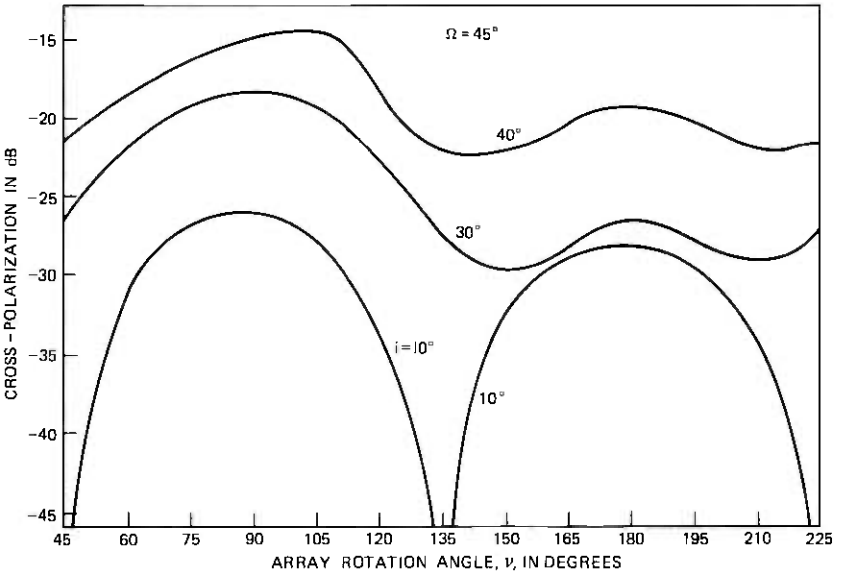


Fig. 16—Continuation of Fig. 14 with the incident polarization at 45 degrees to the normal to the incident plane ($\Omega = 45°$).

beam). The orientation of the grid is selected for minimum depolarization in transmission at 19 GHz. It was noted that operation of a single-grid diplexer under diverging wave conditions does not significantly degrade the system response. The diplexer assembly was scanned within a ± 12-degree angle (± 12 degrees correspond to the -18-dB points of the feed response at 19 GHz). A typical scan is shown in Fig. 10. The worst depolarization is plotted in Fig. 11 as a function of frequency. The incident polarization was at 45 degrees to the incident plane. The depolarization is less than -35 dB from 18 to 20 GHz. The depolarization on reflection, shown in Fig. 12, turns out not to be sensitive to the orientation of the grid in its own plane. This is fortunate, since there is no particular reason to expect the optimum orientation of the grid to be the same on reflection and on transmission. The depolarization on reflection was measured for three incident polarizations, $\Omega = 0°$, $+45°$, and $90°$. Considering only the upper envelope of these curves, we find that the depolarization on reflection is less than -34 dB from 28 to 31 GHz.

In conclusion, we find that the Jerusalem-grid diplexer, operating with a 40-degree angle between reflected and incident beams, gives transmission and reflection losses less than 0.1 dB. The depolarization is less than -34 dB within 10-percent bands centered at 18 and 30 GHz.

VII. THE SELF-SUPPORTED DIPLEXER (Fig. 13)

The most critical dimensions of the self-supported grid are shown in Fig. 14. The equivalent circuit is almost the same as that of the gridded Jerusalem-cross grid previously discussed, but this new grid does not require a mylar backing. The grid was made of beryllium copper, 75-μm thick. The transmission characteristic is shown in Fig. 14 as a function of frequency. Here again, a sharp dip in transmission shows up for some orientations of the grid. The depolarization introduced by the self-supported diplexer is shown in Figs. 15 and 16 for quasi-plane wave excitation. We observe that a very small depolarization can be obtained for a proper orientation of the grid in its own plane, if the angle of incidence does not exceed 20 degrees.

VIII. DOUBLE-SELF-SUPPORTED DIPLEXER

An almost flat response can be obtained from 17 to 19.5 GHz by combining two self-supported grids (shown in Fig. 13) with a 6.3-mm

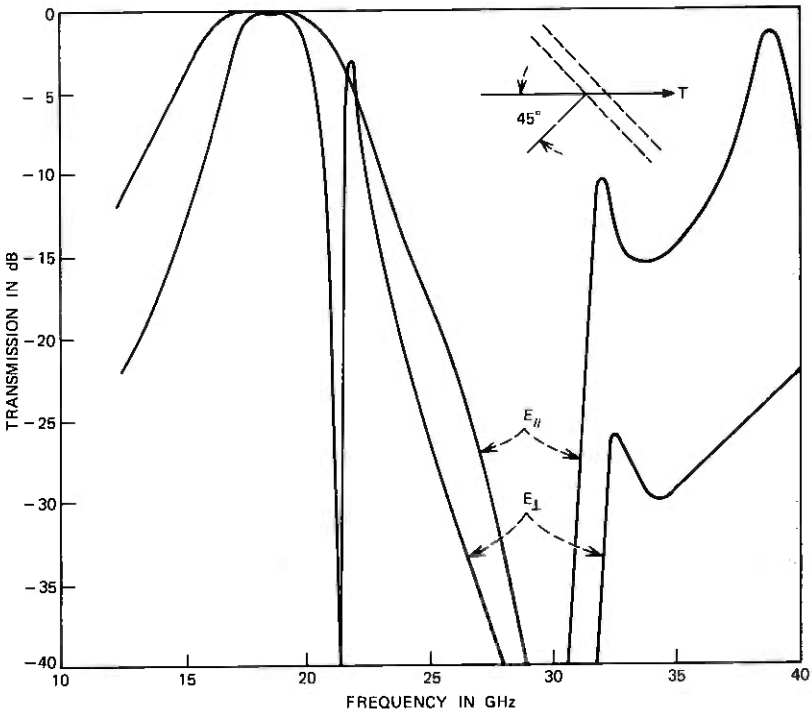


Fig. 17—Transmission of a double-self-supported diplexer with a 6.3-mm spacing at 45-degree angle of incidence, for two polarizations. The grid mask is as shown in Fig. 14. The grids were etched on 76-μm-thick pure copper. Grid size = 300 × 460 mm. Depolarization at 19 GHz is below −30 dB and the peak rejection at 30 GHz exceeds 60 dB.

spacing (Fig. 17). A broad rejection band is also obtained, with a peak rejection of −68 dB.

IX. CONCLUSION

We have shown that a single metallic grid may constitute an efficient diplexer to separate or combine frequency bands in the ratio 1.5:1 or more. The depolarization is low and could be reduced further by selecting unequal periods and dimensions. Because a very accurate theory is not presently available for the complicated patterns that we have investigated, an optimum design would require further testing. For a two-resonant grid diplexer, flat responses are obtained, and the depolarization remains acceptable for most applications. Some of the effects observed: depolarization under normal incidence, sharp dips in transmission at some orientations of the grids, and evanescent coupling, though understandable in principle, remain to be investigated.

X. ACKNOWLEDGMENTS

The authors acknowledge stimulating discussions with A. A. M. Saleh and the assistance of J. T. Ruscio and W. E. Legg. The high-gain dual-mode feeds were provided to us by T. S. Chu and M. J. Gans. Assistance from J. H. Corbin and L. West for the graphic programming of masks is gratefully acknowledged.

REFERENCES

1. J. A. Arnaud and J. T. Ruscio, "Guidance of 100 GHz Beams by Cylindrical Mirrors," to be published in IEEE Trans. on Microwave Theory and Techniques, April 1975.
2. J. A. Arnaud, "Quasi-Optical Channel Dropping Filters at Millimeter Wavelengths," June 1970, unpublished work.
3. J. A. Arnaud, A. A. M. Saleh, and J. T. Ruscio, "Walk-Off Effects in Fabry-Perot Diplexers," IEEE J. of Microwave Theory and Techniques, *MTT22*, No. 5 (May 1974), p. 486. A different type of Fabry-Perot diplexer was reported by A. A. M. Saleh, *op. cit.*, *MTT22*, No. 7 (July 1974), p. 728.
4. J. A. Arnaud, "Hamiltonian Theory of Beam Mode Propagation," *Progress in Optics*, Vol. 11, E. Wolf ed., Amsterdam, The Netherlands: North-Holland, 1973.
5. R. Ulrich, Infrared Physics, *7*, 1967, pp. 37 and 65.
6. J. A. Arnaud and J. T. Ruscio, "A Resonant Grid Quasi-Optical Diplexer," Electronics Letters, *9*, No. 25 (Dec. 13, 1973), p. 589.
7. A. A. M. Saleh, private communication.
8. A. C. Ludwig, "The Definition of Cross-Polarization," IEEE Trans. on Ant. and Propag., *AP21*, No. 1 (January 1973), p. 116.
9. I. Anderson, unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

An Optical-Frequency Pulse-Position-Modulation Experiment

By W. S. HOLDEN

(Manuscript received April 3, 1974)

This paper describes an optical-frequency pulse-position-modulation experiment using a GaAs luminescent diode as the source and either a PIN or an avalanche photodiode as the detector. The experimental system transmits an audio band from 300 Hz to 3.4 kHz at an 8-kb/s repetition rate. Timing synchronization between the transmitter and receiver has been accomplished by two methods: by transmitting a separate clock signal and by recovering the timing from the PPM signal itself. Data show, with each scheme, peak-signal-to-RMS-noise ratios of 70 dB can be achieved with the required average optical power at the receiver being −73 dBm with a PIN detector and −88 dBm with an avalanche detector.

I. INTRODUCTION

Recent advances in low-loss optical fibers,[1] in solid-state photodetectors,[2] and in optical-frequency (O.F.) power sources[3,4] have stimulated interest in O.F. communication systems for numerous applications. Pulse-position modulation (PPM) is particularly attractive for O.F. communications because the optical energy source can be operated at a low, message-independent, duty cycle to extend the lifetime of the device, and the technique affords a high noise immunity to the types of noise that dominate in a well-designed optical receiver. By employing a short pulse width and thereby expanding the bandwidth of the transmitted signal, the effect of detector leakage noise (dark current) and some forms of amplifier noise are reduced.

This paper describes an experiment performed to evaluate the performance that can be achieved in transmitting a single message channel by means of optical PPM.* Included are descriptions of the following:

* In an independent work (Ref. 5), a transmitter and receiver are described for an optical PPM system in which timing information was provided by a reference pulse in each time slot. The signal-to-noise ratio performance of the receiver was not given.

(*i*) The modulator that encodes the audio signal into a PPM signal.
(*ii*) The optical link using a GaAs LED as the source and either a PIN or an avalanche photodetector at the receiver.
(*iii*) The detection circuitry employing high-impedance front-end amplification techniques.[6,7]
(*iv*) The demodulator that transforms the PPM back to the original audio signal.
(*v*) The timing recovery scheme implemented using a phase-locked loop with the phase discriminator being the demodulator itself.
(*vi*) The results of experiments performed to measure the system performance.

II. THE OVERALL SYSTEM

Figures 1a and 1b are block diagrams of the overall transmitter and receiver, respectively. Included are diagrams of the associated waveforms.

An audio signal, filtered to limit the transmitted bandwidth from 300 Hz to 3.4 kHz (3-dB points), is applied to a "sample-and-hold" circuit that transforms it into a staircase waveform having a step width of 125 μs. This staircase and an *inverted* 8-kHz sawtooth are applied to a comparator. The comparator output is high during the interval in which the sawtooth amplitude is greater than the staircase and low when the staircase amplitude exceeds that of the sawtooth. This results in pulse-width modulation (PWM) at an 8-kb/s repetition rate with the falling edge of each pulse varying in position within a time slot; this position relates directly to the audio signal amplitude. By adjusting the DC offset on the sawtooth, the trailing edge of the pulse is set to occur at the midpoint of the time slot when there is no audio input. This position corresponds to the zero reference.

Pulse-position modulation is obtained from PWM by producing a narrow pulse each time the comparator goes from a high to a low state. The timing components for the device were chosen to produce an output pulse width of the order of 0.5 μs. This results in a duty cycle of 0.4 percent, which is compatible with present state-of-the-art, large-optical-cavity, solid-state lasers.[8] (These lasers are capable of being operated with up to a 1-percent duty cycle.)

The PPM signal is amplified and applied to a GaAs light-emitting diode (LED) by means of a driver stage. The driver is capable of applying 500-mA pulses to an LED. During this experiment, however, the LED was driven by pulses having peak amplitudes between 200 and 300 mA.

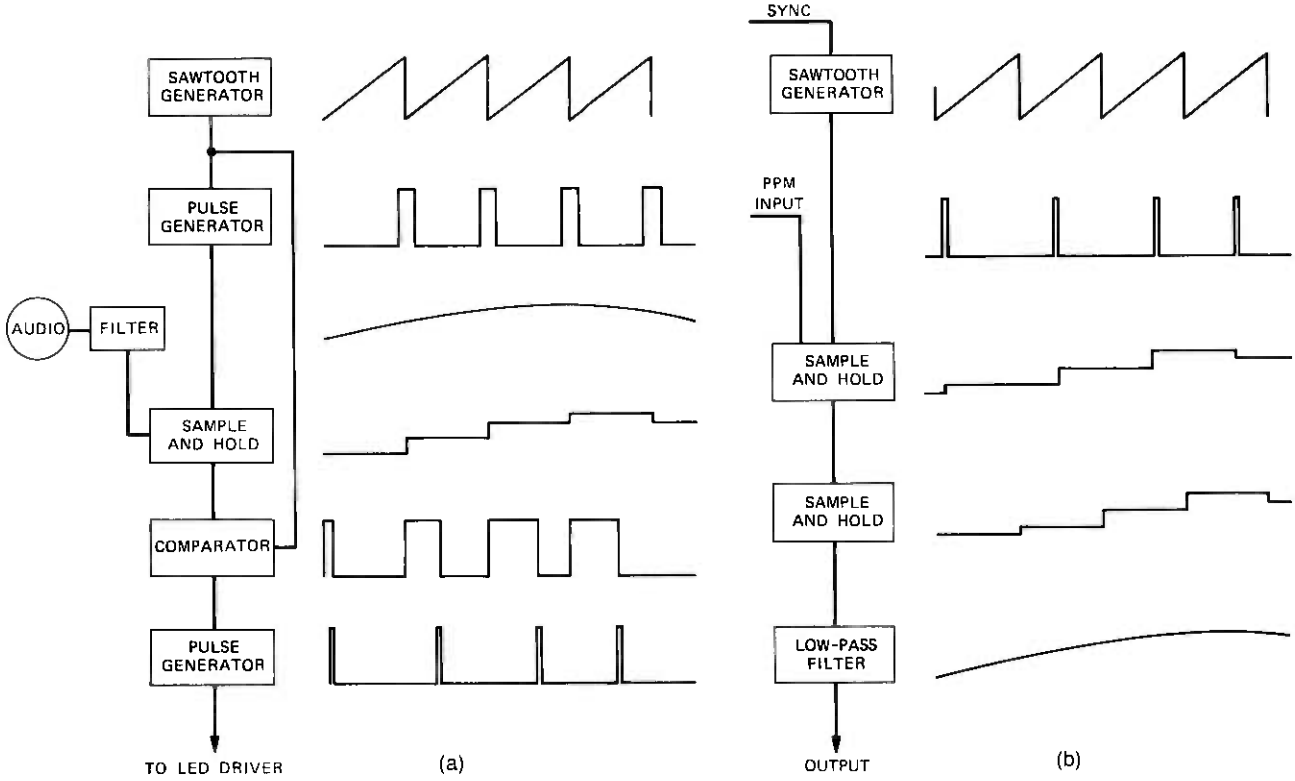


Fig. 1—(a) Block diagram and waveforms associated with the PPM transmitter. (b) Block diagram and waveforms associated with the PPM receiver.

In this experiment, the optical signal traverses an air path. Fiber loss is simulated by neutral density filters (providing large amounts of attenuation) and a pair of crossed polarizers (providing small but precise changes in attenuation). At the receiver, the signal is focused onto a photodetector, either a PIN or an avalanche device. The external quantum efficiency (amperes/watt) for each device has been measured. From this, by monitoring the average current through the device, the received average optical power may be obtained.

The detected signal, before demodulation, is processed by the following three stages:

(*i*) High-impedance front-end amplification—This approach is similar to that taken by J. E. Goell in his 6.3-Mb/s repeater research.[7] Figure 2 is a block diagram of the three-stage amplifier used. The first stage provides a high impedance for the photodetector, the second provides variable gain, and the third, a source follower, is provided to decouple the input amplifier from subsequent circuits.

(*ii*) Equalization—To compensate for the distortion introduced by the long input time constant, the circuit of Fig. 3 is used. The equalizer, a series capacitor and a resistance shunted to ground, is wired between two video amplifiers.

(*iii*) Filter—This stage, a two-section, low-pass Butterworth filter, limits the bandwidth of the signal pulse and noise. Two emitter-follower stages supply the proper impedances at the input and output (Fig. 4).

The PPM signal from the detection circuitry described above is applied to one input of a comparator. The other comparator input, which determines the threshold of the device, is connected to a variable DC source. This threshold was set with the received optical signal at the minimum value for which the demodulated audio output, when viewed on an oscilloscope, showed no distortion or noise. This method is justified because, for signal levels slightly higher than this value, the signal-independent noise is dominant.

The demodulator consists, basically, of two sample-and-hold circuits and a low-pass filter. The first sample-and-hold circuit is triggered by the PPM signal and samples a sawtooth wave. The output is a staircase with the following properties:

(*i*) The amplitude of each step varies as a function of the position of the received pulse in the time slot.

(*ii*) The step width varies according to the PPM signal.

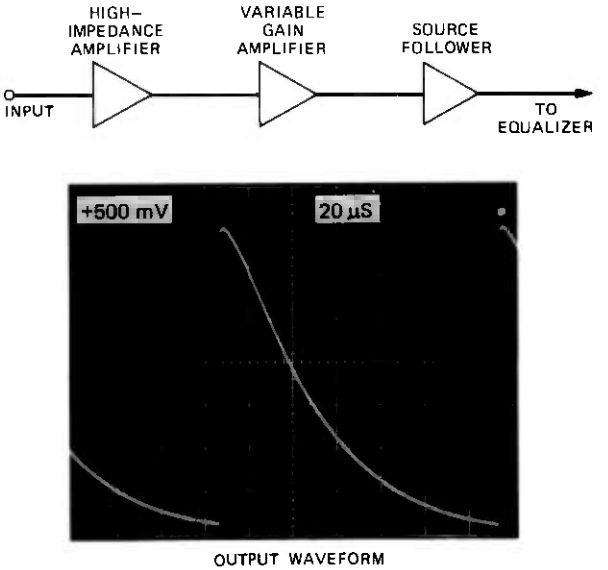


Fig. 2—High-impedance front-end amplifiers and output waveform.

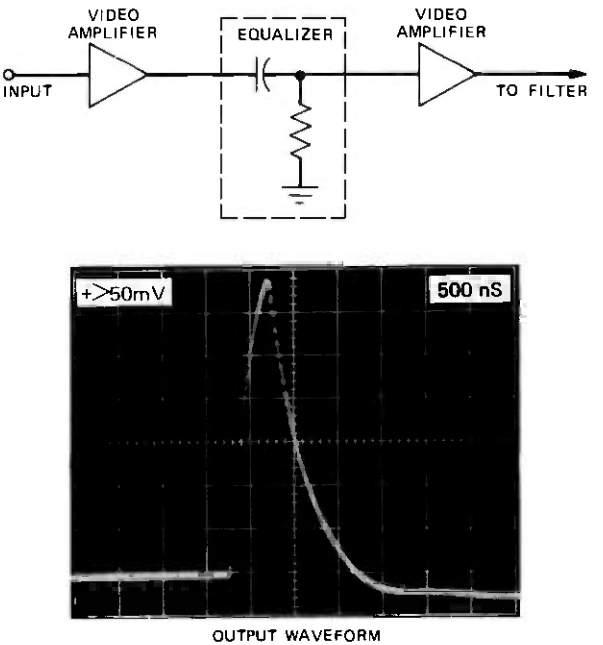


Fig. 3—Equalizer stage and output waveform.

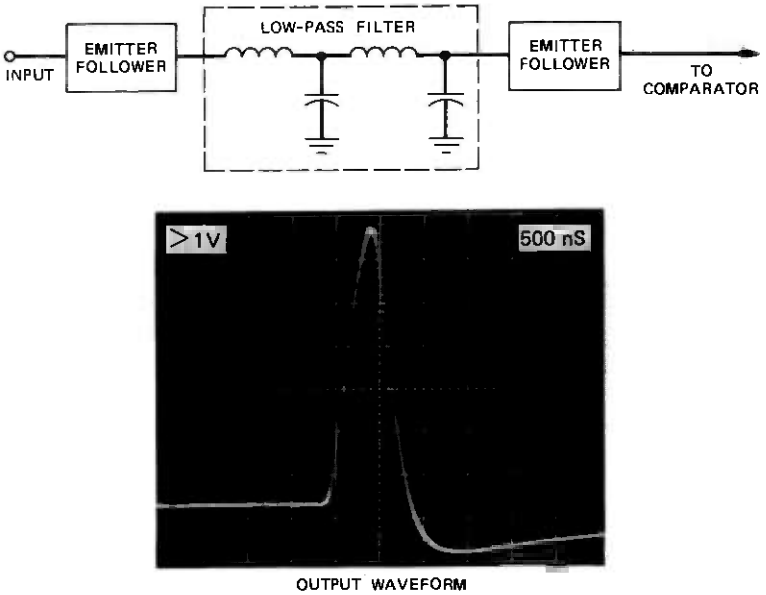


Fig. 4—Low-pass filter stage and output waveform.

Constant step width is achieved by using a second sample-and-hold circuit triggered by a repetitive 8-kb/s pulse train.

Filtering this staircase waveform removes the high-frequency components introduced by the sampling circuits and yields the original audio signal.

The timing recovery scheme incorporated in the receiver is described in the appendix.

III. EXPERIMENTAL PROCEDURE AND RESULTS

This section describes the experiments performed to measure the channel's performance.

For the noise measurements, additional filtering was provided to suppress the third harmonic of the 60-Hz supply voltage, the high frequencies introduced in sampling (in the modulator and demodulator), and out-of-band thermal noise. The filter employed, a Rockland Model 1100, was set for a Butterworth high-pass response and a Bessel low-pass response. The passband (300 Hz to 3.4 kHz) has a ripple of ± 0.5 dB and a 24-dB/octave rolloff.

A RMS voltmeter was used to measure peak-signal-to-RMS noise ratios* at the receiver output. The baseband signal level was measured

* This ratio will be referred to as SNR.

and converted to its corresponding peak value for the largest possible signal pulse excursion transmitted through the channel. The RMS noise was recorded directly, being measured with no audio signal applied. The optical pulse in all cases has a 20-ns rise time and a 3-dB width of 500 ns.

3.1 PIN detector

Figure 5 shows the SNR's in dB plotted as a function of received average optical power. Four bandwidths were implemented by the low-pass filter in the detection circuitry which sets the bandwidth for the detected signal pulse and noise; the 3-dB widths are 5.5 MHz, 1 MHz, 600 kHz, and 200 kHz. These curves exhibit three regions: rapid increase, gradual increase, and no increase in SNR with increasing optical power.

For optical signal levels of the rapidly increasing region, the output noise is primarily due to spurious threshold crossings caused by front-end amplifier thermal noise. This noise may also add to the signal pulse such that it will not exceed the predetermined level.

In the gradually increasing region, the predominant degradation is due to time jitter on the transmitted pulse, originating in the transmitter. The front-end noise also leads to an uncertainty in determina-

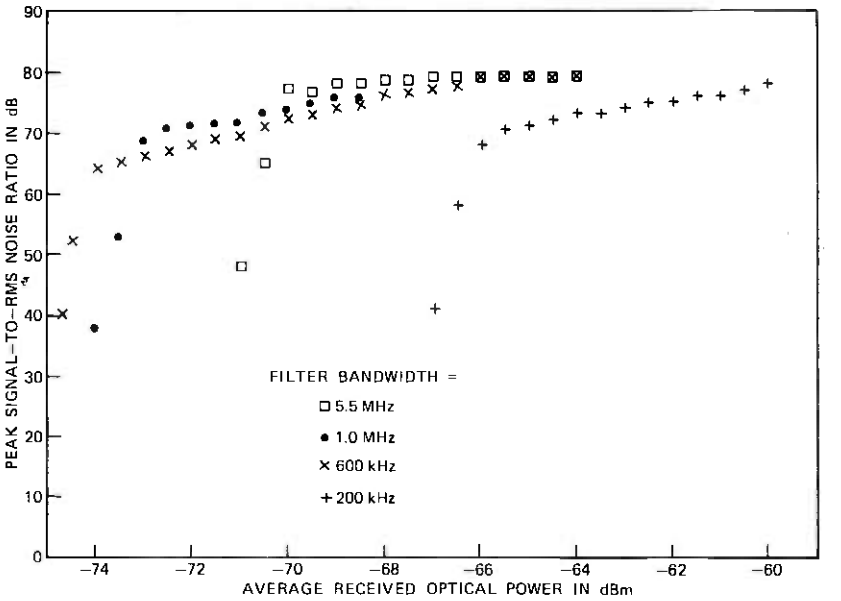


Fig. 5—Peak signal to RMS noise ratio in dB plotted as a function of average received optical power in dBm using a PIN detector.

tion of the time of a threshold crossing. With increasing power, the effect of this front-end jitter is reduced.

For optical signal levels greater than −66 dBm and for filter bandwidths of 5.5 MHz, 1 MHz, and 600 kHz, the noise resulting from the amplifiers and sampling circuits in the transmitter and receiver limits the performance. This effect occurs at −59 dBm for the 200-kHz filter. Increasing the received optical power does not improve the SNR in this region.

As the filter bandwidth decreases from 5.5 MHz to 600 kHz, both the noise and signal decrease. It can be shown that, with the high-impedance front-end, the noise (which is dominated by the FET-series noise source) decreases faster.[9] Thus, the power required to prevent threshold violations decreases with decreasing bandwidth. This effect is evident in Fig. 5. However, if the filter is too narrow, leakage noise will begin to dominate the FET-series noise source, and the required power will increase with decreasing filter width. This occurs in going from the 600-kHz to the 200-kHz filter. In the gradual increase region, increasing the filter width increases the noise voltage more slowly than the pulse slope for all cases. Therefore, increasing the bandwidth increases the SNR at a specified power level until threshold violations become significant.

3.2 Avalanche gain

For this stage of the experiment, the PIN detector was replaced with a Texas Instrument TIXL 56 avalanche detector. The filter bandwidth in all cases was set to 1 MHz.

Figure 6 shows SNR plotted as a function of average received optical power for avalanche gains at 1, 10, 50, 65, and 80. (The accuracy of the measured SNR's is of the order of ±1.5 dB.) We see from this that, by increasing the avalanche gain from 1 to ≈ 60, less optical signal power is required to achieve a given SNR. The optimum gain for this device is ≈ 60. For avalanche gains larger than ≈ 60, the excess avalanche noise, rather than thermal noise, becomes dominant and, since this noise increases faster than the signal as avalanche gain is increased, more signal power is required to achieve a given SNR. Data for an avalanche gain of 80 illustrate this point, as it requires a received power of −83.5 dBm to achieve a 75-dB SNR. However, a 75-dB SNR may be achieved at a level of −85.5 dBm with a gain of 65.

With the avalanche detector biased for optimum gain, ≈15-dB more loss may be tolerated between transmitter and receiver than in the case with unity gain.

The data for Figs. 5 and 6 were recorded for the following two methods employed to synchronize the receiver: by transmitting a

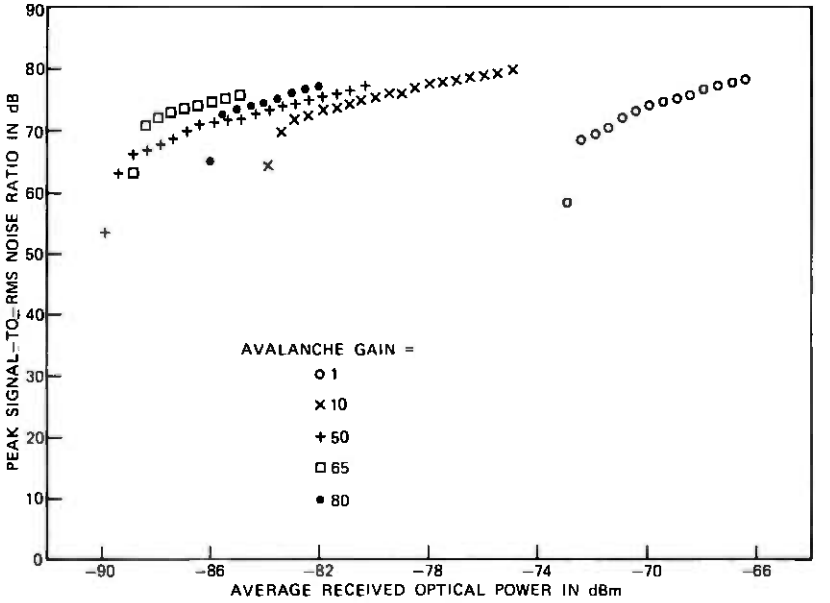
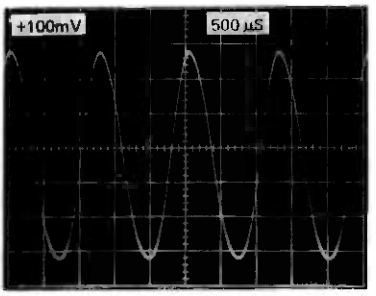


Fig. 6—Peak signal to RMS noise ratio in dB plotted as a function of average received optical power in dBm using an avalanche detector.

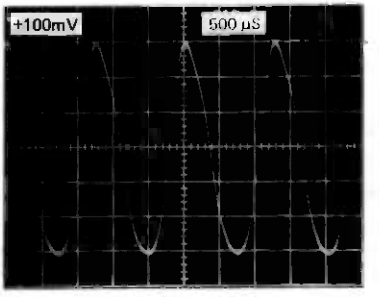
separate clock signal and by recovering the timing from the PPM signal itself. (This is discussed in detail in the appendix.) Differences in performance of the order of ±0.5 dB were observed; for this reason, it is concluded that the timing recovery circuit functioned properly over the range of power levels used in the experiments.

3.3 Distortion

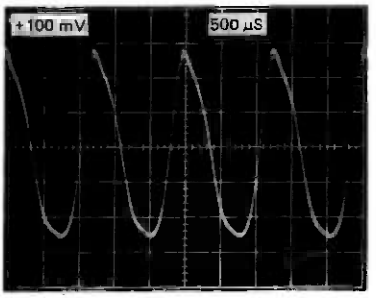
Distortion present on the output signal is shown in Figs. 7a, 7b, and 7c. The photographs show the output for SNR's at 75 dB, 77 dB, and 79 dB, respectively. Because the RMS noise is constant, the SNR here is determined by the signal amplitude. This distortion is introduced by the sampling circuits and amplifiers. It is not a basic limitation of the system. The degree of distortion was investigated using a spectrum analyzer to measure the relative amplitudes of the fundamental, second-harmonic, and third-harmonic components for different audio signal levels. Figure 7d shows the relative difference between the fundamental and second harmonic and between the fundamental and third harmonic plotted as a function of signal power. With full modulation (Fig. 7c), the second-harmonic distortion is 10 dB below the fundamental.



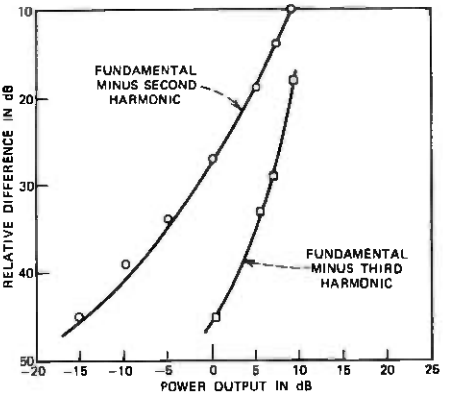
(a)



(b)



(c)



(d)

Fig. 7—Distortion on the output signal for SNR's of (a) 75 dB, (b) 77 dB, and (c) 79 dB. (d) Relative difference between the fundamental and second harmonic and the fundamental and third harmonic in dB plotted as a function of power out in dB.

IV. CONCLUSIONS

An optical PPM communication experiment has been described. The receiver is capable of recovering timing synchronization from the transmitted PPM signal itself. Peak-signal-to-RMS noise ratios greater than 70 dB (the objective of the experiment) have been achieved. By using a PIN detector and a 1-MHz filter, this objective can be achieved at an average received optical power level of −73 dBm. With full modulation, the second-harmonic distortion is 10 dB below the fundamental. Since existing LED's can couple −10 dBm of average optical signal power into an optical fiber, the allowable attenuation would be ≈ 60 dB, which allows a range of 15 km for fibers having 4-dB/km loss. However, with a large-optical-cavity laser as the source, we can contemplate that as much as +20 dBm might be injected into a fiber; this would increase the tolerable attenuation to ≈ 90 dB (corresponding to a range of ≈ 22 km). Data show that an additional improvement

of 15 dB may be obtained by using avalanche gain in the receiver. A PPM system of this nature would be able to tolerate ≈ 105 dB of fiber loss between transmitter and receiver. This corresponds to ≈ 26 km of fiber, which makes this system very attractive.

V. ACKNOWLEDGMENT

The author is grateful to C. A. Burrus for providing the LED, to Rudy Drexler and Paul Fleischer for the active filter design, and to J. E. Goell for his advice and assistance in fabrication of the system.

APPENDIX

A PPM system requires timing synchronization between transmitter and receiver clocks.[10] Two methods have been employed to accomplish synchronization. The first is to simply carry a timing signal from the transmitter to the receiver by a coaxial cable. The second method, which is described in this section, is a scheme that recovers the timing from the transmitted PPM signal itself.

The sawtooth oscillators in the transmitter and receiver have been designed to free run at approximately the same frequency; however, bias-voltage fluctuations and temperature changes may cause one to drift with respect to the other.

A change in relative phase results in a DC level shift at the output of the first sample-and-hold circuit in the demodulator. This DC level is defined as the error signal. To compensate for this frequency drift, the error signal is applied to a voltage-controlled oscillator (VCO) in such a manner as to change the oscillator frequency in opposition to the initial frequency drift. In essence, the demodulator acts as a phase discriminator in a phase-locked loop.

The VCO and associated circuitry, shown in Fig. 8, are used to complete the loop when connected between the first sample-and-hold output and the sawtooth-trigger input. The sampled output contains two components, an audio staircase and the error signal. The error signal is amplified by the operational amplifier, while the audio is suppressed

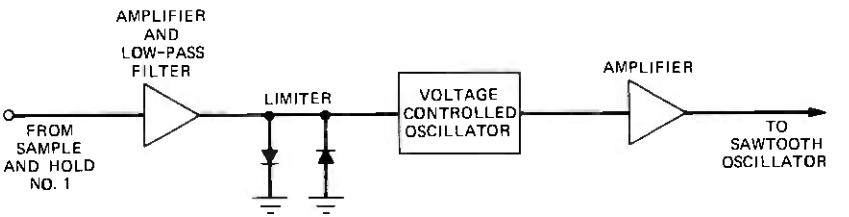


Fig. 8—Voltage-controlled oscillator and associated circuitry.

by the feedback capacitor. The differential input sets the loop reference. Two diodes connected back to back at the operational amplifier output serve to limit the frequency range over which the vco can be varied. The error signal is further amplified and applied across the vco. The final stage amplifies the vco output to a sufficient level to trigger the sawtooth generator.

REFERENCES

1. D. B. Keck, R. D. Maurer, and P. C. Shultz, "On the Ultimate Lower Limit of Attenuation in Glass Optical Waveguides," Appl. Phys. Let., *22*, No. 7 (April 1973), pp. 307–309.
2. H. Melchior, M. B. Fisher, F. R. Arams, "Photodetectors for Optical Communication Systems," Proc. IEEE, *48*, No. 10 (October 1970), pp. 1466–1486.
3. C. A. Burrus and B. I. Miller, "Small-Area, Double Heterostructure Aluminum-Gallium-Arsenide Electroluminescent Diode Sources for Optical-Fiber Transmission Lines," Opt. Commun., *4*, No. 4 (December 1971), pp. 307–309.
4. B. I. Miller, J. E. Ripper, J. C. Dyment, E. Pinkas, and M. B. Panish, "Semiconductor Lasers Operating Continuously in the 'Visible' at Room Temperature," Appl. Phys. Lett., *18*, No. 9 (May 1, 1971), pp. 403–405.
5. B. S. S. Rao, A. Subrahmanyam, and Prem Swarup, "A Technique of Modulating Pulsed Semiconductor Lasers," IEEE Trans. Commun., *COM-21*, No. 4 (April 1973), pp. 284–289.
6. S. D. Personick, "Applications for Quantum Amplifiers in Simple Digital Optical Communication Systems," B.S.T.J., *52*, No. 1 (January 1973), pp. 117–133.
7. J. E. Goell, "An Optical Repeater With High-Impedance Input Amplifier," B.S.T.J., *53*, No. 4 (April 1974), pp. 629–643.
8. H. Kressel, H. F. Lockwood, and F. Z. Hawrylo, "Large Optical Cavity (AlGa) As-GaAs Herojunction Laser Diode: Threshold and Efficiency," J. Appl. Phys., *43* (February 1972), pp. 561–567.
9. S. D. Personick, "Receiver Design for Digital Fiber Optic Communication Systems, I," B.S.T.J., *52*, No. 6 (July–August 1973), pp. 843–874.
10. R. Gagilardi, "Synchronization Using Pulsed Edge Tracking in Optical PPM Communication System," Interim Technical Report (University of Southern California), September 1972.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

An Algorithm for Determining the Endpoints of Isolated Utterances

By L. R. RABINER and M. R. SAMBUR

(Manuscript received June 10, 1974)

An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the endpoint location problem. By accurately detecting the beginning and end of an utterance, the amount of processing of speech data can be kept to a minimum. The algorithm proposed for locating the endpoints of an utterance is based on two measures of the signal, zero crossing rate and energy. The algorithm is inherently capable of performing correctly in any reasonable acoustic environment in which the signal-to-noise ratio is on the order of 30 dB or better. The algorithm has been tested over a variety of recording conditions and for a large number of speakers and has been found to perform well across all tested conditions.

I. INTRODUCTION

The problem of locating the beginning and end of a speech utterance in an acoustic background of silence is important in many areas of speech processing. In particular, the problem of word recognition is inherently based on the assumption that one can locate the region of the speech utterance to be recognized. A further advantage of a good endpoint-locating algorithm is that proper location of regions of speech can substantially reduce the amount of processing required for the intended application.

The task of separating speech from background silence is not a trivial one except in the case of acoustic environments with extremely high signal-to-noise ratio, e.g., an anechoic chamber or a soundproof room in which high-quality recordings are made. For such high signal-to-noise ratio environments, the energy of the lowest-level speech sounds (e.g., weak fricatives, low-level voiced portions, etc.) exceeds the background noise energy and a simple energy measure suffices.[1] However, such ideal recording conditions are not practical for real-world applications of speech-processing systems. Thus, simple energy

measures are not sufficient for separating weak fricatives (such as the /f/ in "four") from background silence. In this paper, we propose a fairly simple algorithm for locating the beginning and end of an utterance, which can be used in almost any background environment with a signal-to-noise ratio of at least 30 dB. The algorithm is based on two measures of speech: short-time energy and the zero crossing rate. The algorithm possesses the feature that is somewhat self-adapting to the background acoustic environment in that it obtains all the relevant thresholds on its decision criteria from measurements made directly on the recorded interval.

The organization of this paper is as follows. In Section II we discuss the major difficulties in locating the beginning and end of an utterance and propose various measurements for distinguishing between speech and no speech in these cases. In Section III we describe the algorithm to locate the endpoints of the utterance. In Section IV we give examples of the use of the algorithm, and give the results of both formal and informal tests on its ability to find endpoints of a corpus of words from several speakers. Finally, in Section V we discuss the general characteristics of the endpoint-location problem and propose alternative methods of solving the problem.

II. EXAMPLES OF SPEECH ENDPOINT-LOCATION PROBLEMS

To arrive at a reasonable algorithm for separating speech from nonspeech, it is necessary first to define the acoustic environment in which the recordings are made. In this paper, we consider two specific modes of recording. In the first mode, the speaker makes recordings on analog tape using a high-quality microphone in a soundproof room. This mode of recording is useful for obtaining reasonably high-quality speech. In the second mode of recording, the speaker records directly into computer memory in a noisy environment (e.g., a computer room) using a noise-reducing, close-talking microphone. This mode of recording is a reasonable approximation to a real-world environment for most man-machine interaction problems. To eliminate 60-Hz hum, as well as any dc level in the speech, it is assumed that the speech is high-pass filtered above 100 Hz; similarly, to keep the processing simple, the speech is low-pass filtered at 4 kHz, thereby allowing a 10-kHz sampling frequency.

Figure 1 shows a comparison of the waveform* of the background silence (on a greatly amplified scale) for these two modes of recording. The top two lines of this figure show the waveform for tape-recorded

* In this and subsequent illustrations, each line shows 25.6 ms of the waveform. Successive lines show successive 25.6-ms segments of the waveform.

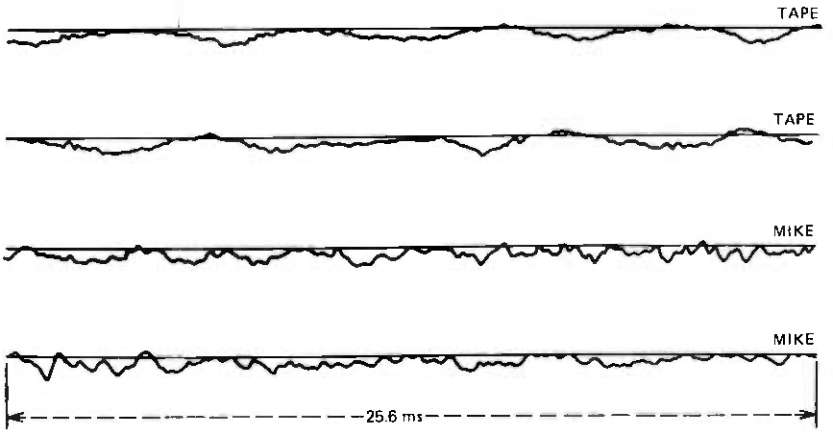


Fig. 1—Acoustic waveforms for the silences from tape and microphone.

silence from a soundproof booth, whereas the lower two lines show the waveform for the silence from the close-talking microphone. It is seen from this figure that the tape-recorded silence has a strong low-frequency component (period ≈ 8 ms) due to the recording process. The waveforms from both the close-talking microphone and the recording process appear to be quite broadband, as one would expect. Figure 2 shows typical frequency spectra of these background silences. The spectra are plotted on a log magnitude scale and are for 512-point Hamming window weighted sections. Except for the strong low-frequency-hum components for the recorded silence, the spectra of these silences are quite similar.

The problem of locating the endpoints of an utterance in these backgrounds of silence essentially is one of pattern recognition. The way one would attack the problem by eye would be to acclimate the eye (and brain) to the "typical" silence waveform and then try to spot some radical change in the pattern. In many cases this is easy to do. Figure 3 shows an example (a waveform of the word "eight") in which the silence pattern (on a reduced amplitude scale) is easily distinguished from the speech which begins just past the beginning of the third line on this figure. What one is observing in this case is a radical change in the waveform energy between the silence and the beginning of the speech.

Figure 4 shows another example (a waveform of the word "six") in which the eye can do an excellent job in locating the beginning of the speech. In this case, the frequency content of the speech is radically different from the frequency content of the background noise as manifested by the sharp increase in the zero crossing (or level crossing)

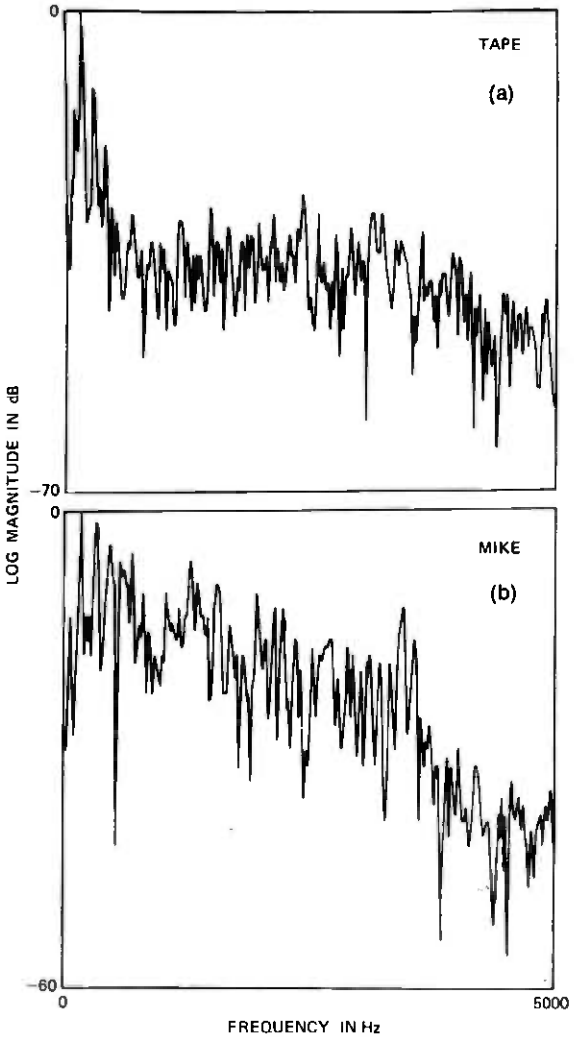


Fig. 2—Log magnitude spectra for the silences from tape and microphone.

rate of the waveform. For this example, the speech energy at the beginning of the utterance is not radically higher than the silence energy; however, other characteristics of the waveform signal the beginning of the speech.

The next set of figures illustrates some of the cases in which the eye can be greatly deceived, even with the use of expanded amplitude scales to aid in the examination of the frequency content of the speech. Figure 5 shows the waveform for the beginning of the utterance "four."

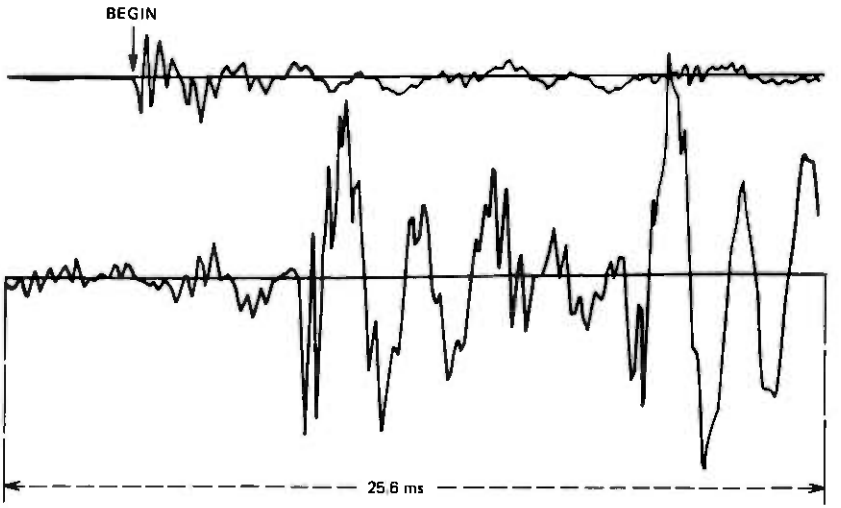


Fig. 3—Waveform for the beginning of the word "eight."

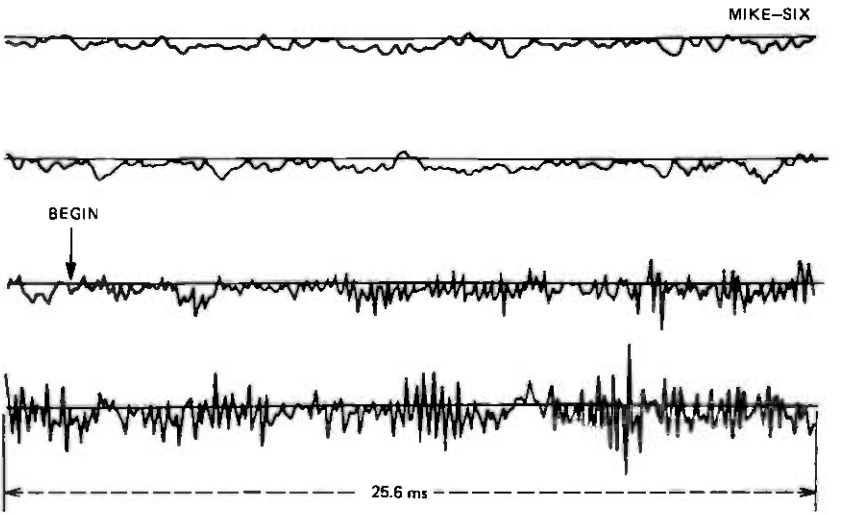


Fig. 4—Waveform for the beginning of the word "six."

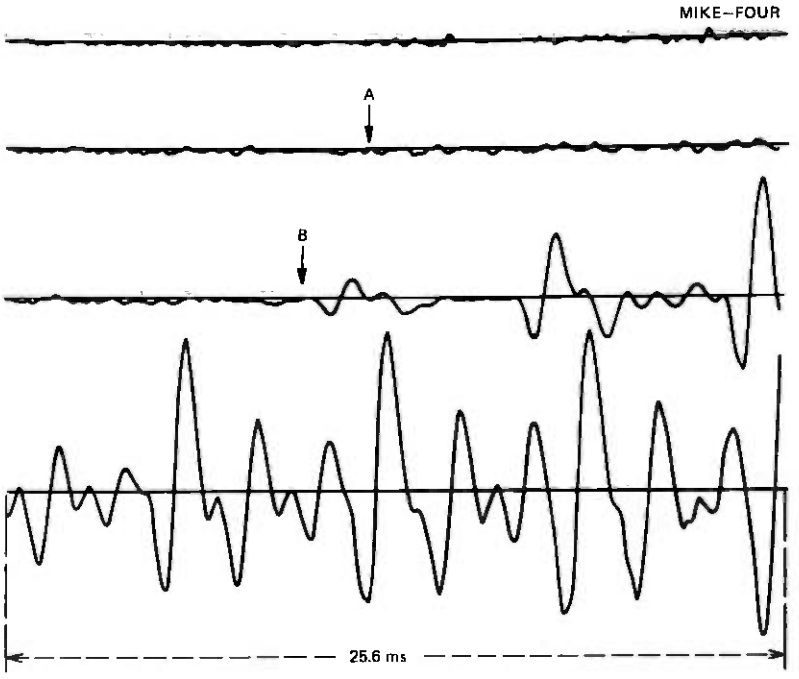


Fig. 5—Waveform for the beginning of the word "four."

This utterance begins with the weak fricative /f/. Without any *a priori* information about the utterance, the eye would select point B as the beginning of the utterance. This is incorrect, however, in that it completely misses the weak fricative /f/ at the beginning. For this example, point A is a better indication of the beginning of the speech.* Thus, one problem to be concerned with is weak fricatives at the beginning (or end) of the utterance.

Figure 6 shows another example of the difficulty in locating the endpoint of an utterance. This figure shows the waveform for the end of the word "five." Without any *a priori* information, point A might be chosen by eye as the endpoint of the utterance. However, the actual endpoint occurs approximately at point B. In this example, the final /v/ in "five" becomes devoiced and turns into an /f/, a weak fricative. Such weak fricatives are difficult to locate by eye (and sometimes even by ear).

* The criterion for deciding the actual beginning and ending points of the utterances was to use a combination of careful listening combined with precise visual examination of the waveform.

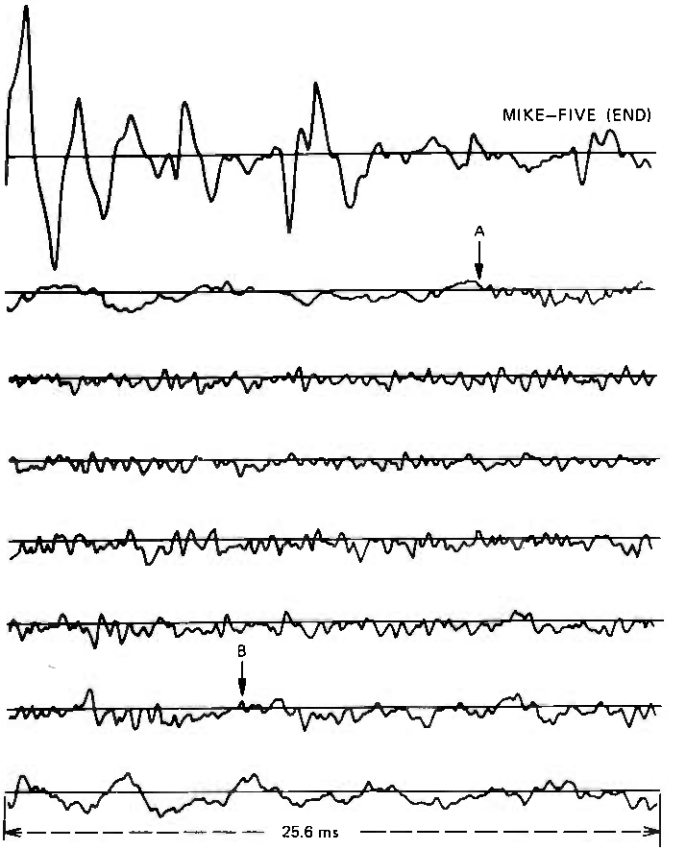


Fig. 6—Waveform for the end of the word "five."

As a final example, Fig. 7 shows the waveform for the end of the word "nine." It is quite difficult to say where the final nasal ends and where the silence begins. A reasonable location for the endpoint is the point marked END in this figure, although it is not clear how accurate this choice actually is.

Rather than give several more examples of situations in which it is difficult to locate either the beginning or the end of an utterance, we list below the broad categories of problems encountered. These include:

(*i*) Weak fricatives (/f, th, h/) at the beginning or end of an utterance.
(*ii*) Weak plosive bursts (/p, t, k/).
(*iii*) Final nasals.

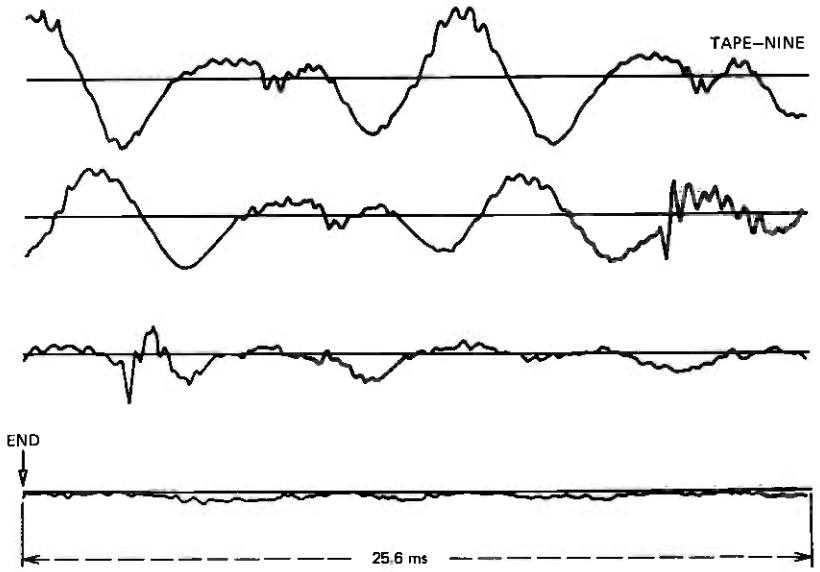


Fig. 7—Waveform for the end of the word "nine."

(*iv*) Voiced fricatives at the ends of words which become devoiced.
(*v*) Trailing off of certain voiced sounds—e.g., the final /i/ becomes unvoiced sometimes in the words "three" (/th-r-i/) or "binary" (/b-aI-n-e-r-i/).

The approach we have taken to solve these problems in an automatic endpoint-location algorithm is a pragmatic one. Our goal is to isolate enough of the word (utterance) so that a reasonable acoustic analysis of what is obtained is sufficient for accurate recognition of the word. Thus, it is not necessary to locate *exactly* the point where the word begins or ends, but instead it is important to include all significant acoustic events within the utterance. For a word like "binary," it is of little consequence if the trailing off unvoiced energy is omitted (in fact, it is probably quite helpful for a "phonetic" word-recognition strategy); however, for a word like "four" it is important to be able to reliably locate and include the initial weak fricative /f/. For this last example, the word "four," it is not necessary to include the entire initial unvoiced interval; in fact, experience has shown that 30 to 50 ms of unvoiced energy is sufficient for most word-recognition purposes. This type of knowledge is of great importance in an endpoint-finding algorithm because it enables you to set conservative values on all decision thresholds (thereby guaranteeing a very low false-alarm rate) and, for the word-recognition application, the concomitant

high miss rate will be of little practical significance. In Section III, we give the details of one practical implementation of an endpoint-location algorithm.

III. THE ENDPOINT-LOCATION ALGORITHM

Based on the preceding discussion, the goals of the endpoint algorithm are:

(*i*) Simple, efficient processing.
(*ii*) Reliable location of significant acoustic events.
(*iii*) Capability of being applied to varying background silences.

The first goal implies that only simple measurements can be made on the speech waveform as a basis for the decision. If speed and simplicity were not major issues, far more sophisticated processing could be used to give a better, more accurate result.

With the above considerations in mind, the endpoint location algorithm that was implemented is based on two simple measurements, energy and zero crossing rate, and uses simple logic in the final decision algorithm. Both energy and zero crossing rate are simple and fast to compute, and, as seen in Section II, can give fairly accurate (although conservative) indications as to the presence or absence of speech. Before proceeding to a description of the algorithm, we first define how the energy and zero crossing rate are measured. The speech "energy," $E(n)$, is defined as the sum of the magnitudes of 10 ms of speech centered on the measurement interval,[2] i.e.,

$$E(n) = \sum_{i=-50}^{50} |s(n+i)|, \tag{1}$$

where $s(n)$ are the speech samples and it is assumed that the sampling frequency is 10 kHz. The choice of a 10-ms window for computing the energy and the use of a magnitude function rather than a squared-magnitude function were dictated by the desire to perform the computations in integer arithmetic and, thus, to increase speed of computation. Further, the use of a magnitude de-emphasizes large-amplitude speech variations and produces a smoother energy function. By way of example, Fig. 8 shows typical energy functions for the words "directive" and "multiply." (The beginning and end of these words is noted on these energy plots.) For this example, the energy function is computed once every 10 ms, or 100 times per second.

The zero (level) crossing rate of the speech, $z(n)$, is defined as the number of zero (level) crossings per 10-ms interval. Although the zero crossing rate is highly susceptible to 60-Hz hum, dc offset, etc., in

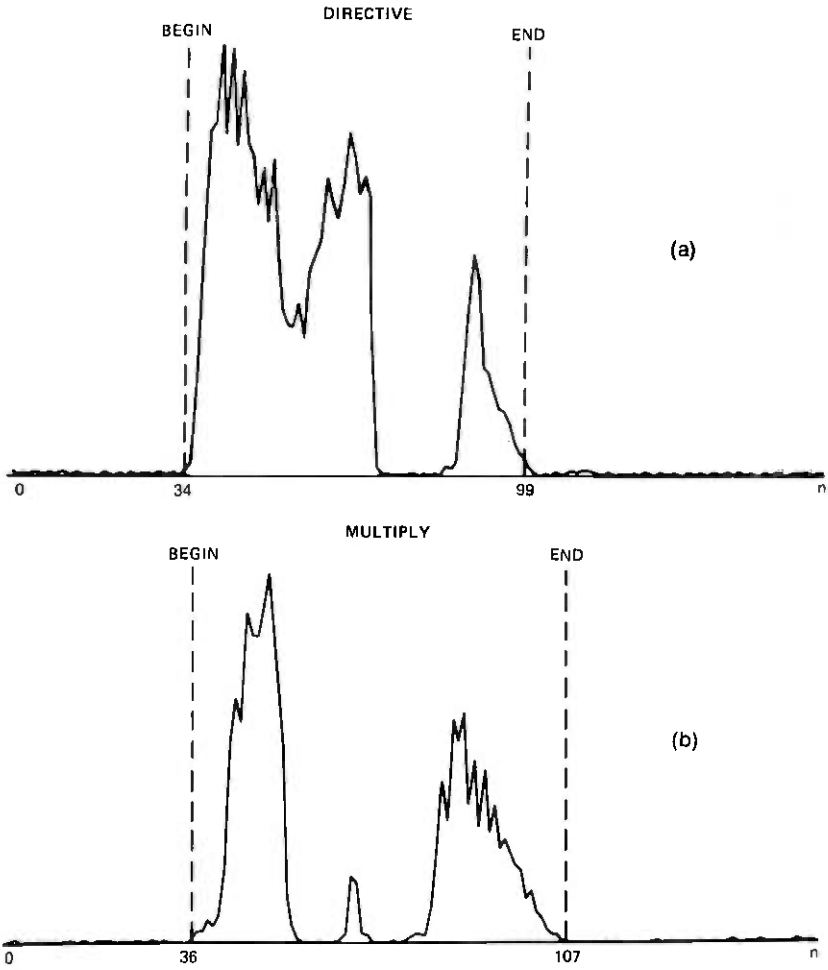


Fig. 8—Typical energy plots for the words "directive" and "multiply" with markers indicating the beginning and end of the utterance.

most cases it is a reasonably good measure of the presence or absence of unvoiced speech.

Figure 9 shows a flowchart of the endpoint-location algorithm. The speech waveform is filtered prior to sampling at 10 kHz by a bandpass filter with a 100-Hz low-frequency cutoff and a 4000-Hz high-frequency cutoff and having 48 dB per octave skirts. It is assumed that during the first 100 ms of the recording interval there is no speech present. Thus, during this interval, the statistics of the background silence are measured. These measurements include the average and

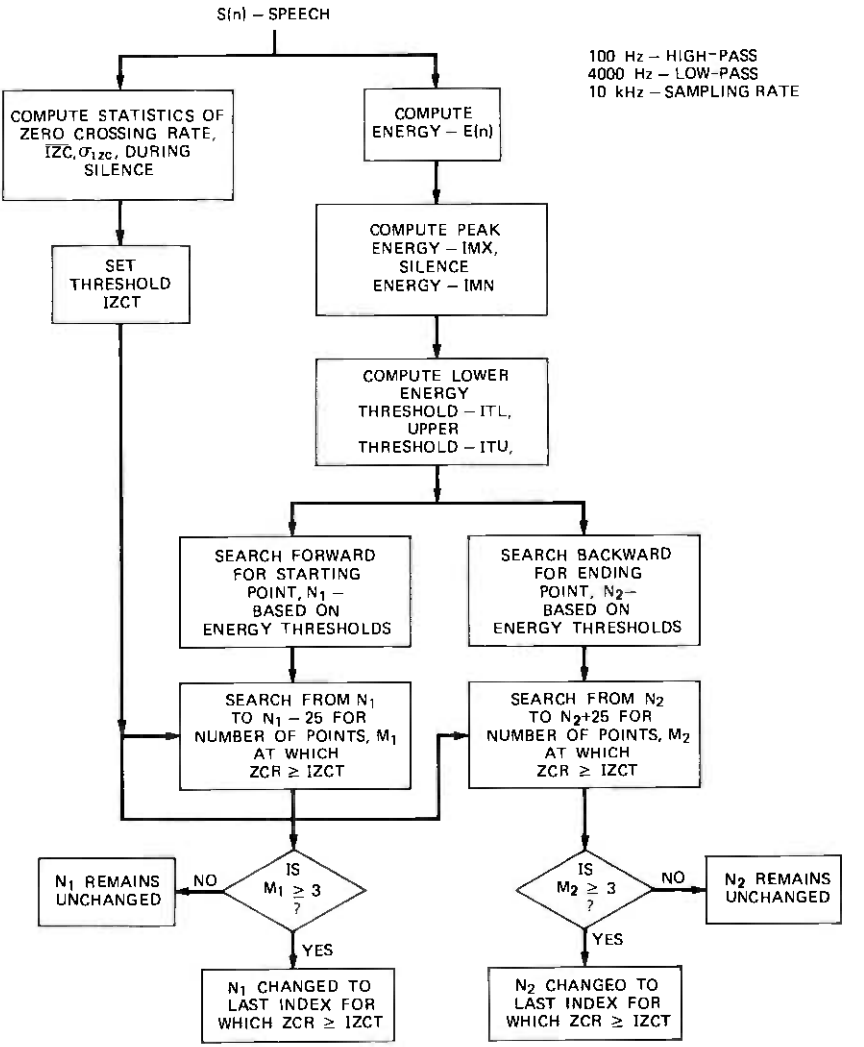


Fig. 9—Flowchart for the endpoint algorithm.

standard deviation of the zero crossing rate and the average energy. If any of these measurements are excessive, the algorithm halts and warns the user. Otherwise, a zero crossing threshold, *IZCT*, for unvoiced speech is chosen as the minimum of a fixed threshold, *IF* (25 crossings per 10 ms), and the sum of the mean zero crossing rate during silence, $\overline{IZC}$, plus twice the standard deviation of the zero crossing rate during silence, i.e.,

$$IZCT = \text{MIN}(IF, \overline{IZC} + 2\sigma_{IZC}). \tag{2}$$

The energy function for the entire interval, $E(n)$, is then computed. The peak energy, IMX, and the silence energy, IMN, are used to set two thresholds, ITL and ITU, according to the rule

$$I1 = 0.03*(IMX - IMN) + IMN \tag{3}$$

$$I2 = 4*IMN \tag{4}$$

$$ITL = \text{MIN}(I1, I2) \tag{5}$$

$$ITU = 5*ITL. \tag{6}$$

Equation (3) shows $I1$ to be a level which is 3 percent of the peak energy (adjusted for the silence energy), whereas (4) shows $I2$ to be a level set at four times the silence energy. The lower threshold, ITL, is the minimum of these two conservative energy thresholds, and the upper threshold, ITU, is five times the lower threshold.

The algorithm for a first guess at the endpoint locations is shown in Fig. 10. The algorithm begins by searching from the beginning of

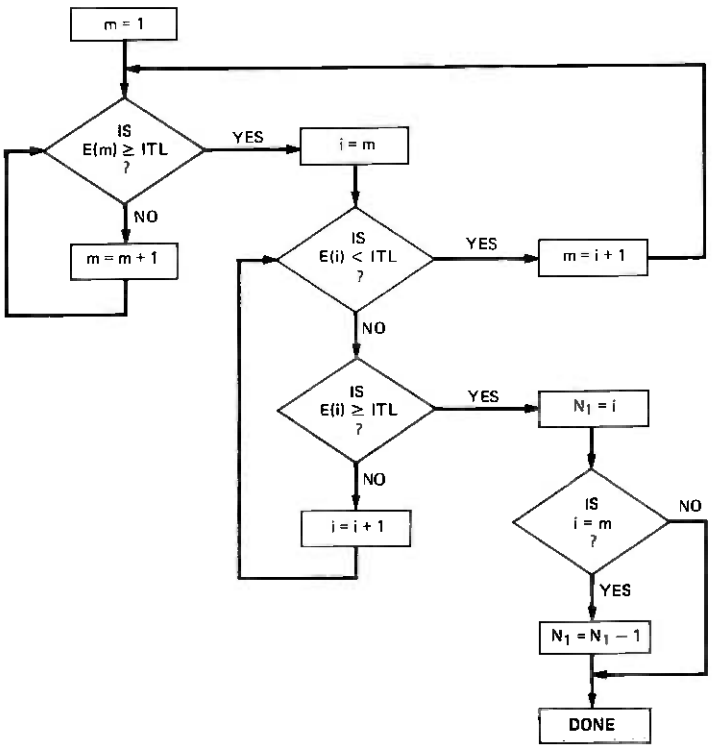


Fig. 10—Flowchart for the beginning point initial estimate based on energy considerations.

the interval until the lower threshold is exceeded. This point is preliminarily labeled the beginning of the utterance unless the energy falls below *ITL* before it rises above *ITU*. Should this occur, a new beginning point is obtained by finding the first point at which the energy exceeds *ITL*, and then exceeds *ITU* before falling below *ITL*; eventually such a beginning point must exist. A similar algorithm (shown in Fig. 11) is used to define a preliminary estimate of the endpoint of the utterance. We call these beginning and ending points N_1 and N_2, respectively.

Until now, we have only used energy measurements to find the endpoint locations; and these endpoint locations are conservative in that fairly tight thresholds are used to obtain these estimates. Thus, at this point, it is fairly safe to assume that, although part of the utterance may be outside the (N_1, N_2) interval, the actual endpoints are not within this interval. In relation to this, the algorithm proceeds to

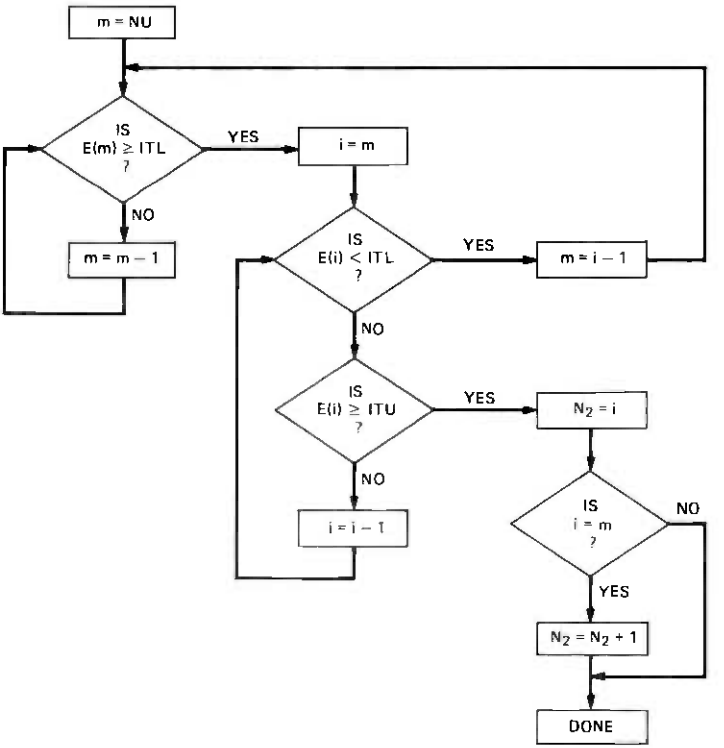


Fig. 11—Flowchart for the ending point initial estimate based on energy considerations.

examine the interval from N_1 to $N_1 - 25$, i.e., a 250-ms interval preceding the initial beginning point, and counts the number of intervals where the zero crossing rate exceeds the threshold *IZCT*. If the number of times the threshold was exceeded was three or more, the starting point is set back to the first point (in time) at which the threshold was exceeded. Otherwise, the beginning point is kept at N_1. The rationale behind this strategy is that for all cases of interest, exceeding a tight threshold on zero crossing rate is a strong reliable indication of unvoiced energy. Of course, it is still possible that a weak fricative will not pass this test, and will be missed. However, in these cases there is no simple, *reliable* method of distinguishing such a weak fricative from background silence.

A similar search procedure is used on the endpoint of the utterance to determine if there is unvoiced energy in the interval from N_2 to $N_2 + 25$. The endpoint is readjusted based on the zero crossing test results in this interval.

To illustrate the use of the endpoint algorithm, Fig. 12 shows representative contours of the energy and zero crossings for an utterance. Using the energy criterion alone, the algorithm chooses the point N_1 as the beginning of the utterance and N_2 as the end of the utterance. By searching the interval from N_1 to $N_1 - 25$, the algorithm finds a large number of intervals with zero crossing rates exceeding the threshold; thus, the beginning point is moved to $\hat{N}_1$, the first point (in time) that exceeded the zero crossing threshold. Similar examination of the interval from N_2 to $N_2 + 25$ shows no significant number of

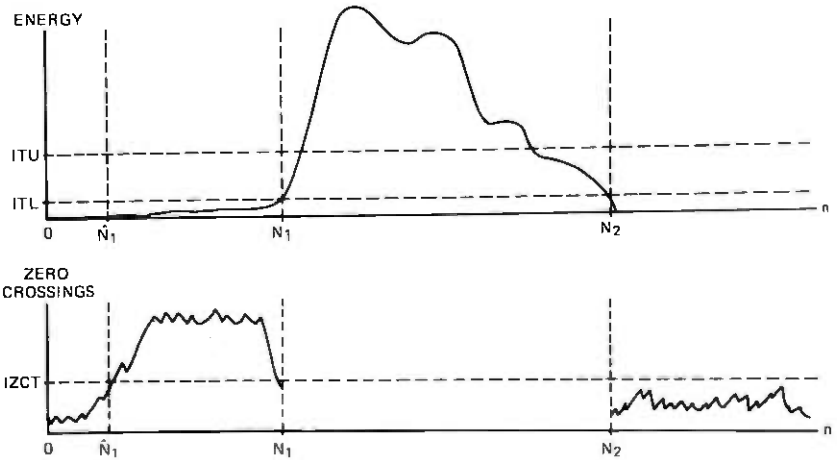


Fig. 12—Typical example of energy and zero crossings data for a word beginning with a strong fricative.

intervals with high zero crossings; thus, the point N_2 is retained as the endpoint of the utterance.

In Section IV, we give examples of the use of the endpoint algorithm for a large number of words with different speakers and different acoustic environments.

IV. EXAMPLES OF THE USE OF THE ENDPOINT ALGORITHM

The endpoint algorithm described in Section III was implemented on the DDP-516 computer facility of the Bell Laboratories Acoustics Research Department. The algorithm was tested using the two modes of recording described in Section II: high-quality tape recordings from a soundproof booth and on-line recordings using a close-talking microphone.

Figures 13 and 14 show examples of how the algorithm worked on typical isolated words. In Fig. 13 there are eight plots of the energy function for eight different words (of two different speakers). Some of the words were recorded on-line (marked MIKE) and others were recorded on tape (marked TAPE) from the soundproof booth. The markers on each plot show the beginning point and ending point of each word, as determined by the automatic algorithm. For the example in Fig. 13a (the word "nine"), the energy thresholds were sufficient to locate the endpoints. For the example in Fig. 13b (the word "replace"), the zero crossing algorithm was used to determine the ending point due to the final fricative /s/. It should be noted that even though the final /s/ has fairly large energy, since the energy thresholds were set conservatively, the energy criterion was not able to find the actual endpoint of the word. Instead, the zero crossing algorithm was relied upon in this case. In Fig. 13c, the final /t/ in the word "delete" was correctly located because of the significant zero crossing rate over the 70-ms burst when the /t/ was released. Thus, even though there was little energy or zero crossing activity for about 50 ms in the stop gap, the algorithm was able to correctly identify the endpoint because of the strength of the burst. On the other hand, if the burst had been weak, the ending point would have been located at the beginning of the stop gap.

Figure 13d is an example in which the energy during the silence was significant in a couple of places prior to the beginning of the word "subtract," yet the algorithm successfully eliminated these places from consideration because of the low zero crossing rates. In this example, a relatively weak burst in the final /t/ was correctly labeled as the endpoint.

Figures 13e through 13h show examples of words with fricatives at either the beginning or end of the word. In all cases, the algorithm was

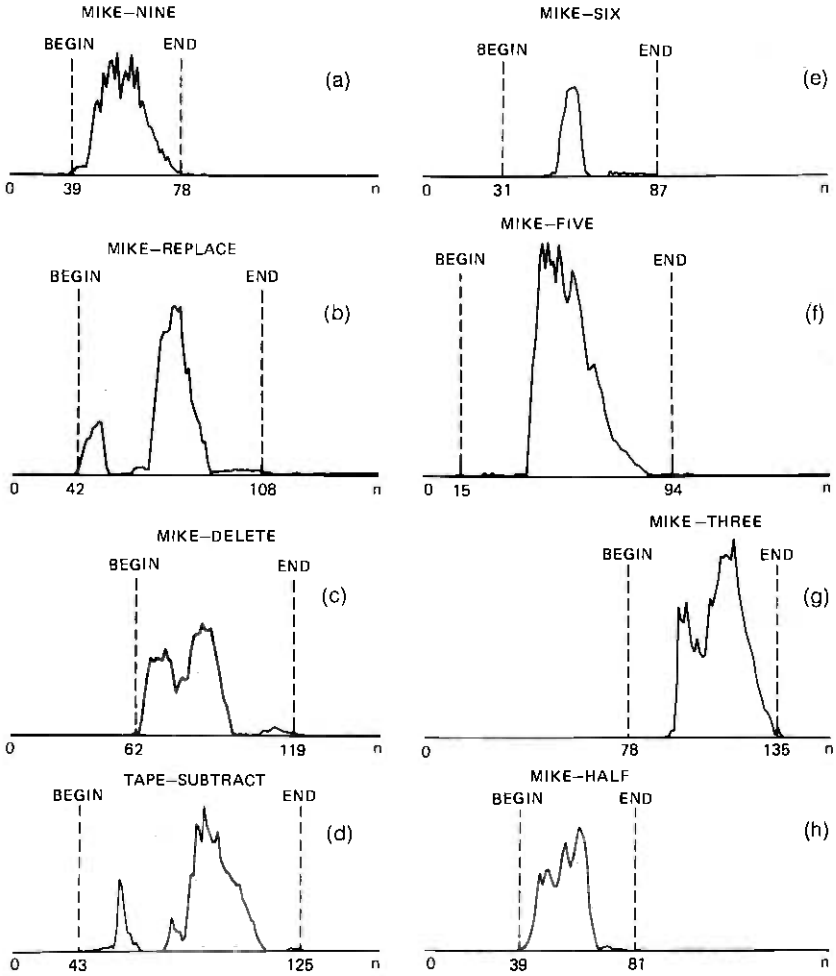


Fig. 13—Sequence of energy plots showing how the endpoint algorithm performed over a variety of words.

able to correctly place the appropriate endpoint so that a reasonable amount of unvoiced duration was included within the boundaries of the word.

Figure 14 shows three examples of how the algorithm performed for the word "four." It can be seen from the location of the beginning point that, although the level of the initial /f/ varied from strong to weak, the zero crossing indicator was able to find positive indications of the frication noise in all three cases. As discussed earlier, there are many examples where initial or final fricatives (mainly /f/ and /th/)

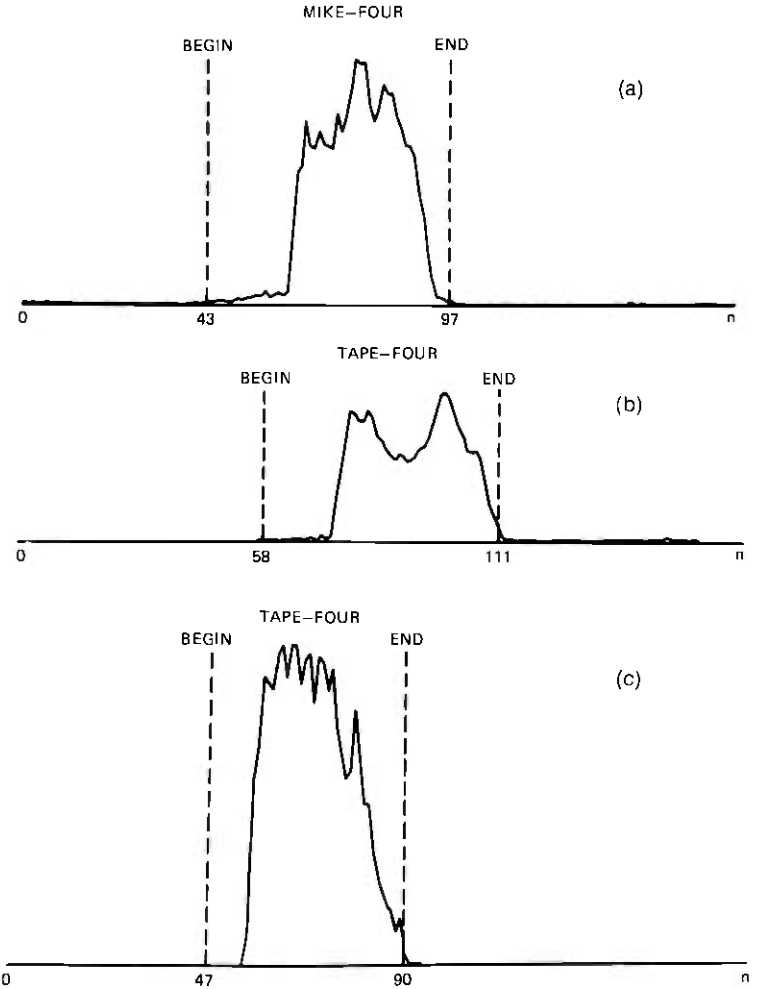


Fig. 14—Energy plots and endpoint assignments for three variations of the word "four."

were so weak they were indistinguishable from the background silence. In Section V, we discuss more sophisticated techniques for distinguishing such weak fricatives from background silence.

Two sets of formal tests were made on the algorithm. In one test, the 54-word vocabulary used by B. Gold in his word-recognition experiments[3] was read by two males and two females. For this vocabulary, the algorithm made no gross errors in locating the beginning and ending points. The algorithm did make a number of small errors of the type discussed earlier, such as losing weak fricatives or releases of

stops; however, none of these errors seriously affected the human recognition (based solely on listening) of the utterance from the portion that the algorithm did locate correctly. Thus, in some pragmatic sense, such errors can be tolerated for word recognition purposes; although for such applications as computer voice response, these small errors would probably be significant.

The second test involved 10 speakers each repeating the 10 digits from zero to nine in 10 separate sessions. (These data were actually measured for a digit-recognition experiment that used this endpoint location algorithm.) For this test, there were essentially no gross errors in locating the endpoints; in fact, it was determined that for purposes of word recognition, the algorithm was essentially error free.

V. DISCUSSION OF THE ENDPOINT-LOCATION PROBLEM

The problem of accurately locating the endpoints of an utterance is actually a specific case of the more general problem of labeling an interval of a signal as silence, unvoiced, or voiced. If one had a perfect technique for this three-level decision, the endpoint-location problem would be trivially solved. However, such an ideal algorithm does not exist as yet. Therefore, we have looked for partial solutions to this more specific problem of isolating speech from a noisy background.

The solution to this problem was based on the premise that somewhere within the given interval there was an utterance and that it would be easy to isolate the broad region in which the speech was located using energy measures alone. From this interval, we set very conservative thresholds on the speech energy (normalized to the maximum speech energy) to get a good first guess at the endpoints of the utterance. The zero crossing rate of the waveform outside these initial estimates of the endpoint was used to provide better estimates as to the existence of unvoiced speech energy in a broad region on either side of the initial endpoints.

The question now arises as to how to make the algorithm work better. One of our key goals in the original formulation was to make the algorithm fast and efficient. To this end, the readily available parameters of short-time energy and zero crossing rate were the only ones used in the decision-making process. To increase the sophistication and thereby the accuracy of the algorithm would require the inclusion of other speech parameters, such as predictor coefficients, autocorrelation coefficients, etc. The use of such additional measurements is predicated upon knowledge of how they differ for silence and for speech. Atal[4] has suggested a reasonable pattern-recognition approach for making the distinction between the three classes of silence, unvoiced speech, or voiced speech. This method, although promising, is much slower in

running and, thus, cannot be relied upon in an on-line environment. It does, however, give good indications that the problems associated with this decision are not totally untractable.

VI. SUMMARY

We have presented a fast, efficient algorithm for locating the endpoints of an utterance in a background of noise. The algorithm is based on two measurements made on the speech: short-time energy and zero crossing rate. Although the algorithm does make small errors in finding the exact endpoints of the utterance, it was designed to minimize the number of gross errors (off by more than 50 ms) in the analysis. The algorithm has been found to be sufficiently reliable and accurate that it is currently being used in on-line experiments on word recognition.

REFERENCES

1. H. F. Silverman and N. R. Dixon, "A Parametrically Controlled Spectral Analysis System for Speech," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-22*, No. 5 (October 1974), pp. 362–381.
2. R. W. Schafer and L. R. Rabiner, "Parametric Representations of Speech," Proc. IEEE Speech Recognition Symposium, Pittsburgh, April 1974.
3. B. Gold, "Word Recognition Computer Program," MIT Research Lab. of Electronics, Technical Report 452, Cambridge, June 1966.
4. B. Atal, personal communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Analysis of Specially Doped Varactors for Direct Frequency Tripling

By S. V. AHAMED

(Manuscript received July 11, 1974)

Specially doped ($N = N_0/\sqrt{x}$) varactors offer a cubic relation between voltage and charge. Frequency tripling is thus possible without an idler frequency excitation at twice the input frequency. Here we have investigated a frequency tripler from 2.115 to 6.345 GHz without an idler frequency at 4.23 GHz, thus hopefully reducing the cost and complexity of the tripler.

The results from such varactors indicate that, although frequency tripling is possible over a wide band, the efficiency and power-handling capacity are considerably lower than conventional frequency tripling with abrupt junction varactors excited at 4.23 GHz. The impedance matching is harder for these specially doped varactors, even though the mechanical construction is greatly simplified.

I. INTRODUCTION

The varactor fabrication technique recently evolved is a notable beneficiary of the IMPATT developments. Such a transfer of technology from IMPATT to varactors[1] has resulted in diodes with a zero bias capacitance of 7.7 pF, a breakdown voltage of 160 V, and a series resistance of 0.66 ohm. These diodes have been utilized in the construction of a coaxial frequency doubler yielding 8.2 W[1] at 3990 mHz and 80-percent efficiency. Further, if these diodes are used for frequency tripling, they yield about 10 W[1] at 6.345 GHz and 72- to 76-percent efficiency. When the doping density is also controlled (as is done in high-low-high profile IMPATT's), it is possible to generate varactor diodes with any predefined relation between the charge and the voltage across the varactor.

Conventional frequency tripling studied by Penfield and Rafuse[2] asserts the presence of an idler frequency excitation at twice the input frequency to mix with the input frequency, thus creating a third harmonic voltage. In the proposed tripler, the doping density is adjusted to directly convert the power at incident frequency to power at

the third harmonic. This particular doping requirement (derived in the appendix) can be relatively easily achieved by the technique used in controlling[3,4] the doping densities of high-low and high-low-high profiles for IMPATT's and for voltage variable capacitors. The concept of varying[5] the doping density for tuning diodes has been reviewed by Norwood and Shatz.[6] The capacitance of P-n junctions has been studied by Chang.[7,8] This paper reports on the study of the efficiency, power capacity, bandwidth, and impedance characteristics of triplers built from varactors formed by the special doping distribution. Computed and experimental data from conventional triplers using an abrupt junction varactor and an idler frequency excitation are also presented to provide a bench mark for comparison.

II. FREQUENCY-TRIPLING MECHANISM

In the abrupt junction varactor, the instantaneous voltage V and charge q are related as

$$(V - V_0) = \alpha(q)^2, \tag{1}$$

giving rise to second harmonic voltages from the exciting frequency charges and currents. Third and fourth harmonic voltages are also generated, and if the currents at these frequencies are suppressed, then stable frequency doubling results. To achieve a third harmonic voltage, a second harmonic (idler) current is essential, and that is the well-established basis of conventional frequency tripling.[2]

Now consider a varactor in which the instantaneous voltage and charge are related as

$$(V - V_0) = \alpha(q)^3, \tag{2}$$

thereby giving rise to third harmonic voltages resulting from charges and currents at the exciting frequency. Stable frequency tripling is possible even if second harmonic currents are not present.

The doping density that leads to such a voltage-charge relation and the corresponding capacitance-voltage relationship is derived in the appendix, Section A.1.

III. DIODE CHARACTERIZATION

The breakdown voltage, the zero bias capacitance per unit area, the depletion layer width, and the series resistance are all influenced by the doping density. These parameters may be calculated if the doping densities at finite distance from the junction are known from the basic requirements that yield the voltage-charge relationship (2) (see appendix). Further, these parameters in turn critically affect the input and output powers, impedances, and the efficiency of the tripler. Only if the diode can be fabricated with existing technology, and if the

input and output characteristics of the overall tripler are compatible with the existing technique of impedance matching, do we have a successful tripler design. Three important characteristics influencing the circuit performance of the diode are: (*i*) the breakdown voltage V_b, (*ii*) the zero bias capacitance C_0, and (*iii*) its series resistance R_s. The breakdown voltage is limited by the maximum electric intensity that the first layer* can withstand. Equation (12) yields the breakdown voltage at different doping densities for various first-layer thicknesses (see Fig. 1). The zero bias capacitance C_0 and the series resistance R_s are inversely related to each other to the first degree of approximation.† Hence, if a typical zero bias capacitance of 7.5 pF is assumed, then the series resistance at different doping densities may be computed. These curves are also shown in Fig. 1.

IV. RESULTS OF THE SIMULATION

4.1 Diode simulation study

Simulation of the results presented in the appendix yields the data necessary to study the performance of the diode from circuit and systems considerations. While the diode is tripling the frequency directly, the circuit parameters may be determined as follows: An impedance across the output is assumed to dissipate a known power; the current and charge at the triple frequency in the diode are derived; and the fundamental frequency current and charge required to sustain the output charge and current are evaluated at different values of static biasing charges across the junction from the fundamental voltage-charge relationship, eq. (2). The limits of the charge excursion across the junction during one cycle at fundamental frequency excitation are compared against the minimum and maximum charges‡ which the diode is capable of withstanding. Only if these minimum and maximum limits are not violated can the diode generate the known power. The input impedance is computed by the voltage-charge relationship at the fundamental frequency. The efficiency and bandwidth are determined by evaluating the losses in the diode at the first and third harmonic frequencies, and by incrementing the input frequency from its nominal value at 2.115 GHz. The various equations governing the distribution of charge, impedances, and efficiency are discussed in the appendix, Section A.3.

---

* For analysis, the doping profile may be approximated by a series of layers in which the doping density is held constant.

† The exact computation of these parameters has been programmed in the simulation developed for the analysis on the HIS 600 computer.

‡ This is known from the value of the breakdown voltage V_b and α in eq. (2). The minimum charge is zero.

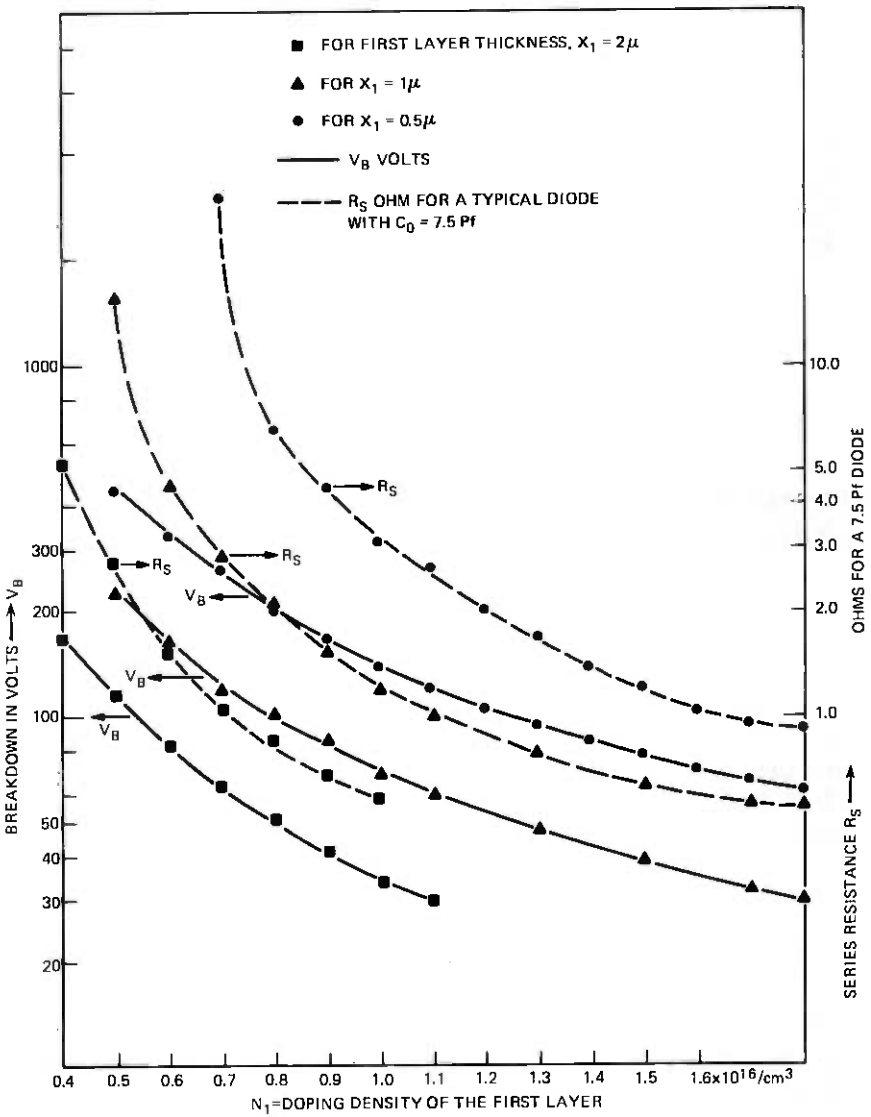
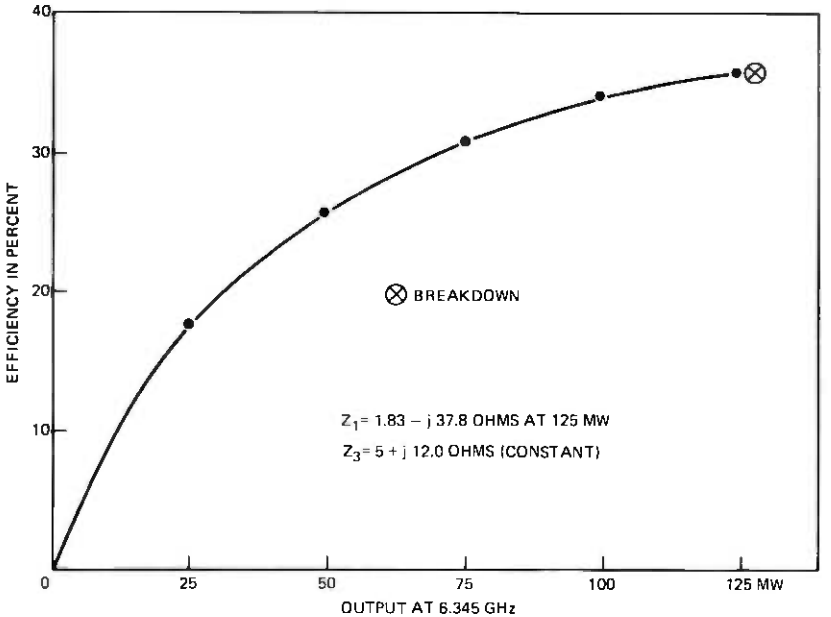


Fig. 1—V_B, R_S characteristics for $N = N_0X^{-\frac{1}{2}}$.

4.2 Single-chip diode performance curves

A single-chip diode can be designed from the basic relationships presented in appendix Sections A.1 and A.2. The diode characteristics, α, V_b, and R_s (discussed in the appendix), lead to the performance curves of the tripler. Figures 2 and 3 depict the performance of two typical diodes with zero bias capacitances of 7.5 pF ($\alpha = 0.61 \times 10^{32}$,



$\alpha = 0.61 \times 10^{32}$, $V_B = 58$ V, $R_S = 1.04$ OHMS
DEPLETION LAYER WIDTH = 4.05 μ, DIAMETER = 6.39 MILS

LAYER	THICKNESS	DOPING DENSITY
1	2 μ	0.74 x 10^{16}/cm^3
2	0.5	0.446 x 10^{16}/cm^3
3	0.5	0.410 x 10^{16}/cm^3
4	0.5	0.382 x 10^{16}/cm^3
5	0.5	0.359 x 10^{16}/cm^3
6	0.5	0.339 x 10^{16}/cm^3

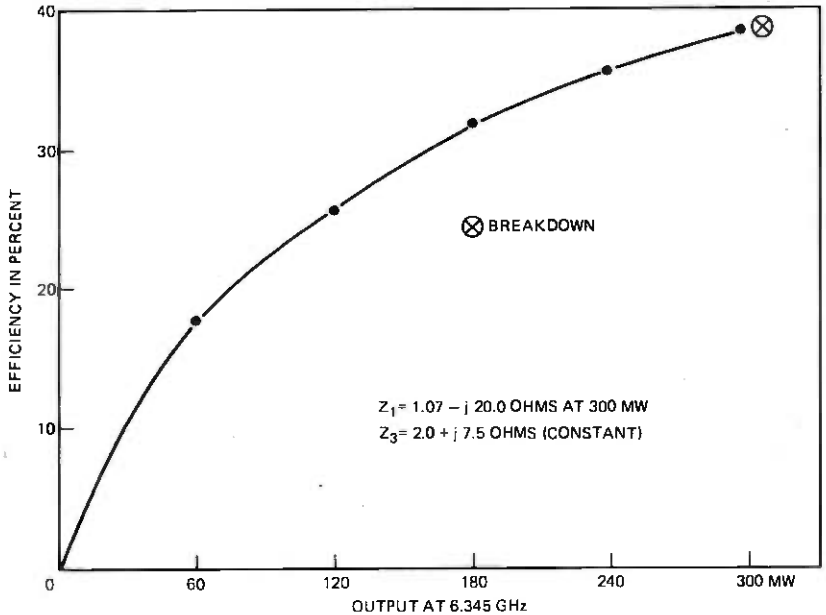
Fig. 2—Efficiency characteristic of single-chip 7.5-Pf varactor diode.

$V_b = 58$ V, $R_s = 1.04$ ohms) and 15 pF ($\alpha = 0.762 \times 10^{31}$, $V_b = 58$ V, and $R_s = 0.54$ ohm).

4.3 Double-stacked diode performance curves

When two single chips are stacked in series across the 2.115-GHz supply, the breakdown voltage and the resistance double, while the capacitance halves. The resistance of each chip is, however, inversely proportional to its capacitance. Hence, a 7.5-pF-stacked diode made from two 15-pF diodes would have a breakdown voltage of 116 V and $R_s = 1.08$ ohms.

The performance curves of two diodes with a zero bias capacitance of 7.5 pF ($\alpha = 0.152 \times 10^{32}$, $V_b = 116$ V, and $R_s = 1.08$ ohms) and 10 pF ($\alpha = 0.643 \times 10^{31}$, $V_b = 116$ V, and $R_s = 0.84$ ohm) are shown in Figs. 4 and 5.



LAYER	THICKNESS	DOPING DENSITY
1	2μ	$0.74 \times 10^{16}/cm^3$
2	0.5	$0.446 \times 10^{16}/cm^3$
3	0.5	$0.410 \times 10^{16}/cm^3$
4	0.5	$0.382 \times 10^{16}/cm^3$
5	0.5	$0.359 \times 10^{16}/cm^3$
6	0.5	$0.339 \times 10^{16}/cm^3$

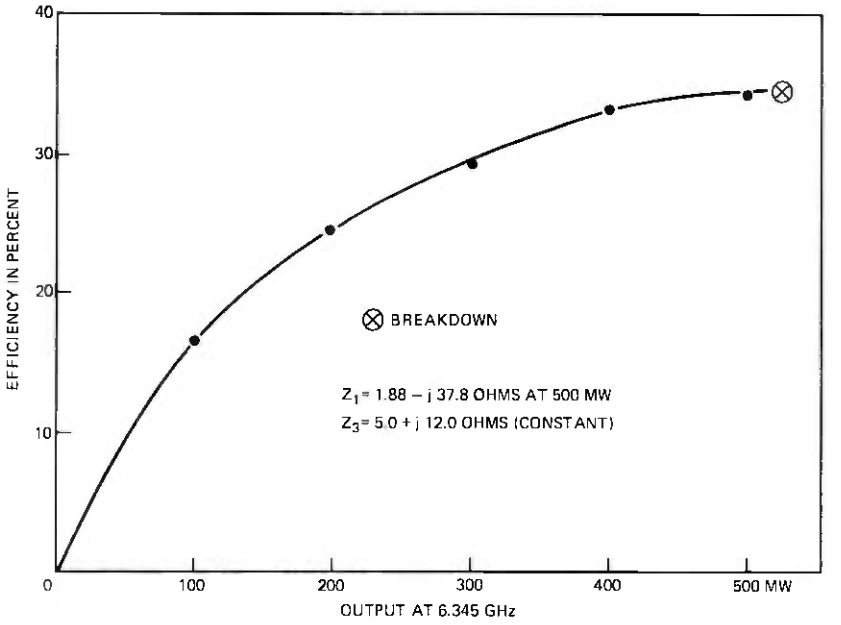
Fig. 3—Efficiency characteristic of single-chip 15.0-Pf varactor diode.

4.4 Triple-stacked diode performance curves

Two typical diodes having zero bias capacitances of 5 and 7.5 pF yield the efficiency-output characteristics shown in Figs. 6 and 7. Three single-chip diodes with zero bias capacitance, $C_0 = 15$ pF, are stacked to obtain the first 5-pF, 174-V, 1.62-ohm diode, and three chips each with 22.5 pF, 58 V, and 0.38 ohm constitute the second diode.

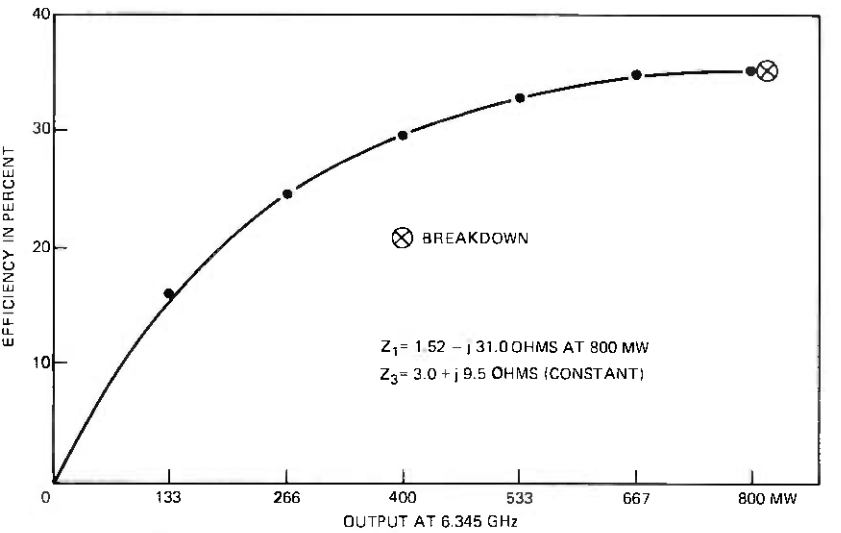
4.5 Effect of changing the doping density

The diode may also be fabricated by altering the concentration densities in the various layers of the diode. The value of N_0 is computed from N_1, the first layer doping density. In most of the simulation presented thus far, the value of N_1 was held at $0.74 \times 10^{16}/cm^3$. It



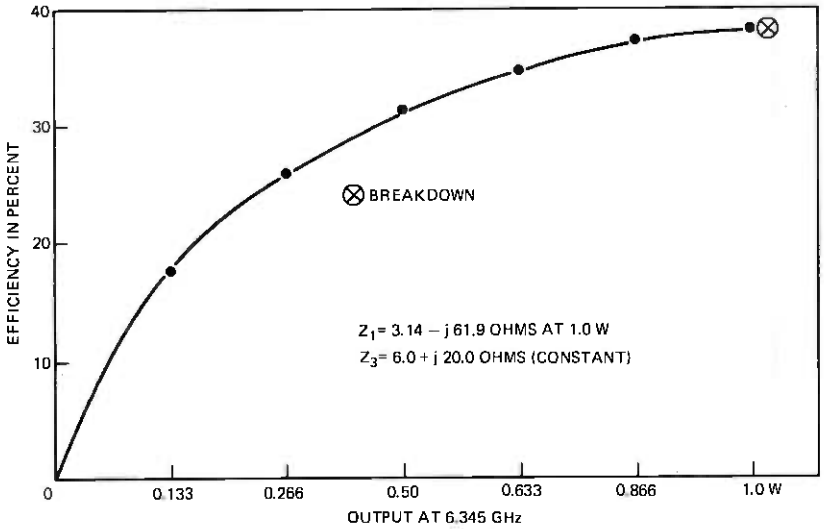
$\alpha = 0.152 \times 10^{32}$, $V_B = 116$ V, $R_S = 1.08$ OHMS
STACKED DIODE OBTAINED BY SERIES CONNECTION OF TWO 15 Pf DIODES (FIG.3)

Fig. 4—Efficiency characteristics of 7.5-Pf stacked diode.



$\alpha = 0.643 \times 10^{31}$, $V_B = 116$ V, $R_S = 0.84$ OHMS
STACKED DIODE OBTAINED BY SERIES CONNECTION OF TWO 20–Pf DIODES WITH DEPLETION LAYER WIDTH = 4.05μ, DIAMETER = 10.4 MILS WITH DOPING DENSITIES SHOWN IN FIG.2

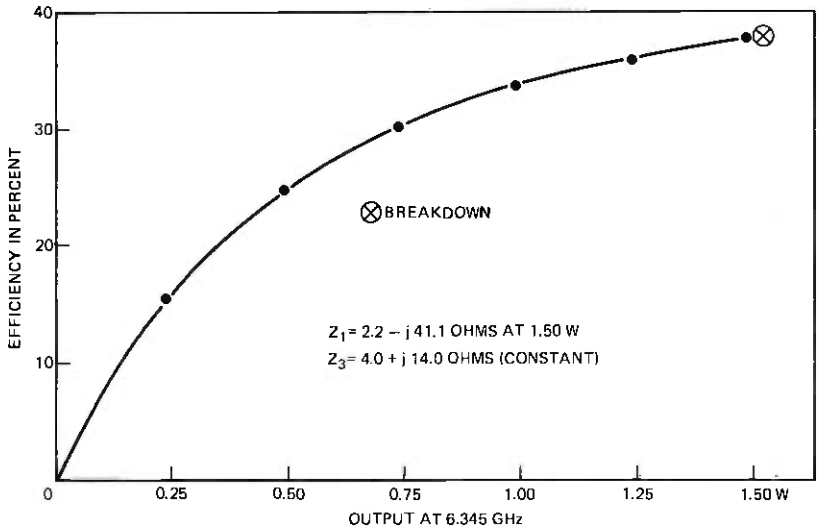
Fig. 5—Efficiency characteristics of 10-Pf stacked diode with $N_1 = 0.74 \times 10^{16}/\text{cm}^3$.



$\alpha = 0.229 \times 10^{32}$, V_B = 174 V, R_S = 1.62 OHMS

TRIPLE–STACKED DIODE OBTAINED BY SERIES CONNECTION OF THREE 15–Pf DIODES. DEPLETION LAYER WIDTH = 4.05 μ, DIAMETER = 9.04 MILS, 6 LAYERS AND DOPING DENSITIES SHOWN IN FIG. 2

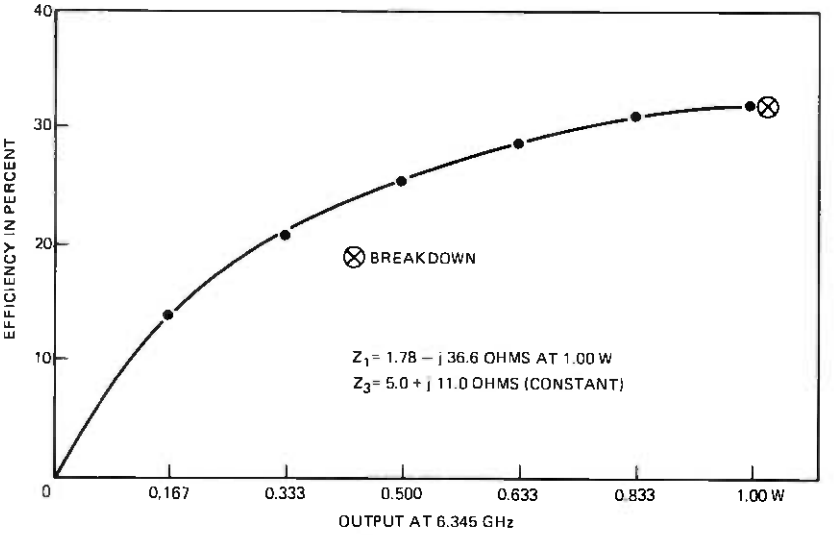
Fig. 6—Efficiency characteristics of 5-Pf triple-stacked diode.



$\alpha = 0.677 \times 10^{31}$, V_B = 174 V, R_S = 1.14 OHMS

TRIPLE–STACKED DIODE OBTAINED BY THREE 22.5–Pf DIODES HAVING DEPLETION LAYER WIDTH = 4.05 μ, DIAMETER = 11.1 MILS, 6 LAYERS WITH DOPING DENSITIES SHOWN IN FIG. 2

Fig. 7—Efficiency characteristics of 7.5-Pf triple-stacked diode.



$\alpha = 0.643 \times 10^{31}$, $V_B = 149$ V, $R_S = 1.09$ OHMS

STACKED DIODE OBTAINED BY TWO 20–Pf DIODES EACH HAVING DEPLETION LAYER WIDTH = 5.27μ, DIAMETER = 11 MILS, 8 LAYERS AS FOLLOWS:

LAYER	THICKNESS	DOPING DENSITY
1	2μ	$N_1 = 0.64 \times 10^{16}/cm^3$
2	0.5	$0.386 \times 10^{16}/cm^3$
3	0.5	$0.355 \times 10^{16}/cm^3$
4	0.5	$0.330 \times 10^{16}/cm^3$
5	0.5	$0.310 \times 10^{16}/cm^3$
6	0.5	$0.294 \times 10^{16}/cm^3$
7	0.5	$0.279 \times 10^{16}/cm^3$
8	0.5	$0.267 \times 10^{16}/cm^3$

Fig. 8—Efficiency characteristic of 10-Pf stacked diode.

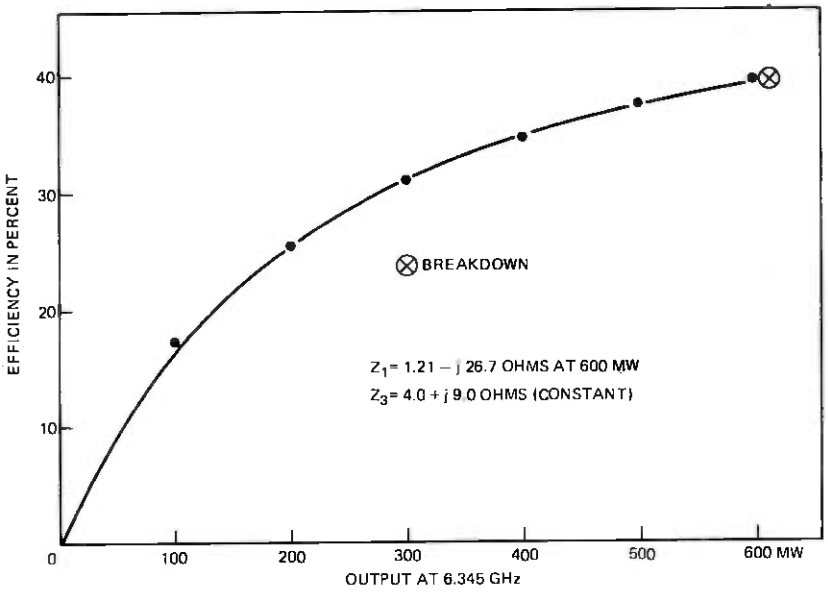
was found by trial that a certain compromise between power and efficiency could be reached at this doping level. However, for the sake of completeness, results with doping densities of $N_1 = 0.64 \times 10^{16}/cm^3$ and $N_1 = 0.84 \times 10^{16}/cm^3$ are also presented in Figs. 8 and 9 for a 10-pF diode obtained by stacking two 20-pF diodes.

V. DISCUSSION OF SIMULATED RESULTS

5.1 Power-handling capacity

When the doping density at contact is in the region of 0.7 to $0.8 \times 10^{16}/cm^3$, the breakdown voltage for specially doped varactors is about 50 percent* lower than that of the abrupt junction varactor.

* For $N_1 = 0.75 \times 10^{16}/cm^3$, $N_0 = 0.75 \times 10^{19}$ with $x_1 = 2\,\mu$ and $V_b = 60.57$ V for specially doped varactor, whereas $V_b = 89.3$ V for abrupt junction varactor. E_{max} is 44×10^4 V/cm.



$\alpha = 0.643 \times 10^{31}$, $V_B = 93$ V, $R_S = 0.65$ OHMS
STACKED DIODE OBTAINED BY TWO 20–Pf DIODES EACH HAVING DEPLETION LAYER WIDTH OF 3.22μ, DIAMETER = 10 MILS, 4 LAYERS AS FOLLOWS:

LAYER	THICKNESS	DOPING DENSITY
1	2μ	$N_1 = 0.84 \times 10^{16}/cm^3$
2	0.5	$0.506 \times 10^{16}/cm^3$
3	0.5	$0.466 \times 10^{16}/cm^3$
4	0.5	$0.434 \times 10^{16}/cm^3$

Fig. 9—Efficiency characteristics of 10-Pf stacked diode with $N_1 = 0.84 \times 10^{16}/cm^3$.

Thus, the total power-handling capacity is severely impaired. Further, the solution of cubic equations relating the charge q_1 at the fundamental frequency and the third harmonic charge q_3 leads to the lowest value of q_1 being approximately six times q_3, whereas the corresponding solutions in the conventional tripler with an idler yield the charges $q_3 : q_2 : q_1$ at the fundamental, the idler, and the third harmonic frequencies to be approximately in the proportion of 1:1.37:2.27. This further reduces the power that can be obtained from the tripler. For a triple-stacked, specially doped varactor, the breakdown voltage is 174 V and the power output is 1.5 W. For a triple-stacked, abrupt junction varactor with an idler circuit, the breakdown voltage is 262 V, and power output is 12.0 W, as shown in Fig. 10.

5.2 *Efficiency*

The specially doped varactor resistance, being a sum of the resistances of the various layers approximating the ($N = N_0 x^{-\frac{1}{2}}$) profile, is

higher than that of the abrupt junction diode. This is especially the case if the depletion layer is wide and if the doping density farther away from the junction is much less than the doping density of the first layer. The efficiency is thus adversely affected. Further, the proportion of charges q_1 and q_3 entail a relatively higher magnitude of q_1 than its corresponding value for a conventional tripler. The fundamental frequency current is much higher, thereby increasing the dissipation in the series resistance of the diode.

While the efficiency of the tripler is limited to a range of 30 to 40 percent with an output of 1 to 1.5 W (see Figs. 2 to 9) at 6.345 GHz, the efficiency (see Fig. 10) of a triple-stacked, abrupt junction tripler is well into the 70- to 76-percent range with an output of 8 to 10 W. For this particular diode, the zero bias capacitance C_0 is 5.0 pF, the breakdown voltage V_b is 262 V, and the series resistance is 0.97 ohm.

5.3 Impedances

The higher value of q_1 required to sustain q_3 reduces the real component of impedance to low values. Most of the power exchange takes place when the inductive component of the output is matched to tune out the average elastance of the diode, and the capacitive component

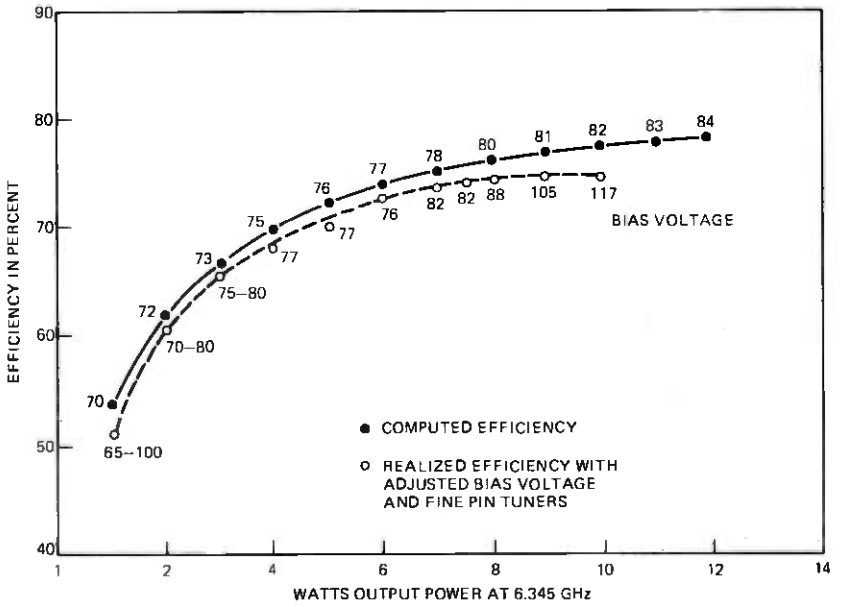


Fig. 10—Efficiency characteristic of a conventional tripler with abrupt junction and idler frequency excitation at 4.23 GHz ($N = 0.76 \times 10^{16}/cm^3$).

of the input impedance becomes approximately three times the inductive component of the output impedance. Typically, the output and input impedances of a triple-stacked diode with a zero bias capacitance of 7.5 pF, breakdown voltage of 174 V, and a series resistance of 1.14 ohms while delivering 1.5 W at 6.345 GHz are $(4 + j14)$ and $(2.21 - j41.5)$ ohms. These impedances are typical of the diode with special doping.

5.4 Bandwidth

The lack of an idler circuit makes this tripler reasonably broadbanded. When the output frequency is varied by approximately 700 mHz, the real component of the input impedance remains at 2.21 ohms, and the imaginary component varies from 43.6 to 38.9 ohms. A 70-mHz variation causes a change from 41.3 to 40.7 ohms. In comparison, the conventional tripler* undergoes a change from 7.28 to 5.71 ohms (real component) and 39.53 to 43.18 ohms (imaginary component) when the output frequency is changed by 35 mHz on either side of this nominal value at 6.345 GHz. Figures 11a and 11b depict the experimentally determined bandwidth characteristics of a conventional tripler at 8- and 5-W output.

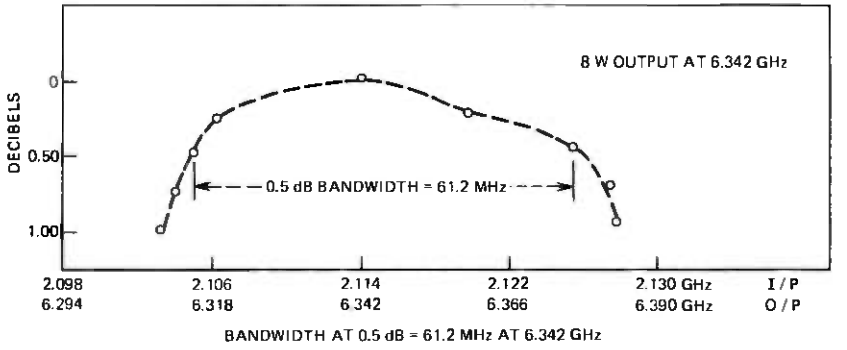
5.5 Effect of diode design variations

The reduction for N_1 from $(0.74 \text{ to } 0.64) \times 10^{16}/\text{cm}^3$ increases the breakdown voltage from 116 to 148.6 V, thus increasing the power from 800 mW to 1 W. Also, the accompanying increase in resistance from 0.84 to 1.09 ohms reduces the efficiency from 35 to 32 percent. Converse results occur when the doping density of the first layer is increased from $(0.74 \text{ to } 0.84) \times 10^{16}/\text{cm}^3$. The power-handling capacity is reduced to 600 mW, and efficiency increases to about 40 percent owing to reduced resistance of 0.648 ohm.

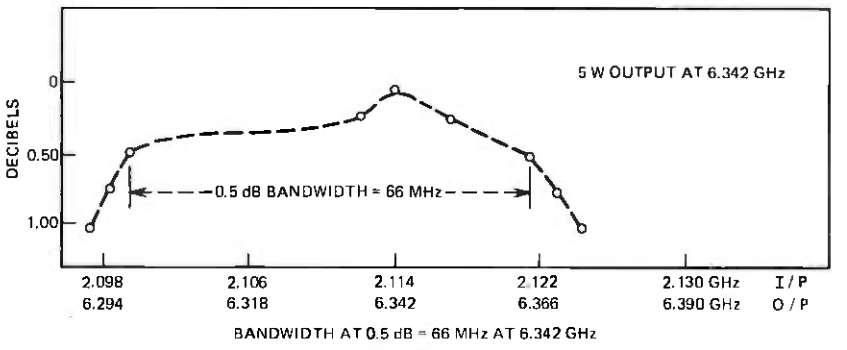
VI. EXPERIMENTAL VERIFICATION

In view of the results obtained from the computer simulation, the experimental investigation has been limited. Instead of actually constructing diodes with the prespecified doping densities, the high-low-high profile IMPATT junction has been used with the high doping density region almost etched out to leave a steeply descending concentration level. This concentration approximates the ideal $(N = N_0/\sqrt{x})$ relation, thus making the voltage-charge relation dominated by a cubic term.

* These results are obtained by the analysis of a triple-stacked, abrupt junction varactor into a zero bias capacitance of 5 pF, breakdown voltage of 262 V, and a series resistance of 1.03 ohms while delivering 6 W into an impedance of $10 + j21.6$ ohms. The variation accounts for the variation in idler circuit impedance because of a change in frequency.



(a)



(b)

Fig. 11—0.5-dB bandwidth data of the conventional tripler built with an abrupt junction diode having $C_0 = 5$ Pf, $V_b = 262$ V, and $R_S = 0.97$ ohm.

Experiments with such varactors for direct frequency tripling have confirmed that the power output is in the region of 100 to 400 mW, depending on the size and breakdown voltage. The efficiency has been in the region of 15 to 27 percent, indicating the lossy components. These diodes prefer to slip into the conventional frequency-tripling mode by utilizing any adjoining tuner circuits for circulating the idler frequency currents. Impedance matching is a sizable problem, causing frequent burnouts of the diodes.

VII. CONCLUSIONS

From power, efficiency, and impedance design considerations, the specially doped varactor without an idler frequency excitation cannot compete with the conventional abrupt junction varactor excited at idler frequency. The decrease in complexity of construction (owing to lack of idler) does not offset the reduction in power-handling capacity, the low efficiency, or the poor impedance. However, for a wideband,

low-power-signal frequency tripling, the specially doped varactor outperforms the abrupt junction varactor. The analytical study presented here, though not completely complemented by experimental results, indicates the power levels, impedances, and efficiencies one may expect from such specially doped GaAs varactors.

VIII. ACKNOWLEDGMENT

The author thanks H. Seidel for the numerous discussions during the analysis and simulation of these specially doped varactors as triplers. R. M. Ryder provided excellent review of the manuscript, and J. C. Irvin supplied the diodes for experimental verification.

APPENDIX

A.1 Basic space-charge equations governing the distribution of the potential

The electric potential ψ and space-charge density [$\rho = eN(x)$] are related* by Poisson equation in the region

$$\frac{d^2\psi}{dx^2} = -\frac{eN(x)}{\epsilon}, \tag{3}$$

where e is the electronic charge ($= 1.602 \times 10^{-19}$ coulombs), ϵ = permittivity of the region [presently, $\epsilon = 110.75 \times 10^{-12}$ F/m, which is the product of $\epsilon_0 (= 8.854 \times 10^{-12}$ F/m) and the relative permittivity $\epsilon_r (= 12.5$ for GaAs)], and finally $N(x)$ is the doping density in concentration per cubic meter, with x being measured in meters.

If the doping density is adjusted to vary by a specific relation such as

$$N(x) = N_0 x^n, \tag{4}$$

then it is possible to evaluate the exponent n to obtain the desired eq. (2) in Section II. Integrating (3) twice, we have

$$\psi = \frac{N_0 e x^{n+2}}{\epsilon(n+1)(n+2)} + C_1 x + C_2,$$

or

$$V = \psi - c_2 = \frac{N_0 e x^{n+2}}{\epsilon(n+1)(n+2)} + C_1 x, \tag{5}$$

when $x = x_d$ = the depletion layer width, then $V = V_a$, the applied voltage, and

$$\frac{dV}{dx} = 0 \quad \text{at} \quad x = x_d,$$

* This basic relationship is discussed in most standard books such as *Microwave Semiconductor Devices and their Circuit Application*, edited by H. A. Walton, New York: McGraw-Hill, 1969.

and

$$c_1 = \frac{N_0 e x_d^{n+1}}{\epsilon(n+1)} \tag{6}$$

$$V_a = \frac{N_0 e x_d^{n+2}}{\epsilon(n+2)} \tag{7}$$

$$x_d = \frac{\epsilon V_a (n+2)^{(1/n+2)}}{N_0 e}. \tag{8}$$

Now Q, charge per unit area, may be evaluated as

$$Q = \epsilon \left[\frac{dv}{dx} \right]_{x=0} = \frac{N_0 e}{n+1} \left[\frac{\epsilon(n+2)}{N_0 e} V_a \right]^{(n+1/n+2)}.$$

If $n = -\frac{1}{2}$, $(n + 1/n + 2) = \frac{1}{3}$, and we have

$$\begin{aligned} Q &= 2\sqrt[3]{(\tfrac{3}{2}) \cdot \epsilon (N_0 e)^2}\, V_a^{\frac{1}{2}} \text{ coulomb/m}^2 \\ &= \alpha_0 (V - V_0)^{\frac{1}{2}}, \end{aligned} \tag{9}$$

where V_0 is the normal contact potential, being about 1.2 V for GaAs contacts. Hence, if the doping density is adjusted to approximate $N = N_0 x^{-\frac{1}{2}}$, we would have the necessary charge-voltage relationship [eq. (2)] for direct tripling of frequency. Here, $\alpha = (\alpha_0 A)^{-3}$, where A is the cross-sectional area of the diode in square meters. Thus,

$$\alpha = 0.02941 \times 10^{36}/(N_0^2 A^3). \tag{9a}$$

A.2 Diode characterization

A.2.1 Breakdown voltage

The maximum voltage gradient permitted for GaAs contacts at various doping densities is known. Therefore, the breakdown voltage for an $N = N_0 x^{-\frac{1}{2}}$ doping distribution at various doping densities may be plotted from the equation

$$V_b = 4 \times 10^{16} [E(N_0)]^3 N_0^{-2} \text{ V}. \tag{10}$$

If the doping density is to be exactly $N = N_0/\sqrt{x}$, then N reaches an infinite value at $x = 0$. To eliminate this situation, the first layer doping density is held as N_1 (per cubic meter) and the value of N_0 is calculated as $N_0 = 0.707\, N_1 x_1^{\frac{1}{2}}$, where x_1 indicates the width of the first layer in meters. The doping densities for the mth adjoining layer is calculated as

$$N_m = N_0 \left(\sum_{i=1}^{m-1} X_i + X_m/2 \right)^{-\frac{1}{2}}. \tag{11}$$

In essence, the doping density distribution is approximated by a series of layers with different doping densities, and the value of the doping

density at the center of each layer corresponds to the necessary distribution of $N = N_0 x^{-\frac{1}{2}}$.

The values of V_b may be plotted against the first layer doping density N_1 for different values of the first layer thicknesses (x_1). Equation (10) now assumes the form

$$V_b = 8 \times 10^{16}[E(N_1)]^3 N_1^{-2} x_1^{-1} \text{ V}, \tag{12}$$

where $E(N_1)$ is in volts per meter, N_1 is in impurity concentration per cubic meter, and x_1 is in meters (see Fig. 1).

A.2.2 Diode resistance

This may be calculated by adding the resistances of the various layers. The resistivities at different values of doping densities are well known,[9] and the total resistance* may be calculated as

$$R_s = R_c + \sum_{i=1}^{i=m} \frac{\rho(N_i) l_i}{A}, \tag{13}$$

where R_c is the contact resistance inversely proportional to the area of cross-section A, $\rho(N_i)$ is the resistivity of the ith layer with a doping density of N_i, and l_i is the width of the ith layer.

A.2.3 Summary of equations

Depletion layer width is

$$x_d = 102.53 \times 10^{-4} N_0^{-\frac{1}{2}} (V + V_0)^{\frac{1}{2}} \text{ m}. \tag{14}$$

Maximum voltage gradient is

$$E_{max} = 0.02924 \times 10^{-4} N_0^{\frac{1}{2}} (V + V_0)^{\frac{1}{2}} \text{ V/m}. \tag{15}$$

Capacitance per unit area is

$$C = 1.079 \times 10^{-16} N_0^{\frac{1}{2}} (V + V_0)^{-\frac{1}{2}} \text{ F/m}^2. \tag{16}$$

Charge per unit area is

$$Q = 3.2394 \times 10^{-16} N_0^{\frac{1}{2}} (V + V_0)^{\frac{1}{2}} \text{ cb/m}^2. \tag{17}$$

A.3 Circuit performance of the diode as a tripler

A.3.1 Loss-less-varactor formulations

The voltages and charges in the basic relationship

$$(V - V_0) = \alpha q^3 \tag{2}$$

* This formulation is correct at a negligible depletion layer width. In practice, the depletion layer is swept during each cycle at the input frequency. In the conventional tripler analysis, the reduction of loss because of sweeping of the epitaxial layer enhanced the efficiency by 2 to 4 percent, while the triple-stacked abrupt junction diode (see Fig. 10) was delivering 6 to 10 W at 6.345 GHz into an impedance of $(8 + j21.4)$ ohms.

may be written in terms of the Fourier components at the first and third harmonics as

$$\left.\begin{aligned}(v - V_0) &= (v_0 - V_0) + v_1 + v_1^* + v_3 + v_3^* \\ q &= q_0 + q_1 + q_1^* + q_3 + q_3^*\end{aligned}\right\}, \tag{18}$$

where the subscript and the star indicate the harmonic and the conjugate. Separating out the two harmonics and ignoring† the currents and voltage at other harmonic frequencies from the equations, we have

$$\frac{v_3}{\alpha} = (q_1^3 + 6q_1q_1^*q_3 + 3q_0^2q_3 + 3q_3^2q_3^*) \tag{19}$$

$$\frac{v_1}{\alpha} = (6q_1q_3q_3^* + 3q_0^2q_1 + 3q_1^2q_1^* + 3q_1^{*2}q_3) \tag{20}$$

$$\frac{v_0 - V_0}{\alpha} = (q_0^3 + 6q_0q_1q_1^* + 6q_0q_3q_3^*). \tag{21}$$

Further, we have the relation between the instantaneous charge q and its Fourier components

$$q_b < q = q_0 + q_1 + q_1^* + q_3 + q_3^* < 0. \tag{22}$$

If the power and impedance at the third harmonic are known, then v_3 and q_3 are known, and for various values of q_0, the values of q_1 may be computed from (19). If the varactor diode is capable of sustaining the assumed output, then the net charge q during any one cycle of oscillation at the exciting frequency should be less than the breakdown charge q_b computed from the relation

$$q_b = \sqrt[3]{(V_b - V_0)/\alpha}. \tag{23}$$

The value of V_b is known from the diode design presented in (3).

A.3.2 Lossy varactor formulations

If R_s is the series resistance of the diode, then eqs. (19) and (20) become

$$\frac{v_3 + I_3R_s}{\alpha} = q_1^3 + 6q_3|q_1|^2 + 3q_0^2q_3 + 3q_3|q_3|^2 \tag{24}$$

$$\frac{v_1 - I_1R_s}{\alpha} = 6q_1|q_3|^2 + 3q_0^2q_1 + 3q_1|q_1|^2 + 3q_3q_1^{*2}. \tag{25}$$

Equation (24) may be solved by rewriting it in terms of q_1^3 and multiplying it by q_1^{*3}, which leads to an equation in terms of $|q_1|^6$, $|q_1|^4$,

† The physical basis for ignoring the other harmonics is that the circuit presents very large impedances at these frequencies, and there is effectively no flow of current or oscillation of charge at these extraneous frequencies.

$|q_1|^2$, and a constant. This resulting equation is a cubic equation in terms of $|q_1|^2$, and only its real and positive root is a valid solution for (24).

If the computed value of q_1 which corresponds to a prechosen value of q_0 and q_3 also satisfies (22), we have the necessary condition for generation of the power P which originally resulted in q_3.

The current I_1, voltage v_1, and bias voltage v_0 are calculated from q_1, eq. (25), and eq. (21), respectively. The efficiency is known by the computation of the power dissipated because of the first and third harmonic currents and their conjugates in R_s.

REFERENCES

1. S. V. Ahamed and J. C. Irvin, "New Frontiers of Varactor Harmonic Power Generation in the C-Band," B.S.T.J., *53*, No. 9 (November 1974), pp. 1839–1843.
2. P. Penfield, Jr., and R. P. Rafuse, *Varactor Applications*, Cambridge, Mass.: M. I. T. Press, 1962.
3. A. Y. Cho and F. K. Reinhart, "Interface and Doping Profile Characteristics with Molecular Beam Epitaxy of GaAs: GaAs Voltage Varactor," J. Appl. Phys., *45*, No. 4 (April 1974), pp. 1812–1817.
4. R. A. Moline and G. F. Foxhall, "Ion-Implanted Hyperabrupt Voltage Variable Capacitors," IEEE Trans. on Electron Devices, *ED-19*, No. 2 (February 1972), pp. 267–273.
5. J. J. Chang, R. M. Ryder, and S. Schonbrun, "A Numerical Analysis of Varactor Frequency Doublers," Microwave Diode Research, Report No. 18, U. S. Army Electronics Laboratory Contract DA36-039SC89205, DA Project No. 1P6-22001-A-056, prepared by Bell Telephone Laboratories, Incorporated.
6. M. H. Norwood and E. Shatz, "Voltage Variable Capacitor Tuning: A Review," Proc. IEEE, *56*, No. 5 (May 1968), pp. 788–798.
7. Y. F. Chang, "Capacitance of P-n Junctions: Space Charge Capacitance," J. Appl. Phys., *37*, pp. 2337–2342.
8. Y. F. Chang, "The Capacitance of P-n Junctions," Solid-State Electronics, *10*, April 1967, pp. 281–287.
9. S. M. Sze and J. C. Irvin, "Resistivity, Mobility and Impurity Levels in GaAs, Ge at 300°K," in *Solid-State Electronics*, Vol. 11, New York: Pergamon Press, 1968, pp. 599–602.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

New Results From a Mathematical Study of an Adaptive Quantizer

By DEBASIS MITRA

(Manuscript received July 17, 1974)

We consider a general class of multibit adaptive quantizers in which the quantizer function is modified at every sampling instant according to a recursive law with the transitions depending on the value of the quantizer output. We obtain a rather comprehensive set of basic properties of the device which explain the interrelationship of different aspects of the device behavior and their dependence on the parameters of the adaptation algorithm. For the quantitative analysis of the device, we give formulas and bounds for the mean time required for the quantizer function to adapt from an arbitrary initial state to the optimal. A feature new with this work is a unified treatment and a common body of results for quantizers with both bounded and unbounded range. This paper extends all the analytical results reported in an earlier paper, which dealt with a restricted class of quantizers having only four levels.

We also present new results from a computational investigation on quantizers up to four bits (sixteen levels). These results indicate, for well-designed examples of the respective classes, the kinds of improvement in performance that can be expected in going from three-bit (eight-level) to four-bit quantizers and from uniform to nonuniform quantizers.

I. INTRODUCTION

In a recent paper[1] we obtained a number of fundamental properties of a class of two-bit (four-level) adaptive quantizers useful for coding speech and other continuous signals with a large dynamic range. We also developed formulas for the quantitative analysis of the device. In the present paper, we consider a general, multibit adaptive quantizer and obtain extensions to all the results previously reported. A feature new with this work is a unified treatment and a common body of results for quantizers with both bounded and unbounded range, the former being the case of practical interest.

In the final section of the paper, Section IV, we present results from a computational investigation on adaptive quantizers up to four bits.

Readers familiar with quantizers and whose primary interest is in the performance of the device may skip the earlier sections that contain the development of the mathematical results. Section IV includes a comparison of the performances of uniform and nonuniform quantizers for normally distributed input sequences.

A quantizer with $2N$ levels is shown in Fig. 1. In the figure, *input* refers to the nth sample of the continuous signal, $x(n)$, where $n = 0, 1, \cdots$; *output* refers to the level that is coded before transmission at that time. We let $\xi_1 = 1$ and call Δ the step size.* In uniform quantizers, $\xi_i = i$ and the vertical axis is also subdivided into equal intervals in the range $(\eta_1\Delta, \eta_N\Delta)$. In adaptive quantizers which are of interest here, the step size, and hence the entire quantizer function, is time-variable, and the step size at the nth sampling instant is denoted by $\Delta(n)$. The parameters $\{\xi_i\}$ and $\{\eta_i\}$ are predetermined and do not change with time.

In this paper, the main algorithm for step-size adaptation is

$$\Delta(n+1) = M_i\Delta(n) \quad \text{if} \quad \xi_{i-1}\Delta(n) \leqq |x(n)| < \xi_i\Delta(n), \tag{1}$$

where $M_1, M_2, \cdots, M_N$, called multipliers, are fixed constants. The following natural restrictions are imposed on the multipliers:

$$M_1 < 1 < M_N \quad \text{and} \quad M_1 \leqq M_2 \leqq \cdots \leqq M_N. \tag{2}$$

Even so, a great deal of the flexibility of the quantizer is incorporated in the multipliers and, to some extent, in the parameters $\{\xi_i\}$ and $\{\eta_i\}$. Observe that the algorithm in (1) utilizes only unit memory and that it is not necessary to transmit to the receiver separate information on the step size.

We shall also be considering the following important variation of (1) in which the step sizes $\{\Delta(n)\}$ are constrained to be within a specific bounded interval $[\bar{K}, \bar{L}]$; suppose $\xi_{i-1}\Delta(n) \leqq |x(n)| < \xi_i\Delta(n)$, then

$$\begin{aligned} \Delta(n+1) &= M_i\Delta(n) && \text{if} \quad \bar{K} \leqq M_i\Delta(n) \leqq \bar{L} \\ &= \bar{K} && \text{if} \quad M_i\Delta(n) \leqq \bar{K} \\ &= \bar{L} && \text{if} \quad \bar{L} \leqq M_i\Delta(n). \end{aligned} \tag{3}$$

We call the associated device the *saturating adaptive quantizer*. There are situations where it is attractive to have the interval $[\bar{K}, \bar{L}]$ relatively small.

The most restrictive assumption that is made about the input sequence $\{x(n)\}$ is that it is a sequence of independent random variables (see Sections 1.1 and 1.2 for a discussion). However, in differential PCM schemes in which the quantizer is used together with a

* For notational convenience, we also let $\xi_0 = 0$ and $\xi_N = \infty$.

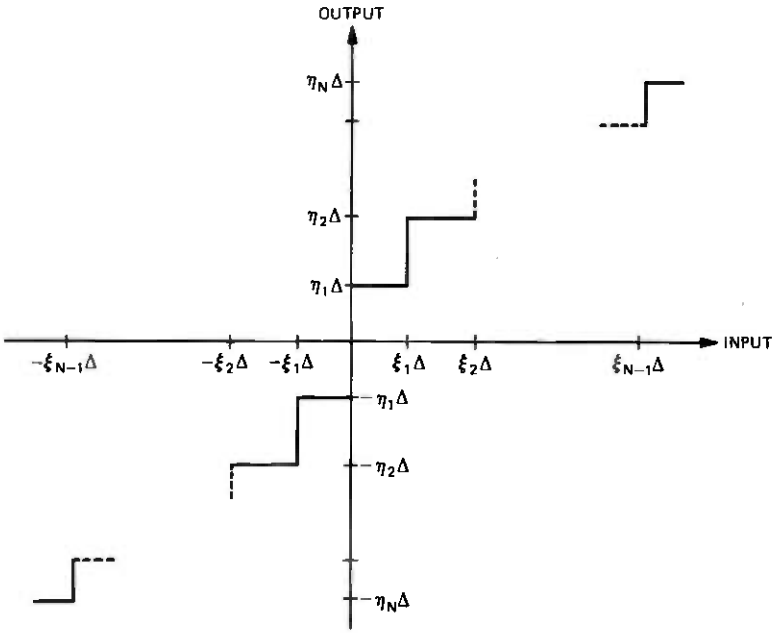


Fig. 1—The quantizer function.

predicting filter in the feedback loop, the effect of the restriction is diminished.

With $\xi_i = i$, the adaptation algorithm in (1) is due to Cummiskey, Flanagan, and Jayant,[2,3] who have also implemented speech coding by a four-bit quantizer. References 1, 2, and 4 may be consulted for a fuller account of the antecedents of the quantizer and related work that has been done in this area. Goodman and Gersho[4] have also examined the general multibit quantizer from a theoretical point of view, and their work complements rather well the work described here.

We briefly summarize here the main features of this paper.

(*i*) The theory that we give here applies to quantizers having bounded range and finite alphabet, with the important properties and relations holding also for quantizers with unbounded range. However, as may be expected, differences do exist between the two types of quantizers. For instance, a key relation in the work of Goodman and Gersho,[4] who do not consider finite range quantizers, called the *design equation*, holds exclusively for the class they consider.

(*ii*) The single most important property of either type of quantizer—ordinary or saturating—that we find is a localization property which states that, for independent identically distributed inputs, there exists a strong localization of the mass of the stationary step-size distribution

about an easily identifiable central value. See Theorem 1, Section 2.2, for a statement of this property. The localization property, together with certain scaling properties of the central state, provides the key to the synthesis of the adaptive quantizers.

(*iii*) A property of the quantizers having important implications is that, under certain conditions, as the range of the multipliers is decreased to approach unity, then the stationary step-size distribution becomes increasingly concentrated about the central step size. A result of this type is given in Ref. 4, where it is shown that a "spread function" has the appropriate behavior. However, the definition of the spread function is novel, and connections, if any, with the dispersion of mass in the distribution are not established. In Section 2.4 we establish the property directly in terms of the mass of the distribution.

(*iv*) In Section III we develop, as design aids, formulas and bounds on the mean adaptation time, i.e., mean time required for the step size to adapt from arbitrary initial values to the central step size.

The mathematical analysis is of a random walk on the integers, in which the state transition probabilities depend on the states. Random walks of the type considered here are encountered in other areas; for instance, in various schemes (up-and-down method, transformed up-and-down method[5–7]) for estimating a quantile of an unknown distribution by using only response, nonresponse data, as is required in bioassay, sensitivity data analysis, and psychological testing. The central properties of the random walk that we obtain here are new and of general interest.

1.1 Assumptions and background

Let $\sigma > 0$ denote a scale parameter and let $\mathcal{G}$ denote an equivalence class of distributions $F_\sigma(z)$, $z \geqq 0$, in which the distributions are identical to within a scaling operation, i.e.,

$$F_\sigma(\sigma z) = F_1(z). \tag{4}$$

For instance, $\mathcal{G}$ may be the class of half normal distributions, in which case σ^2 is the variance and $F_1(z) = \Pr[|x| \leqq z]$, where x is normal with zero mean and unit variance. In what follows we let $\{x_\sigma(n)\}$ denote a sequence of independent random variables, each with the distribution function $\Pr[|x_\sigma(n)| \leqq z] = F_\sigma(z)$.

We recall certain known facts about optimal nonadaptive quantization where $\{x_\sigma(n)\}$ forms the input sequence, $F_\sigma(z)$ is known, and, for some suitable choice of a fidelity criterion such as $E[\{y(n) - x_\sigma(n)\}^2]$ where $\{y(n)\}$ is the output of the quantizer, the optimal step size $\hat{\Delta}_\sigma$ is computed. With the rms criterion and the inputs normally distrib-

uted, Max[8] has computed $\hat{\Delta}_\sigma$ and, for the nonuniform case, the corresponding optimal parameters $\{\hat{\xi}_i\}$, $\{\hat{\eta}_i\}$ for quantizers with various levels, N. A convenient way of presenting such results for any $\mathcal{G}$ is as

$$F_1(\hat{\Delta}_1) = \alpha, \tag{5}$$

where α is some constant, since optimal (nonadaptive) step sizes $\hat{\Delta}_\sigma$ corresponding to the scale parameter σ are obtained from

$$\hat{\Delta}_\sigma = \sigma\hat{\Delta}_1. \tag{6}$$

In this paper we show that, when σ is fixed and $\{x_\sigma(n)\}$ forms the input to the quantizer, then the step size, a random variable evolving according to either (1) or (3), has a natural center C_σ. We show, for instance, that the stationary step-size distribution is localized about C_σ and that the degree of localization may be arbitrarily increased, although at the cost of other aspects of performance. There are two important facts to note about C_σ. First, by virtue of its explicit definition, C_1 can be made to take almost any desired value by suitable choice of the multipliers. Second, as we show in the following section, the central step size has a scaling property similar to (6). We are therefore in a position to incorporate the results of optimal nonadaptive quantization by identifying $\hat{\Delta}_1$ with C_1.

1.2 Central state

We consider only quantizers with multipliers having the following form:

$$M_i = \gamma^{m_i} \quad i = 1, 2, \cdots, N, \tag{7}$$

where γ is some real number greater than 1 and the m_i's take integral values. With (2), this implies

$$m_1 < 0 < m_N \quad \text{and} \quad m_1 \leqq m_2 \leqq \cdots \leqq m_N. \tag{2'}$$

We shall further take the set of m_i's to be relatively prime, i.e., their greatest common divisor is 1. If, as we shall assume, the initial step size is of the form γ^i, i integral, then the step size is always of that form and the space of possible step sizes forms a lattice.

Consider an independent identically distributed input sequence $\{x_1(n)\}$, where $\Pr\{|x_1(n)| \leqq z\} = F_1(z)$ and $F_1(\cdot)$ is an element of $\mathcal{G}$. We drop the subscript that identifies the scaling. For $z \geqq 0$, let*

$$B(z) \triangleq \sum_{r=1}^{N} m_r\{F(\xi_r z) - F(\xi_{r-1} z)\}. \tag{8}$$

* $F(0) = 0$, $F(z) \to 1$ as $z \to \infty$ and $F(z)$ is monotonic, strictly increasing with z.

Since it is also true that

$$B(z) = m_N - \sum_{r=1}^{N-1} (m_{r+1} - m_r)F(\xi_r z), \tag{8'}$$

it is clear that $B(z)$ is a monotonic, strictly decreasing function of z; also, $B(0) = m_N > 0$ and $B(z) \to m_1 < 0$ as $z \to \infty$. Hence, there exists a unique integer $\underline{i}$ with the property that

$$B(\gamma^{\underline{i}-1}) > 0 \geqq B(\gamma^{\underline{i}}). \tag{9}$$

We denote $\gamma^{\underline{i}}$ by C and refer to it as the *central step size*. All step sizes are considered to be of the form $C\gamma^i$, $i = 0, \pm 1, \pm 2, \cdots$.

Remarks:

(*i*) The parameters $\{m_i\}$ and γ may be selected to make the resulting central step size C approximate as closely as desired any given real positive number, $\hat{\Delta}$. First, by making γ close to unity the grid of possible step sizes can be made sufficiently fine. Second, the integral parameters $\{m_i\}$ can be chosen to make $\sum m_r\{F(\xi_r\hat{\Delta}) - F(\xi_{r-1}\hat{\Delta})\}$ sufficiently small.

(*ii*) So far, we have been concerned with the central step size for the probability distribution $F(z)$, corresponding to the particular scale parameter $\sigma = 1$. To demonstrate the behavior of the central step size with various scale parameters, let C_σ denote the central step size corresponding to the input probability distribution, $F_\sigma(z)$, and let $B_\sigma(z)$ be defined like $B(z)$ in (8) with $F(\cdot)$ replaced by $F_\sigma(\cdot)$. Let $\underline{C}_\sigma$ be the unique solution of

$$B_\sigma(\underline{C}_\sigma) = 0, \tag{10}$$

where, of course, $\underline{C}_\sigma$ may not be of the form γ^i, i integral. However,

$$C_\sigma/\gamma < \underline{C}_\sigma \leqq C_\sigma. \tag{11}$$

We observe that $\underline{C}_\sigma$ scales, i.e.,

$$\underline{C}_\sigma = \sigma\underline{C}_1. \tag{12}$$

The above follows from the following property of the functions $\{B_\sigma(\cdot)\}$:

$$B_\sigma(\sigma z) = B_1(z).$$

From (11) and (12),

$$\boxed{C_\sigma/\gamma < \sigma\underline{C}_1 \leqq C_\sigma}, \tag{13}$$

and it is in this sense that we say that the central step size scales.

1.3 Basic equations

We define a Markov chain and obtain the transition equations for the ordinary quantizer with the inputs being $\{x(n)\}$, which are independent identically distributed, and $\Pr\{|x(n)| \leqq z\} = F(z)$. Let

$$\omega(n) \triangleq \log_\gamma \Delta(n) - \log_\gamma C,$$

so that

$$\omega(n+1) = \omega(n) + m_r \quad \text{if} \quad \xi_{r-1} C\gamma^{\omega(n)} \leqq |x(n)| < \xi_r C\gamma^{\omega(n)}, \tag{14}$$

where $1 \leqq r \leqq N$. We have in (14) a Markov chain on $0, \pm 1, \pm 2, \cdots$, with the central step size C corresponding to the 0 state. Let

$$p(i; n) \triangleq \Pr[\omega(n) = i].$$

The state transition equations are

$$p(i; n+1) = \sum_{r=1}^{N} b^{(r)}(i - m_r) p(i - m_r; n), \tag{15}$$

where the transition probabilities are

$$b^{(r)}(i) \triangleq F(\xi_r C\gamma^i) - F(\xi_{r-1} C\gamma^i), \quad 1 \leqq r \leqq N. \tag{16}$$

The qualitative results that we obtain are based on the following two relations that do not depend on the particular distribution $F(z)$.

$$(i)\ 0 \leqq F(\xi_r\gamma^i) < F(\xi_r\gamma^{i+1}) \leqq 1 \quad \text{for all } i \text{ and } 1 \leqq r \leqq (N-1). \tag{17}$$

$$(ii)\ \sum_{r=1}^{N} m_r b^{(r)}(-1) > 0 \geqq \sum_{r=1}^{N} m_r b^{(r)}(0). \tag{18}$$

The latter condition follows from the definition of the central step size.

The 0 state of the random walk has the following important property: There is a net drift to the left (right) from states to the right (left) of the 0 state.

$$\boxed{E[\omega(n+1) \mid \omega(n) = i] - i = \sum_{r=1}^{N} m_r b^{(r)}(i) \begin{cases} < 0 & \text{if } i > 0 \\ > 0 & \text{if } i < 0 \end{cases}}. \tag{19}$$

The above super- and submartingale properties are the basis for the existence of a stochastic Liapunov function (Appendix A) and the bound given in Section 3.2.

1.4 Saturating adaptive quantizer

Any hardware implementation of the quantizers will incorporate some scheme for restricting the range of step sizes. In addition, there are reasons for desiring the step size to be bounded. For instance, by limiting the step sizes at both ends, it is possible to devise automatic schemes for "forgetting" the effects of past channel errors.[9] In such algorithms, the step size may be bounded to fairly small intervals.

For the saturating adaptive quantizer, eq. (3), suppose that

$$\xi_{r-1}C\gamma^{\omega(n)} \leqq |x(n)| < \xi_r C\gamma^{\omega(n)}$$

for some r, $1 \leqq r \leqq N$. We obtain the following equation analogous to (14):

$$\begin{aligned} \omega(n+1) &= \omega(n) + m_r \quad \text{if} \quad -K \leqq \omega(n) + m_r \leqq L \\ &= -K \quad \text{if} \quad \omega(n) + m_r \leqq -K \\ &= L \quad \text{if} \quad L \leqq \omega(n) + m_r, \end{aligned} \tag{20}$$

where K and L are fixed positive integers. The ordinary quantizer is obtained if $K, L \to \infty$.

We observe the following: The central state for the saturating adaptive quantizer may be defined exactly as in the ordinary type of quantizer; the important martingale properties, expressed in eq. (19) for the ordinary quantizer, carry over to the saturating type. The time-dependent transition equations of the saturating quantizer are characterized by numerous involved boundary equations. However, the bulk of the equations are of the form given in (15):

$$p(i; n+1) = \sum_{r=1}^{N} b^{(r)}(i - m_r)p(i - m_r; n)$$

$$-K + m_N \leqq i \leqq L + m_1. \tag{15'}$$

We do not give the remaining equations since we have no direct need for the time-dependent equations. In Appendix B we give, following the method and notation of Section 2.1, a complete set of reduced equations satisfied by the stationary probabilities.

II. STATIONARY DISTRIBUTIONS

Appendix A establishes the existence and uniqueness of a finite stationary distribution for the step size in the quantizers. The following sections establish the main qualitative properties of the stationary distributions for both the ordinary and saturating adaptive quantizers.

If we set $p(i; n+1) = p(i; n) = p(i)$ in the time-dependent equations, then the stationary probabilities are given by $\{p(i)\}$. Thus, the stationary probabilities of the ordinary adaptive quantizer are

obtained from

$$p(i) = \sum_{r=1}^{N} b^{(r)}(i - m_r)p(i - m_r), \quad i = 0, \pm 1, \pm 2, \cdots \tag{21}$$

and the normalization equation,

$$\sum_{-\infty}^{\infty} p(i) = 1.$$

2.1 A useful reduction of the equations for stationary probabilities

In each equation in (21), the maximum difference in the indices of the state probabilities is $(m_N - m_1)$. By exploiting a property of the stationary distribution, we now obtain a set of new equations where the maximum difference in the indices is $(m_N - m_1 - 1)$. The reduced set of equations together with the normalization equation is complete. A simple interpretation and the motivation of the reduced equation is given in Ref. 1; remark (*ii*) below gives an additional probabilistic interpretation. The reduced equations are important to us, as they allow us to consider only a smaller set of solutions.

For any integral j,

$$\begin{aligned}\sum_{-\infty}^{j} p(i) &= \sum_{i=-\infty}^{j} \sum_{r=1}^{N} b^{(r)}(i - m_r)p(i - m_r) \\ &= \sum_{i=-\infty}^{j-m_N} \left\{ \sum_{r=1}^{N} b^{(r)}(i) \right\} p(i) + \sum_{i=j-m_N+1}^{j-m_{N-1}} \left\{ \sum_{r=1}^{N-1} b^{(r)}(i) \right\} p(i) \\ &\qquad + \cdots + \sum_{i=j-m_1+1}^{j-m_1} b^{(1)}(i)p(i).\end{aligned}$$

Since $\sum_{r=1}^{N} b^{(r)}(i) = 1$, the above reduces to

$$\sum_{i=j-m_N+1}^{j} p(i) = \sum_{r=1}^{N-1} \sum_{i=j-m_{r+1}+1}^{j-m_r} \left\{ \sum_{s=1}^{r} b^{(s)}(i) \right\} p(i). \tag{22}$$

Define for $1 \leqq r \leqq N$ and all integral i,

$$\psi^{(r)}(i) \triangleq \sum_{s=1}^{r} b^{(s)}(i). \tag{23}$$

The quantities $\{\psi^{(r)}(i)\}$ may be directly obtained from the input distribution, since $\psi^{(r)}(i) = F(\xi_r C\gamma^i)$. From (22) we obtain the reduced equations

$$\boxed{\sum_{i=j-m_N+1}^{j} p(i) = \sum_{r=1}^{N-1} \sum_{i=j-m_{r+1}+1}^{j-m_r} \psi^{(r)}(i)p(i) \quad j = 0, \pm 1, \pm 2, \cdots} \tag{24}$$

In these equations, the set $[j - m_{r+1} + 1, j - m_r]$ is to be treated as empty if $m_r = m_{r+1}$.

Remarks:

(*i*) The manipulations leading to (24) are justified since they involve bounded quantities, as is implied by the existence of a unique finite stationary distribution.

(*ii*) Equation (24) is equivalent to the following identity, which is intuitively plausible and may be proven independently:

$$\Pr_s [\omega(n) \leqq j \text{ and } \omega(n+1) \geqq j+1] = \Pr_s [\omega(n+1) \leqq j \text{ and } \omega(n) \geqq j+1],$$

where the subscript s is being used to identify stationary probabilities.

(*iii*) Equation (24) may be used to give a simple proof of an identity (called simply an identity in Ref. 1 and "the design equation" in Ref. 4) involving the stationary state probabilities of the ordinary quantizer. Sum both sides of (24) for all integral j:

$$\sum_{j=-\infty}^{\infty} \sum_{i=j-m_N+1}^{j} p(i) = \sum_{j=-\infty}^{\infty} \sum_{r=1}^{N-1} \sum_{i=j-m_r+1}^{j-m_r} \psi^{(r)}(i) p(i).$$

The left-hand side is simply m_N and the right-hand side is

$$m_N - \sum_{r=1}^{N} m_r q_r,$$

where

$$q_r \triangleq \sum_{i=-\infty}^{\infty} \{\psi^{(r)}(i) - \psi^{(r-1)}(i)\} p(i).$$

Hence,

$$\sum_{r=1}^{N} m_r q_r = 0. \tag{25}$$

Equation (25) has a natural interpretation if we recognize that q_r is the stationary rth step occupancy probability, i.e.,

$$q_r = \Pr_s [\xi_{r-1}\Delta(n) \leqq |x(n)| < \xi_r \Delta(n)]. \tag{26}$$

The steps leading to eq. (24) may be repeated for the saturating adaptive quantizer, and a similar reduction may be achieved. These equations are given in Appendix B. The main recursion is identical to that of the ordinary quantizer, namely, eq. (24), and holds for all integral j, $-K + m_N \leqq j \leqq L + m_1 + 1$. Observe that the range over which (24) is valid, for the saturating quantizer, is such that

every state probability is included in at least one component of the recursion.

It may be verified by the reader that the identity in (25), the design equation of Ref. 4, does not hold for the saturating quantizer.

2.2 Localization property of the stationary distribution

We prove a fundamental distribution-free property of the stationary distribution of the step size. For both the ordinary and the saturating adaptive quantizers, we obtain sharp geometric bounds on almost all the stationary state probabilities as a function of the distance of the state from the 0 state. The actual bounds obtained are somewhat stronger than the above statement implies, since the rate parameter in the geometric bound itself decreases monotonically with increasing distance from the 0 state. These bounds show that a strong localization of the mass of the stationary distribution about the 0 state (central step size) is inherent in the random walk. Also, we found that it was necessary to prove a result like Theorem 1 before the effects of the multipliers on the dispersion of the stationary distribution could be quantified.

It is necessary to define certain vectors and matrices of dimensions $(m_N - m_1 - 1)$ and $(m_N - m_1 - 1) \times (m_N - m_1 - 1)$, respectively. Let $\mathbf{P}_i$ denote the column vector with the following components:*

$$\mathbf{P}_i \triangleq [p(i), p(i+1), \cdots, p(i + m_N - m_1 - 2)]^t. \tag{27}$$

Equation (24) may be used to construct matrices $\{\mathbf{A}_i\}$, which govern the transitions of the above vectors in the following manner:

$$\mathbf{P}_{i+1} = \mathbf{A}_i \mathbf{P}_i. \tag{28}$$

By examining (24) we observe that the elements of $\mathbf{A}_i$ depend on the quantities $\psi^{(r)}(i), \cdots, \psi^{(r)}(i + m_N - m_1 - 1)$, $1 \leqq r \leqq N$, and the subscript i indicates this dependence.

Theorem 1 (Localization Property): Let $i > 0$. For both the ordinary and saturating adaptive quantizers, there exists a constant weight vector with positive elements, λ, and a constant, $r > 1$, depending only on $\mathbf{A}_i$ such that, for all $j \geqq i$,

$$(\lambda^t \mathbf{P}_j) \leqq \left(\frac{1}{r}\right)^{j-i} (\lambda^t \mathbf{P}_i). \tag{29}$$

There exists the L_1-norm, $|\mathbf{x}| \triangleq \sum \lambda_k |x_k|$, of the vectors $\{\mathbf{P}_j\}$ which decreases geometrically as $|j - i|$ increases.

---

* The superscript t denotes the transpose.

An identical statement with $|j - i|$ *replacing the index* $j - i$ *in* (29) *is also true for* $i < 0$ *and all* $j \leqq i$.

*Remarks**:

(*i*) When r and λ in (29) are as constructed by us in the proof of the theorem, then the inequality in (29) becomes an equality if $\mathbf{A}_k = \mathbf{A}_i$ for $k = i, i + 1, \cdots, j$. This indicates that it is not possible to obtain tighter geometric bounds without making further assumptions on the distribution $F(z)$.

Using Theorem 1, we can give the following point-wise bound on the stationary state probabilities for both the ordinary and saturating adaptive quantizers:† let $i > 0$; then, for $j \geqq i$

$$\boxed{p(j + m_N - m_1 - 2) \leqq \left(\frac{1}{r}\right)^{j-i} (1^t \mathbf{P}_i) \leqq \left(\frac{1}{r}\right)^{j-i}}. \tag{30}$$

Similarly, for $i < 0$ and all $j \leqq i$,

$$p(j - m_N + m_1 + 2) \leqq \left(\frac{1}{r}\right)^{i-j} (1^t \mathbf{P}_i) \leqq \left(\frac{1}{r}\right)^{i-j}. \tag{30'}$$

The proof of (30) is as follows. Let λ_m denote the largest element of the vector λ occuring in Theorem 1 so that $1 \leqq m \leqq m_N - m_1 - 1$. From Theorem 1,

$$\lambda_m p(j + m - 1) \leqq \lambda^t \mathbf{P}_j \leqq \left(\frac{1}{r}\right)^{j-1} (\lambda^t \mathbf{P}_i) \leqq \left(\frac{1}{r}\right)^{j-i} \lambda_m (1^t \mathbf{P}_i),$$

and the inequalities in (30) follow.

Remarks:

(*ii*) Observe that for the bounds in (29) and (30) we may use any i, $0 < i \leqq j$, as the reference state. The choice of the best reference state depends on the behavior of r with i which, in turn, depends on the distribution $F(z)$. The main distribution-free property of $r(i)$, namely, statement (*iii*) of Lemma 1, indicates an advantage of choosing a large i for the reference state. In Section 2.4, we prove an assertion by implicitly using more than one reference state i.

The proof of Theorem 1 relies on two lemmas that we state here and prove in Appendix C.‡

---

* This remark implies the tightness of the bound in (29), which is lacking for the bound obtained in Ref. 1 for the two-bit quantizer.

† The vector 1 has every element equal to unity.

‡ Observe that neither $\mathbf{A}_i$ nor $\mathbf{A}_i^{-1}$ is a nonnegative matrix so that the usual Frobenius theory does not apply.

Lemma 1: *For every* $i > 0$,

(*i*) $\mathbf{A}_i$ *is nonsingular and* $\mathbf{A}_i^{-1}$ *has a unique positive real eigenvalue, say,* r. *Furthermore,* $r > 1$.

(*ii*) *Every element of the corresponding left eigenvector of* $\mathbf{A}^{-1}$, λ, *is of the same sign and nonzero, hence we may take* λ *to be a positive vector.*

(*iii*) r *which depends on* i *is monotonic, strictly increasing with* i.

Lemma 2: *For* $j \geqq i > 0$,

$$\lambda^t[\mathbf{A}_j^{-1} - \mathbf{A}_i^{-1}]\mathbf{P}_{j+1} \geqq 0. \tag{31}$$

Remarks:

(*iii*) It is not the case that $\lambda^t[\mathbf{A}_j^{-1} - \mathbf{A}_i^{-1}] \geqq 0$, so that (31) is not true if $\mathbf{P}_{j+1}$ is taken to be an arbitrary nonnegative vector.* In proving Lemma 2 it is necessary to take into account the fact that the vector $\mathbf{P}_j$, from which $\mathbf{P}_{j+1}$ evolves according to eq. (28), is itself nonnegative, and this implies that $\mathbf{P}_{j+1}$ is restricted to a cone that is a proper subset of the nonnegative quadrant.

Proof of Theorem 1: For $j \geqq i > 0$,

$$\begin{aligned}\lambda^t\mathbf{P}_j = \lambda^t\mathbf{A}_j^{-1}\mathbf{P}_{j+1} &= \lambda^t[\mathbf{A}_j^{-1} - \mathbf{A}_i^{-1}]\mathbf{P}_{j+1} + \lambda^t\mathbf{A}_i^{-1}\mathbf{P}_{j+1} \\ &= \lambda^t[\mathbf{A}_j^{-1} - \mathbf{A}_i^{-1}]\mathbf{P}_{j+1} + r\lambda^t\mathbf{P}_{j+1} \text{ from Lemma 1} \\ &\geqq r\lambda^t\mathbf{P}_{j+1} \qquad \text{from Lemma 2.}\end{aligned} \tag{32}$$

Hence, $(\lambda^t\mathbf{P}_j) \leqq (1/r)^{j-i}(\lambda^t\mathbf{P}_i)$ for all $j \geqq i$, as was to be proved.

As every element of $\mathbf{P}_j$ is nonnegative, the L_1-norm $|\mathbf{P}_j|$ is equal to $\lambda^t\mathbf{P}_j$. Finally, we may transfer the result that holds for $i > 0$ to the case of $i < 0$ by a simple renumbering of states in the manner that has been indicated in Ref. 1.

The notation common with Ref. 1 conceals some rather significant differences in both the main result (29) and its proof. In Ref. 1, the corresponding result involved λ and r, which were elements of the eigensystem of an additional matrix $\tilde{\mathbf{A}}_i$ obtained in an involved way from $\mathbf{A}_i$. The result in Lemma 2 has no counterpart in Ref. 1. The geometric bound obtained in Ref. 1 is peculiar to two-bit ($N = 2$) quantizers, and does not directly generalize. Also, the bound obtained here is stronger even for the case $N = 2$.

2.3 Lower bounds on the steepness factors, r(i)

Theorem 1 and the subsequent bound in (30) indicates that $r(i)$ is a local measure of the rate with which the stationary probabilities

* A vector is nonnegative if every element is nonnegative. The nonnegative quadrant in R^n is the set of all nonnegative vectors of dimension n.

change, and for this reason we find it natural to call $r(i)$ the local steepness factor. Here we go back to the definition of $r(i)$ as being the unique positive real root of the polynomial $C(\mu)$, eq. (60), to obtain the following bound on $r(i)$, which has the advantages of being explicit and being dependent only on the transition probabilities at state i. We make free use of this bound in the following section.

$$r(i) \geqq \rho(i) \triangleq \left[\frac{\sum_{r=1}^{\mu} (-m_r)\{\psi^{(r)}(i) - \psi^{(r-1)}(i)\}}{\sum_{r=\mu+1}^{N} m_r\{\psi^{(r)}(i) - \psi^{(r-1)}(i)\}} \right]^{1/m_N - m_1 - 1}, \quad (33)$$

where, of the N multipliers, only μ multipliers have values not exceeding unity, i.e.,

$$m_1, m_2, \cdots, m_\mu \leqq 0$$

and

$$m_{\mu+1}, m_{\mu+2}, \cdots, m_N > 0.$$

The bound $\rho(i)$ has certain interesting properties. First, observe that, by virtue of the definition of the central state [eqs. (8) and (9)], $\rho(i) > 1$ for all $i > 0$. Also, the sequence $\rho(i)$, is, like $\{r(i)\}$, monotonic, increasing with i. The numerator and denominator of the bracketed expression have interesting probabilistic interpretations: The numerator (denominator) is the expected change in state conditional on the transition being from state i to all states $i' \leqq i (i' > i)$.

The proof of eq. (33) is involved, and for the sake of brevity we omit giving it.

2.4 Effect of γ on the stationary distribution

We show in this section that the mass of the stationary distribution of the step size can be concentrated about the central step size to an arbitrary extent by making γ sufficiently close to unity. To show this, we first put together, from the results of the preceding two sections, a rather explicit bound on the stationary probability of the step size exceeding a particular value for a given γ, i.e., $\Pr_s [\Delta > C\gamma^i]$. This bound is in a form that allows direct comparison with the corresponding probability arising from the choice of $\gamma' = \sqrt{\gamma}$. By successively taking γ to be the square root of the preceding value, the bound on the probability can be made as small as desired. This procedure for proving the assertion is similar to the one we developed in Ref. 1. We restrict our attention to step sizes that exceed the central step size, i.e., $i > 0$, since a parallel argument holds for $i < 0$.

In the following discussion the quantity $(m_N - m_1 - 2)$ arises frequently, and it is convenient to denote this quantity by the symbol

ν. Clearly, ν is a measure of the spread in the log of the multipliers. For $i > 0$ and $r = r(i)$, we have from eq. (29) that

$$(\Sigma\lambda_i) \sum_{j=i+\nu}^{\infty} p(j) \leqq \sum_{j=i}^{\infty} \boldsymbol{\lambda}^t \mathbf{P}_j \leqq \boldsymbol{\lambda}^t \mathbf{P}_i \sum_{j=0}^{\infty} \left(\frac{1}{r}\right)^j = \boldsymbol{\lambda}^t \mathbf{P}_i \frac{r}{r-1}. \quad (34)$$

Now

$$r \geqq \rho(i), \quad (35)$$

where $\rho(i)$ is defined in eq. (33), and

$$\frac{\boldsymbol{\lambda}^t \mathbf{P}_i}{\Sigma\lambda_i} \leqq \max [p(i), \cdots, p(i+\nu)].$$

Since

$$\Pr_s [\Delta \geqq C\gamma^{i+\nu}] = \sum_{j=i+\nu}^{\infty} p(j),$$

we have, from eq. (34),

$$\boxed{\Pr_s [\Delta \geqq C\gamma^{i+\nu}] \leqq \frac{\rho(i)}{\rho(i)-1} \max [p(i), \cdots, p(i+\nu)]}. \quad (36)$$

Finally, from Eq. (30), for $i \geqq \nu + 1$,

$$\boxed{\max [p(i), \cdots, p(i+\nu)] \leqq \left[\frac{1}{\rho(1)}\right]^{i-\nu-1}}. \quad (37)$$

Equations (36) and (37) together give us the desired bound on the stationary probability of the step size exceeding a given value, which we now compare with a similar bound that holds for $\gamma' = \sqrt{\gamma}$. The prime superscript is used on symbols to denote the functional dependence of the associated quantities on γ'. In establishing the reference, i.e., central, step size corresponding to γ', minor differences exist depending on whether [see eqs. (8) and (9)]

(*i*) $$B(\gamma^{i-1}) > 0 \geqq B(\gamma^{i-\frac{1}{2}})$$

or

(*ii*) $$B(\gamma^{i-\frac{1}{2}}) > 0 \geqq B(\gamma^{i}). \quad (38)$$

We consider only (*ii*), in which case: $\omega'(n) = 2i \Longleftrightarrow \omega(n) = i$, and all the transition probabilities are simply related: $\psi^{(r)}(2i)' = \psi^{(r)}(i)$. As a consequence of the latter property, we have

$$\rho'(2i) = \rho(i). \quad (39)$$

Repeating the arguments leading to eqs. (36) and (37), we have

$$\Pr_s'[\Delta \geqq C\sqrt{\gamma}^{2i+\nu}] \leqq \frac{\rho'(2i)}{\rho'(2i) - 1} \max [p'(2i), \cdots, p'(2i + \nu)] \quad (40)$$

and

$$\max [p'(2i), \cdots, p'(2i + \nu)] \leqq \left[\frac{1}{\rho'(2)}\right]^{2i-\nu-2}. \quad (41)$$

By the fact that $\rho'(2i) = \rho(i)$, we have

$$\Pr_s'[\Delta \geqq C\sqrt{\gamma}^{2i+\nu}] \leqq \frac{\rho(i)}{\rho(i) - 1}\left[\frac{1}{\rho(1)}\right]^{i-\nu-1}\left[\frac{1}{\rho(1)}\right]^{i-1}. \quad (42)$$

Comparison with eqs. (36) and (37) completes the demonstration.

III. TRANSIENT RESPONSE

In this section, we are interested in the random time, called the adaptation time, taken for the step size of the quantizer to adapt from some arbitrary initial value to the central step size. It is necessary to have the adaptation time relatively small if the quantizer is to adequately track the scale variations of the input process. Also, it is reasonable to expect that, as γ is made large, the increased range of the multipliers [eq. (7)] will give the desired tracking. However, as a counterbalance, we already know from the preceding section that, with the correct choice of the log of the multipliers, $\{m_i\}$, the quality of steady-state performance is increasingly impaired as the value of γ is raised. From this brief discussion (see Ref. 1 for a more detailed discussion), it is clear that it is useful to have formulas for the efficient computation of the mean adaptation time and bounds that provide insight on the dependence of the time on the multipliers.

3.1 Mean time for first passage to the central state

We consider only the saturating adaptive quantizer since, as K and L are made large, the quantities obtained for this model approximate corresponding quantities for the ordinary adaptive quantizer. Also, for the usual reason only the case of positive initial states, $\omega(0) > 0$, is considered.

Let the initial step $\omega(0) = i > 0$ and let $T(i)$ denote the mean value of the random time τ where $\omega(\tau) \leqq 0$ and $\omega(n) > 0$ for all $n < \tau$. It can be shown that, as a consequence of the recurrence and irreducibility of the Markov chain (see Appendix A), the mean first passage time, $T(i)$, is finite with probability 1. If the first transition results in a transition to the state $i + m_r$, the process continues as if the initial state had been $i + m_r$. The conditional expectation of the first passage time is therefore $T(i + m_r) + 1$. From this argument, we

deduce that the following recursion is satisfied by the mean first passage time,

$$\boxed{T(i) = \sum_{r=1}^{N} b^{(r)}(i)\{T(i + m_r) + 1\} \quad -m_1 + 1 \leqq i \leqq L - m_N} \quad (43)$$

where, as in eq. (16), $b^{(r)}(i) = F(\xi_r C\gamma^i) - F(\xi_{r-1} C\gamma^i)$. Of course, $\sum_{r=1}^{N} b^{(r)}(i) = 1$. The recursive relation in (43) may be used to generate the entire sequence $\{T(i)\}$, provided $(m_N - m_1)$ initial conditions can be found. Now, by the same argument that led to eq. (43), we have

$$T(1 + m_1) = T(2 + m_1) = \cdots = T(0) = 0. \quad (44)$$

The remaining m_N initial conditions, namely,

$$T(1), T(2), \cdots, T(m_N),$$

are harder to obtain, and it is necessary to look more deeply into the dynamics of the process to obtain these quantities.

For every time instant, we define the L-dimensional vector $\mathbf{z}(n)$ with components $z(j; n)$, $1 \leqq j \leqq L$, where

$$z(j; n) \triangleq \Pr[\omega(n) = j \text{ and } \omega(s) \geqq 1 \text{ for all } s \leqq n]. \quad (45)$$

We show in Appendix D that the vectors $\mathbf{z}(n)$ evolve in time according to the homogeneous equation

$$\mathbf{z}(n + 1) = \mathbf{D}\mathbf{z}(n), \quad n \geqq 0, \quad (46)$$

where $\mathbf{D}$ is an $L \times L$ matrix. Also, in Appendix D we prove the following: For $i \geqq 1$,

$$\boxed{\begin{array}{l} T(i) = \sum_{j \geqq 1} x_j^{(i)}, \\ \text{where} \\ [\mathbf{I} - \mathbf{D}]\mathbf{x}^{(i)} = \mathbf{e}^{(i)} \end{array}} \quad (47)$$

and the elements of the L-vector $\mathbf{e}^{(i)}$ are zero everywhere except at the ith location where the element is unity. It is shown in Appendix D that $[\mathbf{I} - \mathbf{D}]$ is nonsingular.

The simple recursion in (43) may be used to generate the sequence $\{T(i)\}$ after obtaining the nonzero initial conditions via m_N inversions, as in (47). Alternatively, if $T(i)$ is required for only a few particular values of i, it may be easier to obtain them via the inversions in (47).

The bulk of the equations in (47) [see eq. (72)] are in the form encountered in the analysis of the stationary distribution, eq. (21).

Also, the elements of the vectors $x_j^{(0)}$ are all nonnegative. Hence, by applying the techniques and results of the preceding section, we may draw certain conclusions about eq. (47).

First, the bandwidth of the matrix $[\mathbf{I} - \mathbf{D}]$ may be reduced by 1 by carrying out the reduction of the equations described in Section 2.1. For $m_1 = -1$ and arbitrary values of $m_2, \cdots, m_N$, this step is enough to triangularize the matrix $[\mathbf{I} - \mathbf{D}]$ for any countable L and thus substantially simplify the computations. Second, we may conclude from Section 2.2 that, with increasing j, the solution elements $x_j^{(0)}$ decrease at least geometrically. This is a very useful property from the point of view of numerical inversion of $[\mathbf{I} - \mathbf{D}]$ for L large and the approximation of the solution for $L = \infty$ by finite L.

3.2 A bound on the mean first passage time

Let $T(i, j)$, $0 \leqq i < j$, denote the following mean first passage time: the initial state $\omega(0) = j$, first crossing occurs after τ transitions if $\omega(\tau) \leqq i$, and $\omega(n) > i$ for all $n < \tau$, and $T(i, j) = E(\tau)$. The quantity $T(j)$ of the preceding section is equivalent in our present notation to $T(0, j)$. We now give an explicit bound on $T(i, j)$ that provides some insight into the dependence of $T(i, j)$ on the multipliers.

For both the ordinary and saturating adaptive quantizer,

$$T(i, j) \leqq \frac{1}{C(i+1)}[(j - i) - (m_1 + 1)] \quad 0 \leqq i < j,$$

where

$$C(i) = \sum_{r=1}^{N-1} (m_{r+1} - m_r)\psi^{(r)}(i) - m_N \tag{48}$$

From the definition of the central state, eq. (18), and the monotonicity of $\psi^{(r)}(i)$ with respect to i, we observe that for $i > 0$, $C(i)$ is positive, monotonic, increasing with i. We only sketch the proof of (48) because the method of the proof is contained in the proof of the bound that we gave in Ref. 1 for the two-bit quantizer. First, recall [eq. (19)] that a supermartingale property exists that holds for both types of quantizers, according to which there is a net drift to the left from all states $j > 0$. Second, we define a new process in which $\omega'(n) = \omega(n) + nC(i+1)$ and show that the supermartingale property, i.e., $E[\omega'(n+1) \mid \omega'(n)] \leqq \omega'(n)$, is preserved for the range of n of interest. Finally, an application of Doob's theorem on optional stopping of supermartingales[10] on the new process yields the bound in eq. (48).

The bound provides some insight into the dependence of the mean adaptation times on the multipliers, and γ in particular, when the

initial and final step sizes are $C\gamma^j$ and C, respectively. Briefly, consider the effect of making $\gamma' = \sqrt{\gamma}$, i.e., $M_i' = \sqrt{M_i}$ and the spread of the multipliers is reduced. The number of states between the states corresponding to $C\gamma^j$ and C is doubled. Now $C(1)$ is hardly affected by the transformation and, as a consequence of the linear dependence of the bound on $T(i, j)$ on the distance $(j - i)$, we have the bound on the mean adaptation time approximately doubled. For $i = 0$ and $j \gg (-m_1)$, computations amply corroborate this conclusion.

IV. COMPUTATIONAL RESULTS

We present here a sampling of rather extensive computations done on three- and four-bit adaptive quantizers ($N = 4$ and 8, respectively) for independent identically distributed input sequences with gaussian distributions. Both uniform, i.e., $\xi_i = i$, and nonuniform quantizers were considered. Max[8] has shown in the nonadaptive framework that optimal nonuniform quantizers can yield an improvement in the signal-to-noise ratio of about 20 percent over optimal uniform quantizers with the number of bits in the range of interest here. We note that four-bit adaptive quantizers have been breadboarded in Bell Laboratories,[3] and that Jayant's[2] systematic numerical study is restricted to uniform quantizers up to three bits. We also observe that a simple search procedure of the "optimal" set of multipliers grows to be almost unmanageable and expensive when the dimension of the parameter spaces is 8.

Table I lists five quantizers with their respective parameters $\{m_i\}$. The parameter γ is not considered part of the characterization of the quantizer type. Among the quantizers investigated, the following five proved to be the most interesting in their respective classes, specified by number of bits and uniform or nonuniform. The first of the five, with $\gamma \approx 1.12$, is close to what Jayant calls the optimal, three-bit quantizer. The parameters $\{m_i\}$ were arrived at by the procedure described in remark (i), Section 1.2.

Table I — Five quantizers

Specifications			Designation
Uniform or Nonuniform	Number of Bits	$\{\log_\gamma (M_i)\}: m_1, \cdots, m_N$	
Uniform	3	$-1, -1, 2, 5$	*UQ*, 3 bits, No. 1
Uniform	3	$-1, 0, 1, 4$	*UQ*, 3 bits, No. 2
Nonuniform	3	$-2, -1, 2, 8$	*NUQ*, 3 bits
Uniform	4	$-2, -2, 0, 0, 2, 5, 10, 17$	*UQ*, 4 bits
Nonuniform	4	$-2, -2, 0, 0, 1, 2, 5, 16$	*NUQ*, 4 bits

The optimum division of the horizontal axis in Fig. 1, given by ξ_i, $i = 1, 2, \cdots, (N-1)$, was obtained from Max,[8] and we reproduce these parameters for the reader's benefit.

NUQ, 3 bits. $\{\xi_i\} = \{1.0, 2.097, 3.492\}$.
NUQ, 4 bits. $\{\xi_i\} = \{1.0, 2.023, 3.097, 4.256, 5.565, 7.142, 9.299\}$.

Table II lists some statistics of the stationary step-size distribution for unit variance of the input distribution. The stationary distribution was obtained by solving the stationary equations of the saturating adaptive quantizers with suitably large saturating levels ($K + L \approx 100$). We also give the stationary step-occupancy probabilities q_i, where $q_i = \Pr_s [\xi_{i-1}\Delta(n) \leqq |x(n)| < \xi_i\Delta(n)]$, as in eq. (26). Table II also gives, for purposes of comparison, corresponding quantities of the optimal nonadaptive quantizer obtained from Max.[8] In particular, $\hat{\Delta}$ is the optimal, nonadaptive step size.

Figures 2 to 5 show the mean adaptation times for inputs with unit variance. Figures 2 and 3 are concerned with the three types of three-bit quantizers for various values of γ. These figures plot the mean time taken by the quantizers to adapt to the central, and optimal, step size for various values of the initial step size. In Fig. 2, the initial step size exceeds the central step size, while the reverse case is considered in Fig. 3. Similarly, Figs. 4 and 5 plot data on the mean adaptation times for the uniform and nonuniform four-bit quantizers.

The purpose of the remaining tables (III to V) is to give the reader a feel for the relative performance of the five quantizers. We measure performance by the ratio of the input signal energy to the quantization

Table II — Statistics of the stationary step-size distributions

Type	γ	$\hat{\Delta}$ (Max)	$E(\Delta)$	$\sigma(\Delta)$	Step Occupancy Probabilities {adaptive quantizer} {optimal nonadaptive quantizer}
UQ, 3 bits No. 1	1.04	0.586	0.594	0.105	{0.445, 0.310, 0.156, 0.089} {0.442, 0.317, 0.162, 0.078}
UQ, 3 bits No. 2	1.04	0.586	0.613	0.089	{0.458, 0.314, 0.152, 0.075} {0.442, 0.317, 0.162, 0.078}
NUQ, 3 bits	1.04	0.501	0.522	0.114	{0.396, 0.317, 0.198, 0.088} {0.383, 0.323, 0.213, 0.081}
UQ, 4 bits	1.04	0.335	0.366	0.095	{0.285, 0.244, 0.182, 0.121, 0.075, 0.043, 0.024, 0.027} {0.263, 0.235, 0.188, 0.135, 0.086, 0.049, 0.025, 0.019}
NUQ, 4 bits	1.04	0.258	0.279	0.066	{0.219, 0.205, 0.178, 0.145, 0.110, 0.076, 0.045, 0.022} {0.204, 0.195, 0.177, 0.152, 0.121, 0.086, 0.049, 0.016}

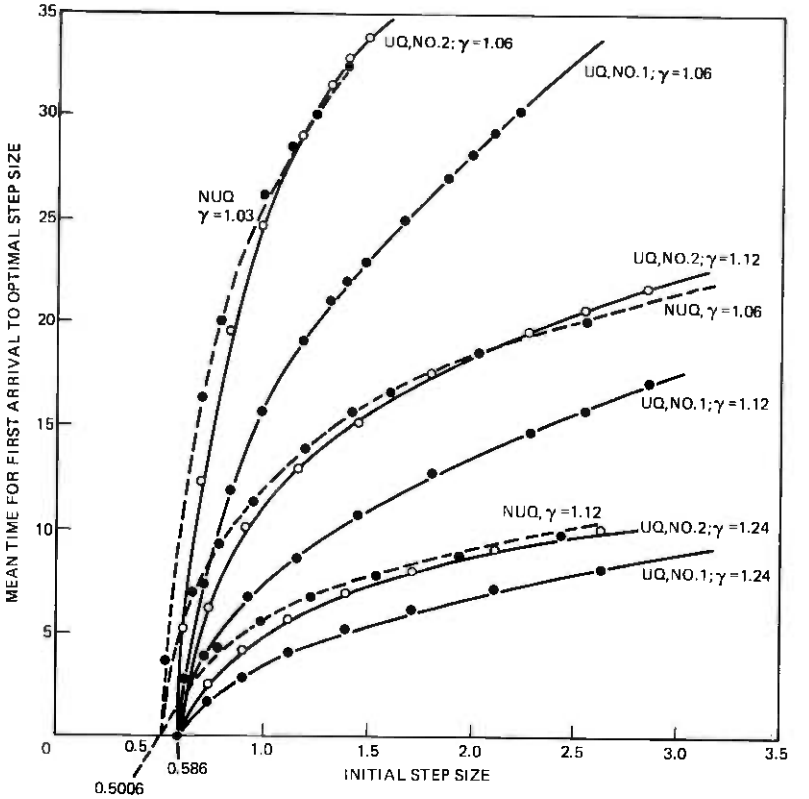


Fig. 2—Transient response of three three-bit quantizers.

error energy. Unlike all previous data, the data for these tables were obtained by Monte Carlo simulation. The interval of time over which performance was monitored is denoted by NA. Thus, signal energy is $\sum_{n=1}^{NA} x^2(n)$. The remaining parameter in the tables is the initial step size, Δ (initial). However, we do not list the raw initial step size, but

Table III* — S/N performance of two uniform three-bit quantizers (Main numbers are for *UQ*, three bits, No. 1; numbers in ( ) for *UQ*, three bits, No. 2)

Log {Δ(initial)/Δ̂}	NA = 10	NA = 100	NA = 1000	NA = 10,000
−1	6.92 (5.84)	14.4 (14.8)	17.4 (19.3)	17.7 (20.1)
0	25.7 (27.6)	19.1 (21.4)	17.9 (20.4)	17.8 (20.2)
1	0.549 (0.549)	3.94 (3.99)	13.1 (14.3)	17.1 (19.2)

* All logarithms in Tables III, IV, and V have base 10.

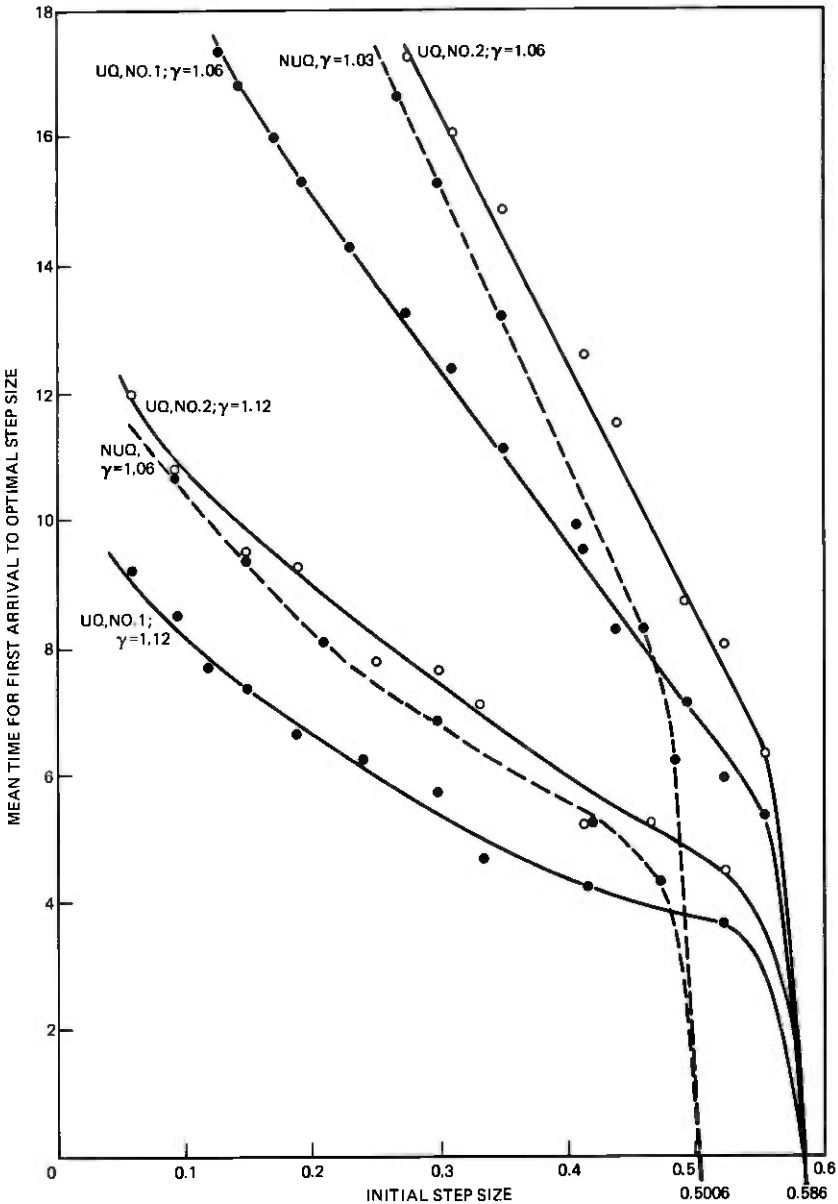


Fig. 3—Transient response of three three-bit quantizers.

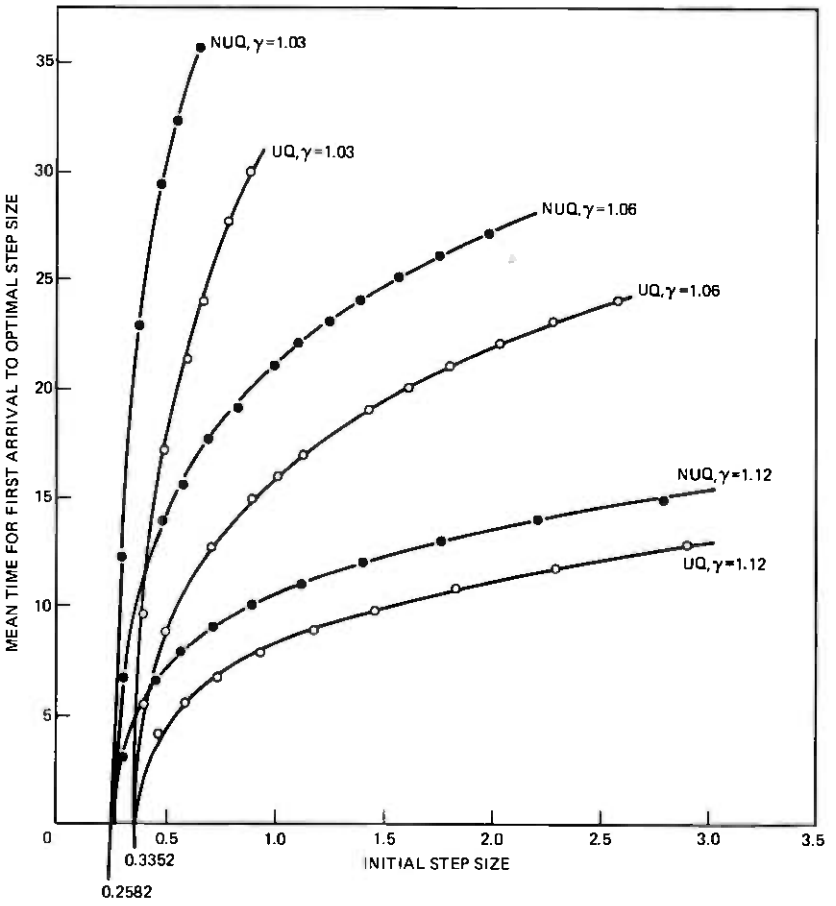


Fig. 4—Transient response of two four-bit quantizers.

the more relevant quantity $\Delta(\text{initial})/\hat{\Delta}$ where $\hat{\Delta}$ is, as usual, the optimal nonadaptive step size. After experimenting, we arrived at the following values of γ for the five quantizers, since they gave a suitable mix of performances over short (NA small) and long (NA large) runs.

Table IV — S/N performance of nonuniform three-bit quantizer (*NUQ*, three bits)

Log $\{\Delta(\text{initial})/\hat{\Delta}\}$	$NA = 10$	$NA = 100$	$NA = 1000$	$NA = 10{,}000$
−1	5.81	16.0	21.2	22.0
0	29.8	23.8	22.4	22.1
1	1.12	7.00	18.2	21.6

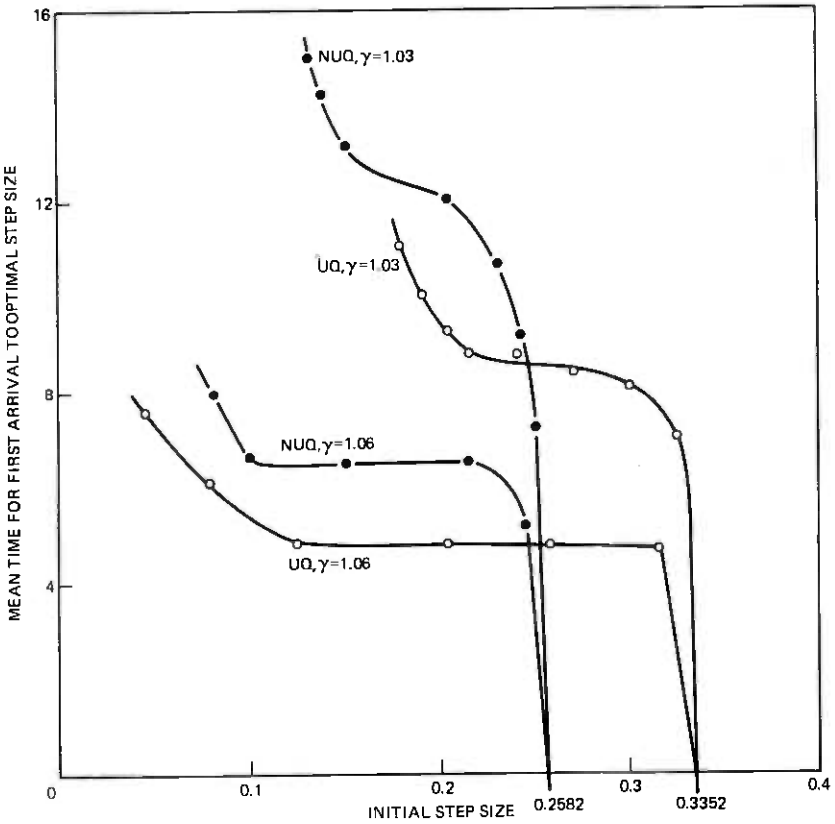


Fig. 5—Transient response of two four-bit quantizers.

For a particular input process, the relative weightings may be quite different, and γ may then be tuned accordingly.

Quantizer	γ
UQ, 3 bits, No. 1	1.12
UQ, 3 bits, No. 2	1.12
NUQ, 3 bits	1.06
UQ, 4 bits	1.06
NUQ, 4 bits	1.06

The following observations may be made on the above results. There is a pronounced asymmetry in performance with respect to $\log \{\Delta(\text{initial})/\hat{\Delta}\}$ over short runs ($NA = 10$ or 100). This is, of

Table V — S/N performance of uniform and nonuniform four-bit quantizers
(Main numbers are for *UQ*, four bits; numbers in ( ) for *NUQ*, four bits)

Log $\lvert\Delta(\text{initial})/\hat{\Delta}\rvert$	$NA = 10$	$NA = 100$	$NA = 1000$	$NA = 10{,}000$
−1	19.62 (21.65)	36.98 (47.30)	48.22 (67.35)	48.97 (71.50)
0	86.2 (111.0)	56.0 (80.1)	50.60 (72.50)	49.20 (71.90)
1	2.97 (4.86)	17.7 (27.6)	42.00 (62.00)	48.10 (70.30)

course, related to the contraction multipliers being grossly smaller than the expansion multipliers in all the quantizers considered (Table I). The s/n when $\Delta(\text{initial})/\hat{\Delta} = 1$ and $NA = 10$ is close to the s/n obtained with the step size optimally tuned to the known level of scaling of the input sequence. The steady but not excessive deterioration in performance with increasing NA is the price paid for adaptability: it is due to the fluctuations in step size arising from the random walk. Finally, we observe from Table V that there is a striking gain from nonuniform quantization, the extent of the gain being somewhat greater than what may be expected from previous results on nonadaptive quantizers.

APPENDIX A

Existence and Uniqueness of the Stationary Distribution

We establish in this appendix that, for independent identically distributed inputs, there exists a unique, finite stationary step-size distribution (invariant measure). The proof given here is via the construction of a stochastic Liapunov function, and it relies on a standard, unified theory of stochastic stability[11,12] that is well-known. The stochastic stability of the adaptive quantizer has been proved by Goodman and Gersho,[4] and the prime reason for including an alternative proof is our belief that familiarity with the method followed here may be beneficial to future workers in adaptive processes. The positive function that is proved to be a stochastic Liapunov function here is identical to the function that worked in Ref. 1 for the two-bit quantizer, and the proof is a straightforward generalization.

We consider in turn two properties of well-behaved Markov chains, namely, irreducibility and recurrence.

A.1 Irreducibility

The Markov chain is irreducible if and only if every state communicates with both neighboring states. This occurs if and only if

there exist nonnegative integers n_i and n_i', $1 \leqq i \leqq N$, such that

$$\Sigma m_i n_i = 1 \tag{49}$$

and

$$\Sigma m_i n_i' = -1. \tag{50}$$

It is an elementary fact from Euclid's theory that this occurs if and only if the integers $\{m_i\}$ are relatively prime, i.e., their greatest common divisor is unity.

A.2 Recurrence

Consider the following nonnegative function of the states

$$V(i) \triangleq |i| \quad i = 0, \pm 1, \pm 2, \cdots. \tag{51}$$

Let $D(i)$ be defined as

$$D(i) \triangleq E[V\{\omega(n+1)\} \,|\, \omega(n) = i] - V(i). \tag{52}$$

Now $D(i)$ is uniformly bounded from above. By the monotonicity of $\psi^{(r)}(i)$ with respect to i and the definition of the central state, (18), we obtain, for all $i \geqq (-m_1)$,

$$\begin{aligned} D(i) &= m_N - \sum_{r=1}^{N-1} (m_{r+1} - m_r)\psi^{(r)}(i) \\ &\leqq m_N - \sum_{r=1}^{N-1} (m_{r+1} - m_r)\psi^{(r)}(-m_1) < 0 \end{aligned} \tag{53}$$

and, for all $i \leqq -m_N$,

$$\begin{aligned} D(i) &= -m_N + \sum_{r=1}^{N-1} (m_{r+1} - m_r)\psi^{(r)}(i) \\ &\leqq -m_N + \sum_{r=1}^{N-1} (m_{r+1} - m_r)\psi^{(r)}(-m_N) < 0, \end{aligned} \tag{54}$$

where, as in eq. (23), $\psi^{(r)}(i)$ denotes $F(\xi_r C \gamma^i)$. Hence, by virtue of eqs. (53) and (54), $D(i) \leqq -\epsilon < 0$ for all but a finite set of states i, and $V(i)$ is a stochastic Liapunov function for the process.

From Kushner's Theorem 7,[11] we have recurrence and we can infer further, from Theorem 4, that there exists at least one finite invariant measure, i.e., stationary distribution. Also, since we have shown earlier that two or more disjoint self-contained subsets of the state space do not exist, we have, from Theorem 5, at most one invariant probability measure. The existence and uniqueness of a finite stationary

distribution for the step size of the ordinary adaptive quantizer is therefore established.

A.3 The saturating adaptive quantizer

The argument leading to irreducibility is intact. In addition, we have here that the end states $(-K)$ and L have period 1 and, since periodicity is a class concept (i.e., every state in a particular communicating class has the same periodicity), the entire Markov chain is aperiodic and, consequently, there is a single ergodic class that includes every state in the chain. Hence, the distribution at time n, $p(n)$ approaches p, the stationary distribution for all initial distributions, and furthermore every component probability of p is strictly positive.

APPENDIX B

The Saturating Adaptive Quantizer

We give in this appendix a set of equations satisfied by the stationary probabilities of the states in the saturating adaptive quantizer. These equations are complete and reduced by the method described in Section 2.1.

Let μ denote the number of contraction multipliers, i.e., multipliers having values less than 1, so that

$$m_1, \cdots, m_\mu < 0 < m_{\mu+1}, \cdots, m_N. \tag{55}$$

The tacit assumption that there are no multipliers exactly equal to unity is by no means necessary, but does lead to a simpler presentation.

The main set of equations is

$$\sum_{i=j-m_N+1}^{j} p(i) = \sum_{r=1}^{N-1} \sum_{i=j-m_r+1}^{j-m_r} \psi^{(r)}(i)p(i), \qquad -K + m_N - 1 \leqq j \leqq L + m_1. \tag{56}$$

The lower boundary equations are*

$$\sum_{i=-K}^{j-1} p(i) = \sum_{r=1}^{s-1} \sum_{i=-K\wedge(j-m_r+1)}^{j-m_r-1} \psi^{(r)}(i)p(i), \tag{57}$$

where $\mu + 1 \leqq s \leqq N$ and $-K + m_{s-1} + 1 \leqq j \leqq -K + m_s$. Finally,

* $x \wedge y = \text{Max}\,[x, y]$ and $x \vee y = \text{Min}\,[x, y]$.

the upper boundary equations are

$$\sum_{i=j-m_N+1}^{j} p(i) = \sum_{r=s}^{N-1} \sum_{i=j-m_r+1}^{L\vee(j-m_r)} \psi^{(r)}(i)p(i), \tag{58}$$

where $1 \leqq s \leqq \mu$ and $L + m_s \leqq j \leqq L + m_{s+1} - 1$.

APPENDIX C

Proofs of Lemmas 1 and 2

C.1 Proof of Lemma 1

(*i*) It can be shown that the determinant of the matrix $\mathbf{A}_i$,

$$\det [\mathbf{A}_i] = (-1)^{m_N - m_1}[1 - \psi^{(N-1)}(i)]/\psi^{(1)}(i + m_N - m_1 - 1).$$

As $\det [\mathbf{A}_i] > 0$, $\mathbf{A}_i^{-1}$ exists.

Since $\mathbf{P}_i = \mathbf{A}_i^{-1}\mathbf{P}_{i+1}$, we observe from the structures of $\mathbf{P}_i$ and $\mathbf{P}_{i+1}$ that the matrix $\mathbf{A}_i^{-1}$ is in companion form in that all rows except the first reflect shift operations, i.e., for $k \geqq 2$,

$$\begin{aligned} [\mathbf{A}_i^{-1}]_{k,l} &= 0 \quad \text{if} \quad l \neq (k-1) \\ &= 1 \quad \text{if} \quad l = (k-1). \end{aligned} \tag{59}$$

The elements of the first row of $\mathbf{A}_i^{-1}$ are obtained from the equation

$$\sum_{l=0}^{m_N-1} p(i+l) - \sum_{r=1}^{N-1} \sum_{l=m_N-m_r+1}^{m_N-m_r-1} \psi^{(r)}(i+l)p(i+l) = 0. \tag{24}$$

As the matrix $\mathbf{A}_i^{-1}$ is in companion form, we know that its characteristic polynomial is equal to within a constant of proportionality to the polynomial obtained by replacing, in eq. (24), $p(i + l)$ by $\mu^{m_N-m_1-1-l}$. That is, where

$$C(\mu) \triangleq (-1)^{m_N-m_1-1} \det [\mathbf{A}_i^{-1} - \mu\mathbf{I}],$$

we have

$$\begin{aligned} [1 - \psi^{(N-1)}(i)]C(\mu) = &\sum_{l=0}^{m_N-1} \mu^{m_N-m_1-1-l} \\ &- \sum_{r=1}^{N-1} \sum_{l=m_N-m_r+1}^{m_N-m_r-1} \psi^{(r)}(i+l)\mu^{m_N-m_1-1-l}. \end{aligned} \tag{60}$$

The quantity $[1 - \psi^{(N-1)}(i)]$ is merely the coefficient of $p(i)$ in eq. (24).

Scanning the coefficients of the polynomial $C(\mu)$, we observe that there is a single-sign alternation and, hence, by Descartes' rule, $C(\mu)$

has at most one real positive root. Since

$$C(0) = -\psi^{(1)}(i + m_N - m_1 - 1)/[1 - \psi^{(N-1)}(i)] < 0$$

and $C(\mu) \to \infty$ as $\mu \to \infty$, there exists exactly one real positive root. Let r denote this root.

Now

$$\begin{aligned}[1 - \psi^{(N-1)}(i)]C(1) &= m_N - \sum_{r=1}^{N-1} \sum_{l=m_N-m_r+1}^{m_N-m_{r-1}} \psi^{(r)}(i + l) \\ &< m_N - \sum_{r=1}^{N-1} \sum_{l=m_N-m_r+1}^{m_N-m_{r-1}} \psi^{(r)}(i) \\ &= \sum_{r=1}^{N} m_r\{\psi^{(r)}(i) - \psi^{(r-1)}(i)\}, \end{aligned} \tag{61}$$

where we have followed the usual convention in setting $\psi^{(N)}(i) = 1$ and $\psi^{(0)}(i) = 0$. So $C(1) < 0$ if $\sum_{r=1}^{N} m_r\{\psi^{(r)}(i) - \psi^{(r-1)}(i)\} \leqq 0$. The latter condition holds for all $i \geqq 0$ [see eqs. (17) and (18)]. Hence, $r > 1$.

(*ii*) Let us denote the elements of the first row of $\mathbf{A}_i^{-1}$ by $\{\alpha_l\}$ and $\{\beta_l\}$ so that the row appears as

$$[-\alpha_1 - \alpha_2 \cdots - \alpha_{m_{N-1}}\beta_1\beta_2 \cdots \beta_{-m_1}]. \tag{62}$$

One reason for expressing the row in this manner is that every α_l and β_l is strictly positive by eq. (24).

The left eigenvector λ of $\mathbf{A}_i^{-1}$ corresponding to the eigenvalue r satisfies, by definition, $\lambda^t\mathbf{A}_i^{-1} = r\lambda^t$. Examining the component equations, we find that

$$\lambda_{l+1} = (r^l + \alpha_1 r^{l-1} + \cdots + \alpha_l)\lambda_1 \quad 1 \leqq l \leqq (m_N - 1). \tag{63}$$

Also, for $1 \leqq l \leqq (-m_1)$,

$$\lambda_{m_N-m_1-l} = \frac{\lambda_{m_N-m_1-1}}{\beta_{-m_1}r^{l-1}}[\beta_{-m_1-l+1}r^{l-1} + \beta_{-m_1-l+2}r^{l-2} + \cdots + \beta_{-m_1}]. \tag{64}$$

Finally,

$$\lambda_{m_N-m_1-1} = \frac{\beta_{-m_1}}{r}\lambda_1. \tag{65}$$

Since the α's and β's are positive quantities, statement (*ii*) of the lemma is true.

(*iii*) The statement may be verified by examining the characteristic polynomial $C(\mu)$ in eq. (60) and observing that the quantities $\psi^{(r)}(i)$ are monotonic, increasing with i.

C.2 Proof of Lemma 2

It is required to prove that, for $j \geqq i > 0$,

$$\lambda_i[\mathbf{A}_j^{-1} - \mathbf{A}_i^{-1}]\mathbf{P}_{j+1} \geqq 0. \tag{66}$$

The matrices $\mathbf{A}_j^{-1}$ and $\mathbf{A}_i^{-1}$ are identical in all except the first row and also $\lambda^1 > 0$. Equation (66) is therefore equivalent to*

$$\mathbf{e}_1^t\mathbf{A}_j^{-1}\mathbf{P}_{j+1} \geqq \mathbf{e}_1^t\mathbf{A}_i^{-1}\mathbf{P}_{j+1}. \tag{67}$$

We prefer to show that

$$\theta^{(N-1)}(i)p(j) \geqq \theta^{(N-1)}(i)\mathbf{e}_1^t\mathbf{A}_i^{-1}\mathbf{P}_{j+1}, \tag{68}$$

where $\theta^{(N-1)}(i) \triangleq \{1 - \psi^{(N-1)}(i)\} > 0$. As $\mathbf{e}_1^t\mathbf{A}_j^{-1}\mathbf{P}_{j+1} = p(j)$, the lemma will then have been proved.

From eq. (24),

$$\theta^{(N-1)}(j)p(j) = p(j) - \psi^{(N-1)}(j)p(j)$$

$$= -\sum_{l=j+1}^{j+m_N-1} p(l) + \sum_{r=1}^{N-1}\sum_{l=j+m_N-m_r}^{j+m_N-m_{r+1}-1} \psi^{(r)}(l)p(l)$$

$$- \psi^{(N-1)}(j)p(j) \tag{69}$$

and

$$\theta^{(N-1)}(i)\mathbf{e}_1^t\mathbf{A}_i^{-1}\mathbf{P}_{j+1} = -\sum_{l=j+1}^{j+m_N-1} p(l)$$

$$+ \sum_{r=1}^{N-1}\sum_{l=j+m_N-m_r}^{j+m_N-m_{r+1}-1} \psi^{(r)}(l-j+i)p(l) - \psi^{(N-1)}(i)p(j). \tag{70}$$

Now

$$\theta^{(N-1)}(i)p(j) - \theta^{(N-1)}(i)\mathbf{e}_1^t\mathbf{A}_i^{-1}\mathbf{P}_{j+1}$$

$$\geqq \theta^{(N-1)}(j)p(j) - \theta^{(N-1)}(i)\mathbf{e}_1^t\mathbf{A}_i^{-1}\mathbf{P}_{j+1}$$

$$\geqq \sum_{r=1}^{N-1}\sum_{l=j+m_N-m_r}^{j+m_N-m_{r+1}-1} \{\psi^{(r)}(l) - \psi^{(r)}(l-j+i)\}p(l)$$

$$- \{\psi^{(N-1)}(j) - \psi^{(N-1)}(i)\}p(j) \geqq 0, \tag{71}$$

because of the monotonicity of $\psi^{(r)}(l)$, and the final term in the expression on the right-hand side of (71) is cancelled by an identical component ($r = N - 1$, $l = j + m_N - m_{r+1}$) of the leading part. The lemma is proved.

---

* The column vector with the leading element equal to unity and all other elements equal to zero is denoted by $\mathbf{e}_1$.

APPENDIX D

Two Equations Concerning Mean First-Passage Times

We prove two assertions made in Section 3.1, eqs. (46) and (47), concerning (*i*) the homogeneous evolution of the vectors $\{\mathbf{z}(n)\}$ via the matrix $\mathbf{D}$ and (*ii*) the explicit formula for the mean first-passage time, $T(i)$.

D.1 Derivation of eq. (46)

Let $X(n)$ denote the event $1 \leqq \omega(\tau) \leqq L$ for all τ, $0 \leqq \tau \leqq n$. Then, by definition,

$$z(j;n) = \Pr[\omega(n) = j \text{ and } X_n] \quad 1 \leqq j \leqq L.$$

Since it is also true that

$$z(j;n) = \Pr[\omega(n) = j \text{ and } X_{n-1}],$$

we have

$$z(j;n) = \sum_{i=1}^{L} \Pr[\omega(n) = j | \omega(n-1) = i, X_{n-1}] z(i;n-1).$$

We have obtained the quantities $\Pr[\omega(n) = j | \omega(n-1) = i, X_{n-1}]$ for $1 \leqq i, j \leqq L$ and, thereby, the following equations. In the following, μ denotes the number of contraction multipliers, that is,

$$m_1, m_2, \cdots, m_\mu < 0 < m_{\mu+1}, \cdots, m_N.$$

The basic recursion is, for $m_N + 1 \leqq j \leqq L + m_1$,

$$z(j;n) = \sum_{r=1}^{N} b^{(r)}(j - m_r) z(j - m_r; n-1). \tag{72}$$

The initial boundary equations are

$$z(j;n) = \sum_{r=1}^{\mu} b^{(r)}(j - m_r) z(j - m_r; n-1) \quad 1 \leqq j \leqq m_{\mu+1} \tag{73}$$

$$= \sum_{r=1}^{s} b^{(r)}(j - m_r) z(j - m_r; n-1) \quad m_s + 1 \leqq j \leqq m_{s+1}$$
$$s = \mu + 1, \mu + 2, \cdots, (N-1). \tag{74}$$

The final boundary equations are

$$z(j;n) = \sum_{r=s}^{N} b^{(r)}(j - m_r) z(j - m_r; n-1)$$
$$L + m_{s-1} + 1 \leqq j \leqq L + m_s, \; s = 2, 3, \cdots, \mu, \tag{75}$$

$$= \sum_{r=\mu+1}^{N} b^{(r)}(j - m_r)z(j - m_r; n - 1)$$

$$L + m_\mu + 1 \leqq j \leqq L - 1, \qquad (76)$$

$$= \sum_{r=\mu+1}^{N} \sum_{i=L-m_r}^{L} b^{(r)}(i)z(i; n - 1) \quad j = L. \qquad (77)$$

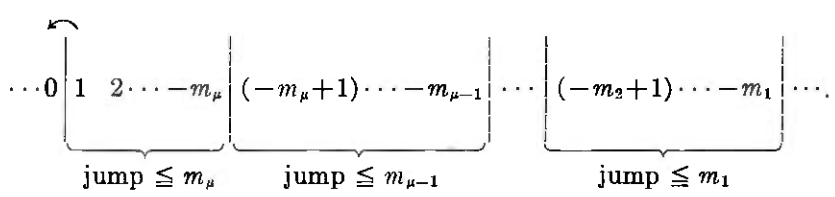
Equations (72) to (77) define the matrix D stated in the main text.

D.2 Derivation of eq. (47)

For $i = 1, 2, \cdots, L$, let

$$f(i; n + 1) \triangleq \Pr[\text{first passage occurs at } (n + 1) | \omega(0) = i]$$
$$= \Pr[\omega(n + 1) \leqq 0, X_n | \omega(0) = i]$$
$$= \sum_{j=1}^{-m_1} \Pr[\omega(n + 1) \leqq 0 | \omega(n) = j] z(j; n), \qquad (78)$$

with $\mathbf{z}(0) = \mathbf{e}^{(i)}$, the vector with every element equal to zero except for the ith element, which is unity. The event $\omega(n + 1) = k \leqq 0$ conditioned on $\omega(n) = j$ is associated with a jump $= k - j$. The following diagram illustrates the magnitudes of the jumps required for passage.



Equation (78) can be explicitly stated, thus,

$$f(i; n + 1) = \sum_{j=1}^{-m_\mu} \psi^{(\mu)}(j)z(j; n) + \sum_{j=-m_\mu+1}^{-m_{\mu-1}} \psi^{(\mu-1)}(j)z(j; n)$$

$$+ \cdots + \sum_{j=-m_2+1}^{-m_1} \psi^{(1)}(j)z(j; n). \qquad (79)$$

In the more convenient vector form,

$$f(i; n + 1) = \mathbf{c}^t \mathbf{z}(n), \qquad (80)$$

where the coefficients of the L-dimensional column vector $\mathbf{c}$ is obtained from (79), and we observe that only the leading $(-m_1)$ elements of $\mathbf{c}$ are nonzero.

The important fact about the vector $\mathbf{c}$ is that

$$\mathbf{c}^t = \mathbf{1}^t[\mathbf{I} - \mathbf{D}], \tag{81}$$

where $\mathbf{1}$ is the vector with every element equal to unity. Equation (81) may be established by either direct verification or by probabilistic reasoning. Now

$$\begin{aligned} T(i) &= \sum_{n \geq 0} (n+1) f(i; n+1), \\ &= \mathbf{c}^t \sum_{n \geq 0} n\mathbf{z}(n) + \sum_{n \geq 0} f(i; n+1), \\ &= \mathbf{c}^t \sum_{n \geq 0} n\mathbf{z}(n) + 1, \qquad (82) \\ &= \mathbf{1}^t[\mathbf{I} - \mathbf{D}] \sum_{n \geq 0} n\mathbf{z}(n) + 1 \qquad \text{from (81)}, \\ &= \mathbf{1}^t \sum_{n \geq 0} \mathbf{z}(n), \qquad (83) \\ &= \mathbf{1}^t \left[\sum_{n \geq 0} \mathbf{D}^n \right] \mathbf{z}(0), \qquad (84) \\ &= \mathbf{1}^t[\mathbf{I} - \mathbf{D}]^{-1}\mathbf{z}(0). \qquad (85) \end{aligned}$$

Equation (82) is obtained by noting that the probability that passage occurs at finite time is unity. In obtaining Eq. (83), we have used $\mathbf{z}(n+1) = \mathbf{D}\mathbf{z}(n)$ and that $\mathbf{1}^t\mathbf{z}(0) = 1$. The convergence of the series $\Sigma\mathbf{D}^n$ is a consequence of the fact that every eigenvalue of the matrix $\mathbf{D}$ is strictly inside the unit circle. We omit the proof of this assertion, as it is similar to the proof given in Ref. 1 in connection with the matrix $\mathbf{D}$ for two-bit quantizers.

Equation (85) with $\mathbf{z}(0) = \mathbf{e}^{(i)}$ is the same as eq. (47) in the main text.

REFERENCES

1. D. Mitra, "Mathematical Analysis of an Adaptive Quantizer," B.S.T.J., *53*, No. 5 (May-June 1974), pp. 867–898.
2. N. S. Jayant, "Adaptive Quantization with a One-Word Memory," B.S.T.J., *52*, No. 7 (September 1973), pp. 1119–1144.
3. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
4. D. J. Goodman and A. Gersho, "Theory of an Adaptive Quantizer," Proc. of December 1973 IEEE Symposium on Adaptive Processes, Decision and Control, pp. 361–365.
5. W. J. Dixon and A. M. Mood, "A Method for Obtaining an Analyzing Sensitivity Data," J. Amer. Statist. Assoc., *43*, 1948, pp. 109–126.
6. G. B. Wetherill, H. Chen, and R. B. Vasudeva, "Sequential Estimation of Quantal Response Curves: A New Method of Estimation," Biometrica, *53*, 1966, pp. 439–454.

7. Herman Chernoff, "Approaches in Sequential Design of Experiments," Technical Report, Stanford University, May 1973.
8. J. Max, "Quantization for Minimum Distortion," Trans. IRE, *IT-6*, March 1960, pp. 7–12.
9. J. C. Candy, "Limiting the Propagation of Errors in 1-Bit Differential Codecs," B.S.T.J., *53*, No. 8 (October 1974), pp. 1667–1676.
10. J. L. Doob, "Stochastic Processes," New York: John Wiley and Sons, 1953, pp. 300–301.
11. H. Kushner, "Stochastic Control," New York: Holt, Rinehart and Winston, 1971, pp. 188–224.
12. R. S. Bucy, "Stability and Positive Super-Martingales," J. on Differential Equations, *1*, 1965, pp. 151–155.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Cyclic Equalization—A New Rapidly Converging Equalization Technique for Synchronous Data Communication

By K. H. MUELLER and D. A. SPAULDING

(Manuscript received June 6, 1974)

A new technique for very fast start-up of adaptive transversal-filter equalizers used in high-speed synchronous data communications is presented. A special training sequence whose period in symbols is equal to the number of equalizer taps is used initially to achieve an open eye pattern. Rapid convergence, even over highly distorted channels, is obtained because an ideal reference sequence is available at the receiver, but it is not necessary to synchronize the ideal reference with the received sequence. The special choice of the training sequence automatically provides the synchronized ideal reference needed for fast convergence, but the resulting equalizer coefficients may be cyclically displaced from their proper positions. After the eye is opened by this process, the equalizer coefficients are rotated to their proper positions, and decision-directed equalization is used with either a longer training sequence or random data to achieve final tap settings. Adjustments during the training period can be made with a gradient-type algorithm or with stochastic adjustment techniques; an exact analysis is possible for both of these schemes. Cyclic equalization is shown to provide perfect equalization at evenly spaced points in the frequency domain.

I. INTRODUCTION

The effective data throughput in polling systems is, to a large degree, dependent on the start-up time of the data modems that are used in the network. Many of these systems operate at high speed and transmit data blocks of comparatively short duration. At 4800 b/s, a 1000-bit block is transmitted in about 200 ms, and to achieve a reasonable overall efficiency, the time needed to condition the modem for transmission (start-up) should be short in comparison to the time required to transmit an average block. This becomes increasingly difficult with

higher modem speeds. Prior to the transmission of the actual data, timing and carrier information must be recovered very accurately, and the adaptive equalizer that is necessary to cope with the linear channel distortion at such high speeds must be trained.

The time required to adjust the equalizer represents the bulk of the modem start-up time; it is thus important to study in detail the problems associated with fast equalizer start-up. The most common structure of such an equalizer consists of a transversal filter with a set of controlled gain coefficients that are spaced at the symbol interval T, and the start-up problem is to find an initial set of "reasonably good" values for these coefficients in a very short time. The purpose of this paper is to present a practical method for doing this.

We first provide some background and discuss some factors that affect equalizer start-up. This leads to the principle of cyclic equalization that is discussed in Section III. Sections IV through VIII discuss the operation of the cyclic equalizer using the mean-square tap-adjustment algorithm where averaging is used to compute the adjustment signals. The optimum tap coefficients are discussed and shown to provide perfect equalization of the channel at equally spaced points in the frequency domain. The relationship is explained between the eigenvalues of the channel correlation matrix, which control the convergence of the adaptive algorithm, and the discrete Fourier transform of the received training signal. Selection of the training sequence and the starting values of the tap coefficients and the effects of noise are also discussed. Finally, in the remaining sections, a more practical implementation is analyzed of the cyclic equalizer that does not use averaging in the tap adjustment algorithm (stochastic adjustment). The analysis of this algorithm is, in general, very difficult but, in the case of the cyclic equalizer, the time-varying difference equation that describes the noiseless equalization process can be solved exactly, and the conditions for this algorithm to be stable can be developed. Again, here the stability of the algorithm is related to the discrete Fourier transform (DFT) of the received signal. It is shown that the algorithm converges if the DFT of the received signal has no zero elements—that is, if the received signal spectrum has no nulls. This material along with a brief discussion of the asymptotic behavior of the algorithm is given in Sections IX through XII.

Within the paper, we also discuss various implementations, including a method to further speed up the tap calculation using an accelerated signal-processing technique. It will be seen that cyclic equalization is very attractive and economical to implement. Actual convergence is presented with some computer simulations that demonstrate the fast start-up capabilities of the new method.

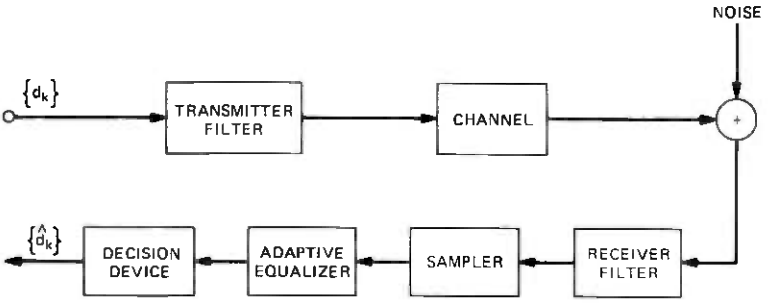


Fig. 1—Block diagram of data transmission system.

II. BACKGROUND

We will consider the pulse-amplitude modulated data system shown in Fig. 1. Data symbols, d_k, are transmitted every T seconds through a transmitter low-pass filter. This signal then passes through a distorting channel that has been made baseband by the modulation-demodulation process inherent in the modem, noise is added, and the composite signal is sampled every T seconds after the receiver filter. The sampled signal vector $\mathbf{x}_k$ is then equalized by a transversal filter with coefficient vector $\mathbf{c}$ (see Fig. 2) to produce an output $y_k = \mathbf{x}_k^T\mathbf{c}$ upon which the decision device operates to produce estimates, $\hat{d}_k$, of the transmitted symbols. The receiver structure has the form of the optimum linear receiver[6] but, because the channel is never precisely known and changes with time, the transversal equalizer is made adaptive to optimize performance.

Our concern in this paper is with the equalizer and ways to make it adapt rapidly from some initial setting to its final setting. A large number of papers, a partial list of which has been included in Refs. 1 to 51, have been written about equalizers, algorithms for adjusting

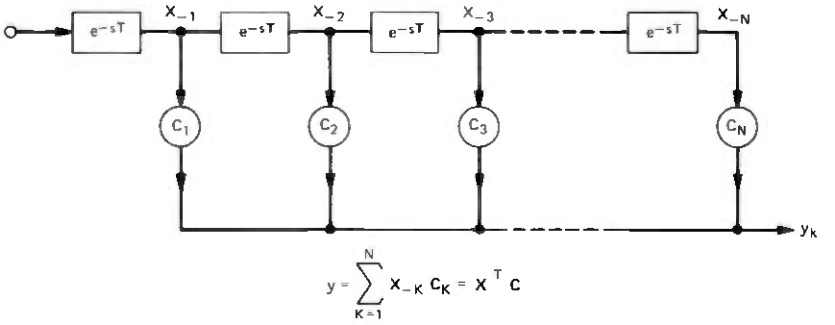


Fig. 2—Nonrecursive transversal filter.

them, and the speed with which these algorithms converge. In developing procedures for adapting the equalizer coefficients, some appropriate performance measure must be defined that will discriminate between good and bad coefficient vectors. Although our goal is to minimize the probability of error, this criterion is too difficult to work with directly. As a result, secondary performance measures such as the peak distortion,[2]

$$D = h_0^{-1} \sum_{k \neq 0} |h_k|, \tag{1}$$

or the mean-square error,

$$\epsilon = E\{|y_k - d_k|^M\}, \qquad M = 2, \tag{2}$$

are used. In (1), h_k is the sampled system impulse response. The peak distortion is related to the "eye opening,"[6] and for binary symbols and noiseless transmission, $D < 1$ implies no decision errors. In (2), $E\{\cdot\}$ is the expectation operation and $y_k - d_k$ is the remaining error at the equalizer output. These performance measures (M could be greater than 2, if desired) can be shown to be convex functions of the equalizer coefficients, thereby proving the existence of a global minimum.

We will work primarily with the mean-square error (MSE) criterion. This criterion includes the effects of noise, whereas the peak distortion criterion does not, it is convenient to work with mathematically, it can be used to bound the probability of error,[5,2] and it leads to adaptive algorithms that are easy to implement. Using the MSE, the optimum coefficient vector for the equalizer can be determined easily. Assuming $E\{d_k^2\} = 1$, we have from (2)

$$\epsilon = \mathbf{c}^T\mathbf{A}\mathbf{c} - 2\mathbf{c}^T\mathbf{v} + 1, \tag{3}$$

where

$$\mathbf{A} = E\{\mathbf{x}_k\mathbf{x}_k^T\} \tag{4}$$

is the signal autocorrelation matrix,

$$\mathbf{v} = E\{d_k\mathbf{x}_k\} \tag{5}$$

defines the signal correlation vector, and $\mathbf{x}_k$ is the vector of tap signals at the kth time instant. Finding the gradient of (3) with respect to the tap gains gives

$$\mathbf{g} = 2E\{(y_k - d_k)\mathbf{x}_k\} = 2(\mathbf{A}\mathbf{c} - \mathbf{v}). \tag{6}$$

Our optimization problem has a unique solution if A^{-1} exists. Setting (6) equal to zero yields

$$\mathbf{c}_{\text{opt}} = \mathbf{A}^{-1}\mathbf{v} \tag{7}$$

$$\epsilon_{\text{opt}} = 1 - \mathbf{v}^T\mathbf{c}_{\text{opt}} = 1 - \mathbf{v}^T\mathbf{A}^{-1}\mathbf{v}. \tag{8}$$

The problem of equalizer start-up is simply to find the solution to (7) in a rapid and economical manner. The economical part of the question is very important. One can imagine a start-up procedure that operates by sending a special training signal for a short period of time. The received signal, $x(t)$, is stored at the receiver. The training sequence is known at the receiver, but its absolute time reference is not known. The receiver contains a very fast high-power computer which now, in essentially no time, computes (8) for a large number of different time references and finds the time reference for the locally stored training sequence that minimizes ϵ_{opt}. The computer has thus accomplished both synchronization and equalization. This hypothetical system achieves a start-up time limited only by the time required to transmit the training signal but, with today's technology, its speed-cost product, if you will, is very poor. It does not represent an economical solution to the problem. Many currently proposed fast start-up equalizers, although not as extreme as this example, still do not present cost-effective solutions.

In addition to the economic aspect, this example illustrates two other important points. The first is the solution of (7). Much of the work on equalization is concerned with efficient algorithms that avoid the direct matrix inversion and obtain an iterative solution. Often, however, the time required to perform the calculations in (4) and (5) is not explicitly considered in evaluating start-up time. Second, synchronizing the stored reference signal in the receiver can take significant time, and that aspect of start-up time seems to be universally ignored.

Now we consider the solution of (7) in more practical terms. A well-known approach for solving (7) is

$$\mathbf{c}_{m+1} = \mathbf{c}_m - \beta_m(\mathbf{A}\mathbf{c}_m - \mathbf{v}), \tag{9}$$

i.e., a first-order steepest-descent gradient algorithm. For appropriate conditions on β_m, $\mathbf{c}_m$ converges to $\mathbf{c}_{\text{opt}}$.

According to (6), the gradient is obtained by correlating the tap-signal vector and the error voltage

$$\mathbf{g} = 2E\{e_k\mathbf{x}_k\}. \tag{10}$$

From an implementation point of view, this is a convenient quantity because the signal vector, $\mathbf{x}_k$, is readily available, and the error, $e_k = y_k - d_k$, can be estimated. A difficulty still remains in that the expected value is not available in real time and must be estimated by averaging over a finite number of symbols. The difference equation (9) then takes the form

$$\left.\begin{aligned}\mathbf{c}_{m+1} &= \mathbf{c}_m - \beta_m \cdot \frac{1}{L}\sum_{k=mL}^{mL+L-1}\mathbf{x}_k(\mathbf{x}_k^T\mathbf{c}_m - d_k)\\ &= \mathbf{c}_m - \beta_m(A_m\mathbf{c}_m - \mathbf{v}_m)\end{aligned}\right\}. \tag{11}$$

Averaging is done over L symbols between succeeding adjustments. If random data are transmitted, A_m and $\mathbf{v}_m$ will depend on the particular signal pattern of each iteration interval and are random variables with mean A and $\mathbf{v}$ and variances decreasing with longer averaging interval L. The analysis of the behavior of (11) is difficult, particularly when we try to determine ways of improving the convergence rate. By reducing L, we can make many more iterations in a given time, but we must use a smaller β-value to take into account the larger variance of the calculated gradient. Longer averaging between each step would take more time but give a better estimate for the gradient, and therefore allow a somewhat higher value of β. Monsen[13] has studied the optimization of the averaging interval, assuming an ideal reference and Gaussian signals. In this special case, the optimum value is $L = 1$; i.e., corrections are made after each symbol and no averaging at all is done. This method is often called "stochastic approximation," because the corrections are stochastic quantities whose means equal the desired gradient.

At this point, it appears that the mean-square algorithm with no averaging, i.e.,

$$\left.\begin{aligned} \mathbf{c}_{m+1} &= \mathbf{c}_m - \beta_m e_m \mathbf{x}_m \\ &= (I - \beta_m \mathbf{x}_m \mathbf{x}_m^T)\mathbf{c}_m + \beta_m d_m \mathbf{x}_m \end{aligned}\right\}, \tag{12}$$

is an attractive scheme to investigate further to obtain fast real-time convergence. There remain, for the moment, two main difficulties that need further discussion. The first one is the problem of obtaining the data values d_k. They can be estimated in the usual way from a threshold decision, but on channels with large distortion the initial error rate may be close to 50 percent and estimates $\hat{d}_k$ are very unreliable in such a situation. An algorithm with a decision-directed reference may thus behave erratically, and convergence cannot be guaranteed. The results of a few simulations will give some further insight.

The channel assumed for the simulation consisted of a 10-percent cosine roll-off baseband filter with parabolic delay distortion [$5.4T$ at the Nyquist frequency $(1/2T)$] and a sampling offset of $0.3T$ from the peak of the response. The resulting channel response from a single pulse is shown in Fig. 3. This same pulse was also used by Hirsch and Wolf.[11] The initial peak distortion is 2.62, resulting in a completely closed eye pattern.

The first simulation is for the algorithm (12), but an estimated reference (obtained from a threshold decision) was used. Figure 4 shows the resulting peak distortion versus the number of symbols for four different step sizes. Decision errors are responsible for random distortion increases rather than reductions. This is avoided in Fig. 5

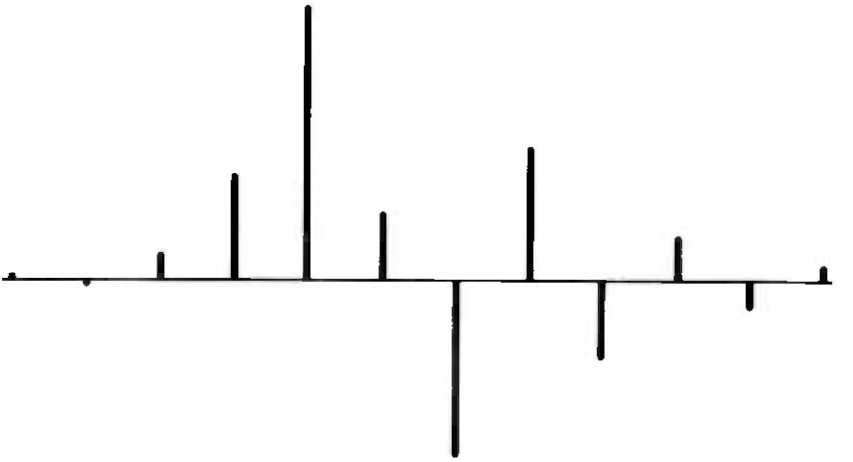


Fig. 3—Impulse response with peak distortion of 2.62.

where we have repeated the same runs with an ideal reference signal. The improvement is significant. Note that the ideal reference signal is really needed only until the peak distortion has decreased sufficiently to yield an open eye pattern; from this time on, the error probability is essentially zero, and a decision-directed reference can be used.

The difficulty in providing an ideal reference signal lies in the synchronization problem. Remember that we require such a signal only in channels with very large amounts of distortion, but achieving reliable synchronization in the presence of severe distortion is a problem in itself that usually calls for time-consuming correlation methods.[5]

A second problem is associated with the choice of the training sequence. Obviously, a strictly random data pattern would be a bad choice, since transitions would only occur on a probabilistic basis and not be guaranteed. The variability of repeated convergence runs would be large. This can be avoided by transmitting a short-period training sequence. Even if the starting point occurred at random, convergence would be more predictable. We know that we cannot make the period of the training sequence shorter than the duration of the impulse response of the equalizer; otherwise, the tap signals would not be linearly independent, the correlation matrix A would be singular, and a unique solution for the optimum tap vector $\mathbf{c}$ would not exist. We will, however, study the limiting case where the period, in symbols, of the training sequence is equal to the number of taps on the equalizer. This will lead directly to the idea of the cyclic equalization. Before we

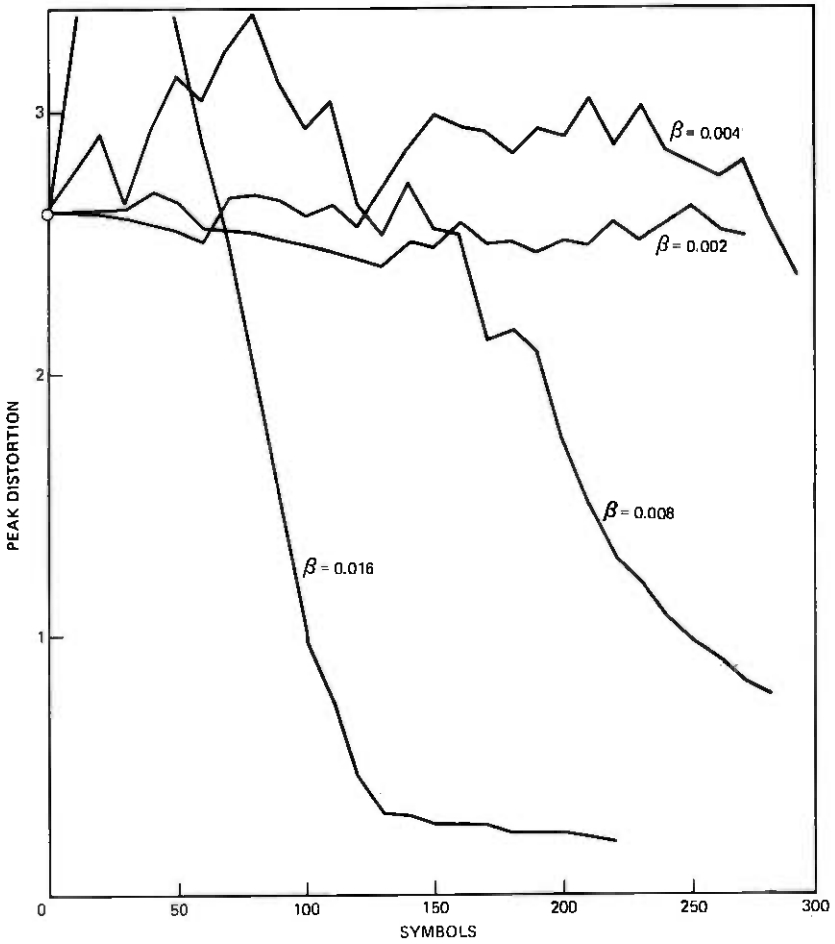


Fig. 4—Convergence behavior of stochastic adjustment algorithm (12) with a decision-directed reference.

do this, we will provide some additional insight by a short discussion of the frequency domain aspects of the equalization problem.

III. PRINCIPLE AND IMPLEMENTATION OF CYCLIC EQUALIZATION

Let the spectrum of the received data signal be $G(\omega)$ and assume that this signal is applied to an N tap equalizer with coefficients c_n, $n = 0, \cdots, N-1$. The resulting output spectrum is

$$X(\omega) = G(\omega) \sum_{n=0}^{N-1} c_n \exp(-j\omega nT), \tag{13}$$

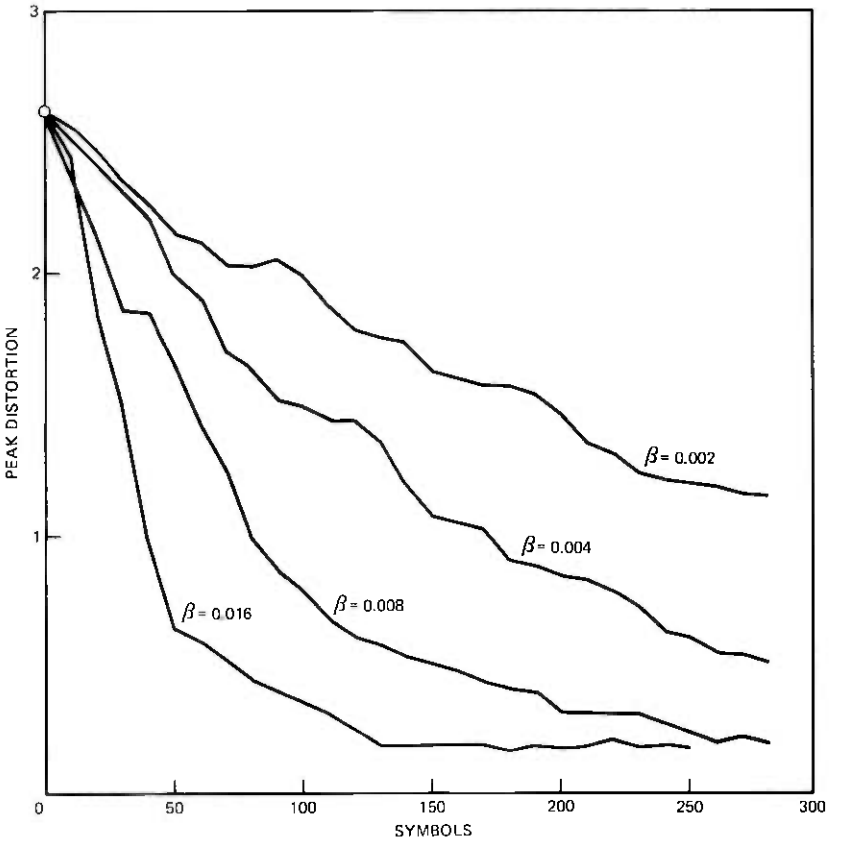


Fig. 5—Convergence behavior of stochastic adjustment algorithm (12) with an ideal reference.

and the overall system would be distortion-free (Nyquist criterion) if

$$\sum_k X\left(\omega + \frac{2\pi k}{T}\right) = \exp(-j\omega\tau), \qquad |\omega| < \frac{\pi}{T}. \tag{14}$$

Combination of (13) and (14) yields the condition

$$\sum_{n=0}^{N-1} c_n \exp(-j\omega nT) \sum_k G\left(\omega + \frac{2\pi k}{T}\right) = \exp(-j\omega\tau), \quad |\omega| < \frac{\pi}{T}. \tag{15}$$

Obviously, (15) cannot be satisfied for a finite N and an arbitrary $G(\omega)$. Usually, the coefficients c_n are chosen according to a minimum mean-square-error (MMSE) criterion in the time domain, which is equivalent to an MMSE criterion of (15) in the frequency domain. The problems of such an approach have been discussed in Section II, and we have seen that the commonly used iterative search schemes can be

very efficient during the tracking mode, but initial training may not be without problems.

A closer look at (15) shows that the left-hand side is a linear combination of the coefficients c_n. Although perfect equalization cannot be achieved at all frequencies, it is possible to obtain zero error at a number of specified frequencies ω_m within the range $|\omega_m| < \pi/T$. This is, of course, also true with an MMSE approach, since the resulting transfer function will oscillate around the desired one; i.e., the error will ripple between positive and negative values. The crossing frequencies are, however, not known and usually not of interest. In the new scheme we propose here, we will do exactly the contrary: We will precisely specify the crossing frequencies, although we realize that such an approach will, in general, not yield MMSE. Specifying the frequencies ω_m where perfect equalization is obtained will transform the condition (15) in a set of linear equations for the coefficients c_n. Obviously, we have to consider two cases:

(*i*) N = even: $N/2$ different frequencies $\omega_m \neq 0$ must be specified.
(*ii*) N = odd: $(N-1)/2$ different frequencies $\omega_m \neq 0$ and $\omega = 0$ must be specified to obtain a unique solution for the c_n's.

Theoretically, a set of reference tones ω_m could be transmitted, $G(\omega_m)$ measured at the receiver, and the coefficients computed from (15). Fortunately, it is possible to propose a much more attractive solution.

The generation of the reference tones can be accomplished in a straightforward way if we select the frequencies ω_m equally spaced across the Nyquist band; a suitable periodic data sequence of length NT will produce such spectral lines at $\omega_m = 2\pi m/NT$. Note that the number of symbols in such a training sequence is equal to the number of taps of the equalizer. This choice is extremely important and provides a number of unique advantages to achieve fast equalizer start-up.

We now discuss in detail such a training procedure. Assume an equalizer where an ideal reference signal is used and the period of the training sequence is equal to the number of taps on the equalizer. Assume for the moment that the channel is distortionless and the ideal reference is synchronized with the incoming signal. If we let the adaptive algorithm adjust the equalizer taps, the center tap on the equalizer will become unity, and all the others will be zero. This is really what we mean when we say the reference is synchronized; that is, the optimum equalizer coefficients are centered on the equalizer rather than shifted off to one end or the other. Now again, for this "ideal" example, if the reference signal is delayed by one symbol from perfect synchronization, the adaptive algorithm will cause the equalizer coefficient one position removed from the center to become unity, and all the

others will be zero. The movement of the unity gain tap by one position indicates a one-symbol delay in synchronization of the ideal reference. In an actual situation, the other taps on the equalizer will be nonzero and, with an unsynchronized reference, the adaptive algorithm will cause tap coefficients to occur that are cyclically rotated from those that would occur if the reference were synchronized.

To say this another way, if the training sequence is periodic with a period equal in symbols to the number of taps of the equalizer, the received signal is then also periodic (neglecting noise effects), and one full period of the sequence is always stored in the equalizer. Each symbol that is shifted out at the end of the delay line is replaced by an identical new symbol at the input. This is more clearly shown in Fig. 6 for a seven-tap equalizer with taps c_1 through c_7 and a seven-bit sequence x_1 through x_7. At time $t_0 + 2T$, it is seen that the stored sequence has been cyclically shifted by two units as compared to the time t_0. But it is also seen that the same output signal $y(t_0 + 2T)$ could have been obtained at time $t = t_0$ if the taps were cyclically shifted back by two positions. Thus, at any given time $t = t_0$, all outputs $y(t = t_0 + kT)$ can be obtained with a suitable cyclic shift of the components of the tap vector.

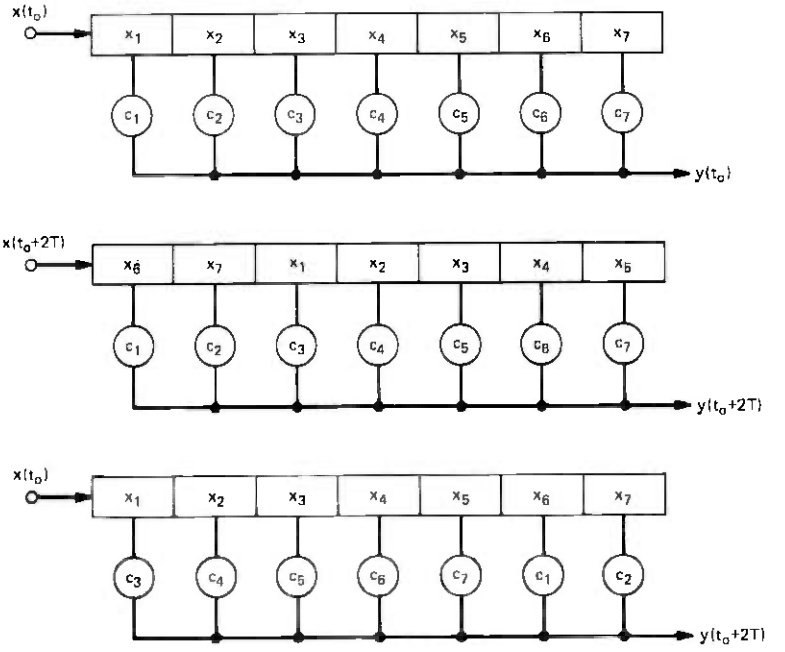


Fig. 6—Basic idea of cyclic equalization.

This feature provides an elegant solution to the synchronization problem. Any cyclic shift between the received sequence and the reference sequence will yield a compensating cyclic shift of the (same) tap coefficients. It is, therefore, not necessary to achieve synchronization prior to equalization, but it is of course necessary to properly shift the tap coefficients after initial training to prepare the equalizer for random data. This can easily be done by cycling them in such a way that the largest coefficient is aligned with a reference position, e.g., the center tap. Because of its particular features just described, we will call this novel start-up scheme "cyclic equalization."[42,43]

The possible structure of such an equalizer is outlined in Fig. 7. An internal word generator produces an ideal reference sequence that need not be synchronized with the received sequence. All taps are initially preset to identical values (since the location of the "center tap" is not known). The equalizer will then produce a set of taps with the particular cyclic shift corresponding to the "synchronization delay." After this initial training, the tap coefficients are cyclically shifted for "alignment," as outlined above. At this point, the equalizer has reasonably good tap coefficient settings and the peak distortion at the output is less than unity; i.e., the eye is open and, in the absence

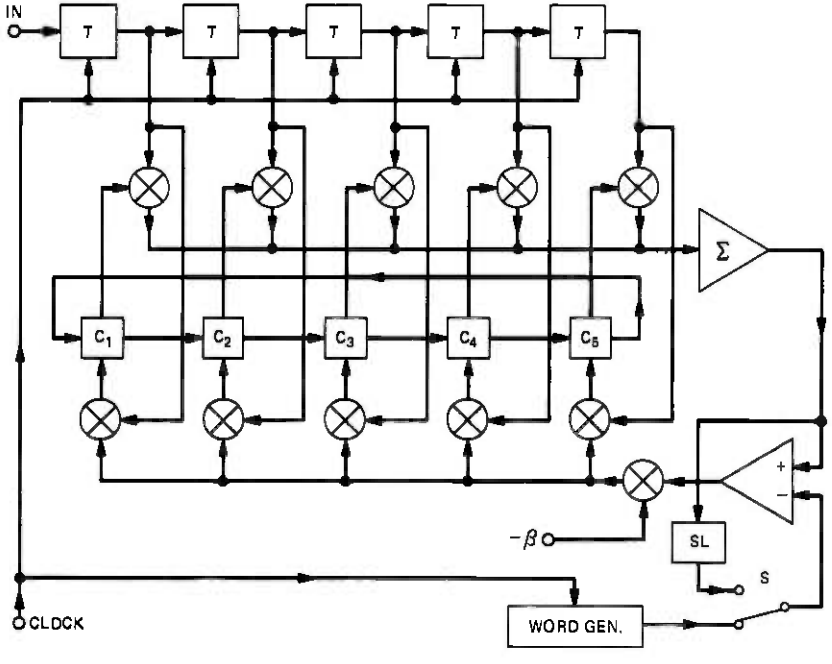


Fig. 7—Block diagram of an equalizer with cyclic start-up.

of noise, errorless decisions can be made. Fast coarse adjustment of the tap coefficients has been achieved without wasting time synchronizing the ideal reference. Once the eye is open, decision-directed equalization can be used with a somewhat longer training sequence or random data to achieve the final fine adjustment of the tap coefficients.

The fact that mean-square equalization with a training sequence period equal to the length of the equalizer can give very fast and very consistent, relative to the starting point of the adaptation, equalization has been demonstrated in numerous simulations. One of these is illustrated in Fig. 8. The same channel is used for this example as was used previously; the peak distortion is 2.62, the signal-to-noise ratio is 30 dB, and the step size is 0.04. In this case, the equalizer has 15 taps and a 15-bit maximum length training sequence is used because of its nice spectral properties. Adjustments are made at the symbol

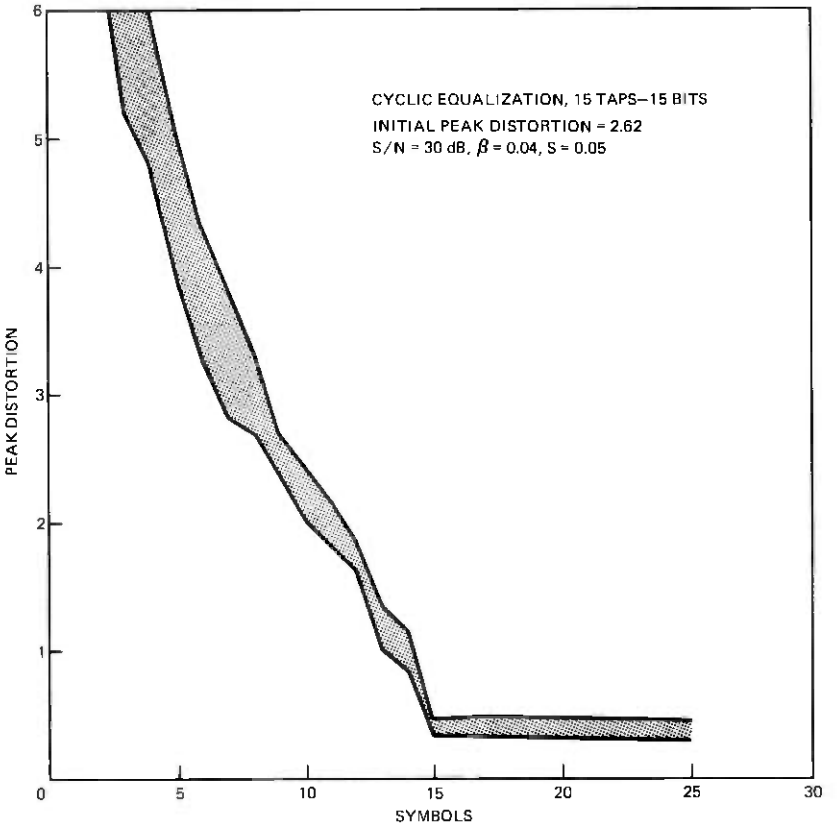


Fig. 8—Start-up behavior with cyclic equalization.

rate. The shaded region in the figure contains all 15 possible convergence curves that correspond to the different starting points for adaptation. Not only are all the convergence curves very similar, but they all achieve a peak distortion of about 0.4 or less in 15 symbols.

A few words are in order about the presetting of the tap coefficients. Because an unsynchronized reference is used, the location of the largest coefficient is *a priori* unknown. It is therefore reasonable to preset all coefficients to identical initial values s, as we have already mentioned above. With most channels, tap coefficients of both polarities will evolve so that one might consider setting $s = 0$ for an unknown channel. The large final value of the center tap would, however, suggest that slightly biased initial conditions might give faster convergence; we will make more precise statements about that in Section VII.

The discussed method of presetting has, of course, some consequences if a channel with low distortion or even an ideal channel were used. In such a situation, a conventional equalizer could do a better job because it would be started with the optimum tap settings ($c_k = \delta_{0k}$) right away and need not make any corrections at all. The cyclic equalizer would have to "converge" even with an ideal input signal; simulations of this case have shown a convergence plot similar to that of Fig. 8.

As a final example, we present the results of a VSB system that is operated over a channel with "parabolic-like" delay (exponent = 2.73) and an s/n of 30 dB. The received and demodulated signal is sampled with different timing phases spaced $T/4$ apart and equalized in a cyclic equalizer with $N = 15$ taps. The distortion values resulting after equalization during only one sequence (i.e., 15 symbols) are summarized in Table I. For comparison, the initial channel distortion D_{Channel} and the minimum distortion $D_{\min}$ that can be achieved with an equalizer of this length are also included. It can be seen that initial training using cyclic equalization achieves a performance that is already close to optimum.

Some comments should be made about the simulation results we have presented. They indicate that initial training with cyclic equalization may only be necessary for a very short time; in some cases, for only one sequence period. This means that the received signal is not

Table I — Distortion for VSB channel and 15-tap equalizer

Timing	D_{Channel}	$D_{\text{Cycl.}}$	$D_{\min}$
0	2.04	0.15	0.06
25%	1.87	0.52	0.21
50%	2.25	0.99	0.98
75%	2.88	0.25	0.12

really periodic and that no spectral lines in the strict sense will occur at equally spaced frequencies ω_m, as we specified earlier in this section. The spectrum will be continuous, showing increasingly concentrated peaks at those frequencies with larger numbers of sequence repetitions. We have not found this to be a disadvantage; in fact, under some circumstances, the tap settings achieved with only a small number of iterations were, for the transmission of random data, preferable to the steady-state solution.

We have shown by example that fast reliable initial convergences can be achieved using an ideal reference signal without spending any time to synchronize the reference. Final fine adjustment of the taps is accomplished in a decision-directed mode using a longer sequence or random data. In the next sections, we will analyze the behavior of the cyclic equalizer during its initial training period. The convergence behavior with the mean-square algorithm with averaging, the choice of the training sequence, and the effect of the initial value of the taps will be considered. Then the exact behavior of the mean-square algorithm without averaging will be analyzed and conditions for convergence will be given.

IV. STEADY-STATE SOLUTION FOR THE TAPS

As was discussed in the previous section, the operation of the cyclic equalizer does not depend upon the synchronization of the reference, and we will not stress the rotation property of the taps unless necessary.

We assume a system with N equalizer taps and let the samples of the received signal be the components of the vector

$$\mathbf{x}^T = (\gamma_N, \gamma_{N-1}, \cdots, \gamma_1). \tag{16}$$

If we neglect the noise components, the tap-signal vector is periodic and successive vectors are cyclic shifts of each other ($\gamma_{N+m} = \gamma_m$). We define a signal matrix

$$S = \begin{bmatrix} \gamma_N & \gamma_{N-1} & \gamma_{N-2} & \cdots & \gamma_1 \\ \gamma_1 & \gamma_N & \gamma_{N-1} & \cdots & \gamma_2 \\ \gamma_2 & \gamma_1 & \gamma_N & \cdots & \gamma_3 \\ \vdots & \vdots & \vdots & & \vdots \\ \gamma_{N-1} & \gamma_{N-2} & \gamma_{N-3} & \cdots & \gamma_N \end{bmatrix}, \tag{17}$$

whose rows consist of all N succeeding sample vectors. The elements of S are given by

$$s_{ik} = \gamma_{(i-k) \text{Mod } N}. \tag{18}$$

At the equalizer output, a sequence of values $\mathbf{x}^T\mathbf{c}$ ($\mathbf{c}$ is the tap-weight vector) appears as the input vector $\mathbf{x}$ is cyclically shifted through its

N states. The resulting output sequence is

$$\mathbf{y} = S\mathbf{c}. \tag{19}$$

Obviously, it is possible to obtain from a given input sequence any arbitrary desired output sequence by a suitable choice of $\mathbf{c}$, provided only that S^{-1} exists. If we define a data vector ξ which contains the reference values associated with $\mathbf{y}$, it is possible to select $\mathbf{c}$ so that

$$\mathbf{y} = \xi = S\mathbf{c}, \tag{20}$$

i.e., the recovered sequence can be perfectly equalized (at least, at the sample points) and there is no residual error. This is even true with nonlinear distortion. Since the error can be reduced to zero, we conclude that the same tap vector

$$\mathbf{c}_0 = S^{-1}\xi \tag{21}$$

is obtained with any equalizer in the steady state, regardless of the particular tap-updating algorithm (as long as it is unbiased).

We now proceed to determine the eigenvalues of the circulant matrix S. Let us first define a set of values r so that

$$r^N = 1 \rightarrow r_k = \exp\left(j\frac{2\pi k}{N}\right). \tag{22}$$

In the next step we form

$$\begin{aligned}
\lambda &= \gamma_N + r\gamma_{N-1} + r^2\gamma_{N-2} + \cdots + r^{N-1}\gamma_1 \\
r\lambda &= \gamma_1 + r\gamma_N + r^2\gamma_{N-1} + \cdots + r^{N-1}\gamma_2 \\
&\vdots \\
r^{N-1}\lambda &= \gamma_{N-1} + r\gamma_{N-2} + r^2\gamma_{N-3} + \cdots + r^{N-1}\gamma_N.
\end{aligned}$$

This may be written in matrix form as

$$\mathbf{r}_k\lambda_k = S\mathbf{r}_k, \tag{23}$$

where we have defined the vector

$$\mathbf{r}_k = \{r_{kn}\}; \quad \text{with } r_{kn} = \frac{1}{\sqrt{N}}\exp\left(j\frac{2\pi}{N}nk\right). \tag{24}$$

The $\mathbf{r}_k$'s are obviously eigenvectors of S. The eigenvalues λ_k are

$$\lambda_k = \mathbf{x}^T\mathbf{r}_k, \quad 0 \leqq k \leqq N-1, \tag{25}$$

and are given by the discrete Fourier transform (DFT) of the input vector $\mathbf{x}$. The signal matrix S can be diagonalized if we introduce a matrix W with

$$\{W\}_{ik} = \frac{1}{\sqrt{N}}\exp\left(j\frac{2\pi}{N}ik\right), \tag{26}$$

whose columns are made up from the vectors $\mathbf{r}_k$. W is also symmetric and unitary; the properties

$$W = W^T, \quad W^* = W^\dagger, \quad W^\dagger W = I \tag{27}$$

are easily established. We may now alternatively either express the eigenvalues as components of the diagonal matrix D

$$D = W^\dagger S W \tag{28}$$

or as components of a vector

$$\boldsymbol{\lambda} = W\mathbf{x}, \tag{29}$$

since multiplication with W transforms a vector into its DFT.

We now give an interpretation in the frequency domain. The received (periodic) sequence can be expanded into a Fourier series

$$x(t) = \frac{1}{\sqrt{N}} \sum_m \chi_m \exp\left(j \frac{2\pi}{NT} mt\right), \tag{30}$$

where the coefficients χ_m correspond to the spectral lines and the range of m is determined by the bandwidth. The components γ_i in $\mathbf{x}$ are given by $x(t = \tau + iT)$; this may be combined with (25), and we obtain for the eigenvalues the frequency domain representation

$$\lambda_k = \sum_l \chi_{k+lN} \exp\left[j \frac{2\pi\tau}{T}\left(l + \frac{k}{N}\right)\right]. \tag{31}$$

In the case where all spectral lines are contained within twice the Nyquist frequency and $\tau = 0$, we have

$$\left.\begin{aligned} \lambda_0 &= \chi_{-N} + \chi_0 + \chi_N \\ \lambda_1 &= \chi_1 + \chi_{1-N} \\ \lambda_2 &= \chi_2 + \chi_{2-N} \\ &\vdots \\ \lambda_{N-2} &= \chi_{N-2} + \chi_{-2} \\ \lambda_{N-1} &= \chi_{N-1} + \chi_{-1} \end{aligned}\right\}. \tag{32}$$

As this represents 100-percent excess bandwidth, we may assume that most practical systems are within this range. If spectral lines are only within the Nyquist limit, (32) is simplified to

$$\left.\begin{aligned} \lambda_k &= \chi_k \quad \text{if} \quad |k| < \frac{N}{2} \\ \lambda_k &= \chi_{N-k} \quad \text{if} \quad |k| \geqq \frac{N}{2} \end{aligned}\right\}. \tag{33}$$

We can now give some comments as to the nature of the resulting tap vector $\mathbf{c}_0$. By combining (21) and (28), $\mathbf{c}_0$ may be expressed as

$$\mathbf{c}_0 = W^\dagger D^{-1} W \xi. \tag{34}$$

Here the term $W\xi$ is the DFT of the ideal samples and establishes a set of reference values at equally spaced points in the frequency domain (discrete Nyquist equivalence). The multiplication with D^{-1} determines the gain of an ideal correction function at these points. The resulting tap-vector $\mathbf{c}_0$ is the inverse DFT of this correction function. The overall transfer characteristic (channel and equalizer) is discrete Nyquist equivalent when $\mathbf{c}_0 = S^{-1}\xi$, i.e., frequency-domain equalization is precise at a set of equidistant points [spacing $(2\pi/NT)$]. This tap vector is, in the general case, not optimum for random data transmission after the training period. Basically, equalization is a mathematical approximation problem. The equalizer approximates the compensation function with a trigonometric polynomial. With a cyclic equalizer, the coefficients are selected to match the desired function at equidistant points. This will generally not give minimum mean-square error at the output, since only discrete frequency information is used and the channel behavior between the sample points is not taken into account. In a recent paper, Chang and Ho[40] briefly discussed this problem from a somewhat different point of view and concluded that the initial approximation $\mathbf{c}_0$ is generally close to the optimum settings for random data. We will not further discuss the approximation problem in this paper.

V. MEAN-SQUARE ALGORITHM WITH AVERAGING

In this section, we are looking at a tap-control system that minimizes the mean-square error between the equalizer output $y(nT)$ and reference symbols ξ_n. We use a steepest descent gradient algorithm of the form

$$\mathbf{c}_{m+1} = \mathbf{c}_m - \beta(A\mathbf{c}_m - \mathbf{v}), \tag{35}$$

where A is the signal-correlation matrix and $\mathbf{v}$ is the signal-correlation vector. The gradient

$$\mathbf{g}_m = E\{\mathbf{x}_i(y_i - \xi_i)\}\,|_{\mathbf{c}=\mathbf{c}_m} \tag{36}$$

is evaluated in the usual way by time averaging.

First we note that, in the noiseless case, because of the cyclic nature of $\mathbf{x}$, both A and $\mathbf{v}$ can be determined by time-averaging over one full sequence length of N symbols. Further, A and $\mathbf{v}$ are constant and well-defined throughout the process. It is easily verified that A can be

expressed in terms of the normal signal matrix S,

$$A = \frac{1}{N} SS^{\dagger} = \frac{1}{N} S^{\dagger}S, \tag{37}$$

and that $\mathbf{v}$ is equivalent to

$$\mathbf{v} = \frac{1}{N} S^{\dagger}\boldsymbol{\xi}. \tag{38}$$

The gradient is zero, and updating stops when

$$\mathbf{c} = \mathbf{c}_0 = A^{-1}\mathbf{v} = S^{-1}\boldsymbol{\xi}. \tag{39}$$

If we introduce the tap error vector $\boldsymbol{\delta}_m = \mathbf{c}_m - \mathbf{c}_o$, (35) takes the form

$$\boldsymbol{\delta}_{m+1} = (I - \beta A)^m \boldsymbol{\delta}_1. \tag{40}$$

The choice of β and the convergence depend on the eigenvalues of A. To guarantee that $\boldsymbol{\delta}_{m+1} \to 0$ for large m, we require $0 < \beta < 2/\mu_{\max}$, where μ_i are the eigenvalues of A. Since

$$S = WDW^{\dagger} \tag{41}$$

and therefore

$$A = \frac{1}{N} SS^{\dagger} = \frac{1}{N} WDD^{\dagger}W^{\dagger}, \tag{42}$$

the eigenvectors of A and S are common (and independent of $\mathbf{x}$). The eigenvalues μ_k of A are related to the eigenvalues λ_k of S by

$$\mu_k = \frac{1}{N} \lambda_k^* \lambda_k. \tag{43}$$

Another interpretation is obtained by realizing that the matrix A is circulant (like S) and symmetric with elements

$$\{A\}_{ik} = a_{i-k} = a_{k-i} = a_n = \frac{1}{N} \sum_{m=l}^{l+N-1} \gamma_m \gamma_{m-n}. \tag{44}$$

By analogy to (25), the eigenvalues are

$$\mu_k = \mathbf{a}^T \mathbf{r}_k; \quad 0 \leqq k \leqq N-1, \tag{45}$$

where $\mathbf{a}$ contains the components a_n. We see that the eigenvalues are given by the DFT of the cyclic autocorrelation values, a_n.

We now express these eigenvalues by the spectral lines in the frequency domain. If we combine (31) and (43), we obtain

$$\mu_k = \frac{1}{N} \sum_m \sum_n \chi_{k+mN} \chi^*_{k+nN} \exp\left[j \frac{2\pi\tau}{T} (m-n)\right] \tag{46}$$

or, in an equivalent form,

$$\mu_k = \frac{1}{N} \sum_m \sum_n \chi_{k-mN} \chi_{(n+m)N-k} \exp\left(j2\pi n \frac{\tau}{T}\right). \tag{47}$$

If the bandwidth is Nyquist-limited, only one single term in (47) contributes to the eigenvalues, namely,

$$\mu_k = \mu_{N-m} = \frac{1}{N} |\chi_k|^2, \qquad k \leqq \frac{N}{2}. \tag{48}$$

The eigenvalues are then independent of timing and carrier parameters and phase distortion of the channel. Only the signaling format, the channel attenuation, and the choice of the sequence ξ determine the eigenvalues. (Note that we have so far not restricted the choice of ξ to a particular class of sequences, such as maximum length sequences.)

In the case of excess bandwidth which is, however, limited to twice the Nyquist frequency (all reasonable pulse-amplitude modulation systems fall in this category), a few more terms in (47) need be considered, and

$$\mu_k = |\chi_k|^2 + |\chi_{N-k}|^2 + \chi_k \chi_{N-k} \exp\left(j\frac{2\pi\tau}{T}\right) + \chi_k^* \chi_{N-k}^* \exp\left(-j\frac{2\pi\tau}{T}\right), \tag{49}$$

which shows the influence of the timing phase τ. Note that only the third and fourth terms depend on phase and timing parameters. This term represents the fold-over around the Nyquist frequency.* The smaller the roll-off, the less the eigenvalues will be affected by this fold-over. In fact, it is even possible to have a small amount of excess bandwidth without any contribution of these terms. This is shown in Fig. 9. We distinguish two cases.

(*i*) N = odd. The Nyquist frequency is located midway between two spectral lines of the training sequence. Fold-over is avoided if we have a normalized roll-off $\alpha \leqq 1/N$.

(*ii*) N = even. The Nyquist frequency coincides with a spectral line of the training sequence. If we choose $\alpha \leqq 1/N$, the eigenvalues will still be phase-invariant, but one of them (for $k = N/2$) will now be dependent on the timing phase t_o.

Most voice-grade telephone channels have very large phase distortion, but only moderate amplitude distortion. Usually, the worst-case gain deviations over a given frequency range are known (e.g., on

* Some crossterms in (49) are zero if the bandwidth is less than twice the Nyquist frequency.

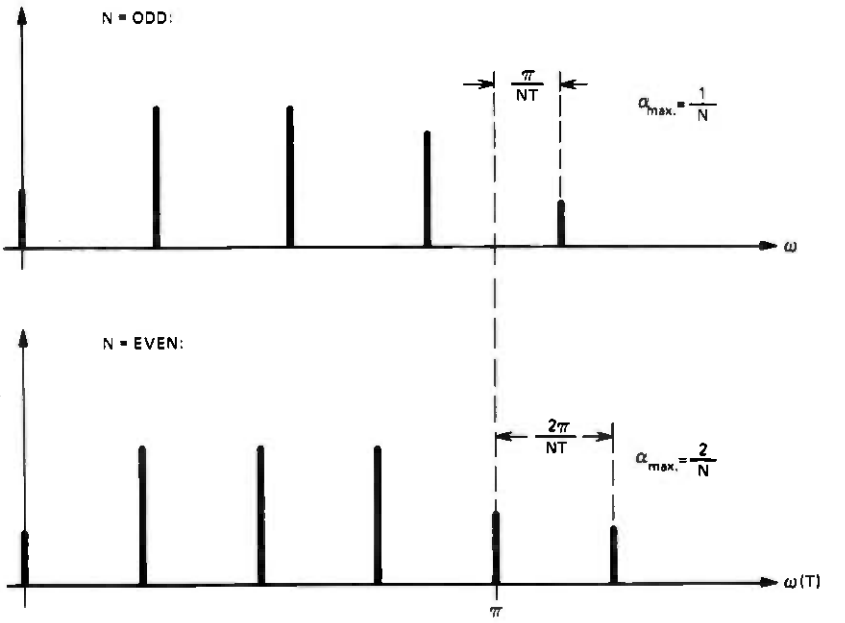


Fig. 9—Spectrum of training sequence.

private channels). If we deal with small excess bandwidth and we know our sequence ξ, it is obviously possible to calculate the spread of the eigenvalues from (46) or (49). We may then choose the value of β in (35) so that

$$0 < \beta < \frac{2}{\lambda_{max}} \tag{50}$$

to insure convergence. In addition, it may, of course, be necessary to normalize the signal power $\mathbf{x}^T\mathbf{x}$ with an automatic gain control to make the eigenvalues dependent only on the relative gain difference between the various frequencies, but independent of the average absolute gain.

VI. CHOICE OF THE TRAINING SEQUENCE

So far, we have not discussed the choice of the training sequence ξ. From the previous study we know that the eigenvalues of S and A are well-behaved as long as the DFT of $\mathbf{x}$, or of the sampled autocorrelation function, respectively, has no zero elements. This is obviously sufficient to guarantee the existence of inverses of S and A and therefore also the existence of a solution $\mathbf{c}_o$. Zero elements can be avoided by selecting a signaling format and a sequence ξ to insure nonzero line amplitudes at all frequencies $f_n = n/NT$ within the transmission bandwidth. If the channel does not have serious attenuation gaps, we have

also nonzero amplitudes at the receiver input. To obtain fast convergence, the eigenvalues should be as equal as possible (minimum spread). This can obviously best be achieved by selecting a sequence ξ which produces lines of equal amplitudes; predistortion for expected attenuation at the band edges is possible. The transmitter is then effectively sending a comb of equally spaced frequencies of approximately equal amplitudes that could obviously also be provided by a number of frequency generators, but are, of course, much more efficiently synthesized as spectral lines of a suitable sequence. Note that the samples of the training sequence ξ need not necessarily be binary; arbitrary numbers (and sequence lengths) can be stored in ROM's in both transmitter and receiver. We will discuss a few special cases for ξ, assuming small excess bandwidth ($<1/N$) and an odd number of taps and flat gain:

(*i*) Single pulse, $\xi^T = (0, \cdots, 0, 1, 0, \cdots, 0)$: This produces a frequency comb of equal amplitudes. We have further

$$A = \frac{1}{N} I, \qquad \lambda_k = \text{const} = 1/N, \qquad \beta_{\text{opt}} = N. \tag{51}$$

Convergence is obtained in a single iteration, independent of the initial settings. See eq. (40).

(*ii*) Single pulse, $\xi^T = (1, \cdots, 1, -1, 1, \cdots, 1)$: This produces a similar frequency comb, but with a much larger amplitude at dc. The eigenvalues are shown to be

$$\lambda_o = (N-2)^2/N; \quad \lambda_1 = \cdots = \lambda_{N-1} = 4/N. \tag{52}$$

(*iii*) Maximum-length pseudorandom sequence: Such sequences have lengths $N = 2^m - 1$ (among others), and were used for the simulations given earlier in Section II. The eigenvalues are

$$\lambda_o = \frac{1}{N}; \quad \lambda_1 = \cdots = \lambda_{N-1} = 1 + \frac{1}{N}. \tag{53}$$

For a given symbol magnitude of the ξ_i's and a given peak power, the maximum-length sequence gives the largest spectral line energy and seems thus to be a good choice, especially for noisy channels. Both in (*ii*) and in (*iii*), λ_o is different from the other $N-1$ identical eigenvalues; we will, however, show that β can be selected according to these $(N-1)$ values and that λ_o does not affect the convergence if the equalizer is properly preset.

VII. PRESETTING THE TAPS

Since the reference sequence is not synchronized with the received signal, the resulting tap vector may have its main tap in any position,

and it would obviously not make sense to preset in the traditional way of having $c_i = \delta_{io}$. Instead, we choose an initial tap vector $s\mathbf{u}$ whose coefficients have equal values s. If we assume $N = 2M + 1$ taps, the initial equalizer transfer function is given by

$$H(\omega) = s \sum_{n=-M}^{M} e^{-j\omega nT} = s \frac{\sin (N\omega T/2)}{\sin (\omega T/2)}, \tag{54}$$

which is a comb filter with period $1/T$ and attenuation poles at $f = k/NT$, i.e., precisely at the frequencies where the spectral lines of the training sequence are located. Only dc information is thus transmitted to the output prior to the first iteration. This is also obvious from the fact that the output $y = s\mathbf{x}^T\mathbf{u}$ does not depend on the cyclic shift of $\mathbf{x}$, since it is the sum of all N sequence samples. If an ideal Nyquist pulse is applied to such a system, the initial distortion of the output signal is very large, i.e.,

$$D_{\text{peak}} = D_{\text{MSE}} = N - 1. \tag{55}$$

This is independent of s. If, however, we look at the average mean-square error of the training sequence, we have with the initial setting

$$\begin{aligned} \epsilon_1^2 &= \frac{1}{N} \sum_i (s\mathbf{x}_i^T\mathbf{u} - \xi_i)^2 \\ &= s^2(\mathbf{x}^T\mathbf{u})^2 - 2\frac{s}{N}(\mathbf{x}^T\mathbf{u})(\xi^T\mathbf{u}) + \frac{1}{N}(\xi^T\xi). \end{aligned} \tag{56}$$

This can be differentiated with respect to s, and we find that the initial mean-square error is minimized if we choose

$$s_{\text{opt}} = \frac{1}{N}\frac{\xi^T\mathbf{u}}{\mathbf{x}^T\mathbf{u}}. \tag{57}$$

The quotient associated with $1/N$ represents the dc gain of the channel and is usually close to unity. Since the dc gain of the equalizer is equal to the sum of the tap coefficients, (57) means that the initial settings should be chosen to have the same sum as the final settings in $\mathbf{c}_o$ (remember that $\mathbf{c}_o$ is the inverse DFT of the correction function).

Some further physical insight is obtained if the mean-square error after m iterations is studied. This mean-square error may be expressed as[36]

$$\epsilon_{m+1}^2 = \sum_{i=0}^{N-1} q_{i,m+1}, \tag{58}$$

where the ith error component, $q_{i,m+1}$, is given by

$$q_{i,m+1} = \mu_i |\delta_1^T \mathbf{r}_i|^2 (1 - \beta\lambda_i)^m. \tag{59}$$

The initial value of each component is proportional to its corresponding eigenvalue and to the square of

$$\delta_1^T \mathbf{r}_i = s\mathbf{u}^T \mathbf{r}_i - \mathbf{c}_o^T \mathbf{r}_i, \tag{60}$$

where

$$\mathbf{u}^T \mathbf{r}_i = \begin{cases} 0 & \text{if } i \neq 0 \\ \sqrt{N} & \text{if } i = 0. \end{cases} \tag{61}$$

The values of $\delta_1^T \mathbf{r}_i$ are then obtained as

$$\delta_1^T \mathbf{r}_i = -\mathbf{c}_o^T \mathbf{r}_i = \{-DFT(\mathbf{c}_o)\} \quad i \neq 0 \tag{62}$$

$$\delta_1^T \mathbf{r}_o = N^{-\frac{1}{2}}(Ns - \mathbf{c}_o^T \mathbf{u}) \quad \text{if} \quad i = 0. \tag{63}$$

Because of what we have said earlier, we see that these coefficients are proportional to the values of the correction function at the line frequencies, except in the case of $i = 0$, where $\delta_1^T \mathbf{r}_o$ is only proportional to the misadjustment at dc. If we select s according to (57), the error component associated with λ_o becomes zero. The constant β is then selected in accordance with the remaining eigenvalues to provide fast convergence.

A few comments are in order for the case of $\mu_o = 0$. This will occur whenever the sequence is dc-free. An example of this property is a maximum-length sequence that is complemented by one additional bit to provide an equal number of ones and zeros (N would then be even). From (59) we see that the error term associated with μ_o is zero; since there is no spectral line at dc, the gain at $\omega = 0$ is obviously immaterial as long as we transmit the training sequence, and convergence and mean-square error are independent of the choice of s. To see how this affects the solution $\mathbf{c}_o$, we write the relation $A\mathbf{c}_o = \mathbf{v}$ in the form

$$DD^\dagger W^\dagger \mathbf{c}_o = NW^\dagger \mathbf{v}. \tag{64}$$

Assume k eigenvalues in the diagonal matrix $DD^\dagger$ are zero.* Therefore, we have only $N - k$ linear independent equations for the N components of $\mathbf{c}_o$. The set of solutions for $\mathbf{c}_o$ can be expressed with k independent linear parameters. In the most important case where only $\mu_o = 0$, this ambiguity can be avoided easily by constraining the sum of the tap values. This sum remains constant throughout the equalization process. We can best show that if we look at the sum of the gradient components,

$$\sum_i \mathbf{u}^T \mathbf{x}_i (\mathbf{x}_i^T \mathbf{c} - \xi_i),$$

* This can happen if the test sequence has zero power at some frequencies $f_k = k/NT$ within the transmission band.

which is obviously zero since $\mathbf{u}^T\mathbf{x}_i$ is zero by definition. We would thus choose s in such a way as to match the desired dc gain (which is no longer immaterial if we transmit random data after the initial training period). This is still in accordance with (57) if the quotient of the right-hand side is replaced by the quotient of the spectral densities at dc when data are random.

VIII. INFLUENCE OF NOISE

So far, we have made the assumption that the received samples are noiseless. We give here a coarse analysis of the effects of noise which will show that its influence is, in fact, quite small and may often be neglected. We assume that the taps are calculated from a single-input signal vector which includes noise; that is, the vector $\mathbf{x}$ in (16) now consists of the received signal values plus noise samples. As the equalizer cannot make any distinction between signal and noise components, a tap vector,

$$\mathbf{c}_{or} = (S + R)^{-1}\xi, \tag{65}$$

will evolve instead of $\mathbf{c}_o = S^{-1}\xi$, where R is a noise matrix defined in accordance with (17). We then write the tap difference vector as

$$\mathbf{c}_o - \mathbf{c}_{or} = S^{-1}R\mathbf{c}_{or} \tag{66}$$

if we combine (21) and (65). If a noiseless test sequence were transmitted over the system, there would be some output error because the vector $\mathbf{c}_{or}$ is different from the optimum $\mathbf{c}_o$. The resulting mean-square error, averaged over the ensemble of $\mathbf{c}_{or}$'s, would be

$$\epsilon^2 = E\{(\mathbf{c}_o - \mathbf{c}_{or})^\dagger A(\mathbf{c}_o - \mathbf{c}_{or})\}, \tag{67}$$

and can be written as

$$\epsilon^2 = E\{\mathbf{c}_{or}^\dagger R^\dagger (S^{-1})^\dagger A S^{-1} R\mathbf{c}_{or}\}. \tag{68}$$

If we make use of the relation (37), this can be simplified to

$$\epsilon^2 = \frac{1}{N} E\{\mathbf{c}_{or}^\dagger R^\dagger R\mathbf{c}_{or}\}. \tag{69}$$

Assuming that succeeding noise samples are uncorrelated and that $|\mathbf{c}_{or}|^2 \approx 1$, we finally obtain for the mean-square error

$$\epsilon^2 = \sigma^2,$$

where σ^2 is the noise power. We conclude that for reasonable s/n's there will be only a small bias introduced because of the superimposed noise.

IX. CYCLIC EQUALIZATION USING A MEAN-SQUARE ALGORITHM WITHOUT AVERAGING

The previous discussion has given an analysis of the process of cyclic equalization using the mean-square algorithm with averaging. Much of the insight developed there regarding the final tap values, the type of training sequence to use, etc., carries over to the equalization process which uses the mean-square algorithm without averaging. However, to be more precise we now will carry out an exact analysis of this algorithm. Because it permits a more simple implementation, it is the algorithm without averaging that will most likely be used in practical situations.

Let the N-component tap-signal vector at time $t_o + kT$ be denoted by $\mathbf{x}_k$. In the absence of noise, succeeding signal vectors will then be related by

$$\mathbf{x}_{k+m} = U^m \mathbf{x}_k, \tag{70}$$

where U is an $N \times N$ cyclic shifting matrix of the form

$$U = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 1 & 0 \end{pmatrix}. \tag{71}$$

Note also that U is orthogonal and

$$U^m = U^{m \pm lN}. \tag{72}$$

The equalizer output at time $t_o + mT$ will be

$$y(t_o + mT) = \mathbf{c}_m^T \mathbf{x}_m = \mathbf{c}_m^T U^m \mathbf{x}_o, \tag{73}$$

where we have expressed the signal vector as a cyclic shift of one fixed state at start-up. We will drop the index on $\mathbf{x}$ from now on.

Let d_k be the reference value of the data signal at $t = t_o + kT$. The mean-square error at $t_o + mT$ is

$$\epsilon_m = E\{(\mathbf{c}_m^T U^m \mathbf{x} - d_m)^2\}, \tag{74}$$

where m indicates any of the equally probable cyclic shifts of $\mathbf{x}$ and d. The expected value in (74) can be obtained by time averaging over $i + 1 \leqq k \leqq i + N$ because of the cyclic nature of the signals under consideration. The gradient with respect to the tap weights is given by

$$\frac{\partial \epsilon}{\partial \mathbf{c}} = 2E\{e_m U^m \mathbf{x}\}. \tag{75}$$

In this section we make adjustments of $\mathbf{c}$ at each symbol interval and

use a nonaveraged approximation of (75) (the product of error and tap signals of the previous baud interval) for updating. Thus, our strategy becomes*

$$\begin{aligned} \mathbf{c}_{m+1} &= \mathbf{c}_m - \beta U^m \mathbf{x}(\mathbf{c}_m^T U^m \mathbf{x} - d_m) \\ &= U^m(I - \beta \mathbf{x}\mathbf{x}^T)U^{-m}\mathbf{c}_m + \beta U^m \mathbf{x} d_m. \end{aligned} \tag{76}$$

For convenience, we define a data vector ξ which contains the reference values d_m. We further define an N-dimensional vector,

$$\mathbf{r} = \{r_i\}, \qquad r_i = \delta_{ik}, \tag{77}$$

containing zeros in all positions except in one reference (kth) position ("center tap"). We observe that

$$d_m = \mathbf{r}^T U^m \xi, \tag{78}$$

and we can write (76) in the form

$$\mathbf{c}_{m+1} = U^m Z U^{-m}\mathbf{c}_m + \beta U^m E U^{-m}\mathbf{r}, \tag{79}$$

where we have introduced for convenience

$$Z = I - \beta \mathbf{x}\mathbf{x}^T \tag{80}$$

$$E = \mathbf{x}\xi^T. \tag{81}$$

By solving the time-varying difference equation (79), the tap vector after m adjustments can be expressed as

$$\mathbf{c}_{m+1} = U^m \left\{ Q^m \mathbf{c}_1 + \beta \sum_{k=0}^{m-1} Q^k E U^{k-m}\mathbf{r} \right\}. \tag{82}$$

The new matrix Q in (82) is defined as

$$Q = ZU^{-1} = (I - \beta \mathbf{x}\mathbf{x}^T)U^{-1}, \tag{83}$$

and will play an important role in further analysis. We can also easily verify the synchronization-invariant properties of (82). In fact, if we replace $\mathbf{x}$ by $U^i\mathbf{x}$ (introducing some arbitrary delay), we obtain

$$\mathbf{c}_{m+1} = U^{m+i} \left\{ Q^m U^{-i}\mathbf{c}_1 + \beta \sum_{k=0}^{m-1} Q^k E U^{k-m}\mathbf{r} \right\}, \tag{84}$$

but since we choose the initial $\mathbf{c}_1$ with equal bias values for all coefficients, $U^{-i}\mathbf{c}_1 = \mathbf{c}_1$, and (84) and (82) are identical except for an i-position cyclic shift of the resulting tap vector.

---

* We are assuming β is constant; in practice, it might be desirable to make β dependent on m.

X. SOLVING THE DIFFERENCE EQUATION

Before we discuss (82) in more detail, we define

$$H_m = \beta \sum_{k=0}^{m-1} Q^k E U^k \tag{85}$$

because sums of this type will be frequently needed in the subsequent analysis. Examples are

$$H_0 = 0$$
$$H_1 = \beta E$$
$$H_N = \beta \sum_{k=0}^{N-1} Q^k E U^k.$$

Further, it is rather straightforward to show that

$$H_{i+j} = H_i + Q^i H_j U^i = H_j + Q^j H_i U^j, \tag{86}$$

and, as a special case,

$$H_{lN+n} = H_{lN} + Q^{lN} H_n = H_n + Q^n H_{lN} U^n. \tag{87}$$

H_{lN} may be expressed in a more convenient form if we introduce a new summation index, $iN + j$,

$$H_{lN} = \beta \sum_{i=0}^{l-1} Q^{iN} \sum_{j=0}^{N-1} Q^j E U^j. \tag{88}$$

The first series can be summed, and we obtain

$$H_{lN} = (I - Q^{lN})(I - Q^N)^{-1} H_N, \tag{89}$$

where we have made the implicit assumption that $I - Q^N$ is nonsingular (we will say more about that in a moment).

We are now ready to study (82), which may be written as

$$\mathbf{c}_{m+1} = U^m \{Q^m \mathbf{c}_1 + H_m U^{-m} \mathbf{r}\}. \tag{90}$$

By setting $m = lN + n$ and combining the first expression in (87) with (89), we obtain

$$\mathbf{c}_{lN+n+1} = U^n \{Q^{lN+n} \mathbf{c}_1 + Q^{lN} H_n U^{-n} \mathbf{r} + (I - Q^{lN})(I - Q^N)^{-1} H_N U^{-n} \mathbf{r}\}. \tag{91}$$

For any nonzero value of β, $\mathbf{c}_m$ will not converge in the usual sense; however, it will reach a steady-state condition that does not depend upon the initial value of $\mathbf{c}_1$. In order that this can occur, we require

$$\lim_{l \to \infty} Q^{lN} = 0, \tag{92}$$

which means that we require the spectral radius $\rho(Q)$ to be less than unity (all eigenvalues inside the unit circle). This will also guarantee the nonsingularity of $I - Q^N$ and thus the existence of (89). The first (transient) term in (91) will then converge to zero, which means that the steady-state solution is independent of the initial tap settings.* The second term will also converge to zero and the steady-state value of $\mathbf{c}$ is

$$\mathbf{c}_{\infty,n} \triangleq U^n(I - Q^N)^{-1}H_N U^{-n}\mathbf{r}. \tag{93}$$

This solution is periodic in n; it is trivial to verify that, owing to the cyclic nature of U,

$$\mathbf{c}_{\infty,n+N} = \mathbf{c}_{\infty,n}. \tag{94}$$

As an important special case we have, if $n = 0$,

$$\mathbf{c}_\infty = (I - Q^N)^{-1}H_N\mathbf{r} \triangleq H_\infty\mathbf{r}. \tag{95}$$

After $m = lN$ iterations, the tap vector is

$$\mathbf{c}_{lN+1} = Q^{lN}\mathbf{c}_1 + (I - Q^{lN})(I - Q^N)^{-1}H_N\mathbf{r}. \tag{96}$$

By combining (95) and (96), we can express the convergence with the error vector $\mathbf{c}_{lN+1} - \mathbf{c}_\infty$,

$$\mathbf{c}_{lN+1} - \mathbf{c}_\infty = Q^{lN}(\mathbf{c}_1 - \mathbf{c}_\infty), \tag{97}$$

as a function of the initial error vector. This is a particularly simple form, which shows how the convergence is directly dependent on the eigenvalues of Q. The error vector is reduced by a factor Q^N with each cycle of iterations. The eigenvalues of Q are functions of β and of the signal format and channel characteristics. We will study this problem in the next section.

XI. CONDITIONS FOR CONVERGENCE

The eigenvalues λ and eigenvectors $\mathbf{z}$ of Q are determined by

$$Q\mathbf{z} = U^{-1}(I - \beta\mathbf{x}\mathbf{x}^T)\mathbf{z} = \lambda\mathbf{z}. \tag{98}$$

We can calculate $(\lambda\mathbf{z})^\dagger(\lambda\mathbf{z})$ and obtain

$$|\lambda|^2\mathbf{z}^\dagger\mathbf{z} = \mathbf{z}^\dagger\mathbf{z} - 2\beta\mathbf{z}^\dagger\mathbf{x}\mathbf{x}^\dagger\mathbf{z} + \beta^2\mathbf{z}^\dagger(\mathbf{x}\mathbf{x}^\dagger)^2\mathbf{z}. \tag{99}$$

Assuming normalization of the eigenvectors, we then require for stability that

$$|\lambda|^2 = 1 - 2\beta|\mathbf{z}^\dagger\mathbf{x}|^2 + \beta^2(\mathbf{x}^\dagger\mathbf{x})|\mathbf{z}^\dagger\mathbf{x}|^2 < 1. \tag{100}$$

---

* If only a small number of iterations are used for training, $\mathbf{c}_1$ should be chosen carefully, since $\mathbf{c}$ will then be a function of the transient term as well.

If we now assume that $\mathbf{z}^\dagger\mathbf{x} \neq 0$, we get the simple condition

$$0 < \beta < \frac{2}{\mathbf{x}^T\mathbf{x}} \tag{101}$$

to ensure convergence of the tap coefficients. Note that the bound (101) depends only on the received signal power, which can be normalized with an automatic gain control.

A completely different situation occurs if $\mathbf{x}$ is orthogonal to an eigenvector $\mathbf{z}$; $\mathbf{z}^\dagger\mathbf{x} = 0$. It is easy to see from (100) that this would imply $|\lambda| = 1$, regardless of β. This case must be avoided and needs some special attention.

We first conclude that $\mathbf{z}^\dagger\mathbf{x} = 0$ implies that $\mathbf{z}$ is an eigenvector of both Q and U; this is evident from (98). The next step is then to determine the eigenvectors $\mathbf{y}$ and eigenvalues μ of the cyclic shifting matrix U. They are defined by the equation

$$U\mathbf{y} = \mu\mathbf{y}. \tag{102}$$

We introduce a unitary matrix W with elements

$$w_{ik} = \frac{1}{\sqrt{N}} \exp\left(-j\frac{2\pi}{N}ik\right), \qquad 0 \leqq i, k \leqq N-1 \tag{103}$$

and observe that

$$\{W^\dagger UW\}_{ik} = \delta_{ik} \exp\left(-j\frac{2\pi i}{N}\right) \tag{104}$$

is diagonal. The eigenvalues of U are thus given by

$$\mu_i = \exp\left(-j\frac{2\pi i}{N}\right) \qquad i = 0, 1, \cdots, N-1 \tag{105}$$

and the corresponding eigenvectors are

$$\mathbf{y}_i^T = (w_{i0}, w_{i1}, w_{i2}, \cdots, w_{i,N-1}). \tag{106}$$

We now define a vector $\mathbf{h}$ with values $\mathbf{y}_i^T\mathbf{x}$,

$$\mathbf{h} = \begin{bmatrix} \mathbf{y}_0^T\mathbf{x} \\ \vdots \\ \mathbf{y}_{N-1}^T\mathbf{x} \end{bmatrix} = W\mathbf{x}. \tag{107}$$

The components of $\mathbf{h}$ are thus simply the components of the discrete Fourier transform of $\mathbf{x}$, and we can finally write our requirement $\mathbf{z}^\dagger\mathbf{x} \neq 0$ in the form

$$\sqrt{N}h_i = \sum_{k=0}^{N-1} x_k \exp\left(-j\frac{2\pi}{N}ik\right) \neq 0 \quad \text{for} \quad i = 0, \cdots, N-1. \tag{108}$$

This is not a serious restriction, since most channels will produce an $\mathbf{x}$ whose DFT will have only nonzero components. Difficulties can arise, however, with frequency gaps of severe attenuation within the passband range, but this is a condition that needs special attention with any equalizer. Partial-response signaling does not satisfy (108), and we conclude that it cannot be used for cyclic equalization without changes in the equalizer structure or tap-updating algorithm.

Before we leave the stability discussion, we would like to point out another aspect of our problem. By setting $n = 1$ and $lN \to \infty$ in (89), we obtain

$$H_\infty - QH_\infty U = \beta E,$$

or, after postmultiplying with U^{-1},

$$QH_\infty - H_\infty U^{-1} = -EU^{-1}. \tag{109}$$

Matrix equations of the above type play an important role in stability and control theory (Lyapunov), and it is known that a unique solution of (109) exists only if Q and U^{-1} have no common eigenvalues. This would obviously also lead to our conditions (101) and (108).

XII. ASYMPTOTIC BEHAVIOR

The coefficient vector that minimizes the mean-square error of the received sequence is given by

$$\mathbf{c}_o = A^{-1}\mathbf{v}, \tag{110}$$

where A and $\mathbf{v}$ are the signal-correlation matrix and the cross-correlation vector between this sequence and the reference. Our current strategy does not use the gradient (75) in a steepest descent algorithm, nor do we assume that β decreases as the iterations proceed. Thus, it is to be expected that we obtain settings that are biased with respect to (110). We first write (95) in the form

$$(I - Q^N)\mathbf{c}_\infty = H_N\mathbf{r}. \tag{111}$$

From (70) and (83) we can express Q^N as

$$\begin{aligned} Q^N &= (I - \beta\mathbf{x}_N\mathbf{x}_N^T) \cdots (I - \beta\mathbf{x}_2\mathbf{x}_2^T)(I - \beta\mathbf{x}_1\mathbf{x}_1^T) \\ &= I - \beta \sum_i \mathbf{x}_i\mathbf{x}_i^T + \beta^2 \sum_{i>k} \mathbf{x}_i\mathbf{x}_i^T\mathbf{x}_k\mathbf{x}_k^T - \cdots (-1)^N\beta^N x_N \cdots \mathbf{x}_1^T. \end{aligned} \tag{112}$$

If the signal matrix,

$$A = E\{\mathbf{x}_i\mathbf{x}_i^T\} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T, \tag{113}$$

is introduced, we have

$$I - Q^N = \beta NA\left\{I - \frac{\beta}{N}A^{-1}\sum_{i>k}\mathbf{x}_i\mathbf{x}_i^T\mathbf{x}_k\mathbf{x}_k^T + \cdots\right\}. \tag{114}$$

We can expand the right-hand side of (111) in a similar way,

$$H_N\mathbf{r} = \beta \sum_{i=1}^{N} \left\{ \prod_{j=i+1}^{N} (I - \beta \mathbf{x}_j \mathbf{x}_j^T) \right\} \mathbf{x}_i d_i$$

$$= \beta N \mathbf{v} - \beta^2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{x}_j \mathbf{x}_j^T \mathbf{x}_i d_i + \cdots, \tag{115}$$

where the signal correlation vector $\mathbf{v}$ is defined as

$$\mathbf{v} = E\{\mathbf{x}_i d_i\} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i d_i. \tag{116}$$

Combining (113) to (116) and writing out only the first-order terms in β/N yield

$$\mathbf{c}_\infty = \left\{ I + \frac{\beta}{N} A^{-1} \sum_{i>k} \mathbf{x}_i \mathbf{x}_i^T \mathbf{x}_k \mathbf{x}_k^T - \cdots \right\} \mathbf{c}_o$$

$$- \frac{\beta}{N} A^{-1} \sum_{i>k} \mathbf{x}_i \mathbf{x}_i^T \mathbf{x}_k d_k - \cdots. \tag{117}$$

The neglected terms in (117) are multiplied with higher powers of β. It is, therefore, always possible to choose β small enough to make the linear term dominant. We can conclude that the resulting asymptotic tap vector differs from the MMSE solution $\mathbf{c}_{opt}$ by a bias which, for sufficiently small β, is directly proportional to β and may be made arbitrarily small. Very fast initial convergence can be obtained by choosing β large; then β may be made smaller for the remaining iterations to reduce the bias error (gear shifting).* This will also reduce the periodic fluctuation of the tap coefficients in the final steady state.

On the other hand, one should always keep in mind that the cyclic process is used only during the training time and that random data are used later on for adaptation. The tap vector that yields MMSE for the training sequence generally does not minimize the mean-square distortion for random data. However, the work of Chang and Ho[40] indicates that (for small β) the results may not be significantly in error. It would be expected that, in the normal data set application, cyclic equalization would be used for enough cycles to achieve a good open eye; then a longer training sequence would be used, decision-directed, to determine the steady-state tap coefficients.

XIII. ACCELERATED PROCESSING

After these theoretical studies, we conclude our paper by discussing some more practical aspects of the signal-processing organization.

* It seems possible that a continuous decrease of β during the iterations would yield superior results; we have, however, not analyzed this case.

More precisely, we present a somewhat modified implementation technique of cyclic equalization that will allow a further reduction in the initial training time. For this, we assume that the received sequence is not substantially corrupted by noise. In a highly dispersive channel with a relatively high s/n, this is a realistic assumption since intersymbol interference is completely dominant and noise is of minor influence at the beginning of equalization. Once the initial transients have settled, the receiver will thus see a train of continuously repeated identical sequences. No information is lost if one sequence length is stored in the receiver for further processing and the input is switched off. Such a system is depicted in Fig. 10.

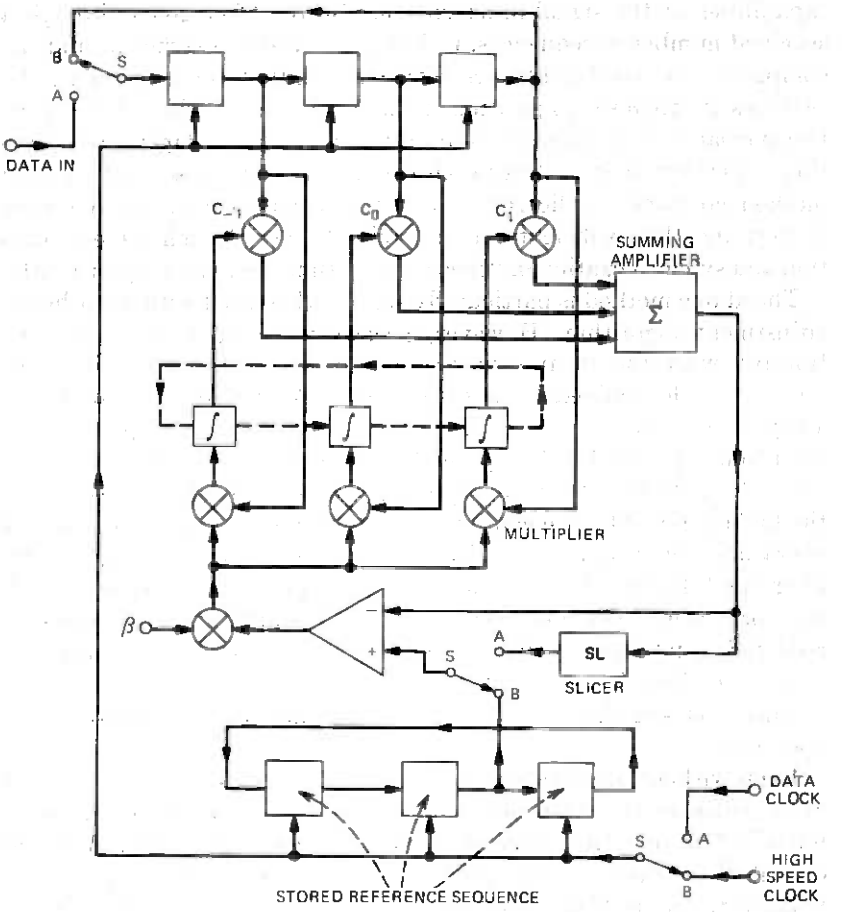


Fig. 10—Block diagram of cyclic equalizer equipped for accelerated processing.

For clarity, only three taps are shown. The samples of the data sequence are entered into the delay line (shift register in the case of a digital equalizer) of the transversal filter while switches S are in position A. As soon as one full sequence is stored, i.e., when samples have reached the end of the delay line, switches S are moved to position B. Thus, a shift-register ring circuit is formed and the stored samples can be shifted cyclically by applying appropriate clock pulses. The stored reference sequence is shifted at the same speed. It is important to realize that this speed need not be related to the actual data rate. The stored signal vector and the reference sequence can be shifted at a much higher rate, thus simulating a "speeded-up" data flow. Initial training can be achieved in a time interval limited only by the speed capabilities of the signal-processing hardware. After going through a specified number of sequences, training is considered sufficient and the computed tap coefficients are cyclically shifted for alignment. All switches are then set to position A, received data are shifted down the transversal filter at the actual data rate, and further adaptive equalization is performed on a decision-directed basis. The described training method combines cyclic rotation of the signal vector, the reference vector, and the coefficient vector to simultaneously achieve equalization and synchronization in "speeded-up" time, i.e., virtually instantly.

The above method is particularly simple when used with a stochastic adjustment algorithm. However, accelerated processing is also attractive with the mean-square gradient-type algorithm. Since the gradient is determined by averaging over N symbols, an additional array is necessary to store the accumulated correlation products of error and tap signals. The speeded-up data flow is again achieved by cyclically shifting either the signal vector or the coefficient vector at the highest possible rate consistent with the required signal-processing operations, only now the coefficient vector remains unchanged until, after one full cycle, the correlator array contains the (suitably scaled) tap corrections. The coefficients are now updated and the process is repeated, if necessary. After a couple of iterations the coefficients are rotated to align the largest of them with the reference position and the equalizer is switched to real-time processing and decision-directed operation.

Even without accelerated processing, the initial training time using cyclic equalization is so short that the delay needed for the signal to initially "fill up" the transversal filter becomes significant. With the described method of accelerated processing, the training time can be reduced to an arbitrary short interval limited only by the speed capabilities of the circuit elements. The "fill-up" time becomes completely

dominant. In the extreme, cyclic training can be achieved within a single symbol interval (after the equalizer is filled up).

XIV. CONCLUSION AND SUMMARY

Cyclic equalization, as presented in this paper, is a new method for initial equalizer training. Its main features are:

(*i*) A special training sequence where the number of symbols equals the number of equalizer taps.
(*ii*) Very fast start-up with provision for futher speeded-up operation, reducing training time theoretically to less than one symbol interval.
(*iii*) Ideal reference operation with no synchronization required. The processes of equalization and synchronization are combined in a unique way.
(*iv*) Perfect equalization at a set of equally spaced points in the frequency domain.
(*v*) Simple and economical implementation.

Cyclic equalization provides a set of tap coefficients that need to be cyclically rotated after initial training. At this time, a coarse equalization is achieved, the eye pattern is open, and the equalizer can switch to a decision-directed mode to achieve final tap settings using random data.

We have shown that the periodic training sequence can always be exactly equalized, so that all unbiased tap-updating algorithms will converge to the same tap settings, namely the inverse DFT of the sampled channel correction function. The mean-square gradient algorithm was analyzed in detail. The channel correlation matrix eigenvalues that influence the convergence are directly related to the lines of the power spectrum of the received sequence. The problem of initial coefficient presetting was discussed, and we made some comments concerning the choice of the training sequence and the influence of noise.

The cyclic equalization process using the mean-square algorithm without averaging was considered, and the difference equation that describes the coefficient convergence was solved. It was proved that the algorithm converges provided that the discrete Fourier transform of the received signal vector has no zero elements, and that the step size is within certain limits related to the number of taps and the received signal power. Finally, it was shown that the tap coefficients for the algorithm without averaging equal those for the algorithm with averaging except for an error term which goes to zero as the step size approaches zero.

The paper has been concluded by presenting a signal processing technique that achieves "accelerated convergence." This allows coefficient calculation in a time interval limited only by the speed capabilities of the equalizer circuitry.

REFERENCES

1. E. Kettel, "Ein automatischer Optimisator fuer den Abgleich des Impulsentzerrers in einer Datenuebertragung," Archiv der Elektrischen Uebertragung, *18*, No. 5 (May 1964), pp. 271–278.
2. R. W. Lucky, "Automatic Equalization for Digital Communication," B.S.T.J., *44*, No. 4 (April 1965), pp. 547–588.
3. R. W. Lucky, "Techniques for Adaptive Equalization of Digital Communication Systems," B.S.T.J., *45*, No. 2 (February 1966), pp. 225–286.
4. H. R. Rudin, "Automatic Equalization Using Transversal Filters," IEEE Spectrum, *4*, No. 1 (January 1967), pp. 53–59.
5. R. W. Lucky and H. R. Rudin, "An Automatic Equalizer for General Purpose Communication Channels," B.S.T.J., *46*, No. 9 (November 1967), pp. 2179–2208.
6. R. W. Lucky, J. Salz, and E. J. Weldon, "Principles of Data Communication," New York: McGraw-Hill, 1968.
7. M. J. DiToro, "Communication in Time-Frequency Spread Media Using Adaptive Equalization," Proc. IEEE, *56*, No. 10 (October 1968), pp. 1653–1679.
8. A. Gersho, "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," B.S.T.J., *48*, No. 1 (January 1969), pp. 55–70.
9. J. G. Proakis and J. H. Miller, "An Adaptive Receiver for Digital Signaling Through Channels with Intersymbol Interference," IEEE Trans. on Information Theory, *IT-15*, No. 4 (July 1969), pp. 484–497.
10. K. Moehrmann, "Implementation of an Adaptive Equalizer for Fast Data Transmission Involving a Minimum of Complexity," Nachrichtentechnische Zeitschrift, *23*, No. 1 (January 1970), pp. 36–42.
11. D. Hirsch and W. J. Wolf, "A Simple Adaptive Equalizer for Efficient Data Transmission," IEEE Trans. on Communication Technology, *COM-18*, No. 1 (February 1970), pp. 5–12.
12. P. Monsen, "Linear Equalization for Digital Transmission Over Noisy Dispersive Channels," Doctoral Thesis, School of Engineering and Applied Science of Columbia University, June 1970.
13. P. Mönsen, "Feedback Equalization for Fading Dispersive Channels," IEEE Trans. on Information Theory, *IT-17*, No. 1 (January 1971), pp. 56–64.
14. A. Lender, "Decision Directed Adaptive Equalization Technique for High Speed Data Transmission," IEEE Trans. on Communication Technology, *COM-18*, No. 5 (October 1970), pp. 625–632.
15. R. W. Chang, "Joint Optimization of Automatic Equalization and Carrier Acquisition for Digital Communication," B.S.T.J., *49*, No. 6 (July-August 1970), pp. 1069–1104.
16. R. W. Chang, "Joint Equalization, Carrier Acquisition, and Timing Recovery for Data Communication," Conference Record, International Conference on Communications, San Francisco, 1970.
17. B. Wendland, "Abtastsysteme zur adaptiven und nichtadaptiven Entzerrung von Kanaelen," Nachrichtentechnische Fachberichte, 37 (1969), pp. 335–352.
18. B. Wendland, "Zur Entzerrbarkeit von Datenkanaelen," Archiv der Elektrischen Uebertragung, *24*, No. 6 (June 1970), pp. 295–300.
19. G. K. McAuliffe, D. M. Montley, and R. A. Northrup, "Operation and Performance of ADEM," Nachrichtentechnische Fachberichte, *37*, 1969, pp. 366–378.
20. P. D. Daniell, "Adaptive Estimation with Mutually Correlated Training Sequences," IEEE Trans. on System Science and Cybernetics, *SSC-6*, No. 1 (January 1970), pp. 12–19.
21. L. D. Davisson, "Steady-State Error in Adaptive Mean Square Minimization," IEEE Trans. on Information Theory, *IT-16*, No. 4 (July 1970), pp. 382–385.
22. C. W. Niessen and D. K. Willim, "Adaptive Equalizer for Pulse Transmission," IEEE Trans. on Communication Technology, *COM-18*, No. 4 (August 1970), pp. 377–395.

23. L. D. Davisson, "Convergence Probability Bounds for Stochastic Approximation," IEEE Trans. on Information Theory, *IT-16*, No. 6 (November 1970), pp. 680–685.
24. R. T. Boyd and F. C. Monds, "Equalizer for Digital Communication," Electronics Letters, *7*, No. 2 (January 1971), pp. 58–60.
25. K. Moehrmann, "Einige Verfahren zur adaptiven Einstellung von Entzerrern für die schnelle Datenuebertragung," Nachrichtentechnische Zeitschrift, *24*, No. 1 (January 1971), pp. 18–24.
26. D. T. Magill, "Optimal Adaptive Estimation of Sampled Stochastic Processes," Technical Report No. 6302-3, December 1963, Stanford University.
27. B. Widrow, "Adaptive Filters I: Fundamentals," Technical Report No. 6764-6, December 1966, Stanford University.
28. B. Widrow, L. Griffiths, P. E. Mantey, and G. Burwell, "Adaptive Antenna Systems," Technical Report No. 6778-3, September 1967, Stanford University.
29. K. D. Senne, "Adaptive Linear Discrete-Time Estimation," Dissertation, Department of Electrical Engineering, Stanford University, June 1968.
30. K. D. Senne, "An Exact Solution to an Adaptive Linear Estimation Problem," Technical Report No. SLR-TR-70-0014, Frank J. Seiler Research Laboratory, U. S. Air Force Systems Command.
31. T. J. Schonfeld and M. Schwartz, "A Rapidly Converging First-Order Training Algorithm for an Adaptive Equalizer," IEEE Trans. on Information Theory, *IT-17*, No. 4 (July 1971), pp. 431–439.
32. T. J. Schonfeld and M. Schwartz, "Rapidly Converging Second Order Tracking Algorithms for Adaptive Equalization," IEEE Trans. on Information Theory, *IT-17*, No. 5 (September 1971), pp. 572–579.
33. D. R. George, R. R. Bowen, and J. R. Sorey, "An Adaptive Decision Feedback Equalizer," IEEE Trans. on Communication Technology, *COM-19*, No. 3 (June 1971), pp. 281–293.
34. P. Monsen, "Linear Estimation in an Unknown Quasi-stationary Environment," IEEE Trans. on Systems, Man, and Cybernetics, *SMC-1*, No. 3 (July 1971), pp. 216–222.
35. E. Y. Ho, "Optimum Equalization and the Effect of Timing and Carrier Phase on Synchronous Data Systems," B.S.T.J., *50*, No. 5 (May-June 1971), pp. 1671–1689.
36. R. W. Chang, "A New Equalizer Structure for Fast Start-up Digital Communication," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1969–2013.
37. G. Ungerboeck, "Theory on the Speed of Convergence in Adaptive Equalizers for Digital Communication," IBM Journal of Research and Development, *16*, No. 6 (November 1972), pp. 546–555.
38. G. Salomonsson, "An Equalizer with Feedback Filter," Ericsson Technics, *28*, No. 2, 1972, pp. 57–101.
39. R. T. Sha and D. T. Tang, "A New Class of Automatic Equalizers," IBM Journal of Research and Development, *16*, No. 6 (November 1972), pp. 556–566.
40. R. W. Chang and E. Y. Ho, "On Fast Start-up Equalization Using Maximum Length Pseudo-Random Training Sequences," B.S.T.J., *51*, No. 9 (November 1972), pp. 2013–2027.
41. R. D. Gitlin, E. Y. Ho, and J. E. Mazo, "Passband Equalization for Differentially Phase-Modulated Data Signals," B.S.T.J., *52*, No. 2 (February 1973), pp. 129–238.
42. K. H. Mueller and D. A. Spaulding, "Fast Start-up Systems for Transversal Equalizers," U. S. Patent No. 3715666, February 6, 1973.
43. K. H. Mueller and D. A. Spaulding, "Cyclic Equalization—A New Rapidly-Converging Equalization Technique for Synchronous Data Communication," Conference Records, ICC74 Minneapolis.
44. R. D. Gitlin, and J. E. Mazo, "Comparison of Some Cost Functions for Automatic Equalization," IEEE Trans. on Communications, *COM-21*, No. 3 (March 1973), pp. 233–237.
45. M. Karnaugh, "Automatic Equalizers Having Minimum Adjustment Time," IBM Journal of Research and Development, *17*, No. 2 (March 1973), pp. 176–179.
46. J. W. Mark and P. S. Budihardjo, "Joint Optimization of Receive Filter and Equalizer," IEEE Trans. on Communications, *COM-21*, No. 3 (March 1973), pp. 264–266.

47. J. W. Mark, "A Note on the Modified Kalman Filter for Channel Equalization," Proc. IEEE, *61*, No. 4 (April 1973), pp. 481–482.
48. R. D. Gitlin, J. E. Mazo, and M. G. Taylor, "On the Design of Gradient Algorithms for Digitally Implemented Adaptive Filters," IEEE Trans. on Circuit Theory, *CT-20*, No. 2 (March 1973), pp. 125–136.
49. S. U. H. Qureshi and E. E. Newhall, "An Adaptive Receiver for Data Transmission Over Time Dispersive Channels," IEEE Trans. on Information Theory, *IT-19*, No. 4 (July 1973), pp. 448–457.
50. J. W. Mark and P. S. Budihardjo, "Performance of Jointly Optimized Prefilter–Equalizer Receivers," IEEE Trans. on Communications, *COM-21*, No. 8 (August 1973), pp. 941–945.
51. J. Salz, "Optimum Mean-Square Decision Feedback Equalization," B.S.T.J., *52*, No. 8 (October 1973), pp. 1341–1373.
52. B. R. Saltzberg, "Intersymbol Interference Error Bounds with Application of Ideal Bandlimited Signaling," IEEE Trans. Information Theory, *IT-14*, No. 4 (July 1968), pp. 563–568.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Applications of Group Theory to Connecting Networks

By V. E. BENEŠ

(Manuscript received May 30, 1974)

Group theory impinges on the combinatorial study of connecting networks in a natural way: the stages, frames, and cross-connect fields from which many existing networks are built provide simple permutations out of which desired, complex assignments are built by composition. Some of the consequences of this interpretation are explored in this paper. In the group-theoretic setting, the action and role of the stages and fields become transparent, and many questions and results regarding networks can be regarded as problems about cosets, subgroups, factorizations, etc. This approach is particularly useful for the study of rearrangeable networks made of stages of square switches; such a network is rearrangeable if and only if the symmetric group of appropriate degree can be factored into products of certain subgroups associated with the network. Or again, the original Slepian-Duguid rearrangeability theorem corresponds to factoring a symmetric group into a product of double cosets of subgroups generated by stages.

I. INTRODUCTION AND SUMMARY

Many connecting networks for telephone switching are constructed of several stages of independently acting rectangular or square switches with suitable cross-connect fields between the stages to allow for "grosser" transitions. The permutations or maps of inlets into outlets that can be realized by the network are obtained in a sequence of steps each corresponding to passage through a stage or a cross-connect field. To put it in mathematical terms, the stages and cross-connect fields provide simple maps out of which desired ones can be built by successive *compositions*. This fact allows one to use the concepts of group theory to study questions about connecting networks. Such a study was initiated in a previous paper,[1] in which we remarked that it always seemed to be easier to obtain results about groups by the few available methods known for networks than vice versa. We are happy to report that this tendency has been in part reversed.[2]

In Sections II and III we describe how the actions of a stage of switching and that of a cross-connect field are to be interpreted in terms of *permutations*. Section IV contains a definition of a general notion of a "stage," and it is explained there how the cascading of successive stages of switching separated by cross-connect fields corresponds to *composition* or multiplication of permutations. In Section V, we indicate how some concepts from group theory can be used to describe the permutations achievable by successive stages of square switches with fields between them. As a result, we can in Section VI pose some questions about (permutation) groups that are relevant to the practical matter of what calls can be carried in a network made of stages of square switches.

Next, Section VII contains a study of how stages can give rise to subgroups, culminating in the result that a stage $\mathcal{S}$ is made of square switches iff $\mathcal{S}$ generates a group and $\mathcal{S}$ is complete in this sense: every crosspoint of $\mathcal{S}$ is used in some permutation that $\mathcal{S}$ can generate. The factoring of groups into products of complexes or subgroups is taken up in Section VIII. This important phenomenon first arose in a group-theoretic interpretation[1] of the rearrangeability theorem of Slepian and Duguid, and has since appeared in other studies[2] of this basic network property. Some half-dozen theorems on factoring a group into a product of complexes or subgroups are given. The special case of factoring by double cosets, exemplified by the Slepian-Duguid results, is considered in Section IX, and it is shown that only the standard "frame" cross-connect field used in that result will give a rearrangeable network when identical square switches are used in a stage.

II. STAGES AND CROSS-CONNECT FIELDS

Two examples of connecting networks are shown in Figs. 1 and 2. They illustrate how networks can be built of stages of (usually square) switches joined by a *link pattern* or *cross-connect field.** These fields are responsible for the distributive characteristics of the network. They afford an inlet many ways of reaching other switches and so many outlets. The examples have the property that the number of inlets, the number of links in a cross-connect field, and the number of outlets are all the same number. We shall restrict our attention to networks with this property, built by alternating stages with link patterns.

* The word "field" in this usage is borrowed from the domain of switching engineering, and has no algebraic significance. Such a field is, of course, usually tantamount to a permutation.

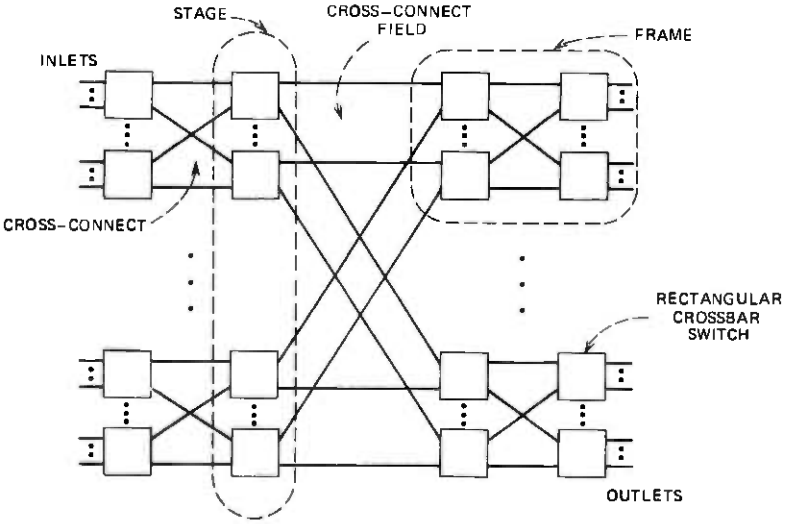


Fig. 1—Network showing stages, frames, and fields.

III. INTERPRETATION IN TERMS OF PERMUTATIONS

Suppose that the inlets and the outlets are both numbered in some arbitrary way from 1 to N. Then it is clear that each link-pattern, and each permitted way of closing N crosspoints in a stage, can be viewed abstractly as a permutation on $\{1, \cdots, N\}$. Here, "permitted," of course, means that no inlet to a stage is connected to more than one outlet, nor is any outlet connected to more than one inlet. Both the examples have the property that any maximal state, i.e., one in which no additional calls can be completed, has exactly N calls in progress; such a state realizes a permutation that is a product of the permutations represented by the link-patterns and the switch settings in the stages.

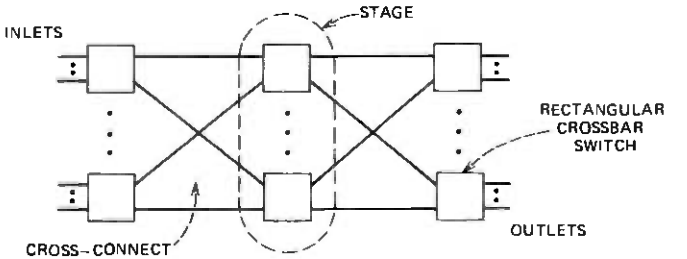


Fig. 2—Three-stage network.

IV. STAGES AND PRODUCTS

It will be convenient to adopt a generalized notion of a "stage" of a switching network. This generalization is based on the view that a stage is essentially a two-sided connecting network in which every call passes through one crosspoint. By a stage of switching we shall mean a connecting network constructed thus: with I the set of inlets and Ω the set of outlets, we choose an arbitrary subset $\mathcal{S}$ of $I \times \Omega$, and we place a crosspoint between all and only those inlets $u \in I$ and outlets $v \in \Omega$ such that $(u, v) \in \mathcal{S}$, and we speak of $\mathcal{S}$ itself as the stage. Logically, $\mathcal{S}$ is a relation indicating between what inlets and outlets there are crosspoints; it thus specifies the structure of the stage in the sense of Ref. 2. Thus,

Definition 1: *A stage is a subset of* $I \times \Omega$.

This terminology is an extension of the usual one, according to which, for example, a column of switches in Fig. 1 or 2 forms a stage, and the network consists of four or three stages joined by three or two cross-connect fields.

Definition 2: *A substage* $\mathcal{S}'$ *of a stage* $\mathcal{S}$ *is a subset of* $\mathcal{S}$.

In view of the discussion in Section III, we henceforth identify $I = \Omega = \{1, \cdots, N\}$.

Definition 3: *A stage* $\mathcal{S}$ *is made of square switches iff there is a partition* Π *of* $\{1, \cdots, N\}$ *such that*

$$\mathcal{S} = \bigcup_{A \in \Pi} (A \times A).$$

Evidently, $\mathcal{S}$ is made of square switches iff it is an equivalence relation. It is easily seen why all the stages illustrated in Figs. 1 and 2 are made of square switches as stated in Definition 1. Consider a stage $\mathcal{S}$ that has N inlets and N outlets. Evidently, such a stage provides ways of connecting some of the inlets simultaneously to some of the outlets. If the stage contains enough cross-points, it can be used to connect every inlet to an outlet, with no inlet connected to more than one outlet and vice versa; such a switch setting corresponds to a permutation on $\{1, \cdots, N\}$. This circumstance motivates the following definition.

Definition 4: *A stage* $\mathcal{S}$ *generates the permutation* π *on* $\{1, \cdots, N\}$ *if there is a setting of* N *crosspoints of* $\mathcal{S}$ *which connects* i *to* $\pi(i)$, $i = 1, \cdots, N$, *that is, if* $[i, \pi(i)] \in \mathcal{S}$ *for* $i = 1, \cdots, N$, *or most simply if* $\pi \subseteq \mathcal{S}$.

Definition 5: $P(\mathcal{S})$ *is the set of permutations generated by* $\mathcal{S}$.

Notice that $P(\mathcal{S})$ may be empty, and that for many π, $\mathcal{S}$ may generate* various submaps of π without generating π itself.

Definition 6: *A network with N inlets and N outlets generates a permutation π if there is a setting of the crosspoints of the network which connects, by mutually disjoint paths, each inlet to a unique outlet such that i is connected to $\pi(i)$, $i = 1, \cdots, N$.*

Multiplication of permutations is defined in the usual way by *composition*. Thus, if π_1 and π_2 are permutations, then $\pi_2\pi_1$ is the permutation defined by

$$\pi_2\pi_1(i) = \pi_2[\pi_1(i)] \quad i = 1, \cdots, N.$$

If two stages $\mathcal{S}_1$ and $\mathcal{S}_2$ are connected by a link pattern corresponding to a permutation π, then together they generate the permutations of the form

$$\varphi_2\pi\varphi_1 \quad \varphi_i \in P(\mathcal{S}_i), \quad i = 1 \text{ and } 2.$$

V. CONNECTION WITH GROUP THEORY

We adopt some concepts and notations from group theory to simplify the presentation. If G is a group, it is customary (although now a little old-fashioned) to speak of a subset $K \subseteq G$ as a *complex*. If $x \in G$, then xK denotes the set of products xy with $y \in K$, and $Kx = \{yx : y \in K\}$. Similarly, for complexes K_1 and K_2, the product K_1K_2 is the set of products yz with $y \in K_1$ and $z \in K_2$.

If a network consists of two stages $\mathcal{S}_1$ and $\mathcal{S}_2$ joined by a cross-connect field corresponding to a permutation π, then it generates exactly the permutations in the complex

$$P(\mathcal{S}_2)\pi P(\mathcal{S}_1).$$

Similarly, a network ν of s stages $\mathcal{S}_i$, $i = 1, \cdots, s$, with a cross-connect field π_i between $\mathcal{S}_i$ and $\mathcal{S}_{i+1}$, $i = 1, \cdots, s-1$, generates the complex

$$P(\nu) = P(\mathcal{S}_s)\pi_{s-1}\cdots P(\mathcal{S}_2)\pi_1 P(\mathcal{S}_1). \tag{1}$$

This complex completely describes the maximal assignments realizable by a network built of stages joined by link-patterns, all of N inlets and N outlets. To ask what simultaneous calls the network can carry is to ask what permutations of the full symmetric group S_N on $\{1, \cdots, N\}$ belong to the complex. This is a question of group theory that can in some cases be answered by its methods.[1,2] It is now possible to formulate a group-theory approach to the analysis and synthesis of

* To use an obvious extension of the terminology of Definition 4.

connecting networks made of stages, for the factors $P(\$_i)\pi_i$, $i = 1, \cdots, b - 1$ and $P(\$_s)$ occurring in (1) are themselves complexes.

VI. QUESTIONS

With this interpretation of the combinatorial power of a network in mind, we can immediately pose several problems of (permutation) group theory that shed light on the practical question, What calls can be carried in a network?

(*i*) What products of complexes are groups?
(*ii*) What groups can be generated by stages?
(*iii*) When can the whole symmetric group S_N be factored into a product of complexes corresponding to stages joined by cross-connect fields? In other words, When does such a product correspond to a *rearrangeable*[1] network?
(*iv*) What relationships and trade-offs exist between the stages and cross-connect fields chosen to build a network and the assignments it can realize?

Aspects of the first three questions will be taken up in the following sections; the fourth is discussed in Ref. 2.

VII. GENERATION OF GROUPS BY STAGES

In studies of rearrangeability of networks, questions have arisen as to (*i*) when the set $P(\$)$ of permutations generated by a stage forms a group and (*ii*) what groups can be got in this way. Only a partial answer has been given.[1] In cases of practical importance, such as those in Figs. 1 and 2, the stages are made of square switches. Clearly, such a stage is capable of effecting or generating only a special class of permutations: for each switch there are numbers m and n with $m < n$ such that the switch can perform all $(n - m + 1)!$ permutations of the numbers k in the range $m \leqq k \leqq n$ among themselves. Since no inlet or outlet is on more than one switch, the permutations generated by a stage form a *subgroup* of the symmetric group S_N of *all* permutations on $\{1, \cdots, N\}$. This subgroup is isomorphic to the direct product $\Pi_i S_{n_i}$, where n_i are the switch sizes; i.e., the subgroup has a property which might be described intuitively by saying that there exist sets on which the subgroup elements can "mix strongly," but which they keep separate. Group theory has some terminology for this situation, and we specialize it as follows.

A group G of permutations is called *imprimitive*[3] iff the objects acted on by the permutations of G can be partitioned into mutually disjoint sets, called the *sets of imprimitivity*, such that every $\pi \in G$

either permutes the elements of a set among themselves, or else carries that set onto another. That is, there is a partition Π of the set X acted upon such that $\pi \in G$ and $A \in \Pi$ imply $\pi(A) \in \Pi$. We shall specialize this terminology as follows:

Definition 7: *G is called strictly imprimitive iff it is imprimitive and the sets of imprimitivity are carried onto themselves by elements of G, i.e., iff there is a partition Π of X such that $A \in \Pi$ implies $\pi(A) = A$ for all $\pi \in G$, so that $\pi \in G$ are nonmixing on Π.*

Remark 1: Let M be a complex, i.e., a set of permutations. Define a stage $\mathbb{S}$ by

$$\mathbb{S} = \{(i, j) : \pi(i) = j \text{ for some } \pi \in M\}.$$

Then $P(\mathbb{S}) \supseteq M$, and no smaller stage has this property.

Remark 2: If $\mathbb{S}$ is made of square switches, then $P(\mathbb{S})$ is a strictly imprimitive group. For $(i, i) \in \mathbb{S}$ for all $1 \leqq i \leqq N$, so that the identity is in $P(\mathbb{S})$. $P(\mathbb{S})$ is closed under multiplication, so it is a group. Its sets of imprimitivity are exactly the sets A of the partition Π such that $\mathbb{S} = \bigcup_{A \in \Pi} A \times A$.

Remark 3: If H is a strictly imprimitive group of permutations on $\{1, \cdots, N\}$ with sets of imprimitivity forming the partition Π, and if $\mathbb{S}$ is the smallest stage with $P(\mathbb{S}) \supseteq H$ (Remark 1), then

$$\mathbb{S} = \bigcup_{A \in \Pi} A \times A,$$

so that $\mathbb{S}$ is made of square switches.

Thus, stages of square switches generate strictly imprimitive groups, and any such group can be generated by a stage made of square switches; there is a correspondence between strictly imprimitive groups and stages made of square switches. It has been shown previously[1] that the permutations generated by a stage include a subgroup *only if* the stage contains a substage made of square switches. This suggests that stages made of square switches arise naturally in switching, not just because designers thought of them, but because the mathematics demands it: to factor the symmetric group efficiently into a product of complexes some of which are subgroups, you must use subgroups that are generated by stages of square switches. We have seen that if $\mathbb{S}$ is made of square switches, then $P(\mathbb{S})$ is a strictly imprimitive group; we now show that (*i*) among the stages we would want to consider, those made of square switches are the only ones that generate groups, and (*ii*) only strictly imprimitive groups can be generated by stages.

Theorem 1: *If* $P(\mathbb{S})$ *contains a group* H, *then* H *is strictly imprimitive, and there is a substage* $\mathfrak{R} \subseteq \mathbb{S} \supseteq H = P(\mathfrak{R})$ *and* $\mathfrak{R}$ *is made of square switches*: $\mathfrak{R} = \bigcup_{A \in \Pi} A \times A$, *and the sets of imprimitivity of* H *are just the* $A \in \Pi$.

Proof: Define a relation $\mathfrak{R}$ on $\{1, \cdots, N\}$ by the condition that $i \mathfrak{R} j$ iff $j = \pi(i)$ for some $\pi \in H$. H must contain the identity, so $i \mathfrak{R} i$ holds for all $i \in \{1, \cdots, N\}$. Let i, j, k be numbers in $\{1, \cdots, N\}$, and φ, ψ permutations in H, such that $j = \varphi(i)$ and $k = \psi(j)$. Then $\psi\varphi \in H$ and $k = \psi\varphi(i)$, so that $i \mathfrak{R} j$; thus $\mathfrak{R}$ is transitive. Finally, if $j = \varphi(i)$ with $\varphi \in H$, we have $i = \varphi^{-1}(j)$ with $\varphi^{-1} \in H$, since H is a group, so $\mathfrak{R}$ is symmetric. Thus $\mathfrak{R}$ is an equivalence relation, and it is made of square switches. Obviously, $\mathfrak{R} \subseteq \mathbb{S}$ and $H = P(\mathfrak{R})$, and the result is proved.

To clarify the situation further, we introduce this property of stages:

Definition 7: $\mathbb{S}$ *is complete iff every crosspoint of* $\mathbb{S}$ *is used in generating some* $\pi \in P(\mathbb{S})$, *i.e., iff* $(i, j) \in \mathbb{S}$ *imply* $\exists \pi \in P(\mathbb{S}) \ni j = \pi(i)$.

The point of introducing this idea of completeness is two-fold: (*i*) it gives rise to clean theorems, and (*ii*) it is a reasonable requirement to impose on stages; for it means that every crosspoint can be used to realize some maximal assignment in the stage.

Theorem 2: $\mathbb{S}$ *is made of square switches iff* $\mathbb{S}$ *is complete and* $P(\mathbb{S})$ *is a group.*

Proof: If $\mathbb{S}$ is made of square switches, there is a partition Π of $\{1, \cdots, N\}$ such that $\mathbb{S} = \bigcup_{A \in \Pi} A \times A$, and $P(\mathbb{S})$ is clearly the largest strictly imprimitive subgroup of permutations whose sets of imprimitivity are just the $A \in \Pi$. From the form of $\mathbb{S}$ it follows that $(i, i) \in \mathbb{S}$ for $1 \leqq i \leqq N$, so that the identity permutation I belongs to $P(\mathbb{S})$ and $I(i) = i$. Thus, $\mathbb{S}$ is complete. Conversely, let $\mathbb{S}$ be complete and $P(\mathbb{S})$ be a group. We show that $\mathbb{S}$ is an equivalence relation. $P(\mathbb{S})$ must contain the identity, so $(i, i) \in \mathbb{S}$ for each $1 \leqq i \leqq N$, and $\mathbb{S}$ is reflexive. Also, if $(i, j) \in \mathbb{S}$ and $(j, k) \in \mathbb{S}$, then by the completeness there are permutations φ, $\psi \in P(\mathbb{S})$ such that $j = \varphi(i)$ and $k = \psi(j)$. Since $P(\mathbb{S})$ is a group, ψ, $\varphi \in P(\mathbb{S})$ with $\psi\varphi(i) = k$. Hence, $(i, k) \in \mathbb{S}$, and we have shown that $\mathbb{S}$ is transitive. Similarly, if $(i, j) \in \mathbb{S}$ there is a $\varphi \in P(\mathbb{S})$ with $j = \varphi(i)$, whence $i = \varphi^{-1} \in P(\mathbb{S})$, so that $(j, i) \in \mathbb{S}$, and $\mathbb{S}$ is symmetric. It is therefore an equivalence relation, and so is made of square switches.

Remark 4: If $\mathbb{S}$ is reflexive and symmetric, then $\mathbb{S}$ is complete. For given i and j, with $(i, j) \in \mathbb{S}$, we consider the permutation π which

interchanges i and j, leaving all else unchanged. Clearly, $\pi \in P(\mathcal{S})$ because $(j, i) \in \mathcal{S}$, and $(k, k) \in \mathcal{S}$ for all $1 \leqq k \leqq N$ so that $\pi \subseteq \mathcal{S}$.

Remark 5: The condition in Theorem 2 that $\mathcal{S}$ be complete can be replaced by symmetry of $\mathcal{S}$. For if $P(\mathcal{S})$ is a group then $I \in P(\mathcal{S})$, so $\mathcal{S}$ is reflexive. By Remark 4, $\mathcal{S}$ will then be complete.

Let G be a subgroup of S_N.

Theorem 3: $G = P(\mathcal{S})$ *for some stage* $\mathcal{S}$ *iff* G *is strictly imprimitive.*

Proof: $\Leftarrow$ is obvious by Remark 3. Let then $G = P(\mathcal{S})$ for a stage $\mathcal{S}$, and define $\mathcal{R}$ by

$$\mathcal{R} = \{(i, j) : j = \pi(i) \text{ for some } \pi \in G\}.$$

$\mathcal{R}$ is the smallest stage that will generate G. We note that $\mathcal{R}$ is complete, since $(i, j) \in \mathcal{R} \Rightarrow j = \pi(i)$ for some $\pi \in G \subseteq P(\mathcal{R})$. Also $\mathcal{R} \subseteq \mathcal{S}$, because $(i, j) \in \mathcal{R} \Rightarrow \exists \pi \in P(\mathcal{S}) \ni j = \pi(i)$, so that $(i, j) \in \mathcal{S}$. Thus, $P(\mathcal{R}) \subseteq P(\mathcal{S}) = G$, so $P(\mathcal{R})$ is a group. By Theorem 2, $\mathcal{R}$ is made of square switches. Hence, by Remark 2, $G[= P(\mathcal{R})]$ is strictly imprimitive.

VIII. FACTORING OF GROUPS

It is known[1] that the Clos three-stage rearrangeable network (Fig. 3) corresponds to a factorization

$$S_{nr} = G\varphi^{-1}H\varphi G, \tag{2}$$

where $G \sim (S_i)^r$ and $H \sim (S_r)^n$. (For "$\sim$" read "is isomorphic to.") There are similar factorizations of S_N, where N has $p > 2$ prime factors, into a product of $2p - 1$ subgroups associated with a re-

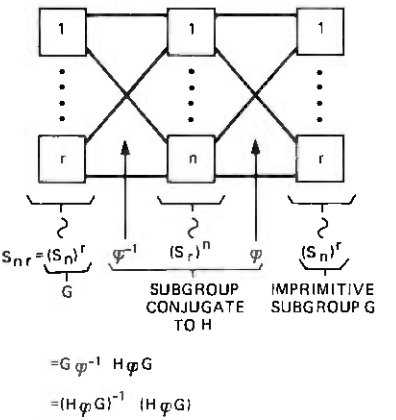


Fig. 3—How the three-stage network factors the group S_{nr}.

arrangeable two-sided network of $2p - 1$ stages symmetrically placed around a center stage. For this reason we shall look at conditions under which a group can be factored into a product of the form

$$H_1H_2\cdots H_k,$$

where some or all H_i may be subgroups.

For $k = 2$ and H_1, H_2 arbitrary complexes, this problem has been studied by S. Stein[4] under the additional condition that each element of H_1H_2 has a unique representation as a product in H_1H_2. Since, for the study of rearrangeability, this kind of uniqueness is of no interest, we shall not impose Stein's condition while we follow, initially, his original lines of reasoning. For A, B subsets of a group G, we let as usual

$$AB = \{xy: x \in A, y \in B\}.$$

Theorem 4: *Let G be a group, and A and B subsets of G. Then $G = AB$ iff for every $x \in G$*

$$A \cap xB^{-1} \neq \phi.$$

Proof: If $G = AB$, then, given $x \in G$, there exist $a \in A$ and $b \in B$ with $x = ab$, so that $a \in xB^{-1}$. Conversely, if $A \cap xB^{-1}$ is not empty, there exist $a \in A$ and $b \in B$ such that $a = xb^{-1}$, whence $x = ab \in AB$.

If B is a subgroup, the necessary and sufficient condition that $AB = G$ is that $A \cap xB \neq \phi$ for each $x \in G$; for then $B = B^{-1}$. This amounts to saying that A intersects every right coset of B, so by Lagrange's theorem

$$|A| \geqq \frac{|G|}{|B|}.$$

In fact, A can be got by choosing an element from each right coset of B. This is the "best" you can do, given B and no further structure. Of course, analogous results hold for left cosets if A is a group.

The factorization $S_{nr} = G\varphi^{-1}H\varphi G$ corresponding to the three-stage network prompts the question: If G and H are (sub) groups, when is GHG a group? The answer is given in the following:

Theorem 5: *Let G, H be groups. Then GHG and HGH are both groups iff they are identical*: $GHG = HGH$.

Proof: If GHG is a group, then $(GHG)^2 = GHG$, so that $GHGHG = GHG$. But $HGH \subseteq GHGHG$, so $HGH \subseteq GHG$. Now interchange G and H. Conversely, if $GHG = HGH$, then $(GHG)^2 = GHGHG = GGHGG = GHG$, so GHG is closed and is a group.

The same form of argument actually proves this apparently stronger result:

Theorem 6: *Let G be a group and H a complex. Then GHG is a group iff $HGH \subseteq GHG$.*

These results are extensions to three factors of the familiar fact that if H, G are (sub) groups, then HG is a group iff $GH \subseteq HG$.

Theorem 7: *Let F, G, H be groups. Then FGH is a group iff $GHFG \subseteq FGH$.*

Proof: If FGH is a group it is closed, so $FGHFGH \subseteq FGH$. But if $I \in F \cap H$, then $GHFG \subseteq (FGH)^2$, so that $GHFG \subseteq FGH$. Conversely, should $GHFG$ be contained in FGH, it would follow that $FGHFGH$ was also a subset of FGH. Thus, FGH would be closed and so a group.

A sufficient condition for Theorem 7 to hold is given in the next result, which then allows extension of the sufficiency part of Theorem 6 to three factors.

Lemma 1: *If G is a group, and F and H are complexes such that $HFG \subseteq GHF \subseteq FGH$, then $GHFG \subseteq FGH$.*

Proof: Left-multiplying the first inclusion by G gives at once that $GHFG \subseteq GGHF = GHF \subseteq FGH$.

Theorem 8: *Let F, G, and H be groups. If either $GHF \subseteq FGH \subseteq HFG$ or $HFG \subseteq FGH \subseteq GHF$, then FGH is a group.*

Proof: Assume the first horn of the dilemma. Then by Lemma 1,

$$(FGH)^2 = FGHFGH \subseteq HFG^2H \subseteq FGH,$$

so FGH is closed. Similarly, for the other horn, interchange the roles of GHF and HFG.

Along the same lines, one can give a sufficient condition for the product $G_1G_2 \cdots G_n$ of n groups to be a group:

Theorem 9: *If $G_1, G_2, \cdots, G_n$ are all subgroups of a given group, and if*

$$\prod_{i=1}^{n} G_{\pi(i)} = G$$

is the same set for every cyclic permutation $\pi \in S_n$, then $G_1G_2 \cdots G_n$ is a group.

$$\begin{aligned} \textit{Proof: } G_1 \cdots G_n G_1 \cdots G_n &= G_1(G_2 \cdots G_{n-1} G_n G_1) G_2 \cdots G_n \\ &= (G_1 G_2 \cdots G_n) G_2 \cdots G_n \\ &= (G_3 G_4 \cdots G_n G_1 G_2) G_2 \cdots G_n \\ &= (G_1 G_2 \cdots G_n) G_3 \cdots G_n \\ &\vdots \\ &= G_1 G_2 \cdots G_n. \end{aligned}$$

Thus, $\prod_{i=1}^{n} G_i$ is closed and so is a group.

It is apparent that the hypothesis of equality of all the sets obtained by cyclically permuting the order of the multiplication could be replaced by a continued inclusion as in Theorem 8.

IX. FACTORING BY DOUBLE COSETS: UNIQUENESS

If G and H are (respectively) the groups generated by a stage of r $n \times n$ switches and a stage of n $r \times r$ switches, and φ is the "canonical" frame cross-connect defined by

$$\varphi: j \to 1 + \left[\frac{j-1}{n}\right] + r((j-1) \bmod n), \quad j = 1, \cdots, nr, \quad (3)$$

then the Slepian-Duguid theorem affords factorizations

$$S_{nr} = G\varphi^{-1} H \varphi G = H \varphi G \varphi^{-1} H$$

corresponding to Clos rearrangeable networks. It is a natural pertinent question whether there are *other* cross-connect fields ψ that can be used instead of φ so as to have

$$S_{nr} = G\psi^{-1} H \psi G.$$

We shall show that any such ψ can differ from φ only in having its links mounted in different places on the switches. In particular, and this is the important property, it must give rise (as φ does) to the complete bipartite graph from n nodes to r nodes, when one lets switches be vertices and links be edges in a "frame" corresponding to $G\psi^{-1}H$ (Fig. 4).

First of all, we notice that the Slepian-Duguid theorem can be viewed as a factorization into a product of double cosets:

$$S_{nr} = (G\varphi^{-1}H)(H\varphi G) = (G\varphi^{-1}H)(G\varphi^{-1}H)^{-1}.$$

So we are really asking this question: For what double cosets $H\psi G$ is it true that $H\psi G$ times its inverse is the whole group S_{nr}? Note that such double cosets are of the form $P(\nu)$ for a conventional frame ν. We make the following convenient definition.

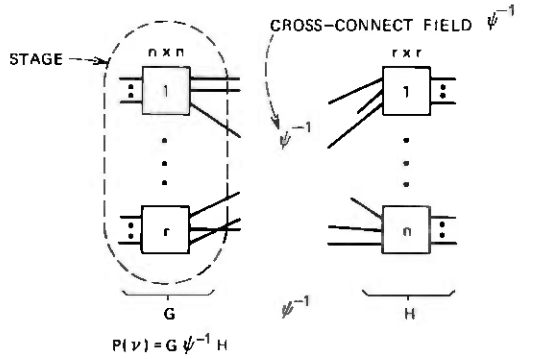


Fig. 4—How a frame generates a double coset.

Definition 8: *A permutation ψ works for G and H iff*

$$(H\psi G)^{-1}(H\psi G) = S_{nr}.$$

Remark 6: The order of occurrence of groups G and H in Definition 8 is of consequence. For it is readily seen from examples that if ψ works for G and H, it may not work for H and G. However, ψ does work for G and H iff ψ^{-1} works for H and G; for by Theorem 4, $(H\psi G)^{-1}(H\psi G) = S_{nr}$ iff $x \in S_{nr}$ implies

$$(H\psi G)^{-1} \cap x(H\psi G)^{-1} \neq 0,$$

i.e., iff $x \in S_{nr}$ implies $G\psi^{-1}H \cap xG\psi^{-1}H \neq \phi$. Thus, also, if $G = H$ $(n = r)$, then ψ works for H and H iff ψ^{-1} does too.

The next concept formalizes what is meant by saying that two cross-connect fields differ only in having their links mounted on different terminals of the same switches.

Definition 9: *Two permutations (cross-connects) ψ and ξ are equivalent with respect to G and H iff*

$$H\psi = \xi G.$$

This amounts to saying that the left ξ-coset of G is exactly the right ψ-coset of H. Intuitively, two cross-connects are equivalent if, when used between columns of switches corresponding to G and H, their links differ only in respect to the terminals on the switches where they attach but not in the switches themselves; thus, exactly the same pairs of switches have links between them, and the cross-connects are in a sense the same except for a renaming of terminals within switches.

What we shall show is that the special, canonical "frame" cross-connect eq. (3) is essentially the only one that works for G and H, in

the sense that any other that does is equivalent to it. Then it follows that there is exactly one double coset $H\psi G$ such that $(H\psi G)^{-1}(H\psi G) = S_{nr}$, namely $H\varphi G$. Thus, there is a unique factorization by double cosets of G and H associated with φ and the complete bipartite graph.

In the next result, G and H are as before the strictly imprimitive subgroups generated respectively by r $n \times n$ switches and n $r \times r$ switches.

Theorem 11: *ψ works for G and H iff ψ is equivalent to φ with respect to G and H, where φ is given by*

$$\varphi\colon j \to 1 + [(j-1)/n] + r((j-1) \bmod n), \quad j = 1, \cdots, nr.$$

Proof: If ψ is equivalent to φ with respect to G and H, then the cosets $H\psi G$ and $H\varphi G$ are identical, and so are $(H\psi G)^{-1}$ and $(H\varphi G)^{-1}$. Thus,

$$(H\psi G)^{-1}(H\psi G) = (H\varphi G)^{-1}(H\varphi G),$$

and, thus, ψ works for G and H. That much is fairly obvious. What is interesting is the converse: to prove that we use network arguments. Consider the network ν obtained by placing ψ between a stage of r $n \times n$ switches (giving G) and a stage of n $r \times r$ switches (giving H), followed by ψ^{-1} to another stage of r $n \times n$ switches. If ψ is not equivalent to φ with respect to G and H, then there are switches (in adjacent stages, left and middle, in fact) between which ψ^{-1} places no links, and some between which ψ places two or more links. Let L be a left (or inlet or first stage) switch and M a middle switch with no link between them. There is, then, a right (or outlet or third stage) switch R with one or more links to M. Since all outer switches are square and identical, no assignment taking the inlets of L onto the outlets of R is realizable, because, at most, $n - 1$ middle switches have links to both L and R. Thus, the network ν is not rearrangeable, and so some permutation in S_{nr} is missing from $(H\psi G)^{-1}(H\psi G)$. Hence, ψ does not work for G and H.

Remark 7: The argument above shows that if there are permutations ψ and ξ such that $(H\psi G)^{-1}(H\xi G) = S_{nr}$, then ψ and ξ are both equivalent to φ with respect to G and H. Thus, there is a unique factorization of S_{nr} into a product of double cosets of G and H.

REFERENCES

1. V. E. Beneš, "Permutation Groups, Complexes, and Rearrangeable Connecting Networks," B.S.T.J., *43*, No. 4 (July 1964), pp. 1619–1640.
2. V. E. Beneš, "Proving the Rearrangeability of Connecting Networks by Group Calculations," B.S.T.J., this issue pp. 421–434.
3. M. Hall, *The Theory of Groups*, New York: Macmillan, 1957, p. 64.
4. S. K. Stein, "Factoring by Subsets," Pacific J. of Math., *22*, No. 3 (September 1967), pp. 523–541.

The Bell System Technical Journal
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Proving the Rearrangeability of Connecting Networks by Group Calculations

By V. E. BENEŠ

(Manuscript received May 30, 1974)

Concepts and calculations from group theory have led to a new way of demonstrating rearrangeability of networks made of stages of square switches and to new factorizations of symmetric groups of composite degree.

I. INTRODUCTION AND BACKGROUND

Telephone connecting networks usually consist of stages of switching that alternate with fixed cross-connect fields; in effect, these two kinds of units are used to build up desired connection patterns out of simpler permutations by *composition* (see Fig. 1). Since permutations form a group under composition, the notions of group theory have become relevant to the study of connecting networks. They are particularly useful for looking at desired combinatorial properties such as rearrangeability, which is the capacity to realize any permutation. This is true because, in the group-theoretic setting, the original Slepian-Duguid rearrangeability theorem[1] provides the possibility of factoring a symmetric group into a product of subgroups, or of double cosets of subgroups generated by stages.

Here we extend a natural notion of "switch permutation" implicit in Duguid's proof to general networks with nr inlets and as many outlets. For such networks μ and ν, we establish a group-theoretic condition on the sets $D(\mu)$ and $D(\nu)$ of switch permutations realized by μ and ν, respectively, under which the larger network obtained by cascading μ and ν alternately between three stages of r $n \times n$ switches is rearrangeable. This result corresponds to factorization of the symmetric group of degree nr into a product of subgroups with the sets $P(\mu)$ and $P(\nu)$ of permutations realized by μ and ν, respectively. The condition given is verified in the examples in Section V by carrying out group multiplications.

It is conceptually useful to regard a connecting network as a quadruple $\nu = (G, I, \Omega, S)$, where G is a graph depicting structure and, in particular, indicates between which terminals (nodes) there is a switch

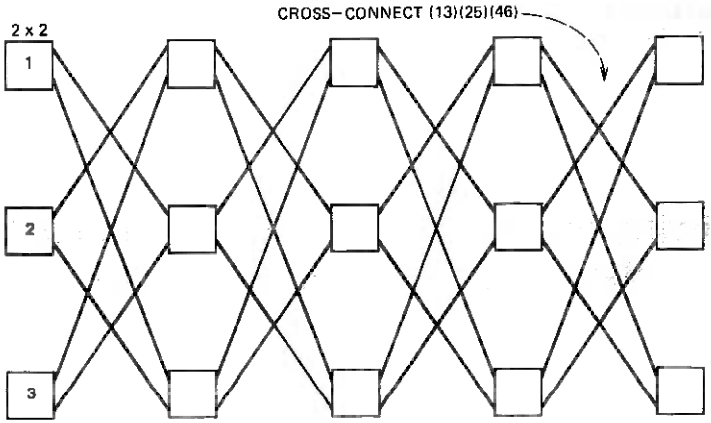


Fig. 1—Switching network with incomplete access between stages.

(edge); I and Ω are respectively the set of inlets (terminals) and the set of outlets, and S is the set of states deemed physically meaningful, that is, the set of allowed ways of closing switches so as to connect I to Ω by paths through G. We shall assume ν to be two-sided: $I \bigcap \Omega = \phi$ and $|I| = |\Omega| = nr$, where n and r are integers $\geqq 2$. The set A of assignments is the set of correspondences of subsets of I into Ω, each correspondence being interpreted as a particular way that terminals could ask to be connected together in pairs. Of course, there may or may not be a state in S realizing such a desired assignment. In any case, there is a natural map $\gamma: S \to A$ such that $\gamma(x)$ is the assignment realized by state x; in effect, $\gamma(x)$ tells us who is talking to whom when the network is in state x.

To put our questions into their natural group-theoretic setting, we shall identify both I and Ω with the integers $\{1, 2, \cdots, nr\}$, and the set of *maximal* assignments (everybody wanting to talk to somebody) with S_{nr}, where

$$S_k = \{k - \text{permutations}\} = \text{symmetric group of degree } k.$$

The set $P(\nu)$ of maximal assignments or permutations realized by ν is then expressible as

$$P(\nu) = \gamma(S) \bigcap S_{nr}.$$

A connecting network is called *rearrangeable* iff for every assignment $a \in A$ there is a state $x \in S$ such that x realizes a, i.e., $\gamma(x) = a$. Thus, the basic problem of the rearrangeability of ν can be cast in the following equivalent questions: When can every assignment be realized? When is $\gamma(S) = A$? Under our assumptions, these questions take the

form: When can the symmetric group on $\{1, \cdots, nr\}$ be realized? When is $P(\nu) = S_{nr}$?

The latter, group-theoretic form of the question begins to assume interest and importance when we note that many of the usual ways of constructing networks from stages of square switches correspond to *factoring* S_{nr} into factors that are subgroups. How this happens is explained next.

II. FACTORING S_{nr}

If X and Y are sets of group elements (complexes, in the old terminology) then XY is the set of products xy with $x \in X$ and $y \in Y$. We drop the notation I for the set of inlets, and use it henceforth for the identity permutation. Also, it is convenient to use exponent notation both for *products of complexes* with themselves, as X^2 for XX, and for the *direct product* of a group with itself some number of times. Thus, we establish the convention that if X is a complex, X^2 is XX as defined above; but if X is a group, then X^k means the k-fold direct product of X with itself.

It is readily seen, and has been pointed out before,[2] that a stage of square switches realizes an imprimitive subgroup of permutations. For example, the column of r $n \times n$ switches shown in the top half of Fig. 2 realizes the (imprimitive) subgroup that permutes the first n inlets among themselves, the second n among *themselves*, etc., up to the last n among each other. This subgroup is isomorphic to the direct product of S_n with itself r times, that is to $(S_n)^r$, and will be denoted by the same notation. In short, if ν is a stage of r $n \times n$ switches, then $P(\nu) = (S_n)^r$.

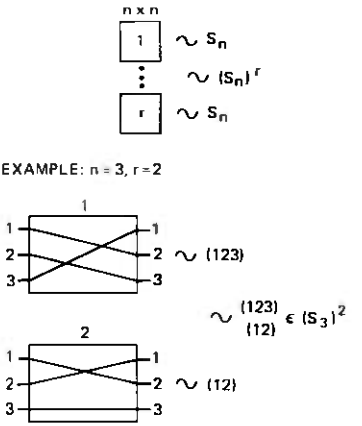


Fig. 2—Direct product group interpretation of a stage of square switches.

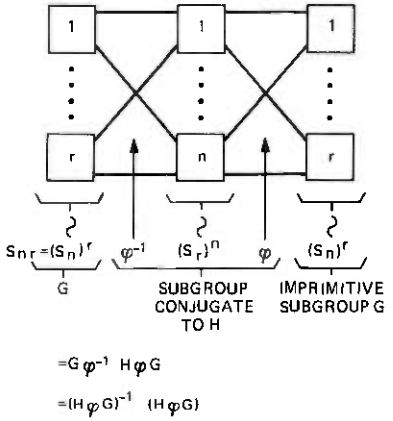


Fig. 3—Manner in which the three-stage network factors group S_{nr}, with $H=(S_r)^n$

Passing now to the three-stage network depicted in Fig. 3, we recall that by the classical result[3] of Slepian and Duguid, it is a rearrangeable network. We denote by φ the permutation corresponding to the standard cross-connect field between stages that defines a frame, namely,

$$\varphi\colon j \to 1 + [(j-1)/n] + r((j-1) \bmod n) \quad j = 1, \cdots, nr,$$

and we see that in Fig. 3 the middle and right stages have φ between them. (An alternative description of φ is that it takes the jth outlet on switch i into the ith outlet of switch j, for $j = 1, \cdots, n$ and $i = 1, \cdots, r$.) The original rearrangeability theorem can now be stated as a factorization, as follows (Fig. 3):

Classical Theorem (Slepian and Duguid): The symmetric three-stage network of square switches, in which switches on adjacent stages are connected by exactly one link, is rearrangeable and corresponds to a factorization

$$S_{nr} = (S_n)^r \varphi^{-1} (S_r)^n \varphi (S_n)^r. \tag{1}$$

The three middle factors above define a conjugate subgroup, so we have factored S_{nr} into a product of three subgroups. The remaining sections of this paper are devoted to finding alternative factorizations of S_{nr} that are associated with rearrangeable networks. We prove a factorization like (1) but with φ replaced by $P(\nu)$ for suitable ν, and then describe some applications.

III. SWITCH PERMUTATIONS

Now the essence of Duguid's proof of Slepian's result from Hall's theorem is contained in what we shall call a *switch-permutation*: he

decomposes any nr-permutation into a union of n submaps, each of which, because it corresponds basically to permuting outer switches, can be realized on a single middle switch. This idea is made precise as follows: define the function

$$\mathrm{sw}\colon \{1, \cdots, nr\} \to \{1, \cdots, r\}$$

by

sw_i = the switch (inlet or outlet) i is on in a stage of r $n \times n$ switches
= the $k (1 \leqq k \leqq r)$ such that $nk - n + 1 \leqq i \leqq nk$.

Let π be a permutation of S_{nr}. A *Hall decomposition* of π is a partition $\pi = \bigcup_{l=1}^{n} p_l$ of π into n submaps p_l such that for $l = 1, \cdots, n$, the set

$$q_l = \{(\mathrm{sw}_i, \mathrm{sw}_j)\colon (i, j) \in p_l\}$$

is an r-permutation, i.e., $q_l \in S_r$. The intuitive meaning of this property of the p_l is that each one maps exactly one inlet from each consecutive set of n onto outlets that are on distinct consecutive sets of n outlets. Hall's theorem on distinct representatives of subsets implies:

Fact: Every $\pi \in S_{nr}$ has a Hall decomposition.

We can now define the switch-permutations generated by a network ν as follows: an element

$$\begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix} \in (S_r)^n$$

is a switch permutation generated by ν iff there exists $\pi \in P(\nu)$ with a Hall decomposition $\pi = \bigcup_{l=1}^{n} p_l$ such that

$$q_l = \{(\mathrm{sw}_i, \mathrm{sw}_j)\colon (i, j) \in p_l\}. \tag{2}$$

Remark 1: If $\begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix}$ is a switch permutation generated by ν, then so is

$$\begin{pmatrix} q_{\pi(1)} \\ \vdots \\ q_{\pi(n)} \end{pmatrix}, \text{ for any } \pi \in S_n.$$

Intuitively, the q_l associated by (2) with the p_l of a Hall decomposition are just the settings of the successive middle switches that come out of Duguid's rearrangeability argument. The remark above is a reflection of the fact that submaps p_l of the decomposition can be assigned to the middle switches in an arbitrary way.

IV. FACTORIZATION

We let $D(\nu)$ be the set of switch permutations generated by ν. The new factorization-rearrangeability result we prove is as follows:

Theorem 1: *If μ and ν are networks with nr inlets and nr outlets, such that*

$$(S_r)^n \subseteq D(\nu)D(\mu),$$

then the network (Fig. 4) obtained by cascading ν and μ alternately between three stages of r $n \times n$ switches is rearrangeable, and corresponds to a factorization

$$S_{nr} = (S_n)^r P(\nu)(S_n)^r P(\mu)(S_n)^r.$$

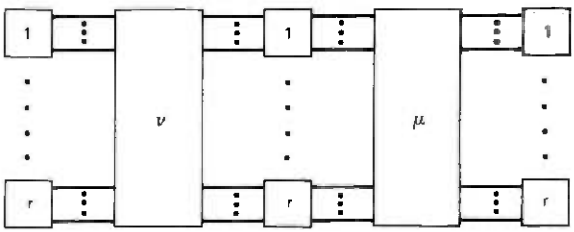
Proof: Take $\pi \in S_{nr}$ to be realized. It has a Hall decomposition $\pi = \bigcup_{l=1}^{n} p_l$ inducing a switch permutation

$$\begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix} \in (S_r)^n \subseteq D(\nu)D(\mu)$$

via $q_l = \{(\mathrm{sw}_i, \mathrm{sw}_j): (i, j) \in p_l\}$ as before. Thus, for each $l = 1, \cdots, n$ there exist a_l and b_l each in S_r such that $q_l = b_l a_l$, with

$$\alpha = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \in D(\mu) \text{ and } \beta = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in D(\nu).$$

The desired permutation can now be obtained by setting μ and ν to generate switch permutations α and β, respectively. For $(i, j) \in p_l$ we look at how sw_i and sw_j are connected to the middle stage and claim that they are connected to the same middle-stage switch! This is because μ connects sw_i to $a_l(\mathrm{sw}_i)$, and ν connects sw_j to $b_l^{-1}(\mathrm{sw}_j)$.



REARRANGEABLE NETWORK WHEN $(S_r)^n \subseteq D(\nu)\ D(\mu)$ YIELDING FACTORIZATION

$$S_{nr} = (S_n)^r\ P(\nu)\ (S_n)^r\ P(\mu)\ (S_n)^r$$

Fig. 4—Rearrangeable network when $(S_r)^n \subseteq D(\nu)\ D(\mu)$, yielding factorization $S_{nr} = (S_n)^r P(\nu)(S_n)^r P(\mu)(S_n)^r$.

Since, by construction,

$$\begin{aligned} \mathrm{sw}_j &= q_l(\mathrm{sw}_i) \\ &= b_l[a_l(\mathrm{sw}_i)], \end{aligned}$$

we have $a_l(\mathrm{sw}_i) = b_l^{-1}(\mathrm{sw}_j)$. It remains to route i to $a(\mathrm{sw}_i)$, to route j to $b_l^{-1}(\mathrm{sw}_j)$ and to complete the connection in the middle switch. This recipe works for all pairs $(i, j) \in \pi$, and the theorem is proved.

We next note that the hypothesis $(S_r)^n \subseteq D(\nu)\ D(\mu)$ of the theorem can be replaced by a stronger, more complicated condition that is less work to verify by calculation.

Remark 2: If M, N are subsets of $D(\mu)$, $D(\nu)$, respectively, such that for any $q_1, \cdots, q_n \in S_r$ there is some $\varphi \in S_n$ such that

$$\begin{pmatrix} q_{\varphi(1)} \\ \vdots \\ q_{\varphi(n)} \end{pmatrix} \in NM, \tag{3}$$

then $(S_r)^n \subseteq D(\nu)D(\mu)$. For if (3) holds, then there are a_l, b_l in M, N, respectively, and, hence, in $D(\nu)$ such that $q_{\varphi(l)} = b_1 a_l$, i.e., $q_l = b_{\varphi^{-1}(l)} a_{\varphi^{-1}(l)}$. But

$$\begin{pmatrix} a_{\varphi^{-1}(1)} \\ \vdots \\ a_{\varphi^{-1}(n)} \end{pmatrix} \in D(\mu), \quad \begin{pmatrix} b_{\varphi^{-1}(1)} \\ \vdots \\ b_{\varphi^{-1}(n)} \end{pmatrix} \in D(\nu)$$

by the remark following the definition of switch permutation. Hence,

$$\begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix} \in D(\nu)D(\mu).$$

V. EXAMPLES

Figure 5 and Tables I through III illustrate an application to the network of Fig. 1 to prove it rearrangeable. Here $\mu = \nu$, the network ν being just a stage of three 2×2 switches preceded and followed by the permutation (13) (25) (46) induced by the cross-connect field that links successive stages. Figure 5 illustrates two of the switch permutations generated by a copy of ν; the three stages shown in Fig. 5 are either the first three or the last three stages of the network of Fig. 1. Table I gives all eight possibilities; these form sets M, N (with $M = N$) of the form described in Remark 2, as can be verified from the product table, Table III, using the multiplication table for S_3 given in Table II. The entries of the product table that are shown form a subset C

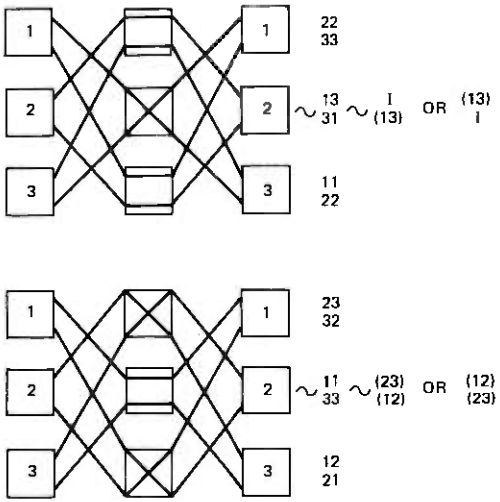


Fig. 5—Switch permutations generated by states of the middle stage.

of M such that either

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} b \\ a \end{pmatrix}$$

belongs to C for any choice of a and b in S_3; thus, property (3) of Remark 2 holds, and Theorem 1 is applicable.

Tables IV and V show the same kind of calculation for the network with a cyclic cross-connect field (Fig. 6) that induces the permutation (5432). Tables VI and VII show the same results for the network (Fig. 7) based on (23) (45). Asterisks in Table VII define a subset

Table I — Direct product elements corresponding to switch settings for cross-connect (13) (25) (46) used in Fig. 1

sw #								
1st	22	23	22	22	23	23	22	23
	33	32	33	33	32	32	33	32
2nd	11	11	13	11	13	11	13	13
	33	33	31	33	31	33	31	31
3rd	11	11	11	12	11	12	12	12
	22	22	22	21	22	21	21	21
Elements of K	I	(23)	I	(12)	(13)	(23)	(13)	(123)
	I	I	(13)	I	(23)	(12)	(12)	(132)
	1	2	3	4	5	6	7	8

Table II — Multiplication table for S_3

		1st Operator					
		I	(12)	(13)	(23)	(123)	(132)
2nd Operator	I	I	(12)	(13)	(23)	(123)	(132)
	(12)	(12)	I	(132)	(123)	(23)	(13)
	(13)	(13)	(123)	I	(132)	(12)	(23)
	(23)	(23)	(132)	(123)	I	(13)	(12)
	(123)	(123)	(13)	(23)	(12)	(132)	I
	(132)	(132)	(23)	(12)	(13)	I	(123)

with the property (3) of Remark 2, except that neither

$$\begin{matrix}(13)\\(132)\end{matrix} \quad \text{nor} \quad \begin{matrix}(132)\\(13)\end{matrix}$$

is in the subset; nevertheless $(S_3)^2 \subseteq D(\nu)^2$.

Table III — Partial table of M^2 for cross-connect corresponding to the permutation (13) (25) (46) and showing that condition of Remark 2 is satisfied

		M							
		1	2	3	4	5	6	7	8
M	1	I I							
	2	(23) I						(123) (12)	(13) (132)
	3	I (13)					(23) (123)	(13) (123)	
	4	(12) I				(132) (23)		(132) (12)	
	5	(13) (23)					(132) (132)	I (132)	(12) (12)
	6	(23) (12)				(123) (123)			(13) (13)
	7	(13) (12)				I (123)			
	8	(123) (132)						(23) (23)	

Table IV — Direct product elements corresponding to switch settings for cyclic cross-connect (5432)

sw #								
1st	11	21	11	21	11	21	21	11
	23	13	23	13	23	13	13	23
2nd	21	31	31	21	21	31	21	31
	32	22	22	32	32	22	32	22
3rd	12	32	12	32	32	12	12	32
	33	13	33	13	13	33	33	13
Elements of M	(12)	(132)	I	(132)	(23)	(13)	(12)	(13)
	(23)	(13)	(123)	(132)	(132)	(12)	(132)	(23)
	1	2	3	4	5	6	7	8

VI. CONJECTURE ABOUT NUMBER OF STAGES NEEDED TO GIVE REARRANGEABILITY WHEN A GIVEN CROSS-CONNECT FIELD IS USED

From Fig. 1 it is evident that an input switch on the left does not reach all the switches of the second stage, but can reach all the switches of the third stage by passing through the second stage. Thus, regarding switches as vertices and links as edges, we can say that no input switch is farther away from a third-stage switch than $d = 2$ units, in the usual metric of the graph defined by the vertices and edges. Furthermore, the number R of stages necessary and sufficient for rearrangeability is $5 = 2d + 1$. Similarly, in the three-stage network of Fig. 3, the distance from any input switch on the left to a middle

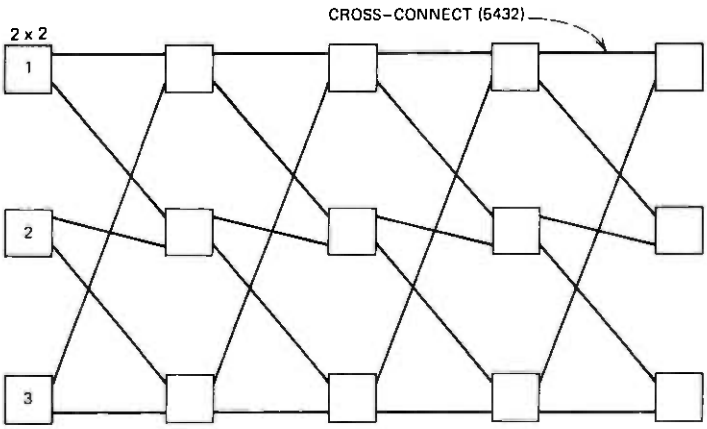


Fig. 6—Network based on cyclic cross-connect field corresponding to the permutation (5432).

Table V — Partial table of M^2 for cyclic cross-connect, showing that condition of Remark 2 is satisfied

M \ M	1	2	3	4	5	6	7	8
1	I I	(13) (123)				(132) (132)		
2	(23) (132)						(23) (23)	
3	(12) (12)		I (132)					
4	(23) (13)	(123) (12)	(132) (123)					
5	(132) (13)		(23) I					
6	(123) (123)							
7	I (13)	(13) (12)	(12) I					
8	(123) I	(23) (123)	(13) (13)	(23) (12)	(132) (12)			

switch is, of course, $d = 1$, and the number of stages R (necessary and sufficient for rearrangeability) is $3 = 2d + 1$. This leads us to suspect that there is a connection between the number of links one must go through to reach all switches of a stage and the number of stages needed to get a rearrangeable network.

To pose the question another way, let S be a stage of square switches, and φ a cross-connect field (permutation), and consider the natural sequence of networks such that

$$\begin{aligned} P(\nu_2) &= S\varphi S \\ P(\nu_3) &= S\varphi S\varphi S \\ P(\nu_4) &= S\varphi S\varphi S\varphi S \\ &\vdots \end{aligned}$$

We ask for what value $s = R$ will ν_s first be rearrangeable, and how does this number R depend on φ?

Going back now to the graph defined by the switches as vertices and the links as edges, we shall say that an inlet switch or vertex has access to a switch in a given stage iff there is a path on the graph from

Table VI — Direct product elements corresponding to switch settings for cross-connect represented by (23) (45) (see Fig. 7)

sw #								
1st	11	12	11	11	12	11	12	12
	22	21	22	22	21	22	21	21
2nd	11	11	13	11	13	13	11	13
	33	33	31	33	31	31	33	31
3rd	22	22	22	23	23	23	23	22
	33	33	33	32	32	32	32	33
Elements of M	I	I	I	I	(123)	(23)	(12)	(12)
	I	(12)	(13)	(23)	(132)	(13)	(23)	(13)
	1	2	3	4	5	6	7	8

Table VII — Complete table of M^2 for cross-connect corresponding to the permutation (23) (45)

		M							
		1	2	3	4	5	6	7	8
M	1	I*	I*	I*	I*	(123)*	(23)*	(12)*	(12)*
		I	(12)	(13)	(23)	(132)	(13)	(23)	(13)
	2	I	I	I*	I*	(123)*	(23)*	(12)*	(12)*
		(12)	I	(132)	(123)	(13)	(132)	(123)	(132)
	3	I	I	I	I	(123)*	(23)	(12)*	(12)
		(13)	(123)	I	(132)	(23)	I	(132)	I
	4	I	I	I	I	(123)	(23)	(12)	(12)
		(23)	(132)	(123)	I	(12)	(123)	I	(123)
	5	(123)	(123)	(123)	(123)	(132)	(12)*	(13)*	(13)
		(132)	(23)	(12)	(13)	(123)	(12)	(13)	(12)
	6	(23)	(23)	(23)	(23)	(13)	I	(132)*	(132)
		(13)	(123)	I	(132)	(23)	I	(132)	I
	7	(12)	(12)	(12)	(12)	(23)	(123)*	I	I
		(23)	(132)	(123)	I	(12)	(123)	I	(123)
	8	(12)	(12)	(12)	(12)	(23)*	(123)	I	I
		(13)	(123)	I	(132)	(23)	I	(132)	I

Note: Neither $\begin{matrix}(13)\\(132)\end{matrix}$ nor $\begin{matrix}(132)\\(13)\end{matrix}$ occurs, so that condition of Remark 2 fails, although condition of Theorem 1 holds because

$$\begin{matrix}\mathrm{I}\\(13)\end{matrix} \cdot \begin{matrix}(13)\\(23)\end{matrix} = \begin{matrix}(13)\\(132)\end{matrix}$$

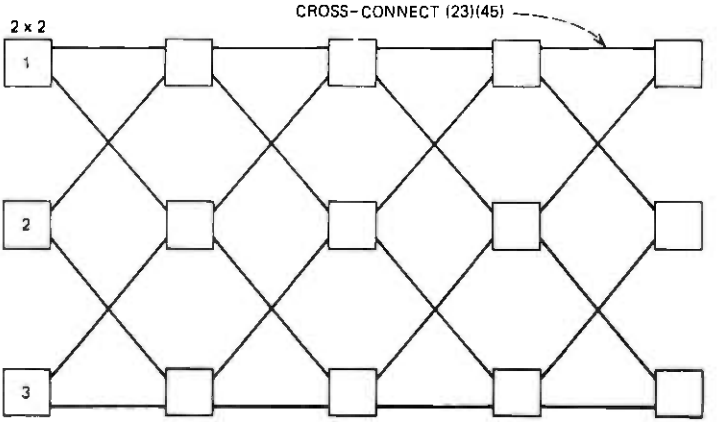


Fig. 7—Network based on cross-connect field corresponding to the permutation (23) (45).

the first switch to the second containing exactly one switch from every intermediate stage. Typically, the set of outlet switches to which an inlet switch has access will grow with the number of stages, and, for rearrangeability, it is, of course, necessary that every inlet switch have access to every outlet switch. Roughly speaking, the more access the field φ provides for switches from one stage to those of the next, the smaller will be the number of stages required for rearrangeability. It would therefore be of interest to relate this "amount of access" available with a given number of stages to the number of stages required for rearrangeability.

To this end, let us say that ν_s has "full access" if every inlet switch has access to every outlet switch, and define

$$d = \min \{s: \nu_{s+1} \text{ has full access}\}$$
$$R = \min \{s: \nu_s \text{ is rearrangeable}\}.$$

To return to the examples, if φ is the permutation (13) (25) (46) corresponding to the cross-connect field of Fig. 1, and S is a stage of three 2×2 switches, then $d = 2$ and $R = 5 = 2d + 1$. In Fig. 3, φ consists of a link between every pair of switches in successive stages, and so $d = 1$ and clearly $R = 3 = 2d + 1$. Again, in Fig. 6, using the cyclic cross-connect corresponding to (2345), it can be seen that $d = 2$ and $R = 5 = 2d + 1$.

All of these cases induce the following conjectures:

(*i*) ν_{2d+1} is rearrangeable.
(*ii*) $R = 2d + 1$.

It is easy to find additional confirming examples, especially necessity arguments for $R \geqq 2d + 1$, but to give a general proof seems to be very difficult.

REFERENCES

1. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, New York: Academic Press, 1965, p. 86.
2. Ref. 1, p. 100.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Some Effects of Measurement Errors on Rain Depolarization Experiments

By D. C. COX

(Manuscript received February 19, 1974)

Measurements of rain-produced depolarization for linearly polarized 20- and 30-GHz satellite signals oriented vertically (V) and horizontally (H) and a few degrees either side of V and H are required to determine both the proper polarization orientation for future systems and the depolarization at that orientation. The polarization received from area-coverage satellite beacons will vary considerably for different measuring sites within the coverage area. Calculation of rain-produced depolarization of one pair of orthogonally polarized signals from measurements of depolarization, differential attenuation, and differential phase of a different pair of orthogonally polarized signals will be required. Some effects of measurement errors on these calculations are shown. Accuracies on the order of ±0.5 dB in differential attenuation and ±2 degrees in differential phase are required in the measurements. Cross-polarization isolation of 25 dB in the measuring system is inadequate.

I. INTRODUCTION

It is desirable to use two orthogonal polarizations to double the number of radio channels available in future 20- and 30-GHz satellite repeaters. Since rain-produced depolarization (cross-polarization coupling) is more severe for circular polarizations than for linear polarizations[1] oriented parallel to the raindrop axes (see Fig. 1), two properly oriented orthogonal (near vertical and horizontal) linear polarizations are the logical choice for such systems. The only measurements of rain-produced depolarization[1–4] above 10 GHz have been made on terrestrial propagation paths. Therefore, measurements of rain-produced depolarization for linearly polarized 20- and 30-GHz satellite signals oriented vertically (V) and horizontally (H) and a few degrees either side of V and H are required to determine both the proper polarization orientation for future systems and the rain-produced depolarization at that orientation.

There will be considerable variation, however, in the polarization orientation of linearly polarized signals received at different measuring sites within the coverage area of 20- and 30-GHz area-coverage satellite beacons.[5] For example, the polarization orientation will vary on the order of 40 degrees over the continental United States for a satellite in synchronous orbit. Thus, the orientation cannot be optimum at all receiving sites from the standpoint of collecting data for system design. Also, since the angular orientation of the raindrop axes with respect to V and H at a given site varies from storm to storm,[4] statistics of the polarization orientation that produces the minimum depolarization and of the minimum value of that depolarization are needed, as well as statistics of the depolarization for several fixed polarization orientations. These requirements imply the need for calculating depolarization at orientations other than the polarization transmitted by a satellite beacon to a given receiving site. In principle, this is readily done if signal attenuation, phase shift, and depolarization parameters are measured with sufficient accuracy. This paper shows some effects of measurement errors on the calculation of rain-produced depolarization for one pair of orthogonally polarized signals from measurements made on a different pair of orthogonally polarized signals that have a different angular orientation. This information is needed in determining the accuracy required for earth-based instrumentation used in depolarization-determining satellite-beacon propagation experiments.

II. PROBLEM DEFINITION

A propagation experiment for determining rain-produced depolarization can be defined as follows.

Transmit a linearly polarized signal, E_{1t}, through the rain. Measure the attenuation and phase of (*i*) E_{11}, the signal received with the same polarization as E_{1t}, and (*ii*) E_{12}, the depolarized signal received with polarization orthogonal to E_{1t}. Then transmit E_{2t}, the signal orthogonal to E_{1t}, and measure the attenuation and phase of (*iii*) E_{22}, the received signal with E_{2t} polarization, and (*iv*) E_{21}, the depolarized signal orthogonal to E_{2t}.

From these measurements, calculate (*v*) E_{y1}, the orthogonal depolarized signal that would be received through the same rain from E_{xt}, a linearly polarized signal transmitted at some orientation other than that of E_{1t} or E_{2t} and (*vi*) E_{x2}, the corresponding depolarized signal received from E_{yt}, a signal transmitted orthogonal to E_{xt}. Figure 1 illustrates the spatial orientation of the polarization vectors, E_1, E_2 and E_x, E_y, on a coordinate system referred to the axes of an elliptical raindrop.[6]

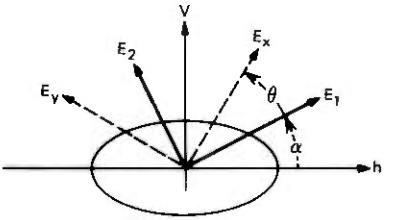


Fig. 1—Signal polarization vectors referred to elliptical raindrop axes.

A coefficient matrix for relating signals transmitted with polarizations 1 and 2, i.e., E_{1t} and E_{2t}, to signals received with the same polarization, E_{1r} and E_{2r}, can be defined as

$$\begin{bmatrix} E_{1r} \\ E_{2r} \end{bmatrix} = A_{12} \begin{bmatrix} 1 & b_{12} \\ c_{12} & d_{12} \end{bmatrix} \begin{bmatrix} E_{1t} \\ E_{2t} \end{bmatrix}, \tag{1}$$

where all coefficients and signals are complex and the phase reference for the transmission coefficients, b_{12}, c_{12}, and d_{12}, is the E_{1t} to E_{1r} coefficient.

The coefficients are obtained from the measurements described above as follows.

$A_{12} = \dfrac{E_{11}}{E_{1t}}$ is the absolute transmission coefficient (attenuation and phase) for the 1-oriented signal.

$b_{12} = \dfrac{E_{21}}{E_{2t}} \cdot \dfrac{1}{A_{12}} = \dfrac{E_{21}}{E_{11}} \cdot \dfrac{E_{1t}}{E_{2t}}$ is the depolarization coefficient for transmit 2 receive 1 normalized to A_{12}.

$c_{12} = \dfrac{E_{12}}{E_{1t}} \cdot \dfrac{1}{A_{12}} = \dfrac{E_{12}}{E_{11}}$ is the depolarization coefficient for transmit 1 receive 2 normalized to A_{12}.

$d_{12} = \dfrac{E_{22}}{E_{2t}} \cdot \dfrac{1}{A_{12}} = \dfrac{E_{22}}{E_{11}} \cdot \dfrac{E_{1t}}{E_{2t}}$ is the transmission coefficient for the 2-oriented signal normalized to A_{12}.

The coefficient E_{1t}/E_{2t} can be measured in clear air (or before launch) so, in principle, all coefficients can be determined. Note that the absolute transmission attenuation and phase appear only in A_{12} so the absolute phase, which is not measurable in the beacon experiment, is not involved in the other coefficients. (It is not necessary for computing the rotated polarization coefficients either, as will be shown.) The coefficient d_{12} contains the differential attenuation and phase between the two polarizations, 1 and 2.

Table I — Representative differential attenuation and phase for vertical-horizontal polarization

Total Attn. (dB)	$\lvert A_{12}\rvert$	Differential Amplitude Ratio	Differential Phase (degrees)
10	0.32	1.25	17
20	0.10	1.6	33.5

The corresponding matrix relation for signals transmitted and received with orthogonal polarizations x and y (see Fig. 1) is

$$\begin{bmatrix} E_{xr} \\ E_{yr} \end{bmatrix} = A_{xy} \begin{bmatrix} 1 & b_{xy} \\ c_{xy} & d_{xy} \end{bmatrix} \begin{bmatrix} E_{xt} \\ E_{yt} \end{bmatrix}. \tag{2}$$

The problem stated at the beginning of this section then reduces to

Given: A_{12}, b_{12}, c_{12}, d_{12}, and θ, the angle between polarizations 1, 2 and x, y,
Find: A_{xy}, b_{xy}, c_{xy}, d_{xy}.

This is easily done since, from Fig. 1,

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = [\theta_{12}] \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}$$

and (3)

$$\begin{bmatrix} E_1 \\ E_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} E_x \\ E_y \end{bmatrix} = [\theta_{xy}] \begin{bmatrix} E_x \\ E_y \end{bmatrix},$$

where, of course, $[\theta_{12}]^{-1} = [\theta_{xy}]$.

Table II — Rain-produced matrix coefficients determined from Table I and eqs. (6)

$\alpha = 0°$ = Vertical and Horizontal

Attn. (dB)	$\lvert A_{12}\rvert$	α deg.	d_{12}		b_{12}		c_{12}	
			Mag.	Ang.	Mag.	Ang.	Mag.	Ang.
10	0.32	0	1.25	17	0	0	0	0
10	0.32	45	1.0	0	0.1862	52.42	0.1862	52.42
10	0.32	20	1.188	13	0.1301	59.48	0.1301	59.48
20	0.1	0	1.6	33.5	0	0	0	0
20	0.1	45	1.0	0	0.3783	48.55	0.3783	48.55
20	0.1	20	1.448	25.38	0.2906	63.59	0.2906	63.59

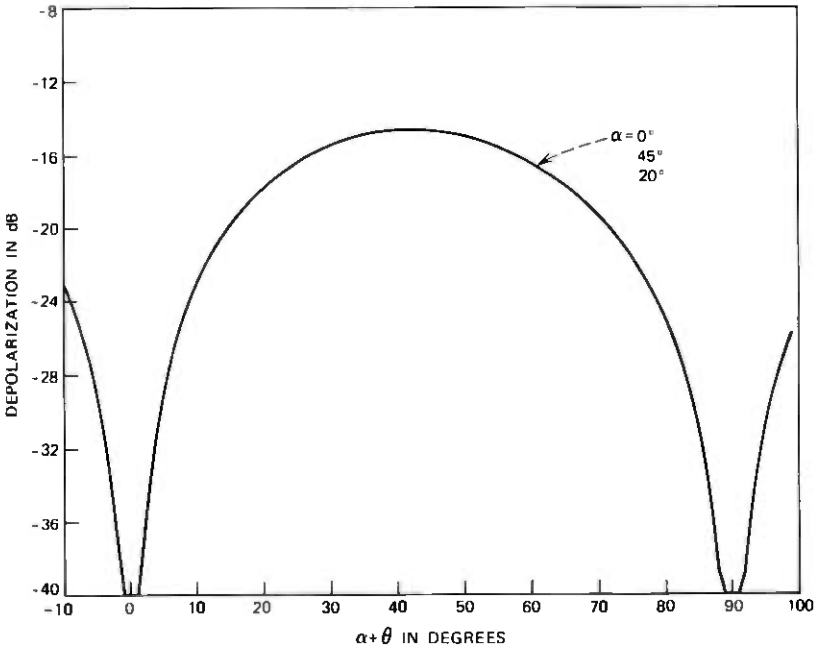


Fig. 2—Depolarization vs angle for 10-dB attenuation. No measurement errors for measurements at $\alpha = 0$, 20, and 45 degrees.

After combining (1) and (3), comparing with (2), completing the simple matrix multiplication, and defining

$$v = \cos^2\theta + d_{12}\sin^2\theta + (b_{12} + c_{12})\sin\theta\cos\theta, \tag{4}$$

then

$$\begin{aligned} A_{xy} &= vA_{12} \\ b_{xy} &= [(d_{12} - 1)\cos\theta\sin\theta + b_{12}\cos^2\theta - c_{12}\sin^2\theta]/v \\ c_{xy} &= [(d_{12} - 1)\cos\theta\sin\theta - b_{12}\sin^2\theta + c_{12}\cos^2\theta]/v \\ d_{xy} &= [\sin^2\theta + d_{12}\cos^2\theta - (b_{12} + c_{12})\sin\theta\cos\theta]/v. \end{aligned} \tag{5}$$

The absolute transmission attenuation and phase contained in A_{12} affects only the absolute attenuation and phase in A_{xy} and not the relative coefficients, b_{xy}, c_{xy}, and d_{xy}. The question of interest in the depolarization-determining propagation experiment is: What effect do errors in b_{12}, c_{12}, and d_{12} for specified α (orientation with respect to raindrop axes) have in determining the magnitudes of the depolarization coefficients, b_{xy} and c_{xy}, for $0 \leq \theta \leq 45°$? To proceed further, estimates of b_{12}, c_{12}, and d_{12} and of errors in measuring these quantities are needed.

III. ESTIMATES OF COEFFICIENTS AND MEASUREMENT ERRORS

Since direct measurements of most of the coefficients are not available, estimates must be obtained analytically using known properties of rain,[7–9] electromagnetic theory,[6,10] and the measurement data that are available.[1–11] The theoretical calculations are for paths through uniform rain. Setzer[10] calculates attenuation (db/km) and phase (degrees/km) constants for different rain rates assuming a Laws-and-Parsons drop-size distribution. Morrison et al.[6] calculate differential attenuation (db/km) and differential phase (degrees/km) for oblate spheroidal raindrops[7,8] with a Laws-and-Parsons drop-size distribution for two orthogonal linear polarizations oriented parallel to the axes of the elliptical cross section of the raindrops (see Fig. 1) at 18 GHz. The obvious difficulties in using this work are that actual rain is not uniform over the satellite path, the length of the rain-filled path is not known, the radio wave fronts will not be incident on the raindrops perpendicular to the raindrop cross-section axes with maximum ellipticity (as is more often the case for line-of-sight paths and is assumed in Ref. 6), and, of course, the drop-size distribution and raindrop ellipticity also vary from storm to storm.

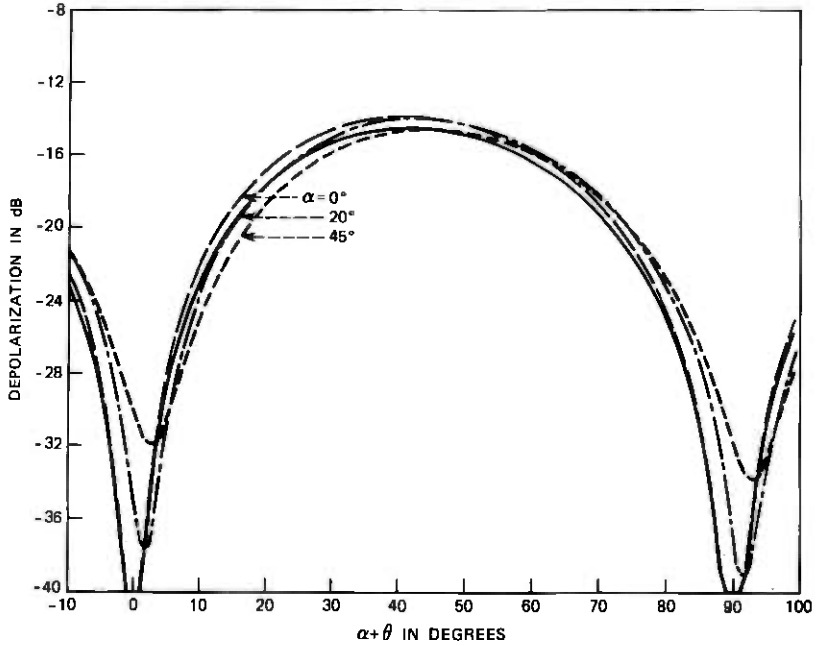


Fig. 3—Depolarization vs angle for 10-dB attenuation. Measurement error in d_{12} of +0.06 or +0.5 dB for measurements at α = 0, 20, and 45 degrees.

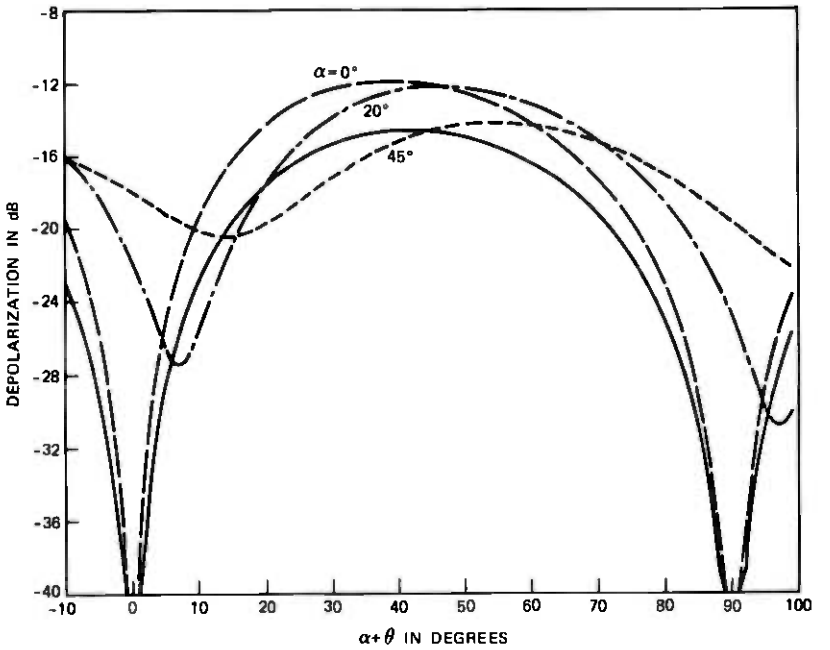


Fig. 4—Depolarization vs angle for 10-dB attenuation. Measurement error in d_{12} of +0.26 or +2 dB for measurements at α = 0, 20, and 45 degrees.

The experimental data yield path attenuation only for satellite paths[11] or differential attenuation for terrestrial paths.[2,3] The procedure that was followed in estimating satellite-path differential attenuation and differential phase was

(*i*) Choose total path attenuations,[11] 10 and 20 dB at 20 GHz, that represent rather severe conditions but that are exceeded for a significant time, $\approx$ 10 hours and $\approx$ 3 hours during a year.

(*ii*) Assume a uniform rain rate R falling over a path of length L and calculate a few R, L pairs that yield total attenuations of 10 and 20 dB using Setzer's[10] attenuation constants.

(*iii*) Calculate differential attenuation and phase between two linearly polarized signals oriented parallel to raindrop axes[6] (V and H in Fig. 1) using the calculated R and L.

(*iv*) Assume all raindrops oriented with their elliptical cross-section axes parallel and perpendicular to the signal polarizations, i.e., vertical V and horizontal H, so that, because of symmetry, the depolarization is 0.

The representative sets of estimated differential attenuation and phase

(i.e., Section II matrix coefficients) obtained by this procedure are summarized in Table I.

These representative sets were selected from the different R and L combinations for total attenuations of 10 and 20 dB. The V and H matrix coefficients were then used in the polarization-rotation equations (5) to calculate coefficients for the two signal polarizations rotated 20 and 45 degrees with respect to the elliptical raindrop axes (i.e., $\alpha = 20°$ and $45°$ in Fig. 1). These coefficients used in later calculations are tabulated in Table II.

Methods available for estimating potential measurement errors are even less rigorous than those used in estimating the matrix coefficients, since measurement errors will depend both on carrier-to-noise ratios and on equipment-measuring accuracy. The three general types of measurement errors that are distinguishable are errors in (*i*) differential attenuation or phase between the two signals transmitted with different polarizations, d_{12}, (*ii*) relative attenuation and phase between depolarized signals and direct signals, b_{12} and c_{12}, and (*iii*) absolute amplitude and phase A_{12}. Since errors in absolute measurement do

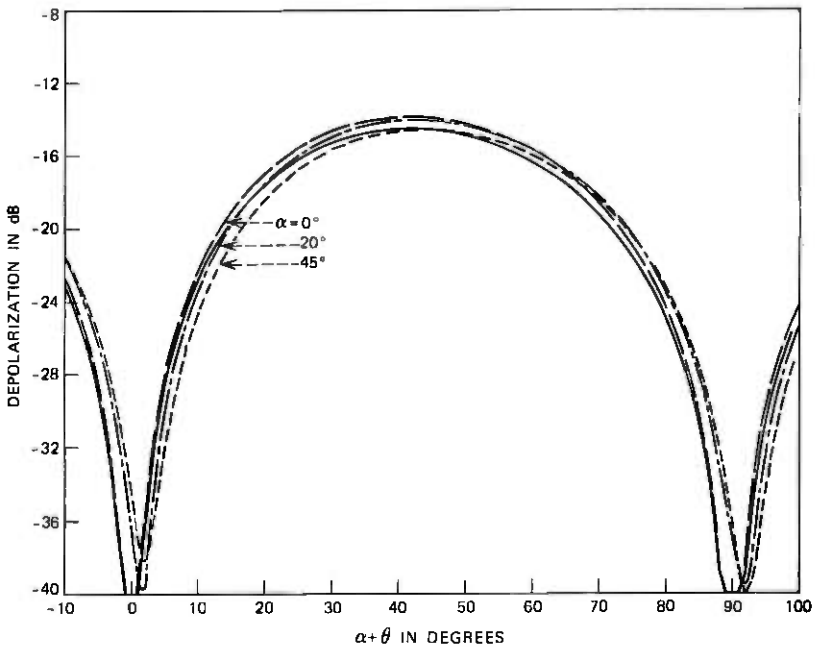


Fig. 5—Depolarization vs angle for 10-dB attenuation. Measurement error in d_{12} of +2 degrees for measurements at $\alpha = 0$, 20, and 45 degrees.

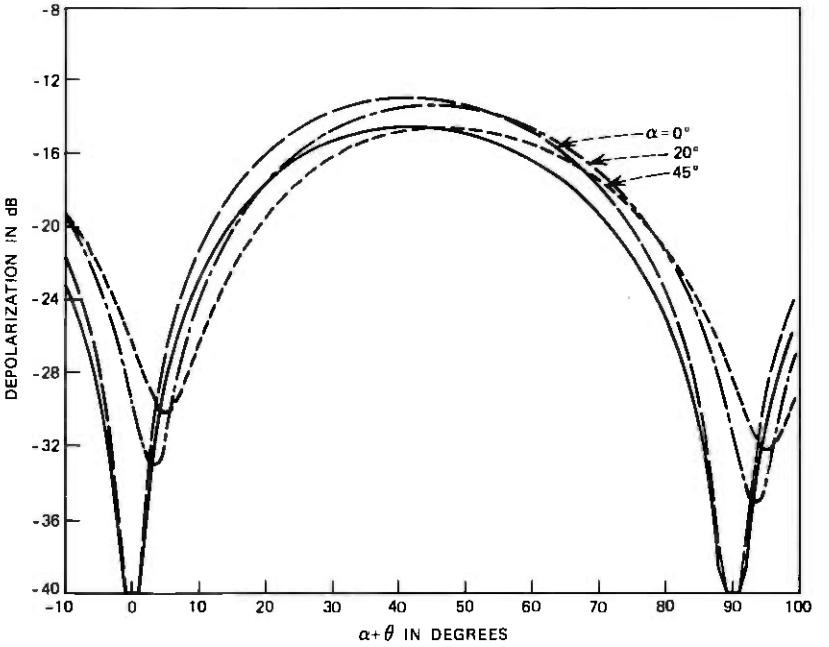


Fig. 6—Depolarization vs angle for 10-dB attenuation. Measurement error in d_{12} of $+5$ degrees for measurements at $\alpha = 0$, 20, and 45 degrees.

not affect the accuracy in calculating rotated coefficients, they will not be considered here.

Additive errors to d_{12} of ± 0.06 and ± 0.26 in magnitude and ± 2 and ± 5 degrees in angle were selected for assessing the effects of direct-signal differential attenuation and phase errors on rotated polarization coefficients. These values correspond to differential amplitude measuring errors of ≈ 0.5 and 2 dB. Such error values are likely to be contributed to by the measuring equipment or by noise, if the receiver carrier-to-noise ratio degrades sufficiently with attenuation of the signal by rain. Computing the rotated depolarization coefficients, b_{xy} and c_{xy}, in eq. (5) uses only the difference, $d_{12} - 1$, between the transmission coefficient so the magnitude error was added to d_{12} only.

Estimating errors for depolarization coefficients, b_{12} and c_{12}, is somewhat different for the V and H case than for the case of $\alpha = 45$ and 20 degrees because, for the assumed symmetric raindrop orientation, the theoretical V and H depolarization coefficients, b_{12} and c_{12}, are 0. Antenna cross-polarization isolation may be as low as 25 dB,

resulting in a contribution of 0.06 to b_{12} or c_{12}. Since this overshadows noise contributions at reasonable signal levels, a spurious magnitude of 0.06 was assumed for b_{12}. In the VH case, this is the only contribution to b_{12}, and its phase angle is unknown. Error phase angles of 45, 90, 180, and 270 degrees were selected for b_{12}.

In the 45-degree case with rain attenuation of 10 dB, the rain contribution to b_{12} is $0.1862\ \underline{/52.42}$ degrees (see Table II). Since the phase angle of the assumed error, 0.06, is unknown, several cases were considered for b_{12}: (*i*) no phase error, magnitude errors of ± 0.06, (*ii*) no magnitude error, phase error of ± 18 degrees, and (*iii*) phase error of ± 5 degrees, magnitude error of 0.06.

Similar considerations for the case of $\alpha = 20$ degrees resulted in the choice of errors in b_{12} of (*i*) magnitudes of ± 0.06, phase of ± 5 and ± 10 degrees, and (*ii*) magnitude of 0, phase of ± 30 degrees.

IV. RESULTS

It is obvious from Fig. 1 and eq. (2) that letting θ range over $0 \leqq \theta \leqq 90$ degrees and looking at $|b_{xy}|$ is equivalent to letting θ range over $0 \leqq \theta \leqq 45$ degrees and looking at both $|b_{xy}|$ and $|c_{xy}/d_{xy}|$.

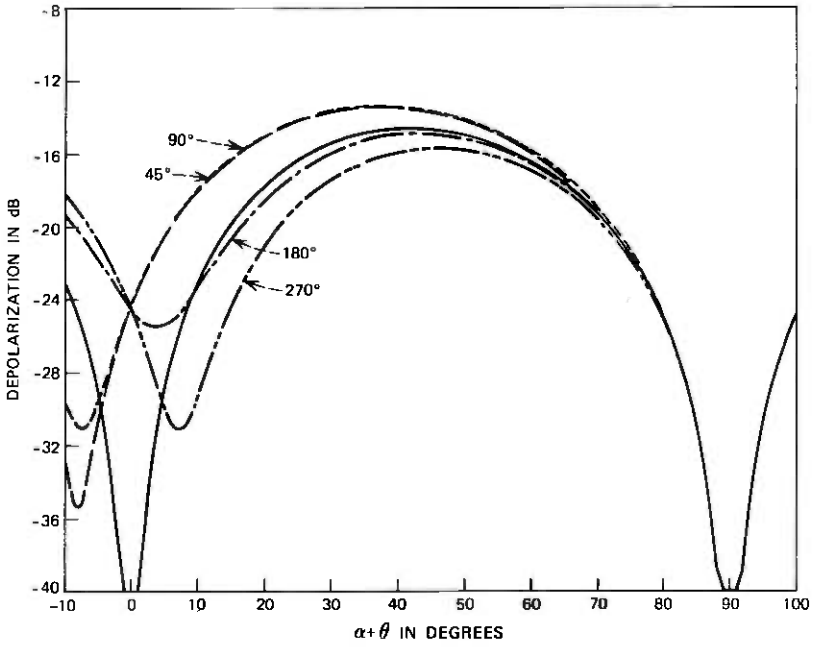


Fig. 7—Depolarization vs angle for 10-dB attenuation. Measurement error in b_{12} of $+0.06$ at phase angles of 45, 90, 180, and 270 degrees corresponding to cross-polarization contamination of ≈ -25 dB for $\alpha = 0$ degrees, i.e., V and H.

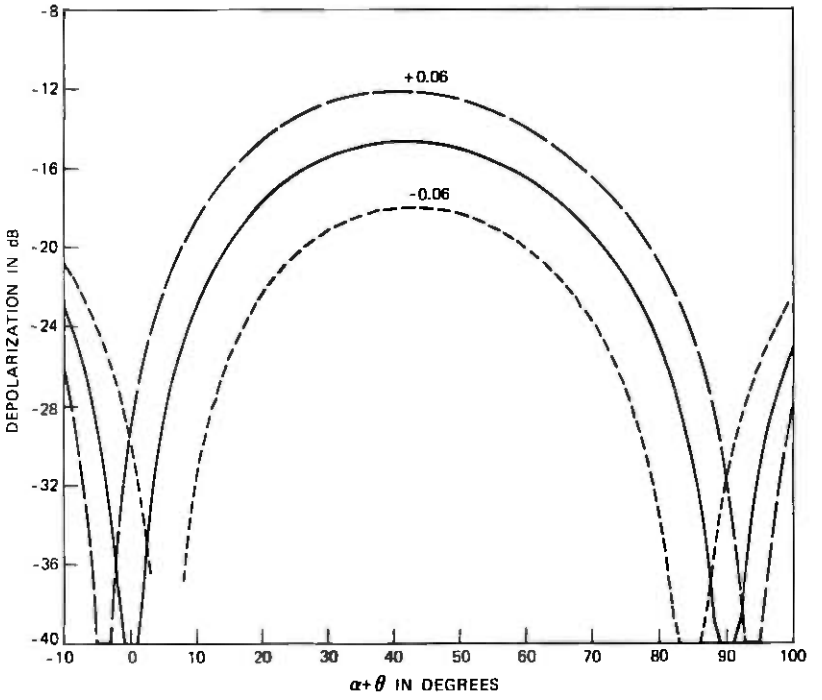


Fig. 8—Depolarization vs angle for 10-dB attenuation. Measurement error in b_{12} of ± 0.06 for measurements at $\alpha = 45$ degrees. Corresponds to cross-polarization contamination of ≈ -25 dB.

Also, any continuous 90-degree range of θ contains all values of $|b_{xy}|$. For the graphs in this section, then, $|b_{xy}|$ is plotted for the range of $0 \leqq \alpha + \theta \leqq 90$ degrees for all three initial polarization angles, $\alpha = 0$, 20, and 45 degrees. (Note that α is the orientation angle of the "measured" coefficients, 12, with respect to the axis of the elliptical raindrops and $\alpha + \theta$ is the angle of the calculated coefficients, xy, referred to the same elliptical axes and rotated an angle θ with respect to the measured coefficients, as in Fig. 1.)

Figure 2 is a plot of cross-polarization coupling (depolarization), $|b_{xy}|$, for $0 \leqq \alpha + \theta \leqq 90$ degrees computed using as starting points each of the three sets of coefficients in Table II for 10-dB attenuation with no measurement errors included. Without measurement errors, the rotated coefficients can be calculated without significant computation error. The 0 depolarization at 0 and 90 degrees is a result of the assumed perfect raindrop symmetry and orientation. In actual rain, random orientations and asymmetries will cause these nulls to fill in to some as-of-now unknown level.

Figures 3 to 10 are plots of depolarization, $|b_{xy}|$, for $0 \leqq \alpha + \theta \leqq 90$ degrees computed from the coefficients in Table II for 10-dB attenuation. The solid curve on each figure is the "no errors" curve from Fig. 2. Figures 3 to 6 each include measurement angles of $\alpha = 0$, 20, and 45 degrees and different errors in differential attenuation, $|d_{12}|$, and differential phase, $/\underline{d_{12}}$, indicated in the captions. In general, errors with the opposite sign of those in the figures produce error curves with approximately the same magnitudes but shifted in the opposite angular direction from $\alpha + \theta = 0$ degrees. Figures 7 to 10 include different measurement errors in depolarization, b_{12}, indicated in the captions. Each of these figures is for a specific α as indicated. Asymmetries in the error curves are a result of alloting all the measurement error to the one coefficient, b_{12}.

Figures 11 and 12 are similar plots of $|b_{xy}|$ computed from coefficients in Table II for 20-dB attenuation. They are for the different errors in d_{12}, as indicated.

Three overall effects of measurement errors on the calculated $|b_{xy}|$ near the regions of minimum depolarization that are indicated in

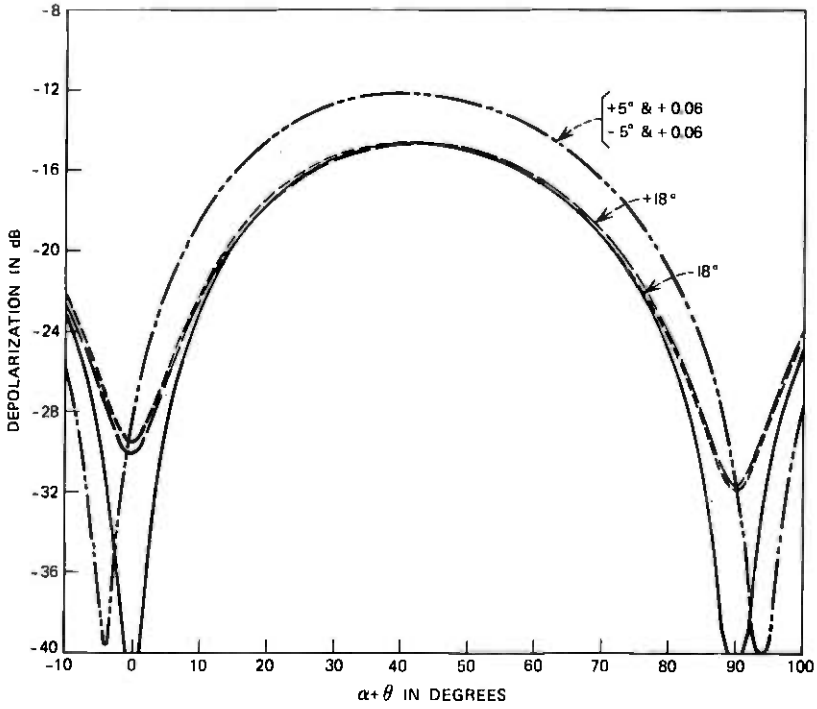


Fig. 9—Depolarization vs angle for 10-dB attenuation. Measurement error in b_{12} of ± 18 degrees and of $+0.06$ with ± 5 degrees. Corresponds to cross-polarization contamination of ≈ -25 dB.

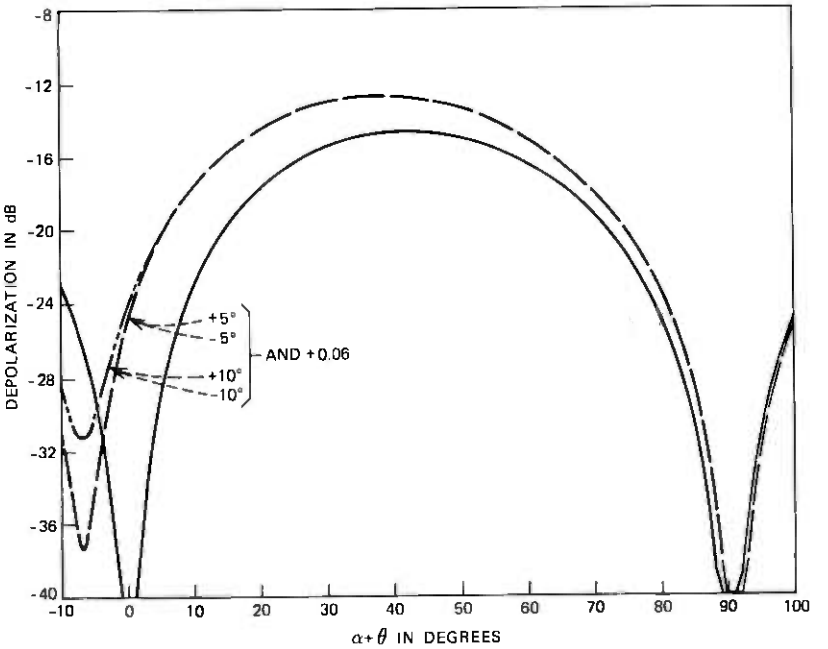


Fig. 10—Depolarization vs angle for 10-dB attenuation. Measurement error in b_{12} of $+0.06$ with ± 5 degrees and of $+0.06$ with ± 10 degrees for measurements at $\alpha = 20$ degrees. Corresponds to cross-polarization contamination of ≈ -25 dB.

Figs. 3 to 12 are (*i*) the $|b_{xy}|$ at $\alpha + \theta = 0$ and 90 degrees become nonzero in all cases, (*ii*) the minima of the $|b_{xy}|$ vs $\alpha + \theta$ curves shift in angle from the no-error 0 and 90-degree positions, and (*iii*) at the minima the $|b_{xy}|$ also become nonzero. These effects are tabulated for each separate error for initial angles, α, of 0 and 45 degrees and total attenuation of 10 dB in Table III.

From the figures and Table III, it appears that errors in measuring $|d_{12}|$ at $\alpha = 45$ degrees on the order of ± 2 dB or ± 5 degrees produce unacceptable errors (min > -30 dB and offset of min $\geqq 5°$) in $|b_{xy}|$ at the vertical and horizontal orientation, $\alpha + \theta = 0$ and 90 degrees, from the standpoint of use in evaluating future systems performance. An error of ± 2 degrees in $|d_{12}|$ at $\alpha = 45$ degrees produces acceptable errors in $|b_{xy}|$ at $\alpha + \theta = 0$ and 90 degrees. An error of ± 0.5 dB in $|d_{12}|$ at 10-dB attenuation produces errors in $|b_{xy}|$ at 0 degrees that are marginally acceptable.

V. SUMMARY

The deficiencies in the methods used to estimate the rain-produced differential attenuation and phase and the measurement errors are

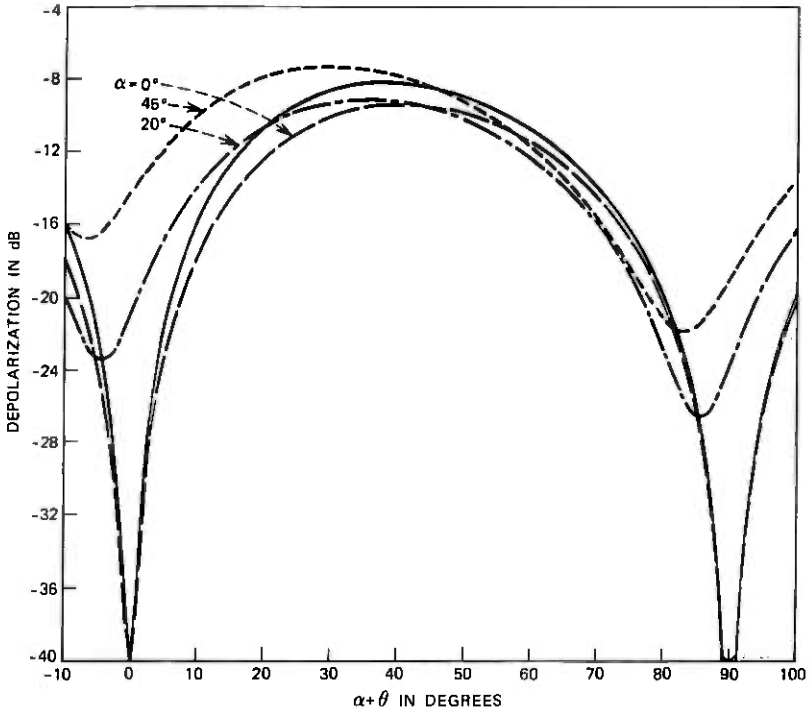


Fig. 11—Depolarization vs angle for 20-dB attenuation. Measurement error in d_{12} of -0.26 or -2 dB for measurements at $\alpha = 0$, 20, and 45 degrees.

obvious. However, a range of 20-GHz attenuation and rain rate and of measurement errors were considered, so the estimates of the effects of the errors illustrated in the figures and in Table III should be representative of those to be encountered in an actual experiment.

Calculation of rain-produced depolarization of one pair of orthogonally polarized signals from measurements of depolarization and differential attenuation and phase of a different pair of orthogonally polarized signals is quite sensitive to measurement errors. Therefore, it is better to measure the propagation parameters for the polarization orientation for which the parameters are desired. Considering future system applications, the optimum polarizations are linear, oriented horizontally and vertically (i.e., perpendicular to horizontal and the propagation path) at the receiving site, since this combination is expected to produce the minimum depolarization on the average. Measurement at the desired orientation produces the best results at the desired orientation and can produce at least partial results during partial equipment failure. Useful depolarization information can be

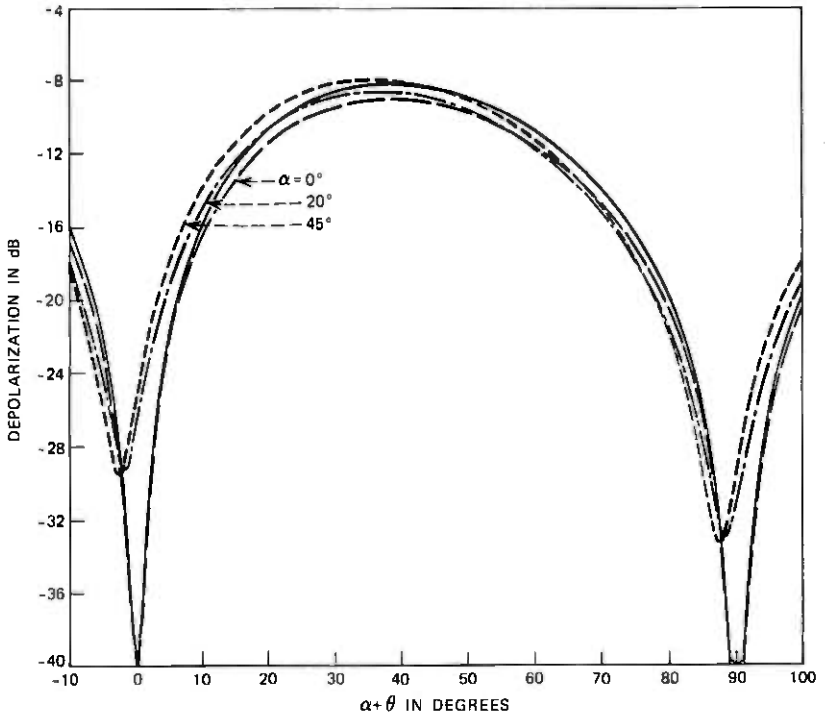


Fig. 12—Depolarization vs angle for 20-dB attenuation. Measurement error in d_{12} of −5 degrees for measurements at α = 0, 20, and 45 degrees.

Table III — Summary of effects of measurement errors on $|b_{xy}|$ for 10-dB rain attenuation and $\alpha = 0°$ and 45°

Error	Max $\lvert b_{xy}\rvert$ at $\alpha + \theta = 0°$ or 90°		Max Angular Shift of Min of $\lvert b_{xy}\rvert$		Largest Min of $\lvert b_{xy}\rvert$	
	$\alpha = 0°$	$\alpha = 45°$	$\alpha = 0°$	$\alpha = 45°$	$\alpha = 0°$	$\alpha = 45°$
	dB	dB	deg.	deg.	dB	dB
$\lvert d_{12}\rvert$						
±0.5 dB	0	−30	0	4	0	−32
±2.0 dB	0	−18	0	15	0	−20
$\angle d_{12}$						
±2 degrees	0	−34	0	2	0	38
±5 degrees	0	−26	0	5	0	30
Cross-polarization of −25 dB	−25	−29				
at max shift of min			7°	4°	−31	−40
at largest min			4°	0°	−26	−29

obtained by calculation if the satellite configuration requires measurement of propagation parameters at a polarization orientation other than the optimum discussed above.

The accuracies required in the measuring system to ensure adequate accuracy in the calculated propagation parameters (particularly depolarization) are (*i*) error in differential attenuation between the transmitted polarizations $< \pm 0.5$ dB, and (*ii*) error in differential phase for the same two signals $\leqq \pm 2$ degrees. Cross-polarization isolation of 25 dB in the measuring system is inadequate.

VI. ACKNOWLEDGMENTS

I wish to thank M. J. Gans for helpful discussions on this material and D. Vitello for programming the equations and plotting the figures used in this paper.

REFERENCES

1. R. A. Semplak, "The Effect of Rain on Circular Polarization at 18 GHz," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 1029–1031.
2. R. A. Semplak, "Attenuations Induced by Oblate Raindrops at Centimeter and Millimeter Wavelengths," unpublished work.
3. R. A. Semplak, "Effect of Oblate Raindrops on Attenuation at 30.9 GHz," Radio Science, *5*, March 1970, pp. 559–564.
4. R. A. Semplak, "Measurement of Rain-Induced Polarization Rotation at 30.9 GHz," Radio Science, *9*, April 1974, pp. 425–429.
5. T. L. Duffield, "Communications and Propagation Ground Receiving Station for the ATS-F Millimeter Wave Experiment," Conference Record of IEEE International Conference on Communications (ICC 74), June 17–19, 1974, Minneapolis, Minnesota, pp. 27F1–27F5, and D. C. Cox, "Design of the Bell Laboratories 19 and 28 GHz Satellite Beacon Propagation Experiment," op. cit., pp. 27E2–27E5.
6. J. A. Morrison, M. -J. Cross, and T. S. Chu, "Rain-Induced Differential Attenuation and Differential Phase Shift at Microwave Frequencies," B.S.T.J., *52*, No. 4 (April 1973), pp. 599–604.
7. D. M. A. Jones, "The Shape of Raindrops," Journal of Meterology, October 16, 1959, pp. 504–510.
8. M. Kumai and K. Itagaki, "Shape and Fall Velocity of Raindrops," Journal of Meteorological Society, Japan, *32* (2), 1954, pp. 11–18.
9. J. O. Laws and D. A. Parsons, "The Relation of Raindrop-Size to Intensity," Transactions of American Geophysical Union, *24*, 1943, pp. 452–461.
10. D. E. Setzer, "Computed Transmission Through Rain at Microwave and Visible Frequencies," B.S.T.J., *49*, No. 8 (October 1970), pp. 1873–1892.
11. R. W. Wilson, "Sun Tracker Measurements of Attenuation by Rain at 16 and 30 GHz," B.S.T.J., *48*, No. 5 (May-June 1969), pp. 1383–1404.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Permanent Multiple Splices of Fused-Silica Fibers

By F. W. DABBY

(Manuscript received April 15, 1974)

A method for obtaining low-loss splices of multimode optical fibers has been devised that uses no adhesives, mechanical clamps, or adjustable alignment tools. Simultaneous splices were made of three-fiber as well as single-fiber pairs. The coupling efficiency was over 95 percent for single-fiber pairs and 93 percent for three-fiber pairs. The strength of the joint was two-thirds the breaking strength of the fibers.

I. INTRODUCTION

Recently, several techniques for connecting both multimode and single-mode fibers have been developed.[1–5] The problems associated with obtaining good optical ends of the fibers to be spliced have also been the subject of a recent work.[6] These efforts have been motivated by the fact that, for optical communications to become a reality, permanent low-loss splices of fibers must be achieved.

In this paper, a method of permanently joining single and multiple pairs of fibers is described. The method used is summarized here, and a complete description of the technique and results is given in the following sections. It should be noted that the results reported for single bonds were obtained with eight splices and for the multiple bonds with 30-fiber pairs. In brief, coupling efficiency is over 95 percent, and the strength of a joint does not fail until a force of over 300 grams is applied to the end of the fiber. The fiber breaks at a force of approximately 450 grams. Alignment of the fibers is achieved by feeding the fiber into flared tunnels over a fused-silica substrate and is easily done by hand.[4] After the fibers are aligned, an aluminum washer $\frac{1}{16}$-inch in diameter is placed against the aligned fibers. The washer is centered so that the splice is in the open center of the washer. The washer is then compressed against the aligned fibers so that the aluminum yields around a portion of each fiber. The pressure required is below the breaking strength range of the fiber, and the aluminum is

permanently bonded to the fibers without bonding to the silica substrate. The yielding of the aluminum protects the fiber from the ram pressure and acts as a control, making the process independent of the heated ($\approx 300°C$) ram pressure. The process is schematically illustrated for a single fiber in Fig. 1, and photographs of the spliced single and multiple fibers and washer are shown in Fig. 2. The bonding process occurs in less than five seconds. After splicing, the washer weight is easily supported by a single fiber. This splicing technique differs from results reported previously in that the splice is a consequence of a metal oxide-glass bond and is not a result of end fusing,[1,2] mechanical clamping,[3] glass-to-glass bonds,[4] or crimping.[5]

The coupling efficiency is measured at a light wavelength of 6328 Å. The index-matching fluid used between the fibers is glycerol having a refraction index of approximately 1.48, and the fiber cores have an index of 1.45. The core diameters are approximately 100 μm and the fiber diameter 142 μm. The ends of the fibers are cut using the diamond cutter.[6]

The washer provides a convenient container for the index-matching fluid.

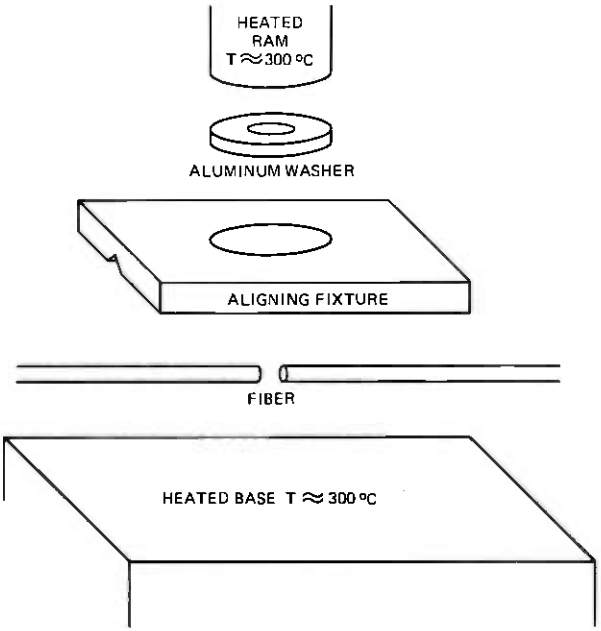
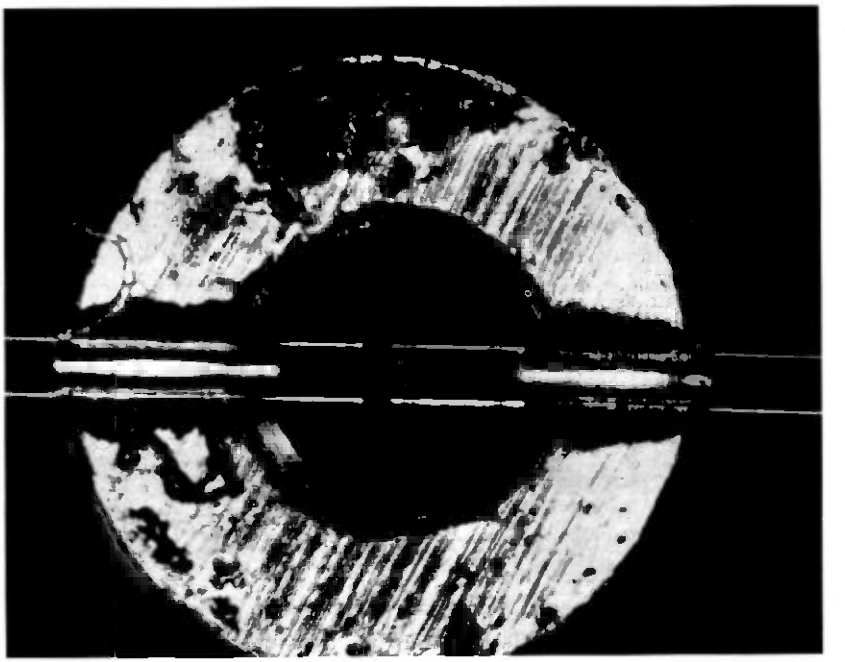
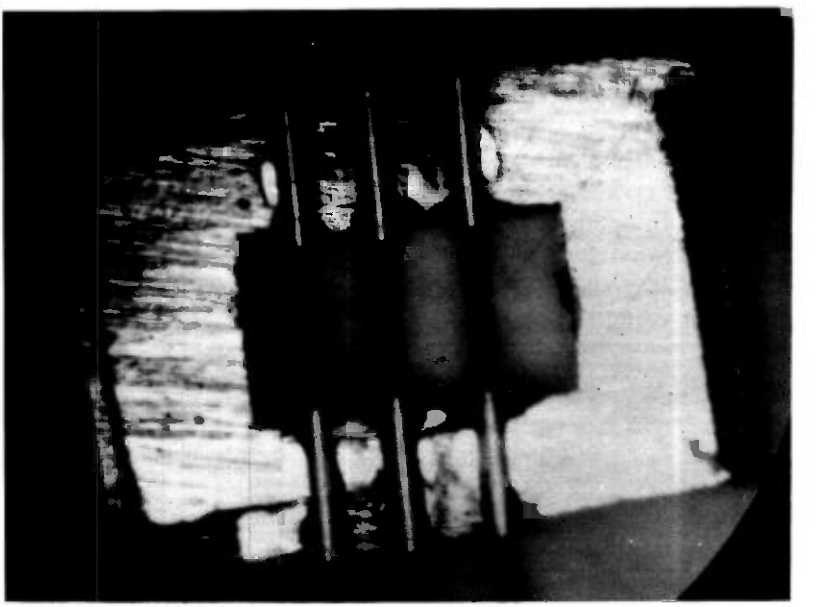


Fig. 1—Schematic of fiber splicing apparatus.



(a)



(b)

Fig. 2—Spliced optical fibers embedded in an aluminum washer. The fiber diameter in both photographs is 142 μm. (a) Single-fiber pairs. (b) Multiple fiber-pairs.

II. JOINING TECHNIQUE

To obtain a joint as shown in Fig. 2, the alignment tool shown schematically in Fig. 1 is used. Variations of this tool, using no adjustable alignment tools, have been used to splice fibers having outer diameters as low as 18 μm, and even at this small diameter the fibers were hand-fed.[4] A hole in the center of the alignment tool allows access for the bonding ram. The fibers are fed by hand from both tunnels and meet in the access hole of the ram. The ends of the fibers are cut flat by means of a diamond-cutting technique,[6] giving a measurable improvement in the coupling efficiency over fibers whose ends are not cut flat.

An aluminum washer is placed over the aligned fibers. The entire apparatus is then placed over a heated ($\approx$300°C) base, and a heated ($\approx$300°C) ram applies pressure to the washer, which yields around the fibers. The pressure applied by the ram is generated by hand, and the total force is quite low. The spliced fibers are then removed from the alignment tool by unclamping the fused-silica substrate. The final spliced fibers are shown in Fig. 2.

III. OPTICAL MEASUREMENTS

The optical measurements used apparatus and techniques that are similar to those described previously[3] and are made at a wavelength of 6328 Å. The results of the various measurements are given in Table I, which gives the average and worst results. The single-fiber-pair transmissions data exclude only one instance in which a clearly identified "mistake" occurred that was a burr of aluminum remaining in the washer and blocking the passage between the splices. The results of the multiple splices are also given in Table I, and are based on ten consecutive three-fiber splices.

The index-matching fluid used between the fibers is glycerol. It should be noted that if the ends are properly cut and the alignment maintained, coupling efficiency depends on when the glycerol is added.

Table I — Experimental results

No. of Fiber Pairs	Transmission			Breaking Strength		
	Worst (%)	Ave. (%)	No. of Meas.	Worst (grams)	Ave. (grams)	No. of Meas.
1	90.5	94.5	8	235	284	7
3	89.5*	92.7	30†	290	331	11†

* Lowest average of three-fiber pair.
† Number of fiber pair measured.

If the splices remain outside the index-matching fluid for a day, splicing efficiencies in the low-90-percent range were recorded. If the glycerol is added approximately 4 hours after splicing, the efficiency is approximately in the mid-90-percent range. Measurements made immediately after splicing showed no measurable loss (less than 2 percent).

IV. BONDS

The strength of the bonds is measured by vertically suspending a fiber splice pair and pulling the fibers apart. The results are given in Table I and exclude a single instance where a fiber failed at a point considerably removed from the bond area. The fiber failed when a force of approximately 450 grams was applied to the end of a single fiber.

The bonds are probably oxide bonds and arise when the fibers break through the alumina (Al_2O_3) coating on the washer leaving the fused-silica fiber embedded in the aluminum. The bonds could not be made at room temperature, but no research has been conducted to determine the optimum bonding temperature.

It is interesting to note that bonds of up to three-fiber pairs using a single washer have been completed. No degradation of the strength of the splice has been observed in these multiple-fiber bonds.

V. CONCLUSION

Single and multiple permanent optical connections between fibers using aluminum as a joining medium have been made. The results are strong bonds with high efficiency.

VI. ACKNOWLEDGMENT

The author is grateful for the expert technical assistance of C. M. Schroeder and technical discussions with A. Coucoulas.

REFERENCES

1. D. L. Bisbee, "Optical Fiber Joining Technique," B.S.T.J., *50*, No. 10 (December 1971), pp. 3153–3158.
2. R. B. Dyott, J. R. Stern, and J. H. Stewart, "Fusion Junction for Glass Fiber Waveguides," Elec. Letters, *8*, No. 11 (June 1, 1972), pp. 290–292.
3. C. G. Somedha, "Simple Low-Loss Joints Between Single-Mode Optical Fibers," B.S.T.J., *52*, No. 4 (April 1973), pp. 583–596.
4. A. Coucoulas and F. W. Dabby, "Glass to Glass Bonding of Optical Fibers," unpublished work.
5. A. Coucoulas and C. M. Schroeder, private communication.
6. D. Gloge, P. W. Smith, D. L. Bisbee, and E. L. Chinnock, "Optical Fiber End Preparation for Low-Loss Splices," B.S.T.J., *52*, No. 9 (November 1973), pp. 1579–1588.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 2, February 1975
Printed in U.S.A.

Microbending Loss in Optical Fibers

By W. B. GARDNER

(Manuscript received June 11, 1974)

The loss induced in optical fibers by random bends in the fiber axis is studied by winding fibers under constant tension onto a drum surface that is not perfectly smooth. The tension forces the fibers to conform to slight surface irregularities, which can result in an increase in the optical loss on the order of 100 dB/km. This microbending loss may be a significant design consideration in system applications of low-loss optical fibers. Data are presented on the reduction of the effect by means of coatings and increased numerical aperture.

I. INTRODUCTION

For the full potential of presently available optical fibers to be realized, care will have to be taken to minimize any perturbations that affect the fiber's transmission. One such perturbation is random bends in the axis of the fiber. Gloge[1] and Marcuse[2] have shown that such bends need not be of large amplitude to cause losses of a few decibels per kilometer. We have found this "microbending loss" to be common in multifiber structures. The worst of these structures add as much as 500 dB/km to the loss of the fibers. Although several decibel-per-kilometer added loss is more typical, the effect clearly poses a danger to system performance unless proper steps are taken to minimize it. The following experimental study of microbending loss shows how it can be reduced by means of coatings and increased fiber numerical aperture.

II. EXPERIMENTAL TECHNIQUE

To obtain quantitative data on microbending loss, fibers were wound under controlled tension onto a drum whose surface was not perfectly smooth. The tension forced the fiber to partially conform to the surface roughness. The resulting random bending of the fiber axis caused a measurable increase in the optical loss. The drums were 10-in. diameter cast acrylic, and no roughening of the polished surface was necessary to obtain measurable microbending loss.

The technique is illustrated in Fig. 1, where a continuous 455-m length of fiber is excited with a He-Ne laser. The left half was wound under 0.7 kg/mm^2 tensile stress and the right half under 7 kg/mm^2. It is seen from the scattering that the light is decaying much more rapidly in the right half. In fact, the microbending loss is 15 dB/km in the left half and 145 dB/km in the right half.

The method for winding the fibers is shown in Fig. 2. The pay-out shaft rides on low friction bearings and is splined to a hysteresis brake which generates a torque that is approximately independent of revolutions per minute. The torque is set by the current to the brake, and the resulting tension in the fiber is monitored with the polariscope to assure its constancy during winding. The polariscope is calibrated before each run by hanging weights on the pay-out drum while the brake is disconnected.

III. RESULTS

To determine the length dependence of the microbending loss, a Corning Glass Works (CGW) fiber with an inherent loss of 15 dB/km at 632.8 nm was wound with uniform pitch (20 turns/cm) at 2-kg/mm^2 tensile stress. A He-Ne laser beam was launched into the fiber, and the forward scattering was detected with a 1-cm-wide solar cell (appropriately baffled), whose edge was about 3 mm from the windings. Thus, the detector integrated the scattering from about 20 turns of

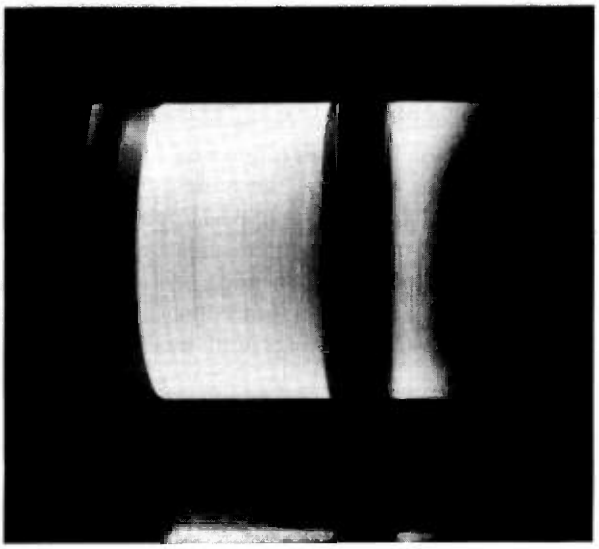


Fig. 1—632.8-nm scattering from a CGW step-profile fiber wound under 0.7 kg/mm^2 (1 kpsi) tensile stress (left half) and 7 kg/mm^2 (10 kpsi) (right half).

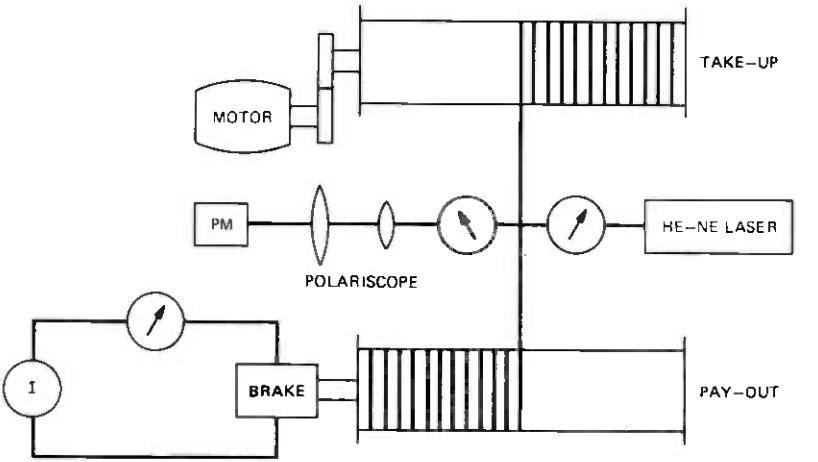


Fig. 2—Apparatus for winding fibers onto a drum under controlled tension.

the fiber. The detector was then translated parallel to the drum axis to generate the solid dots in Fig. 3. The fiber was then rewound under 0 kg/mm² and 4 kg/mm², and the scan repeated for each of these cases. Following the transient condition at the launching end, the curves become linear to within experimental error, and the slopes yield the attenuation coefficients shown. These numbers agree within experimental error with the total loss measured in the conventional way (by breaking a 1-ft length at the input end). Launching into the opposite end of the fiber did not alter the results. The linearity of the data in Fig. 3 shows the microbending attenuation coefficient γ to be independent of position along the fiber. This is to be expected when the statistics of the bending are not a function of position, and the energy distribution among the modes has reached equilibrium.

It has been shown that the microbending loss should decrease with increasing fiber numerical aperture for both parabolic[2] and step[3] index profiles. Experimental data for step-profile CGW fibers are shown in Fig. 4. The two fibers were similar except for their numerical apertures, and the microbending loss is plotted against the tensile winding stress. In a recent paper,[4] Gloge derived expressions for the microbending loss γ in both step and parabolic index fibers, assuming the spectral density of the drum roughness to be of the form

$$P(K) = P(0)/(1 + l^2K^2)^{\mu}. \tag{1}$$

Here, K is the mechanical wave number $2\pi/\Lambda$, and l has the physical significance of a correlation distance. Gloge has derived[4] a general expression for γ in terms of the parameters l and μ. This expression is

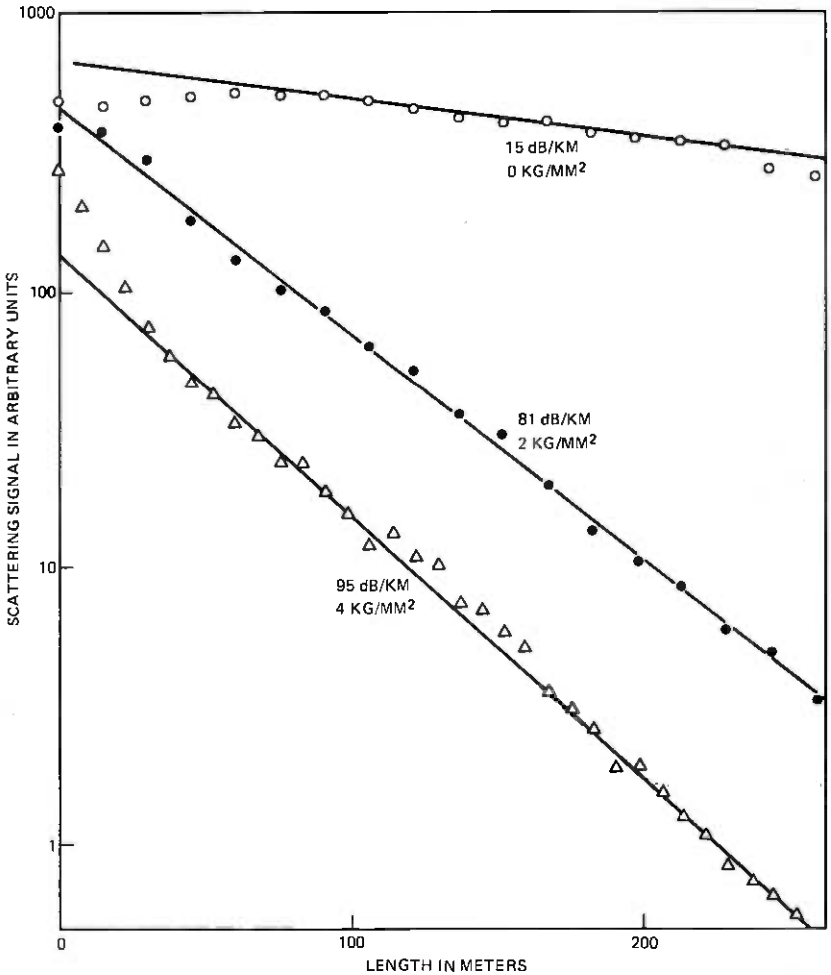


Fig. 3—632.8-nm scattering from a CGW step-profile fiber with $a_c = 44$ μm, $a_0 = 66$ μm, N.A. = 0.120.

consistent with the $\gamma \propto (\text{N.A.})^{-4.3}$ dependence manifested in Fig. 4, when $\mu = 3.1$. Setting $\mu = 3$ in the expression gives the following upper limit for the microbending loss in a step profile fiber:

$$\gamma \leqq \frac{3\sigma^2 a_c^2 / 2\pi l^5 \Delta^2}{(1 + 144\Delta^4 H^2 / 25 a_c^8 D^2)(1 + 64\sigma^4 H^4 D^4 / 225 f_0^4 l^{10})^{\frac{1}{4}}}, \tag{2}$$

where

σ = rms drum roughness
a_c = core radius

Δ = (core index − cladding index)/(core index)
= (N.A.)$^2/2n^2$
H = flexural rigidity
D = lateral rigidity
f_0 = normal force per unit length of fiber
= (tensile winding force)/(drum radius).

For an uncoated fiber of Young's modulus E_f and outer radius a_0 wound onto a drum of Young's modulus E_d, we have

$$H = \pi E_f a_0^4/4 \quad \text{and} \quad D \cong E_d. \tag{3}$$

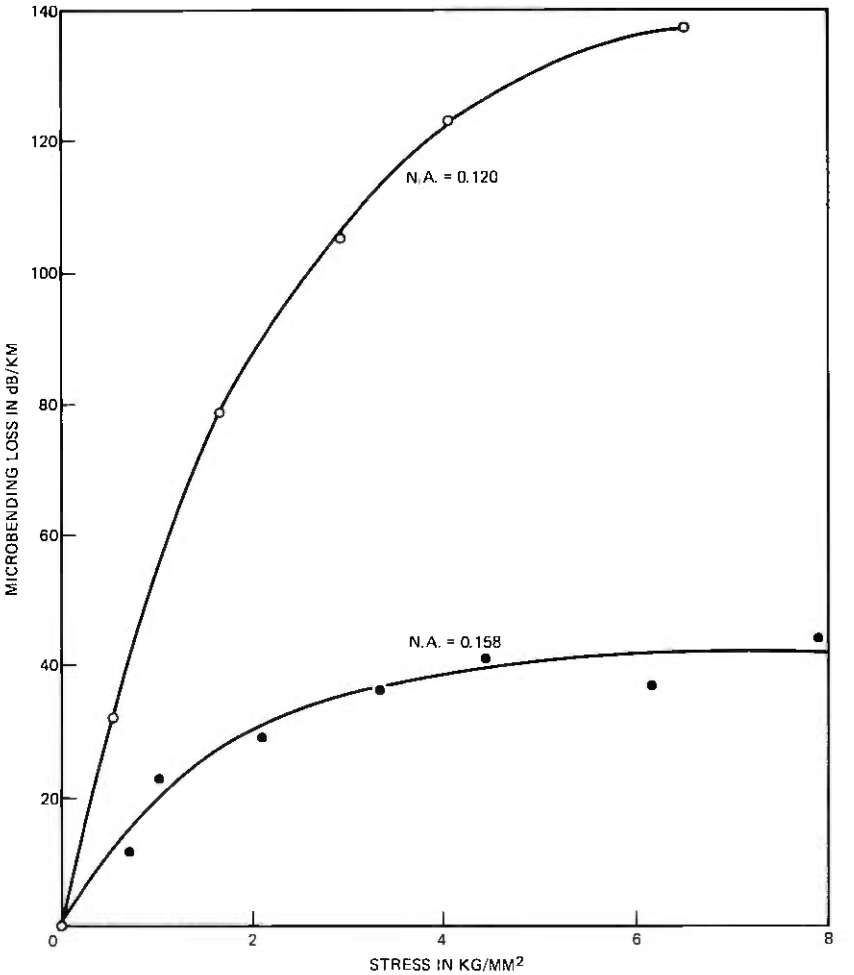


Fig. 4—γ vs winding stress at 632.8 nm for two CGW step-profile fibers which are similar except for their N.A.'s. a_c = 44 μm, a_0 = 66 μm, and the lengths were about 200 m.

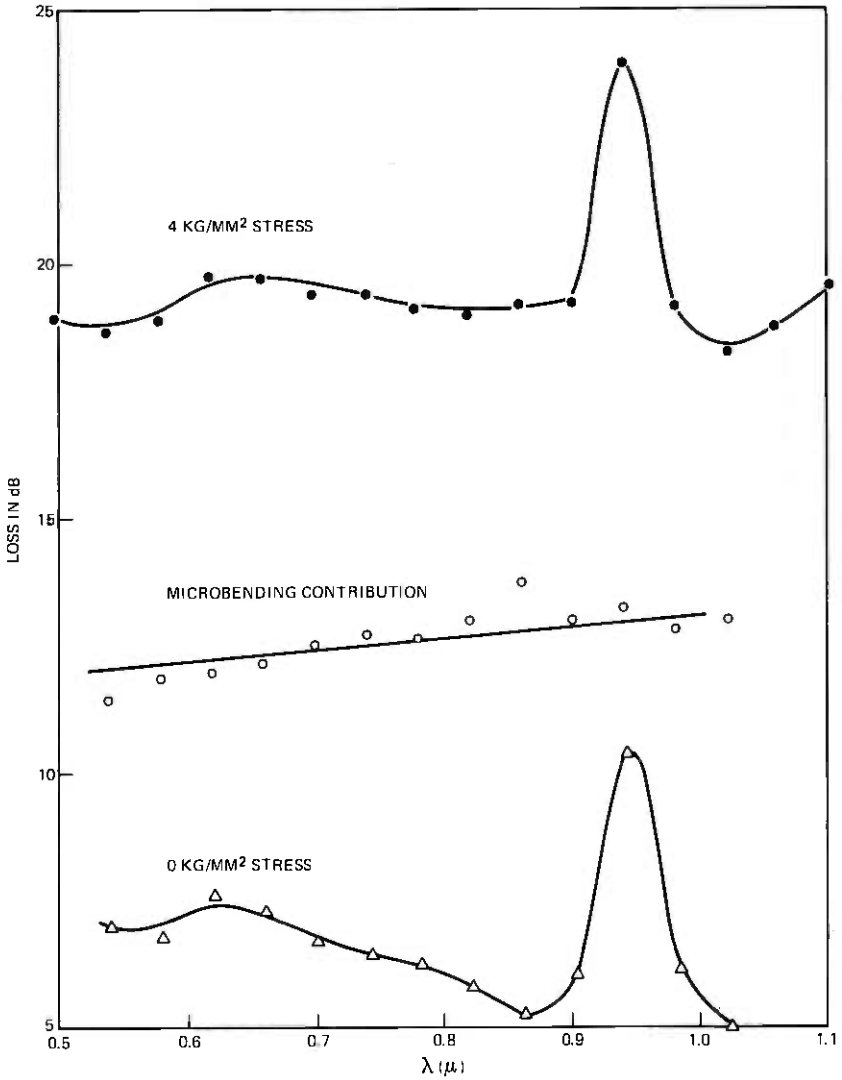


Fig. 5—Spectral loss curves showing that the added loss under stress is almost wavelength-independent.

Equation (2) assumes that l is large compared to both $K_c^{-1} = a_c/(2\Delta)^{\frac{1}{2}}$ and $(H/D)^{\frac{1}{2}}$, which are typically a few tenths of a millimeter.

Although the spectrum (1) with $\mu = 3.1$ leads to $\gamma \propto (\text{N.A.})^{-4.3}$, any other roughness (whether from coatings, drums, packaging, or whatever) will likely have a different spectral density and hence cause a different dependence of γ on numerical aperture.

There is no explicit dependence on optical wavelength in eq. (2). As a test of this, a spectral loss curve was obtained for a 160-m length CGW fiber wound first under 4-kg/mm^2 stress and then 0 kg/mm^2. Subtracting the lower curve from the upper curve in Fig. 5 gives the microbending contribution, which is indeed almost wavelength-independent. The very slight dependence on λ may be a result of dispersion in the fiber's relative index difference Δ. Similarly, measurements of the γ induced in 14 different fibers by multifiber structures were the same (within experimental error) at 0.64 μm as at 0.84 μm.

According to eq. (2), γ should be proportional to f_0 for small f_0. As $f_0 \to \infty$, however, the fiber fully conforms to the roughness, and γ becomes independent of f_0. Since f_0 is proportional to the winding stress, the shape of the Fig. 4 curves is consistent with this prediction. The value of f_0 corresponding to the transition between these two regimes can be predicted from eq. (2). The predicted value is several times larger than the measured value (which corresponds to about 3 kg/mm^2 stress) from Fig. 4. This may be because adjacent turns of the fiber are not isolated, a fact which is evident from a measured decrease in γ with increasing winding pitch. For this reason, the same pitch (20 turns/cm) was used for all measurements.

Letting $f_0 \to \infty$ in (2) and setting γ equal to the asymptotic value of the N.A. = 0.158 curve in Fig. 4 yields $\sigma^2/l^5 = 0.4 \times 10^{-6}$ mm^{-3}. A correlation distance of $l = 1$ mm, for example, would then imply $\sigma = 0.6$ μm. The existence of roughness of this magnitude is not surprising, despite the polished appearance of the surface. A 0.6-μm variation over a distance of 1 mm would be difficult to measure.

For the acrylic drum used, $D = 280$ kg/mm^2 (400 kpsi), and with the fibers used, $144\Delta^4 H^2/25a_c^8 D^2 \ll 1$, so that, in the limit of small f_0, (2) becomes

$$\gamma_0 \leqq \frac{0.76\sigma a_c^2 f_0}{l^{\frac{5}{2}}\Delta^2 a_0^{\frac{1}{2}} E_f^{\frac{5}{8}} D^{3/8}}. \tag{4}$$

From this expression, it is evident that a small core radius and large outer radius is desirable for minimizing microbending loss. The minimum usable core radius may be determined by splicing considerations, and the maximum outer radius by the bending which the fiber is required to withstand without breaking. The microbending also increases the penetration of the evanescent wave into the cladding,[5] thus possibly making thicker cladding necessary for adequate optical isolation.

In addition to maximizing Δ and a_0 and minimizing a_c and σ, a further option is available for minimizing γ. This is to encapsulate the

fiber in a compliant medium. The requirements for the encapsulant are that it be thick and uniform. The case of a homogeneous coating is illustrated in Fig. 6. The linearity of these curves is probably due to a larger σ for the acrylic drum surface used here than for the one used in Fig. 4. The coating applied to the fiber was DuPont *Elvax*® 265, a co-polymer of ethylene and vinyl acetate. The coating was 50 μm thick, with a modulus of $E_c = 1.4$ kg/mm^2 (2000 psi), and was applied

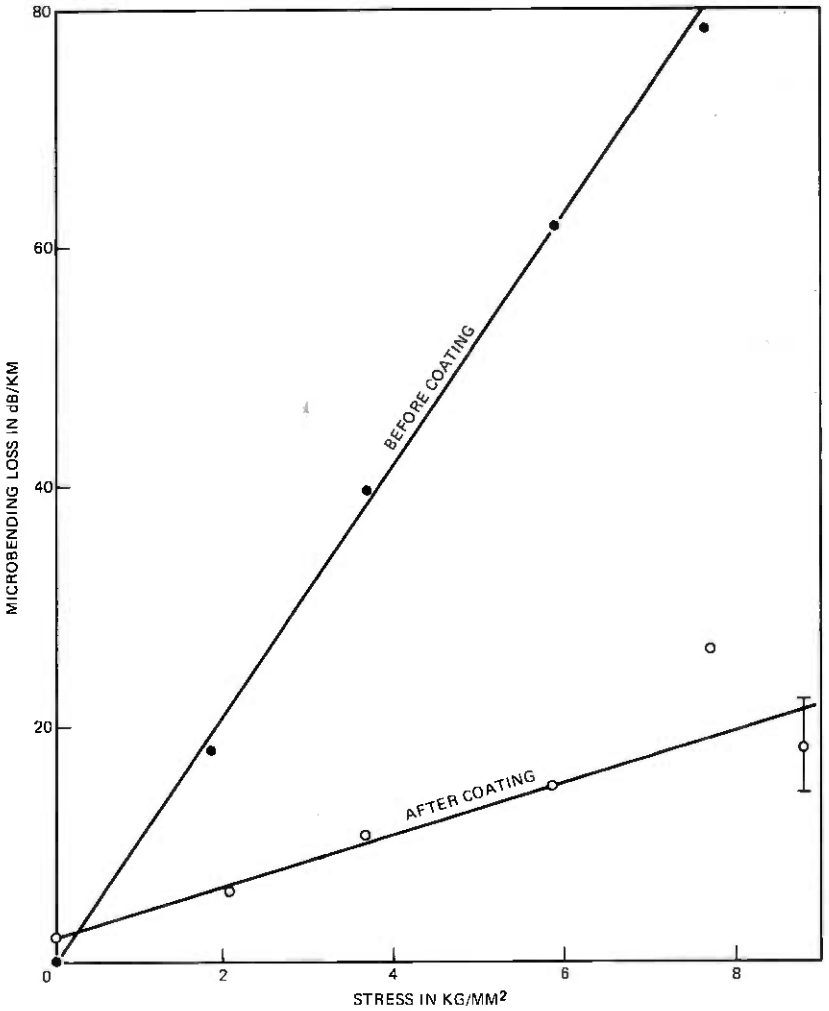


Fig. 6—γ vs winding stress at 632.8 nm before and after coating a CGW step profile fiber with a 50-μm thickness of DuPont *Elvax*® 265. The fiber was 180 meters long with $a_c = 43$ μm, $a_0 = 66$ μm, and N.A. = 0.160.

with a die technique. After the fiber is coated, D in (2) becomes $(E_d^{-1} + E_c^{-1})^{-1} \cong E_c$, while H is only slightly changed. In order for Gloge's general expression[4] for γ to predict the observed reduction in γ of a factor of 4.3 in the small f_0 regime owing to a uniform coating, μ must equal 4.4. The discrepancy between this and the value $\mu = 3.1$ (deduced from the N.A. dependence) may be an indication that the assumption of perfect coating uniformity is invalid. In that case, the coating thickness variation spectrum would add to the drum roughness spectrum, creating a new composite spectrum. Also, despite careful cleaning of the drum, it is possible that foreign material with a modulus different from E_d may make some contribution to the microbending.

IV. SUMMARY AND CONCLUSIONS

The microbending caused when an optical fiber is forced to conform to small irregularities is shown to be capable of causing sufficient optical loss to affect the performance of a communication system. Studies involving the winding of fibers under tension onto drums show significant reduction in the effect by means of coatings and increased fiber numerical aperture. Studies of multifiber structures are currently in progress and suggest that, with proper care and knowledge in design, the effect can be reduced to an acceptable level.

V. ACKNOWLEDGMENTS

The author would like to thank M. I. Schwartz for suggesting the winding procedure for creating constant stress in long lengths of fiber, B. R. Eichenbaum for creating the *Elvax*® coating, M. J. Saunders for devising the polariscope, and D. Gloge for many helpful discussions.

REFERENCES

1. D. Gloge, "Bending Loss in Multimode Fibers with Graded and Ungraded Core Index," Appl. Opt., *11*, No. 11 (November 1972), pp. 2506–2513.
2. D. Marcuse, "Losses and Impulse Response of a Parabolic Index Fiber with Random Bends," B.S.T.J., *52*, No. 8 (October 1973), pp. 1423–1437.
3. D. Marcuse, *Theory of Dielectric Optical Waveguides*, New York: Academic Press, 1974, p. 235.
4. D. Gloge, "Optical-Fiber Packaging and its Influence on Fiber Straightness and Loss," B.S.T.J., this issue, pp. 243–260.
5. D. Marcuse, "Bent Optical Waveguide with Lossy Jacket," B.S.T.J., *53*, No. 6 (July-August 1974), pp. 1079–1101.

Contributors to This Issue

Syed V. Ahamed, B.E., 1957, University of Mysore, India; M.E., 1958, Indian Institute of Science; Ph.D., 1962, University of Manchester, U. K.; Post Doctoral Research Fellow, 1963, University of Delaware; Assistant Professor, 1964, University of Colorado; Bell Laboratories, 1966—. Mr. Ahamed has worked in computer-aided engineering analysis and software design. He has applied algebraic analysis to the design of domain circuits and investigated computer aids to the design of bubble circuits. Since 1972, he has been investigating microwave devices.

Jacques A. Arnaud, Dipl. Ing., 1953, Ecole Supérieure d'Electricité, Paris, France; Docteur Ing., 1963, University of Paris; Docteur es Science, 1972, University of Paris; Assistant at E.S.E., 1953–1955; CSF, Centre de Recherche de Corbeville, Orsay, France, 1955–1966; Warnecke Elec. Tubes, Des Plaines, Illinois, 1966–1967; Bell Laboratories, 1967—. At CSF, Mr. Arnaud was engaged in research on high-power traveling-wave tubes and supervised a group working on noise generators. He is supervisor of a group currently studying microwave quasi-optical devices and the theory of optical wave propagation. Senior Member, IEEE; Member, Optical Society of America.

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Laboratories, 1953—. Mr. Beneš has pursued mathematical research on traffic theory, stochastic processes, frequency modulation, combinatorics, servomechanisms, and stochastic control. From 1959 to 1960, he was visiting lecturer in mathematics at Dartmouth College. In 1971, he taught stochastic processes at SUNY Buffalo, and from 1971 to 1972, he was Visiting MacKay Lecturer in electrical engineering at the University of California in Berkeley. He is the author of two books in his field. Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mathematical Association of America, Mind Association.

Donald C. Cox, B.S. (E.E.), 1959, and M.S. (E.E.), 1960, University of Nebraska; Ph.D. (E.E.), 1968, Stanford University; U. S. Air Force Research and Development Officer, Wright-Patterson Air Force Base, Ohio, 1960–1963; Bell Laboratories, 1968—. After coming

to Bell Laboratories from Stanford where he was engaged in microwave transhorizon propagation research, Mr. Cox was engaged in microwave propagation research in mobile radio environments and in high-capacity mobile radio systems studies until 1973. He is now doing millimeter-wave satellite propagation and systems research. Senior Member, IEEE and member, Commissions II and VI of USNC/URSI, Sigma Xi, Sigma Tau, Eta Kappa Nu, Pi Mu Epsilon.

Franklin W. Dabby, B.S. 1965, M.E.(E.E.), 1966, Cornell University; Ph.D., 1969, Acting Assistant Professor, 1969, University of California; Western Electric, 1969–1974; Bell Laboratories, 1974—. Mr. Dabby's fields of interest have included quantum electronics and optical communications. Member, IEEE, OSA, Sigma Xi.

Francis P. Duffy, B. A., 1965, King's College; M.S., 1968, Stevens Institute of Technology; Bell Laboratories, 1965—. Mr. Duffy has been involved in conducting transmission performance surveys of the toll telephone network. His activities have centered on computer applications and data analysis. Currently he is involved in studying the customer call attempt and network completion characteristics of the telephone network.

William B. Gardner, B.S., 1961, University of Alabama; Ph.D., 1968, Johns Hopkins University; Bell Laboratories, 1968—. Mr. Gardner has worked on nonlinear infrared devices and the electrical properties of backplane wiring. His present interest is in optical fiber communication systems. Member, IEEE, OSA.

D. Gloge, Dipl. Ing., 1961, Dr. Ing., 1964, Technical University of Braunschweig, Germany; Bell Laboratories, 1965—. Mr. Gloge's work has included the design and field testing of various optical transmission media and the application of ultra-fast measuring techniques to optical component studies. He is presently engaged in transmission research related to optical fiber communication systems.

Wayne S. Holden, Electronics Technology, 1970, RCA Institutes; Bell Laboratories, 1970—. Mr. Holden has been involved in the evaluation of optical fiber parameters and the design of electronic circuitry for optical fiber communication systems.

Gary K. McNees, B.S.(E.E.), 1960, University of Missouri, School of Mines and Metallurgy; M.S.(E.E.), 1962, New York University;

Bell Laboratories, 1960—. Mr. McNees has been concerned with characterization of performance of the telephone network. He is presently engaged in exploratory studies of mechanized call disposition detection in network performance monitoring.

Debasis Mitra, B.Sc. (E.E.), 1964, and Ph.D. (E.E.), 1967, University of London; United Kingdom Atomic Energy Authority Research Fellow, 1965–1967; University of Manchester, U.K., 1967–1968; Bell Laboratories, 1968—. Mr. Mitra, a member of the Mathematics of Physics and Networks Department, is interested in the application of mathematical methods to physical problems.

Kurt H. Mueller, E.E. Diploma, 1961, Ph.D. 1967, both from Swiss Federal Institute of Technology; Bell Laboratories, 1969—. Mr. Mueller has worked on various problems in the field of high-speed data communication. During 1972 and 1973 he was on leave of absence at the Swiss Federal Institute of Technology and was a member of the Executive Body of the European Informatics Network. He is presently involved in digital signal processing techniques for data transmission. Member, IEEE.

Ingemar Nåsell, Civilingenör, 1955, Royal Institute of Technology, Stockholm, Sweden; M.E.E., 1962, M.S. (Mathematics), 1965, and Ph.D., 1971, New York University; Research Institute of National Defense, Stockholm, Sweden, 1955–1960; Bell Laboratories, 1960—. Mr. Nåsell has been concerned with characterizing the transmission performance of the telephone network for systems engineering purposes, most recently as supervisor of the Surveys Planning group concerned with statistical designs of surveys. He presently is on leave of absence to the Royal Institute of Technology, Stockholm. Member, Svenska Teknologföreningen, American Statistical Association, Eta Kappa Nu.

F. A. Pelow, A.A.S., 1972, Mohawk Valley Community College; United States Air Force Communication Center, 1966–1970; Bell Laboratories, 1972—. Mr. Pelow is currently working with experiments on quasi-optical devices.

Lawrence R. Rabiner, S.B., S.M., 1964, PhD., 1967, Massachusetts Institute of Technology; Bell Laboratories, 1962—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques.

Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; President, IEEE G-ASSP Ad Com; member, G-ASSP Technical Committee on Digital Signal Processing, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Technical Committee on Speech Communication of the Acoustical Society; former Associate Editor of the G-ASSP Transactions.

Marvin R. Sambur, B.E.E., 1968, City College of New York; S.M., 1969, Ph.D., 1972, Massachusetts Institute of Technology; Bell Laboratories, 1968—. At present, Mr. Sambur is engaged in automatic speaker verification and automatic speech recognition research. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

David A. Spaulding, A.B., 1959, M.S., Dartmouth College; M.S., 1961, Ph.D., 1965, Stanford University; Bell Laboratories, 1966–1973. From 1966 to 1969 Mr. Spaulding was involved in studies of active networks, pulse-shaping networks, and adaptive equalizers for data transmission systems. From 1969 to 1973 he was supervisor of the Digital Data Sets Group which was concerned with the study and development of modulation techniques, adaptive equalizers, timing and carrier recovery systems, and digital filters for use in high-speed voice-band data transmission systems.

T. W. Thatcher, Jr., B.S.E.E., 1942, Iowa State University; Bell Laboratories, 1942—. Mr. Thatcher formerly supervised a group working on objectives, field trials, and application information for short-haul FDM carrier telephone systems. For the last few years, he has supervised field measurement activities and reporting of results of telephone network performance surveys. Member, IEEE.