THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV NOVEMBER 1965 NUMBER 9

Copyright © 1965, American Telephone and Telegraph Company

Experimental 224 Mb/s PCM Terminals

By J. S. MAYO

(Manuscript received July 14, 1965)

Experimental 224 Mb/s terminal equipment for a toll-grade, long-haul PCM transmission system has been designed, constructed, and successfully operated. Transmission of television and frequency-multiplexed mastergroups of voice channels is emphasized. Fundamental design considerations are explored. Detailed designs and performance levels are briefly mentioned, but are covered in detail in companion articles.

I. INTRODUCTION

During the late 1950's, the feasibility of pulse code modulation for the transmission of voice over cable pairs was established,[1] and since 1962 there have been substantial installations of such equipment. More recently, attention has been focused on extending PCM to much higher bit rates and to long-haul, toll quality systems. This effort has produced experimental 224 Mb/s PCM terminals capable of transmitting broadband signals such as television and mastergroups of voice channels with a signal quality that meets Bell System transmission objectives.

High-performance, high-speed terminals having provision for all major functions required in a commercial system were constructed in order to demonstrate the applicability of PCM to the broadband, long-haul network. Satisfactory solutions to all the major technical problems associated with high-speed PCM terminals have been demonstrated. Two important results of the studies are: the feasibility of

precise encoding of broadband signals such as television and mastergroups by means of all solid-state circuits, and the feasibility of adding and dropping channels by digital means without locking the coder sampling frequencies to the line transmission frequency. Thus, a satisfactory solution to the PCM network synchronization problem has been demonstrated. Thorough analysis and experimental demonstration of satisfactory operation of jitter removal equipment indicates confidence that high-speed PCM systems can operate in the presence of time jitter on the received pulse train.

The experimental terminal has been an invaluable asset in obtaining experimental verification of analytical work done over the years on impairments introduced into broadband signals as a result of quantization, overload, time jitter, and digital transmission errors. The analysis was verified without significant discrepancies.[2] The experimental terminal has also been a valuable vehicle for high-speed circuit studies. At the inception of this work there was considerable question as to the technical feasibility of operating thousands of transistors in a circuit environment that required switching times as low as a fraction of a nanosecond. Although isolated circuits had been previously operated at these speeds, there was considerable uncertainty regarding the feasibility of interconnecting and reliably operating large amounts of circuitry at nanosecond speeds.

II. PROBLEM AREAS — PCM TERMINALS

2.1 *Coding and Decoding*

Rendering broadband signals such as color television and frequency division multiplex (FDM) mastergroups into pulse sequences with high precision is a difficult technical task. It is relatively easy to achieve sufficient coder precision so that moderately good picture and voice transmission (6 to 7 digit quality) may be demonstrated over a single codec (coder-decoder). It is a much larger task to build broadband codecs of sufficient precision that a very high-quality signal is delivered after passage through numerous codecs in tandem (9 digit quality). Tandem codecs will be required until digital transmission is available on all routes within the country, for the signal must be decoded each time it passes an interface between a PCM transmission link and an analog transmission link. Also, in the case of coded FDM mastergroups, the signal must be decoded to get access to channels within the mastergroup. With coded mastergroups, if 120 channels are to be dropped and added along a PCM route, the mastergroup must be

decoded, the 120 channels (two supergroups) demultiplexed in the frequency domain, two new supergroups multiplexed back into the frequency slots previously occupied by the dropped supergroups, and then the new master group coded for transmission by PCM. It is apparent that the through channels experience an additional coding and decoding each time channels are added to the mastergroup.

2.1.1 *Television Signals*

A minimum bandwith of approximately 4.5 Mc/s is required for adequate transmission of black and white or color television. The PCM coder must, therefore, sample the television signal at at least a 9-Mc/s rate, and the sample must be coded with sufficient precision to avoid significant impairment of the pictures.[3] The number of levels or codes required in the coder depends on the signal-to-noise requirement, the number of codecs to be operated in tandem, and whether the full sync pulse is encoded along with the video component of the signal. Performance also depends on whether the video signal is dc clamped ahead of the coder or whether the signal may "drift" with dc content. Assuming the bottom of the sync pulse is clamped to the first code and the peak video excursion extends to the highest code, the quantization noise performance of Table I is achieved. It is seen that with a 1-db framing impairment (impairment resulting from a particular method of word framing), and recognizing that a 51-db peak-to-peak signal to rms noise ratio is required for a high-quality picture, seven binary digits per code will render a completely satisfactory signal. Allowing for reasonable departures from theoretical performance, up to nine digits per code are required when half a dozen codecs are operating in tandem. Although 8-digit coding may be satisfactory in the present television network, attention has been directed toward 9-digit coding in order to provide increased flexibility.

Assuming 9-digit encoding and a 12-Mc/s sampling rate (in order to

TABLE I — PCM-TV PERFORMANCE LEVELS

Number Digits	$(S/N)_T$	$(S/N)_A$	n
7	52	51	1
8	58	56	3
9	64	59	6

$(S/N)_T$ = Theoretical peak-to-peak signal to rms noise ratio (including 1 db framing impairment).
$(S/N)_A$ = S/N ratio expected in codec at end of maintenance period.
n = number tandem codecs to give S/N = 51 db.

make the television bit rate twice that required for a coded mastergroup) leads to the time scale shown on Fig. 1, which is typical for a color television signal. The signal is sampled every 83 ns, and the signal in the vicinity of a sample changes at a maximum rate of about six steps per nanosecond. Analysis shows that for adequate performance (impairment small compared to quantizing noise) the mastergroup or television signal must be sampled in about 0.5 ns (gate switching time), and the value of the sample must be held constant during the coding interval to an accuracy of one part in several thousand.[4]

The television signal must be bandlimited to approximately half the sampling frequency. It is also important that the filters that accomplish this objective exhibit good transient response. The design of the bandlimiting filters at the input to the coder and the output of the decoder is established by balancing foldover distortion against transient response. Sharp cutoffs produce excessive ringing, while a gradual cutoff produces excessive foldover distortion. The filters may also employ pre-emphasis and de-emphasis.[5]

2.1.2 *Mastergroup Signals*

A mastergroup such as is transmitted in the Bell System's L3 carrier system is made up of 600 voice channels, frequency-division mul-

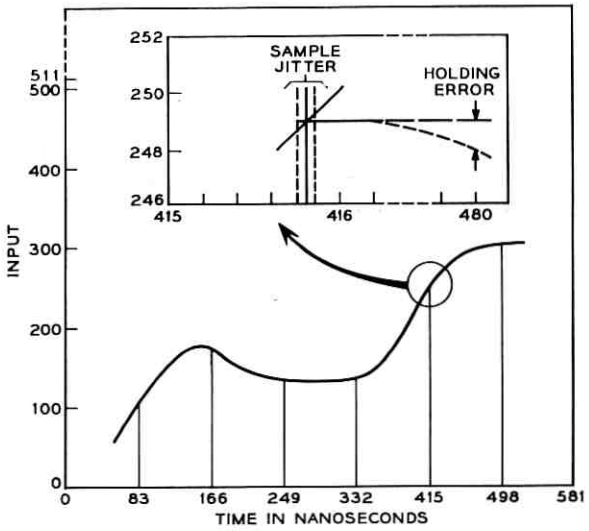


Fig. 1 — High-speed PCM sample and hold. (Expanded-scale insert shows sensitivity to sampling jitter and effect of "droop" on held sample waveform.)

tiplexed by single sideband techniques in a frequency band extending from 564 to 3084 kc/s.[6] It is well known that the mastergroup signal has a Gaussian amplitude distribution.

The mastergroup signal must be applied to a PCM coder in such a way that all code levels are exercised to the greatest extent possible, yet peak signal excursions beyond the range of available codes should occur very infrequently.[7] The theoretical noise performance of a mastergroup codec has been computed, experimentally verified, and is shown in Fig. 2. Minimum noise results when the amplitude of the mastergroup signal applied to the coder is set so that the rms signal

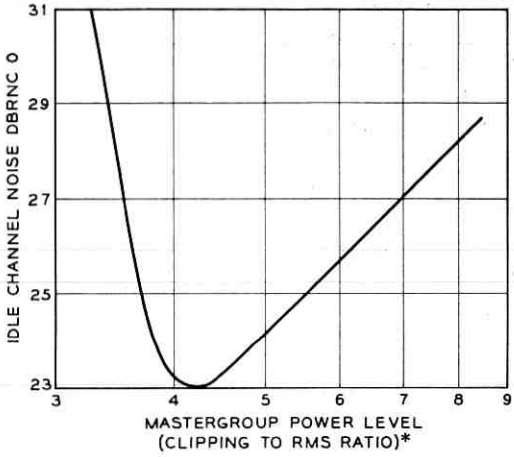


Fig. 2 — Theoretical noise performance of 9-digit mastergroup codec.

voltage is approximately one-eighth of the peak-to-peak voltage required to exercise all codes. For higher mastergroup power levels (small clipping to rms ratio), excessive noise is generated by frequent overloads, while for smaller mastergroup power levels the signal-to-noise ratio decreases because the quantization noise remains constant but signal amplitude drops. Under optimum signal amplitude conditions, 8-digit quantization results in approximately 29-dBrnc0 noise while 9-digit quantization yields a 23-dBrnc0 noise level. To obtain actual noise levels for a given telephone connection, allowance must be made for the operation of several codecs in tandem, imperfections in codec operation, and noise in the FDM equipment. It results that linear 9-digit coding is entirely satisfactory.

With the mastergroup in its normal frequency assignment of 564-

* Ratio of input amplitude at which coder clips to applied signal rms amplitude.

3084 kc/s, the minimum sampling rate is 6.168 Mc/s. This sampling rate may be lowered somewhat if the mastergroup is shifted down in the frequency domain prior to coding. When the mastergroup is coded directly the frequency band from dc to 564 kc/s is wasted. It is worth noting that the master group signal is severely bandlimited as a result of the frequency domain channel stacking, and only modest filters are required before and after the codec.

2.1.3 *Voice Channels*

The D1 channel bank of the T1 carrier system codes 24 voice channels into a 1.5-Mb/s pulse train.[8] The experimental terminal, therefore, does not provide for direct coding of voice. However, the design of the multiplex equipment allows the 224-Mb/s pulse train to be made up of various types of signal components, including 1.5-Mb/s streams of the type transmitted in T1 carrier.

2.1.4 *PICTUREPHONE* Signals*

A broadband signal that may be of considerable importance in future carrier systems is that produced by the Bell System's *PICTUREPHONE* set. The 0.5-Mc/s bandwidth signal has been coded into PCM and transmitted over the experimental equipment. The exact bit rate required for transmission of the *PICTUREPHONE* signal is still under study but is probably in the range of 3 to 6 Mb/s. Rather than develop special coding equipment for *PICTUREPHONE* signals, the sampling rate and the number of digits per code of the mastergroup codec were reduced in order to convert the *PICTUREPHONE* signal into PCM form. A coded *PICTUREPHONE* signal was also transmitted over the 224-Mb/s stream in the time slots normally allotted to two T1 line signals.

2.2 *Multiplexing and Demultiplexing*

The bit sequences from the various coders and bits from other sources such as data sets may be readily interleaved to form a high-speed pulse train provided the various input bit rates are exact submultiples of the high-speed line rate. Assume for the time being that the sampling rates for the various coders are locked to some master frequency, so judicious selection of sampling rates and number of digits per sample gives coder output rates that are harmonically related. A

* "*PICTUREPHONE*" is a service mark of American Telephone and Telegraph Company.

firm requirement is that the system be capable of accepting the T1 line signal (1.544 Mb/s). Also, a rather firm requirement for 9-digit encoding of mastergroups and a minimum sampling rate of 6.168 Mc/s for the unshifted mastergroup has been established. By selecting a sampling rate of 6.176 Mc/s (four times the T1 rate) for the mastergroup, the resulting codec output rate is precisely 36 times the T1 line rate. The multiplex equipment, therefore, sees a mastergroup as equivalent to 36 T1 line signals.

It is very desirable to have an integral relationship between the bit rate required for mastergroups and the bit rate required for television. It does not appear feasible to transmit high-quality television over the bit rate required for a coded mastergroup. It is, therefore, convenient to sample the television signal at twice the mastergroup sampling rate or 12.352 Mc/s, and it has been shown desirable to code television to nine digits also. The multiplex, therefore, sees the television signal as equivalent to two mastergroups or 72 T1 line signals.

There are various fundamental approaches to multiplexing lower speed bit streams into a high-speed stream. These differ primarily in the amount of digital storage provided in the multiplex. Since high-speed digital storage is very expensive, bit-at-a-time multiplexing is presently the most attractive approach for high-speed PCM.

The bit stream organization chosen for the experimental terminal is shown in Fig. 3 which gives the format when transmitting one television signal, one mastergroup signal, and one T1 line signal over the 224-Mb/s line. Ignore the synchronizing pulse for the time being and note that the multiplex circuit generates a basic frame of 145 pulse positions. The frame rate is 1.544 Mb/s, and the last time slot of each frame is devoted to a framing pulse which marks reference time. The

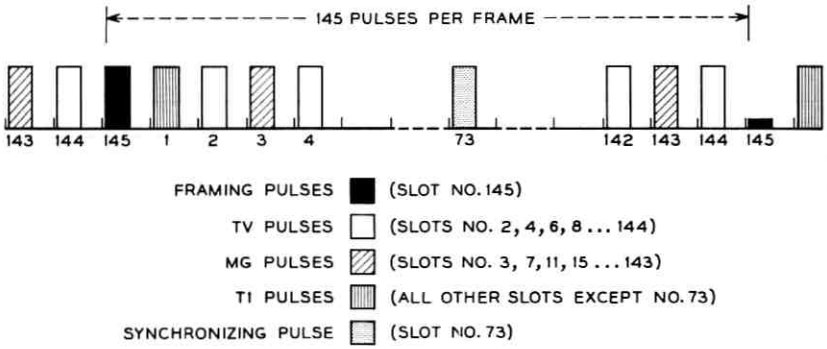


Fig. 3 — 224-Mb/s line organization.

framing pulse is given on ON-OFF statistic (alternating ONE-ZERO) which is very improbable for any other pulse, and the demultiplexer searches for a pulse which repeats at the frame rate and obeys the alternate ON-OFF statistic. A coded television signal is multiplexed into alternate time slots within the frame and a coded mastergroup multiplexed into every fourth time slot. The T1 line signal is transmitted on the 224-Mb/s line as a single pulse per frame, so the 224-Mb/s line may simultaneously transmit up to two television signals, or up to four mastergroups, or 144 T1 line signals. It can accommodate any combination of these signals totaling 144 units, where a television signal requires 72 units, a mastergroup 36 units and a T1 line signal requires one unit.

2.3 *Framing*

The need for an occasional framing pulse on the high-speed line in order to identify reference time at the demultiplexer is now obvious. Since bit multiplexing was chosen, additional framing is required at the various decoders to identify the most-significant digit of each code group. Had word multiplexing been chosen the 224-Mb/s line framing signal would also have identified the most-significant digit of each word. On the other hand, word-at-a-time multiplexing requires an excessive amount of high-speed storage, and leads to an inflexible multiplex based on a particular word length.

There are fundamentally two approaches to word framing of the various decoders. One approach makes use of known signal statistics and compares the statistics of the signal (either before or after decoding) to the known signal statistics.[9] When the decoder is operating on improperly grouped codes, the input-digit and output-signal statistics are altered, this fact is detected, and the phasing of the decoder is shifted until the output obeys the proper statistic. In the experimental terminal, this type of framing is used in the mastergroup codec. This coder operates in the reflected binary or Gray code, and it can readily be shown that such a coder operating on a Gaussian signal of the rms value previously shown to be optimum for coding of mastergroups produces output digits where the probability of a one in various digit positions is essentially 0.5 for all digits except the second, which has a 0.95 probability of being a one. Code groups are identified within the mastergroup decoder, therefore, by searching for the digit of each code that has high probability of being a one.[10] Many other approaches to statistical framing have been investigated.

A second approach to word framing a decoder is to add a distinguishable statistic to the signal being coded. This may be added in analog form before the coding process or in digital form after the coding process. The approach is demonstrated in the experimental system in the television codec. The least-significant digit of each ninth coded word from the television coder is forced to an alternate 1,0 sequence. The framing circuit for the television decoder, therefore, looks at every 81st received time slot and searches for a phase position where each 81st pulse follows an alternate 1,0 pattern. It can be shown that the quantizing noise impairment introduced by robbing the least-significant digit every m words is

$$10 \log \left(1 + \frac{3a}{m}\right) \text{dB},$$

where $a = 2$ if the decoder operates on the robbed bits just as if they were signal bits, and $a = 1$ if the decoder does not operate on the robbed bits. For $m = 9$, the increase in theoretical quantization noise is 2 dB if no special action is taken at the decoder or 1 dB if the robbed bits are not decoded.

Framing by means of occasional digits has been selected as preferable to utilizing clusters of framing bits in order to minimize the amount of high-speed digital storage required. Samples delivered by the decoder must be at a uniform rate, and clusters of framing digits generally interfere with the uniformity of receipt of codes unless sufficient digital storage is provided to bridge the time gaps introduced by the framing pattern. Also, it may be difficult to completely remove the time jitter introduced by large time gaps.

2.4 *Synchronization*

Multiplexing was stated to be a relatively simple operation if the bit streams entering the multiplexer are precise submultiples of some master frequency. This is readily accomplished if the sampling clocks for the various coders are locked to the same master frequency. In a nationwide network, however, coders are spread all about the country, and it is not a simple matter to frequency lock all coders to a master clock. If the various sampling clocks in a PCM network are not submultiples of the same frequency, multiplexing may still be readily accomplished provided techniques are developed for shifting a coder output from one rate to a slightly different rate. Both bit-rate shifting techniques and locked-frequency techniques have been examined.

2.4.1 *Master Clock*

The most obvious approach to frequency locking the sampling rates of coders at various geographic locations is to distribute a master clock to all coder locations. Sampling clocks for all coders are then derived from this master clock, so the outputs of the various coders are harmonically related in frequency. Then multiplexing and demultiplexing, including facility for digital adding and dropping, and the gating of certain portions of a high-speed bit stream from one line to another, are readily accomplished.

Such a system has numerous undesirable features. Since one clock serves the whole nation, the clock and its distribution system must be extremely reliable — protected by redundancy against technical failures as well as man-made or natural disasters. Since precise relative phasing must be maintained on all bit streams entering a high-speed multiplex, this approach requires that all transmissions delay variations be built out by variable delays located ahead of the multiplexer. The delay variation in the clock distribution system itself will amount to approximately 20 μs for a 1000-mile low-frequency radio link, approximately 2 μs for a coaxial link of the same length and approximately 0.2 μs for a 1000-mile microwave radio link. Even 1 μs of delay variation may necessitate the need of 100 bits of high-speed storage. Such a system, therefore, requires relatively large digital stores at each multiplex point. The storage time must be variable or of the "elastic delay" type, because the sum of transmission delay and delay through the store must be held constant to approximately 1 ns.[11,12] Two signals arriving at a point over separate 1000-mile links of cable must be phased to ± 0.5 ns. This is to be compared to the total delay of each 1000-mile circuit of approximately 8 ms.

The master clock plan has appreciable "start up" costs and is especially unattractive for the early days when there are relatively few PCM systems to share the relatively large costs of the master clock. Also, the technique is "brittle" in that delays must be precisely matched — precise length of cable interconnecting equipment is important, and there is appreciable equipment shared over otherwise independent transmission links. At the time of introduction of the master clock, existing PCM systems, such as T1, would also have to be modified in order to lock their clock rates to the master clock rate.

2.4.2 *Phase Averaging*

A technique has been studied which allows sampling clocks of all coders to be frequency locked, yet does not establish any individual

clock as a master.[13] This technique is known as phase- or frequency-averaging. It makes use of the fact that an interconnected network of digital systems, especially two-way systems, has pulse streams entering and leaving every codec location. One may, therefore, establish a reference phase or frequency for each codec location which is the average of all phases or frequencies entering that location. If each location transmits the same phase or frequency to all other connected locations, it can be shown that the resulting reference frequencies established at the various locations are identical. For a particular implementation, where the reference phase at each location is established by an oscillator whose phase is locked to the average phase of all signals entering the location, it can be shown that the resulting sampling frequencies throughout the network are not only identical, but also bounded by the lowest and highest free-running frequencies of the various oscillators in the network. It can also be shown that the sampling frequency established by such a network responds to transients in a damped manner. If the network is disturbed (such as by addition or removal of equipment), the sampling frequency will settle in a well-controlled manner to the new frequency.

Networks synchronized by phase averaging do not have as serious a reliability problem as those synchronized by a master clock. On the other hand, much of the "brittleness" of the master clock system remains, i.e., precise control of phase, much equipment common to otherwise independent transmission links and necessary modification of existing systems. The scheme has been analyzed in considerable detail by Bell Telephone Laboratories' Systems Research Department, however, further analysis of specific embodiments of the phase averaging technique should be completed before committing a large system to this approach.

2.4.3 *Stable Clocks*

The use of very stable oscillators as sampling clocks allows asynchronous operation of a PCM network. If C bits of digital storage are provided at each multiplex interface, and a bit stream of frequency f_0 is multiplexed to a frequency $f_0 + \Delta f$, then the digital store will be exhausted (contents depleted or storage capacity exceeded) every $C/\Delta f$ seconds. Defining $s = \Delta f/f_0$ as the clock stability factor, the store is exhausted every C/sf_0 seconds. When operating at the highest pulse-rate of interest i.e., 100 Mb/s for television, store exhaustion period is approximately $C/s \times 10^{-13}$ days. Each time the store is exhausted, a group of pulses is either repeated or lost, and the system

may have to reframe to re-establish proper timing sequence. If the store capacity and store resetting mechanism repeats or drops complete frames, then exhaustion of store capacity at the multiplex interface will result only in the loss of information bits and will not initiate a long reframing sequence. Nevertheless, if bits are to be lost or inserted into the pulse stream not more often than once per day, then the size of store required is $s \times 10^{+13}$ for the 100-Mb/s signal, as shown in Fig. 4. With the clock stability factor of 10^{-13} only one time slot of storage is required for once-a-day reframing. On the other hand, a clock stability factor of 10^{-10} dictates 10^3 bits of storage to produce once-a-day reframing.

There does not seem to be a question of technical feasibility of operating even high-speed PCM networks asynchronously. Use of a cesium beam clock with a long-term stability of 10^{-12} results in a very small storage requirement at all multiplex interfaces except at very high speeds and then only tens of bits (at 100 Mc/s). On the other hand, very stable clocks are expensive and, therefore, the stable clock must be shared over a number of systems — resulting in reliability problems. However, the most serious drawback of this approach to synchronization results from the fact that occasionally bits are lost or are repeated.

Although the long-term digital error rate due to store exhaustion is equal to the stability factor, s, the errors generally occur in bursts and/or produce reframing. The time of occurrence of these bursts may

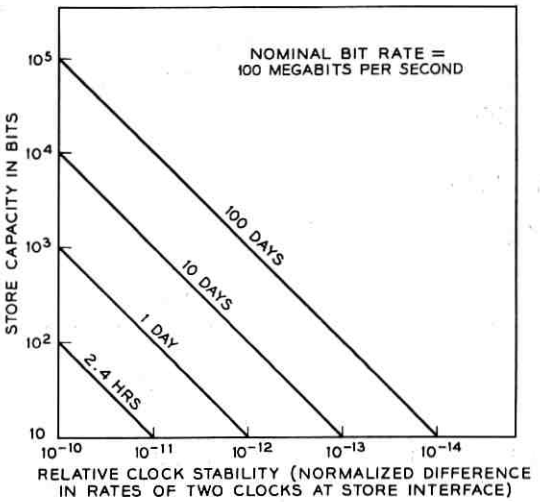


Fig. 4 — Contours of constant store exhaustion periods.

be controlled by command resetting of the store. This approach to synchronization may be entirely satisfactory from the point of view of analog signals, but may be disastrous for certain data signals. Although simple in concept, this approach is not as economical as might appear to be the case at first glance, unless frequent reframings are allowed.

2.4.4 *Pulse Stuffing*

A technique for asynchronous multiplexing has been developed which does not require that all coder clocks be synchronized, yet does not periodically lose bits. In this approach a coder does not provide as many pulses per second as the multiplexer needs, and the multiplexer is arranged to skip over occasional time slots so as to make up the frequency difference.[14,15] The multiplexer also communicates to the demultiplexer the precise locations of the "stuffed" time slots. The demultiplexer removes the stuffed slots from the pulse train, closes the time gaps occupied by the stuffed slots, and thus returns the pulse stream to its original form.

The basic operation is shown in Fig. 5. Pulses entering the multiplex point are written into a 3-bit digital store, and are read from store by

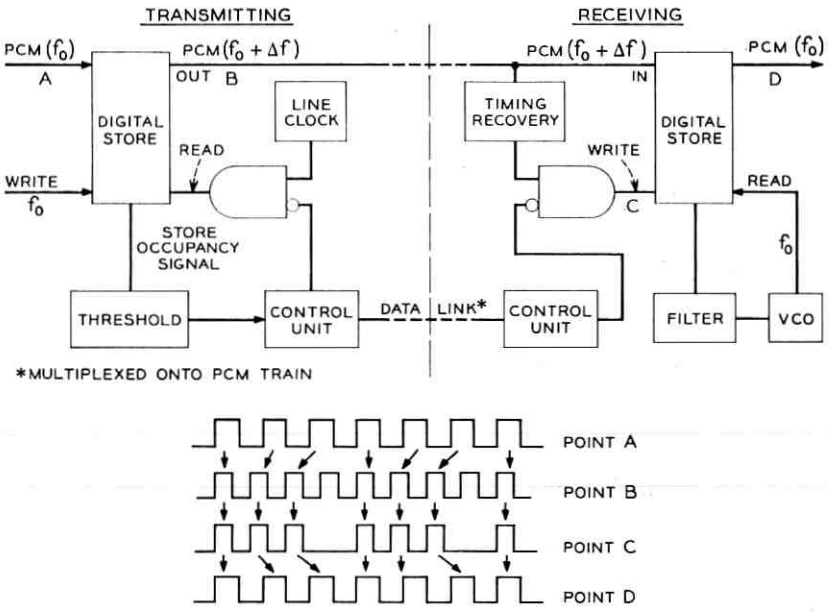


Fig. 5 — Pulse stuffing synchronization.

a submultiple of the line clock which is slightly faster than the input bit rate. There is a built-in tendency to exhaust stored bits, and the number of stored bits is monitored by a store occupancy signal. When this signal exceeds a critical value, the threshold circuit signals that a time slot should be stuffed into the output train. Prior to the actual stuffing operation, a control unit communicates to the receiving equipment the precise location of the stuffed time slot, then at the appropriate time provides the signal to inhibit reading of the store, allowing transmission of a stuffed time slot.

At the receiving end of the system the pulse train is written into a small store, but the stuffed time slots are not allowed to enter the store. Only the original coder output pulses flow through the store and these are read from the store at a smooth rate by using the store occupancy signal (after appropriate filtering) to drive a voltage controlled oscillator.

The store occupancy signal is a measure of the difference between phases of input and output pulses, and the closed loop around the oscillator is a rather conventional phase-locked loop.[16] This loop should have a cutoff frequency which is low compared to the average stuffing rate. In this way, time jitter on the output signal resulting from the removal of the stuffed pulses will be small and this jitter will not accumulate very rapidly on signals passing through tandem multiplexers.

The transmitting equipment must be capable of signaling the location of the stuffed pulse positions to the receiving equipment, and do this precisely, even in the presence of high transmission error rates. Error free signaling is most readily accomplished by utilizing redundant coding in the "data-link" of Fig. 5, but adequate signaling may also be accomplished by a predictive receiving circuit which operates on the basis that stuffing rates vary slowly with time and a future stuffed pulse position may be predicted on the basis of time occurrence of all past stuffed pulse positions.

Signaling the stuffed slot position information over the PCM line (multiplexing the "data-link" of Fig. 5 into the PCM stream) is attractive, and may be accomplished by devoting an occasional pulse on the PCM line entirely to this purpose. Another alternative is to multiplex the signaling information into the bit stream from the coder (without increasing bit speeds). J. W. Pan has proposed a "statistical subcarrier" approach based on this latter technique wherein certain codes are forbidden at the coder, and these forbidden codes are substituted for certain probable codes in order to transmit stuffed slot

location information. At the receiving end receipt of the forbidden codes is noted, and the codes are changed back into the original probable code.

The most promising approach, however, appears to be the technique of adding occasional additional bits into the pulse stream as a "data-link" to communicate "stuffed" time slot positions. This approach was taken in the experimental terminals where a single pulse per frame was devoted to this task. This pulse was transmitted over a pulse position normally allotted to a T1 line signal, but normally a frame of 146 pulses would be required to transmit a frame of 144 information pulses. The last pulse per frame is used for framing, and the 73rd used for the synchronization data link, as shown in Fig. 3.

2.5 *Jitter Reduction*

The delay a PCM pulse train experiences in passing through a self-timed regenerative repeater is necessarily a function of pattern density being transmitted.[17] The amount of jitter introduced by each repeater is small, perhaps a few degrees of phase shift at the pulse repetition frequency.[18] For random pulse patterns, it is probable that the jitter introduced by a given repeater is random with an rms value of a few degrees. This jitter is introduced at each repeater and accumulates along a string of repeaters proportional to the square root of the number of repeaters.[19] A repeater design that results in an rms jitter of five degrees per repeater will result in the accumulation of 150° of rms jitter in a string of 1000 such repeaters.

The bandwidth of the jitter introduced into the pulse stream by a number of self-timed repeaters is determined by the effective Q of the timing extraction circuit of each repeater. Since the transmission pulse rate is usually very high compared to the bandwidth of a signal before coding, and since extremely high repeater Q's are difficult to achieve and result in uneconomical repeater designs, then it is quite likely that a long string of regenerative repeaters will introduce timing jitter of sufficient amplitude and broad enough frequency spectrum that it will impair the quality of any broadband signal being transmitted. Time jitter on the PCM pulse train results in pulse postition modulation of the decoder output, and this introduces noise components into the signal.[20]

Both television and mastergroup signals are quite vulnerable to pulse jitter. The amount of jitter that one might tolerate in a coded mastergroup has been computed, experimentally verified, and is

shown in Fig. 6. Results are plotted for the most vulnerable channel (the one multiplexed to the highest FDM frequency slot), and two constraints have been placed on the resulting signal impairment. For large jitter bandwidths, the effect is to produce crosstalk between channels within the FDM package, and this has been constrained to equal theoretical 9-digit quantizing noise (if the jitter is random the crosstalk will be unintelligible, but if the jitter is sinusoidal, and a multiple of 4 kc/s, intelligible crosstalk would be produced). For jitter bandwidths much less than 4 kc/s, the effect of jitter is to shift frequency components within a given channel, which results in signal distortion. The curve of Fig. 6 constrains low-frequency jitter to a level which produces a signal-to-distortion ratio of 30 dB.

Jitter introduced into a PCM pulse train does not constitute an irremovable signal impairment. It is relatively easy to remove jitter from a pulse train, particularly bothersome high-frequency jitter. This is accomplished by writing the jittered pulse train into a digital memory and reading the pulses from memory at a smoothed rate. A circuit for accomplishing this, a dejitterizer, is shown in Fig. 7.[11] The input pulses are read sequentially into memory cells in a jittered fashion. However, the bits are read from memory by the controlled oscillator, which is locked to the fundamental pulse rate. The flip-flop phase comparator keeps the store "half full" on the average by supplying a control signal for the oscillator. By use of a suitable low-pass filter ahead of the controlled oscillator, the resulting phase-locked loop may be designed to have a high Q. Obviously, the pulse stream out of the dejitterizer will

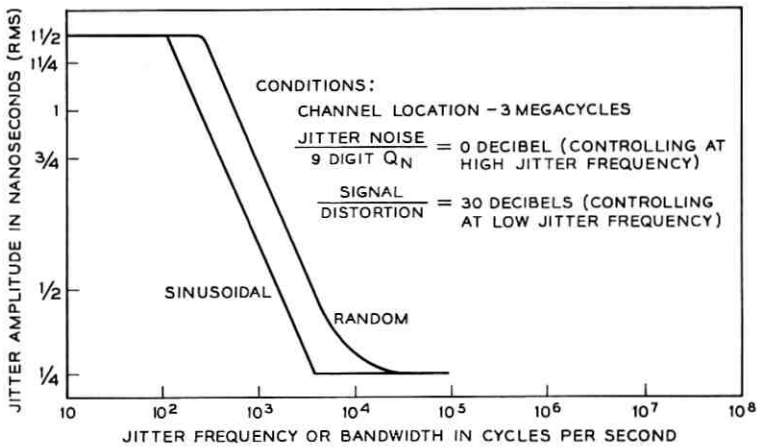


Fig. 6 — Jitter allowed in FDM-PCM system.

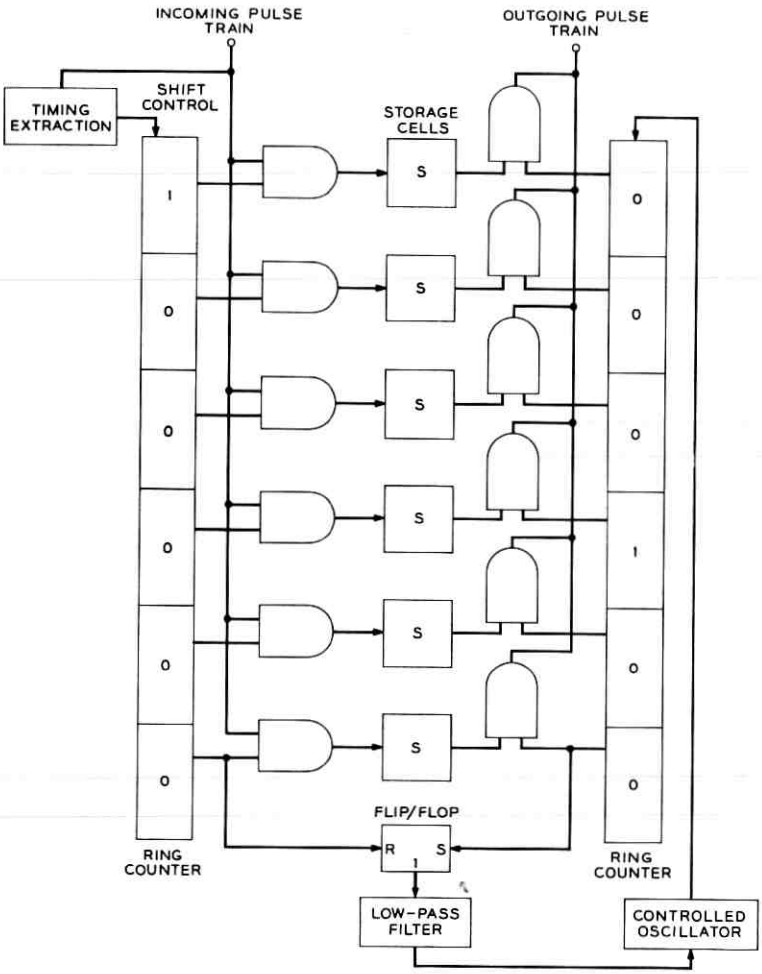


Fig. 7 — Jitter reducer circuit.

contain some low-frequency jitter components, as determined by the frequency response characteristic of the phase-locked loop. Very high effective Q's can be realized by reducing loop gain, but this results in large static phase shifts (a large correction signal must be supplied from the phase detector to overcome any drifts in natural frequency of the oscillator). Large static phase shift "wastes" storage capacity in the memory.

Although the dejitterizer does not remove very low-frequency jitter components, very low-frequency jitter components do not significantly

impair the coded signals. For example, very low-frequency jitter is introduced by delay variations in the transmission media, especially as a function of temperature, and such slow variations could be removed only by the introduction of very large stores. The amount of very low-frequency jitter that one can tolerate depends on the types of signals being transmitted, and the jitter requirements shown in Fig. 6 were deliberately not shown in the frequency range below ten cycles. It is apparent that the curve of Fig. 6 can be extended from 1.5 ns at ten cycles to an infinite amount of jitter at zero frequency. For speech, the subjective impairment of large amounts of low-frequency jitter is small, so the low-frequency end of Fig. 6 is determined by certain narrowband special service signals that may be transmitted over a voice channel. Subjective tests of the effect of jitter on color television suggest the signal is not significantly impaired by Gaussian random jitter of up to 1 ns rms amplitude.

III. THE EXPERIMENTAL TERMINAL

The experimental terminal design focused attention on the important PCM problem areas to be overcome prior to the design of a commercial high-speed system. These areas include:

— Coding and decoding of black and white and color television

— Coding and decoding of mastergroups of voice channels

— Multiplexing of lower-speed digital signals into high-speed PCM

— Ability to organize the bit stream in such a way as to accommodate a wide range of input signal mixtures

— Ability to control information flow through the PCM network with adequate reframe times for each circuit

— Ability to operate the various coders with sampling rates independent of the line transmission rate

— Ability to reduce the expected amount of pattern jitter accumulation to levels that may be tolerated by the signal components.

An important overall objective was that the various signal components be processed with the precision required for a commercial system.

The resulting terminal block diagrams are shown in Figs. 8 and 9. A commercial television signal is sampled at 12 Mc/s and each sample coded with 9-digit precision. A mastergroup signal is sampled at 6 Mc/s and also coded with 9-digit precision. Capability for transmission of two 1.544-Mb/s PCM signals is also provided, and the additional unused time slots are filled with random pulses. The master-group coder and the 1.544-Mb/s line signals are not frequency locked to the 224-Mb/s

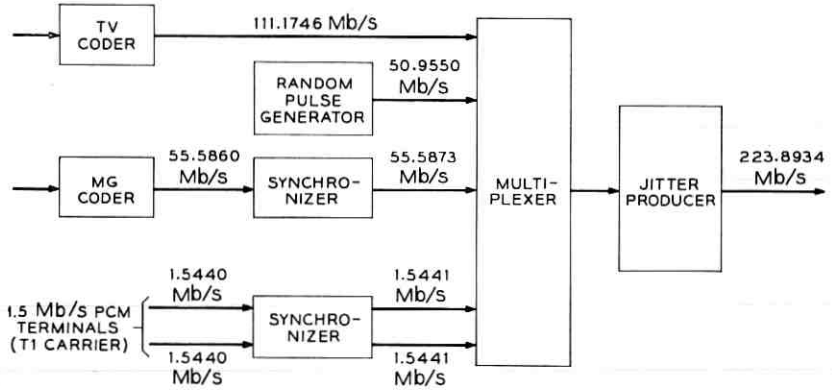


Fig. 8 — Transmitting terminal.

line, and the frequency difference is made up by pulse stuffing-type synchronizers. The multiplexer interleaves the various signal components and the jitter producer simulates transmission of up to 4000 miles of repeatered line. The received signal is dejitterized and demultiplexed. The television signal is decoded, and thereby restored to analog form. The mastergroup signal must have the stuffed time slots removed by the desynchronizer prior to decoding. The stuffed time slots are also removed from the 1.544-Mb/s signals, which may then be transmitted over T1 carrier lines to T1 carrier terminals. The T1 line signals have also been used to transmit high-speed data and coded *PICTUREPHONE* signals.

Although the experimental equipment appears to be a point-to-point system, all the features have been provided that are required

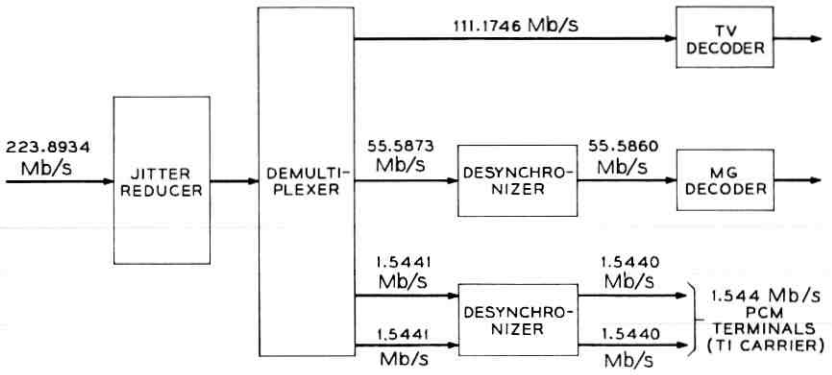


Fig. 9 — Receiving terminal.

for terminal equipment located along the PCM route to add and drop portions of the 224 Mb/s stream.

In addition to the terminals described above, a 224-Mb/s repeatered line (for use with coaxial cables) has been developed. This line will be the subject of a future paper by the group responsible for repeater development.

IV. PREFERRED CIRCUIT APPROACHES

4.1 *Coding*

Work on the broadband experimental PCM system was initiated at a time when the only sure road to success in video coding involved use of the beam encoding tube. This device was, therefore, further perfected and extended to 9-digit capability.[21] The basic device consists of cathode and lens structure for generating a ribbon beam which is focused onto a code plate. The signal sample to be coded is applied to the deflection plates. Beam deflection is proportional to the sample voltage, and apertures on the code plate allow current to be collected on output wires in an on-off pattern defining a binary number proportional to the sample. The major sources of signal impairment in such a coder are nonuniformity of current density across the beam, tilt of the beam relative to the code plate structure and an adverse ratio of smallest aperture width on the code plate to a standard deviation of the thickness of the electron beam. These parameters have been sufficiently controlled and sufficiently precise external solid-state circuits have been constructed so that the resulting coder performs with the quantizing noise level within a few dB of theoretical performance when operating at 12 Mc/s sampling rate and 9-digit coding.

An effort parallel to that devoted to the perfection of the tube coder was directed toward all solid-state encoding. A survey was made of possible coding approaches, and a "folding" encoder was selected for development. This coder consists of tandem operational amplifiers (one for each digit) where each operational amplifier has the input-output characteristic shown in Fig. 10. The transfer characteristic has a slope of precisely $+2$ for negative input signals and a slope of precisely -2 for positive input signals. The coder operates directly in the reflected binary or Gray code. The digits are obtained from each stage — a zero if the gain (slope) is in the $+2$ state and a one if the gain is in the -2 state. The performance of such a coder is limited by the fundamental accuracy (static and dynamic) with which the input-output characteristic can be established.

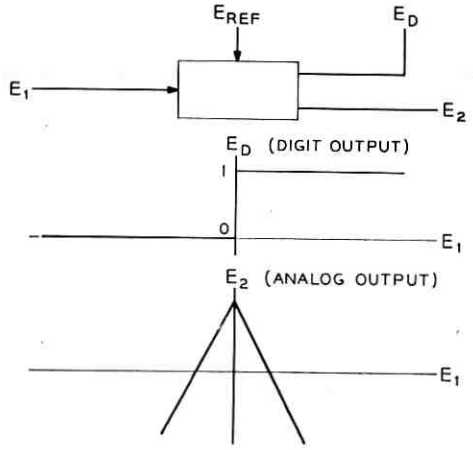


Fig. 10 — Characteristic of one stage of folding coder.

The basic technique for realizing the characteristics of Fig. 10 uses the operational amplifier with nonlinear feedback shown in Fig. 11. It is seen that positive input currents are routed to the E_B output and negative input currents are routed to the E_A output. The E_D output undergoes an abrupt transition as the input current goes through zero and provides a convenient point for extracting the coded digit. Various techniques may be applied to combining E_A, E_B, and a reference so as to produce the desired characteristic of Fig. 10.

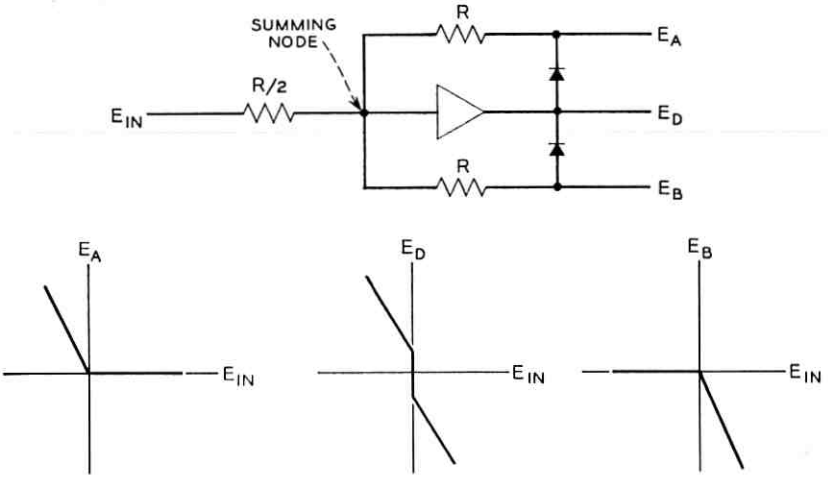


Fig. 11 — Realization of characteristic for solid-state coder.

Precise 9-digit encoding of television signals requires operational amplifiers that settle to an accuracy of one part in several thousand in a few nanoseconds. This level of performance is presently achievable, and the folding coder approach is a sufficiently satisfactory solution to the precise high-speed coding problem that further development of the coding tube does not appear warranted. The solid-state coder is more economical than the beam tube coder (for the production rates envisioned) and, in addition, is much smaller, and does not require high voltage power supplies. There are other potential advantages such as reliability, but it remains to be proven that these advantages exist in reality.

Both the beam tube coder and the solid-state coder have been used for coding mastergroup and television signals. The resulting coder assemblies are shown in Figs. 12 and 13. Both arrangements require precise sample-and-hold circuits and equipment to convert the parallel code available from the coder to a serial line code. The television coder also includes the framing pattern generator which forces the least-significant digit of each ninth code word to an ON-OFF pattern. The mastergroup coder transmits no special digits for identification of word framing.

4.2 *Decoder*

A standard resistor ladder network decoder is employed for decoding of broadband signals. Precise reference currents are gated into a resistive ladder network with 6-dB attenuation between current injection points. Extreme care must be exercised in designing the decoder, how-

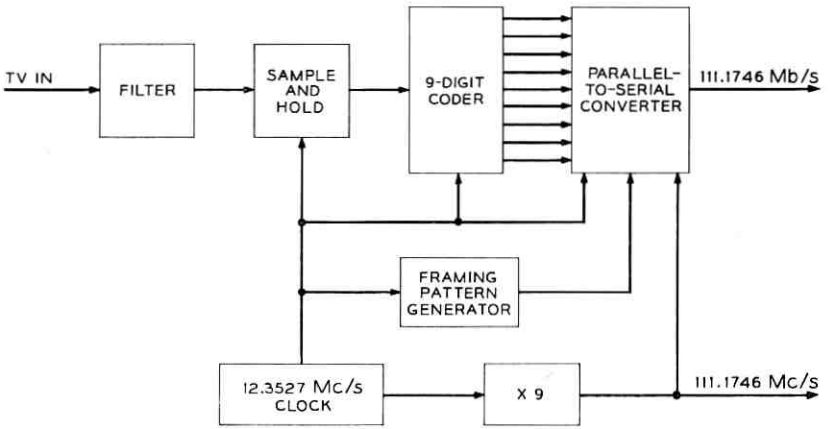


Fig. 12 — Television coder.

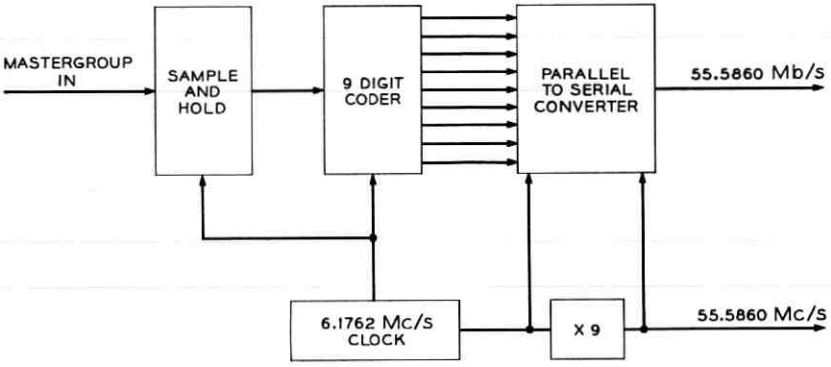


Fig. 13 — Coder for FDM mastergroups.

ever, to prevent the digital control signals from crosstalking into the analog output. Also, precise broadband resistors with end-of-life tolerance of better than 0.1 per cent are required for 9-digit decoders. Broadband precision resistors for the experimental terminals have been realized by the use of nitrided tantalum thin-film. Also, low-capacitance, negligible-storage-time diodes are required for gating of the reference currents. Both gallium arsenide and hot-carrier silicon diodes have served this purpose. Resulting decoder arrangements for television and mastergroups are shown in Figs. 14 and 15. Note that the transmitted code must be converted from Gray to binary, the serial train converted to parallel form, and the output of the decoder must be resampled to remove "spikes" generated during the time of change

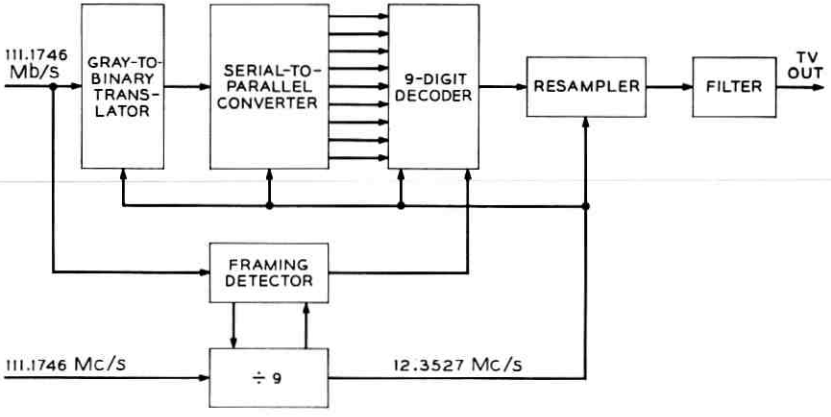


Fig. 14 — Television decoder.

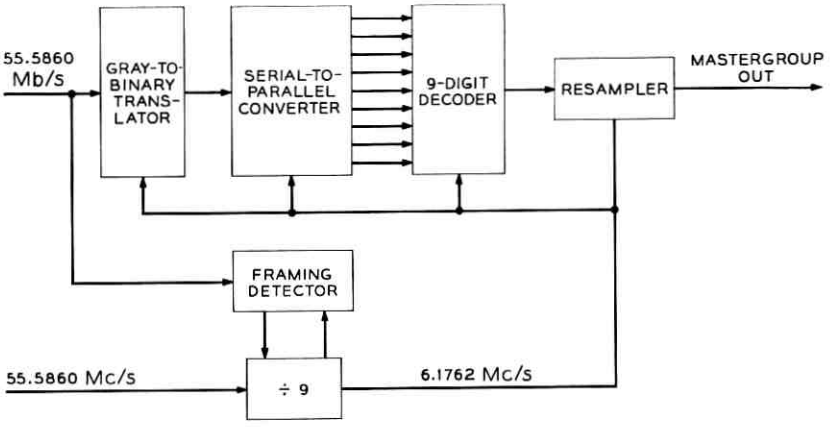


Fig. 15 — Decoder for FDM mastergroups.

of code. In the television decoder, the forced least-significant digit must be searched out to lock the decoder timing so as to properly group incoming words. In the case of the FDM decoder, word framing is accomplished by searching out the second digit in the Gray code, a digit which has a much higher probability of being a ONE than any of the other digits.

4.3 *High-Speed Multiplexer-Demultiplexer*

The multiplexer for combining the various low-speed trains into a 224-Mb/s pulse stream, and the associated demultiplexer are organized in such a way as to accommodate the various options required for a given signal package. Television equipment may be replaced by the equipment required for two mastergroups, and mastergroup equipment may be replaced by equipment to handle 36 T1 line signals.

In the experimental terminals the television signal is handled synchronously, and the repeatered line clock is derived from the video signal. This is accomplished by a two-bit elastic-store, phased-locked-loop arrangement shown in Fig. 16 as the video gap inserter. This unit not only derives the appropriate line frequency but also provides momentary storage for the video signal while the multiplex framing pulse is being transmitted.

Sampling clocks for the T1 line signals and the coded mastergroup operate independently of the 224-Mb/s line rate. The frequency difference is made up by the pulse stuffing technique previously described. The sync signal transmitter generates the basic pulse format

for the data link interconnecting the multiplex and demultiplex. This data link is time shared among all asynchronous inputs. The output of the sync signal generator is a 1.544-Mb/s signal which is multiplexed into the transmitting bit stream. This 1.544-Mb/s data link signal is made up of repeating sequences, and each sequence begins with a redundantly coded marker code. The first three time slots after the marker code are devoted to the first signal component to be synchronized, the second three time slots to the second component, etc. When an input signal train is to be stuffed, three ones are transmitted in the appropriate time position. Receipt of two or more ones in the three time slots assigned to a given signal signifies stuffing has taken place, and the receiver drops out a time slot in a predetermined position, which has been arranged to be precisely the position of the stuffed slot.

The demultiplex equipment (Fig. 17) operates in the manner similar to the multiplexer. There is a frame detector that searches for the 224-Mb/s framing pulse and locks the receiving clock to the appropriate phase. The sync signal receiver detects the synchronizing marker code and demultiplexes the stuffing commands so as to provide the proper

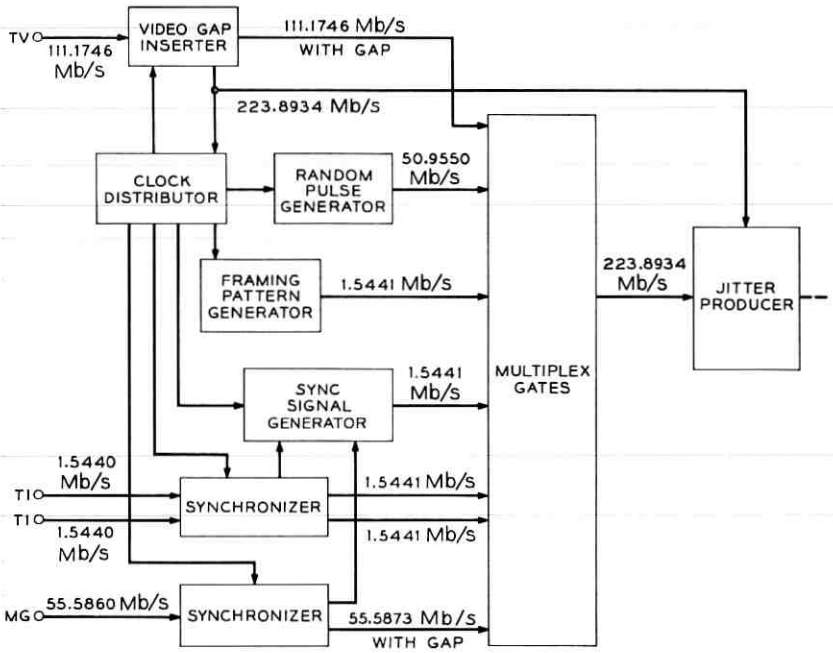


Fig. 16 — High-speed multiplex.

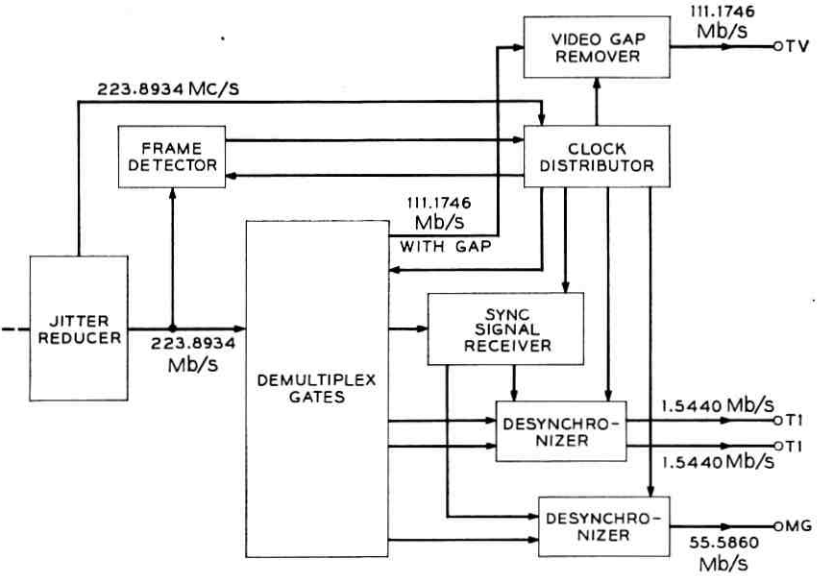


Fig. 17 — High-speed demultiplex.

write-inhibit signals to the elastic stores of the desynchronizers. The video gap remover closes the time gap introduced into the video signal by the multiplexer and thus delivers the same bit sequence to the television decoder as had been generated by the television coder.

A large portion of the multiplexer-demultiplexer circuits operate at high bit speeds. For rates up to about 120 Mb/s, logic is performed by high-frequency planar silicon transistors — often operating in the current routing or pulse routing mode.[22] The 224-Mb/s speeds have been achieved by using tunnel diodes and tunnel diodes in combination with a high-frequency germanium transistor ($f_t \approx 2.5$ Gc/s). Both silicon point-contact and hot-carrier diodes have been used with success in gates of fractional nanosecond rise time. Care has been exercised in the layout of circuit blocks and in construction techniques. All circuit functions have been realized by conventional construction techniques augmented by the use of a few integrated circuits and thin-film components. Of course, stripline and coaxial interconnections are frequently used.

4.4 *Jitterizer and Dejitterizer*

High-speed jitter reduction equipment of the type shown in Fig. 7 has been used to remove the effects of pattern jitter. Eight storage

slots were provided in the elastic delay, and an effective Q as high as 10^7 has been realized in the design of the phase-locked loop, which also exhibits satisfactory lock-in, pull-out, and static phase shift performance.

A second dejitterizer was constructed and operated as a jitter producer. By introducing bandlimited noise in the phase-locked loop, it is possible to simulate the expected jitter performance of a repeatered line. This has been used as a vehicle for observing signal impairments resulting from jitter as well as demonstrating the ability of the dejitterizer to remove the expected amounts of pattern jitter.

V. SYSTEM PERFORMANCE

Companion papers give detailed design and performance information for the various parts of the experimental terminal.[23,24] It has been established that signals transmitted through the experimental terminals meet Bell System transmission objectives for long-haul systems, and in a signal deterioration sense, a commercial design need not perform at a significantly higher level than has been achieved in the experimental equipment. The beam-tube encoder has been used at both television and mastergroup speeds and more than meets the objectives that have been tentatively set for in-service performance of a broadband codec. The solid-state coder also meets expected performance level requirements when coding mastergroups. However, the quantizing noise falls 2 dB short of present objectives when coding television. Performance is currently being improved, but even at the present performance level one should be able to operate at least four codecs in tandem and yet meet transmission objectives for television.

It has been shown that the pulse stuffing technique (for synchronization) is an effective solution to the synchronization problem. Performance of the T1 carrier terminal and the mastergroup codec when transmitting over the 224-Mb/s line has been shown to be independent of the precise sampling frequency, and these frequencies may drift over a range which is quite adequate for run-of-the-mill crystal oscillators. This system is also capable of simulating the jitter expected in 4000 miles of repeatered coaxial line, and, as a result of the action of the dejitterizer, there is no noticeable signal impairment introduced by the expected amounts of jitter.

The experimental terminal utilizes approximately 1300 transistors, and many of these are operated at nanosecond speeds. In spite of the serious problems presented by high-speed, high-accuracy circuits, all these devices have been satisfactorily interconnected, and the result-

ing system has proven to be quite reliable. It has operated for approximately a year without major difficulties, and its operation has been satisfactorily demonstrated on numerous occasions.

VI. CONCLUSIONS

Experimental equipment has demonstrated the technical feasibility of the transmission of high-quality broadband signals by means of pulse code modulation. The experimental terminals demonstrate a satisfactory solution to all fundamental system problems. Signal impairment introduced by the experimental terminals does not differ significantly from analytical predictions, nor does the signal impairment differ significantly from that expected from a commercial system.

VII. ACKNOWLEDGMENTS

The work reported herein has been carried out by the PCM Terminal Department under the direction of the author. Early work of a group under C. W. Rosenthal, the support of Systems Engineering and Device Development, and the guidance of R. A. Kelley are gratefully acknowledged. The work is, of course, based on foundations established long ago by the Research Department.

REFERENCES

1. Davis, C. G., An Experimental Pulse Code Modulation System for Short-Haul Trunks, B.S.T.J., *41*, Jan., 1962, pp. 1–24.
2. Bender, W. G., An Experiment in PCM Transmission of Multiplexed Channels, Bell Laboratories Record, July/August, 1964, pp. 240–246.
3. Carbrey, R. L., Video Transmission over Telephone Cable Pairs by Pulse Code Modulation, Proc. IRE, *48*, Sept., 1960, pp. 1546–1561.
4. Gray, J. R., and Kitsopoulos, S. C., A Precision Sample-and-Hold Circuit with Subnanosecond Switching, IEEE Trans. Circuit Theor., *CT-11*, Sept., 1964, pp. 389–396.
5. Bruce, R. A., Optimum Pre-Emphasis and De-Emphasis Networks for Transmission of Television by PCM, IEEE Trans. on Commun. Tech., *COM-12*, Sept., 1964, pp. 91–96.
6. Hallenbeck, F. J., and Mahoney, J. J. Jr., The New L Multiplex — System Description and Design Objectives, B.S.T.J., *42*, March, 1963, pp. 207–221.
7. Mayo, J. S., An Experimental Broadband PCM Terminal, Bell Laboratories Record, May, 1964, pp. 152–157.
8. Hoth, D. F., The T1 Carrier System, Bell Laboratories Record, Nov., 1962, pp. 358–363.
9. Mayo, J. S., and Trantham, R. J., Statistical Framing of Code Words in a Pulse Code Receiver, U.S. Patent No. 3,175,157, 1965.
10. Gray, J. R., and Pan, J. W., Using Digit Statistics to Word-Frame PCM Signals, B.S.T.J., *43*, Nov., 1964, pp. 2985–3007.
11. Byrne, C. J., and Scattaglia, J. V., A Buffer Memory for Synchronous Digital Networks, Sixth Mil-E-Con Convention Record, 1962.
12. Geigel, A. A., and Witt, F. J., Elastic Stores in High-Speed Digital Systems, NEREM Record, 1964.

13. Runyon, J. P., Reciprocal Timing of Time-Division Switching Centers, U.S. Patent No. 3,050,586, 1962.
14. Graham, R. S., Pulse Transmission System, U.S. Patent No. 3,042,751, 1962.
15. Mayo, J. S., PCM Network Synchronization, U.S. Patent No. 3,136,861, 1964.
16. Byrne, C. J., Properties and Design of the Phase-Controlled Oscillator with a Sawtooth Comparator, B.S.T.J., *41*, March, 1962, pp. 559–602.
17. Rowe, H. E., Timing in a Long Chain of Regenerative Binary Repeaters, B.S.T.J., *37*, Nov., 1958, pp. 1543–1598.
18. Aaron, M. R., PCM Transmission in the Exchange Plant, B.S.T.J., *41*, Jan., 1962, pp. 99–141.
19. Byrne, C. J., Karafin, B. J., and Robinson, D. B., Jr., Systematic Jitter in a Chain of Digital Regenerators, B.S.T.J., *42*, Nov., 1963, pp. 2679–2714.
20. Bennett, W. R., Statistics of Regenerative Digital Transmission, B.S.T.J., *37*, Nov., 1958, pp. 1501–1542.
21. Cooper, H. G., Crowell, M. H., and Maggs, C., A High-Speed PCM Coding Tube, Bell Laboratories Record, Sept., 1964, pp. 266–272.
22. Koehler, D., A 110-Megabit Gray Code to Binary Code Serial Translator, Int. Solid-State Circuits Conf. Digest, Feb., 1965, pp. 84–85.
23. Witt, F. J., An Experimental 224 Mb/s PCM Multiplexer-Demultiplexer Using Pulse Stuffing Synchronization, B.S.T.J., This Issue, pp. 1843–1885.
24. Edson, J. O., and Henning, H. H., Broadband Codecs for an Experimental 224 Mb/s PCM Terminal, B.S.T.J., This Issue, pp. 1887–1940.

An Experimental 224 Mb/s Digital Multiplexer-Demultiplexer Using Pulse Stuffing Synchronization

By F. J. WITT

(Manuscript received July 27, 1965)

Solid-state device and circuit technology has advanced to the point where processing of digital signals with bit rates as high as 224 Mb/s may be accomplished. As a specific demonstration of this fact, an experimental multiplexer-demultiplexer has been developed which combines the following signals into a 224 Mb/s binary pulse train for transmission over a digital transmission network and which furnishes the necessary processing to reconstitute the original signal components:

i) A PCM-coded commercial color video signal (111.2 Mb/s)

ii) A PCM-coded FDM mastergroup signal (55.6 Mb/s)

iii) Two T1 carrier (24 voice channel TDM) signals (1.544 Mb/s each)

iv) Word generator signals to occupy the remaining time slots.

The line bit rate is derived from and is therefore synchronous with the coded video signal; however, the coded mastergroup and T1 carrier signals are derived from independent clocks. Synchronization of the latter two signal types has been achieved through the use of pulse stuffing synchronization with added-bit synchronization signaling.

The ability of a 224 Mb/s buffer memory coupled with a phase-locked loop to attenuate adequately the timing jitter which accumulates in a long digital transmission network has also been demonstrated.

All of the experimental results have indicated that the realization of a commercial high-speed digital multiplexer-demultiplexer is feasible.

I. INTRODUCTION

Elsewhere in this issue is a description of an experimental digital transmission terminal which has been used to demonstrate, among other things, the feasibility of multiplexing a variety of asynchronous* digital

* The term asynchronous as used here means a lack of synchronism both among the various signals being multiplexed and with the multiplexing clock. The individual digital signals taken separately are synchronous following the normal usage of the word in data transmission in that pulses occur at regular clock intervals.

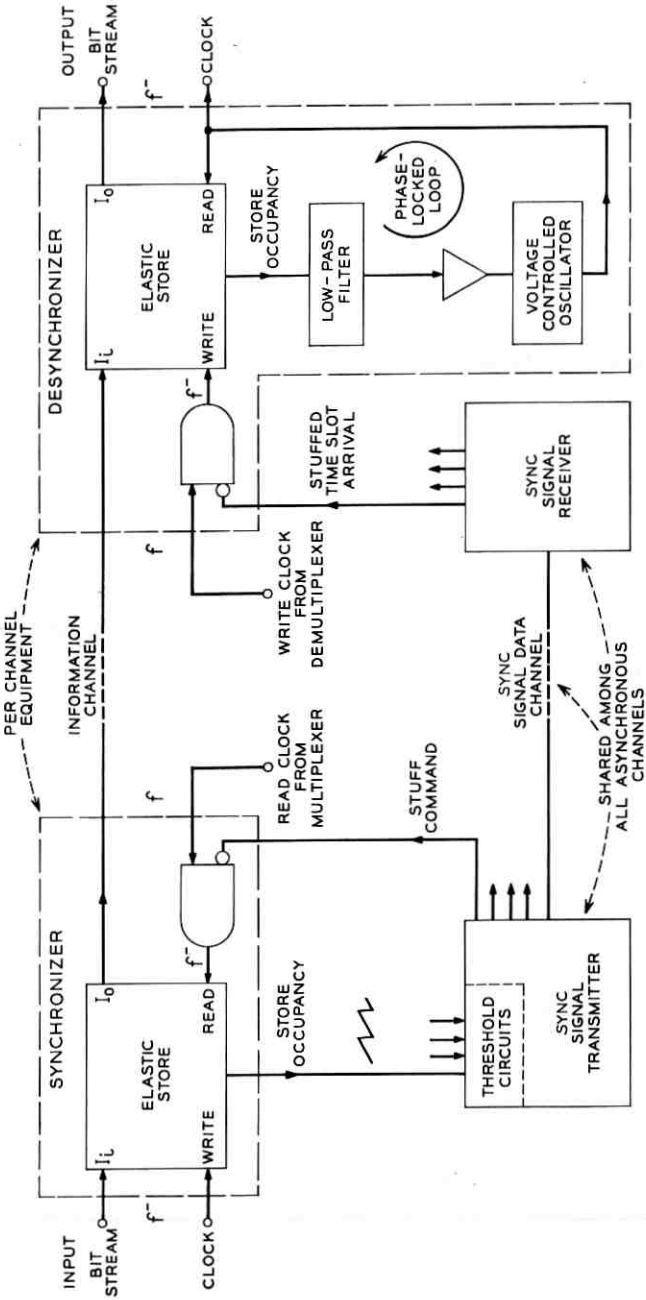


Fig. 1 — Pulse stuffing synchronization.

signals (PCM-video, PCM-SSB-FDM, data, coded *PICTUREPHONE** signals, and lower bit-rate PCM) into a 224 Mb/s binary pulse train.[1] This paper gives a more detailed description of the techniques used for multiplexing and demultiplexing, synchronization and timing jitter reduction. Also, a number of the high-speed digital circuit techniques which were employed are disccused.

II. PULSE STUFFING SYNCHRONIZATION

A variety of solutions have been proposed for the problem of synchronizing a large digital transmission network.[1] The method used in this system is that of pulse stuffing.[2,3] In concept, the technique is simply to constrain the bit rates of all multiplexed asynchronous signals to be slightly less than the bit rates needed at the multiplexer. These asynchronous signals are each passed through processing circuits, called "synchronizers," which sense the amount by which the bit rates must be increased to be synchronous with the transmitted clock. A buffer memory in each synchronizer, called an elastic store,[4,5,6] allows time slots to be added at a rate equal to the difference between the synchronous and asynchronous bit rates;† this difference is referred to as the stuff rate. The position of the stuffed time slots is signaled to the demultiplexer over a data channel which contains the synchronization information for all asynchronous channels. At the receiving terminal, the data channel is decoded, and the stuffed time slots are removed from the various signals which had been synchronized by "desynchronizers." This synchronization by digital signal processing is independent of the format of the signals which are multiplexed, and it achieves complete integrity of the information pulse streams. The only change which a digital signal undergoes is the addition of some timing jitter, and it will be shown that this jitter can be held to a sufficiently small amplitude so that the performance is quite suitable for application to a commercial system.

A simplified block diagram showing synchronization by pulse stuffing is shown in Fig. 1. The incoming pulse train at bit rate f^- is written into the synchronizer elastic store using the associated timing for a "write" clock. The store is read with a "read" clock of frequency f which is slightly higher than f^- and synchronized with the transmitting clock. By means of a phase detector, the elastic store furnishes an output which is

* A service mark of the American Telephone and Telegraph Company.

† Asynchronous bit rate and synchronous bit rate are terms defined here to describe the bit rates of the signal before and after it has been processed by the synchronizer.

proportional to the occupancy of the store. Since f is greater than f^- the store would become depleted, but before this happens a threshold circuit in the "sync signal transmitter" causes an inhibition of the read clock for one time slot. This action enables the elastic store to "recover" and results in an output time slot which corresponds to none of the input time slots, i.e., contains no message information. The sync signal transmitter also transmits over a data channel the position of the "stuffed" time slot.

At the receiving end of the system, the sync signal receiver decodes the signal from the data channel and inhibits the writing of the stuffed time slots into the desynchronizer elastic store. After inhibition, the bit rate of the signal is f^-, the asynchronous bit rate, but the timing contains abrupt phase discontinuities one time slot in amplitude. These phase discontinuities are reduced in amplitude by using a phase-locked loop to provide a read clock which is a smoothed version of the write clock.

III. FORMAT SELECTION*

Selection of the digital line bit rate was based on the bit rate of the digital signal components and the economic and technical aspects of regenerative repeater design. The signals which played a dominant role in the format selection were the PCM mastergroup (600 SSB-FDM voice channels) (~3 mc bandwidth × 2 samples/cycle × 9 bits/sample = 54 Mb/s), the PCM color video signal (~6 mc bandwidth × 2 samples/cycle × 9 bits/sample = 108 Mb/s)[2] and the T1 carrier 24-channel PCM signal (1.544 Mb/s).[7] A composite bit rate was chosen which would allow the multiplexing of four PCM mastergroups or two PCM color video signals or combinations of these and other signals. A bit rate slightly in excess of 216 Mb/s, but a multiple of 1.544 Mb/s, would seem appropriate. There are, however, other factors, including multiplex framing and synchronization signal signaling, which influence the bit rate selection.

3.1 *Multiplex Framing*

Of the several ways in which the various line signal components can be identified at the demultiplexer, the method of added-bit framing was used because it places very few restrictions on line format and it is relatively simple to implement. This technique, similar to the one used in the T1 system, allocates one time slot every 145 high-speed time slots to

* For a more detailed description of the line format selection, see Ref. 1.

an alternating ONE-ZERO pattern.* Thus, a "multiplex frame" is 145 high-speed time slots long. This selection yields a further system simplification: The multiplex frame rate is chosen to be the synchronous T1 carrier bit rate [(1.544 Mb/s) + δ ppm†]. The quantity δ is added to allow for the frequency offset required for pulse stuffing of T1 carrier signals.‡ The synchronous coded mastergroup and coded video bit rates are [(1.544 × 36 = 55.584) Mb/s + δ ppm] and [(1.544 × 72 = 111.168) Mb/s + δ ppm], respectively. The line bit rate becomes [(1.544 × 145 = 223.880) mc + δ ppm]. Fig. 2 shows several ways in which the 224 Mb/s line can be loaded. Bit interleaving is used because it simplifies the processing circuitry and allows the use of relatively small buffer memories.

The time slots for all signals to be multiplexed are uniformly spaced. Therefore, a phase discontinuity must be introduced into any signal which is not arriving at the frame rate (or some submultiple of it) in order to accommodate the framing pattern. This discontinuity will be referred to as the framing gap and, expressed in time, it must be precisely one high speed time slot in amplitude. To generate the framing gap in the information pulse stream, some form of memory is required.

The basic technique for accomplishing framing gap insertion and, at the same time, synchronizing the line frequency to a separate clock, is shown in Fig. 3. In the case shown, the line bit rate is synchronized to the incoming bit rate of a video pulse train which is to be multiplexed onto the line. Basically, three distinct functions are accomplished by the circuit shown. First, the framing gap is formed in the multiplexer clock circuits by means of the divide-by-145 counter containing an inhibit gate. Secondly, the multiplexer clock is synchronized to the video coder clock by means of a phase-locked loop. Lastly, in order to establish the phase discontinuity without errors in the information pulse train, a two-bit elastic store is used. Write and read clock signals for the store have a frequency of one-half the video signal bit rate since an n-bit store requires clock signals at $1/n$th of the bit rate. The reason for use of this frequency will become more apparent in Section 5.2.1 where the elastic store is described in more detail.

An explanation of the gap generation follows: Assume that the output

* A more complicated pattern could have been used at some expense in circuit complexity, but the reframe time performance with the simple scheme described above is adequate.

† ppm = parts per million.

‡ In the experimental terminal, the nominal offset δ for T1 carrier signals was chosen to be 60 ppm. Note that the offset need not be the same for other services; in fact, for the coded mastergroup, it was chosen to be 24 ppm, i.e., the nominal bit rate for the incoming coded mastergroup is [(1.544 × 36) Mb/s + (60 − 24)ppm] = [55.584 Mb/s + 36 ppm].

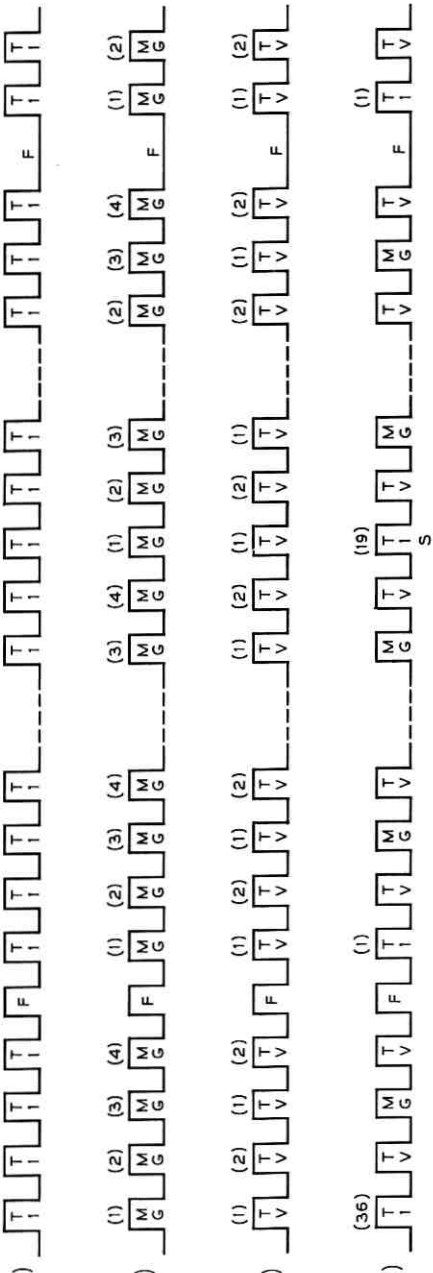


Fig. 2 — Some high-speed digital line loading possibilities.

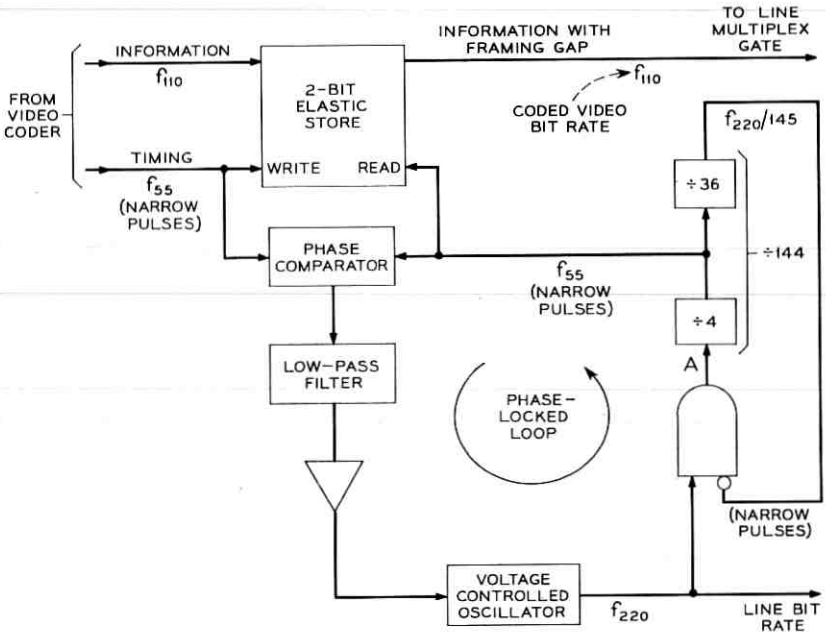


Fig. 3 — Framing gap insertion.

signal of the VCO is at the desired line bit rate. Entry of every 145th pulse into the divide-by-144 circuit is prevented by the inhibit gate. The net result is that at point A in Fig. 3 a pulse train exists which occurs at the line bit rate, but has every 145th pulse missing — thus, the framing gap. Part way through the counter (after division by four) the desired read clock rate is available with the framing gap. The phase-locked loop forces the average pulse repetition rate of the write and read clocks to be the same, and it synchronizes the line bit rate to the video bit rate.

The alternating ONE-ZERO pattern, which is placed in the framing gaps, yields a maximum average reframe time of 188 μs. This theoretical result is based on an assumption that the 144 information time slots contain a random pattern and that, in the search mode, the time slots are examined on a one-per-frame basis. To the above figure must be added a "flywheel effect" which results from the fact that, in order to control the misframe rate, the "frame detector" has been designed so that it does not begin searching until several framing pattern errors have been observed. The flywheel effect, adjusted for moderate line error situations (error rate $<10^{-6}$), adds a negligible amount to the reframe time.

3.2 *Synchronization Signaling*[8]

As mentioned in Section II, the signaling which identifies the inserted synchronizing time slots is sent over a data channel. A variety of methods exists for establishing the synchronization signaling data channel. To achieve independence of the format of the signals which are being synchronized, however, a separate time slot per frame is devoted to sync signaling. It will be seen that the information handling capability of this channel (1.544 Mb/s) is more than adequate to handle the synchronization of a full complement of asynchronous channels on the high-speed line. Because this channel contains data which is needed by all synchronized signal channels, certain steps were taken to make it relatively immune to line errors and to provide rapid recovery in the event multiplex framing, and thus the sync signaling channel, is lost. To these ends, within the sync signaling channel, stuff occurrence is redundantly coded and a code highly immune to errors is used for sync signal frame identification. Another feature of the format chosen is the one-for-one correspondence between stuff signal organization and line organization; hence, this organization lends itself to convenient dropping and adding of channels at points along a route.

In an operational system patterned after the experimental system described herein, the signaling pulses might occupy a single time slot per frame located midway between multiplex framing time slots and would be called "added-bit signaling." Thus, the sync signaling bit rate would be 1.544 Mb/s + δ ppm. For simplicity, however, the sync signaling for the experimental system was transmitted over the 73rd time slot of the frame and appears as a single T1 carrier channel. This time slot is identified by an "S" in Fig. 2(*d*), and the sync signal will be referred to as the "S bit" in what follows.

Fig. 4 shows the S-bit format. The C words (000 when no stuffing has occurred and 111 when it has) redundantly signal the presence of inserted time slots. At the demultiplex point, two out of three ONES are interpreted as a stuff indication. When an all-T1 carrier loading is used,

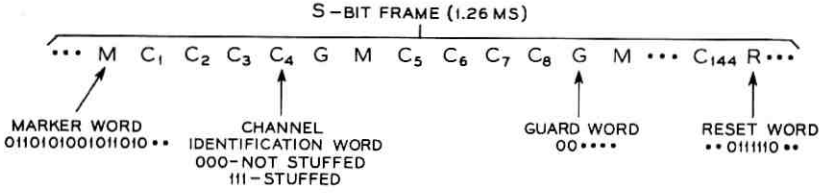


Fig. 4 — S-bit format (simplified).

each C word of an S-bit frame is allocated to a corresponding T1 carrier channel; for each mastergroup, 36 C words per S-bit frame are used, and for video, 72.

A key part of the sync signaling is the M word which is used to frame the S bit. It is the 16-bit word shown in Fig. 4 which has the characteristic that it can be identified exactly in an S-bit pulse train even though two errors exist anywhere within it. In many cases, more than two errors can exist and identification is still unique. An M word is inserted every fourth C word to allow for rapid reframing of the S bit for the higher bit rate synchronized channels. Reframing of the S bit for T1 carrier synchronization can take as long as a complete S-bit frame since the reset code, R, must be decoded to allow proper C-word association. The guard words, G, are strings of ZEROS which allow a simplification of the M word.

The format for the S bit as described above is inefficient in bandwidth usage because the statistical variations of the stuff rates used cause the C words to be 000 much of the time. A more efficient queuing system which signals stuffing by means of a channel address could have been used, but such a system would not yield the easy drop-add capability of the cyclic system described. The sync signaling for each digital channel is readily identifiable, and replacement would be a simple operation at drop-add points.

A disadvantage of the cyclic system is the waiting period between the time stuffing is demanded and the time it occurs. Since stuff signaling for a T1 carrier channel occurs once per S-bit frame (which is 1944 T1 carrier time slots long), the waiting period for T1 carrier can be as long as 1.26 msec, and hence, the maximum allowable T1 carrier stuff rate equals 794 c/s. Since one C word is allotted per line frame bit, the waiting time for higher bit rate constituents is proportionately less. For example, the maximum waiting interval for a video channel is 17.5 μs (1260/72 μs). The effect of this waiting interval is illustrated in Fig. 5. The phase of the signal at the output of the synchronizer (relative to the phase of the input signal) for the case where the stuff interval is large with respect to the waiting time is shown in Fig. 5(a). Note that it is approximately a sawtooth waveform of amplitude 2π (one time slot) and fundamental frequency equal to the stuff rate. Figs. 5(b) and 5(c) illustrate the nature of the output phase as the stuff rate increases: another component is added to that which would occur if demand stuffing could take place. This "waiting time effect" can include frequency components between dc and the stuff rate, and, as the stuff interval approaches the waiting time interval, the total jitter amplitude can be as great as 4π

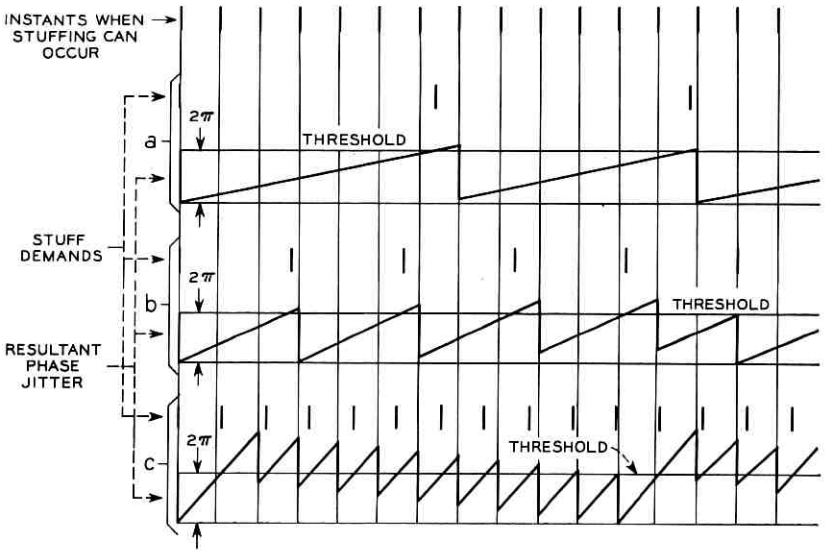


Fig. 5 — Pulse stuffing synchronization — waiting time effect.

(two time slots). Since the desynchronizer has a low-pass jitter attenuation characteristic, care must be exercised in the selection of the nominal stuff rate relative to the maximum stuff rate to assure that the residual output jitter after the desynchronizer is within tolerable limits.

Since the location of the inserted time slots is carried by the data channel, this method of synchronization is vulnerable to line errors. Analysis reveals, however, that misframes of an information channel due to sync signaling channel errors would never occur more frequently than several times per year for a line error rate of 10^{-6}. This performance could be improved, if necessary, through modification of the S-bit format; however, the format described is convenient to realize economically and is adequate in performance.

IV. PHASE JITTER REMOVAL

In a long digital transmission system there is an accumulation of phase jitter which arises from the dependence of the phase of the timing signal at each regenerative repeater on pulse pattern.[9] The contribution of each repeater is small, but the over-all effect can cause a significant transmission impairment.* Consequently, in a long system circuits

* See Ref. 1, this issue, p. 1827.

would be installed at intervals along the transmission route to reduce the accumulated jitter.

Fig. 6 is the block diagram of such a circuit, called a "dejitterizer," which consists of two closely associated parts: a phase-locked loop, which is driven by the jittered timing signal providing a timing signal with greatly reduced jitter, and an elastic store. After regeneration, the jittered line pulse train is sequentially written into the store using the jittered timing signal as a write clock. The smoothed timing is used as a read clock to produce a pulse train whose timing jitter has been substantially attenuated.

For the experimental system an eight-bit elastic store was developed. Tunnel diodes biased in a bistable mode proved to be highly satisfactory as storage elements. The effective store size is seven bits because the smallest spacing between write and read operations is about one-half bit at the line bit rate. Commutation was achieved by forming pulse trains at one-eighth the write and read clock rates and launching them down tapped delay lines. A ring counter could have been used to perform the function of the frequency divider and the delay lines, but it is a more expensive solution. It should be noted that application of delay line commutation is limited by the fact that misalignment of pulses at the AND gates occurs if phase jitter is present. This limitation may be expressed as follows for sinusoidal jitter with frequency $f_j \ll 1/(\mathrm{N} - 1)\tau$:

$$mf_j < \frac{M}{(N-1)\tau} \tag{1}$$

where

m = peak deviation of the phase jitter [radians]
f_j = jitter frequency [c/s]
M = allowable misalignment at the gates [time slots]
N = number of storage cells
τ = delay between taps [seconds].

For example, if $M = 0.1$ time slots for an eight-bit store operating at 224 Mb/s, then the maximum allowable phase jitter amplitude would be one radian at a jitter frequency of 3.2 Mc/s or ten radians at a jitter frequency of 320 kc/s. In the applications for which delay line commutation is used in the experimental system, the constraint expressed in inequality (1) is never limiting.

Since absolute phase information is not available at a repeater, the jitter attenuation characteristic must be a low-pass function, and very low frequency jitter will not be affected by the dejitterizer. Fortunately,

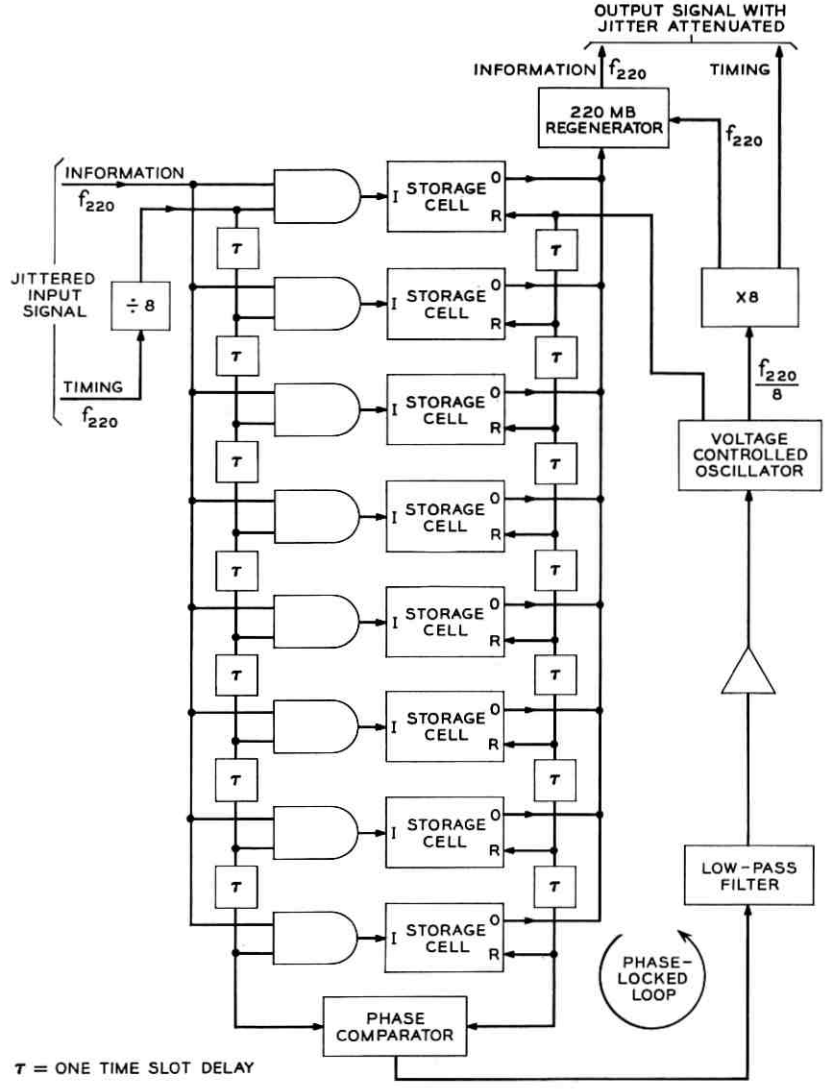


Fig. 6 — 224 Mb/s dejitterizer.

large amounts of low frequency jitter can be tolerated. Considerable flexibility exists for shaping the jitter attentuation characteristic through design parameters in the loop transmission function.[10] For all cases there is a tradeoff between the tolerable net oscillator instability (instability of line clock plus instability of the voltage controlled oscillator)

and the amount of store capacity allotted to frequency drift for a fixed jitter bandwidth of the dejitterizer. For the simplest case where there is no shaping network in the phase-locked loop (so-called no filter case), it may be shown that

$$(\delta_{VCO} + \delta_B) = \frac{B_j \cdot \pi n \cdot 10^6}{2\omega_B} = \frac{\pi n \cdot 10^6}{2Q} \tag{2}$$

where

δ_{VCO} = peak VCO frequency drift [ppm]
δ_B = peak line frequency drift [ppm]
B_j = jitter bandwidth [radians/sec]*
n = number of storage cells allocated to frequency drift
ω_B = line bit rate [radians/sec]
Q = quality factor of phase-locked loop considered as a bandpass filter.

As an example, for a net oscillator instability of 4 ppm and a Q of 10^6 (parameters used in the experimental system), a total of 2.5 storage cells would be allocated to frequency drift. If, however, the net oscillator instability were doubled, about five storage cells would be taken up by frequency drift alone.

The phase-locked loop is adjusted so that the elastic store is half full on the average, and the elastic store is made large enough so that overflow does not occur often enough to be a problem.

In order to examine the feasibility of a dejitterizer operating at a bit rate of 224 Mb/s, the eight-bit elastic store and associated phase-locked loop were designed to function as the last dejitterizer in a simulated long system. If it is assumed that the Chapman model† can be used for the regenerative repeaters,‡ and if the following assumptions are made:

Number of repeaters in system = 3600
Q of each repeater timing tank = 80
rms phase jitter per repeater = 11.2°§
One dejitterizer every 360 repeaters
No filter case with $Q = 10^6$ for each dejitterizer

then the spectrum of the jitter into the last dejitterizer would be that shown in Fig. 7.[11] This spectrum, approximated by the dashed curve,

* $B_j = 2\alpha_o$, where α_o = the loop gain [radians/sec/radian].
† Ref. 9, p. 2681.
‡ Insufficient experimental data exist to establish the validity of this model for a 224 Mb/s repeater, but experimental evidence has demonstrated its usefulness for estimating the jitter performance of strings of T1 carrier repeaters.[9]
§ Experimental models of regenerative repeaters have exhibited slightly better jitter performance than the 11.2° rms used here.

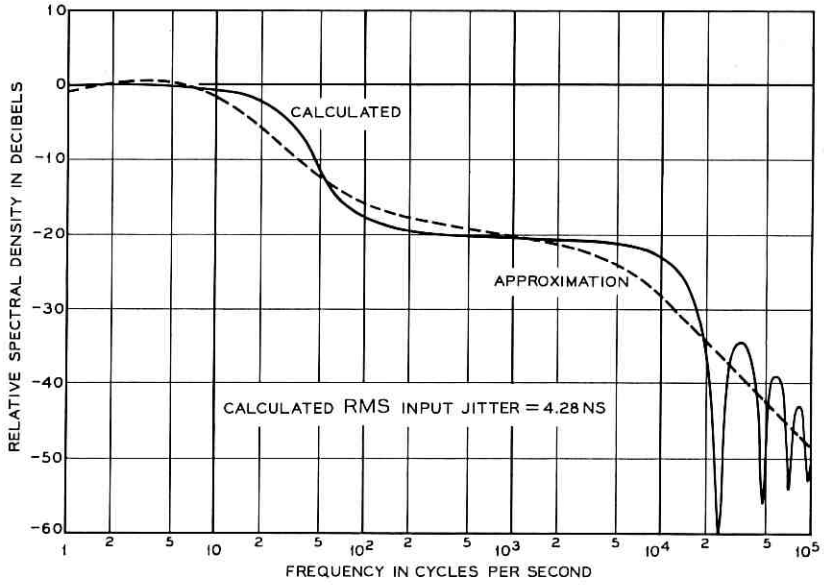


Fig. 7 — Spectrum of dejitterizer; input jitter.

was used to phase modulate the input signal of a dejitterizer having a $Q = 10^6$. The calculated and experimental results are shown in Fig. 8. No digital errors occurred during this experiment because the input phase deviation was limited (by clipping) to a level which prevented store overflow. An analysis has shown[11] that for a net oscillator instability of 4 ppm (used in the experimental system) and for the case described above, the effective size of store required for one overflow per minute, week and century would be 13, 15 and 17, respectively. Note how rapidly the overflow rate varies with store size. Since it has been determined that the required store size is not decreased rapidly as the number of dejitterizers is increased, it is expected that more than the eight cells used in the experimental dejitterizer would be needed in a commercial system. The store size required is still quite reasonable, however.

An indication of the effectiveness of the dejitterizer is provided by Fig. 9, where 1 kc/s sinusoidal jitter has been applied. That the effective store size is at least seven time slots is seen from the waveforms shown.

V. MULTIPLEXER-DEMULTIPLEXER DESCRIPTION

In order to demonstrate the feasibility of performing the variety of digital processing steps necessary to multiplex and demultiplex the

digital signals described earlier, the system shown in Fig. 10 was developed. The arrangement shown was chosen to allow the transmission of the following signal types over the 224 Mb/s digital line:

(1.) Synchronous PCM video.
(2.) Synchronous or asynchronous PCM mastergroup.
(3.) Two T1 carrier channels, synchronous with each other, but asynchronous with the line clock, which can be used separately or can be combined to handle a 3 Mb/s coded *PICTUREPHONE* signal. The individual T1 carrier channels also carried data-on-T1 carrier signals as one experiment.[12]
(4.) Pulses from a synchronous word generator to fill the unused time slots with a restricted set of pulse patterns which includes all-ZEROS and all-ONES.
(5.) The synchronous sync signaling signal (S bit).

One feature which was not demonstrated was the processing of an asynchronous PCM video signal; i.e., a 111 Mb/s synchronizer and desynchronizer was not developed. It is clear that development of these

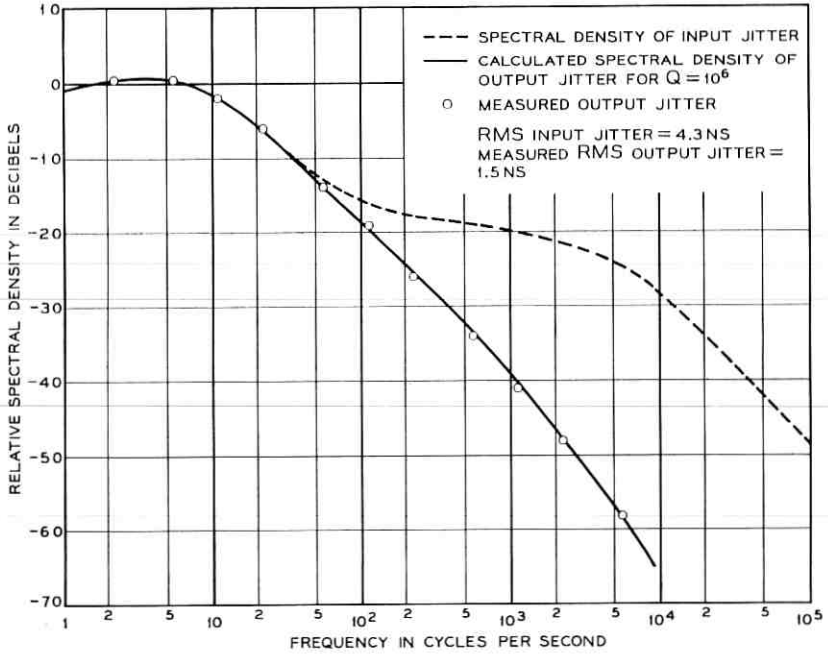


Fig. 8 — Spectrum of dejitterizer; output jitter.

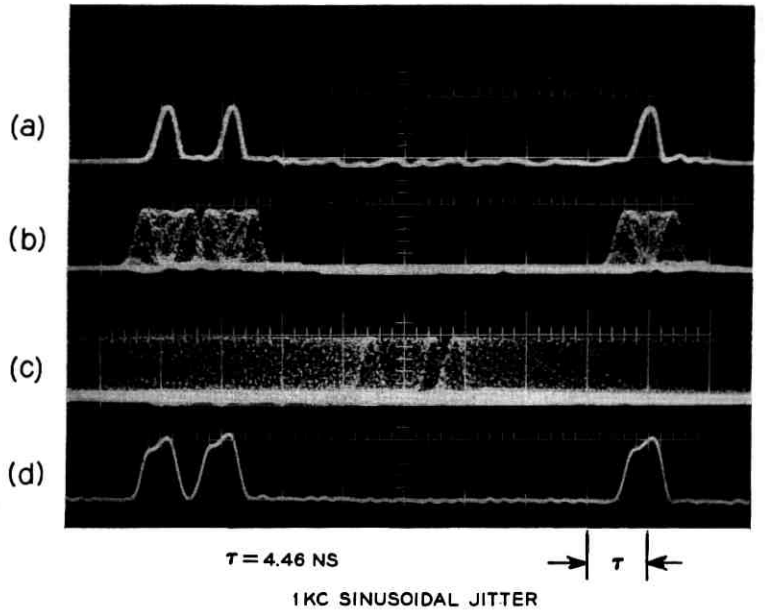


Fig. 9 — Performance of dejitterizer.

system blocks can be accomplished through the use of circuit design techniques which have been applied elsewhere in the experimental terminal. Also, the two T1 carrier channels were synchronized with each other. This was a requirement for processing a 3 Mb/s coded *PICTUREPHONE* signal,* and it was apparent that this simplification would not cause the omission of the development of a crucial circuit function. The format chosen highlights all of the significant developments needed for this study.

The line format is shown in Fig. 11 along with a timing diagram which defines all of the signals depicted in Fig. 10 and in the figures which follow. Table I defines the frequencies shown in Fig. 10. In the experi-

* Coding and decoding of the *PICTUREPHONE* signal into a 3 Mb/s digital signal was accomplished through the use of equipment developed within Bell Telephone Laboratories.

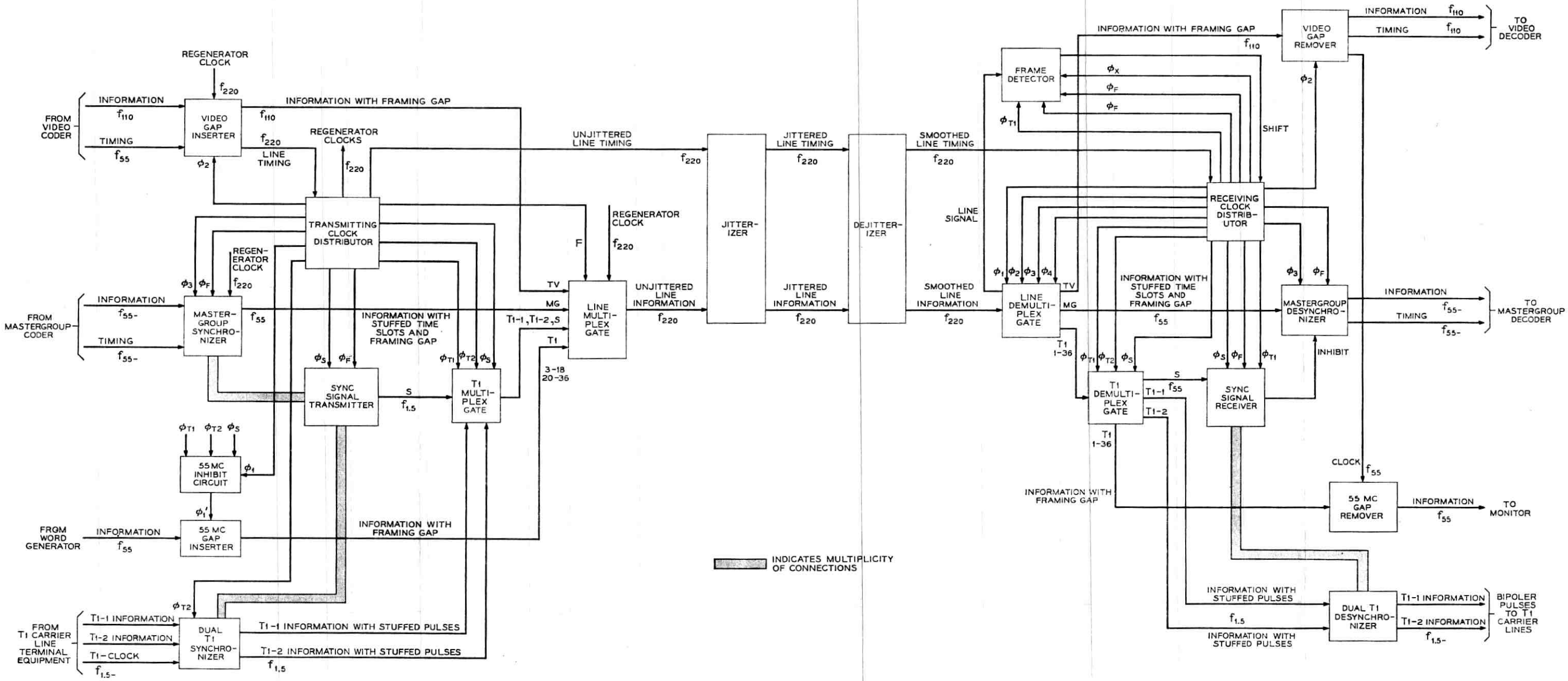


Fig. 10 — Block diagram of multiplexing and jitter portions of experimental terminals.

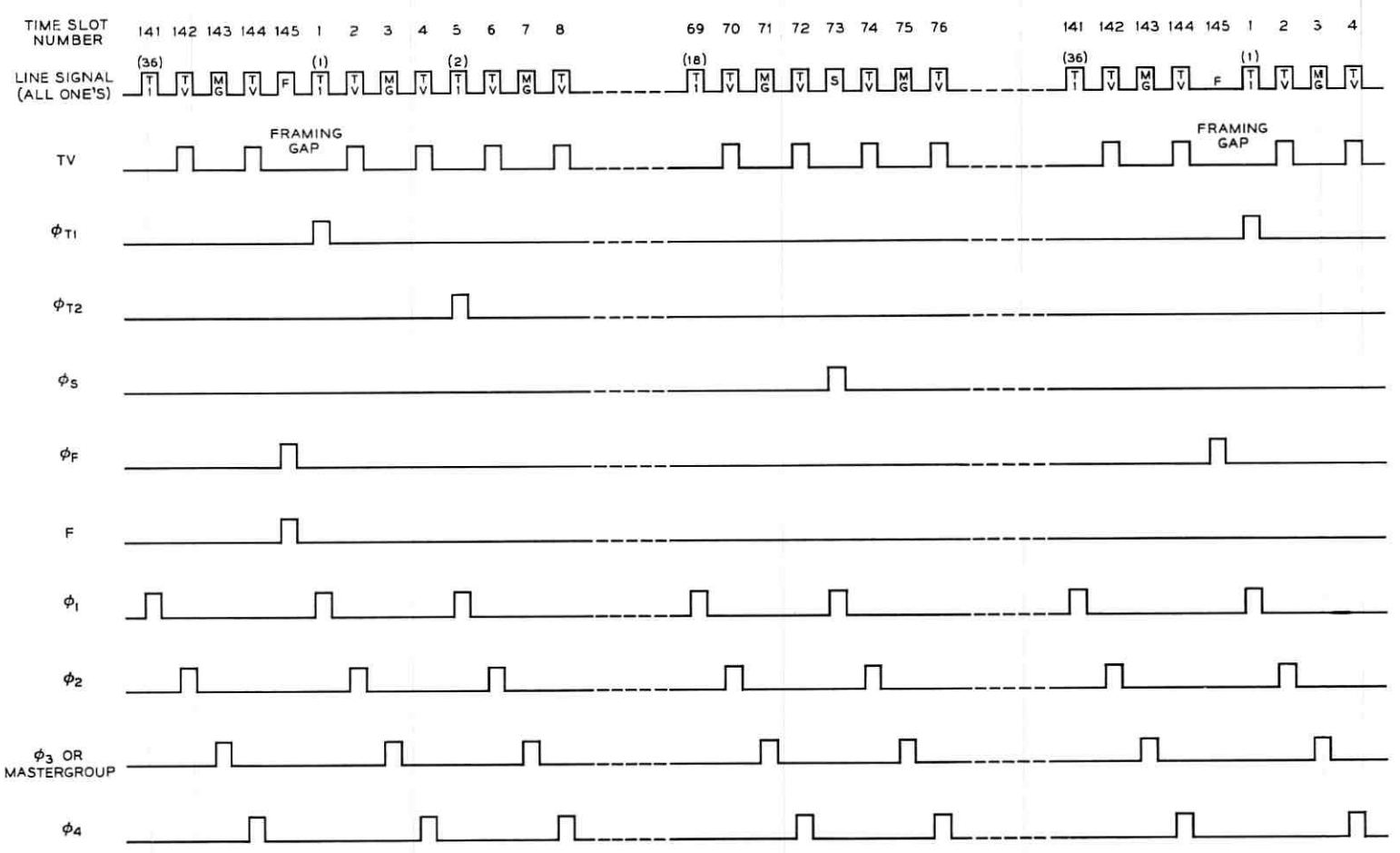


Fig. 11 — Time organization of the experimental terminal.

TABLE I — FREQUENCIES USED IN THE EXPERIMENTAL DIGITAL TRANSMISSION TERMINAL

$f_{1.5-}$ = Asynchronous T1 carrier bit rate	= 1.544 Mb/s ± 50 ppm
$f_{1.5}$ = Synchronous T1 carrier bit rate	= 1.544 Mb/s + (60 ± 2) ppm
f_{55-} = Asynchronous coded mastergroup bit rate	= (1.544 × 36 = 55.584) Mb/s + (36 ± 2) ppm
f_{55} = Synchronous coded mastergroup bit rate	= 55.584 Mb/s + (60 ± 2) ppm
f_{110} = Synchronous coded video bit rate	= (1.554 × 72 = 111.168) Mb/s + (60 ± 2) ppm
f_{220} = Line bit rate	= (1.544 × 145 = 223.880) Mb/s + (60 ± 2) ppm

mental terminal, the clock stabilities were taken to be ±2 ppm (readily attainable with a crystal in a simple oven) except for the T1 carrier clock which was assumed to be ±50 ppm.

Processing of the pulse trains from the various sources is required either to form the framing gap or to achieve a pulse train in synchronism with the line rate, or both. This processing is accomplished by the "video gap inserter," the "mastergroup synchronizer," the "55-Mc/s gap inserter," and the "dual T1 synchronizer."

The various sine wave and pulse clocks for the multiplex circuits are furnished by the "transmitting clock distributor." Synchronism between the line signal and the video bit rate is achieved by means of a phase-locked loop whose elements are portions of the video gap inserter and the transmitting clock distributor.

The sync signal transmitter monitors the mastergroup and dual T1 synchronizers for stuff-demand indications, clocks the stuffing operation, and signals with the S bit where and when stuffing has occurred.

Multiplexing of the S bit and each of the T1 carrier signals is accomplished in the "T1 multiplex gate." All signals are combined to form a binary line signal in the "line multiplex gate." *

The output of the line multiplex gate drives the "jitterizer" which is essentially the same circuit as the dejitterizer described in Section IV except that the phase-locked loop parameters are modified. The loop gain is increased substantially and the desired phase modulating signal is introduced at the output of the phase comparator.†

At the receiving end of the system the jittered signal is smoothed by the action of the dejitterizer. Frame synchronization of the "receiving clock distributor" is forced by the frame detector. The receiving clock distributor furnishes the variety of clocks required by the demultiplex

* Simultaneous use of signals from the 55-Mc/s gap inserter and the T1 multiplex gate is permitted even though their signals may coincide. The "55-Mc/s inhibit circuit" eliminates the pulses from ϕ_1 which fall in the time slots occupied by the T1 carrier signals and the S bit.

† Ref. 10, p. 571.

circuits. The "line demultiplex gate" separates the individual components of the line signal; as at the transmitting end, the S bit and the T1 carrier signals pass through another level of multiplex — the "T1 demultiplex gate." Framing gap and/or stuffed pulse removal is accomplished in the "video gap remover," the "mastergroup desynchronizer," the "55-Mc/s gap remover" and the "dual T1 desynchronizer." Decoding of the S bit is done by the sync signal receiver which signals the mastergroup and dual T1 desynchronizers exactly which time slots are to be dropped.

5.1 *Clock Distribution*

5.1.1 *Transmitting Clock Distributor*

A block diagram of this circuit is shown in Fig. 12. Its function is to derive from a sinusoidal line clock signal all of the clock pulse patterns

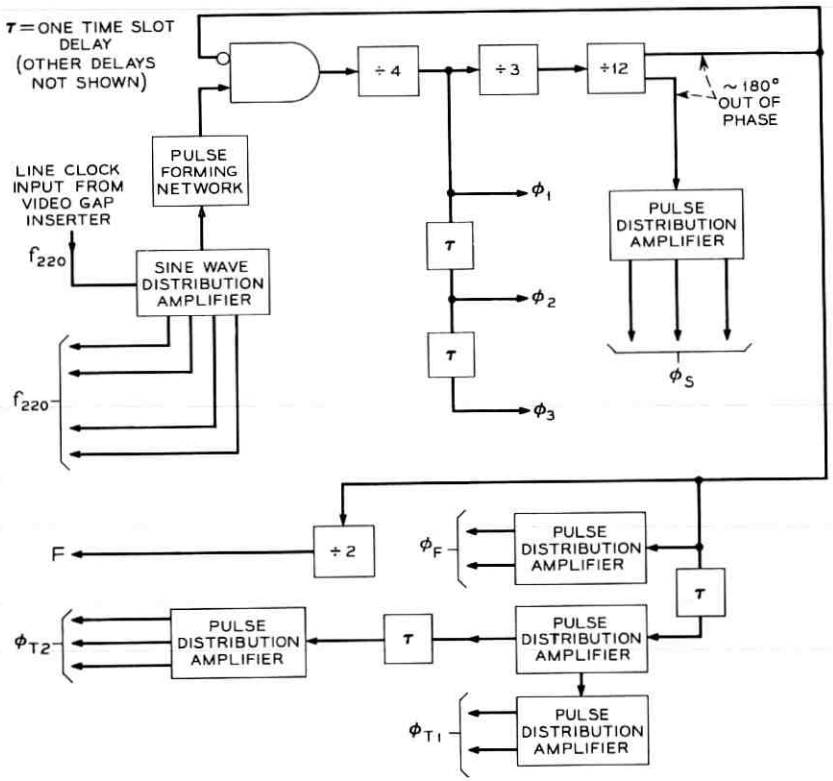


Fig. 12 — Transmitting clock distributor.

(ϕ's) shown in Fig. 11 and to distribute sinusoidal line clock signals. The framing gap, which is 4.47 ns wide and occurs once for every 145 line time slots, is created with a divide-by-144 counter preceded by an inhibit gate. The output of the counter inhibits the line rate input to the counter; hence, the framing gap exists on signals at each stage of the counter. Furthermore, the output of the counter occurs at the frame rate. Note that the S bit clock pulses, which are located midframe, are derived by selecting the "negative" phase at the output of the divide-by-12 counter, thereby avoiding the use of one-half frame of delay.

5.1.2 *Receiving Clock Distributor*

Note from Fig. 13 that this circuit is almost identical to the transmitting clock distributor, the main exceptions being the provision for inhibition by a signal from the frame detector and the furnishing of a special clock signal to the frame detector.

5.1.3 *Clock Distributor Performance*

The most difficult problems associated with the realization of the clock distributors were the counting function (because of the speed of the input signal and the delay stability dicated by the fact that such precise inhibition (one out of 145) was required), and the nature of the required output pulse trains. The circuit used for counting is discussed in Section 6.2.3. The nature of the pulse clock signals is shown in Fig. 14 where ϕ_1, ϕ_2, ϕ_3, and ϕ_4 are shown around the framing gap.

5.2 *Framing Gap Insertion and Removal*

5.2.1 *Video Gap Inserter*

The block diagram of Fig. 15 shows the components of the video gap inserter. The need for write and read clocks at one-half the video bit rate becomes apparent from this diagram. Pulses at that rate are launched down delay lines to accomplish sequential accessing of the storage cells. Complete regeneration of the 50 mV pulses at the output of the storage cells is performed by a 220 Mb/s regenerator. Use of the high speed regenerator at this point permits the use of a simple untimed line multiplex gate.

To allow for crystal control of the VCO, $f_{220}/18$ was generated and applied to a frequency multiplier, which is considered as an integral part of the VCO in Fig. 15.

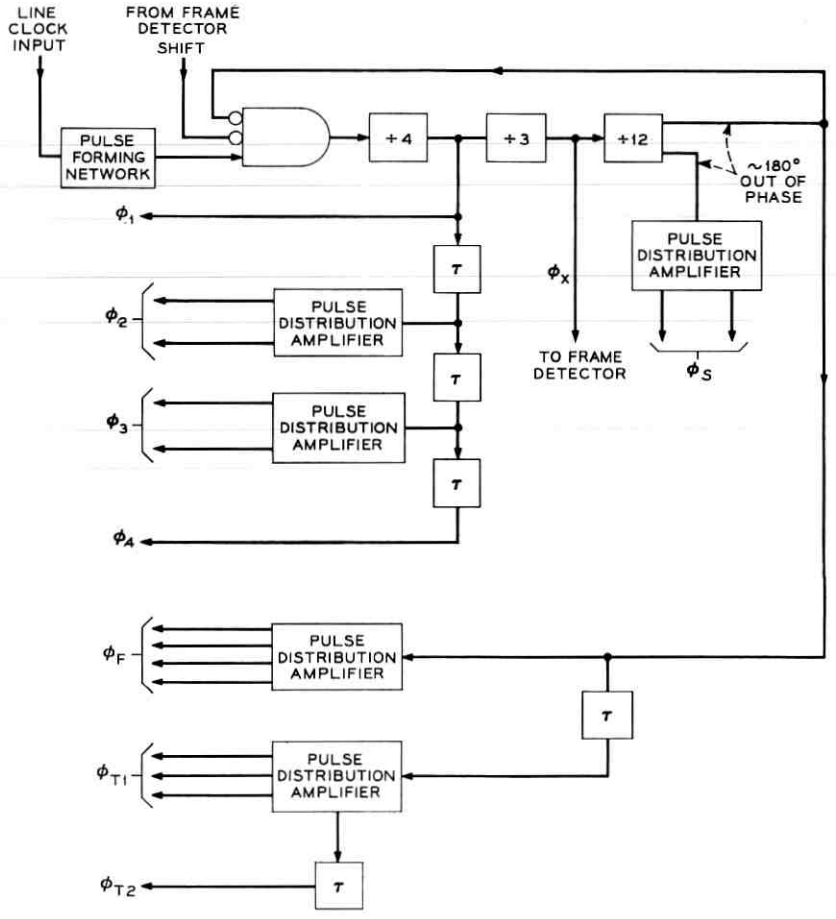


Fig. 13 — Receiving clock distributor.

5.2.2 *Video Gap Remover*

The phase discontinuity due to the framing gap can be thought of as phase jitter. Attenuation of this jitter is accomplished by the video gap remover shown in Fig. 16, which operates on the same principle as that used in the dejitterizer. The information containing the framing gap is written into the store under the control of ϕ_2, the write clock. The jitter on this clock is smoothed by the action of a phase-locked loop. The resultant smoothed timing signal is used as the read clock for the store. Note the similarity with the video gap inserter of Fig. 15. Many of the

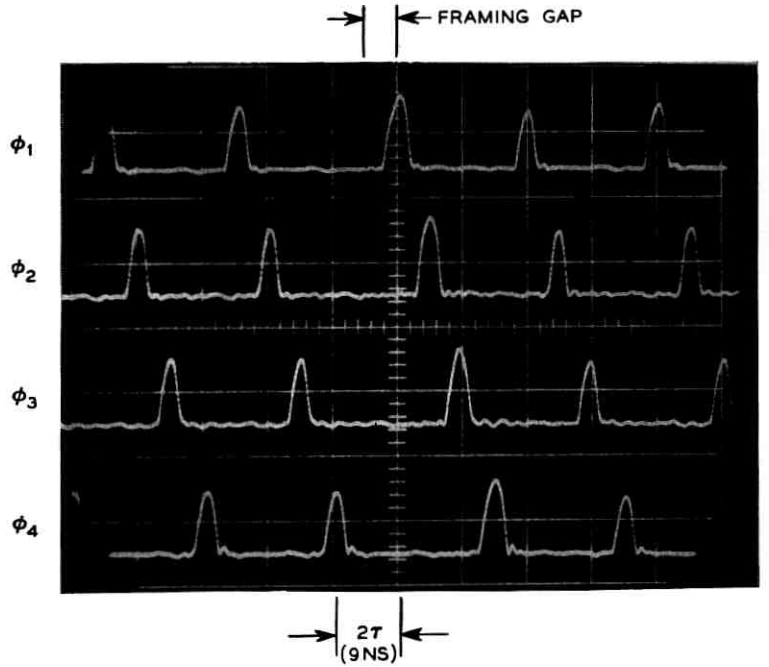


Fig. 14 — Oscilloscope trace showing ϕ_1, ϕ_2, ϕ_3, and ϕ_4.

circuit blocks are duplicated in the two equipment blocks. Also notice that a one-shot multibrator is used to widen the video information pulses from the line demultiplex gate to improve the operating margin of the store.

5.2.3 *55-Mc/s Gap Inserter*

This circuit, shown diagrammatically in Fig. 17, enables pules from a 55-Mb/s word generator (synchronized with the video coder clock) to be multiplexed onto the high-speed line. These pulses can be used to fill the time slots not occupied by video, mastergroup, T1 carrier, synchronization signaling (S) or framing pulses. This function could be performed by an elastic store as in the video gap inserter, but the lower bit rate being processed (f_{55}) allows the use of a simple stretching-sampling technique. The 55-Mb/s pulses are stretched to occupy three 224-Mb/s time slots. These are sampled by a modified ϕ_1 pulse train (called ϕ_1') to establish the framing gap. See Fig. 17. Pulse train ϕ_1' is ϕ_1 acted upon the 55-Mc/s inhibit circuit of Fig. 18 so that digits from the 55-Mc/s

word generator are not written into time slots occupied by the two T1 carrier channels and the S bit.

5.2.4 *55-Mc/s Gap Remover*

This circuit, shown in Fig. 19, applies the same stretch-sample technique employed in the 55-Mc/s gap inserter. The circuit is slightly more complicated since the narrow input pulses from the T1 demultiplex gate and the requirement for 50 per cent duty cycle output pulses dictate that pulse stretching take place at both input and output. The sampling clock is provided by the video gap remover. For the experimental system,

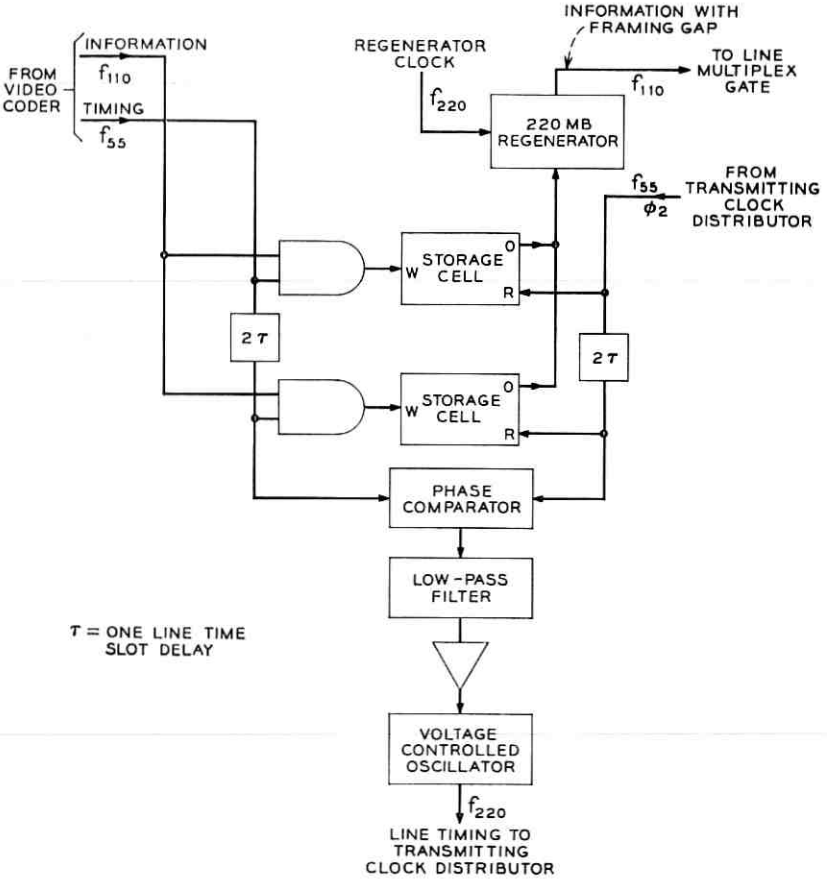


Fig. 15 — Video gap inserter.

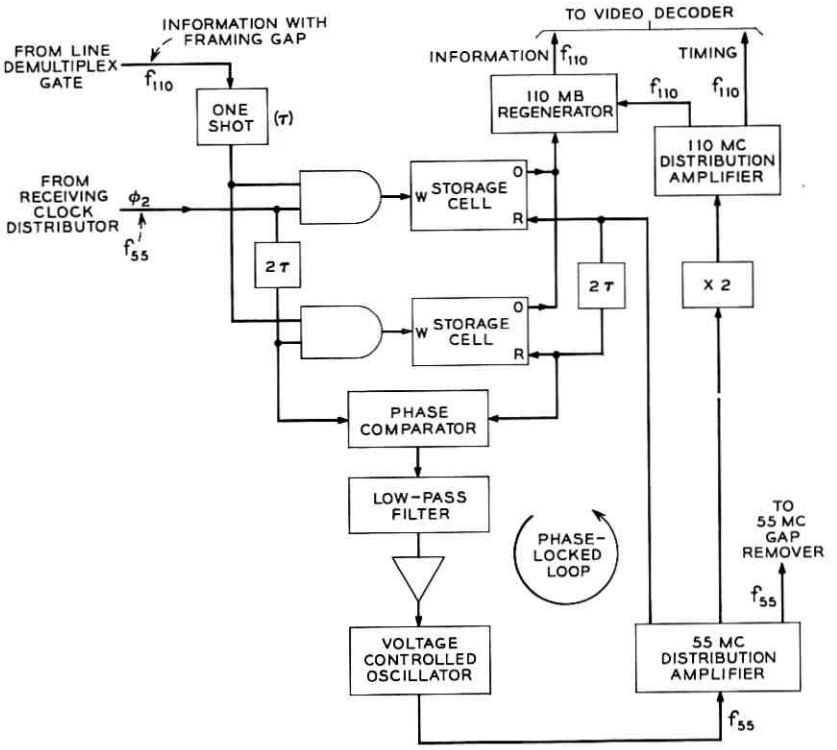


Fig. 16 — Video gap remover.

all 36 pulses per multiplex frame (T1 carrier, S bit, word generator) were processed as though they were from a synchronous coded master-group in order to establish the feasibility of this processing technique.

5.2.5 *Gap Insertion and Removal Performance*

Fig. 20 shows the performance of the video gap inserter and video gap remover. An all-ONES pattern was used to emphasize the phase discontinuity. The waveforms at the 55-Mc/s circuits are similar, but at one-half the bit rate.

5.3 *Pulse Stuffing Synchronization*

5.3.1 *Sync Signal Transmitter and Receiver*

These circuits provide a central control function for all synchronized channels. In addition to the coding and decoding of the sync signal (S

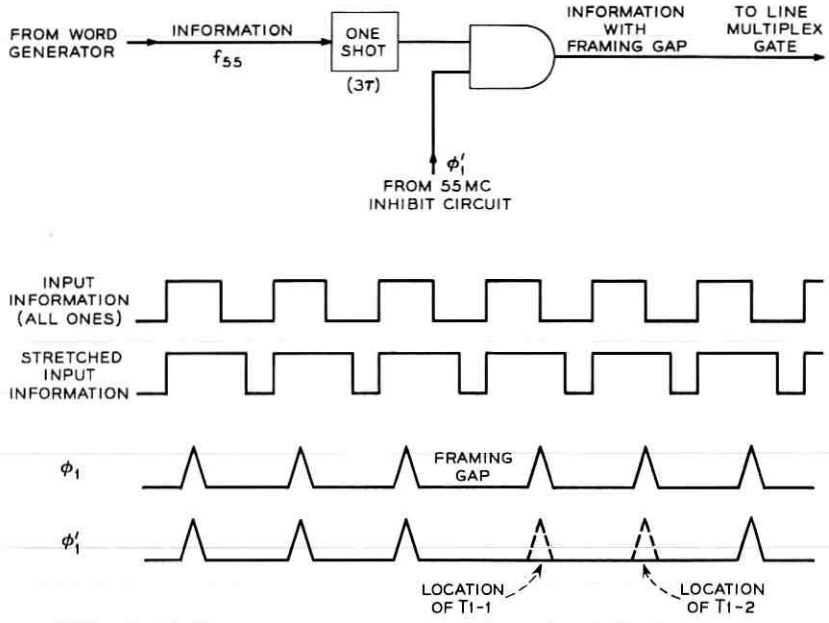


Fig. 17 — 55-Mc/s gap inserter.

bit), they control pulse stuffing and extraction at the synchronizers and desynchronizers. Each unit is a 1.5-Mc/s logic circuit which has been realized as a straightforward design using mostly ESS logic modules.[13] The S bit has been organized so that a single 18-stage shift register in each unit is used for generation and recognition of the M, R and C words.[8] In a commercial design, monolithic integrated circuits would probably be used for these circuits.

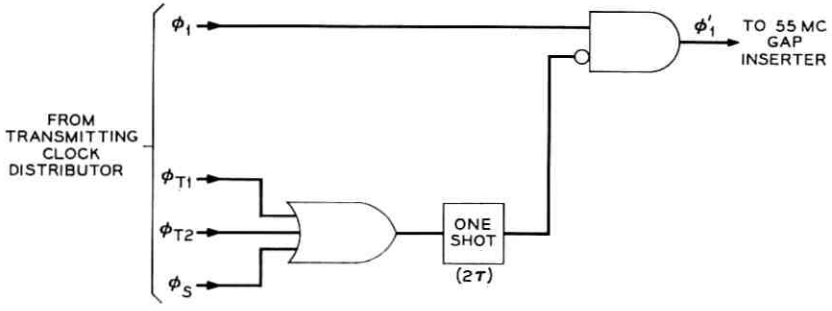


Fig. 18 — 55-Mc/s inhibit circuit.

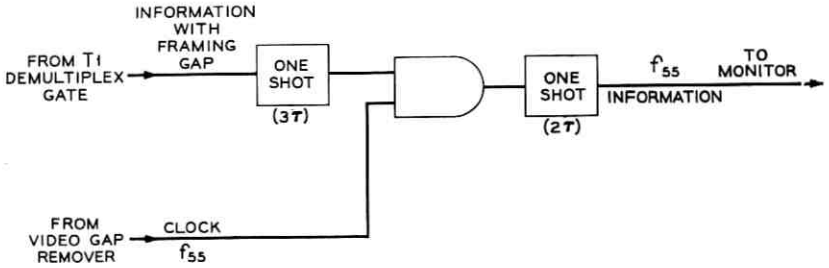


Fig. 19 — 55-Mc/s gap remover.

(a) VIDEO GAP INSERTER INPUT
(b) VIDEO GAP INSERTER OUTPUT
(c) VIDEO GAP REMOVER OUTPUT

Fig. 20 — Framing gap insertion and removal waveforms.

5.3.2 *Mastergroup Synchronizer*

This circuit, shown in block diagram form in Fig. 21, uses a three-cell elastic store with delay line commutation at the input and ring counter commutation at the output. Delay line commutation, which was used in the elastic stores of the dejitterizer, and video gap inserter and remover could not be used at the output because the stuff gap positioning was not synchronized with the multiplex clock. A three-cell store size is dictated by the following: one cell for the stuffed time slot, up to one more cell for waiting time when stuff rates near the maximum stuff rate are used, one-fourth cell for the framing gap (one 224 Mb/s time slot), one-fourth cell allowance for the fact the write and read operations cannot occur

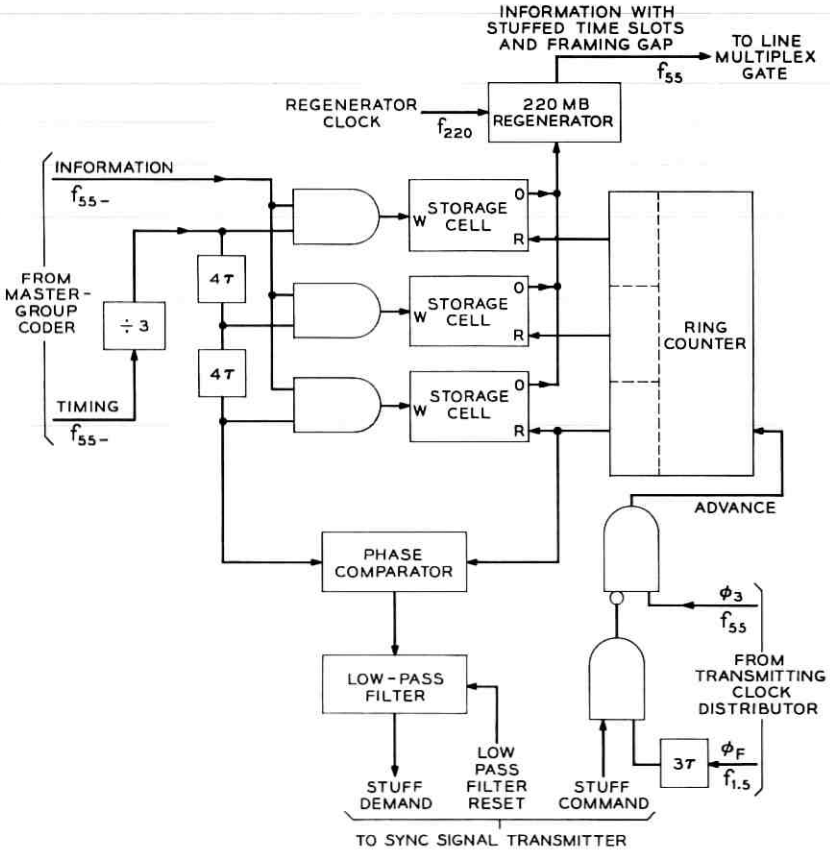


Fig. 21 — Mastergroup synchronizer.

simultaneously, and one-half cell for margin. Of course, lower stuff rates will increase this margin.

The low-pass filter reset signal is provided to assure that the need-to-stuff signal to the sync signal transmitter does not linger after stuffing has occurred.

5.3.3 *Mastergroup Desynchronizer*

It may be seen from Fig. 22 that the mastergroup desynchronizer is very similar to the mastergroup synchronizer except that ring counter commutation is required on the write side of the elastic store and that the read clock is obtained through the use of a phase-locked loop. Below

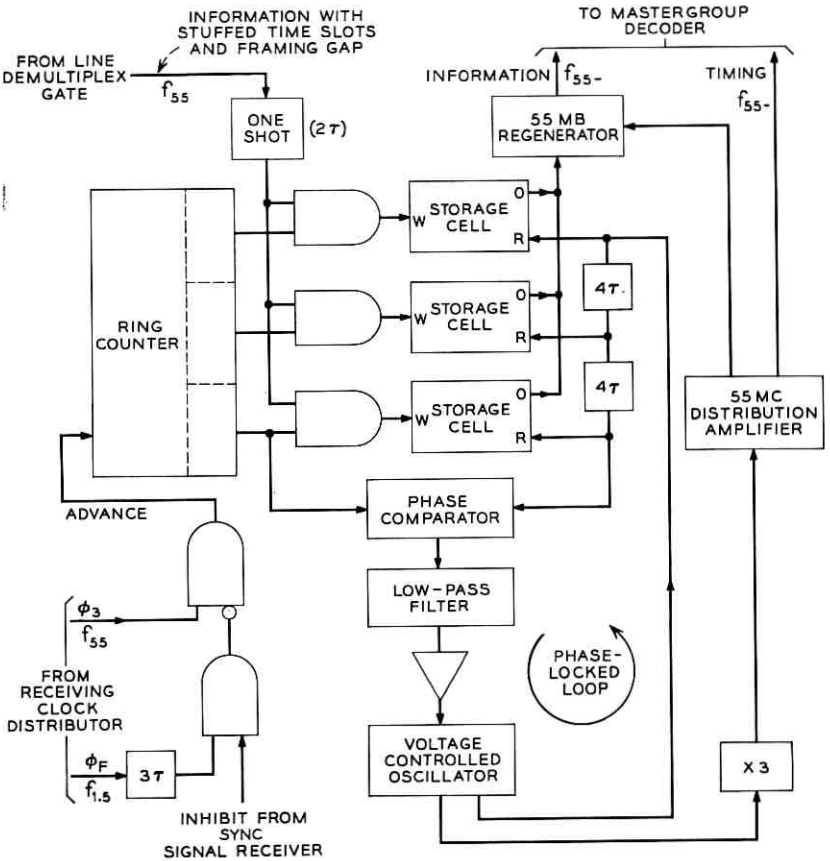


Fig. 22 — Mastergroup desynchronizer.

are calculations which show why the nominal offset of the synchronous and asynchronous 55-Mb/s rates was set at 24 ppm (see Table I).

It is apparent from the previous section that one-half cell is the maximum that is available for frequency drift if the maximum stuff rate is used. It was not anticipated that the maximum stuff rate be used during normal operation of the system, however, provision was made for operation at that rate to check out the circuits and to observe the effects of the resultant low-frequency phase jitter. Therefore, only one-half cell was allotted to frequency drift. The assumed ± 2 ppm clock stability (Table I) yields a low frequency mastergroup desynchronizer phase-locked loop gain, α_o, of

$$2\pi \cdot 142 \left(= \frac{(2 + 2)2\pi\ 55.}{\pi \cdot 1/2} \right)$$

radians/s/radian. (Equation (2) holds with α_o substituted for $B_j/2$.) With a simple one-pole RC filter with parameters set to yield critical damping, the jitter attenuation characteristic of Fig. 23 results. Note that if a 24 ppm offset (stuff rate = 1335 c/s) is used, the attenuation at the stuff rate is 27 dB, or the fundamental component of the residual phase jitter is 0.18 ns rms. This quantity of jitter can be seen from Fig. 7 of Ref. 1 to be acceptable. Furthermore, it may be demonstrated that since the stuff rate is about one-twentieth of the maximum stuff rate, the waiting time component of the jitter also meets the transmission requirements. The 24 ppm offset is by no means optimum, but the

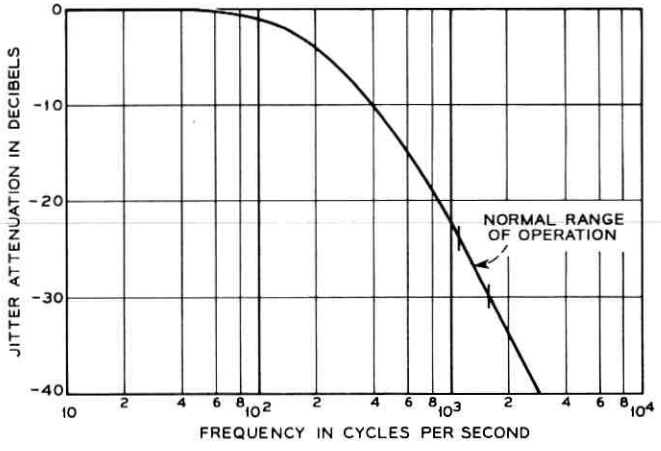


Fig. 23 — Jitter attenuation characteristic of mastergroup desynchronizer.

performance it yields indicates that pulse stuffing synchronization is indeed feasible. Note that almost a full cell of storage is available as margin for the stabilities indicated in Table I.

5.3.4 *Dual T1 Synchronizer*

This circuit, realized with monolithic integrated circuit modules, shares write and read circuits between two three-cell elastic stores. Whereas the tunnel diode is used elsewhere as the storage cell element, flip-flops are used for the T1 carrier processing circuits.

5.3.5 *Dual T1 Desynchronizer*

An interesting aspect of this circuit was that the phase-locked loop was made to have high-loop gain and the phase jitter was reduced only enough to prevent improper operation of a T1 carrier repeatered line. The motivation here was to reduce the amount of storage needed to accommodate the ± 50 ppm T1 carrier clock variations and to allow the use of a noncrystal VCO. From Table I it may be seen that the maximum required stuff rate for a T1 carrier channel is 112 $(=50 + 60 + 2)$ ppm or 173 c/s. The phase jitter due to pulse stuffing may be roughly represented by a sawtooth wave of frequency 173 c/s and amplitude 648 ns (one T1 carrier time slot) since the stuff rate is only about one-fifth of the maximum allowable stuff rate. Calculations revealed that the distortion in a T1 carrier system used for voice or data service which would result from this amount of phase modulation is entirely negligible. Measurements on the system bore this out. Since the distortion was even less for lower stuff rates, the full ± 50 ppm variation of the T1 carrier clock could be tolerated.

Unfortunately this simplification of the dual T1 desynchronizer could not be tolerated for *PICTUREPHONE* transmission over two T1 lines. It was determined, however, that a three-bit store was adequate if the nominal offset were increased from 60 to 100 ppm and the phase-locked loop design were modified to provide some jitter attenuation. The sacrifice paid is a tighter stability requirement on the VCO; crystal control was necessary in order to make the T1 carrier clock the primary contributor to the net oscillator instability.

5.3.6 *Synchronizing Circuit Performance*

Clock frequencies were varied substantially more than those shown in Table I without any observable impairment of transmission of a coded mastergroup or T1 carrier signals through the system.

The mastergroup coder clock could be synchronized with the multi-

plex clock and with no circuit adjustment from the asynchronous arrangement, error-free performance was achieved. All of the circuits used to achieve synchronization operated over the full stuff frequency ranges, and the calculated quantities of phase jitter were observed.

5.4 *Multiplexing and Demultiplexing Circuits*

Two levels of multiplex are employed. The two T1 carrier channels and the S bit are first combined in the T1 multiplex gate. The output of this circuit and the other digital signals, now all synchronous, are combined in the line multiplex gate. At the demultiplexer, the two T1 carrier channels and the S bit are identified in the T1 demultiplex gate, but the entire "ϕ_3" digital channel is passed on to the 55-Mc/s gap remover to show the ability of the technique used therein to remove framing gap phase jitter.

5.4.1 *Line Multiplex Gate*

This circuit, shown diagrammatically in Fig. 24, is a simple five-input untimed combining gate followed by a regenerator. This simplicity is made possible by the shaping given to the pulses at the output of each processing circuit. Provision was made for clocking each input to the gate but it was not necessary.

5.4.2 *T1 Multiplex Gate*

This circuit is very similar to the line multiplex gate except that each input must be clocked because the T1 carrier and S-bit processing circuits furnish 50 per cent duty cycle pulses at 1.5 Mb/s. No regenerator is needed at the output of this gate.

5.4.3 *Line Demultiplex Gate*

Functions performed by this block, shown in Fig. 25, include the distribution of the incoming pulse train to the frame detector and the

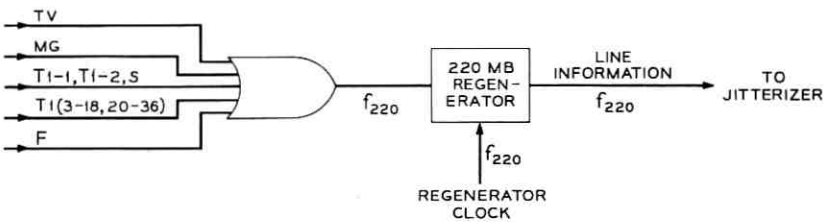


Fig. 24 — Line multiplex gate.

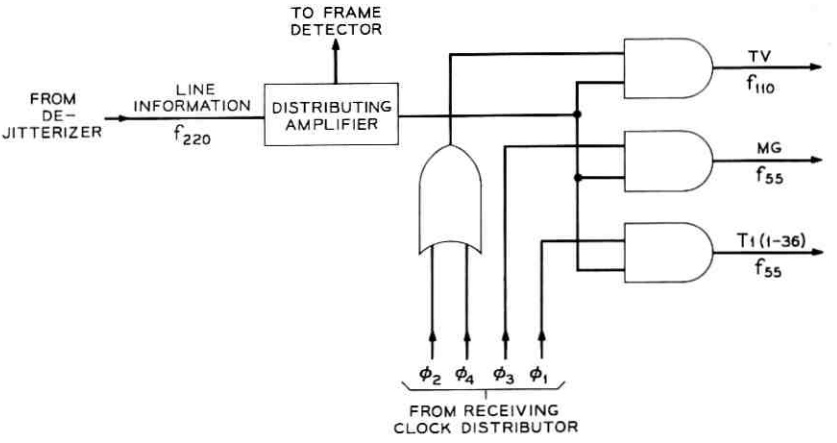


Fig. 25 — Line demultiplex gate.

separation of the coded video and the two 55-Mb/s pulse trains. Note that ϕ_2 and ϕ_4, the clock pulse trains for coded video, are first combined before entering the gate, thereby minimizing the fan-out of the gate driving amplifier to three, an important consideration for logic operating at this speed.

5.4.4 *T1 Demultiplex Gate*

As one sees in Fig. 10, the role played by this circuit is very similar to that of the line multiplex gate. The same basic circuit is used for both applications.

5.4.5 *Frame Detector*

The function performed by this circuit is similar to that performed by the frame detectors in the robbed bit coded video signal framing[14] and in the T1 carrier system.[7] However, a feature has been added which results in a circuit simplification.

In each system mentioned above and for the multiplex framing scheme, the framing pattern is alternating ONE's and ZERO's. During the reframe or search mode, the condition of a time slot is compared with the status of another exactly one frame away. If the comparison yields other than the alternation, the next later time slot is examined. To reduce the reframe time, not only is the time slot under consideration observed, but so is its neighbor. The previous methods mentioned above use the first half of a frame to make one comparison and the

second half to make the comparison for the neighboring time slot. This means that at least one-half frame of storage must be provided and that the comparator must be reset and must sense the condition of that neighboring time slot. A simpler comparator circuit, shown in Fig. 26 as a part of the frame detector, uses the neighboring time slot condition to route the error indicating pulse to either SET or RESET the flip-flop (whichever is appropriate), thereby readying the comparator for the next frame. Thus, one comparison per frame is made as with the other technique.

The gates and one shot multivibrators of Fig. 26 are required to provide pulse selection and to allow the use of slower logic circuits within the frame detector. The shift signal producing circuit contains an ad-

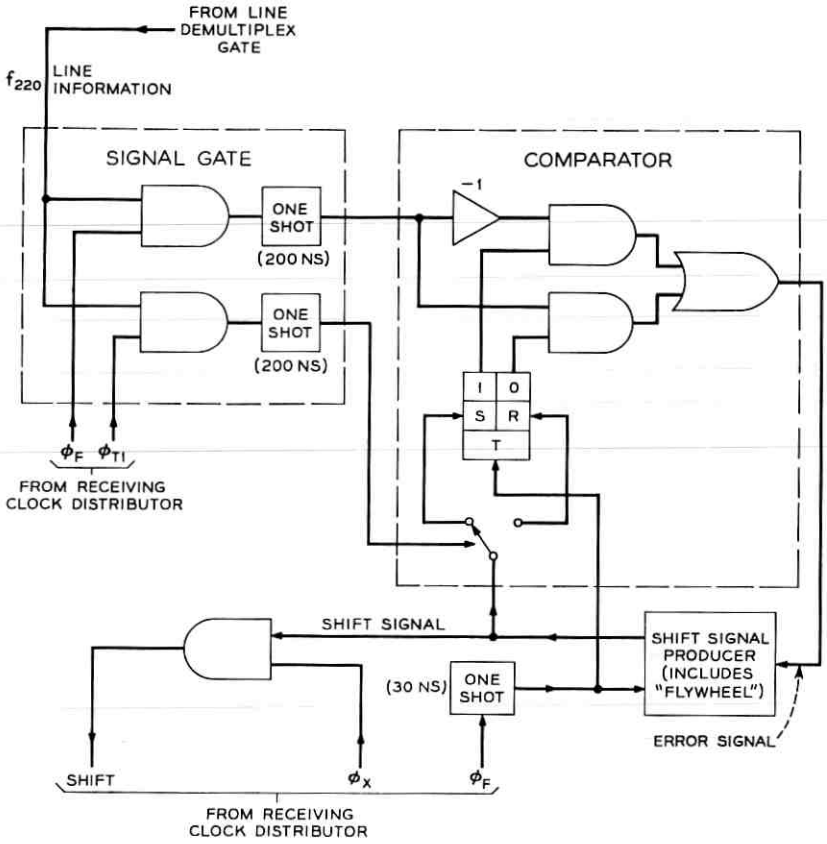


Fig. 26 — Frame detector.

justable flywheel which is realized with an integrator followed by a threshold circuit.

5.4.6 *Multiplexing and Demultiplexing Circuit Performance*

The multiplexing and demultiplexing functions provided complete isolation of the individual digital channels. These circuits performed satisfactorily for all signal formats which could be checked, which included fixed patterns ranging from all-ZEROS to all-ONES from word generators and typical continuously changing pulse patterns from the mastergroup, video and T1 carrier signal sources.

A photograph of an oscilloscope trace of a fully loaded line signal at the output of the line multiplex gate is shown in Fig. 27. Note the absence of intersymbol interference.

One test of the frame detector performance consists of forcing the out-of-frame condition to occur by forcing a single SHIFT output signal to be generated by the frame detector. Since searching takes place over an entire frame, the resulting reframe time is an upper bound on the

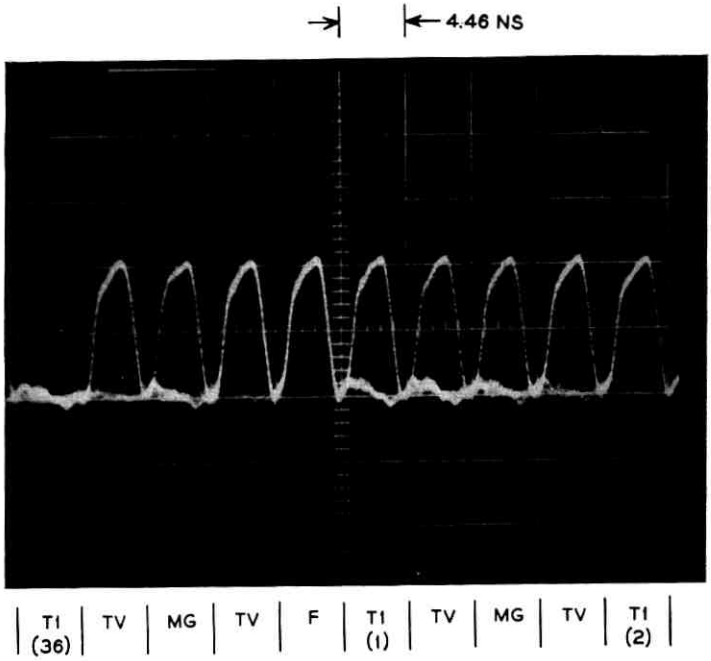


Fig. 27 — Line waveform.

actual reframe time for the particular line loading present at the time of the test. It was found that for an all-ZEROS line loading the reframe time was 93 μs. It was not convenient to make measurements with a random loading pattern, however with the loading shown in Fig. 27 the maximum average reframe time measured in 1000 trials was 134 μs. These numbers are consistent with the 188 μs calculated maximum average reframe time for a random pattern, after the flywheel effect was taken into account.

VI. CIRCUIT TECHNIQUES[15]

The intention of this section is to convey in general terms some of the more significant high speed circuit concepts which led to the realization of the multiplexer and demultiplexer in the 224 Mb/s digital transmission system.

6.1 *Solid-State Devices*

To achieve reliable logic circuits in a 224 Mb/s system, it was necessary to have available transistors which could switch in less than one nanosecond. Such a transistor was developed for this and other projects[16,17,18] which has a gain-bandwidth product in excess of 2.5 Gc/s, a dissipation rating of 50 mW at 25°C ambient and a maximum C_{ob} (direct collector-base capacitance with the emitter open circuited) of 0.8 pF. It is a pnp diffused base epitaxial mesa germanium transistor. All transistors shown in the diagrams which follow are this device.

In applications where higher power handling capability and/or lower speed requirements dictate, npn planar silicon devices are used. These units typically have a gain bandwidth product of 1.0 Gc/s, a 25°C ambient dissipation limit of 200 mW and a C_{ob} of 1.2 pF.

A number of tunnel and step-recovery diodes were used throughout the system. Many Schottky-barrier (metal-semiconductor) diodes were used because of their excellent low capacitance and small minority carrier storage characteristics. In the following figures, all diode locations not identified as tunnel or step-recovery diodes are applications for the Schottky-barrier diode.

6.2 *Specific Circuits*

6.2.1 *Current and Pulse Routing*

Fig. 28 illustrates how the current routing and pulse routing concepts[19] were utilized in the demultiplex gate. Of all the gating circuits realized,

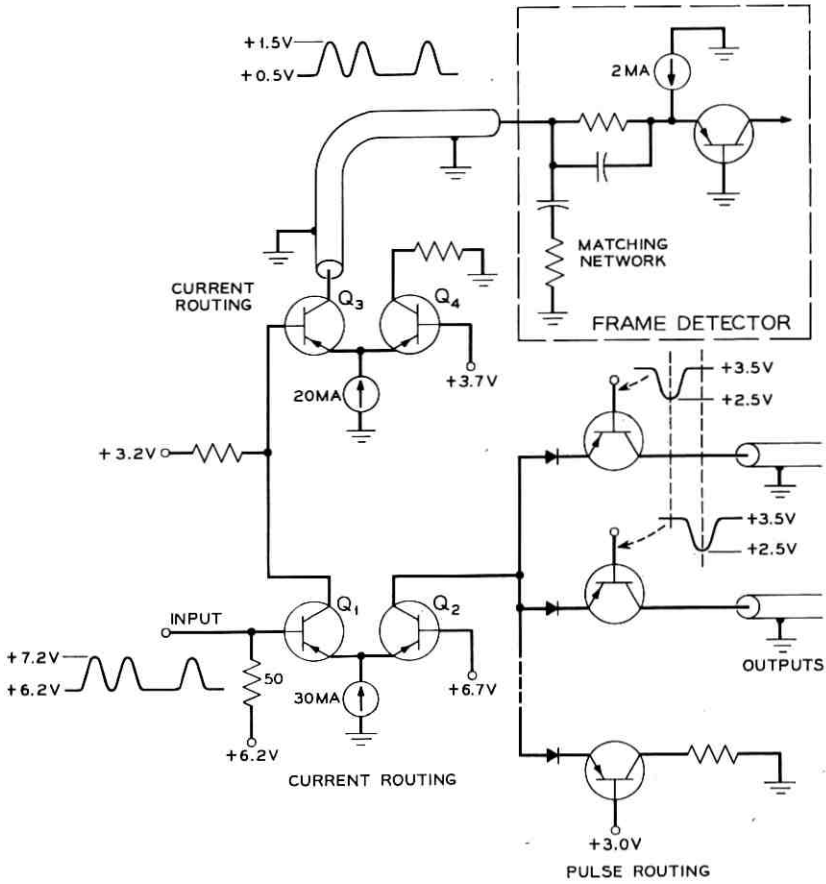


Fig. 28 — Current and pulse routing.

this one was found to be most demanding of the high speed characteristics of the available devices. The current routing stage with transistors Q_1 and Q_2 provides a good termination for the incoming pulse train, furnishes a noninverted signal for the gate transistors and drives another current routing stage (Q_3Q_4). Notice that the bias on the "cold" base is midway between the levels on the "signal" base.

The stage containing Q_3 and Q_4 is used as an inverter to send a replica of the input signal to the frame detector. One principle applied extensively throughout the system is the transmission of signals over moderate distances using coaxial cables with a far end termination only. The far end also provides the dc bias path for the line driving transistor.

The common-base stage shown, quiescently biased slightly ON and with a network to build out the input impedance, provides an excellent line termination with less than 10 per cent reflection and negligible memory.

The desired pulse trains are stripped off by applying the composite signal as a current drive to all emitters of the gate transistors and by clocking at each base (pulse routing).

Current routing and pulse routing are used extensively throughout the system, especially for circuits operating at the line bit rate.

6.2.2 *Tunnel Diode Storage Cell*

All elastic stores in the system except for those used to process the T1 signals use the storage cell and associated gates shown in Fig. 29.[6] The tunnel diode, biased in a bistable mode, furnishes a simple, fast memory element. As described earlier, write and read commutation comes from tapped delay lines except where ring counters are required. The write AND gate and read AND gate route current to the tunnel diode to change its state. The use of this type gate results in a controlled and predictable loading of the tapped write and read transmission lines. The "differentiator-clipper" is required to form the negative output pulse. Gating and clipping functions are performed by diodes rather than transistors for reasons of economy. However, the relatively small output level (50 mV across 50Ω) requires the use of regeneration; hence, in a

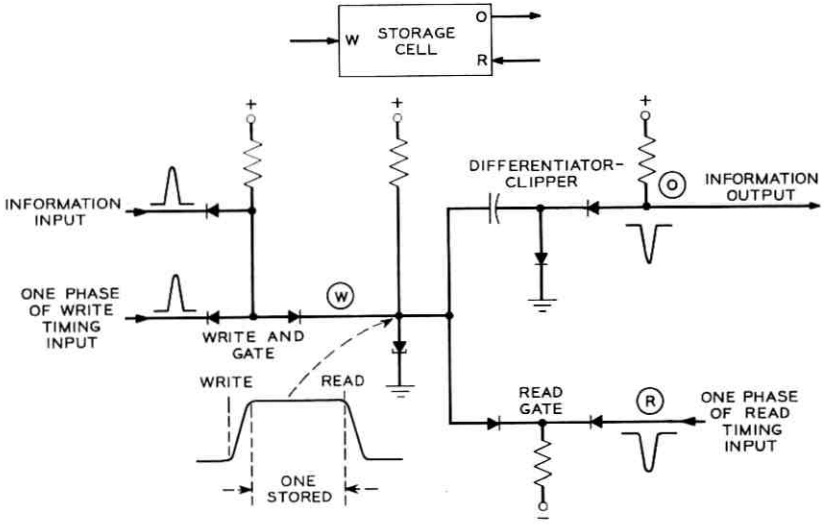


Fig. 29 — Tunnel diode storage cell.

commercial system, it is quite possible that for small stores a cost savings could result by using a transistor in each differentiator-clipper and thereby simplifying the regenerator design.

6.2.3 *Counter Stage*

Another use of the tunnel diode is in the binary counter stage shown in Fig. 30. The basic counting stage is that described by Chow.[20] Reliable cascading of these stages has been achieved by using a transistor with a small emitter bypass capacitor as a differentiator-clipper. In cases where feedback is applied to count by factors other than powers of two, the reset pulse is applied as a current drive to the point shown.

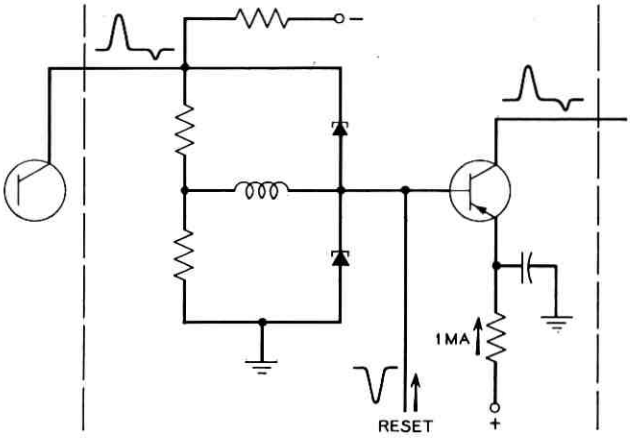


Fig. 30 — Basic counter stage.

A critical characteristic of the binary stage used in this system is its delay stability (see Section 5.1.3). A typical realization of the type indicated in Fig. 30 resulted in the performance shown in Table II for a single binary stage. For this circuit the tunnel diode bias voltage was obtained from the voltage drop of a forward biased diode. These data show that the delay stability performance of this circuit is quite adequate for use in this system.

6.2.4 *Tunnel Diode — Transistor Stage*

A circuit which has proven to be quite useful is the gallium arsenide tunnel diode - 2.5-Gc/s transistor circuit shown in Fig. 31.* The composite

* Ref. 20, pp. 275–277.

TABLE II — DELAY VARIATIONS FOR A SINGLE BINARY STAGE

±0.03 ns for ±20% power supply voltage variations
±0.04 ns for ±5 dB input pulse amplitude variations
0.08 ns over −10°C to +60°C temperature range

input I-V characteristic is very similar to that of a tunnel diode. A GaAs diode was chosen because it has the largest available valley voltage and is available with parameters which are good enough to make the transistor the speed limiting component. In earlier investigations, reliability problems were encountered with GaAs tunnel diodes when the valley voltage was exceeded. This circuit has the feature that the transistor input characteristic protects against this possibility, and furthermore, the tunnel diode, with its "backward diode" reverse characteristic, removes the strain from the transistor for a reversed polarity excitation.

As a bistable circuit, the combination is used to realize a sawtooth phase comparator (Fig. 32) and a pulse regenerator (Fig. 33). In the former application it can be seen that the dc value of the output waveform is proportional to the phase difference of the two input pulse trains. With moderately narrow trigger pulses (<3 ns at the base) this circuit

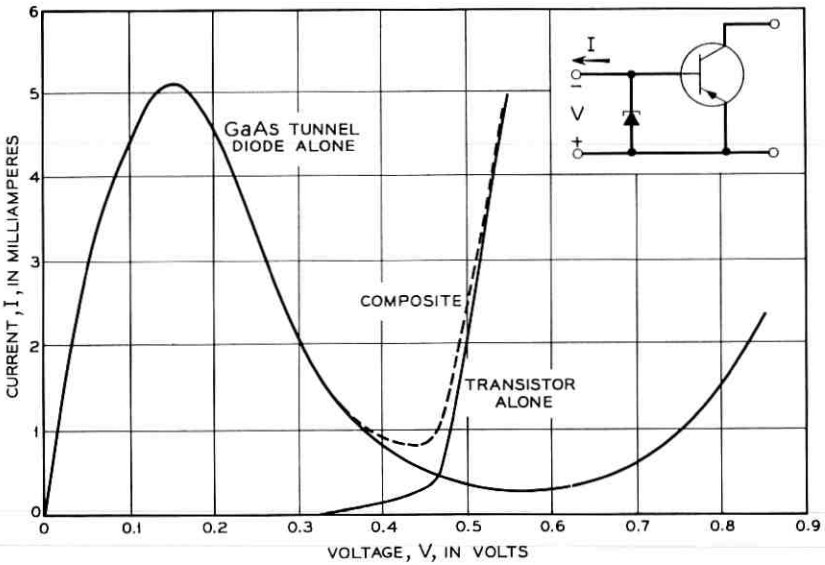


Fig. 31 — Tunnel diode transistor stage.

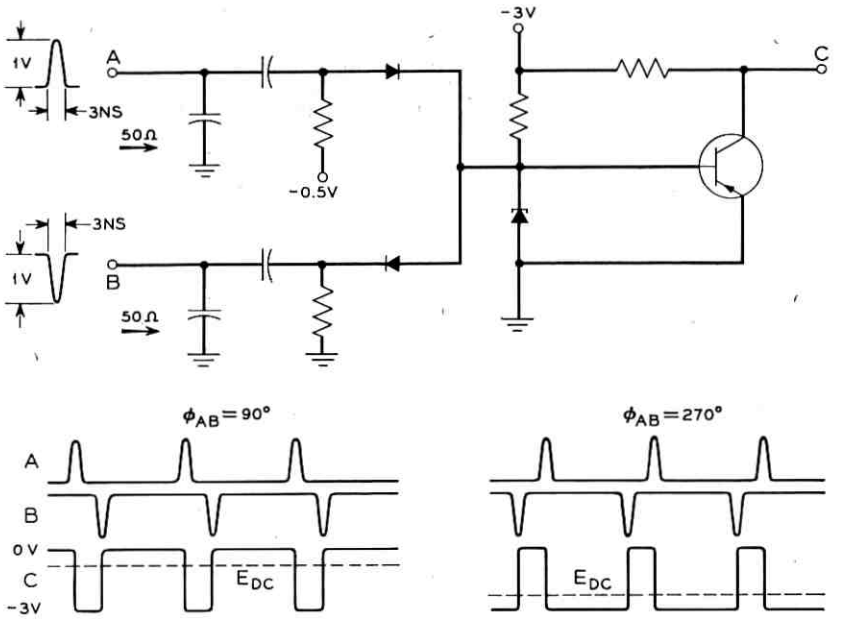


Fig. 32 — Sawtooth phase comparator.

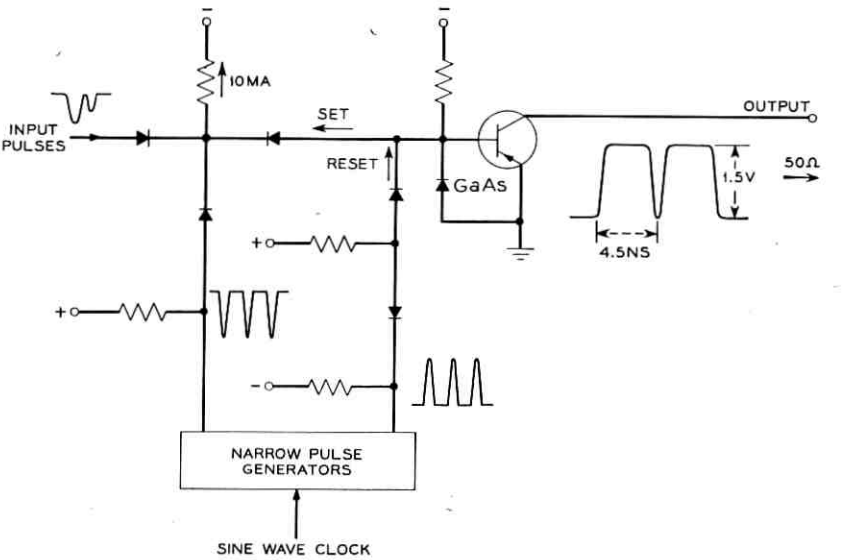


Fig. 33 — 224 Mb/s regenerator.

has been used up to 55 Mc/s (in the video gap inserter and remover) with excellent results.

In the pulse regenerator, the pulse to be regenerated is gated with a clock pulse generated from a sine wave in a step-recovery diode circuit. The resultant pulse turns the transistor ON. A second clock pulse of opposite polarity turns the stage OFF. This circuit is the one which produces the 224-Mb/s pulse train shown in Fig. 27.

Another application of the basic tunnel diode-transistor circuit is in the various one shot multivibrators in the system. Both kinds of R-L load lines with one stable point (quiescent ON and quiscent OFF) have been used with satisfactory results. Output pulse widths range from 4.5 to 300 ns.

VII. EQUIPMENT DESIGN

Since the primary goal of the investigation was to establish the feasibility of processing high-speed digital signals in the manner described, a conservative equipment design approach was used. For the most part, circuits shown as blocks in the preceding figures are constructed as separate units using point-to-point wiring boards mounted in an enclosure which provides good shielding and power supply decoupling. The construction technique for the elastic stores is described elsewhere.[6] Interconnection is accomplished by 50-ohm coaxial cable. Each functional block, such as the mastergroup synchronizer, is realized as a single 19-inch panel. All T1 and sync signaling circuitry is made from standard ESS plug-in modules and mounting hardware.[13] Standard relay racks house the equipment. The multiplexing and demultiplexing bays are shown in Fig. 34(a), and the bay containing the jitterizer and dejitterizer is shown in Fig. 34(b).

The equipment design approach used, although quite adequate from an electrical performance standpoint, would not be suitable for a commercial system. It is quite possible that a mother-board approach similar to that used in the coder[14] would yield suitable electrical performance. Also, integrated and thin-film circuit techniques would be applicable to some circuits.

VIII. SUMMARY

It has been demonstrated that solid-state device and circuit technology has advanced to the point where digital signals with bit rates up to 224 Mb/s can be readily processed. Through the use of the pulse stuffing synchronization technique, with added-bit sync signaling, a

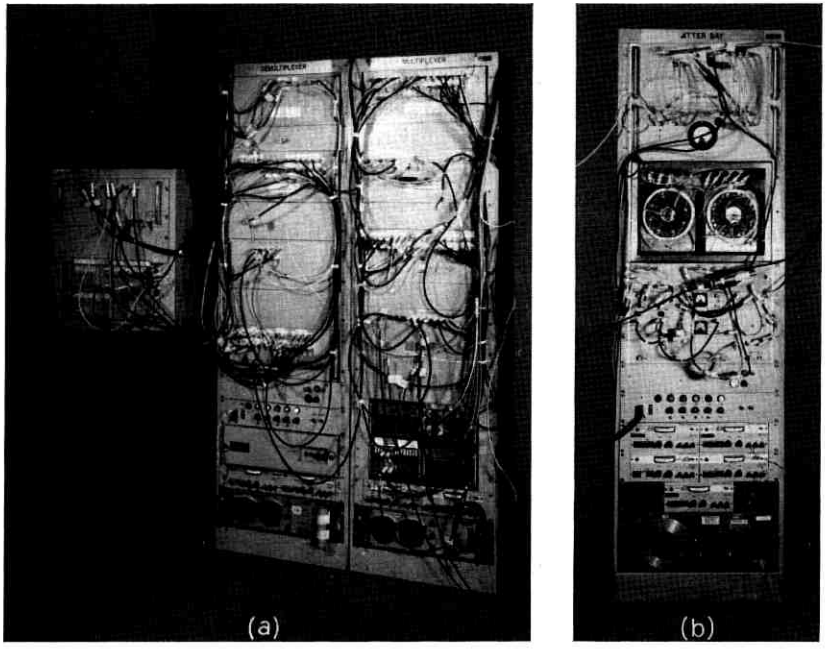


Fig. 34(a) — Multiplexer-demultiplexer bays.
(b) — Jitter bay.

multiplicity of asynchronous signals with a variety of bit rates can be combined into a 224-Mb/s pulse stream, and each signal can be recovered at the other end of the system. The techniques described also furnish the capability of dropping portions of the bit stream at an intermediate point and adding other signals in the vacated time slots. These latter signals can of course also be asynchronous. All of this can be accomplished without highly stable clocks and without large digital storage.

Also, it has been shown that accumulated phase jitter can be controlled sufficiently to avoid signal impairment.

A large number of presentations have been given where the features and performance reported herein were demonstrated. Experience during the course of these demonstrations has confirmed the conclusion that realization of an elaborate 224-Mb/s digital multiplexer-demultiplexer is feasible.

IX. ACKNOWLEDGMENTS

The evolution and development of the system described herein are obviously due to the efforts of many members of Bell Telephone Labora-

tories. The system plan was proposed by J. S. Mayo. Major sections were designed by A. A. Geigel, V. I. Johannes, D. Koehler, W. E. Ballentine and L. D. Owens. Noteworthy contributions were made by R. H. McCullough, V. R. Saari, and H. A. Hageman. R. J. Kirkpatrick and A. W. Busler carried out the equipment design and construction.

REFERENCES

1. Mayo, J. S., Experimental 224 Mb/s PCM Terminals, B.S.T.J. This Issue, pp. 1813–1841.
2. Graham, R. S., Pulse Transmission System, U. S. Patent No. 3,042,751, 1962.
3. Mayo, J. S., PCM Network Synchronization, U. S. Patent No. 3,136,861, 1964.
4. Byrne, C. J., Karnaugh, M., and Scattaglia, J. V., Retiming of Digital Signals with a Local Clock, NEREM Record, 1960.
5. Byrne, C. J., and Scattaglia, J. V., A Buffer Memory for Synchronous Digital Networks, Sixth Mil-E-Con Convention Record, 1962.
6. Geigel, A. A., and Witt, F. J., Elastic Stores in High-Speed Digital Systems, NEREM Record, 1964, pp. 122, 123.
7. Fultz, K. E. and Penick, D. B., The T1 Carrier System, B.S.T.J., *44*, Sept., 1965, pp. 1405–1451.
8. Johannes, V. I., and McCullough, R. H., Multiplexing of Asynchronous Digital Signals Using Pulse Stuffing with Added-Bit Signaling, NEREM Record, 1965, pp. 168, 169.
9. Byrne, C. J., Karafin, B. J., and Robinson, D. B., Jr., Systematic Jitter in a Chain of Digital Regenerators, B.S.T.J., *42*, Nov., 1963, pp. 2679–2714.
10. Byrne, C. J., Properties and Design of the Phase Controlled Oscillator with a Sawtooth Comparator, B.S.T.J., *41*, March, 1962, pp. 559–602.
11. Rubin, P. E., private communication.
12. Travis, L. F., and Yaeger, R. E., Wideband Data on T1 Carrier, B.S.T.J., *44*, October, 1965, pp. 1567–1604.
13. Cagle, W. B., Meene, R. S., Skinner, R. S., Staehler, R. E., and Underwood, M. D., No. 1 ESS Logic Circuits and their Application to the Design of the Central Control, B.S.T.J., *43*, Sept., 1964, pp. 2055–2095.
14. Edson, J. O., and Henning, H. H., Broadband Coders for an Experimental 224 Mb/s PCM Terminal, B.S.T.J., This Issue, pp. 1887–1940.
15. Koehler, D., Semiconductor Switching at High Pulse Rates, IEEE Spectrum, Nov., 1965.
16. Hamasaki, J., A Wideband High-Gain Transistor Amplifier at L-Band, International Solid-State Circuits Conference, 1963, Digest, pp. 46–47.
17. Englebrecht, R. S., and Kurokawa, K., A Wideband Low-Noise L-Band Balanced Transistor Amplifier, Proc. IEEE, *53*, March, 1965, pp. 237–247.
18. Eisele, K. M., Engelbrecht, R. S., and Kurokawa, K., Balanced Transistor Amplifiers for Precise Wideband Microwave Applications, International Solid-State Circuits Conference, 1965, Digest, pp. 18–19.
19. Koehler, D., A 110-Megabit Gray-Code to Binary-Code Serial Translator, International Solid-State Circuits Conference, 1965, Digest, pp. 84–85.
20. Chow, W. F., *Principles of Tunnel Diode Circuits*, John Wiley & Sons, New York, 1964, pp. 307–309.

Broadband Codecs for an Experimental 224 Mb/s PCM Terminal

By J. O. EDSON and H. H. HENNING

(Manuscript received August 5, 1965)

High-speed PCM codecs (coders and decoders) have been constructed to handle broadband signals such as a mastergroup of telephone channels or an NTSC color TV signal. Two widely different approaches to the realization of the coding function are described. The first and earliest version utilizes a beam coding tube to convert the analog signal to digital form. The second approach employs a tandem array of solid-state stages to perform the required conversion. Associated circuits to implement the filtering, sampling, holding, timing, translation, framing, and decoding functions in a PCM codec are also described. Analysis results, experience gained in design, and measured performance of the coders and the associated circuits that make up a PCM codec demonstrate that they can be produced to meet stringent performance objectives.

I. INTRODUCTION

The design of the coding and decoding complexes of the PCM terminals discussed by Mayo[1] will be covered herein. The term "codec" will be used to denote the coder (analog to digital converter), the decoder (digital to analog converter), and all of the associated circuits, such as sample and hold, parallel to serial converted, Gray to binary translator, resampler, and framing and timing circuits. This definition specifically excludes the multiplexing and demultiplexing portions of the terminal covered in the paper by Witt.[2]

Before the detailed circuit designs are examined, it will be profitable to delineate the two approaches taken to the realization of the encoder. In both coder approaches ideal uniform quantization was sought. The additional quantizing noise advantage attendant to nonuniform quantization was left for future development.

When this project was undertaken only the beam coding tube[3] appeared to satisfy the need for high performance mastergroup and TV

encoding.[4] Even then it was recognized that this was by no means an easy task. A much improved coding tube was required[5] and high-speed, high-accuracy, solid-state circuitry had to be designed to couple the tube to the outside world.

At the same time, several other general approaches to coding were examined to determine an approach that could be realized exclusively with solid-state devices. This study phase led to a proposal for utilizing a novel circuit realization[6] of the folding coder.[7] Considerable uncertainty remained since implementation of the scheme called for the use of solid-state devices on the very fringe of the state of the art.

As expected from the outset, the coder using the beam coding tube achieved the desired performance first. Realization of the solid-state coder proceeded more slowly as each new obstacle was recognized and overcome. To date, the tube coder has outperformed its solid-state counterpart in both speed and accuracy; though this gap is being narrowed. Indeed, improved devices, circuit techniques, and equipment layouts (beyond those described in this paper) are coming into being to support the conclusion that this gap is destined to disappear. There is little doubt *today* that the solid-state coder can be made more economically than the tube coder, and be realized in a considerably smaller package, and can be made more compatible from the equipment standpoint with the rest of the terminal. Though it is clear that the future lies with the all solid-state system, the beam coding tube has served as a very useful stepping stone. Its mere existence was an essential vehicle for exercising other terminal circuits as they came into being. More important, the early realization of the tube coder was used in a field experiment to verify the predicted high quality performance obtainable with a PCM terminal transmitting live mastergroups.[8,9] This was an integral and valuable link in the over-all development program.

Section II covers the beam coding tube and its associated deflection and readout circuitry. Section III is devoted to the solid-state coder design, realization, and performance. Section IV is compartmentalized into subsections describing the remaining circuits that complete the coder. In Section V the circuits in the decoder are described. A brief section on general equipment principles concludes the paper.

II. BEAM TUBE CODER

2.1 *General*

Coding tubes of the ribbon-beam type are particularly well suited to analog to digital conversion where precision and relatively fast coding

times (in the order of several tens of nanoseconds) are required. Tube coders fall into the class of word-at-a-time coders which are inherently fast. All possible code words are stored on the code plate, the code word that corresponds to the analog input is selected by deflecting the beam, and binary decisions on all digits are made simultaneously and independently.

The accuracy of the coding operation is primarily a function of mechanical tolerances within the tube. Considerable progress was made in this area by the Tube Development Laboratory of Bell Telephone Laboratories culminating in the design of the nine-digit Gray code coding tube used in the system.[5]

The tube coder has been operated at a sampling rate of about 12 Mc/s for coding of standard black and white as well as color television signals and at approximately a 6-Mc/s sampling rate for coding of 600-channel single-sideband frequency multiplexed mastergroup signals.

The system block diagram of the tube coder is shown in Fig. 1. The signal is sampled and held and applied to the deflection amplifier which delivers a balanced signal to the two deflection plates of the tube. During the latter part of the hold period the beam is turned on by the grid driver. During this time the parallel PCM output is amplified by linear sense amplifiers and directed to the parallel to serial converter. The serial PCM signal is finally regenerated by the output regenerator.

In the following sections the design considerations of these blocks are discussed in greater detail.

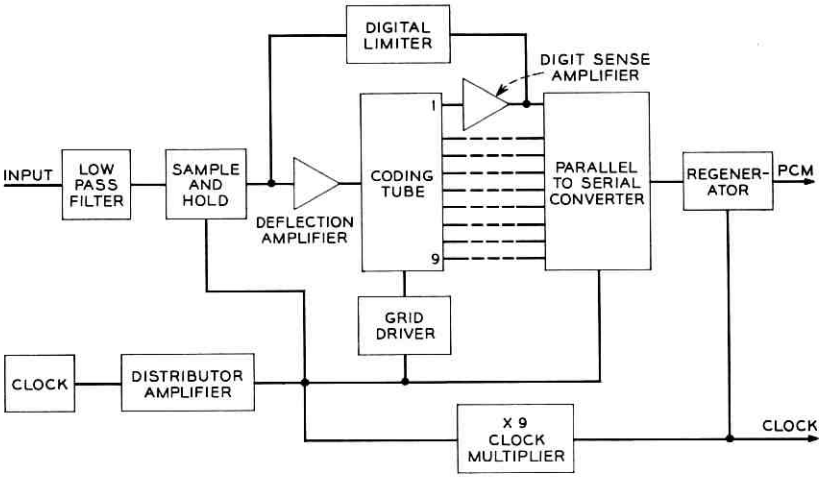


Fig. 1 — Block diagram of tube coder.

2.2 *Coding Tube Description*

As illustrated in Fig. 2(a), a triode electron gun generates a ribbon-shaped beam about one-half inch wide. An electrostatic objective lens system focuses the beam to an average thickness of 2 mils. By means of a pair of tilt electrodes the horizontal orientation of the beam can be controlled by application of an external correction voltage. The analog sample is applied to a pair of vertical deflection plates to direct the beam to the corresponding code position on the code plate. The code plate is perforated with nine vertical columns of apertures (Fig. 2b). The pattern of apertures in each column represents a digit position in a nine-digit Gray code. That portion of the beam which intersects the code plate at an aperture position in a particular column penetrates to the target block and generates secondary electrons. These electrons are

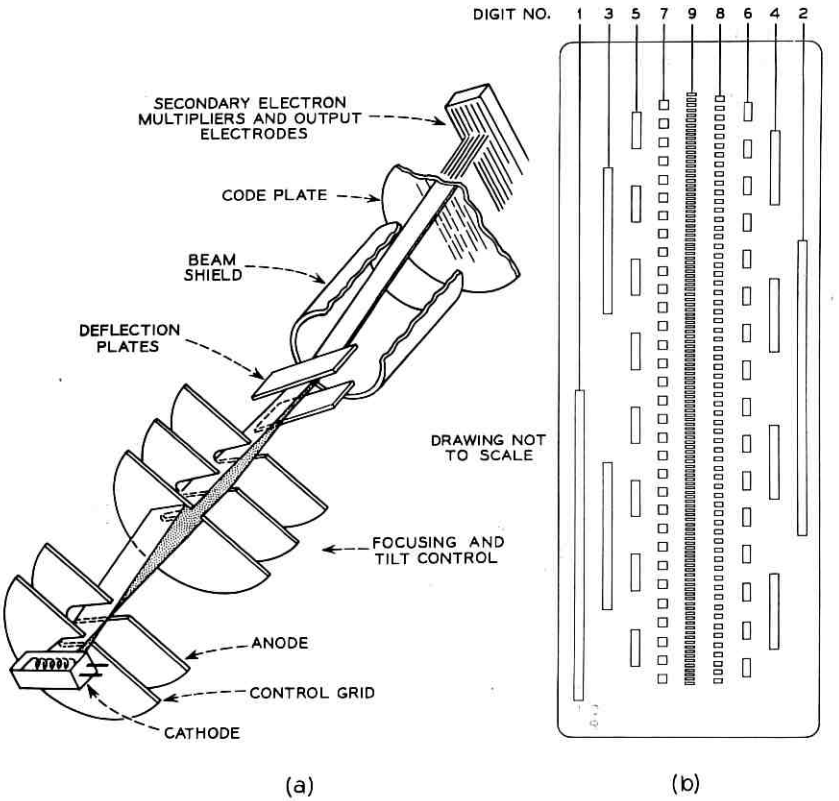


Fig. 2 — Schematic representation of (a) coding tube, (b) 9-digit code plate.

collected by the vertical collector wire assigned to each column. Current flows into an external load and represents a binary ONE in that digit.

A more extensive description of the mechanical details and assembly techniques is given elsewhere.[5]

2.3 *Critical Parameters and Impairments*

2.3.1 *Deflection System*

The two deflection plates form a balanced electrostatic deflection system. Drive voltages of equal magnitudes and opposite polarities are applied to the plates. A peak-to-peak magnitude of 30 volts is required at each plate in order to deflect the beam over the entire code plate. The equivalent capacitance of each plate (measured from plate to ground) is 15 pF.

To avoid coding errors due to external stray electric fields, the beam is shielded in the area between the deflection system and the code plate by a concentric shield. The entire tube is externally encased by a Mu-metal envelope to eliminate interference from stray external magnetic fields.

2.3.2 *Electron Beam*

The accuracy of the coding process is adversely affected by imperfections in the electron beam. Beam thickness, uniformity of current density distribution across the width of the beam, shape, and horizontal orientation are critical characteristics to be considered. Impairments in these parameters are referred to as static or dynamic imperfections. A static impairment implies a fixed deviation regardless of vertical position of the beam on the code plate, while the magnitude of a dynamic imperfection is dependent on the beam position.

2.3.2.1 *Focusing*

The beam has a normal current density distribution with standard deviation, σ. An important parameter is the ratio W/σ where W is the length of the smallest apertures in the code plate, i.e., those corresponding to the ninth digit (Fig. 2b). When the center of the beam is positioned in the center of the aperture, the digit output current I_0 is a maximum and, conversely, I_0 is a minimum when the beam is positioned midway between two adjacent apertures. In Fig. 3, the digit output current is plotted for the ninth digit as a function of beam position with W/σ as parameter. As W/σ increases, $I_{0_{max}}$ reaches a saturated value

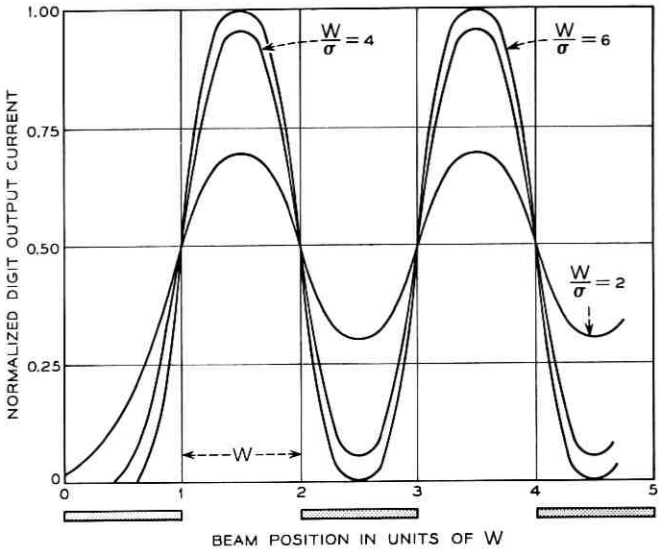


Fig. 3 — A plot of digit output current as a function of beam position. The aperture pattern is sketched below the abscissa.

and $I_{0_{min}}$ approaches zero with an increasingly steep transition between the two. This affects the final binary decision process which is performed by the output regenerator following the parallel to serial converter.

This decision circuit has a threshold level which, referred to the output level of the tube, is positioned half-way between $I_{0_{max}}$ and $I_{0_{min}}$. In practical regenerators this threshold level is the center of a finite uncertainty region bounded by $\frac{1}{2}(I_{0_{max}} - I_{0_{min}}) \pm \Delta I$. Decisions for input amplitudes which cause the digit output to fall in this region are uncertain and may result in partial output pulses, i.e., pulses having low amplitude or deteriorated rise, fall, and duration times. The likelihood of errors is further increased if the noise contributed by the digit output amplifiers to the input signal of the regenerator is considered. The implications of these effects are twofold.

First, at each transition between adjacent codes, there exists a narrow region in the input signal range over which errors are possible in those digits that undergo a change from ZERO to ONE or vice versa. A code which has the property that only one digit is changing at the transition between any two adjacent code words exhibits errors limited to less than one quantum step. For this reason, the Gray code was chosen over the straight binary code. The error is limited to less than one step and occurs

with equal probability at all 512 code transitions. The noise generated has essentially flat frequency spectrum, similar in nature to quantizing noise. The enhancement in quantizing noise has been computed for a signal with Gaussian statistics, for a regenerator with an uncertainty region of $2\Delta I/(I_{0_{max}} - I_{0_{min}}) = 0.1$, 20-dB peak signal to rms noise ratio in the digit output amplifiers, and various values of W/σ. For $W/\sigma \geqq 4$, a performance level that was readily achieved, the increment in quantizing noise is less than 0.5 dB.

A second consequence of indecision in the regenerator, and partial outputs in particular, relates to Gray code to binary code translation errors and is discussed in Section 5.6.3.

A practical coding tube usually exhibits some degree of dynamic defocusing. Consequently, at initial installation, σ is measured at various positions of the beam on the code plate for a range of external focusing bias voltages. From these data, a focusing voltage is established such that the average focus over the entire code plate is optimum. For the models of the coding tube that were used in the experimental system, average W/σ ratios in excess of 4.5 were achieved.

The beam can be focused most accurately in the center region. Defocusing effects near the edges of the beam are minimized by making the code plate narrower than the width of the beam. Also, the digits on the code plate are arranged such that those digits undergoing the most frequent transitions (digits nine, eight, etc.) are positioned in the center while digits one and two, which change less frequently, are located near the edges (Fig. 2b).

2.3.2.2 *Uniformity of Beam Current Density*

If the current density varies across the width of the beam, the digit output currents will exhibit different amplitudes. Static nonuniformity is compensated by an adjustment in each of the digit output amplifiers. This adjustment assures that the threshold for the common serial regenerator can always be set at one-half peak output current for all digits. In the tubes for the experimental system the average value of peak output current was 2.5 μA and never varied more than 10 per cent from digit to digit; usually the ninth-digit output current was lowest in peak-to-peak amplitude due to the small aperture of digit nine in the code plate. External compensation for dynamic nonuniformity in current density was not provided. The digit output current of any particular digit never varied more than ± 5 per cent as the beam was swept across the entire code plate.

2.3.2.3 *Beam Tilt*

The electron beam intersection on the code plate forms a line which should be precisely at right angles to the aperture columns. Any deviation from this angle is tilt. Static tilt can be corrected by applying an external bias voltage to the tilt correction electrodes. External dynamic tilt correction was not provided. A procedure similar to the one used for focusing is followed. A tilt correction voltage is established which minimizes the average tilt over the entire code range. After this correction is applied the average tilt usually turns out to be zero.

2.3.2.4 *Beam Bowing*

This term denotes any deviation from a straight line for the intersection of the beam with the code plate. The tubes have no provision for external correction of static or dynamic beam bowing. However, in the tilt adjustment procedure the effects of bowing can be included in the averaging process since both tilt and bowing are geometric beam imperfections and cause similar coding impairments.

2.3.3 *The Collector System*

The collector system consists of the target block with nine vertical slots in which the collector wires are located coaxially. The slots serve as an electric shield between the collector wires guarding against interdigit crosstalk.

Electrically, the digit output is represented by a current generator of 2.5 μA shunted by capacitance of 5 pF. The coupling capacitance between adjacent digits is less than 0.25 pF. The interdigit crosstalk into a particular digit from the two adjacent digits is down by more than 20 dB.

2.4 *Deflection Amplifier*

2.4.1 *General*

The design specifications for this amplifier are more severe for video coding than for mastergroup coding. For this application the principal design objectives were as follows:

(1.) The amplifier should have single-ended input and two equal-magnitude, opposite-polarity outputs, each driving a deflection plate represented by a 15-pF capacitance to ground.

(2.) Voltage excursion on each plate for full end-to-end beam deflection must be 30 V peak-to-peak.

(3.) Voltage gain from input to each output should be 18 dB and stable to within ± 0.05 dB.

(4.) The transient response of the amplifier must be such that the output settles to within about 0.02 per cent (1/10 quantizing step) of its final value 60 ns after application of the sampled and held input.

(5.) Gain must be maintained to dc in order to utilize full coding range for television signals that are clamped at the horizontal sync pulse.

(6.) The input impedance of the amplifier should be 500 Ω.

For mastergroup coding (about 6-Mc/s sampling rate), the settling time could be relaxed to 100 ns and the amplifier may be ac-coupled as long as sufficient low frequency gain is provided to prevent any appreciable droop-off during the holding period.

2.4.2 *Circuit Realization*

The simplified circuit configuration of the amplifier is shown in Fig. 4. It consists of two identical forward amplifiers with common input and different feedback connections, such that the two outputs are equal in amplitude and opposite in phase.

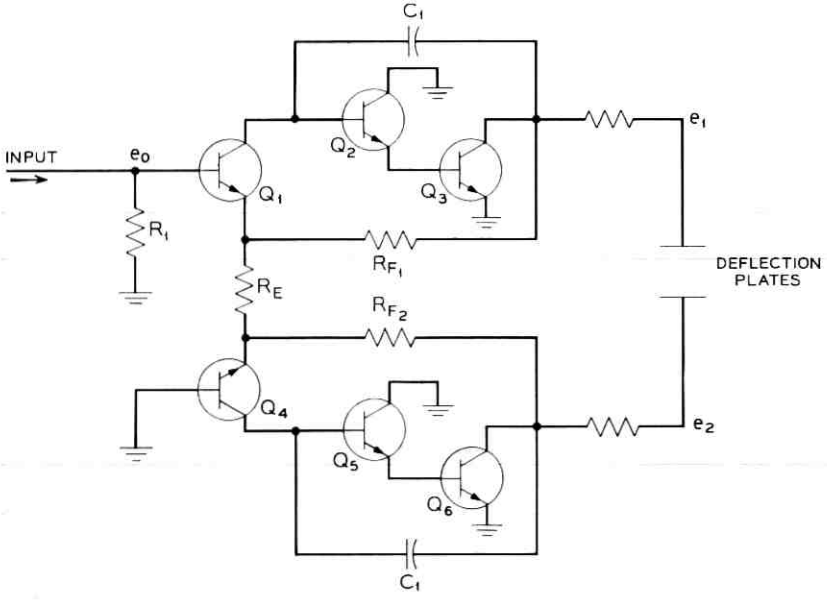


Fig. 4 — Simplified schematic of coding tube deflection amplifier.

From the feedback network the voltage gains from the input to the two outputs are established:

$$A_{V1} = \frac{e_1}{e_0} \doteq 1 + \frac{R_{F1}}{R_E}$$

$$A_{V2} = \frac{e_2}{e_0} \doteq -\frac{R_{F2}}{R_E}.$$

To make $e_2 = -e_1$,

$$R_{F2} = R_{F1} + R_E.$$

The open-loop gain-frequency characteristic is shaped with the transient response requirements of the amplifier in mind. The settling time objective is met by designing the frequency response of the open-loop current gain to have a well controlled slope of 20 dB/decade to at least an octave beyond a unity gain crossover frequency at 25 Mc/s.

For minimum gain transistors, the amplifier has an open-loop current gain of 54 dB at dc. This assures that the output settles to a steady state value with an accuracy of better than ± 0.05 dB. This loop gain is maintained up to about 50 kc/s and then falls off with the desired 20 dB/decade slope, controlled by the local feedback capacitor C_1. This is readily verified by considering the loop current gain of the lower amplifier in Fig. 4.

Breaking the feedback loop at the emitter of Q_4, the loop gain in the frequency range from 30 kc/s to about 50 Mc/s is

$$A\beta \doteq \frac{A_i Z_T}{R_{F2}}$$

where

A_i = common base current gain of Q_4
Z_T = transfer impedance from base of Q_5 to collector of Q_6.

Since Q_4 is a transistor having a gain-bandwidth product of greater than 800 Mc/s, A_i is not dependent on frequency in the range of interest and has a value near unity. The transistors Q_5 and Q_6 have gain-bandwidth products in excess of 500 Mc/s. Hence, Z_T is determined by C_1 and $A\beta$ becomes

$$A\beta(p) \doteq \frac{1}{pC_1R_{F2}}.$$

This produces the desired asymptotic performance with a unit gain frequency crossing at

$$\omega_0 = \frac{1}{C_1 R_{F2}}.$$

The 20 dB/decade slope was maintained up to about 50 Mc/s and the open loop gain then fell off with a final slope of 60 dB/decade. Since series feedback is used at the input, the input impedance of the amplifier is set primarily by resistor R_1.

The quiescent dc voltage at the output of one amplifier was fixed at -6 V. The other amplifier contains a bias control which varies the output dc voltage from -5.5 V to -6.5 V and permits external correction for any vertical misalignment of the beam. By means of this control the beam is positioned at the center code on the code plate with no input signal. Differential dc stability of the deflection amplifier outputs is important to maintain the beam at this center position. Long term differential dc stability corresponding to ± 3 quantizing steps has been achieved in the amplifier.

The physical separation between the amplifier outputs and the deflection plates is critical. Excessive lead inductance causes undesirable resonance effects. The amplifier was constructed with this in mind and installed adjacent to the tube deflection plates so that the external lead length did not exceed one inch.

2.5 *Digit Output Amplifiers*

2.5.1 *General*

The coding tube delivers a relatively low level digit output signal. The outputs are cosine-squared current pulses of 2.5-μA peak amplitude and 40-ns base width. The maximum repetition rate is about 12 Mc/s for video coding. Since it is undesirable to make binary decisions at such low levels, the output signals are amplified by linear digit output amplifiers ahead of the decision circuit.

The digit output amplifier is comprised of a low level preamplifier, which raises the signal level to 50 mV, followed by a postamplifier which delivers a 2-V peak-to-peak signal to the parallel to serial converter. Because of the low input current, it was desirable to ac-couple the stages in the preamplifier and to include a dc-restoration circuit in the postamplifier.

2.5.2 *Preamplifier*

The preamplifier is a conventional feedback amplifier consisting of three common-emitter stages and an over-all shunt-shunt feedback

network which stabilizes the input-output transresistance to 20 kΩ. The collector wires from the tube are each directly connected to the low-impedance input summing node of the corresponding amplifier. The output drives a 93-Ω coaxial cable which is terminated at the input of the postamplifier.

The amplifier has more than 32-dB loop gain at midband. The 3-dB closed-loop bandwidth of the amplifier is 25 Mc/s. This is adequate to amplify the cosine-squared pulses with only minor deterioration in wave shape.

As shown in Fig. 5, the preamplifiers are mounted directly at the head of the tube, where the collector wire output pins are located, to avoid interference and noise problems.

2.5.3 *Postamplifier*

This amplifier is similar to the preamplifier except that it is designed for higher levels and contains provisions for dc restoration. The voltage gain is feedback stabilized to 32 dB over a 3-dB bandwidth of 40 Mc/s. The input, which has an impedance of 93 Ω, is ac-coupled to the preamplifier. The amplifiers drive negative-going pulses with 2 V peak-to-peak amplitude into a 2-mA diode AND gate in the parallel to serial converter.

The base line of the output signal is restored by the dc-restoration

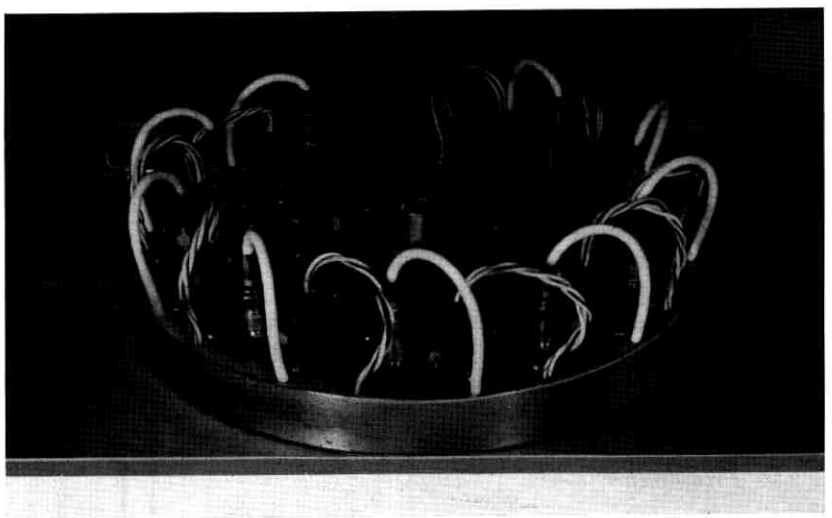


Fig. 5 — Digit output amplifier assembly.

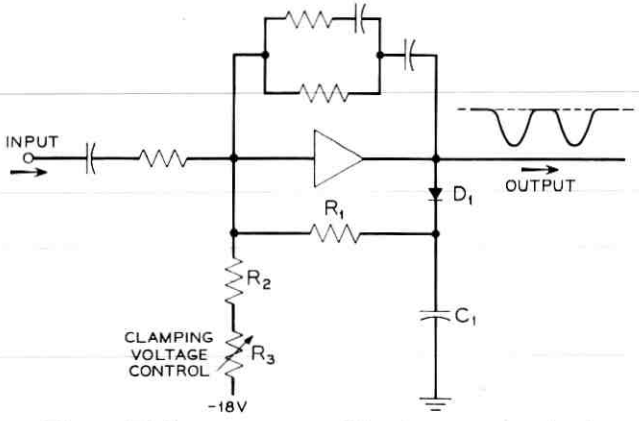


Fig. 6 — Digit output postamplifier dc-restoration circuit.

circuit to +1 V. This circuit, shown in Fig. 6, consists of diode $D1$, capacitor C_1, and resistors R_1, R_2, and R_3. The operation of this circuit is as follows:

In the absence of an input signal the dc value of the output is at the clamping voltage of +1 V. Diode D_1 is conducting and delivers positive current through R_1 into the summing node of the amplifier so as to maintain this steady-state condition. When an input signal is applied, output voltage excursions above +1 V are integrated by the capacitor C_1. This increases the dc feedback current to keep the output signal excursions below the clamping voltage.

The clamping voltage can be shifted slightly about +1 V by means of potentiometer R_3. With this control each digit output can be positioned symmetrically about a common threshold voltage which in this system is zero volts. As mentioned in a previous section, nonuniform beam current density across the width of the beam causes slightly different peak-to-peak amplitudes in the digit output pulses. The R_3 adjustment keeps the decision threshold level halfway between peak and base line on each digit.

2.6 *Associated Tube Coder Circuits*

2.6.1 *Dc Biasing*

The dc bias voltages for the tube are derived from a conventional voltage divider circuit including resistors and zener diodes. A single high-voltage supply and a filament supply are the main sources of power.

In order to avoid isolation problems in the digit output amplifiers, the digit collector wires are at ground potential and the cathode is negative. The voltage divider supplies bias voltages to the following electrodes:

(1.) cathode
(2.) control grid
(3.) focus electrodes
(4.) target block.

The electron acceleration voltage is approximately 800 V. The control grid and focus electrode voltages are externally variable. The heater power supply is floating at −800 V to prevent filament to cathode breakdown.

2.6.2 *Control Grid Driver*

The control grid is turned on towards the latter part of the holding interval by means of a 12-V peak-to-peak drive signal generated by the grid driver circuit from the sampling clock. The positive peaks of the drive signal are held flat for a period of 20 ns. The flat portion of the drive signal is desirable because at this voltage level the beam focus, which is dependent on the grid voltage, is optimized.

The input capacitance of the control grid is 16 pF. The driving circuit is a conventional two-transistor saturated amplifier which is coupled to the grid by means of a capacitor. Because the grid is at a high negative potential, the coupling capacitor has high-voltage breakdown requirements.

2.6.3 *Digital Limiter Circuit*

A PCM coder is expected to exhibit perfect limiting for an input signal which extends beyond either of the two extreme code levels. For a Gray code, one extreme level is represented by nine ZEROS, and the other by a ONE followed by eight ZEROS. In the tube coder there exists a problem with the latter. For overloads limited to several quantizing steps in magnitude, the tube coder continues to generate a ONE followed by eight ZEROS code because the aperture corresponding to digit ONE is extended somewhat beyond the extreme level. However, for excessive overload peaks the beam is deflected completely off the code plate and, in effect, generates the nine ZEROS code, which is in error by the full coding range. The resulting input-output characteristic is illustrated in Fig. 7.

This undesirable feature is eliminated by the digital limiter which is

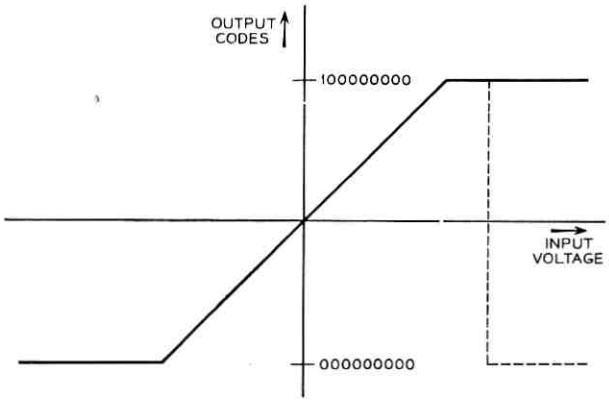


Fig. 7 — Input-output characteristic of tube coder. The dotted curve shows the transfer characteristic that results if the digital limiter is omitted.

connected as shown in Fig. 1. This circuit effectively bypasses the coding tube and forces the first digit to a ONE in the range where the problem exists.

In the Gray code the first digit is essentially the sign digit and is a ONE for the entire positive range of the signal, from level 256 to 512. The digital limiter is a wide-threshold circuit which is connected to the sample and hold output and determines whether the signal is above or below the midpoint of the positive range. If the signal is above positive half range (level 384), it forces a ONE in the output; if the signal is below, the digital limiter is effectively out of the circuit. The critical decision operation which takes place when digit one is in transition, i.e., when the signal is just changing polarity, is still performed exclusively by the coding tube.

2.7 *Performance*

The only meaningful tests of the performance of a coder are: (1.) a measurement of quantizing noise, and (2.) a measurement of errors that have large amplitudes but may occur infrequently enough so that quantizing noise is not appreciably affected. These errors may have severe subjective effects such as the appearance of occasional black or white dots on a television display.

Quantizing noise tests have to be executed in real time in conjunction with a decoder and all the associated circuitry. It is, therefore, rather difficult to establish precisely the contribution from the individual circuits to the over-all impairments. This is especially true when a

performance level very close to theoretical is reached. What circuit has to be improved when the actual performance is only a few dB below theoretical? This posed challenging questions during the experimental PCM program. However, it is most likely that the major contribution to the deviations from theoretical performance originate in the coder. Not only does this circuit perform the most difficult and critical operation, but it is also the circuit where individual tests are least likely to have meaningful results.

The tube coder was operated at 6-Mc/s and 12-Mc/s sampling rates. The theoretical and measured noise performance for 6-Mc/s mastergroup operation are shown in Fig. 8. For high input signal levels overload is controlling, and for low levels quantizing noise diminishes. Optimum loading is attained when the rms value of the mastergroup signal corresponds to about one-eighth of the peak-to-peak coding range. Under this condition, the signal-to-noise ratio is maximum and is less than 2 dB away from theoretical performance. The measurements were performed with a noise loading test set which loads the coder with Gaussian distributed bandlimited noise, and measures the total quantizing noise falling in an initially signal-free 3-kc/s channel slot. At the 12-Mc/s sampling rate, the tube coder noise performance was less than 2 dB above theoretical quantizing noise. At both sampling rates, the tube coder met the performance objective which was set at the outset of the experimental program. The performance level of the coder in regard to single errors, that are subjectively undesirable, is discussed in Section 5.6.3.

III. THE SOLID STATE CODER

3.1 *Basic Plan of Operation*

The general plan of the solid state coder was suggested by F. D. Waldhauer[6] and is an improved design based on an earlier coder described by B. D. Smith.[7] A Gray code is generated by a cascade arrangement of binary stages. Each of the binary stages must perform two functions: it must (1.) decide and report whether the applied signal is positive or negative, and (2.) deliver a residue signal to the next stage for further processing. Full-wave rectification of the input, and addition of a suitable bias so the residue ranges between equal plus and minus values, is a desirable way of presenting the residue.

Fig. 9 illustrates this process. Fig. 9(a) shows the transfer characteristic of a full-wave rectifier with gain of two. In the (b) portion of the

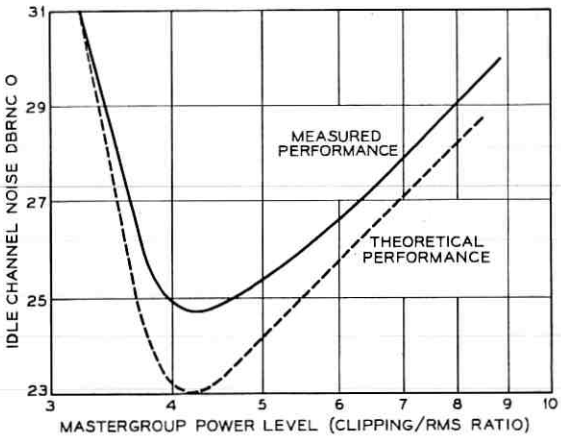


Fig. 8 — Noise performance of mastergroup codec.

figure, a one unit reference has been subtracted from the characteristic shown in (a). The digit output is a ONE when the input is positive and a ZERO when the input is negative. The amplitude of the residue ranges from -1 to $+1$. Similar stages in cascade can generate successive Gray-code digits. As in the parlor game of twenty questions, each binary decision narrows the range of uncertainty. When nine digits are known, the input can be specified within one of 512 small ranges, often called steps.

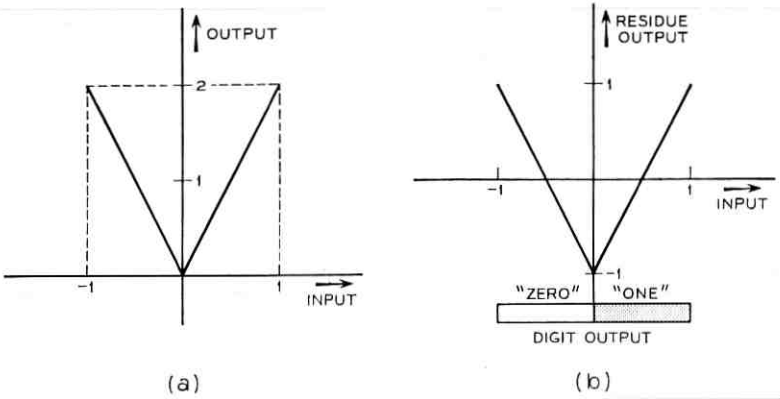


Fig. 9 — (a) Full-wave rectifier characteristic with gain of two, (b) characteristic biased with reference of one unit.

3.2 *Need for Sample and Hold*

If the input signal changes slowly with respect to the speed of the coder stages, the code corresponding to the input signal can be read out of the coder stages at any time. However, when the signal changes rapidly, the stages are unable to follow accurately and incorrect codes are generated. In order to avoid excessive speed requirements on the coder stages, the signal is sampled briefly and the sample is held constant at the input of the coder for the remainder of the interval available for coding. About one-fourth of the period (21 ns) is allowed for sampling the signal and charging a capacitor to a proportional value. The capacitor holds this value for the remaining 60 ns and applies it to the input of the coder.

3.3 *General Method of Coding*

Waldhauer[6] proposed that nearly ideal rectification can be attained by use of diodes in the feedback path of a high-gain operational amplifier. Two feedback paths with oppositely poled diodes give two half-wave rectified outputs. If the feedback is large and the diodes do not conduct in the reverse direction, the rectification characteristic is independent of diode forward drop. To obtain full-wave rectification, an inverting amplifier can be used to invert one of the two outputs. When this plan is used, the number of cascade amplifiers traversed by a signal varies with the value of the sample to be coded. This causes difficulties at high speed. Waldhauer saw that this problem was eliminated if balanced signals were fed to balanced coder stages. A typical stage is shown as Fig. 10. Two operational amplifiers are provided with rectifiers in the feedback paths. Equal and opposite input currents are applied to the two amplifiers.

Combining the outputs and adding reference currents provides balanced residues suitable for driving a following stage. A digit output circuit connected to the outputs of the operational amplifiers senses the polarity of the signal and generates a clearly defined positive or negative output with negligible chance of indecision. Figs. 10 and 11 illustrate the currents in the coder stage for positive and negative inputs.

A different circuit (Figs. 12 and 13) is used for the first stage. It is capable of taking a single ended input and producing a balanced residue suitable for driving a stage such as Figs. 10 or 11. This circuit uses partial cancellation to yield the full-wave rectifier characteristic. Figs. 12 and 13, show the circuit conditions when the maximum positive and negative inputs, $+E_M$ and $-E_M$, are applied. The delay cable is

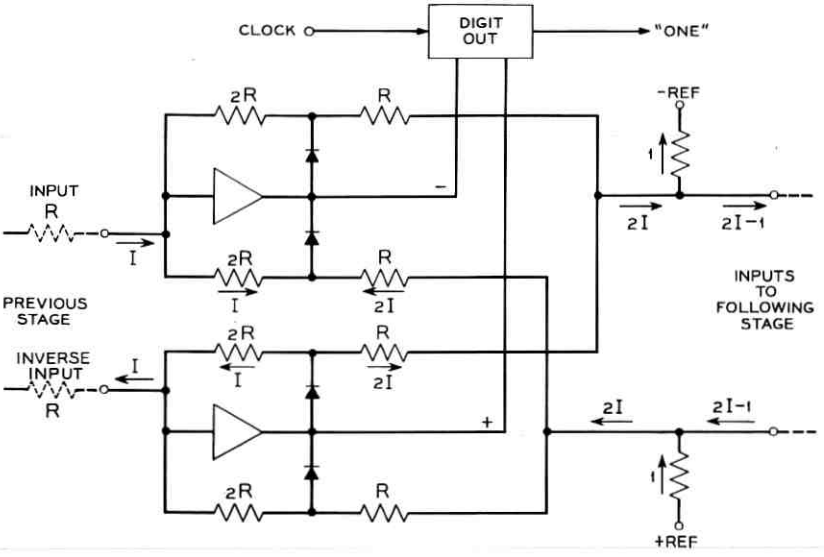


Fig. 10 — Typical solid-state coder stage, positive input.

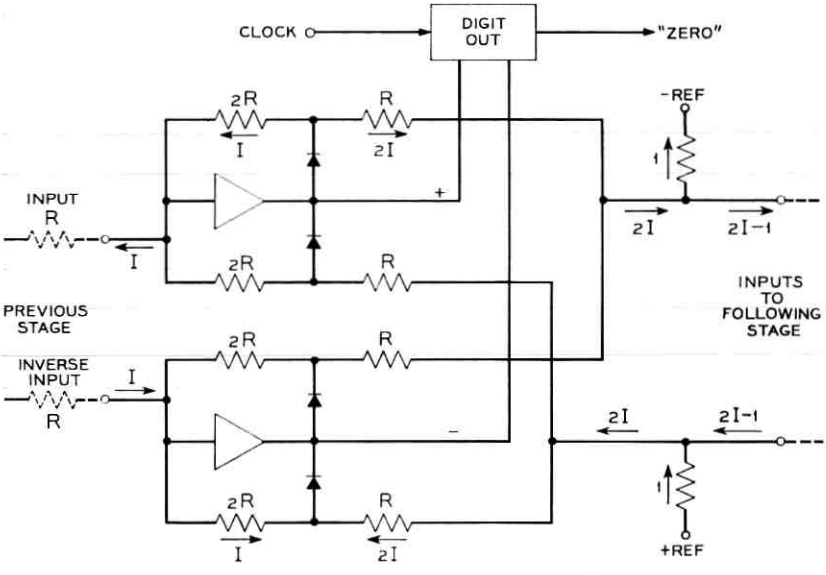


Fig. 11 — Typical solid-state coder stage, negative input.

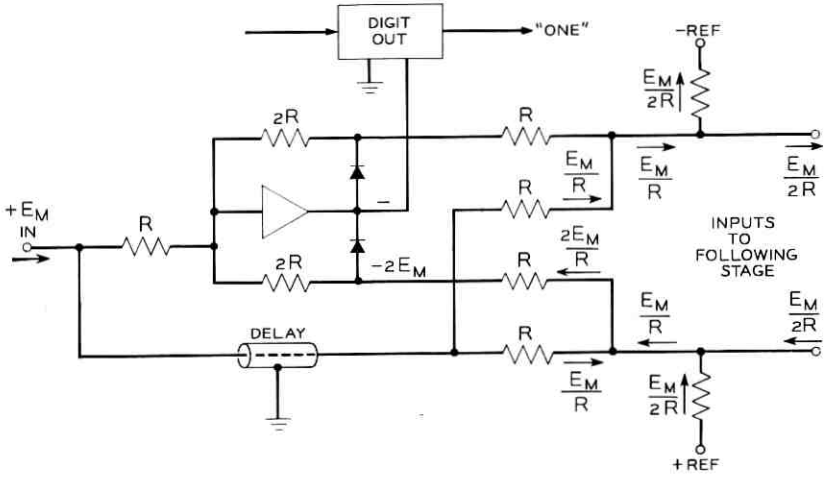


Fig. 12 — First coder stage giving balanced residue output from unbalanced, positive input.

selected empirically to match the delay of the operational amplifier. Best balance of the total output is the criterion used for delay adjustment.

When this stage is used at high speed, the operational amplifier must have a transient response such as to settle very nearly to the final steady-

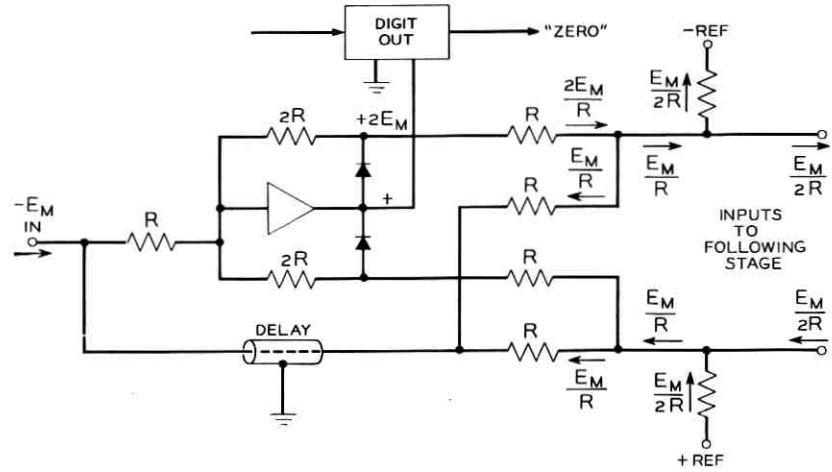


Fig. 13 — First coder stage giving balanced residue output from unbalanced, negative input.

state value in the allowed 60-ns interval. This requires a loop gain which slopes off uniformly at 20 dB/decade and has a unity-gain crossover frequency as high as can be attained. A unity-gain crossing at 80 to 90 Mc/s has been realized. This gives a settling time constant of 2 ns. The gain may roll off more rapidly above the unity-gain crossover frequency without seriously deteriorating the transient response. The amplifiers used in this coder have a minus three slope (60 dB/decade) which sets in about half a decade above the unity-gain crossover frequency. Deviations from the desired unit slope in the three decades below the unity-gain crossover frequency are more serious. They can produce relatively slow transients that are not of negligible amplitude.

An additional high-speed problem is introduced by the transition from one feedback path to the other when the output changes sign. The forward drop required to cause the decision diodes to conduct requires the amplifier output voltage to change by twice this value before current can be diverted from one diode to the other. If the input changes from a large value to a small value of opposite polarity, the amplifier output moves exponentially until one diode cuts off. Next, the output moves linearly at the same rate until the opposite diode conducts. Then the exponential approach to final value is resumed. Time required to cross the midregion may be longer than a sampling interval if the diode drop is large in relation to the signal swing. This difficulty can be reduced only at the expense of some reduction of static accuracy. The means for improvement is to bias the diodes in the forward direction so that, with no input to the coder stage, some current flows in each of the decision diodes. The feedback path does not fully open and the change of output voltage required to cross over is much reduced. The rectification characteristic (Fig. 9a) is rounded at the point of the "VEE" and does not reach zero. This makes the decision problem more difficult for the digit output circuit, and inputs near the decision point deliver a less accurate residue to the succeeding stages.

3.4 *Computer Analysis of Coder*

The result of forward bias on coder performance was studied by simulation on the digital computer assuming amplifiers with an ideal unit slope which becomes a three slope starting half a decade above the unity gain point. The diodes were assumed to follow the ideal exponential law. These studies show that forward bias improves transient response at the expense of static accuracy and makes a very large reduction in rms error of coding. All stages were assumed to have the same bias. Because of coder stage gain, this means that maximum error

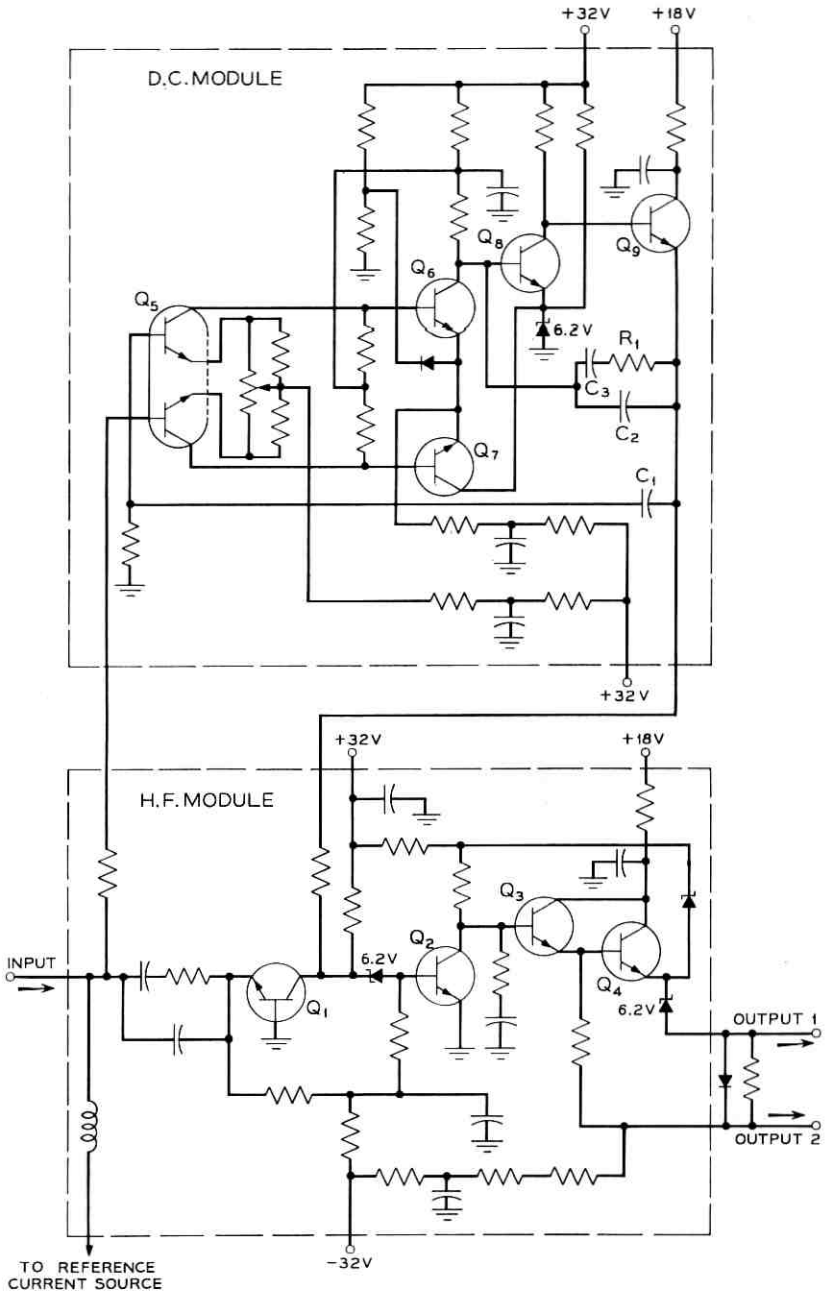


Fig. 14 — The composite operational amplifier.

occurs in the two decision points adjacent to the center transition. Bias in the first stage does not inherently introduce error in the center decision or most significant digit, though it does make the decision problem more difficult and so increases the random error and drift at this decision point. The errors caused by bias, expressed as a fraction of a step, tend to be halved in each successive stage. There are, of course, twice as many opportunities for error in each successive stage if all input amplitudes are equally likely. Thus, for a signal with flat probability distribution, the error power contribution of all succeeding stages would equal that of the first. With a normal distribution, as provided by a mastergroup of single sideband telephone channels, the first stage error would dominate.

For the normal distribution of input signals, the optimum bias is such as to move the decision points adjacent to the center by about one-half step. Increase of bias to eliminate the two center codes, or even four center codes, does not cause serious deterioration of system performance if judged by rms error.

3.5 *Circuit Design of the Operational Amplifiers*

Circuit design features of the operational amplifiers were discussed by F. D. Waldhauer.[10] If the desired accuracy of transient settling is to be attained, the open-loop gain must maintain substantially unit slope over a range of about six decades below the frequency of unity transmission. Some pole-zero pair cancellations may be tolerated in the lower three decades but the upper three decades must be extremely clean. The amplifier must also be nearly free of drift and noise. The design chosen is such that a current of 13 μA applied at the input represents a change of one code step. Drift resulting from temperature and aging over a reasonable maintenance interval must be less than one-tenth of this value.

Stable, low-drift, dc transistors do not have good wideband performance. Thus, the problems are solved separately and suitable circuits are joined as shown on Fig. 14. The high-frequency transmission path is through a four-stage amplifier comprising a common-base stage, a common-emitter stage, and a compound common-collector stage. This ampfilier has the desired single phase reversal at low frequency and from about 100 kc/s to over 100 Mc/s the amplification varies inversely with frequency. The collector-to-base capacitance of Q_2 is dominant in controlling the gain slope. Bootstrapping the power feed resistors to the output preserves the very high input impedance of the compound common-collector output stage.

In broadband amplifiers, excess phase caused by transit time is very important. Each transistor tends to add about 0.1 ns. In the cordwood modules used in the first coder, wiring length added about three inches which translates to 0.3 ns. Later models used multiple-chip integrated packages containing the four high-speed transistors and two breakdown diodes providing a 6.2-V drop. The shorter path length and better controlled stray capacitance resulted in substantially smoother high frequency transmission characteristics.

To continue the gain slope below 100 kc/s and to provide low dc drift, a dc amplifier takes over transmission from the common-base stage. The first stage is a differential pair of matched transistors contained in a single encapsulation. Mounting the two transistor chips in one package minimizes differences of V_{BE} with variation of temperature. This stage is followed by a second balanced stage using large common-emitter resistance to maintain constant total current. One collector of the second stage drives the base of a common-emitter third stage. The output stage is common-collector.

Variations of V_{BE} in transistors Q_6 and Q_7 are not very important, provided they are less than 100 mV. Therefore, mounting the transistors in a common encapsulation is not necessary. Drift in the balance of current input required by the second stage is very important, because any such unbalanced input must be provided by the first stage. For this reason the operating current is reduced to about 6 μA for each transistor. Reduction of the total current also reduces the possible unbalance. The transistor was chosen to maintain large current gain at very low current. The first stage operates from a relatively low source impedance so current gain is less important and V_{BE} is very important. Current to each half of Q_5 is about 50 μA. With this design, the initial balance takes care of dissymetry in transistors and other components and subsequent drift is small in terms of a peak input of 3.3 mA to a coder stage.

The feedback capacitor C_1 gives the dc amplifier a unit slope for about four decades from 100 c/s to over 1 Mc/s. The gain value is such that the transmission through the dc amplifier, and that of the common-base stage of the high-frequency amplifier merge at about 100 kc/s and produce a zero to match the pole of the high-frequency amplifier.

If the gain of the dc amplifier continued to fall off smoothly with a slope of 20 dB/decade, no further compensation would be needed in the mid-frequency region. The transistors used in the first stages are of limited bandwidth so the gain slope must increase to 60 dB/decade at about 2 Mc/s. Even though the high-frequency path dominates in this frequency region, the change of slope of the low-frequency path

produces a small doublet in the over-all transmission characteristic and can cause noticeable degradation of transient behavior. Local feedback by capacitors C_2 and C_3, and resistor R_1 reduces this effect, but it may still be significant. Further study is needed.

3.6 *Digit Output Circuits*

When the coder transients have settled, it is necessary to read out the states of the individual stages and generate a corresponding code word. The code is read out on nine parallel lines approximately simultaneously. These outputs are brought into time coincidence by suitable lengths of cable and directed to the parallel to serial converter.

It is important that the code pulses delivered shall be definitely binary in nature. Indecisive or partial pulses can result in serious errors in Gray to binary translation and decoding. The process of forcing a decision is commonly called regeneration. It is usually done by the use of trigger action involving negative resistance in some form. For best results, the negative resistance circuit should be very fast compared to the time allotted for decisions. For this reason, it has been placed in the digit output circuits, where the code is handled in parallel and more time is available for decision.

The circuit now in use is shown is Fig. 15. The differential input stage can be driven balanced or single sided. Suitably biased limiter diodes in the collector circuit prevent saturation of the transistors on the one excursion and limit the positive swing on the other. The input impedance presented to the coder stage is a few hundred ohms for small signals near the decision point and much higher for inputs of 50 mV or more regardless of polarity. It does not load the coder stage.

The second stage (Q_3 and Q_4) serves the dual purposes of amplifier and gate. The emitters are held at a static potential of about +6.4 V by a bias regulator circuit. This is comparable to the maximum base voltage permitted by the catcher diodes in the first stage so current cannot flow to Q_3 or Q_4 in the absence of additional emitter drive. The 12-Mc/s clock signal is supplied to Q_7, which runs as a limiter, and the small pulse-forming capacitor of 10 pF drives the emitters of Q_3 and Q_4 in a negative direction for a period of a few nanoseconds. A pulse of current flows to one or both of transistors Q_3 and Q_4, depending on the input to the first stage.

The tunnel diode has a peak current of 5 mA and a valley current of about 0.5 mA. Resistors connected to the +32-V and −12-V power supplies together with a center tapped pair of resistors across the tunnel

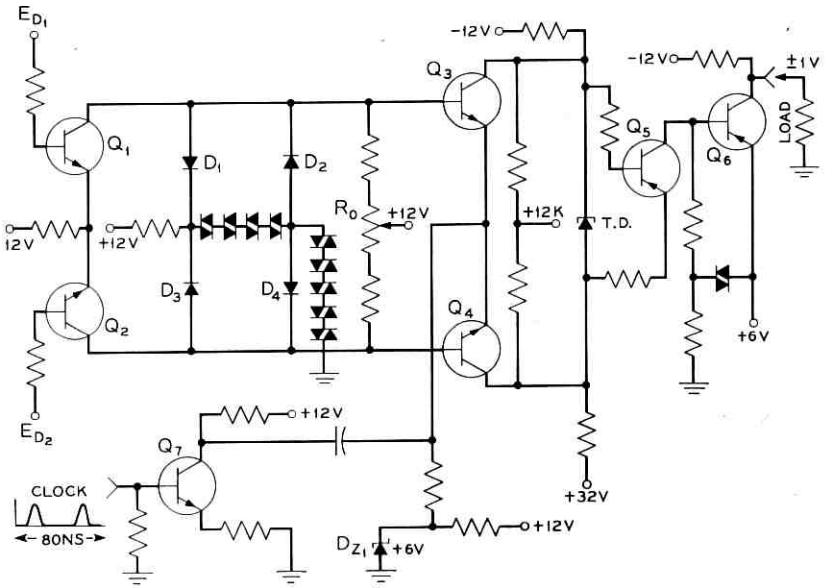


Fig. 15 — Solid-state coder digit output circuit.

diode provide bias equivalent to an 870 Ω, 2.56 V source which is balanced with respect to +12 V. The negative slope resistance of the tunnel diode is much less than 870 Ω so the circuit is strongly bistable.

When the clock pulse turns on Q_3 alone, the tunnel diode is forced into its high-current state. If Q_4 alone is turned on, the tunnel diode goes to the low-current state. If the currents of Q_3 and Q_4 are equal, or differ by too small an amount, the state of the tunnel diode remains unchanged after the clock pulse passes. These marginal inputs do produce a disturbance; but it dies out rapidly. The timing is arranged so that the output is gated into the serializing line in a 5-ns interval immediately preceding application of the next following clock pulse to Q_3 and Q_4. Thus, the tunnel diode circuit has about 70 ns in which to form a decision. The probability of residual indecision is very small. However, to meet system requirements it is necessary to use an additional serial regenerator (as in the case of the tube coder) to further resolve decision uncertainties (see Section 5.6.3).

3.7 *Experimental Results*

The nine-digit coder was built using precision resistors with ±0.02 per cent tolerance in the feedback, feed forward, and reference bias

positions. The dc amplifiers were trimmed by means of the potentiometer between emitters of first stage so that the voltage at the summing node of every stage was less than 50 μV. Static alignment was then attained by trimming the reference resistors of all stages. This was done by applying specified dc inputs corresponding to the successive decision levels and observing the state of the coder stage being adjusted. The work progresses from the first-digit stage to the ninth-digit stage. Initial adjustment was made to a precision of 1/20 step or better. Static accuracy is maintained over a period of months with errors no more than ± 0.1 step in most cases and a maximum of ± 0.15 step. This source of error would not degrade the noise performance by more than 1 dB from theoretical nine-digit performance.

When the coder is operating at a sampling rate of 12 Mc/s into a decoder whose imperfections are known to be small, the measured peak-to-peak signal to rms quantizing noise is 57 dB which is 7 dB more noise than would be expected from a perfect nine digit system. When the sampling rate was cut to 6 Mc/s, corresponding to operation on a mastergroup, the performance was 28 dBrnc0, which is within 5 dB of theoretical nine-digit quantizing noise. The rms error was reduced to about three-quarters of its former value because twice as much time was allowed for the transients to settle. It is believed that imperfect matching of the dc amplifier and the high-frequency amplifier is an important contributor to the noise impairments. A second coder with improved amplifiers is in process of construction.

IV. ASSOCIATED CODER CIRCUITS

4.1 *Sample and Hold*

4.1.1 *General Description of Sample and Hold Circuit*

Fig. 16 shows in block form the method used to carry out the sampling and holding process. The input signal passes through an amplifier whose output impedance is only a few ohms. The amplifier output is switched to the holding capacitor C by driving balanced currents through a diode bridge. When the bridge current is cut off, the charge on the capacitor is held. The capacitor provides an input to an amplifier whose input impedance is about 1 MΩ. With 100-pF capacitance the time constant is 100 μs; so the droop of the held voltage is about one part in a thousand in an 80-ns interval. In the following discussion, the more difficult 12-Mc/s sampling operation for the solid-state coder will be emphasized.

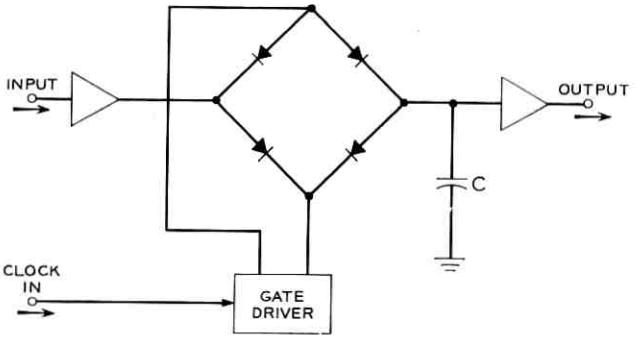


Fig. 16 — Sample and hold.

4.1.2 *Analysis and Requirements*

If the gate driver could supply a constant current to the diode bridge for the sampling period (about 21 ns) and then instantly switch the diodes off and into a state of back bias greater than the signal voltage, perfect operation would depend only on diodes with negligible storage time and very high back resistance. Hot-carrier diodes satisfy these requirements well enough. If the driver changes the bridge voltage slowly, the diodes cut off at different times unless the signal is zero. This results in some change in the charge stored on the holding capacitor during the turn off time. This, in turn, aggravates the transient settling problem of the coder and also produces some nonlinearity in the sampling process. To keep the nonlinear distortion well below quantizing noise, the driver needs to switch the gate off in less than 1 ns. This requirement is documented in the paper by Gray and Kitsopoulos.[11]

4.1.3 *Preamplifier and Postamplifier*

The preamplifier is of straightforward design using shunt feedback to provide constant gain, linearity, and low output impedance. Its load impedance is switched from open circuit to 100-pF capacitance. The preamplifier must be stable under both load conditions and have good transient performance.

The postamplifier is more critical in design. High input impedance is obtained with a transistor amplifier by using series feedback at the input and designing for unity voltage gain. The loop gain varies inversely with frequency (unit slope) up to about 50 Mc/s. This results in good transient response and in an equivalent input impedance of 1 MΩ in parallel with a 10-pF capacitance.

4.1.4 *Gate Driver Pulse Forming Network*

The pulse forming portion of the driver is shown on Fig. 17. Balanced direct currents I_0 are applied through resistors from the plus and minus 32-V supplies. The magnitude of I_0 is 15 mA. This current will flow through the bridge or be diverted through diodes $D5$ and $D6$ depending on the voltage generated by the driver.

Sinusoidal current at the desired sampling frequency of 6 Mc/s or 12 Mc/s is introduced through blocking capacitors C_1 and C_2 and flows through inductors L_1 and L_2 to the pulse-forming diode network. The peak value is 0.6 A. Direct bias current is introduced through L_3 and L_4 with a value of about 0.4 A so that the total current tends to forward bias $D12$ for 25 per cent of the time and to forward bias $D11$ for 75 per cent of the time. These diodes are of the type which exhibit charge storage and snap recovery properties. While $D12$ is conducting in its forward direction, with a peak current of about 200 mA, stored carriers accumulate in the semiconductor. As the current reduces to zero and reverses direction, these carriers tend to maintain conductivity and hold the voltage drop almost constant at the value that existed when it was conducting in the forward direction. This situation persists for about 4 ns with reverse current sweeping out the stored charges. Then, within a period of 0.2 or 0.3 ns, conductivity "snaps" off and reverse current no longer flows. By this time the reverse current has built up to a value of about 160 mA and is still growing since the inductors L_1 and L_2 insure

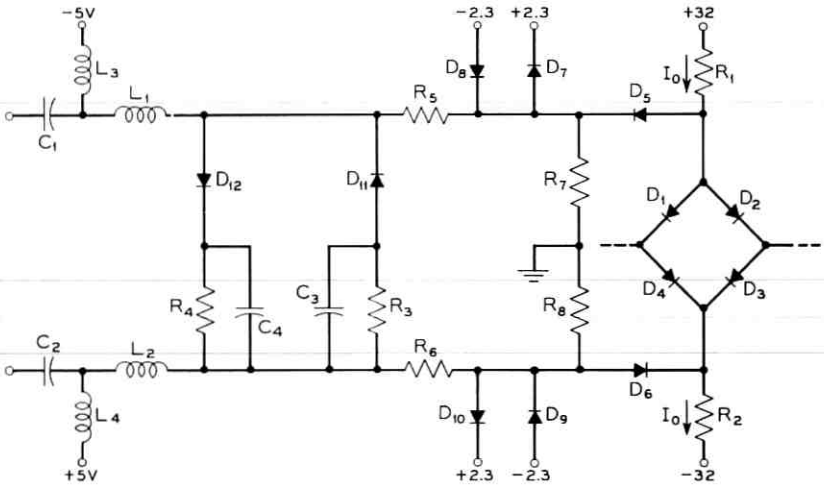


Fig. 17 — Sample and hold gate driver pulse forming network.

that the fundamental component dominates and the harmonics are small. When $D12$ snaps off, the large current is immediately available to charge the stray capacitance (about 4.5 pF) and to overcome the bias voltage and diode drop establishing current in $D11$. $D11$ is also a snap diode but capable of working with larger average current. It operates in the same manner as diode $D12$ except that it serves to steepen the step that takes place when current switches from diode $D11$ to diode $D12$. Average currents flowing through R_3 and R_4, bypassed by capacitors C_3 and C_4, provide back bias on the diodes of about 6 V. Thus, the voltage from line to line must swing about 14 V to transfer conduction from diode $D12$ to $D11$ or vice versa. The more important transition from $D12$ to $D11$ (the onset of the holding portion of the cycle) takes almost 1 ns; the other is a little slower. Diodes $D7$–10, supported by resistors R_5 and R_6, cut the output slightly so the voltage presented across R_7 and R_8 to ground swings ± 3 V. These diodes clean up the top portions of the waveform and serve to establish a low impedance at high frequencies from each side of the bridge to ground while the bridge diodes are in the cut off state. This aids in establishing a large loss from the source to the holding capacitor during the holding interval so that the held voltage is maintained constant.

4.1.5 *Power Amplifier for Gate Driver*

The power amplifier that provides the sine-wave drive is shown in Fig. 18. Design need not be covered in detail. The tuned circuits are not very high Q; so variations of phase shift have not proven troublesome. Shielding is essential in the interest of preventing unwanted oscillations and to keep the large circulating currents from getting into the coder. There is quite a bit of heat generated but it has caused no serious problem.

The common-base first stage is used only to provide a good impedance to the clock source. The push-pull common-emitter second stage provides power gain to drive the final amplifier. The push-pull output stage operates in the common-base mode. This may not be immediately apparent because the collectors are grounded. It was necessary to solidly ground the cans (and thereby the collectors) for heat transfer. Using isolation windings on the interstage transformer permits the base and emitter to swing. The stage is operated class B for convenience and efficiency.

4.1.6 *Mechanical Layout*

As is the case in all high-speed circuits, the mechanical layout of the sample and hold circuit is an important part of the total electrical de-

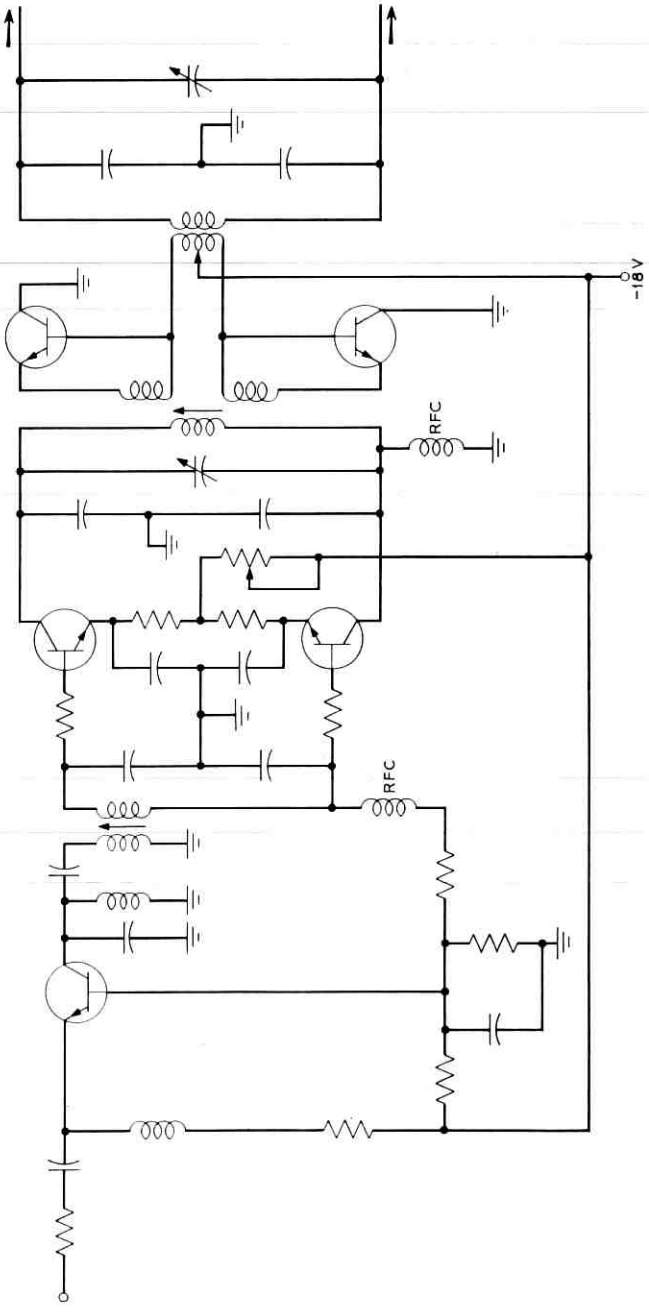


Fig. 18 — Sample and hold sine wave power amplifier.

sign. The photograph (Fig. 19) shows one view of this circuit. Complete symmetry of the bridge is important. The preamplifier and postamplifier are mounted close to the bridge as are the final pulse forming parts of the driver. The entire circuit is enclosed in a solid shield.

4.1.7 *Performance of Sample and Hold*

When tested separately with the resampler only, the sample and hold circuit gives a noise performance at least 6 dB better than an ideal nine-digit system. It does not appear to limit system performance at present. It is somewhat difficult to adjust for good performance but once adjusted has performed well over long periods of time. Simpler driving circuits are under investigation.

4.2 *Parallel to Serial Converter*

The digit output circuits of both the tube and the solid-state coder present the PCM code words as plus or minus voltages on nine separate leads. All outputs are in phase and occur periodically at the sampling rate. The parallel to serial converter combines these signals into a serial pulse train. This process is carried out by means of a tapped delay line.

Fig. 20 illustrates this operation for the case of the 12-Mc/s converter. Each digital output is sampled by a 4-ns clock pulse and gated onto the respective tap of the delay line. To avoid reflections, the line is terminated at both ends in its characteristic impedance. The total end-to-

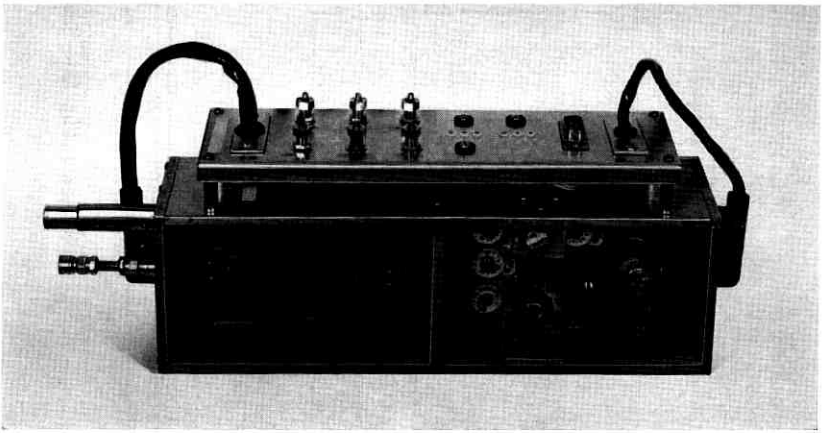


Fig. 19 — Sample and hold.

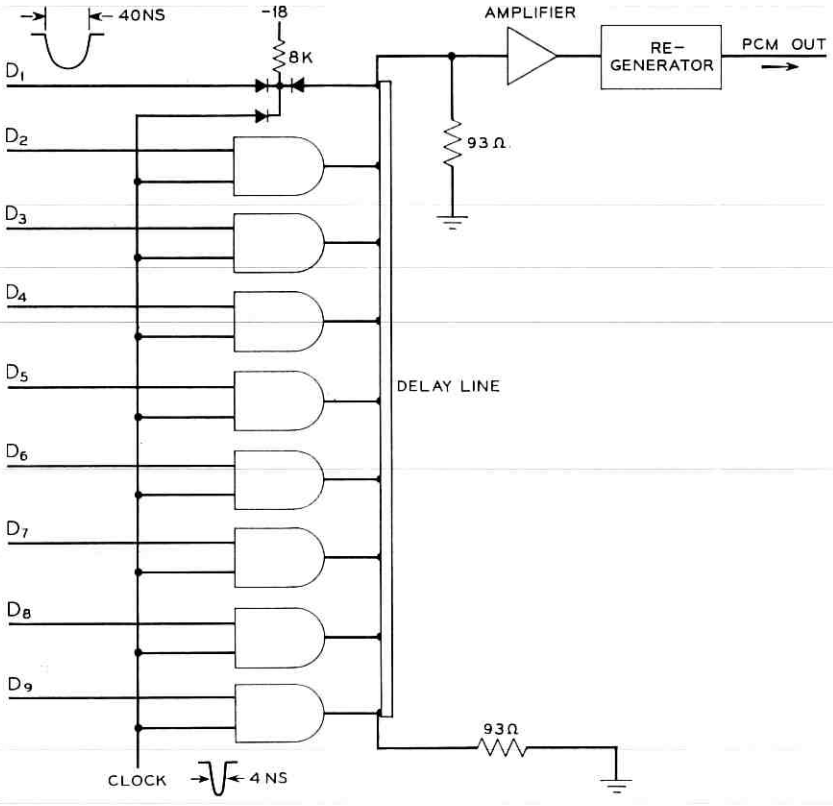


Fig. 20 — Parallel to serial converter.

end delay of the line is 72 ns, divided into eight sections of equal length. At the upper termination of the line the signal appears in serial form; digit one first, digit nine last. A linear dc amplifier of 200-Mc/s bandwidth raises the output signal to a level of 2 V peak-to-peak.

A partially shielded microstrip delay line of somewhat unconventional construction was used. Fig. 21 is a photograph of this line. A dielectric material is sandwiched between a ground plane and a conductor plane. However, the conductor plane contains an additional ground area that is interleaved with the conductors. This construction reduces coupling between adjacent sections of the conductor, permits a closer spacing between conductors, and hence larger amounts of delay are realizable in a given area. The line is photoetched on a 1/16-in. thick copper clad Tellite board, whose dimensions are 12 in. by 16.5 in.

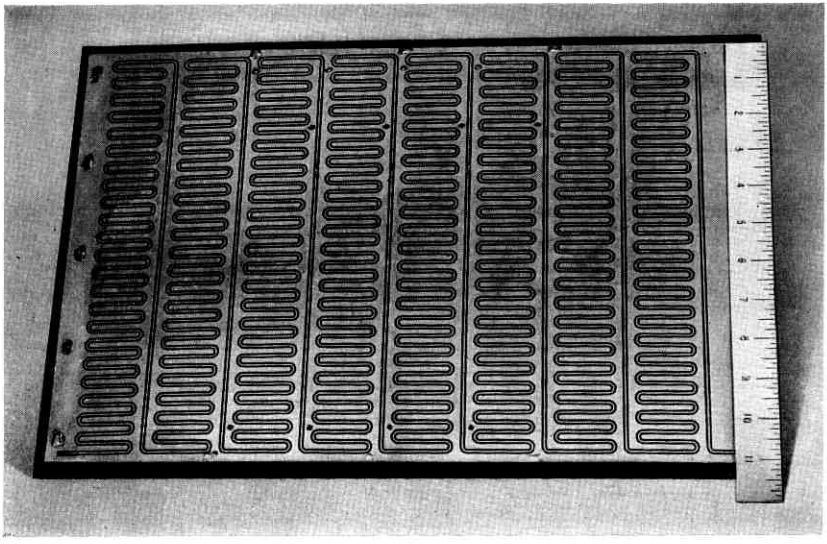


Fig. 21 — Partially shielded microstrip delay line.

The characteristic impedance of the line is 93 Ω. The response to a 4-ns pulse is excellent. The total end-to-end attenuation of 2 dB is somewhat higher than the attenuation of a coaxial cable of the same electrical length. However, this line has a definite advantage over a coaxial line because the conductor is exposed and, therefore, the taps can be connected without any difficulties and without the introduction of appreciable discontinuities in the line.

A source of reflections on the line is the load that the gates present to their respective tapping points. When a gate is conducting, the load is the 8-kΩ gate resistor; when a gate is not conducting the load is the capacitance of the output diode, which is less than 0.5 pF. The two load conditions represent impedances that are high compared to the 93-Ω line impedance. The total reflections on the line are 20 dB below the signal.

In the 6-Mc/s converter, which requires a delay of 144 ns, a coaxial cable line was used. A microstrip delay line with that much delay would have unwieldy dimensions and excessive end-to-end transmission loss.

4.3 *Output Regenerator*

The function of the regenerator is twofold. It makes the final binary decision on the output digits, i.e., it resolves the remaining partial, or

undecided digits that may still be present at this point. Second, it retimes and reshapes the serial output pulses. It must be a relatively fast circuit, since it executes these functions on 4.5-ns pulses. As will be discussed in a later section, the threshold uncertainty region must be extremely small. Achievement of such speed and accuracy dictates the use of 2.5-Gc/s transistors in conjunction with tunnel diodes.

Fig. 22 is a partial circuit schematic which shows the threshold and retiming circuit. Fig. 23 illustrates the composite current-voltage characteristic of the tunnel diode $D1$ and transistor Q_1. It includes the actual operating points of the circuit for various input conditions. These points are readily derived from the circuit parameters.

Transistor Q_1 can be turned on only when both the signal and the clock are negative to their respective thresholds (operating point A). Even if the signal subsequently rises positive with respect to its above threshold (as may be the case for a partial undecided digit pulse), the transistor remains in the "on" condition (point B) until the clock signal crosses its threshold. Q_1 then turns off (point C). Under the remaining two input conditions (D and E), Q_1 never turns on. The timing of the regenerator is shown in Fig. 24.

This circuit ensures that the output pulsewidth is determined exclusively by the clock, regardless of the width of the input signal. The use of the tunnel diode as a threshold element makes this circuit extremely fast and results in a very narrow threshold region.

The threshold performance of this circuit cannot be measured directly.

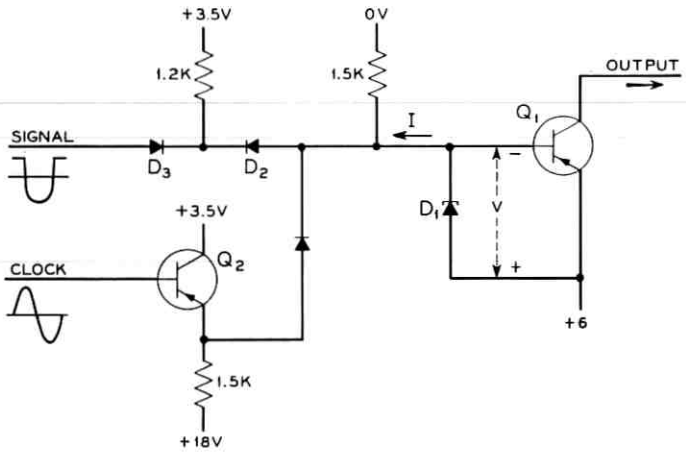


Fig. 22 — Regenerator threshold and retiming circuit.

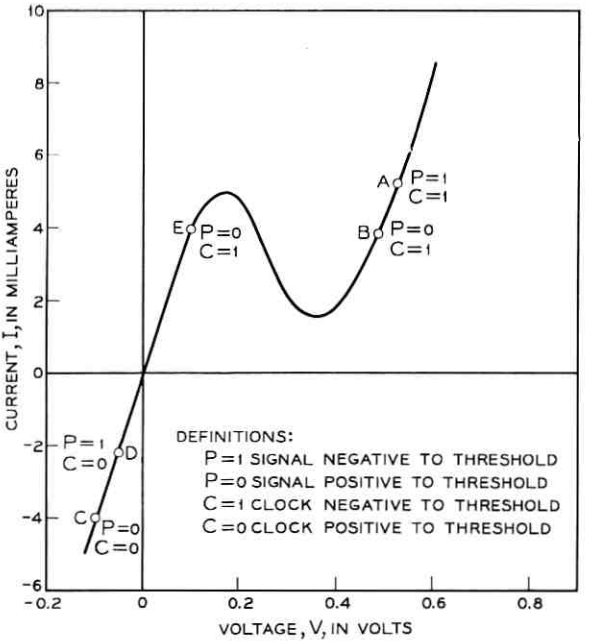


Fig. 23 — V-I characteristics of threshold circuit shown in Fig. 22. Operating points for various input conditions are shown.

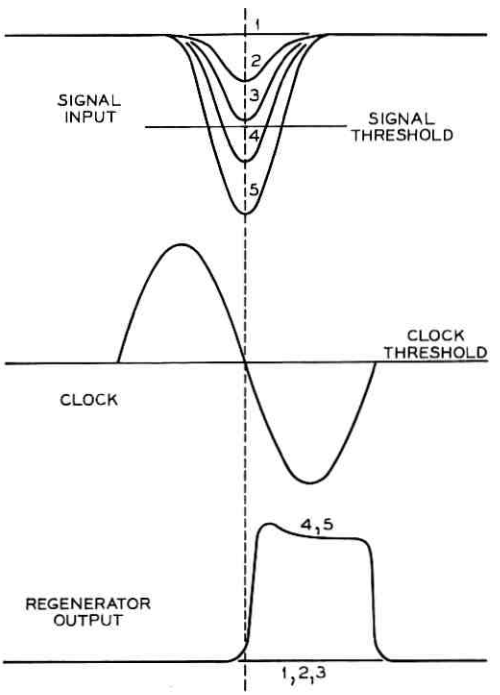


Fig. 24 — Timing of regenerator. The input signal wave shape illustrates the effect of partial pulses.

All attempts to do this have lead to the erroneous conclusion that this regenerator is ideal; which it is not. However, the error performance of the entire system will give a clue to the actual performance. This will be discussed further in Section 5.6.3.

4.4 *Clock Circuits*

The video coder as well as the mastergroup coders have individual sine-wave clocks. The clock signal is generated by means of a crystal-controlled Colpitts oscillator. The desired stability is achieved by mounting the crystal in a temperature-controlled oven.

The sine-wave clock is distributed, via 93-Ω coaxial cable, to the various points in the system from conventional emitter-follower distribution amplifiers. Proper timing is obtained by suitable choice of line-length.

A times-nine multiplier circuit (Fig. 1) generates the clock signal used for timing the output regenerator. The video coder has an additional circuit that divides the sampling clock by a factor of eighteen for purposes of forced-bit framing (Section 5.5.1).

4.5 *Bandlimiting and Reconstruction Filters*

The mastergroup signal is inherently sharply bandlimited and therefore the filter design is straightforward and presents no problems. On the other hand, the television signal falls off rather slowly with frequency and the filter design for this application is more difficult. It is controlled by the transmission objectives for television. Both the bandlimiting filter and the reconstruction filter for television were designed by balancing foldover distortion against transient behavior and making minor refinements to satisfy the over-all transmission objectives. The bandlimiting filter is implemented by a seventh-order Butterworth filter with delay equalization, and the reconstruction filter is a fifth-order inverse Tchebycheff filter also with delay equalization.

V. DECODING TERMINAL CIRCUITS

5.1 *General*

The decoder executes the digital to analog conversion process at the receiving terminal. Fig. 25 shows a block diagram of the receiving video terminal. The Gray-code serial PCM pulse and its timing signal are received from the demultiplex terminal. By means of a countdown circuit the frequency of the received high-speed (approximately 110

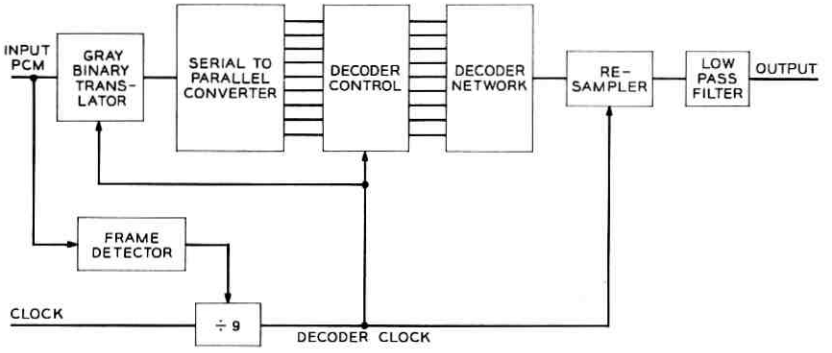


Fig. 25 — Decoding terminal.

Mc/s) clock signal is divided by a factor of nine. This process yields the 12-Mc/s clock signal which is used to time the various decoding circuits. The terminal framing detector controls the countdown circuit and thereby ensures that the clock signal has the correct phase relationship with respect to the information pulse train.

The information digits, after being processed by the Gray to binary translator and the serial to parallel converter, are read into the decoder control circuitry.

The actual digital-to-analog conversion process takes place in the decoding network. The output of this network is then fed to the resampler and the low-pass filter.

The mastergroup decoder is identical with the video decoder except all operations are executed at one-half the speed and the framing method is different.

5.2 *Decoder*

The decoder consists of: (1.) the decoder control circuits, (2.) the weighting network, and (3.) the resampler. The decoder control circuits are nine storage flip-flops into which the parallel binary PCM word is read. The weighting network (Fig. 26) is a resistive ladder network with eight sections, each of which has a transmission ratio of two. Under control of its corresponding storage flip-flop, each network node is supplied with a current of $+I$ or $-I$, depending on wether the stored digit is a ONE or a ZERO. The voltage generated at the output of the network, as a result of the individual node currents, has a magnitude which corresponds to the PCM word. The network output is then resampled.

Maintenance of constant loss ratio in the ladder network depends on keeping a constant resistance at each node independent of the value of the digit. The gate shown on Fig. 26 presents a resistance R when the flip-flop is in either the ONE or the ZERO state. Use of $+E$ or $-E$, rather than E and ground, gives network outputs with equal plus and minus swings. In the design of the weighting network the shunting effect of resistance R must be considered.

E. F. Kovanic[12] has discussed the design and operation of the video decoder in detail. The reader is referred to his paper for further information on this subject. Only minor modifications have been made in the decoder control circuitry; but, the resampler was redesigned entirely and is discussed in Section 5.3.

The mastergroup decoder utilizes an integrated tantalum-nitride thin-film network. The design requirements and the fabrication of this network are described elsewhere.[13] The diodes that switch the network currents are encapsulated in groups of four in a single package. These are appliqued to the thin-film network. A photograph of the mastergroup decoder, including the network, is shown in Fig. 38.

The integrated thin-film approach to the realization of the network has several advantages over the discrete resistor approach. The geometry of the integrated network permits a reduction in parasitic elements, especially inductance; thereby reducing ringing in the analog output. The entire package is considerably smaller in size and use of the common substrate yields better temperature tracking of the resistors.

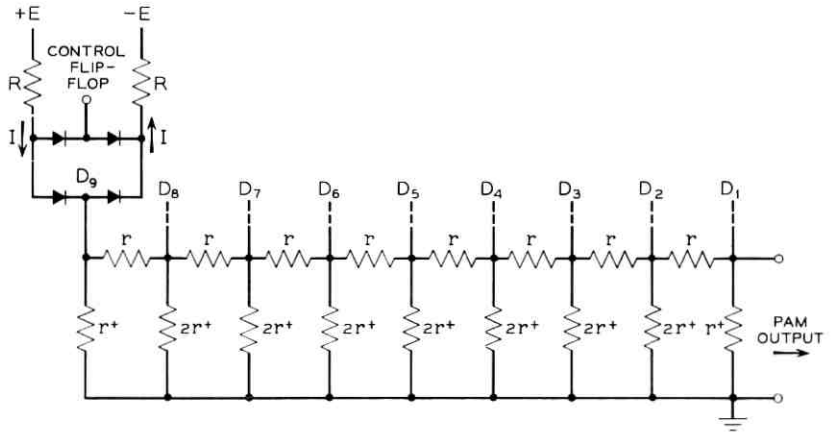


Fig. 26 — Decoder weighting network.

5.3 *Resampler*

5.3.1 *Function*

The resampler substantially eliminates all switching transients that are present in the decoder PAM output signal. These transients occur in the form of spikes when the PAM signal changes from one quantizing level to another. In a practical decoder, these transition spikes are unavoidable because the nine network currents, that are under the control of the digits, cannot be switched simultaneously in an infinitessimal time. These spikes are not linearly correlated with the signal. Their energy is dependent on which and how many digits are being switched. For example, a transition from code 011111111 to 100000000, which represents a change of only one quantizing level, at the center of the coding range in the binary code causes the largest spike because all digit currents are switched. If these spikes were not removed, the signal impairment would be substantial.

5.3.2 *Requirements*

The resampler is basically a transmission gate which conducts only during that portion of the PAM sample that is free of transients. For video decoding the gate periodically conducts for a 28-ns interval.

To avoid further degradation in over-all system performance as a result of the resampling process, the total signal distortion due to the resampler is held to a value which is better than 10 dB below quantizing noise. Transmission impairments in the resampler are caused by: (1.) distortion in the input and output amplifiers, (2.) nonlinearities in the gate, and (3.) signal compression which is a result of finite switching time. This latter condition is quite important because the output signal, after it has passed through the low-pass filter, is proportional to the area under the resampled PAM pulses. For infinitely fast rise and fall times this area is linearly proportional to the pulse amplitude; but for finite gate switching times this is not the case. For a PAM pulse width of 28 ns, gate switching time must be less than 4 ns.

5.3.3 *Circuit Description*

A schematic of the resampler is shown in Fig. 27. It consists of an input amplifier, a balanced diode gate, an output amplifier, and a gate driver. The gate is a balanced diode bridge, similar to the sample and hold gate, which works into the summing node of the output amplifier. Since this amplifier is a shunt-feedback amplifier with more than 40 dB of

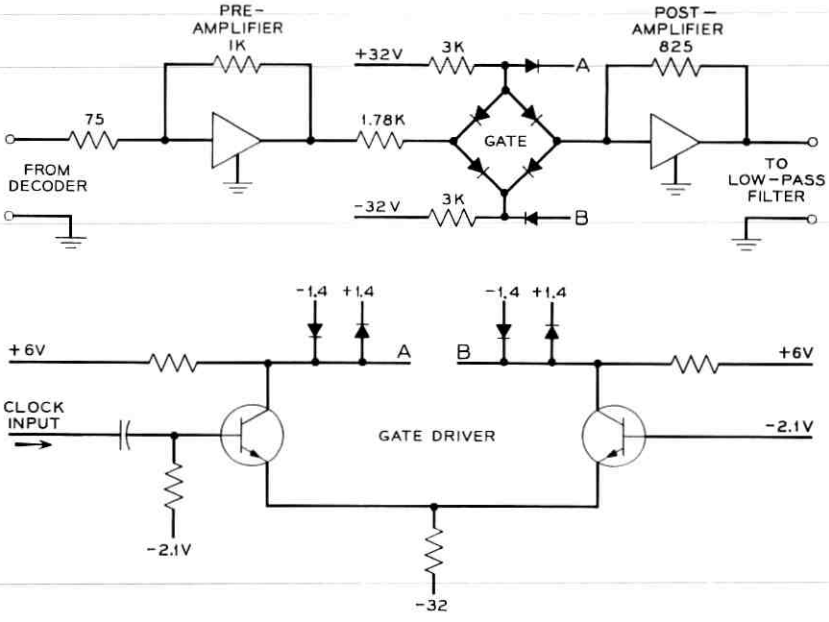


Fig. 27 — Resampler circuit.

loop gain, the impedance looking into the summing node is very low. On the other side, the gate is driven from the input amplifier through a 1.78-kΩ resistor. This impedance is high relative to the summing node impedance that the gate sees at its output. As a result, the gate is switching signal current into a virtual short circuit, rather than switching a signal voltage into a finite impedance. This has several advantages.

This mode of operation makes the gate relatively insensitive to the effects of an error voltage which may exist between the input and output of the gate as a result of mismatched diodes. Because of the relatively high impedance in the gate transmission path, the error voltage has negligible effect on the gated signal current. This considerably reduces the magnitude of a pedestal which an error voltage would introduce into the gated signal if the gate were to operate in the voltage mode. By the same argument, the impedance of the bridge, which is nonlinear because of the diodes, introduces negligible degradation to the transmission performance.

The current mode of operation has further advantages in regard to the switching performance. The gate is switched in a similar fashion to the sample and hold gate. However, since the resampler, in contrast

to the sample and hold gate, works into a virtual short circuit, the voltage excursions at the input and output nodes of the bridge are extremely small and centered around the quiescent dc voltage which is adjusted to a value of zero volts and has less than 10-mV drift. Consequently, if the gate is driven by control pulses that are reasonably symmetrical in shape with respect to each other, all diodes switch very nearly simultaneously (Fig. 28). The rise time of the control waveforms may be relaxed provided symmetry and summing-node voltage drift is under control.

The circuit that generates the control pulses is a current routing circuit (Fig. 27). This circuit generates pulses which have opposite phase and excellent symmetry. The pulses are clamped to ± 1.7 V and have a rise time of 10 ns. Ideal current-mode operation of the gate is achieved if (1.) the dc voltage at the postamplifier summing node is zero volts, (2.) the two gate control pulses cross zero at the same instant, and (3.) the postamplifier summing node is a short circuit so that the voltage excursions of the gated signal are zero. In the system, deviations from these ideal conditions were held to within ± 100 mV. In this case, the theoretical switching time of the gate is 0.6 ns. Because of the finite switching times of the diodes, the actual gate switches somewhat more slowly. The gate diodes are gallium arsenide switching diodes with less than 1-pF junction capacitance and negligible storage time.

The noise performance of the resampler was measured with a noise-loading test set. The noise level is more than 10 dB below the quantizing noise of a theoretical nine-digit coder.

5.4 *Serial to Parallel Converter*

This circuit is similar to the parallel to serial converter. Identical delay lines are used. The serial pulse train is fed into one end of the line and the parallel outputs are taken from the tapping points by means of single-stage emitter followers.

Strictly speaking, the outputs do not contain the digits in parallel

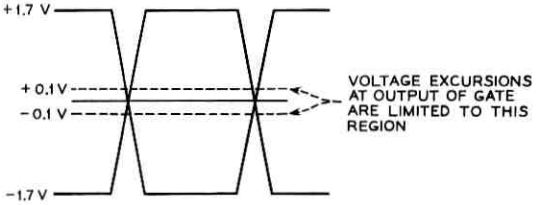


Fig. 28 — Resampler gate driver wave shapes.

form, but instead, each output lead contains the serial pulse train delayed one time slot from the preceeding output lead. At the decoder control circuit all nine outputs are sampled simultaneously by a single clock pulse. In this way the proper digits are gated to their respective decoder circuits.

5.5 *Framing Detector*

5.5.1 *Television Framing*

The television codec is framed by removing the ninth digit of every ninth word at the coder and substituting ONES and ZEROS alternately. This introduces a small degradation in over-all system noise performance.[1] At the decoder the framing detector identifies the framing pulses and ensures that the decoder timing clock has the proper phase relationship with respect to the received digits. The framing detector is patterned after the framing circuit used in the T1 carrier system.[14]

A simplified functional block diagram is shown in Fig. 29. The approximately 110-Mc/s clock received from the demultiplex terminal is divided down by a factor of 81 to produce a clock at the frame rate. The frame clock gates a pulse out of the high-speed line. This pulse is compared with a reference flip-flop which generates the framing pattern. If the system is out of frame, violations occur and are sent to the store. After an average of about three violations have occurred, the threshold

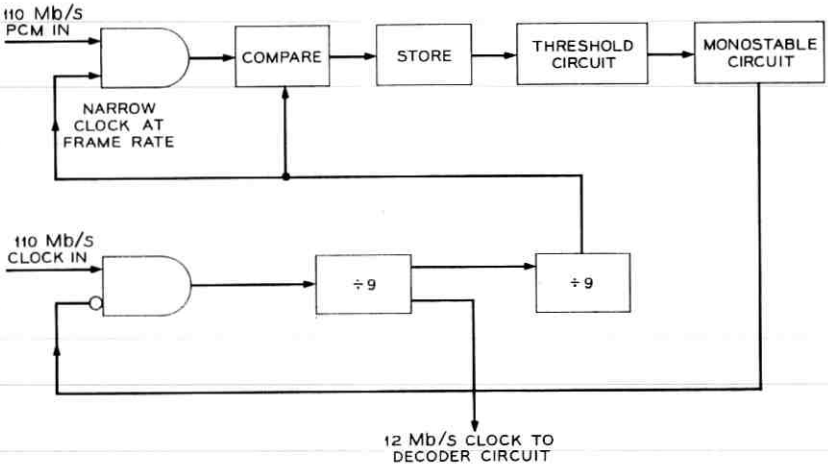


Fig. 29 — Video codec frame detector.

circuit recognizes an out of frame condition and sends an error pulse which in turn inhibits the high-speed clock and shifts the counter by one time slot. This process is repeated until the system is in frame.

The worst case exists when the decoder timing has slipped by only one time slot. The decoder framing detector has to advance the decoder timing a slot at a time through 80 time slots until it is in frame again. It checks for a framing digit every $9 \times 81 = 729$ ns. If it is assumed that any information digit has a 50 per cent probability of being in the same state as the framing digit, it takes on the average about 1.5 μs to advance one time slot. Since 80 time slots have to be traversed before the decoder is in frame again, the average reframe time is $1.5 \times 80 = 120$ μs. Fig. 30 shows experimental data on the distribution of reframe time in the worst case.

5.5.2 *Mastergroup Framing*[15]

The mastergroup codec is framed by observing the statistics of the digits in the received Gray-code pulse train. If the coder is loaded with

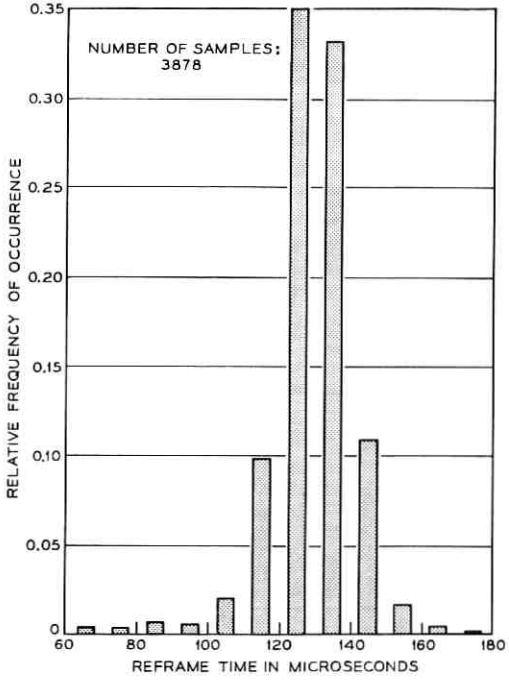


Fig. 30 — Distribution of video codec reframe time in worst case.

a mastergroup signal, which has Gaussian amplitude distribution and an rms value of one-eighth the peak-to-peak coding range, digit two has a 95 per cent probability of being a ONE while the probability of a ONE in any other digit position is nearly 50 per cent (Fig. 31).

The block diagram of the framing detector is shown in Fig. 32. The approximately 55-Mc/s clock is divided by a factor of nine. This process yields the basic decoder clock at the sampling rate. This clock periodically extracts a digit from the signal pulse train. If the extracted digit is a ZERO, the clock pulse also charges a leaky integrator. If the extracted digits have the value ZERO frequently enough, the integrator reaches the threshold value and sends out a shifting command to the countdown circuit. When the receiver is finally in frame, the integrator will remain below the threshold because of the infrequent occurrence of zeros in digit two.

Measured data of the reframing time statistic in the worst case (when timing slips only one digit) is given in Fig. 33.

5.6 *Gray to Binary Translator*[16]

5.6.1 *General*

The coders generate the Gray code because, as previously mentioned, the probability of gross error in coding is thereby made negligibly small. Since the binary code is more simply decoded, the code must be trans-

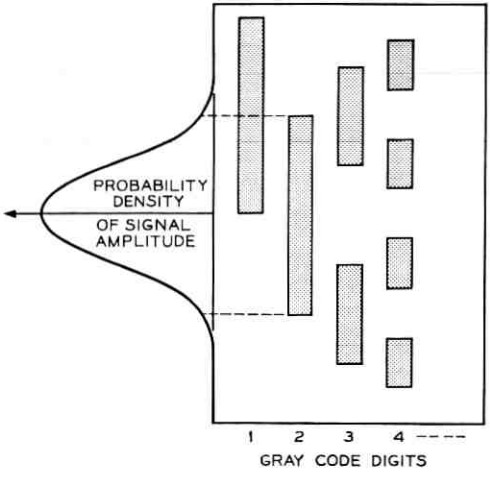


FIG. 31 — Gray-code digit assignments.

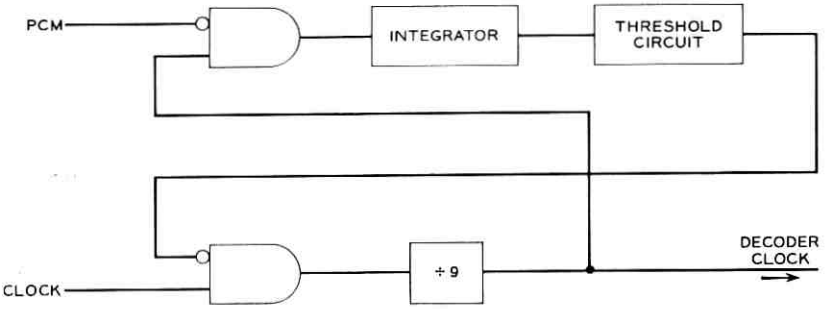


Fig. 32 — Mastergroup codec framing detector.

lated. As will be discussed in Section 5.6.3, the translator produces gross errors when partial pulses are presented at the input. In order to take advantage of the circuits in the multiplex and demultiplex terminal and the repeatered line to further resolve undecided pulses, the translation is performed at the receiving end. The amount of circuitry required is considerably less when the translation is performed on the serial pulse

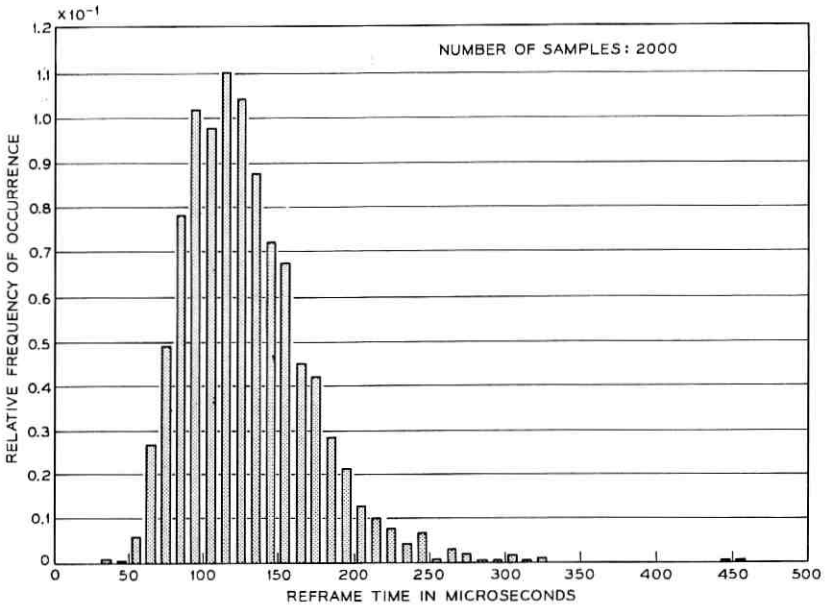


Fig. 33 — Distribution of mastergroup codec reframe time in worst case.

stream rather than in parallel, although the speed requirements are quite stringent.

5.6.2 *Circuit Realization*

Let b_k and g_k represent the kth digits in the binary and the Gray-pulse streams respectively. Let n be the number of digits in a serial word. The translation logic is

$$b_1 = g_1$$
$$b_k = b_{k-1} \cdot \overline{g_k} + \overline{b_{k-1}} \cdot g_k$$

for

$$2 \leqq k \leqq n.$$

The second equation states that binary digits other than the first repeat the preceding binary digit if the Gray-code digit is ZERO and are inverse to the preceding binary digit if the Gray-code digit is ONE. This function is represented by the output of a flip-flop, or binary counter, which can be forced to agree with the first digit and is then free to be triggered by subsequent ONES in the Gray-code word.

The desired logic function is realized as shown in the block diagram of Fig. 34. The actual circuit schematic is shown in Fig. 35. The circuit operates with a speed of 110 Mc/s. It employs germanium mesa transistors which have a gain-bandwidth product of 2.5 Gc/s. The duration and position of the clock pulses is critical and must be controlled to a small fraction of a time slot.

5.6.3 *Effect of Partial Pulse on the Translation Process*

A cause of translation errors is the presence of an unresolved partial pulse at the input of the translator. A partial pulse has deteriorated wave shape and represents neither a binary ONE nor a ZERO. The magnitude of the error depends on which digit is unresolved. The largest error occurs in the case of a partial first digit and the error magnitude decreases for subsequent unresolved digits. To illustrate the mechanism by which these translation errors are generated, the case of a partial first digit is considered.

Suppose the sample at the input to the coder has an amplitude level which is precisely midway in the coding range. The coder is expected to process this sample into the code word 010000000 (quantizing level 255) or 110000000 (quantizing level 256). However, because the coder as well as subsequent threshold circuits are not ideal, the state of the

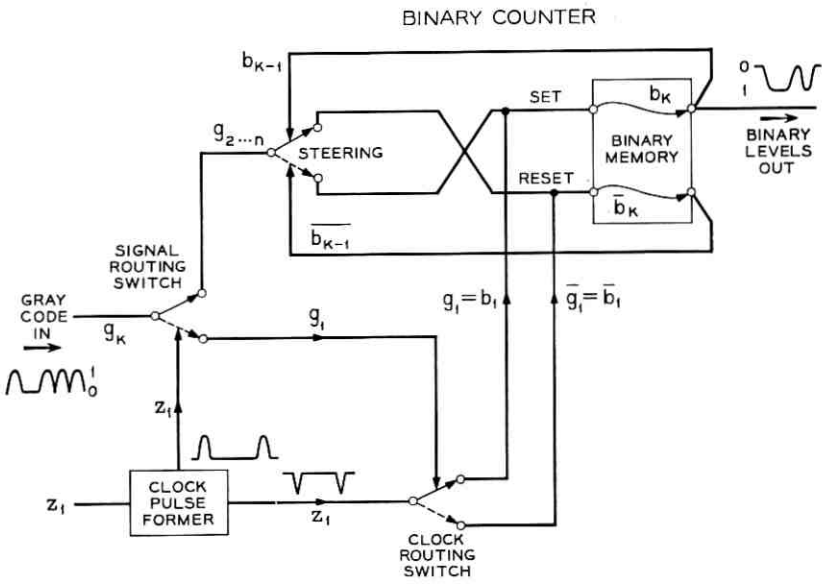


Fig. 34 — Serial Gray to binary translation.

first digit may not be completely resolved and a code which may be called $\frac{1}{2}10000000$ and has a partial pulse in the first digit position appears at the translator input.

The first digit controls the clock routing switch which directs the clock pulse Z_1 to the set or reset inputs of the binary memory (Fig. 34). Since the first digit has partial energy, this operation may not be executed properly. The binary cell may momentarily change state, but, because of insufficient trigger energy, will return to its previous condition. Therefore, the partial pulse has propagated to the output of the translator. The output of the translator is simultaneously directed to two different destinations: (1.) the decoder control circuit, and (2.) the steering switch. An undecided translator output may result in one of the following situations:

(1.) Both the decoder and the steering switch interpret the translator output as a ONE (or a ZERO). In this case no error is made and the word is correctly decoded as quantizing level 256 (or 255).

(2.) The decoder interprets the output as a ONE while the steering switch interprets a ZERO (or vice versa). In this case the resulting binary code is effectively 111111111, which corresponds to quantizing level 512 (or the binary code is 000000000, which corresponds to quantizing level ZERO).

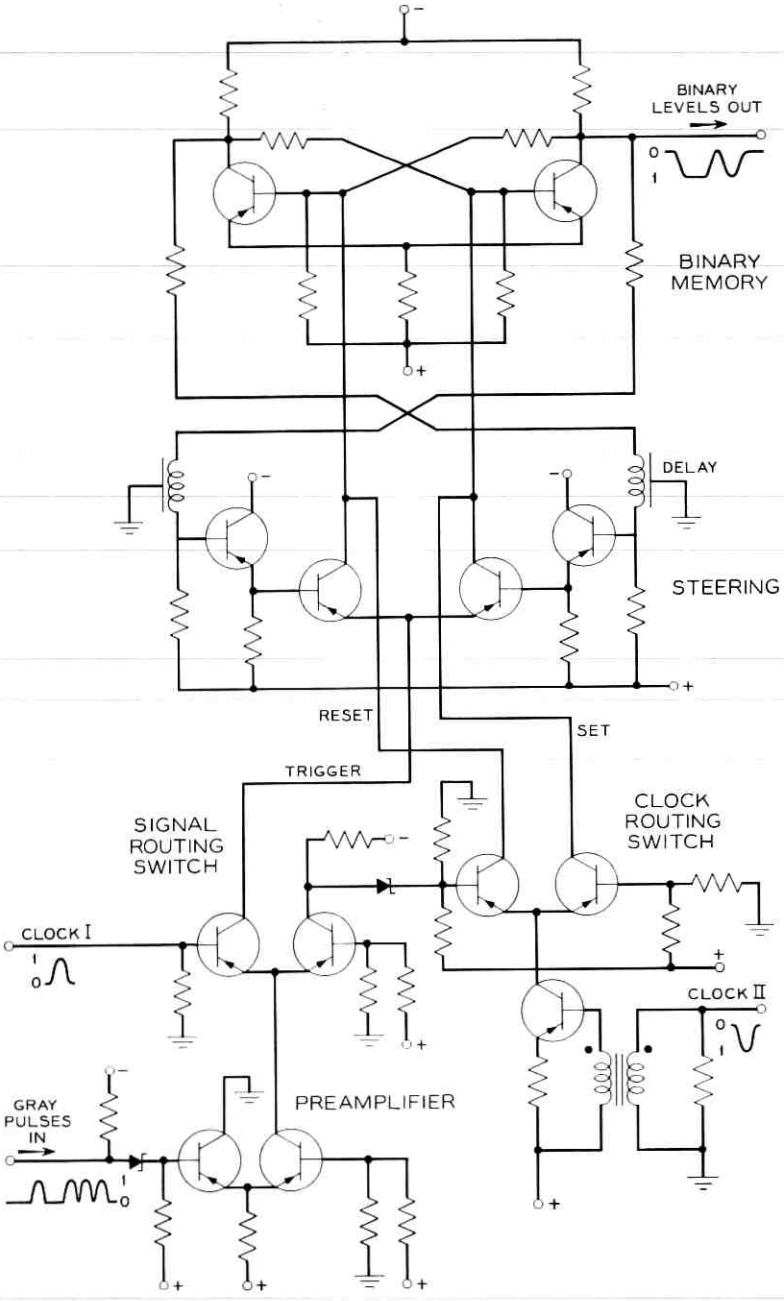


Fig. 35 — Gray to binary translator.

Because of the inconsistency in the decisions reached by the two circuits, an error with a magnitude of one-half the coding range is made. A similar error mechanism exists for partial pulses in the second-digit time slot. These errors have a magnitude of one-quarter the amplitude of the coding range.

These errors occur relatively infrequently and therefore do not affect the noise appreciably. However, when they occur, they are, because of their large amplitude, clearly visible in the form of occasional white or black dots in a decoded television display.

In order to see how sharp the effective threshold of the entire system must be for a tolerable error rate, the width of the threshold uncertainty region over which first digit errors occur is computed. For a standard television signal, the video signal excites about 366 codes; the remaining 146 codes are coding the sync pulses. Of the 720-million samples coded every minute only 600 million contain video information; the remainder are used for the sync pulses. If it is assumed that the video information has uniform amplitude distribution, the width of the uncertainty region over which first-digit errors occur must be held to within six parts in 10^7 for an error rate of one error per minute.

This level of performance has been achieved in the experimental system.

VI. EQUIPMENT DESIGN

The codec equipment is mounted on four standard 19-in. relay racks. The major equipment features are depicted in the photographs (Figs. 36 and 37). Except for the coding tube and the video decoder, the circuits are mounted on small plug-in cards and in 16.5-in. by 12-in. sliding drawers. The latter appears to be the most promising approach for a design for manufacture. Fig. 38 is a photograph of the drawer that houses the decoder. The basic design encompasses a large printed circuit motherboard onto which small circuit packages and other components are mounted. By this approach the system can be divided into major functional blocks which have relatively few interconnections. Within each major block the more numerous interconnections between the subsystems are done with little difficulty on the motherboard. The high-frequency signals are brought in from the outside via RG 180/U coaxial cable and are transmitted to their destination by means of microstrip lines on the motherboard. The extensive use of microstrip lines for interconnection of subsystems reduces the amount of bulky coaxial cables and associated connectors.

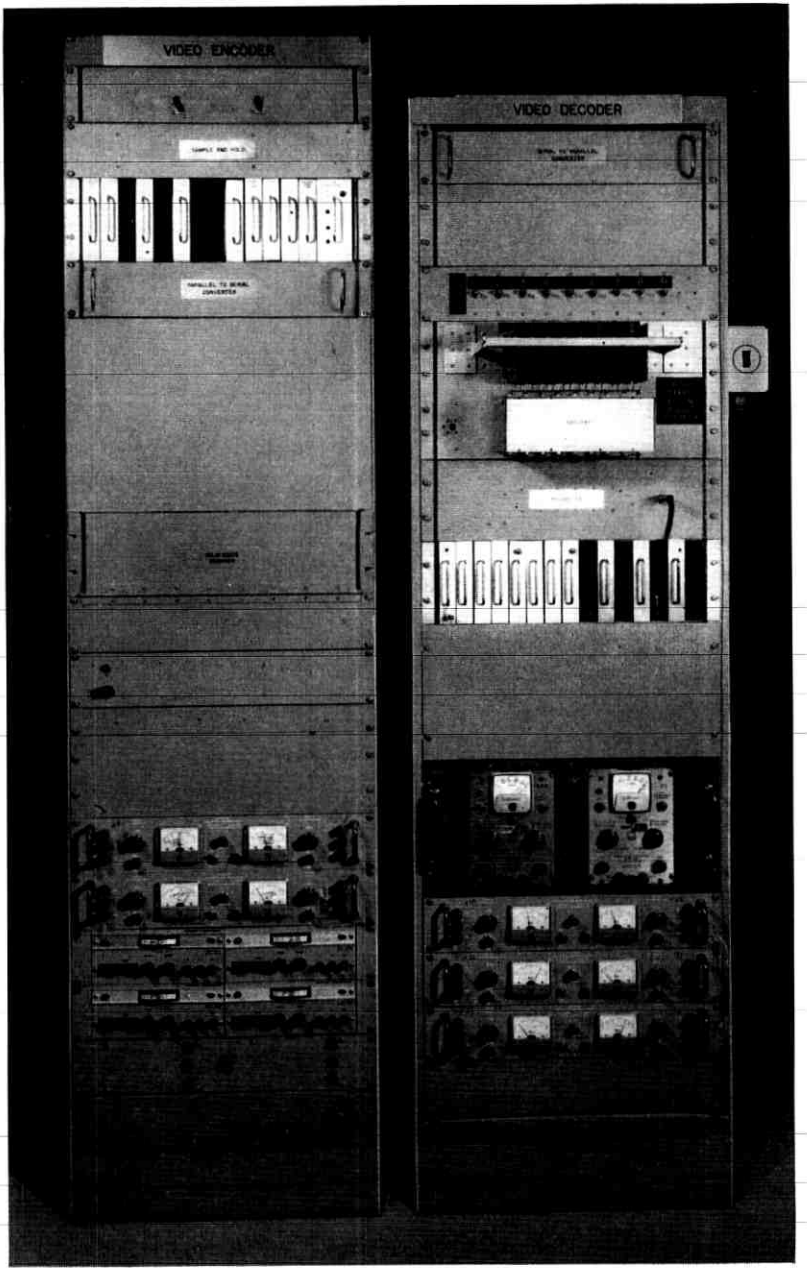


Fig. 36 — Video codec.

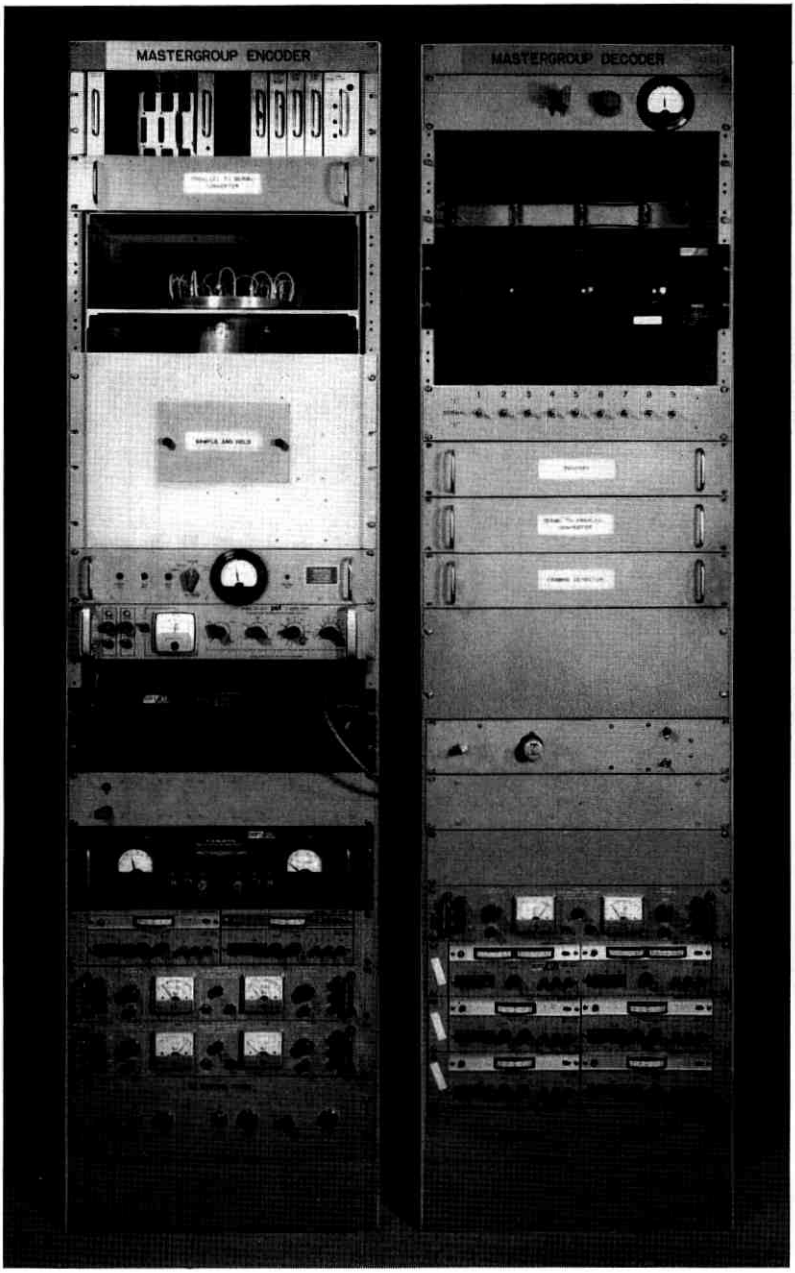


Fig. 37 — Mastergroup codec.

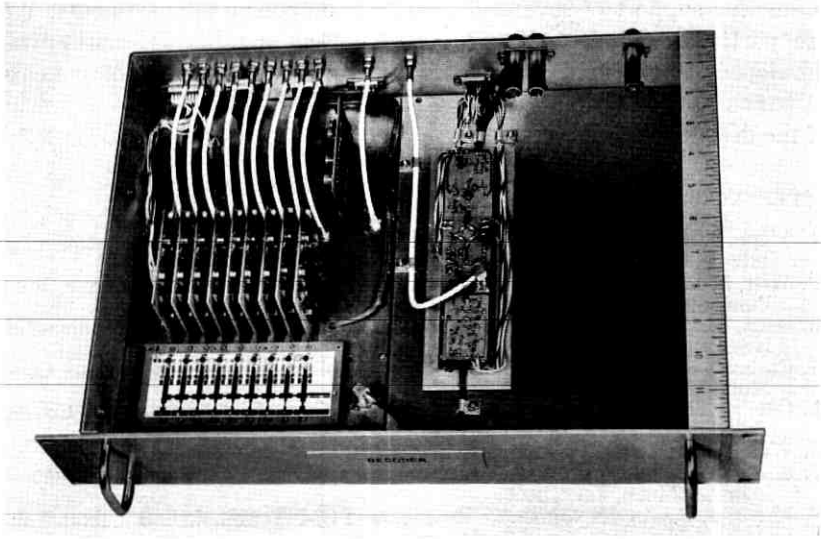


Fig. 38 — Decoder.

VII. CONCLUSION

Codecs for the PCM coding and decoding of a 600-voice channel, frequency division multiplexed, mastergroup and of a standard color or black and white television signal have been built and operated successfully. The practicability of commercial designs of such codecs and their associated circuitry has been demonstrated.

For mastergroup coding (6-Mc/s sampling rate), the theoretical quantizing-noise level under optimum load conditions is 23 dBrnc0. The measured noise of the over-all systems was 25 dBrnc0 with the tube coder and 28 dBrnc0 with the solid-state coder.

For television coding (12-Mc/s sampling rate) the theoretical peak-to-peak signal to rms noise ratio is 64 dB (including 1-dB framing impairment). The measured peak-to-peak signal to rms noise is 62 dB for the tube coder and 57 dB for the solid-state coder. Improvement in the performance of the solid-state coder should be possible with additional work.

VIII. ACKNOWLEDGMENTS

This work represents the efforts of several people and was carried out in the PCM Terminal Department under the direction of J. S. Mayo. The success of this project would not have been possible without

the active support of several device and components departments. We are particularly indebted to M. H. Crowell and his associates who developed the nine-digit coding tube and to the Thin Film Components Department under the direction of D. A. McLean for the development of the decoder thin-film network.

REFERENCES

1. Mayo, J. S., Experimental 224 Mb/s PCM Terminals, B.S.T.J., This Issue, pp. 1813–1842.
2. Witt, F. J., An Experimental 224 Mb/s PCM Multiplexer-Demultiplexer Using Pulse Stuffing Synchronization, B.S.T.J., This Issue, pp. 1843–1886.
3. Sears, R. W., Electron Beam Deflection Tube for Pulse Code Modulation, B.S.T.J., *27*, Jan., 1948, pp. 44–57.
4. Carbrey, R. L., Video Transmission over Telephone-Cable Pairs by Pulse Code Modulation, Proc. IRE, *48*, Sept., 1960.
5. Cooper, H. G., Crowell, M. H., and Maggs, C., A High-Speed PCM Coding Tube, Bell Laboratories Record, *48*, Sept., 1964, pp. 267–272.
6. Waldhauer, F. D., U.S. Patent 3-187-325, 1965.
7. Smith, B. D., An Unusual Analog-Digital Conversion Method, IRE Trans. Instrum. Meas., June, 1956.
8. Mayo, J. S., An Experimental Broadband PCM Terminal, Bell Laboratories Record, May, 1964, pp. 152–157.
9. Bender, W. G., An Experiment in PCM Transmission of Multiplexed Channels, Bell Laboratories Record, July/August, 1964, pp. 240–246.
10. Waldhauer, F. D., Latest Approach to Integrated Amplifier Design, Electronics, May, 1963.
11. Gray, J. R., and Kitsopoulos, S. C., A Precision Sample-and-Hold Circuit with Subnanosecond Switching, IEEE Trans. Circuit Theor., *CT11*, Sept., 1964, pp. 389–396.
12. Kovanic, E. F., A High Accuracy Nine-Bit Digital-to-Analog Converter Operating at 12 mc, IEEE Trans. Commun. Electron., March, 1964, pp. 185–191.
13. Jackson, W. H., and Moore, R. T., A High Accuracy Thin Film PCM Decoder Network Operating at 12 mc, IEEE Trans., *PMP-1*, June, 1965, p. 45.
14. Davis, C. G., An Experimental Pulse Code Modulation System for Short-Haul Trunks, B.S.T.J., *41*, Jan., 1962, pp. 1–24.
15. Gray, J. R., and Pan, J. W., Using Digit Statistics to Word Frame PCM Signals, B.S.T.J., *43*, Nov., 1964, pp. 2985–3008.
16. Koehler, D., A 110 Megabit Gray Code to Binary Code Serial Translator, 1965 Int. Solid-State Circuits Conf. Digest Techn. Papers.

Some Inequalities in the Theory of Telephone Traffic

By V. E. BENEŠ

(Manuscript received May 26, 1965)

The dynamical theory of telephone traffic in connecting networks, initiated by A. K. Erlang, has long lacked satisfactory ways of making approximations and deriving inequalities. These would reduce the fantastic computational burden implicit in the "statistical equilibrium" equations while still controlling accuracy. It is the aim of this paper to present a start in such a direction, in the form of inequalities (valid for wide classes of networks) for moments, probabilities, and ratios of expectations, among these last being the loss. The bounds in one series of these inequalities all depend on the known distribution of the number of calls in progress in a nonblocking network associated with the network under study. In a second series of cognate, simpler, but weaker inequalities, these bounds depend on Erlang's loss function or more generally on the terms of the Poisson distribution.

I. INTRODUCTION

Determining the grade of service of a telephone connecting network, as measured, for example, by the probability of blocking, continues to be a major outstanding problem of telephone traffic theory. Two principal methods are available for solving this problem. The first, simulation of a mathematical model of the operating system, has, with the advent of large high-speed computers, become very much less arduous than it used to be. The second, *calculation* of desired probabilities and expectations from [the] statistical equilibrium equations [of a mathematical model for network operation], is still hampered by the astronomical order of the equations, and in spite of its apparent promise, and its success with trunking problems early in the century, cannot be said to have reached fruition as far as connecting networks are concerned.

Indeed, it is taking so long for the strictly analytical approach to develop beyond its trunking and delay applications that in practical

engineering circles its value is in serious question. The problem is not so much the lack of a suitable basic theory, for the models provided by the "statistical equilibrium" approach have been available since the time of Erlang. The real problem is a lack of approximate methods for collecting and reducing the information available in these models in a manageable way to desired quantities *without losing track of the accuracy* of the approximations along the way. It is no trick to dream up approximate ways of calculating loss. But who can meet a challenge to show *theoretically* that his approximate method is not off by more than fifty per cent?

Some basic studies of the combinatorial and probabilistic features of connecting systems have been undertaken in previous work.[1] From these emerged a broad class of Markov stochastic processes suitable as mathematical descriptions of operating connecting networks. The statistical equilibrium equations for these models have been solved in principle with complete rigor, and the probability of blocking defined and calculated in principle. The results obtained were valid for arbitrary networks, and so were of necessity rather complex. Subsequent effort has been concentrated on reducing the rigorous results to practice by finding bounds and inequalities, and by making suitable approximations.

It is the aim of this paper to present, as the first step in such a program, a number of inequalities involving such quantities of interest to the traffic engineer as the probability of blocking, the mean and variance of the number of calls in progress, and the probability of more than k calls in progress. These inequalities have several noteworthy features:

(*i*) They are simple.

(*ii*) Most of them are consequences of one analytical "basic lemma".

(*iii*) The bounds they give are couched in terms of the distribution of the number of calls in progress in a corresponding nonblocking network of comparable size, or in terms of the Poisson or truncated Poisson (Erlang) distribution. (These distributions are familiar in traffic theory, but they have not been exploited systematically to give rigorous bounds for large classes of connecting networks.)

(*iv*) They afford ways of directly converting combinatorial information about network structure into probabilistic information about the chance of loss, the load carried, the attempt rate, etc.

In casting about for approximations and inequalities in a subject such as the present one, it is reasonable to collect first those that are valid for wide classes of connecting networks, and then those that depend on special combinatorial features of certain connecting networks. Only the first task has been attempted here; a start on the second appears in a later paper.[2]

II. THE THEORETICAL MODEL AND ITS PRINCIPAL PROPERTIES

Before summarizing results, we describe the assumptions on which they are based, the notations in which they are couched, and the salient properties of the theoretical model used to represent network operation. The results themselves will be given and derived only for the important case of "one-sided" networks,[3] for which all inlets are also outlets; it will be easy to see that analogous results (with similar proofs) are valid in the "two-sided", and other, cases.

With S the set of permitted (i.e., physically meaningful) states of the network ν (of T terminals) under study, we recall[1] that S is partially ordered by inclusion $\leqq$, where $x \leqq y$ means that state x can be obtained from state y by removing zero or more calls. If x is a state, the notation $| x |$ will denote the number of calls in progress in state x.

The Markov stochastic process x_t (taking values on S) studied in previous work[1,3] is used as a mathematical description of an operating connecting network subject to random traffic. This process is based on two simple probabilistic assumptions:

(*i*) Holding-times of calls are mutually independent variates, each with the negative exponential distribution of unit mean.

(*ii*) If u is an inlet idle in state x, and $v \neq u$ is any outlet, there is a (conditional) probability

$$\lambda h + o(h), \qquad \lambda > 0$$

that u attempt a call to v in $(t,t + h)$ if $x_t = x$, as $h \to 0$. All terminals have the same traffic characteristics.

The choice of unit mean for the holding-times merely means that the mean holding-time is being used as the unit of time, so that only the traffic parameter λ needs to be specified.

It is assumed that attempted calls to busy terminals are rejected, and have no effect on the state of the system; similarly, blocked attempts to call an idle terminal are refused, with no change of state. Successful attempts to place a call are completed instantly with some choice of route.

To describe how routes are assigned to calls, we introduce a *routing matrix* $R = (r_{xy})$, with the following properties: For each x, with A_x the set of states accessible from x by new calls, let Π_x be the partition of A_x induced by the equivalence relation of "having the same calls up", or satisfying the same "assignment" of inlets to outlets; then for each $Y \in \Pi_x$, r_{xy} for $y \in Y$ is a probability distribution over Y; in all other cases $r_{xy} = 0$.

The interpretation of the routing matrix R is this: Any $Y \in \Pi_x$ repre-

sents all the ways in which a particular call c not blocked in x (between an inlet idle in x and an outlet idle in x) could be completed when the network is in state x; for $y \in Y$, r_{xy} is the chance that if this call c is attempted, it will be routed through the network so as to take the system to state y. That is, we assume that if c is attempted in x, then a state y is drawn at random from Y with probability r_{xy}, independently each time c is attempted in x; the state y so chosen indicates the route c is assigned. The distribution of probability $\{r_{xy}, y \in Y\}$ thus indicates how the calling-rate λ due to the call c is to be spread over the possible ways of putting up the call c. It is apparent that

$$\sum_{y \in A_x} r_{xy} = \text{number of calls each of which can actually be put up in state } x = s(x), \text{ ("successes" in } x),$$

the second equality defining $s(\cdot)$ on S. This account of the method of routing completes the description of the traffic models to be studied.

The "statistical equilibrium" equations for the stationary probabilities $\{p_x, x \in S\}$ have the simple form

$$[\,|\,x\,|\, + \lambda s(x)]p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y r_{yx}, \qquad x \in S$$

where

A_x = set of states accessible from x by placing a new call,
B_x = set of states accessible from x by a hangup.

The probability of blocking, or call-congestion, written in the mnemonic form Pr{bl} is just

$$\Pr\{\text{bl}\} = \frac{\sum_{x \in S} p_x \beta_x}{\sum_{x \in S} p_x \alpha_x} \tag{1}$$

where

β_x = number of idle inlet-outlet pairs that are blocked in state x.
α_x = number of idle inlet-outlet pairs in state x.

The mean of the number of calls in progress is

$$m = \sum_{x \in S} |\,x\,|\, p_x$$

and its variance is

$$\sigma^2 = \sum_{x \in S} (|\,x\,|\, - m)^2 p_x .$$

It has been shown[4] that for a "one-sided" network ν of T terminals we have

$$1 - \Pr\{\text{bl}\} = \frac{1}{\lambda} \frac{2m}{(T - 2m)^2 - T + 2m + 4\sigma^2}. \tag{2}$$

This formula relates the important parameters of the system, limiting their possible values to a surface in five dimensions.

III. TRAFFIC IN NETWORKS

It is important to equip a reader with intuitive motivation for mathematical procedures and results, and to do this at an early enough stage in the exposition of work for it to be helpful. With this motivation in mind we proceed with a discussion of certain traffic theory topics, to which the ensuing mathematics is most directly relevant.

The best-known and most widely used results in telephone traffic theory are undoubtedly those deduced in Erlang's classical model for a finite trunk group: c trunks, Poisson arrivals at rate $a > 0$, negative exponential holding-times, and blocked calls cleared without retrials. As is familiar, the probability of k calls in progress in equilibrium in this model is

$$p_k = \frac{\frac{a^k}{k!}}{\sum_{j=0}^{c} \frac{a^j}{j!}}, \quad k = 0, \cdots, c,$$

the probability of blocking is just $E(c,a) = p_c$, the load offered is a, the load carried is $m = a(1 - p_c)$, and the load variance is

$$\sigma^2 = m - ap_c(c - m).$$

It is important to note precisely just what is given, and what is calculated, in this model. The attempt rate and the number of trunks are given, and all else is calculated from a and c. This is because there is no "finite-source effect" here, no diminution of the instantaneous calling rate when many calls are in progress.

In a telephone connecting network model with a finite number of terminals, however, the finite source effect is inescapable. The attempt rate (or offered load, if the mean holding time is the unit of time) is not given *a priori*, but must itself be determined from the statistical equilibrium equations. This fact is sometimes overlooked. The same circumstance applies to the carried loads, whether the total load or simply the loads on particular parts (e.g., junctors or links, or groups

thereof) within the networks; all these loads are functions of network structure and operation (e.g., routing) and are not given *a priori*. It is, nevertheless, a common practice to assume that such loads are known.[4,5]

For the reasons cited in the foregoing paragraph, the fact that the basic relationship (2) obtains between m, $\Pr\{\text{bl}\}$, σ^2, T, and λ assumes additional importance over and above its value as an aid to rough calculation.

Also, while it is inescapable, the finite source effect may nevertheless be demonstrably negligible. For example, T may be so large and λ so small that the finite source effect is virtually absent and can be neglected: almost everyone is idle almost all of the time. On the other hand, this may not happen, and thus it is important to be able to foretell to some extent when it does. Our analyses provide (among other results) some upper bounds on how large the finite source effect in a particular model actually is, and thus are of use in deciding whether or not it can be ignored.

It is known that there are respects in which either the Poisson distribution $e^{-a}(a^j/j!)$, $j = 0, 1, 2, \cdots$, or sometimes the truncated Poisson or Erlang distribution

$$\frac{\frac{a^j}{j!}}{\sum_{i=0}^{c} \frac{a^i}{i!}} \; j = 0,1, \cdots, c,$$

plays a boundary or limiting role for the equilibrium distribution of the number of calls in progress in various stochastic telephone traffic models. Examples of this phenomenon abound. In Palm's "infinite trunk" model[1] this equilibrium distribution is exactly the Poisson; in Erlang's classical model[1] for c trunks, Poisson arrivals, and lost calls cleared, it is exactly the truncated Poisson.

Further, it has been shown[6] that if the present model is used to describe the operation of a nonblocking network, then as $\lambda \to 0$ and $T \to \infty$ with $\frac{1}{2}\lambda T^2 = a = \text{constant}$, the distribution of the number of calls in progress approaches the Poisson with mean a. Finally, it is suggestive but perhaps less directly relevant that in the present model the expansion of $\Pr\{|x_t| = k\}$ in powers of λ has the form

$$\Pr\{|x_t| = k\} = p_0(\lambda^k/k!)u_k + o(\lambda), \qquad \lambda \to 0$$

where p_0 is the probability that no calls are in progress (p_0^{-1} = normalization constant) and u_k is a constant depending only on the structure of the network and on the routing rule R used.[3]

All these facts suggest that the relationships of the distribution $\{p_k\}$ of the number of calls in progress to various possible distributions similar in algebraic character to the truncated Poisson should be explored in a systematic way, with special efforts to establish rigorous inequalities for quantities of interest in terms of truncated Poisson distributions. Such inequalities are obtained in the sequel.

Applications of the inequalities are numerous. A particularly important one provides a precise form of the following natural approximation procedure: In most telephone systems, the chance of having many more calls in progress than the average will be small; hence little error will be incurred in the calculation of loss if the states with many more calls in progress than the average are omitted from the sums defining [cf. (1)] the loss.

IV. SUMMARY AND CONCLUSIONS

Some of the problems, ideas, and observations that motivated the inequalities to be presented here were considered informally in Section III. The problem of estimating the extent of the "finite source effect", and the fact that distributions related to the Poisson or Erlang distributions give rigorous and useful bounds in traffic theory, were both mentioned.

In Section V we discuss the distribution of the number of calls in progress, and remark on the basic inequality

$$\Pr\{k \text{ calls in progress}\} \leqq \Pr\{\text{no calls in progress}\} \cdot \frac{\lambda^k}{k!} \prod_{j=0}^{k=1} \binom{T-2j}{2}. \tag{3}$$

The distribution of the number of calls in progress, it is to be recalled,[1] entirely determines the load carried and the load offered in a "one-sided" network; thus it also determines the probability of loss, by (2).

Section VI contains two analytical lemmas on which all the ensuing inequalities are based. The first merely observes that all extrema of a bilinear functional on a polyhedron must be achieved at the vertices. The second lemma is used over and over again in the sequel and for this reason it is called the "basic lemma." For certain special convex polyhedra and bilinear functionals, it pinpoints that extreme point of the polyhedron at which the functional assumes its maximum. Many problems of traffic theory lead to polyhedra and functionals of just these special types, whence their relevance.

A network ν, together with a routing rule R for ν, is called a *system*. There is a natural map μ which takes a system (ν,R) into the distribu-

tion of the number of calls in progress induced by (ν,R) (for a stochastic process x_t describing the operation of ν under R and under the traffic assumptions of Section II). Section V is devoted to proving a basic preliminary result to the effect that if ν carries at most w calls then for any R the induced distribution of the number of calls in progress belongs (if normalized so that $\Pr\{x_t = 0\} = 1$) to a special convex set of $(w + 1)$ dimensions, describable in terms of w, λ, and T, and closely related to the factors

$$b_k = \frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \binom{T - 2j}{2}, \qquad k = 1, \cdots, w, \tag{4}$$

appearing in (3) and in the theory of traffic in nonblocking networks.[6]

All the preceding preliminary results are combined in Section VIII to prove a principal inequality for ratios of expectations: For nondecreasing nonnegative $f(\cdot)$ and positive nonincreasing $g(\cdot)$,

$$\max_{\nu,R} \frac{E\{f(\,|\,x_t\,|\,)\}}{E\{g(\,|\,x_t\,|\,)\}} = \frac{\sum_{j=0}^{w} f(j)\,b_j}{\sum_{j=0}^{w} g(j)\,b_j},$$

where b_k are as in (4), and the maximum is over ν and R appropriate to ν such that ν has T terminals and carries at most w calls. In Section IX we make direct applications of this result to the mean load carried and to the attempt rate.

The extent of the finite source effect is estimated in Section X in terms of the quantities b_k of (4), or, more roughly, in terms of the Erlang loss function $E(c,a)$, with $a = \frac{1}{2}\lambda T^2$. Section XI next considers the problem of estimating the equilibrium chance that more than k calls are in progress; again, this is done in terms of the b_k, and also by means of Erlang's function, using the basic lemma. Estimates of this probability have important applications to studying the error incurred in omitting states with more than k calls in progress in the sums in (1), defining loss.

It is natural to expect that in most telephone systems the probability that an immoderately large number (about twice the average number) of calls be in progress is small. This expectation suggests omitting states with more than k calls in progress from the sums defining loss, for some suitable k, as an approximation. The next two sections, XII and XIII, are concerned with the magnitude and the sign, respectively, of the error in this approximation. Two of the results are simple enough to paraphrase: Theorem 5: In virtually all cases of practical interest, if $\Pr\{|\,x_t\,| > k\} \leqq [p/(1 + p)]$, then omitting states with more than k

calls in progress in calculating Pr{bl} [by (1)] will not result in an error of more than $100p$ per cent. Theorem 6: If $\Pr\{|x_t| > k\} \leqq \epsilon$, then omitting states ... [etc., as above] will not result in an absolute error of more than ϵ.

Because the loss, Pr{bl}, is a bilinear (or linear fractional) functional of the state probabilities, determining the sign of the error incurred in the approximation under discussion is usually not simple. This sign depends (Theorems 7, 8, 9) on whether or not the fraction of *hangups* made with more than $k + 1$ calls in progress exceeds the fraction of *attempts* made with more than k calls in progress. In particular, if the fraction of *attempts* made with at most k calls in progress is not more than a certain expression (15) involving k and Erlang's loss function the approximation is an underestimate; whereas if the fraction of *hangups* made with at most $k + 1$ calls in progress exceeds another similar expression (16), then the approximation overestimates loss.

Other approximations are considered in Section XIV. A natural one is omission of states with more than k calls in progress in (2) for loss in terms of the mean and variance of the load. The basic lemma implies that this approximation is *always* an upper bound.

The final section, XV, exhibits a simple upper bound on the loss in terms of a bound on the number of blocked idle terminal-pairs in a state with k calls up, i.e., a bound of the form

$$\beta_x \leqq f_{|x|}, \qquad f(\cdot) \text{ increasing.} \tag{5}$$

This result, to be developed in a later paper,[2] provides a reasonably manageable way of converting combinatorial information about network structure directly into probabilistic inequalities about loss. The search for bounds of the form (5) for various classes and kinds of network is now one of the next most important tasks of congestion theory.

Some of the conclusions to be drawn from the present work are set down in the following list; many others will occur to those skilled in the art.

(*i*) The terms b_k, given by (4), of the distribution of the number of calls in progress in nonblocking networks can be used to give inequalities for the mean load carried, the attempt rate, the loss, and other quantities of interest arising in the study of traffic in blocking networks.

(*ii*) Terms of the Poisson or Erlang distribution, long used in trunking theory and in certain limiting cases of no congestion, can be used to give inequalities similar to, but always weaker and simpler than, those of (*i*), for the same quantities of interest.

(*iii*) The inequalities of (*i*) become those of (*ii*) in the "infinite

source" limit $\lambda \to 0$, $T \to \infty$, $\frac{1}{2}\lambda T^2 = a$ = constant, with λ the calling rate per pair of idle lines, and T the total number of lines. (This limit is interesting and relevant to practical matters.)

(*iv*) Of the networks with T terminals, the nonblocking ones carry the most load and have the smallest attempt rate; among those that can carry at most w calls, the networks which are nonblocking up to w calls in progress, and block completely at w calls in progress, carry the most load and have the smallest attempt rate.

(*v*) If a network carries at most w calls, its load per line is at most

$$\lambda T[1 - E(w,a)],$$

and the equilibrium chance that it have more than k calls in progress is at most

$$1 - a^{w-k}\frac{k!}{w!}\frac{E(k,a)}{E(w,a)},$$

where $E(\cdot,a)$ is Erlang's loss function and $a = \frac{1}{2}\lambda T^2$.

(*vi*) In almost all cases, omitting states with more than three times the average number of calls in progress from the sums in (1) defining loss will result in at most a 50 per cent error in the loss.

(*vii*) Omission, in calculating loss by (1), of all states with more than k calls in progress will result in an underestimate if k is low enough. If (2) is used, this omission always overestimates loss.

(*viii*) If $\Pr\{x_t > k\} < \epsilon$, the above omission makes an absolute error of at most ϵ, if formula (1) is used.

(*ix*) Any bound $\beta_x \leqq f_{|x|}, f_{|}$, on the number of blocked idle terminal pairs in a state x at once yields the inequality

$$\Pr\{\text{bl}\} \leqq \frac{\sum_{j=0}^{w} f_j \frac{a_j}{j!}}{\sum_{j=0}^{w} \alpha_j \frac{a^j}{j!}}.$$

When the right-hand side is within an order of magnitude of the left, this result puts a large premium on combinatorial studies in networks of the rate at which blocking goes up with number of calls in progress.

V. THE DISTRIBUTION OF THE NUMBER OF CALLS IN PROGRESS

The calculation of the call-congestion, or probability of blocking $\Pr\{\text{bl}\}$, reduces in general to that of the stationary state-probabilities

$\{p_x, x \in S\}$. In the case of one-sided connecting networks, however, the basic formula (2) shows that a knowledge of the equilibrium mean and variance of the number of calls in progress is sufficient to determine the call-congestion. It is particularly important, then, to study the distribution of the number of calls in progress very carefully, because: (a) it contains all the information necessary to calculate congestion; (b) being a distribution of probability over a finite subset of the integers, it is a much simpler object than the distribution $\{p_x\}$ over S; (c) without doubt, it is much easier to approximate than $\{p_x\}$ itself, and so is much more likely to be useful. Various properties, inequalities, etc., pertaining to this distribution are studied in this section.

We use the notation

$$p_k = \sum_{|x|=k} p_x$$

for the probability that k calls are in progress in equilibrium. We know from Lemma 1 of Ref. 3 that for $1 \leqq k \leqq w = \max_{x \in S} |x|$,

$$kp_k = \lambda \sum_{|x|=k-1} p_x s(x). \tag{6}$$

This formula expresses the fact that in equilibrium the average rate of entrances into the set $\{x: |x| \geqq k\}$ must equal the average rate of exits from this set.

Unfortunately, (6) does not in general permit an actual calculation of $\{p_k\}$, because it depends, on the right, on the actual distribution of probability over $\{x: |x| = k - 1\}$, and not merely on p_{k-1}. However, let us observe (*i*) that if it takes more than $k - 1$ calls in progress to block any call at all then

$$s(x) = \binom{T - 2|x|}{2}, \qquad \text{for} \qquad |x| = \leqq k - 1,$$

and (*ii*) that in any case

$$s(x) \leqq \binom{T - 2|x|}{2}.$$

Thus, if n is the minimum number of calls which must be in progress in order that there by any blocked calls at all, we find that

$$kp_k = \lambda p_{k-1} \binom{T - 2k + 2}{2}, \qquad 1 \leqq k \leqq n$$

$$kp_k \leqq \lambda p_{k-1} \binom{T - 2k + 2}{2}, \qquad 1 \leqq k \leqq w.$$

Iteration of these relations then gives

$$\begin{aligned} p_k &= p_0 \frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \binom{T-2j}{2}, \qquad 1 \leqq k \leqq n \\ p_k &\leqq p_0 \frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \binom{T-2j}{2}, \qquad 1 \leqq k \leqq w. \end{aligned} \tag{7}$$

The bound on the right of this last inequality has the same form as the exact formula for p_k for a nonblocking network.[6] This implies that for λ fixed the maximum possible value of the ratios

$$(p_k/p_0) \qquad k = 1, \cdots, w$$

is achieved by nonblocking networks, and it is achieved by a blocking network at a particular value of k only if

$$\frac{1}{p_{k-1}} \sum_{|x|=k-1} s(x) p_x = \binom{T-2k+2}{2}$$

i.e., only if the conditional expectation of $s(\cdot)$ given that $k-1$ calls are in progress is the number

$$\binom{T-2k+2}{2}.$$

Since this is an upper bound for $s(\cdot)$ over all x with $|x| = k-1$, this means that all the probability is concentrated on the nonblocking states, so that the bound (7) is also achieved for $k-1$. (This observation will be fundamental in the proof of Lemma 3.)

Reasoning from (6) leads to the inequalities

$$\lambda p_{k-1} \min_{|y|=k-1} s(y) \leqq k p_k \leqq \lambda p_{k-1} \max_{|y|=k-1} s(y)$$

and thence by iteration to the

Remark:

$$\frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \min_{|y|=j} s(y) \leqq \frac{p_k}{p_0} \leqq \frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \max_{|y|=j} s(y).$$

This result indicates (to a first approximation) how the values assumed by the "success" function $s(\cdot)$ on S affect the distribution of the number of calls in progress, and through it, the congestion or probability of blocking. Obviously, the nearer the network is to being nonblocking, i.e., the nearer $s(\cdot)$ comes to assuming the value

$$\binom{T-2\,|\,x\,|}{2}$$

for state x, the closer p_k will be to its upper bound (7), and the less will be the congestion.

VI. TWO PRELIMINARY RESULTS

Lemma 1: Let P be a polyhedron in n-dimensional Euclidean space, and let

$$F(x) = \frac{c_1 + (a,x)}{c_1 + (b,x)}, \qquad (\cdot\,,\cdot) = \textit{inner product}$$

be a bilinear (or linear fractional) function of the n-vector x, such that the plane $c_2 + (b,x) = 0$ does not intersect P. Then the extreme values of $F(\cdot)$ on P are assumed at the vertices of P.

Proof: Let x be a point interior to P. Since the sign of

$$\frac{\partial F}{\partial x_i} = \frac{a_i[c_2 + (b,x)] - b_1[c_1 + (a,x)]}{[c_2 + (b,x)]^2}$$

does not depend on x_i, we can find another point $y \in \partial P$ (the boundary of P) such that $F(x) \leqq F(y)$. The point y will be on a face P_1 of P determined by a linear condition $(c,x) = \alpha$ which can be used to eliminate one of the variables from $F(\cdot)$ to get a new bilinear function $F_1(\cdot)$ of $(n-1)$ variables agreeing with $F(\cdot)$ on P_1. Except for dimension, the problem of maximizing $F_1(\cdot)$ over P_1 is of exactly the same form as that maximizing $F(\cdot)$ over P. The result is true for $n = 1$, and hence for all $n \geqq 1$. The argument for minima is dual.

Basic Lemma (Lemma 2): Let $\lambda = (\lambda_0, \lambda_1, \cdots, \lambda_n)$ be a vector of $(n+1)$ positive numbers, and let Λ be the closed convex hull of the points

$$\begin{array}{lllll}
\lambda_0, & 0, & 0, & \cdots, & 0 \\
\lambda_0, & \lambda_1, & 0, & \cdots, & 0 \\
\lambda_0, & \lambda_1, & \lambda_2, & \cdots, & 0 \\
 & & \vdots & & \\
\lambda_0, & \lambda_1, & \lambda_2, & \cdots, & \lambda_n .
\end{array}$$

Let $f(\cdot)$ be nondecreasing and nonnegative, and let $g(\cdot)$ be nonincreas-

ing and positive, on $\{0,1, \cdots, n\}$. *Then with* $(\cdot\ ,\ \cdot) =$ *inner product,*

$$\max_{v \in \Delta} \frac{(f,v)}{(g,v)} = \frac{(f,\lambda)}{(g,\lambda)}.$$

Proof: It follows from Lemma 1 that the maximum is assumed at a vertex. However, as can be verified,

$$\frac{\sum_{j=0}^{k} f_j\lambda_j}{\sum_{j=0}^{k} g_j\lambda_j} \leqq \frac{\sum_{j=0}^{k+1} f_j\lambda_j}{\sum_{j=0}^{k+1} g_j\lambda_j}, \quad k = 0, 1, \cdots n-1.$$

VII. BASIC INCLUSION

Let I be the set of inlets, and Ω that of outlets, of a possible or intended connecting network to be used for making calls from I to Ω. By a *network* ν *for* I *and* Ω we mean a quadruple

$$\nu = (G,I,\Omega,S)$$

where G is a linear graph indicating network structure, I and Ω are respectively the inlets and outlets of the network, and S is the set of permitted states.[1] The letter w is used to stand for the largest possible number of calls in progress; thus

$$w = \max_{x \in S} |\, x \,|\,.$$

If ν is one-sided, $I = \Omega$ and $w \leqq [\frac{1}{2} \mid I \mid]$. If ν is two-sided, $I \cap \Omega = \theta$ and $w \leqq \min \{ \mid I \mid, \mid \Omega \mid \}$.

By a *system for* I *and* Ω we mean a pair (ν,R) with ν a network for I and Ω and R a routing rule defined on the states $x \in S = S(\nu)$ and satisfying the conditions of Section II. It follows from the theoretical assumptions made in Section II that, together with a value of the traffic parameter $\lambda > 0$, ν and R determine a stochastic process x_t taking values on S with a stationary distribution

$$\{p_x = p_x(\nu,R,\lambda),\ x \in S(\nu)\}$$

determined by the equilibrium condition (cf. Section II).

We shall assume that I, Ω, and λ are fixed, and shall omit indications of dependence on these notions or numbers.

Let $\mathfrak{S}_n$ denote the set of systems for I and Ω such that $w \leqq n$. With

$$p_k = p_k(\nu,R) = \sum_{\substack{|x|=k \\ x \in S(\nu)}} p_x(\nu,R)$$

the map $\mu(\cdot\ ,\ \cdot)$ is defined on $\mathcal{S}_n$ for each value of $\lambda > 0$ by

$$\mu\colon (\nu,R) \to \frac{(p_0, p_1, \cdots, p_n)}{p_0}, \qquad p = p(\nu,R),$$

i.e., its value for (ν,R) is the distribution of the number of calls in progress in the associated stochastic process x_t, normalized so that $p_0 = 1$.

Lemma 3: Let $n \leqq [\frac{1}{2}T]$, *and let*

$$b_0 = 1$$

$$b_k = \frac{(\frac{1}{2}\lambda)^k}{k!}\frac{T!}{(T-2k)!}, \qquad k = 1, \cdots, n. \tag{8}$$

Let C be the closed convex hull of the $(n+1)$*-dimensional points*

$$\begin{aligned} c_0 &= (1,0,0, \cdots, 0) \\ c_1 &= (1, b_1, 0, \cdots, 0) \\ c_2 &= (1, b_1, b_2, \cdots, 0) \\ &\vdots \\ c_n &= (1, b_1, b_2, \cdots, b_n). \end{aligned}$$

Then

$$\mu(\mathcal{S}_n) \subseteq C,$$

i.e., C includes the μ*-image of* $\mathcal{S}_n$.

Proof: We show first that each c_i, $i = 0, \cdots, n$ is in fact in the image of $\mathcal{S}_n$ under $\mu(\cdot)$. Let ν_0 be the trivial network containing no crosspoints, and let R_0 be the trivial rule that says nothing. Then

$$\mu\colon(\nu_0,R_0) \to c_0,$$

and $c_0 \in \mu(\mathcal{S}_n)$. Now let ν_k, $k = 1, \cdots, n$, be a "one-sided" network consisting (*i.*) of a concentrator taking T terminals to $2k$ in a nonblocking manner, and (*ii.*) of a nonblocking "one-sided" network on those $2k$ terminals. In such a network, obviously, a state is nonblocking if fewer than k calls are in progress; the network blocks up completely as soon as k calls are in progress. It follows from the arguments for Theorem 1 of Ref. 6 that for any routing rule R_k appropriate to ν_k,

$$\mu\colon(\nu_k,R_k) \to c_k.$$

From formula (7) of Section V we know that

$$\frac{p_k(\nu,R)}{p_0(\nu,R)} = \frac{p_k}{p_0} \leqq b_k.$$

Hence, to show that $\mu(\mathcal{S}_n)$ is contained in C, it suffices to show that if for some $(\nu,R) \in \mathcal{S}_n$ and some $1 \leqq k \leqq n$,

$$(p_k/p_0) = b_k\,, \qquad p = p(\nu,R)$$

then for all $1 \leqq j \leqq k$,

$$(p_j/p_0) = b_j\,.$$

Suppose then that $p_k/p_0 = b_k$. Using (6) we find

$$\begin{aligned} kp_k = \lambda \sum_{|y|=k-1} p_y s(y) &= \lambda p_{k-1} E\{s(x_t) \mid |x_t| = k-1\} \\ &= p_0 k b_k \\ &= p_0 \lambda \binom{T-2k+2}{2} b_{k-1}\,. \end{aligned}$$

Hence,

$$\frac{p_{k-1}}{p_0} = b_{k-1} \frac{\binom{T-2k+2}{2}}{E\{s(x_t) \mid |x_t| = k-1\}}\,.$$

But,

$$\max_{|y|=k-1} s(y) \leqq \binom{T-2k+2}{2}$$

so the ratio is $\geqq 1$. But we know from (7) that $p_{k-1}/p_0 \leqq b_{k-1}$. Hence, equality holds, and by iteration,

$$p_j/p_0 = b_j\,, \qquad 1 \leqq j \leqq k.$$

VIII. PRINCIPAL INEQUALITY FOR RATIOS OF EXPECTATIONS

We now combine Lemmas 2 and 3 to obtain a basic inequality for ratios of expectations. Applications of this result to the quantities of interest in traffic engineering appear in the following Sections IX through XIV.

Theorem 1: If $f(\cdot)$ is nondecreasing and nonnegative, and $g(\cdot)$ is nonincreasing and positive, on $\{0,1,\cdots,n\}$, then

$$\max_{(\nu,R)\in\mathcal{S}_n} \frac{E\{f(|x_t|)\}}{E\{g(|x_t|)\}} = \frac{\sum_{j=0}^{n} f(j)b_j}{\sum_{j=0}^{n} g(j)b_j}\,,$$

(the expectation being calculated with respect to the stationary probabilities associated with ν and R.)

Proof: By Lemma 3, the image of $\mathcal{S}_n$ under $\mu(\cdot\ ,\ \cdot)$ is contained in the closed convex hull C of the points $c_0\ ,\ c_1\ ,\ \cdots\ ,\ c_n$. By the basic lemma, the maximum of the functional

$$\xi(r) = \frac{(f,r)}{(g,r)} = \frac{\sum_{j=0}^{n} f(j)r_j}{\sum_{j=0}^{n} g(j)r_j}$$

for $r \in C$ is assumed at $r = c_n$.

IX. INEQUALITIES FOR THE MEAN AND THE ATTEMPT RATE

Let ν and ν' be two connecting networks with the same number of terminals, and the same offered traffic λ per idle pair. If ν is nonblocking, it is our intuitive expectation that it will carry at least as great a load as ν', and that (since more lines of ν are busy on the average than of ν') the attempt rate for ν' will be at least as great as that for ν. It is being assumed here, of course, that in each case the operation of the network is being represented by a stochastic process x_t of the type described in Section II, with

$$m = \text{carried load} = \frac{\sum_{k=1}^{w} kp_k/p_0}{1 + \sum_{k=1}^{w} p_k/p_0}$$

$$\lambda E\{\alpha_{|x_t|}\} = \text{attempt rate} = \lambda \frac{\sum_{k=0}^{w} p_k/p_0 \binom{T-2k}{2}}{1 + \sum_{k=1}^{w} p_k/p_0}.$$

Arguments ν,R are used in the next three results to indicate dependence on the network ν and the routing rule R under discussion.

Theorem 2: Let ν be a one-sided network of T terminals, and let $w = w(\nu) = \max_{x \in S(\nu)} |x|$, $a = \frac{1}{2}\lambda T$.[2] Then for any R such that $(\nu,R) \in \mathcal{S}_w$

$$m(\nu,R) \leqq \frac{\sum_{j=1}^{w} jb_j}{1 + \sum_{j=1}^{w} b_j} \leqq a[1 - E(w,a)].$$

Proof: For the first inequality let $f(j) = j$ and $g \equiv 1$ in Theorem 1; for the second, use the basic lemma and $b_j < a^j/j!$.

Corollary 1: Let (ν,R) *and* (ν',R') *belong to* $\mathcal{S}_w$ *for some integer* w, *with both* ν *and* ν' *one-sided. If* ν' *is nonblocking until* w *calls are in progress, then with* $a = \frac{1}{2}\lambda T^2$

$$m(\nu,R) \leqq m(\nu',R') = \frac{\sum_{j=1}^{w} jb_j}{1 + \sum_{j=1}^{w} b_j} \leqq a[1 - E(w,a)].$$

The following properties of the Erlang loss function

$$E(c,a) = \frac{\frac{a^c}{c!}}{\sum_{j=0}^{c} \frac{a^j}{j!}}$$

are used:

$$\frac{\sum_{j=1}^{c} j \frac{a^j}{j!}}{\sum_{j=0}^{c} \frac{a^j}{j!}} = a[1 - E(c,a)],$$

$$\frac{\sum_{j=1}^{c} j^2 \frac{a^j}{j!}}{\sum_{j=0}^{c} \frac{a^j}{j!}} = a[1 - E(c,a)] - a^2E(c,a)\left[\frac{c}{a} - 1 + E(c,a)\right] + a^2[1 - E(c,a)]^2,$$

$$= (a + a^2)[1 - E(c,a)] - acE(c,a).$$

Theorem 3: Let ν *be a one-sided network of* T *terminals. Then for* $w = w(\nu)$, $a = \frac{1}{2}\lambda T^2$, *and any* R *such that* $(\nu,R) \in \mathcal{S}_w$,

$$E\{\alpha_{|x_t|}\}_{\nu,R} \geqq \frac{\sum_{j=0}^{w} \binom{T-2j}{2} b_j}{1 + \sum_{j=1}^{w} b_j} \geqq \frac{\sum_{j=0}^{w} \binom{T-2j}{2} \frac{a^j}{j!}}{\sum_{j=0}^{w} \frac{a^j}{j!}}$$

Proof: The first inequality follows from taking $f \equiv 1$ and $g(j) = \alpha_j$ in Theorem 1; the second, from the basic lemma. The last term on the right is expressible in terms of Erlang's loss function as

$$\binom{T}{2} - a(2T - 3 + 2a)[1 - E(w,a)] - 2awE(w,a).$$

X. APPLICATIONS TO ESTIMATING THE "FINITE-SOURCE EFFECT"

It is reasonable to expect that in a telephone system with a large number T of terminals, each one contributing only a small amount of traffic, the "finite-source effect" will be small. Since the finite-source effect is a diminution of the instantaneous calling rate due to busy terminals, it is properly measured by the *fraction of busy terminals*, i.e., in our model, the quantity

$$q = 2m/T = \text{load per customer's line in erlangs.}$$

When T is so large and λ so small that q is very small we might with justifiable confidence replace our finite-source model with an infinite source model. One way of doing this is to consider a sequence of connecting networks which concentrate traffic from more and more terminals into a sub-connecting network of fixed structure. However, we do not here digress into a detailed consideration of this transition; the basic idea has been at the heart of applications of the Poisson arrival process in telephone traffic theory since its beginning. Instead, we obtain an upper bound on q in terms of T and λ; this bound provides a conservative estimate of the negligibility of the finite source effect.

Corollary 2: With $a = \frac{1}{2}\lambda T^2$, $E(c,a)$ *the (first) Erlang loss function, and* $b_0, b_1, \cdots, b_w$ *as in formula (8),*

$$q \leqq \frac{2}{T}\frac{\sum_{j=1}^{w} jb_j}{\sum_{j=0}^{w} b_j} \leqq \frac{2a}{T}\{1 - E(w,a)\} \qquad (w = \max_{x \in S} |x|)$$

$$\leqq \frac{2a}{T}\left(1 - \frac{e^{-a}a^w}{w!}\right)$$

$$\leqq \frac{2a}{T} = \lambda T.$$

Proof: The first inequality follows from Theorem 1, the second from the basic lemma, and the third from

$$E(w,a) = \frac{\frac{a^w}{w!}}{\sum_{j=0}^{w} \frac{a^j}{j!}} \geqq \frac{e^{-a}a^w}{w!}.$$

Alternatively, since the finite-source effect is a diminution of the calling rate due to busy terminals, one can also estimate it in terms of

the difference between the *maximal* calling rate $\lambda\binom{T}{2}$ when no terminals are busy and the *average* calling rate

$$\lambda E\{\alpha_{|x_t|}\} = \lambda \sum_{j=0}^{w} p_j \binom{T-2j}{2}.$$

This estimate is covered in the

Corollary 3: The average diminution

$$D = \lambda \binom{T}{2} - \lambda E\{\alpha_{|x_t|}\}$$

in calling rate due to busy terminals satisfies the inequality

$$\begin{aligned} D &\leqq \lambda a(2T - 3 + 2a)[1 - E(w,a)] + \lambda a w E(w,a) \\ &\leqq O(T^{-1}) \quad \text{as} \quad \lambda \to 0, \qquad T \to \infty, \qquad a = \tfrac{1}{2}\lambda T^2. \end{aligned}$$

Proof: Theorem 3 and the known properties of $E(\cdot,a)$.

XI. ESTIMATE OF THE CHANCE OF MORE THAN k CALLS IN PROGRESS

The chance $\Pr\{|x_t| > k\}$ is a quantity that is useful in estimating the extent of the finite source effect, and the error incurred in ignoring states with more than k calls in progress in calculating loss. (See Section XII.) Upper bounds for it are given in

Theorem 4: If ν is one-sided, $w = w(\nu)$, and $a = \frac{1}{2}\lambda T^2$, then

$$\begin{aligned} \Pr\{|x_t| > k\} &\leqq \frac{\sum_{j=k+1}^{w} b_j}{1 + \sum_{j=1}^{w} b_j} \\ &\leqq 1 - \frac{a^{w-k} k!}{w!} \frac{E(k,a)}{E(w,a)}. \end{aligned}$$

Proof: For the first inequality, choose

$$f(j) = \begin{cases} 0 & j \leqq k \\ 1 & j > k \end{cases}$$

and $g(\cdot) \equiv 1$ in Theorem 1; the second follows from the basic lemma.

XII. APPROXIMATION THEOREMS FOR THE PROBABILITY OF BLOCKING

In any telephone system that provides adequate service the probability of a substantially larger than average number of calls in progress

will be small. Thus, in using the formula for probability of blocking,

$$\Pr\{\text{bl}\} = \frac{\sum_{x \epsilon S} p_x \beta_x}{\sum_{x \epsilon S} p_x \alpha_x},$$

it should be possible to omit states with more than k calls in progress from the sums, without incurring too much error. It is the purpose of this section to examine this possibility rigorously. In particular, we wish to answer the following very important question: If p is a given positive number, how large must k be so that the omission of states with more than k calls in progress, i.e., the approximation

$$\Pr\{\text{bl}\} \approx \frac{\sum_{|x| \leqq k} p_x \beta_x}{\sum_{|x| \leqq k} p_x \alpha_x}$$

results in an error of at most $100p$ per cent?

In what follows we shall make systematic use of the following abbreviations:

$$r = \sum_{|x| \leqq k} p_x s(x) \tag{9}$$

$$s = \sum_{|x| \leqq k} p_x \alpha_x \tag{10}$$

$$u = \sum_{|x| > k} p_x s(x) \tag{11}$$

$$v = \sum_{|x| > k} p_x \alpha_x .$$

(The notation b for the probability of blocking, used in Ref. 3, e.g., is being avoided in favor of $\Pr\{\text{bl}\}$.) It can be seen that

$$r + u = \frac{\text{success rate}}{\lambda} = \frac{m}{\lambda}$$

$$s + v = \frac{\text{attempt rate}}{\lambda}$$

$$1 - \Pr\{\text{bl}\} = \frac{r + u}{s + v}.$$

Thus, omitting states with more than k calls in progress in calculating $1 - \Pr\{\text{bl}\}$ is equivalent to approximating it by

$$r/s. \tag{12}$$

We note that

$$v/(s + v)$$

is the fraction of attempts made with more than k calls in progress.

Lemma 4: $v/(s + v) \leqq Pr\{|\, x \,| > k\}$.

Proof: We have

$$\frac{v}{s + v} = \frac{\sum_{|x|>k} p_x \alpha_x}{\sum_{|x| \leq k} p_x \alpha_x + \sum_{|x|>k} p_x \alpha_x}.$$

Let $\alpha = \max\{\alpha_x : |\, x \,| = k\}$. Then because $\alpha_{(\cdot)}$ is antitone on S

$$\frac{v}{s + v} \leqq \frac{v}{\alpha \Pr\{|\, x \,| \leqq k\} + v}.$$

Since for $\mu > 0$

$$\frac{d}{dt} \frac{t}{\mu + t} = \frac{\mu}{(\mu + t)^2} > 0$$

we can replace v in the last inequality by its majorant

$$\alpha \Pr\{|\, x \,| > k\},$$

which proves the lemma.

Also, it is seen that

$$\frac{v - u}{s + v - r - u} = \frac{\sum_{|x|>k} p_x \beta_x}{\sum_{x \in S} p_x \beta_x}$$

is the fraction of blocked attempts that occur when more than k calls are in progress.

Theorem 5: If, simultaneously, the fraction of attempts that are blocked with more than k calls in progress is at most $p/(1 + p)$, *and* $Pr\{|\, x \,| > k\} \leqq p/(1 + p)$, *then omitting states with more than k calls in progress when using* (1) *for* $Pr\{bl\}$ *will not result in an error of more than* 100p *per cent.*

Proof: It suffices to show that

$$\left| \frac{r + u}{s + v} - \frac{r}{s} \right| \leqq p \left(1 - \frac{r + u}{s + v} \right).$$

The hypothesis and Lemma 4 imply that

$$\frac{v}{s+v} \leqq \frac{p}{1+p}$$

and hence that

$$v \leqq ps,$$

$$v\left(1 - \frac{r}{s}\right) \leqq p\,(s - r)$$

$$(1 + p)\,(r + v) \leqq p\,(s + v) + \frac{r}{s}(s + v).$$

Since $u \leqq v$, we have

$$(1 + p)\,(r + u) \leqq \left(p + \frac{r}{s}\right)(s + v)$$

or

$$\frac{r+u}{s+v} - \frac{r}{s} \leqq p\left(1 - \frac{r+u}{s+v}\right),$$

which is one half of the requisite inequality. For the other half, the hypothesis gives

$$\frac{v-u}{s-r} \leqq p$$

$$v - u \leqq p(s - r)$$

$$\frac{r}{s}v \leqq u + p\,(s - r)$$

$$\leqq pv + (1 - p)u + p(s - r)$$

$$= u + p(s + v - r - u)$$

so that

$$\frac{v\frac{r}{s} - u}{s+v} \leqq p\left(1 - \frac{r+u}{s+v}\right)$$

or

$$\frac{r+u}{s+v} - \frac{r}{s} \geqq -p\left(1 - \frac{r+u}{s+v}\right).$$

Since $\beta \leqq \alpha$,

$$\frac{v - u}{s + v - r - u} \leqq \frac{\alpha_{k+1} \Pr\{| x_t | > k\}}{s + v - r - u}.$$

Hence, the hypotheses of Theorem 5 are satisfied if both

$$\Pr\{| x_t | > k \} \leqq \frac{p}{1 + p}$$

and

$$\Pr\{| x_t | > k\} \leqq \frac{p}{1 + p} \frac{s + v - r - u}{\alpha_{k+1}} = \frac{p}{p + 1} \frac{\Pr\{\text{bl}\}}{\alpha_{k+1}/(s + v)}$$

or if both

$$\left.\begin{aligned} \Pr\{| x_t | > k\} &\leqq \frac{p}{1 + p} \\ \Pr\{\text{bl}\} &\leqq \frac{\binom{T - 2k - 2}{2}}{s + v}. \end{aligned}\right\} \tag{13}$$

We now show that the first inequality in (13) is easily met by a choice of k that depends very simply on p and on the carried load m, and that with this choice of k, the second inequality holds for virtually all cases of interest. By Chebyshev's inequality, the first inequality is satisfied if

$$k \geqq \frac{p + 1}{p} m - 1,$$

for then

$$\Pr\{| x_t | > k\} \leqq \frac{m}{k + 1} \leqq \frac{p}{1 + p}.$$

As for the second, we have

$$s + v = \binom{T - 2m}{2} + 2\sigma^2,$$

$$0 \leqq 4\sigma^2/T^2 \leqq 1.$$

Thus,

$$\frac{\binom{T - 2k - 2}{2}}{s + v} \geqq \frac{\binom{T - 2k - 2}{2}}{\binom{T - 2m}{2}} \cdot \frac{\binom{T - 2m}{2}}{\binom{T - 2m}{2} + \frac{1}{2}T^2}$$

Since $m \gg 2p$, choosing $k = m + (m/p) - 1$ makes the first factor greater than unity. The second factor is, with $q = 2m/T$ = line usage

$$\frac{1}{1 + \frac{\frac{1}{2}T^2}{\binom{T - 2m}{2}}} = \frac{1}{1 + \frac{T^2}{(T - 2m)^2 - T + 2m}} = \frac{1}{1 + \frac{1}{(1 - q)^2} + o(1)}$$

as T becomes large. As q assumes values in the representative range 0 to 0.2, the second factor varies between 0.5 and about 0.39, if the $o(1)$ term is ignored. The strongest form of the second inequality in (13) is then roughly

$$\Pr\{\text{bl}\} \leqq 0.4,$$

and is virtually always fulfilled in cases of practical interest. Thus, for example, to obtain an error of at most 25 per cent, it is sufficient to consider only states with at most

$$5m - 1$$

calls in progress. For a 50 per cent error, only states with at most

$$3m - 1$$

calls in progress need be considered.

In many cases, especially in those in which very little is known about the actual value of the probability $\Pr\{\text{bl}\}$ of blocking, it may be desirable to assess the effect of neglecting states with more than k calls in progress on the *absolute* error rather than the *percentage* error. This situation is covered by the following simple result:

Theorem 6: Let $\epsilon > 0$ *be any positive real number. If* $Pr\{|x| > k\} \leqq \epsilon$, *then omitting states with more than k calls in progress in calculating* $\Pr\{\text{bl}\}$ *by* (I) *will not result in an absolute error of more than* ϵ.

Proof: It is sufficient to establish that

$$\left| \frac{r + u}{s + v} - \frac{r}{s} \right| \leqq \epsilon.$$

By Lemma 4, we have

$$\frac{v}{s + v} \leqq \Pr\{|x| > k\} \leqq \epsilon$$

and hence, using $u \leqq v$,

$$\frac{v}{s+v} \leqq \frac{r}{s} + \epsilon - \frac{r}{s+v}$$

$$\frac{r+u}{s+v} - \frac{r}{s} \leqq \epsilon.$$

For the other half of the requisite inequality, we observe that

$$\epsilon + \frac{u}{s+v} \geqq \epsilon \frac{r}{s}$$

$$\frac{u}{s+v} - \frac{r}{s} + \frac{r}{s}(1-\epsilon) \geqq -\epsilon.$$

From the hypothesis we have

$$\frac{v}{s+v} \leqq \epsilon$$

$$\frac{s}{s+v} \geqq 1 - \epsilon$$

$$\frac{r}{s+v} \geqq \frac{r}{s}(1-\epsilon),$$

and hence,

$$\frac{r+u}{s+v} - \frac{r}{s} \geqq -\epsilon,$$

which proves the result.

XIII. THE SIGN OF THE ERROR

In the two preceding theorems we have studied the approximation $1 - (r/s)$ to Pr{bl} (obtained by omitting from the sums in (1) states with more than k calls in progress) without considering whether this approximation will tend to overestimate or underestimate Pr{bl}. This question is now taken up.

Various intuitive arguments why $1 - (r/s)$ should lie on one side or the other of Pr{bl} come readily to mind. The number β_x of blocked idle inlet-outlet pairs in state x tends first to grow with $|x|$, but then as $|x|$ becomes large enough it must again decrease to zero, because $\beta_x \leqq \alpha_x$ = number of idle inlet-outlet pairs in state x. However, if the network cannot carry more than w calls with $w \ll \frac{1}{2}T$, it is possible that β_x is actually monotone increasing (or isotone) with respect to the

partial ordering $\leqq$ of S; since α_x is definitely monotone decreasing (or antitone) on $(S, \leqq)$, one might in this case expect that omitting the states where β_x is largest and α_x is smallest would tend to make the congestion seem to be *less* than it actually is.

Similarly, viewing (r/s) as an approximation to

$$\frac{\sum_{x \epsilon S} p_x s(x)}{\sum_{x \epsilon S} p_x \alpha_x} = 1 - \Pr\{\text{bl}\} \tag{14}$$

and noting that $s(\cdot)$ is antitone on $(S, \leqq)$, one might expect that omitting the states where $s(\cdot)$ is smallest would tend to make $1 - \Pr\{\text{bl}\}$ larger than it is.

In fact, neither of the above intuitions is always correct; omission of states x with more than k calls in progress from the sums in the ratio (14) defining $1 - \Pr\{\text{bl}\}$ sometimes gives an underestimate, and at others gives an overestimate. Roughly, if k is large enough, $1 - (r/s)$ will be an overestimate of the loss, whereas if it is too small, it will be an underestimate.

Theorem 7: If the fraction of hangups made with more than $k + 1$ calls in progress exceeds the fraction of attempts made with more than k calls in progress, then omitting states with more than k calls in progress in the calculation of $\Pr\{\text{bl}\}$ *results in an overestimate; in the opposite case, the omission results in an underestimate.*

Proof: For $t \epsilon [0,1]$, let

$$U(t) = \frac{r + ut}{s + vt}$$

so that $U(0) = r/s$ and $U(1) = 1 - \Pr\{\text{bl}\}$. It can be seen that

$$\frac{\sum_{j>k+1} jp_j}{\sum_{i=0}^{w} ip_i} = \frac{\sum_{|x|>k} s(x)p_x}{\sum_{x \epsilon S} s(x)p_x} = \frac{u}{r + u},$$

$$\frac{\sum_{|x|>k} \alpha_x p_x}{\sum_{x \epsilon S} \alpha_x p_x} = \frac{v}{s + v},$$

and that the following inequalities are all equivalent:

$$rv \geqq us$$

$$r(s + v) + ut(s + v) \geqq s(r + u) + tv(r + u)$$

$$U(t) \geqq U(1)$$

$$\frac{v}{s + v} \geqq \frac{u}{r + u}.$$

Theorem 8: If the fraction $s/(s + v)$ of attempts made with at most k calls in progress is at most

$$\frac{1 - E(k + 1,a)}{1 - E(w,a)} \frac{E(w,a)}{E(k + 1,a)} \frac{w!}{k + 1!} a^{k+1-w} \tag{15}$$

then omitting states with more than k calls in progress in calculating $Pr\{bl\}$ results in an underestimate:

$$Pr\{bl\} \geqq 1 - (r/s).$$

Proof: It can be verified that

$$1 - \frac{1 - E(k + 1,a)}{1 - E(w,a)} \frac{E(w,a)}{E(k + 1,a)} \frac{w!}{k + 1!} a^{k+1-w}$$

$$= \frac{\sum_{j=k+2}^{w} j \frac{a^j}{j!}}{\sum_{j=0}^{w} j \frac{a^j}{j!}} \geqq \frac{\sum_{j=k+2}^{w} jb_j}{\sum_{j=0}^{w} jb_j} \geqq \frac{\sum_{j>k+1}^{w} jp_j}{\sum_{j=0}^{w} jp_j} = \frac{\lambda \sum_{|x|>k} p_x s(x)}{m} = \frac{\lambda u}{m}.$$

The first equality follows from known properties of Erlang's function, the two inequalities follow from the basic lemma with $g \equiv 1$ and

$$f_j = \begin{cases} 0 & j \leqq k + 1 \\ 1 & j > k + 1, \end{cases}$$

and the last two equalities follow from (6) and the definition (11) of u, respectively. Thus the hypothesis gives

$$\frac{v}{s + v} > \lambda \frac{u}{m}$$

$$\frac{u}{v} < \frac{1}{\lambda} \frac{m}{s + v} = \frac{r + u}{s + v} = 1 - \Pr\{\text{bl}\},$$

and the argument now proceeds as in Theorem 7.

Remark:

$$\frac{s}{s+v} \leqq \frac{\sum_{j>k}^{w} \alpha_j b_j}{\sum_{j=0}^{w} \alpha_j b_j} \leqq \frac{\sum_{j>k}^{w} \alpha_j \frac{a^j}{j!}}{\sum_{j=0}^{w} \alpha_j \frac{a^j}{j!}}.$$

Proof: Basic lemma, with $\lambda_j = \alpha_j b_j$.

It follows that

$$\frac{v}{s+v} = 1 - \frac{s}{s+v} \geqq 1 - \frac{\sum_{j>k}^{w} \alpha_j b_j}{\sum_{j=0}^{w} \alpha_j b_j} \geqq 1 - \frac{\sum_{j>k}^{w} \alpha_j \frac{a^j}{j!}}{\sum_{j=0}^{w} \alpha_j \frac{a^j}{j!}}.$$

Theorem 9: If the fraction $\lambda u/m$ of hangups made with $k+1$ or fewer calls in progress is at least

$$1 - a^{k-w} \frac{w!}{k!} \frac{E(w,a)}{E(k,a)} \cdot \frac{\binom{T}{2} - a(2T - 3 + 2a)[1 - E(k,a)] - 2akE(k,a)}{\binom{T}{2} - a(2T - 3 + 2a)[1 - E(w,a)] - 2awE(w,a)}, \tag{16}$$

then omitting states with more than k calls in progress in calculating $Pr\{bl\}$ results in an overestimate:

$$Pr\{bl\} \leqq 1 - (r/s).$$

Proof: It can be verified, using the formula, for integers $c \leqq w$,

$$\sum_{j=0}^{c} \binom{T-2j}{2} \frac{a^j}{j!} = \frac{a^c}{c!} \frac{\binom{T}{2} - a(2T - 3 + 2a)[1 - E(c,a)] - 2acE(c,a)}{E(c,a)},$$

that

$$(16) = \frac{\sum_{j=k+1}^{w} \alpha_j \frac{a^j}{j!}}{\sum_{j=0}^{w} \alpha_j \frac{a^j}{j!}} \geqq \frac{\sum_{j=k+1}^{w} \alpha_j b_j}{\sum_{j=0}^{w} \alpha_j b_j} \geqq \frac{\sum_{j=k+1}^{w} \alpha_j p_j}{\sum_{j=0}^{w} \alpha_j p_j} = \frac{v}{s+v},$$

where the first equality follows from the stated formula, the two inequalities follow from the basic lemma, and the last equality from the definitions of s and v. Hence, the hypothesis implies

$$\frac{\lambda u}{m} \geqq \frac{v}{s+v},$$

$$\frac{u}{r+u} \geqq \frac{v}{s+v},$$

and the argument again proceeds as in Theorem 7.

XIV. OTHER APPROXIMATIONS

Since

$$1 - \Pr\{\text{bl}\} = \frac{1}{\lambda} \frac{\sum_{j=1}^{w} jp_j}{\sum_{j=0}^{w} \alpha_j p_j}, \tag{17}$$

one can envisage an approximation

$$\Pr\{\text{bl}\} \approx 1 - \frac{1}{\lambda} \frac{\sum_{j=1}^{k} jp_j}{\sum_{j=0}^{k} \alpha_j p_j}, \tag{18}$$

obtained by omitting states with more than k calls in progress from the sums in (17). The basic lemma implies that this approximation is *always an overestimate*. We have

Theorem 10: For each $k = 1, \cdots, w$,

$$1 - Pr\{bl\} \geqq \frac{\lambda^{-1} \sum_{j=1}^{k} jp_j}{\sum_{j=0}^{k} \alpha_j p_j} = \frac{\sum_{|x| \leqq k-1} s(x) p_x}{\sum_{|x| \leqq k} \alpha_x p_x}.$$

Proof:

$$1 - \Pr\{\text{bl}\} = \frac{1}{\lambda} \frac{\sum_{j=1}^{w} jp_j}{\sum_{j=0}^{w} p_j \alpha_j}$$

$$\geqq \frac{1}{\lambda} \frac{\sum_{j=1}^{k} jp_j}{\sum_{j=0}^{k} p_j \alpha_j},$$

the inequality following from the basic lemma. We now use formula (6):

$$jp_j = \lambda \sum_{|y|=j-1} p_y s(y), \qquad j = 1, \cdots, w.$$

We note that the bound given in Theorem 10 is exactly the estimate r/s of (12) with the top term in the numerator sum omitted. This term is just $\lambda^{-1}(k+1)p_{k+1}$, and we have shown that

$$\begin{aligned}
\Pr\{\text{bl}\} &- 1 + \frac{r}{s} \\
&\leqq \frac{(k+1)p_{k+1}}{\lambda s} \\
&\leqq \frac{(k+1)b_{k+1}}{\lambda \sum_{j=0}^{k} b_j \alpha_j} \\
&\leqq a \frac{\frac{a^k}{k!}}{\lambda \sum_{j=0}^{k} \frac{a^j}{j!} \binom{T-2j}{2}}, \qquad \left(\text{with } a = \frac{\lambda T^2}{2}\right) \\
&\leqq \frac{aE(k,a)}{\lambda \binom{T}{2} + \lambda a k - \lambda a(2T - 2a^2 + 3a + k)[1 - E(k,a)]}.
\end{aligned}$$

The last bound goes to $E(k,a)$ as $\lambda \to 0$, $T \to \infty$, with $2a = \lambda T^2$. In this limit, then, if $1 - r/s$ underestimates Pr{bl} at all, it does so by at most $E(k,a)$.

Another result of the same character is

Theorem 11: With $a = \frac{1}{2}\lambda T^2$ *and* $k + 2 \leqq w$,

$$Pr\{bl\} \leqq 1 - \frac{r}{s} + \frac{k+2}{a} \frac{E(k+2,a)}{1 - E(k+2,a)}.$$

Proof:

$$\begin{aligned}
1 - \Pr\{\text{bl}\} &= \frac{1}{\lambda} \frac{\sum_{j=1}^{w} jp_j}{\sum_{j=0}^{w} \alpha_j p_j} \geqq \frac{1}{\lambda} \frac{\sum_{j=1}^{k+1} jp_j}{\sum_{j=0}^{k+1} \alpha_j p_j} \\
&= \frac{r}{s + \alpha_{k+1} p_{k+1}},
\end{aligned}$$

the inequality coming from the basic lemma, and the second identity

from formula (6) and the definitions (10) and (11) of r and s respectively. Writing the last term on the right as

$$\frac{r}{s} \cdot \frac{s}{s + \alpha_{k+1}p_{k+1}} = \frac{r}{s} \cdot \left(1 - \frac{\alpha_{k+1}p_{k+1}}{s + \alpha_{k+1}p_{k+1}}\right)$$

we find, since $r/s < 1$,

$$1 - \Pr\{\text{bl}\} - \frac{r}{s} \geqq - \frac{\alpha_{k+1}p_{k+1}}{\sum_{j=0}^{k+1} \alpha_j p_j}.$$

The basic lemma gives, using $jb_j = \lambda\alpha_{j-1}b_{j-1}$,

$$\frac{\alpha_{k+1}p_{k+1}}{\sum_{j=0}^{k+1} \alpha_j p_j} \leqq \frac{\alpha_{k+1}b_{k+1}}{\sum_{j=0}^{k+1} \alpha_j b_j} = \frac{(k+2)b_{k+2}}{\sum_{j=1}^{k+2} jb_j}$$

$$\leqq (k+2) \frac{\frac{a^{k+2}}{(k+2)!}}{\sum_{j=1}^{k+2} j \frac{a^j}{j!}} = \frac{k+2}{a} \frac{E(k+2,a)}{1 - E(k+2,a)}.$$

Theorem 12: Let K be a set of integers j all satisfying $j > \lambda\alpha_j[1 - Pr\{bl\}]$. Then omission of all the states in $\bigcup_{j \in K} L_j$ in the calculation of $Pr\{bl\}$ as defined by (2) results in an overestimate.

Proof: Let $\xi = (\xi_0, \cdots, \xi_w)$ be a $(w+1)$ dimensional vector variable taking values in the positive orthant, and consider the function $V(\xi)$ defined by

$$V(\xi) = \frac{1}{\lambda} \frac{\sum_{j=1}^{w} j\xi_j}{\sum_{j=0}^{w} \alpha_j \xi_j}.$$

It is apparent that if $\xi = (p_0, \cdots, p_w) = p$ = distribution of the number of calls in progress, then

$$V(p) = 1 - \Pr\{\text{bl}\}.$$

Now,

$$\frac{\partial V}{\partial \xi_j} = \frac{1}{\lambda} \frac{j - \lambda\alpha_j V(\xi)}{\sum_{j=0}^{w} \alpha_i \xi_j}.$$

Hence, $V(\xi) \leqq V(p)$ and $j \in K$ imply

$$\frac{\partial V}{\partial \xi_j} > 0.$$

Consider a path of integration Γ along which

$$\xi_j = p_j \qquad j \notin K$$

and which runs from the point ξ_s with coordinates

$$\begin{cases} 0, & \text{for } j \in K \\ p_j & \text{for } j \notin K \end{cases}$$

to the point $\xi = p$ in such a way that $d\xi_j/ds > 0$ for $j \in K$ along Γ. $V(\xi_s)$ is the approximation to $1 - \Pr\{\text{bl}\}$ resulting from omitting from (17) states having j calls in progress for $j \in K$.

It is apparent that there is a segment of Γ in the neighborhood of p on which $V(\xi) \leqq V(p)$. Since $V(\cdot)$ is continuous the set $A = \{\xi: V(\xi) \leqq V(p)\}$ is closed. If Γ first intersects ∂A at some point $q \neq p$ we have

$$V(p) = V(q) = V(p) - \int_q^p \sum_{j \in K} \frac{\partial V}{\partial \xi_j} \frac{d\xi_j}{ds}\, ds$$

which is impossible since the integral does not vanish. Thus

$$V(\xi_s) \leqq V(p).$$

It is easy to see that the condition $j > \lambda\alpha_j[1 - \Pr\{\text{bl}\}]$ in Theorem 12 occurs for relatively low values of j. For it is enough that

$$j > \frac{\lambda T^2}{2}[1 - \Pr\{\text{bl}\}] = \frac{m}{(1-q)^2 - T^{-1}(1-q) + 4\sigma^2 T^{-2}},$$

and thus it suffices that

$$j > \frac{m}{(1-q)^2 + T^{-1}(1-q)}.$$

The second term in the denominator is negligible for all but uninterestingly small values of T, so roughly j can be any integer larger than

$$\frac{m}{(1-q)^2}.$$

With $q = 0.1$ erlang, a representative value, the condition is approximately $j > 1.22m$.

The method used in Theorem 12 also proves

Theorem 13: Let X be a set of states such that $x \in X$ implies

$$\frac{s(x)}{\alpha_x} > 1 - Pr\{bl\}.$$

Then the approximation

$$Pr\{bl\} \approx 1 - \frac{\sum_{x \epsilon S-X} s(x)p_x}{\sum_{x \epsilon S-X} \alpha_x p_x}$$

is an overestimate.

XV. INEQUALITY FOR PROBABILITY OF BLOCKING

Last, yet we hope not least, we give a basic inequality for the probability of blocking itself. The result to be given clearly shows how *combinatorial knowledge* about the connecting network of interest (in this case information about how fast the number of blocked pairs goes up with the number of calls in progress) can be used to give an upper bound on the loss.

Theorem 14: Let $\beta_x \leqq f_{|x|}$ for nondecreasing $f(\cdot)$. Then

$$Pr\{bl\} \leqq \frac{\sum_{j=0}^{w} f_j b_j}{\sum_{j=0}^{w} \alpha_j b_j} \leqq \frac{\sum_{j=0}^{w} f_j \frac{a^j}{j!}}{\sum_{j=0}^{w} \alpha_j \frac{a^j}{j!}}.$$

Proof:

$$\begin{aligned} \Pr\{\text{bl}\} &= \frac{\sum_{x \epsilon S} \beta_x p_x}{\sum_{x \epsilon S} \alpha_x p_x} \leqq \frac{\sum_{j=0}^{w} f_j p_j}{\sum_{j=0}^{w} \alpha_j p_j} \\ &\leqq \frac{\sum_{j=0}^{w} f_j b_j}{\sum_{j=0}^{w} \alpha_j b_j} \\ &\leqq \frac{\sum_{j=0}^{w} f_j \frac{a^j}{j!}}{\sum_{j=0}^{w} \alpha_j \frac{a^j}{j!}}, \end{aligned}$$

with $a = \frac{1}{2}\lambda T^2$. The first inequality follows from the hypothesis and the assumed one- or two-sided nature of the network; the ensuing two inequalities follow from the basic lemma.

The foregoing theorem makes it plain that much is to be learned about congestion in a connecting network from a study of the rate at which the special function $\{\beta_x, x \in S\}$ changes with $|x|$. The search for bounds of the form

$$\beta_x \leqq f_{|x|}, \qquad x \in S$$

(with f_j increasing) for various kinds or classes of connecting networks now becomes one of the next most important problems in the endeavor to bring, by purely analytical methods, A. K. Erlang's dynamical theory of telephone traffic to belated but final fruition. This problem is beyond the scope of this paper; some elementary phases of it are considered in a later paper.[2]

XVI. ACKNOWLEDGMENT

The author is indebted to W. S. Hayward, Jr., for a careful reading of the draft, and for pointing out an error in the proof of an earlier version of Theorem 5.

REFERENCES

1. Beneš, V. E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.
2. Beneš, V. E., Some Bounds for Loss in Crossbar Networks, to be published.
3. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, B.S.T.J., *42*, 1963, pp. 2795–2838.
4. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., *34*, 1955, pp. 1287–1315.
5. Grantges, R. F., and Sinowitz, N. R., NEASIM: A General-Purpose Computer Simulation Program for Load-Loss Analysis of Multistage Central Office Switching Networks, B.S.T.J., *43*, 1964, pp. 965–1004.
6. Beneš, V. E., Properties of Random Traffic in Nonblocking Telephone Connecting Networks, B.S.T.J., *44*, 1965, pp. 509–525.

An 80-Megabit 15-Watt Transistor Pulse Amplifier

By L. U. KIBLER

(Manuscript received June 28, 1965)

A transistor pulse amplifier delivering a 1.1-ampere, 11-nanosecond pulse at an 80-mc rate into a 15-ohm load has been designed. This amplifier was developed for use with an optical modulator. This paper describes the performance and gives the design and construction details of the amplifier.

I. INTRODUCTION

High-power broadband amplifiers operating into impedances of less than 50 ohms have not been previously available. With the advent of optical modulators a need has arisen for such amplifiers. In particular an optical modulator[1] designed for experiments on optical communication systems requires a one-ampere peak signal to produce a one-radian phase shift. This modulator consists of a one meter long strip line partially loaded with a KDP (potassium dihydrogen phosphate) dielectric. The line impedance is 15 ohms. A PCM system was proposed requiring an 80-mc pulse rate and 10 to 12 nanosecond (ns) raised cosine pulses. The use of raised cosine pulses was dictated by the low-level vacuum tube pulse generator in the PCM system. A suitable driver for this optical modulator must deliver a one-ampere peak pulse into the 15-ohm line over a band extending from near dc to at least 100 mc.

The discussion of the amplifier that realizes these requirements is divided into several parts. Section II is a general description of the completed amplifier, including the configuration used and the results obtained. Section III gives the performance that was obtained, including photographs of the response to both square wave pulses and raised cosine pulses at rates up to 80 mc. The design considerations are taken up in detail in Section IV, and the mechanical construction is described in Section V. A discussion of the results is contained in Section VI.

II. GENERAL DESCRIPTION

The complete amplifier consists of a power amplifier and a preamplifier. The power amplifier consists of a pair of VHF silicon power transistors coupled at the input and output with broadband transformer hybrids. This amplifier has a current gain of 3.2 and delivers a 1.02-ampere pulse 11-ns wide at the base into an output impedance of 15 ohms.

A three-stage transformer-coupled transistor preamplifier is used to drive the power amplifier. The preamplifier has a current gain of 5.3 and has a bandwidth larger than that of the power amplifier so that it has negligible effect on the pulse response. An input pulse of 60 ma into 50 ohms is required to achieve the full peak output power of 15 watts.

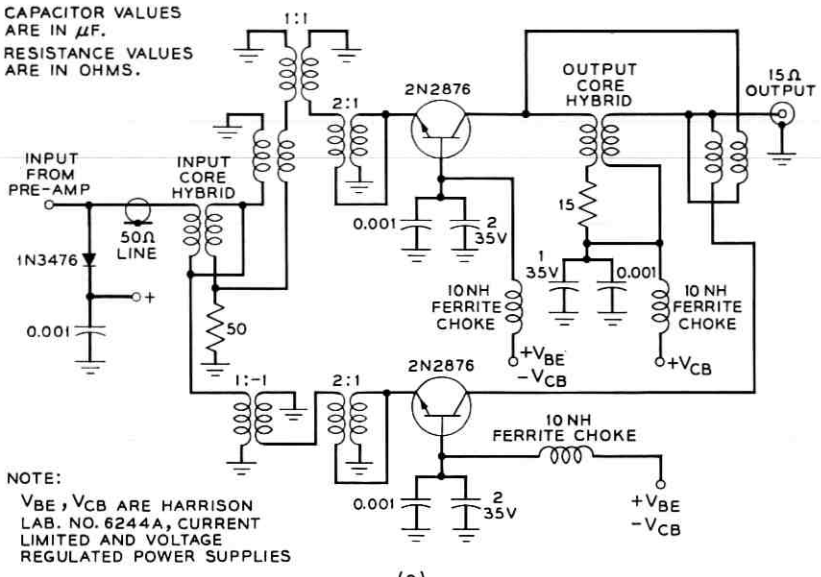
These amplifiers use the common-base configuration with current gain obtained by transformer coupling. The pulse response of the combined amplifier is limited by the parameters of the transistors used in the output stage. Under these circumstances, the common-base configuration is the only one which can be used to deliver the required power with the desired pulse response.

The complete circuit diagram is shown in Figs. 1(a) and 1(b). The transistors used in the output stage are RCA 2N2876; those in the preamplifier the RCA TA2307 (2N3375). The transformers and the hybrids are of the type described by Ruthroff.[2]

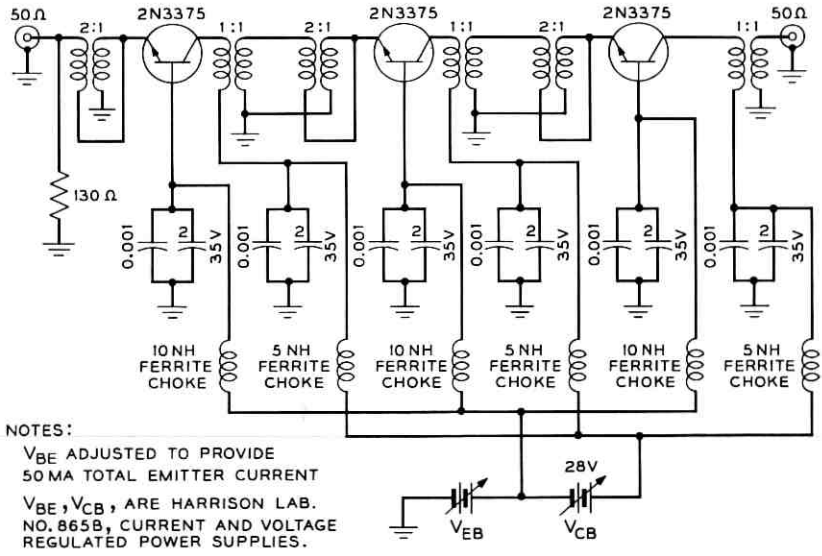
The total dc power required is 13 watts: 0.46 ampere at 28 volts. The power amplifier occupies a space of approximately 3-inches wide by 1-inch deep by $1\frac{3}{4}$-inches high. The preamplifier occupies a space of 1 by 1 by 4 inches. The total volume of the complete amplifier excluding the heat sinks is $9\frac{1}{4}$-cubic inches. A photograph of the complete amplifier is shown as Figs. 2(a), (b) and (c).

III. PEFORMANCE

The amplifier was tested with 8-ns raised cosine pulses recurring to rates varying from 10 mc to 80 mc and with 50-ns square wave pulses at 100 cps rate. The raised cosine pulses were generated by tube pulse generators supplied by A. F. Dietrich of Bell Telephone Laboratories. The pulse amplitude was variable up to a maximum of 15 volts into 50 ohms. All waveforms were observed on a Hewlett-Packard 187A sampling oscilloscope. The arrangement of the test circuit is shown in Fig. 3. A coaxial probe for one channel of the sampling scope was placed in the input line with a 3-db General Radio coaxial pad as iso-

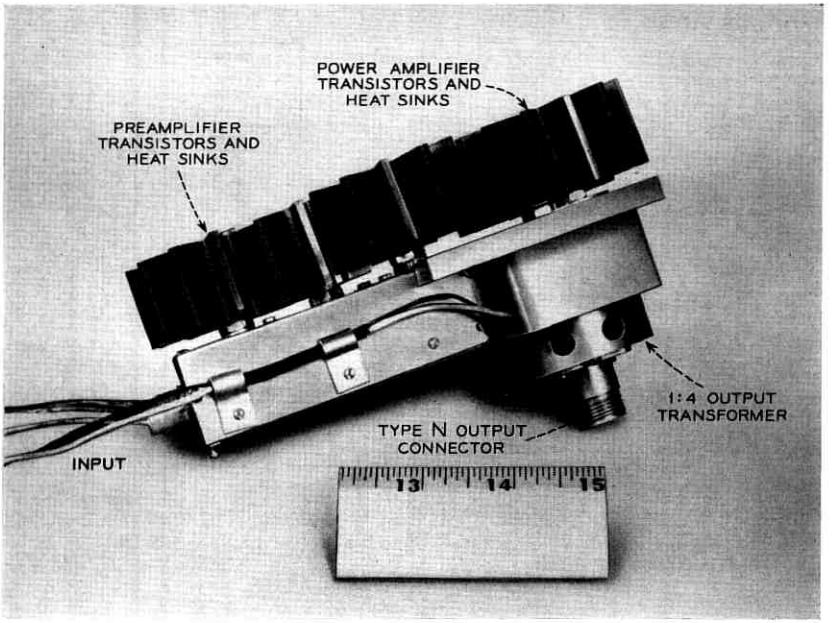


(a)

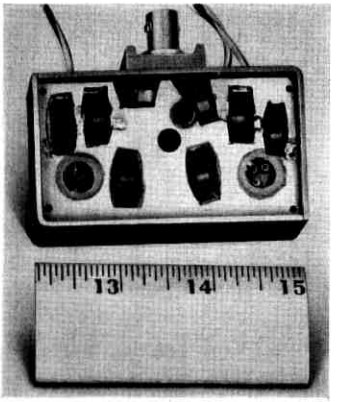


(b)

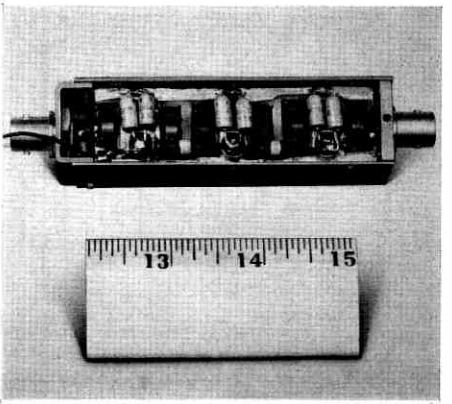
Fig. 1 — (a) Power amplifier schematic diagram; (b) preamplifier schematic diagram.



(a)



(b)



(c)

Fig. 2 — (a) Side view of complete amplifier; (b) top view of interior of power amplifier; (c) bottom view of interior of preamplifier.

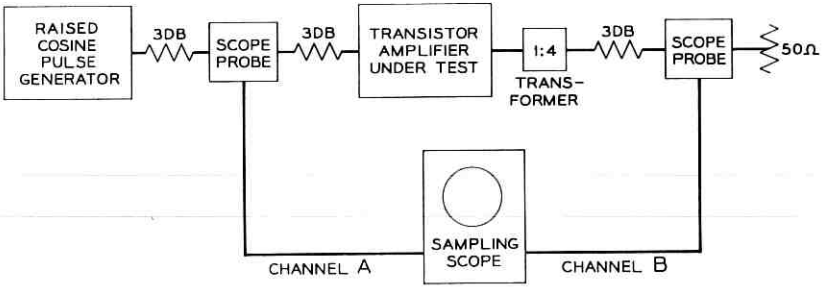


Fig. 3 — Test circuit for transistor amplifier.

lation between the pulse generator and the amplifier input. The coaxial probe of the second scope channel was placed in the amplifier output between a 3-db isolation pad and a 50-ohm, 10-watt termination. Coupling of the 15-ohm output of the amplifier to the 50-ohm test circuit was accomplished with a 4:1 core transformer and a length of tapered strip line. The 15-ohm side of this transformer was connected to the output hybrid of the amplifier through a 0.01-μf capacitor to provide dc isolation. The collector-base and base-emitter circuits of the power amplifier were supplied with dc bias from separate Harrison Lab 36-volt, 3-ampere current- and voltage-regulated power supplies. The same circuits in the preamplifier were separately supplied by Harrison Lab 40-volt, 0.5-ampere power supplies. This division of power supplies was used to allow a variety of bias conditions for test purposes.

The power amplifier was tested separately with an 8-ns pulse at a 10-mc rate. The pulse in the input line and the output pulse are shown in Fig. 4. The collector was biased with +28 volts and the zero pulse emitter bias was 20 ma. The output pulse amplitude corresponds to a 0.9-ampere pulse in 15 ohms. There is a large reflection evident in the input line as seen in Figs. 4(a) and (b). This is a consequence of designing the transformers to match the 50-ohm input line to the transistor input impedance of 6 ohms at the peak pulse amplitude. At lower amplitudes, the transistor input impedance is approximately $2\frac{1}{2}$ times greater and varies both with drive and frequency. The frequency variation of the input impedance is less at high drive than at low drive. The reflection in the input produces the small pulse following the main output pulse shown in Fig. 4(a).

To suppress the reflected pulse from the power amplifier, and prevent further reflection from the preamplifier with the attendant

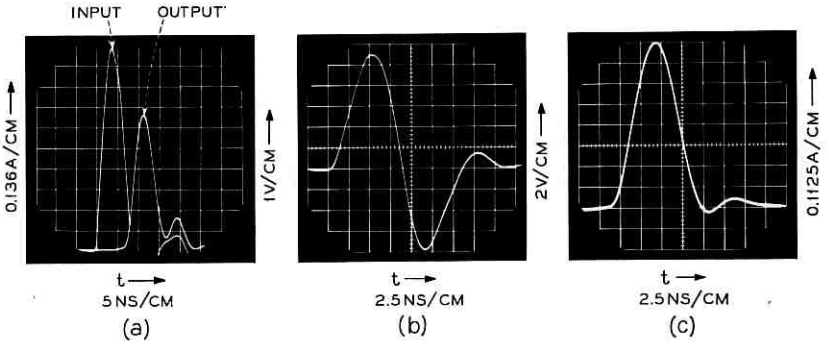


Fig. 4 — Power amplifier pulse response: (a) power amplifier input and output without compensating diode; (b) expanded view of amplifier input without diode; (c) amplifier output with diode compensation.

degradation in the output of the power amplifier, a Western Electric 1N3476 "pin head" diode was placed across the short section of 50-ohm line in the power amplifier input (Fig. 1). The diode anode was connected to the center conductor. The cathode was connected to a positive bias supply through a 0.001-μf feed-through capacitor. This diode has a 1-pf capacitance and will dissipate 200 mw. Thus the negative-going input pulse to the power amplifier will cause no diode conduction. The 1-pf capacitance will cause negligible loading. The reflected positive pulse will cause the diode to conduct. By adjusting the diode bias voltage, the amount of conduction and hence the impedance of the diode can be controlled. This arrangement damps out the reflected pulse. The effect of this damping is evident in the output pulse of Fig. 4(c). The pulse following the main pulse in Fig. 4(a) has been eliminated.

Fig. 5 shows a comparison between the measured wave shape and the calculated wave shape for this amplifier. The plot of the measured response was normalized so that the peak amplitudes of the measured and calculated responses were equal. The beginning of the output pulse was taken to coincide with the calculated pulse (thus eliminating the amplifier delay) so that a comparison of the pulse shapes could be readily seen.

The response of the power amplifier with the compensating diode to a 50-ns pulse with 0.5-ns rise time from the Spencer-Kennedy pulse generator is shown in Fig. 6. The input is shown in Fig. 4(a) with the amplifier output in Fig. 4(b). The rise time of the amplifier is approximately that shown in Fig. 5, and there is no indication of sag.

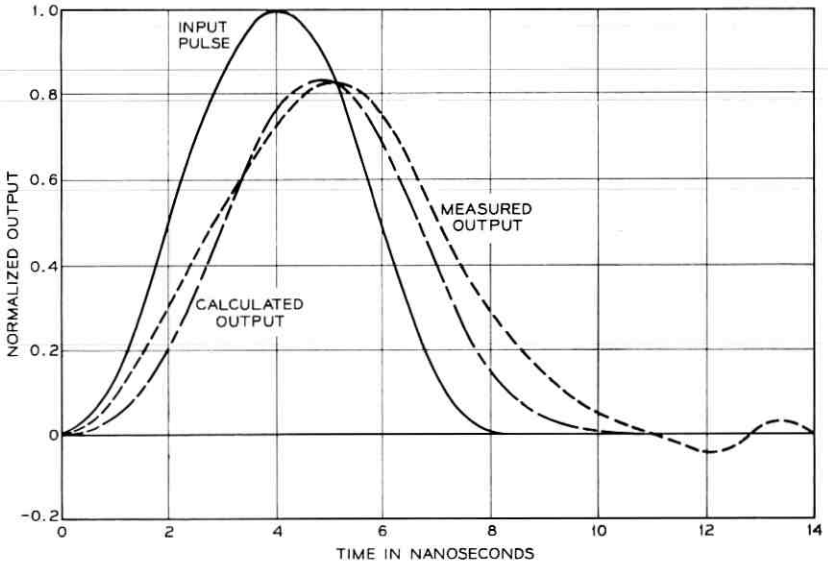


Fig. 5 — Comparison of calculated and measured pulse response of power amplifier.

The pulse response of the complete amplifier is shown in Fig. 7. Fig. 1(a) shows the amplifier output with 8-ns driving pulse at a 40-mc rate. The 35-ma input pulses (in 50 ohms) are 8-ns long. The 0.9-ampere output pulses measured in 15 ohms are approximately 11-ns long. Fig. 7(b) shows an expanded scale of Fig. 7(a). The output base line ripple is 0.076 ampere for a 0.008-ampere ripple in the input pulses.

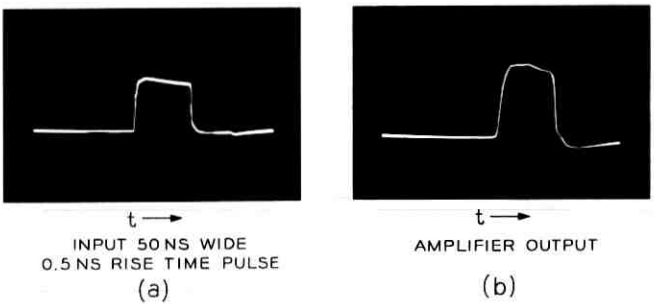


Fig. 6 — Amplifier step response.

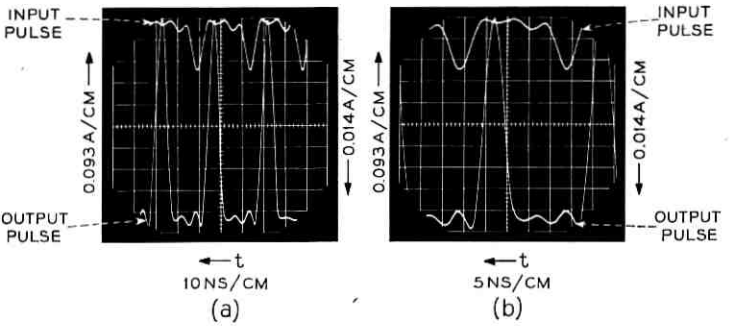


Fig. 7 — Amplifier pulse response, 40-mc pulse rate.

Fig. 8 shows the response to a group of pulses at an 80-mc rate. The output of the amplifier to a group of 4 pulses occurring at an 80-mc rate (12.5-ns pulses) is shown in Fig. 8(a). It was not possible to obtain sufficient output from the pulse generator to drive the amplifier with all 8 pulses. Fig. 8(b) shows an expanded view of the 4 pulses.

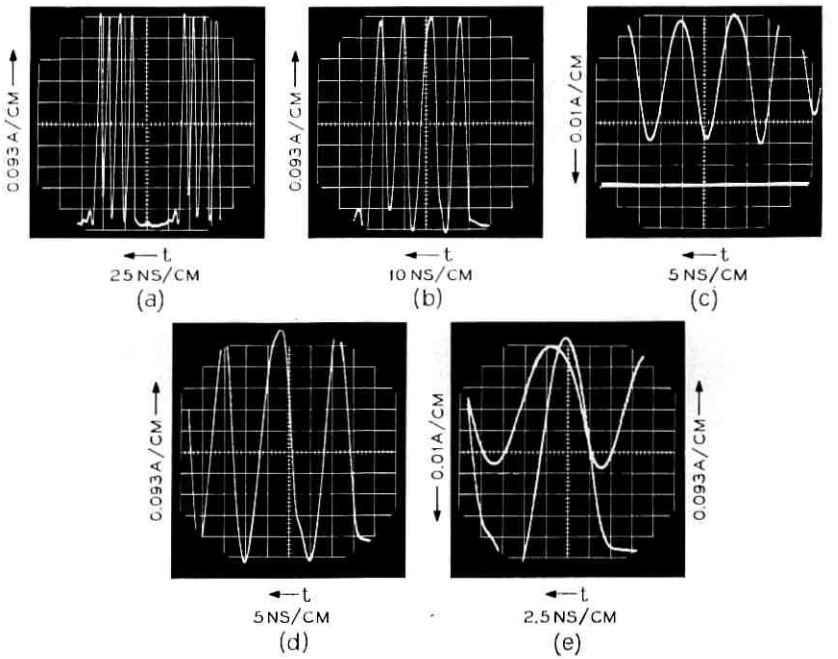


Fig. 8 — Amplifier pulse response, 80-mc rate.

These 0.93-ampere pulses measured in 15 ohms are approximately 12 ns wide. The 60-ma driving pulses and the 1.02-ampere output pulses measured in 15 ohms are shown in Fig. 8(c) and 8(d), respectively. An expanded view of these pulses is shown in Fig. 8(e). Here a 56-ma driving pulse produces a 0.93-ampere output pulse about 12-ns long. This corresponds to a total amplifier gain of 19.2 db.

IV. DESIGN CONSIDERATIONS

The design of the amplifier with the required specifications was divided into two stages. Since the power amplifier has the most difficult power and bandwidth requirements, it was considered first. With the power amplifier design completed, a preamplifier was designed so that the overall amplifier would have a gain of 20 to 23 db.

A number of low-level broadband transistor amplifiers have been built. These have been of the distributed type[3,4] and of the staggered tuned[5] variety. Both types have delivered only a few milliwatts of power into 50 ohms. These amplifiers have been operated in a linear region of the transistor characteristic. In addition to the linear range of operation, however, transistors can be operated in a pulsed mode. In this type of operation, the transistor is normally off, drawing no collector current until a driving pulse turns the transistor on.[6] In many applications, the input pulse is sufficient to drive the transistor into the saturation region where the collector current is dependent on the dc supply voltage and the load impedance only. This type of operation decreases the turn-on time of the output pulse, but increases the pulse length due to charge storage in the collector-base region. Such transistor switching amplifiers are usually operated in the common-emitter configuration with a passive network in the base lead to compensate for the fall-off in gain above the β cutoff frequency. Current gain can also be realized in the common-base configuration if suitable broadband transformers are available to match the transistor input and output impedances to the source and load.

Gartner[6] points out that the common-base power amplifier produces the same output power as the common-emitter amplifier but with less distortion. However, the power gain of the common-base amplifier is less than that of the common-emitter amplifier. Since this power amplifier must deliver a large peak pulse power with a minimum distortion of the pulse shape, and since sufficient over-all gain can be obtained in the lower-level preamplifier, the common-base configuration using broadband coupling transformers is best suited for this application.

The required broadband transformers have been developed by Ruthroff.[2] These transformers consist of a bifilar winding about a toroidal core of ferrite material. Hybrids, 4:1 impedance transformers, and 1:1 reversing transformers can all be assembled by combinations of the basic core transformer with appropriate interconnection of windings. Schematic representations of these three coupling devices are shown in Fig. 9.

Of the commercially available VHF power transistors, the RCA 2N2876 and TA2307 provided the best compromise in power dissipation, α-cutoff frequency and output capacitance for broadband high-power operation. These NPN transistors are of the triple-diffused silicon planar construction using a novel overlay construction. Either transistor will handle the required 3 watts of average pulse power. The use of a single common-base transistor with input and output transformers to achieve current gain would result in peak voltages that are close to or exceed the breakdown conditions of the transistors. To prevent this possibility and to allow cooler operation, two transistors are used in a hybrid coupled power amplifier.

The operation of the hybrid coupled amplifier can be understood with reference to the schematic diagram in Fig. 1(a). We can consider a unit amplitude current pulse incident upon the 50-ohm terminal of the input hybrid. The current pulse out at the conjugate terminals of the hybrid will be of unit amplitude at a 25-ohm impedance level. The polarity of the pulse at the conjugate arms will be opposite. In order to convert the positive pulse in one conjugate arm to the required negative pulse, to drive the transistor, a 1:1 reversing transformer is used. Since we want both transistor input circuits to have the same time delay, a 1:1 nonreversing transformer was inserted in the opposite conjugate arm. The negative unit amplitude current pulses are coupled to the emitter of each transistor through 4:1 impedance transformers. This arrangement provides a 2-unit amplitude negative pulse at a $6\frac{1}{4}$-ohm impedance level to each emitter-base circuit. The collector current of each transistor will be a 1.8-unit amplitude pulse since the α's of these power transistors are approximately 0.9. The output of the two transistors are combined in the output hybrid to produce a 3.6-unit amplitude negative pulse into the 15-ohm load. With a 15-ohm load, the collector load impedance of each transistor is 30 ohms. The current gain of the amplifier from the 50-ohm input to the 15-ohm output is 3.6 (a power gain of 5.9 db). Thus a one-ampere peak pulse in the 15-ohm load requires a 0.28-ampere driving pulse in the 50-ohm input line.

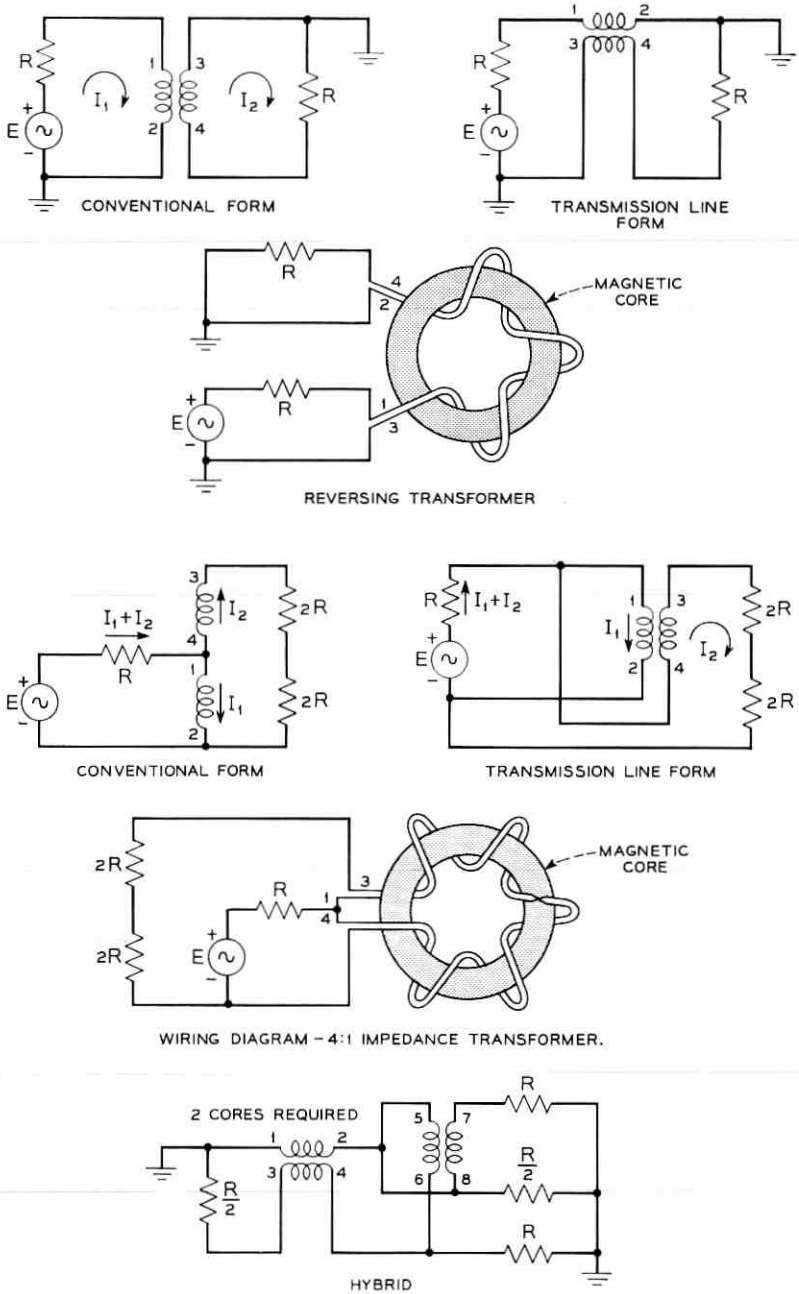


Fig. 9 — Typical core transformers used in amplifier.

The theoretical response of the hybrid coupled amplifier to a raised cosine pulse was determined by a combination of experimental and analytic investigation of the component parts. Ruthroff[2] has shown that the core transformers have a flat response from less than 1 mc to greater than 400 mc at low current levels. To determine what saturation effects might be present at peak currents of one ampere and greater, we tested a reversing transformer and a pair of 4:1 transformers placed back-to-back at various current levels from 0.1 ampere to 3 amperes. Transformers, bifilar wound with #20, #22, and #24 Formex wire, were tested with 0.5-ns rise time, 50-ns long pulses of varying amplitudes generated by a Spencer-Kennedy pulse generator operated at a 100-cps rate. Representative pictures taken from the Edgerton Germeshausen & Grier traveling-wave oscilloscope of the input and output pulses are shown in Fig. 10. It is evident from these pictures that there is no significant change in the pulse shape in going from a 0.1-ampere pulse to a 3-ampere pulse.

The amplitude and phase of the α of the 2N2876 transistor was measured as a function of frequency for two collector currents, 0.25 ampere and 0.5 ampere at 28 volts. The rather novel method of measurement is described in the Appendix A. Representative response curves are shown in Fig. 11. At 0.25 ampere, the α-cutoff frequency is approximately 200 mc while at 0.5 ampere it is 170 mc. These measurements confirm the manufacturer's nominal f_α of 200 mc.

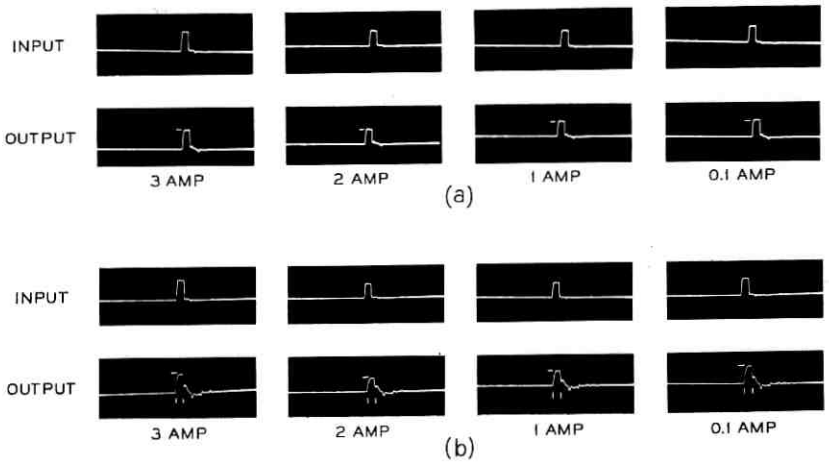


Fig. 10 — Top: reversing transformer, #20 wire; bottom: two 4:1 transformers back-to-back.

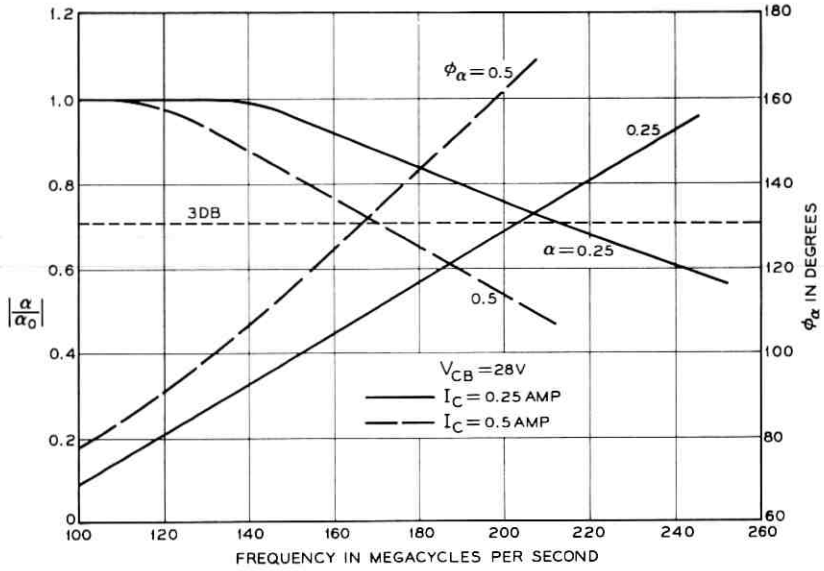


Fig. 11 — 2N2876 transistor measured magnitude and phase of short circuit current gain, alpha, as a function of frequency.

From these measurements it is evident that the response of the amplifier to a raised cosine pulse will be determined by the α of the transistor rather than the connecting circuitry. Gartner[6] has a general analysis of the step response of the common-base transistor operated in the nonsaturating switching mode. The response to a raised cosine pulse can be obtained by using a step approximation to the convolution integral.

The effect of the collector circuit RC cutoff frequency has been analyzed by Easley.[7] For the common-base configuration where $\omega_\alpha RC_c$ is less than one, but of the order of one, the effect of RC_c is to alter the shape of the waveform and slow the response for times small compared to $1/\omega_\alpha$, i.e., in this case less than 1 ns. The details of this analysis are contained in Appendix B. Since some 400 calculations were required for each point of the response curve, the solution was programmed on an IBM 7090 computer. The response of a transistor with an α of 0.9, an ω_α of 200 mc, a collector capacity of 20 pf, and a load of 30 ohms to a raised cosine pulse 8 ns long is plotted in Fig. 12. This figure shows that the output pulse is broadened to 11 ns and decreased in amplitude by 17 per cent. Applying these results to the hybrid amplifier of Fig. 1, we see that the amplifier using the

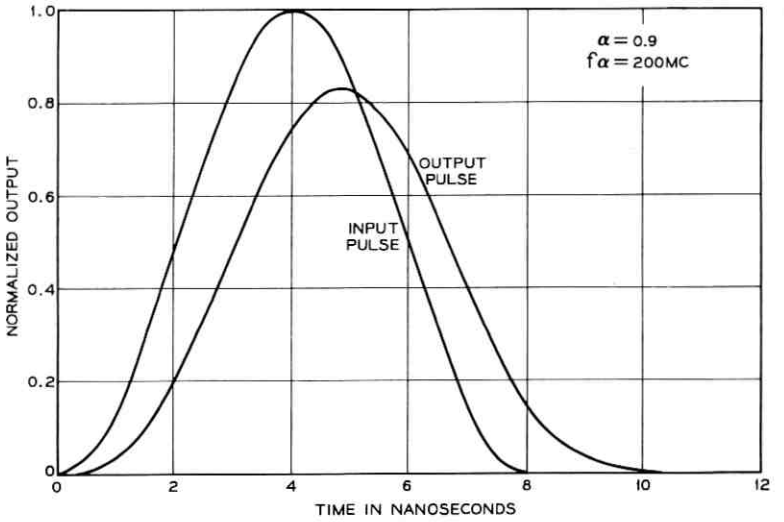


Fig. 12 — Calculated pulse response of 2N2876 transistor.

2N2876 transistor will have a current gain of 3.32 or 5.2 db operating from 50 ohms to 15 ohms. Thus for a 1-ampere output pulse we require a 0.3-ampere driving pulse. A preamplifier with 16-db gain will require a 48-ma peak current pulse to drive the power amplifier to full output.

The schematic diagram of a three-transistor preamplifier that will supply the required gain is shown in Fig. 1(b). This amplifier uses the common-base configuration with the ferrite core transformers to provide the current gain. To maintain a 200-mc bandwidth in the presence of the bandwidth reduction that results from cascading identical amplifiers, the TA2307 (2N3375) transistor with its α-cutoff frequency of 500 mc was used. Since the preamplifier operates at lower power levels than the power amplifier, the 11.5-watt dissipation of the TA2307 is adequate. As in the power amplifier, the transistors in the preamplifier are operated in a nonsaturating switching mode.

The pulse response of a single 2N3375 transistor was calculated in the same manner as above (see Appendix B). The width of a unit amplitude pulse 8 ns wide is increased to 8.3 ns and the amplitude is decreased by 12 per cent, yielding a current gain of 0.88 per transistor.

The operation of the preamplifier between 50-ohm impedances can be understood with reference to the schematic diagram, Fig. 1(b). A negative unit amplitude pulse incident upon the input of the first 4:1

impedance transformer produces a 2-unit pulse to the emitter of the first transistor at a 12.5-ohm impedance level. The real part of the input impedance of these transistors ranges from 12 to 20 ohms, depending on pulse amplitude. The collector current, using the effective α of 0.88, is then 1.76 units. A 1:1 transformer is used to isolate the collector dc bias from the 4:1 input transformer of the next stage. Since each succeeding stage is identical, the collector current of the output of the preamplifier is 5.43 units. The current gain is 5.43, or a power gain of 14.7 db. For the 0.3-ampere output pulse, a driving pulse of 55 ma in 50 ohms is required. This corresponds to a peak power of 150 mw per pulse which is available from most pulse generators. Since the transistors and the core transformers have bandwidths of at least 400 mc, the bandwidth of the amplifier will be limited to less than 400 mc by the effect of cascading. The reduction in bandwidth due to transformer coupling will be less than the reduction caused by cascading RC coupled amplifiers. Three RC coupled stages using these transistors would have a bandwidth of 204 mc.

V. MECHANICAL CONSTRUCTION

The basic amplifier construction uses a quasi-stripline to interconnect the core transformers and transistors. The input and output circuits are placed on opposite sides of a center board in nonoverlapping areas. Ground plane boards are placed on top and bottom of the common board. This results in a 3-layer sandwich construction. The center board is $\frac{1}{8}$-inch thick glass-loaded Teflon with 2-oz copper on each side. The ground plane boards are $\frac{1}{16}$-inch thick glass-loaded teflon with 2-oz copper on each side. Transistor sockets were made by placing a spring contact for each pin on the transistor case into holes drilled in the center board. These contacts were made using the center conductor of BNC female coaxial connectors. After being placed into the center board, these contacts were soldered to the copper. Matching rectangular holes were cut in each board to accommodate the core transformers. The copper on the board surfaces was cut away so as to leave stripline connections between the cores and transistors. The boards were separated by placing $\frac{3}{16}$-inch thick brass spacers around the periphery of the boards. Where this spacing was not adequate for the proper stripline impedance, brass pieces were attached to the center board to decrease the spacing in a local region. The terminations for the input and output hybrids, 50 ohms and 15 ohms, respectively, are 1 watt 1 per cent metalized film microwave rod resistors made by

Film Ohm Corp. The base by-pass capacitors consist of a 0.001-μf feed-through capacitor mounted in the ground plane with the center connected to the base terminal of the center board. Two 2-μf miniature tantalum capacitors were used in parallel with the feed-through capacitor. The collector dc connection is by-passed with a 0.01-μf postage stamp type capacitor in parallel with a 2-μf miniature tantalum capacitor. Collector bias is applied between collector-base terminals. Emitter bias is applied between base terminals and ground.

The layout of the three circuit boards for the power amplifier is shown in Fig. 13. Both sides of each board are shown. The layout of the 3-transistor circuit board of the preamplifier is shown in Fig. 14. The positions of the parts and of the sections of stripline are evident upon examination of these figures.

The outside dimensions of the circuit boards were dictated by the space available inside the KDP modulator structure and by a need to keep all leads as short as possible. To keep the connection between the power amplifier and the KDP stripline to a minimum, the power amplifier had to be placed in a space 1-inch deep, 3-inches wide, and $1\frac{5}{8}$-inches high. A mounting box of these outside dimensions was made of $\frac{1}{16}$-inch brass and the circuit board sized so as to fit inside this box. The transistor heat sinks — 2-inch diameter, 1-inch high, 12-fin aluminum cylinders — were extended out through the end plate of the KDP mounting box. The heat sinks are mounted on the transistor package stud and are kept isolated from ground. Since the collector of the 2N2876 and TA2307 are isolated from the case (collector-to-case capacitance 6 pf), the mounting of the heat sinks in this manner reduced the collector circuit capacitance to a value close to the transistor collector capacitance of 20 pf for the 2N2876 and to 10 pf for the TA2307.

The cores in the output hybrid of the power amplifier are bifilar wound with 6 turns of #22 Formex wire. The transformer cores and the input hybrid cores are bifilar wound with 6 turns of #24 Formex wire. All the transformer cores in the preamplifier except the output transformer are bifilar wound with 6 turns of #24 Formex wire. The output transformer core is bifilar wound with 5 turns of #24 Formex wire.

The size of the preamplifier was also dictated by the dimension of the KDP modulator structure. A narrow 1-inch by 1-inch passage existed between the top of the power amplifier and the top wall of the modulator structure. The preamplifier was mounted in a $\frac{3}{64}$-inch thick brass box, 1 by 1 by 4 inches, mounted on top of the power

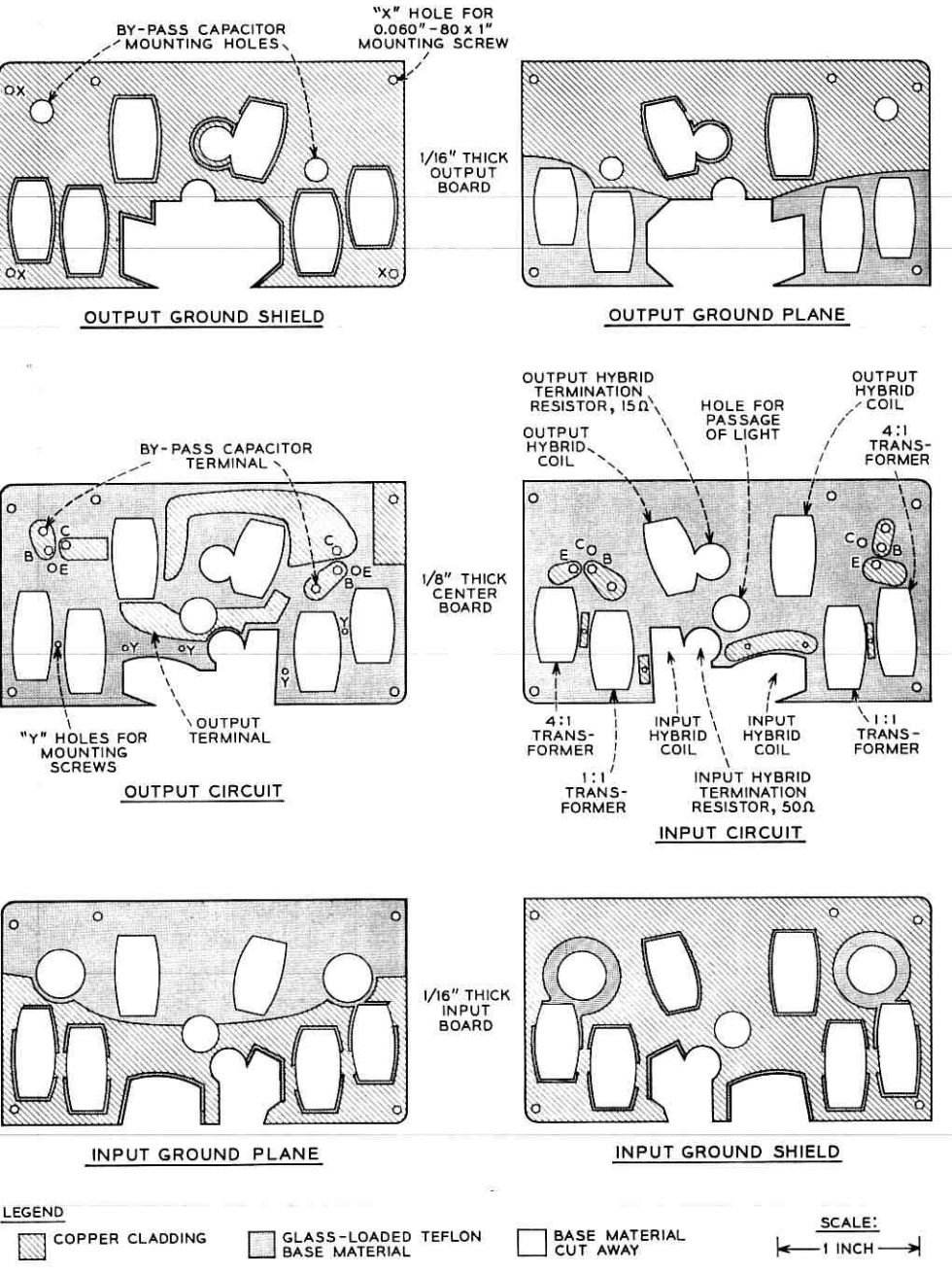


Fig. 13 — Power amplifier circuit boards.

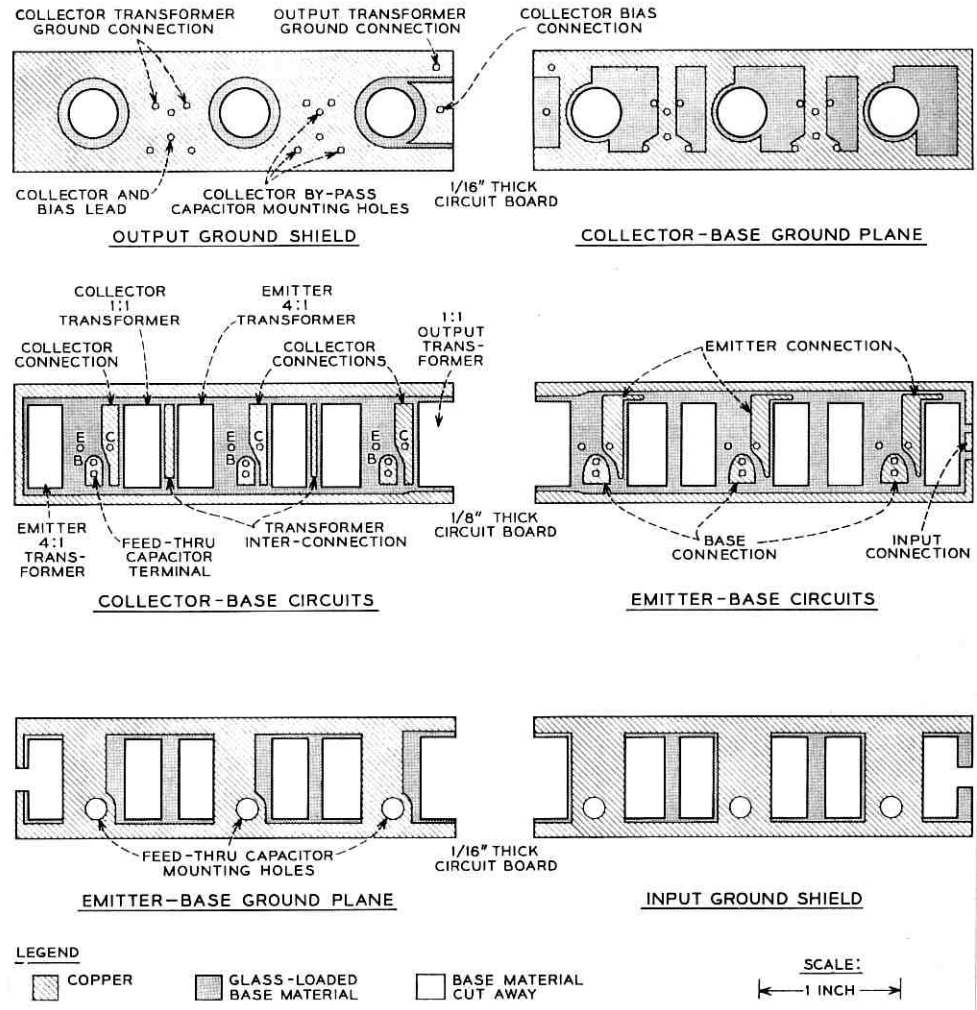


Fig. 14 — Preamplifier circuit boards.

amplifier and extended through the top of the modulator structure. Connection of the preamplifier with the coaxial output of the modulator source was made through a BNC coaxial connector. The circuit boards of the preamplifier were cut to dimensions to fit this box. The heat sinks for the preamplifier are of the same type as those of the power amplifier. Certain fins of each heat sink were cut away so that the heat

sinks of successive preamplifier transistors could interleave. The details of the complete amplifier can be seen in Fig. 2.

VI. CONCLUSIONS

This experimental transistor amplifier has demonstrated that broadband high-power transistor pulse amplifiers operating into impedances less than 50 ohm can be designed and built with existing commercial transistors. In the common-base configuration with broadband transformer coupling, the amplifier design depends on the transistor switching analysis and the RC product in the output circuit. In addition, the experimental work with these bifilar wound core transformers shows that these transformers can handle short rise time pulses of several amperes without saturation effects.

Several amplifiers of this type have been constructed. Their performance is equivalent to the amplifier reported in this paper. One of these amplifiers has been in daily use for six months with no degradation in performance.

These results also support Easley theory on the effect of ω_α RC_c products that are of order unity but less than unity. Easley in his paper[7] discusses this effect theoretically but only examines experimentally the effect of ω_α RC_c on the common-emitter and common-collector configurations. For these latter transistor configurations, the effect of ω_α RC_c is to lengthen the rise time by $(1 + \omega_\alpha RC_c)$. If this theory were valid for the common-base configuration, the output pulse width would be doubled.

It is evident from the experimental results obtained with this common-base amplifier that the pulse width is not broadened appreciably from that due to the transistor alpha. Since the ω_α RC product for this amplifier is 0.75, the results are in accord with the theory. Further support of Easley's theory is evident in the small distortion present in the measured output.

The results obtained with this amplifier justify the use of the common-base configuration when the highest frequency of operation is of the same order of magnitude as the f_α of the transistor. The use of a common-emitter configuration would result in the pulse-broadening noted above and would require additional circuits in the base or emitter leads to compensate for the frequency variation of the transistor beta. While in theory such compensation is possible, the addition of lumped circuit elements in high-frequency circuits introduces parasitics which would adversely effect the amplifier performance.

A problem of conditional stability is associated with the use of the common-base amplifier. The possibility of oscillation can be avoided by careful isolation of the input and output circuits and by the proper choice of the bias conditions and the load resistance. The isolation of the input and output of each transistor in the amplifier was achieved with the three-layer sandwich construction. The choice of the load resistance and the bias eliminated any tendency to oscillate.

The use of a hybrid-coupled transistor amplifier for the power amplifier increases the maximum power by 3 db for the same transistor rating without paying a penalty in the RC cutoff frequency in the individual transistor output circuit. It is evident that 2N transistors could be coupled by tandem hybrids to obtain a 3N db increase in output power without a significant bandwidth reduction over the single amplifier circuit. As an example, consider the circuit for a power amplifier delivering a 2-ampere pulse into a 15-ohm load. Four of the present hybrid coupled power amplifiers would have their inputs and output coupled by core hybrids. Since the input and output currents are additive, we require a 0.15-ampere pulse to produce a 2-ampere, 11-ns pulse at an 80-mc rate.

The success of these amplifiers is due in part to the ability of J. W. Batton to accurately construct and assemble the strip line connections.

APPENDIX A

Measurements of Transistor Alpha Cutoff Frequency

To determine the response of a transistor pulse amplifier in the common-base configuration, it is necessary to know the variation of the short circuit current gain alpha, with frequency. The manufacturer lists the f_T, a gain bandwidth product, of the 2N2876. Since this transistor is of a new design, the relation between f_T and f_α was not known.

A coaxial test jig was designed to measure both the magnitude and phase of alpha as a function of frequency. The drawing of the test jig is in Fig. 15. The diameters of the inner and outer conductors are the same as the General Radio 50-ohm air coaxial line. A phenolic transistor socket was mounted in a hole cut in the outer conductor. The base terminal was grounded directly to the inner wall of the outer conductor. The emitter terminal was connected to the center conductor through a parallel RL circuit with a cutoff of 140 mc. The collector terminal was connected directly to the center conductor with a short

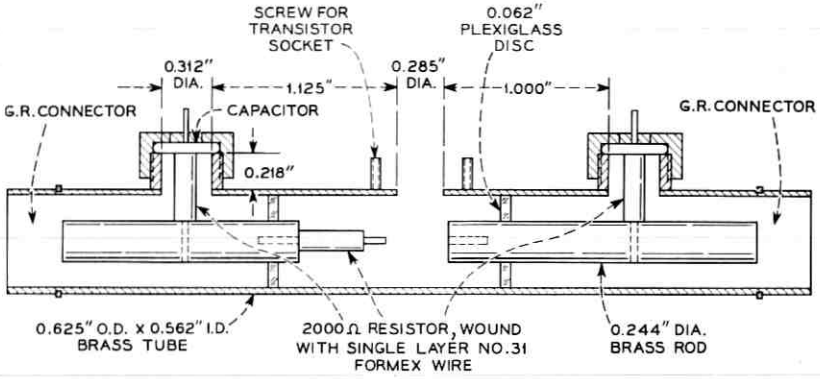


Fig. 15 — Transistor —test jig.

length of #22 wire. Bias connections to the center conductor were supplied by by-passed terminals in the outer wall of the test jig. Silver mica button capacitors, 0.001 μf, were soldered to the outer wall to provide RF by-passing. The bias terminal of these capacitors was connected to the center conductor through a parallel RL circuit with 140-mc cutoff frequency.

The arrangement of the test circuit is shown in Fig. 16. The emitter side of the test jig is connected through a 20-db pad to the coaxial probe of one channel of a Hewlett Packard 185/187A sampling oscilloscope. The coaxial probe was connected through a 3-db pad to a General Radio mixer rectifier. The low-frequency output of this mixer (dc to 30 mc) was connected to the sync terminal of the scope.

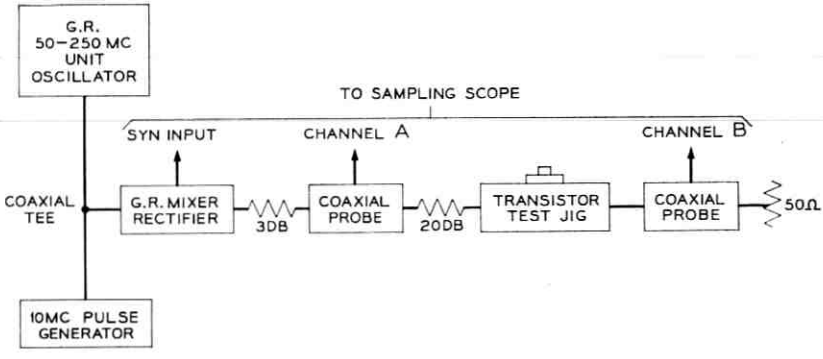


Fig. 16 — Test circuit for measurement of alpha.

The mixer input was fed from a 10-mc pulse generator and a General Radio 50- to 250-mc unit oscillator through a coaxial tee. The oscillator output was much greater than the pulse amplitude. This arrangement provides a 50- to 250-mc sine wave to the test jig and a sync signal in the 100-kc to 30-mc range that the scope will accept. The output of the test jig is connected to a 50-ohm termination through the coaxial probe for the other scope channel.

A measurement of the magnitude and phase of alpha is made in the following manner. The emitter-collector terminals are shorted together with a low inductance plate that covers the transistor socket except for the base terminal region. The input and output waveforms were displayed on the dual traces of the scope. The amplitudes of the two waves and the difference between the zero crossings were noted. The difference in zero crossing in centimeters was converted to degrees. The short circuit was removed and the transistor was inserted in the socket. The required bias conditions were set up, normally a collector-base voltage of 28 volts and a collector current of 250 ma or 500 ma. The RF input amplitude was adjusted to the value noted on the scope with the short present. The output amplitude and the difference in zero crossing between the input and output wave were determined from the scope. The magnitude of alpha is the ratio of the output amplitude to the input amplitude. The phase of alpha is the difference of the difference in zero crossings with the short and with the transistor. This procedure was repeated at a number of frequencies so that a complete curve was obtained.

The validity of this measurement procedure was checked by measuring Western Electric transistors whose alpha vs frequency curves were available. The measurements checked within 10 per cent of the nominal value curves of the transistors. The accuracy is determined by the impedance of the socket short circuit and the ability to read vertical and horizontal dimensions on the waveform displayed on the scope.

APPENDIX B

Analysis of Transistor Pulse Response

The analysis of the transistor pulse amplifier in the common-base configuration follows directly from the work of Gartner.[6] From the physics of the transistor, the short-circuit current gain is defined as

$$\alpha = \operatorname{sech} \frac{w}{L_{pB}} (1 + j\omega\tau_{pB})^{\frac{1}{2}} \tag{1}$$

where

ω = frequency
w = base width
L_{pB} = base diffusion length
τ_{pB} = base lifetime.

We define

$$a = \frac{w}{L_{pB}}$$

$$b = \frac{w}{L_{pB}} \sqrt{\tau_{pB}}$$

$$s = j\omega$$

and rewrite (1) as

$$\alpha = \operatorname{sech} (a^2 + b^2 s)^{\frac{1}{2}}. \tag{2}$$

The low-frequency α is determined from (2) by letting $\omega = 0$.

$$\alpha_{N0} = \operatorname{sech} a. \tag{3}$$

The short-circuit current gain is also defined as

$$-i_c(t) = \alpha i_E(t).$$

Applying (2) and taking reverse Laplace transforms leads to

$$-i_c(t) = \mathfrak{L}^{-1}[(\operatorname{sech} [a^2 + b^2 s]^{\frac{1}{2}}) \ \mathfrak{L}(i_e[t])]. \tag{4}$$

If $i_e(t)$ is a step of emitter current, i_{E1}, the collector current becomes

$$\frac{-i_c(t)}{\alpha_{N0} i_{E1}} = \frac{1}{\alpha_{N0}} \mathfrak{L}^{-1} \left[\frac{\operatorname{sech} (a^2 + b^2 s)^{\frac{1}{2}}}{s} \right]. \tag{5}$$

To solve (5) we write the hyperbolic secant in a series expansion and apply the tables of integral transforms. This yields

$$\frac{-i_c(t)}{\alpha_{N0} i_{E1}} = \frac{1}{\alpha_{N0}} \sum_{n=0}^{\infty} (-1)^n \left\{ \exp [(2n+1)a] \operatorname{erfc} \left[\frac{(2n+1)}{2} \frac{b}{\sqrt{t}} + \frac{a}{b} \sqrt{t} \right] + \exp [-(2n+1)a] \operatorname{erfc} \cdot \left[\frac{(2n+1)}{2} \frac{b}{\sqrt{t}} - \frac{a}{b} \sqrt{t} \right] \right\} \tag{6}$$

where erfc is the complementary error function.

We define the alpha cutoff frequency ω_N as

$$\frac{|\alpha|}{a_{N0}} = \frac{1}{\sqrt{2}}.$$

Using (1) for α we obtain

$$\omega_N = K/b^2$$

where $K = 2.64$ for $\alpha_{N0} = 0.98$.

This step response of the transistor can be used to find the response to arbitrary waveforms. The arbitrary waveform can be approximated by a stair case function — i.e., a series of finite steps. This step-wise approximation can be expressed mathematically as

$$i_{eN}(t) = \sum_{\alpha=0}^{n} \delta(i_\alpha)\, S_{-1}(t - \alpha\delta t) \tag{7}$$

where $S_{-1}(t)$ is a symbol for the unit step and n is the largest integer, so that

$$t - n\delta t > 0.$$

Equation (6) can be written as

$$-i_c(t) = A(t) i_{E1}(t).$$

The approximate collector current for an arbitrary emitter current becomes

$$-i_c(t) = \sum_{\alpha=0}^{\infty} \delta(i_{e\alpha}) A(t - \alpha\delta t) S_{-1}(t - \alpha\delta t). \tag{8}$$

For a raised cosine emitter pulse defined by

$$i_e(t) = I_0 \cos^2 \pi \left(t - \frac{t_0}{2}\right)$$

equation (8) becomes

$$-i_c(t) = \sum_{\alpha=0}^{t_0} \delta[i_{e\alpha}(t)] A(t - \alpha\delta t) S_{-1}(t - \alpha\delta t)$$

where $\delta[i_{e\alpha}(t)]$ is

$$\delta[i_{e1}(t)] = i_e(\delta t) - i_e(0)$$

$$\delta[i_{e2}(t)] = i_e(2\delta t) - i_e(\delta t), \text{ etc.,}$$

and $A(t - \alpha\delta t)$ is written in a like manner. The solution of (8) was ob-

tained using an IBM 7090 computer with $\delta t = 0.1t_0$ and several values of α_{N0}.

REFERENCES

1. Peters, C. J., Gigacycle Bandwidth Coherent Light Traveling-Wave Phase Modulator, Proc. IEEE, *51*, January, 1963, pp. 147–153.
2. Ruthroff, C. L., Some Broadband Transformers, Proc. IRE, *47*, August, 1959, pp. 1337–1342.
3. Enloe, L. H., and Rogers, P. H., Wideband Transistor Distributed Amplifiers, 1959, Solid-State Circuits Conference Digest, pp. 44–45.
4. Beneteau, P. J., and Blaser, L., The Design of Distributed Amplifiers Using Silicon Double-Diffused Transistors, The Solid State Journal, *2*, March, 1961, pp. 38–43.
5. Ballentine, W. E., and Blecher, F. H., Broad Band Transistor Video Amplifiers, 1959, Solid-State Circuits Conference Digest, pp. 42–43.
6. A discussion of this type of operation is given by W. W. Gartner in his book, *Transistors: Principles, Design, and Applications,* Van Nostrand Co., Inc., 1960.
7. Easley, J. W., The Effect of Collector Capacity on the Transient Response of Junction Transistors, IRE Trans. on Electron Devices, *ED-4*, January, 1957.

A 14-Watt Transistor CW Amplifier with a 50-mc Bandwidth

By L. U. KIBLER

(Manuscript received June 24, 1965)

A transistor amplifier capable of delivering 14 watts of CW power into a 15-ohm load over a band of frequencies from 46 to 90 mc was designed for use with an FM optical modulator. This paper describes the performance and details the design and construction of the amplifier.

I. INTRODUCTION

An experimental FM optical communication system under active study required an amplifier capable of delivering a one-ampere rms signal to an optical phase modulator over a 60-mc band centered at 70 mc. The optical phase modulator[1] consists of a one-meter long 15-ohm strip line partially filled with KDP (potassium dihydrogen phosphate). The amplifier must amplify the FM input signal with a minimum harmonic distortion. Since the signal is FM modulated, the amplitude linearity of the amplifier is unimportant.

Rather than embark on a new design using stagger tuned amplifier sections, it was decided to modify a high-power transistor pulse amplifier[2] to operate as a CW amplifier. The general description of this modified amplifier appears in Section II of this paper. Amplifier performance and the modifications made to it are then described in Section III. Its design is presented in Section IV and a discussion of results is contained in Section V.

II. GENERAL DESCRIPTION

The CW amplifier consists of a power amplifier driven by a preamplifier. The power amplifier uses two UHF silicon power transistors, RCA type 2N2876, in a common-base configuration coupled at the input and output with broadband transformer hybrids. The preamplifier is a three-stage amplifier using broadband transformer

coupling. It uses a UHF silicon transistor, RCA type TA2307 (2N3375), in the common-base configuration.

The complete amplifier delivers a 0.98-ampere rms signal to a 15-ohm load impedance over a band ranging from 44 mc to 90 mc and centered on 67 mc. The amplifier has a gain of 21 db and requires a 60-ma signal in the 50-ohm input line for a full output of 14.5 watts. The circuit diagram of the power amplifier and the preamplifier is shown in Figs. 1(a) and (b). The transformer hybrids and the coupling transformers are of the type described by Ruthroff.[3]

The total dc power required is 33 watts: a V_{CB} of 28 volts and a total I_C of 0.8 ampere for the power amplifier, and a V_{CB} of 32 volts and a total I_C of 0.32 ampere for the preamplifier. At the full output of 14.5 watts, the efficiency of the complete amplifier is 44 per cent.

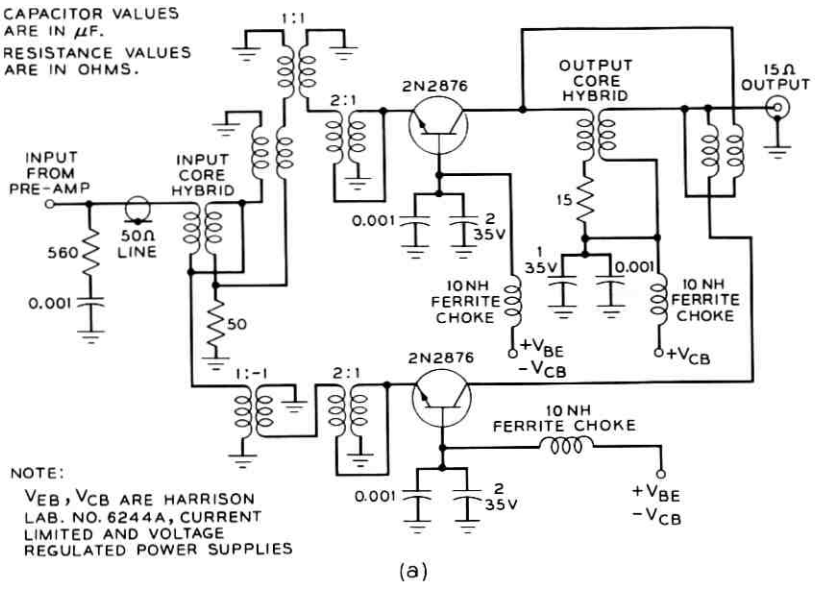
The power amplifier is 3 inches wide, 1 inch deep and $1\frac{3}{4}$ inches high. The preamplifier is 1 by 1 by 4 inches. The total volume of the complete amplifier excluding the heat sinks is $9\frac{1}{4}$ cubic inches. The amplifier is shown in Figs. 2(a), (b), and (c).

III. PERFORMANCE*

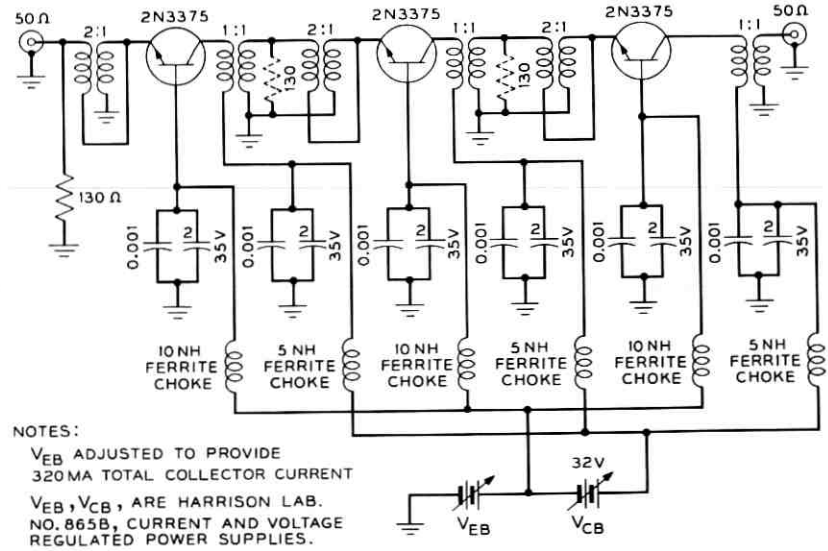
The maximum power dissipation of the power amplifier transistors limits the output power that can be obtained from CW operation of the pulse amplifier. The 14.5-watt output in the 15-ohm load impedance was obtained by experimentally adjusting the drive and the transistor bias conditions to obtain a class AB mode of operation. This mode of operation as used in this amplifier over the frequency band of interest gives the maximum power output with minimum distortion.

The frequency response of the complete amplifier consisting of the preamplifier and power amplifier is shown in Fig. 3. The maximum power output was obtained at 50 mc with upper and lower 3-db frequencies of 90 mc and 44 mc, respectively. Examples of the output wave shape for a sine wave input are shown in Fig. 4. The first scope picture, Fig. 4(a), shows some distortion present at 46 mc. The remaining two photographs, Figs. 4(b) and (c), show no distortion present at 70 mc and 90 mc, respectively. The distortion evident at 46 mc becomes smaller and disappears at 55 mc. A fall off in response from 60 mc to 100 mc of 6 db per octave was required by the characteristics of the FM modulator.

* These measurements were taken with a Hewlett Packard 185A sampling scope with the coaxial probe across a 50-ohm load. The 15-ohm output terminal of the power amplifier was connected to the 50-ohm line through a 0.01-μf coupling capacitor and a 1:4 broadband transformer. The location of this transformer and the 50-ohm type N output connector are shown in Fig. 2(a).

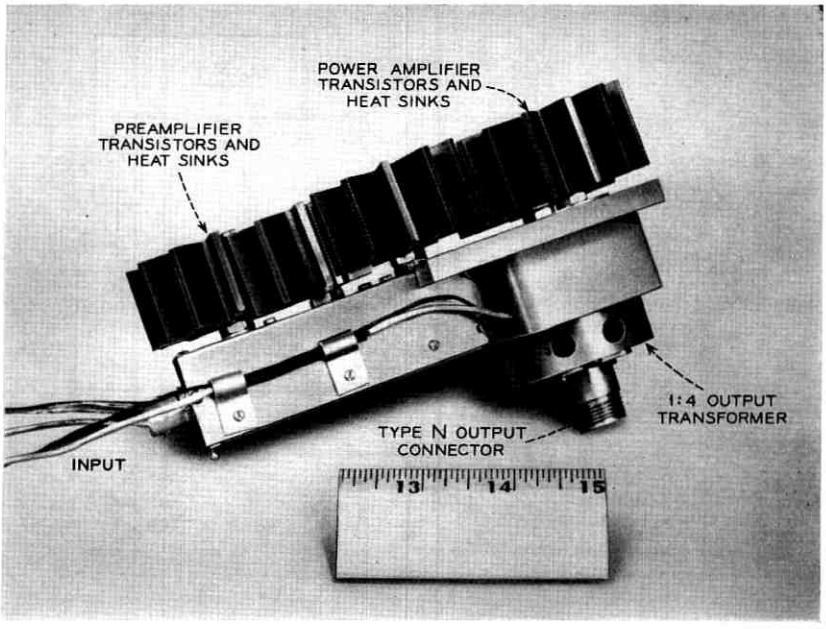


(a)

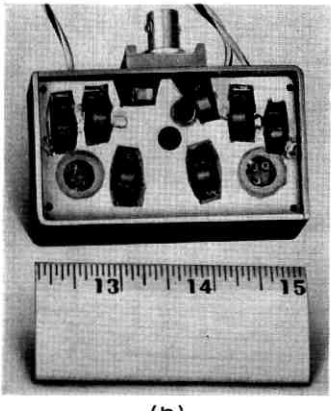


(b)

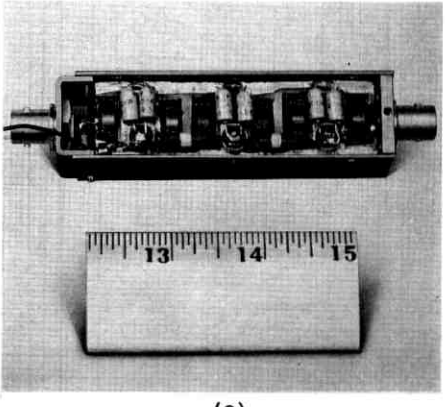
Fig. 1 — (a) Power amplifier schematic diagram; (b) preamplifier schematic diagram.



(a)



(b)



(c)

Fig. 2 — (a) Side view of complete amplifier; (b) top view of interior of power amplifier; (c) bottom view of interior of preamplifier.

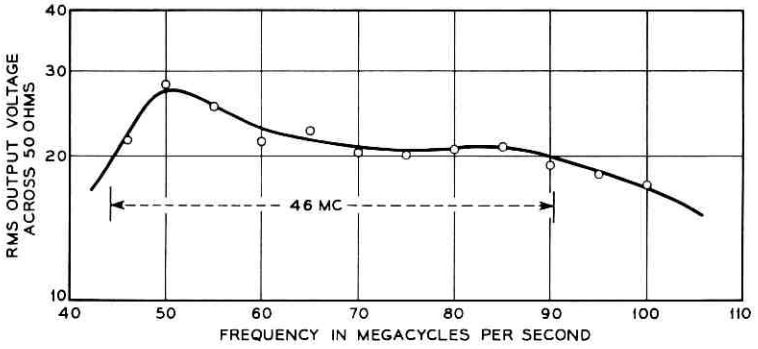


Fig. 3 — Frequency response of CW power amplifier for a 1.6-volt rms drive.

The distortion present in the output of the power amplifier results from the use of the class AB operation. In this bias condition, the transistor is operated at a collector current between that necessary for class A operation and that necessary for cutoff. Since the two transistors in the power amplifier are operated in a push-push configuration rather than a push-pull one, the second harmonic distortion is not cancelled in the output. However, by operating only over a band of frequencies from the upper 3-db frequency, $\omega_{\mu\,3\,db}$ to one half this frequency, the second harmonic distortion is filtered out by the amplifier itself.

In addition to harmonic distortion, class AB operation causes rectification of the input signal. This can produce additional direct current in the collector and emitter of the transistor with an increase in the dc power dissipation in the transistor.

Under class AB bias condition, an increase in the emitter drive

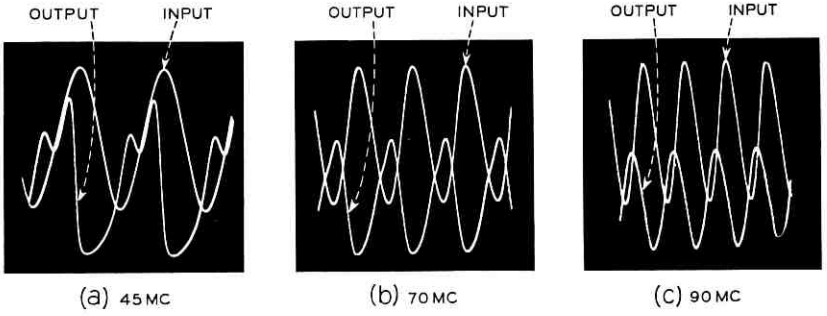


Fig. 4 — CW power amplifier output wave shapes.

causes the emitter-base junction of the transistor to become reversed biased over a portion of the RF cycle. The effect of the reversed biased emitter-base junction is to cause rectification of the emitter signal and the increase in dc emitter current. In order to prevent an increase in the collector current with drive, both the emitter and collector bias supplies were current-limited in the absence of signal to a predetermined level.

A further increase in the emitter drive causes the transistor to be driven into the saturation region where the collector current depends only on the collector supply voltage and the load. The nonlinearity associated with the transition to this region causes an increase in the dc collector current. The use of the current-limited bias supplies prevents this increase.

The bias and drive conditions for class AB operation of the power amplifier were determined experimentally. The total collector current to the two transistors was limited to 0.9 ampere with a collector voltage of 28 volts. The emitter bias was adjusted to provide a limited total current of 0.9 ampere in the absence of signal. The drive was increased until a maximum output power with minimum distortion was obtained. Under these conditions the total collector current was 0.8 ampere.

These bias conditions correspond to a dc power dissipation of 11.2 watts per transistor. Approximately 5.5 watts of RF drive were required for full output. To obtain the 14.5-watts output, each transistor must supply 4.5 watts of RF power. The total power dissipation, the sum of the RF and dc power, is 15.7 watts per transistor. The maximum allowable dissipation of the 2N2876 transistor at a 25°C case temperature is 17.5 watts. A derating factor of 1 watt per 10°C rise in case temperature is required. The case temperature rise of the transistors in the power amplifier was held at 10°C or less through the use of heat sinks and copious quantities of cooling air.

The preamplifier is operated in class AB under current-limited bias conditions also. The three 2N3375 transistors are operated from a common-collector bias supply and a common-emitter bias supply. The total collector current was limited to 0.32 ampere at a collector voltage of 32 volts. The emitter current was limited to 0.34 ampere under no-signal conditions. An input signal of 60 ma to the preamplifier in the 50-ohm input line is required to drive the power amplifier to the full output of 14.5 watts.

The need for class AB operation rather than class A operation is evident when one examines the transistor characteristic curves and the

region of second breakdown. For class A linear operation of the power amplifier, an rms collector current of 0.5 ampere is required to obtain the 15-watt output. Thus the peak collector current is 0.71 ampere. With the nominal α of 0.9, we need a peak emitter current of 0.79 ampere. The peak collector voltage across the 30-ohm load seen by each transistor is 22 volts. The class A bias points must be at a collector current and voltage somewhat greater than these values to prevent distortion due to the nonlinear transistor characteristics at the peak value of current. For a collector voltage of 26 volts and a bias current of 0.75 ampere, the transistor is in a region of second breakdown for class A operation.

The 60-ma input signal required to drive the power amplifier to its full output could not be supplied by the IF amplifier used in the FM system. A second preamplifier was designed to amplify the +5 dbm output of the IF amplifier. This amplifier was of the same basic construction as the preamplifier used to drive the power amplifier. The original preamplifier had a rising output with frequency to compensate for the fall off of the power amplifier. In order to flatten the frequency response of the second preamplifier, it was necessary to add a 130-ohm, $\frac{1}{8}$-watt resistor across the output of each collector transformer. These resistors are shown dotted in the circuit diagram of the preamplifier in Fig. 1. The measured frequency response of this second amplifier is shown in Fig. 5.

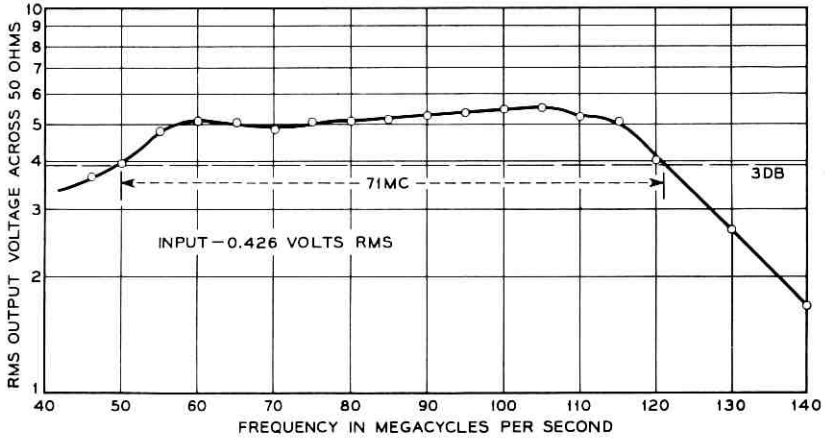


Fig. 5 — Second preamplifier response.

IV. AMPLIFIER DESIGN AND CONSTRUCTION

The operation of the power amplifier can be seen by reference to Fig. 1(a). Consider a unit current signal at a 50-ohm impedance level incident on the input transformer hybrid. Unit current signals appear at the two conjugate arms at a 25-ohm impedance level 180° out of phase. A 1:1 reversing transformer in one arm reverses the phase of one signal. A 1:1 transformer in the opposite arm preserves the phase of the signal but introduces a delay equal to that of the reversing transformer. These in-phase signals are applied to the emitter of the respective transistors through 4:1 impedance transformers. These two unit current signals applied to the emitters are at a $6\frac{1}{4}$-ohm impedance level. This level is close to the 6-ohm input impedance of the transistor in the common-base configuration. The transistor collector current using nominal short current gain, α, of 0.9 is approximately 1.8 units. The output signals of the two transistors at a 30-ohm impedance level are combined in the output hybrid to produce a 3.6-unit amplitude current signal in the 15-ohm output impedance. The power amplifier has a current gain of 3.6. Due to the class AB operation, the actual current gain will be somewhat less.

The operation of the preamplifier can be seen from the circuit diagram of Fig. 1(b). Consider a unit current signal incident at the input. The first 4:1 transformer produces a 2-unit amplitude signal at the emitter of the first transistor. Under large signal conditions the input impedance of these transistors, 2N3375, is approximately 8 ohms. Using a nominal α of 0.9, the collector current will be 1.8 units. The output of the transistor is coupled to the 4:1 transformer at the input of the next transistor through a 1:1 isolation transformer. The use of this transformer provides a path for the collector bias. The operation of the two following stages is identical with the first stage. The overall current gain is 5.83. Because of the class AB operation, the actual current gain will be less.

The current gain of a complete amplifier between the 32-ohm input impedance of the preamplifier and the 15-ohm output of the power amplifier is 23 db. Since the measured gain was 21 db, approximately 2 db of gain is lost in using the class AB bias conditions.

The basic amplifier construction uses a quasi-stripline to interconnect the core transformers and transistors. The input and output circuits are placed on opposite sides of a center board in nonoverlapping areas. Ground planes boards are placed on top and bottom of the common board. This results in a 3-layer sandwich construction. The center board is $\frac{1}{8}$-inch thick glass-loaded Teflon with 2-oz copper on

each side. The ground plane boards are $\frac{1}{16}$-inch thick glass-loaded Teflon with 2-oz copper on each side. Transistor sockets were made by placing spring contacts for each pin on the transistor case into holes drilled in the center board. These contacts were made using the center conductor of a BNC female coaxial connector. After being placed into the center board, these contacts were soldered to the copper. Matching rectangular holes were cut in each board to accommodate the core transformers. The copper on the board surfaces was cut away so as to leave stripline connections between the cores and transistors. The boards were separated by placing $\frac{3}{16}$-inch thick brass spacer around the periphery of the boards. Where this spacing was not adequate for the proper stripline impedance, brass pieces were attached to the center board to decrease the spacing in a local region. The terminations for the input and output hybrids — 50 ohms and 15 ohms, respectively — were 1 watt, 1 per cent metalized film microwave rod resistors made by Film Ohm Corporation. The base by-pass capacitors consisted of a 0.001-μf feed-through capacitor mounted in the ground plane with the center conductor connected to the base terminal of the center board. Two 2-μf miniature tantalum capacitors were used in parallel with the feed-through capacitor. The collector dc connection was bypassed with a 0.01-μf postage stamp type capacitor in parallel with a 2-μf miniature tantalum capacitor. Collector bias was applied between collector-base terminals. Emitter bias was applied between base terminals and ground.

The layout of the three circuit board for the power amplifier is shown in Fig. 6. Both sides of each board are shown. The layout of the 3-transistor circuit board of the preamplifier is shown in Fig. 7. The position of the parts and the sections of stripline are evident upon examination of these figures.

The outside dimensions of the circuit boards were dictated by the space available inside the KDP modulator structure and by a need to keep all leads as short as possible. To keep the connection between the power amplifier and the KDP stripline to a minimum, the power amplifier had to be placed in a space 1-inch deep, 3-inches wide and $1\frac{5}{8}$-inches high. A mounting box of these outside dimensions was made of $\frac{1}{16}$-inch brass, and the circuit board was sized so as to fit inside this box. The transistor heat sinks, 2-inch diameter, 1-inch high 12-fin aluminum cylinders, were extended out through the end plate of the KDP mounting box. The heat sinks were mounted on the transistor package stud and kept isolated from ground. Since the collectors of the 2N2876 and TA2307 are isolated from the case (collector-to-case capacitance 6 pf), the mounting of the heat sinks in this manner reduced the col-

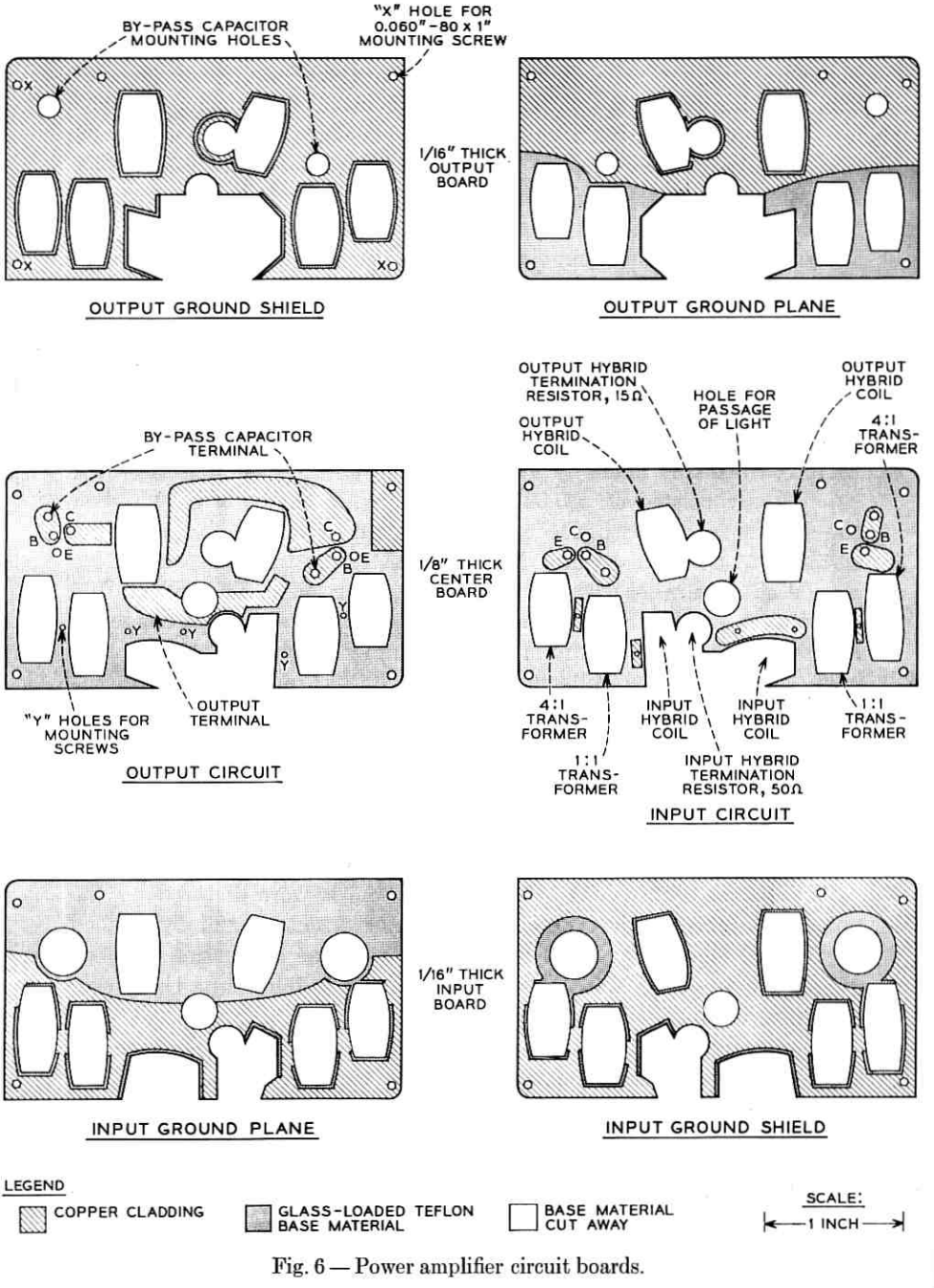


Fig. 6 — Power amplifier circuit boards.

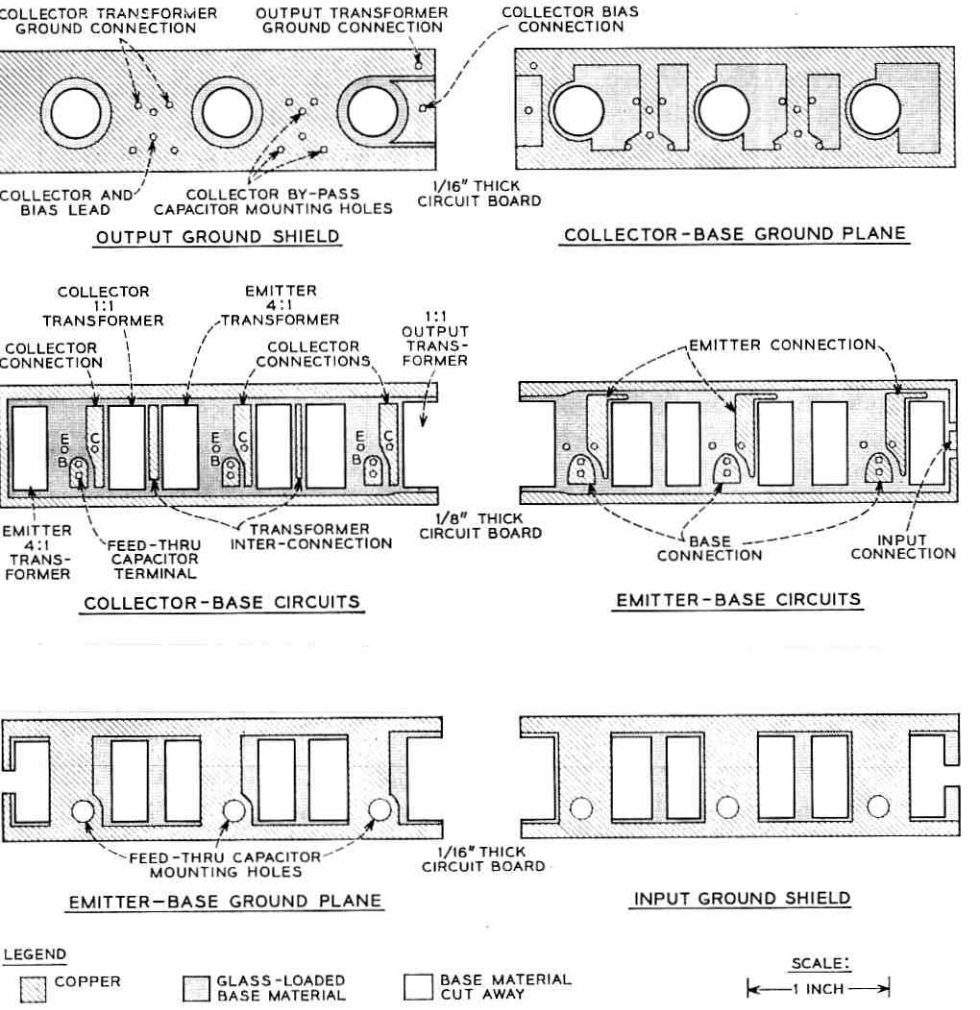


Fig. 7 — Preamplifier circuit boards.

lector circuit capacitance to a value close to the transistor collector capacitance, 20 pf for the 2N2876 and 10 pf for the TA2307.

The cores in the output hybrid of the power amplifier were bifilar wound with 6 turns of number 22 Formex wire. The transformer cores and the input hybrid cores were bifilar wound with 6 turns of number 24 Formex wires. All the transformer cores in the preamplifier except the output transformer were bifilar wound with 6 turns of number 24

Formex wire. The output transformer core was bifilar wound with 5 turns of number 24 Formex wire.

The size of the preamplifier was also dictated by the dimension of the KDP modulator structure. A narrow 1-inch by 1-inch passage existed between the top of the power amplifier and the top wall of the modulator structure. The preamplifier was mounted in a $\frac{3}{64}$-inch thick brass box 1 by 1 by 4 inches mounted on top of the power amplifier and extended through the top of the modulator structure. Connection of the preamplifier with the coaxial output of the modulation source was made through a BNC connector. The circuit boards of the preamplifiers were cut to dimensions to fit this box. The heat sinks for the preamplifier were of the same type as those of the power amplifier. Certain fins of each heat sink were cut away so that the heat sinks of successive preamplifier transistors could interleave. The details of the complete amplifier can be seen in Figs. 2(a), (b) and (c).

V. DISCUSSION

The broadband transformer-coupled transistor pulse amplifier has been shown capable of high-power CW operation. CW output power up to 14 watts into a 15-ohm load was obtained over a 46-mc band extending from 44 to 90 mc. CW operation of the pulse amplifier was obtained by using class AB bias with current-limited power supplies. The distortion generated by this bias condition was eliminated from the output by using the amplifier as its own filter. This scheme of filtering is useful for operation over a band of frequencies extending from the upper 3 db frequency of the amplifier to one half this frequency. A second amplifier with selected transistors, matched in amplitude and phase of the α response, delivered 9 watts ± 1.5 db with a 3-db bandwidth extending from 70 mc to 140 mc. As a comparison, a similar amplifier operated with class A bias can deliver 5 watts to the 15-ohm load with some distortion over a band extending from 19 mc to 86 mc. Several of these amplifiers have been in daily use for several months. There has been no degradation in their performance during this period.

With the advent of this family of high-power UHF and VHF silicon power transistors using a novel overlay construction, it is possible to design broadband high-power CW and pulse amplifiers that operate into low impedance loads. With a staggered tuned amplifier design it should be possible to obtain output powers of 10 or more watts over wide bands centered in the 100- to 300-mc region. Future effort could well be directed in this direction.

The success of the amplifier described in this paper is due to large part to the ability of J. W. Batton. He was able to accurately construct and align the stripline segments that make up the major interconnection network of the amplifier.

REFERENCES

1. Peters, C. J., Gigacycle Bandwidth Coherent Light Traveling-Wave Phase Modulator, Proc. IEEE, *51*, January, 1963, pp. 147–153.
2. Kibler, L. U., B.S.T.J., This Issue, pp. 1977–2001.
3. Ruthroff, C. L., Some Broadband Transformers, Proc. IRE, *47*, August, 1959, pp. 1337–1342.

Light Propagation in Generalized Lens-Like Media

By S. E. MILLER

(Manuscript received July 23, 1965)

This paper provides a preliminary assessment of electromagnetic wave propagation in focusing media which departs from those previously studied, ideal lenses and continuous media with square-law index variation. New approximate methods are described for obtaining the transverse beam width, phase constant, and ray trajectory in continuous lens-like media, and for determining stability conditions in lens waveguides, where the lenses contain spherical aberration and the continuous media contain fourth-order or higher-order terms of variation in index of refraction.

It is proven that only in aberrationless-lens waveguides or in a continuous medium with square-law index variation will the shape of a beam injected off axis or with an angle to the medium's axis remain constant about a beam axis which oscillates about the axis of the medium. In non-square law media the beam will spread, but knowledge of the coefficients describing the medium and the position and angle of the injected beam enables one to specify the maximum radius within which all of the energy will be confined.

The following is an example of the type of solution obtained for non-square law media: for a medium characterized by the transverse index variation

$$n = n_a\,(1 - \tfrac{1}{2}a_4x^4)$$

and assuming no index variation exists in the direction of propagation, the radius to the $1/e$ *point in field is approximately*

$$w_e = 0.666\,\frac{\lambda^{1/3}}{a_4^{1/6}}$$

and the phase constant is

$$\beta = \frac{2\pi}{\lambda} - \frac{0.256\,(a_4\lambda)^{1/3}(m+1)^2}{\left(\dfrac{m+2.5}{2.5}\right)^{2/3}}$$

where the free-space wavelength $\lambda_0 = n_a\lambda$ *and* m *is the order of the mode,* $m = 0, 1, 2 \cdots$ *. Quite generally, non-square law media show dispersion, and unlike the square-law media the various modes travel with different group velocities. Expressions are given to allow these effects to be evaluated for small perturbations on a square-law medium as well as for higher-order index variations.*

The transverse beam shape associated with any law of index variation is shown to be as well approximated (in the region of significant power density) by a cosine function or Gaussian function as is the field for an ideal lens-waveguide approximated by a Gaussian function in the presence of typical diffraction losses.

Normal mode shapes are obtained for resonators with fourth-order and eighth-order mirrors by the method of Fox and Li; diffraction losses for a few Fresnel numbers are also given. In a certain range of Fresnel numbers, fourth- and eighth-order mirrors give lower diffraction losses than spherical mirrors.

An approximate method for solving the paraxial ray equation for rather general (non-square-law) media is outlined. Requiring only reciprocity and symmetry about the medium's axis, it is shown that the radial position of the ray (x) *is related to distance* (z) *along the axis of the medium by*

$$x = \Sigma b_m \cos m\beta z$$

where $m = 1, 3, 5, 7 \cdots$ *. Moreover, it is shown that this series converges very rapidly, making it possible to get a good approximate representation with only a few terms. For example, for the fourth-order medium described above, an approximate solution is*

$$x = x_0 \{0.959 \cos \beta z + 0.041 \cos 3\beta z\}$$

where $\beta = x_0 \sqrt{1.44a_4}$ *. It is characteristic of all non-square law media to have a ray period* $2\pi/\beta$ *which is a function of the peak ray displacement,* x_0 *.*

Lens waveguides with fourth- or higher-order terms in the focusing or index function can of course exhibit increasingly strong focusing for energy departing farther from the medium's axis, and if the lenses are suitably spaced might be useful in reducing the magnitude of beam wander due to imperfections or guide-axis curvature. However, for a given beam spot size, non-square-law lenses must be placed closer together than square-law lenses.

Use of non-square law lenses or distributed media in transmission systems will most probably require repeater-system techniques which are operable with multi-mode signals at the receiver input.

I. INTRODUCTION

Although a great deal of work has been done to describe electromagnetic wave guidance using a sequence of ideal aberrationless lenses, little has been done with more general lens-like media. The objective of this paper is to provide a beginning understanding of what happens to the important descriptive parameters of a light waveguide—wave phase constant, wave spot size, ray trajectory, stability restrictions—for general lens-like media, either continuous or formed from a sequence of focusing elements. Exact solutions for these parameters are difficult or impossible, a possible reason for little having been done.* A second result presented herein is a series of analytical techniques for obtaining useful approximate solutions representing a broad class of lens-like media; some of these techniques are most clearly presented by giving examples, which unfortunately leads to a rather large number of equations. To aid the reader in finding the section dealing with a particular topic, an outline and brief resume is given below.

As pointed out by J. W. Tukey, there is no *a priori* reason why ideal aberrationless lenses are best for use as a communication medium. We would like to know what does happen to wave guidance as we depart from aberrationless lenses, which is the case studied extensively in earlier work. Gas lenses have nearly constant focal length but their "principal planes" are actually curved surfaces.[1] E. A. Marcatili and D. H. Ring have pointed out that these curved principal planes have effects similar to those expected in plane lenses with spherical aberration. Thus the present work relates to existing gas lenses as well as to the question of whether or not to attempt to create a different form of inhomogeneous medium for light-wave guidance.

Some of the present work dates back more than a year; impetus to putting it on paper was given by recent calculations planned by E. A. Marcatili and D. Marcuse.[2] These calculations showed that ray optics can accurately predict the loss even when half of the beam falls off the edge of the lens. Many of the conclusions drawn here are based on ray optics but since the controlling transverse variations take place over distances that are large compared to the wavelength, ray optics should give a correct conclusion.

The subject is developed in this paper in the following manner. Section II covers ray optics for aberrationless lenses of any thickness, in-

* As this paper goes into print, S. J. Buchsbaum points out the existence of an exact solution for a particular non-square-law $f(x)$ originally derived with reference to a quantum-mechanical problem. J. P. Gordon will report on this in a later publication.

cluding the case of a continuous medium. A ray-optic method for determining stability conditions applicable (using results of later sections) with generalized focusing elements is also covered here. Proof that the transverse field distribution of a normal mode beam injected off the medium's axis (or which goes there due to curvature, or displacement of the medium) is preserved only when the index of refraction decreases as the square of the distance off axis is presented in Section III. Section IV develops the proof that in a wide class of continuous focusing media, the ray paths are representable by a series of odd-harmonic cosine or sine terms. A definition of the focal length of a segment of an arbitrary medium is covered in Section V. Section VI discusses a new technique for obtaining approximately the wave phase constant and spot size for generalized lens-like media, stated generally and illustrated for fourth-order and eighth-order media. A characteristic ray angle and a characteristic ray period is defined for the generalized medium. Calculations for field distribution and diffraction losses in resonators and lens waveguides using non-square-law elements are derived in Section VII. These calculations are the only ones available on fields and losses in non-square-law resonators. Section VIII discusses a solution for ray paths in pure fourth-order, or a mixture of second- and fourth-order, media; the approach for extension to other media is indicated. Some further discussion and acknowledgements are presented in Section IX.

II. TRANSMISSION IN AN IDEAL LENS-LIKE MEDIUM

The medium referred to as ideal is one in which the index of refraction has the form[3] (see Fig. 1a)

$$n = n_a\,(1 - \tfrac{1}{2}a_2x^2) \tag{1}$$

where

n_a = index of refraction on axis, $x = 0$
x = transverse dimension
a_2 = a constant.

The index is independent of z,

$$\partial n/\partial z = 0. \tag{2}$$

We can obtain a solution for the path of a light ray in such a medium using the well known paraxial ray equation

$$\frac{\partial^2 x}{\partial z^2} = \frac{1}{n}\frac{\partial n}{\partial x}. \tag{3}$$

Here, x is the position of the ray at some value of z. Equation (3) is

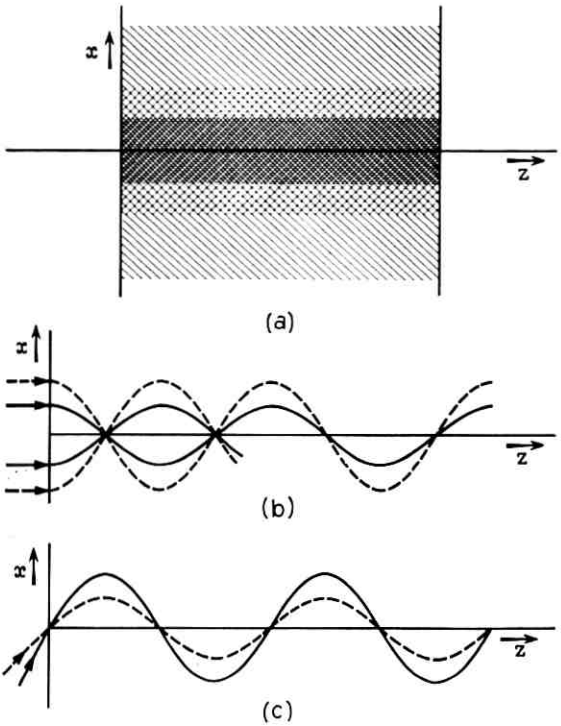


Fig. 1 — Ideal lens-like medium, (a)—two-dimensional continuous medium with index of refraction independent of z and varying with x according to (1), (b)—light ray paths in medium a for parallel input rays, (c)—light ray paths in medium a with non-parallel input rays.

valid over a broad range of conditions provided only that the ray path makes a small angle to the z-axis. For a medium of the form of (1)

$$\frac{\partial n}{\partial x} = -n_a a_2 x \tag{4}$$

and with the restriction

$$|\tfrac{1}{2} a_2 x^2| \ll 1 \tag{5}$$

we can write

$$\frac{\partial^2 x}{\partial z^2} = \frac{1}{n}\frac{\partial n}{\partial x} = -a_2 x. \tag{6}$$

The general form of (6) indicates an exponential solution for x as a function of z, and it can be verified that

$$x = A \cos \sqrt{a_2}\, z + B \sin \sqrt{a_2}\, z \tag{7}$$

is a solution, A and B being constants to be determined.

Using the boundary conditions at $z = 0$,

$$x = r_i = \text{input ray position} \tag{8}$$

$$\frac{\partial x}{\partial z} = r_i' = \text{input ray slope.} \tag{9}$$

We have

$$x = r_i \cos \sqrt{a_2}\, z + r_i' \sqrt{1/a_2} \sin \sqrt{a_2}\, z. \tag{10}$$

This is very interesting in that it corresponds exactly to the ray behavior in a sequence of equally spaced aberrationless lenses.[4,5] It is important that all rays have the same oscillatory period regardless of input displacement or input slope, as sketched in Figs. 1(b) and 1(c). It is likewise significant that the effects of input ray slope and position are separable with respect to x, subsequent ray position.

If a_2 is negative, the cosine and sine of (10) become cosh and sinh and a divergent medium results.

A short segment of the distributed medium characterized by (1) may be assigned an equivalent focal length, even when it is not a "thin" or "weak" lens. With reference to Fig. 2, the most general case is one in which the index on axis, n_a, is different from that of the surrounding

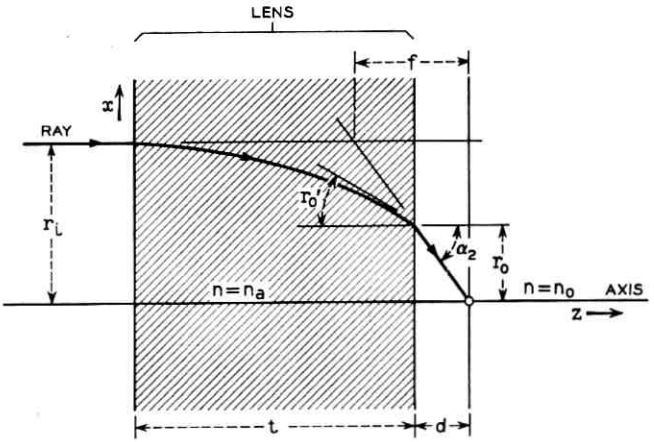


Fig. 2 — Diagram defining focal length and principal plane for an ideal distributed lens.

medium n_0. The path of input ray with zero slope and displacement r_i can be obtained from (10).

The general equation for the ray slope, derived from (10), is

$$\frac{dx}{dz} = -r_i \sqrt{a_2} \sin \sqrt{a_2}\, z + r_i' \cos \sqrt{a_2}\, z. \tag{11}$$

At the lens output the displacement is

$$r_0 = r_i \cos \sqrt{a_2}\, t \tag{12}$$

and the ray slope is, from (11),

$$r_0' = -r_i \sqrt{a_2} \sin \sqrt{a_2}\, t. \tag{13}$$

Because (5) holds, the refraction at the lens output surface is (see Fig. 2)

$$\frac{r_0'}{\alpha_2} = \frac{n_0}{n_a}. \tag{14}$$

Hence,

$$d = \left| \frac{r_0}{\alpha_2} \right| = \left| \frac{n_0 \cot \sqrt{a_2}\, t}{n_a \sqrt{a_2}} \right|. \tag{15}$$

The distance from the focal point to the principal plane where an equivalent thin lens may be placed (Fig. 2) is

$$f = \frac{r_i}{\alpha_2} = \frac{n_0}{n_a \sqrt{a_2} \sin \sqrt{a_2}\, t}. \tag{16}$$

This result was excerpted from the present work and given previously[6] with the approximation appropriate to gas lenses $n_0 \cong n_a$.

With the continuous focusing medium according to (1), there is no limit to the strength of the focusing effect; a stronger index gradient merely confines the normal mode energy more closely to the axis. However, when a series of segments of the distributed medium are used as a waveguide, with gaps of a homogeneous medium in-between, there is a cut-off effect which occurs if the lenses are too strong in relation to the lens spacing. This is a well-known effect with thin lenses,[4] and the limiting conditions were derived from wave optics for the medium according to (1) by J. R. Pierce[3] and E. A. Marcatili.[7] A ray-optic derivation is given here for the medium (1) to show the method which can be extended to arbitrary focusing media, including fourth-order or higher-order terms in x, which has not yet been handled by wave optics.

With reference to Fig. 3(a), we assume an input ray of zero slope and a displacement r_i at the center of the first lens, where the longitudinal axis $z_1 = 0$. At $z_1 = t/2$ and restricting our attention momentarily to the region $0 < \sqrt{a_2}\, t < \pi/2$, the output of lens #1 is

$$r_{01} = r_i \cos(\sqrt{a_2}\, t/2) \tag{17}$$

$$r_{01}' = -r_i \sqrt{a_2} \sin(\sqrt{a_2}\, t/2). \tag{18}$$

Taking the approximation appropriate to gas lenses, $n_0 \cong n_a$, the ray output of lens 1 intersects the axis at a distance $b/2$ from the end of lens #1 (Fig. 3a).

$$b/2 = |\, r_{01}/r_{01}' \,|. \tag{19}$$

Using our knowledge that (1) is symmetrical about $x = 0$ and reciprocity holds, we can construct the remainder of Fig. 3(a) with a ray maximum r_i at the center of the second lens. The value of b corresponding to this condition is

$$b = \frac{2}{\sqrt{a_2}} \cot(\sqrt{a_2}\, t/2). \tag{20}$$

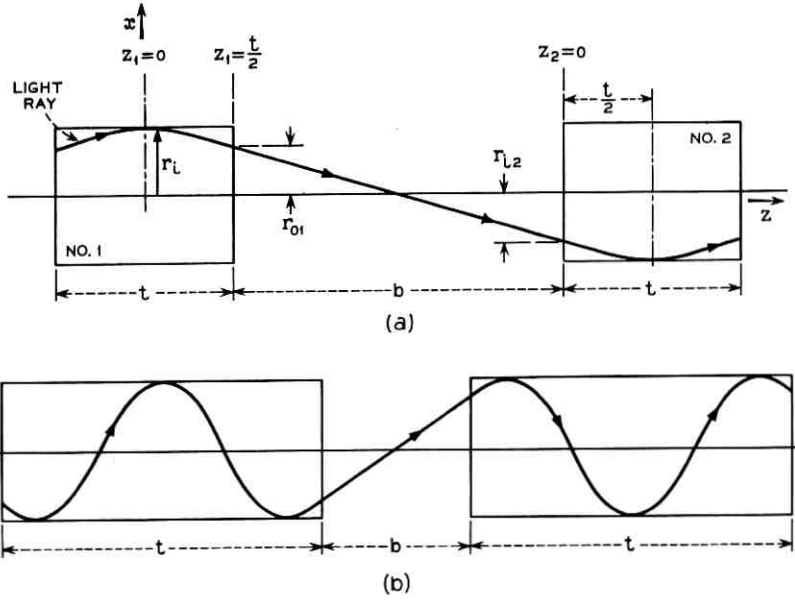


Fig. 3 — Ray path at one maximum lens spacing, (a)—weakest lens case, (b)—second stability band (stronger lenses than 3a).

It is clear that this is the largest permissible ray displacement at the center of lens #2, since a larger value there would represent displacement growth without limit as $z \to \infty$. We can show that larger values of b violate this stability condition by writing the expression for the displacement at the center of the second lens. The input to the second lens is

$$\text{slope} \qquad r_{i2}' = r_{01}' \tag{21}$$

$$\text{displacement} \quad r_{i2} = r_{01} + br_{01}'. \tag{22}$$

Then, from (10) and using a new origin of the longitudinal axis $z_2 = 0$ at the input to the second lens,

$$x_2 = r_{i2} (\cos \sqrt{a_2}\, z_2) + \frac{r_{i2}'}{\sqrt{a_2}} \sin \sqrt{a_2}\, z_2. \tag{23}$$

Using (21) and (22)

$$\frac{x_2}{r_i} = \cos (\sqrt{a_2}\, z) \{\cos (\sqrt{a_2}\, t/2) - b \sqrt{a_2} \sin (\sqrt{a_2}\, t/2)\} - \sin (\sqrt{a_2}\, z_2) \sin (\sqrt{a_2}\, t/2). \tag{24}$$

Making the substitution

$$b = \frac{2}{\sqrt{a_2}} \cot (\sqrt{a_2}\, t/2) + \delta \tag{25}$$

and evaluating (24) at $z_2 = t/2$ we find

$$\left.\frac{x_2}{r_i}\right|_{z_{i2}=t/2} = -[1 + \delta \sin (\sqrt{a_2}\, t/2) \cos (\sqrt{a_2}\, t/2)]. \tag{26}$$

In the region $0 < \sqrt{a_2}\, t/2 < \pi/2$ both $\sin (\sqrt{a_2}\, t/2)$ and $\cos (\sqrt{a_2}\, t/2$ are positive; hence for δ positive, the ray at the center of the second lens is displaced more than the input ray and a divergent path is followed in such a sequence of lenses. For δ negative the propagation is stable. For $\pi < \sqrt{a_2}\, t/2 < 3\pi/2$ another passband occurs with the limiting value of b having a ray path as in Fig. 3(b). This leads to the stability conditions

$$b \leqq \frac{2}{\sqrt{a_2}} \cot (\sqrt{a_2}\, t/2) \tag{27}$$

applicable in the regions of lens length

$$n\pi < \sqrt{a_2}\, t/2 < n\pi + \frac{\pi}{2} \tag{28}$$

where $n = 0, 1, 2, 3$, etc.

For other permissible values of lens length we derive the stability condition with reference to Fig. 4. The input ray at $z_1 = 0$, center of lens #1, has zero displacement and a slope of r_i'. Fig. 4(*a*) shows the ray path for the limiting value of b when $\pi/2 < \sqrt{a_2}\, t/2 < \pi$, constructed using symmetry and reciprocity as before. This yields

$$b = -\frac{2}{\sqrt{a_2}} \tan\left(\sqrt{a_2}\, t/2\right). \tag{29}$$

This gives a positive value for b because the tangent is negative. Following the lines indicated above we can derive an expression for the ray path in lens #2

$$x_2 = \cos\left(\sqrt{a_2}\, z_2\right)\left\{\frac{r_i'}{\sqrt{a_2}} \sin\left(\sqrt{a_2}\, t/2\right) + b r_i' \cos\left(\sqrt{a_2}\, t/2\right)\right\} + \frac{r_i'}{\sqrt{a_2}} \cos\left(\sqrt{a_2}\, t/2\right) \sin\left(\sqrt{a_2}\, z\right). \tag{30}$$

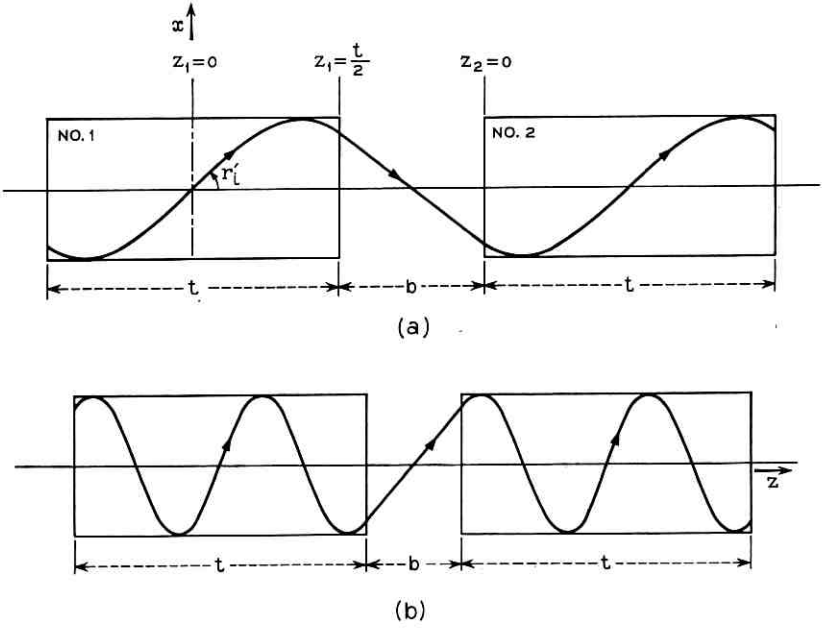


Fig. 4 — Ray path at another maximum lens spacing, (a)—weak lens case (b)—second stability band (stronger lenses than 4a).

We want the slope at the middle of this lens, which is obtained by putting $z_2 = t/2$ in

$$\frac{dx_2}{dz_2} = -\sqrt{a_2}\sin(\sqrt{a_2}\,z)\left\{\frac{r_i'}{\sqrt{a_2}}\sin(\sqrt{a_2}\,t/2) + br_i'\cos(\sqrt{a_2}\,t/2)\right\} + r_i'\cos(\sqrt{a_2}\,t/2)\cos(\sqrt{a_2}\,z)\,. \tag{31}$$

Letting

$$b = -\frac{2}{\sqrt{a_2}}\tan(\sqrt{a_2}\,t/2) + \delta \tag{32}$$

we obtain

$$\left.\frac{1}{r_i'}\frac{dx_2}{dz_2}\right|_{z_2=t/2} = 1 - \delta\sqrt{a_2}\sin(\sqrt{a_2}\,t/2)\cos(\sqrt{a_2}\,t/2). \tag{33}$$

In the region $\pi/2 < (\sqrt{a_2}\,t/2) < \pi$, the sine term is positive and the cosine term is negative; hence a positive δ indicates instability, a slope at the center of the second lens greater than the slope at the center of the first lens and (as may be verified in (30)) a displacement at $z_2 = t/2$ adding to the subsequent ray divergence.

Higher order passbands occur, one sketched in Fig. 4(b). Thus, in the region

$$(n\pi + \pi/2) < (\sqrt{a_2}\,t/2) < (n+1)\pi$$

$n = 0, 1, 2, 3, 4$, etc., the permitted range of lens spacing b is

$$b \leqq -\frac{2}{\sqrt{a_2}}\tan(\sqrt{a_2}\,t/2). \tag{34}$$

Equations (27) and (34) are identical to the relations obtained from wave theory.[7]

The above method for determining stability of lens waveguides can be applied for arbitrarily shaped lenses, as long as symmetry about the z axis and reciprocity exist. Numerical integration can be employed where the lens cannot easily be represented by functions in closed form.[1]

III. WAVE BEHAVIOR IN INHOMOGENEOUS MEDIA

We will now consider media which are of the form

$$n = n_a[1 + f(x)] \tag{35}$$

where

$$f(x) = -\tfrac{1}{2}a_2x^2 - \tfrac{1}{2}a_4x^4 - \tfrac{1}{2}a_6x^6 \cdots \tag{36}$$

in which a_2, a_4, $a_6 \cdots$ are constants which may be positive or negative. Positive values represent convergent focusing and negative values represent divergent focusing. This represents a general medium restricted only to symmetry about the axis $x = 0$ and uniform with respect to the direction of propagation.

The normal mode for such a medium is characterized by the index variation. We will presently discuss the field shapes and losses for several special cases of (35); for now it is sufficient to note that a *different* equation for the index n means a different normal mode.

An interesting general conclusion can be drawn by expanding the index n of (35) about some off-axis radius r_1 as in Fig. 5:

$$\begin{aligned}\frac{n(x - r_1)}{n_a} = \frac{n(x')}{n_a} = \frac{n(x = r_1)}{n_a}\\ + x'\{-a_2r_1 - 2a_4{r_1}^3 - 3a_6{r_1}^5 \cdots\}\\ + \frac{(x')^2}{2}\{-a_2 - 6a_4{r_1}^2 - 15a_6{r_1}^4 \cdots\}\\ + \frac{(x')^3}{6}\{-12a_4r_1 - 60a_6{r_1}^3 \cdots\}\\ + \frac{(x')^4}{24}\{-12a_4 - 180a_6{r_1}^2 \cdots\}\end{aligned} \tag{37}$$

in which

$$x' = (x - r_1).$$

With the approximation

$$f(x) \ll 1 \tag{38}$$

the first term of (37) is unity. Then for a_4, a_6, and all higher order coefficients equal to zero, (37) becomes

$$n(x') = n_a\{1 - a_2r_1x' - \tfrac{1}{2}a_2(x')^2\}. \tag{39}$$

This is sketched in Fig. 5. Thus, for the medium according to (1), the effect of entering off axis at $x = r_1$ instead of at $x = 0$ is to introduce a term in the index which is linear in x. This term has the effect of delaying or advancing every region of the transverse cross section an amount proportional to the displacement from the axis; this term

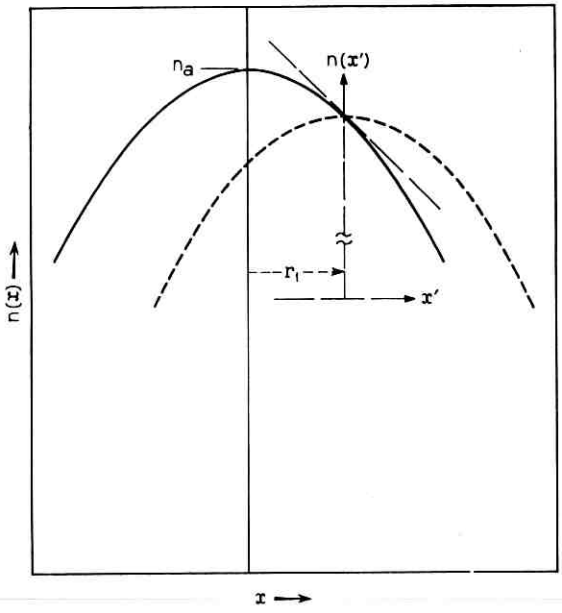


Fig. 5 — Expansion of $n(x)$ about point $x = r_1$.

acts the same as a dielectric wedge and tilts the wave front. The path of any ray is described by ray optics in the continuous medium according to (10). The remaining term of (39) is identical to the original equation at $r_1 = 0$; hence the normal mode for a wave entering at $x = r_1$ is the same as at $x = 0$, with a direction change superimposed.* Thus, the beam follows a path sketched as in Fig. 6(a); a pure mode introduced into the square-law medium travels without change of shape if the change in position and direction of the beam axis is taken into account. If a plane wave is introduced, Fig. 6(b), the field is concentrated periodically at the points where the beam axis crosses the axis of the medium.

We can also see immediately that these properties are *not* characteristic of any medium in which a_4 or higher order terms are present. In the expansion (37) there are, in addition to the term linear in x', additional terms in $(x')^2$ and $(x')^3$ brought in by a_4, a_6, etc., which change the normal modes in a manner dependent on the particular value of r_1. When any one or more of a_4, a_6, etc., are non-zero, the

* This was previously proven for a sequence of ideal lenses independently by H. E. Rowe and J. P. Gordon.

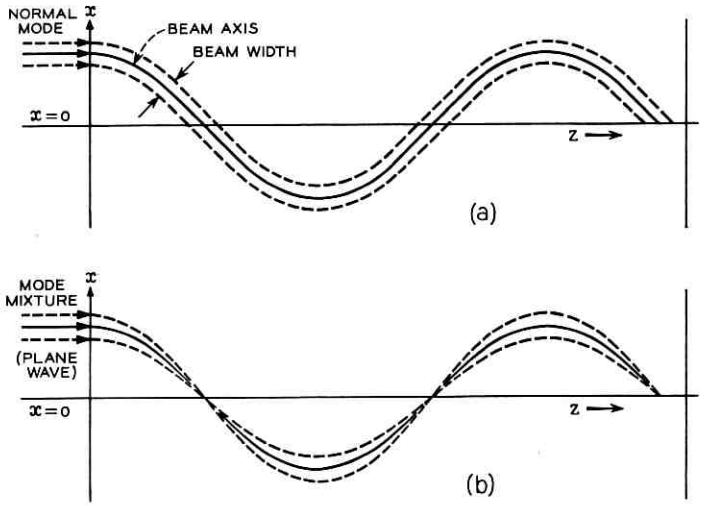


Fig. 6 — Ray path representation of beam flow in the ideal square law-medium, (a)—normal mode off-axis input, (b)—mode mixture off-axis input.

modes for a region off the axis of the medium are different from those on axis and the rather simple and attractive situation sketched in Fig. 6 no longer exists. It also follows that ray paths also differ importantly from those sketched in Figs. 1(b) and 1(c).

IV. RAY PATHS IN NON-SQUARE-LAW MEDIA

When $f(x)$ of (35) and (36) contain non-zero a_4 or higher order terms, the general shape of $f(x)$ can be something like that of Fig. 5, even though a_4 or some other terms are negative yielding a defocusing tendency. We shall be concerned now with arbitrary values of a_2, a_4, a_6, etc., within the limits on x such that

$$\begin{aligned} \frac{df(x)}{dx} \quad &< 0 \quad \text{for } x > 0 \\ &> 0 \quad \text{for } x < 0 \end{aligned} . \tag{40}$$

This assures that all rays will be bent toward the axis at any x, per (3).

We can observe several additional features of the general case:

(1.) As a consequence of (40), any ray will monotonically go through a maximum and return to the axis, as sketched in Fig. 7. Due to symmetry about the z axis and reciprocity, the curve $x = g(z)$ will have even symmetry about z_2, z_4, etc. and odd symmetry about z_3, z_5, etc. The period of $g(z)$, $(z_5 - z_1)$, will depend on the coefficients a_2, a_4, $\cdots$ in (36).

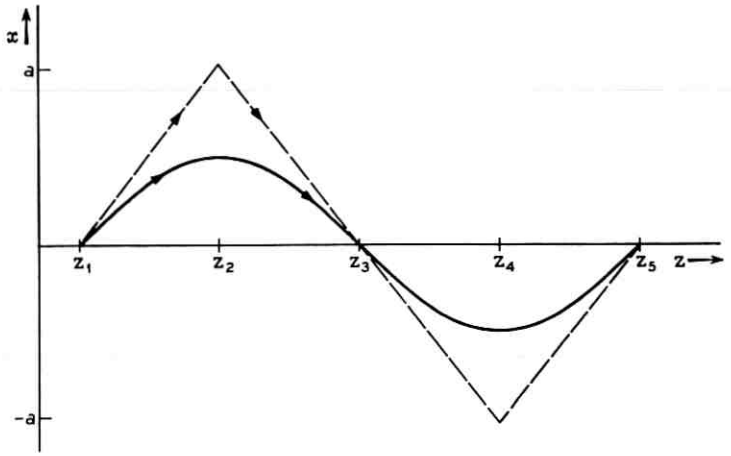


Fig. 7 — Ray paths in a generalized lens-like medium.

(2.) It will be instructive to note that

$$f(x) = -\left|\frac{x}{a}\right|^m \tag{41}$$

with $m \to \infty$ is the special case of (36) which corresponds to a step change in index at $|x| = a$:*

$$\begin{aligned} f(x) &\to 0 \quad \text{and} \quad n = n_a \quad \text{for } |x| < |a| \\ -f(x) &\gg 0 \quad \text{and} \quad (n-1) \ll (n_a - 1) \quad \text{for } |x| \to |a|. \end{aligned} \tag{42}$$

We will see that this is a convenient bounding condition on non-ideal focusing media. In Fig. 7, the dotted line represents the ray path for such a medium.

(3.) As already noted, the paraxial ray equation (3) is valid. Since $f(x)$ in (36) is to be kept very small compared to unity, the solution for the position of the ray as a function of z, $x = g(z)$, is related

$$\frac{d^2}{dz^2}[g(z)] = \frac{1}{n}\frac{\partial n}{\partial x} = \frac{d}{dx}[f(x)] \tag{43}$$

or

$$f(x) = \int \frac{d^2}{dz^2}[g(z)]\, dx. \tag{44}$$

* We will want to continue to use the paraxial ray equation and the associated condition, $f(x) \ll 1$. We shall use the region $0 < |x| < a$ but allow $|x|$ to approach a so closely that $f(x) \gg 0$.

(4.) Due to symmetry about the z axis,

$$\frac{d^2}{dz^2}[g(z)] = 0 \text{ at values of } z \text{ where } x = g(z) = 0. \tag{45}$$

Obviously,

$$\frac{d}{dz}[g(z)] = 0 \tag{46}$$

periodically at maxima for $g(z)$, which will occur at values of z midway between the places where $g(z) = 0$, Fig. 7.

(5.) It follows from the preceding notes that the function $g(z)$ must have even symmetry about the values of z for which (46) holds, and odd symmetry about the values of z for which $g(z) = 0$, a consequence of reciprocity and the known symmetry about $x = 0$ for $f(x)$. Hence, the most general ray trajectory will be of the form

$$x = \sum_{1}^{\infty} b_m \cos m\beta z \tag{47}$$

where $m = 1, 3, 5, 7, \cdots$ etc. We will proceed presently to find a particular solution of the form (47).

In all functions (35) where a_4 or some higher a_n is present, the period of the ray trajectory depends on the peak amplitude. That is, if the medium characterized by (35) starts at $z = 0$ (as in Fig. 8) and if a series of rays enter parallel to the z axis but at different values of x, the ray trajectories will have different periods. If a_2, a_4, etc. are all positive, the rays farther from the z axis will be bent more sharply and the picture qualitatively will be as sketched in Fig. 8. However, if a_2

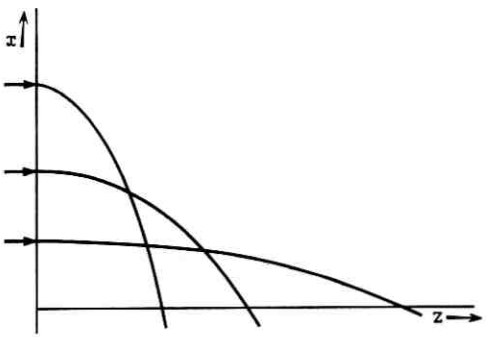


Fig. 8 — Ray paths in a generalized focusing medium for various parallel input rays.

is positive and a_4 negative (other $a_n = 0$) the rays entering farther from the z axis will have a period *longer* than those entering near the z axis. As previously noted, all rays have the same period (Fig. 1) only when all a_n except a_2 are zero.

V. FOCAL LENGTH OF A SEGMENT OF AN ARBITRARY MEDIUM

Suppose a ray at radius x passes through a section of arbitrary medium (Fig. 1a) of length t. We can ascribe a focal length to each x position as follows: the phase difference $\Delta\varphi$ between a ray *on*-axis and a ray at x is given by

$$\Delta\varphi = \frac{2\pi}{\lambda_0} n_a t\{\tfrac{1}{2}a_2x^2 + \tfrac{1}{2}a_4x^4 \cdots\}. \tag{48}$$

We require t in (48) to be so small that

$$\Delta\varphi \ll \frac{\pi x}{\lambda} \tag{49}$$

and we represent the segment of focusing medium as a thin lens; for any thin lens the focal length f is

$$f = \frac{\pi x^2}{\lambda_0 \Delta\varphi}. \tag{50}$$

Then, combining (50) and (48),

$$f = \frac{1}{n_a t\{a_2 + a_4x^2 + a_6x^4 \cdots\}}. \tag{51}$$

If $a_2 = 0$, the focal length approaches infinity as $|\, x \,| \to 0$.

If the medium is represented mainly by the square law term a_2 with a small a_4 perturbation

$$f = \frac{1}{n_a t a_2(1 + R)} \tag{52}$$

where

$$R = \frac{a_4x^2}{a_2}. \tag{53}$$

The ratio (53) appears repeatedly in describing slightly non-ideal media.

VI. NORMAL MODE PROPERTIES IN ARBITRARY FOCUSING MEDIA

It would be nice to have solutions to Maxwell's equations for a medium according to (35) and (36) but this has not yet been achieved. E. A.

Marcatili has solved Maxwell's equations and found normal mode field properties for the ideal square-law medium (only a_2 present),[7] and for a perturbation thereon (small a_4 in a medium principally characterized by a_2).[8]

In this section, an approximate method for obtaining some of the properties of the modes of an arbitrary medium (pure fourth order, for example) is described. This is of interest as guidance on whether or not to put the effort into getting better solutions.

We use as limiting cases the solutions found by Marcatili when only a_2 is present, and the known solutions for a step transition in dielectric constant, which as previously noted in connection with (42) corresponds to a very high exponent on the x^n term of (36). These two cases can be viewed as limits to the possible solutions for arbitrary (36), within the bounds of (40), and useful results inferred about the intervening cases.

We note that the transverse field distribution for the first-order mode with the step-change of index, Curve I of Fig. 9, is of the form

$$E = \cos\left(\frac{\pi x}{2a}\right) \tag{54}$$

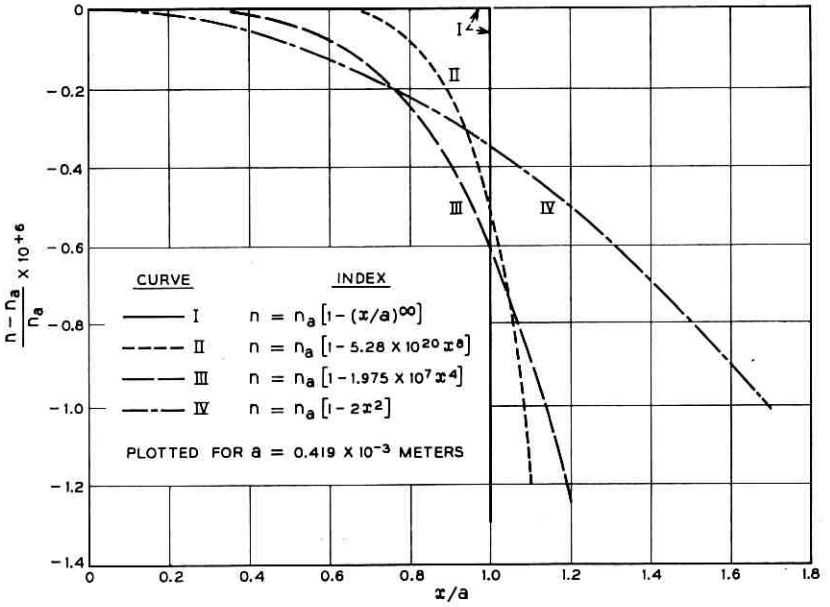


Fig. 9 — Index of refraction versus normalized transverse position (x/a) for several media.

where the index change occurs at $x = \pm a$. From Marcatili's wave solution,[7] the transverse field shape for the lowest order mode in a medium characterized by

$$n = n_a (1 - \tfrac{1}{2} a_2 x^2) \tag{55}$$

is given by

$$E = \exp\left[-\left(\frac{x}{\bar{x}}\right)^2\right] \tag{56}$$

where

$$\begin{aligned} \bar{x} &= \sqrt{\frac{\lambda}{\pi}} \frac{1}{(a_2)^{\frac{1}{4}}} \\ \lambda &= \lambda_0 / n_a \\ \lambda_0 &= \text{free space wavelength.} \end{aligned} \tag{57}$$

We now match the fields given by (54) and (56) by setting (54) equal to (56) at $x = \bar{x}$; this yields

$$a_e = 1.315\, \bar{x} \tag{58}$$

where a_e is an "equivalent" or effective half-width for the square-law medium. The corresponding field shapes and plots of index variation are given by Curves I and IV in Figs. 9 and 10.

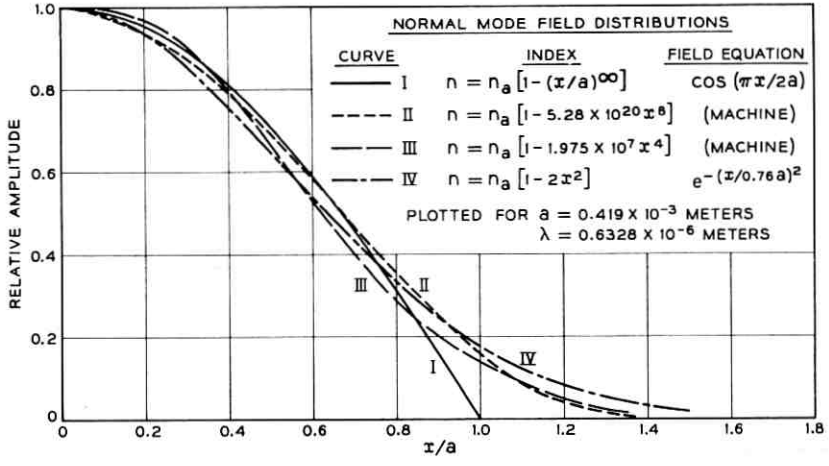


Fig. 10 — Normal mode amplitude vs normalized transverse dimension (x/a) for media of Fig. 9.

As already indicated, these two cases are limits on the index variation. It is inferred that an arbitrary index distribution (35) will have a field distribution similar to that of Curves I and IV of Fig. 10 when x is related to the effective width a_e to yield an index variation approximating Curves I and IV of Fig. 9. This idea has been carried through to specify a procedure for defining a_e for arbitrary media, and to yield particular solutions for several specific index variations. Approximate phase constants for the various modes of arbitrary media are also found.

To document the inference on normal mode field shape for the non-ideal medium, two specific cases were considered in a little more detail. These were represented by Curves II and III of Figs. 9 and 10. Pure fourth order and pure eighth-order index variations were taken, the constant a_n in (36) being selected to give an "eye-ball" fit to Curves I and IV of Fig. 9; the actual numbers for $(a_n/2)$ are on Fig. 9. Then, using the computer-simulation of a resonator as first employed by Fox and Li,[9] the normal mode field distributions were computed. More will be said about these calculations presently, but for now we note the conclusion on the normal mode field distributions for the eighth-order and fourth-order index variations, Curves II and III of Fig. 10. The choice of a_n for these curves was a first guess, not an optimized choice. Nevertheless, the fields do correspond very well to the limiting cases, Curves I and IV. We shall see that differences between Curves I, II, III, and IV in the region $x < 0.9a$ are comparable to the differences between the true fields in the system of ideal lenses and the Gaussian approximations commonly used to represent them.

The effective medium width a_e can be defined for a more general non-ideal medium as follows: the normal mode in the medium characterized by (35), (36) and (40) will have a shape similar to that of the associated square-law medium or step-change medium when a_e is defined by

$$f(x = a_e) = -r \frac{(1.315)^4}{2\pi^2} b^4 \left(\frac{\lambda}{a_e}\right)^2. \tag{59}$$

This comes from making $f(x)$ of (35) for the arbitrary medium equal to r times the same quantity for an equivalent square-law medium, both at $x = a_e$, and then using (57) and (58) to eliminate the a_2 of the equivalent square-law medium. The constant b is unity for the lowest order mode, and is shown in Appendix I to be

$$b = \sqrt{\frac{m + 2.5}{2.5}} \tag{60}$$

for the TEM_m mode.* The constant r varies slightly depending on the exact form of $f(x)$; in the examples plotted in Fig. 9, r is 1.72 for the fourth order case and 1.43 for the eighth order case. When considering a medium which is very nearly square law, r may be taken as unity, and for fourth or higher order media a value $r = 1.5$ is a good value.

Following known theory for metallic waveguides, the normal mode for the step-change index can be represented by two plane waves travelling at an angle α to the axis of the medium, where

$$\sin \alpha = \frac{\lambda}{\lambda_c} \tag{61}$$

and λ_c is the cutoff wavelength in the guide containing everywhere the index n_a. Restricting our interest to the region $\lambda/\lambda_c \ll 1$, the phase constant is given by

$$\beta = \frac{2\pi}{\lambda}\sqrt{1 - \sin^2 \alpha} \cong \frac{2\pi}{\lambda}\{1 - \tfrac{1}{2}\alpha^2\}. \tag{62}$$

Higher modes are accounted for by noting

$$\lambda_c = \frac{4a}{(m+1)} \tag{63}$$

giving

$$\begin{aligned} \beta &= \frac{2\pi}{\lambda}\left\{1 - \frac{1}{32}\left(\frac{\lambda}{a}\right)^2 (m+1)^2\right\} \\ \lambda &= \lambda_0/n_a . \end{aligned} \tag{64}$$

Since the principal part of the field distribution has been shown to be essentially the same for all $f(x)$ restricted by (40) provided a_e is defined as in (59), we can expect that the expression (64) will give a good approximation for the phase constant in a medium characterized by any $f(x)$.

We proceed to write down the expressions which result.

For the medium described by

$$n = n_a (1 - \tfrac{1}{2}a_2 x^2 - \tfrac{1}{2}a_4 x^4) \tag{65}$$

the application of (59) leads to

$$a_e = 1.315b\left[\frac{r}{a_2(1+R)}\right]^{\frac{1}{4}}\sqrt{\frac{\lambda}{\pi}} \tag{66}$$

* Following Fox and Li and Boyd and Gordon[10] the mth order mode has $(m+1)$ field maxima in the transverse cross section.

with $R = a_4 a_e^2/a_2$. The ratio R is the ratio of the fourth order term $0.5\ a_4x^4$ to the second order term $0.5\ a_2x^2$ at $x = a_e$, the equivalent half-width of the medium. Thus the expression (66) gives a_e implicitly.

When $R \ll 1$, we can rewrite $(1 + R)^{\frac{1}{2}} = 1 + \frac{1}{2} R$; under this condition, putting (66) into (64) gives for the medium characterized principally by a square-law distribution with small fourth-order variation, (taking $r = 1$),

$$\beta = \frac{2\pi}{\lambda} - 0.357 \sqrt{a_2} \left\{ \frac{(m+1)^2}{\left(\frac{m+2.5}{2.5}\right)} + \frac{0.275(m+1)^2 a_4 \lambda}{a_2^{3/2}} \right\}. \quad (67)$$

We can compare this result to Marcatili's direct wave solution[7] in the limit of $a_4 = 0$. The functional dependence on λ and a_2 is identical; the coefficient of $\sqrt{a_2}$ in (67) compares to the correct one as follows:

	Coefficient of $\sqrt{a_2}$	
m	Marcatili[7]	Eq. (67)
0	0.5	0.357
1	1.5	1.02
2	2.5	1.78
3	3.5	2.6
4	4.5	3.43
$m \to \infty$	m	$0.892\ m$

Thus, the approximate value of the phase constant is correctly given for all modes at any wavelength by the approximate theory outlined above, but the unique integral relationship for the phase differences between the modes in the square-law medium is not given.

A comparison between (67) and some unpublished results Marcatili obtained using a direct perturbation solution of Maxwell's equations shows identical dependence upon a_4, a_2 and λ with comparable but somewhat different constants.

The above comparisons lead one to have confidence in other results obtained from (59) and (64) for which there is no previous information. For example, we can get the phase constant corresponding to medium (65) with $a_2 = 0$. Equation (66) leads to

$$\begin{aligned} a_e &= \frac{r^{1/6}(1.315)^{2/3}}{\pi^{1/3}} \frac{b^{2/3} \lambda^{1/3}}{a_4^{1/6}} \\ &= 0.876 \left(\frac{m+2.5}{2.5}\right)^{1/3} \frac{\lambda^{1/3}}{a_4^{1/6}} \end{aligned} \quad (68)$$

for the equivalent half-width of the medium, and

$$\beta = \frac{2\pi}{\lambda} - 0.256(a_4\lambda)^{1/3} \frac{(m+1)^2}{\left(\frac{m+2.5}{2.5}\right)^{2/3}} \tag{69}$$

for the phase constant.

Likewise, for a medium

$$n = n_a(1 - \tfrac{1}{2}a_8x^8) \tag{70}$$

the application of (59) and (64) gives

$$\begin{aligned} a_e &= \left(\frac{r}{a_8}\right)^{0.1} \frac{(1.315)^{0.4}}{\pi^{0.2}} b^{0.4}\lambda^{0.2} \\ &= 0.925 \frac{\lambda^{0.2}}{a_8{}^{0.1}} \left(\frac{m+2.5}{2.5}\right)^{0.2} \end{aligned} \tag{71}$$

and

$$\beta = \frac{2\pi}{\lambda} - 0.323\lambda^{0.6}a_8{}^{0.2} \frac{(m+1)^2}{\left(\frac{m+2.5}{2.5}\right)^{0.4}}. \tag{72}$$

For the limiting case of a step change in index

$$\begin{aligned} a_e &= a \\ \beta &= \frac{2\pi}{\lambda} - \frac{\pi}{16}\frac{\lambda}{a^2}(m+1)^2. \end{aligned} \tag{73}$$

Looking at the change in β as the exponent in variation of index of refraction changes from 2 to 4 to 8 to ∞, (see (67) with $a_4 = 0$, (69), (72) and (73)), we observe that the β dependence on λ goes smoothly from λ^0 to $\lambda^{\frac{1}{2}}$, $\lambda^{0.6}$, and $\lambda^{1.0}$. It seems certain that the square-law medium is unique in having beats between modes being independent of λ.

There is a physical explanation for the increasing β dependence on λ for increasing exponent in the index variation. For the step change in refractive index the modes are all contained in the same transverse space, illustrated by Curves I of Figs. 11 and 12; by definition no energy can exist at $x > a$. However, with the square-law variation in refractive index, Curve IV of Fig. 9, energy can and does exist at $x > a_e$ and the penetration of the field at $x > a_e$ increases as the mode index m increases. This is illustrated for $m = 0$ and $m = 2$ as Curves II in Figs. 11 and 12 respectively. Since the higher order modes occupy more transverse width for increasing m, λ_c decreases more slowly for

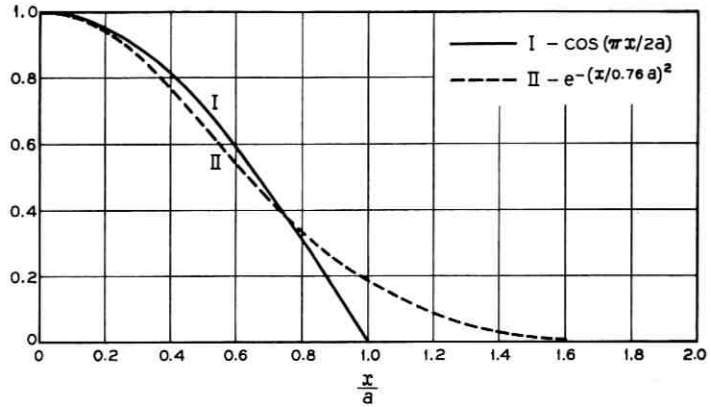


Fig. 11 — Normal mode amplitude versus normalized transverse dimension (x/a) for step change refractive index and square-law refractive index.

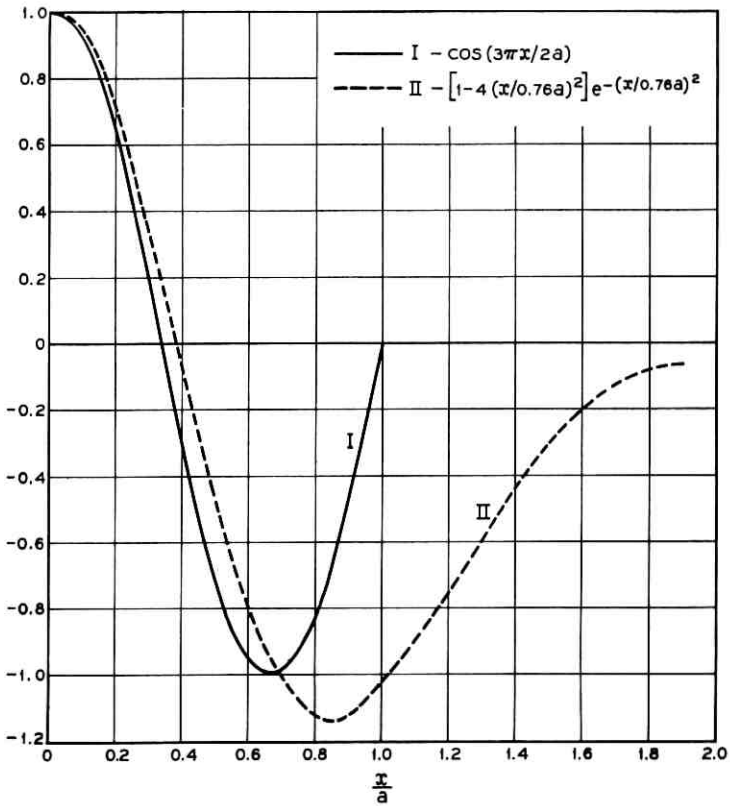


Fig. 12 — Normal mode amplitude versus normalized transverse dimension (x/a) for a higher-order mode in the step change index and square-law index media.

the square-law medium than for the step-change refractive index. That is the basic reason for the factor b, (60) appearing in (59). Similarly, the increased field penetration at $x > a_e$ for the square law compared to the step-change index makes the angle α smaller for the square-law medium than in the step-change medium, and consequently the β is less altered from the value $2\pi/\lambda$. As the exponent in the terms of $f(x)$ increases from 2 toward ∞ the behavior in β and λ_c approach those for the step-change medium because the more rapid variation in index at $x > a_e$ prevents the field from penetrating that region as much.

One can write explicitly an expression for the half-width w_e of the lowest order mode's field for the various non-square-law media using the relation

$$w_e = \frac{a_e}{1.315} \tag{74}$$

where a_e is again defined by (59). As a consequence of the method of defining a_e, (74) gives the exact spot size for the pure square-law medium. For the medium which is principally square law with a small fourth-order term, (65) with $R < 1$,

$$w_e = \sqrt{\frac{\lambda}{\pi}} \frac{1}{a_2^{1/4}} \left\{1 - 0.55 \frac{\lambda a_4}{a_2^{3/2}}\right\}. \tag{75}$$

For the pure fourth-order medium,

$$w_e = 0.666 \frac{\lambda^{1/3}}{a_4^{1/6}} \tag{76}$$

and for the pure eighth-order medium,

$$w_e = 0.703 \frac{\lambda^{0.2}}{a_8^{0.1}}. \tag{77}$$

It is possible now to specify unique rays to consider characteristic of the particular mode in each of the media. These rays are the normals to the two plane waves which in combination give approximately the transverse field distributions in the manner outlined above. Each of these normals makes an angle α to the axis of the medium as in equation (61). Again confining our interest to waves far from "cut-off"

$$\alpha \cong \frac{\lambda(m+1)}{4a_e} \tag{78}$$

where a_e is again defined by (59). For the medium (65) with $R < 1$

$$a_e \cong \frac{1.315}{a_2^{1/4}} \sqrt{\frac{m + 2.5}{2.5}} \sqrt{\frac{\lambda}{\pi}} \left\{1 - 0.55 \left[\frac{m + 2.5}{2.5}\right] \frac{\lambda a_4}{a_2^{3/2}}\right\} \quad (79)$$

and

$$\alpha = \frac{\sqrt{\lambda}\, a_2^{1/4}}{2.97} \frac{(m + 1)}{\sqrt{\frac{m + 2.5}{2.5}} \left\{1 - 0.55 \frac{\lambda a_4}{a_2^{3/2}} \left(\frac{m + 2.5}{2.5}\right)\right\}}. \quad (80)$$

This expression also gives the characteristic ray angles for the square law medium by letting $a_4 = 0$.

For the fourth-order medium, (68) gives a_e and the characteristic ray angle is

$$\alpha = 0.285 \lambda^{2/3} a_4^{1/6} \frac{(m + 1)}{\left[\frac{m + 2.5}{2.5}\right]^{1/3}}. \quad (81)$$

For the eighth-order medium, (71) gives a_e and the characteristic ray angle is

$$\alpha = 0.271 \lambda^{0.8} a_8^{0.1} \frac{(m + 1)}{\left[\frac{m + 2.5}{2.5}\right]^{0.2}}. \quad (82)$$

Having a characteristic half-width a_e and ray angle α for each medium, we can easily calculate a characteristic ray period for each medium. With reference to Fig. 13, the ray at an angle α with the z axis has a period λ_p

$$\lambda_p = \frac{4a_e}{\alpha}. \quad (83)$$

For the pure square-law medium, using (66) and (80)

$$\lambda_p = \frac{4 \times 1.315 \times 2.97}{\sqrt{\pi}} \frac{1}{a_2^{\frac{1}{2}}} \left[\frac{m + 2.5}{2.5}\right] \frac{1}{(m + 1)} \quad (84)$$

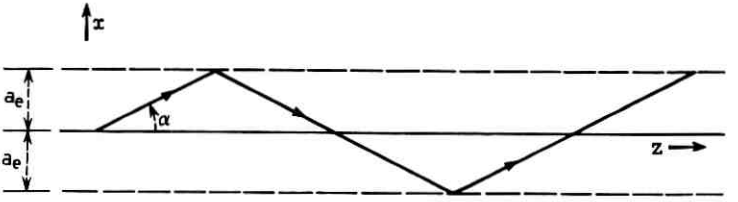


Fig. 13 — Diagram for determining an equivalent ray period in non-square-law media.

or

$$\lambda_p = \frac{8.8}{a_2^{\frac{1}{2}}} \frac{(m + 2.5)}{2.5(m + 1)}. \tag{84a}$$

We can compare this ray period to that derived from the paraxial ray equation; that solution is (10) from which the ray period is seen to be $2\pi/\sqrt{a_2}$. Equation (84a) corresponds quite well to $2\pi/\sqrt{a_2}$ for the lowest order mode ($m = 0$) including independence of λ! Equation (84a) is also relatively independent of mode order, m.

For the pure fourth-order medium the ray period from (83) is

$$\lambda_p = \frac{12.3}{\lambda^{1/3} a_4^{1/3}} \left[\frac{m + 2.5}{2.5}\right]^{2/3} \frac{1}{(m + 1)}. \tag{85}$$

There is no previous theory for comparison and, as already noted, ray theory alone does not define a unique period; unlike the square-law case, it depends on the initial ray slope.

VII. FIELD DISTRIBUTION AND DIFFRACTION LOSSES IN RESONATORS AND LENS WAVEGUIDES USING NON-SQUARE-LAW ELEMENTS

In this section there are recorded new computations of normal mode field distributions and diffraction losses for a few resonators using non-spherical mirrors. The results are applicable to lens waveguides using non-square-law focusing elements. No other information is known to be available on these configurations.

The motivation for the work was to determine the normal-mode field distributions for non-square-law continuous waveguides to support the approximations made in the preceding section. Thus, most of the conditions selected represent relatively close spacing of weak lenses. A few instances are cited where cutoff effects associated with too great spacing of lenses were observed.

The resonator whose descriptive dimensions are given in Fig. 14 is to be used to represent the lens guidance system of Fig. 15. The equivalence is exact if one includes the absorbing screens in Fig. 15; if the diffraction losses in the resonator are not large very little effect is expected from omitting the absorbing screens. Fig. 15, in turn, can represent a continuous guidance medium when the focal length of the lenses is appreciably greater than the lens spacing. Then the beam size is very nearly the same at the lens and at the plane midway between lenses, and the electrically thin lenses of Fig. 15 can be representations of segments of the continuous medium of length s. For the square-law medium, this approximation is that the focal length expression (16) is

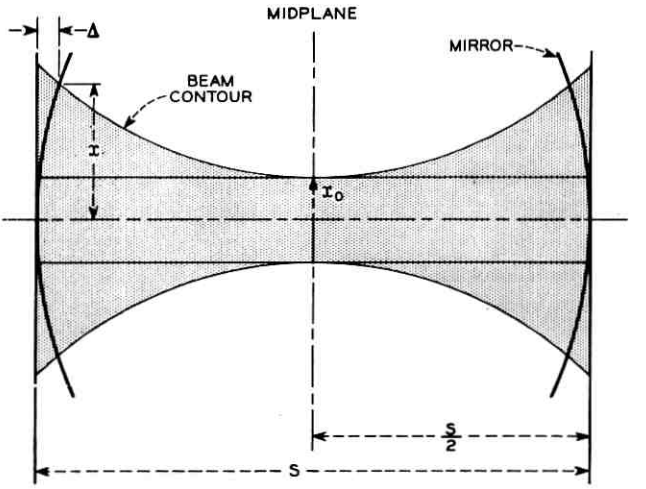


Fig. 14 — Definition of parameters in a resonator for use in Fox-Li type calculations.

well represented by

$$f = \frac{n_0}{n_a a_2 l} \tag{86}$$

when $\sqrt{a_2}\, t \ll 1$.

In representing a continuous medium by a resonator it is important to have a measure of the accuracy of the approximation. We obtain this and observe an interesting outcome by comparing the phase shifts along equi-power contours (i.e., contours within which the fraction of the total power in the beam is constant) for the resonator with spherical

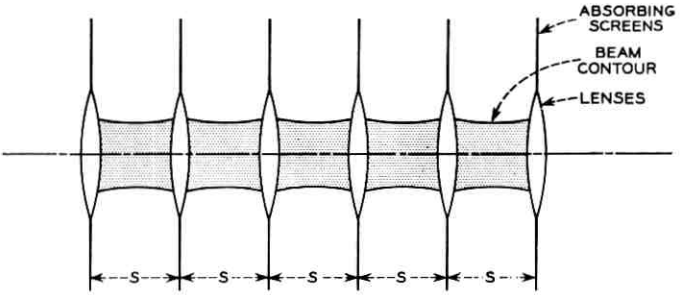


Fig. 15 — Transmission medium analogous to Fig. 14.

mirrors and for the square-law continuous medium,

$$n = n_a\ (1 - \tfrac{1}{2}a_2x^2). \tag{87}$$

For the square-law medium of length s, the phase difference for a ray following the axis and one at radius x_0 is

$$\Delta\varphi_c = \frac{2\pi}{\lambda_0}\, s(-\tfrac{1}{2}a_2x_0^{\,2}). \tag{88}$$

For the resonator with spherical mirrors the phase difference between a ray on axis and one following an equi-power contour is

$$\Delta\varphi_m \cong \frac{2\pi}{\lambda_0}\,(-2\Delta). \tag{89}$$

By making

$$\Delta = \frac{a_2 s}{4}\,x^2 \tag{90}$$

it is apparent that (88) and (89) are identical when $x = x_0$. It may be shown that the beam spot size at the center of the resonator is the same as the spot size of the continuous medium when

$$w_0 = \sqrt{\frac{\lambda}{\pi}\,\frac{1}{a_2^{\frac{1}{2}}}} = \sqrt{\frac{R_c\lambda}{2\pi}} \tag{91}$$

where R_c is the field curvature at the confocal spacing of mirrors associated with the mid-plane spot size w_0.* The equi-power contours are given by

$$\frac{x}{x_0} = \frac{w_s}{w_0} = \sqrt{1 + \xi^2} \tag{92}$$

$$\xi = \frac{s}{R_c} \tag{93}$$

w_s = spot size at the mirrors.

Using (91), (92), and (93), the value of *both* (88) and (89) is

$$\Delta\varphi_m = \Delta\varphi_c = \frac{4\pi}{\lambda_0}\,x_0^{\,2}\,\frac{s}{R_c^{\,2}}. \tag{94}$$

Thus by choosing the continuous medium and spherical mirror system to have the same mid-plane spot size, the phase shift along any contour enclosing equal powers is identical, regardless of lens spacing.†

* See Ref. 10.

† Within, of course, the region of stability for the resonator.

For reference it may be noted that the mirror radius of curvature b' and focal length f are related to R_c and s by[10]

$$b' = 2f = \frac{(1 + \xi^2)R_c}{2\xi} \tag{95}$$

$$R_c = \frac{2}{\sqrt{a_2}} = \sqrt{2sb' - s^2}. \tag{96}$$

Useful forms giving the mid-plane spot size w_0 and spot size at the lens w_s for the lens system of Fig. 15 are:

$$w_0 = \sqrt{\frac{\lambda}{2\pi}}\,(4fs - s^2)^{\frac{1}{4}} \tag{97}$$

$$w_0 = \sqrt{\frac{\lambda}{\pi}}\,(fs)^{\frac{1}{4}}\left(1 - \frac{s}{4f}\right)^{\frac{1}{4}} \tag{98}$$

$$w_s = \sqrt{\frac{\lambda}{\pi}}\,\frac{(fs)^{\frac{1}{4}}}{\left(1 - \frac{s}{4f}\right)^{\frac{1}{4}}}. \tag{99}$$

It seems certain from (94) that the normal mode field distribution for a continuous medium will be well represented by the normal mode of a resonator having the proper correspondence between Δ (Fig. 14) and the coefficients of (36), provided that the mid-plane of the resonator field does not differ appreciably from the field at the mirror.

The proper value for Δ of Fig. 14, to represent a general medium (35) and (36), is obtained by taking $x = x_0$ (Fig. 14) and letting $\Delta\varphi_m = \Delta\varphi_c$ as was done in connection with (88) and (89). This yields

$$\Delta = \frac{s}{2}\left(\frac{1}{2}\,a_2x^2 + \frac{1}{2}\,a_4x^4 + \frac{1}{2}\,a_6x^6 \cdots\right). \tag{100}$$

Selection of a mirror spacing suitably small can be done for the square-law medium with the aid of (92). For example, with

$$a_2 = 4 \text{ (meters)}^{-2}$$

$$s = 0.2 \text{ meters}$$

$$w_0 = 0.319 \times 10^{-3} \text{ meters}$$

$$R_c = 1 \text{ meter}$$

one computes $w_s/w_0 = \sqrt{1.04}$. This is based on the Gaussian function approximations for the actual fields. A computer determination of the

actual field distributions* has been carried out using the method of Fox and Li and the results are plotted in Fig. 16 for a Fresnel number $N = 1.38$.† These are true normal mode field distributions for the resonator; some perturbation from the Gaussian shape is caused by the 5.27 per cent power diffraction loss on each reflection but this does *not* account for the fact that the midplane $1/e$ width is approximately 7.5 per cent less than the mirror field instead of 2 per cent less as predicted by (92).

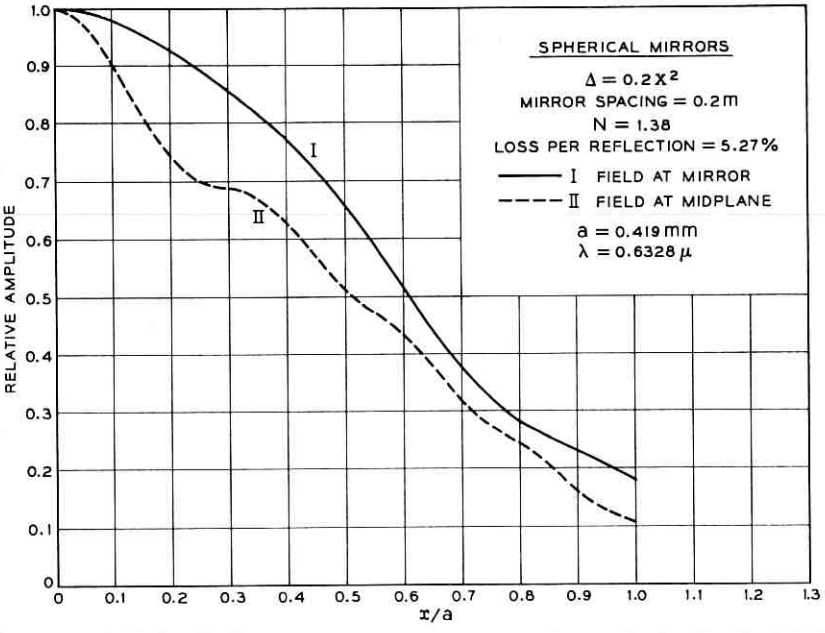


Fig. 16 — Computed normal mode fields at the mirrors and at midplane for a spherical mirror resonator.

See Fig. 17, where the diffraction loss is increased by reducing the mirror size; the mirror spot size changes very little. Also see Fig. 18, where the fields, both the real fields developed on the computer simulation and the Gaussian theoretical approximation thereto, are plotted. We see here, with $N = 1$ where diffraction losses are less than 0.1 per cent per reflection, that the Gaussian approximation is too narrow at the mirrors

* In all of the computer determinations of field distributions and losses, 100 radial intervals were employed in representing the transverse field distribution.

† In the terminology of this paper, $N = r^2/s\lambda$ where $2r$ is the mirror diameter and s is the mirror spacing.

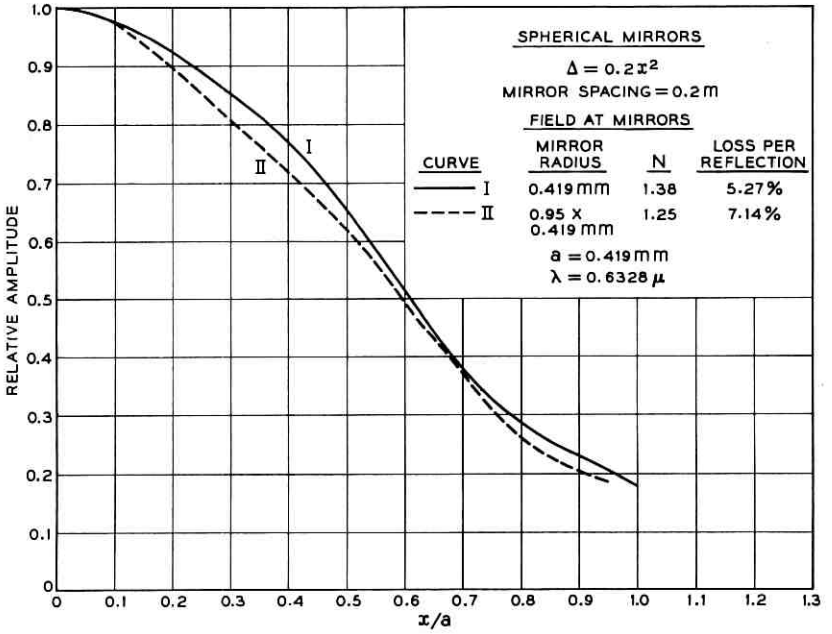


Fig. 17 — Computed normal mode fields at the mirrors for the resonator of Fig. 16 with two mirror diameters.

and too broad at the midplane; whereas $w_s/w_0 = \sqrt{2}$ from the Gaussian approximation, the actual ratio of widths at $1/e$ is about 1.58. The prolate spheroidal wave functions give computed results in much better agreement with the Fox-Li machine computation[10,11] and we presume the latter to be more accurate than the Gaussian approximation.

The transverse fields corresponding to the fourth-order index variation, Curve III of Fig. 9, is shown in Figs. 19 and 20. The mirror spacing, losses and Fresnel numbers are listed on the figures. The midplane field corresponds as well to the mirror field as did that for the square-law case, Fig. 16, and the spacing s is judged adequately short. For the same Fresnel number, $N = 1.38$ and the same spot size, the fourth-order medium had 3.91 per cent loss per reflection compared to 5.27 per cent for the square-law medium. This is because the field at $x = a_e$ falls off more sharply as shown on Fig. 10. The field plotted in Fig. 10 as Curve III was obtained from the computer data which produced Fig. 20, with the reflection number selected to minimize the amplitude of the higher modes still present in the data for Curve I of Fig. 20.

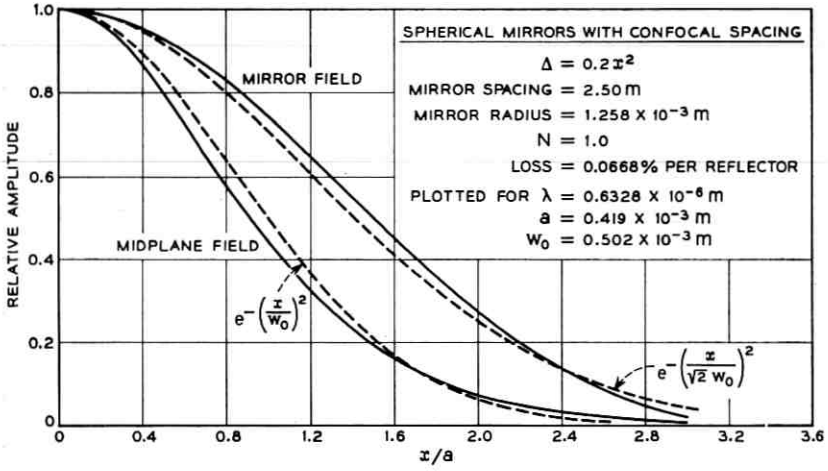


Fig. 18 — Comparison of normal mode fields in a confocal square law resonator as determined by (1.) Fox-Li type calculations and (2.) Gaussian analytical approximation.

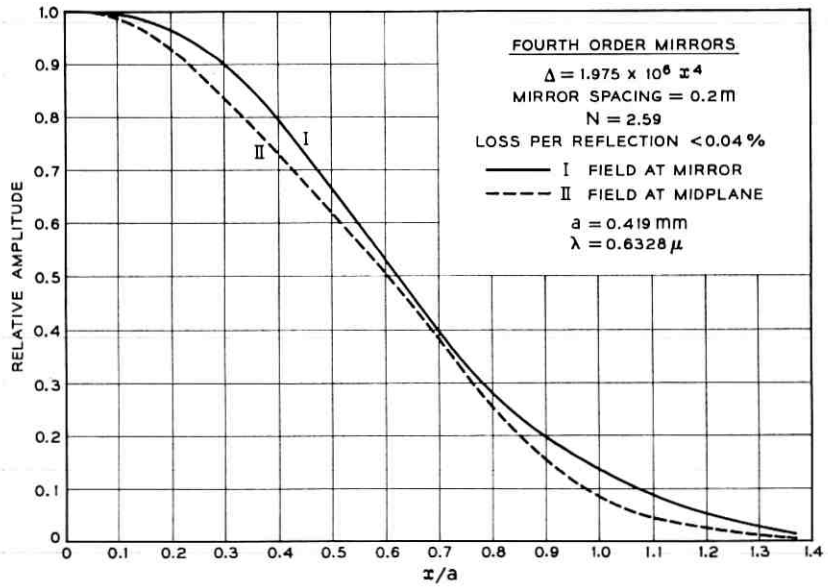


Fig. 19 — Computed normal mode fields at the mirrors and at midplane for a fourth-order mirror resonator.

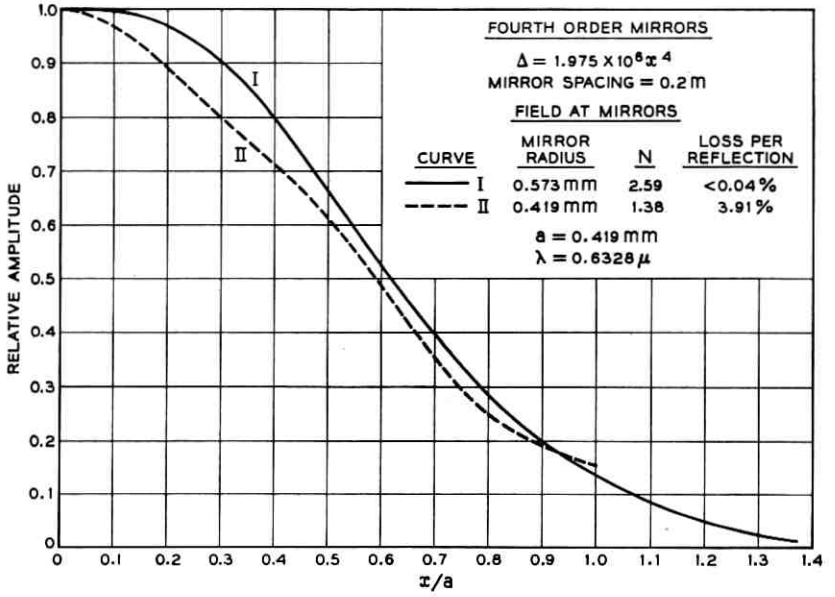


Fig. 20 — Computed normal mode fields at the mirrors for the resonator of Fig. 19 with two mirror diameters.

Figs. 21 and 22 show computed field distributions for a resonator with eighth-order mirrors chosen to match the medium of Curve II, Fig. 9, at the same mirror spacing used for Figs. 16–20, 0.2 meters. The value of Δ is computed from (100). In Fig. 21 we note the midplane field is quite unlike the mirror field, and in Fig. 22 we note the loss does *not* decrease smoothly for increasing mirror size, but remains more or less independent of mirror size. This is due to the fact that the focal length, which is a function of radial position on the mirror according to (51), is too small compared to the mirror spacing at the outer edge of the mirror. For example, for the Curve III of Fig. 22, the focal length is 0.077 meters at the edge of the mirror. For larger mirrors than those represented in Fig. 22, it was found that both the loss and field distribution became constant, independent of mirror size. The fields striking the outer edges of these mirrors are radiated because they are reflected through the axis of the resonator at such a sharp angle as to miss the opposite reflector. Fig. 23 shows a rough sketch of this situation. (Note that the midplane fields resulting would be composed of a mixture of a propagating field and a radiated field.) An analogy can be made to the unstable region which occurs for all rays in a resonator with spherical

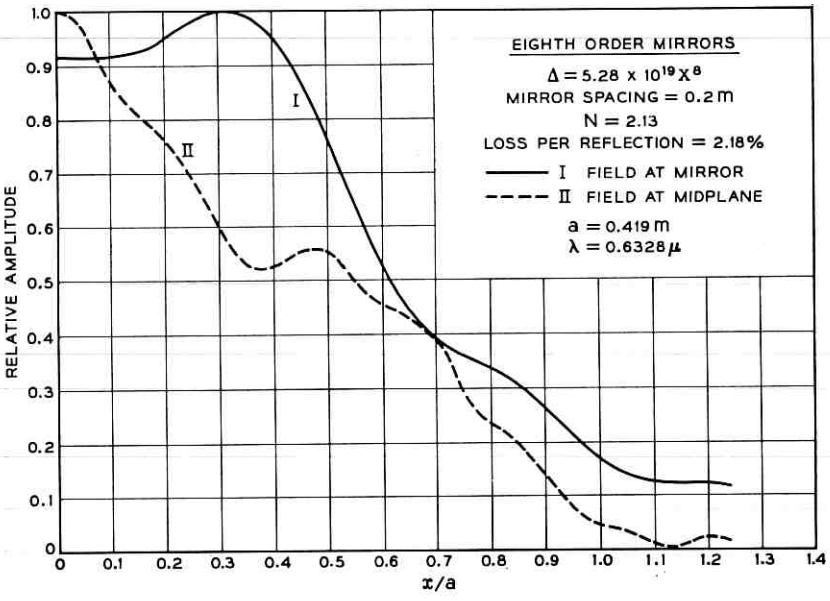


Fig. 21 — Computed normal mode fields at the mirrors and at midplane for an eighth-order mirror resonator.

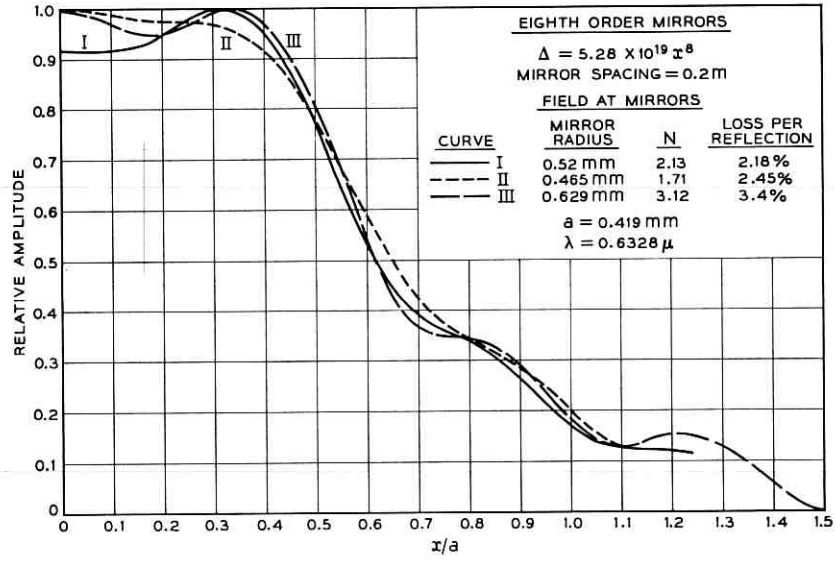


Fig. 22 — Computed normal mode fields at the mirrors for the resonator of Fig. 21 with three mirror diameters.

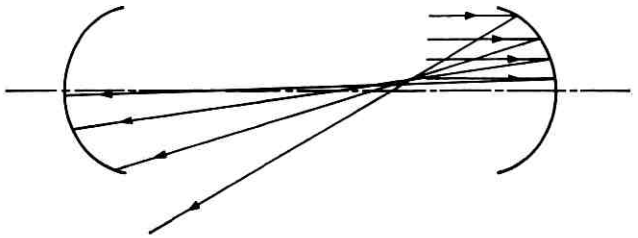


Fig. 23 — Diagram illustrating over-focusing losses in a resonator.

mirrors when $f < s/4$ or $f < 0.05$ meter when $s = 0.2$ meters as in Figs. 21 and 22.

The corresponding situation in a lens guidance system is shown in Fig. 24. The energy near the beam axis is weakly focused leaving a rather large spot size. The energy extending to larger and larger radii is eventually focused too strongly and misses the next focusing element, illustrated by the dotted lines of Fig. 24. Thus, ordinary field spreading, causing the usual diffraction losses, is mixed with over-focusing losses in a certain region of the parameters for a non-square-law resonator. It is believed that this accounts for the loss differences tabulated on Fig. 22.

The objective of determining the normal mode field for Curve II of Fig. 9 can be met by making the mirror spacing in the equivalent resonator smaller. This has been done in Figs. 25 and 26. In Fig. 25, the midplane field and mirror field are shown to agree quite well. In Fig. 26, some variation in field shape results with reduction in mirror size, but this is in part due to higher order modes still present in the field plotted for Curve I where the loss is extremely low. For the normal mode shape plotted as Curve II in Fig. 10, a careful selection of the resonator reflection number was made to minimize the higher-order mode amplitudes.

Using very closely spaced mirrors with so little curvature over the area where most of the beam energy is located, as in Curve II of Fig. 26

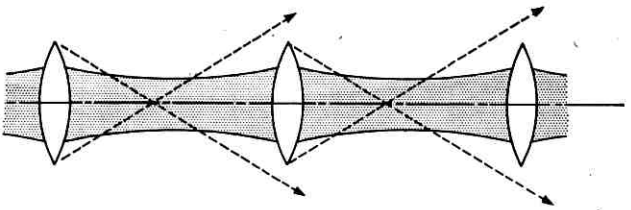


Fig. 24 — Diagram illustrating over-focusing losses in a lens guidance system.

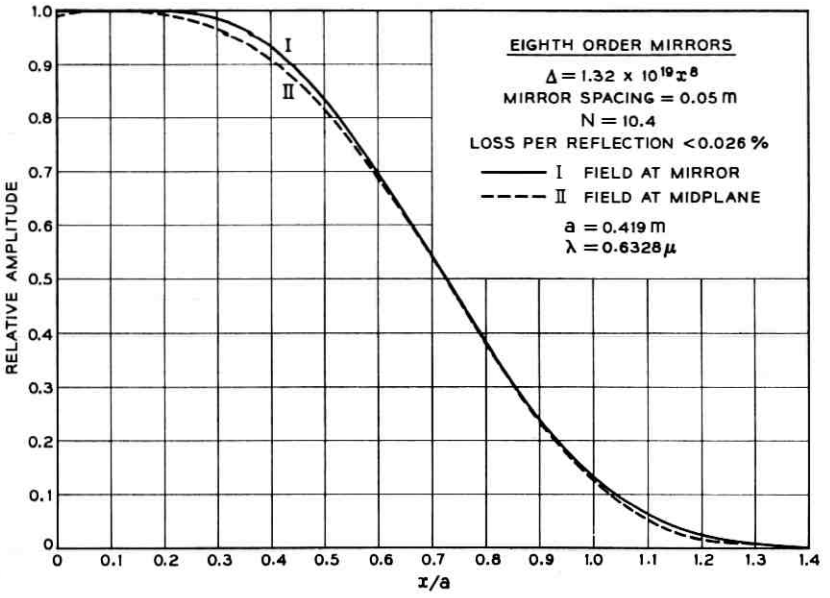


Fig. 25 — Computed normal mode fields at the mirrors and at midplane for an eighth-order mirror resonator representing a shorter segment of the same medium as in Fig. 21.

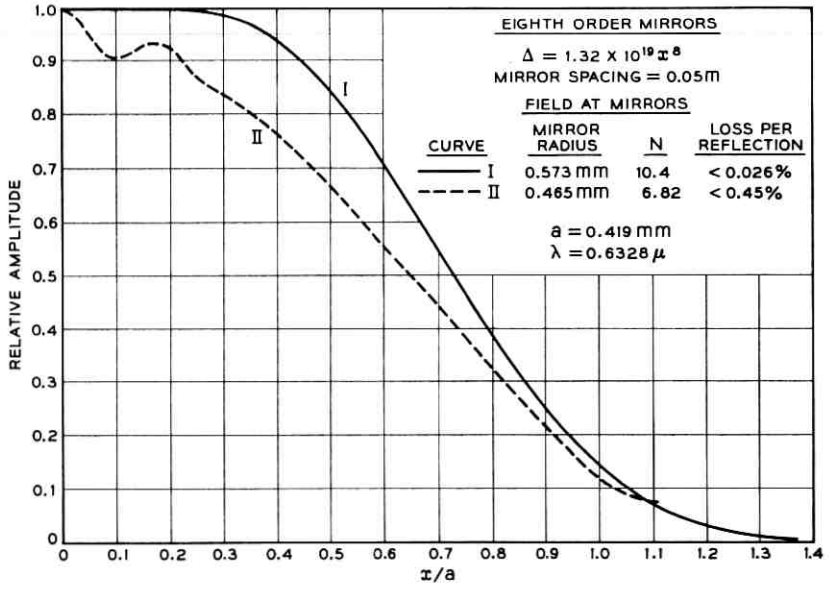


Fig. 26 — Computed normal mode fields at the mirrors for the resonator of Fig. 25 with two mirror diameters.

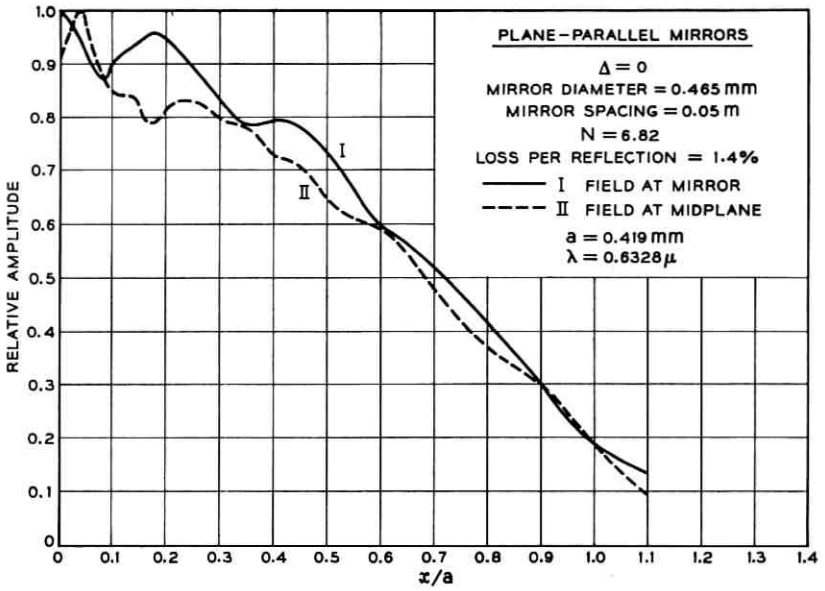


Fig. 27 — Computed normal mode fields at the mirrors and at midplane for a plane parallel mirror resonator analogous to the resonator of Fig. 25.

and the associate medium curvature curve II of Fig. 9, one might wonder whether the guidance is caused by diffraction as with plane parallel mirrors or really by the curvature of the mirrors. Fig. 27 shows the fields and loss for the plane-parallel mirrors corresponding exactly to Curve II of Fig. 26 for the eighth-order mirrors. The loss is <0.45 per cent for the eighth-order mirrors and 1.4 per cent for the plane parallel case. Evidently the guidance is importantly determined by the curvature.

VIII. RAY PATHS IN NON-SQUARE LAW MEDIA

In this section, the general properties of media described by (35), (36) and (40) are exploited to obtain approximate solutions for the paraxial ray equation (3) in several non-square-law media, and to indicate an approach which reduces to straightforward algebra the problems of getting similar solutions for other media.

It has been shown that (47) gives the general form of the ray path for the media of interest, and it was noted in connection with Fig. 7 that the limiting case of a step-shift in index of refraction results in a triangular waveform for the ray path. For a triangular waveform the coefficients b_m in (47) are in the ratio 1, 1/9, 1/25, etc. for $m = 1, 3, 5,$

etc. For the square law medium, a simple $\cos \beta z$ represents the ray path, as given in (7). Hence, we know that for any medium of interest the series (47) will converge very rapidly. We can hope to get a good approximation with only a few terms of (47). We now outline the results of such a solution for a medium described by (65). We set

$$x = c_1 \cos \beta z + c_2 \cos 3\beta z \tag{101}$$

which is equivalent to

$$\left.\begin{aligned} x &= x_0 \left\{ \left(1 + \frac{A}{4}\right) \cos \beta z - \frac{A}{4} \cos 3\beta z \right\} \\ x &= x_0 \{1 + A \sin^2 \beta z\} \cos \beta z \end{aligned}\right\} \tag{102}$$

with

$$A = - \frac{4 \frac{c_2}{c_1}}{\left(1 + \frac{c_2}{c_1}\right)} \tag{103}$$

$$\frac{c_2}{c_1} = - \frac{A}{4 + A} \tag{104}$$

$$c_1 + c_2 = x_0 . \tag{105}$$

In form (102) it is seen that the maximum x is x_0, occurring at $z = 0$, for any value of A; this is a useful form in visualizing the effect of a_4 as a perturbation on a medium mainly controlled by a_2.

Using (101) we find

$$\frac{d^2}{dz^2}(x) = -c_1\beta^2 \cos \beta z - 9c_2\beta^2 \cos 3\beta z. \tag{106}$$

Also, from (65)

$$\frac{d}{dx}[f(x)] = -a_2 x - 2a_4 x^3. \tag{107}$$

We can get the solution to the paraxial ray equation in the form (43) by equating (106) to (107) with x replaced by (101). Equating the coefficients of the $\cos \beta z$ terms yields

$$\beta = \sqrt{a_2}\, \{1 + R(1.5 + 0.376A + 0.1875A^2)\}^{\frac{1}{2}} \tag{108}$$

where

$$R = \frac{a_4 x_0^2}{a_2} . \tag{109}$$

We note that the restriction (40) requires (107) to be greater than zero, or

$$R \geqq -\tfrac{1}{2}.$$

This is a limitation on x_0 if a_4 is negative.

Equating the coefficients of the cos $3\beta z$ terms gives

$$R = \frac{-0.89A}{\left[1.15A + 0.222\left(1 + \frac{A}{4}\right)^3\right]}. \tag{110}$$

Thus, with known R which is fully defined by the medium (a_2 and a_4) and the point of entry (x_0) for the ray, one can compute A and β. The plots of Figs. 28–30 show the interrelations between A, R, and β as given by (108) and (110). For $R \ll 1$, $A \cong -R/4$ and $\beta \cong \sqrt{a_2}\,(1 + 1.5R/2)$. The principal effect of a small a_4 term in the index variation is to change the period of the ray oscillation. This is illustrated in Fig. 31 for several values of R compared to the $R = 0$ (square law) case. Positive R means a_4 has a focusing effect, and the ray period is shortened.

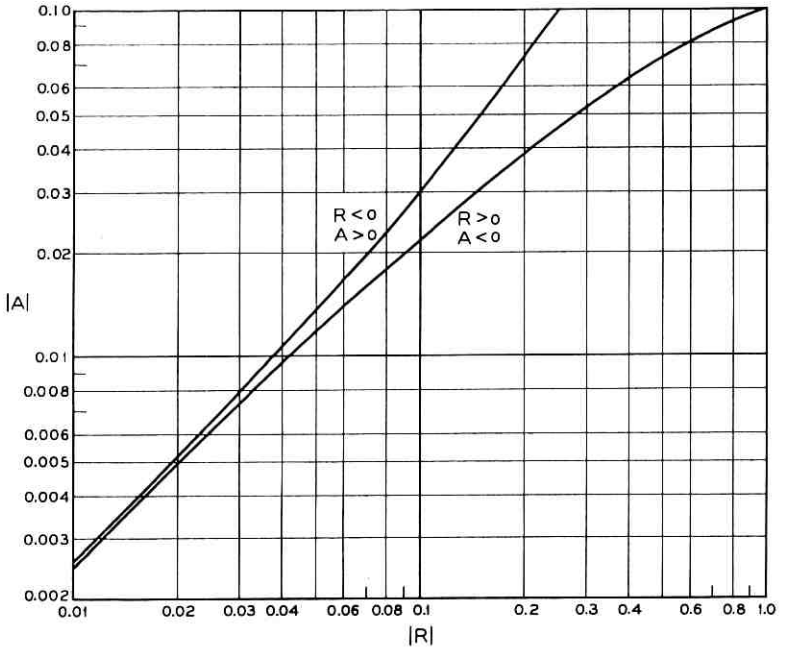


Fig. 28 — R vs A according to (110).

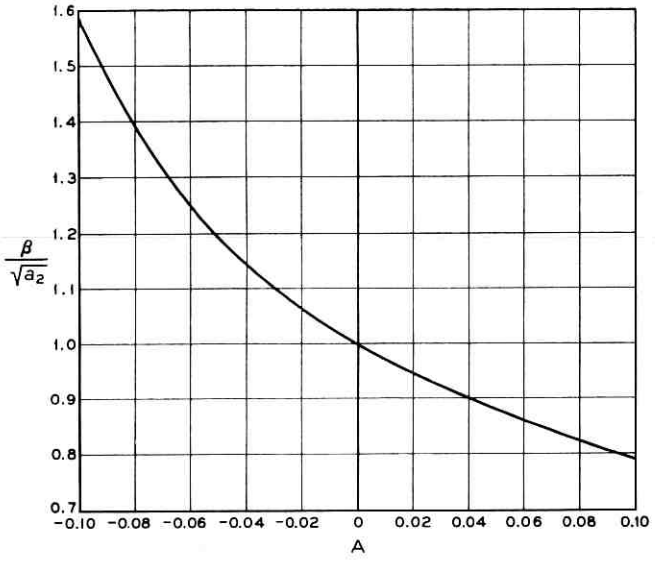


Fig. 29 — Normalized phase constant $\beta/\sqrt{a_2}$ vs A according to (108) and (110).

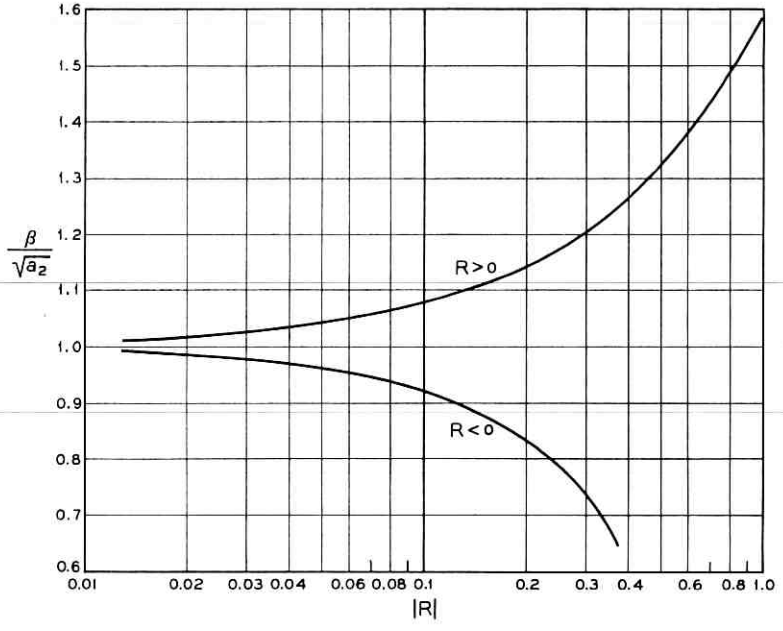


Fig. 30 — Normalized phase constant $\beta/\sqrt{a_2}$ vs R according to (108) and (110).

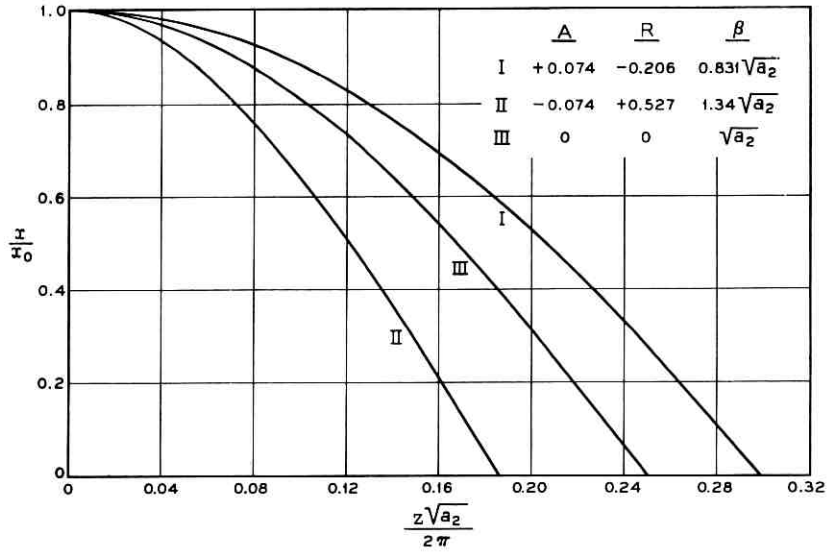


Fig. 31 — Ray position path x/x_0 vs normalized distance for a particular non-square-law medium.

With a renormalization of the abscissa, the curves of Fig. 31 are replotted in Fig. 32 to show that the cos $3\beta z$ term is indeed small and the ray path differs little from a sine wave.

However, the period $2\pi/\beta$ does depend on the peak ray-path amplitude (or ray-path slope at the axis), and hence the nice separation between the input-ray slope and input-ray position which was found in (10) does not exist for non-square law media.

Because the ray period depends on the peak ray-path amplitude, a group of rays entering a non-square law medium at $z = 0$ as in Fig. 33 will fall out of step and at some large z one ray will be at a positive maximum when another is at a negative maximum. These rays can represent parts of a beam of light injected into the medium off-axis when the beam spot size is very large compared to a wavelength. This shows that the injected off-axis beam will spread out and occupy the region $\pm w_t$ about the guide axis. However, the beam will never occupy any more than the region $\pm w_t$ if it is injected parallel to the guide axis.

One can obtain solutions analogous to the one given above if the input ray is on axis but at some slope x by noting

$$x = \Sigma c_n \sin n\beta z \tag{111}$$

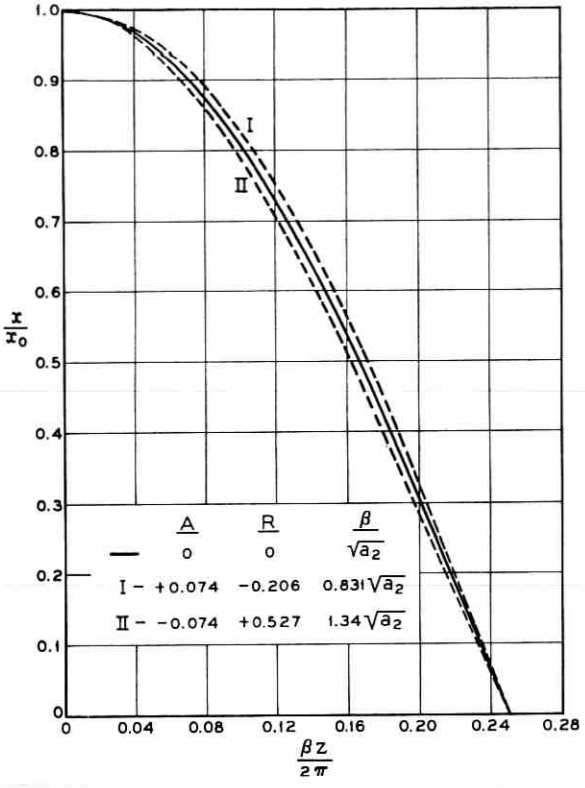


Fig. 32 — Ray position x/x_0 vs $\beta_2/2\pi$ for the medium of Fig. 31 showing basic similarity in shape of ray path.

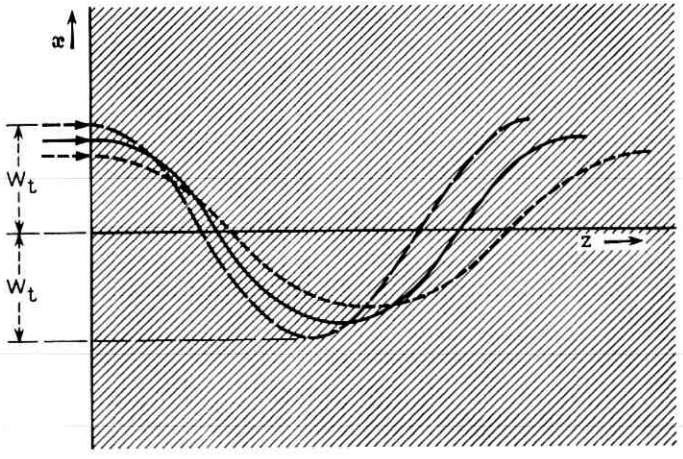


Fig. 33 — Ray position x vs distance z in a non-square-law medium illustrating the way an input beam breaks up as the wave propagates.

$n = 1, 3, 5$, etc. This follows from the same considerations which led to (47). Since we already have solved for the coefficients c_1 and c_2 of (101), we can get the solution for (111) by transforming (101), letting $x_1(z') = x(z - \pi/2\beta)$ which yields

$$x_1 = c_1 \sin \beta z - c_2 \sin 3\beta z. \tag{112}$$

The parameters c_1 and c_2 are related to x_0, the maximum of x_1, as in (105) and we define a new parameter B

$$B = \frac{12\left(\frac{c_2}{c_1}\right)}{1 - 3\left(\frac{c_2}{c_1}\right)} \tag{113}$$

which allows the ray slope to be written

$$\begin{aligned} \frac{dx_1}{dz} &= c_1\beta \cos \beta z - 3c_2\beta \cos 3\beta z \\ &= x_1' \left\{\left(1 + \frac{B}{4}\right) \cos \beta z - \frac{B}{4} \cos 3\beta z\right\} \\ &= x_1'\{1 + B \sin^2 \beta z\} \cos \beta z. \end{aligned} \tag{114}$$

The initial slope x_1' is the ray slope at $z = 0$, and it is related to c_1 and c_2 by

$$x_1' = c_1\beta \left\{1 - \frac{3c_2}{c_1}\right\}. \tag{115}$$

The previous interrelations between c_1, c_2, A, β, and R still hold. We seek a method for getting β and x_0, knowing only x_1, a_2 and a_4. To do so x_1', is rewritten in the form

$$x_1' = \frac{a_2}{\sqrt{a_4}} \sqrt{R}\,(1 + A)\{1 + R(1.5 + 0.376A + 0.1875A^2)\}^{\frac{1}{4}}. \tag{116}$$

Hence, given x_1', (116) determines $x_1' \sqrt{a_4}/a_2$ in terms of R (since R vs A is given in (110)); knowing R, a_2 and a_4, we can compute x_0 and β using (109) and (108). Fig. 34 shows $x_1' \sqrt{a_4}/a_2$ vs R to facilitate this process.

For an input ray with both slope and displacement, the proper matching to (101) with a suitable transformation can in principle be done, but has not been attempted.

An approximate solution for the pure fourth-order medium

$$n = n_a(1 - \tfrac{1}{2}a_4x^4) \tag{117}$$

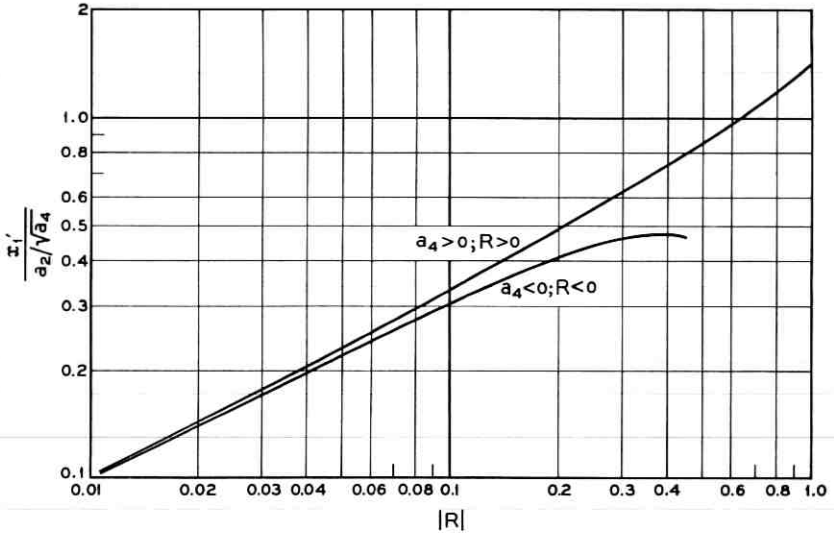


Fig. 34 — Normalized input ray slope x_1' vs R according to (116) and (110).

is obtained by letting $a_2 \to 0$ in the equations which led to (108) and (110). This yields

$$\beta = x_0 \sqrt{1.44 a_4} \tag{118}$$

$$\frac{c_1}{c_2} = 23.3 \tag{119}$$

$$A = -0.165 \tag{120}$$

for the solution corresponding to (101) and

$$x_1' = x_0^2 \sqrt{a_4} \tag{121}$$

for the solution corresponding to (114). Equation (121) gives the ray slope at the axis $x = 0$.

Better approximations would of course be obtained by including higher order terms in (101) but the large c_1/c_2 ratio produced by including only the cos $3\beta z$ terms suggests that the cos $5\beta z$ term would have a negligible coefficient.

We can combine (121) and (118) to express the ray period for the fourth-order medium in terms of the ray slope at the x-axis; this gives

$$\beta = a_4^{\frac{1}{4}} \sqrt{1.44 x_1'}\,. \tag{122}$$

We can now compare the ray period from (122) with that previously derived, (85). We let x_1' equal the characteristic ray angle defined by (81) with $m = 0$. Then the ray period from (122) becomes

$$\frac{2\pi}{\beta} = \frac{9.81}{\lambda^{\frac{3}{4}} a_4^{\frac{1}{4}}}. \tag{123}$$

The constant in (123) differs slightly from the value 12.3 found in (85) but the dependence on λ and a_4 is identical.

IX. DISCUSSION AND ACKNOWLEDGEMENT

The abstract contains a summing up review of the contents of this paper. One might add that a remarkably small step change in index is required to contain completely (for practical purposes) a light beam. As shown in Fig. 9, at $\lambda = 0.6328\ \mu$, a step change in index of refraction of a few parts in 10^6 is adequate for a beam radius of $a = 0.419$ mm, and this change need only be maintained from $x = a$ to $x \cong 2a$ where the energy is certain to be too small for that region to influence the wave propagation.

The author would like to thank Mr. Tingye Li for the use of his computer program which was modified to make the computations represented in Figs. 16–27. Without Mr. Li's previous work the author would not have included those figures which help justify the inferences leading to (59). Mrs. C. L. Beattie made the modifications and very effectively saw through all of the computations, for which the author is most appreciative. Mr. E. A. Marcatili on numerous occasions gave learned reactions to the newly developed ideas.

APPENDIX

We seek here to account for the expression (60)

$$b = \sqrt{\frac{m + 2.5}{2.5}}. \tag{124}$$

As noted in connection with Figs. 11 and 12, in a medium with a continuous variation in index of refraction, the fields of the higher order modes extend farther from the axis than do the fields of the lowest order mode. The technique used here to determine the phase constant and characteristic ray angle for rather general media is to establish an equivalent width of medium in which the energy is completely confined —as in guides with perfectly conducting walls or with zero permittivity

walls. It is clear then that this equivalent width must be different for the lowest-order modes than for higher-order modes.

When the author was casting about for a method of expressing this change, E. A. Marcatili pointed out that there is a characteristic radial distance at which the function describing the square-law medium's modes changes from an oscillating function to an exponentially decaying function. These functions are the parabolic cylinder functions and the value of x for the transition is*

$$x_t = 2\sqrt{m + \frac{1}{2}} \tag{125}$$

where m is the mode index.

This was tried as an equivalent width, but was found to change much more rapidly at small m than the actual increase in extent of the field illustrated in Figs. 11 and 12. By examining the radii at which the field decreased to about 1 per cent of the peak value for various low-order modes, it was found that (124) represents the variation quite well. Thus, the general form (125), which is supported by the function theory for the square-law medium, was modified to fit actual known field width variations at small m. For large m, (124) and (125) do have the same variation.

The method of defining the equivalent width, outlined in the body of the paper in connection with (59), causes media with higher than square-law variations in index to merge smoothly into the known behavior of the step-change index variation including the gradual disappearance of the factor b.

REFERENCES

1. Marcuse, D., Theory of a Tubular Gradient Gas Lens, Trans. MTT, Nov., 1965.
2. Marcuse, D., Private communication.
3. Pierce, J. R., Modes in Sequences of Lenses, Proc. Natl. Acad. Sci., *47*, 1961, pp. 1808–1813.
4. Pierce, J. R., *Theory and Design of Electron Beams*, second edition, D. Van Nostrand Co., 1954.
5. Miller, S. E., Alternating Gradient Focusing and Related Properties of Convergent Lens Focusing, B.S.T.J., *43*, July, 1964, pp. 1741–1758.
6. Marcuse, D., and Miller, S. E., Analysis of Tubular Gas Lens, B.S.T.J., *43*, July, 1964, pp. 1759–1782.
7. Marcatili, E. A. J., Modes in a Sequence of Thick Astigmatic Lens-Like Focusers, B.S.T.J., *43*, Nov. 1964, pp. 2887–2904.

* See *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U. S. Dept. of Commerce, National Bureau of Standards, Applied Mathematics Series, *55*, June, 1964, p. 690.

8. Marcatili, E. A. J., to be written.
9. Fox, A. G., and Li, T., Resonant Modes in a Maser Interferometer, B.S.T.J., *40*, March, 1961, pp. 453–488.
10. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers, B.S.T.J., *40*, March, 1961, pp. 489–508.
11. Slepian, D., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-IV: Extensions to Many Dimensions; Generalized Prolate Spheroidal Functions, B.S.T.J., *43*, Nov., 1964, p. 3009.

Statistical Treatment of Light-Ray Propagation in Beam-Waveguides

By D. MARCUSE

(Manuscript received July 15, 1965)

It is well known that uncorrelated, transverse displacements of the lenses of a beam-waveguide cause the light beam to deviate from its axis and that the tolerance requirements on the accuracy of the transverse lens positions are very stringent.

This paper extends the statistics of beam waveguides to include correlations between displacements of different lenses and studies the effects of a succession of uncorrelated bends.

It can be concluded from this work that the rms deviation of the light beam is proportional to the square root of the length of the waveguide if the correlation between lens displacements extends only over a limited range.

The amplitude of the Fourier component of the waveguide axis whose period equals the oscillation period of the ray has to be less than 0.2 micron if the deviation of a light beam passing through 10,000 lenses of a confocal waveguide is to be kept less than 2 mm. This requirement means that the average radius of curvature of a waveguide composed of independent circular sections of an average length of 20 m with lenses spaced 1 m apart has to be more than 10 km. The comparison between two model guides, one composed of circular section and the other of sections shaped like $\sin^2 \beta x$, indicates that the beam deflection depends only on the average radius of curvature and average length of the sections but not on their particular shape.

I. INTRODUCTION

Hirano, Fukatsu and Rowe[1] have studied the behavior of a light beam in a beam-waveguide whose lenses are randomly displaced from a perfectly straight line. The first two authors considered also a waveguide with sinusoidal axis displacements. The behavior of light beams in bent lens-waveguides was studied in Ref. 2. These two papers represent two extreme cases of completely uncorrelated departures of the waveguide axis from perfect straightness on the one hand and perfectly correlated departures from a straight line on the other hand.

This paper describes the statistics of a light ray by introducing a correlation function connecting different points on the waveguide axis.[7]

We show how the ray position at the nth lens depends on one Fourier component of the curvature function of the waveguide axis, while the rms value of the beam displacement depends on one frequency component of the "power spectrum" of the curvature function.

The dependence of the ray position on the Fourier component of the curvature function of the waveguide axis is analogous to the mode conversion loss of multimode waveguides.[3]

The description of the beam deflection in terms of the correlation function of the guide axis is used to draw some general conclusions. It is found that the rms value of the light beam deflection depends on the square root of the number of lenses in the guide provided that the correlation length is much less than the length of the guide. This fact can be used to deduce that the contributions of different, uncorrelated sections of the waveguide to the mean square of the beam deflection simply add up, so that this value can be computed by computing the average value of the beam deflection of one section only.

It is pointed out that apparently plausible models for the correlation function can lead to widely varying results. For this reason no attempt was made to describe the statistics of the waveguide in terms of the correlation function model.

The results of this paper can be applied to alternating gradient focusing systems since it is known[4] that the beam deviation in such systems is of the same order of magnitude as that of a system of positive lenses.

The results of this paper are of particular significance for beam waveguides composed of gas lenses since such a waveguide would use closely spaced lenses so that the number of lenses for a given length of waveguide would be very large. The tolerance requirements are proportional to the square root of the number of lenses in the guide so that the tolerances of waveguides with closely spaced gas lenses become more stringent than those of waveguides using lenses spaced far apart.

II. RELATION TO FOURIER SERIES

We use the ray description of Ref. 2 to study the statistical behavior of a light ray in a lens-waveguide.

The position of the light ray is given by its distance r_n from the lens centers. The inhomogeneous difference equation[2]

$$r_{n+2} - (2 - \kappa)r_{n+1} + r_n = Y_{n+2} \tag{1}$$

with

$$\kappa = L/f \tag{2}$$

(L = lens spacing, f = focal length) connects the ray positions at three successive lenses. The quantity Y_{n+2} at the right hand side of (1) is the distance from the center of the $(n + 2)$th lens to the point at which the straight line through the centers of the nth and $(n + 1)$th lens intersects the $(n + 2)$th lens (Fig. 1). If the lens spacing L becomes infinitesimal, $L \to 0$, Y/L^2 assumes the meaning of the inverse radius of curvature R of the waveguide axis. However, even for finite lens spacing one can define a radius of curvature R_{n+1} by the relation[2] (Fig. 1)

$$Y_{n+2} = \frac{L^2}{R_{n+1}} \tag{3}$$

so that Y_n is a measure of the curvature of the waveguide axis.

The waveguide axis can also be described by the distance S_n of the nth lens from an arbitrary straight line.[1] Between S_n and Y_n exists the following approximate relationship which can easily be derived from Fig. 1

$$Y_{n+2} = -S_{n+2} + 2S_{n+1} - S_n . \tag{4}$$

The solution of (1) can be given in the form[2]

$$r_n = r_n{}^h + r_n{}^i \tag{5}$$

with the solution of the homogeneous equation

$$r_n{}^h = r_0 \cos n\theta + \frac{r_1 - r_0 \cos\theta}{\sin\theta} \sin n\theta \tag{5a}$$

and the definition

$$\cos\theta = 1 - \tfrac{1}{2}\kappa \tag{6}$$

and the solution of the inhomogeneous equation

$$r_n{}^i = \frac{1}{\sin\theta} \sum_{\nu=1}^{n-1} Y_{\nu+1} \sin (n - \nu)\theta, \qquad n \geqq 2. \tag{7}$$

We will calculate the ray's departure r_n from the axis at the end of the waveguide assuming that the ray entered the waveguide on-axis, $r_0 = r_1 = 0$.

Equation (7) can be rewritten

$$r_n = \frac{1}{\sin\theta} \left\{ \sin n\theta \sum_{\nu=1}^{n-1} Y_{\nu+1} \cos \nu\theta - \cos n\theta \sum_{\nu=1}^{n-1} Y_{\nu+1} \sin \nu\theta \right\}. \tag{8}$$

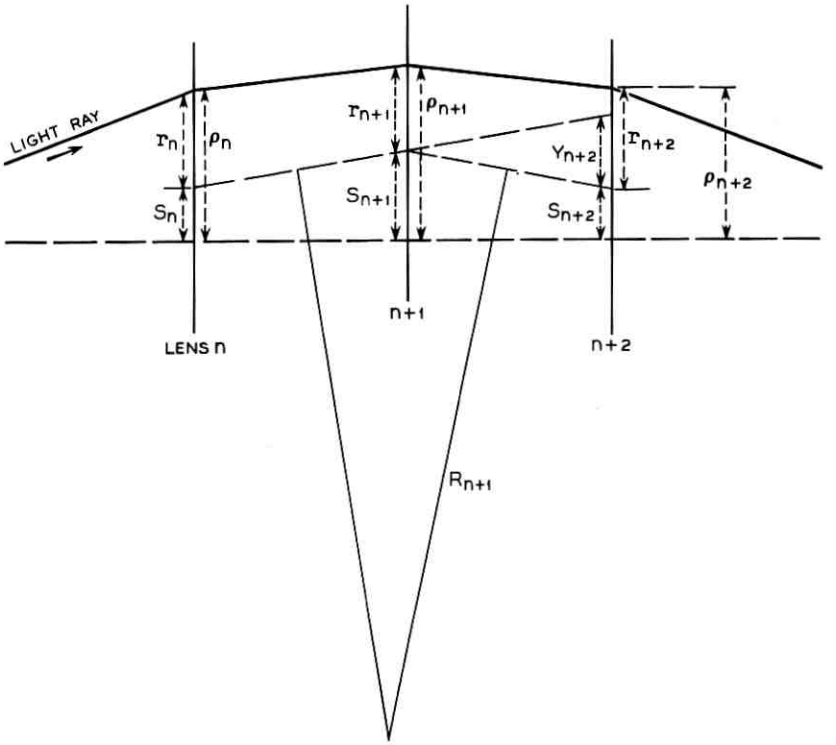


Fig. 1 — Definition of various parameters of the beam-waveguide.

The sums in (8) are Fourier coefficients as can be shown in the following way. The N-discrete points $Y_{\nu+1}$ can be represented by a Fourier series

$$Y_{\nu+1} = \sum_{\mu=-(N-1)/2}^{(N-1)/2} a_\mu \exp [i(2\pi/N)\mu\nu] \qquad \nu = 1, 2, \cdots, N. \quad (9)$$

Equation (9) is a system of N simultaneous equations for N unknown quantities a_μ. The Fourier coefficients a_μ can be calculated by multiplying (9) by $\exp [-i(2\pi/N)\sigma\nu]$ and summing over ν from 1 to N

$$a_\sigma = \frac{1}{N} \sum_{\nu=1}^{N} Y_{\nu+1} \exp [-i(2\pi/N)\sigma\nu]. \quad (10)$$

The representation (9) works if N is an odd integer. Separating a_σ into its real and imaginary part

$$a_\sigma = \alpha_\sigma + i\beta_\sigma \quad (11)$$

we obtain

$$\alpha_\sigma = \frac{1}{N}\sum_{\nu=1}^{N} Y_{\nu+1} \cos (2\pi/N)\sigma\nu \tag{12a}$$

$$\beta_\sigma = -\frac{1}{N}\sum_{\nu=1}^{N} Y_{\nu+1} \sin (2\pi/N)\sigma\nu. \tag{12b}$$

If we choose σ such $\theta = 2\pi\sigma/N$ (this may be possible only approximately but $2\pi\sigma/N$ will approximate θ closely if N is large) and denote the corresponding values of α and β by α_θ and β_θ, we get from (8)

$$r_n = \frac{n-1}{\sin\theta}\{\alpha_\theta \sin n\theta + \beta_\theta \cos n\theta\}. \tag{13}$$

This is an oscillatory function with an amplitude r_{max} which is a slowly varying function of n,

$$r_{max} = \frac{n-1}{\sin\theta}\sqrt{\alpha_\theta^2 + \beta_\theta^2} = \frac{n-1}{\sin\theta}\,|\,a_\theta\,|. \tag{14}$$

The maximum deviation of the ray in the vicinity of the nth lens is determined by the magnitude of the Fourier coefficient of $Y_{\nu+1}$ belonging to the *frequency* θ. The amplitude $|\,a_\theta\,|$ has to be extremely small to keep r_{max} within reasonable limits.

If S_ν is strictly a sinusoidal function

$$S_\nu = A \sin\theta\nu \tag{15}$$

we obtain with the help of (4)

$$Y_\nu = \kappa A \sin\theta\,(\nu - 1) \tag{16}$$

and from (10) with $N \gg 1$

$$|\,a_\theta\,| = \tfrac{1}{2}\kappa A.$$

Equation (14) leads to

$$r_{max} = \frac{\kappa A(n-1)}{2\sin\theta}. \tag{17}$$

Using (3) we can also write

$$r_{max} = \frac{L^2(n-1)}{2R\sin\theta}$$

where R is the minimum radius of curvature.

If we take $\kappa = 2$, $L = 1$ m, $n = 10{,}000$ we obtain a waveguide of 10

km length. Requiring $r_{max} \leqq 2$ mm we find that the amplitude A, the maximum departure of the waveguide from a straight line, has to be

$$A \leqq 0.2 \text{ micron}$$

or

$$R = 2500 \text{ km}.$$

These numbers show how extremely small the θ-Fourier component has to be!

If we take $Y_{\nu+1} = B \sin \vartheta\nu$ and substitute into (10) with $\theta = (2\pi/N)\sigma$ we get nonvanishing values for $|a_\theta|$ even if $\vartheta \neq \theta$ and $N \to \infty$. These nonvanishing values appear for all ϑ values satisfying the equation

$$\vartheta = \theta + 2\pi p \ (p = 0,1,2,3, \cdots).$$

It appears, therefore, as if we obtain Fourier components a_θ for all harmonics $\vartheta = \theta + 2\pi p$. This apparent discrepancy is resolved if we consider the period length $\lambda_\vartheta = 2\pi L/\vartheta$ of the oscillation $\sin (\vartheta/L)\nu L$. The above equation leads to the solutions for the period length

$$\lambda_\vartheta = \frac{L\lambda_\theta}{L + p\lambda_\theta} = \begin{cases} \lambda_\theta & \text{if } p = 0 \\ < L & \text{if } p \neq 0 \end{cases}$$

with $\lambda_\theta = 2\pi L/\theta$. The period length λ_ϑ is therefore either equal to λ_θ, the natural period of the ray oscillations, or it is less than L. Since the lenses are spaced a distance L apart a period of length less than L is meaningless!

III. RANDOM DISPLACEMENTS OF THE GUIDE AXIS

Our considerations so far have been limited to a definite shape of the waveguide axis. However, they can easily be extended to a statistical theory. Equation (14) can be used to obtain the rms value of the maximum beam displacement.

$$\Delta = \sqrt{\langle r_{max}^2 \rangle} = \frac{n-1}{\sin \theta} \sqrt{\langle |a_\theta|^2 \rangle}. \tag{18}$$

The symbol $\langle\ \rangle$ designates an ensemble average. The quantity $\langle |a_\theta|^2 \rangle$ is the expected value of the θ component of the "power spectrum" of the waveguide curvature. It is also possible to express Δ not by means of the power spectrum but by the correlation functions of Y_ν. For this purpose we write, with the help of (10),

$$\langle |a_\theta|^2 \rangle = \frac{1}{N^2} \sum_{\nu=1}^{N} \sum_{\mu=1}^{N} \langle Y_{\nu+1} Y_{\mu+1} \rangle \exp [i\theta(\mu - \nu)]. \tag{19}$$

It is reasonable to assume that $\langle Y_{\nu+1}Y_{\mu+1}\rangle$ depends only on the difference $\mu - \nu$ so that we can write

$$\langle Y_{\nu+1}Y_{\mu+1}\rangle = f_{\mu-\nu} \tag{20}$$

and to assume, furthermore,

$$f_{\mu-\nu} = f_{\nu-\mu}\,. \tag{21}$$

The factor f_λ is the correlation function of the curvature function of the waveguide axis.

Equation (19) can be rewritten in the following way:

$$\begin{aligned}\langle |\, a_\theta \,|^2\rangle = \frac{1}{N^2}\Bigg\{ &\sum_{s=-\frac{1}{2}(N-1)}^{\frac{1}{2}(N-1)} \sum_{t=|s|+1}^{N-|s|} f_{2s} e^{2is\theta} \\ &+ \sum_{s=1}^{\frac{1}{2}(N-1)} \sum_{t=|s|+1}^{N+1-|s|} f_{2s-1} \exp\left[i\theta(2s-1)\right] \\ &+ \sum_{s=-1}^{-\frac{1}{2}(N-1)} \sum_{t=|s|+1}^{N+1-|s|} f_{2s+1} \exp\left[i\theta(2s+1)\right]\Bigg\}.\end{aligned}$$

The summation over t can be carried out since the terms under the summation signs are independent of t. Using (21) to simplify our expression further we obtain

$$\langle |\, a_\theta \,|^2\rangle = \frac{1}{N^2}\left\{Nf_0 + 2\sum_{\lambda=1}^{(N-1)} (N-\lambda) f_\lambda \cos\lambda\theta\right\} \tag{22}$$

so that (18) becomes

$$\Delta = \frac{\sqrt{n-1}}{\sin\theta}\left\{f_0 + 2\sum_{\lambda=1}^{n-2}\left(1 - \frac{\lambda}{n-1}\right) f_\lambda \cos\lambda\theta\right\}^{\frac{1}{2}}. \tag{23}$$

If f_λ decreases with increasing λ so that the upper limit in the sum of (23) becomes immaterial the equation shows that Δ is proportional to $\sqrt{n-1}$ and not to $n-1$ itself as one might have suspected by looking at (18). If the waveguide's curvature contains a sinusoidal component which persists throughout its length, so that no correlation length less than n exists, then the sum is proportional to its upper limit n and Δ becomes truely proportional to n itself.[1] Equation (23) expresses the rms-beam deviation in terms of the correlation function f_λ of the curvature of the waveguide axis. It is easy to rewrite this equation as an expression depending on the correlation function of the waveguide axis displacement itself. The displacement of the nth lens from a straight line is S_n. Defining

$$\langle S_\nu S_\mu\rangle = G_{\nu-\mu} = G_{\mu-\nu} \tag{24}$$

we get with the help of (4) and (20),

$$f_\lambda = 6G_\lambda - 4G_{\lambda-1} - 4G_{\lambda+1} + G_{\lambda-2} + G_{\lambda+2}. \tag{25}$$

Substituting (25) into (23), rearranging terms leads to

$$\Delta = \frac{\sqrt{n-1}\,\kappa}{\sin\theta}\left\{G_0 + 2\sum_{\lambda=1}^{\infty}[1-\lambda/(n-1)]G_\lambda\cos\lambda\theta + \frac{2}{(n-1)\kappa^2}\left[2(1+\kappa-\tfrac{1}{2}\kappa^2)G_0 - (2-\kappa)G_1 - 4\kappa\sqrt{\kappa-\tfrac{1}{4}\kappa^2}\sum_{\lambda=1}^{\infty}G_\lambda\sin\lambda\theta\right]\right\}^{\frac{1}{2}}$$

where we assumed that the correlation length $\Lambda \ll n$. If we assume that $n \gg 1$ (26) simplifies

$$\Delta \approx \frac{\sqrt{n}\,\kappa}{\sin\theta}\left\{G_0 + 2\sum_{\lambda=1}^{\infty}G_\lambda\cos\lambda\theta\right\}^{\frac{1}{2}}. \tag{27}$$

This approximate equation shows that Δ is very nearly proportional to the θ component of the Fourier coefficient of the correlation function.

If the lens displacements S_ν are uncorrelated, $G_\lambda = 0$ for $\lambda \neq 0$, we have

$$\Delta = 2\sqrt{\frac{\kappa}{4-\kappa}}\,\delta\sqrt{n} \tag{28}$$

with

$$\delta = \sqrt{G_0} = \sqrt{\langle S_n^2\rangle}.$$

Equation (28) differs from Ref. 1 (17) (for large values of n) by a factor of $\sqrt{2}$. The reason for the occurrence of this additional factor in our theory is that our Δ is the rms value for the amplitude of the oscillatory beam trajectory while (17), Ref. 1 describes the rms value of all points of the oscillatory beam trajectory.

Let us assume that the waveguide axis is composed of sections of a given average length and that the curvature function of one section is uncorrelated to that of any of the other sections. All sections are assumed to be of the same type. For example, they may all be circular bends which differ only in length and radius of curvature. A waveguide of this type has a finite correlation length and if n, the total number of lenses, is

large we get from (23)

$$\Delta_n^2 = An \tag{30}$$

with

$$A = \left(\frac{1}{\sin\theta}\right)^2 \left\{f_0 + 2\sum_{\lambda=1}^{\infty} f_\lambda \cos\lambda\theta\right\}. \tag{31}$$

Adding one more section with m lenses changes n to $n + m$ in (30). The increase of Δ^2 due to the addition of a section with m lenses is given by $\Delta_m^2 = Am$ so that we get

$$\Delta_{n+m}^2 = \Delta_n^2 + \Delta_m^2 \tag{32}$$

for a guide with $n + m$ lenses.

Each section of the guide can be thought of as such an additional section so that

$$\Delta^2 = \sum_{\nu=1}^{M} \Delta_{m_\nu}^2$$

with $\Delta_{m_\nu}^2$ being the contribution of the νth section and M the number of sections. Introducing the average value $\Delta_{\mathbf{m}}^2$, we obtain

$$\Delta = \Delta_{\mathbf{m}}\sqrt{M}. \tag{33}$$

The rms beam deviation of a waveguide composed of sections can be obtained by calculating the contribution of the rms beam deviation of each section, computing their rms values and applying (33).

IV. EXAMPLES

As a first example we consider a waveguide composed of circular sections which are connected so that the first derivative is continuous. The departure of the light beam which goes through the center of the first two lenses of one of the circular arcs is given by (7)

$$r_n = Y \frac{\sin n\frac{\theta}{2} \sin (n-1)\frac{\theta}{2}}{\sin\theta \sin\frac{\theta}{2}} \tag{34}$$

with $Y = Y_\nu = \text{const}$. If the circular sections have an average number of $\mathbf{m}$ lenses which vary with a Gaussian distribution with variance σ_m we get with the help of (33)

$$\Delta = \frac{\sqrt{\tfrac{1}{2}M\langle Y^2\rangle}}{\sin\theta \sin\frac{\theta}{2}}\left\{1 + \tfrac{1}{2}\cos\theta - 2\cos\frac{\theta}{2}\cos(\mathbf{m} - \tfrac{1}{2})\theta \exp(\tfrac{1}{2}\theta^2\sigma_m{}^2) + \tfrac{1}{2}\cos(2\mathbf{m} - 1)\theta \exp(-2\theta^2\sigma_m{}^2)\right\}^{\frac{1}{2}} \tag{35}$$

with M circular sections per waveguide. It was assumed that Y and m are statistically independent.

The validity of (35) was checked by a computer simulated *experiment*. We constructed 30 different waveguides composed of a series of circular arcs. The quantities Y and m were computed as Gaussian random variables with mean value $\langle Y \rangle = 0$ and $\langle m \rangle = \mathbf{m}$. Each waveguide contained 10,000 lenses. One light ray was traced through each guide with the use of (1) and the rms value of the values r_n with $n =$ 10,000 was computed which, multiplied by $\sqrt{2}$, should equal the value Δ of (35). This experiment was repeated three times with different values of $\mathbf{m}$. In all experiments we considered the confocal case, $\kappa = 2$, $\cos\theta = 0$.

Table I shows how (35) compares to the computer results. The σ_m values in Table I were chosen for the convenience of the computer calculations.

The agreement between the theoretical and *experimental* values is quite good considering that the rms value was computed from only 30 samples.

We can relate $\langle Y^2 \rangle$ to an average radius of curvature $\mathbf{R}$ since according to (3)

$$\langle Y^2\rangle = L^4\left\langle\frac{1}{R^2}\right\rangle = \frac{L^4}{\mathbf{R}^2}. \tag{36}$$

TABLE I

$n = 10{,}000$			
m	σ_m	$\frac{\Delta}{\sqrt{\langle Y^2\rangle}}$ (equation (35))	$\frac{\Delta}{\sqrt{\langle Y^2\rangle}}$ Computer Experiment
3	0.92	77.9	60.2
20	6.1	22.4	21.9
100	30.0	10.0	8.7

TABLE II

m	R
3	39 km
20	11.2 km
100	5.0 km

We may allow $\Delta = 0.2$ cm at the end of the waveguide of 10 km length ($L = 1m$). The permissable values of **R**, computed from the theoretical values of Table I, are listed in Table II.

As a second example we consider a guide which is built up of sections of tapered bends formed according to Fig. 2

$$S_\nu = A \sin^2 (\pi/m)(\nu - 1) \qquad \nu = 1,2 \cdots m. \tag{37}$$

The amplitudes A and the number of lenses m are random variables. Substituting (37) into (4) and (7) we obtain

$$r_{n+1} = -A \frac{\sin^2 \frac{\pi}{m}}{\sin \theta} \cdot \frac{\left(\sin \theta n\right)\left(\cos \frac{2\pi}{m} - \cos \theta\right) + \left(\sin \theta\right)\left(\cos \frac{2\pi}{m} n - \cos \theta n\right)}{2 \sin\left(\frac{\theta}{2} - \frac{\pi}{m}\right) \sin\left(\frac{\theta}{2} + \frac{\pi}{m}\right)}. \tag{38}$$

The lens numbered $n = m + 1$ is the last lens on the bend. According to (5a) the amplitude of the oscillations is given by

$$r_{\max}^2 = r_0^2 + \left(\frac{r_1 - r_0 \cos \theta}{\sin \theta}\right)^2. \tag{39}$$

We restrict ourselves to the confocal case $\kappa = 2$, $\cos \theta = 0$ and set $r_0 = r_m$, $r_1 = r_{m+1}$ because the oscillation caused by the bend is taken as the

Fig. 2 — Beam-waveguide composed of sections of tapered bends.

initial condition of the ray. We obtain from (38) and (39)

$$r_{max}^2 = A^2 \frac{\left(\sin^4 \frac{\pi}{m}\right)\left(1 + \cos^2 \frac{2\pi}{m}\right)\left(1 - \cos \frac{\pi}{2} m\right)}{2\left[\sin\left(\frac{1}{4} - \frac{1}{m}\right)\pi \sin\left(\frac{1}{4} + \frac{1}{m}\right)\pi\right]^2}. \tag{40}$$

To compute Δ we must average r_{max}^2 over A^2 and m. We assume that A and m are uncorrelated and that m follows a Gaussian distribution with average value $\mathbf{m}$ and variance σ_m. To simplify the averaging π/m is replaced by $\pi/\mathbf{m}$ and these terms are taken out of the integral. This procedure is valid only if $\pi/\mathbf{m} \ll 1$ so that we may also take $\sin \pi/\mathbf{m} = \pi/\mathbf{m}$ and $\cos 2\pi/\mathbf{m} = 1$. Remembering that $M = n/\mathbf{m}$, we obtain from (33)

$$\frac{\Delta}{\sqrt{\langle A^2 \rangle}} = \frac{\pi^2 \sqrt{M} \sqrt{1 - \cos \frac{\pi}{2} \mathbf{m} \exp\left(-\frac{1}{8}\pi^2 \sigma_m^2\right)}}{\mathbf{m}^2 \cdot \sin\left(\frac{1}{4} - \frac{1}{\mathbf{m}}\right)\pi \cdot \sin\left(\frac{1}{4} + \frac{1}{\mathbf{m}}\right)\pi} \tag{41}$$

with n being the total number of lenses of the guide and $\mathbf{m}$ the average number of lenses per section.

The equation (41) is valid if $\mathbf{m} \gg 1$ and $\sigma_m/\mathbf{m} \ll 1$. It shows that Δ decreases as the number of lenses per bend is increased.

Some numerical results are shown in Table III. The column labeled "computer result" again contains data calculated by tracing rays through simulated beam-waveguides. The waveguides were "constructed" of arcs according to (37) with A and m being Gaussian random variables. Thirty random waveguides were constructed for each value of $\mathbf{m}$ and the rms values of the ray position r_n with $n = 10{,}000$ was computed from these 30 samples.

The computer results agree with the theoretical values to the order of magnitude. The agreement in this second example is poorer than that of Table I. However, the 30 values of $r_{10,000}$ of this example scatter more widely than those of the first example. Omitting the largest of the 30 values for $\mathbf{m} = 50$ changes Δ by a factor of 0.65 which shows that the statistics of these 30 samples is not very reliable. By comparison omitting the largest value of the sample with $\mathbf{m} = 20$ of the first example changes Δ only by a factor of 0.95.

The results of Table III can be used to get an impression of the line tolerances required for a nominally straight waveguide. Let us assume that the 10,000 lenses of our model guide are spaced 1 m apart ($L = 1m$)

TABLE III

n = 10,000

m	σ_m	$\frac{\Delta}{\sqrt{\langle A^2 \rangle}}$ equation (41)	Computer Result
10	2.8	7.70	12.6
50	15	$1.12\ 10^{-1}$	$8.8\ 10^{-1}$
250	77	$2.00\ 10^{-3}$	$2.7\ 10^{-3}$

which results in a guide 10 km long. If this is a guide built of gas lenses of 0.65 cm diameter (about $\frac{1}{4}$ in.) we might allow $\Delta = 0.2$ cm to be reasonably sure that the light beam will get through the pipe. Fixing Δ allows us to calculate the rms value of the amplitudes $\sqrt{\langle A^2 \rangle}$ of the deviation from straightness and the average radius of curvature of the guide. An average radius of curvature $\mathbf{R}$ at the peak of the arc of (37) is given by

$$\frac{1}{\mathbf{R}} = \left\langle \left| \frac{1}{R} \right| \right\rangle \approx \frac{2\pi^2 \sqrt{\langle A^2 \rangle}}{\mathbf{m}^2 L^2} \approx 2\Delta \frac{\sqrt{\mathbf{m}}}{\sqrt{n}\, L^2} \sin\left(\frac{1}{4} - \frac{1}{\mathbf{m}}\right)\pi \sin\left(\frac{1}{4} + \frac{1}{\mathbf{m}}\right)\pi. \tag{42}$$

The second half of this equation was obtained by substituting (41) for $\langle A^2 \rangle^{\frac{1}{2}}$ assuming that $\frac{1}{8}\pi^2 \sigma_m{}^2 \gg 1$. Table IV shows the values of the rms amplitudes and the average permissable radius of curvature which were computed from the theoretical values of Table III. The tolerance requirements are rather stringent as Table IV shows. For short bends, departures from straightness of only a fraction of a millimeter can be allowed while this tolerance moves up into the meter range as the average length of the bend exceeds hundreds of meters. In case of uncorrelated random wiggles our present example of $\Delta = 2$ mm and $n = 10{,}000$ lenses leads to an rms value for the position tolerances of $\delta = 0.01$ mm according to (28). This is a very real tolerance requirement

TABLE IV

mL	$\sqrt{\langle A^2 \rangle}$	$\mathbf{R}$
10 m	$2.6\ 10^{-2}$ cm	19.4 km
50 m	1.78 cm	7.1 km
250 m	$1.00\ 10^{2}$ cm	3.16 km

since there will always be random lens displacements superimposed on longer bends so that the uncorrelated, random component of lens displacements has to be kept below 10 microns.

A comparison of Table II with Table IV shows that the average radii of curvature permissable for our two examples are nearly the same. This is true even though the waveguides of the two examples are of very different construction. This may support the belief that the average radius of curvature and the length of the sections determine the deflection of the ray regardless of how the waveguide is shaped in detail. The agreement between the values of Table II and IV is improved if one corrects for the different length of the sections.

It may be in order to add a remark concerning models for the correlation function G_λ. Since the correlation has to be of finite length one might be tempted to try a correlation function of the form[6]

$$G_\lambda = G_0 \exp\left(-\frac{|\lambda|}{q}\right) \tag{43}$$

with q being the number of lenses within the correlation distance. The correlation number q must be of the same order of magnitude as the average number $\mathbf{m}$ of lenses per section of our model waveguide. Substituting (43) into (27) we obtain

$$\Delta = \frac{\sqrt{n}\,\kappa}{\sin\theta}\sqrt{G_0} \cdot \sqrt{1 - 2\exp(-1/q)\,\frac{\exp(-1/q) - \cos\theta}{1 + \exp(-2/q) - 2\exp(-1/q)\cos\theta}} \tag{44}$$

or if $1/q \ll 1$

$$\Delta = \frac{\sqrt{n}\,\kappa}{\sin\theta}\sqrt{\frac{G_0}{q(1-\cos\theta)}}. \tag{45}$$

This correlation function is obviously a poor model for a waveguide with random bends since Δ of (45) decreases with q only like $q^{-\frac{1}{2}}$ while Δ of (41) decreases like $\mathbf{m}^{-2}$. For $q = 250$ we obtain from (45), setting $\sqrt{G_0} \approx \sqrt{\langle A^2\rangle}$ ($G_0 \approx \frac{1}{2}\sqrt{\langle A^2\rangle}$ for the arcs of (37)),

$$\frac{\Delta}{\sqrt{\langle A^2\rangle}} = 6.4$$

which is three orders of magnitude larger than corresponding values of Table III.

Another possible choice for a correlation function may be

$$G_\lambda = G_0 \exp(-\lambda^2/q^2). \tag{46}$$

Using the identity[5]

$$\sum_{\lambda=-\infty}^{\infty} \exp(-\lambda^2/q^2) \cos \lambda\theta = \sqrt{\pi}\, q \sum_{\nu=-\infty}^{\infty} \exp(-\pi^2 q^2[(\theta/2\pi) - \nu]^2) \quad (47)$$

we obtain from (27)

$$\frac{\Delta}{G_0} = \frac{\pi^{\frac{1}{4}} \kappa \sqrt{n}}{\sin \theta} \sqrt{q}\, \{\exp\{-\pi^2 q^2[(\theta/2\pi) - 1]^2\} + \exp(-\tfrac{1}{4} q^2\theta^2) + \exp\{-\pi^2 q^2[(\theta/2\pi) + 1]^2\}\}^{\frac{1}{2}} \quad (48)$$

where all but three terms of the sum on the right hand side of (47) are neglected which is justified if $q > 1$.

The maximum rms beam deviation of (48) decreases with increasing q like

$$\sqrt{q} \exp(-\tfrac{1}{8} q^2\theta^2)$$

that is much faster than (45). These two examples indicate how critically Δ depends on the shape of the correlation function. It appears that more insight can be gained by choosing models for the random deviation of the waveguide curvature rather than by trying to guess at model correlation functions.

The reason for the critical dependence of Δ on the shape of the correlation function can be seen from the following argument.

In our second example, (37) we obtain $G_0 \approx 0.5 \langle A^2 \rangle$. However the ratio $\Delta/\sqrt{\langle A^2 \rangle}$ is, for example, in the order of 0.1 according to the second line of Table III. From (27) we obtain, with $\kappa = 2$ and $n = 10^4$,

$$\frac{\Delta}{\sqrt{\langle A^2 \rangle}} \approx 2\, 10^2 \sqrt{0.5 + 2 \sum_{\nu=1}^{\infty} \frac{G_\lambda}{\langle A^2 \rangle} \cos \lambda\theta}.$$

If $\Delta/\sqrt{\langle A^2 \rangle}$ is to be 0.1, the sum under the square root sign must be very nearly equal to -0.25 so that the two terms under the square root sign cancel to a term of the order of magnitude 10^{-6}. A very slight variation of the value of the sum gives rise to a large variation of Δ.

As a last example we consider a waveguide with M random tilts. If each tilt is located at a lens, Ref. 2 (31), gives for the beam amplitude caused by one tilt with angle α

$$r_{\max} = 2\alpha \frac{L}{\sqrt{4\kappa - \kappa^2}} \quad (49)$$

so that (33) leads to

$$\Delta = 2 \sqrt{\langle \alpha^2 \rangle} \frac{L}{\sqrt{4\kappa - \kappa^2}} \sqrt{M}. \quad (50)$$

If $n = 10{,}000$, $L = 1m$, $\kappa = 2$, and $M = 100$ we find that

$$\sqrt{\langle \alpha^2 \rangle} \leqq 2\ 10^{-4} \text{ radians} = 0.0115°$$

if $\Delta \leqq 2$ mm is required.

V. CONCLUSION

Is it more advantageous to space the lenses of a beam-waveguide closely or farther apart? The answer to this question depends on our ability to control tolerances. Equation (41) shows that the rms beam deviation due to random bends decreases rapidly with increasing number of lenses, while (28) indicates that the rms beam deviation due to uncorrelated lens displacements increases slowly with increasing lens number. Only practical experience can tell how to compromise between these two conflicting requirements.

APPENDIX

Equivalence of Two Representations

The problem of ray propagation in a bent lens-waveguide has been treated in Ref. 2 and by Hirano and Fukatsu.[1] The treatments of the problem in these two papers differ in the way the ray is described. In Ref. 2 the ray position r_n is measured from the center of the lenses and the position of the lenses with respect to each other is described by a quantity Y_n (Fig. 1). In Ref. 1, a straight reference line is used to determine the position ρ_n of the ray as well as the lens displacements S_n. The ray position r_n at the nth lens in the representation of Ref. 2 is given by

$$r_n = \frac{1}{\sin \theta} \sum_{\nu=1}^{n-1} Y_{\nu+1} \sin (n - \nu)\theta \qquad n \geqq 2. \tag{51}$$

The values of r_n at $n = 0$ and $n = 1$ are $r_0 = r_1 = 0$.

In the representation of Ref. 1 the solution of the inhomogeneous difference equation reads

$$\rho_n = \frac{\kappa}{\sin \theta} \sum_{\nu=1}^{n-1} S_\nu \sin (n - \nu)\theta \qquad n \geqq 2 \tag{52}$$

with $\rho_0 = \rho_1 = 0$.

With the help of Fig. 1 it is easy to see that approximately

$$S_{n+2} = 2S_{n+1} - S_n - Y_{n+2}\sqrt{1 - (S_{n+1} - S_n)^2/L^2}, \tag{53}$$

or if

$$\frac{S_{n+1} - S_n}{L} \ll 1$$

$$Y_{n+2} = 2S_{n+1} - S_{n+2} - S_n\,, \tag{53a}$$

and

$$\rho_n = r_n + S_n\,. \tag{54}$$

The substitution of (53a) and (54) into (51) leads to

$$\begin{aligned}\rho_m &= S_n + \frac{1}{\sin\theta}\sum_{\nu=1}^{n-1}(2S_\nu - S_{\nu+1} - S_{\nu-1})\sin(n-\nu)\theta \\ &= S_n + \frac{1}{\sin\theta}\Big\{S_1\sin(n-1)\theta - S_n\sin\theta - S_0\sin(n-1)\theta \\ &\quad + \sum_{\nu=1}^{n-1} S_\nu[2\sin(n-\nu)\theta - \sin(n-\nu+1)\theta - \sin(n-\nu-1)\theta]\Big\}\end{aligned}$$

which can be simplified to

$$\begin{aligned}\rho_n = {} & (S_0 - S_1)\cos n\theta - (S_0 - S_1)\cos\theta\sin n\theta \\ & + \frac{\kappa}{\sin\theta}\sum_{\nu=1}^{n-1} S_\nu \sin(n-\nu)\theta.\end{aligned} \tag{55}$$

The first two terms with $S_0 - S_1$ are solutions of the homogeneous difference equation and appear here because $r_0 = r_1 = 0$ does not coincide with $\rho_0 = \rho_1 = 0$ if S_0 and S_1 are unequal to zero.

Since the first two terms can always be removed by adding a suitable solution of the homogeneous equation the equivalence of (51) and (52) has been shown.

REFERENCES

1. Hirano, J., and Fukatsu, Y., Stability of a Light Beam in a Beam Waveguide, Proc. IEEE, *52*, 11, Nov., 1964, pp. 1284–1292. Rowe, H. E., unpublished work.
2. Marcuse, D., Propagation of Light Rays Through a Lens Waveguide with Curved Axis, B.S.T.J. *43*, March, 1964, pp. 741–753.
3. Rowe, H. E., and Warters, W. D., Transmission in Multimode Waveguide with Random Imperfections, B.S.T.J., *41*, May, 1962, pp. 1031–1170.
4. Steier, W. H., Alternating Gradient Focusing of Light Beams, to be published.
5. Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, Interscience Publishers, 1962, *2*, p. 200, equation (8).
6. Unger, H. G., Light Beam Propagation in Curved Schlieren Guides, Arch. Elektr. Übertr., *19*, April, 1965, pp. 189–198.
7. Berreman, D. W., Growth of Oscillations of a Ray about the Irregularly Wavy Axis of a Lens Light Guide, B.S.T.J., This Issue, pp. 2117–2064.

Properties of Periodic Gas Lenses

By D. MARCUSE

(Manuscript received June 18, 1965)

Gas lenses are being considered as focusing elements of beam-waveguides. Since very many lenses are needed to form a long waveguide, it is reasonable to consider periodic arrangements of gas lenses. Such periodic structures might operate with a gas stream which flows through all of the lenses in succession. A periodic temperature distribution in the gas results which is different from that of single gas lenses considered in two earlier papers.[3,4]

This paper analyses the ray optics properties, such as focal length and principal surfaces of the gas lenses, of two types of alternating gradient focusing systems. One system consists simply of a succession of hot and cold tubes. The other system results from the first by insertion of heat insulating tubes of equal length between the hot and cold tubes.

I. INTRODUCTION

The beam-waveguide described by Goubau[1] appears as a promising device to transmit light over long distances. However, to reduce the power loss due to absorption and reflection, which is inevitable with lenses made of solid dielectrics, gas lenses have been proposed[2,3] instead of the solid lenses used by Goubau.

Two earlier papers[3,4] discussed the properties of a particular type of gas lens. This tubular gas lens consists of a warm tube into which a cooler gas is blown. The thermal gradients in the gas lead to density gradients which give the structure the properties of a positive lens.

Ref. 4 discusses the focal length and principal surface of this gas lens for the case that the gas enters the lens at a constant temperature. It was shown that this device, when operated under optimum conditions, acted as an optically rather thin lens with moderate lens distortions.

The present paper extends the earlier analysis in several ways. We consider periodic lens structures. Such a structure results if hot and cold tubes are alternated to form a long, periodic structure. The gas is heated and cooled periodically giving rise to periodically arranged positive and negative lenses. A periodic structure of this type represents an alter-

nating gradient focusing system.[5] In general, the temperature of the gas entering the hot or cold tubes will not be constant over the cross section of the tube so that the earlier results are no longer applicable. The gas temperature in the periodic structure will also be periodic and will depend on the temperatures of the hot and cold tubes as well as on the flow velocity.

It is our aim to compute the temperature distribution in such periodic structures and use it to determine the properties of the equivalent lenses which describe the ray optics of the alternating gradient focusing systems.

The equivalent lenses are rather complicated. They are neither optically thin nor free of distortions. Further investigations are required to determine the guidance properties of an alternating gradient focusing system with imperfect lenses of this type.

We discuss two types of periodic gas lens systems. In one case we assume that hot and cold tubes of equal length are directly adjacent to each other. The other type is an alternating gradient beam-waveguide which consists of hot and cold tubes which are separated by tube sections made of an ideally heat insulating material. For simplicity it is assumed that the insulating sections are as long as the hot and cold tubes. The assumption of a perfectly heat insulating material is an overidealization since hardly any material conducts heat more poorly than gases. It is intended as an approximation to the real situation of imperfect heat insulators.

In the insulating tube sections, the gas has a chance to equalize its temperature. As it does so rather rapidly, we again have the case of hot and cold tubes being fed by an input gas at a constant temperature. However, the insulating sections act also as lenses in the same sense as the hot or cold tubes by which they are preceeded. Therefore, it is not surprising that some improvement of efficiency results if heat insulating tube sections are used to separate the hot and cold tubes. But this advantage is not very striking; and, since this analysis assumes ideally insulating tubes, it is not certain how much of a real advantage can be gained by using this construction. Considerably more experience is needed before a decision can be made.

This analysis again neglects all convection effects in the gas lenses.

To be able to distinguish which of the two structures is being discussed we will call the structure using hot and cold tubes without heat insulating tube sections the *simple periodic structure* while the second case which includes insulating sections will be called the *extended periodic structure*.

II. RAY TRACING

Before entering into a discussion of the simple and extended periodic structures, the ray tracing technique used to determine the focal length and principal surface of the lenses will be explained.

The trajectory of the light ray in the gas lens is given by the ray equation[6]

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \text{grad } n \tag{1}$$

$\mathbf{r}$ = the position vector leading from an arbitrary origin to points on the ray.
n = index of refraction.
s = length coordinate measured along the ray.

We limit ourselves to rays which are very nearly parallel to the axis of the structure which is used as the z-coordinate so that we can replace s by z.* Assuming angular symmetry it is sufficient to consider the vector component in radial direction r perpendicular to the z-axis. Finally, we neglect the term

$$(\partial n/\partial z)\ (dr/dz)$$

because $dr/dz \ll 1$ for rays which are nearly parallel to the z-axis and also because the variation of n in the z-direction will generally be smaller than that in the r-direction.

$$(\partial n/\partial z) \ll (\partial n/\partial r).$$

With these assumptions we obtain from (1)

$$\frac{d^2r}{dz^2} = \frac{1}{n}\frac{\partial n}{\partial r}. \tag{2}$$

However, since we are only interested in gases where $n - 1 \ll 1$ we can safely write

$$\frac{d^2r}{dz^2} = \frac{\partial n}{\partial r}. \tag{3}$$

The index of refraction depends on temperature in the following way:

$$n - 1 = (n_0 - 1)\frac{T_0}{T} \tag{4}$$

T_0 is the absolute temperature at which n_0 is measured while T is the

* The error caused by this approximation is estimated in Ref. 4.

absolute temperature for which we want to determine n. It follows from (4) that

$$\frac{\partial n}{\partial r} = -(n_0 - 1)\frac{T_0}{T^2}\frac{\partial T}{\partial r}.$$

The value of the absolute temperature varies only slightly throughout the gas so that T can be replaced by a suitable average temperature. It is convenient to choose T_0 equal to this average temperature which should be chosen as

$$T_0 = \tfrac{1}{2}(T_h + T_c) \tag{5}$$

T_h = temperature of hot tubes
T_c = temperature of cold tubes

This leads us to the ray equation

$$\frac{d^2r}{dz^2} = -\frac{n_0 - 1}{T_0}\frac{\partial T}{\partial r}. \tag{6}$$

Equation (6) is our starting point for the ray optics of the gas lenses. Since $\partial T/\partial r$ is a complicated function of r and z, it is difficult to solve (6) analytically so that we content ourselves with numerical solutions obtained by means of an electronic computer.

Rather than expressing our results as functions of z, we want to obtain them as functions of the *on*-axis gas velocity v_0 normalized by a suitable constant V. (This representation was also used in Refs. 3 and 4.) We define $V(L)$ by

$$v_0/V(L) = a/\sigma L \tag{7}$$

with

$$\sigma = k/av_0\rho c_p \tag{8}$$

a = tube radius
L = tube length
k = heat conductivity of the gas
ρ = (average) gas density
c_p = specific heat at constant gas pressure.

Equation (7) shows that $v_0/V(L)$ is inversely proportional to the length of the tube. Therefore, it is convenient to introduce a variable

$$u(z) = a/\sigma z \tag{9}$$

which at $z = L$ equals

$$W = u(L) = \frac{v_0}{V(L)}. \tag{10}$$

Using

$$x = r/a \tag{11}$$

equation (6) becomes

$$\frac{d^2x}{du^2} + \frac{2}{u}\frac{dx}{du} = -\frac{1}{u^4}\frac{n_0 - 1}{\sigma^2 T_0}\frac{\partial T}{\partial x}. \tag{12}$$

Equation (12) is used to obtain x and dx/du as a function of W from which the focal length f and principal surface p can be computed.

As shown in Fig. 1(a), we follow the ray from a point $z = z_1$ to $z = z_2$, corresponding to $u = u_1$ and $u = u_2$, through the tube anticipating the case of lenses which are not bounded by planes through the ends of the tube but by surfaces inside of the tube to be defined later.

The definition of focal length and principal surface can be seen from Fig. 1(a). The principal surface is obtained by following a ray, which is incident parallel to the axis ($dx/dz = 0$ at $z = z_1$), through the lens. If we extend the direction of the ray entering the lens and the direction of the ray leaving the lens at $z = z_2$ by straight lines back into the lens we obtain a cross-over point which defines a point on the principal surface. The distance p of this point measured from the beginning of the tube as a function of x_1, the input position of the ray, describes the principal surface. The distance p_+ for rays traveling in the same direction as the gas flow is given by (Fig. 1(a))

$$p_+ = z_1 + L - \frac{x(z_2) - x(z_1)}{\left(\frac{dx}{dz}\right)_{z=z_2}} \tag{13}$$

or, expressed in terms of u and W rather than L

$$\frac{p_+}{L} = \frac{z_1}{L} + 1 + \frac{x(u_2) - x(u_1)}{u_2^2\left(\frac{dx}{du}\right)_{u=u_2}} W. \tag{14}$$

We define the focal length as the distance from the intersection of the incoming ray (extended in a straight line) with the principal surface to the point at which the outgoing ray crosses the axis of the structure.

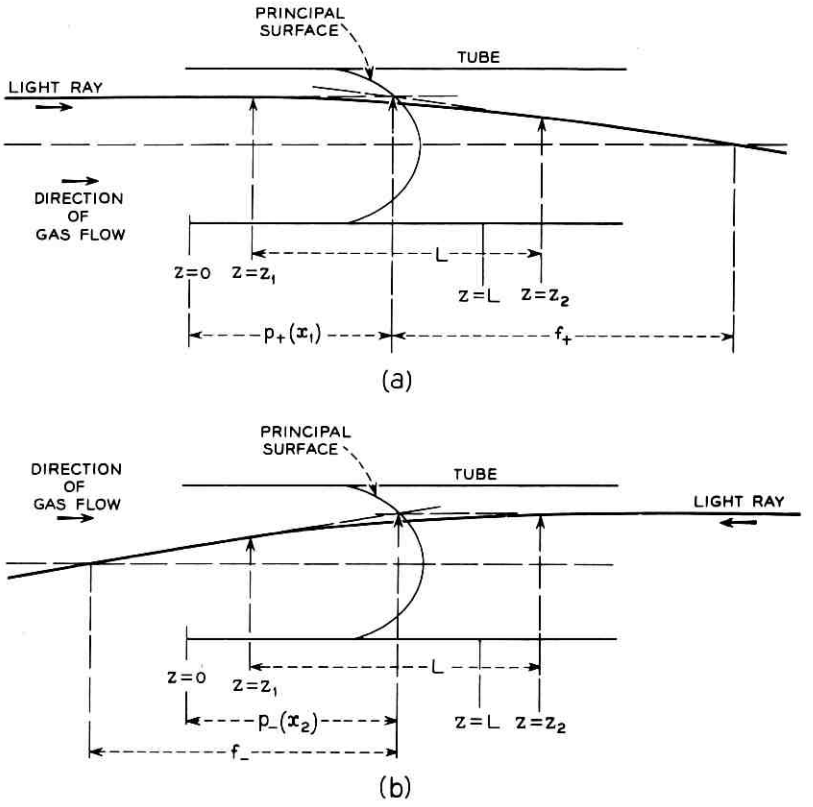


Fig. 1 — Geometry of a single gas lens showing the definition of focal length and principal surfaces.

Therefore, we have

$$f_+ = -\frac{x(z_1)}{\left(\frac{dx}{dz}\right)_{z=z_2}} \tag{15}$$

or, in terms of u and W,

$$\frac{f_+}{L} = W\frac{x(u_1)}{u_2^2\left(\frac{dx}{du}\right)_{u=u_2}}. \tag{16}$$

Similarly, we obtain from Fig. 1(b) the principal surface and focal

length for the ray traveling in opposite direction to the gas flow

$$\frac{p_-}{L} = \frac{z_1}{L} - W \frac{x(u_2) - x(u_1)}{u_1^2 \left(\frac{dx}{du}\right)_{u=u_1}} \tag{17}$$

and

$$\frac{f_-}{L} = -W \frac{x(u_2)}{u_1^2 \left(\frac{dx}{du}\right)_{u=u_1}} . \tag{18}$$

The principal surfaces p_+ and p_- do not, in general, coincide. If they are identical the lens is called optically thin. The separation between the two principal surfaces is an indication of the optical thickness of the lens.

III. THE SIMPLE PERIODIC STRUCTURE

3.1 *Temperature Distribution*

The simple periodic structure consists of alternating hot and cold tubes (Fig. 2). In order to compute the equivalent positive and negative lenses of this structure we first have to determine the temperature distribution.

The temperature distribution is given by a series expansion[7] which, in the hot tube, reads

$$T_1(x,u) = T_h - \sum_{n=0}^{\infty} A_n R_n(x) \exp\left(-\beta_n^2/u\right) \tag{19}$$

$$\infty > u > W = \frac{v_0}{V(L)}$$

with T_h being the wall temperature of the hot tube and $u = a/\sigma z$; and

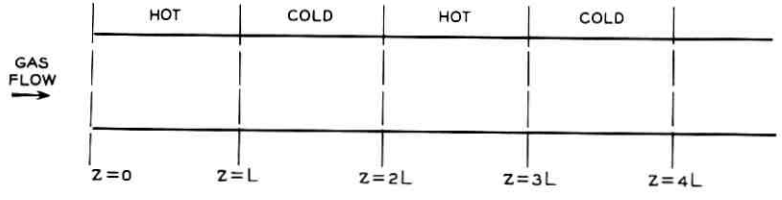


Fig. 2 — Sequence of hot and cold tubes comprising the "simple periodic structure".

in the cold tube

$$T_2(x,u) = T_c - \sum_{n=0}^{\infty} B_n R_n(x) \exp\left\{-\beta_n^2 \cdot \left(\frac{1}{u} - \frac{1}{W}\right)\right\} \qquad (20)$$

with

$$W > u > \tfrac{1}{2}W$$

T_c = wall temperature of cold tube
$W = a/\sigma L$.

The functions R_n and the eigenvalues β_n are discussed in the appendix. The coefficients A_n and B_n have to be determined so that the temperature is a periodic function in the simple periodic structure of Fig. 2.

To simplify the determination of A_n and B_n we limit ourselves to sufficiently long tubes or slow enough flow velocities so that the first term of the series expansions (19) and (20) are sufficient to describe the temperature distribution at the end of each tube accurately. This condition is expressed by the requirements

$$W = \frac{v_0}{V(L)} < 10. \qquad (21)$$

The periodicity condition requires that the temperature at the end of the cold tube ($z = 2L$ or $u = \frac{1}{2}W$) equals the temperature at the beginning of the hot tube ($z = 0$ or $u = \infty$)

$$T_1(x,\infty) = T_2(x,\tfrac{1}{2}W). \qquad (22)$$

In addition, we have to require that the gas temperature passes continuously from the hot to the cold tube

$$T_1(x,W) = T_2(x,W). \qquad (23)$$

The conditions (21) to (23) allow the determination of the constants A_n and B_n

$$A_n = -\frac{2(T_h - T_c)}{\beta_n \left(\frac{\partial R_n}{\partial \beta}\right)_{x=1}} \left\{1 - \frac{\delta_{on}}{1 + \exp\left(\beta_0^2 \frac{1}{W}\right)}\right\} \qquad (24)$$

and

$$B_n = -A_n \qquad (25)$$

with

$$\delta_{on} = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0. \end{cases}$$

The eigenvalues β_n and the derivation $\partial R/\partial \beta$ are given in the appendix.

The temperature distribution in the hot tube is shown in Fig. 3(a) for $v_0/V(L) = 5$ and in Fig. 3(b) for $v_0/V(L) = 10$ as a function of normalized radium $x = r/a$. The various curves in each figure correspond to different positions along the tube axis. At $z = 0$ the temperature distribution is identical to that at the end of the cold tube. The temperature is cold on the wall and warmer in the center of the tube. As we follow the temperature distribution deeper into the hot tube we see that the temperature on the wall changes instantly from its value equal to the wall temperature of the cold tube to that of the hot tube. However, the slope of the temperature distribution close to the tube axis remains negative for quite some length. This means that the gas in the hot tube acts like a negative lens close to the input end of the tube. It takes some distance to reverse the negative temperature gradient which the cold tube imparted to the gas. In fact, there exists a neutral surface in the hot as well as the cold tube which is defined by the points where the temperature gradient $\partial T/\partial x = 0$. On this surface the gas acts neither as a positive nor negative lens. The neutral surface separates the region

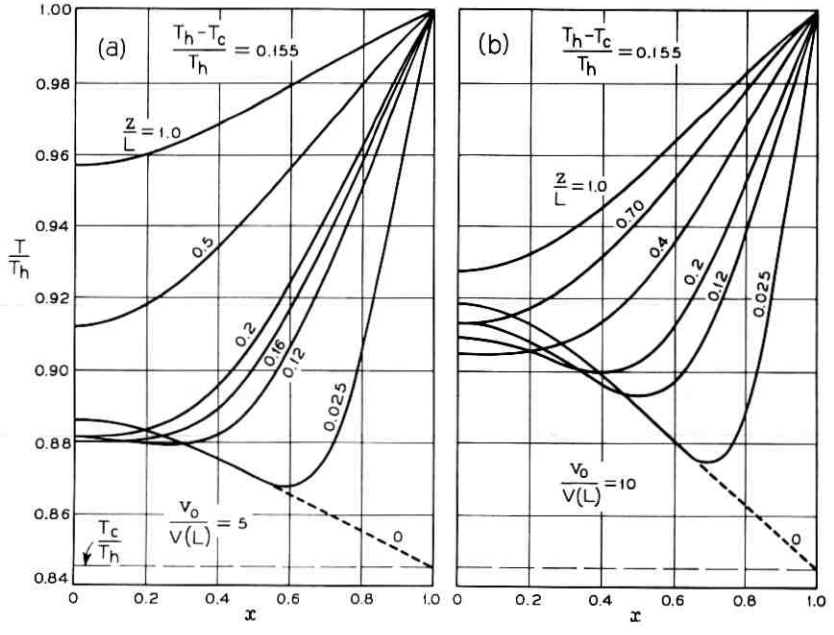


Fig. 3(a) — Temperature distribution in the simple periodic structure as a function of normalized radius $x = r/a$ at various cross sections z/L in the tube. The normalized flow velocity is $v_0/V(L) = 5$ and $(T_h - T_r)/T_h = 0.155$.

Fig. 3(*b*) — Same as Fig. 3(a) with $v_0/V(L) = 10$.

of the positive from the negative lens. It has the same shape in the hot as well as the cold tube and the distance between corresponding points of these surfaces in either tube is L, the length of the hot and cold tubes. The temperature distribution T/T_h in the cold tube is obtained by reflecting each point of the temperature distribution of Fig. 3(a) or 3(b) on the line parallel to the x-axis at $T/T_h = T_h + T_c/2T_h$. The neutral surfaces, z/L as a function of x, for various values of flow velocity are obtained by rotating the curves of Fig. 4 around the z-axis. At high gas velocities ($v_0/V(L) = 10$) the neutral surface extends almost to the half way point into the hot and cold tubes.

3.2 *Focal Length and Principal Surface*

To calculate effective lenses which describe the ray optics properties of the hot and cold tubes it is not permissible to trace rays through each tube and compute focal length data from the ray trajectory since each tube functions as a combination of positive and negative lenses. It is more reasonable to trace rays from one neutral surface to the next since the gas between two neutral surfaces acts entirely in one sense either as a positive or negative lens.

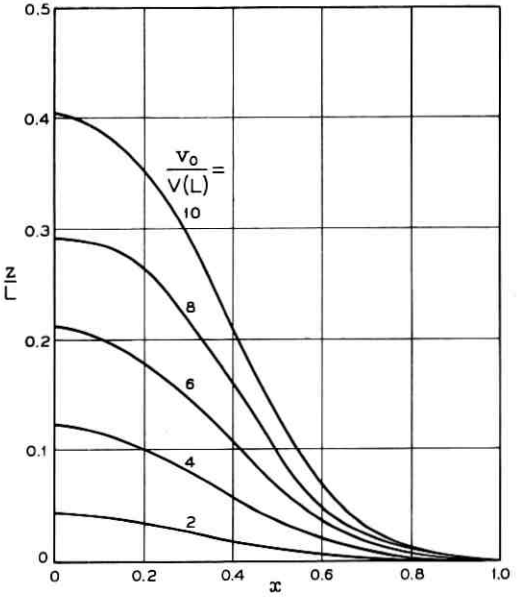


Fig. 4 — Shape of the neutral surfaces for various values of $v_0/V(L)$.

We inject a ray at a point z_1 , $x(z_1)$ on the neutral surface with a slope $(dx/dz)_{z=z_1} = 0$ and follow it to the point $z_2 = z_1 + L$, $x(z_2)$. This point does not, in general, lie on the next neutral surface since the ray moves from its entrance position, $x(z_1) \neq x(z_2)$. However, this point z_2 , $x(z_2)$ lies sufficiently close to points on the next neutral surface that this ray tracing procedure seems justified.

Our present discussion explains the meaning of the points z_1 and z_2 (or correspondingly u_1 and u_2) introduced in (13) through (18) and shown in Fig. 1.

The slope and positions of rays entering at $z = z_1$ with $dx/dz = 0$ were computed at $z = z_2$ by numerical integration of (12). The temperature distribution entering into (12) is given by (19) and (20). The values of the slope and the ray position were then used to calculate the focal length and principal surface from (14) and (16). The rays traveling in the direction opposite to the gas flow were launched at $z = z_2$, $x(z_2)$ on the neutral surface with the slope $(dx/dz)_{z=z_2} = 0$ and their slope and position at $z_1 = z_2 - L$, $x(z_1)$ was used to calculate p_-/L and f_-/L from (17) and (18).

It is apparent from (12) and (24) that all of our results depend on a parameter

$$D = \frac{n_0 - 1}{\sigma^2} \frac{T_h - T_c}{T_0}. \tag{26}$$

However, we like to plot our results as functions of $W = v_0/V(L)$ which is contained in σ. It is therefore convenient to write

$$D = \left(\frac{v_0}{V(L)}\right)^2 C\left(\frac{L}{a}\right) \tag{27}$$

and use

$$C\left(\frac{L}{a}\right) = (n_0 - 1) \frac{T_h - T_c}{T_0} \left(\frac{L}{a}\right)^2 \tag{28}$$

to characterize the focusing power of the lens.* Fig. 5 shows the focal length f divided by the length L of the tubes as a function $v_0/V(L)$. The solid curves represent the positive, the broken curves the negative lens. The focal length of the negative lens is shown as a positive quantity. These curves were computed setting $x(z_1) = 0.1$. The positive and negative lenses have almost equal focusing power for small values of $C(L/a)$. The negative lens has more focusing power for larger values

* In reference 4 $C(L/a)$ was defined slightly differently. There, $T_h - T_c$ was replaced by $T_h - T_i$ (T_i temperature of input gas).

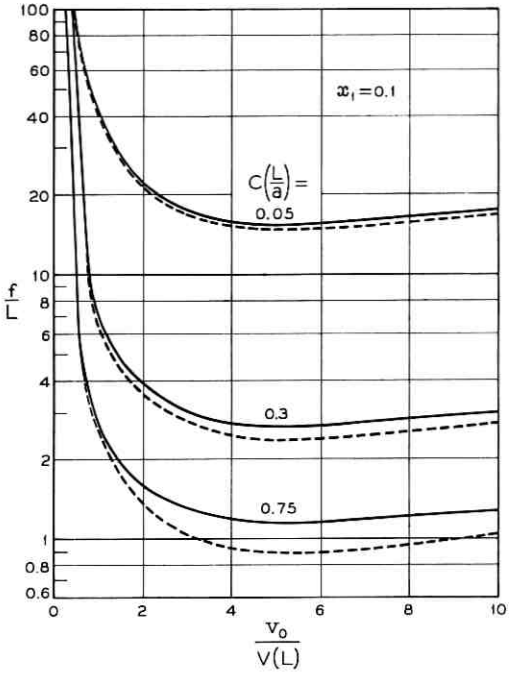


Fig. 5 — Normalized focal length f/L of the positive (solid lines) and negative (broken lines) lenses of the simple periodic structure as functions of the normalized flow velocity $v_0/V(L)$.

of this parameter. The temperature gradients are equal but opposite in sign for either of these lenses so that one might expect that they should have equal focusing power. The discrepancy is explained by considering that the rays travel through different parts of the lenses. In the positive lens the ray starting at $x(z_1) = 0.1$ moves closer towards the lens axis while the ray in the negative lens, starting at the same point, moves away from the axis and toward the wall.

The minima of the focal length curves are explained by the fact that we have no lens action if the gas is stationary, $v_0/V(L) = 0$. The lens begins to function with increasing gas flow. But, if the gas finally flows so fast that the *on*-axis temperature does not have time to follow, lens action ceases again. Interpolation of curves for parameter values other than those shown in Fig. 5 is facilitated by noting that the focal length is nearly proportional to $[C(L/a)]^{-1}$.

The focal length of an ideal lens does not depend on the input position $x(z_1)$ of the ray. Plotting f/L as a function of x should result in a straight

line parallel to the x-axis. That gas lenses are not ideal lenses is shown in Figs. 6(a) through 6(e). These figures contain three different types of curves. The solid curves represent the positive lens for rays traveling in the same direction as the gas while the dotted curves represent rays traveling opposite to the gas flow. The dash-dotted curves give the results for the negative lens and rays in the positive gas direction. The rays opposite to the gas flow in the negative lens have been ommitted. They can be visualized by the fact that the curves in the two directions coincide at $x = 0$. At $x = 0.9$ the curves for the negative lens join up with the dotted lines of the positive lens. The lines for the negative lens do not all extend to $x = 0.9$ because the ray in the negative lens moves toward the wall and may hit it before it travels its full length if the lens is too strong and if the ray started out sufficiently close to the wall.

Figs. 6(a) through 6(e) show that there are focal length distortions for smaller values of $v_0/V(L)$. For $v_0/V(L) = 6$ and 8 the focal length curves are substantially parallel to the x-axis. (We see, furthermore, that the focal length for the two directions of propagation coincide more closely for smaller values of $C(L/a)$ and x.

The principal surfaces are shown in Figs. 7(a) through 7(e). The meaning of the solid, dotted and dash-dotted curves is the same as explained above. The dash-dotted lines for the negative lens for the low values of $C(L/a)$ coincided very nearly with the solid line for the positive lens and was omitted. Also not shown are the corresponding curves for the negative lens for rays traveling against the gas flow. The principal surfaces are far from being plane. It is also apparent that for most values of $C(L/a)$ and x, the two principal surfaces for the two directions of the beams don't coincide too closely. This shows not only that the lenses comprising the simple periodic structure have considerable distortions but also that they are not optically thin under all conditions.

Fig. 8 shows the dependence of the point $x = 0.1$ of the principal surfaces on the flow velocity $v_0/V(L)$. The principal surfaces move to $z = 0$ for vanishing flow velocities and extend far into the tube for large values of $v_0/V(L)$.

IV. THE EXTENDED PERIODIC STRUCTURE

4.1 *Temperature Distribution*

The extended periodic structure is shown in Fig. 9. It consists of alternating positive and negative lenses which are separated by pieces of insulating tubes of equal length.

The temperature distribution, T_3, in the hot or cold tubes is well

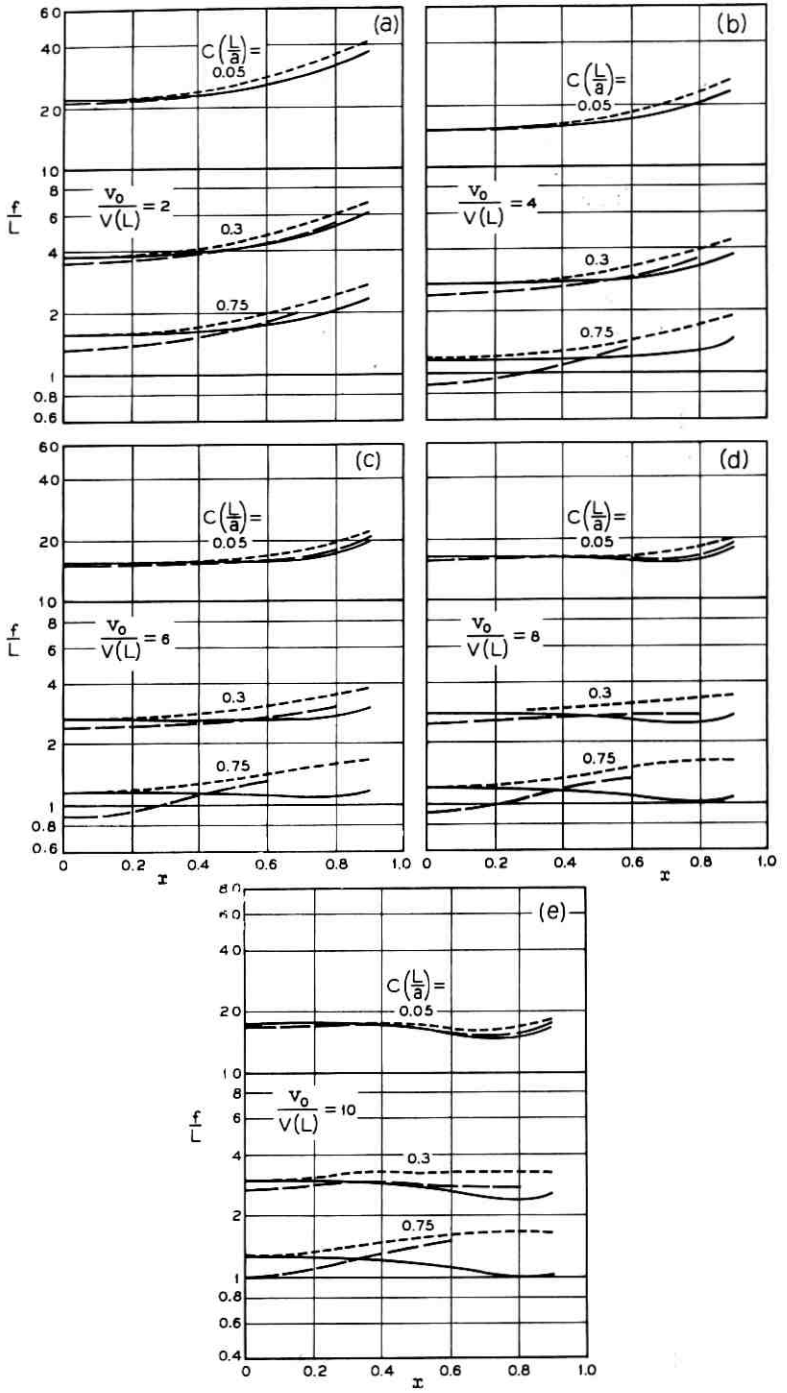


Fig. 6 — Normalized focal length f/L as a function of the ray's position $x = r/a$ for rays in the positive lens traveling (1.) with the gas stream (solid lines), and (2.) against the gas stream (dotted lines); and for rays in the negative lens traveling with the gas stream (dash-dotted lines).

known[3,7] if we can assume that the input gas is at a constant temperature.

$$T_3(x,u) = T_W + 2(T_W - T_i) \sum_{n=0}^{\infty} \frac{R_n(x)}{\beta_n \left(\frac{\partial R_n}{\partial \beta}\right)_{\substack{x=1 \\ \beta=\beta_n}}} \exp\left(-\beta_n^2/u\right) \qquad (29)$$

$$\infty > u > W$$

T_W = either T_h, temperature of hot tube, or T_c, temperature of cold tube.
T_i = input temperature to the hot or cold tubes.
W = $v_0/V(L)$ with L length of hot or cold tubes.

The temperature distribution, T_4, in the insulating sections is given in terms of U-functions and their eigenvalues γ which are defined in the appendix.

$$T_4(x,z) = A_0 + \sum_{n=1}^{\infty} A_n U_n(x) \exp\left[-\gamma_n^2 \left(\frac{1}{u} - \frac{1}{W}\right)\right] \qquad (30)$$

$$W > u > \tfrac{1}{2}W.$$

The expansion coefficients have to be determined from the condition

$$T_3(x,W) = T_4(x,W). \qquad (31)$$

Since the exponential functions appearing in (29) and (30) decrease very rapidly with decreasing values of u, it is sufficient to consider only the first term in the expansion at the end of each tube. This is justified if

$$0 \leqq W < 10. \qquad (32)$$

Condition (31) leads to

$$A_0 = T_W - 8(T_W - T_i) \frac{R_0'(1) \exp\left(-\beta_0^2/W\right)}{\beta_0^3 \left(\frac{\partial R_0}{\partial \beta}\right)_{\substack{x=1 \\ \beta=\beta_0}}} \qquad (33a)$$

and for $n \neq 0$

$$A_n = -4(T_W - T_i) \frac{\gamma_n R_0'(1) \exp\left(-\beta_0^2/W\right)}{\beta_0(\gamma_n^2 - \beta_0^2) \left(\frac{\partial R_0}{\partial \beta_0} \cdot \frac{\partial U'}{\partial \gamma_n}\right)_{x=1}} \qquad (33b)$$

where we have used the notation $F' = dF/dx$.

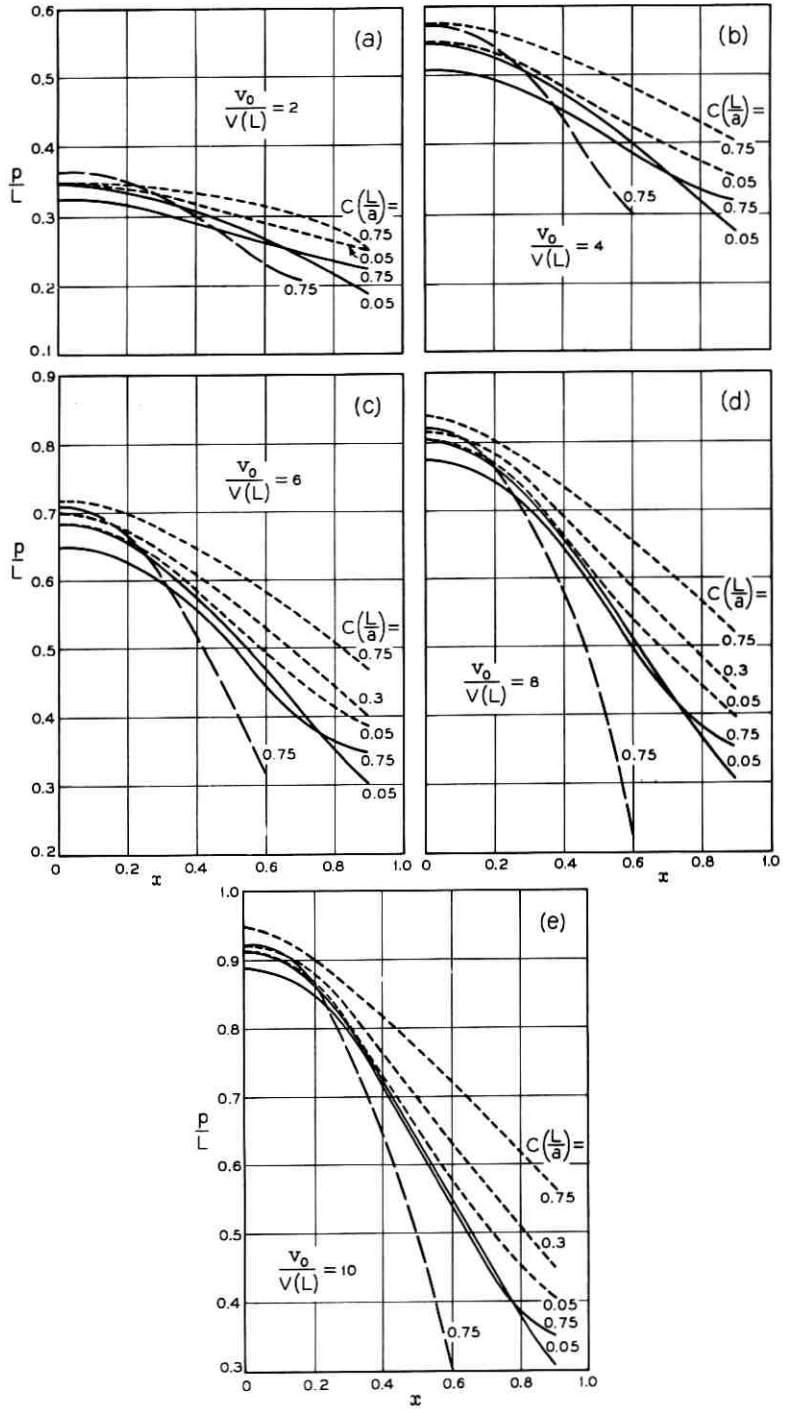


Fig. 7 — The principal surface for rays in the positive lens traveling (1.) with the gas stream (solid lines), and (2.) against the gas stream (dotted lines); and for rays in the negative lens traveling with the gas stream (dash-dotted lines).

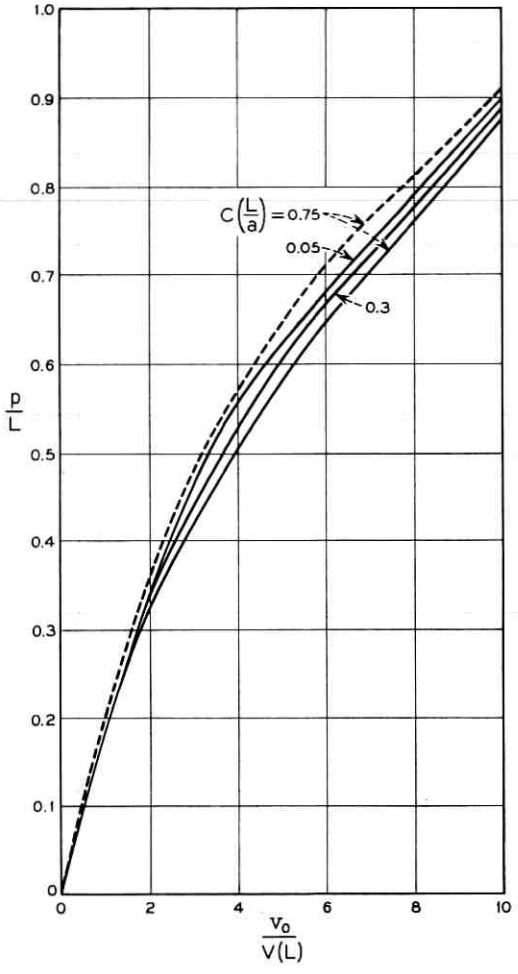


Fig. 8 — The dependence of the point of the principal surface at $x = 0.1$ on the gas velocity. The solid lines represent the positive lens, the dash-dotted lines the negative lens for rays traveling with the gas stream.

The temperature distribution in the hot and cold tubes can be inferred from curves shown in Ref. 3. The temperature distribution in the insulating tubes is shown in Fig. 10(a) for $\sigma(L/a) = 0.15$ ($W = 6.67$) and in Fig. 10(b) for $\sigma(L/a) = 0.05$ ($W = 20$) for various values of $\sigma[(z - L)/a]$. $z - L$ is the length coordinate counting from the beginning of the insulating tube. The curves show the temperature as a function of radius at different distances from the input to the insulating

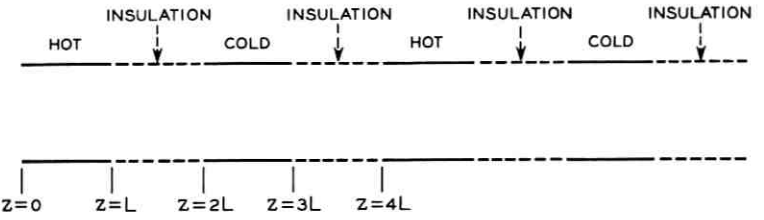


Fig. 9 — The hot, cold and insulating tubes comprising the "extended periodic structure".

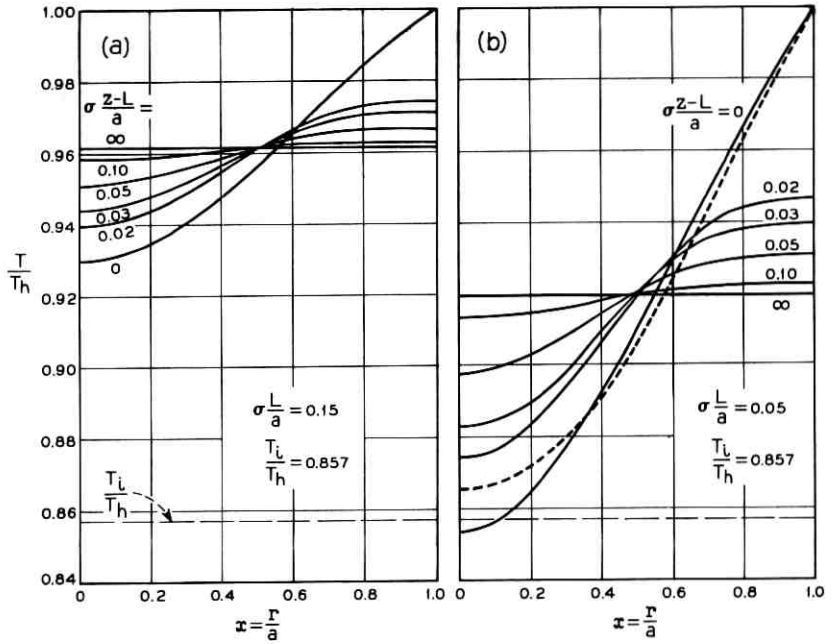


Fig. 10(*a*) — Temperature distribution in the insulating tubes of the extended periodic structure following a hot tube. $\sigma[(z - L)/a]$ is the normalized length measured from the beginning of the insulating tube. The ratio of input temperature to the preceeding hot tube T_i over its wall temperature T_h is $T_i/T_h = 0.857$. The normalized length of the tubes is $\sigma L/a = 0.15$.

Fig. 10(*b*) — Same as Fig. 10(*a*). $\sigma L/a = 0.05$. The dotted line is the actual temperature distribution at the end of the hot tube while the line with $\sigma\,[\,(z - L)/a] = 0$ is the temperature distribution at the end of the hot tube which results from dropping all but the first term in the series expansion of (29).

tube. At $z - L = 0$, the temperature distribution is identical to that at the output end of the hot tube feeding the insulating tube. It is apparent that the temperature equalizes rather rapidly. For practical purposes we can say that the temperature has reached a constant value at $\sigma[(z - L)/a] \geqq 0.1$. If we consider insulating tubes of length L, equal to the length of the hot or cold tubes feeding them, we obtain constant output temperatures of the insulating tubes for values $0 \leqq W < 10$. The hot and cold tubes are fed by gas at a constant temperature as long as these conditions are met. Therefore, we are justified in using (29) which has been derived for the case that the gas at the input end of the tube is at a constant temperature.

In the periodic structure of Fig. 9, the input temperature T to the hot and cold tubes are not arbitrary. They adjust themself to satisfy the periodicity condition

$$(T_3(x,u))_{z=0} = (T_4(x,u))_{z=4L}. \tag{34}$$

With the help of (34) we can calculate the input temperature T_{ih} of the hot or T_{ic} of the cold tube from (29), (30) and (33).

We obtain from (30) and (33a), for the constant output temperature of the insulating tube following the hot tube, ($W < 10$ is assumed so that all exponential terms $\exp(-\gamma_n^2/W)$ can be neglected)

$$(T_4(x))_{z=2L} = T_{ic} = T_h - 8(T_h - T_{ih}) \frac{R_0'(1) \exp(-\beta_0^2/W)}{\beta_0^3 \left(\frac{\partial R_0}{\partial \beta}\right)_{\substack{x=1 \\ \beta=\beta_0}}}$$

and also

$$(T_4(x))_{z=4L} = T_{ih} = T_c - 8(T_c - T_{ic}) \frac{R_0'(1) \exp(-\beta_0^2/W)}{\beta_0^3 \left(\frac{\partial R_0}{\partial \beta}\right)_{\substack{x=1 \\ \beta=\beta_0}}}.$$

Here we have two equations which allow us to determine the two quantities T_{ih} and T_{ic}.

It is convenient to express them in the form

$$\frac{T_h - T_{ih}}{T_h - T_c} = \left\{1 + 8 \frac{R_0'(1) \exp(-\beta_0^2/W)}{\beta_0^3 \left(\frac{\partial R}{\partial \beta}\right)_{\substack{x=1 \\ \beta=\beta_0}}}\right\}^{-1} \tag{35a}$$

and

$$\frac{T_{ic} - T_c}{T_h - T_c} = \frac{T_h - T_{ih}}{T_h - T_c}. \tag{35b}$$

A plot of (35a) as a function of $W = v_0/V(L)$ is shown in Fig. 11.

4.2 *Focal Length and Principal Surface*

The following graphical representations show the focal length and principal surface of one hot and insulating or one cold and insulating tube. The extended periodic structure is thus transformed into a system of equivalent positive and negative lenses in the same way as in the case of the simple periodic structure. Fig. 12 shows the dependence of the normalized focal length f/L on the normalized flow velocity $v_0/V(L)$ for a ray entering at $r/a = 0.1$. The length L is that of the hot tube and not the total length of the combination of hot and insulating tubes which has the length $2L$. The solid curve in Fig. 12 shows the focal length of the combination of hot and insulating tubes while the dotted curve shows the focal length of the hot tube alone for comparison. It is obvious that the insulating tube adds to the focusing power of the gas lens. We terminated the curves at $v_0/V(L) = 10$ since we wanted to remain in the domain of (32) where our simplifying assumptions used to calculate the temperature distribution are valid. Fig. 13 shows the corresponding

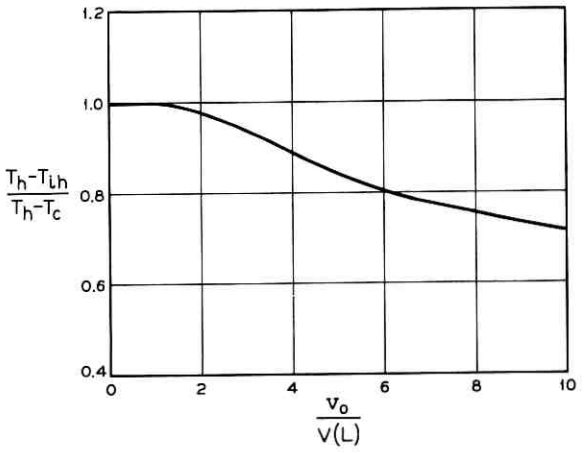


Fig. 11 — Temperature difference between the hot tube T_h and the input temperature to the hot tube T_{ih} divided by the temperature difference $T_h - T_c$ between hot and cold tube as a function of normalized gas velocity $v_0/V(L)$.

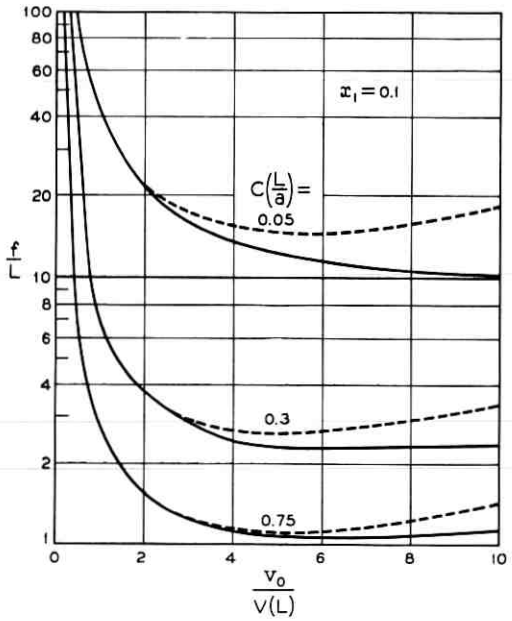


Fig. 12 — Normalized focal length f/L of the positive lenses of the extended periodic structure (solid lines) and of the hot tubes alone (dotted lines) as functions of normalized flow velocity $v_0/V(L)$.

curves for the cold tube resulting in a negative lens. A comparison of the two figures shows that the negative lens is more powerful than the positive lens for corresponding values of $C(L/a)$.

Figs. 14(a) and (b) show the dependence of the focal length (measured from the principal surface) on the input position, $x = r/a$, of the ray for the hot plus insulating tubes. The solid lines represent rays traveling in the same direction as the gas while the dotted curves show the focal length of rays traveling in opposite direction to the gas flow.

The shape of the principal surfaces for the hot plus insulating tube are shown in Fig. 15(a) and (b). The distance p of points on the principal surface is measured from the gas input end of hot tube. The solid and dotted lines again represent rays traveling with and against the gas flow respectively. The principal surface is far from being a plane for larger values of $v_0/V(L)$. The two principal surfaces do not coincide exactly which means that the lens has some optical thickness for larger values of $C(L/a)$.

The corresponding negative lens shows very similar distortions and has therefore been omitted.

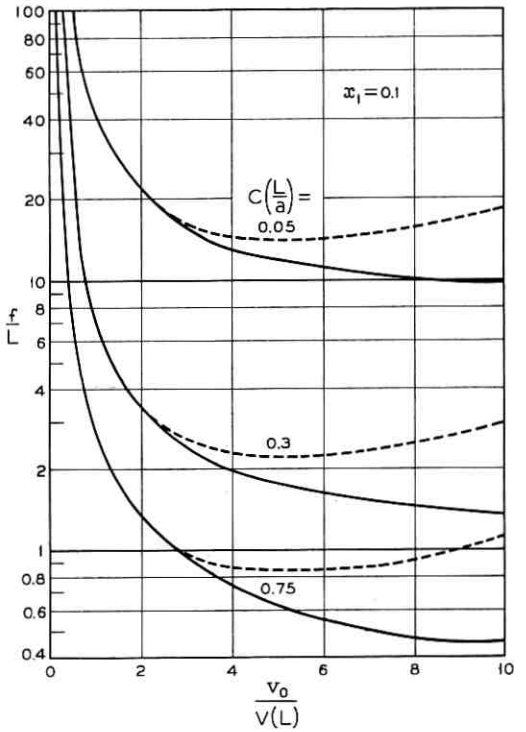


Fig. 13 — Normalized focal length f/L of the negative lenses of the extended periodic structure (solid lines) and of the cold tubes alone (dotted lines) as functions of the normalized flow velocity $v_0/V(L)$.

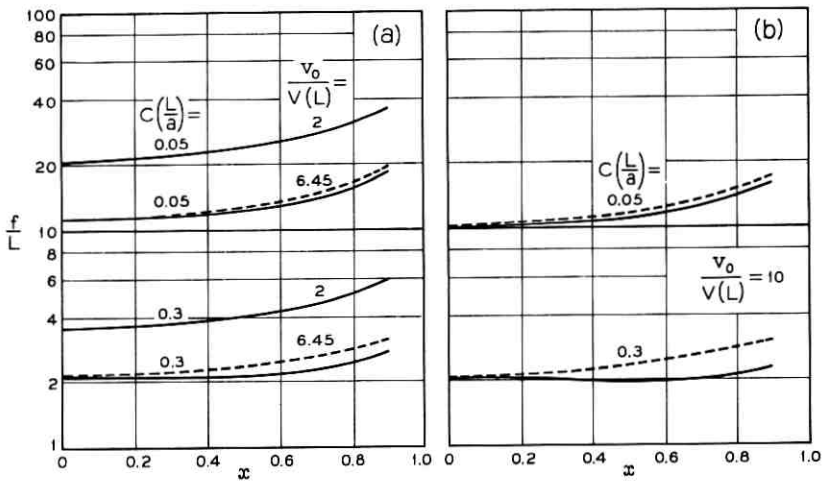


Fig. 14 — Dependence of focal length on the ray's input position for the positive lenses of the extended periodic structure for rays traveling with the gas flow (solid lines) and against the gas flow (dotted lines).

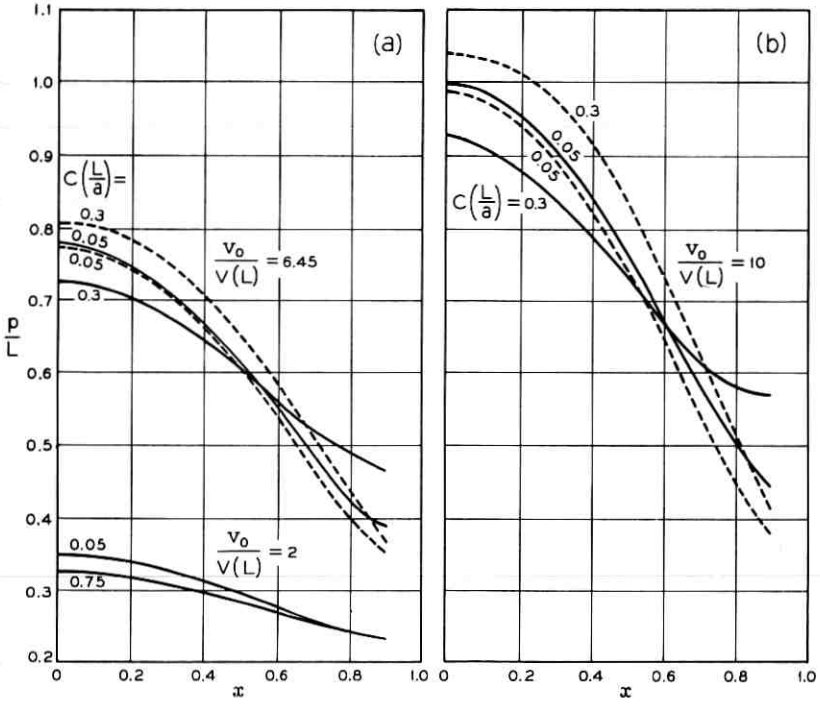


Fig. 15 — The principal surface of the positive lenses of the extended periodic structure for rays traveling with the gas flow (solid lines) and against the gas flow dotted lines.

V. COMPARISON OF THE TWO PERIODIC STRUCTURES

In order to compare the two periodic structures let us assume that their equivalent lenses are spaced at the same distance D. For the simple periodic structure D, the distance between a positive and the next negative lens is equal to the length L of the individual tubes. In the extended periodic structure $D = 2L$. It seems fair to compare both structures under the condition that the actual gas velocities in either one of them are identical. However, this assumption requires some rescaling of the data of the extended structure. If we operate the simple structure at a certain value of $v_0/V(D)$ the corresponding value for the extended structure will be different since v_0 is the same but in the extended structure $D = 2L$. It is apparent that

$$\left(\frac{v_0}{V(L)}\right)_{\text{extended}} = \frac{v_0}{V\left(\frac{D}{2}\right)} = 2\,\frac{v_0}{V(D)}\,. \tag{36}$$

A similar transformation has to be done on C. A certain value of $C(D/a)$ of the simple periodic structure corresponds to a value of

$$C\left(\frac{L}{a}\right)_{\text{extended}} = C\left(\frac{D}{2a}\right) = \frac{1}{4}C\left(\frac{D}{a}\right). \tag{37}$$

Since f/L is nearly proportional to C^{-1} for small values of C it is convenient to compare the values of

$$C\left(\frac{D}{a}\right)\cdot\frac{f}{D} = 2\left[C\left(\frac{L}{a}\right)\frac{f}{L}\right]_{\text{extended}}.$$

This comparison is shown in Fig. 16. For the same values of $T_h - T_c$ and v_0, the extended structure has the longer focal length because its active hot (or cold) tube is only half as long as that of the simple structure. The curve of Fig. 16 for the extended structure is not very

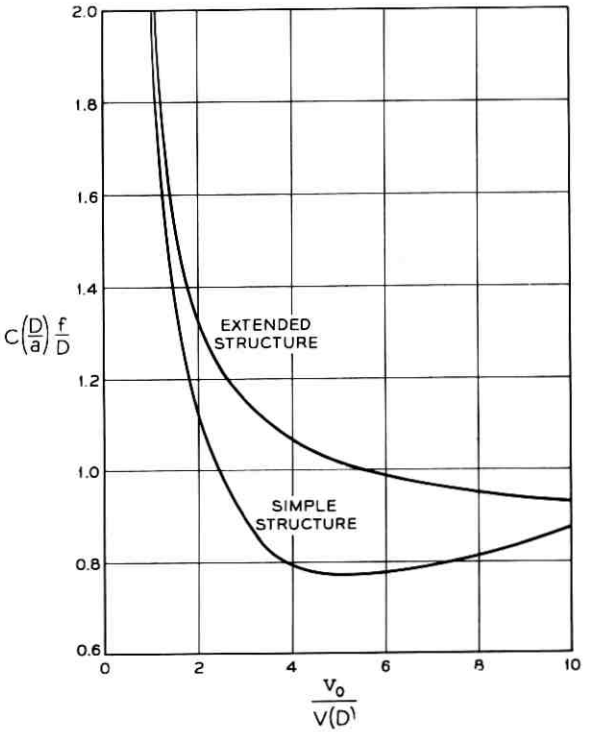


Fig. 16 — Comparison between the focal lengths of positive lenses of the simple and extended periodic structures as functions of the same normalized gas velocity $v_0/V(D)$.

accurate for values of $v_0/V(D) > 5$ because the value of $v_0/V(D) = 10$ corresponds to $v_0/V(L) = 20$ for the extended structure. For such a large value of the normalized flow velocity, our assumption of a constant input temperature to the hot (or cold) tube is incorrect.

To be able to compare the efficiencies of the two systems we need to know the power consumption of one positive lens of either of them assuming that it requires no additional power expenditure to keep the cold tubes at the temperature T_c.

This power consumption is given by

$$P = 2\pi\rho c_p a^2 \int_0^1 \{[T(x,u)]_{z=L} - [T(x,u)]_{z=0}\} x v(x)\, dx \qquad (38)$$

$v(x)$ is the gas velocity as a function of x. For viscous, laminar flow

$$v(x) = v_0(1 - x^2). \qquad (39)$$

Using (19) and (24), we obtain for the power consumption per hot tube of the simple periodic structure

$$\frac{2P_s}{\pi kD(T_h - T_c)} = \frac{v_0}{V(D)} \left\{ 1 - \frac{16R'(1)}{\beta_0^3 \left(\frac{\partial R_0}{\partial \beta}\right)_{x=1} (1 + \exp(\beta_0^2/W))} \right\} \qquad (40)$$

and with the help of (29) for the corresponding power consumption in the structure composed of extended tubes (assuming that the first term in the series of (29) describes the temperature distribution at $z = L$ sufficiently well), we obtain

$$\frac{2P_{ext}}{\pi kD(T_h - T_c)} = \frac{v_0}{V(D)} \frac{T_h - T_{ih}}{T_h - T_c} \left\{ 1 - \frac{8R_0'(1)}{\beta_0^3 \left(\frac{\partial R_0}{\partial \beta}\right)_{x=1}} \exp(-\beta_0^2/W) \right\}. \qquad (41)$$

The expression $(T_h - T_{ih})/(T_h - T_c)$ has to be substituted from (35a). The quantities represented by (40) and (41) are plotted in Fig. 17 as functions of $v_0/V(D)$. The positive lenses of the extended structure consume less power than those of the simple structure for equal amounts of gas flowing through them. This is not surprising considering that the hot tubes of the extended structure are only half as long as those of the simple structure.

To compare the two structures we require that their lenses have equal focal length which we achieve by adjusting the temperature difference between the hot and cold tubes. We then plot the ratio of the resulting power consumptions. This plot is shown in Fig. 18. The extended structure requires less power than the simple one and this ratio improves as

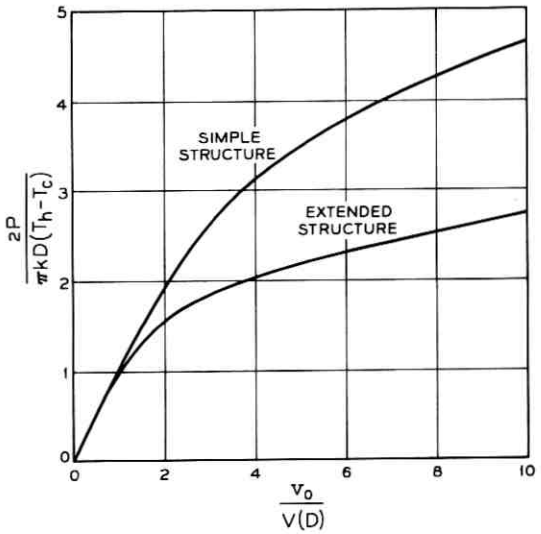


Fig. 17 — Comparison between the power consumptions of the positive lenses of the simple and extended periodic structures as functions of the same normalized gas velocities $v_0/V(D)$.

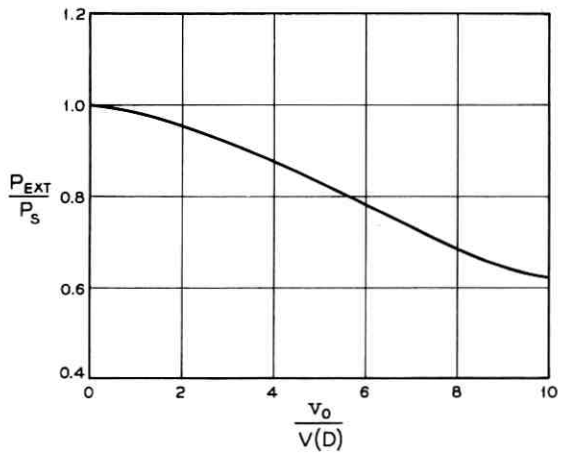


Fig. 18 — The ratio of the power consumption of the positive lenses of the extended periodic structure over the power consumption of the positive lenses of the simple periodic structure for equal focal length as a function of normalized gas velocity $v_0/V(D)$.

the flow velocity increases. It should be remembered, however, that the focal length of the extended structure (Fig. 16) is inaccurately represented for values $v_0/V(D) > 5$. This inaccuracy carries over to Fig. 18. Nevertheless, it is clear that the efficiency of the extended structure is more than 20 per cent higher than that of the simple structure for $v_0/V(D) > 6$. The additional focusing which the insulating tube sections provide pays off to some extent to make the extended structure more efficient. The improvement is not as large, however, as one might have hoped and it may be considerably poorer if the insulating tubes have the finite heat conductivity of a real material.

VI. ACKNOWLEDGMENT

The author gratefully acknowledges the help of Mrs. C. L. Beattie who wrote the machine programs for the calculation of the R- and U-functions and for part of the ray tracing.

APPENDIX

The functions to be discussed in the appendix are solutions of the differential equation

$$\frac{d^2F}{dx^2} + \frac{1}{x}\frac{dF}{dx} + k^2(1 - x^2)F = 0. \tag{42}$$

The solutions of (42) are related to Whittaker's functions $W_{\kappa,\mu}$ by

$$F = \frac{1}{x} W_{(k/4),0}(kx^2).$$

The differential equation (42) stems from an approximate formulation of a heat transfer problem.[7] Assume that a gas at a different temperature is blown into a tube with a circular cross section. The gas is supposed to flow as a viscous fluid in laminar flow with a velocity profile

$$v(r) = v_0\left(1 - \frac{r^2}{a^2}\right) \tag{43}$$

r = distance from tube axis
a = tube radius.

The stationary state of the temperature distribution T is obtained from

$$\alpha\nabla^2 T = \mathbf{v}\cdot\boldsymbol{\nabla}T \tag{44}$$

with α being a constant which contains the heat conductivity, density

and specific heat at constant pressure — all of which are assumed to be temperature independent which, strictly speaking, is not true.

The velocity has only a longitudinal component v_z given by (43). In polar coordinates r, φ, z taking $\partial/\partial\varphi = 0$ (44) can be written

$$\alpha\left\{\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{\partial^2 T}{\partial z^2}\right\} = v_0\left(1 - \frac{r^2}{a^2}\right)\frac{\partial T}{\partial z}. \tag{45}$$

It is often permissible to neglect $\partial^2 T/\partial z^2$ compared to the term on the right hand side of (45). Taking

$$r = ax \tag{46}$$

and

$$T(r,z) = F(x)\exp(-\lambda z) \tag{47}$$

we get from (45)

$$\frac{d^2F}{dx^2} + \frac{1}{x}\frac{dF}{dx} + \frac{a^2 v_0 \lambda}{\alpha}(1 - x^2)F = 0. \tag{48}$$

Finally, taking

$$\lambda = k^2\frac{\alpha}{a^2 v_0} \tag{49}$$

we recognize that (48) is identical to (42).

We are interested in two types of heat transfer problems:

(1.) The tube through which the gas flows may be kept at a constant temperature T_w. We then have the boundary condition $T(a,z) = T_w$. It is more convenient to introduce a new variable

$$\theta(r,z) = T_w - T(r,z). \tag{50}$$

We can replace T by θ without changing any of equations (44) through (47). However, the boundary condition now becomes simply

$$\theta(a,z) = 0. \tag{51}$$

The functions satisfying this boundary condition are designated by

$$F(x) = R(x) \tag{52}$$

and (51) becomes

$$R(1) = 0. \tag{53}$$

The R functions have been studied in some detail.[7,8]

(2.) The second type of problem involves a tube which is a perfect

heat insulator. That means that no heat flows into or out of the walls. This assumption requires that

$$\left(\frac{\partial T}{\partial r}\right)_{r=a} = 0. \tag{54}$$

A second class of functions is obtained by setting $F(x) = U(x)$. The U-functions satisfy the same differential equation but are defined by the boundary condition

$$\left(\frac{dU}{dx}\right)_{x=1} = 0. \tag{55}$$

For convenience, both functions are normalized so that

$$R(0) = U(0) = 1. \tag{56}$$

The U-functions have not been studied to the knowledge of the author.

For series expansions of arbitrary heat distributions in terms of either the R or the U functions we need their orthogonality relations and any numerical evaluation of heat transfer problems requires the knowledge of the eigenvalues and numerical values of these functions.

The eigenvalues belonging to the R functions $R_n(x)$ will be designated as

$$k_n = \beta_n \tag{57}$$

and those of $U_n(x)$ will be designated by

$$k_n = \gamma_n . \tag{58}$$

Orthogonality Relations

Let F_n with eigenvalue k_n designate either an R_n or a U_n function. We proceed to show that

$$\int_0^1 x(1 - x^2)F_n(x)F_m(x)\,dx = 0 \qquad \text{for } n \neq m. \tag{59}$$

We have

$$F_n'' + \frac{1}{x}F_n' + k_n^2(1 - x^2)F_n = 0 \tag{60}$$

and

$$F_m'' + \frac{1}{x}F_m' + k_m^2(1 - x^2)F_m = 0. \tag{61}$$

We multiply (60) by xF_m and (61) by xF_n, subtract and integrate:

$$(k_n^2 - k_m^2) \int_0^1 x(1 - x^2)F_n(x)F_m(x)\,dx$$
$$= -\int_0^1 x\left\{\left(F_m\left(F_n'' + \frac{1}{x}F_n'\right)\right) - F_n\left(F_m'' + \frac{1}{x}F_m'\right)\right\}dx.$$

We perform partial integrations and obtain for the right hand side of this equation

$$-[xF_mF_n' - xF_nF_m']_0^1 + \int_0^1 \{F_mF_n' - F_nF_m' + x(F_m'F_n' - F_n'F_m')$$
$$- (F_mF_n' - F_nF_m')\}\,dx = (F_mF_n' - F_nF_m')_{x=1} = 0.$$

The last part of the equation follows from the boundary condition (53) or (55), depending whether F stands for an R or a U function. This calculation proves (59).

Next, we calculate the value of (59) in the case $n = m$.

$$\int_0^1 x(1 - x^2)[F_n(x)]^2\,dx = \lim_{k_m \to k_n}\left(\frac{F_mF_n' - F_nF_m'}{k_n^2 - k_m^2}\right)_{x=1}$$
$$= \left(\frac{\dfrac{\partial F_n}{\partial k}F_n' - F_n\dfrac{\partial}{\partial k}F_n'}{2k_n}\right)_{\substack{x=1\\k=k_n}}.$$

For $F_n = R_n$ we get

$$\int_0^1 x(1 - x^2)[R_n(x)]^2\,dx = \frac{1}{2\beta_n}\left[\frac{\partial R}{\partial \beta}R_n'\right]_{\substack{x=1\\\beta=\beta_b}} \tag{62}$$

and for $F_n = U_n$

$$\int_0^1 x(1 - x^2)[U_n(x)]^2\,dx = -\frac{1}{2\gamma_n}\left[U_n\frac{\partial}{\partial \gamma}U_n'\right]_{\substack{x=1\\\gamma=\gamma_n}}. \tag{63}$$

Finally, we need to determine the value of the integral over the products R_nU_m:

$$R_n'' + \frac{1}{x}R_n' + \beta_n^2(1 - x^2)R_n = 0 \tag{64}$$

$$U_m'' + \frac{1}{x}U_m' + \gamma_n^2(1 - x^2)U_m = 0, \tag{65}$$

$$(\beta_n^2 - \gamma_m^2) \int_0^1 x(1 - x^2) R_n(x) U_m(x) dx$$

$$= \int_0^1 \{x(R_n U_m'' - U_m R_n'') + R_n U_m' - U_m R_n'\} dx$$

$$= [R_n U_m' - U_m R_n']_{x=1}$$

or using (53) and (55),

$$\int_0^1 x(1 - x^2) R_n(x) U_m(x) dx = \frac{1}{\gamma_m^2 - \beta_n^2} [U_m R_n']_{x=1}. \qquad (66)$$

Calculation of the R and U Functions and Their Eigenvalues

We make the series expansion

$$F(x) = \sum_{v=0}^{\infty} C_{2v} x^{2v}. \qquad (67)$$

For the problem of interest to us $F(x)$ has to be an even function of x, for that reason only even powers of x appear in (67). The normalization

$$F(0) = 1$$

requires that (68)

$$C_0 = 1.$$

The substitution of (67) into (42), using (68), leads to

$$C_2 = -\tfrac{1}{4} k^2$$

and

$$C_{2v} = \frac{k^2}{(2v)^2} \{C_{2v-4} - C_{2v-2}\} \qquad \text{for } v \geqq 2. \qquad (69)$$

The parameter k has to be chosen so that either $F(1) = 0$ or $F'(1) = 0$ results, depending whether k and F shall represent β and R or γ and U respectively.

The fact that k^2 enters all coefficients C_{2v} makes the determination of β_n and γ_n very tedious.

A further difficulty results from the fact that the coefficients C_{2v} grow to very large values particularly for the larger values of β_n and γ_n before they decrease again. The series (67) does not converge readily for values of x close to 1. In fact, it proved impossible to compute more than the first eight R and U functions from (67) on the IBM 7094 com-

puter even using double precision since the absolute value of R and U remains between zero and one but the coefficients C_{2v} grow to values above 10^{20}. The series (67) can be used to compute R_n and U_n for x in the range $0 \leqq x \leqq 0.5$ since the powers of x decrease rapidly enough to keep the value of the product $C_{2v}\, x^{2v}$ within manageable proportions.

However, in order to cover the whole range $0 \leqq x \leqq 1$ it proved necessary to use the following series expansion:

$$F(x) = \sum_{v=0}^{\infty} D_v y^v \quad \text{with} \quad y = 1 - x \tag{70}$$

to calculate R and U in the range $0.5 \leqq x \leqq 1$.

Equation (42) expressed in terms of y reads

$$(1 - y)\frac{d^2F}{dy^2} - \frac{dF}{dy} + k^2(2y - 3y^2 + y^3)F = 0. \tag{71}$$

The coefficients D_0 and D_1 have to be properly chosen to satisfy the boundary conditions at $x = 1$. For $F = R$ we require $R(1) = 0$ so that

$$D_0 = 0$$

results. For $F = U$ we require $U'(1) = 0$ so that

$$D_1 = 0$$

results.

The substitution of (70) into (71) yields

$$\begin{aligned} &D_2 = \tfrac{1}{2}D_1\,, \qquad D_3 = \tfrac{1}{3}D_1 - \tfrac{1}{3}k^2D_0\,, \qquad D_4 = (\tfrac{1}{4} - \tfrac{1}{6}k^2)D_1 \\ &D_v = \frac{1}{v(v-1)}\{(v-1)^2 D_{v-1} - k^2(2D_{v-3} - 3D_{v-4} + D_{v-5})\}. \end{aligned} \tag{72}$$

The eigenvalue $k = \beta$ or $k = \gamma$ and the coefficient D_1 or D_0 must be chosen so that F as well as F' are continuous at $x = 0.5$ where both series expansions should coincide.

By breaking the range of x into two parts and using different series expansions to cover both parts of the range it was possible to compute the function and their eigenvalues. Table I shows the eigenvalues β_n and γ_n as well as the values of $\partial R_n/\partial\beta$, $-D_1 = R_n'$. $(\partial/\partial\gamma)U_n'$ and $D_0 = U_n$ all taken at $x = 1$. These values are needed to evaluate the integrals (62), (63) and (66).

The values of $\partial R_n/\partial\beta$ and $\partial U_n'/\partial\gamma$ were obtained from differentiation of the series (67) and evaluating it at $x = 1$. The terms of the differ-

entiated series grow very large so that only the first eight values of $\partial R/\partial\beta$ and the first six values of $\partial U'/\partial\gamma$ could be obtained. The remaining values of $\partial R/\partial\beta$ were calculated from the approximation[8]

$$\left(\frac{\partial R}{\partial \beta}\right)_{\substack{x=1\\ \beta=\beta_n}} = (-1)^n \frac{\pi}{6^{2/3}\,\Gamma\left(\frac{2}{3}\right)\beta_n^{1/3}} \tag{73}$$

which is in good agreement with the values obtained by machine calculation for larger values of n. An approximation of $\partial U/\partial\gamma$ can be obtained by using approximations similar to the ones used for the R functions in Ref. 8. One gets

$$\left(\frac{\partial U'}{\partial \gamma}\right)_{\substack{x=1\\ \gamma=\gamma_n}} = -(-1)^n \frac{\pi\gamma^{\frac{1}{3}}}{6^{\frac{1}{3}}\Gamma\left(\frac{1}{3}\right)}\,. \tag{74}$$

However, this approximation is not very good for $n \leqq 15$ so that we used the equation

$$\left(\frac{\partial U'}{\partial \gamma}\right)_{\substack{x=1\\ \gamma=\gamma_n}} = -(-1)^n \frac{\pi\gamma^{\frac{1}{3}}}{6^{\frac{1}{3}}\Gamma\left(\frac{1}{3}\right)} + \sum_{\mu=1}^{6} \frac{A_\mu}{\gamma^\mu}\,. \tag{75}$$

The coefficients A_μ were determined from the first six values of $\partial U'/\partial\gamma$ which were obtained from a machine calculation. Their values are given at the bottom of Table I.

TABLE I

n	β_n	$R_n'(1)$	$\left(\frac{\partial R(1)}{\partial\beta}\right)_{\beta=\beta_n}$	γ_n	$U_n(1)$	$\left(\frac{\partial U'(1)}{\partial\gamma}\right)_{\gamma=\gamma_n}$
0	2.70436	−1.01430	−0.50090	0	1	
1	6.67903	1.34924	0.37146	5.06750	−0.492517	0.97816
2	10.67338	−1.57232	−0.31826	9.15760	0.395509	−1.24720
3	14.6711	1.74600	0.28648	13.1972	−0.345874	1.43522
4	18.6699	−1.89090	−0.26449	17.2202	0.314047	−1.58486
5	22.6691	2.01647	0.24799	21.2355	−0.291252	1.71127
6	26.6686	−2.12814	−0.23491	25.2465	0.273806	−1.82164
7	30.6682	2.22038	0.22485	29.2549	−0.259853	1.92042
8	34.6679	−2.32214	−0.21548	33.2615	0.248332	−2.01037
9	38.6676	2.40274	0.20779	37.2669	−0.238591	2.09330
10	42.6667	−2.48992	−0.20108	41.2714	0.230199	−2.17045
11	46.6667	2.56223	0.19516	45.2752	−0.222863	2.24275
12	50.6667	−2.64962	−0.18988	49.2785	0.216371	−2.31088
13	54.6667	2.70216	0.18513	53.2813	−0.210569	2.37539
14	58.6667	−2.76421	−0.18083	57.2837	0.205216	−2.43671

$A_1 = -4.881355$ $\quad A_4 = 2.838701\ 10^4$
$A_2 = 1.536461\ 10^2$ $\quad A_5 = -1.420240\ 10^5$
$A_3 = -2.838383\ 10^3$ $\quad A_6 = 2.728875\ 10^5$

An approximate formula for β is[8]

$$\beta_n = 4n + \tfrac{8}{3}. \tag{76}$$

The β values of Table I for $n \geqq 11$ have been computed from (76).

The lowest order U function is a constant

$$U_0 = 1 \quad \text{with} \quad \gamma_0 = 0. \tag{77}$$

The γ values should converge to

$$\gamma_n = 4n + \tfrac{4}{3}. \tag{78}$$

This expression can be derived by methods analogous to those used in Ref. 8. For unknown reasons this approximation is much poorer than that for β_n. However, it appears from the values of Table I that γ_n will converge to (78) for very high values of n.

REFERENCES

1. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, IRE Trans. *AP-9*, May, 1961, pp. 248–256.
2. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43*, July, 1964, pp. 1469–1475.
3. Marcuse, D., and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J. *43*, July, 1964, pp. 1759–1782.
4. Marcuse, D., Theory of a Thermal Gradient Gas Lens. Paper delivered at the 1965 IEEE-G-MTT Symposium, to be published in the IEEE Trans-G-MTT, Nov., 1965.
5. Miller, S. E., Alternating-Gradient Focusing and Related Properties of Convergent Lens Focusing, B.S.T.J., *43*, July, 1964, pp. 1741–1758.
6. Born, M., and Wolff, E., *Principles of Optics*, Pergamon Press, New York, 1959, p. 121 (Equation (2)).
7. Jakob, M., *Heat Transfer, 1*, John Wiley, New York, 1949, pp. 451–464.
8. Sellers, J. R., Tribus, M., and Klein, J. S., Heat Transfer to Laminar Flow in a Round Tube or Flat Conduit — The Graetz Problem Extended, Trans. Am. Soc. Mec. Eng., *78*, 1956, pp. 441–448.

Growth of Oscillations of a Ray about the Irregularly Wavy Axis of a Lens Light Guide

By D. W. BERREMAN

(Manuscript received July 20, 1965)

If a ray is launched in a direction coincident with the axis of a lens light guide whose axis is somewhat wavy, the ray will soon begin to oscillate about the axis. The amplitude of these oscillations will grow in proportion to the square root of the product of the distance from the origin and the natural wave number of the oscillations, on the average. The growth rate is proportional to the amplitude of the components of the spectrum of the waves in the guide axis in the immediate vicinity of the natural wave number of the ray oscillations. The rest of the spectrum of waves in the guide does not contribute appreciably to the growth of ray oscillations after the first few oscillations. A tractable analytic expression with four adjustable parameters is used to approximate the wave spectrum of the shape of the guide axis. The expression is used to illustrate the relations between various factors such as mechanical stiffness of the guide, spectrum of external forces, over-all amplitude of waves in the guide, etc., and the rate of growth of the ray oscillations. Estimates of the order of magnitude of the oscillation growth rate in some realistic models of light guides are made from these relations.

1. INTRODUCTION

The possibility of guiding a beam of light over a long distance through a series of lenses or a continuous lens-like medium has recently received considerable attention because such a system might be useful in communication. One problem that has only recently come under investigation is the statistical growth rate of the oscillation of an initially paraxial ray about the axis of such a guide when the axis is crooked in a somewhat random but partially coherent way.[1,2] In this paper I will show that the ray will oscillate about the axis with ever-increasing amplitude, on the average.

The oscillations of the ray about the axis are analogous to the os-

TABLE I

Input parameters and results of computations in M.K.S. units: kilograms, (kg), meters, (m), and Newtons, (Nt). The first four input parameters describe the pipe; the next three, the nature of the support; and the last three, the optical elements.

Input	*Continuous*	*Either*	*Lenses*	*Continuous*	*Either*	*Lenses*
(1) ρ (kg/m^3)		7.8×10^3			same	
(2) M (Nt/m^2)		2.0×10^9			same	
(3) OD (m)		0.1			same	
(4) ID (m)		0.09			same	
(5) m (number)		11			same	
(6) n (number)		8			same	
(7) λ_p (m)		1.0			4.0	
(8) C (m^{-2})	0.25		—	same		—
(9) c (m^{-1})	—		0.637	—		same
(10) L (m)	—		3.14	—		same
Results						
(1) $\sqrt{C}$ (m^{-1})	0.5		—	0.5		—
(2) k_c (m^{-1})	—		0.5	—		0.5
(3) k_0 (m^{-1})		3.27			0.819	
(4) λ_y (m)		3.68			14.73	
(5) σ (Nt·m^2)		3380			same	
(6) $\bar{\bar{p}}$ (Nt/m)		114			same	
(7) $\bar{\bar{y}}$ (m)		0.407×10^{-2}			10.43×10^{-3}	
(8) P_0 (Nt/m)		222			same	
(9) Y ($\sqrt{C}$) (m)	0.572×10^{-3}		—	146.1×10^{-3}		—
(10) $S(k_c)$ (m)	—		1.040×10^{-3}	—		146.2×10^{-3}
(11) $\delta/\sqrt{x}$ (m$^{\frac{1}{2}}$)	0.358×10^{-3}		0.830×10^{-3}	91.6×10^{-2}		116.7×10^{-3}

cillations of an undamped mechanical oscillator subject to a particular steady "noise spectrum" of forces. Spherical aberration in the guide would be analogous to anharmonicity in the mechanical oscillator. Distance in the optical problem corresponds to time in the mechanical problem.

The dominant term in the expression for the amplitude of ray oscillations is proportional to the square root of the distance from the origin, if the amount and spectral form of the crookedness is constant and if the lenses have no spherical aberration. The oscillations of the ray about the axis have a natural wave number, k_c. The growth rate of ray oscillations is also proportional to the square root of k_c. The third factor governing the growth rate is the amount of crookedness of the axis. A major portion of this paper is devoted to a mathematical description of the amount of crookedness of the axis of the guide. Only those components of the crookedness having approximately the same wave number as the natural oscillations of the beam have an appreciable influence on the growth of the oscillations beyond the first few oscillations.

Some numerical examples of growth rates of oscillations in guides having reasonable amounts of crookedness or waviness are given at the end of the paper and in Table I. The amount of crookedness in the examples was obtained by estimating the variation of forces on a pipe of reasonable stiffness lying on a rough surface. See Fig. 1. The pipe was assumed to be relatively straight before the irregular forces due to the rough surface were applied. This specific method of estimating crookedness does not limit the generality of the relations between the crookedness spectrum and the growth rate of the oscillations.

The results indicate that, unless such a light guide could be kept extremely straight or free of waves on a scale approximating the wavelength of natural oscillations of the beam, the beam would have to be recentered at frequent intervals* or the guide would have to have a very large aperture to avoid vignetting of an initially paraxial ray.

II. EQUATIONS OF THE RAY TRAJECTORY

In order to keep the analysis simple, I will suppose that the axis of the guide is wavy only in one dimension, y. The distance from a straight

* There is no physical reason why we could not redirect the beam down the axis of the guide at intervals with almost no loss of beam energy. For example, the position and direction of the main beam could be located at one point by reflecting a very small fraction of it into a photoelectric analyzing device with a very weakly reflecting mirror. The analyzing device could be used to control some prisms which would change the direction and displacement of the beam. The analyzing device could be a simple null device if the beam passed through the prisms shortly before reaching the analyzer. L. U. Kibler of Bell Telephone Laboratories described this idea in an unpublished memorandum in March, 1962.

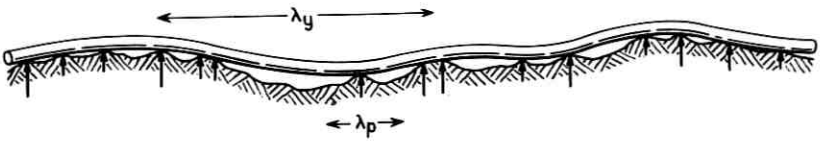


Fig. 1 — A guide on a rough surface that produces an irregular distribution of forces with a characteristic separation λ_p. This results in waves with a characteristic length λ_y.

coordinate axis, x, to the axis of the guide will be called $y(x)$ and the distance from the axis of the guide to the ray or any point in the guide will be called $\delta(x)$. See Fig. 2.

For the case of the continuous lens-like medium, suppose that the refractive index can be described as a function only of distance, δ, from the axis of the guide. A medium free of spherical aberration will be defined as one obeying the relation

$$n(\delta) = n_0 \exp(-C\delta^2/2) \approx n_0(1 - C\delta^2/2). \quad (1)$$

C is the *specific convergence*[3] of the guide, which is approximately the convergence, in diopters or reciprocal meters of a one-meter segment of the guide if that convergence is small compared to one diopter.

The equation of the trajectory of a ray in a medium of slowly, smoothly varying isotropic refractive index is[4]

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \text{grad } n. \quad (2)$$

In this equation ds is an element of length along the path of the ray, $\mathbf{r}$ is the vector position of a point on the ray and n is the (isotropic) refractive index at that point.

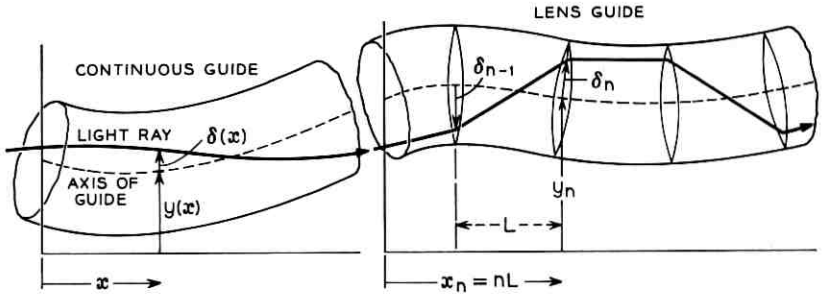


Fig. 2 — Symbols used to describe a continuous guide (left) and a thin lens guide (right).

If the slope of the ray, $d\delta/dx$, and the slope of the guide axis, dy/dx, both remain small, then (1) and (2) may be combined to give the following differential equation for the trajectory of the ray,

$$\frac{d^2\delta}{dx^2} + C\delta = -\frac{d^2y}{dx^2}. \tag{3}$$

This is like the familiar forced, undamped harmonic oscillator equation.

For a guide composed of a series of thin, aberration-free lenses of equal convergence, c, separated by a distance L, we can obtain the following analogous difference equation for the trajectory from simple geometrical construction.

$$\tfrac{1}{2}(\delta_{n+1} + \delta_{n-1}) - \delta_n + \frac{cL}{2}\delta_n = -(\tfrac{1}{2}(y_{n+1} + y_{n-1}) - y_n) \tag{4}$$

δ_n is the displacement of the ray from the axis of the nth lens and y_n is the displacement of that lens axis from the straight x-coordinate axis. (See Fig. 2.) Note that the product nL is equal to x. Hence, once we define a shape, $y(x)$, for the guide axis, the right-hand side of either (3) or (4) can be written down.

III. SOLUTIONS IN TERMS OF GUIDE SHAPE

Suppose we pretend that the shape of the guide is periodic with a length, Λ, for the periodicity. The path of the axis of the pipe can be represented by the Fourier series:

$$y(x) = \sum_{j=1}^{\infty} Y\left(\frac{2\pi j}{\Lambda}\right) \sin\left(\frac{2\pi jx}{\Lambda} - \varphi_j\right). \tag{5}$$

The phase factor φ_j is used to avoid writing cosine terms in the series. We can avoid a constant term ($j = 0$) by proper choice of the position $y = 0$. Making the substitution $k_j = 2\pi j/\Lambda$, we may write

$$y(x) = \sum_{j=1}^{\infty} Y(k_j) \sin(k_jx - \varphi_j). \tag{6}$$

3.1 *The Case of a Continuous Lens-Like Medium*

By substituting a trial solution of the form

$$\delta(x) = \sum_{j=1}^{\infty} A(k_j) \sin(k_jx - \psi_j) + B(k_j) \sin(\sqrt{C}x - \xi_j) \tag{7}$$

into (3), along with the shape spectrum (6), we obtain the results

$$A(k_j) = \frac{k_j^2 Y(k_j)}{C - k_j^2} \qquad \text{and} \qquad \psi_j = \varphi_j . \tag{8}$$

If we add the boundary conditions

$$\delta(0) = (d\delta/dx)_{x=0} = 0 \tag{9}$$

we find that

$$B(k_j) = -A(k_j) \sin \varphi_j / \sin \xi_j \tag{10}$$

where

$$\cot \xi_j = \frac{k_j}{\sqrt{C}} \cot \varphi_j . \tag{11}$$

Note that $\sqrt{C}$ is the natural angular wave number of oscillations of the ray about the axis of the guide if the guide is straight.

Inserting these results into (7) we finally obtain

$$\delta(x) = \sum_{j=1}^{\infty} \frac{k_j^2 Y(k_j)}{C - k_j^2} \left[\cos \varphi_j \left(\sin k_j x - \frac{k_j}{\sqrt{C}} \sin \sqrt{C} x \right) - \sin \varphi_j (\cos k_j x - \cos \sqrt{C} x) \right]. \tag{12}$$

At this point in the derivation we introduce the random, statistical character of the problem by asserting that the phase factors, φ_j, are random. Then we can no longer specify $y(x)$ or $\delta(x)$ but we can specify a mean square value of each and relate one to the other. Since the mean square value of a sinusoidal function is half the square of its absolute value, we can rewrite (5) as

$$\langle y^2(x) \rangle \equiv \bar{\bar{y}}^2 = \tfrac{1}{2} \sum_{j=1}^{\infty} Y^2(k_j). \tag{13}$$

Similarly, from (12) we obtain

$$\langle \delta^2(x) \rangle \equiv \bar{\bar{\delta}}^2(x) = \frac{1}{2} \sum_{j=1}^{\infty} \left(\frac{k_j^2 Y(k_j)}{C - k_j^2} \right)^2 \left[\left(\sin k_j x - \frac{k_j}{\sqrt{C}} \sin \sqrt{C} x \right)^2 + (\cos k_j x - \cos \sqrt{C} x)^2 \right]. \tag{14}$$

If we make the substitution

$$\epsilon_j = \frac{k_j}{\sqrt{C}} - 1 \tag{15}$$

and do quite a lot of algebra and trigonometry, we find that (14) can be written in the following form: (Remember that ϵ_j ranges from -1 to $+\infty$.)

$$\bar{\bar{\delta}}^2(x) = \frac{1}{2}\sum_{j=1}^{\infty} Y^2(\epsilon_j)\frac{(1+\epsilon_j)^4}{(2+\epsilon_j)^2}\left[\frac{4\sin^2(\epsilon_j\sqrt{C}x/2)}{\epsilon_j^2} + \frac{4\sin(\epsilon_j\sqrt{C}x/2)}{\epsilon_j} - \frac{\sin\epsilon_j\sqrt{C}x\sin 2\sqrt{C}x}{\epsilon_j} + \sin^2\sqrt{C}x\right]. \tag{16}$$

If the length, Λ, of the periodicity of the shape of the guide is allowed to become very large, we may replace the summation signs in (13) and (16) by integrations over the variable ϵ, and we may drop the index j.

If $\epsilon^4 Y^2(\epsilon)$ remains finite as ϵ approaches infinity, then large values of ϵ will not cause the integral for $(\bar{\bar{\delta}}(x))^2$ to diverge. (Slightly weaker conditions could be applied.)

For large values of x, the first of the four terms in square brackets in (16) will dominate if $Y(\epsilon = 0)$ is nonzero. Thus we obtain the following result

$$\bar{\bar{\delta}}^2(x) \approx \frac{1}{2}\int_{-1}^{\infty} Y^2(\epsilon)\frac{(1+\epsilon)^4}{(2+\epsilon)^2}\cdot\frac{4\sin^2\epsilon\sqrt{C}x/2}{\epsilon^2}\,d\epsilon \to x\cdot\frac{\pi\sqrt{C}}{4}Y^2(\epsilon=0) \quad \text{for large } x. \tag{17}$$

The proper normalization of the function $Y(\epsilon)$ can be obtained from the following relation:

$$\bar{\bar{y}}^2 = \frac{1}{2}\int_{-1}^{\infty} Y^2(\epsilon)d\epsilon\,. \tag{18}$$

(Compare (13). It is assumed that the relative spectral distribution for $Y(\epsilon)$ is known, and we may also assume that the mean square value of y is known. See Section V for examples.)

3.2 *The Case of a Series of Lenses*

An exactly analogous but slightly more complicated procedure can be followed for the case of a series of lenses. We can keep the same function, (6), to describe the displacements of the lens centers if we let x be an integral multiple, n, of the lens separation, L. (See Fig. 2.) Equation (7) for the ray trajectory may similarly be retained with the same expression replacing x.

However, when (7) is substituted into (4) rather than (3), we obtain

a somewhat different expression for $A(k_j)$. The phase factors ψ_n are the same.

$$A(k_j) = \frac{(1 - \cos k_j L)Y(k_j)}{\frac{cL}{2} - (1 - \cos k_j L)} \quad \text{and} \quad \psi_j = \varphi_j . \tag{8'}$$

We use boundary conditions analogous to (9),

$$\delta(n = 0) = \delta(n = 1) = 0 \tag{9'}$$

and we obtain expressions for B and ξ_j :

$$B(k_j) = - A(k_j) \frac{\sin \varphi_j}{\sin \xi_j} \tag{10}$$

and

$$\cot \xi_j = \frac{\cos kL - \cos k_c L + \sin kL \cot \varphi_j}{\sin k_c L} . \tag{11'}$$

In (11′), k_c is the natural angular wave number of oscillations of the ray about the axis, corresponding to $\sqrt{C}$ in the case of a continuous lens. In this case it is the first positive real solution to the following equation, if such a solution exists.

$$\cos k_c L = 1 - \frac{cL}{2} \tag{19}$$

The denominator in (8′) may be written $\cos k_j L - \cos k_c L$. (When $cL > 4$ the lenses are separated by more than four times their focal length and they will not guide the beam.)

The analog of (12) would be like (12) except that terms from (8′), (11′) and (19) replace the terms from (8) and (11). It is not worthwhile to write it out here.

Equation 19 has an infinite number of real roots if $cL < 4$. We have chosen to describe the ray trajectory using only the first root, but we cannot ignore the rest entirely. If the pipe containing the lenses has an appreciable amplitude of waviness near one or more of these higher wave-number roots, the ray oscillations will grow because the lenses will be displaced as if there were additional amplitude in the fundamental, long wave-length component. These higher roots correspond to waves in the pipe with length less than $2L$ or twice the lens separation. In choosing the lowest root for k_c we are following Brillouin's example[5] and restricting values of k to the "first Brillouin zone" of the periodic

array of lenses. Remember that $k_j = 2\pi j/\Lambda$ where Λ is the length of the periodicity in the shape of the guide, which will ultimately be made very large. We can rewrite (6) in the following form in order to limit the range of k_j from zero to π/L.

$$y(nL) = \sum_{j=1}^{\Lambda/2L} \sum_{h=0}^{\infty} \left[Y\left(\frac{2\pi h}{L} + k_j\right) \sin\left(k_j nL + \varphi_{h,j,1}\right) + Y\left(\frac{2\pi(h+1)}{L} - k_j\right) \sin\left(k_j nL + \varphi_{h,j,2}\right) \right] \tag{6'}$$

(nL is the x coordinate of lens number n. See Fig. 2.) Then the analog of (13) may be written as

$$\bar{\bar{y}}^2 = \frac{1}{2} \sum_{j=1}^{\Lambda/2L} \sum_{h=0}^{\infty} \left[Y^2\left(\frac{2\pi h}{L} + k_j\right) + Y^2\left(\frac{2\pi(h+1)}{L} - k_j\right) \right] \equiv \frac{1}{2} \sum_{j=1}^{\Lambda/2L} S^2(k_j). \tag{13'}$$

When written in terms of S rather than Y, the analog of (14) looks like this:

$$\bar{\bar{\delta}}^2(x) = \frac{1}{2} \sum_{j=1}^{\Lambda/2L} S^2(k_j) \left(\frac{1 - \cos k_j L}{\cos k_j L - \cos k_c L}\right)^2 \cdot \left[\left(\sin k_j x - \frac{\sin k_j L \sin k_c x}{\sin k_c L} \right)^2 + \left(\cos k_j x - \cos k_c x + \frac{\sin k_c x}{\sin k_c L} (\cos k_c L - \cos k_j L) \right)^2 \right]. \tag{14'}$$

It is easy to show yourself that if L approaches zero then $k_c \to \sqrt{C}$ and (14′) approaches (14).

Next, we make the substitution

$$\epsilon_j = \frac{k_j}{k_c} - 1 \tag{15'}$$

in (14′) and neglect terms of order $(\epsilon_j k_c L)^2$ or smaller in the result. The neglect of these terms is justified because the result we are seeking, as before, turns out to depend only on extremely small values of ϵ when x is large. The result of this and a large amount of manipulation is the analog of (16) which follows:

$$\bar{\delta}^2(x) \approx \frac{1}{2}\sum_{j=1}^{\Lambda/2L} \frac{S^2(\epsilon_j)}{k_c^2L^2}\left(\frac{1-\cos k_cL}{1+\cos k_cL}\right)$$
$$\cdot\left[\frac{4\sin^2(\epsilon_jk_cx/2)}{\epsilon_j^2} + \frac{k_cL}{\epsilon_j}\left(\cot k_cL\left(1+\cos\frac{k_cL}{2}\right.\right.\right.$$
$$\left.\left.\left. - 2\sin k_cx\right) - \tan k_cL\left(1+\cos\frac{k_cL}{2} - 2\cos k_cx\right)\right)\right] \tag{16'}$$
$$\equiv \frac{1}{2k_c^2L^2}\left(\frac{1-\cos k_cL}{1+\cos k_cL}\right)\sum_{j=1}^{\Lambda/2L} S^2(\epsilon_j)\left(\frac{u(\epsilon_j)}{\epsilon_j^2} + \frac{v}{\epsilon_j}\right).$$

As before, we let Λ go to infinity and change the sum to an integral.

$$\bar{\delta}^2(x) \approx \frac{1}{2k_c^2L^2}\left(\frac{1-\cos k_cL}{1+\cos k_cL}\right)\int_{-1}^{(\pi/k_cL)-1} S^2(\epsilon)\left(\frac{u(\epsilon)}{\epsilon^2} + \frac{v}{\epsilon}\right)d\epsilon$$

Although the integral of v/ϵ has a logarithmic divergence at $\epsilon = 0$, the divergent part is an odd function of ϵ and the integrand over a small finite range around zero is small. The dominant term in the integration, for large values of x, comes from the $u(\epsilon)/\epsilon^2$ part. Thus we find that for large x

$$\bar{\bar{\delta}}^2(x) \approx \frac{1}{2k_c^2L^2}\left(\frac{1-\cos k_cL}{1+\cos k_cL}\right)\int_{-1}^{(\pi/k_cL)-1} S^2(\epsilon)\cdot\frac{4\sin^2(\epsilon k_cx/2)}{\epsilon^2}\,d\epsilon$$
$$\approx \frac{1}{2k_c^2L^2}\left(\frac{1-\cos k_cL}{1+\cos k_cL}\right)\cdot 2\pi k_cxS^2(\epsilon = 0) \tag{17'}$$

where, from (13'),

$$S^2(\epsilon = 0) = \sum_{h=0}^{\infty} Y^2\left(\frac{2\pi}{L}h + k_c\right) + Y^2\left(\frac{2\pi}{L}(h+1) - k_c\right).$$

Again, when $L \to 0$ we obtain the same result as from (17).

IV. THE SHAPE OF THE AXIS OF THE GUIDE

We might arbitrarily guess a spectral distribution for the Fourier components, $Y(k)$, of the shape of the somewhat irregular pipe. It may be somewhat more meaningful, however, to guess a spectrum of forces on the pipe and to obtain the shape from that. The latter procedure enables us to see the effect of stiffness of the pipe, and of the spectrum of applied forces, on its shape. For simplicity we will assume that the pipe would be perfectly straight, or at least relatively very straight, in the absence of the forces that are to be applied to it.

Suppose the pipe has a modulus of rigidity $\sigma = EI$ where E is Young's

modulus and I is the geometric second moment of a cross section of the pipe about a transverse axis. The curvature of the pipe is related to the local moment of torque, $\mathbf{M}(x)$, through the equation[6]

$$\frac{d^2y}{dx^2} = \frac{\mathbf{M}(x)}{\sigma}. \tag{20}$$

The second derivative of the torque moment is the transverse force per unit length, p, applied to the pipe, so that

$$\frac{d^4y(x)}{dx^4} = \frac{1}{\sigma}p(x). \tag{21}$$

If the net force per unit length on the pipe is represented by the expression

$$p(x) = \sum_{j=1}^{\infty} P(k_j) \sin (k_j x - \varphi_j), \tag{22}$$

we obtain the following expression for the Fourier transform of the shape of the pipe by comparing (6), (21) and (22):

$$Y(k) = \frac{P(k)}{k^4\sigma}. \tag{23}$$

A convenient analytic expression that can be used to approximate a reasonable spectrum of forces is the following, where P_0, k_0, m and n are adjustable parameters.

$$P(k) = P_0((k/k_0)^n/(1 + (k/k_0)^m))^{\frac{1}{2}}. \tag{24}$$

In this case (23) becomes

$$Y(k) = (P_0/k_0^4\sigma)((k/k_0)^{n-8}/(1 + (k/k_0)^m))^{\frac{1}{2}}. \tag{23'}$$

Curves of Y and P as functions of k/k_0 are shown in Fig. 3 for the specific values $m = 11$ and $n = 8$. The factors outside the square roots are omitted. (If necessary we could use a sum of several such expressions with different parameters in each without greatly complicating the results, but we shall restrict the analysis to one.)

We will make use of the following formula several times in the following analysis.[7]

$$\int_0^{\infty} \frac{K^r}{1 + K^s}\,dK = \frac{\pi}{s \sin\left(\dfrac{(r+1)\pi}{s}\right)}. \tag{25}$$

Let K represent the ratio k/k_0.

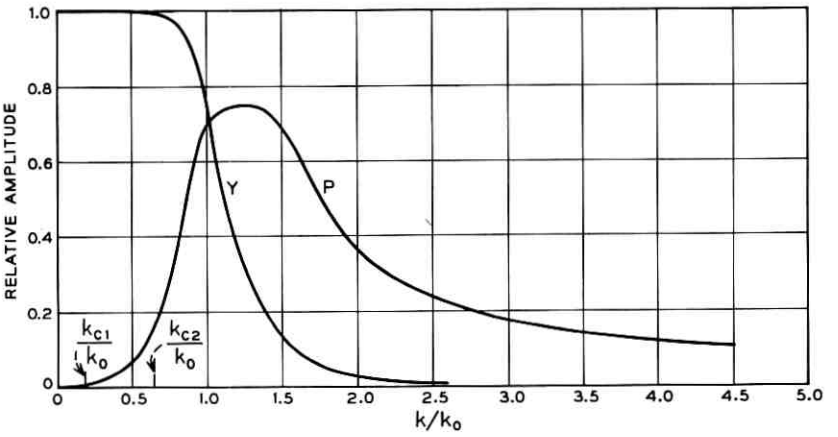


Fig. 3 — Relative amplitudes of sinusoidal components of forces applied to the guide (P), and of the shape of the pipe (Y), as a function of relative wave number, k/k_0 . k_{c1}/k_0 is the ratio of k_c to k_0 used in the examples on the left side of Table I, while k_{c2}/k_0 is the ratio used in the examples on the right.

The mean square value of $P(K)$ is

$$\bar{\bar{p}}^2 = \frac{1}{2}\int_0^\infty P^2(K)dK = \frac{P_0^2}{2}\int_0^\infty \frac{K^n}{1+K^m}\,dK$$
$$= \frac{\pi P_0^2}{2m \sin\left((n+1)\pi/m\right)}. \tag{26}$$

The mean square displacement on the guide axis from the straight x axis is (cf. (23))

$$\bar{\bar{y}}^2 = \frac{1}{2}\int_0^\infty Y^2(K)dK = \frac{1}{2}\left(\frac{P_0}{k_0^4\sigma}\right)^2\int_0^\infty \frac{K^{n-8}}{1+K^m}\,dK$$
$$= \left(\frac{P_0}{k_0^4\sigma}\right)^2 \frac{\pi}{2m \sin\left((n-7)\pi/m\right)}. \tag{27}$$

Hence $\bar{\bar{y}}$ and $\bar{\bar{p}}$ are related through the expression

$$\bar{\bar{y}} = \frac{\bar{\bar{p}}}{k_0^4\sigma}\sqrt{\sin\frac{(n+1)\pi}{m} \Big/ \sin\frac{(n-7)\pi}{m}}. \tag{28}$$

We will also define characteristic wavelengths, λ_p and λ_y for $p(x)$ and $y(x)$ since they are more tangible than the parameter k_0. The characteristic wavelengths will be taken as 2π divided by the angular wave number corresponding to the "center of gravity" of the integrands of

(26) and (27). Thus we obtain

$$\lambda_p = \frac{2\pi}{k_0} \int_0^\infty \frac{K^n}{1 + K^m}\, dK \Big/ \int_0^\infty \frac{K^n}{1 + K^m} \cdot K\, dK$$
$$= \frac{2\pi}{k_0} \frac{\sin\left((n + 2)\pi/m\right)}{\sin\left((n + 1)\pi/m\right)} \tag{29}$$

and similarly

$$\lambda_y = \frac{2\pi}{k_0} \frac{\sin \dfrac{(n - 6)\pi}{m}}{\sin \dfrac{(n - 7)\pi}{m}}. \tag{30}$$

For very narrow force spectra λ_y and λ_p are both equal to $2\pi/k_0$, but when the spectra are broad, λ_y tends to be larger than λ_p, since short wavelength force components are unable to bend the stiff pipe much. See Fig. 1.

V. NUMERICAL ILLUSTRATION

The following numerical illustration serves mainly to show how the preceding results could be used. The parameters chosen are the ones that one is likely to know in a real situation. The actual numbers might vary considerably, but it is not hard to estimate how much changes in various parameters would affect the results. What is most important is to know the order of magnitude of the effects using reasonable parameters. The following example forms the left half of Table I.

5.1 *Shape of the Guide Axis*

Let us suppose we have a guide composed of a series of lenses or a lens-like medium perfectly centered in a cylindrical pipe. We will suppose that curvature of the pipe axis is due largely to strain from externally applied forces. For purposes of illustration, we will suppose that the main force on the pipe is due to its own weight. We shall suppose that it is lying on an irregular bed which supports the weight of the pipe at more or less random intervals averaging about λ_p meters apart. See Fig. 1. (The spread in the intervals is determined by m and n in (24).)

Suppose the pipe is made of steel whose density is $\rho = 7.8 \times 10^3$ kg/meter^3 and whose Young's modulus of rigidity is $E = 2.0 \times 10^9$ newtons/meter^2.

Let the pipe be 10 cm OD and 9 cm ID. Let the distance λ_p be one meter.

Let us use the smallest integer values of m and n that are consistent with the conditions on (25) for all the applications of that equation (see (26) to (29)). We find that these values are $m = 11$ and $n = 8$. (This gives a rather broad spectrum of wavelengths of applied force. See Fig. 3. Larger values would give narrower spectra, which would be appropriate if the pipe were supported at regular intervals.)

Using these numbers in (29) we find that $k_0 = 3.27$ meters^{-1}.

Equation (30) then gives $\lambda_y = 3.68$ meters.

The stiffness of a cylindrical pipe is related to its OD and ID through the equation

$$\sigma = \frac{\pi}{64} E((\text{OD})^4 - (\text{ID})^4) \tag{31}$$

which gives $\sigma = 3.38 \times 10^3$ newton meters^2 for our example.

We shall assume for simplicity that the mean square amplitude of force variations, $\bar{\bar{p}}$, is equal to the linear density of the pipe times the gravity constant, which is correct to within a small numerical factor when the weight of the pipe determines all the forces. We thus obtain

$$\bar{\bar{p}} = \rho g((\text{OD})^2 - (\text{ID})^2)\, \pi/4. \tag{32}$$

For our pipe we get $\bar{\bar{p}} = 114$ newtons/meter.

Equation (28) then gives us $\bar{\bar{y}} = 0.407 \times 10^{-3}$ meters.

We shall need to know P_0 when we compute the beam oscillation growth rate. Equation (26) gives us $P_0 = 222$ newtons/meter.

5.2 *Beam Oscillation Growth Rate*

Let us first suppose we have a continuous lens-like guide with specific convergence C and the shape defined in part A of this section. We use (23′) to evaluate Y at $k = \sqrt{C} = 0.5$ meters^{-1}, which is the value at $\epsilon = 0$. If we let $C = 0.25$ diopters per meter, we get $Y(\sqrt{C}) = 0.572 \times 10^{-3}$ meters. The ratio $\sqrt{C}/k_0$ is 0.153 in this example. It is labeled k_{c1}/k_0 on Fig. 3.

Inserting the value of Y into (17) gives $\bar{\bar{\delta}}(x) = \sqrt{x} \cdot (0.358 \times 10^{-3})$ meters. In 100 meters, the root mean square amplitude of oscillation is 3.58 mm and in 10 kilometers it is 3.58 cm.

Next let us consider a series of thin lenses that are separated by twice their individual focal lengths, $(L = 2/c)$, and that give the same angular wave number for ray oscillations as in the preceding example,

$k_c = 0.5$ meters^{-1}. From (19) we find that $\cos k_c L = 0$, which gives $L = \pi$ or 3.14 meters and $c = 2/\pi$ or 0.637 diopters.

Now we evaluate the sum for $S^2(\epsilon = 0)$ in (17′), using (23′), as in the first example. The first term of the sum is equal to Y^2 at 0.5 meters^{-1}, whose square root we already evaluated in the first example. The next term is Y^2 evaluated at 1.5 meters^{-1}, the next at 2.5 meters^{-1}, etc. The terms rapidly diminish in size. Equation (17′) gives $\bar{\bar{\delta}} = \sqrt{x} \cdot (0.830 \times 10^{-3})$ meters, which is not much larger than the result for the continuous lens-like medium with the same ray oscillation wave number.

The preceding examples illustrate the fact that the results are essentially the same for a series of lenses or a continuous lens-like medium of equal k_c when the characteristic wavelength, λ_y, of irregularities in the pipe axis is longer than the separation of the lenses and when the lens separation is not too near to the limiting value of four focal lengths.

The input parameters and results of the preceding examples are listed on the left side of Table I.

The right side of Table I shows what happens when the characteristic distance between bumps on the ground, λ_p, is increased to 4 meters, keeping other parameters the same. There is a catastrophic increase in rate of growth of beam oscillations because the root mean square amplitude of waves in the pipe is greatly increased while the value of Y at k_c remains near its maximum. See Fig. 3. The ratio $\sqrt{C}/k_0$ or k_c/k_0 in these examples is 0.611 and is labeled k_{c2}/k_0 on Fig. 3.

A short FORTRAN computer program is available upon request to anyone who may wish to enlarge on Table I using other values of the input parameters.

VI. CONCLUSIONS

The results are strongly dependent on some of the input parameters, but we hope the examples represent the general magnitude of the factors one might have to work with if such a light guide were built. By using a suitably spaced periodic set of supports for the guide, we could probably make Y^2 or S^2 very small at $\sqrt{C}$ or k_c. The peak or peaks in the function $Y^2(k)$ or $S^2(k)$ could lie elsewhere. Thus, the contribution to wave growth due to forces on the pipe might be considerably reduced.

The computations did not consider waves in the pipe axis due to tolerance limits that would arise in manufacture of the guide or in linking sections, but these waves could obviously be incorporated as an additional term in $Y^2(k)$ if they could be estimated or measured.

It seems probable that the beam would have to be recentered at rather frequent intervals along the guide unless the aperture were very large or the pipe were extremely stiff and straight.

REFERENCES

1. Hirano, J., and Fukatsu, Y., Stability of a Light-Beam in A Beam-Waveguide, Proc. IEEE, *52*, Nov. 1964, p. 1284. The authors have started the attack on the problem by discussing beam instability in the case of a series of lenses with incoherent, random displacements and with perfectly sinusoidal displacements. H. E. Rowe of Bell Telephone Laboratories has also investigated the effects of random displacements (unpublished memorandum, June 19, 1963.)
2. Marcuse, D., Statistical Treatment of Light-Ray Propagation in Beam-Waveguides, B.S.T.J., This Issue, p. 2065–2081 and Unger, H. G., Light Beam Propagation in Curved Schlieren guides, Arch. Electr. Ubertr., *19*, April 1965, pp. 189–198. Marcuse has been working concurrently on the problem for a series of lenses from the point of view of correlated displacements. Unger studied the problem for a continuous medium from that viewpoint.
3. Berreman, D. W., Convective Gas Light Guides or Lens Trains for Optical Beam Transmission, J. Opt. Soc. Am., *55*, Mar. 1965, p. 239.
4. Born, M., and Wolf, E., *Principles of Optics*, Second Edition, Pergamon Press, Oxford, England, 1964, Equation (2), p. 122.
5. Brillouin, L., *Wave Propagation in Periodic Structures*, Dover Publications, New York, 1953.
6. cf. Joos, G., *Theoretical Physics*, Second Edition, Hafner Publishing Company, New York, 1950, Equation (60), p. 176.
7. Dwight, H. B., *Tables of Integrals and Other Mathematical Data*, 4th Edition, Macmillan Company, New York, 1961, Equation (856.08.)

Some Stability Results Related to Those of V. M. Popov*

By I. W. SANDBERG

(Manuscript received July 15, 1965)

In a recent paper by this writer, some new techniques are described for obtaining sufficient conditions for the $\mathfrak{L}_2$-boundedness and $\mathfrak{L}_\infty$-boundedness of solutions of nonlinear functional equations. In this paper, these techniques are developed further and are used to prove some stability results for large classes of feedback systems and electrical networks that contain subsystems which are not necessarily representable in terms of ordinary differential equations.

I. NOTATION

Let $\mathfrak{IC}(0, \infty)$ denote the set of real-valued measurable functions of the real variable t defined on $[0, \infty)$. Let

$$\mathfrak{L}_p(0, \infty) = \left\{ f \mid f \in \mathfrak{IC}(0, \infty), \int_0^\infty |f(t)|^p \, dt < \infty \right\}$$

for $p = 1$ and $p = 2$. Let

$$\mathfrak{L}_\infty(0, \infty) = \{ f \mid f \in \mathfrak{IC}(0, \infty), \sup |f(t)| < \infty \}.$$

Let $y \in (0, \infty)$, and define f_y by

$$f_y(t) = f(t) \qquad \text{for } t \in [0, y]$$
$$= 0 \qquad \text{for } t > y$$

for all $f \in \mathfrak{IC}(0, \infty)$, and let

$$\mathcal{E}(0, \infty) = \{ f \mid f \in \mathfrak{IC}(0, \infty), f_y \in \mathfrak{L}_2(0, \infty) \text{ for all } y \in (0, \infty) \}$$

[i.e., $\mathcal{E}(0, \infty)$ denotes the set of real-valued *locally* square-integrable functions defined on $[0, \infty)$].

* This paper was presented at the Symposium on Network Theory, Cranfield-Bedford, England, September, 1965.

Finally, the integral

$$\int_0^y f(t)g(t)\,dt$$

is denoted by $\langle f_y, g\rangle$ (or by $\langle g, f_y\rangle$) for all $f \in \mathcal{E}(0,\infty)$, all $g \in \mathcal{E}(0,\infty)$, and all $y \in (0,\infty)$; and $\| h \|$ denotes

$$\left(\int_0^\infty |h(t)|^2\,dt\right)^{\frac{1}{2}}$$

for all $h \in \mathcal{L}_2(0,\infty)$.

II. INTRODUCTION

To a considerable extent, Ref. 1 is a summary of certain results of a recent study by this writer of the input-output properties of a large class of time-varying nonlinear systems. The properties of a vector nonlinear Volterra integral equation of the second kind that frequently arises in the study of physical systems are considered in detail,* and some conditions are presented for the norm-boundedness of solutions of a functional equation of similar type defined on an abstract space. Much of the material presented in Ref. 1 is drawn from Refs. 2 and 3.

In Ref. 1, some techniques other than those of Refs. 2 and 3 are described for obtaining sufficient conditions for the $\mathcal{L}_2$-boundedness and $\mathcal{L}_\infty$-boundedness of solutions of functional equations. In this paper, these techniques are developed further and are used to prove some stability results, related to those of V. M. Popov,[8] for large classes of feedback systems and electrical networks that contain subsystems which are not necessarily representable in terms of ordinary differential equations.†

III. THE FEEDBACK SYSTEM AND THE MAIN RESULTS

Consider the system of Fig. 1. We shall restrict our discussion throughout to cases in which g, f, u, r, v, and w denote functions belonging to $\mathcal{E}(0,\infty)$. The block labeled ψ represents a memoryless time-invariant nonlinear element that introduces the constraint $w(t) = \psi[v(t)]$ for $t \geqq 0$.

* The results for the Volterra equation are of direct engineering interest because of the central role played by a certain "critical-disk" frequency-domain condition. "Critical-disk" frequency-domain conditions were encountered in connection with related analytical questions in Refs. 4, 5, and 6. Some material related to the results of Refs. 1, 4, 5, and 6 has been written up by G. Zames.[7]

† Some interesting "Popov-like" stability theorems for systems governed by ordinary differential equations are proved in Refs. 9, 10, and 11.

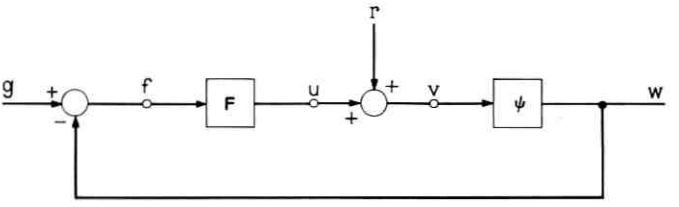


Fig. 1 — Nonlinear feedback system.

Assumption 1: $\psi(x)$ is defined and continuous on $(-\infty, \infty)$, $\psi(0) = 0$, and there exist real constants $\alpha > 0$ and $\beta < \infty$ such that

$$\alpha \leqq x^{-1}\psi(x) \leqq \beta$$

for all $x \neq 0$.

The block labeled **F** represents a (not necessarily linear or time-invariant) subsystem that introduces the constraint $(\mathbf{F}f)(t) = u(t)$ for $t \geqq 0$.

Assumption 2: The operator **F** can be written as $\mathbf{F}_2\mathbf{F}_1$ with

(1.) $\mathbf{F}_1$ a (not necessarily linear or time-invariant) mapping of $\mathcal{E}(0, \infty)$ into itself, and

(2.) $\mathbf{F}_2$ the linear mapping of $\mathcal{E}(0, \infty)$ into itself defined by the condition:

$$(\mathbf{F}_2 q)(t) = \int_0^t \exp\left[-\int_\tau^t \delta(x)dx\right] q(\tau)\, d\tau, \qquad t \geqq 0$$

for all $q \in \mathcal{E}(0, \infty)$, in which δ is a real measurable function defined on $[0, \infty)$ such that there exist real constants $c_1 > 0$ and $c_2 < \infty$ with the properties that $c_1 \leqq \delta(x) \leqq c_2$ for all $x \in [0, \infty)$.

We note that $\mathbf{F}_2 q$ denotes the convolution of q with the impulse-response function of a positive-element parallel resistor-capacitor combination with time-varying resistance.

In Fig. 1, g denotes an input and r takes into account the effect of initial conditions at $t = 0$. The relation between f, g, and r is

$$g = f + \psi[\mathbf{F}_2\mathbf{F}_1 f + r]. \tag{1}$$

Equation (1) also governs the behavior of a large class of active time-varying nonlinear networks. A network analog of the feedback system of Fig. 1 is shown in Fig. 2, where ψ denotes a nonlinear conductance.

Assumption 3: $r \in \mathcal{E}(0, \infty)$, $\dot{r}$ exists on $[0, \infty)$ and $\dot{r} \in \mathcal{E}(0, \infty)$.

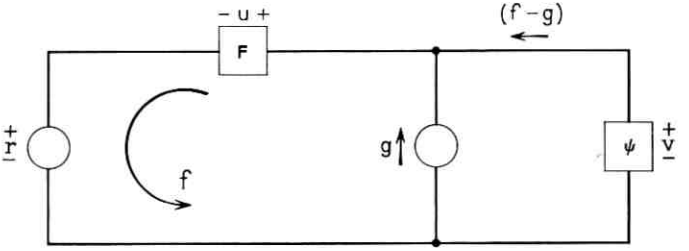


Fig. 2 — Equivalent network.

3.1 *The Main Results*

The principal contributions of this paper are believed to be the techniques used to prove the following results.

Theorem 1: Let Assumptions 1, 2, and 3 be satisfied. Let $\mathbf{F}_1$ *be such that there exist a real constant* k, *a nonnegative constant* σ, *and a positive constant* c *with the properties that*

$$(i) \qquad \sigma < 2\alpha^2\beta^{-1}(\beta^{-1}c_1 + k)$$

$$(ii) \qquad \langle e^{\sigma t}(\mathbf{F}_1 q)_y , q\rangle \geqq k \parallel e^{\frac{1}{2}\sigma t} q_y \parallel^2$$

$$\parallel e^{\frac{1}{2}\sigma t}(\mathbf{F}_1 q)_y \parallel \leqq c \parallel e^{\frac{1}{2}\sigma t} q_y \parallel$$

for all $q \in \mathcal{E}(0,\infty)$ *and all* $y \in (0,\infty)$.

Let $f \in \mathcal{E}(0,\infty)$, *and let*

$$g = f + \psi[\mathbf{F}_2\mathbf{F}_1 f + r].$$

Then there exists a positive constant λ, *depending only on* k, σ, c_1, α, β, *and* c, *such that*

$$\lambda \parallel e^{\frac{1}{2}\sigma t} f_y \parallel \leqq \parallel e^{\frac{1}{2}\sigma t} g_y \parallel + \parallel e^{\frac{1}{2}\sigma t}(\dot{r} + \delta r)_y \parallel + \left[\int_0^{r(0)} \psi(\eta)\, d\eta\right]^{\frac{1}{2}}$$

for all $y \in (0,\infty)$.

Corollary 1: If the hypotheses of Theorem 1 are satisfied with $\sigma = 0$, i $g \in \mathcal{L}_2(0,\infty)$, *and if* $(\dot{r} + \delta r) \in \mathcal{L}_2(0,\infty)$, *then* $f \in \mathcal{L}_2(0,\infty)$, *and there exists a constant* $\lambda > 0$, *depending only on* k, c_1, α, β, *and* c *such that*

$$\lambda \parallel f \parallel \leqq \parallel g \parallel + \parallel \dot{r} + \delta r \parallel + \left[\int_0^{r(0)} \psi(\eta) d\eta\right]^{\frac{1}{2}}.$$

Remarks: If $\mathbf{F}_1$ denotes the mapping of $\mathcal{E}(0,\infty)$ into itself defined by

$$(\mathbf{F}_1 u)(t) = v_0 u(t) + \int_0^t v(t-\tau)u(\tau)\,d\tau, \qquad t \geqq 0$$

for all $u \in \mathcal{E}(0,\infty)$, where v_0 is a real constant and $v \in \mathcal{L}_1(0,\infty)$, then

$$\| (\mathbf{F}_1 q)_y \| \leqq \left[|v_0| + \int_0^\infty |v(t)|\,dt \right] \| q_y \|$$

for all $y \in (0,\infty)$ and all $q \in \mathcal{E}(0,\infty)$, and $\langle (\mathbf{F}_1 q)_y , q \rangle \geqq k \| q_y \|^2$ for all $y \in (0,\infty)$ and all $q \in \mathcal{E}(0,\infty)$ provided that

$$v_0 + \operatorname{Re} \int_0^\infty e^{-i\omega t} v(t)\,dt \geqq k$$

for all $\omega \in (-\infty,\infty)$.

Corollary 2: If the hypotheses of Corollary 1 are satisfied, if $g(t) \to 0$ as $t \to \infty$, and if $r(t) \to 0$ as $t \to \infty$, then $f(t) \to 0$ as $t \to \infty$.

Corollary 3: If the hypotheses of Theorem 1 are satisfied with $\sigma > 0$, if $g \in \mathcal{L}_\infty(0,\infty)$, if r and $\dot{r}$ belong to $\mathcal{L}_\infty(0,\infty)$, and if there exists a constant γ such that, with $\mathbf{F} = \mathbf{F}_2\mathbf{F}_1$,

$$|(\mathbf{F}q)(y)| \leqq \gamma e^{-\frac{1}{2}\sigma y} \| e^{\frac{1}{2}\sigma t} q_y \|$$

for all $q \in \mathcal{E}(0,\infty)$ and all $y \in (0,\infty)$, then $f \in \mathcal{L}_\infty(0,\infty)$, and there exists a positive constant λ_1, depending only on k, σ, γ, c_1, α, β, and c such that

$$\lambda_1 \sup_{t \geqq 0} |(\mathbf{F}_2\mathbf{F}_1 f)(t)| \leqq \sup_{t \geqq 0} |g(t)| + \sup_{t \geqq 0} |(\dot{r} + \delta r)(t)| + \left[\int_0^{r(0)} \psi(\eta)\,d\eta \right]^{\frac{1}{2}}.$$

Remarks: Suppose that $\mathbf{F}_1$ is the mapping of $\mathcal{E}(0,\infty)$ into itself defined by

$$(\mathbf{F}_1 u)(t) = v_0 u(t) + \int_0^t v(t-\tau)u(\tau)\,d\tau, \qquad t \geqq 0$$

for all $u \in \mathcal{E}(0,\infty)$, where v_0 is a real constant and $e^{\frac{1}{2}\sigma t} v \in \mathcal{L}_1(0,\infty)$ for some positive constant σ. Then

$$\langle e^{\sigma t}(\mathbf{F}_1 q)_y, q\rangle = \int_0^\infty e^{\frac{1}{2}\sigma t} q_y(t) \left[v_0 e^{\frac{1}{2}\sigma t} q_y(t) + \int_0^t v(t-\tau) e^{\frac{1}{2}\sigma(t-\tau)} e^{\frac{1}{2}\sigma\tau} q_y(\tau) d\tau \right] dt$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[v_0 + \int_0^\infty v(t) e^{-(i\omega - \frac{1}{2}\sigma)t} dt \right] \left| \int_0^\infty e^{\frac{1}{2}\sigma t} q_y(t) e^{-i\omega t} dt \right|^2 d\omega$$

and

$$\| e^{\frac{1}{2}\sigma t}(\mathbf{F}_1 q)_y \| \leqq \| e^{\frac{1}{2}\sigma t} \mathbf{F}_1 q_y \| = \left\| e^{\frac{1}{2}\sigma t} v_0 q_y + \int_0^t v(t-\tau) e^{\frac{1}{2}\sigma(t-\tau)} e^{\frac{1}{2}\sigma\tau} q_y(\tau)\, d\tau \right\|$$

$$\leqq \left[| v_0 | + \int_0^\infty | v(t) e^{\frac{1}{2}\sigma t} |\, dt \right] \| e^{\frac{1}{2}\sigma t} q_y \|$$

for all $y \in (0, \infty)$ and all $q \in \mathcal{E}(0, \infty)$. Thus Assumption (ii) of Theorem 1 is satisfied if

$$v_0 + \operatorname{Re} \int_0^\infty v(t) e^{-(i\omega - \frac{1}{2}\sigma)t} dt \geqq k$$

for all $\omega \in (-\infty, \infty)$.

Concerning the key hypothesis of Corollary 3, if $\mathbf{F} = \mathbf{F}_2\mathbf{F}_1$ is the mapping of $\mathcal{E}(0, \infty)$ into itself defined by

$$(\mathbf{F}u)(t) = \int_0^t w(t-\tau) u(\tau) d\tau, \qquad t \geqq 0$$

for all $u \in \mathcal{E}(0, \infty)$, where $e^{\frac{1}{2}\sigma t} w \in \mathcal{L}_2(0, \infty)$, then

$$| (\mathbf{F}q)(y) | = e^{-\frac{1}{2}\sigma y} \left| \int_0^y w(y-\tau) e^{\frac{1}{2}\sigma(y-\tau)} e^{\frac{1}{2}\sigma\tau} q_y(\tau)\, d\tau \right|$$

$$\leqq e^{-\frac{1}{2}\sigma y} \left(\int_0^\infty | w(t) e^{\frac{1}{2}\sigma t} |^2\, dt \right)^{\frac{1}{2}} \| e^{\frac{1}{2}\sigma t} q_y \|$$

for all $q \in \mathcal{E}(0, \infty)$ and all $y \in (0, \infty)$.

3.2 *Related Results*

The results stated above can be extended in many different directions by exploiting the techniques of Section 4.1. For example, similar results can be obtained (see Section 4.5) for the case in which the nonlinear element ψ of Fig. 1 is replaced by a linear time-varying element that introduces the constraint $w(t) = m(t)v(t)$ for $t \geqq 0$, in which $m(\cdot)$ is a

positive bounded measurable function. In that case

$$\tilde{g} = f + m\mathbf{F}_2\mathbf{F}_1 f,$$

in which $\tilde{g} = g - mr$.

A specific application of the material of Section 3.1 is considered in the appendix. In particular, the result proved there implies that a rather general type of (not necessarily lumped) time-invariant physical system containing a single nonlinear element is "bounded-input bounded-output stable" if the so-called Popov inequality[8] is satisfied.

Some material related to the content of Theorem 1 can be found in Ref. 7. Our results differ in many respects from those stated in Ref. 7. In particular, there the effect of the initial condition function r is not taken into consideration.

The idea of using an inequality of the form stated in Corollary 3 in order to establish the boundedness of solutions of nonlinear functional equations evolved from the techniques of Ref. 3 and was presented in Ref. 1. This idea has also been considered by G. Zames in very recent independent unpublished work.

IV. PROOFS

4.1 *Proof of Theorem 1*

Lemma 1: Suppose that Assumptions 1, 2, and 3 are satisfied. Let σ be a nonnegative constant. Then

$$\langle e^{\sigma t}(\psi[\mathbf{F}_2 q_y + r])_y, q_y\rangle \geqq \left(\frac{c_1}{\beta} - \frac{\sigma\beta}{2\alpha^2}\right) \| e^{\frac{1}{2}\sigma t}(\psi[\mathbf{F}_2 q_y + r])_y \|^2$$

$$- \| e^{\frac{1}{2}\sigma t}(\dot{r} + \delta r)_y \| \cdot \| e^{\frac{1}{2}\sigma t}(\psi[\mathbf{F}_2 q_y + r])_y \| - \int_0^{r(0)} \psi[\eta]\, d\eta$$

for all $y \in (0,\infty)$ and all $q \in \mathcal{E}(0,\infty)$.

Proof of Lemma 1: Let $y \in (0,\infty)$, let $q \in \mathcal{E}(0,\infty)$, and let $z = \mathbf{F}_2 q_y$. Then $\dot{z}(t) + \delta(t)z(t) = q_y(t)$ for almost all $t \in (0,\infty)$, and, with $\sigma \in [0,\infty)$,

$$\begin{aligned}\langle e^{\sigma t}(\psi[\mathbf{F}_2 q_y + r])_y, q_y\rangle &= \langle e^{\sigma t}(\psi[z + r])_y, \delta(z + r)\rangle \\ &\quad + \langle e^{\sigma t}(\psi[z + r])_y, \dot{z} + \dot{r}\rangle \\ &\quad - \langle e^{\sigma t}(\psi[z + r])_y, \dot{r} + \delta r\rangle.\end{aligned}$$

Thus, since

$$\langle e^{\sigma t}(\psi[z + r])_y , \delta(z + r)\rangle \geqq (c_1/\beta) \| e^{\frac{1}{2}\sigma t}(\psi[z + r])_y \|^2$$

(we have used the fact that $x/\psi(x) \geqq \beta^{-1}$ for all real $x \neq 0$), and (by the Schwarz inequality)

$$| \langle e^{\sigma t}(\psi[z + r])_y , \dot{r} + \delta r\rangle | \leqq \| e^{\frac{1}{2}\sigma t}(\dot{r} + \delta r)_y \| \cdot \| e^{\frac{1}{2}\sigma t}(\psi[z + r])_y \|,$$

we have

$$\begin{aligned}\langle e^{\sigma t}(\psi[\mathbf{F}_2 q_y + r])_y , q_y\rangle \geqq\ & (c_1/\beta) \| e^{\frac{1}{2}\sigma t}(\psi[z + r])_y \|^2 \\ & - \| e^{\frac{1}{2}\sigma t}(\dot{r} + \delta r)_y \| \cdot \| e^{\frac{1}{2}\sigma t}(\psi[z + r])_y \| \\ & + \langle e^{\sigma t}(\psi[z + r])_y , \dot{z} + \dot{r}\rangle.\end{aligned}$$

With $\eta = z + r$, we find that

$$\begin{aligned}\langle e^{\sigma t}(\psi[z + r])_y, \dot{z} + \dot{r}\rangle &= \langle e^{\sigma t}(\psi[\eta])_y, \dot{\eta}\rangle = \int_0^y e^{\sigma t}\psi[\eta]\dot{\eta}\, dt \\ &= e^{\sigma t}\int_{\eta(0)}^{\eta(t)} \psi[\eta]d\eta \Big|_0^y - \sigma \int_0^y \int_{\eta(0)}^{\eta(t)} \psi[\eta]d\eta e^{\sigma t}\, dt \\ &= e^{\sigma y}\int_0^{\eta(y)} \psi[\eta]d\eta - \int_0^{\eta(0)} \psi[\eta]d\eta \\ &\quad - \sigma \int_0^y \int_0^{\eta(t)} \psi[\eta]d\eta e^{\sigma t}\, dt.\end{aligned}$$

Thus,

$$\langle e^{\sigma t}(\psi[\eta])_y, \dot{\eta}\rangle \geqq -\int_0^{\eta(0)} \psi[\eta]d\eta - \sigma \int_0^y \int_0^{\eta(t)} \psi[\eta]d\eta e^{\sigma t}\, dt$$

Since

$$0 \leqq \int_0^y \int_0^{\eta(t)} \psi[\eta]d\eta e^{\sigma t} dt \leqq \beta \int_0^y \int_0^{\eta(t)} \eta d\eta e^{\sigma t}\, dt,$$

and

$$\begin{aligned}\beta \int_0^y \tfrac{1}{2}|\eta(t)]^2 e^{\sigma t} dt &\leqq \frac{\beta}{2\alpha^2}\int_0^y \{\psi[\eta(t)]\}^2 e^{\sigma t}\, dt \\ &= \frac{\beta}{2\alpha^2} \| e^{\frac{1}{2}\sigma t}(\psi[\eta])_y \|^2,\end{aligned}$$

we have, using the fact that $\eta(0) = r(0)$,

$$\langle e^{\sigma t}(\psi[z+r])_y, \dot{z}+\dot{r}\rangle \geqq -\int_0^{r(0)} \psi[\eta]d\eta - \frac{\sigma\beta}{2\alpha^2} \| e^{\frac{1}{2}\sigma t}(\psi[z+r])_y \|^2.$$

Upon combining our bounds, we obtain the inequality stated in the lemma.

Lemma 2: Let **A** *and* **B** *denote mappings of* $\mathcal{E}(0,\infty)$ *into itself. Let* σ *be a real constant. Let* $f \in \mathcal{E}(0,\infty)$, $h = \mathbf{B}f$, *and* $g = f + \mathbf{A}h$. *Then*

$$|\langle e^{\sigma t}(\mathbf{A}h)_y, h_y\rangle + \langle e^{\sigma t}(\mathbf{B}f)_y, f_y\rangle| \leqq \| e^{\frac{1}{2}\sigma t}g_y \| \cdot \| e^{\frac{1}{2}\sigma t}h_y \|$$

for all $y \in (0,\infty)$.

Proof of Lemma 2: It is clear that

$$\begin{aligned}\langle e^{\sigma t}(\mathbf{A}h)_y, h_y\rangle + \langle e^{\sigma t}(\mathbf{B}f)_y, f_y\rangle &= \langle e^{\sigma t}(\mathbf{A}h)_y + e^{\sigma t}f_y, h_y\rangle \\ &= \langle e^{\sigma t}g_y, h_y\rangle \\ &= \langle e^{\frac{1}{2}\sigma t}g_y, e^{\frac{1}{2}\sigma t}h_y\rangle\end{aligned}$$

for all $y \in (0,\infty)$. Therefore, by the Schwarz inequality,

$$|\langle e^{\sigma t}(\mathbf{A}h)_y, h_y\rangle + \langle e^{\sigma t}(\mathbf{B}f)_y, f_y\rangle| \leqq \| e^{\frac{1}{2}\sigma t}g_y \| \cdot \| e^{\frac{1}{2}\sigma t}h_y \|$$

for all $y \in (0,\infty)$.

Lemma 3: Let **A** *and* **B** *denote mappings of* $\mathcal{E}(0,\infty)$ *into itself. Let* σ *be a real constant. Let* $f \in \mathcal{E}(0,\infty)$, *and let* $g = f + \mathbf{AB}f$. *Suppose that*

(*i*) *there exists a real constant* k_1' *such that*

$$\langle e^{\sigma t}(\mathbf{B}q)_y, q_y\rangle \geqq k_1' \| e^{\frac{1}{2}\sigma t}q_y \|^2$$

for all $q \in \mathcal{E}(0,\infty)$ *and all* $y \in (0,\infty)$

(*ii*) *there exist a positive constant* k_1, *and nonnegative functions* $k_2(y)$ *and* $k_3(y)$ *such that*

$$\langle e^{\sigma t}(\mathbf{A}q)_y, q_y\rangle \geqq (k_1 - k_1')\| e^{\frac{1}{2}\sigma t}(\mathbf{A}q)_y \|^2 - k_2(y)\| e^{\frac{1}{2}\sigma t}(\mathbf{A}q)_y \| - k_3(y)$$

for all $q \in \mathcal{E}(0,\infty)$ *and all* $y \in (0,\infty)$

(*iii*) *there exists a constant* $k_4 > 0$ *such that* $\| e^{\frac{1}{2}\sigma t}(\mathbf{B}q)_y \| \leqq k_4 \| e^{\frac{1}{2}\sigma t}q_y \|$ *for all* $q \in \mathcal{E}(0,\infty)$ *and all* $y \in (0,\infty)$.

Then there exists a positive constant λ, *depending only on* k_1', k_1 *and* k_4 *such that*

$$\lambda \| e^{\frac{1}{2}\sigma t}f_y \| \leqq \| e^{\frac{1}{2}\sigma t}g_y \| + k_2(y) + [k_3(y)]^{\frac{1}{2}}$$

for all $y \in (0,\infty)$.

Proof of Lemma 3: Let $y \in (0, \infty)$. Using Lemma 2, we have, with $h = \mathbf{B}f$,

$$\langle e^{\sigma t}(\mathbf{A}h)_y, h_y\rangle + \langle e^{\sigma t}(\mathbf{B}f)_y, f_y\rangle \leqq |\langle e^{\sigma t}(\mathbf{A}h)_y, h_y\rangle + \langle e^{\sigma t}(\mathbf{B}f)_y, f_y\rangle| \leqq \| e^{\frac{1}{2}\sigma t} g_y \| \cdot \| e^{\frac{1}{2}\sigma t} h_y \|.$$

Thus,

$$(k_1 - k_1')\| e^{\frac{1}{2}\sigma t}(\mathbf{A}h)_y \|^2 - k_2(y) \| e^{\frac{1}{2}\sigma t}(\mathbf{A}h)_y \| - k_3(y) + k_1' \| e^{\frac{1}{2}\sigma t} f_y \|^2 \leqq k_4 \| e^{\frac{1}{2}\sigma t} g_y \| \cdot \| e^{\frac{1}{2}\sigma t} f_y \|.$$

Using the fact that $(\mathbf{A}h)_y = g_y - f_y$, we have

$$k_1 \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \|^2 + 2k_1'\langle e^{\frac{1}{2}\sigma t} g_y, e^{\frac{1}{2}\sigma t} f_y\rangle - k_1' \| e^{\frac{1}{2}\sigma t} g_y \|^2 - k_2(y) \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \| - k_3(y) \leqq k_4 \| e^{\frac{1}{2}\sigma t} g_y \| \cdot \| e^{\frac{1}{2}\sigma t} f_y \|.$$

Therefore,

$$\begin{aligned} k_1 \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \|^2 &\leqq k_2(y) \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \| + k_3(y) + k_1' \| e^{\frac{1}{2}\sigma t} g_y \|^2 \\ &\quad + (2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t} g_y \| \cdot \| e^{\frac{1}{2}\sigma t} f_y \| \\ &\leqq k_2(y) \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \| + k_3(y) \\ &\quad + (2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \| \cdot \| e^{\frac{1}{2}\sigma t} g_y \| \\ &\quad + (2|k_1'| + k_1' + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|^2. \end{aligned}$$

Let $\rho = \| e^{\frac{1}{2}\sigma t}(g_y - f_y) \|$. Then

$$k_1\rho^2 \leqq [k_2(y) + (2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|]\rho + k_3(y) + (2|k_1'| + k_1' + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|^2,$$

and hence,

$$\begin{aligned} 2\rho &\leqq [k_1^{-1}k_2(y) + k_1^{-1}(2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|] \\ &\quad + \{[k_1^{-1}k_2(y) + k_1^{-1}(2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|]^2 \\ &\quad + 4[k_1^{-1}k_3(y) + k_1^{-1}(2|k_1'| + k_1' + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|^2]\}^{\frac{1}{2}}. \end{aligned}$$

Since $(a^2 + b)^{\frac{1}{2}} \leqq a + b^{\frac{1}{2}}$ for any positive constants a and b,

$$\begin{aligned} \rho &\leqq [k_1^{-1}k_2(y) + k_1^{-1}(2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|] \\ &\quad + [k_1^{-1}k_3(y) + k_1^{-1}(2|k_1'| + k_1' + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|^2]^{\frac{1}{2}} \\ &\leqq [k_1^{-1}k_2(y) + k_1^{-1}(2|k_1'| + k_4) \| e^{\frac{1}{2}\sigma t} g_y \|] \\ &\quad + [k_1^{-1}k_3(y)]^{\frac{1}{2}} + [k_1^{-1}(2|k_1'| + k_1' + k_4)]^{\frac{1}{2}} \| e^{\frac{1}{2}\sigma t} g_y \|. \end{aligned}$$

Using $\| e^{\frac{1}{2}\sigma t} f_y \| \leqq \rho + \| e^{\frac{1}{2}\sigma t} g_y \|$, we see that

$$\| e^{\frac{1}{2}\sigma t} f_y \| \leqq \{1 + k_1^{-1}(2 \mid k_1' \mid + k_4) + [k_1^{-1}(2 \mid k_1' \mid + k_1' + k_4)]^{\frac{1}{2}}\} \| e^{\frac{1}{2}\sigma t} g_y \| + k_1^{-1} k_2(y) + k_1^{-\frac{1}{2}} k_3(y)^{\frac{1}{2}}.$$

This proves the lemma.*

Theorem 1 follows at once from Lemmas 1, 2, and 3 with $\mathbf{B} = \mathbf{F}_1$, $\mathbf{A}$ defined by

$$\mathbf{A}h = \psi[\mathbf{F}_2 h + r]$$

for all $h \in \mathcal{E}(0, \infty)$, $k_1' = k$, $k_4 = c$, $k_1 = c_1\beta^{-1} - \frac{1}{2}\sigma\beta\alpha^{-2} + k$, $k_2(y) = \| e^{\frac{1}{2}\sigma t}(\dot{r} + \delta r)_y \|$, and

$$k_3 = \int_0^{r(0)} \psi(\eta)\, d\eta.$$

4.2 *Proof of Corollary 1*

Since $\| g_y \| \leqq \| g \|$ and $\| (\dot{r} + \delta r)_y \| \leqq \| \dot{r} + \delta r \|$,

$$\lambda \| f_y \| \leqq \| g \| + \| \dot{r} + \delta r \| + \left[\int_0^{r(0)} \psi(\eta) d\eta \right]^{\frac{1}{2}}.$$

The right-hand side is finite and is independent of y. The conclusion of the corollary follows at once.

An equally simple argument establishes the following useful result.

Proposition 1: Suppose that the hypotheses of Lemma 3 are satisfied with $\sigma = 0$. Let $g \in \mathcal{L}_2(0, \infty)$, and let $k_2(y)$ and $k_3(y)$ be uniformly bounded on $[0, \infty)$. Then $f \in \mathcal{L}_2(0, \infty)$, and there exists a positive constant λ, depending only on k_1', k_1 and k_4, such that

$$\lambda \| f \| \leqq \| g \| + \sup_{y \geqq 0} k_2(y) + \sup_{y \geqq 0} [k_3(y)]^{\frac{1}{2}}.$$

4.3 *Proof of Corollary 2*

Since $f \in \mathcal{L}_2(0, \infty)$, we have $\mathbf{F}_1 f \in \mathcal{L}_2(0, \infty)$. Thus, for $t \geqq 0$,

$$| (\mathbf{F}_2 \mathbf{F}_1 f)(t) | \leqq \int_0^t \exp\left[-\int_\tau^t \delta(x) dx \right] | (\mathbf{F}_1 f)(\tau) |\, d\tau$$

$$\leqq \int_0^t e^{-c_1(t-\tau)} | (\mathbf{F}_1 f)(\tau) |\, d\tau$$

* For results directly related to Lemmas 2 and 3, see Section 5.3 of Ref. 1.

in which the last integral approaches zero as $t \to \infty$ (see the proof of Theorem 6 of Ref. 2). Hence,

$$g(t) - (\psi[\mathbf{F}_2\mathbf{F}_1 f + r])(t) \to 0 \quad \text{as} \quad t \to \infty.$$

4.4 *Proof of Corollary 3*

Since $\| e^{\frac{1}{2}\sigma t} g_y \| \leqq \sigma^{-\frac{1}{2}} e^{\frac{1}{2}\sigma y} \sup_{t \geqq 0} | g(t) |$, and

$$\| e^{\frac{1}{2}\sigma t} (\dot{r} + \delta r)_y \| \leqq \sigma^{-\frac{1}{2}} e^{\frac{1}{2}\sigma y} \sup_{t \geqq 0} | (\dot{r} + \delta r)(t) |,$$

we have

$$| (\mathbf{F}_2\mathbf{F}_1 f)(y) | \leqq \frac{\gamma}{\lambda} \left[\sigma^{-\frac{1}{2}} \sup_{t \geqq 0} | g(t) | + \sigma^{-\frac{1}{2}} \sup_{t \geqq 0} | (\dot{r} + \delta r)(t) | + \left[\int_0^{r(0)} \psi(\eta) d\eta \right]^{\frac{1}{2}} \right]$$

for all $y \in (0, \infty)$. This establishes the last inequality of the corollary. Since

$$| f(t) | \leqq \sup_{t \geqq 0} | g(t) | + \sup_{t \geqq 0} | (\psi[\mathbf{F}_2\mathbf{F}_1 f + r])(t) |,$$

it is evident that $f \in \mathcal{L}_\infty(0, \infty)$.

4.5 *Proof of Proposition 2*

As was stated in Section 3.2, results similar to those of Section 3.1 can be obtained for the case in which the nonlinear element ψ in Fig. 1 is replaced by a linear time-varying element that introduces the constraint $w(t) = m(t)v(t)$ for $t \geqq 0$, in which $m(\cdot)$ is a positive bounded function. For that case the proposition that plays the role of Lemma 1 is

Proposition 2: Let $m(\cdot)$ denote a positive bounded measurable function defined on $[0, \infty)$. Let $\dot{m}(t)$ exist on $[0, \infty)$ with $\dot{m} \in \mathcal{L}_\infty(0, \infty)$, and let $\mathbf{F}_2$ be as defined in Assumption 2. Let σ be a real constant. Then

$$\langle e^{\sigma t} m(\mathbf{F}_2 q)_y , q \rangle \geqq \inf_{t \geqq 0} \left[\frac{m(t)\delta(t) - \frac{1}{2}\dot{m}(t) - \frac{1}{2}\sigma m(t)}{m(t)^2} \right] \| e^{\frac{1}{2}\sigma t} m(\mathbf{F}_2 q)_y \|^2$$

for all $y \in (0, \infty)$ and all $q \in \mathcal{E}(0, \infty)$.

Proof of Proposition 2: Let $z = \mathbf{F}_2 q_y$ with $y \in (0, \infty)$. Then

$$\langle e^{\sigma t} m(\mathbf{F}_2 q)_y , q\rangle = \langle e^{\sigma t} m z_y , \dot{z} + \delta z\rangle$$
$$= \langle e^{\sigma t} m z_y , \delta z\rangle + \langle e^{\sigma t} m z_y , \dot{z}\rangle,$$

and

$$\langle e^{\sigma t} m z_y , \dot{z}\rangle = \int_0^y m(t) e^{\sigma t} z(t) \dot{z}(t) dt$$
$$= \tfrac{1}{2} m(t) e^{\sigma t} z(t)^2 \,|_0^y - \tfrac{1}{2} \int_0^y [\dot{m}(t) + \sigma m(t)] e^{\sigma t} z(t)^2 \, dt.$$

Therefore, since $z(0) = 0$,

$$\langle e^{\sigma t} m(\mathbf{F}_2 q)_y , q\rangle = \tfrac{1}{2} m(y) e^{\sigma y} z(y)^2$$
$$+ \int_0^y e^{\sigma t} [m(t)\delta(t) - \tfrac{1}{2}\dot{m}(t) - \tfrac{1}{2}\sigma m(t)] z(t)^2 \, dt.$$

This establishes Proposition 2.

Comment:

The case in which $\mathbf{F}_2$ is the *identity* operator, $m(\cdot)$ is a positive bounded measurable function, and $\dot{m}(t)$ does not necessarily exist, is also of some interest.[1,2] Then

$$\langle e^{\sigma t} m(\mathbf{F}_2 q)_y , q\rangle \geqq \inf_{t \geqq 0} m(t)^{-1} \, \| e^{\frac{1}{2}\sigma t} m(\mathbf{F}_2 q)_y \|^2$$

for all $y \in (0, \infty)$ and all $q \in \mathcal{E}(0, \infty)$.

V. APPENDIX

As a specific application of the material of Section 3.1, we shall prove the following result.

Theorem 2: Let ψ satisfy Assumption 1 of Section III. Let g and r belong to $\mathcal{E}(0, \infty)$. Let w and $\dot{w}$ belong to $\mathcal{L}_1(0, \infty)$ with $w(t) \to 0$ as $t \to \infty$. Suppose that there exists a positive constant ξ such that

$$\inf_{0 \leqq \omega < \infty} \operatorname{Re} [(1 + \xi i\omega) W(i\omega) + \beta^{-1}] > 0,$$

in which $W(i\omega) = \int_0^\infty w(t) e^{-i\omega t} dt$. Let $f \in \mathcal{E}(0, \infty)$ satisfy

$$g(t) = f(t) + \psi\left[\int_0^t w(t - \tau) f(\tau) d\tau + r(t)\right], \qquad t \geqq 0$$

Then

(1.) *if* $g \in \mathcal{L}_2(0,\infty) \cap \mathcal{L}_\infty(0,\infty)$ *with* $g(t) \to 0$ *as* $t \to \infty$, *if* $r \in \mathcal{L}_2(0,\infty) \cap \mathcal{L}_\infty(0,\infty)$ *with* $r(t) \to 0$ *as* $t \to \infty$, *and if* $\dot{r} \in \mathcal{L}_2(0,\infty)$, *then* $f \in \mathcal{L}_2(0,\infty) \cap \mathcal{L}_\infty(0,\infty)$, *there exists a positive constant* λ, *depending only on* ξ, α, β, *and* w, *such that*

$$\lambda \| f \| \leqq \| g \| + \| (\dot{r} + \xi^{-1} r) \| + \left[\int_0^{r(0)} \psi(\eta) d\eta \right]^{\frac{1}{2}},$$

and $f(t) \to 0$ *as* $t \to \infty$

(2.) *if there exists a positive constant* ρ *such that* $e^{\rho t} w \in \mathcal{L}_1(0,\infty) \cap \mathcal{L}_2(0,\infty)$ *and* $e^{\rho t} \dot{w} \in \mathcal{L}_1(0,\infty)$ *with* $e^{\rho t} w(t) \to 0$ *as* $t \to \infty$, *if* g, r, *and* $\dot{r}$ *belong to* $\mathcal{L}_\infty(0,\infty)$, *then* $f \in \mathcal{L}_\infty(0,\infty)$, *and there exists a positive constant* λ_1, *depending only on* ρ, ξ, α, β, *and* w *such that*

$$\lambda_1 \sup_{t \geqq 0} \left| \int_0^t w(t-\tau) f(\tau) d\tau \right| \leqq \sup_{t \geqq 0} | g(t) | + \sup_{t \geqq 0} | (\dot{r} + \xi^{-1} r)(t) | + \left[\int_0^{r(0)} \psi(\eta) d\eta \right]^{\frac{1}{2}}.$$

Proof of Theorem 2:

Let $\mathbf{F}$ be defined by

$$(\mathbf{F}q)(t) = \int_0^t w(t-\tau) q(\tau) d\tau, \qquad t \geqq 0.$$

Then $\mathbf{F} = \mathbf{F}_2 \mathbf{F}_1$, where $\mathbf{F}_2$ is as defined in Assumption 2 of Section III with $\delta(x) = \xi^{-1}$, and

$$(\mathbf{F}_1 q)(t) = \int_0^t [\dot{w}(t-\tau) + \xi^{-1} w(t-\tau)] q(\tau) d\tau + w(0+) q(t), \qquad t \geqq 0.$$

Let

$$\zeta = \inf_{0 \leqq \omega < \infty} \operatorname{Re} [(1 + \xi i\omega) W(i\omega) + \beta^{-1}].$$

Then

$$\langle (\mathbf{F}_1 q)_y, q \rangle \geqq \xi^{-1} (\zeta - \beta^{-1}) \| q_y \|^2$$

for all $q \in \mathcal{E}(0,\infty)$ and all $y \in (0,\infty)$ [see the remark following Corollary 1]. Thus conditions (i) and (ii) of Theorem 1 are satisfied for $\sigma = 0$.

Therefore, by Corollary 1, the hypotheses of (i) of Theorem 2 imply that $f \varepsilon \mathfrak{L}_2(0,\infty)$ and that $\| f \|$ is bounded as indicated. By Corollary 2, we have $f(t) \to 0$ as $t \to \infty$. Further, since $f \varepsilon \mathfrak{L}_2(0,\infty)$, we have $\mathbf{F}_1 f \varepsilon \mathfrak{L}_2(0,\infty)$, and, by the Schwarz inequality, $\mathbf{F}_2\mathbf{F}_1 f \varepsilon \mathfrak{L}_\infty(0,\infty)$. Therefore $g - \psi[\mathbf{F}f + r] \varepsilon \mathfrak{L}_\infty(0,\infty)$.

Suppose now that the hypotheses of (ii) are satisfied. Then since both $| W(i\omega - x) - W(i\omega) |$ and $| (i\omega - x)W(i\omega - x) - i\omega W(i\omega) |$ approach zero uniformly in ω as $x \to 0+$, there exists a positive constant σ such that $\sigma < \min [2\rho,\alpha^2\zeta(\beta\xi)^{-1}]$ and

$$\inf_{0\leqq\omega<\infty} \operatorname{Re} \{[1 + \xi(i\omega - \tfrac{1}{2}\sigma)]W(i\omega - \tfrac{1}{2}\sigma) + \beta^{-1}\} \quad > \tfrac{1}{2}\zeta.$$

Hence,

$$\langle e^{\sigma t}(\mathbf{F}_1 q)_y, q\rangle \geqq \xi^{-1}(\tfrac{1}{2}\zeta - \beta^{-1}) \| e^{\frac{1}{2}\sigma t} q_y \|^2$$

for all $y \varepsilon (0,\infty)$ and all $q \varepsilon \mathcal{E}(0,\infty)$. Thus, by Corollary 3 and the remarks following Corollary 3, we have $f \varepsilon \mathfrak{L}_\infty(0,\infty)$ with

$$\sup_{t\geqq 0} \left| \int_0^t w(t - \tau)f(\tau)d\tau \right|$$

bounded as stated in Theorem 2.

Comments:

Our assumption that $f \varepsilon \mathcal{E}(0,\infty)$ is satisfied if f is locally (Lebesgue) integrable on $(0,\infty)$, since then (under the stated assumptions on g, ψ, w, and r):

$$\int_0^t w(t - \tau)f(\tau)d\tau$$

is continuous on $[0,\infty)$ and hence,

$$g - \psi\left[\int_0^t w(t - \tau)f(\tau)d\tau + r\right] \varepsilon \mathcal{E}(0,\infty).$$

A result closely related to the first part of Theorem 2 has been proved by Desoer.[12]

REFERENCES

1. Sandberg, I. W., Some Results on the Theory of Physical Systems Governed by Nonlinear Functional Equations, B.S.T.J., *44*, May–June, 1965, p. 871.
2. Sandberg, I. W., On the $\mathfrak{L}_2$-Boundedness of Solutions of Nonlinear Functional Equations, B.S.T.J., *43*, July, 1964, p. 1581.
3. Sandberg, I. W., On the Boundedness of Solutions of Nonlinear Integral Equations, B.S.T.J., *44*, March, 1965, p. 439.

4. Sandberg, I. W., On the Properties of Some Systems that Distort Signals — II, B.S.T.J., *43*, January, 1964, p. 91.
5. Sandberg, I. W., On the Response of Nonlinear Control Systems to Periodic Input Signals, B.S.T.J., *43*, May, 1964, p. 911.
6. Sandberg, I. W., On the Stability of Solutions of Linear Differential Equations with Periodic Coefficients, J. Soc. Indust. Appl. Math., *12*, June, 1964, p. 487.
7. Zames, G., On the Stability of Nonlinear Time-Varying Feedback Systems, Proc. Nat. Electronics Conference, *20*, October, 1964, p. 725.
8. Aizerman, M. A., and Gantmacher, F. R., *Absolute Stability of Regulator Systems*, (trans. by E. Polak), Holden-Day, San Francisco, 1964.
9. Brockett, R. W., and Forys, L. J., On the Stability of Systems Containing a Time-Varying Gain, Proc. Second Annual Allerton Conference on Circuit and System Theory, Monticello, Illinois, September, 1964.
10. Brockett, R. W., and Willems, J. W., Frequency Domain Stability Criteria, Part I; to appear in IEEE PGAC, July, 1965.
11. Dewey, A. G., and Jury, E. I., A Stability Inequality for a Class of Nonlinear Feedback Systems, to appear.
12. Desoer, C. A., A Generalization of the Popov Criterion, IEEE Trans. Automatic Control, *AC-10*, April, 1965, p. 182.

Spectra for a Class of Asynchronous FM Waves

By R. R. ANDERSON, J. E. MAZO and J. SALZ

(Manuscript received July 6, 1965)

The spectral density of a frequency modulated carrier is evaluated for the case when the modulating baseband wave is a "quantized" random facsimile signal. By this nonsynchronous form of modulation we mean holding one of the allowed set of transmitted frequencies for a finite, but randomly distributed, time before switching to another frequency while maintaining phase continuity. Emphasis is given to the Poisson case of exponentially distributed intervals between transitions, and some typical curves for discrete level and continuous level situations are included.

I. INTRODUCTION

In facsimile data transmission systems, printed or pictorial information is converted into electrical signals by optical means. At any instant the signal corresponds to a definite grey level of the facsimile copy. The resulting electrical wave is an analog signal and can be transmitted as such. In some applications only black-and-white images need be transmitted and therefore the electrical signal may be quantized into only two levels. If more detail is desired, multilevel quantization can be applied.

It is possible to model such a quantized signal by considering a random sequence of points on the time (t) axis. At each point a transition may occur in the signal. The value of the signal between transitions is a constant, taking on one of N different values. For the black-and-white case, there are only two permissible values, either $+1$ or -1. The quantized facsimile signal is statistically characterized by specifying the distribution of the points on the t axis and the distribution of the amplitudes between transitions.

In this paper we concern ourselves with the spectral density of a carrier wave whose frequency is modulated by quantized facsimile signals. The spectral density is a useful item in the statistical description of such

a signal in that it furnishes an estimate of bandwidth requirements. It often is also used to evaluate mutual interference between channels.

So far as is known, the amplitude modulation case is the only one hitherto covered in literature. However, we were prompted to examine the FM case as FM is currently used in facsimile data sets. From a practical point of view the most interesting case is that in which the phase is continuous at the transitions, as may be obtained from keying a single oscillator. This case differs from previous results[1] in that the transitions occur at random times.

The present paper gives a complete solution for the spectrum for an arbitrary distribution of the interval between transitions as well as arbitrary distribution of amplitudes. We treat an important special case of Poisson transitions, for which we present our results graphically in terms of the important parameters of the process.

An interesting feature is the rapidity with which the spectral density falls off with frequency measured from midband as compared with the AM case. The extent to which spectral peaking occurs at the average signaling frequency for some range of parameters is another curious feature. As would be expected from the asynchronous nature of the modulation there can be no steady sine-wave components in the process and therefore there are no discrete components in the spectrum.

II. ANALYSIS

The baseband facsimile signal is constructed in the following form. Pick a finite set $\{\mathbf{t}\}$ at random and arrange the points such that

$$0 = t_0 < t_1 < t_2 < \cdots t_N = T. \tag{1}$$

Define a set of functions

$$g_{\Delta_n}(t - t_n) = \begin{cases} 1, & t_n \leqq t \leqq t_{n+1} \\ 0, & \text{elsewhere} \end{cases} \tag{2}$$

where

$$\Delta_n = t_{n+1} - t_n .$$

In terms of (1) and (2), construct the baseband signal $x(t)$ as the following time series

$$x(t) = \sum_{n=0}^{n=N-1} a_n g_{\Delta_n}(t - t_n) \tag{3}$$

where $\mathbf{a} = (a_0, a_1, \cdots, a_{N-1})$ is an additional arbitrary set of identically distributed random variables.

The instantaneous phase $\psi(t)$ is represented as follows

$$\psi(t) = \omega_c t + \omega_d \int_0^t x(t')dt' + \varphi \equiv \psi_1(t) + \varphi, \qquad 0 \leqq t \leqq T \tag{4}$$

where φ is uniformly distributed on $[0,2\pi]$, giving the value of the phase at $t = 0$. The instantaneous frequency is $d\psi(t)/dt = \omega_c + \omega_d x(t)$, where ω_c is the angular carrier frequency and ω_d is the minimum angular frequency deviation.

The FM wave whose spectrum we wish to examine has the following representation

$$\begin{aligned} S(t) &= A \cos [\psi(t)] \\ &= (A/2) \exp \{i\psi(t)\} + (A/2) \exp \{-i\psi(t)\}, \end{aligned} \tag{5}$$

where A is a real amplitude. The spectral density of $S(t)$, $G(\omega)$, defines the average power in a unit bandwidth. Formally, it may be obtained as

$$G(\omega) = \lim_{T\to\infty} (2/T) \langle | S(\omega,T) |^2 \rangle, \qquad \omega > 0,$$

where $S(\omega,T)$ is the Fourier transform of $S(t)$ given by

$$S(\omega,T) = \int_0^T S(t) \exp (-i\omega t)\, dt \tag{7}$$

and the symbol $\langle \cdot \rangle$ denotes the ensemble average over all the random variables in $S(\omega,T)$.

One may write $S(\omega,T)$ as

$$S(\omega,T) = (A/2)e^{i\varphi}W_1(\omega,T) + (A/2)e^{-i\varphi}W_2(\omega,T) \tag{8}$$

where W_1 and W_2 are the Fourier transforms of $\exp [i\psi_1(t)]$ and $\exp [-i\psi_1(t)]$ respectively.

The ensemble average of $| S(\omega,T) |^2$ over φ is readily carried out to obtain

$$\langle | S(\omega,T) |^2 \rangle_\varphi = (A^2/4) | W_1(\omega,T) |^2 + (A^2/4) | W_2(\omega,T) |^2. \tag{9}$$

We proceed to evaluate $\langle | W_1(\omega,T) |^2 \rangle$:

$$\begin{aligned} W_1(\omega,T) &= \int_0^T \exp i\{\psi_1(t) - \omega t\} dt \\ &= \sum_{k=0}^{k=N-1} \int_{t_k}^{t_{k+1}} \exp i\{\psi_1(t) - \omega t\} dt. \end{aligned} \tag{10}$$

We observe from (4) that in the interval $[t_k, t_{k+1}]$,

$$\psi_1(t) = \omega_d \sum_{n=0}^{n=k-1} a_n\Delta_n + \omega_d a_k(t - t_k) + \omega_c t. \tag{11}$$

Inserting this expression into (10), and using

$$t_k = \sum_{n=0}^{n=k-1} \Delta_n ,$$

we find that

$$W_1(\omega,T) = \sum_{k=0}^{k=N-1} \frac{1}{i\lambda_k}\left[\exp\left\{i\sum_{n=0}^{n=k}\lambda_n\Delta_n\right\} - \exp\left\{i\sum_{n=0}^{n=k-1}\lambda_n\Delta_n\right\}\right] \tag{12}$$

where

$$\lambda_n = \omega_d a_n - \omega + \omega_c .$$

Multiplying (12) by its complex conjugate we obtain

$$\begin{aligned} |\, W_1(\omega,T)\,|^2 = 2\mathrm{Re}\Bigg[&\sum_{k=0}^{k=N-1} \frac{1 - \exp i\lambda_k\Delta_k}{\lambda_k^2} \\ &+ \sum_{\substack{k,s=0 \\ k>s}}^{N-1} \frac{1}{\lambda_k\lambda_s}\left(\exp\left\{i\sum_{n=s+1}^{k}\lambda_n\Delta_n\right\} + \exp\left\{i\sum_{n=s}^{k-1}\lambda_n\Delta_n\right\}\right. \\ &\left. - \exp\left\{i\sum_{n=s+1}^{n=k-1}\lambda_n\Delta_n\right\} - \exp\left\{i\sum_{n=s}^{n=k}\lambda_n\Delta_n\right\}\right)\Bigg], \end{aligned} \tag{13}$$

where $\mathrm{Re}(\cdot)$ denotes the real part.

At this point we must specify in more statistical detail the sets of random vectors $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, \cdots, \Delta_N)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_N)$. In our original representation (3), the Δ_n's are the intervals between transitions. We adopt the reasonable assumption that these intervals are independent. On the other hand the random variables λ_n which are related to the amplitude a_n by (12) are not independent if we consider only observable transitions. Clearly for observable transitions one insists that $a_n \neq a_{n-1}$, and thus adjacent amplitudes are dependent. To remedy this awkwardness in the analysis, we construct an alternate random process with independent a_n's by admitting virtual transitions. That is, we do not require that the signal change at every t_n given in the sequence (1). We will show later how such a process, entirely equivalent to the original, may be constructed. For the present, we shall merely assume that the λ_n's are independent.

For fixed N the average of (13) with respect to $\boldsymbol{\Delta}$ and $\boldsymbol{\lambda}$ becomes

$$\langle |\, W_1(\omega,T)\,|^2\rangle_{\Delta,\lambda} = 2\mathrm{Re}\left[N\left\langle \frac{1}{\lambda^2}\right\rangle - N\left\langle \frac{\exp i\lambda\Delta}{\lambda^2}\right\rangle \right.$$
$$+ \sum_{\substack{k,s=0\\ k>s}}^{k,s=N-1} \left\{ \left\langle \frac{\exp i\sum_{s+1}^{k}\lambda_n\Delta_n}{\lambda_s\lambda_k}\right\rangle \right.$$
$$+ \left\langle \frac{\exp i\sum_{s}^{k-1}\lambda_n\Delta_n}{\lambda_s\lambda_k}\right\rangle \qquad (14)$$
$$- \left\langle \frac{\exp i\sum_{s+1}^{k-1}\lambda_n\Delta_n}{\lambda_s\lambda_k}\right\rangle$$
$$\left.\left. - \left\langle \frac{\exp i\sum_{s}^{k}\lambda_n\Delta_n}{\lambda_s\lambda_k}\right\rangle \right\}\right].$$

The respective averages in (14) may be expressed in terms of the characteristic function of Δ. A typical calculation yields

$$\left\langle \frac{\exp i\sum_{s+1}^{k}\lambda_n\Delta_n}{\lambda_s\lambda_k}\right\rangle_{\lambda,\Delta} = \left\langle \frac{1}{\lambda_s\lambda_k}\prod_{n=s+1}^{n=k}\exp i\lambda_n\Delta_n\right\rangle_{\lambda,\Delta}$$
$$= \left\langle \frac{1}{\lambda_s\lambda_k}\left\langle \prod_{n=s+1}^{n=k}\exp i\lambda_n\Delta_n\right\rangle_{\Delta}\right\rangle_{\lambda} \qquad (15)$$
$$= \left\langle \frac{1}{\lambda_s\lambda_k}\prod_{n=s+1}^{n=k}C_\Delta(\lambda_n)\right\rangle_{\lambda}$$
$$= \left\langle \frac{1}{\lambda}\right\rangle_{\lambda}\left\langle \frac{C_\Delta(\lambda)}{\lambda}\right\rangle_{\lambda}[\langle C_\Delta(\lambda)\rangle_\lambda]^{k-s-1}$$

where $C_\Delta(\lambda) = \langle \exp(i\lambda\Delta)\rangle_\Delta$ is the characteristic function of the random variable Δ.

Using the same procedure as above on every term in (14) we obtain

$$\langle |\, W_1(\omega,T)\,|^2\rangle_{\Delta,\lambda} = 2\mathrm{Re}\left[N\left\langle \frac{1-C_\Delta(\lambda)}{\lambda^2}\right\rangle_{\lambda}\right.$$
$$\left. - \left(\left\langle \frac{1-C_\Delta(\lambda)}{\lambda}\right\rangle\right)^2 \sum_{\substack{k,s\\ k>s}} \rho^{k-s-1}\right] \qquad (16)$$

where

$$\rho \equiv \langle C_\Delta(\lambda)\rangle_\lambda .$$

Since $|\rho| < 1$, the series in (16) can be summed. If we designate the average of N by $\bar{N}$, divide (16) by T and take the limit as $T \to \infty$ such that

$$\nu = \lim_{T\to\infty} (\bar{N}/T)$$

we obtain

$$\lim_{T\to\infty} \frac{1}{T} \langle |\, W_1(\omega,T)\,|^2 \rangle_{\Delta,\lambda,N} = 2\nu \mathrm{Re}\left[\left\langle \frac{1 - C_\Delta(\lambda)}{\lambda^2} \right\rangle_\lambda - \left(\left\langle \frac{1 - C_\Delta(\lambda)}{\lambda} \right\rangle_\lambda\right)^2 \frac{1}{1-\rho}\right], \tag{17}$$

where we made use of the identity

$$\sum_{\substack{k,s=0\\k>s}}^{N-1} \rho^{k-s-1} = \sum_{n=1}^{N-1} (N-n)\rho^{n-1}.$$

We can repeat the identical operations on $W_2(\omega,T)$ in (9) that we have just concluded on $W_1(\omega,T)$ and obtain an identical expression except that $\omega - \omega_c$ in W_1 will be replaced by $\omega + \omega_c$.

Combining (17) with (9) and (6), we write down the positive image spectrum as our general result, namely

$$G_+(\omega) = A^2\nu \mathrm{Re}\left\langle \left[\frac{1 - C_\Delta(\lambda)}{\lambda^2}\right\rangle_\lambda - \frac{1}{1 - \langle C_\Delta(\lambda)\rangle_\lambda} \cdot \left(\left\langle \frac{1 - C_\Delta(\lambda)}{\lambda} \right\rangle_\lambda\right)^2\right]. \tag{18}$$

This result is general and applies when the choice of amplitudes is made independently at every t_n point. As remarked earlier, in a real facsimile process the choice of amplitudes is constrained. If a transition is to occur at every t_n point, the adjacent amplitudes must be correlated. We now show that the real process can indeed be represented in terms of independent amplitudes by the expediency of introducing virtual transitions. We write down the following equality

$$\sum_{n=0}^{n=N-1} a_n g_{\Delta_n}(t - t_n) = \sum_{n=0}^{n=N'-1} b_n g_{\Delta'_n}(t - t_n'). \tag{19}$$

The process on the left is the artificial process with independent a_n's and Δ_n's, whereas the process on the right has the b_n's correlated, as observation would require, and a different set of t_n''s representing the real process. Clearly the two processes are equivalent if one can find a transformation from the primed set of variables on the right to the unprimed set on the left.

We observe the following characteristics of the two representations. The representation on the right demands that a transition occur at every $t = t_n'$ for all n. To accomplish this b_n must be different from b_{n-1}, thus restricting the choice of the b_n's. The representation on the left admits independent choices of the a_n's, thus giving rise to masking of some transitions, since in fact, if $a_{n-1} = a_n$ there cannot be a transition at t_n. Furthermore, the set of t_n''s is a proper subset of the set of t_n's. To make the representation on the left useful, we must find how the two parameter sets transform. Toward this end define the following set of random variables

$$X_n = f(a_n, a_{n-1}) = \begin{cases} 1, & a_n = a_{n-1} \\ 0, & a_n \neq a_{n-1} \end{cases} \tag{20}$$

for $n = 1, 2, 3 \cdots$.

Let P_j be the probability of obtaining a sequence of exactly jX's out of $N + j$ taking on the value unity. Then the probability $P(N)$ of obtaining exactly N real transitions in T seconds in the process on the left of (19) is

$$P(N) = \sum_{j=0}^{j=\infty} f(N + j)P_j, \tag{21}$$

where $f(N + j)$ is the probability of exactly $N + j$ transitions in the process on the left of (19). Equation (21) is a linear summand equation from which we would like to find a suitable $f(\cdot)$ from the knowledge of $P(\cdot)$ and P_j.

Not intending to make a general study of solutions of (21), one simple solution is presented in the next section for the case of exponentially distributed intervals Δ. In preparation for this discussion, we point out that if a_n is a discrete multilevel random variable with equally likely probabilities the set of random variables $\{X_n\}$ in (20) is independent and therefore P_j is the binomial probability distribution. To demonstrate that indeed the set $\{X_n\}$ is independent we have to show that the conditional probability distribution of X_n given X_{n-1} does not depend on X_{n-1}.

With this in mind consider the joint characteristic function $C(\omega_1, \omega_2)$ of X_n and X_{n+1} :

$$\begin{aligned} C(\omega_1, \omega_2) &= \langle \exp\ (i\omega_1 X_n + i\omega_2 X_{n+1}) \rangle_{x_n, x_{n+1}} \\ &= \langle \exp\{i\omega_1 f(a_n, a_{n-1})\} \cdot \exp\{i\omega_2 f(a_n, a_{n+1})\} \rangle_{a_n, a_{n-1}, a_{n+1}} . \end{aligned} \tag{22}$$

From (20), fixing a_n and averaging first over a_{n+1} and then over a_{n-1}, we obtain

$$\begin{aligned} C(\omega_1, \omega_2) = \langle \langle \exp\{i\omega_1 f(a_n, a_{n-1})\} \rangle_{a_{n-1}} \\ \cdot \langle \exp\{i\omega_2 f(a_n, a_{n+1})\} \rangle_{a_{n+1}} \rangle_{a_n} . \end{aligned} \tag{23}$$

Since $P\{a_n = y_k\} = (1/M)$, $k = 1, 2, \cdots, M$, independent of k we can write the last equation as

$$\begin{aligned} C(\omega_1, \omega_2) = (1/M) \sum_{k=1}^{k=M} \langle \exp\{i\omega_1 f(y_k, a_{n-1})\} \rangle_{a_{n-1}} \\ \cdot \langle \exp\{i\omega_2 f(y_k, a_{n+1})\} \rangle_{a_{n+1}} . \end{aligned} \tag{24}$$

Now

$$\langle \exp\{i\omega_1 f(y_k, a_{n-1})\} \rangle_{a_{n-1}} = \frac{1}{M} \exp(i\omega_1) + \left(1 - \frac{1}{M}\right), \tag{25}$$

and likewise

$$\langle \exp\{i\omega_2 f(y_k, a_{n+1})\} \rangle_{a_{n+1}} = \frac{1}{M} \exp(i\omega_2) + \left(1 - \frac{1}{M}\right).$$

Since neither of the above averages under the summation sign in (24) depend on k, the joint characteristic function $C(\omega_1, \omega_2) = C(\omega_1)C(\omega_2)$ which says that the random variables X_n and X_{n-1} are independent for all n. Clearly the above arguments still hold if a_n is allowed to take on a continuum of values.

III. POISSON TRANSITIONS

As a special case we assume that the number N of t points in a fixed interval T obeys the Poisson probability law; consequently the probability density of the intervals Δ between transitions is exponentially distributed, namely

$$P(\Delta) = \begin{cases} \nu e^{-\nu\Delta} & \Delta \geqq 0 \\ 0, & \Delta < 0. \end{cases} \tag{26}$$

The characteristic function of Δ is then

$$C_\Delta(\lambda) = \langle \exp i\lambda\Delta \rangle = [\nu/(\nu - i\lambda)]. \tag{27}$$

In particular it follows from (27) that

$$\lambda = \frac{\nu}{i} \frac{C_\Delta(\lambda) - 1}{C_\Delta(\lambda)}. \tag{28}$$

When (28) is substituted into (18) we obtain a very simple result for the spectrum, namely

$$G_+(\omega) = \frac{A^2}{\nu} \operatorname{Re}\left[\frac{\langle C_\Delta(\lambda)\rangle_\lambda}{1 - \langle C_\Delta(\lambda)\rangle_\lambda}\right], \tag{29}$$

where we made use of the fact that

$$\left\langle \frac{C_\Delta(\lambda)}{1 - C_\Delta(\lambda)} \right\rangle_\lambda = i\nu \left\langle \frac{1}{\lambda} \right\rangle_\lambda$$

which is purely imaginary.

For black and white transmission $a_n = \pm 1$ with equal probability, and the spectrum reduces to

$$G_+(\omega) = \frac{A^2}{\nu} \operatorname{Re}\left[\frac{\frac{1}{2}\frac{\nu}{\nu - i\lambda_1} + \frac{1}{2}\frac{\nu}{\nu - i\lambda_2}}{1 - \frac{1}{2}\frac{\nu}{\nu - i\lambda_1} - \frac{1}{2}\frac{\nu}{\nu - i\lambda_2}}\right], \tag{30}$$

where $\lambda_1 = \omega_d - \omega + \omega_c$ and $\lambda_2 = -\omega_d - \omega + \omega_c$ from (12). By algebraic manipulation (30) is reduced to

$$G_+(\omega) = \frac{A^2}{\nu} \frac{\nu^2\omega_d^2}{[\omega_d^2 - (\omega - \omega_c)^2]^2 + \nu^2(\omega - \omega_c)^2}. \tag{31}$$

We see from this expression that the spectrum falls off as the fourth power of frequency. In a forthcoming section we shall present graphs of the various spectra.

It is instructive to examine the physical meaning of the parameter ν. This parameter is the average number of transitions per unit time of the virtual process. In fact the average number of transitions in the real process is $\nu[1 - (1/M)]$, with the Poisson form of the density being preserved. To show that this is so, we observe that a solution of (21) is a Poisson probability distribution. In general, for M levels with Pr $[a_n = k] = 1/M$, $k = 1, 2, \cdots, M$ we have, from the previous section,

$$P_j = \frac{(N + j)!}{j!N!} \left(\frac{1}{M}\right)^j \left(1 - \frac{1}{M}\right)^N.$$

If $f(N + j)$ is assumed to be

$$f(N + j) = \frac{e^{-\nu}\nu^{N+j}}{(N + j)!}, \tag{32}$$

we find from (21) that

$$P(N) = \sum_{j=0}^{j=\infty} f(N + j)P_j = \frac{\nu_1^N e^{-\nu_1}}{N!} \tag{33}$$

where $\nu_1 = \nu[1 - (1/M)]$. Thus Poisson transitions with parameter ν in the simple representation correspond to Poisson transitions with parameter ν_1 in the real process.

IV. GRAPHICAL REPRESENTATION

In this section we present graphical results for the case of Poisson distributed transitions. It is important to bear in mind the distinction between the parameters for the virtual transitions, with which the calculations are done, and the parameters of the real process, for which the results are reported. Here, we shall reverse the convention of (19) and use primes to distinguish "virtual" parameters. Frequency and fequency deviation are normalized to the average transition rate, i.e.,

$$\begin{aligned} \beta' &= \frac{M - 1}{M}\beta = \frac{M - 1}{M} \cdot \frac{\omega - \omega_c}{2\pi\nu} \\ K' &= \frac{M - 1}{M}K = \frac{M - 1}{M}\frac{\omega_d}{\pi\nu}. \end{aligned} \tag{34}$$

The normalized characteristic function, with λ_n defined as in (12), may then be written as

$$C_\Delta(a) = \frac{1}{1 - 2\pi i\left(\frac{aK'}{2} - \beta'\right)}. \tag{35}$$

As our first example we consider a to be a discrete random variable taking on the possible values $2n - (M + 1)$, $n = 1, 2, \cdots, M$ with equal probability $(1/M)$. Using these facts

$$\langle C_\Delta(a)\rangle_a = \bar{x}(1 - iZ\pi\beta') + i\pi k'\bar{y}, \tag{36}$$

where

$$\bar{x} = \frac{1}{M}\sum_{n=1}^{n=M} \frac{1}{1 + 4\pi^2\left(\frac{a_nK'}{2} - \beta'\right)^2},$$

$$\bar{y} = \frac{1}{M} \sum_{n=1}^{n=M} \frac{a_n}{1 + 4\pi^2 \left(\frac{a_n K'}{2} - \beta' \right)^2},$$

where we let $a_n = 2n - (M + 1)$. Using (36) in (29) we obtain, for the virtual process

$$\frac{\nu G_+(\beta')}{A^2} = \frac{1 - \bar{x}}{(1 - \bar{x})^2 + \pi_2(2\bar{x}\beta' - K'\bar{y})^2} - 1, \tag{37}$$

and for the real process

$$\frac{\nu G_+(\beta)}{A^2} = \frac{M - 1}{M} \cdot \frac{\nu G_+(\beta')}{A^2}. \tag{38}$$

We have plotted this normalized spectral density (38) as a function

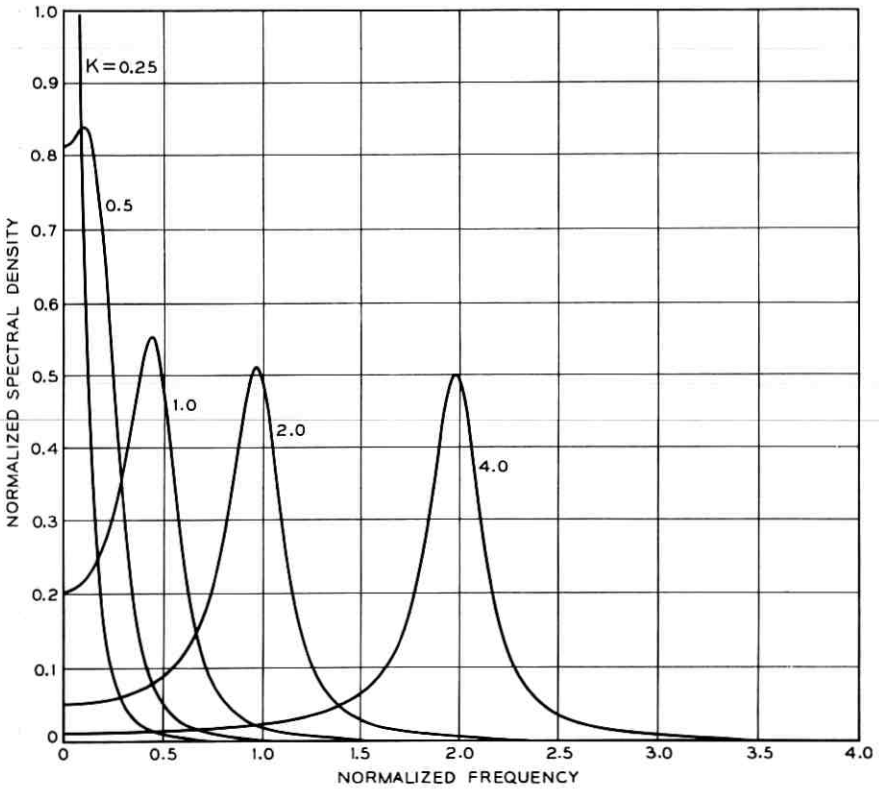


Fig. 1 — Spectra for 2-level FM FAX signals.

of the normalized frequency β for several values of normalized frequency deviation. Only the positive spectrum is shown since it is symmetrical about the normalized carrier, $\beta = 0$.

The binary case is shown on Fig. 1. We note that these spectra contain none of the spectral lines which appear with synchronous modulation. There is, nevertheless, a tendency for the spectrum to be concentrated about the frequency $\beta = k/2$. Unnormalized, this is $\omega = \omega_c \pm K\pi\nu$.

Higher level cases, $M = 4$ and 8, are shown in Figs. 2 and 3, respectively. These are similar to the binary case except that they have M levels and therefore M frequencies where concentration tends to occur. These frequencies are approximately $\beta = (2n - 1)K/2$, $n = 1, 2, \cdots, M/2$.

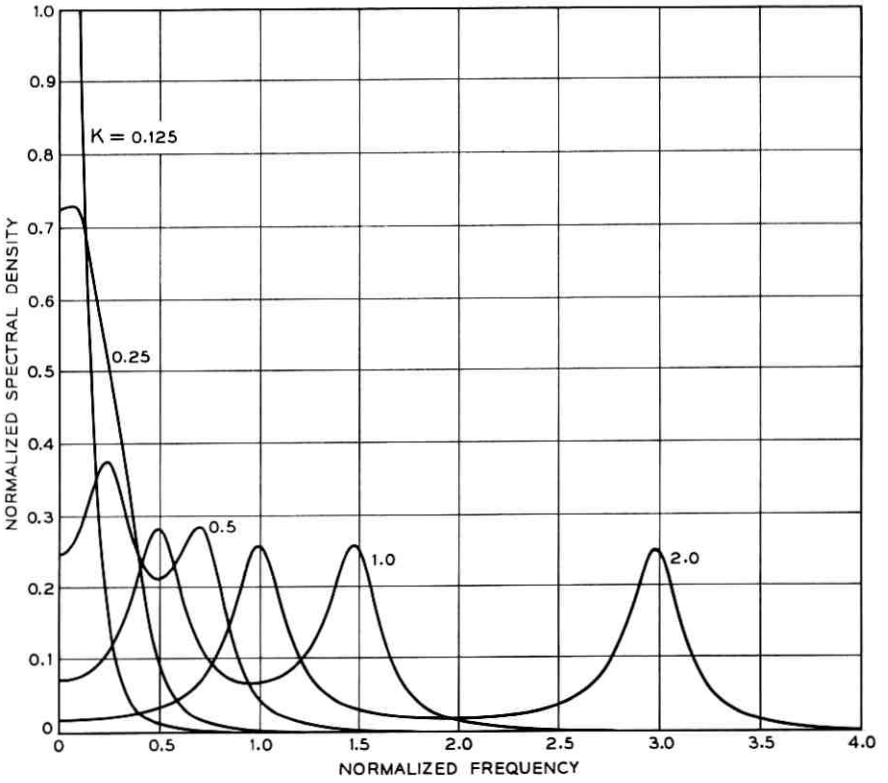


Fig. 2 — Spectra for 4-level FM FAX signals.

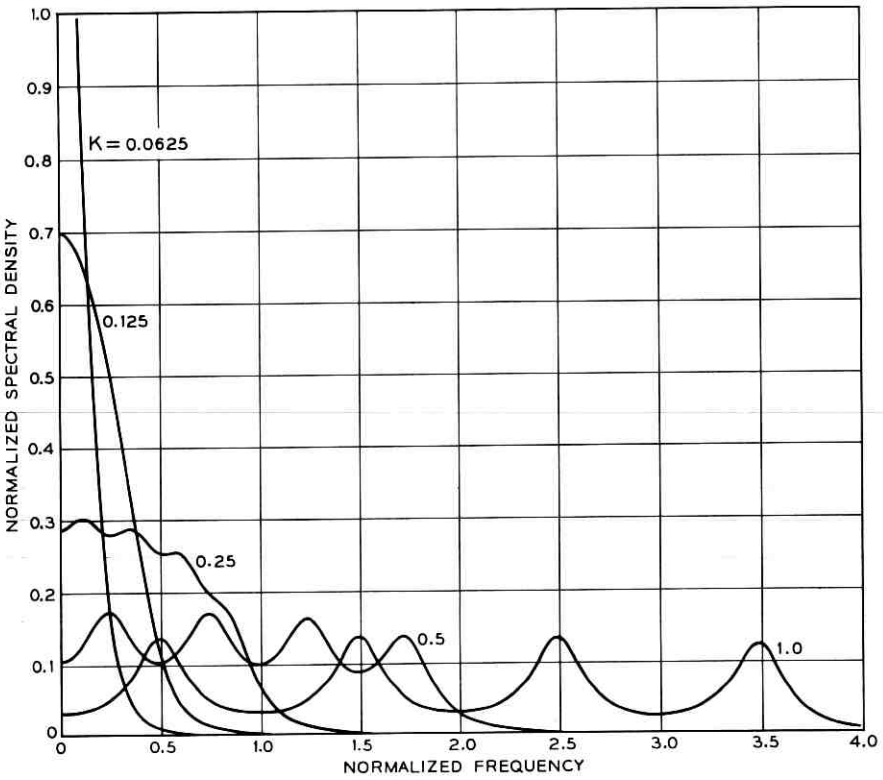


Fig. 3 — Spectra for 8-level FM FAX signals.

As a second example, we shall retain the Poisson distribution of transitions times, but allow the amplitudes to be continuously distributed over the interval $[-r,r]$. This is not true analog representation, but corresponds to "sample-and-hold" operation with exponential holding times. For this case the probability density of a is

$$P(a) = 1/2r, \qquad -r \leqq a \leqq r \tag{39}$$

and the expected value of the characteristic function becomes

$$\langle C \rangle_a = \frac{\nu}{2r} \int_{-r}^{r} \frac{da}{\nu + i\omega - i\omega_d a}$$

$$= i \frac{\nu}{2r\omega_d} \ln \frac{\nu + i\omega - i\omega_d r}{\nu + i\omega + i\omega_d r}.$$

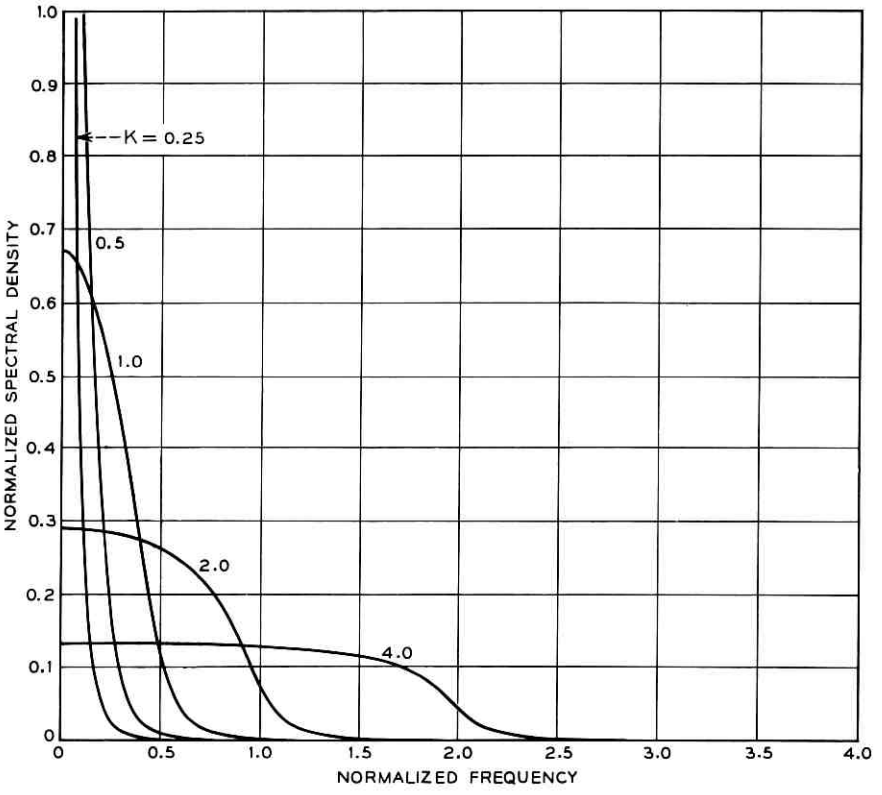


Fig. 4 — Spectra for continuous distribution FM FAX signals.

Inserting this into (29) we obtain

$$\frac{G_+(\beta)}{A^2} = \frac{R^2 + (\varphi_0 + 2\pi K)\varphi_0}{R^2 + (\varphi_0 + 2\pi K)^2}, \tag{41}$$

where

$$R = \ln\left[\frac{1 + (2\pi\beta - \pi K)^2}{1 + (2\pi\beta + \pi K)^4}\right]^{\frac{1}{4}}$$

$$\varphi_0 = \arctan\,(2\pi\beta - \pi K) - \arctan\,(2\pi\beta + \pi K).$$

The normalized frequency deviation K is now modified to include r, namely

$$K = r\omega_d/\pi\nu. \tag{42}$$

For this continuous case, spectra for various K are shown on Fig. 4. The shape of these is very nearly rectangular, with height of $1/2K$ and width of $K/2$, for the displayed positive spectrum.

Considering the above results, together with results from a previous study on digital FM,[1] it is interesting to observe that in all the plots the shape of the spectrum is approximately the same as the first order probability density function of the baseband modulation process when K is large, in accordance with the adiabatic theorem.[2]

REFERENCES

1. Anderson, R. R., and Salz, J., Spectra of Digital FM, B.S.T.J., *44*, July, 1965, pp. 1165–1189.
2. Blachman, N. M., Limiting Frequency-Modulation Spectra, Information and Control *1*, 1957, pp. 26–37.

Probability of Error for Quadratic Detectors

By J. E. MAZO and J. SALZ

(Manuscript received July 28, 1965)

A procedure is presented for evaluating the performance of a general class of digital detectors which square or multiply signal waves contaminated by gaussian noise. In addition to simplifying and unifying the treatment of a number of previously solved problems and some hitherto unsolved ones, the method achieves a considerable advance toward a complete evaluation of postdetection filtering. In contrast to most of the earlier related work, which is typically restricted to filters described as accepting low-frequency difference products and rejecting high-frequency sum products, the present analysis offers a tractable inclusion of filters which do significant selective processing of the detected low-frequency signal and noise components.

A principal goal sought is the asymptotic form approached by the error probability expressed as a function of the signal-to-noise ratio when the latter is large. This is a primary region of interest in digital data transmission over the telephone network, and the applicable results give a basis for comparing performance of different systems. The mathematical problem is one of calculating the probability that a quadratic form in a set of gaussian variables with arbitrary means and variances will assume values critically far removed from that obtained when each variable is at its mean. The mean values represent signal contributions unperturbed by noise and for good performance should dominate over the noise except at the tails of the distribution. Concentration of attention on the infrequent large noise peaks calls for an approach inherently different from the conventional series expansions appropriate near the center of the distribution. The results are of importance not only in detection theory but also in general statistical analysis of rare events.

I. INTRODUCTION

A general class of data receivers have decision logic based on observing at the output the sign of a quadratic form

$$q = \sum_{i,j=1}^{n} Q_{ij} w_i w_j \equiv w^+ Q w. \tag{1}$$

Examples of such quadratic detectors and the reduction of their output to the form (1) are given in Sections IV and V, and a schematic representation in Fig. 1. Let it suffice for the present to say that the real symmetric matrix Q is determined by the system filters, while the real N-dimensional vector w and its transpose w^+ are related to the received signal plus noise.

During a considerable portion of this paper we shall be concerned with the asymptotic evaluation of the probability of error for such receivers when the noise is additive gaussian. To develop the tools

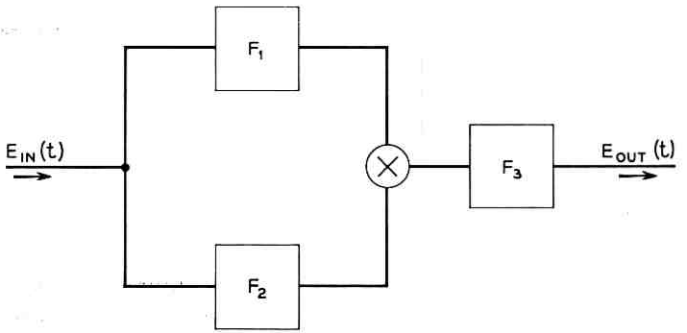


Fig. 1 — Model of the quadratic detector.

for this end, and also partly for general mathematical interest, Sections II and III are devoted to the following general problem: Given that the real gaussian vector w whose components have means $\bar{w}_i$ and a positive definite* covariance matrix M,

$$M_{ij} = \langle (w_i - \bar{w}_i)\ (w_j - \bar{w}_j) \rangle = M_{ji}\,, \tag{2}$$

is such that the quadratic form $\bar{w}^+ Q \bar{w}$ is positive, what is the probability P_e that for small noise (i.e., that the w_i have small variances) the quadratic form (1) is negative? We refer to P_e as the asymptotic probability of error.† Various special cases of this problem have been treated in the literature; some references may be found in Section IV where we reproduce and generalize some of these special results.

* Without any real loss of generality strict positive definiteness is assumed.

† As pointed out later our techniques may also be used to obtain the distribution function in certain regions, not only the probability of error.

A systematic treatment of the general problem has hitherto been lacking. Progress has been inhibited due to the fact that the exact probability distribution of arbitrary gaussian quadratic forms is too complicated to be useful. Experience has shown that the curves for the probability of error vs signal-to-noise ratio tend to be roughly parallel for different data receivers and to be characterized sufficiently well by their asymptotic slopes. We, therefore, do not strive to obtain exact error rates but rather deal with asymptotic forms for large signal-to-noise ratios. In other words, we will be concerned with the behavior of the distribution on the tails of the density functions of our quadratic forms. For high error rates or low signal-to-noise ratios, other approximations may be obtained by using various well-known moment series such as the Edgeworth and Gram-Charlier expansions.

We show in Section V how our results may be used to attack the problem of the distribution of the filtered response of a product detector. It is primarily this consideration of postdetection filtering that has been absent from earlier discussions. Finally in Section VI, a simple model of a fairly representative class of quadratic detectors is analyzed in some detail by means of a rapidly convergent expansion in prolate spheroidal wave functions.

II. GENERAL ANALYSIS FOR QUADRATIC FORM

In the introduction we defined the problem of obtaining the probability of error for the quadratic form q. We approach this problem via the characteristic function $C(\omega)$ of (1), which is well known to be given by[1]

$$C(\omega) \equiv \langle e^{i\omega q} \rangle_q = \frac{\exp \tfrac{1}{2}[\bar{w}^+ M^{-1}(I - 2i\omega MQ)^{-1}\bar{w} - \bar{w}^+ M^{-1}\bar{w}]}{\sqrt{\det\,(I - 2i\omega MQ}}. \quad (3)$$

The symbol $\langle \cdot \rangle_q$ denotes the average with respect to q; "det" means determinant and I is the identity matrix. Since M^{-1} is the inverse of a real symmetric and positive definite matrix, it itself has all of these properties. We now note the following theorem:[2] Let A and B be real symmetric matrices, and further let A be positive definite. Then there exists a real matrix S such that

$$S^+ AS = I \quad (4)$$

$$S^+ BS = \sigma^2 D, \quad (5)$$

where D is some diagonal matrix and σ^2 is a positive parameter introduced for later convenience.

Equation (4) implies, in particular, that S^{-1} exists. If we identify A with M^{-1}, and B with Q, (4) and (5) tell us that we may write

$$M = SS^{+} \tag{4a}$$

$$Q = (S^{+})^{-1}\sigma^{2}DS^{-1}. \tag{5a}$$

If we substitute (4a) and (5a) into (3), and further introduce a new real gaussian vector y by the linear transformation

$$y = S^{-1}w, \tag{6}$$

we arrive at a simpler expression for the characteristic function, namely,

$$C(\omega) = \frac{\exp\left[\frac{1}{2}\bar{y}^{+}(I - 2i\omega\sigma^{2}D)^{-1}\bar{y} - \frac{1}{2}\bar{y}^{+}\bar{y}\right]}{\sqrt{\det\,(I - 2i\omega\sigma^{2}D)}}. \tag{7}$$

The principal simplification is now that D is a diagonal matrix. We note that

$$w^{+}Qw = \sigma^{2}y^{+}Dy \tag{8}$$

and

$$\bar{w}^{+}Q\bar{w} = \sigma^{2}\bar{y}^{+}D\bar{y}. \tag{9}$$

We find it more convenient to deal with q when it is expressed in terms of the variables y_i.

The probability of error defined earlier may be expressed in terms of the characteristic function by the formula

$$\Pr\{q < 0\} \equiv P_e = -\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{C(\omega)}{\omega + i\epsilon}\,d\omega. \tag{10}$$

The $i\epsilon$ ($\epsilon > 0$) appearing in the denominator of the integrand and in (10) is used to signify that in the complex ω-plane, the contour of integration implied in (10) goes above the singularity at $\omega = 0$. Making use of (7), we write (10) in more detail*

$$P_e = -\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{d\omega}{\omega + i\epsilon}\,\frac{1}{\sqrt{\prod_{j=1}^{N}(1 - 2i\omega\sigma^{2}d_j)}} \cdot \exp\left[\sum_{j=1}^{N}\frac{i\omega d_j y_j^{2}}{1 - 2i\omega\sigma^{2}d_j}\right] \tag{11}$$

* Henceforth, we will not use bars to denote the mean of a random variable.

$$= -\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{d\omega}{\omega + i\epsilon}\frac{1}{\sqrt{\prod_{j=1}^{N}(1 - i\omega d_j)}} \cdot \exp\left[-\frac{\omega}{2\sigma^2}\sum_{j=1}^{N}\frac{y_j^2}{\omega + \frac{i}{d_j}}\right], \tag{12}$$

where d_j is the jth element of the real diagonal matrix D. Equation (12) follows from (11) by a simple change of variable. We observe in (12) that the quantity $1/\sigma^2$ appears only in the exponent, and therefore we may obtain an asymptotic expansion of P_e valid for small σ^2 by considering only the exponent. One will see in later sections that this asymptotic result for small σ^2 corresponds to the asymptotic result for large signal-to-noise ratios.

Proceeding with the analysis of (12), we remark that the integrand obviously falls off sufficiently rapidly at infinity to allow one to close or distort the contour at infinity without changing the value of the integral. We further note the singularities of the integrand of (12). There is a simple pole at $\omega = 0$ which has already been discussed. In addition to the simple pole at $\omega = 0$, the exponent has simple poles at $\omega = -i/d_i$; these all lie along the imaginary axis. Also at these points the denominator of the integrand has, in general, branch points due to the square root. For doubly degenerate eigenvalues, the branch points become simple poles.

We now concentrate our attention on performing the integration in (12) for small σ^2 by the saddle point method.[3,4] To locate the saddle points, let $\omega = i\eta$, and consider the solutions to

$$-\frac{d}{d\eta}\left[\eta\sum_{i=1}^{N}\frac{y_i^2}{\eta + \frac{1}{d_i}}\right] = 0, \tag{13}$$

or

$$F(\eta) \equiv -\sum_{i=1}^{N}\frac{\frac{y_i^2}{d_i}}{\left(\eta + \frac{1}{d_i}\right)^2} = 0. \tag{14}$$

Consider $F(\eta)$ defined by (14) and let η be real. We shall show that there always exists a saddle point for positive η (ω positive imaginary),

which occurs *before* the first singularity of the exponent on the positive imaginary axis.* We separate the sum (14) into two parts

$$F(\nu) = -\sum_{d_i \text{pos}} \frac{\frac{y_i^2}{d_i}}{\left(\nu + \frac{1}{d_i}\right)^2} - \sum_{d_i \text{neg}} \frac{\frac{y_i^2}{d_i}}{\left(\nu + \frac{1}{d_i}\right)^2}. \tag{15}$$

We note that when $\nu \approx -(1/d_i)$, $F(\nu)$ is large and positive for $d_i < 0$ and $F(\nu)$ is large and negative for $d_i > 0$. Also note that $F(0) = -\sum d_i y_i^2 \leqq 0$ by (9) and our assumption that the noise-free signal is positive. Hence by continuity, there must be at least one $\nu > 0$ for which $F(\nu) = 0$, which locates a saddle point for us. Clearly by the monotonicity of the two parts of expression (15) there can be only one such saddle point between the origin and the first singularity of the exponent on the positive imaginary axis.

The relations in the complex ω-plane described above are illustrated in Fig. 2. The dotted line labeled L_1 is the original contour of integration, and the small circle labeled "*S.P.*" is the (unique) saddle point on the positive imaginary axis lying before the nearest singularity $\omega = -i/d_i$, $d_i < 0$, which has nonvanishing residue y_i^2 in the exponent of the integrand. Also drawn on the imaginary axis in Fig. 2 are the branch cuts of the denominator, indicated by the hatched bars in the figure. We assume in this section (and Fig. 2 is so drawn) that the saddle point occurs before the nearest singularity $\omega = -i/d'$ on the positive imaginary axis.† As discussed above, this is guaranteed to be the case if only $(y')^2 = 0$.‡

If the saddle point is situated as shown in Fig. 2, the contour may then be shifted from the real axis (line L_1 in Fig. 2) to a contour which is a straight line parallel to the real axis and passing through the saddle point (the line L_2 in Fig. 2). The integrand drops off sufficiently rapidly for large $|\omega|$ so that the ends of the contour connecting L_1 and L_2 (in accordance with Cauchy's theorem) give no contribution. We will now, for large signal-to-noise ratios (small σ^2), approximate the integral along the contour L_2 by the contribution in the immediate neighborhood of the saddle point. It is shown in the appendix that the magnitude

* This assumes that at least one of the d_i is negative with nonvanishing residue y_i^2. We also, of course, are assuming that the quadratic form q in the absence of noise is positive.

† For definiteness we have based our analysis on the assumption that the noiseless signal is positive. We emphasize that the role of upper and lower half planes would be interchanged if one assumed that the noiseless signal were negative.

‡ We have used d' to denote that particular d_i which corresponds to the nearest singularity on the positive imaginary axis; y' is the associated y_i.

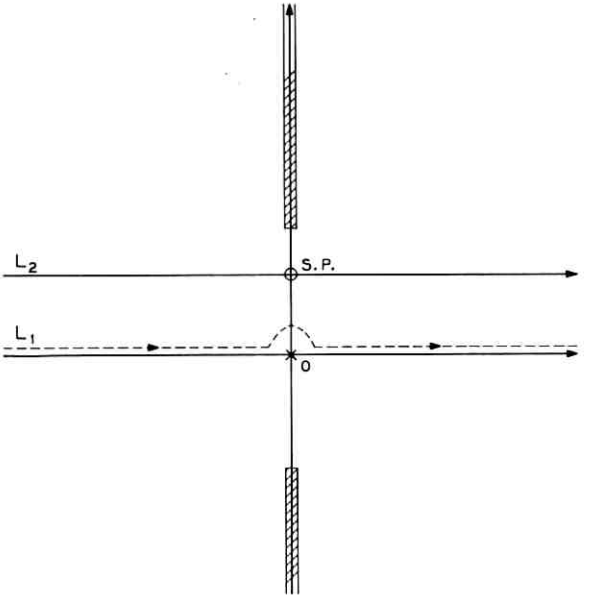


Fig. 2 — Contours and singularities in the ω-plane.

of the exponential term in the integrand of (12) is a monotonically decreasing function of ω as one recedes from the saddle point on the contour L_2, and therefore this saddle point evaluation is asymptotically correct.[3,4] One might also add that it can be shown that the contour L_2 is in fact tangent to the path of steepest descent at the saddle point.

To obtain an explicit formula for our asymptotic evaluation of P_e under these conditions, write (12) as

$$P_e = -\frac{1}{2\pi i}\int_{L_2} g(z)\,\exp -\left[\frac{f(z)}{\sigma^2}\right] dz, \tag{16}$$

with

$$g(z) = \frac{1}{z}\,\frac{1}{\sqrt{\prod(1 - izd_i)}} \tag{17}$$

and

$$f(z) = \frac{z}{2}\sum_{j=1}^{N}\frac{y_j^2}{z + \dfrac{i}{d_j}}. \tag{18}$$

Then, by a standard saddle point evaluation, we have

$$P_e \sim \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\sqrt{\prod(1+\Gamma d_j)}} \frac{\sigma}{\sqrt{f''(i\Gamma)}} \exp\left[-\frac{\Gamma}{2\sigma^2}\sum \frac{y_j^2}{\Gamma + \frac{1}{d_j}}\right], \quad (19)$$

where, to repeat, Γ is the smallest positive root of the equation

$$\frac{d}{dz} f(z) \mid_{z=i\Gamma} = 0. \quad (20)$$

The notation $f''(i\Gamma)$ in (19) means the second derivative of $f(z)$ evaluated at $z = i\Gamma$.

We finally wish to note that although we have concentrated in this section on the evaluation of P_e, one could use entirely analogous techniques to find an asymptotic expression for the probability P_K that the quadratic form q is less than some number K as long as K is less than the value of q in the absence of noise. The analog of (10) is (for any value of K)

$$P_K = -\frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{C(\omega)}{\omega + i\epsilon} e^{-i\omega K} d\omega. \quad (21)$$

For a K satisfying the stated condition, the integrand in (21) possesses a saddle point such that a discussion for asymptotically calculating P_K may be given which is entirely analogous to that already given for P_e. For a given σ^2, the accuracy of such an approach will depend on K.

III. VANISHING OF CRITICAL RESIDUE IN UPPER HALF-PLANE

We wish, in this section, to review a particularly trivial example to illustrate a violation of a condition necessary for the guaranteed applicability of our method, namely the nonvanishing of the residue associated with the nearest pole in the upper half-plane. In the example we consider here, the exact answer can be obtained by a simple contour integral. Consider the quadratic form

$$z_1 = x_1^2 - x_2^2 + x_3^2 - x_4^2, \quad (22)$$

where the x's are independent gaussian variables, all with the same variance σ^2. Further, assume that x_2 and x_4 have zero means. Then the characteristic function of z is

$$C(\omega) = \frac{\exp i\omega \frac{x_1^2 + x_3^2}{1 - 2i\omega\sigma^2}}{(1 - 2i\omega\sigma^2)(1 + 2i\omega\sigma^2)}. \quad (23)$$

The contour integral implied in (10) may be closed in the upper half-plane to yield immediately by exact methods

$$P_e = \tfrac{1}{2}\exp\left[-\frac{x_1^2 + x_3^2}{4\sigma^2}\right]. \tag{24}$$

The exponent in (24) is simply obtained by evaluating the exponent in (23) at the singularity $\omega = i/2\sigma^2$ in the upper half-plane. Now consider instead the expression

$$z_2 = z_1 + \sum_{n=2}^{N} x_{2n+1}^2 - \sum_{n=2}^{N} x_{2n}^2, \tag{25}$$

where z_2 in the absence of noise is assumed positive, the additional gaussian variables are assumed to have variances which are less than σ^2, and the variable x_{2n} $(n \geqq 3)$ with smallest variance has nonvanishing mean. The exponent $[-f(iy)/\sigma^2]$ for the characteristic function of z_2 as one travels up the positive imaginary axis appears as in Fig. 3, where we have written $\omega = iy$. Two situations are clearly possible. First $y_0 < 1/2\sigma^2$; in this case the contour of integration for (10) may be distorted to pass through the saddle point, and the previous discussion then applies. Since $x_2^2 + x_4^2 = 0$ this could not be guaranteed *a priori*. The second case is $y_0 > 1/2\sigma^2$; in order to distort the contour to pass through the saddle point in this situation, one must first sweep the contour past the singularity at $\omega = i/2\sigma^2$. Since this singularity is, for the present case, only a simple pole, the result of pushing the contour past is simply to pick up the residue of this pole. As an asymptotic answer one then has this residue plus the saddle point contribution. However, it is clear from Fig. 3 that the residue term (recall $y_0 > 1/2\sigma^2$) has a less negative

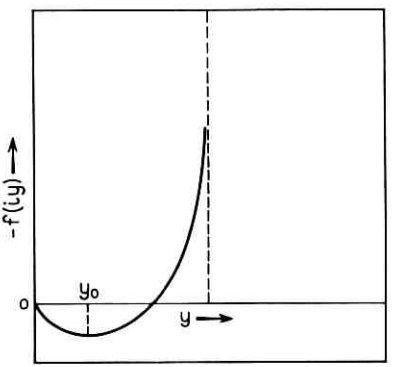


Fig. 3 — Behavior of exponent discussed in section III.

exponent, and thus in the limit of large signal-to-noise ratio will be exponentially dominant over the saddle point. If, for this second case, one has instead one or more branch points before the saddle point, the analysis is not as simple, but realizing that one may distort up to the first singularity (which, of course, has vanishing residue in the exponent), and keeping in mind Fig. 3, one might argue that the exponential behavior for this case is determined by the value of the exponent evaluated at the nearest singularity in the upper half-plane.

IV. SIMPLE COMMUNICATION APPLICATION

In this section we apply our saddle point technique to a number of problems whose asymptotic forms have appeared previously in the literature with derivations based on techniques different from ours. The reproduction of previous results helps to establish confidence in our methods, and arrives at these answers in a more straightforward manner than previously.

The problems considered here may all be put into the form where

$$q = u_1^2 + u_2^2 - v_1^2 - v_2^2 \equiv u^2 - v^2. \tag{26}$$

That is, q is the difference of the squared lengths of two two-dimensional gaussian vectors. We let the variance of both components of the vector $\mathbf{u}$ be equal to σ_2^2 and of both those of $\mathbf{v}$ equal to σ_1^2. Further, all components are independent. We shall see later that analysis of binary or multilevel FM using discrimination detection and differential detection of FM[9] reduces to this case with, in general, $\sigma_1 \neq \sigma_2$, while the analysis for differential phase modulation is also of this form, with $\sigma_1 = \sigma_2$.

The characteristic function for q defined by (26) is simply

$$C(\omega) = \frac{\exp i\omega\left[\dfrac{u^2}{1 - 2i\omega\sigma_2^2} - \dfrac{v^2}{1 + 2i\omega\sigma_1^2}\right]}{(1 - 2i\omega\sigma_2^2)(1 + 2i\omega\sigma_1^2)}, \tag{27}$$

and the probability of error is

$$P_e = -\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{d\omega}{\omega + i\epsilon}\frac{\exp -\dfrac{\omega}{2\sigma_1^2}\left[\dfrac{v^2}{\omega - iA} + \dfrac{u^2}{A^2\left(\omega + \dfrac{i}{A}\right)}\right]}{(\omega - iA)\left(\omega + \dfrac{i}{A}\right)}, \tag{28}$$

where

$$A = \sigma_2/\sigma_1. \tag{29}$$

Letting, again, $\omega = iy$, we find that the saddle point of interest is at

$$y_0 = A \frac{u^2 + A^2v^2 - uv(1 + A)^2}{u^2 - A^4v^2}. \tag{30}$$

Using (19), we then obtain for the probability of error*

$$P_e \sim \frac{1}{\sqrt{2\pi}} \frac{\sigma_1}{\sqrt{1 + A^2}} \frac{1}{\sqrt{uv}} \frac{(u - A^2v)(u + A^2v)}{[u^2 + A^2v^2 - uv(1 + A^2)]} \cdot \exp\left[\frac{(u - v)^2}{2\sigma_1^2(1 + A^2)}\right]. \tag{31}$$

We might note in passing that if one has two independent N-dimensional gaussian vectors $\mathbf{u}_N$ and $\mathbf{v}_N$ and if all the components of $\mathbf{u}_N$ ($\mathbf{v}_N$) are independent and have the same variance $\sigma_2^2(\sigma_1^2)$, then the probability of error for the difference of the square of these vectors, $u_N^2 - v_N^2$, will have different multiplicative factors from these in (31) but the exponent in (31) will still be correct when u and v are reinterpreted to be the lengths of the vectors $\mathbf{u}_N$ and $\mathbf{v}_N$, respectively. The result (31) and its generalizations for higher dimensions may also be viewed as giving the asymptotic form of the cumulative probability distribution for the doubly noncentral F-distribution. Exact, but not very transparent, formulas for this problem have recently been published by Price.[5]

Analysis of the error performance of binary FM and differential phase modulation leads one to consider the probability that the inner product q of two independent 2-dimensional gaussian vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$,

$$q = 2\boldsymbol{\alpha} \cdot \boldsymbol{\beta} \tag{32}$$

is negative when the inner product of the means is positive.[6,7,8] Here again the components of each vector are independent, and those of $\boldsymbol{\alpha}$ have the same variance and those of $\boldsymbol{\beta}$ have possibly a different variance. However, it is clear that multiplication of (32) by a positive constant can adjust these two variances to be identical (with appropriate adjustment of the means) without affecting the probability of error. Hence, it suffices to choose the variances to be equal to the same constant σ^2. Further, introduce

$$\begin{aligned} \mathbf{u} &= (\boldsymbol{\alpha} + \boldsymbol{\beta})/\sqrt{2} \\ \mathbf{v} &= (\boldsymbol{\alpha} - \boldsymbol{\beta})/\sqrt{2}, \end{aligned} \tag{33}$$

so that

* If we write (31) as $P_e = f(u,v,\sigma_1,\sigma_2)$, then, if the noisefree form (26) is *negative*, the probability P_e' that q is positive is $P_e' = -f(u,v,\sigma_1,\sigma_2)$.

$$q = u^2 - v^2. \tag{34}$$

The conditions assumed in the derivation of (31) are satisfied by (34), and in particular we have

$$\sigma_1^2 = \sigma_2^2 = \sigma^2. \tag{35}$$

Equation (31) for the probability of error then simplifies to

$$P_e \sim \frac{\sigma}{2\sqrt{\pi}} \frac{1}{\sqrt{uv}} \frac{u+v}{u-v} \exp\left[-\frac{(u-v)^2}{4\sigma^2}\right]. \tag{36}$$

By making extensive use of (33), it is readily verified that our expression (36) for the asymptotic (large S/N) probability of error for q defined by (32) is identical with the expression given in Bennett and Davey,[6] (9–56). The considerably more complicated form of the Bennett-Davey result is solely due to the fact that they express their answer in terms of variables α and β instead of u and v.

We now consider another application of our formula (31), namely to the analysis of errors in multilevel FM data transmission using discrimination detection. In this application, one is concerned with the probability that the instantaneous frequency $\dot{\psi}$ is in error, where $\dot{\psi}$ is given in terms of in-phase and quadrature components, $x(t)$ and $y(t)$ respectively, by the equation[10,11]

$$\dot{\psi} = \frac{x\dot{y} - y\dot{x}}{x^2 + y^2}. \tag{37}$$

The quantities x, y, $\dot{x}$, $\dot{y}$ are gaussian variables with arbitrary means, and with variances equal to σ^2, σ^2, $\dot{\sigma}^2$, $\dot{\sigma}^2$ respectively. We also use the notation

$$\begin{aligned} R^2 &= x^2 + y^2 \\ R\dot{R} &= x\dot{x} + y\dot{y}. \end{aligned} \tag{38}$$

Suppose for the noise-free situation $\dot{\psi} > z$. It is of interest in the multilevel situation to know the probability P that $\dot{\psi} < z$ in the presence of gaussian noise. To put this problem into a form to which (31) is directly applicable, we made use of the following chain of equalities:

$$\begin{aligned} P &= \Pr[\dot{\psi} \leqq z] = \Pr[\dot{\psi} - z \leqq 0] \\ &= \Pr[x(\dot{y} - zx) + y(-\dot{x} - zy) \leqq 0] \\ &= \Pr[xa + yb \leqq 0], \end{aligned}$$

where

$$\begin{aligned} a &= k(\dot{y} - zx) \\ b &= -k(\dot{x} + zy), \end{aligned} \tag{39}$$

and k is any *positive* constant.

We define

$$\begin{aligned} u_1 &= x + a \\ u_2 &= y + b \\ v_1 &= x - a \\ v_2 &= y - b \end{aligned} \tag{40}$$

and further choose k so that

$$k^2 = \frac{1}{z^2 + \rho^2}, \tag{41}$$

where

$$\rho^2 = \frac{\dot{\sigma}^2}{\sigma^2}. \tag{42}$$

Then we see that

$$P = \Pr\,(u_1^2 + u_2^2 - v_1^2 - v_2^2 \leqq 0), \tag{43}$$

where the set of variables (u_1, u_2, v_1, v_2) are all independent (due to this choice of k) and further,

$$\sigma_{u_1}^2 = \sigma_{u_2}^2 = \sigma^2(1 - kz)^2 + k^2\dot{\sigma}^2 = \sigma_2^2 \tag{44a}$$

$$\sigma_{v_1}^2 = \sigma_{v_2}^2 = \sigma^2(1 + kz)^2 + k^2\dot{\sigma}^2 = \sigma_1^2. \tag{44b}$$

It is useful to write $u^2 = u_1^2 + u_2^2$ and $v^2 = v_1^2 + v_2^2$ in terms of the original FM variables. Using (37) – (40), we find that

$$u^2 = R^2\left\{\left[1 + kz\left(\frac{\dot{\psi}}{z} - 1\right)\right]^2 + k^2\left(\frac{\dot{R}}{R}\right)^2\right\} \tag{45a}$$

$$v^2 = R^2\left\{\left[1 - kz\left(\frac{\dot{\psi}}{z} - 1\right)\right]^2 + k^2\left(\frac{\dot{R}}{R}\right)^2\right\}. \tag{45b}$$

A convenient simplification results if we restrict ourselves to constant amplitude FM waves, i.e., $\dot{R} = 0$. For this specialization we have

$$u = R[1 + k(\dot{\psi} - z)] \tag{46a}$$

$$v = R\begin{cases} 1 - k(\dot{\psi} - z) & \text{if } k(\dot{\psi} - z) < 1 \\ k(\dot{\psi} - z) - 1 & \text{if } k(\dot{\psi} - z) > 1. \end{cases} \tag{46b}$$

If we set $z = 0$, we immediately have that $k = \sigma/\dot{\sigma}$, $\sigma_1^2 = \sigma_2^2 = 2\sigma^2$, and, distinguishing the two cases in (46b), (46) and (31) reproduce exactly (38) and (39) of Ref. (6).

Equations (46) and (31) together provide general formulas for the asymptotic evaluation of multilevel FM.

V. GENERAL QUADRATIC DETECTOR

In this section we consider the probability of error for the filtered response of the product detector given in Fig. 1 when the noise input is white gaussian. The signal $s(t)$ plus the added noise $n(t)$ is divided into two branches, each of which has a filter (denoted by F_1 and F_2 for the two branches). The outputs of these filters are multiplied and the product is passed through the filter F_3, whose output is the final system output. This product detector is a generalization of the square-law detector considered by Kac and Siegert,[12] and by Emerson.[13] We first parallel Emerson's treatment and express the problem as in expression (1), except now $N = \infty$.

One familiar with Refs. 12 and 13 will not be surprised that the results amount to solving an integral equation, one which there is little hope of solving in practice. Therefore, we feel that an important point is made when we show in the next section that for a particular model of considerable practical interest, one can effectively approximate the system function by a kernel of finite rank, expressible in terms of known functions; the functions we have in mind for this purpose are the prolate spheroidal wave functions.[1] We wish to emphasize that although the results of earlier sections are applied here, we regard the rapidly converging approximations to the system function as important in the treatment of this problem.

Following Emerson, let $f_i(t)$ be the impulse response of the ith filter. Then the output at time t, $E_{\text{out}}(t)$, may be written in terms of the input wave

$$E_{\text{in}}(t) = s(t) + n(t) \tag{47}$$

as

$$E_{\text{out}}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} du\, dv\, E_{\text{in}}(t-u)g(u,v)E_{\text{in}}(t-v), \tag{48}$$

where the system function $g(u,v)$ is given by

$$g(u,v) = \int_{-\infty}^{\infty} f_1(u-z)f_3(z)f_2(v-z)dz. \tag{49}$$

One can immediately see from (49) that if the filters F_1 and F_2 are not identical, then $g(u,v)$ is not a symmetric function of u and v. How-

ever, it is apparent from (48) that if we write the identity

$$g(u,v) = \tfrac{1}{2}[g(u,v) + g(v,u)] + \tfrac{1}{2}[g(u,v) - g(v,u)], \tag{50}$$

then when (50) is used in (48), the second term in (50) gives no contribution, and we have

$$E_{\text{out}}(t) = \iint du\, dv\, E_{\text{in}}(t-u)G(u,v)E_{\text{in}}(t-v), \tag{51}$$

where

$$G(u,v) = \tfrac{1}{2}[g(u,v) + g(v,u)]. \tag{52}$$

The kernel $G(u,v)$ is now hermitian, and all its eigenvalues are guaranteed to be real; we also assume that G is square integrable. If λ_n denotes its nth eigenvalue and $\varphi_n(t)$ its nth eigenfunction, then we may write the well-known operator result*

$$G(u,v) = \Sigma\lambda_n\varphi_n(u)\varphi_n(v). \tag{53}$$

Thus,

$$E_{\text{out}}(t) = \Sigma\lambda_n e_n^{\,2}(t), \tag{54}$$

where

$$e_n(t) = \int_{-\infty}^{\infty} \varphi_n(v)E_{\text{in}}(t-v)\, dv\,. \tag{55}$$

Upon using (47) and the fact that $n(t)$ represents white gaussian noise with correlation function

$$\langle n(t)n(t')\rangle = N_0\delta(t-t'), \tag{56}$$

and suppressing the t-dependence in (55), we see that the e_n are independent gaussian variables with variances given by

$$\text{var}\,\{e_n\} = N_0\,. \tag{57}$$

The probability P_e may now, in accordance with Section II, be written as

$$P_e = -\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{d\omega}{\omega + i\epsilon}\prod\frac{1}{\sqrt{1 - i\omega\lambda_n}} \cdot \exp\left[-\frac{1}{2N_0}\sum_n \frac{\omega e_n^{\,2}}{\omega + i/\lambda_n}\right], \tag{58}$$

* We assume without loss of generality that the $\varphi_n(t)$ are real.

or

$$P_e \sim \frac{\sqrt{2N_0}}{\sqrt{2\pi}} \frac{1}{y_0} \frac{1}{\sqrt{\prod(1 + y_0\lambda_n)}} \frac{\exp\left[-\frac{1}{2N_0} f(y_0)\right]}{\sqrt{-f''(y_0)}}. \tag{59}$$

In (59) we have written

$$(y) = \sum \frac{y e_n^2}{y + 1/\lambda_n} = \sum \frac{y\lambda_n e_n^2}{1 + y\lambda_n} \tag{60}$$

and y_0 is determined by

$$f'(y_0) = 0. \tag{61}$$

We will pay particular attention to the function $f(y)$, since it is this function which determines the exponential behavior for large S/N. We note that in terms of the system operator G (60), which defines the function $f(y)$, may be written

$$f(y) = \left(s, \frac{yG}{1 + yG} s\right) \tag{62}$$

where we use the usual notation (a,b) for the inner product of two vectors in Hilbert space. In (62) we have used s to denote the vector $s(t - v)$, t fixed. If one knew explicitly the resolvent operator $(1 + yG)^{-1}$ appearing in (62), one might perhaps calculate the required integrals in (62) and search for the maximum of this function of y, thus determining y_0 which is implicitly a functional of the signal s. In general, however, approximation methods must be used.

Before mentioning some approximation schemes, we would like to demonstrate an interesting result. The saddle point y_0 is determined by (61), or, using (62), we have the implicit relationship

$$y_0 = \frac{\left(s, \frac{G}{1 + y_0G} s\right)}{\left(s, \frac{G^2}{(1 + y_0G)^2} s\right)}. \tag{63}$$

Thus, from (62) and (63),

$$f(y_0) = y_0^2 \left(s, \frac{G^2}{(1 + y_0G)^2} s\right), \tag{64}$$

or

$$f(y_0) = \frac{(s,Ks)^2}{(s,K^2s)} , \tag{65}$$

where

$$K = \frac{G}{1 + y_0 G} . \tag{66}$$

Now clearly,

$$f(y_0) = \frac{(s,Ks)^2}{(s,K^2s)} \leqq \frac{(s,s)(Ks,Ks)}{(Ks,Ks)} = (s,s) \tag{67}$$

by hermiticity of K and Schwarz's inequality. Thus, if the signal energy $E = (s,s)$ is fixed, the best performance would be achieved if $(1/E)f(y_0) = 1$, provided that this is possible. Certainly the equality in (67) is satisfied if s is an eigenfunction of K, and hence of the system function G, and at first glance this would appear to be an optimizing solution. However, such a solution violates the necessary condition that the function $f(y)$ have poles at the nearest eigenvalue for both positive and negative y, and hence our basic relation (65) does not hold for such a choice of s. In fact, under the assumptions for which (65) was derived, no function s, subject to the constant energy constraint, will yield a stationary value for the exponent (65).* In fact, it is not difficult to convince oneself that the best function s to take (for a positive output) is the eigenfunction φ_0 with the largest positive eigenvalue. Note that (65) is not applicable here. However, the exponent may be estimated according to the discussion of Section III to be

$$f\left(y = \left|\frac{1}{\lambda_-}\right|\right) = e_1^2 \frac{\left|\frac{\lambda_+}{\lambda_-}\right|}{1 + \left|\frac{\lambda_+}{\lambda_-}\right|} \tag{68}$$

where λ_+ (λ_-) is the positive (negative) eigenvalue of G with largest magnitude. If λ_+ is large compared with $|\lambda_-|$, the equality in (67) may be approached (note $e_1^2 = (s,s)$) for positive pulses. However, for negative pulses the factor in the exponent will be

$$\frac{\left|\frac{\lambda_-}{\lambda_+}\right|}{1 + \left|\frac{\lambda_-}{\lambda_+}\right|} \tag{69}$$

* Even though the exponent is a function of y_0 which implicitly depends on s this dependence can be ignored in a first-order variation because of (61).

and will become arbitrarily bad. Thus, if one wants symmetry between positive and negative pulses, one must take λ_+ and λ_- to have essentially the same magnitude, leading to a result that is a factor of two worse than suggested by the equality in (67). We might point out that the value of the exponent that would be obtained if the equality in (67) held, is the same as that for the probability of error for an ideal correlator, the optimum detector for this binary situation. Since, as just described, the optimum binary scheme here uses orthogonal signals, the factor of two worse than ideal binary correlation detection is not surprising. Since the ideal binary system (again a correlator) for *orthogonal* signals also gives an exponent worse by a factor of two, no exponent could be better.

VI. EXAMPLE OF QUADRATIC DETECTOR

A particular specialization of the general quadratic detector given schematically in Fig. 1 is a differential detector. By this term we shall mean that the filters F_1 and F_2 in Fig. 1 are identical except that F_2, in addition to representing the channel as F_1 does, has a delay of one bit interval. The filter F_3 is a low-pass filter. We treat the simplified baseband case where F_1 is an ideal low-pass filter with cutoff Ω rad/s., and F_2 is identical to F_1 except for a delay T_1, and finally the postdetection filter F_3 is an ideal integrator with integration time T seconds. The analysis is also relevant to the situation where F_1 and F_2 are bandpass filters, symmetrical with respect to some carrier frequency, provided one neglects the double carrier-frequency terms at the input to F_3.

We begin by considering first the alternative version of the output (51), namely

$$E_{\text{out}}(t) = \tfrac{1}{2}[(s_d, \bar{g}s) + (s, \bar{g}s_d)], \tag{70}$$

where in (70) we have for convenience included the delay T_1 directly in the signal rather than in the system function, and have denoted the delayed version of s by s_d. In (70), $\bar{g}$ denotes the system function for two identical filters F_1 and F_2. Equation (49) now reads

$$\bar{g}(u,v) = \int_{-T/2}^{T/2} \frac{\sin \Omega\,(u - z)}{\pi(u - z)} \frac{\sin \Omega\,(v - z)}{\pi(v - z)}\, dz. \tag{71}$$

We would like to evaluate the integral for u and v on the entire real line. We shall first present a formal evaluation for u and v both restricted to the interval $(-T/2, T/2)$, and shall then invoke analytic continuation to claim that this restriction on the evaluation of (71) can be dropped. Let $\psi_n(t)$ be the prolate spheroidal function normalized to unity in the

infinite interval, and hence to λ_n on the interval $(-T/2, T/2)$. Also let $\psi_n'(t)$ be the same function normalized to unity on the interval $(-T/2, T/2)$ and hence to $1/\lambda_n$ on the infinite interval. The functions $\psi_n'(t)$ satisfy the equation[14]

$$\int_{-T/2}^{T/2} \frac{\sin \Omega (u - z)}{\pi(u - z)} \psi_n'(z)\, dz = \lambda_n \psi_n'(u), \tag{72}$$

where the eigenvalues λ_n are real and positive, and $\psi_n(z)$ satisfy the identical equation. On the Hilbert space of square-integrable functions on the interval $(-T/2, T/2)$, it then follows that we may write

$$\frac{\sin \Omega (u - z)}{\pi(u - z)} = \sum_{n=0}^{\infty} \lambda_n \psi_n'(u)\psi_n'(z). \tag{73}$$

Therefore,

$$\bar{g}(u,v) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \lambda_n \lambda_m \psi_n'(u)\psi_m'(v) \int_{-T/2}^{T/2} \psi_n'(z)\psi_m'(z)\, dz,$$

or

$$\bar{g}(u,v) = \sum_{n=0}^{\infty} \lambda_n^2 \psi_n'(u)\psi_n'(v). \tag{74}$$

Reinterpreting (74) to hold on the infinite interval, we have finally

$$\bar{g}(u,v) = \sum_{n=0}^{\infty} \lambda_n \psi_n(u)\psi_n(v). \tag{75}$$

Rewriting (70) as

$$E_{\text{out}} = \tfrac{1}{2}[(s, U^+(T_1)\bar{g}s) + (s, \bar{g}U(T_1)s)], \tag{76}$$

where $U(T_1)$ represents the unitary operator of time translation by amount T_1, and comparing (76) with (51), we see that the symmetrized system function $G(u,v)$ is given by

$$G(u,v) = \tfrac{1}{2} \sum_n \lambda_n \psi_n(u + T_1)\psi_n(v) + \tfrac{1}{2} \sum_n \lambda_n \psi_n(u)\psi_n(v + T_1). \tag{77}$$

Important simplifications in the result (77) obtain when one makes use of the fact that the λ_n tend to zero rapidly after a few terms. Thus, one would expect that the infinite sum for the system function (77) may be effectively truncated after a few terms, and the problem of finding the resolvent kernel $(1 + yG)^{-1}$ in (62) is thereby reduced to inverting a finite dimensional matrix. The rapidity with which the λ_n decrease

depends on the parameter[14] $c = \Omega T/2$. Some fairly typical situations correspond roughly to $c = \pi$ and $T_1 = T$. The first six eigenvalues for $c = \pi$, which are 0.981, 0.749, 0.243, 0.025, 0.001, $\sim 10^{-5}$, illustrate this behavior well. Furthermore, it usually happens that the shape of the signal during the integration time of the filter, i.e., the pulse shape, bears a great deal of resemblance to $\psi_0(t)$, which one may crudely visualize as having a $(\sin t)/t$ shape. This tends to emphasize the $n = 0$ term in (77) even more. Hence we should not be far wrong if we simply write (for $T_1 = T$)

$$G(u,v) = \tfrac{1}{2}\lambda_0\psi_0(u + T)\psi_0(v) + \tfrac{1}{2}\lambda_0\psi_0(u)\psi_0(v + T), \tag{78}$$

or equivalently if we define

$$\epsilon = \int_{-T/2}^{T/2} \psi_0(u)\psi_0(u + T)\, du, \tag{79}$$

we have that

$$\begin{aligned} G(u,v) = {} & \frac{\lambda_0(1 + \epsilon)}{2}\left[\frac{\psi_0(u) + \psi_0(u + T)}{\sqrt{2(1 + \epsilon)}}\right]\left[\frac{\psi_0(v) + \psi_0(v + T)}{\sqrt{2(1 + \epsilon)}}\right] \\ & - \frac{\lambda_0(1 - \epsilon)}{2}\left[\frac{\psi_0(u) - \psi_0(u + T)}{\sqrt{2(1 - \epsilon)}}\right]\left[\frac{\psi_0(v) - \psi_0(v + T)}{\sqrt{2(1 - \epsilon)}}\right]. \end{aligned} \tag{80}$$

The latter form (80) explicitly exhibits the eigenvalues and eigenfunctions to this approximation. For $T_1 = T$, ϵ is quite small and in the spirit of our approximation may be neglected. We note in closing that we have found the representations (78) through (80) extremely useful in evaluating the effects of an added external tone on differential phase detection of a signal accompanied by gaussian noise. We emphasize that the dependence of the degradation on the frequency of the tone in such a problem is strongly influenced by the presence of the postdetection filter and hence its inclusion (aside from its role of selecting only difference frequencies) was essential. This, in fact, motivated much of the present work.

APPENDIX

Proof of Monotonicity of Exponent on the Contour

The function of interest is

$$\exp\left[-f(\omega)/\sigma^2\right] \equiv \exp\left[-\frac{\omega}{2\sigma^2}\sum_{j=1}^{N}\frac{y_j^2}{\omega + \frac{i}{d_j}}\right]. \tag{81}$$

If we let $\omega = x + i\Gamma$ with x real and with the saddle point occurring at $\omega = i\Gamma$, we have on the contour L_2

$$|\exp\{-f(\omega)/\sigma^2\}| = \exp\left\{-\frac{1}{2\sigma^2}\sum_j \frac{y_j^2\left[x^2 + \Gamma\left(\Gamma + \frac{1}{d_j}\right)\right]}{x^2 + \left(\Gamma + \frac{1}{d_j}\right)^2}\right\}. \quad (82)$$

Therefore, it is sufficient for us to prove for those j such that $y_j^2 \neq 0$, that

$$\frac{x^2 + \Gamma\left(\Gamma + \frac{1}{d_j}\right)}{x^2 + \left(\Gamma + \frac{1}{d_j}\right)^2} > \frac{\Gamma\left(\Gamma + \frac{1}{d_j}\right)}{\left(\Gamma + \frac{1}{d_j}\right)^2}, \quad (83)$$

which in turn amounts to showing

$$\frac{\frac{1}{d_j}}{\Gamma + \frac{1}{d_j}} > 0. \quad (84)$$

Now, if we merely recall that $\Gamma > 0$, then the inequality is obviously true for $d_j > 0$. If we also recall that if $d_j < 0$ then $(-1)/d_j > \Gamma$, (84) obviously holds for $d_j < 0$.

REFERENCES

1. Turin, G., Error Probabilities for Binary Symmetric Ideal Reception through Nonselective Slow Fading and Noise, Proc. IRE, *46*, Sept., 1958, pp. 1603–1619.
2. Friedman, B., *Principles and Techniques of Applied Mathematics*, John Wiley & Sons, Inc., New York, 1956, p. 109.
3. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, pp. 744–749.
4. DeBruijn, M. G., *Asymptotic Methods in Analysis*, North-Holland Publishing Company, Amsterdam, 1958.
5. Price, R., Some Noncentral F-Distributions Expressed in Closed Form, Biometrika *51*, 1 and 2, 1964, pp. 107–122.
6. Bennett, W. R., and Davey, J. R., *Data Transmission*, McGraw-Hill Book Company, New York, 1965.
7. Bennett, W. R., and Salz, J., Binary Data Transmission by FM over a Real Channel, B.S.T.J., *42*, Sept., 1963, pp. 2387–2426.
8. Stein, S., Unified Analysis of Certain Coherent and Noncoherent Binary Communications Systems, IEEE Trans., *IT-10*, Jan., 1964, pp. 43–51.
9. Anderson, R. R., Bennett, W. R., Davey, J. R., and Salz, J., Differential Detection of Binary FM, B.S.T.J., *44*, Jan., 1965, pp. 111–159.
10. Salz, J., and Stein, S., Distribution of Instantaneous Frequency for Signal Plus Noise, IEEE Trans., *IT-10*, Oct., 1964, pp. 272–274.
11. Balakrishnan, A. V., and Abrams, I. J., Detection Levels and Error Rates in

PCM Telemetry Systems, 1960 IRE Int'l. Convention Record, pt. 5, pp. 32–55.
12. Kac, M., and Siegert, A. J. F., On the Theory of Noise in Radio Receivers with Square Law Detectors, J. Appl. Phys., *18*, 1947, pp. 383–397.
13. Emerson, R. C., First Probability Densities for Receivers with Square Law Detectors, J. Appl. Phys. *24*, 1953, pp. 1168–1180.
14. Slepian, D., and Pollak, H. O., Prolate Spheroidal Wave Functions — I, B.S.T.J., *40*, Jan., 1961, pp. 43–63.

Optimum Reception of M-ary Gaussian Signals in Gaussian Noise

By T. T. KADOTA

(Manuscript received May 28, 1965)

The problem of optium reception of M-ary Gaussian signals in Gaussian noise is to specify, in terms of the observable waveform, a scheme for deciding among M alternative mean and covariance functions with minimum error probability. Although much literature on the problem exists, a mathematically rigorous solution has yet to appear. By formulating the problem as optimum discrimination of M Gaussian measures in function space induced by the mean and covariance functions, this paper presents such a solution.

Let $m_k(t)$ *and* $r_k(s,t)$, $k = 1, \cdots, M$, *be the alternative mean and covariance functions of the Gaussian signal, and let* $m_0(t)$ *and* $r_0(s,t)$ *be the mean and covariance functions of the Gaussian noise. If, for each* $k = 1, \cdots, M$, *the integral equations,*

$$\int r_0(s,t)g_k(s)\,ds = m_k(t)$$

and

$$\iint r_0(s,u)h_k(u,v)[r_0(v,t) + r_k(v,t)]\,du\,dv = r_k(s,t),$$

admit a square-integrable solution $g_k(t)$ *and a symmetric, square-integrable solution* $h_k(s,t)$, *then the following decision scheme is optimum: given an observable waveform* $x(t)$,

choose $m_k(t)$ *and* $r_k(s,t)$ *if* $I_k(x)$ *is the largest among all* $I_j(x)$, $j = 1, \cdots, M$,

where I_k *is defined by*

$$I_k(x) = \frac{1}{2}\iint x(s)h_k(s,t)x(t)\,ds\,dt + \int x(t)f_k(t)\,dt + c_k$$

in which

$$f_k(t) = g_k(t) - \int h_k(s,t)[m_0(s) + m_k(s)]\, ds$$

and c_k is a constant determined by the mean and covariance functions $m_0(t)$, $m_k(t)$, $r_0(s,t)$ and $r_k(s,t)$ as well as the a priori probability associated with $m_k(t)$ and $r_k(s,t)$.

The first section introduces and defines the problem and the second presents the solution with pertinent discussions while a precise mathematical treatment is left to the appendix.

I. INTRODUCTION

Before we formulate the general problem of optimum reception of M-ary signals in noise, let us review a simplified version of the classical problem: optimum reception of M-ary *sure* signals in Gaussian noise. Suppose there are M sure signals $m_k(t)$, $k = 1, \cdots, M$, with *a priori* probabilities α_k, $0 < \alpha_k < 1$ and $\sum_{k=1}^{M} \alpha_k = 1$, for transmission. The received waveform $x(t)$ consists of one of these M signals and an additive Gaussian noise $n(t)$, i.e.

$$x(t) = m_k(t) + n(t).$$

In order to simplify the problem, we "represent" the signals and noise by certain finite sequences $m_{k1}, \cdots, m_{kn}$, $k = 1, \cdots, M$, and $n_1, \cdots, n_n$ respectively.* Then the representing sequence $x_1, \cdots, x_n$ of the received waveform is given by

$$x_i = m_{ki} + n_i, \qquad i = 1, \cdots, n. \tag{1}$$

It is assumed that the signal sequences are linearly independent vectors in an n-dimensional space R_n while the elements, $n_1, \cdots, n_n$ of the noise sequence are statistically independent, identically distributed Gaussian variables with mean zero and variance one. The task of the receiver is to observe the received sequence $x_1, \cdots, x_n$ and to decide which one of M signal sequences must have been transmitted. For each possible erroneous decision, there is an associated probability, and the average of all these probabilities weighted by the *a priori* probabilities is the so-called (average) error probability. Then, the problem of optimum reception in this simplified form is to specify in terms of the observable sequence $x_1, \cdots, x_n$ a scheme for choosing the value of the index k such that the error probability is minimum over all possible decision schemes.

* These sequences may be regarded as the sample values of the waveforms or the Fourier coefficients of certain orthonormal expansions of the waveforms.

Now, from the assumption concerning the noise sequence, the joint probability density function of $n_1, \cdots, n_n$ is

$$p_0(\nu_1, \cdots, \nu_n) = (2\pi)^{-(n/2)} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \nu_i^2\right). \tag{2}$$

Hence, through the use of the relation (1), the joint density function of $x_1, \cdots, x_n$ becomes

$$p_k(\nu_1, \cdots, \nu_n) = (2\pi)^{-(n/2)} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} (\nu_i - m_{ki})^2\right]. \tag{3}$$

Note that since the observable sequence is a point in R_n, a decision scheme in terms of it is equivalent to a division of R_n into M non-overlapping regions $\hat{\Lambda}_1, \cdots, \hat{\Lambda}_M$ so that $k = j$ is chosen if $(x_1, \cdots, x_n)$ belongs to $\hat{\Lambda}_j$. Then, the error probability associated with such a division $(\hat{\Lambda}_1, \cdots, \hat{\Lambda}_M)$ is given by

$$P_e = 1 - \sum_{k=1}^{M} \alpha_k \int_{\hat{\Lambda}_k} p_k(\nu_1, \cdots, \nu_n)\, d\nu_1 \cdots d\nu_n,$$

which can be rewritten as

$$P_e = 1 - \int_{R_n} \left[\sum_{k=1}^{M} \chi_{\hat{\Lambda}_k}(\nu_1, \cdots, \nu_n)\alpha_k p_k(\nu_1, \cdots, \nu_n)\right] d\nu_1 \cdots d\nu_n, \tag{4}$$

where $\chi_{\hat{\Lambda}_k}$ is the indicator function of the set $\hat{\Lambda}_k$, $k = 1, \cdots, M$. Clearly, P_e is minimum over all divisions if the integrand in the above is maximum at every point $(\nu_1, \cdots, \nu_n)$ in R_n. This can be achieved by arranging a division $(\hat{\Lambda}_1, \cdots, \hat{\Lambda}_M)$ so that, in the region $\hat{\Lambda}_k$, the $\alpha_k p_k$ with the same index is the largest among all $\alpha_1 p_1, \cdots, \alpha_M p_M$. Namely, the region with index k consists of a set of points at which $\alpha_k p_k$ is the largest. If more than one $\alpha_k p_k$ is the largest at the same point, that point is to be assigned to the region whose index is the smallest of all those k's so that all the regions remain disjoint. That is, a division $(\hat{S}_1, \cdots, \hat{S}_M)$ of R_n specified by

$$\left.\begin{aligned} \hat{S}_k = \{(\nu_1, \cdots, \nu_n): \alpha_k p_k(\nu_1, \cdots, \nu_n) &> \alpha_j p_j(\nu_1, \cdots, \nu_n), \\ &\qquad j < k, \\ \alpha_k p_k(\nu_1, \cdots, \nu_n) &\geqq \alpha_j p_j(\nu_1, \cdots, \nu_n), \\ &\qquad j > k \end{aligned}\right\}, \tag{5}$$

is the division which minimizes P_e. Or the optimum decision scheme is the following:

Given an observable sequence $(x_1, \cdots, x_n)$, choose the minimum value of k for which $\alpha_k p_k(x_1, \cdots, x_n)$ is maximum as a function of k.*

Now by expanding the exponent in (3),

$$\alpha_k p_k(x_1, \cdots, x_n) = \alpha_k \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i m_{ki} - \frac{1}{2}\sum_{i=1}^{n} m_{ki}^2\right).$$

Since the first sum in the bracket is common to all $\alpha_k p_k$, if we put

$$\hat{I}_k = \sum_{i=1}^{n} x_i m_{ki} - \frac{1}{2}\sum_{i=1}^{n} m_{ki}^2 + \log \alpha_k, \tag{6}$$

then the optimum decision scheme is equivalent to choosing the minimum value of k for which $\hat{I}_k$ is maximum.

In the special case of "equi-probable and equi-energetic" transmitted signal sequences, i.e.

$$\alpha_i = \cdots = \alpha_M \qquad \text{and} \qquad \sum_{i=1}^{n} m_{1i}^2 = \cdots = \sum_{i=1}^{n} m_{Mi}^2,$$

$\hat{I}_k$ can be effectively replaced by its first term, i.e. $\sum_{i=1}^{n} x_i m_{ki}$. In other words, the optimum decision scheme in this case consists in performing correlation of the observable sequence with M signal sequences and choosing the smallest of the k values corresponding to the largest correlation sums.†

In the general problem of optimum reception of M-ary Gaussian signals in Gaussian noise, the observable waveform (received waveform) $x(t)$ is expressed by

$$x(t) = y_k(t) + n(t)$$

where $y_k(t)$ is one of M possible Gaussian signals which are characterized by mean and covariance functions just as $n(t)$ is. We assume that each Gaussian signal is statistically independent of the noise and it cannot be detected "perfectly" in the presence of this noise.‡ Again, the task of the receiver is to observe the waveform $x(t)$ for a finite time, say $0 \leqq t \leqq 1$, and to decide which one of M Gaussian signals must have been received. Then, by defining the error probability as before, the problem of optimum reception becomes that of specifying, in terms of the

* That is, suppose for a given $(x_1, \cdots, x_n)$, $\alpha_k p_k(x_1, \cdots, x_n)$ as a function of k assumes its maximum at $k = k_1, \cdots, k_j$ where $k_1 < \cdots < k_j$. Then, we choose $k = k_1$.

† For classical references, see Refs. 1 and 2.

‡ This is the assumption of "non-singular detection". A necessary and sufficient condition of non-singular detection is given by (13) in Appendix.

observable waveform $x(t)$, a scheme for choosing the index k such that the error probability is minimum over all possible decision schemes.

One mathematical idealization of the above problem is the following: Let $\{y_t, 0 \leqq t \leqq 1\}$ be a Gaussian process whose mean and covariance functions are one of M possible pairs of $m_k(t)$, $0 \leqq t \leqq 1$, and $r_k(s,t)$, $0 \leqq s, t \leqq 1$, $k = 1, \cdots, M$, where $m_k(t)$ and $r_k(s,t)$ are assumed to be continuous.* Similarly, let $\{n_t, 0 \leqq t \leqq 1\}$ be a Gaussian process whose mean and covariance functions are $m_0(t)$, $0 \leqq t \leqq 1$, and $r_0(s,t)$, $0 \leqq s,t \leqq 1$, where $m_0(t)$ is assumed to be continuous while $r_0(s,t)$ is positive-definite as well as continuous. It is further assumed that $\{y_t, 0 \leqq t \leqq 1\}$ and $\{n_t, 0 \leqq t \leqq 1\}$ are mutually independent for every $k = 1, \cdots, M$. Now define a new process $\{x_t, 0 \leqq t \leqq 1\}$ by $x_t = y_t + n_t$. Then, from the mutual independence assumption, the mean and covariance functions of $\{x_t, 0 \leqq t \leqq 1\}$ are one of M possible pairs of $m_0(t) + m_k(t)$ and $r_0(s,t) + r_k(s,t)$, $k = 1, \cdots, M$. Let P_k, $k = 1, \cdots, M$, be the Gaussian (probability) measure corresponding to the pair $m_0(t) + m_k(t)$ and $r_0(s,t) + r_k(s,t)$, and let P_0 be the one corresponding to $m_0(t)$ and $r_0(s,t)$. It is assumed that $m_0(t)$, $m_k(t)$, $r_0(s,t)$ and $r_k(s,t)$ are such that the two measures P_0 and P_k are equivalent, i.e. $P_0 \equiv P_k$, $k = 1, \cdots, M$.† Denote by H_k and α_k, $k = 1, \cdots, M$, the hypothesis and *a priori* probability that $m_k(t)$ and $r_k(s,t)$ are the pertinent mean and covariance functions of $\{y_t, 0 \leqq t \leqq 1\}$. Let $x(t)$ be the sample function of $\{x_t, 0 \leqq t \leqq 1\}$. Then, specification of the decision scheme amounts to dividing the space Ω of all sample functions $x(t)$ into M disjoint sets, $\Lambda_1, \cdots, \Lambda_M$, so that, if $x(t) \in \Lambda_k$, then H_k is to be chosen. Moreover, the error probability associated with such a division (or a decision scheme) is given by

$$P_e = 1 - \sum_{k=1}^{M} a_k P_k(\Lambda_k). \tag{7}$$

Thus, the problem of optimum reception is to specify in terms of $x(t)$ such a division of Ω that its associated error probability is minimum over all possible divisions.

Unlike the previous simple case, the sample (the observable) in this general case is the sample *function* $x(t)$ of the Gaussian process $\{x_t, 0 \leqq t \leqq 1\}$ instead of the sample *sequence* $x_1, \cdots, x_n$ of the finite sequence of Gaussian variables. Thus, the sample space which is to be divided into M non-overlapping regions, is the function space Ω instead of the n-dimensional sequence space R_n. Hence, we no longer have at our

* Note: $r_k(s,t) = E_k\{(x_s - m_k(s))(x_t - m_k(t))\}$, $k = 1, \cdots, M$.

† This corresponds to the assumption of non-singular detection.

disposal the joint density functions $p_k(\nu_1, \cdots, \nu_n)$, $k = 1, \cdots, M$, through which the optimum decision scheme is constructed. Nevertheless, there exists a certain generalization to this basic approach. Note from (2) that $p_0(\nu_1, \cdots, \nu_n) > 0, -\infty < \nu_i < \infty, i = 1, \cdots, n$. Hence, (4) can be rewritten as

$$P_e = 1 - \int_{R_n} \left[\sum_{k=1}^{M} \chi_{\hat{\Lambda}_k}(\nu_1, \cdots, \nu_n) \alpha_k \frac{p_k(\nu_1, \cdots, \nu_n)}{p_0(\nu_1, \cdots, \nu_n)} \right] p_0(\nu_1, \cdots, \nu_n) \times d\nu_1 \cdots d\nu_n .$$

Then, the optimum division of R_n is specified in terms of $\alpha_k[p_k(x_1, \cdots, x_n)/p_0(x_1, \cdots, x_n)]$ instead of $\alpha_k p_k(x_1, \cdots, x_n)$, though the two are obviously equivalent. Now, in the general case where the sample space is Ω instead of R_n, the likelihood ratio $p_k(x_1, \cdots, x_n)/p_0(x_1, \cdots, x_n)$ is replaced by its generalized version dP_k/dP_0, the Radon-Nikodym derivative (of P_k with respect to P_0), which is a function of $x(t)$, and $p_0(\nu_1, \cdots, \nu_n)d\nu_1 \cdots d\nu_n$ is replaced by dP_0. Thus, the error probability in the general case can be expressed as

$$P_e = 1 - \int_{\Omega} \left[\sum_{k=1}^{M} \chi_{\Lambda_k}(x) \alpha_k \frac{dP_k}{dP_0}(x) \right] dP_0(x),$$

where $(\Lambda_1, \cdots, \Lambda_M)$ is a nonoverlapping division of Ω. Then an optimum division of Ω, which is analogous to (5), can be specified by $\alpha_k(dP_k/dP_0)$, $k = 1, \cdots, M$, and dP_k/dP_0 can in turn be expressed in terms of certain functionals of $x(t)$.

II. SUMMARY OF MAIN RESULTS AND DISCUSSIONS

The foundation for solution of the general problem stated in the preceding section consists of the two following facts:[3,4]

(*i*) If two Gaussian measures P_0 and P_k are equivalent for each $k = 1, \cdots, M$, then there exists random variables dP_k/dP_0 so that the optimum division $(S_1, \cdots, S_M)$ of the sample space Ω can be specified by

$$S_k = \left\{ x(t) : \alpha_k \frac{dP_k}{dP_0}(x) > \alpha_j \frac{dP_j}{dP_0}(x), j < k; \quad \alpha_k \frac{dP_k}{dP_0}(x) \geqq \alpha_j \frac{dP_j}{dP_0}(x), j > k \right\}. \tag{8}$$

(*ii*) If the integral equations

$$\int_0^1 r_0(s,t) g_k(s)\, ds = m_k(t), \qquad 0 \leqq t \leqq 1, \tag{9}$$

and

$$\int_0^1 \int_0^1 r_0(s,u)h_k(u,v)[r_0(v,t) + r_k(v,t)]\,du\,dv = r_k(s,t), \qquad 0 \leqq s,t \leqq 1, \tag{10}$$

have a square-integrable solution $g_k(t)$ and a symmetric, square-integrable solution $h_k(s,t)$ respectively,* then

$$\frac{dP_k}{dP_0}(x) = \beta_k^{\frac{1}{2}} \exp\left[\frac{1}{2}\int_0^1\int_0^1 [x(s) - m_0(s) - m_k(s)]h_k(s,t) \cdot [x(t) - m_0(t) - m_k(t)]\,ds\,dt + \int_0^1 [x(t) - m_0(t) - \tfrac{1}{2}m_k(t)]g_k(t)\,dt\right], \tag{11}$$

for almost all sample functions under all hypotheses H_k, $k = 1, \cdots, M$, where $\beta_k^{-1} = \prod_{i=1}^{\infty} \lambda_i^{(k)}$ and $\lambda_i^{(k)} > 0$, $i = 1,2, \cdots$, are the eigenvalues of the operators $R_0^{-\frac{1}{2}}(R_0 + R_k)R_0^{-\frac{1}{2}}$ and R_0 and R_k are integral operators whose kernels are $r_0(s,t)$ and $r_k(s,t)$ respectively.

Then, upon combination of (i) and (ii), the optimum decision scheme can be specified as follows:

Choose the minimum value of k for which I_k is maximum where

$$I_k = \frac{1}{2}\int_0^1\int_0^1 x(s)h_k(s,t)x(t)\,ds\,dt + \int_0^1 x(t)f_k(t)\,dt + \frac{1}{2}\int_0^1\int_0^1 [m_0(s) + m_k(s)]\,h_k(s,t)\,[m_0(t) + m_k(t)]\,ds\,dt - \int_0^1 [m_0(t) + \tfrac{1}{2}\,m_k(t)]\,g_k(t)\,dt + \log \alpha_k \beta_k^{\frac{1}{2}}, \tag{12}$$

in which $f_k(t)$, $k = 1, \cdots, M$, are defined by

$$f_k(t) = g_k(t) - \int_0^1 h_k(s,t)[m_0(s) + m_k(s)]\,ds, \qquad 0 \leqq t \leqq 1.$$

Needless to say, the condition for the above decision scheme to be optimum is the existence of such $g_k(t)$ and $h_k(s,t)$, $k = 1, \cdots, M$, as described in (ii). It should be remarked that the existence of $g_k(t)$

* If such solutions exist, they are necessarily unique. It should be remarked that square-integrability of $h(s,t)$ is in the sense of

$$\int_0^1\int_0^1 h^2(s,t)ds\,dt < \infty.$$

and $h_k(s,t)$ implies that P_0 and P_k are equivalent for each k. That is, if these $g_k(t)$ and $h_k(s,t)$ exist, then none of the Gaussian signals $y_k(t)$ can be detected perfectly in the presence of noise $n(t)$.

Physical interpretation of the above optimum decision scheme is straightforward, at least in principle. Suppose given $m_0(t)$, $m_k(t)$, $r_0(s,t)$ and $r_k(s,t)$, $k = 1, \cdots, M$, are such that the integral equations (9) and (10) admit a square-integrable solution $g_k(t)$ and a symmetric, square-integrable solution $h_k(s,t)$, then the optimum decision scheme consists in performing the single and the double integrals involving the received waveform $x(t)$ as specified by (12), and adding to these integrals the predetermined constants, the last three terms of (12), and finally choosing the minimum value of k for which the sum of these integrals and the constants is maximum.

It is instructive to consider the following two special cases:

Case 1 $\quad m_0(t) \equiv 0, \qquad r_k(s,t) \equiv 0, \qquad k = 1, \cdots, M.$

This is the case of "M-ary sure signals in noise", a simplified version of which has already been discussed in the introduction. Here the integral equation (10) always has a symmetric, square-integrable solution for each $k = 1, \cdots, M$, namely, the trivial solution:

$$h_k(s,t) \equiv 0.$$

Furthermore, $\lambda_i^{(k)} = 1$, $i = 1,2, \cdots$; $k = 1, \cdots, M$. Thus,

$$\beta_k = 1, \qquad k = 1, \cdots, M.$$

Hence, I_k of (12) is reduced to

$$I_k' = \int_0^1 x(t)g_k(t) - \frac{1}{2}\int_0^1 m_k(t)g_k(t)\,dt + \log \alpha_k,$$

provided the square-integrable solutions $g_k(t)$, $k = 1, \cdots, M$, exist for the integral equations (9).* Note that I_k' is the function-space counterpart to $\hat{I}_k$ of (6) in the sequence-space case. With additional conditions that

$$\alpha_1 = \cdots = \alpha_M \quad \text{and} \quad \int_0^1 m_1(t)g_1(t)\,dt = \cdots = \int_0^1 m_M(t)g_M(t)\,dt,\dagger$$

the optimum decision scheme is reduced to choosing the minimum value

* The form of I_k' agrees with the result formally obtained in Ref. 5.

† For example, choose $m_k(t) = \sqrt{\sigma_k}\,\psi_k(t)$, $k = 1, \cdots, M$, where $\sigma_1 \geqq \sigma_2 \geqq \cdots$, and $\psi_1(t)$, $\psi_2(t)$, $\cdots$ are the eigenvalues and the orthonormalized eigenfunctions of R_0.

of k for which the "correlation integral"

$$\int_0^1 x(t)g_k(t)\,dt$$

is maximum.

Case 2 $\quad m_0(t) \equiv 0, \quad m_k(t) \equiv 0, \quad k = 1, \cdots, M.$

This is the case commonly termed as "M-ary Gaussian signals in noise". Here, the integral equation (9) always has a square-integrable solution, namely, the trivial solution:

$$g_k(t) \equiv 0, \quad k = 1, \cdots, M.$$

Hence, I_k is reduced to

$$I_k'' = \frac{1}{2}\int_0^1\int_0^1 x(s)h_k(s,t)x(t)\,ds\,dt + \log \alpha_k\beta_k^{\frac{1}{2}}.$$

Thus, the optimum decision scheme consists in choosing the minimum value of k for which I_k'' is maximum, provided that the symmetric, square-integrable solutions $h_k(s,t)$, $k = 1, \cdots, M$, of (10) exist.

(Remark)

It is interesting to note that formal substitution of $r_0(s,t) = \delta(s - t)$ into (10) yields the result which is consistent with those obtained previously by Price.[6]

APPENDIX

Mathematical Supplement

In the preceding, mathematical precision has been somewhat compromised for intuitive appeal. It is the purpose of this appendix to clarify the content of the preceding sections by supplying a brief mathematical summary with pertinent remarks.

Let Ω be the space of all real-valued functions on [0,1] and let $\tilde{P}_0$ and $\tilde{P}_k$, $k = 1, \cdots, M$, be the Gaussian measures induced by m_0 and r_0 and by $m_0 + m_k$ and $r_0 + r_k$ respectively on a σ-field generated by the class of all intervals in Ω, where m_0 and m_k, $k = 1, \cdots, M$, are real-valued, continuous functions on [0,1], while r_0 and r_k, $k = 1, \cdots, M$, are real-valued, symmetric, positive-definite, continuous functions on [0,1] $\times$ [0,1].* Then, without loss of generality, there exists a real,

* r_k, $k = 1, \cdots, M$, can be only nonnegative-definite.

separable (with respect to $\tilde{P}_0$) and measurable process $\{x_t, 0 \leqq t \leqq 1\}$.* Let $\mathcal{B}$ be the minimal σ-field with respect to which every x_t, $t \in [0,1]$, is measurable, and P_0 and P_k, $k = 1, \cdots, M$, be the restrictions on $\mathcal{B}$ of $\tilde{P}_0$ and $\tilde{P}_k$ respectively. We assume that $P_0 \equiv P_k$, $k = 1, \cdots, M$, which immediately implies that $P_i \equiv P_j$; $i,j = 1, \cdots, M$, and $\{x_t, 0 \leqq t \leqq 1\}$ is separable with respect to P_k, $k = 1, \cdots, M$, also. For a necessary and sufficient condition for $P_0 \equiv P_k$, we cite the following:†

$$\lim_{n\to\infty} \operatorname{tr} [R_0^{(n)}(R_0^{(n)} + R_k^{(n)})^{-1} - 2I + (R_0^{(n)} + R_k^{(n)})(R_0^{(n)})^{-1}] < \infty,$$

$$\lim_{n\to\infty} \operatorname{tr} (R_0^{(n)})^{-1} M_k^{(n)} < \infty, \tag{13}$$

where $R_0^{(n)}$, $R_k^{(n)}$ and $M_k^{(n)}$, $k = 1, \cdots, M$, are $n \times n$-matrices defined by

$$\begin{aligned} (R_0^{(n)})_{ij} &= r_0(t_i, t_j), \\ (R_k^{(n)})_{ij} &= r_k(t_i, t_j), \qquad i,j = 1, \cdots, n, \\ (M_k^{(n)})_{ij} &= m_k(t_i) m_k(t_j), \end{aligned}$$

and t_i, $i = 1, \cdots, n$, are a finite subset of any sequence dense in $[0,1]$. Now the two fundamental facts for solution are:

(*i*) Let dP_k/dP_0, $k = 1, \cdots, M$, be the Radon-Nikodym derivatives of P_k with respect to P_0, and let $(S_1, \cdots, S_M)$ be the partition of Ω defined by

$$S_k = \left\{\omega\colon \alpha_k \frac{dP_k}{dP_0}(\omega) > \alpha_j \frac{dP_j}{dP_0}(\omega), j < k\right\} \cap \left\{\omega\colon \alpha_k \frac{dP_k}{dP_0}(\omega) \geqq \alpha_j \frac{dP_j}{dP_0}(\omega), j > k\right\}.$$

Then, for any partition $(\Lambda_1, \cdots, \Lambda_M)$ of Ω,

$$\sum_{k=1}^{M} \alpha_k P_k(S_k) \geqq \sum_{k=1}^{M} \alpha_k P_k(\Lambda_k).$$

(*ii*) For each $k = 1, \cdots, M$, if the integral equations (9) and (10) admit a square-integrable solution g_k and a solution h_k satisfying

$$h_k(s,t) = h_k(t,s), \qquad \int_0^1 \int_0^1 h_k^2(s,t)\, ds\, dt < \infty,$$

* For a detailed justification, see Ref. 4.
† See Ref. 4.

then

(a) such solutions g_k and h_k are unique,

(b) $P_0 \equiv P_k$, $k = 1, \cdots, M$,

(c) there exist eigenvalues $\lambda_i^{(k)} > 0$, $i = 1, 2, \cdots$, of

$$R_0^{-\frac{1}{2}}(R_0 + R_k)R_0^{-\frac{1}{2}}$$

such that $\prod_{i=1}^{\infty} \lambda_i^{(k)}$ converges to β_k^{-1}, $0 < \beta_k < \infty$, and

$$\frac{dP_k}{dP_0} = \beta_k^{\frac{1}{2}} \exp\left[\frac{1}{2}\int_0^1\int_0^1 [x_s - m_0(s) - m_k(s)]h_k(s,t)[x_t - m_0(t) - m_k(t)]\,ds\,dt + \int_0^1 [x_t - m_0(t) - \tfrac{1}{2}m_k(t)]g_k(t)\,dt\right], \quad \text{a.e. } (P_0).$$

Hence, specification of S_k, $k = 1, \cdots, M$, is obtained by combining (i) and (ii), and the hypothesis in (ii) is the condition that such a partition $(S_1, \cdots, S_M)$ exists and minimizes (7).

REFERENCES

1. Woodward, P. M., *Probability and Information Theory with Applications to Radar*, Pergamon Press, London, 1953.
2. Kotel'nikov, V. A., *The Theory of Optimum Noise Immunity*, McGraw-Hill Book Co., New York, 1959.
3. Kadota, T. T., Generalized Maximum Likelihood Test and Minimum Error Probability, (to appear in IEEE Trans. on Information Theory, Jan., 1966).
4. Kadota, T. T., Optimum Reception of Binary Sure and Gaussian Signals, B.S.T.J., *44*, Oct., 1965, pp. 1621–1658.
5. Thomas, J. B., Wolf, J. K., On the Statistical Detection Problem for Multiple Signals, IRE Trans., *IT-8*, pp. 274–280.
6. Price, R., Optimum Detection of Stochastic Signals in Noise with Applications to Scatter-Multipath Communications, IRE Trans., *IT-2*, pp. 125–135.

Contributors to This Issue

Richard R. Anderson, B.S.M.E., 1949, Northwestern University; M.S.E.E., 1960 Stevens Institute of Technology; Bell Telephone Laboratories, 1949—. Mr. Anderson first engaged in research on electronic switching systems for telephone central offices. In 1956 he joined the data transmission exploratory development department and made several prototype magnetic-tape transports for storing digital data. He has conducted theoretical studies of data transmission systems by computer simulation. Member, AAAS, Sigma Xi, Tau Beta Pi.

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory, and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. He is the author of General Stochastic Process in the Theory of Queues (Addison-Wesley, 1963), and of Mathematical Theory of Connecting Networks and Telephone Traffic (Academic Press, 1955). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mind Association, Phi Beta Kappa.

Dwight W. Berreman, B.S., 1950, University of Oregon; M.S., Ph.D., 1955, California Institute of Technology; Stanford Research Institute 1955–56; Physics Faculty, University of Oregon, 1956–61; Bell Telephone Laboratories, 1961—. Mr. Berreman has worked in X-ray, visible and infrared optics. As a member of the Chemical Physics Research Department, he has recently been engaged in infrared spectroscopy of crystal films and in the invention and study of gas lenses. Member, Phi Beta Kappa, Sigma Xi, American Physical Society, Optical Society of America.

James O. Edson, B.S.E.E., 1929, University of Kansas; Bell Telephone Laboratories, 1929—. Mr. Edson has worked on D-2 Carrier and K-1 and K-2 Carrier Repeater development. He also developed the terminal amplifier for type J carrier. His wartime activities included short assignments in radar and underwater sound and was primarily concerned with pulse modulation for communication systems. Mr. Edson

spent several years in transmission research working on laminated transmission lines. He has engaged in development of a transistorized vocoder system, pitch detectors, variolossers, and other equipment for use in the Data-Phone service. He is now concerned with development of high-speed solid-state coders. Senior Member, IEEE.

HANSJUERGEN H. HENNING, B.E.E., 1955, Polytechnic Institute of Brooklyn; M.E.E., 1961, New York University; Bell Telephone Laboratories, 1955—. Mr. Henning has been primarily engaged in circuit design for PCM systems, including the T1 exchange carrier system and a 224 Mb/s experimental PCM system. He is presently responsible for a group concerned with the development of a new PCM carrier terminal. Member, Sigma Xi.

T. T. KADOTA, B. S., 1953, Yokahama National University (Japan); M.S., 1956, Ph.D., 1960, University of California (Berkeley); Bell Telephone Laboratories, 1960—. Mr. Kadota has been engaged in the study of noise theory with application to optimum detection theory. Member, Sigma Xi, SIAM, IMS.

L. U. KIBLER, B.S., 1950, U. S. Coast Guard Academy; M.S.E.E., 1956, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1956—. At present, Mr. Kibler is working toward the Ph.D. degree in Electrophysics at Polytechnic Institute of Brooklyn. His work has included studies of microwave logic and diode use, parametric and tunnel diode microwave amplifiers, high-speed optical detector diodes and light emitting diodes, and most recently high-power, high-frequency transistor amplifiers. Member, IEEE, Eta Kappa Nu.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Telephone Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Telephone Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He is presently working on the transmission aspect of a light communications system. Member, IEEE.

JOHN S. MAYO, B.S., 1952, M.S., 1953, Ph.D., 1955, North Carolina State University; Bell Telephone Laboratories, 1955—. Mr. Mayo was

first engaged in computer research, including studies relating to the use of digital computers for radar and military weapons control systems. He was subsequently involved in the development of repeaters for an exchange carrier PCM system, and high-speed PCM terminals for an experimental high-speed PCM system. In 1960 he assumed his present responsibilities as Head, Pulse Code Modulation Terminal Department, where he is in charge of the coding and processing of broadband signals for transmission by pulse code modulation. Member, IEEE, Sigma Xi, Phi Kappa Phi.

James E. Mazo, B.S., 1958, Massachusetts Institute of Technology; M.S., 1960, and Ph.D., 1963, Syracuse University; Research Associate, University of Indiana, 1963–64; Bell Telephone Laboratories, 1964—. At Indiana University, Mr. Mazo was engaged in work on quantum scattering theory. At present, he is engaged in theoretical analysis of data systems. Member, American Physical Society, IEEE, Sigma Xi.

Stewart E. Miller, B.S. and M.S., 1941, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1941—. Mr. Miller first worked on coaxial carrier repeaters and later shifted to microwave radar systems development. At the close of World War II, he returned to coaxial carrier repeater development until 1949, when he joined the radio research department. There his work has been in the field of circular electric waveguide communication, microwave ferrite devices, and other components for microwave radio systems. As Director, Guided Wave Research Laboratory, he heads a group engaged in research on communication techniques for the millimeter wave and optical regions. Fellow, IEEE.

J. Salz, B.S.E.E., 1955, M.S.E., 1956, Ph.D., 1961, University of Florida; the Martin Company, 1958–60; Bell Telephone Laboratories, 1961—. Mr. Salz first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems. Member, IEEE; associate member, Sigma Xi.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the

signal-theoretic properties of nonlinear systems. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

FRANCIS J. WITT, B.S.E.E., 1953, M.S.E.E., 1955, Johns Hopkins University; Bell Telephone Laboratories, 1954–55, 1957—. Mr. Witt has engaged in active and sampled-data network exploratory research and in solid-state circuit development for short-haul carrier systems. Later he was in charge of a group responsible for the development of some of the solid-state circuits in the TELSTAR experimental communications satellite. He was concerned with the development of digital processing circuitry for a high-speed digital transmission system. At present, as Head, Transmission Analysis Department, he is responsible for analysis and computing support for the development of coaxial cable and radio relay transmission systems.