THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV JANUARY 1965 NUMBER 1

A Technique for Investigating On-Off Patterns of Speech

By PAUL T. BRADY

(Manuscript received July 6, 1964)

A study is made of certain properties of speech which are concerned with determining the presence of speech on a telephone circuit. A speech detector is constructed to yield an output of spurts and gaps, corresponding to the presence or absence of energy above a threshold. A computer program then attempts to correct this pattern for spurious noise operation and for gaps due to stop consonants, eventually yielding a pattern of talkspurts and pauses. Data reported here include the distributions of the spurts and gaps resulting from the detector as well as the distributions of talkspurts and pauses from the computer program. Studied here are the influence on these distributions of detector threshold variations as well as of parameter variations within the computer program. The gaps occurring within talkspurts retain their distribution over a range of thresholds, but the spurts do not. It appears that 200 msec forms a boundary between intersyllabic gaps and listener-detected pauses.

The detection technique developed here is considered to be an improvement over conventional methods, but still yields data whose significance is uncertain. It may be that a simple automatic speech detecting technique using fixed parameters is inadequate for some purposes.

I. INTRODUCTION

The object of this study is to investigate certain properties of speech pertinent to the problem of establishing the pattern of its presence and

absence on a telephone circuit. Recent developments in telephony, such as the introduction of long circuits with appreciable delay and the increasing use of voice-operated devices, have prompted learning more about speech patterns, especially as they occur in conversation.[1,2,3]

Although the task of detecting the presence of speech may seem at first to be an almost trivial problem, it is in fact very difficult. Speech has a large dynamic range, and its level frequently falls into the noise, even during segments audible to a listener. In addition, momentary interruptions due in part to stop consonants (/p/, /t/, /k/, etc.) might cause a speech detector to indicate a silent interval whereas a listener would indicate a continuing flow of speech.

Most existing designs of speech detectors employ a slow release, or hangover, to bridge such gaps, but an error equal to the hangover time is made every time the person actually stops talking. The method of detection which was investigated in this study will hopefully avoid some of the pitfalls of conventional detectors.

II. THE DETECTION TECHNIQUE

The detection technique used here is a two-step process. Speech is first played through a speech detector, whose output is then processed by a computer program. These steps will be discussed separately.

2.1 *The Speech Detector — Spurts and Gaps*

A block diagram of the speech detector used in this study is shown in Fig. 1. The incoming signal is first amplified and then full-wave rectified. A threshold detector is set at this point to detect the presence of a voltage above some fixed value. The threshold detector triggers a flip-flop which is cleared 200 times per second by a clock. If the flip-flop is triggered in between clock pulses, a pulse will appear on the output when the flip-flop is cleared. That is, an output pulse indicates that at some time during the last 5 msec the speech energy crossed the threshold. A pulse is therefore an indication of an "on interval," and the absence of a pulse indicates an "off interval." The pulse train from the detector serves as the data for computer analysis.

The threshold width is the difference, in db, between the 1000-cps signal level just required to cause pulses to appear sporadically at the output, and the signal level required to maintain a constant train of pulses. It is about 1 db in this detector. The frequency response of the detector is flat over the voice range.

The detector is thus able to resolve speech into 5-msec segments, this

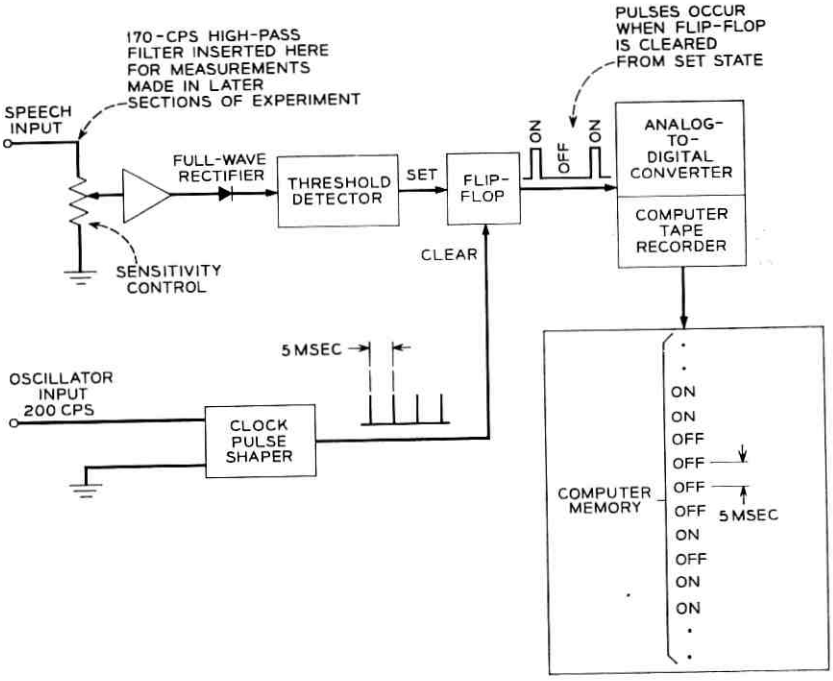


Fig. 1 — Apparatus to convert speech to on-off states.

interval being considered long enough to avoid audio-frequency fluctuations in the energy pattern. This is not exactly the same as a linear *RC* smoothing operation, but there is a rough correspondence. An *RC* network simply smoothes out fluctuations in the energy pattern, while our sampling circuit divides the energy pattern into 5-msec segments and for each segment produces a yes-no output, depending on the maximum amplitude in the segment. The *RC* circuit is much simpler to instrument, but the sampler is more suitable for subsequent computer analysis.

An *on interval* is defined as a 5-msec interval during which the speech energy exceeds the threshold at some time. In an *off interval*, the energy remains below the threshold during the entire interval.

A *spurt* is defined as an unbroken sequence of on intervals. A *gap* is an unbroken sequence of off intervals. By these definitions, therefore, speech can be transformed into a *spurt-gap pattern*. If a 1 represents an on interval and a 0 represents an off interval, the spurt-gap pattern could look like $\cdots$ 00111100011 $\cdots$.

2.2 *Talkspurts and Pauses*

We have already indicated that even while a person is talking his speech can still contain many gaps due to stop consonants and slight hesitations. To obtain a correspondence to the presence or absence of speech, we define a talkspurt and a pause.

A *talkspurt* is a time period which is judged by a listener to contain a sequence of speech sounds unbroken by a pause.

A *pause* is a time period which is judged by a listener to be a period of nontalking, other than one caused by a stop consonant, a slight hesitation, or a short breath.

2.3 *The Computer Program*

Spurts and gaps, as they come from the speech detector, are by definition physically measurable events, while talkspurts and pauses are determined subjectively. It is the function of the computer program to attempt to transform the spurt-gap pattern into a talkspurt-pause pattern. The program which was used actually performs two distinct functions:

(1) An attempt is made to obtain a talkspurt-pause pattern from the speech detector output. The manner in which this is done is described below. The original speech data are thus converted into "corrected data."

(2) The cumulative distribution functions of the durations of talkspurts and pauses are tabulated and plotted. Also computed are the per cent time speech is present and its converse, the per cent time speech is not present, as well as other data such as the mean and median talkspurt and pause lengths.

The procedure used by the computer program to obtain the corrected data is best described by an example. In Fig. 2, pattern (a) is a typical spurt-gap pattern produced by the speech detector. Each spurt and gap has, of course, a duration which is an integral multiple of 5 msec. The first step in data processing is to throw out all spurts which are less than or equal to a *throwaway time*. This is done because noise occasionally operates the speech detector for short periods, and the resulting spurts should be discarded. The throwaway operation produces pattern (b).

At this point, gaps less than or equal to a *fill-in-time* are filled in and considered as speech. This is an attempt to correct for the gaps due to stop consonants and other brief interruptions. It is hoped that pattern (c), obtained after fill-in, will correspond to talkspurts and pauses rather

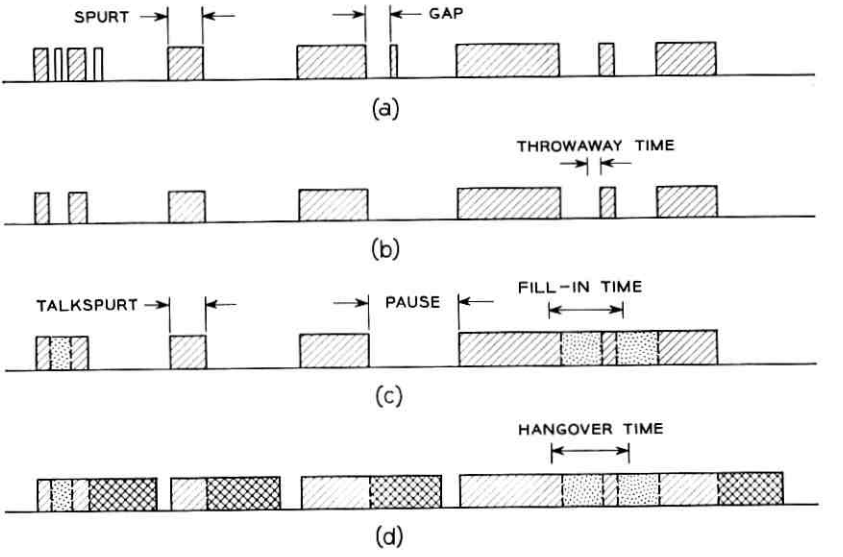


Fig. 2 — Process used by the computer to obtain talkspurts and pauses: (a) data from speech detector, (b) speech pattern after throwaway, (c) final pattern after fill-in, and (d) speech pattern if hangover were used instead of fill-in.

than to spurts and gaps. For convenience, this pattern will be labeled "talkspurts and pauses" even though such use of these terms is imprecise.

Notice that fill-in is not the same as hangover. With hangover, the beginning of *every* gap is filled in, regardless of the ultimate length of the gap. Pattern (d) would result if hangover were applied.

If the fill-in operation had preceded, rather than followed the throwaway operation, the resulting pattern would not have corresponded with pattern (c) of Fig. 2. The order of operation is, therefore, important. Now, the fill-in time, used to bridge stop consonants, should intuitively be much longer than the throwaway time, used to reduce noise effects. If the speech were filled in first, errors of fill-in time magnitude could occur as noise pulses were bridged to the adjacent speech. To avoid this problem, the throwaway operation is performed first.

2.4 *Discussion of Technique*

The process described above of using a speech detector with a fixed threshold in conjunction with a computer program with fixed throwaway and fill-in times was adopted here simply because it appeared to be a

reasonable technique to use. No formal study was made in which several schemes were compared, as such an investigation did not seem warranted. The suitability of this technique for obtaining a talkspurt-pause pattern will be discussed later, after some of the resulting data have been examined.

III. PROCEDURE TO OBTAIN PARAMETER VALUES

3.1 *Outline of Method*

Values had to be determined for the speech detector threshold, throwaway time, and fill-in time. It was decided to make up a recording of "continuous speech," that is, speech which contained no noticeable pauses. This speech would be played many times into the detector, with the threshold set lower for each time. It was hoped that eventually the spurt and gap distributions would stabilize so that continued lowering of the threshold would contribute little to the detection of speech. This would establish a threshold.

To determine a throwaway time, the original (unedited) tapes would be examined for spurious noise. This noise would hopefully be of some maximum short duration and could be discarded with a throwaway time.

The fill-in time would finally be established by again processing the continuous speech, this time using the fixed threshold and throwaway time, and since the computer should ideally view the continuous speech as one long talkspurt, the fill-in time would be chosen equal to the longest observed gaps.

3.2 *Source of Speech — the Telephone Conversations*

Eight pairs of subjects were asked to hold telephone conversations over a special circuit. The circuit, illustrated in Fig. 3, was a four-wire circuit which had losses which simulated the effect of a long distance connection. Delay could be switched into one of the paths.*

The conversations were recorded on a two-channel tape recorder connected, as shown, to a level point representative of the zero TL point.† The two members of each pair of subjects were good friends

* The delay was included for use in a separate study. Some speech recorded on the delay circuit is analyzed here because, by doing so, twice as much continuous speech becomes available than if only the "standard" circuit were used. Also, note that although in this case a delay of 400 msec is inserted from B to A, the subjects cannot distinguish this condition from a 200-msec delay in each path, provided that they have no common time reference.

† The zero transmission level point is a point to which all level points in a toll system can be referred. It is analogous to citing altitude by referring to height above sea level.

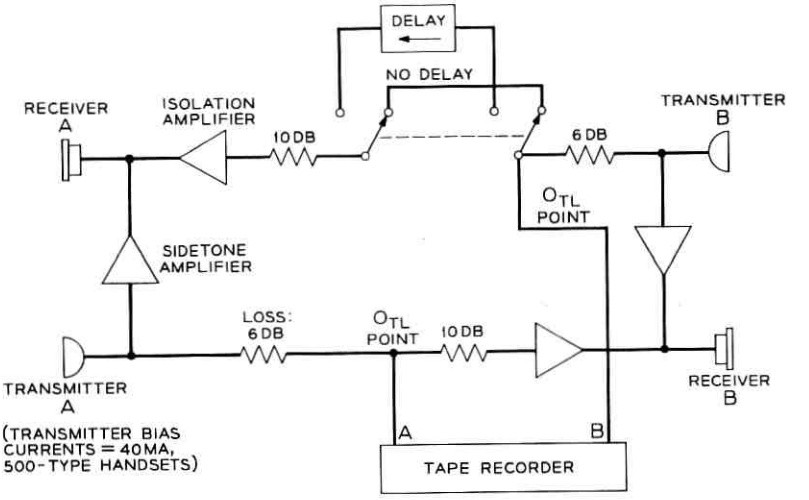


Fig. 3 — Circuit over which subjects talked.

who had many mutual interests. They received instructions to engage in active conversation about any topic they desired. They were told that their conversations would be recorded for use in "speech analysis work," but no other details were furnished. In the opinion of the experimenter, the subjects conversed readily and easily. There is no apparent evidence that their knowledge they were being recorded had a constraining effect on their conversations, but the reader should be aware that the experimental conditions differed in certain respects from those encountered in real life. The speech of the subjects was, however, certainly of a conversational nature and did not result from any formal preparation, as would occur with readings from printed matter.

The eight pairs of subjects consisted of four pairs of women and four of men. Each pair was allowed about a minute of warm-up time before the recorder was started. Then they talked for seven minutes over a standard circuit, followed by seven minutes over the 400-msec delay circuit. (One pair of girls spoke for only 3.3 minutes on the standard circuit and did not talk on the delay circuit.)

3.3 *Preparing Continuous Speech Tapes*

Each recording of each person's speech was edited by the author so that only segments of continuous speech, containing no pauses, were allowed to remain. This procedure was, of course, arbitrary and the

edited tape could not possibly be exactly duplicated, even by the experimenter. There were, however, certain rules which the experimenter tried to follow:

(1) In the edited tape there were no obvious long conversational pauses.

(2) Pauses for breathing were excluded, unless the breaths were very short and embedded in a train of speech. This event occurred only a few times in all the samples.

(3) Brief comments ("of course," "really?") and other short sounds ("uh-huh," "um") were excluded, as those were too difficult to handle in the splicing operation.

(4) Similarly, stuttering and other meaningless speech was excluded.

(5) Very low-level speech, which was difficult to hear, was excluded. This was an infrequent occurrence in speech which was part of an attempt to converse. Most very low-level speech consisted of sighs and other such remarks and could be ruled out by (3) and (4), above.

On the average, seven minutes of conversation of one person was reduced to about 55 seconds of continuous speech. The average length of a speech segment uninterrupted by tape splices was 2.13 seconds. In all, there were 27.4 minutes of continuous speech, made up of 773 tape segments.

To show that the edited tape did indeed consist only of continuous speech, a panel of six people was invited to listen to all of the samples. The listeners were provided with tally sheets, and were requested to make a checkmark whenever they felt that the speaker paused, or that there was any break in his conversation. The panel members were asked to be severe — if they detected any hesitancy in the speaker's voice, or any gap which they felt could be used as an opportunity to interrupt, they were told to count this as a pause.

On the average each listener indicated 14.7 pauses in the 27.4 minutes of recorded speech. Computer analysis of this speech (discussed later) showed that thousands of gaps did occur in the continuous speech. Thus, a negligible number of the gaps in the edited tape were judged by the listeners to be conversational pauses.

Although the continuous speech tapes contain virtually no pauses, the tapes do not by any means contain *all* of the continuous speech which occurred in the conversation. The data of this study are therefore not representative of *all* the spurts and gaps that were present in the original speech, but rather of a large number of them. Gaps which occurred in the neighborhood of pauses were usually excluded, since the editing process tended to select speech away from the beginnings and endings of talkspurts.

IV. RESULTS USED TO SET PARAMETERS

4.1 *Threshold*

The speech detector threshold should ideally be chosen low enough to pick up almost all of the speech signal and high enough to avoid noise operation. To see how low a threshold was required for the first criterion, the effect of threshold variation on the continuous speech gap and spurt distributions was studied. If a point were reached where the data remain substantially unchanged as the threshold is lowered, then that threshold would be considered sufficiently low to cause operation on most of the speech.

Volume measurements were made of the 16 samples of continuous speech taken from the conversations made on the no delay circuit. A Daven volume level indicator, Model 1866, was used to obtain VU readings, a commonly accepted indication of speech volume.[4,5] The readings for the samples ranged from −29.9 VU (weakest talker) to −16.5 VU (loudest talker). The loudest, softest, and median speakers were selected from both the male and female talkers, thus providing six speech samples for analysis. Each of these samples was played through the speech detector four times, with the threshold set at −44, −40, −36, and −32 dbm,* for each time respectively. The lowest (most sensitive) threshold was chosen as −44 dbm, since greater sensitivity would have aggravated noise problems resulting from low-level tape hiss.

The four samples from each speaker were analyzed to see the effect of threshold variation on gap and spurt distributions. A fill-in time of 10 msec was provided to eliminate the introduction of gaps due to tape splices.† The throwaway time was zero.

The set of gap distributions of subject AD, volume = −24.1 VU, is typical of the subjects, and is shown in Fig. 4. The abscissa is the length of the gap and the ordinate is the per cent of gaps which are less than the abscissa. For example, when the threshold was set at −44 dbm, 50 per cent of the gaps were less than 28 msec long. There are no gaps equal to or less than 10 msec, because those that existed were filled in.

The curves are very much alike, except possibly the −32-dbm curve. At first glance, it appears that for reasonably low thresholds, the threshold value is not critical for measuring speech. A fill-in time of about 130 msec would bridge all the gaps for any of the chosen threshold values.

---

* That is, db re 1 milliwatt into 600 ohms, so that zero dbm equals 0.775 volts rms.

† A typical splice in tape traveling at 15 ips causes the level of a tone to drop about 6 db for about 8 msec. A fill-in time of 10 msec bridges any gap caused by this momentary level drop.

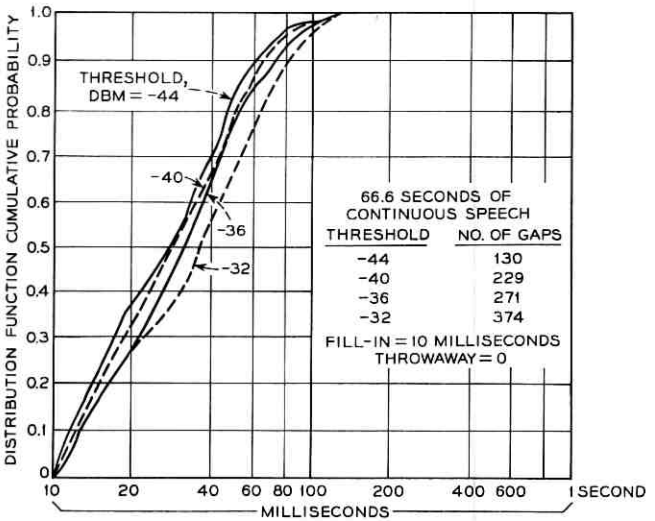


Fig. 4 — Gap distribution for subject AD.

But the gap distribution tells only part of the story. The total number of gaps almost triples as the threshold is raised from −44 to −32 dbm, going from 130 to 374 gaps in a continuous speech sample 66.6 seconds long.

The spurt distributions are plotted in Fig. 5. These change appreciably as the threshold is varied, showing shorter spurts with the higher thresholds. Also, whereas the low threshold indicates energy present 94 per cent of the time, the higher threshold yields only 75 per cent, as found from other results of the computer analysis.

Examination of the gap distributions of the six talkers whose speech was analyzed at four different thresholds indicates that gap distributions remain fairly stable for any threshold which is at least 12 db below the VU level.* As the threshold increases above this level, the gaps generally become noticeably longer. The 12-db value is only an estimate, which was arrived at by visual inspection of the data. It is a conservative estimate; a somewhat higher threshold would probably suffice for most of the speakers. However, until more data can be obtained for better analysis, 12 db will be used as a rule of thumb.

Although the gap distribution may stabilize as the threshold is lowered, the spurt distribution does not, and neither does the per cent time

* A similar analysis of the speech samples of all speakers also shows very little effect on gap distributions as the threshold is varied, as long as the threshold remains fairly low.

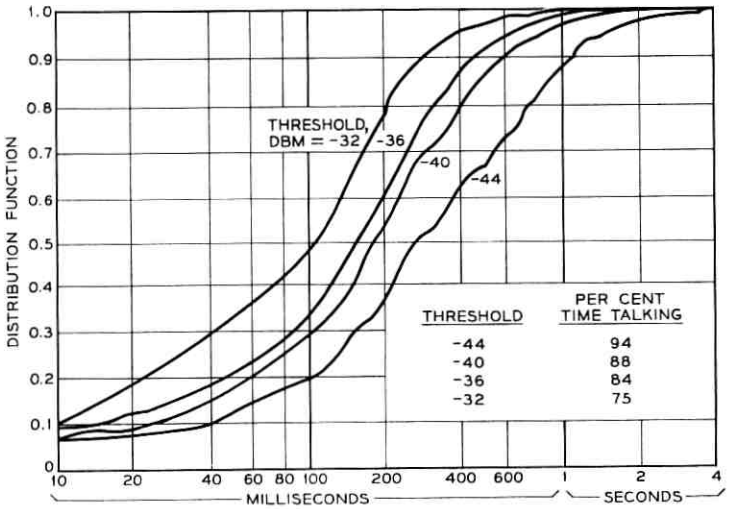


Fig. 5 — Spurt distribution for subject AD.

talking. It appears, then, that the threshold should be set as low as possible in order to pick up most of the speech.

The threshold cannot be set so low as to cause excessive noise operation. To estimate a lower bound, the original unedited tapes were played and the experimenter compared the detector output with his impressions of the sound on the tape. For thresholds much below −40 dbm there was noticeable detector triggering on breathing, moving the telephone handset, and other spurious noises. It was found that passing the speech through a high-pass filter helped considerably in reducing such noise operation. An SKL filter was used with a cutoff frequency of 170 cps. The filter has unity gain at 1000 cps.

Even with the filter in place, however, setting the threshold much below −40 db still seemed to result in some unwanted noise operation, and −40 dbm became the arbitrary choice for the fixed threshold. This is still a fairly sensitive value; echo suppressors, for example, commonly operate at −32-dbm sensitivity and adequately detect speech.* (Of course, the suppressors are equipped with hangover, which helps bridge subthreshold gaps. The comparison is still within reason, since fill-in will later be applied to our present data.)

The threshold is thus an unavoidable compromise between too much noise and too much lost speech. The value was chosen arbitrarily be-

* Echo suppressors are voice-operated devices which insert losses into transmission paths to alleviate the irritating effects of echoes.[2]

cause the data did not indicate an optimum setting, contrary to our initial expectations.

4.2 *Throwaway Time*

The throwaway operation is intended to remove spurts generated by short bursts of noise. The throwaway time should be long enough to eliminate most of the spurious noise operation, but short enough to avoid discarding much speech. Electrical disturbances from various sources, such as the tape recorder, constitute a type of spurious noise. From listening to the tapes, it appears that this noise consists of an occasional impulse of fairly low level. This would normally register as one pulse on the speech detector and show up as a 5-msec spurt. The impulse of noise is, of course, smeared out by the tape recording process, and if the impulse occurs near the end of a sampling interval, it may be wide enough to bridge two intervals, causing a 10-msec spurt to register. Visual observation of the detector output shows that every so often, one or two speech detector pulses appear when the talker appears to be silent. As a first guess, 10 msec seems reasonable as a throwaway time.

To see if a throwaway time could be determined from the data, the original unedited speech tapes were processed with a detector sensitivity of -40 dbm, using the high-pass filter. An arbitrary fill-in time of 150 msec was chosen to bridge over most of the short gaps occurring within talkspurts, leaving long pauses suitable for searching for spurious noises. Thus, if a 5-msec spurt is observed it is a relatively isolated event, for the fill-in operation has presumably bridged over such short spurts occurring within talkspurts. The eight conversations analyzed lasted a total of 51.56 minutes, representing 103.12 minutes of speech by the 16 subjects.

The results showed that of a total of 3375 spurts observed, 953, or 27 per cent, were 5 or 10 msec long. A 10-msec throwaway time would eliminate this excess of very short spurts. Should the throwaway time be greater? It turned out that an excess of spurts did not occur in other short spurt regions, such as 15 to 20 msec (61 spurts) 25 to 30 msec (45 spurts), or 35 to 40 msec (41 spurts). Since there is no reason to believe that spurts longer than 10 msec are due to circuit noise rather than the conversants, a 10-msec throwaway time was chosen.

Now, there are several types of noise which a 10-msec fill-in time could never eliminate. A great deal of the noise on the tapes was caused by the conversants; breathing, coughing, etc., generally operated the speech detector, sometimes for long durations (at least 100 msec). There seems to be little possibility of distinguishing talker-generated

noises from speech solely on the basis of their on-off patterns. However, it is reasonable to include such sounds in the speech analysis because:

(1) These noises, although they carry little speech information, are generally audible to the listener and are a legitimate part of the conversational exchange.

(2) If there are any voice-operated devices on a communications circuit, these will be influenced just as much by coughing, etc., as by speech. If the results of this study are ever to be used to predict the behavior of these devices, then the data must include all sounds, speech or otherwise.

4.3 *Fill-In Time*

The fill-in time should be chosen just long enough to bridge the longest gaps in the continuous speech, having applied the throwaway operation. The distribution of the long gaps in the continuous speech gap distribution is shown in Table I.

From the data of Table I, there appears to be no obvious setting for the fill-in time which should be used to bridge the gaps. A fill-in time in excess of 250 msec would be required to bridge *all* the gaps, but it seems unreasonable to pick a value based on the one or two longest gaps. Because the distribution trails off smoothly without an obvious breakpoint, we arbitrarily select 200 msec as a fill-in time. There are two justifications for this choice:

(1) 200 msec is an easily remembered number. By its very nature, it appears to be rounded off, and therefore an approximation, which of course it is.

(2) When the panel of listeners monitored the continuous speech tapes, each member detected, on the average, 15 pauses in the speech.

TABLE I — DISTRIBUTION OF GAPS GREATER THAN 150 MSEC FOR ALL TALKERS*

Length, msec	Number	Per Cent
155,160	20	0.44
165,170	12	0.26
175,180	8	0.18
185,190	4	0.09
195,200	8	0.18
205–225	5	0.11
230–250	3	0.07
>250	2	0.04

Total number of gaps = 4537

* Continuous speech, 16 talkers, both circuit conditions. Throwaway = 10 msec, fill-in = 10 msec, threshold = −40 dbm; 170-cps high-pass filter.

If the 15 longest gaps are thrown out of the Table I distribution, the longest remaining gap is, by happy coincidence, 200 msec long. (This does not imply that the 15 longest gaps are those particular ones which the subjects called pauses.)

4.4 *Additional Gap and Spurt Data*

Although the continuous speech data already reported were sufficient for purposes of setting threshold and fill-in, additional data were obtained which may be of interest to some researchers. Fig. 6 is a plot of

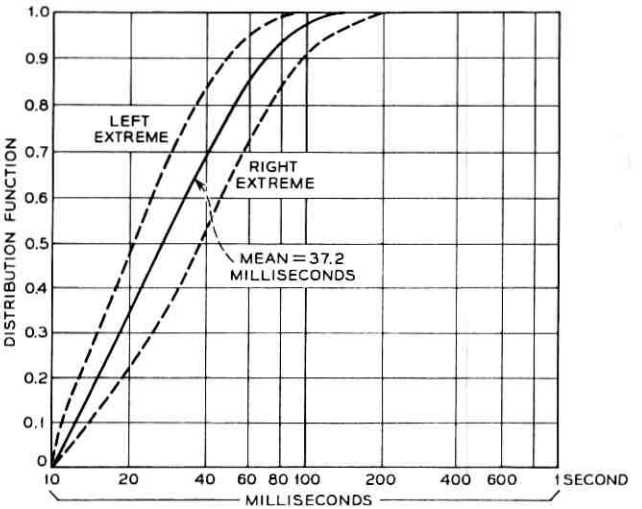


Fig. 6 — Continuous speech gap distribution: solid curve is the distribution for the entire speech sample for all talkers; dashed curves are composites of the boundaries of the curves for individual speech samples. Conditions: 16 speakers; 27.4 minutes of continuous speech; both delay and no-delay circuit conditions; speech detector threshold −43 dbm with no filter; throwaway time 0; fill-in 10 msec; 3754 gaps.

the gap distribution for the 27.4 minutes of continuous speech for all 16 talkers over both circuit conditions. The speech detector threshold was set at −43 dbm. No filter was used prior to the detector.

To determine the variations in the gap distribution among the different talkers, the individual gap distributions for all 30 samples* were plotted on one graph. The left boundary of this graph is shown as the "left extreme" curve of Fig. 6. This curve is not necessarily the curve for a single sample but is a composite of all curves which happen to fall on the boundary. The same applies to the "right extreme" curve.

* Sixteen talkers on the standard circuit, fourteen on 400-msec delay.

Visual inspection of the curves for individual talkers shows no apparent over-all differences between male and female talkers.

Fig. 7 shows the continuous speech spurt distribution. The spurt distribution might be considered representative of syllabic bursts, but care must be taken in drawing conclusions from these data since the graph is strongly influenced by the threshold setting (as shown in Fig. 5) and also because of a possible tape splice effect, described as follows.

A probabilistic analysis of the influence of tape splices on the speech reveals that of the more than 770 splices which were present, about 130 of them involved a gap. Thus, 130 of the 3754 gaps (or 3.5 per cent) were affected by splices, probably only slightly influencing the gap distribution of Fig. 6.

It turns out, however, that almost all of the splices had a spurt on one side or both (virtually all of the 130 splices involving a gap had a spurt on the other side). This means that over 20 per cent of the total number of spurts were affected in some way by tape splices.

Because of the influence of artifacts on the spurts, further analysis of the spurt distribution (such as obtaining the range among subjects) was not carried out.

V. TALKSPURT AND PAUSE DISTRIBUTIONS FOR CONVERSATIONAL SPEECH

5.1 *Data from Conversations*

This section is an illustration of the results obtained from analysis of conversational speech. Fig. 8 is a plot of the distribution of talkspurts and pauses for all 16 speakers engaging in eight conversations over the standard circuit. The conversations lasted almost 52 minutes, yielding about 103 minutes of conversation data for all subjects. Some of the more interesting statistical measures of these data are shown in Table II.

The results shown here are included only to illustrate the speech measuring technique developed in this study. The conversations were artificially induced in a test-room atmosphere in which the subjects knew they were being recorded. There is no assurance that the statistical measures shown in Table II and in Fig. 8 are representative of those which would be obtained on real telephone calls.

5.2 *A Comparison with Other Studies*

Many other studies have been concerned with measuring some of the statistical properties of speech patterns. Three of these studies are selected for comparison with our results.

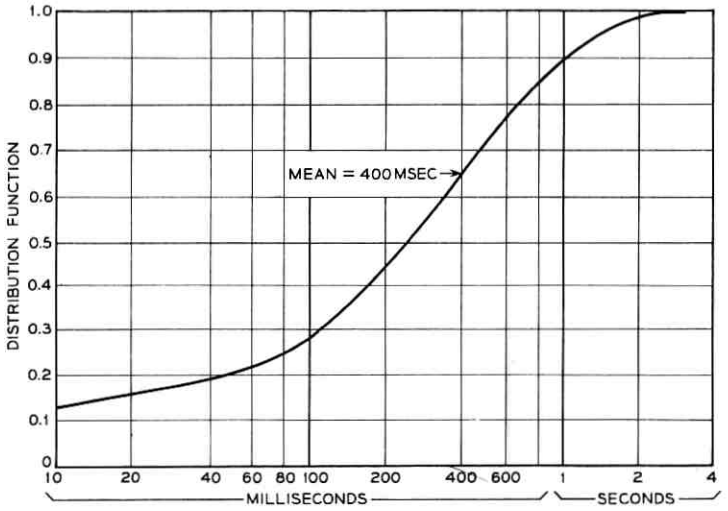


Fig. 7 — Continuous speech spurt distribution. Conditions: 16 speakers; 27.4 minutes of continuous speech; both delay and no-delay circuit conditions; threshold −43 dbm; throwaway time 0; fill-in 10 msec; 3754 spurts.

Norwine and Murphy[6] observed "talkspurts" and "pauses" in oscillograph recordings of telephone conversations made on a special circuit between New York and Chicago. The circuit had a round-trip delay of almost 600 msec and was equipped with echo suppressors. Detailed observations were made from about 4400 feet of graph paper, representing 51 calls with a total duration exceeding 13,000 seconds. The Norwine and Murphy definition of talkspurt is significantly different from ours and is included here.

"A talkspurt is speech by one party, *including his pauses*, which is preceded and followed . . . by speech from the other party perceptible to the one producing the talkspurt." (Italics mine.)

Since Norwine and Murphy include pauses in their talkspurts, their talkspurts are much longer than ours, and a direct comparison of their distributions with ours is inappropriate. It is possible, however, to apply a correction on one or two statistics and make a fair comparison. Considering the mean talkspurt length, Norwine and Murphy indicate a value of 4.3 seconds for 2845 talkspurts. These talkspurts include, however, 2811 pauses with a mean of 0.73 seconds. One may then calculate that there were about 10,200 seconds of speech composed of 2845 + 2811 = 5656 "shortened talkspurts." (For each pause inserted, a new talkspurt is created.) The new average talkspurt length is 1.8 seconds, which compares with our average of 1.34 seconds.

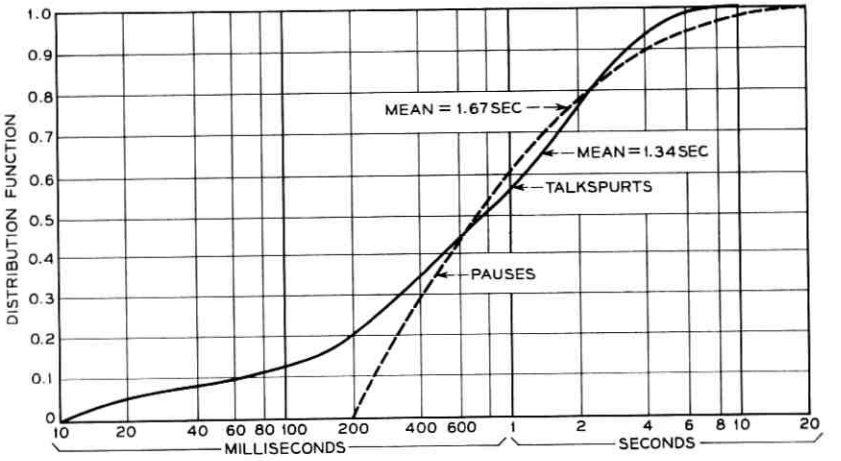


Fig. 8 — Talkspurt and pause distributions for all speakers combined; there were 51.56 minutes of conversational speech: therefore 16 speakers talked for 103.12 minutes. Conditions: no delay; threshold −40 dbm; throwaway 10 msec; fill-in 200 msec; 170-cps high-pass filter in detector circuit; 2042 talkspurts, 2054 pauses.

More recently, measurements were made on calls placed on the Atlantic cable via TASI.* "Spurts" and "gaps" were determined by the TASI speech detector operation, and results were tabulated for about 600 commercial calls. The TASI speech detector has an operate (pickup) time of 5 msec and a deferred hangover of 240 msec. A deferred hangover means that the full hangover value is applied only for longer spurts (60 msec or greater) but for very short spurts the hangover is shorter (minimum of 25 msec). The sensitivity is −40 dbm re 0TL.

Although the TASI detector uses hangover rather than fill-in and a delayed operate time rather than throwaway, the values used for the measuring parameters are very similar to ours, especially since the sensitivity is the same. The TASI data indicate an activity period (fraction of time the TASI detector is operated) of 48 per cent averaged over all calls. This compares with 44.3 per cent obtained in our study, as shown in Table II. In addition, the mean TASI "spurt" length is 1.3 seconds, compared with our value of 1.34 seconds.

A distribution of TASI talkspurts is shown in Fig. 9 and was taken from an unpublished report by H. Miedema,[8] printed here with the author's permission. Fig. 9(a) shows both the talkspurt distribution

* TASI is essentially a bank of voice-operated switches which connects a subscriber to a channel only when he is actually talking. Thus 72 people may talk over 36 channels.[7]

TABLE II — SOME STATISTICAL MEASURES OF EIGHT CONVERSATIONS

Total conversation time	51.56 min
Per cent time both speakers are silent	18.2%
Per cent time circuit in use by either or both speakers	81.8%
Per cent time double talking	7.2%
Per cent time the average conversant talks (obtained by dividing the total time all speakers A talk plus the total time of all B talk by 103.12 minutes)	44.3%
Number of talkspurts (all 16 speakers)	2042
Median talkspurt length	0.77 sec
Mean talkspurt length	1.34 sec
Number of pauses	2054*
Median pause	0.72 sec
Mean pause	1.67 sec

Circuit conditions: 500-type sets, four-wire connections, 16-db loss between speakers, speech recorded at 0TL point, 6 db from each transmitter.
Measurement parameters: speech detector threshold set at −40 dbm re 0TL, throwaway time = 10 msec, fill-in time = 200 msec, 170-cps high-pass filter.

* There are more pauses than talkspurts because several of the eight conversations began and ended with a pause for each of the speakers.

obtained in our study and the distribution of TASI speech detector spurts for U. S. talkers. Also included is a "corrected" curve, in which the 10-msec pickup time is added to each talkspurt and the hangover time is subtracted. Since the hangover time is variable, the correction depends on the talkspurt length. The net effect is to shift the curve left 15 msec for short spurts (25-msec hangover minus 10-msec pickup) and 230 msec for long spurts. This widens the discrepancy between our results and the TASI data, making the TASI talkspurts appear shorter. Some of the short TASI spurts, however, may have been due to line noise on the trunk. The magnitude of this effect cannot be determined, since there are at present no available data on TASI noise operation.

Fig. 9(b) shows the pause distributions. Again, the original pause curve must be corrected for hangover. We will assume that a hangover of 240 msec preceded each gap (and it did for all gaps following the 90 per cent of the talkspurts exceeding 60 msec). The 10-msec pickup time will reduce the correction to 230 msec. The resulting curve is in this case very close to ours.

Finally, J. F. Agnello of Ohio State University made an analysis of the gaps which occur in spoken text, and he studied the effect on the distributions of varied text (prose, poetry, single sentences).[9] He differentiated between *intraphase pauses* (gaps) and *interphase pauses* (pauses) and had the speakers listen, as a group, to their own speech, indicating on the printed text whenever they detected a pause. A "pause timer"

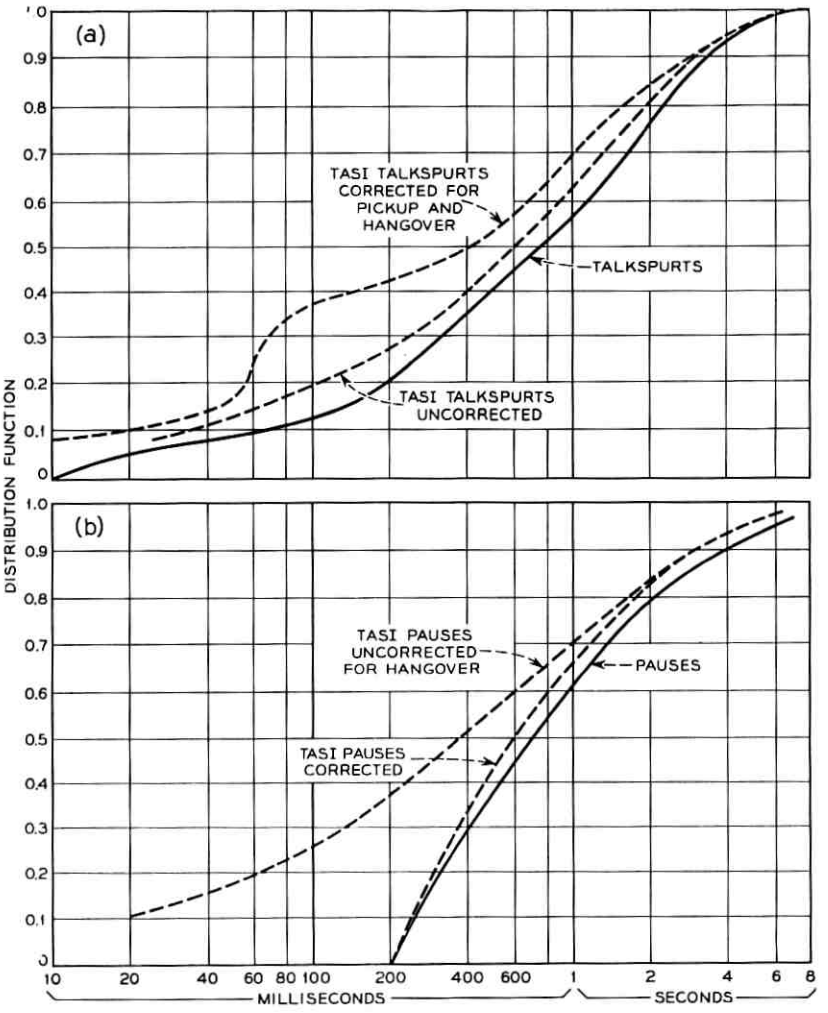


Fig. 9 — A comparison of talkspurt and pause data with TASI data: (a) talkspurts, (b) pauses.

detected and classified by duration 887 pauses exceeding 150 msec in about 40 minutes of speech, and the listeners indicated 710 pauses. Agnello concludes that "the minimal detected pause was estimated . . . to be 190 msec." He also concludes on other grounds that the longest *intraphase pause* was 200 msec long. These results are in excellent agreement with our choice of 200 msec as a fill-in time.

VI. DISCUSSION

6.1 *Suitability of Technique*

The question which now presents itself is whether or not the method developed here for speech detecting is a good technique. In answering this question, we are immediately faced with another one, that of defining what is meant by a "good technique." The answer to the latter question is, of course, a function of the intended purpose of the detector.

It was originally hoped that the computer would yield speech patterns such that segments of talkspurts would be bridged together and spurious activity would be ignored. One way of evaluating this criterion is to look at the curves of Fig. 8 and see if these results are consistent with our goals.

On one hand, the detection technique seems to be a meaningful indication of perceived speech because of the many points of correspondence between the Fig. 8 data and that of other studies. For example, although our data do not exactly duplicate the TASI detector operation, there are many similarities, and people are able to converse fluently over TASI.

On the other hand, the talkspurt data of Fig. 8 seem unreasonable because of the presence of numerous short spurts. Note that 10 per cent are less than 60 msec long. It is unlikely that an utterance of such short duration would be considered by a listener to be a "talkspurt." What, then, are these sounds?

On listening to the tapes, one observes that there *are* occasional short utterances, such as those produced when a person begins to talk but suddenly realizes that he has been interrupted. These are sometimes as short as a few pitch periods. Short sounds are often produced by parting the lips quickly, or by clicking the tongue against the roof of the mouth. Another source of short sounds is the low-level speech which triggers the speech detector only intermittently. An informal experiment was carried out in which a speech detector was hooked up to cut out any speech exceeding a threshold, leaving only the subthreshold speech audible. (The detector consisted of a relay with 5-msec pickup and hangover times.) The three people who listened to the speech noticed that occasionally whole phrases would get through. This was an infrequent event, however, for sensitivities as low as -40 dbm.

The author is convinced that these short sounds are indeed produced by the talkers and are not extraneous noises. They may be fragmentary parts of talkspurts, or clicks, or whatever you will, but their sources are indistinguishable to the computer. Within the framework of the de-

tection technique adopted here there are two things which can be done with them:

(1) Throw them out with some arbitrary throwaway time. This introduces the possibility of discarding real speech, as well as the difficulty of establishing a suitable throwaway time.

(2) Keep them as is, realizing that there is some doubt that they represent meaningful speech, and hope that they will not hamper further work with speech analysis. The author has chosen to adopt this second approach.

6.2 *Relationship to the Acoustics of Speech Production*

The curves presented in Fig. 8 are obtained from gross statistics of over 100 minutes of speech. No effort has been made to refine and classify the data into subheadings such as "plosives" or "glottal consonants," etc. Accordingly, it is not possible, given only the curves shown here, to explain the shape of the curves in terms of the acoustics of speech production. This could easily be a study unto itself, and, indeed, studies of this nature presently constitute an important phase of speech research by acousticians and phoneticians who are expert in this field.

If anyone wishes to examine our results in the light of data in the speech literature, he must note that our speakers were not trained and made no effort to talk clearly and precisely. In addition, they spoke over telephones, not high-quality microphones. This procedure is quite different from that used in most speech studies, in which the speech is carefully manicured to approach an "ideal" sound.

VII. CONCLUSION

This study has shown that even small changes in the measuring technique can produce noticeable effects on the results. Data on speech dynamics must therefore include a detailed description of the technique used to measure the speech. Without this, the results must be regarded as unrepeatable.

Many unforeseen difficulties arose with the proposed technique of using a speech detector with a fixed threshold followed by a computer program with simple throwaway and fill-in operations. The setting of the detector threshold is a compromise between excessive noise and too little speech operation, and errors of both kinds must be expected. Although a fill-in time of 200 msec seems fairly well established, the throwaway time of 10 msec is not adequate to remove many short spurts, some of which may represent legitimate talkspurts, while others

do not. Indeed, it may turn out that a simple automatic speech detecting technique involving fixed parameters is inadequate for some purposes, and a considerably more sophisticated method must be employed.

What therefore originally began as a seemingly straightforward attempt to build a speech detector has instead exposed a problem of far more difficulty than was first imagined. The data obtained here have hopefully shed light on some aspects of the problem, but further study is required before a completely satisfactory solution is obtained.

VIII. ACKNOWLEDGMENTS

I wish to thank Mrs. Nancy Shrimpton for writing the computer program used to process the speech and, in addition, for assisting in the data analysis. I would also like to thank Miss Kathryn McAdoo for making the VU measurements on the continuous speech samples.

REFERENCES

1. Emling, J. W., and Mitchell, D., The Effects of Time Delay and Echoes on Telephone Conversations, B.S.T.J., **42,** Nov., 1963, p. 2869.
2. Brady, P. T., and Helder, G. K., Echo Suppressor Design in Telephone Communications, B.S.T.J., **42,** Nov., 1963, p. 2893.
3. Riesz, R. R., and Klemmer, E. T., Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications, B.S.T.J., **43,** Nov., 1963, p. 2919.
4. McAdoo, K. L., Speech Volumes on Bell System Message Circuits — 1960 Survey, B.S.T.J., **42,** Sept., 1963, p. 1999.
5. Beranek, L. L., *Acoustic Measurements*, Wiley, New York, 1949, p. 504.
6. Norwine, A. C., and Murphy, O. J., Characteristic Time Intervals in Telephone Conversation, B.S.T.J., **17,** April, 1938, p. 281.
7. Miedema, H., and Schachtman, M. G., TASI Quality — Effect of Speech Detectors and Interpolation, B.S.T.J., **41,** July, 1962, p. 1455.
8. Miedema, H., unpublished report presenting TASI speech detector data, 1960.
9. Agnello, J. G., A Study of Intra- and Inter-Phrasal Pauses and Their Relationship to the Rate of Speech, Ohio State University Ph.D. Thesis, 1963.

Interaction of Adaptive Antenna Arrays in an Arbitrary Environment

By SAMUEL P. MORGAN

(Manuscript received July 22, 1964)

This paper deals with adaptive transmitting arrays in which the excitations of the elements are varied in response to a pilot field incident on the array from a distant source. General theorems, some quite simple, are obtained relating to optimal power transfer from an adaptive array in an arbitrary reciprocal medium to either a single receiver or a receiving array. We assume first that the amplitudes and phases of the transmitting elements are separately adjustable, and afterward that only the phases are adjustable. The results involve in particular the matrix which represents the pilot fields produced at the elements of the transmitting array by currents at the locations of the receiving elements. In some important special cases, optimal power transfer results from making the phase of each transmitting element equal to the negative of the phase of the pilot field at that element.

We also consider the dynamic behavior of two adaptive arrays which simultaneously transmit and receive, the phases on transmission being made equal to the negatives of the received phases. Analysis of an idealized model indicates that the arrays will reach a unique steady state which is in practical cases identical with or very close to the condition for optimal power transfer. Some numerical simulations of 2- and 3-element interacting arrays have been made to show how such arrays approach an essentially steady state under moderately realistic assumptions.

I. INTRODUCTION AND SUMMARY

A number of recent papers[1,2] have dealt with adaptive antenna arrays, also called self-steering or retrodirective arrays. In an adaptive transmitting array, the excitations of the individual elements are electronically varied in response to a pilot field incident on the array from a distant terminal, in order to steer the beam to the terminal which is originating the pilot signal. It is easy to see that in free space the required steering can be accomplished by making the phase of each element equal

to the negative of the phase of the pilot beam at the given element. Cutler, Kompfner, and Tillotson[1] and others[2] have shown how phase reversal can be obtained using frequency conversion techniques.

This paper deals with general adaptive transmitting arrays in an arbitrary environment. The transmission medium need not be homogeneous or isotropic, but it is assumed to be linear and symmetric and to be time-invariant, at least over intervals comparable to the propagation time between transmitter and receiver. We are concerned in particular with the conditions for optimal power transfer from an adaptive transmitting array to either a single receiver or a receiving array. We shall also investigate the transient and steady-state behavior of two interacting adaptive arrays, each of which simultaneously transmits and adjusts the excitations of its elements in response to the field received from the other array.

In Section II we consider an adaptive array in which the amplitudes and phases of the excitations of the individual elements are separately adjustable, but the total radiated power is fixed. Such an array is easier to treat mathematically than one in which the excitation amplitudes are all fixed and only the phases are variable, even though the latter array might be easier to build. In the most general case, the power radiated by the transmitting array is a positive definite Hermitian form in the element excitations, and the received power is a positive definite Hermitian form in the electric fields at the elements of the receiving array. The distribution of excitations which maximizes the ratio of received to radiated power is the eigenvector corresponding to the largest eigenvalue of a certain pencil of Hermitian matrices. The matrices in question are constructed from the impedance matrix of the transmitting array, the admittance matrix of the receiving array, and a Green's function matrix of pilot fields produced at the transmitting elements by currents at the receiver locations. The results simplify considerably if the elements of each array are uncoupled and are identical among themselves. In particular, if the receiver consists of but a single element, and if the transmitter elements are identical and uncoupled, then the optimal excitation of each element is merely proportional to the complex conjugate of the pilot field at that element.

Section III contains a brief discussion of the problem of maximizing the power transferred from an arbitrary transmitting array to an arbitrary receiving array when the excitation amplitudes are fixed and only the phases are adjustable. Maximum power is always conveyed to a *single* receiver by reversing the phase of the pilot field at each element of the transmitting array. This phase reversal principle, first recognized for

free-space transmission, is thus shown to be valid for an arbitrary transmission medium. For multielement receivers an explicit solution is not given; but an example shows that even when each array consists of identical, uncoupled elements, maximum power transfer does *not* generally correspond to reversing the phase of the total pilot field at the transmitter elements.

Section IV deals with the interaction of two adaptive arrays or, in principle, the interaction of an adaptive radar array with itself. A mathematical model is set up, in which each array transmits constant power and continuously adjusts the excitations of its own elements to be proportional to the complex conjugate of the incident field. A single delay time is taken to represent the transmission delay between the two arrays. The transient behavior of this model turns out to be quite simple, and it is shown that in general, excluding mathematically pathological cases, the two arrays reach an equilibrium configuration which depends only on the Green's function (pilot field) matrix corresponding to the given geometry and transmission medium. In the most general case the equilibrium configuration is not the same as the condition for optimal power transfer derived in Section II; but it *is* the condition for optimal power transfer in the important special case when the elements of each array are identical among themselves and the interelement coupling is zero. If the elements are nearly identical and the mutual impedances are small compared to the self-impedances, then the equilibrium configuration should be nearly the same as the configuration for optimal power transfer.

Numerical simulations of the transient behavior of 2- and 3-element interacting adaptive arrays are described in Section V, both for the case of simultaneous phase and amplitude variations, and for the case of phase variations only. The simulations also include the effects of small differences in the interelement delay times compared to the average delay between the arrays. Random choices are made for the elements of the Green's function matrix and for all pairs of interelement delays. Simulations of 50 pairs of 2-element arrays and 25 pairs of 3-element arrays indicate that arrays with only phase adjustment approach a steady state about as quickly as arrays with both phase and amplitude adjustment (of course, the two steady states are not the same). Interelement delay differences which are small compared to the average interelement delay produce small fluctuations about the steady state which would be achieved for equal delays.

The results obtained in this paper depend only on the linearity, symmetry, and time-invariance of the transmission medium; in particular, they do not involve calculating any antenna patterns. Pattern calcula-

tions would be necessary if one wished to get numerical values for maximal power transfer, or to estimate the radiated fields in unwanted directions. Furthermore, the analysis is essentially for a single frequency; variations in phase and amplitude are assumed to be very slow compared to the transmission times involved. It would be worthwhile to study the behavior of adaptive arrays over a finite frequency band, but such a study is outside the scope of the present paper.

II. OPTIMAL POWER TRANSFER BETWEEN ARBITRARY ANTENNA ARRAYS

Consider a transmitting array and a receiving array embedded in an arbitrary linear, time-invariant medium, as in Fig. 1. The medium may be inhomogeneous and anisotropic, but the permeability, permittivity, and conductivity tensors at any point are assumed to be symmetric. (This rules out ferrites and plasmas in the presence of a magnetic field.) All fields are assumed to be time-harmonic with angular frequency ω, the time dependence exp $i\omega t$ being suppressed. For simplicity the individual radiators and receivers are taken to be elemental electric dipoles, although they could equally well be elemental current loops. The assumption of dipole sources is not a major restriction, since the dipoles could be used, for example, together with microwave circuitry to feed aperture-type radiators such as elemental horns.

Let the transmitting array have M elements and let the complex excitation of the ith element be $I_{1,i}$. Physically $I_{1,i}$ may be regarded as the electric moment of an elemental current, having the dimensions of ampere-meters. The M-component vector

$$\mathbf{I}_1 = (I_{1,1}, I_{1,2}, \cdots, I_{1,M}), \tag{1}$$

whose components are complex scalars, will be called the excitation of Array 1. Similarly let the receiving array have N elements, and let the

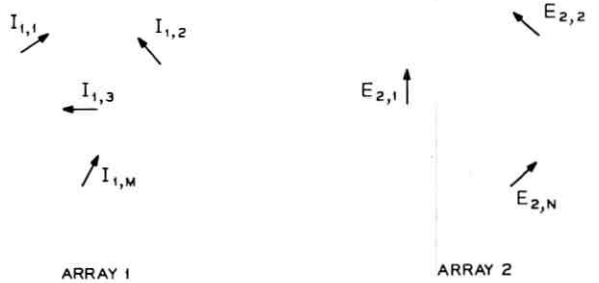


Fig. 1 — Schematic representation of arbitrary transmitting and receiving arrays of electric dipoles.

complex electric field component at the location and in the direction of the jth element be $E_{2,j}$. Then the electric field at Array 2 is the N-component vector

$$\mathbf{E}_2 = (E_{2,1}, E_{2,2}, \cdots, E_{2,N}). \tag{2}$$

If Array 2 is transmitting and Array 1 is receiving, we define the vectors $\mathbf{I}_2$ and $\mathbf{E}_1$ in an analogous way.

The total power P_T radiated by the transmitter over all space is given by the Hermitian form

$$P_T = \tfrac{1}{2}(\mathbf{Z}_1\mathbf{I}_1, \mathbf{I}_1), \tag{3}$$

where $\mathbf{Z}_1$ is an $M \times M$ positive definite Hermitian matrix and $(\mathbf{x},\mathbf{y})$ represents the scalar product of two vectors $\mathbf{x}$ and $\mathbf{y}$. A fuller discussion of notation and of the properties of Hermitian forms is given in Appendix A.

In principle the radiation impedance matrix $\mathbf{Z}_1$ may be determined from experimental measurements, or it may be calculated from the fields of the radiating elements. For example, if the field due to unit excitation of the ith element at a great distance R from all currents and material media is $\mathcal{E}_{1,i}/R$, then by integrating the Poynting vector due to the whole array over a large sphere we find that the total radiated power is given by an expression of the form (3), with

$$Z_{1,ij} = \frac{1}{\eta}\int \mathcal{E}_{1,i}\cdot\mathcal{E}_{1,j}^{*}\,d\Omega, \qquad i = 1,2,\cdots,M; \qquad j = 1,2,\cdots,M, \tag{4}$$

where η is the characteristic impedance of free space and $d\Omega$ is an element of solid angle.

We now assume that the power P_R received by Array 2 is given by a Hermitian form in $\mathbf{E}_2$, the electric field which would exist at Array 2 if its elements were open-circuited. Thus we write

$$P_R = \tfrac{1}{2}(\mathbf{Y}_2\mathbf{E}_2, \mathbf{E}_2), \tag{5}$$

where $\mathbf{Y}_2$ is an $N \times N$ positive definite Hermitian matrix. Equation (5) is equivalent to the assumption that the transmitter field is independent of whether or not currents are flowing in the elements of Array 2, i.e., that the back reaction of Array 2 on Array 1 is negligible. This will be a very good approximation in the practical case where the arrays are far apart, so that P_R is a very small fraction of P_T.

The field at Array 2 is related to the excitation of Array 1 by the Green's function matrix $\boldsymbol{\Gamma}$; thus

$$\mathbf{E}_2 = \boldsymbol{\Gamma}\mathbf{I}_1, \tag{6}$$

where $\mathbf{\Gamma}$ is an $N \times M$ matrix and Γ_{ij} represents the field at the ith element of Array 2 due to unit excitation of the jth element of Array 1. A basic reciprocity theorem for linear, time-invariant, symmetric media, proved in Appendix B, guarantees that Γ_{ij} also represents the field at the jth element of Array 1 due to unit excitation of the ith element of Array 2. Thus if Array 2 has the excitation $\mathbf{I}_2$, the field at Array 1 is given by

$$\mathbf{E}_1 = \mathbf{\Gamma}'\mathbf{I}_2, \tag{7}$$

where $\mathbf{\Gamma}'$ is the transpose of $\mathbf{\Gamma}$.

From (5) and (6), the power received by Array 2 is

$$P_R = \tfrac{1}{2}(\mathbf{Y}_2\mathbf{\Gamma}\mathbf{I}_1, \mathbf{\Gamma}\mathbf{I}_1) = \tfrac{1}{2}(\mathbf{\Gamma}^\dagger\mathbf{Y}_2\mathbf{\Gamma}\mathbf{I}_1, \mathbf{I}_1), \tag{8}$$

where $\mathbf{\Gamma}^\dagger$ is the adjoint (= conjugate transpose) of $\mathbf{\Gamma}$. We wish to maximize the ratio of received power to transmitted power, which is

$$\frac{P_R}{P_T} = \frac{(\mathbf{\Gamma}^\dagger\mathbf{Y}_2\mathbf{\Gamma}\mathbf{I}_1,\mathbf{I}_1)}{(\mathbf{Z}_1\mathbf{I}_1,\mathbf{I}_1)}. \tag{9}$$

But the right side is the quotient of two Hermitian forms in which the denominator is positive definite; and it is well known (see Appendix A) that the maximum value of the quotient is the largest eigenvalue λ_M of the pencil of matrices $\mathbf{\Gamma}^\dagger\mathbf{Y}_2\mathbf{\Gamma} - \lambda\mathbf{Z}_1$. The desired eigenvalue is the largest root of the equation

$$\det(\mathbf{\Gamma}^\dagger\mathbf{Y}_2\mathbf{\Gamma} - \lambda\mathbf{Z}_1) = 0; \tag{10}$$

and the corresponding eigenvector, which maximizes the right side of (9), is any nonzero solution of the system of equations

$$\mathbf{\Gamma}^\dagger\mathbf{Y}_2\mathbf{\Gamma}\mathbf{I}_1 - \lambda_M\mathbf{Z}_1\mathbf{I}_1 = 0. \tag{11}$$

The foregoing equations simplify in a special case which will be important in what follows, namely when all the self-impedances and self-admittances are equal and all the mutual impedances and admittances are zero. In this case we may write

$$\mathbf{Z}_1 = R_1\mathbf{1}_M, \qquad \mathbf{Y}_2 = G_2\mathbf{1}_N, \tag{12}$$

where R_1 and G_2 are real scalars and $\mathbf{1}_M$ and $\mathbf{1}_N$ are unit matrices of orders M and N respectively. Then (9) becomes

$$\frac{P_R}{P_T} = \frac{G_2}{R_1}\frac{(\mathbf{\Gamma}^\dagger\mathbf{\Gamma}\mathbf{I}_1,\mathbf{I}_1)}{(\mathbf{I}_1,\mathbf{I}_1)}, \tag{13}$$

and the maximum value of the ratio is proportional to the largest eigenvalue of the matrix $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$, that is, the largest root λ_M of

$$\det\,(\mathbf{\Gamma}^\dagger\mathbf{\Gamma} - \lambda\mathbf{1}_M) = 0. \tag{14}$$

The excitation corresponding to maximum power transfer is any nonzero solution of

$$\mathbf{\Gamma}^\dagger\mathbf{\Gamma}\mathbf{I}_1 - \lambda_M\mathbf{I}_1 = 0. \tag{15}$$

The optimal transmitter excitation given by (14) and (15) is one which can exist when both arrays are transmitting and the excitation of each element of each array is proportional to the complex conjugate of the field incident on the element from the other array. Suppose, for example, that

$$\mathbf{I}_1 = M_1\mathbf{E}_1^*, \qquad \mathbf{I}_2 = M_2\mathbf{E}_2^*, \tag{16}$$

where M_1 and M_2 are complex scalars and $\mathbf{E}_1$ and $\mathbf{E}_2$ are related to $\mathbf{I}_2$ and $\mathbf{I}_1$ by the Green's function matrix, as in (6) and (7). Then it is easy to show that $\mathbf{I}_1$ must satisfy

$$\mathbf{I}_1 = M_1M_2^*\mathbf{\Gamma}^\dagger\mathbf{\Gamma}\mathbf{I}_1\,. \tag{17}$$

A nonvanishing solution of (17) exists if and only if

$$M_1M_2^* = 1/\lambda, \tag{18}$$

where λ is an eigenvalue of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$, that is, a root of the determinantal equation (14). Although steady-state excitations satisfying (16) are mathematically possible when λ is any eigenvalue of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$, it is shown in Section IV that the system is unstable unless λ is the largest eigenvalue, and that the excitations corresponding to the largest eigenvalue are in fact the excitations toward which two interacting adaptive arrays tend.

We shall now consider optimal power transfer in the special case where Array 2 consists of but a single receiver. Then $\mathbf{Z}_1$ is an $M \times M$ matrix, $\mathbf{Y}_2$ is a 1×1 matrix, i.e., a scalar quantity G_2, and $\mathbf{\Gamma}$ is a $1 \times M$ matrix. Equation (10) is therefore equivalent to

$$\begin{aligned}\det\,(\lambda G_2^{-1}\mathbf{Z}_1 - \mathbf{\Gamma}^\dagger\mathbf{\Gamma}) &= \det\,(\lambda G_2^{-1}\mathbf{Z}_1)\,\det\,(\mathbf{1}_M - \{\lambda G_2^{-1}\mathbf{Z}_1\}^{-1}\mathbf{\Gamma}^\dagger\mathbf{\Gamma}) \\ &= \lambda^M G_2^{-M}\det\mathbf{Z}_1\det\,(\mathbf{1}_1 - \mathbf{\Gamma}\{\lambda G_2^{-1}\mathbf{Z}_1\}^{-1}\mathbf{\Gamma}^\dagger) \\ &= (\lambda/G_2)^{M-1}(\lambda/G_2 - \mathbf{\Gamma}\mathbf{Z}_1^{-1}\mathbf{\Gamma}^\dagger)\,\det\,\mathbf{Z}_1 \\ &= 0,\end{aligned} \tag{19}$$

where in the second step we have made use of a lemma due to Sandberg,[3]

which is stated and proved in Appendix C. It follows that the only nonzero eigenvalue of (19) is

$$\lambda_M = G_2 \mathbf{\Gamma} \mathbf{Z}_1^{-1} \mathbf{\Gamma}^\dagger. \tag{20}$$

The corresponding eigenvector is

$$\mathbf{I}_1^{(M)} = \mathbf{Z}_1^{-1} \mathbf{\Gamma}^\dagger \tag{21}$$

up to a constant factor, since it is easy to see that

$$\begin{aligned} \mathbf{\Gamma}^\dagger G_2 \mathbf{\Gamma} \mathbf{I}_1^{(M)} - \lambda_M \mathbf{Z}_1 \mathbf{I}_1^{(M)} &= \mathbf{\Gamma}^\dagger G_2 \mathbf{\Gamma} \mathbf{Z}_1^{-1} \mathbf{\Gamma}^\dagger - \lambda_M \mathbf{Z}_1 \mathbf{Z}_1^{-1} \mathbf{\Gamma}^\dagger \\ &= \mathbf{\Gamma}^\dagger \lambda_M - \lambda_M \mathbf{\Gamma}^\dagger = 0. \end{aligned} \tag{22}$$

If the elements of the transmitting array are uncoupled, then the mutual radiation impedances vanish, and $\mathbf{Z}_1$ and $\mathbf{Z}_1^{-1}$ are diagonal matrices. The optimal excitation of the jth element is then

$$I_{1,j} = \Gamma_{1j}^*/R_{jj}, \qquad j = 1, 2, \cdots, M, \tag{23}$$

up to a constant factor, where Γ_{1j}^* is the complex conjugate of the pilot field produced at the jth element by a dipole at the location of the receiver, and R_{jj} is the radiation resistance of the jth element. If all the radiation resistances are equal, then since the eigenvector is determined only up to a multiplicative constant, we may take

$$I_{1,j} = \Gamma_{1j}^*, \qquad j = 1, 2, \cdots, M; \tag{24}$$

in other words, the excitation is merely proportional to the complex conjugate of the pilot field.

We have shown that if the mutual radiation impedances of the transmitter elements are zero, then the field at the receiver is maximized, for constant radiated power, when the phase of the excitation of each transmitter element is the negative of the phase of the pilot field. If, however, the transmitter elements are coupled by their radiation fields, so that the impedance matrix $\mathbf{Z}_1$ is not diagonal, the optimal excitations are given by (21) and do not generally satisfy the phase reversal condition.

III. OPTIMAL POWER TRANSFER WITH PHASE ADJUSTMENTS ONLY

If the amplitudes of the transmitter excitations are fixed but the phases are adjustable, we wish to maximize the received power,

$$P_R = \tfrac{1}{2}(\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma} \mathbf{I}_1, \mathbf{I}_1) = \tfrac{1}{2}(\mathbf{Y}_2 \mathbf{\Gamma} \mathbf{I}_1, \mathbf{\Gamma} \mathbf{I}_1), \tag{25}$$

when

$$I_{1,j} = r_j e^{i\theta_j}, \qquad j = 1, 2, \cdots, M, \tag{26}$$

and the r_j are fixed but the θ_j are at our disposal.

If there is only one receiver ($N = 1$), the solution is immediate. We have to maximize

$$P_R = \tfrac{1}{2} G_2 \left| \sum_{j=1}^{M} \Gamma_{1j} I_{1,j} \right|^2 \tag{27}$$

by adjusting the phases of the $I_{1,j}$; and it is clear that the modulus of the sum will be greatest when the phases of all the summands are equal, that is, when

$$\arg I_{1,j} \equiv \theta_j = -\arg \Gamma_{1j} + \text{constant}, \qquad j = 1, 2, \cdots, M. \tag{28}$$

In other words, the phase of the jth transmitting element should be the negative of the phase of the pilot field produced at that element by a radiating element at the position of the receiver. This result is independent of the nature of the transmission medium, subject only to the requirements of linearity, time-invariance, and symmetry, and it is independent of the position of the receiver relative to the transmitting array.

For a two-element transmitter ($M = 2$) and an arbitrary receiving array, we have

$$\begin{aligned} 2P_R = {} & (\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma})_{11} r_1^2 + (\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma})_{12} r_1 r_2 e^{i(\theta_2 - \theta_1)} \\ & + (\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma})_{21} r_1 r_2 e^{i(\theta_1 - \theta_2)} + (\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma})_{22} r_2^2. \end{aligned} \tag{29}$$

Since $\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma}$ is Hermitian, the right side of (29) is maximized by taking

$$\arg I_{1,j} - \arg I_{2,j} \equiv \theta_1 - \theta_2 = \arg (\mathbf{\Gamma}^\dagger \mathbf{Y}_2 \mathbf{\Gamma})_{12}, \tag{30}$$

and this is the condition for optimal power transfer if the transmitting array has only two elements.

A complete analytic solution of the problem of maximizing P_R for an arbitrary transmitter with $M \geqq 3$ and an arbitrary receiver with $N \geqq 2$ has not been found, although since P_R is a continuous, periodic function of each of the θ_j, it is obvious that a maximum exists and could be located as accurately as desired by an iterative numerical procedure.

In contrast to the situation for arrays with both amplitudes and phases adjustable, the condition for optimal power transfer between multielement arrays with fixed excitation amplitudes is not generally satisfied by making the phase of each element equal to the negative of the phase of the field incident from the other array, even if it be assumed that the array elements are uncoupled and are identical among themselves. As a

counterexample, consider the case in which each array has two elements, and $\mathbf{Y}_2$ is a multiple of the unit matrix. Let the element currents be

$$\begin{aligned} I_{1,1} &= re^{i\theta_1}, & I_{2,1} &= \rho e^{i\varphi_1}, \\ I_{1,2} &= re^{i\theta_2}, & I_{2,2} &= \rho e^{i\varphi_2}, \end{aligned} \tag{31}$$

where r and ρ are real and positive, and the phases are at our disposal. Equations (6) and (7) give

$$\begin{aligned} E_{2,1} &= r(\Gamma_{11}e^{i\theta_1} + \Gamma_{12}e^{i\theta_2}), \\ E_{2,2} &= r(\Gamma_{21}e^{i\theta_1} + \Gamma_{22}e^{i\theta_2}), \\ E_{1,1} &= \rho(\Gamma_{11}e^{i\varphi_1} + \Gamma_{21}e^{i\varphi_2}), \\ E_{1,2} &= \rho(\Gamma_{12}e^{i\varphi_1} + \Gamma_{22}e^{i\varphi_2}). \end{aligned} \tag{32}$$

The phase reversal condition leads to the following pair of simultaneous equations:

$$\begin{aligned} \theta_1 - \theta_2 &= \arg E_{1,2} - \arg E_{1,1} = \arg \frac{\Gamma_{12}e^{i(\varphi_1-\varphi_2)} + \Gamma_{22}}{\Gamma_{11}e^{i(\varphi_1-\varphi_2)} + \Gamma_{21}} \\ \varphi_1 - \varphi_2 &= \arg E_{2,2} - \arg E_{2,1} = \arg \frac{\Gamma_{21}e^{i(\theta_1-\theta_2)} + \Gamma_{22}}{\Gamma_{11}e^{i(\theta_1-\theta_2)} + \Gamma_{12}}. \end{aligned} \tag{33}$$

On the other hand, the condition (30) for maximum power transfer reduces to

$$\theta_1 - \theta_2 = \arg (\mathbf{\Gamma}^\dagger\mathbf{\Gamma})_{12} = \arg (\Gamma_{11}^*\Gamma_{12} + \Gamma_{21}^*\Gamma_{22}). \tag{34}$$

Since $\mathbf{\Gamma}$ is an essentially arbitrary complex matrix, equations (33) are not equivalent to (34), although it is possible that in practical cases the two conditions will yield values of $\theta_1 - \theta_2$ which do not differ by very much.

IV. DYNAMIC BEHAVIOR OF INTERACTING ADAPTIVE ARRAYS

In this section we set up a simple model of the dynamic behavior of two adaptive arrays, each of which continuously adjusts the excitations of its own elements in response to the fields from the other array. In principle the same equations would apply to a single array interacting with itself, as a combined radar transmitter and receiver. The fundamental assumption is that the amplitudes and phases of the element currents vary so slowly, compared with the transmission time between the arrays, that a single-frequency analysis is valid.

Since in this model the excitation of an adaptive array is indeterminate

in the absence of an external field, we have to use an auxiliary antenna or beacon to turn the system on. The steps are as follows: First the beacon is turned on, illuminating at least Array 1. Then Array 1 is turned on, Array 2 is turned on, and the beacon is turned off, leaving Arrays 1 and 2 to interact only with each other. It is convenient to assume that when a transmitter is switched on or off, its radiated power changes continuously, during a finite time interval, from one steady-state value to another.

First we consider arrays in which the excitation of each element is proportional to the complex conjugate of the field incident on that element, and the total radiated power is a prescribed function of time. Thus let $\mathbf{I}_1(t)$ and $\mathbf{I}_2(t)$ be the (slowly varying) complex excitations of the two arrays. We assume that the dynamic behavior of the arrays is described by the following equations:

$$I_{1,j}(t) = \mu_1(t)e^{i\vartheta_1}\left[\sum_{k=1}^{N}\Gamma_{kj}{}^{*}I_{2,k}{}^{*}(t - \tau_{kj} - \tau_1) + B_{1,j}{}^{*}(t - \tau_1)\right], \quad j = 1, 2, \cdots, M; \tag{35}$$

$$I_{2,k}(t) = \mu_2(t)e^{i\vartheta_2}\left[\sum_{j=1}^{M}\Gamma_{kj}{}^{*}I_{1,j}{}^{*}(t - \tau_{kj} - \tau_2) + B_{2,k}{}^{*}(t - \tau_2)\right], \quad k = 1, 2, \cdots, N. \tag{36}$$

In these equations, $\mathbf{B}_1(t)$ and $\mathbf{B}_2(t)$ are the beacon fields, if any, at Arrays 1 and 2, τ_{kj} is the transmission delay between the jth element of Array 1 and the kth element of Array 2, τ_1 and τ_2 are constant time delays in the amplifiers of Arrays 1 and 2, ϑ_1 and ϑ_2 are constant phase shifts, and $\mu_1(t)$ and $\mu_2(t)$ are real normalization factors determined by

$$\tfrac{1}{2}(\mathbf{Z}_1\mathbf{I}_1(t),\mathbf{I}_1(t)) = P_1(t), \tag{37}$$

$$\tfrac{1}{2}(\mathbf{Z}_2\mathbf{I}_2(t),\mathbf{I}_2(t)) = P_2(t), \tag{38}$$

where the radiated powers $P_1(t)$ and $P_2(t)$ are given functions of time.

Similarly, the equations describing two arrays in which the excitation amplitudes $|I_{1,j}(t)|$ and $|I_{2,k}(t)|$ are prescribed functions of time, while the phases are continuously adjusted to satisfy the phase reversal condition, are as follows:

$$\arg I_{1,j}(t) = \vartheta_1 - \arg\left[\sum_{k=1}^{N}\Gamma_{kj}I_{2,k}(t - \tau_{kj} - \tau_1) + B_{1,j}(t - \tau_1)\right], \quad j = 1, 2, \cdots, M; \tag{39}$$

$$\arg I_{2,k}(t) = \vartheta_2 - \arg\left[\sum_{j=1}^{M} \Gamma_{kj} I_{1,j}(t - \tau_{kj} - \tau_2) + B_{2,k}(t - \tau_2)\right], \quad k = 1, 2, \cdots, N. \tag{40}$$

Before undertaking numerical simulations of the dynamic behavior of adaptive arrays, we consider an example which can be handled analytically, namely the special case of two power-limited arrays in which all the interelement delay times are equal. We obtain this case from (35) and (36) by setting

$$\tau_{kj} = \tau_3 = \text{constant}. \tag{41}$$

If we assume for simplicity that the beacon has been turned off, (35) and (36) take the form

$$\mathbf{I}_1(t) = \mu_1(t) e^{i\vartheta_1} \mathbf{\Gamma}^\dagger \mathbf{I}_2^*(t - \tau_3 - \tau_1), \tag{42}$$

$$\mathbf{I}_2(t) = \mu_2(t) e^{i\vartheta_2} \mathbf{\Gamma}^* \mathbf{I}_1^*(t - \tau_3 - \tau_2). \tag{43}$$

Eliminating $\mathbf{I}_2$ yields

$$\mathbf{I}_1(t) = \mu_1(t) e^{i\vartheta} \mathbf{\Gamma}^\dagger \mathbf{\Gamma} \mathbf{I}_1(t - \tau), \tag{44}$$

where

$$\vartheta = \vartheta_1 - \vartheta_2, \qquad \tau = \tau_1 + \tau_2 + 2\tau_3, \tag{45}$$

and the normalizing factor $\mu_1(t)$ may be expressed, if needed, in terms of the radiated power $P_1(t)$ by (37).

The Hermitian matrix $\mathbf{\Gamma}^\dagger \mathbf{\Gamma}$ is at least positive semidefinite and will have M real eigenvalues. We suppose that the eigenvalues are numbered in order of increasing size and that the largest eigenvalue is *unique;* that is,

$$0 \leqq \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_{M-1} < \lambda_M. \tag{46}$$

The corresponding eigenvectors $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots, \mathbf{z}^{(M)}$ satisfy

$$\mathbf{\Gamma}^\dagger \mathbf{\Gamma} \mathbf{z}^{(i)} = \lambda_i \mathbf{z}^{(i)}, \qquad i = 1, 2, \cdots, M, \tag{47}$$

and may be taken as orthonormal, i.e.,

$$(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}) = \delta_{ij}. \tag{48}$$

Let $\mathbf{I}_1(t)$ be expanded in terms of the $\mathbf{z}^{(i)}$, with coefficients depending, of course, on time; thus

$$\mathbf{I}_1(t) = \sum_{i=1}^{M} c_i(t) \mathbf{z}^{(i)}. \tag{49}$$

Repeated application of (44) and (47) gives

$$\mathbf{I}_1(t + n\tau) = N_n(t)e^{in\vartheta} \sum_{j=1}^{M} c_j(t)\lambda_j{}^n\mathbf{z}^{(j)}, \tag{50}$$

where $N_n(t)$ is again a normalization factor chosen to satisfy (37).

Now suppose that there is an interval of length τ, say $t_0 \leqq t < t_0 + \tau$, in which all the $c_i(t)$ are bounded and $c_M(t)$ is bounded away from zero. It follows from (46) that the term in $\lambda_M{}^n$ in (50) will eventually dominate all the others, and we shall have

$$\mathbf{I}_1(t + n\tau) \underset{n\to\infty}{\rightarrow} \frac{[2P_1(t)]^{\frac{1}{2}} \exp\,[in\vartheta + \arg\, c_M(t)]}{(\mathbf{Z}_1\mathbf{z}^{(M)},\mathbf{z}^{(M)})^{\frac{1}{2}}}\,\mathbf{z}^{(M)}, \tag{51}$$

$$\text{for} \quad t_0 \leqq t < t_0 + \tau.$$

It is easy to verify that the phase of $\mathbf{I}_1(t)$, as given by (51), is continuous at $t = t_0 + (n + 1)\tau$ if the phase of $c_M(t)$ is continuous at $t = t_0 + \tau$.

We have just proved that two power-limited adaptive arrays with equal interelement delays will reach an equilibrium state in which the excitation of Array 1 is proportional to the eigenvector belonging to the largest eigenvalue of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$. Similarly, the equilibrium excitation of Array 2 is proportional to the eigenvector belonging to the largest eigenvalue of $\mathbf{\Gamma}\mathbf{\Gamma}^\dagger$ (again λ_M). But it was shown in Section II that the eigenvector belonging to the largest eigenvalue of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ corresponds to maximum power transfer when $\mathbf{Z}_1$ and $\mathbf{Y}_2$ are multiples of the unit matrix; that is, when all the mutual impedances and admittances are zero and all the self-impedances and self-admittances are equal. If this condition is approximately satisfied, as will often be the case in practice, then the equilibrium excitation should be nearly the same as the excitation for optimal power transfer.

We observe that the equilibrium excitation is unique except in the pathological case where $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ has two or more equal eigenvalues which are larger than all the rest. Steady states in which the current distribution corresponds to one of the smaller eigenvalues of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ are mathematically possible, but are unstable. Also, if the arrays are moved with respect to each other or if the transmission medium changes (either case would correspond to changing the Green's function matrix), the final equilibrium state depends only on the final positions of the arrays and the properties of the transmission medium, and not at all on how the situation was reached.

It should be pointed out that the foregoing argument does not apply, at least in its present form, to fixed-amplitude arrays with only phases

adjustable. Clearly there are extreme cases in which the behavior of fixed-amplitude arrays will be qualitatively different from that of power-limited arrays. For example, if Γ is diagonal, so that each element of Array 1 is coupled to only one element of Array 2, then power-limited arrays will ultimately cut out all of the elements except for the pair which is most closely coupled; but the elements of fixed-amplitude arrays will go on indefinitely talking to each other in pairs, with no particular phase relationship between the elements of different pairs. Nevertheless, the numerical simulations of the next section indicate that in typical cases fixed-amplitude arrays do settle down to a steady state about as quickly as power-limited ones. As yet, however, no mathematical theorem has been proved about the steady-state behavior of fixed-amplitude arrays.

V. NUMERICAL SIMULATIONS

Because the equations describing the dynamic behavior of interacting adaptive arrays generally do not lend themselves to analytic treatment, we have made a few numerical simulations of 2- and 3-element arrays on an IBM 7094, in order to get some feeling for the possible behavior of interacting adaptive arrays in practice. Since these simulations were only computational experiments, no physical significance is to be attached to the specific numerical results.

We shall first describe the method of simulation, then show the outcome of a typical calculation, and finally summarize the results of the whole study.

For each simulation we selected the elements of the 2×2 or 3×3 matrix Γ according to the following scheme: We set

$$\Gamma_{jk} = G_{jk} \exp\,(-i\pi n_{jk}/5), \tag{52}$$

where G_{jk} was a random number selected with equal probability from the set $\{1,2,4,8\}$, and n_{jk} was selected with equal probability from the set $\{-5, -4, \cdots, 5\}$. For the interelement delay times we took

$$\tau_{jk} = 20 + n_{jk}\,. \tag{53}$$

As a matter of interest, we also computed the Hermitian matrix $\Gamma^{\dagger}\Gamma$ and its eigenvalues and eigenvectors.

If the time delays are all commensurable, (35) and (36) or (39) and (40) can easily be solved recursively on a digital computer. To start the system off, Array 1 was supposed to be illuminated initially by a constant beacon field. Array 2 was turned on linearly during a period of 20 time units, and the beacon was turned off linearly during a similar period. Four cases were run with each choice of Γ.

Case I. Power limited, equal delays. The condition

$$(\mathbf{I}_1 , \mathbf{I}_1) = (\mathbf{I}_2 , \mathbf{I}_2) = 1 \tag{54}$$

was imposed, and all interelement delays were set equal to 20 units. This case must approach a steady state, according to Section IV, provided only that the largest eigenvalue of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ is unique.

Case II. Power limited, unequal delays. Same as Case I, except that the time delays given by (53) were used.

Case III. Fixed amplitudes, equal delays. The condition

$$| I_{1,j} | = | I_{2,k} | = 1 \tag{55}$$

was imposed, and all interelement delays were set equal to 20 units.

Case IV. Fixed amplitudes, unequal delays. Same as Case III, except that the time delays given by (53) were used.

In a typical run, the random number generator produced:

$$(n_{jk}) = \begin{pmatrix} -1 & -1 & -2 \\ 0 & -5 & -5 \\ -1 & 1 & 3 \end{pmatrix}, \tag{56}$$

$$(\Gamma_{jk}) = \begin{pmatrix} 8/\underline{36^\circ} & 2/\underline{36^\circ} & 8/\underline{72^\circ} \\ 2/\underline{0^\circ} & 8/\underline{180^\circ} & 1/\underline{180^\circ} \\ 8/\underline{36^\circ} & 8/\underline{-36^\circ} & 2/\underline{-108^\circ} \end{pmatrix}. \tag{57}$$

It follows that

$$\mathbf{\Gamma}^\dagger\mathbf{\Gamma} = \begin{pmatrix} 132.0/\underline{0^\circ} & 64.0/\underline{-72.0^\circ} & 46.4/\underline{37.5^\circ} \\ 64.0/\underline{72.0^\circ} & 132.0/\underline{0^\circ} & 26.5/\underline{-12.7^\circ} \\ 46.4/\underline{-37.5^\circ} & 26.5/\underline{12.7^\circ} & 69.0/\underline{0^\circ} \end{pmatrix}; \tag{58}$$

$$\lambda_1 = 20.6, \qquad \lambda_2 = 109.8, \qquad \lambda_3 = 202.6. \tag{59}$$

The eigenvector corresponding to λ_3 is

$$\mathbf{z}^{(3)} = (0.719/\underline{0^\circ}, 0.656/\underline{64.5^\circ}, 0.229/\underline{-6.2^\circ}). \tag{60}$$

Figs. 2 through 5 show the results of running Cases I through IV over the time interval $0 \leqq t \leqq 400$. In the figures the phases are referred to the phase of the first element of each array, and the following notation is used:

$$\mathbf{I}_1 = (A_1, A_2/\underline{\alpha_2}, A_3/\underline{\alpha_3}), \tag{61}$$

$$\mathbf{I}_2 = (B_2, B_2/\underline{\beta_2}, B_3/\underline{\beta_3}). \tag{62}$$

The initial behavior of the two arrays depends on the particular way in which they were turned on and is of no great importance; what we are really interested in is the behavior at large times. In Cases I and III (Figs. 2 and 4), under the assumption of equal interelement delay times, the system appears to settle down to a perfectly steady state. It is easy to verify that in Case I the steady-state excitation of Array 1 corresponds to the eigenvector $\mathbf{z}^{(3)}$ given by (60). On the other hand, in Cases II and IV (Figs. 3 and 5), where the delays are not all equal, the array excitations continue to show small residual fluctuations about the steady-state solutions of Cases I and III. These fluctuations are quite apparent in the original plots from which the present figures were redrawn.

In the numerical study, 50 pairs of 2-element arrays were simulated and four cases run for each pair. The ratio of eigenvalues λ_2/λ_1 of $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ ranged from 293.5 to 1.385. As expected, the larger values of λ_2/λ_1 generally produced quicker convergence; but only one case, out of all those tried, failed to reach essentially steady values by $t = 400$. In this particular example λ_2/λ_1 was 7.37, and the interelement delays happened to range all the way from 15 to 25. Cases I, II, and III settled down relatively quickly, but Case IV (fixed amplitudes, unequal delays) went into a large-amplitude oscillation which was obviously not dying out at $t = 1000$. A similar, subsequent run in which the extreme interelement delay times were changed to 16 and 24 settled down normally.

Twenty-five pairs of 3-element arrays were simulated, with eigenvalue ratios λ_3/λ_2 ranging from 26.02 to 1.458. Every one of these cases appeared to have reached an essentially steady state at $t = 400$. The example shown in Figs. 2 through 5 is entirely typical.

From the numerical simulations it is clear that sufficiently large delay differences (perhaps ± 25 per cent of the average delay time) can make a pair of interacting adaptive arrays fail to settle down. We conjecture, however, that the arrays will always reach an essentially steady state if the delay differences are a sufficiently small fraction of the average delay. Conceivably one could put bounds on the fluctuations as a function of the deviations of the delays from the mean delay, but a more practical approach might be to do some experiments with real adaptive arrays.

VI. ACKNOWLEDGMENTS

I am indebted to W. C. Jakes and C. C. Cutler for bringing this problem to my attention, and for stimulating discussions. I. W. Sandberg made significant contributions to the analysis of Section II. My thanks go to Mrs. Marie Dolan for all of the numerical simulations.

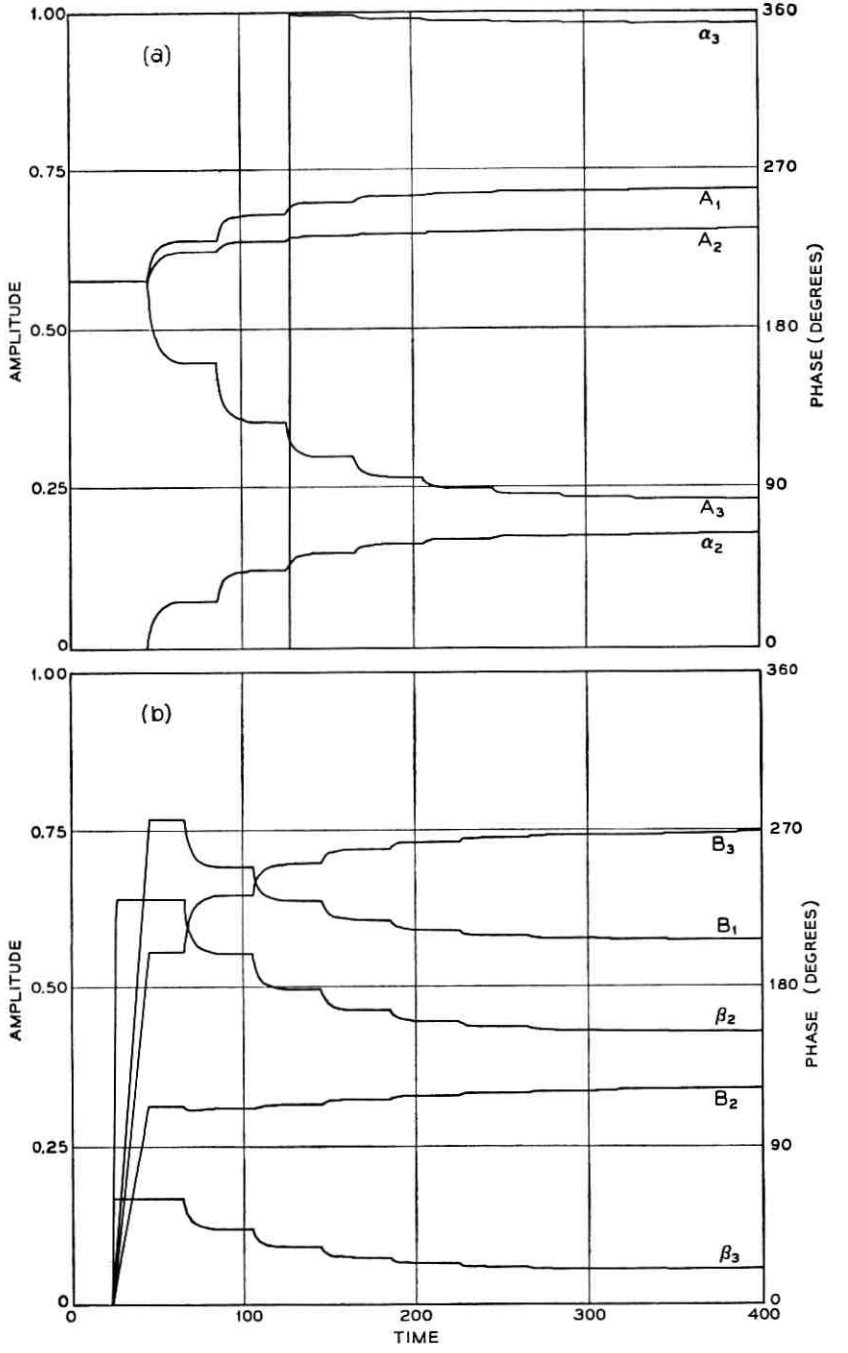


Fig. 2 — Transient behavior of amplitudes and phases in power-limited adaptive arrays with equal interelement delays: (a) Array 1, (b) Array 2.

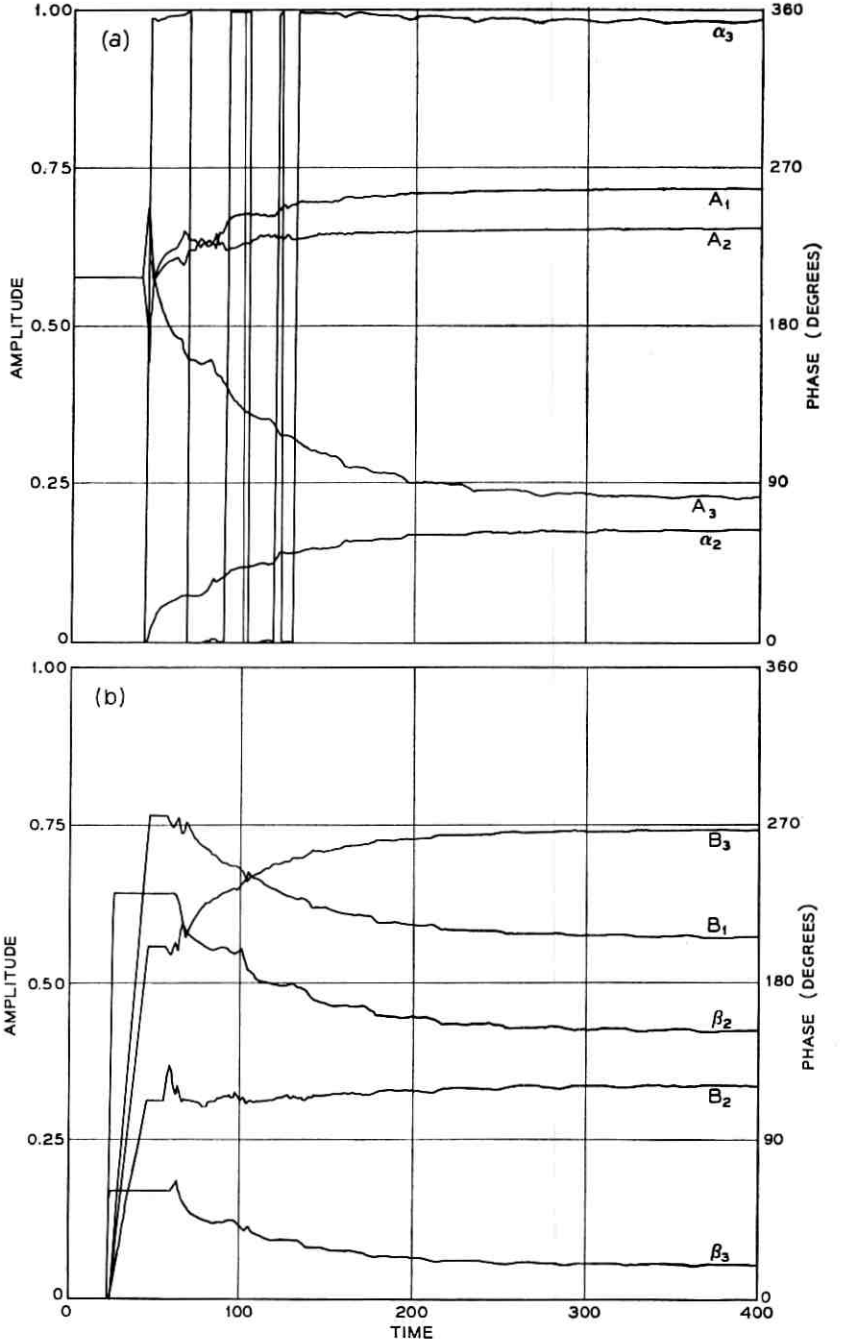


Fig. 3 — Transient behavior of amplitudes and phases in power-limited adaptive arrays with unequal interelement delays: (a) Array 1, (b) Array 2.

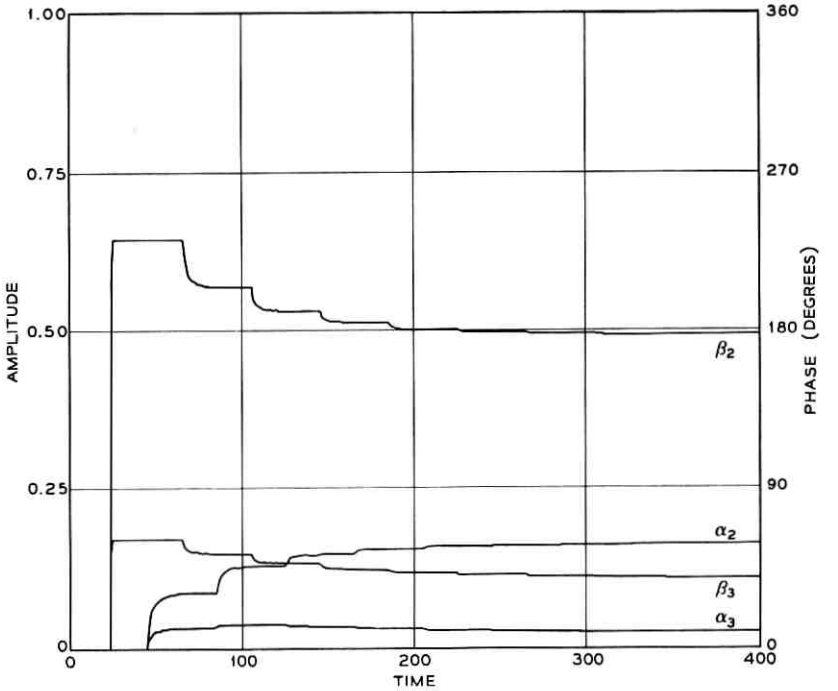


Fig. 4 — Transient behavior of phases in fixed-amplitude adaptive arrays with equal interelement delays.

APPENDIX A

Vectors, Matrices, and Hermitian Forms

We summarize here the notation used in this paper, as well as some properties of Hermitian forms which are proved in textbooks like that of Gantmacher.[4]

A *vector* in n-dimensional complex space is an ordered array of n complex numbers:

$$\mathbf{x} = (x_1, x_2, \cdots, x_n). \tag{63}$$

The *scalar product* of the vectors $\mathbf{x}$ and $\mathbf{y}$ is written $(\mathbf{x},\mathbf{y})$ and is defined by

$$(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} x_i y_i^*, \tag{64}$$

where the asterisk denotes complex conjugate.

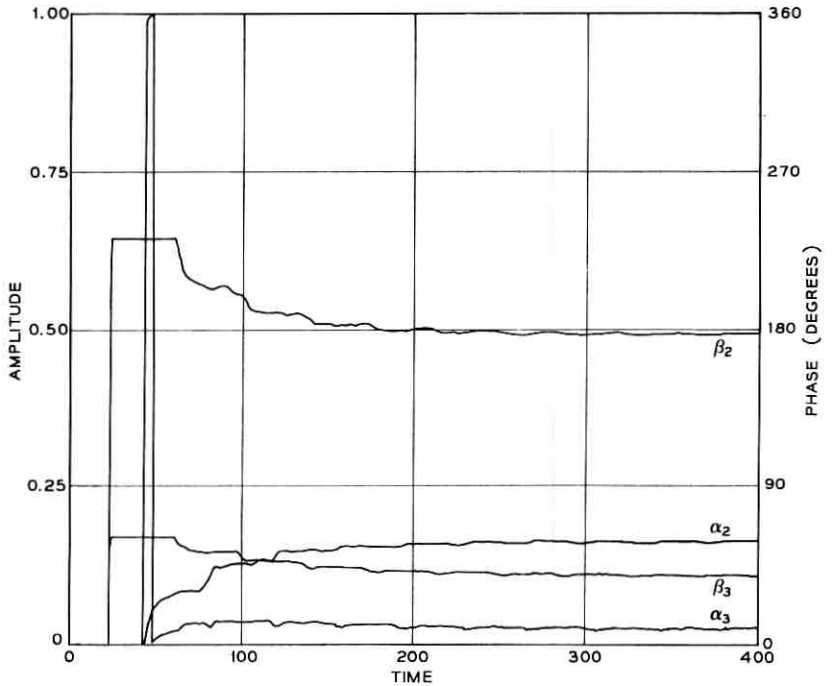


Fig. 5 — Transient behavior of phases in fixed-amplitude adaptive arrays with unequal interelement delays.

A *matrix* is an $m \times n$ array of complex numbers:

$$\mathbf{A} = (A_{ij}), \qquad i = 1, 2, \cdots, m; \qquad j = 1, 2, \cdots, n. \tag{65}$$

Associated with a given matrix are the following matrices:

$$\begin{aligned} &\textit{Conjugate} && (\mathbf{A}^*)_{ij} = A_{ij}^*, \\ &\textit{Transpose} && (\mathbf{A}')_{ij} = A_{ji}, \\ &\textit{Adjoint} && (\mathbf{A}^\dagger)_{ij} = A_{ji}^*. \end{aligned} \tag{66}$$

Note that this definition of the adjoint, while in accord with modern usage, differs from the definitions given in some older textbooks.

A *Hermitian matrix* is one which is equal to its own adjoint:

$$\mathbf{H} = \mathbf{H}^\dagger \quad \text{or} \quad H_{ij} = H_{ji}^*. \tag{67}$$

The *product* of an $m \times n$ matrix and an n-component vector is an m-component vector, written

$$\mathbf{y} = \mathbf{Ax} \quad \text{or} \quad y_i = \sum_{j=1}^{n} A_{ij}x_j, \qquad i = 1, 2, \cdots, m. \tag{68}$$

If $\mathbf{x}$ is an m-component vector, $\mathbf{y}$ an n-component vector, and $\mathbf{A}$ an $m \times n$ matrix, then from (64) and (66),

$$(\mathbf{x},\mathbf{Ay}) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i A_{ij}{}^* y_j{}^* = (\mathbf{A}^\dagger\mathbf{x},\mathbf{y}). \tag{69}$$

A *Hermitian form* is the scalar product of $\mathbf{Hx}$ with $\mathbf{x}$, where $\mathbf{H}$ is a Hermitian matrix:

$$(\mathbf{Hx},\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i{}^* H_{ij} x_j . \tag{70}$$

Hermitian forms are *real-valued*, since in view of (67),

$$(\mathbf{Hx},\mathbf{x})^* = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i H_{ij}{}^* x_j{}^* = \sum_{i=1}^{n}\sum_{j=1}^{n} x_j{}^* H_{ji} x_i = (\mathbf{Hx},\mathbf{x}). \tag{71}$$

A Hermitian form is *positive definite* if

$$(\mathbf{Hx},\mathbf{x}) > 0 \quad \text{whenever} \quad (\mathbf{x},\mathbf{x}) \neq 0. \tag{72}$$

If the $>$ sign is replaced by $\geqq$, the form is called *positive semidefinite*.

The *product* of an $m \times n$ matrix $\mathbf{A}$ and an $n \times p$ matrix $\mathbf{B}$ is an $m \times p$ matrix $\mathbf{C}$ whose elements are given by

$$C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}, \qquad i = 1, 2, \cdots, m; \qquad j = 1, 2, \cdots, p. \tag{73}$$

A square matrix whose determinant vanishes is called *singular*. If the determinant does not vanish, the matrix is called *nonsingular*. The matrix of a positive definite Hermitian form is nonsingular.

The *inverse* of a nonsingular $n \times n$ square matrix $\mathbf{A}$ is the $n \times n$ square matrix $\mathbf{A}^{-1}$ which satisfies

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{1}_n , \tag{74}$$

where $\mathbf{1}_n$ is the $n \times n$ *unit matrix* with 1's on the main diagonal and 0's elsewhere. The elements of $\mathbf{A}^{-1}$ are given by

$$(\mathbf{A}^{-1})_{ij} = \frac{\mathfrak{a}_{ji}}{\det \mathbf{A}}, \tag{75}$$

where $\mathfrak{a}_{ji}$ is the cofactor of the element A_{ji} in the determinant of $\mathbf{A}$.

If $\mathbf{\Gamma}$ is an arbitrary $m \times n$ matrix, $\mathbf{\Gamma}^\dagger\mathbf{\Gamma}$ is an $n \times n$ Hermitian matrix, since

$$(\mathbf{\Gamma}^\dagger\mathbf{\Gamma})_{ij} = \sum_{k=1}^{m} \Gamma_{ki}^*\Gamma_{kj} = \sum_{k=1}^{m} \Gamma_{kj}\Gamma_{ki}^* = (\mathbf{\Gamma}^\dagger\mathbf{\Gamma})_{ji}^*, \qquad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, n. \tag{76}$$

If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ Hermitian matrices and λ is a complex parameter, then $\mathbf{A} - \lambda\mathbf{B}$ is called a *pencil* of matrices. If $\mathbf{B}$ is positive definite, the pencil is called *regular*. The characteristic equation of a regular pencil, namely

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0, \tag{77}$$

always has n real roots $\lambda_1, \lambda_2, \cdots, \lambda_n$, which are called the *eigenvalues* of the pencil. The eigenvalues correspond to *eigenvectors* $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots, \mathbf{z}^{(n)}$, which satisfy the homogeneous equations

$$\mathbf{A}\mathbf{z}^{(k)} = \lambda_k\mathbf{B}\mathbf{z}^{(k)}, \qquad k = 1, 2, \cdots, n. \tag{78}$$

The eigenvectors may be chosen to satisfy

$$(\mathbf{B}\mathbf{z}^{(i)}, \mathbf{z}^{(j)}) = \delta_{ij}, \tag{79}$$

where δ_{ij} is the Kronecker delta.

The largest eigenvalue λ_n of the regular pencil $\mathbf{A} - \lambda\mathbf{B}$ satisfies

$$\lambda_n = \max_{\mathbf{x}\neq 0} \frac{(\mathbf{A}\mathbf{x},\mathbf{x})}{(\mathbf{B}\mathbf{x},\mathbf{x})}, \tag{80}$$

and this maximum is assumed only for eigenvectors of the pencil corresponding to the eigenvalue λ_n.

APPENDIX B

Reciprocity Theorem for Time-Harmonic Fields

We shall prove the reciprocity theorem in a form convenient for use in the present paper, following an approach similar to that of Harrington.[5]

Consider a linear, time-invariant medium characterized by the permittivity tensor $\boldsymbol{\varepsilon}$, the permeability tensor $\boldsymbol{\mu}$, and the conductivity tensor $\boldsymbol{\sigma}$. All three tensors are assumed to be symmetric, although they may be functions of the space coordinates. Let

$$\mathcal{Y} = \boldsymbol{\sigma} + i\omega\boldsymbol{\varepsilon}, \qquad \mathcal{Z} = i\omega\boldsymbol{\mu}, \tag{81}$$

where ω is the angular frequency of the time-harmonic fields.

Consider two sets of electric current densities, $\mathbf{J}^a$ and $\mathbf{J}^b$, which are

vector functions of position and which give rise to the fields $\mathbf{E}^a$, $\mathbf{H}^a$ and $\mathbf{E}^b$, $\mathbf{H}^b$ respectively. Maxwell's equations are

$$\begin{aligned} \nabla \times \mathbf{H}^a &= \mathcal{Y}\mathbf{E}^a + \mathbf{J}^a, & \nabla \times \mathbf{H}^b &= \mathcal{Y}\mathbf{E}^b + \mathbf{J}^b, \\ -\nabla \times \mathbf{E}^a &= \mathcal{Z}\mathbf{H}^a, & -\nabla \times \mathbf{E}^b &= \mathcal{Z}\mathbf{H}^b. \end{aligned} \tag{82}$$

From the first and fourth equations,

$$\begin{aligned} -\nabla\cdot(\mathbf{E}^b \times \mathbf{H}^a) &= \mathbf{E}^b\cdot\nabla \times \mathbf{H}^a - \mathbf{H}^a\cdot\nabla \times \mathbf{E}^b \\ &= \mathbf{E}^b\cdot\mathcal{Y}\mathbf{E}^a + \mathbf{E}^b\cdot\mathbf{J}^a + \mathbf{H}^a\cdot\mathcal{Z}\mathbf{H}^b, \end{aligned} \tag{83}$$

and from the second and third,

$$\begin{aligned} -\nabla\cdot(\mathbf{E}^a \times \mathbf{H}^b) &= \mathbf{E}^a\cdot\nabla \times \mathbf{H}^b - \mathbf{H}^b\cdot\nabla \times \mathbf{E}^a \\ &= \mathbf{E}^a\cdot\mathcal{Y}\mathbf{E}^b + \mathbf{E}^a\cdot\mathbf{J}^b + \mathbf{H}^b\cdot\mathcal{Z}\mathbf{H}^a. \end{aligned} \tag{84}$$

Subtracting (83) from (84) and using the symmetry of $\mathcal{Y}$ and $\mathcal{Z}$, we obtain

$$\nabla\cdot(\mathbf{E}^b \times \mathbf{H}^a - \mathbf{E}^a \times \mathbf{H}^b) = \mathbf{E}^a\cdot\mathbf{J}^b - \mathbf{E}^b\cdot\mathbf{J}^a. \tag{85}$$

Now integrate over a large spherical volume V bounded by the surface S, which contains all sources and matter in its interior. The divergence theorem yields

$$\int_S (\mathbf{E}^b \times \mathbf{H}^a - \mathbf{E}^a \times \mathbf{H}^b)\cdot\mathbf{n}\, dS = \int_V (\mathbf{E}^a\cdot\mathbf{J}^b - \mathbf{E}^b\cdot\mathbf{J}^a)\, dV, \tag{86}$$

where $\mathbf{n}$ is the outward normal to S. The individual fields fall off as $1/r$, where r is the radius of V, but for large r the leading terms satisfy

$$\mathbf{E}^a = \eta\mathbf{H}^a \times \mathbf{n}, \qquad \mathbf{E}^b = \eta\mathbf{H}^b \times \mathbf{n}, \tag{87}$$

where η is the characteristic impedance of free space.

Hence for the leading terms,

$$\begin{aligned} &\eta^{-1}[\mathbf{E}^b \times \mathbf{H}^a - \mathbf{E}^a \times \mathbf{H}^b]\cdot\mathbf{n} \\ &= [(\mathbf{H}^b \times \mathbf{n}) \times \mathbf{H}^a - (\mathbf{H}^a \times \mathbf{n}) \times \mathbf{H}^b]\cdot\mathbf{n} \\ &= [\mathbf{n}(\mathbf{H}^a\cdot\mathbf{H}^b) - \mathbf{H}^b(\mathbf{n}\cdot\mathbf{H}^a) - \mathbf{n}(\mathbf{H}^a\cdot\mathbf{H}^b) + \mathbf{H}^a(\mathbf{n}\cdot\mathbf{H}^b)]\cdot\mathbf{n} = 0. \end{aligned} \tag{88}$$

It follows that if all sources and matter are of finite extent, then

$$\int \mathbf{E}^a\cdot\mathbf{J}^b dV = \int \mathbf{E}^b\cdot\mathbf{J}^a dV, \tag{89}$$

where each integral is taken over the region in which the source currents

are different from zero. If the medium is not symmetric, the theorem is still true provided that $\mathbf{E}^b$ represents the field produced by $\mathbf{J}^b$ in the "transposed" medium; but this generalization is not very useful in the present context.

Now let $\mathbf{J}^a$ correspond to an electric dipole of unit moment in the direction $\mathbf{u}^a$ at the point P_a, and let $\mathbf{J}^b$ correspond to an electric dipole of unit moment in the direction $\mathbf{u}^b$ at P_b. Equation (89) takes the form

$$\mathbf{u}^b \cdot \mathbf{E}^a(P_b) = \mathbf{u}^a \cdot \mathbf{E}^b(P_a), \tag{90}$$

where the left side represents the components of electric field due to source A at the location and in the direction of source B, and the right side represents the component due to source B at the location and in the direction of source A. This is the desired reciprocity theorem.

APPENDIX C

Sandberg's Lemma

We reproduce Sandberg's proof[3] of the following result.

Lemma. If $\mathbf{A}$ and $\mathbf{B}$ respectively are $n \times m$ and $m \times n$ matrices, then

$$\det (1_n + \mathbf{AB}) = \det (1_m + \mathbf{BA}).$$

Proof. First consider the case in which $\mathbf{A}$ and $\mathbf{B}$ are square $p \times p$ matrices. Then, if $\mathbf{A}$ is nonsingular,

$$\begin{aligned} \det [1_p + \mathbf{AB}] &= \det [\mathbf{A}^{-1}(1_p + \mathbf{AB})\mathbf{A}] \\ &= \det [1_p + \mathbf{BA}]. \end{aligned} \tag{91}$$

If $\mathbf{A}$ is singular, it has a zero characteristic root, and hence there exists a positive number λ_0 such that $\mathbf{A} + \lambda 1_p$ is nonsingular for all real λ satisfying $0 < |\lambda| < \lambda_0$. Thus when $0 < |\lambda| < \lambda_0$,

$$\det [1_p + (\mathbf{A} + \lambda 1_p)\mathbf{B}] = \det [1_p + \mathbf{B}(\mathbf{A} + \lambda 1_p)]. \tag{92}$$

Both sides of (92) are polynomials in λ of degree at most p. Furthermore these polynomials must be identical since they agree throughout the real interval $(0,\lambda_0)$. Therefore (92) is valid when $\lambda = 0$.

Consider now the case in which $\mathbf{A}$ and $\mathbf{B}$ are not square. Let $p = m + n$,

$$\tilde{\mathbf{A}} = \begin{matrix} \\ n \\ m \end{matrix}\!\!\begin{matrix} \begin{matrix} m & n \end{matrix} \\ \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{matrix}\!\!\begin{matrix} \\ n \\ m \end{matrix}, \qquad \tilde{\mathbf{B}} = \begin{matrix} \\ m \\ n \end{matrix}\!\!\begin{matrix} \begin{matrix} n & m \end{matrix} \\ \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{matrix}\!\!\begin{matrix} \\ m \\ n \end{matrix}; \tag{93}$$

and let the symbol $\dotplus$ denote a direct sum of matrices. Observe that

$$\begin{aligned} \det [\mathbf{1}_p + \tilde{\mathbf{A}}\tilde{\mathbf{B}}] &= \det [(\mathbf{1}_n + \mathbf{AB}) \dotplus \mathbf{1}_m] = \det [\mathbf{1}_n + \mathbf{AB}], \\ \det [\mathbf{1}_p + \tilde{\mathbf{B}}\tilde{\mathbf{A}}] &= \det [(1_m + \mathbf{BA}) \dotplus \mathbf{1}_n] = \det [\mathbf{1}_m + \mathbf{BA}], \end{aligned} \tag{94}$$

which proves the lemma.

REFERENCES

1. Cutler, C. C., Kompfner, R., and Tillotson, L. C., A Self-Steering Array Repeater, B.S.T.J., **42,** Sept., 1963, pp. 2013–2032.
2. Special Issue on Active and Adaptive Arrays, IEEE Trans. on Antennas and Propagation, **AP-12,** Mar., 1964, pp. 140–233.
3. Sandberg, I. W., On the Theory of Linear Multi-Loop Feedback Systems, B.S.T.J., **42,** Mar., 1963, p. 379.
4. Gantmacher, F. R., *Theory of Matrices*, trans. K. A. Hirsch, Vol. I, Chelsea, New York, 1959. Gantmacher's notation for Hermitian forms differs slightly from that adopted here.
5. Harrington, R. F., *Time-Harmonic Electromagnetic Fields*, McGraw-Hill, New York, 1961, pp. 116–118.

Optimum Design of a Gravitationally Oriented Two-Body Satellite

By E. Y. YU

(Manuscript received July 15, 1964)

Optimum ranges of the inertia ratios, the spring constants, and the damping constants have been obtained for the design of a gravitationally oriented two-body satellite with satisfactory over-all damping performance. In the case of viscous damping, optimum damping constants can be simply chosen from diagrams of complex root loci. It is found impossible to convert the optimum viscous damping constants into optimum magnetic hysteresis damping constants, and the latter have to be obtained from computer solutions. The result of this optimization work makes possible a better design of the satellite with lighter attitude control weight, shorter rod lengths, and smaller earth-pointing error than previously reported in articles in the Bell System Technical Journal.

I. INTRODUCTION

The dynamics analysis by Fletcher, Rongved, and Yu[1] has shown that a two-body satellite will achieve an earth-pointing motion from an initial tumbling as a result of energy dissipation in the hinge joint through the relative motion between the two bodies. For a practical application, we shall consider the earth-pointing body to be like a dumbbell and the auxiliary body like a sheet, the two being connected to each other through a hinge mechanism of universal-joint type to allow a two-degree-of-freedom relative motion. When the satellite is in an earth-pointing motion, the auxiliary body is parallel to the local horizontal in its unstable orientation, and the two axes of relative motions are aligned with the roll and pitch direction (see Fig. 1). The over-all aspect of such a passive gravitational attitude control system has been studied by Paul, West, and Yu[2] employing the magnetic hysteresis damping mechanism. Certain designs of the satellite in terms of moments of inertia, spring constants, and damping constants have been given in both Refs. 1 and 2 for an altitude of 6000 nautical miles (nm). They

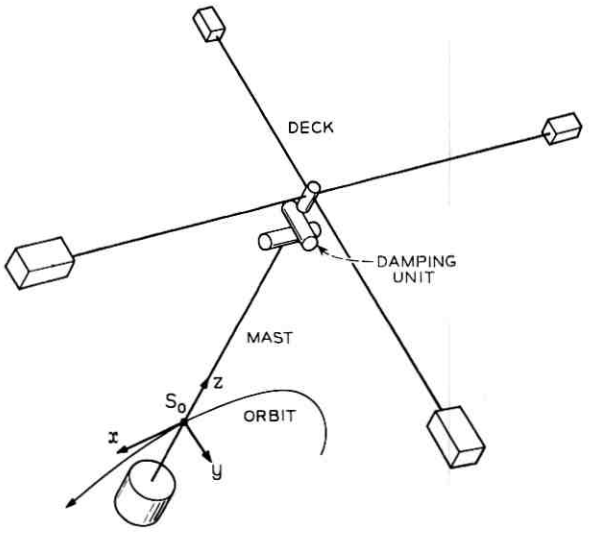


Fig. 1 — Two-body satellite configuration.

are found workable, but by no means "optimum," which means the achievement of the best over-all damping performance in tumbling motion, large-angle motion, torque-free librations, and forced librations under environmental torques of various frequencies.

The present work concerns a parameter optimization of the aforementioned two-body satellite employing two types of damping: the linear velocity of viscous type, and the nonlinear magnetic hysteresis type. With damping of the viscous type, analytical results can be obtained for the librational motions which may be used as a basis of parameter design in the case of magnetic hysteresis damping. The employment of magnetic hysteresis damping even makes the (linearized) equations of librational motions highly nonlinear, such that analytical treatment becomes intractable and results have to be obtained by numerical means.

II. GENERAL EQUATIONS OF MOTION

The equations of rotational motion which have been derived in Ref. 1 will be repeated here. The coordinate systems, as indicated in Fig. 2, are defined as follows. Let O-XYZ be a nonrotating frame, with its origin at the geocenter O, its Z-axis passing through the perigee of the orbit, and its Y-axis parallel to the orbital angular momentum vector.

Let S_o-xyz be an earth-pointing frame, with its origin S_o at the center of mass of the composite satellite, with z axis parallel to OS_o (the local vertical) making an angle ψ with OZ and with the y axis parallel to OY. The body coordinates of body 1, S_1-$x_1y_1z_1$, are defined along its principal axes, with adjusted[1] moments of inertia (I_1, I_2, I_3). Euler parameters (ξ, η, ζ, χ) are employed to describe the motion of S_1-$x_1y_1z_1$ relative to S_o-xyz. The matrix of transformation from the latter to the former frame is given as

$$[a_{ij}] = \begin{bmatrix} \xi^2 - \eta^2 - \zeta^2 + \chi^2 & 2(\xi\eta + \zeta\chi) & 2(\xi\zeta - \eta\chi) \\ 2(\xi\eta - \zeta\chi) & -\xi^2 + \eta^2 - \zeta^2 + \chi^2 & 2(\xi\chi + \eta\zeta) \\ 2(\xi\zeta + \eta\chi) & 2(-\xi\chi + \eta\zeta) & -\xi^2 - \eta^2 + \zeta^2 + \chi^2 \end{bmatrix}. \quad (1)$$

Among the Euler parameters the relation $\xi^2 + \eta^2 + \zeta^2 + \chi^2 = 1$ holds.

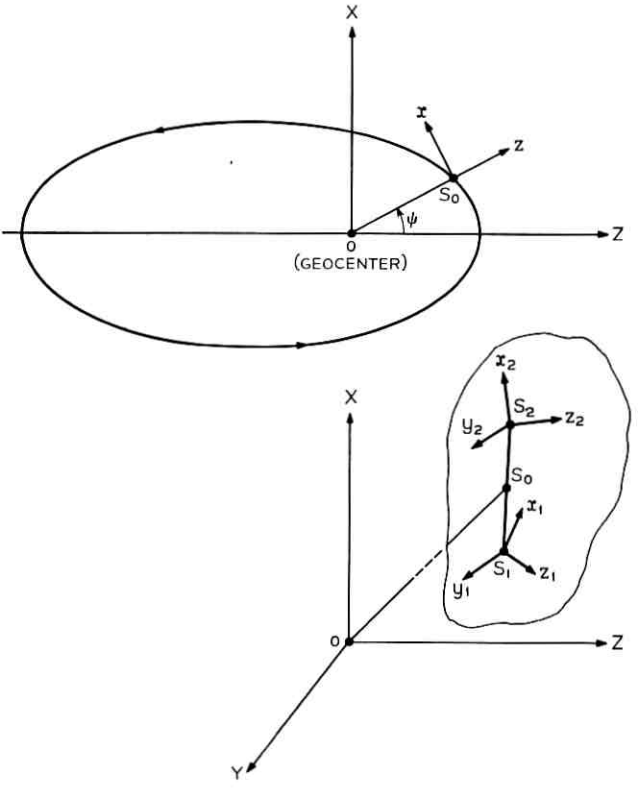


Fig. 2 — Coordinate systems.

The coordinates S_2-$x_2y_2z_2$ are defined along the principal axes of body 2, with moments of inertia (I_4, I_5, I_6). The rotation of body 2 relative to body 1 is specified by an angle, α, about the x_1 axis and then an angle, β, about the y_2 axis. The transformation matrix from S_1-$x_1y_1z_1$ to S_2-$x_2y_2z_2$ is

$$[b_{ij}] = \begin{bmatrix} C\beta & S\alpha S\beta & -C\alpha S\beta \\ 0 & C\alpha & S\alpha \\ S\beta & -S\alpha C\beta & C\alpha C\beta \end{bmatrix}, \tag{2}$$

where C and S are abbreviations of cosine and sine, respectively.

The equations of motion of the two-body satellite with the hinge joint situated at the center of mass of the auxiliary body and at the earth-pointing z_1 axis of the satellite body are:

$$I_1\dot{\omega}_1 = (I_2 - I_3)(\omega_2\omega_3 - Gn_2n_3) + T_{r1} + T_{d1}, \tag{3a}$$

$$I_2\dot{\omega}_2 = (I_3 - I_1)(\omega_3\omega_1 - Gn_3n_1) + (T_{r2} + T_{d2})C\alpha - T_cS\alpha, \tag{3b}$$

$$I_3\dot{\omega}_3 = (I_1 - I_2)(\omega_1\omega_2 - Gn_1n_2) + (T_{r2} + T_{d2})S\alpha + T_cC\alpha, \tag{3c}$$

$$I_4\dot{\omega}_4 = (I_5 - I_6)(\omega_5\omega_6 - Gn_5n_6) - (T_{r1} + T_{d1})C\beta + T_cS\beta, \tag{3d}$$

$$I_5\dot{\omega}_5 = (I_6 - I_4)(\omega_6\omega_4 - Gn_6n_4) - T_{r2} - T_{d2}, \tag{3e}$$

$$I_6\dot{\omega}_6 = (I_4 - I_5)(\omega_4\omega_5 - Gn_4n_5) - (T_{r1} + T_{d1})S\beta - T_cC\beta. \tag{3f}$$

In the above, $(\omega_1, \omega_2, \omega_3)$ are the components of the total angular velocity of body 1 along S_1-$x_1y_1z_1$, and $(\omega_4, \omega_5, \omega_6)$ are those of body 2 along S_2-$x_2y_2z_2$. The coefficient G involves orbital elements: i.e.,

$$G = 3\Omega^2(1 - \epsilon^2)^{-3}(1 + \epsilon C\psi)^3,$$

where ϵ = orbital eccentricity, and $\Omega = 2\pi$ divided by the orbital period. Also,

$$n_i = a_{i3}, \quad n_{i+3} = \sum_{k=1}^{3} b_{ik}a_{k3}, \qquad i = 1, 2, 3,$$

with a_{ij} and b_{ij} the elements of the transformation matrices (1) and (2). The constraint torque, T_c, is given as

$$
\begin{aligned}
T_c = \left(\frac{S^2\alpha}{I_2} + \frac{C^2\alpha}{I_3} + \frac{S^2\beta}{I_4} + \frac{C^2\beta}{I_6}\right)^{-1} \\
\cdot\Bigg\{\frac{S\alpha}{I_2}[(I_3 - I_1)(\omega_1\omega_3 - Gn_1n_3) + (T_{r2} + T_{d2})C\alpha] \\
- \frac{C\alpha}{I_3}[(I_1 - I_2)(\omega_1\omega_2 - Gn_1n_2) + (T_{r2} + T_{d2})S\alpha] \\
- \frac{S\beta}{I_4}[(I_5 - I_6)(\omega_5\omega_6 - Gn_5n_6) - (T_{r1} + T_{d1})C\beta] \\
+ \frac{C\beta}{I_6}[(I_4 - I_5)(\omega_4\omega_5 - Gn_4n_5) - (T_{r1} + T_{d1})S\beta] \\
+ \dot{\alpha}(\omega_2C\alpha + \omega_3S\alpha) - \dot{\beta}(\omega_4C\beta + \omega_6S\beta)\Bigg\}.
\end{aligned}
\tag{4}
$$

The restoring torques T_{r1} and T_{r2} acting on body 1 along the x_1 and y_2 axes, respectively, are linear with the relative angle of rotation: i.e., $T_{r1} = k_1\alpha$, $T_{r2} = k_2\beta$, where k_1 and k_2 are spring constants. The damping torques acting on body 1, T_{d1} and T_{d2}, along the x_1 and y_2 axes, respectively, are defined in the following. For viscous damping, $T_{d1} = C_1\dot{\alpha}$, and $T_{d2} = C_2\dot{\beta}$, where C_1 and C_2 are viscous damping coefficients. For magnetic hysteresis damping, the torque is dependent on the history of motion and is defined in regions I and III of Fig. 3 for T_{d2} as

$$
T_{d2} = T_{d2}^* + \bar{T}_{d2}\frac{\beta - \beta^*}{\bar{\beta}}, \tag{5}
$$

as long as $|T_{d2}| < \bar{T}_{d2}$, where $\bar{\beta}$, $\bar{T}_{d2}$ are constants, and β^*, T_{d2}^* are the values of β, T_{d2} when $\dot{\beta}$ last changed sign. After $|T_{d2}|$ reaches $\bar{T}_{d2}$ then T_{d2} remains at $\bar{T}_{d2}$ as long as $\dot{\beta}$ does not change sign, as represented in regions II and IV of Fig. 3. The magnitudes of $\bar{\beta}$ and $\bar{T}_{d2}$ will be given in Section IV, where the minor loops will be described. According to the major loop in Fig. 3, no energy dissipation will result if the amplitude of oscillation is less then $\bar{\beta}$. T_{d1} is defined by replacing β by α and subscript 2 by 1 in (5).

It can be shown that the Euler parameters and the relative angles of rotation are related to the ω_i's, $i = 1, \cdots, 6$, as follows:

$$
\dot{\xi} = \tfrac{1}{2}(\chi\lambda_1 - \zeta\lambda_2 + \eta\lambda_3), \tag{6a}
$$

$$
\dot{\eta} = \tfrac{1}{2}(\zeta\lambda_1 + \chi\lambda_2 - \xi\lambda_3), \tag{6b}
$$

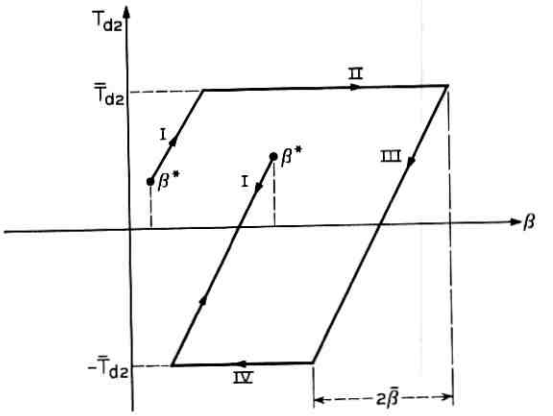


Fig. 3 — Magnetic hysteresis damping torque vs amplitude, $T_{d2} - \beta$.

$$\dot{\zeta} = \tfrac{1}{2}(-\eta\lambda_1 + \xi\lambda_2 + \chi\lambda_3), \tag{6c}$$

$$\dot{\chi} = \tfrac{1}{2}(-\xi\lambda_1 - \eta\lambda_2 - \zeta\lambda_3), \tag{6d}$$

$$\dot{\alpha} = -\omega_1 + \omega_4 C\beta + \omega_6 S\beta, \tag{6e}$$

$$\dot{\beta} = -\omega_2 C\alpha - \omega_3 S\alpha + \omega_5, \tag{6f}$$

where

$$\lambda_i \equiv \omega_i - a_{i2}\dot{\psi}, \qquad i = 1,2,3, \tag{7}$$

and

$$\dot{\psi} = \Omega(1 - \epsilon^2)^{-\frac{3}{2}}(1 + \epsilon C\psi)^2. \tag{8}$$

If the ω_i's in (3a) through (3f) are treated as dependent variables, then (3a) through (3f) together with (6a) through (6f) form a system of 12 first-order equations in the 12 unknowns ξ, η, ζ, χ, α, β and ω_i, $i = 1, \cdots, 6$. This system of equations has been programmed on an IBM 7090 for numerical solutions for any given initial conditions and for given dimensionless parameters: satellite moment of inertia ratios I_i/I_2, $i = 1,3, \cdots, 6$, spring constants $k_i/I_2\Omega^2$, damping constants $\bar{C}_i = C_i/I_2\Omega$ (viscous damping) or $\bar{T}_{di}/I_2\Omega^2$, $i = 1,2$, $\bar{\alpha},\bar{\beta}$ (hysteresis damping). Some numerical results have been given in Ref. 1, while more numerical results are summarized in Section IV for parameter optimization in the case of hysteresis damping.

III. PARAMETER OPTIMIZATION WITH VISCOUS DAMPING

In the case of librational motion, the S_1-$x_1y_1z_1$ and S_2-$x_2y_2z_2$ axes oscillate about the earth-pointing rotating frame S_o-xyz with infinitesimal angles ξ_1, η_1, ζ_1 and ξ_2, η_2, ζ_2, respectively. Because of the universal joint constraint, we have $\alpha = \xi_2 - \xi_1$, $\beta = \eta_2 - \eta_1$, and $\zeta_2 = \zeta_1$. The position vector of the hinge in the S_1-$x_1y_1z_1$ coordinate system is $-L\hat{z}_1$, while in the S_2-$x_2y_2z_2$ it is zero. Let us assume that the orbital eccentricity, ϵ, is small compared with unity: i.e., ϵ is assumed to be of the same order as the infinitesimal angles, ξ_1, η_1, ζ_1, α, and β. Hence, from (8) one obtains $\dot{\psi} = \Omega + 2\epsilon\Omega C\Omega t + 0(\epsilon^2)$, $\psi = \Omega t + 2\epsilon S\Omega t + 0(\epsilon^2)$, and $\ddot{\psi} = -2\epsilon\Omega^2 S\Omega t + 0(\epsilon^2)$. Equations (3a) through (3f) can then be linearized in the case of viscous damping to two sets of equations in pitch and roll-yaw librations. Upon transformation of the independent variable from t to $\psi = \Omega t$, these equations become:

$$\frac{d^2\eta_1}{d\psi^2} - \frac{1}{I_2} T_{d2} + 3p_1\eta_1 - \bar{k}_2\beta = 2\epsilon S\psi, \tag{9a}$$

$$\frac{d^2\beta}{d\psi^2} + \frac{d^2\eta_1}{d\psi^2} + \frac{1}{I_5} T_{d2} + (3p_2 + \lambda\bar{k}_2)\beta + 3p_2\eta_1 = 2\epsilon S\psi. \tag{9b}$$

$$\frac{d^2\xi_1}{d\psi^2} - \frac{1}{I_1} T_{d1} + (1 - q_1)\frac{d\zeta_1}{d\psi} + 4q_1\xi_1 - \bar{k}_1\alpha = 0, \tag{10a}$$

$$\frac{d^2\alpha}{d\psi^2} + \frac{d^2\xi_1}{d\psi^2} + \frac{1}{I_4} T_{d1} + (1 - q_2)\frac{d\zeta_1}{d\psi} + (4q_2 + \mu\bar{k}_1)\alpha + 4q_2\xi_1 = 0, \tag{10b}$$

$$\frac{d^2\zeta_1}{d\psi^2} + (1 - f_1 - f_2)\zeta_1 - (f_1 + f_2)\frac{d\xi_1}{d\psi} - f_2\frac{d\alpha}{d\psi} = 0, \tag{10c}$$

where

$$T_{d1} = C_1\Omega^{-1}\frac{d\alpha}{d\psi}, \qquad T_{d2} = C_2\Omega^{-1}\frac{d\beta}{d\psi},$$

$$\bar{k}_1 = k_1/I_1\Omega^2, \qquad \bar{k}_2 = k_2/I_2\Omega^2,$$

$$\lambda = I_2/I_5, \qquad \mu = I_1/I_4,$$

$$p_1 = (I_1 - I_3)/I_2, \qquad p_2 = (I_4 - I_6)/I_5,$$

$$q_1 = (I_2 - I_3)/I_1, \qquad q_2 = (I_5 - I_6)/I_4,$$

$$f_1 = (I_1 + I_3 - I_2)/(I_3 + I_6), \quad \text{and} \quad f_2 = (I_4 + I_6 - I_5)/(I_3 + I_6).$$

Note that (10c) is obtained by adding up the linearized versions of (3c) and (3f).

3.1 *Free Librational Motions*

For the free pitch libration, i.e., with $\epsilon = 0$ in (9a) and (9b), the stability criteria have been obtained in Ref. 1. To ensure that the earth-pointing frame, S_o-xyz, be the equilibrium position, it has been found that the spring constant should be larger than the following critical value:

$$\bar{k}_2^* = -\frac{3p_1p_2}{\lambda p_1 + p_2}. \tag{11a}$$

Thus, the pitch spring constant $\bar{k}_2$ is chosen as

$$\bar{k}_2 = a\bar{k}_2^* = -\frac{3ap_1p_2}{\lambda p_1 + p_2}, \tag{11b}$$

where $a > 1$. Similarly, the roll critical spring constant is found to be

$$\bar{k}_1^* = -\frac{4q_1q_2}{\mu q_1 + q_2}, \tag{12a}$$

and the roll spring constant is chosen to be

$$\bar{k}_1 = b\bar{k}_1^* = -\frac{4bq_1q_2}{\mu q_1 + q_2}, \tag{12b}$$

where $b > 1$. The characteristic equations in the complex variable s for the free pitch and roll-yaw librations are, respectively, of fourth order and sixth order:

$$\begin{aligned} s^4 + \bar{C}_2(1+\lambda)s^3 + [3(p_1+p_2) - 3a(\lambda+1)p_1p_2/(\lambda p_1 + p_2)]s^2 \\ + 3\bar{C}_2(\lambda p_1 + p_2)s + 9p_1p_2(1-a) = 0, \end{aligned} \tag{13}$$

and

$$\begin{aligned} s^6 + \bar{C}_1(1+\mu)s^5 + \left[\frac{(\mu-\lambda) + \lambda(1-q_2)^2 + \mu\lambda(1-q_1)^2}{\mu(\lambda+1) - \lambda(\mu q_1 + q_2)}\right. \\ \left. + \frac{4q_2[q_2 - \mu(b-1)q_1] + 4q_1[\mu q_1 - (b-1)q_2]}{\mu q_1 + q_2}\right] s^4 \\ + \bar{C}_1\left[\frac{(1+\mu)(\mu-\lambda) + \lambda(1-q_2)^2 + \mu\lambda(1-q_1)(1-q_2)}{\mu(\lambda+1) - \lambda(\mu q_1 + q_2)}\right. \end{aligned}$$

$$+ 4(\mu q_1 + q_2) + (1 - q_1)\mu\lambda \frac{\mu(1 - q_1) + (1 - q_2)}{\mu(\lambda + 1) - \lambda(\mu q_1 + q_2)}\Bigg] s^3$$

$$+ 4\Bigg[\frac{(\mu - \lambda)q_2[q_2 - \mu(b - 1)q_1] + q_1[\mu q_1 - (b - 1)q_2]\{(\mu - \lambda) + \lambda(1 - q_2)^2\}}{(\mu q_1 + q_2)\{\mu(\lambda + 1) - \lambda(\mu q_1 + q_2)\}} \tag{14}$$

$$- \frac{b\mu\lambda q_1 q_2(1 - q_1)(1 - q_2)}{(\mu q_1 + q_2)\{\mu(\lambda + 1) - \lambda(\mu q_1 + q_2)\}} - 4(b - 1)q_1 q_2$$

$$- (1 - q_1)\mu\lambda \frac{bq_1 q_2(1 - q_2) - q_2(1 - q_1)[q_2 - \mu(b - 1)q_1]}{(\mu q_1 + q_2)\{\mu(\lambda + 1) - \lambda(\mu q_1 + q_2)\}}\Bigg] s^2$$

$$+ 4\bar{C}_1 \frac{(\mu - \lambda)(\mu q_1 + q_2)}{\mu(\lambda + 1) - \lambda(\mu q_1 + q_2)} s + \frac{16 q_1 q_2(\mu - \lambda)(1 - b)}{\mu(\lambda + 1) - \lambda(\mu q_1 + q_2)} = 0.$$

In deriving (14) the satellite body has been assumed to be axisymmetric. The optimization problem for the free librational motions is such that by varying the parameters in the coefficients of (13) and (14) one gets the largest possible negative real part of the complex roots for the most slowly damped mode of librations. These parameters are $\bar{C}_2$, a, λ, p_1, and p_2 for the pitch motion, and $\bar{C}_1$, b, μ, q_1, and q_2 for the roll-yaw motion. The parameter λ contained in the roll-yaw characteristic equation is determined from the pitch optimization and is not treated as a varying parameter in the roll-yaw optimization. From the optimized values of the six parameters λ, μ, p_1, p_2, q_1, and q_2, one can determine the optimized inertia ratios I_i/I_2, $i = 1,3, \cdots, 6$. The value of I_2 is chosen from the magnitude of the gravitational torque at a given orbital altitude such that, in the presence of various environmental disturbing torques, the forced libration amplitude will not be larger than a specified value.

For a two-body satellite with geometric configuration of a dumbbell-sheet combination, we have

$$p_1 = 1, \quad p_2 = -1, \quad q_1 = 1, \quad \text{and} \quad q_2 = -1;$$

the characteristic equations are simplified to

$$s^4 + (1 + \lambda)\bar{C}_2 s^3 + 3a\frac{\lambda + 1}{\lambda - 1} s^2 + 3(\lambda - 1)\bar{C}_2 s + 9(a - 1) = 0, \tag{15}$$

with parameters $\bar{C}_2$, a, and λ, and

$$
\begin{aligned}
&s^6 + (1+\mu)\bar{C}_1 s^5 + \left[\frac{4\lambda + (\mu - \lambda)}{\mu + \lambda} + \frac{4b(1+\mu)}{\mu - 1}\right] s^4 \\
&+ \left[\frac{(1+\mu)(\mu - \lambda) + 4\lambda}{\mu + \lambda} + 4(\mu - 1)\right] \bar{C}_1 s^3 \\
&+ 4\left[\frac{b(\mu - \lambda)(1+\mu) + 4\mu\lambda + 4\lambda(b-1)}{(\mu - 1)(\mu + \lambda)} + 4(b-1)\right] s^2 \\
&+ 4\bar{C}_1 \frac{(\mu - \lambda)(\mu - 1)}{\mu + \lambda} s + \frac{16(\mu - \lambda)(b-1)}{\mu + \lambda} = 0,
\end{aligned}
\tag{16}
$$

with parameters $\bar{C}_1$, b, and μ. These two equations have been programmed on a digital computer for computation of complex roots with $a,b = 1.0$ to 2.0, $\bar{C}_1$, $\bar{C}_2 = 0$ to 7.0, $\lambda = 1.25$ to 6.00, and $\mu = 2.0$ to 24.0. The case with $a,b < 1.0$ will give rise to positive real parts of the roots and is of no interest to us.

The root loci of (15) and (16) are plotted in Figs. 4 and 5, respectively, with fixed spring constants ($a = 1.2$, $b = 1.2$) and varying damping constants and inertia ratios. The roots are plotted only in the second quadrant of the complex plane, as they are complex conjugates with negative real parts. In the case of critical and overcritical damping, roots degenerate into the negative real axis. In the pitch case, with $\lambda = 2/[3 - 5^{\frac{1}{2}}] \approx 2.61804$, the two distinct modes will, at $\bar{C}_2 \approx 1.2805$, coalesce into a single point, $-n = -\mathrm{Re}(s) = -1.16$, on the negative real axis, corresponding to a $1/e$ damping time of 0.137 orbit at large t or ψ. This is the result given by Zajac.[3] For other values of λ, there always exists a pair of complex conjugates for the two modes if $\bar{C}_2$ is not too large. It is noted from Fig. 4 that when λ is in the range of 2.5 to 4.0, a proper choice of the damping constant $\bar{C}_2$ will make $-n_1$ of the least damped mode not less than 0.7, which corresponds to a $1/e$ damping time of 0.23 orbit. Values of λ in the above range and the corresponding optimum damping constants are given in Table I. This gives the satellite designer a wide choice of the inertia ratio, I_2/I_5, and the damping constant, $C_2/I_2\Omega$, and still the $1/e$ damping time is not greater than 0.23 orbit in the free pitch libration case.

From the complex root-locus plot for the roll-yaw free libration (Fig. 5 at $b = \bar{k}_1/\bar{k}_1^* = 1.2$), it is noted that the highest and the intermediate modes coalesce into a single point at $\mu \approx 6.425$, and $\bar{C}_1 \approx 0.3625$. The corresponding lowest mode has a poor damping. The intermediate and the lowest modes coalesce into a single point at $\mu \approx 18$ and $\bar{C}_1 \approx 0.0935$, which gives a poor damping for the highest mode. A close examination of the plot indicates that for μ lying in the range of 8 to 10,

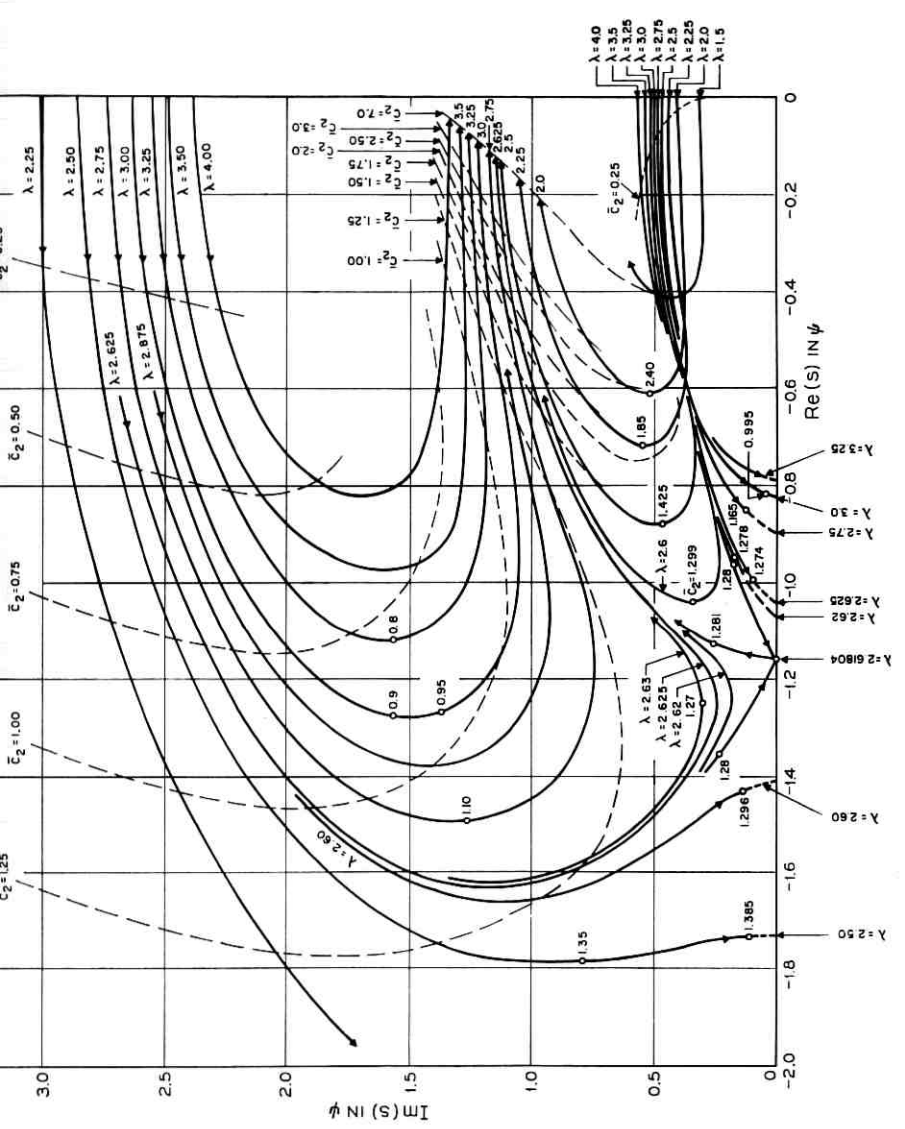


Fig. 4 — Complex root loci for pitch optimization with viscous damping ($a = \bar{k}_2/\bar{k}_2^* = 1.2$).

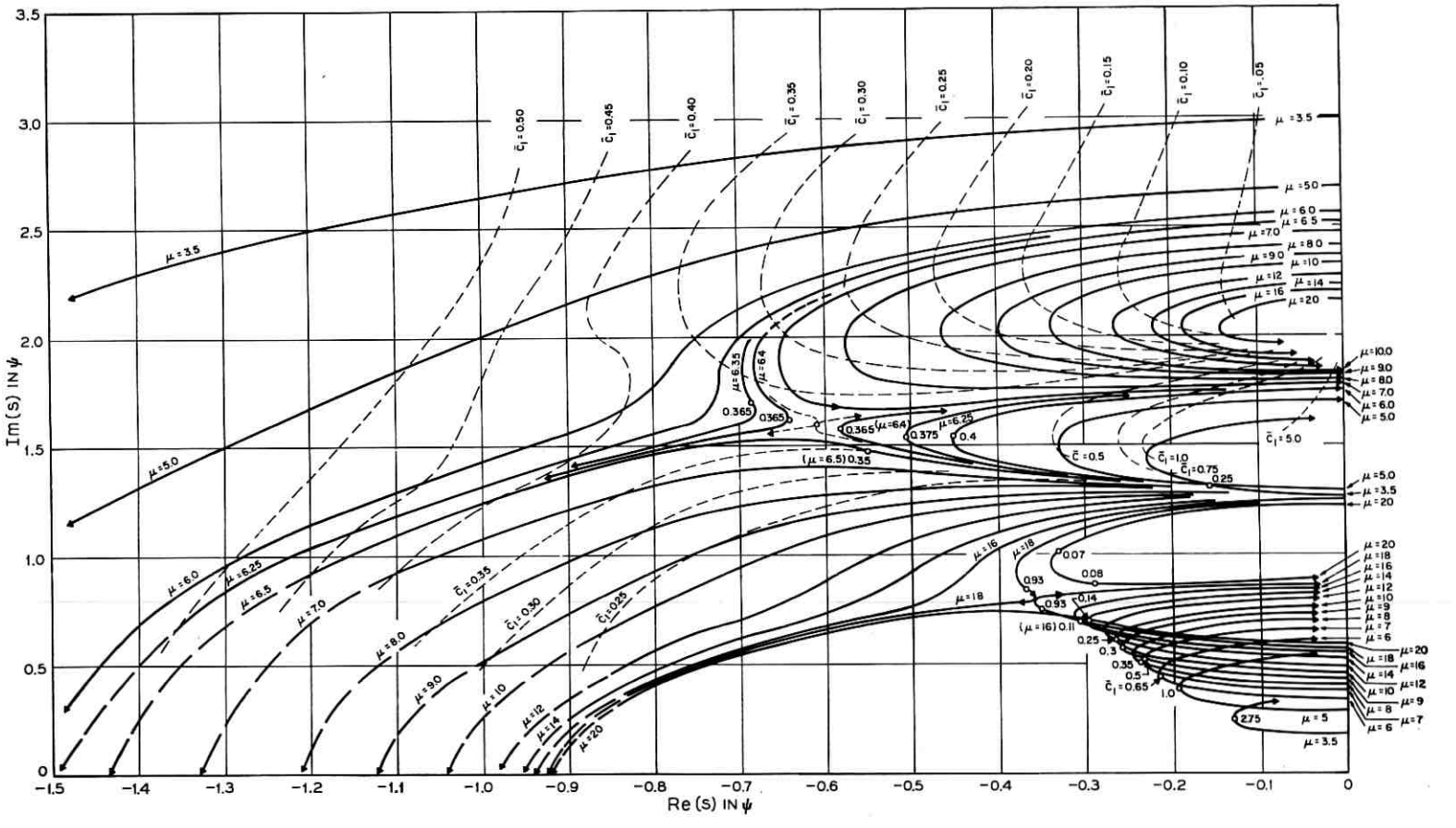


Fig. 5 — Complex root loci for roll-yaw optimization with viscous damping ($b = \bar{k}_1/\bar{k}_1^* = 1.2$).

TABLE I — OPTIMUM DAMPING CONSTANT OF PITCH LIBRATION

$\lambda \left(= \frac{I_2}{I_3}\right)$	$\bar{C}_2$	Negative Real Part in ψ, $-\mathrm{Re}(s)$	
		Lower Mode	Higher Mode
2.5	1.425	0.88	critically
2.6	1.299	1.04	damped
2.61804	1.2805*	1.16†	1.16
2.75	1.165	0.85	1.33
3.0	0.995	0.81	1.17
3.25	0.870	0.78	1.00
3.5	0.80	0.74	0.90
4.0	0.69	0.70†	0.75

* Both modes coalesce into a single point on the negative real axis.
† $\mathrm{Re}(s) = -1.16\psi$ corresponds to $1/e$ settling time of 0.137 orbit. $\mathrm{Re}(s) = -0.70\psi$ corresponds to $1/e$ settling time of 0.228 orbit.

a proper choice of the damping constant, $\bar{C}_1$, will make the negative real part of the least damped mode, $-\mathrm{Re}(s)_{\min}$, larger than 0.24, which corresponds to a $1/e$ damping time of 0.66 orbit. If the lower bound of $-\mathrm{Re}(s)_{\min}$ is relaxed to 0.20 ($1/e$ time $\approx$ 0.796 orbit), the range of μ becomes much wider, i.e., 6 to 14, as indicated in Table II with the corresponding optimum damping constant.

The variation of $-\mathrm{Re}(s)$ with the spring constants is such that the pitch optimum spring constant (for the least damped mode) depends on the choice of damping constant, $\bar{C}_2$, at a given inertia ratio, λ, as shown in Fig. 6 for a few chosen cases. In the roll-yaw case, the

TABLE II — OPTIMUM DAMPING CONSTANT OF ROLL-YAW LIBRATIONS

$\mu \left(= \frac{I_1}{I_4}\right)$	$\bar{C}_1$	Negative Real Part in ψ, $-\mathrm{Re}(s)$		
		Lowest Mode	Intermediate Mode	Highest Mode
6	0.65	0.216	0.211	critically damped
7	0.49	0.233	0.225	"
8	0.375	0.247	0.250	"
9	0.30	0.258	0.254	0.990
10	0.25	0.265	0.250	0.715
12	0.175	0.280	0.230	0.660
14	0.13	0.271	0.212	0.495

Note: $\mathrm{Re}(s) = -0.2\psi$ corresponds to $1/e$ settling time of 0.796 orbit. $\mathrm{Re}(s) = -0.28\psi$ corresponds to $1/e$ settling time of 0.568 orbit.

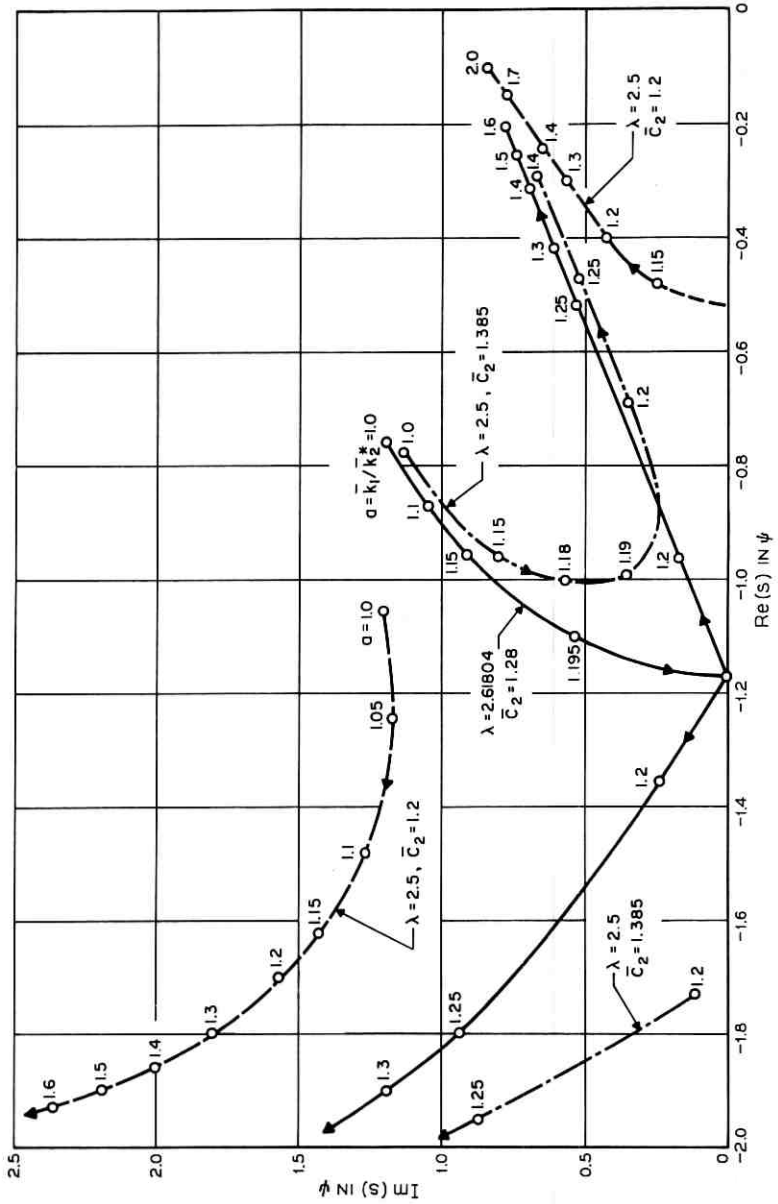


Fig. 6 — Variation of pitch damping with spring constants at fixed λ and $\bar{C}_2$.

optimum spring constant for the lowest mode is found always smaller than that for the intermediate mode, as indicated in Fig. 7, while the highest mode gets better damping for increased $\bar{k}_1$ or b. For example, at $\mu = 8.0$, and $\bar{C}_1 = 0.35$, optimum $\bar{k}_1$ equals 1.13 $\bar{k}_1^*$ for the lowest mode but equals 1.31 $\bar{k}_1^*$ for the intermediate mode. From a practical consideration of the spring design, one should not choose the spring constant too close to the critical value because of possible decrease due to vibrations, thermal effects, etc. Furthermore, an analysis of pointing errors resulting from deviations in geometric configuration (due to rod deflections,[1,2] etc.) indicates the advantage of employing larger spring constants. In view of these conflicting results, one may have to settle for some compromise values, e.g., $a,b \approx 1.2$ to 1.8.

3.2 *Forced Pitch Libration by Orbital Eccentricity*

The steady-state solution of (9a) and (9b) for η_1 is found to be

$$\eta_1 = F_1 C\psi + G_1 S\psi. \tag{17}$$

Here $F_1 = 2\epsilon\Delta^{-1}\bar{C}_2 P_2(P_1 - P_2)$, and

$$G_1 = 2\epsilon\Delta^{-1}[P_1 P_2^2 + (1 + 2\lambda)\bar{k}_2 P_1 P_2 + \bar{k}_2 P_2^2 + (1 + \lambda)(\lambda P_1 + P_2)(\bar{C}_2^2 + \bar{k}_2^2)],$$

where $P_1 = 3p_1 - 1$, and $P_2 = 3p_2 - 1$. The ratio of the amplitude, $\bar{\eta}_1 = (F_1^2 + G_1^2)^{\frac{1}{2}}$, to the eccentricity, ϵ, is plotted in Fig. 8 versus the spring constant, a, with varying $\bar{C}_2$ and λ. This plot indicates the advantage of using a lower spring constant, and, for $1.0 < a \leqq 1.4$, the advantage of employing a smaller damping constant. This latter advantage is further reflected in the plot of $\bar{\eta}_1/\epsilon$ versus $\bar{C}_2$ ($a = 1.2$) (Fig. 9), especially when λ is in the range of 2.5 to 3.5. It is noted from both Figs. 8 and 9 that at $\bar{C}_2 = 0.5$ ($a = 1.2$), $\bar{\eta}_1/\epsilon$ is relatively independent of λ. If the optimum damping constant for the free librational motion, as given in Table I, for $\lambda = 2.5$ to 4.0 is used, the average value of $\bar{\eta}_1/\epsilon$ is approximately 2.25, which is only about 30 per cent larger than that given by $\bar{C}_2 = 0.5$.

IV. PARAMETER OPTIMIZATION WITH MAGNETIC HYSTERESIS DAMPING

4.1 *Energy-Fitting Method*

In the case of magnetic hysteresis damping, the equations of librational motions are the same as (9a) and (9b) for pitch, and (10a)

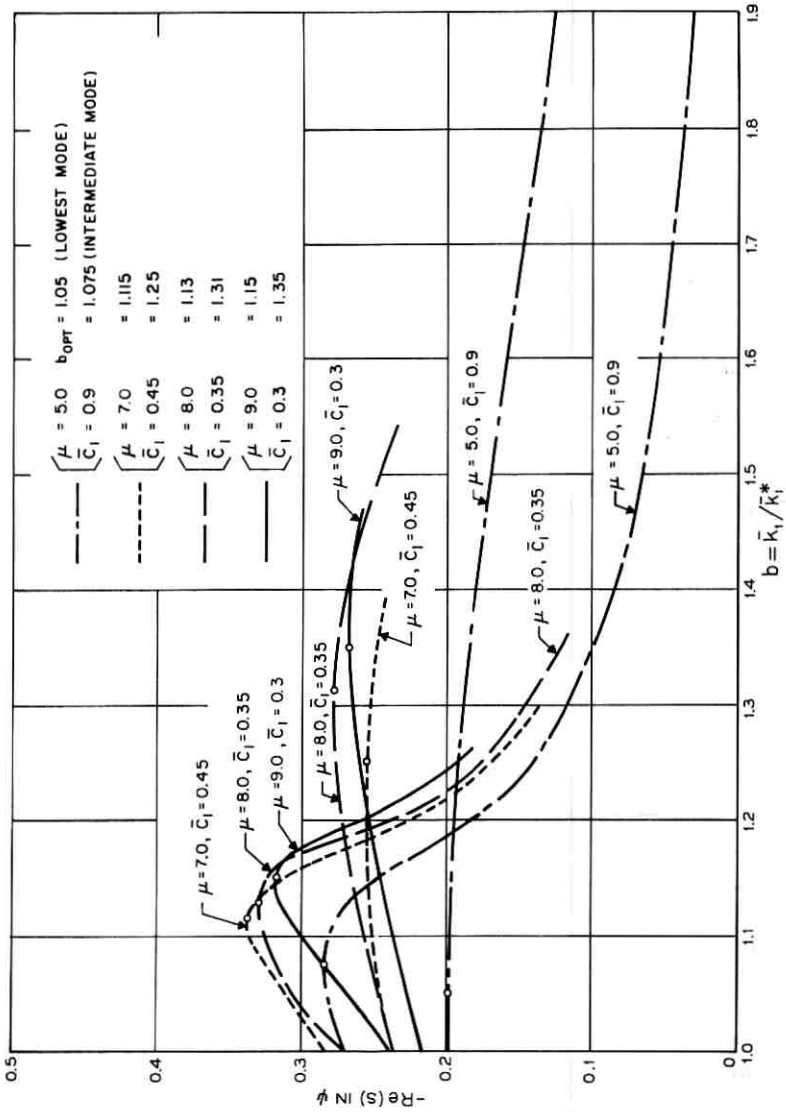


Fig. 7 — Variation of roll-yaw damping with spring constants at fixed μ and $\bar{C}_1$.

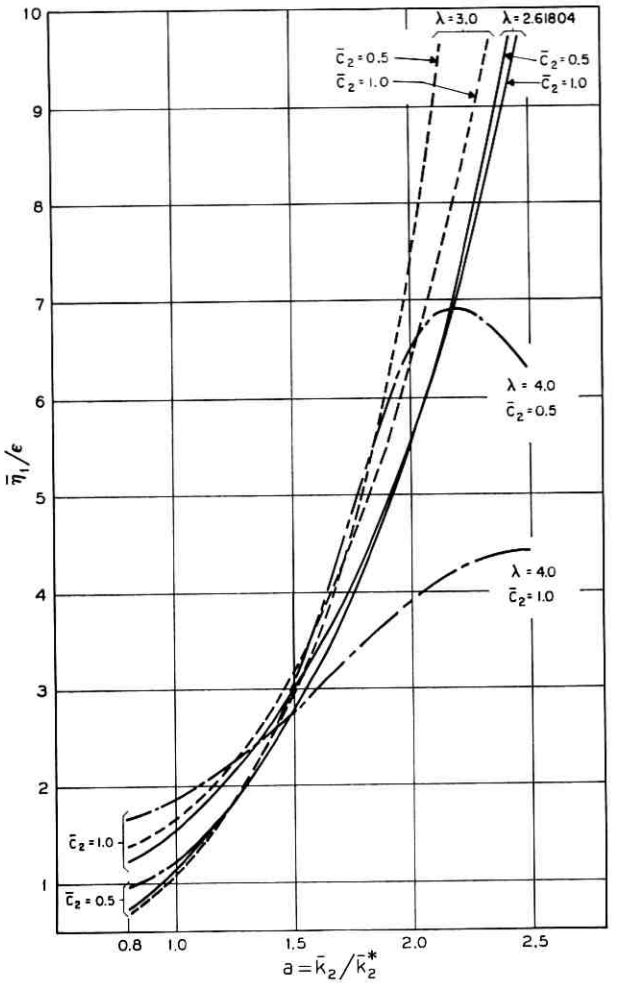


Fig. 8 — $\bar{\eta}_1/\epsilon$ vs a with varying λ and $\bar{C}_2$.

through (10c) for roll and yaw, except that the damping torques T_{d1} and T_{d2} become nonlinear and may be defined as follows. If, for example, the pitch relative angle of rotation from a neutral position, β, is larger then $\bar{\beta}$, then the major hysteresis loop (see Fig. 3) will be traced, and (5) will be applied for the pitch damping torque. If however, β is less than $\bar{\beta}$ after a change of sign of $\dot{\beta}$, then the torque will in general trace a minor loop as, for example, that shown in Fig. 10. The maximum

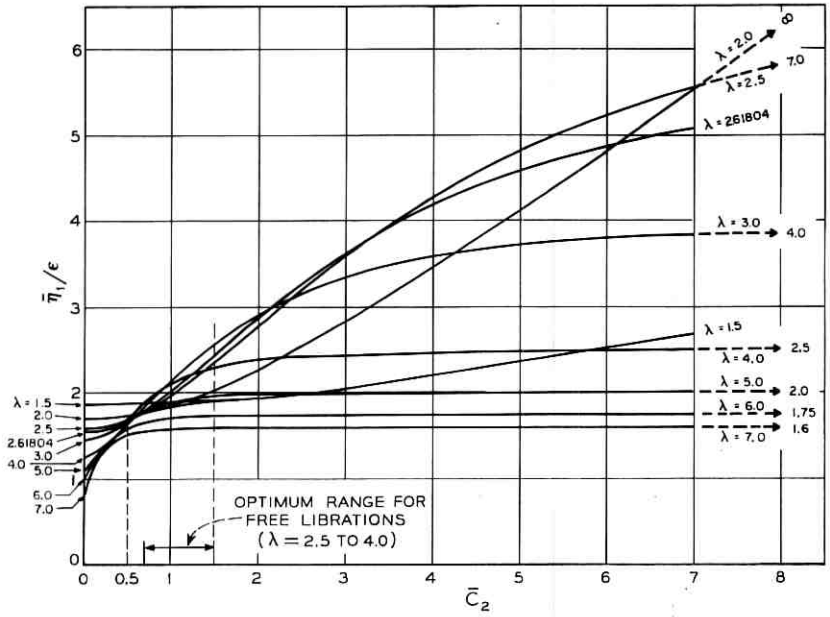


Fig. 9 — $\bar{\eta}_1/\epsilon$ vs $\bar{C}_2$ with varying λ ($a = \bar{k}_2/\bar{k}_2^* = 1.2$).

magnitude of the torque in the minor loop may be a function of the angular excursion. Since the torque always opposes the motion, the plot of the torque versus time is a broken curve (see Fig. 10). Each section of the curve can be represented by an analytical expression if it is approximated as a section of a rhomboid or an ellipse. Such an approximation can be used for numerical integration of the equations of motion.

If the T_{d2}-β curve is approximated as a closed loop for a slowly damped system, the loop area which represents the energy of dissipation per cycle of oscillation, E_h, is related to the angular amplitude of oscillation β as follows:

$$E_h = 4\bar{T}_{d2}(\beta - \bar{\beta}) \qquad \text{for} \qquad \beta \geqq \beta_0 , \tag{18a}$$

and

$$E_h = K\bar{T}_{d2}\beta^m \qquad \text{for} \qquad \beta < \beta_0 . \tag{18b}$$

In the above, $\bar{\beta}$, β_0, K, and m are constants depending on the characteristics of the damping material, etc., where $\bar{\beta}$ and β_0 are found usually to be very small, e.g., $\bar{\beta} = 1°$ to $4°$, and $\beta_0 = 3°$ to $7°$. If one equates E_h to the energy dissipation per cycle with the viscous damping,

$E_v = \pi C_2 \omega \beta^2$ (ω = circular frequency of one of the principal modes) at a certain value of β, then a relation between $\bar{C}_2$ and $\bar{T}_{d2}$ results. For example, when $E_h = E_v$ at $\beta = \beta_0$, then

$$\frac{\bar{T}_{d2}}{I_2 \Omega^2} = \left[\frac{\pi}{4} \frac{\beta_0^2}{(\beta_0 - \bar{\beta})} \frac{\omega}{\Omega}\right] \bar{C}_2 . \tag{19}$$

A similar relation can be obtained between $\bar{C}_1$ and $\bar{T}_{d1}/I_1\Omega^2$. Thus, from the "optimum" viscous damping constants obtained in Section III, one may find the equivalent "optimum" hysteresis damping torque. Nevertheless, it is pointed out that the equivalent "optimum" $\bar{T}_d$ obtained in such a way could be erroneous for the following reasons. First, this is not a slowly damped system, as can be observed from the computer solutions[1] of equations (3) with hysteresis damping, and also, as can be noted from Figs. 4 and 5, Re(s) is of the same order as Im(s) in the case of viscous damping. Hence, a closed-loop approximation for computing E_h and E_v based on a particular frequency is obviously a very poor one. Second, since the quadratic curve for E_v can fit the E_h-curve ($m \approx 1.5$ power for $\beta \leqq \beta_0$ and linear for $\beta \geqq \beta_0$) at only one point (see Fig. 11), (19) gives a much worse approximation at other values of β than at the point of fit.

4.2 *Numerical Method — Computer Solutions*

From the foregoing, it is apparent that "optimum" magnetic hysteresis damping constant cannot be accurately evaluated from the "opti-

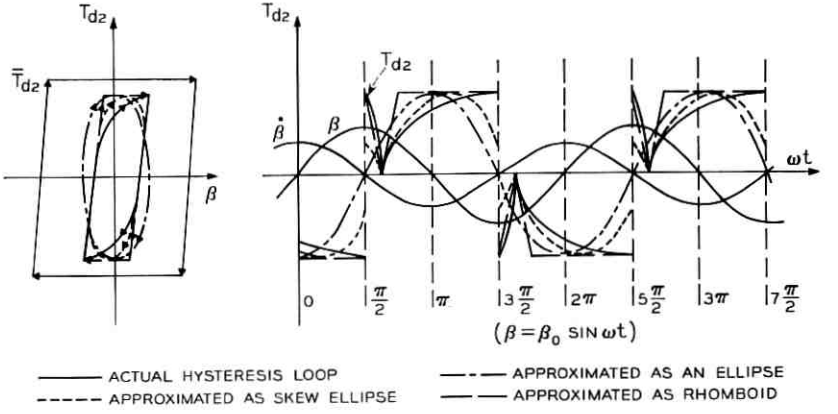


Fig. 10 — Minor magnetic hysteresis loops: torque-time diagram.

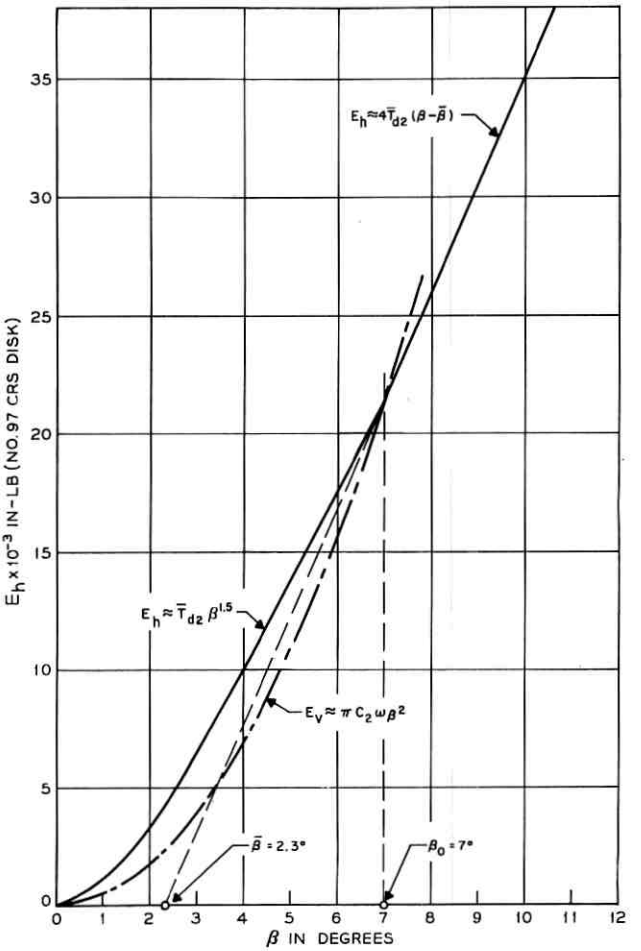


Fig. 11 — E_h-β curve.

mum" viscous damping constant by means of the energy method or other approximation methods, such as, the describing function method* (for example, see Ref. 4). Therefore, in the case of hysteresis damping, it is necessary to resort to numerical methods for the parameter optimi-

* If one defines the equivalent gain as the ratio of the first harmonics of the output, hysteresis damping torque, to the amplitude of the oscillation angle, assumed to be sinusoidal, and equates it to $C_1 s$ or $C_2 s$ in the characteristic equations, then the equations will only have terms of even power with, however, complex coefficients. This method is actually the same as the energy-fitting method and clearly offers no advantages.

zation. The results obtained in Section III for the viscous damping case may, however, be used as a guide. It is found from the computer solutions that the critical spring constants (11a) and (12a) obtained from the linearized equations (9a), (9b), and (10a) through (10c) with the viscous damping apply also for the case of hysteresis damping. It has also been demonstrated by computer solutions that the values of the inertia ratios, λ and μ, in the ranges given in Tables I and II, will give better damping for the same hysteresis damping torques and spring constants. We shall numerically integrate (3a) through (3f) and (6a) through (6f) with the hysteresis damping torque defined in (5) by putting $\bar{\alpha}$ and $\bar{\beta}$ equal to 1° to 4°. This eliminates the unnecessary complex programming of the minor loops of the hysteresis damping, though no damping will result when α and β are smaller than $\bar{\alpha}$ and $\bar{\beta}$, respectively. If the initial condition is a tumbling motion, the numerical results will cover tumbling motion, large-angle motion, and librational motion. Only in the librational motion does the solution get less and less accurate when the relative angles get closer to $\bar{\alpha}$ and $\bar{\beta}$.

A number of computer runs have been made employing hysteresis damping constants in a wide range. The results indicate: (a) the number of orbits after which the satellite will stop tumbling from the initial condition, $\omega_1 = \Omega$, $\omega_2 = 5\Omega$, $\omega_5 = \Omega$, $\omega_3 = \omega_4 = \omega_6 = 0$; and (b) the librational angles after 20 orbits. Computer runs have also been made (for the case of $\lambda = 3.333$, $\mu = 7.4$ only) to determine the librational angles after 10 orbits from the initial condition: at $t = 0$, $\theta_x = 36°$, $\theta_y = 43°$, $\theta_z = 27°$, $\alpha = 1°$, $\beta = 10°$, $\omega_1 = -0.84\Omega$, $\omega_2 = 1.17\Omega$, $\omega_3 = -0.3\Omega$. These angles are defined as $\theta_x = C^{-1}(\hat{x}_1 \cdot \hat{x})$, $\theta_y = C^{-1}(\hat{y}_1 \cdot \hat{y})$, and $\theta_z = C^{-1}(\hat{z}_1 \cdot \hat{z})$, (the caret denotes a unit vector), where θ_z is the earth-pointing error angle. The results of these computer runs are summarized in Figs. 12 through 14 for three sets of inertia ratios.

It is noted from Figs. 12–14 that the damping constants which give better damping in librations are, in general, smaller than those for better tumbling damping. Poor damping will result in both tumbling and librational motions when $\bar{T}_{d1}$ and $\bar{T}_{d2}$ are either too small or too large. In the intermediate wide range of the damping constants there is relatively small variation in the damping time. The "optimum" damping constants may be chosen from this wide range as, for example,

$$\bar{T}_{d2}/I_2\Omega^2 = \begin{cases} 0.19 - 0.38 & \text{for } \lambda = 2.62, \quad \mu = 6.28 \\ 0.16 - 0.32^* & \text{for } \lambda = 3.33, \quad \mu = 7.4 \\ 0.08 - 0.16 & \text{for } \lambda = 4.0, \quad \mu = 14, \end{cases}$$

* The upper limit is taken to be smaller than that given in Ref. 2.

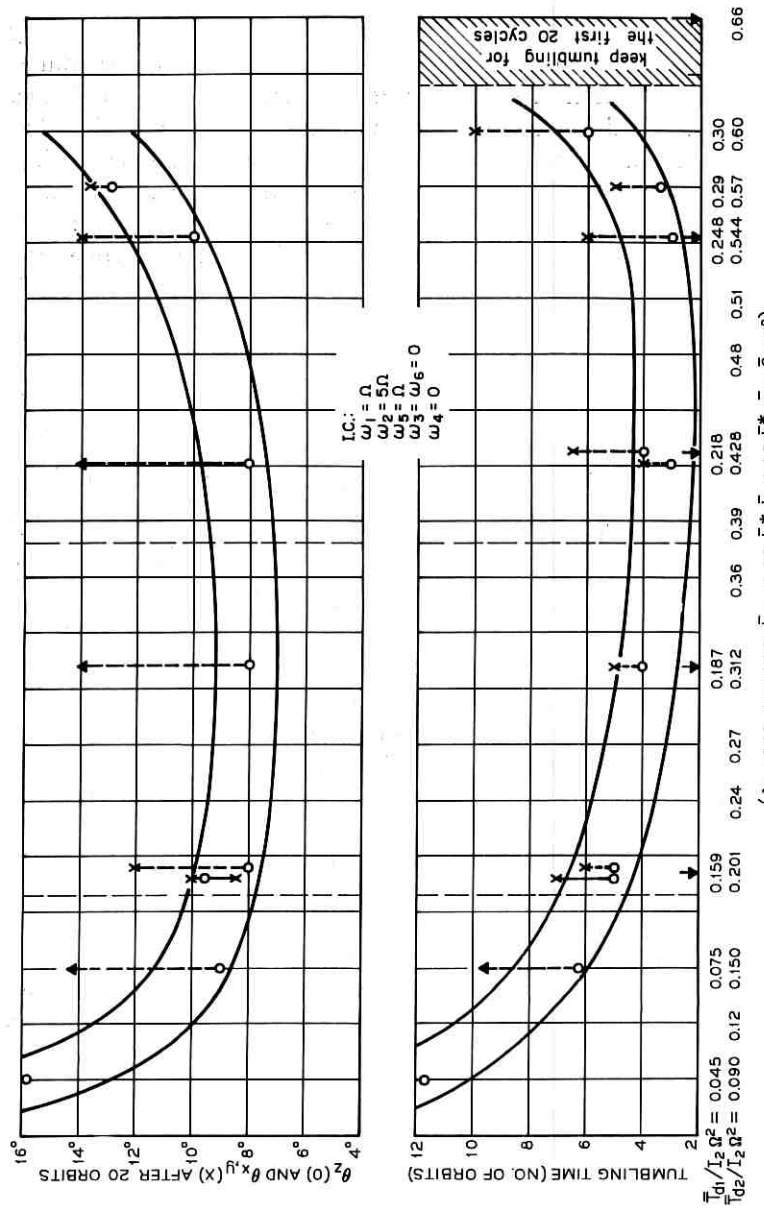


Fig. 12 — Computer results with magnetic hysteresis damping, $\lambda = 2.62$, $\mu = 6.28$.

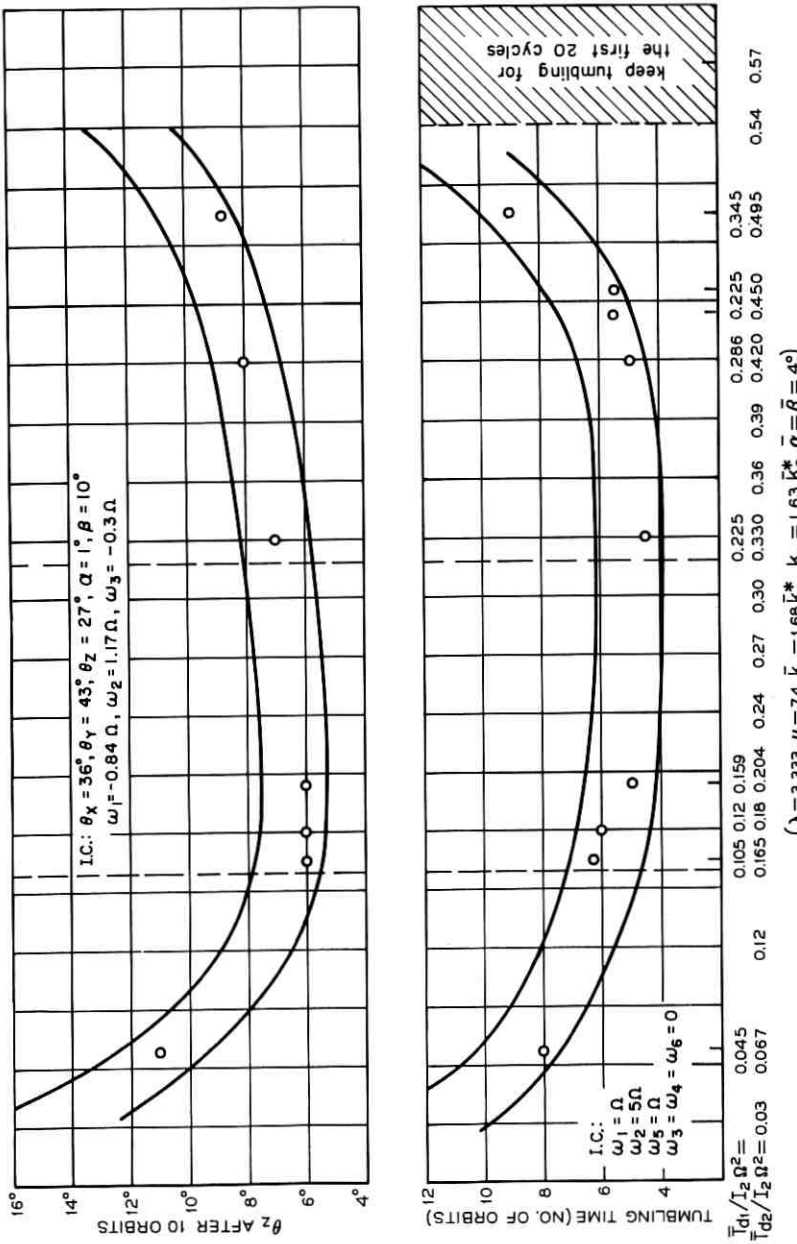
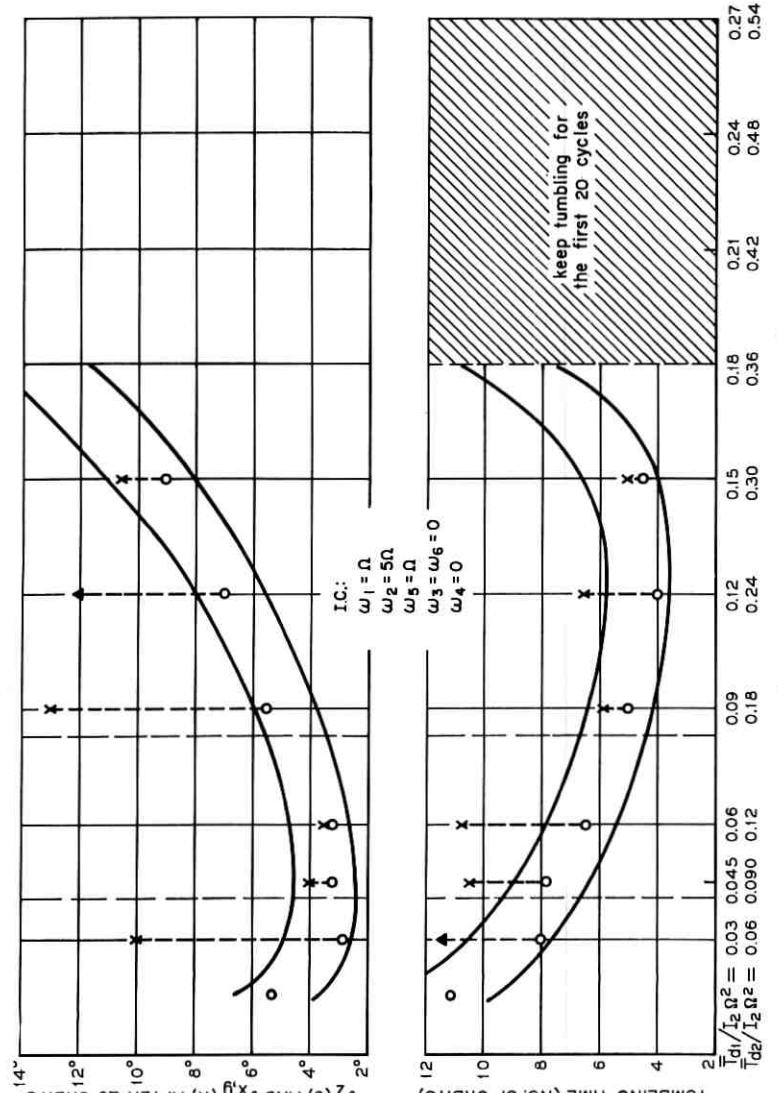


Fig. 13 — Computer results with magnetic hysteresis damping, $\lambda = 3.333$, $\mu = 7.4$.



$(\lambda = 4, \mu = 14, \bar{K}_{1,2} = 1.4\,\bar{K}^*_{1,2}, \bar{\alpha} = \bar{\beta} = 2°)$

Fig. 14 — Computer results with magnetic hysteresis damping, $\lambda = 4, \mu = 14$.

and $\bar{T}_{d1}$ is chosen to be one-half of $\bar{T}_{d2}$. Since only a limited number of computer runs have been made, it is not possible to display $\bar{T}_{di}/I_2\Omega^2$ $(i = 1,2)$ as a function of $\lambda(= 2.5$ to $4.0)$ and $\mu(= 6$ to $14)$. The "optimum" damping constants corresponding to other values of λ and μ have yet to be determined from computer runs.

The effects of spring constants have also been investigated on the computer by varying the spring constants under the same hysteresis damping constants. It is found that the variation in damping time is relatively small for $\bar{k}_i/\bar{k}_i^* = 1.2$ to 1.8 $(i = 1,2)$. The lower limit is chosen to guarantee stability.[1] To ensure torsional fatigue strength of the torsion wire, it is preferable to choose high spring constants. Furthermore, employment of larger spring constants will reduce the error angle in case of rod bending, as indicated in the error analysis.

The relation between the forced librational amplitude and the orbital eccentricity has been found from the computer runs to be

$$\theta_z \approx 2\epsilon \text{ to } 3\epsilon$$

for $\epsilon \leqq 0.10$. At $\epsilon = 0.2$, the satellite starts tumbling from an earth-pointing position after 1.5 orbits, whereas at $\epsilon = 0.4$ tumbling begins immediately after the start.

Due to the nonlinear characteristics, the effectiveness of the magnetic hysteresis damping cannot be measured by the $1/e$ settling time as with viscous damping. However, an equivalent $1/e$ settling time for the hysteresis damping may be obtained in the following way. Take, for example, a typical computer run as plotted in Fig. 15 for $\theta_z[\lambda = 4, \mu = 14, \bar{T}_{d1}/I_2\Omega^2 = 0.06, \bar{T}_{d2}/I_2\Omega^2 = 0.12, \bar{k}_i = 1.4\bar{k}_i^*$ $(i = 1,2)$, $\bar{\alpha} = \bar{\beta} = 2°]$. From the envelope curve of the amplitudes of the damped large-angle motion, the logarithmic decrements can be approximately evaluated by taking the motion as exponentially damped. It is found that the equivalent $1/e$ time ranges from one to three orbits with an average of two orbits. Within the validity of the librational motion, this represents an average $1/e$ time for both pitch and roll-yaw librations with the above parameters.

V. SUMMARY — ILLUSTRATION OF A PRACTICAL DESIGN

The results given in the previous sections indicate that there does not exist a single set of optimum parameters which give the best over-all damping performance. However, there have been found relatively wide optimum ranges of the inertia ratios, the spring constants, and the damping constants for the design of a two-body satellite. The optimum

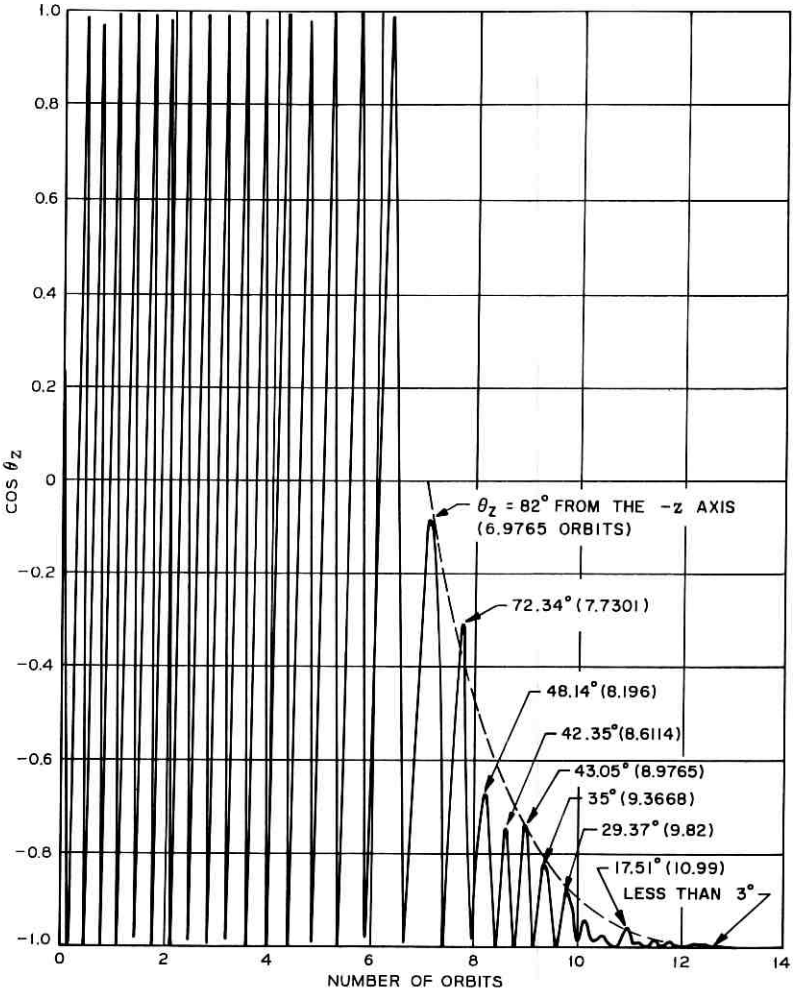


Fig. 15 — Variation of θ_z with magnetic hysteresis damping.

ranges of the inertia ratios and the spring constants are found applicable to both viscous and magnetic hysteresis damping. They are λ ($= I_2/I_5$) $= 2.5$ to 4.0, μ ($= I_1/I_4$) $= 6$ to 14, and $\bar{k}_{1,2}/\bar{k}_{1,2}^* = 1.2$ to 1.8. The optimum viscous damping constants, as tabulated in Tables I and II, can be read directly from the complex root-locus plots, whereas the optimum ranges of the magnetic hysteresis damping constants corresponding to a few inertia ratios are obtained from the computer runs.

As far as the inertia ratios are concerned, it is advantageous to employ large λ and μ so that the deck rods[2] can be made shorter or the tip mass can be smaller. By employing shorter deck rods and furthermore by shortening the mast rod,[2] it is possible to balance out the solar torque, which was the largest disturbing torque for the satellite described in Ref. 2. As long as the solar torque is reduced to a small magnitude, the lowered gravity torque level resulting from the shortening of the mast rod will not substantially increase the error angles produced by other relatively small disturbing torques. In view of the foregoing, we could improve the design previously given in Ref. 2. The parameters may now be chosen as follows:

$I_1 = I_2 = 600, I_3 = 10$ slug-ft^2
pitch gravity torque, $T_g = 0.23 \times 10^{-5}$ ft-lb/deg $= 31$ dyne-cm/deg
$\lambda = 4, I_5 = 150; \mu = 14, I_4 = 42.8$ slug-ft^2
$\bar{T}_{d1}/I_2\Omega^2 = 0.06, \bar{T}_{d1} = 0.446 \times 10^{-4}$ ft-lb
$\bar{T}_{d2}/I_2\Omega^2 = 0.12, \bar{T}_{d2} = 0.892 \times 10^{-4}$ ft-lb
$\bar{k}_1 = 1.8\,\bar{k}_1^* = 0.55, k_1 = 0.247 \times 10^{-4}$ ft-lb/rad
$\bar{k}_2 = 1.8\,\bar{k}_2^* = 1.8, k_2 = 0.804 \times 10^{-4}$ ft-lb/rad
length of mast rod $= 30$ ft (assuming 160-lb satellite proper)
length of deck rod $= 16.5$ ft (4 deck rods)
roll tip mass $= 2 \times 2.5 = 5$ lbs
pitch tip mass $= 2 \times 9 = 18$ lbs
total mass for attitude control (excluding mast motor and support) $= 26$ lbs.

The solar torque error angle is now only 1°, and the total error angle is found to be about 4° excluding the effect of the orbit eccentricity:

magnetic dipole moment (TELSTAR's)	2°
solar torque	1°
rod bending (silver-plated rod)	1°
orbital eccentricity, ϵ	2ϵ to 3ϵ
total	$4° + 2.5\epsilon$.

In the preceding, the error angle due to the residual magnetic dipole moment can be easily reduced to 1° or less by further refinement of the cancellation techniques used on the TELSTAR satellite.

REFERENCES

1. Fletcher, H. J., Rongved, L., and Yu, E. Y., Dynamics Analysis of a Two-Body Gravitationally Oriented Satellite, B.S.T.J., **42**, Sept., 1963, pp. 2239–2266.

2. Paul, B., West, J. W., and Yu, E. Y., A Passive Gravitational Attitude Control System for Satellites, B.S.T.J., **42,** Sept., 1963, pp. 2195–2238.
3. Zajac, E. E., Damping of a Gravitationally Oriented Two-Body Satellite, ARS Jour., **32,** pp. 1871–1875, Dec., 1962.
4. Gibson, J. E., *Nonlinear Automatic Control,* McGraw-Hill, New York, 1963, Chap. 9.

The Existence of Eigenvalues for the Integral Equations of Laser Theory

By J. ALAN COCHRAN

(Manuscript received July 15, 1964)

In this paper the general integral equations governing the mode spectra of optical masers are investigated from a point-of-view based upon certain theoretical results for Hölder continuous kernels. Using an estimation originally performed by Fredholm, it is proved that the homogeneous integral equation

$$\varphi(x) = \lambda \int_a^b K(x,y)\, \varphi(y)\, dy$$

has at least one eigenvalue for Hölder continuous kernels K with exponent $\alpha > \frac{1}{2}$ and with nonvanishing trace. All the integral equations which have been treated in laser theory so far can be "factored" into one-dimensional equations with continuously differentiable kernels, to which this result applies directly.

Although in practice the vanishing of the trace is the exception rather than the rule, the later sections of this paper are devoted to demonstrations of the nonvanishing character of the trace of several of the common "laser kernels" associated with practical reflector configurations. These results provide in almost all cases the first rigorous proofs of the existence of eigenvalues and eigenfunctions for the integral equations of the optical maser.

I. INTRODUCTION

Homogeneous linear Fredholm integral equations with nonsingular kernels of normal type (which includes Hermitian kernels as a special case), i.e., kernels for which

$$\int_a^b K(x,z)\, \bar{K}(y,z)\, dz = \int_a^b \bar{K}(z,x)\, K(z,y)\, dz,$$

have been extensively studied. Within the framework of complex-valued $\mathfrak{L}^2$ functions, questions of existence, uniqueness, and representation of

solutions can be largely answered for such equations. Recently, however, a number of integral equations involving kernels which are neither Hermitian nor normal have arisen in laser theory. These kernels, of which

$$K(x,y) = e^{ik(x-y)^2}, \tag{1}$$

with k a given complex constant, may be considered representative, do have the seemingly beneficial property of being complex-symmetric, viz.

$$K(x,y) = K(y,x).$$

Unfortunately, due to the lack of a sufficient analytic theory for such kernels, this "advantage" has yet to be adequately exploited. Thus, although it is well-known that every Hermitian kernel distinct from the zero transformation has at least one eigenvalue, the existence of eigenvalues for general complex-symmetric or other non-normal kernels still remains an open mathematical question.

These remarks are not meant to imply that there has been a paucity of theoretical investigation to complement the widespread experimental work with masers and lasers of various geometries. Quite the contrary! Boyd and Gordon,[1] for instance, have shown that for the confocal geometry the resultant integral equation is equivalent to that considered earlier by Slepian and Pollak[2] and has prolate spheroidal wave functions as eigenfunctions. Somewhat later, Boyd and Kogelnik[3] generalized this work to resonators with unequal reflector apertures and curvatures. Moreover, iterative computational methods have been applied by Fox and Li[4,5] and Li[6] to integral equations arising from a wide range of interferometer geometries. Their techniques have produced plausible numerical descriptions of the characteristic low-order modes and eigenvalues for the configurations considered.

Even with the contributions represented by the above papers, however, there still remains a dearth of knowledge, in a mathematical sense, about the eigenfunctions and eigenvalues (if any) of the homogeneous integral equations encountered in the general theory of the optical maser. The nature of some of the mathematical questions yet to be answered in this area was considered in an early 1963 paper by S. P. Morgan.[7] Since that time some progress has been made regarding the existence of eigenvalues for certain "laser kernels." Newman and Morgan,[8] by means of lengthy Taylor series techniques, have proved that kernels of the form

$$K(x,y) = G(x)\, F(xy)\, H(y)$$

with rather general G, F, H and with nonvanishing trace possess at

least one nontrivial eigenvalue. Other recent work[9,10] has centered around the use of the natural Hilbert-Schmidt expansions of "planar" and "near-confocal" kernels in terms of their singular systems [Ref. 11, p. 142 ff.].

In this paper we want to assume a somewhat different approach based upon certain theoretical results for Hölder continuous kernels. We first prove (in Section III) the following

Theorem: Let the kernel $K(x,y)$ be Hölder continuous in either variable, with exponent $\alpha > \frac{1}{2}$, for $a \leqq x,y \leqq b$. Then if the trace of K does not vanish, the homogeneous integral equation

$$\varphi(x) = \lambda \int_a^b K(x,y)\, \varphi(y)\, dy$$

has at least one eigenvalue.

The essential step in the proof is an estimation of the coefficients in the classical series representation for the Fredholm determinant of the kernel $K(x,y)$, an estimation originally carried out by Fredholm himself.[12]

We next observe that all the integral equations which have actually been treated in laser theory so far can be "factored" into one-dimensional equations with continuously differentiable kernels, to which the above theorem applies directly. Although we expect that in practice the vanishing of the trace is the exception rather than the rule, we devote the latter sections of this paper to demonstrations of the nonvanishing character of the trace of several of the common "laser kernels" associated with practical reflector configurations. These examples are indicative of the ease with which the existence of eigenvalues and eigenfunctions can be rigorously established for many of the one-dimensional kernels arising in the theory of the optical maser.

II. MATHEMATICAL PRELIMINARIES

In general we shall consider complex-valued kernels $K(x,y)$ defined on the bounded real domain $a \leqq x,y \leqq b$. Thus, where limits of integration on integrals are not specified, the integrations are to be performed over the interval $[a,b]$. We shall also assume that K belongs to the class $\mathfrak{L}^2$, i.e.,

$$\text{norm } K = \| K \| = \left[\int_a^b \int_a^b |K(x,y)|^2 \, dx\, dy \right]^{\frac{1}{2}} < \infty ,$$

and that $K(x,y)$ is a square-summable function of y for each value of x and conversely.

Our notation for composite kernels shall be

$$KL = \int_a^b K(x,z)\, L(z,y)\, dz.$$

Iterated kernels will be denoted by

$$\begin{aligned} K^\nu &= KK^{\nu-1} \\ &= \int_a^b K(x,z)\, K^{\nu-1}(z,y)\, dz \qquad \nu \geqq 2 \end{aligned}$$

with $K^1 = K(x,y)$. In the same manner

$$\text{trace } K = \text{tr}(K) = \int_a^b K(x,x)\, dx$$

and

$$\begin{aligned} k_\nu = \text{tr}(K^\nu) &= \int_a^b K^\nu(x,x)\, dx \\ &= \int_a^b \int_a^b K(x,z)\, K^{\nu-1}(z,x)\, dz\, dx. \end{aligned}$$

Reference should be made to Smithies[11] for further definitions and standard theorems on integral equations as needed.

Certain notions regarding the characterization of entire or integral functions will also be of value in our work. In particular, recall that the *order* μ of entire $f(z) = \sum_{n=0}^{\infty} a_n z^n$ is given by

$$\mu = \limsup_{n\to\infty} \frac{n \log n}{\log (1/|\, a_n \,|)}. \tag{2}$$

Other definitions and results may be found in texts such as Boas.[13]

The property of Hölder continuity is indicative of the smoothness of a given function. Kernels for which there exist positive constants A and α such that either

$$|\, K(x,y) - K(z,y) \,| < A\, |\, x - z \,|^{\alpha} \qquad \text{for all } x,y,z \text{ in } [a,b]$$

or

$$|\, K(x,y) - K(x,z) \,| < A\, |\, y - z \,|^{\alpha} \qquad \text{for all } x,y,z \text{ in } [a,b]$$

are termed Hölder continuous in x or y respectively with exponent α. If $\alpha = 1$ the functions are said to satisfy a Lipschitz condition, and thus Hölder continuity is occasionally designated Lip_α. It should be noted

that continuously differentiable functions automatically satisfy Lipschitz conditions with $\alpha = 1$.

III. THE MAIN THEOREM

Theorem: Let the kernel $K(x,y)$ be Hölder continuous in either variable, with exponent $\alpha > \frac{1}{2}$, for $a \leqq x,y \leqq b$. Then if the trace of K does not vanish, the homogeneous integral equation

$$\varphi(x) = \lambda \int_a^b K(x,y)\, \varphi(y)\, dy \tag{3}$$

has at least one eigenvalue.

Proof: The eigenvalues of (3) are the zeros, if any, of the Fredholm determinant $D(\lambda)$ associated with the kernel $K(x,y)$. The classical series representation[12] for this entire function $D(\lambda)$ is

$$D(\lambda) = \sum_{\nu=0}^{\infty} d_\nu \lambda^\nu \tag{4}$$

where $d_0 = 1$ and

$$d_\nu = \frac{(-1)^\nu}{\nu!} \iint \cdots \int K\begin{pmatrix} s_1, s_2, \cdots, s_\nu \\ s_1, s_2, \cdots, s_\nu \end{pmatrix} ds_1 ds_2 \cdots ds_\nu \qquad (\nu \geqq 1) \tag{5}$$

with

$$K\begin{pmatrix} s_1, s_2, \cdots, s_\nu \\ s_1, s_2, \cdots, s_\nu \end{pmatrix} = \det\,(K(s_i, s_j)) = \begin{vmatrix} K(s_1,s_1) & K(s_1,s_2) & \cdots & K(s_1,s_\nu) \\ K(s_2,s_1) & K(s_2,s_2) & \cdots & K(s_2,s_\nu) \\ \cdots & \cdots & \cdots & \cdots \\ K(s_\nu,s_1) & K(s_\nu,s_2) & \cdots & K(s_\nu,s_\nu) \end{vmatrix}. \tag{6}$$

We want to determine the order μ of $D(\lambda)$ under the above hypotheses on the kernel $K(x,y)$. Let us assume, therefore, that K is uniformly Hölder continuous with respect to the second variable, that is

$$|K(x,y) - K(x,z)| < A\,|y - z|^{\alpha} \tag{7}$$

with $\alpha > \frac{1}{2}$.

To estimate the coefficients d_ν,* we first transform the determinant in (6) by subtracting the second column from the first, the third column from the second, etc., thus obtaining

* This estimation was originally performed by Fredholm in 1903.[12]

$$\det (K(s_i, s_j)) = [(s_1 - s_2)(s_2 - s_3) \cdots (s_{\nu-1} - s_\nu)]^\alpha \cdot \begin{vmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1,\nu-1} & K(s_1, s_\nu) \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2,\nu-1} & K(s_2, s_\nu) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \epsilon_{\nu 1} & \epsilon_{\nu 2} & \cdots & \epsilon_{\nu,\nu-1} & K(s_\nu, s_\nu) \end{vmatrix}$$

where

$$\epsilon_{mn} = \frac{K(s_m, s_n) - K(s_m, s_{n+1})}{(s_n - s_{n+1})^\alpha}$$

$$m = 1,2, \cdots, \nu; \qquad n = 1,2, \cdots, \nu - 1.$$

Since $| \epsilon_{mn} | < A$ by (7), and $K(x,y)$ itself is bounded by continuity, a simple application of Hadamard's inequality yields

$$| \det (K(s_i, s_j)) | < C^\nu \nu^{\nu/2} | (s_1 - s_2)(s_2 - s_3) \cdots (s_{\nu-1} - s_\nu) |^\alpha, \quad (8)$$

where C is a constant. Inasmuch as the determinant of (6) is symmetric in the s_i, we may assume in deriving a further upper bound that

$$b \geqq s_1 \geqq s_2 \geqq \cdots \geqq s_\nu \geqq a.$$

The right side of (8) is then maximized by spacing the s_i uniformly between a and b. As a consequence we obtain

$$| \det (K(s_i, s_j)) | < C^\nu \nu^{\nu/2} \left[\frac{b - a}{\nu - 1}\right]^{\alpha(\nu-1)} \qquad (\nu > 1)$$

from which it follows that the estimate*

$$| d_\nu | < (\text{const.})^\nu \, \nu^{-\nu(\alpha+\frac{1}{2})} \quad (9)$$

is valid for the coefficients of the power series of (4).

The relation (9) implies that the order of the entire function $D(\lambda)$ satisfies

$$\mu < \frac{1}{\alpha + \frac{1}{2}}$$

which becomes less than 1 for $\alpha > \frac{1}{2}$. Since

$$d_1 = -\int_a^b K(s_1, s_1)\, ds_1 = -\text{tr}(K) \quad (10)$$

* We have used Stirling's expansion for the factorial function.

does not vanish by hypothesis, $D(\lambda)$ must be a *nonconstant* entire function of order less than 1 and hence must have at least one zero (see Ref. 13, p. 22 ff). It follows then that the integral equation has at least one eigenvalue.* Q.E.D.

For entire functions of finite order a general product expansion follows from the Hadamard factorization theorem. In view of the above results, therefore, $D(\lambda)$ may be written as the canonical product

$$D(\lambda) = \prod_{\nu=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_\nu}\right) \tag{11}$$

where $\lambda_1, \lambda_2, \cdots$ are the zeros of $D(\lambda)$ arranged in order of increasing modulus. This expansion converges absolutely and uniformly on compacta.

If we differentiate the two representations of $D(\lambda)$ given by (4) and (11), set $\lambda = 0$, and make use of (10) we obtain

$$\mathrm{tr}(K) = \sum_{\nu=1}^{\infty} \frac{1}{\lambda_\nu}. \tag{12}$$†

Thus kernels Hölder continuous in either variable with exponent $\alpha > \frac{1}{2}$ are one of the overlapping categories of nondegenerate kernels for which the expansion (11) and hence the relation (12) is valid. Other classes include

(*i*) Hermitian kernels with only a finite number of eigenvalues of one sign or the other (Mercer's Theorem), and

(*ii*) composite kernels of the form $K = K_1 K_2$ (Lalesco's result; see Ref. 15).

IV. ONE-DIMENSIONAL EQUATIONS FOR THE OPTICAL MASER

Careful analysis of an idealized diffraction model for the optical maser results in the following coupled integral equations for typical field quantities, such as the current densities:[5]

$$\begin{aligned} \Phi_1(x_1,y_1) &= \lambda_1 \int_{S_2} K_{12}(x_1,y_1;x_2,y_2)\, \Phi_2(x_2,y_2)\, dx_2 dy_2, \\ \Phi_2(x_2,y_2) &= \lambda_2 \int_{S_1} K_{21}(x_2,y_2;x_1,y_1) \Phi_1(x_1,y_1)\, dx_1 dy_1, \end{aligned} \tag{13}$$

* If the order μ of the Fredholm determinant $D(\lambda)$ satisfies $0 < \mu < 1$ then we can conclude that there exists a countably infinite set of eigenvalues of the equation (3) since entire functions of nonintegral order have an infinite set of zeros.

† This relation is not generally valid for arbitrary $\mathfrak{L}^2$ kernels. In fact it may be inferred from results of Salem[14] that there are continuous symmetric kernels with denumerably many eigenvalues for which $\sum_{\nu=1}^{\infty} (1/\lambda_\nu)$ does not even exist.

where

$$K_{12}(x_1, y_1; x_2, y_2) = K_{21}(x_2, y_2; x_1, y_1)$$
$$= \exp ik\,[(x_1 - x_2)^2 + (y_1 - y_2)^2 + h_1(x_1, y_1) + h_2(x_2, y_2)]. \quad (14)$$

In these relations, S_1 and S_2 are the mirror surfaces, $h_1(x_1, y_1)$ and $h_2(x_2, y_2)$ represent their respective departures from parallel planes, and k is a dimensionless parameter containing the wavelength as well as various geometrical dimensions such as the average mirror separation.

The solutions $\Phi_{1,2}$ of (13) are eigenfunctions which describe the field distributions at the reflectors of the possible normal modes of laser oscillation; λ_1 and λ_2 are the corresponding eigenvalues that specify the loss and phase shift which a propagating wave undergoes between the reflecting surfaces. Note in particular that these coupled equations (13) are single-transit relations; that is, they give the field at each mirror in terms of the reflected field at the other. They can, of course, be combined into a single integral equation with a composite kernel.

In the derivation of the above relations, the active maser material is assumed to be infinite in extent, homogeneous and isotropic. Diffraction effects at the edge of the reflecting surfaces are neglected. Moreover, the separation between the reflectors, as well as the radii of curvature of these surfaces, is taken to be large compared to typical transverse dimensions.

Although the integral equations (13) are two-dimensional, the preponderance of analytic work on this problem has been under the additional assumption that the laser kernels (14) can be adequately approximated by a product of functions of single variables. If the field quantities are correspondingly decomposable, namely if

$$\Phi(x,y) = X(x)\, Y(y)$$

or

$$\Phi(r,\theta) = R(r)\, \Theta(\theta),$$

the general problem may be reduced to consideration of integral equations in one independent variable. These equations take the form

$$\varphi_1(x_1) = \lambda_1 \int_{a_2}^{b_2} K_{12}(x_1,x_2)\, \varphi_2(x_2)\, dx_2,$$
$$\varphi_2(x_2) = \lambda_1 \int_{a_1}^{b_1} K_{21}(x_2,x_1)\, \varphi_1(x_1)\, dx_1, \quad (15)$$

with the single-transit kernels K given, for instance, by

$$K_{12}(x_1, x_2) = K_{21}(x_2, x_1) = \exp ik\,[(x_1 - x_2)^2 + p_1(x_1) + p_2(x_2)]. \quad (16)$$

Since these kernels are in general continuously differentiable functions of their arguments, we can use the main theorem of the preceding section directly to show the existence of eigenvalues whenever the trace is nonvanishing.

V. ANALYSIS OF SPECIAL REFLECTOR CONFIGURATIONS

In this section we examine the kernels associated with several well-known practical interferometer geometries and establish the general nonvanishing character of the traces.

5.1 *Plane Reflectors*

For rectangular plane reflecting surfaces which are mirror images of each other, the integral equation of interest is

$$\varphi(x) = \lambda \int_{-1}^{1} K(x,x')\,\varphi(x')\,dx'$$

with kernel

$$K(x,x') = \exp\,[ik(x - x')^2].$$

It is a trivial matter to verify that the trace of this differentiable function is 2, from which we infer the existence of eigenvalues of the above integral equation. In fact, for this kernel one can use parity arguments to show there are at least two eigenvalues for $k \neq 0$.[8]

5.2 *Circular Plane Reflectors*

When the plane reflecting surfaces have circular cross section the integral equation kernel becomes[4]

$$K(r,r') = J_n(krr')\,r' \exp\,[ik(r^2 + r'^2)/2] \qquad r,r' \text{ in } [0,1]$$

where J_n is a Bessel function of the 1st kind and nth order. The integer index n is indicative of an angular variation $e^{in\theta}$ which has been suppressed. The trace of the kernel K is proportional to

$$T_n = \int_0^1 J_n(k\tau)\,e^{ik\tau}\,d\tau,$$

and thus the nonvanishing of T_n will imply the existence of normal

modes of oscillation for the system. Using an integral representation of J_0, for example, we obtain

$$\begin{aligned} T_0 &= \int_0^1 \frac{1}{\pi} \int_{-1}^1 e^{ikrt}(1-t^2)^{-\frac{1}{2}} e^{ikr}\, dtd\tau \\ &= \frac{1}{\pi} \int_{-1}^1 (1-t^2)^{-\frac{1}{2}} \int_0^1 e^{ikr(1+t)}\, d\tau dt \\ &= \frac{1}{ik\pi} \int_{-1}^1 (1-t^2)^{-\frac{1}{2}} \left[\frac{e^{ik(1+t)}-1}{1+t}\right] dt. \end{aligned}$$

It is easy to see that for Im $(k) \geqq 0$, for instance, the real part of the integrand is negative almost everywhere. Hence the trace of the kernel does not vanish in this particular case, and we may draw our conclusion as to the existence of eigenvalues.

5.3 *Other Reflector Configurations*

For certain kernels, of course, there is little to be learned from application of our results on Hölder continuous functions. Such is the situation regarding the kernel

$$K(x,x') = e^{ikxx'}$$

associated with mirror image reflectors of square cross section, each having the curvature of a sphere centered at the center of the other reflector. As noted previously, this particular kernel gives rise to eigenvalues and eigenfunctions which may be expressed in terms of prolate spheroidal wave functions.[2]

At the same time, however, it is advantageous that the eigenvalue existence question can be easily settled for more general reflector geometries which do not exhibit as beneficial analytic properties as the confocal configuration. In particular, the kernel

$$K(x,x') = \exp ik[(x-x')^2 + p(x) + p(x')],$$

which pertains to mirror image square reflectors with shape function proportional to $p(x)$, has eigenvalues if

$$T = \int K(x,x)\, dx = \int e^{2ik\, p(x)}\, dx \neq 0.$$

The vanishing of T for practical geometries would certainly seem to be the exception rather than the rule.

5.4 *Composite Kernels*

The above examples show how our main theorem can be used to establish simply yet rigorously the existence of eigenvalues and eigenfunctions for the one-dimensional laser kernels arising when the reflectors are mirror images of each other. For more generalized configurations in which the reflecting surfaces may be of unequal size and curvature, the applicable kernels are of a composite nature (see Refs. 3 and 5). In view of this one might choose to reason from Lalesco's results on composite kernels mentioned earlier rather than from our main theorem. This would be an acceptable method of attack. However, since (12), relating the trace to the sum of reciprocal eigenvalues, is valid in both situations, a verification of the nonvanishing character of the kernel traces is needed in either case. As a last illustrative example we shall provide this verification for the integral equations associated with asymmetric spherical reflectors of arbitrary curvature.

Let $a_1 = -b_1$, $a_2 = -b_2$, $p_1(x_1) = \alpha x_1^2$ and $p_2(x_2) = \beta x_2^2$. The one-dimensional integral equations (15) then become appropriate for analysis of an idealized interferometer having two rectangular mirrors of unequal size and unequal curvatures. As usual, these two coupled equations (15) can be combined into a single integral equation for either φ_1 or φ_2. Moreover, this new integral equation may then be split apart into two subsidiary equations according to whether the eigenfunction modes are even or odd. The kernels resulting from this division are given by

$$K_{e,o}(x,y) = 2\int_{-b_2}^{b_2} \begin{Bmatrix} \cos 2kx'y \\ i \sin 2kx'y \end{Bmatrix} \exp\{ik[(x^2 + y^2)(1 + \alpha) + 2x'^2(1 + \beta) - 2x'x]\}\, dx' \tag{17}$$

and have traces

$$\mathrm{tr}(K_{e,o}) = 2\int_0^{b_1}\int_{-b_2}^{b_2} \begin{Bmatrix} \cos 2kx'x \\ i \sin 2kx'x \end{Bmatrix} \exp\{2ik[x^2(1 + \alpha) + x'^2(1 + \beta) - x'x]\}\, dx'dx. \tag{18}$$

It is easy to show that at least one of these traces is different from zero for real k and arbitrary curvatures α, β.

Note first that

$$\mathrm{tr}(K_e) - \mathrm{tr}(K_o) = 2\int_0^{b_1}\int_{b_2}^{b_2} \exp\{2ik[x^2(1 + \alpha) + x'^2(1 + \beta)]\}\, dx'dx$$

$$= 4\left[\int_0^{b_1} \exp\{2ik(1+\alpha)x^2\}dx\right] \cdot \left[\int_0^{b_2} \exp\{2ik(1+\beta)x'^2\}\,dx'\right].$$

Now neither of the two bracketed terms on the right-hand side vanishes, since the real parts of these Fresnel integrals are positive for real k, α, β. Thus, the difference between the two traces, and hence at least one of the traces itself, is different from zero [one suspects, of course, that both of the kernels (17) have nonvanishing traces]. Although this argument gives no measure of the loss to be expected with any individual eigenmode, it does show that normal modes of oscillation exist for this arbitrary asymmetric spherical configuration.

VI. ACKNOWLEDGMENTS

We are indebted to Menahem M. Schiffer of Stanford University, who first introduced us to the study of integral equations, and to Samuel P. Morgan, whose constructive comments were beneficial in this work.

REFERENCES

1. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter through Optical Wavelength Masers, B.S.T.J., **40,** Mar., 1961, pp. 489–508.
2. Slepian, D., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — I:, B.S.T.J., **40,** Jan., 1961, pp. 43–63.
3. Boyd, G. D., and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., **41,** July, 1962, pp. 1347–1369.
4. Fox, A. G., and Li, Tingye, Resonant Modes in a Maser Interferometer, B.S.T.J., **40,** Mar., 1961, pp. 453–488.
5. Fox, A. G., and Li, Tingye, Modes in a Maser Interferometer with Curved and Tilted Mirrors, Proc. IEEE, **51,** Jan., 1963, pp. 80–89.
6. Li, Tingye, Mode Selection in an Aperture-Limited Concentric Maser Interferometer, B.S.T.J., **42,** Nov., 1963, pp. 2609–2620.
7. Morgan, S. P., On the Integral Equations of Laser Theory, IEEE Trans. on Microwave Theory and Techniques, **MTT-11**, May, 1963, pp. 191–193.
8. Newman, D. J., and Morgan, S. P., Existence of Eigenvalues of a Class of Integral Equations Arising in Laser Theory, B.S.T.J., **43,** Jan., 1964, pp. 113–126.
9. Swann, D. W., Generalization of a Theorem of Hammerstein and the "Natural" Expansion of exp $[ik(s-t)^2]$, to appear.
10. Streifer, W., and Gamo, H., On the Hilbert-Schmidt Expansion for Optical Resonator Modes, Proc. Polytechnic Inst. of Brooklyn Symp. on Quasi-Optics, June, 1964.
11. Smithies, F., *Integral Equations*, Cambridge University Press, 1962.
12. Fredholm, I., Sur une classe d'équations fonctionnelles, Acta Math., **27,** 1903, pp. 365–390.
13. Boas, R. P., *Entire Functions*, Academic Press, New York, 1954.
14. Salem, R., On a Problem of Smithies, Nederl. Akad. Wetensch. Proc., Ser. A, **57,** 1954, pp. 403–407.
15. Hille, E., and Tamarkin, J., On the Characteristic Values of Linear Integral Equations, Acta Math., **57,** 1931, pp. 1–76.

A Model of the Switched Telephone Network for Data Communications

By E. O. ELLIOTT

(Manuscript received August 20, 1964)

The error statistics from data-transmission field tests on the telephone network may be compactly represented by about one dozen parameters. These relate to a model of the telephone network in which there are three distinct channels. The errors in binary data on each channel are produced by a renewal process in which a bit-error is a renewal event. The mixture of three such channels allows a close fit to the error statistics for a large range of block lengths. It is not implied that the errors on the telephone network are actually produced by such processes, but merely that they may be conveniently and compactly represented by them.

Use of this model simplifies the analysis of error-control systems and the determination of error rates for error-control codes. In this paper the model is applied to study the effect that interleaving (time division multiplexing) has on the effectiveness of error-correcting codes.

I. INTRODUCTION

In the study of errors on data communication channels, several mathematical models of the error process have been proposed. Gilbert[1] proposed a two-state Markov process, and Berger and Mandelbrot[2] have employed a Pareto distribution to fit data collected from the German telephone network. Sussman[3] has applied a Pareto fitting to part of the Alexander-Gryb-Nast data.[4] Common to the models of Gilbert and Berger-Mandelbrot is the assumption that the error process is of the renewal type wherein the state of a bit-error is the renewal event whose occurrence frees the process from dependence upon past history and starts it anew. In such models the distributions of lengths of error-free intervals (gaps) determine the processes, because the lengths of the gaps before and after an error are independently distributed. One may calculate from this distribution the probabilities $P(m,n)$ that m bit-errors occur in a block of n consecutive bits. These probabilities are useful

in the analysis of error-control methods for data communications systems.[5]

On the telephone network, as exemplified by the Alexander-Gryb-Nast (AGN) data, the $P(m,n)$ for individual calls are, as noted in Ref. 5, quite diverse in nature. This suggests that there are several error processes involved. In fact, the combined AGN data cannot be described by just one error process of the renewal type but require the mixture of several such processes for a satisfactory description. The mixture of a collection of processes is determined by a corresponding set of channels on which the separate error processes act and specification of the probabilities of being assigned the various channels when placing a call (the assignment, and hence the error process acting, is fixed for the duration of the call).

The purpose of the present paper is to show that in this way the use of three distributions for gap lengths can yield satisfactory approximations to the probabilities $P(m,n)$ for the telephone network. This cannot be accomplished with just one or two distributions, and conceivably four or more might be required in some cases. The $P(m,n)$ for renewal processes depend heavily on the first few values of the gap-length distribution. Because of this there is some choice of distributions for satisfactory models of the telephone network. Noting how various gap-length distributions affect the $P(m,n)$ distributions, we select three gap-length distributions with appropriate weightings to represent the combined switched telephone network. This selection gives excellent agreement to $P(m,n)$ over a wide range of block lengths n. To exemplify further this means of representing the telephone network we apply it also to the Townsend-Watts (TW) data given in Ref. 6.

In the present paper we do not attempt to optimize the degree to which the models represent the telephone network. Rather, we attempt to demonstrate the feasibility of such representations and note that a better accuracy of fit would hardly affect the applications suggested. Also, no attempt is made to associate the parameters with particular causes, such as type of exchange or calling distance, etc. These goals would be the object for future work. We do suggest, however, that dropouts (momentary open-line conditions) and the test words used in field tests are at least partially the cause for the hump in the $P(m,n)$ curves at m near $n/2$.

II. RENEWAL-TYPE ERROR PROCESSES

For a renewal error process, the lengths of successive gaps are independent and distributed according to a common distribution. Let $p(k)$

be the probability that a gap length is $k - 1$, i.e., $p(k) = \Pr(0^{k-1}1 \mid 1)$ where 1 denotes an error-bit, 0 a correct bit and 0^i denotes i consecutive 0's.

Let

$$P(k) = \sum_{m=k}^{\infty} p(m)$$

so that $P(k)$ is the probability that at least $k - 1$ 0's follow a given error [i.e., $P(k) = \Pr(0^{k-1} \mid 1)$]. Now, if p_1 is the unconditional probability of a bit-error, then $p_1P(k) = \Pr(1)\Pr(0^{k-1} \mid 1)$ is the probability of 10^{k-1}. But, because of the independence among gap-lengths of a renewal process, order is irrelevant and it is clear that the events 10^{k-1} and $0^{k-1}1$ are equiprobable. Hence, $p_1P(k) = \Pr(0^{k-1}1)$. To obtain the value of p_1, note that $p_1 = 1/\bar{k}$ where $\bar{k}$, the average distance to the next error, is equal to

$$\sum_{k=1}^{\infty} kp(k).$$

The probabilities of individual error patterns of a renewal process are easily calculated (but we do not make use of these here). For example, consider a block ζ of n consecutive bits which contains m bit-errors and, as in Ref. 5, p. 1985, let a be the number of 0's before the first 1 in ζ, c the number of 0's following the last 1 in ζ and b_i ($i = 1, \cdots, m - 1$) be the number of 0's between consecutive 1's in ζ. Then, the probability of ζ's occurrence is given by

$$\Pr(\zeta) = p_1P(a + 1)\left\{\prod_{i=1}^{m-1} p(b_i + 1)\right\}P(c + 1).$$

Calculations of the above sort may be of use in evaluating both error-correcting and error-detecting codes on renewal-type channels. For more general, but approximate, applications, the probabilities $P(m,n)$ that m bit-errors occur in a block of length n are of use. To calculate these we may use recurrence relations or generating functions as follows.

First, let $R(m,n)$ be the probability that $m - 1$ errors occur in the next $n - 1$ bits following an error. Thus, $R(1,n) = P(n)$ for $n \geqq 1$, and

$$R(m,n) = \sum_{k=1}^{n-m+1} p(k)R(m - 1,n - k)$$

for $2 \leqq m \leqq n$.

Now,

$$P(m,n) = \sum_{k=1}^{n-m+1} p_1 P(k) R(m, n-k+1)$$

whenever $1 \leqq m \leqq n$.

Computer programs for computing $P(m,n)$ from the above recurrence relations have been written and used in obtaining the data presented later in this paper.

An alternate approach to the above relation is through generating functions. If we let $g(z)$ and $G(z)$ be the generating functions associated with $p(k)$ and $P(k+1)$ respectively

$$\left(\text{i.e., } g(z) = \sum_{k=1}^{\infty} p(k) z^k \quad \text{and} \quad G(z) = \sum_{k=0}^{\infty} P(k+1) z^k\right)$$

then, from Ref. 7, p. 249, we have

$$G(z) = \frac{1 - g(z)}{1 - z}.$$

Letting

$$H_m(z) = \sum_{n=m}^{\infty} P(m,n) z^n$$

we obtain that

$$H_m(z) = p_1 z G(z) g(z)^{m-1} G(z)$$

considering that m errors involve $m-1$ gaps in a total number of bits adding up to n and that the generating function for a convolution of variables is the product of their associated generating functions. [$p_1 z G(z)$ is the generating function for the probabilities of the events $0^{k-1}1$, $g(z)$ is that of $0^{k-1}1 \mid 1$ and $G(z)$ is that of $0^k \mid 1$.]

Thus, we obtain

$$H_m(z) = p_1 z \left\{\frac{1 - g(z)}{1 - z}\right\}^2 g(z)^{m-1}.$$

Calculation of $P(m,n)$ from this generating function is rather inconvenient. The recurrence equations are generally preferred in practice.

III. A REPRESENTATION OF THE TELEPHONE NETWORK

In both of the data-transmission field-test programs on the telephone network, data calls were placed on a variety of circuits and bit-errors were recorded. In Refs. 5 and 6 the composite effects of these are represented by the $P(m,n)$ probabilities. Fig. 1 shows $P(m,31)$ for these two

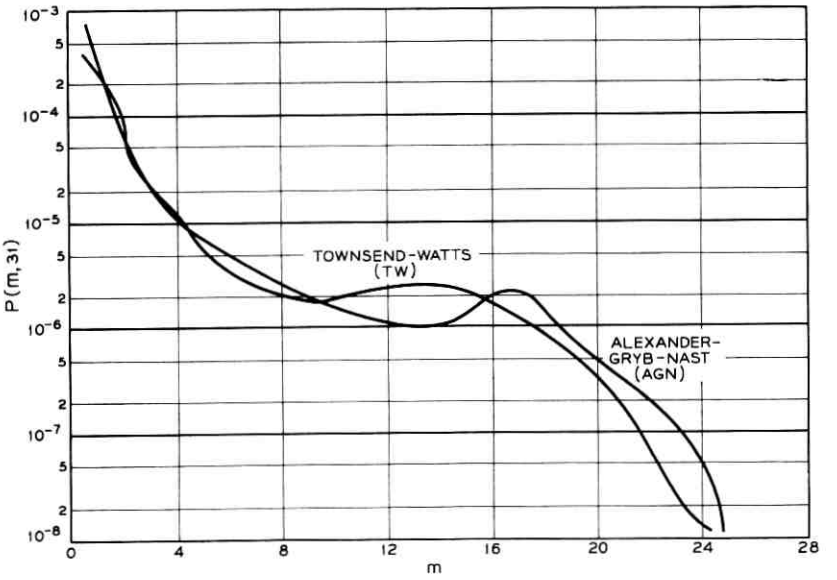


Fig. 1 — $P(m,31)$ for field test data.

field tests and is typical of $P(m,n)$ for other intermediate block lengths n. These curves appear to have three separate segments: an initial segment with a steep slope, an intermediate segment with a smaller slope, and a terminal hump and tail. Using the recurrence equations of the previous section, the $P(m,31)$ curves for the three gap-length distributions (determined by trial and error) given in Table I were calculated and are displayed in Fig. 2. The unconditional bit-error rate for each curve is taken to be that of the AGN data. In so doing we essentially assume that the tails of the gap-length distributions are appropriately tailored. The tails of these distributions, of course, do not influence $P(m,n)$ when n is not too large. Each of these three curves is more-or-less parallel to the respective first, second or third segment of the $P(m,31)$ curve for the AGN data. They are to be added together, after multiplication by suitable weighting factors, to produce our approximation to the AGN curve. By trial and error, we find that weighting factors of 50, 25, and 25 per cent respectively give the close fit which is shown in Fig. 3 for block length 31 and again in Figs. 4–7 for some other block lengths. This trial and error procedure of finding curves of the right shape and then appropriate weighting factors represents a simple attempt to approximate the $P(m,n)$ curves of the AGN data

TABLE I — GAP-LENGTH DISTRIBUTIONS $p(k)$ FOR AGN MODEL

k =	1	2	3	4	5	>5
Initial segment	0.12	0.06	0.0	0.0	0.0	≈0
Intermediate segment	0.40	0.20	0.10	0.05	0.0	≈0
Hump	0.56	0.24	0.06	0.0	0.12	≈0

over a wide range of block lengths n. It would be desirable to use analytic methods instead of trial and error in obtaining such approximations, but the number of parameters involved is large and the analytic expressions for $P(m,n)$ are cumbersome. Undoubtedly, appropriate programming techniques can be developed to improve upon our trial and error method.

The gap-length distributions and weighting factors given in Table II furnish an approximate model for the TW data. Figs. 8–10 compare the original and the model at block lengths 10, 31, and 63.

IV. CHOICE AND INTERPRETATION OF COMPONENT DISTRIBUTIONS

Suppose the field test data we wish to represent by a mixture of different renewal processes have an unconditional bit-error rate p_1 and suppose that gap-length distributions $p^i(k)$ (the superscript is not an

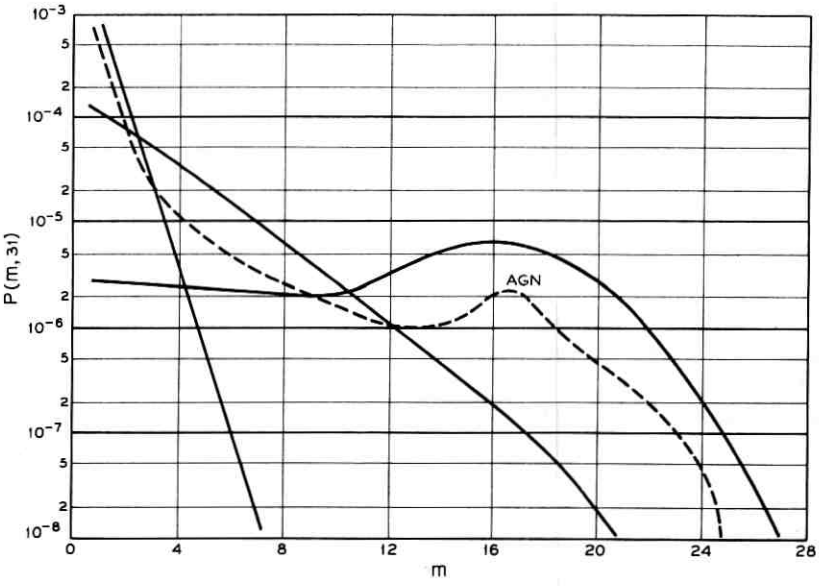


Fig. 2 — $P(m,31)$ for components of model.

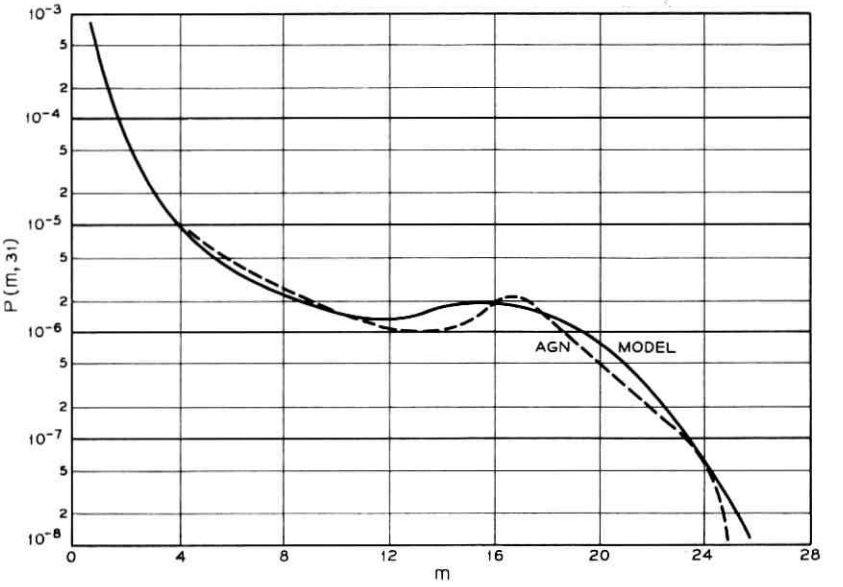


Fig. 3 — $P(m,31)$ for model of AGN data.

exponent) and weighting factors λ_i ($i = 1, \cdots, J$) have been decided upon. (Determination of $p^i(k)$ and λ_i will be discussed subsequently.) Then, let $P^i(m,n)$ be calculated for each distribution $p^i(k)$ using the prescribed error rate p_1. For the model we then take

$$P(m,n) = \sum_{i=1}^{J} \lambda_i P^i(m,n) .$$

Use of the common value p_1 in computing $P^i(m,n)$ does not imply that each distribution $p^i(k)$ has this bit-error rate, but is just a device to assure that the model has the same unconditional bit-error rate as the field-test data. In fact, since $p^i(k)$ is generally specified for only a few small values of k, we assume that $p^i(k)$ for larger values of k is distributed so that its real bit-error rate $p_{1,i}$ may be different from p_1, and it is not used explicitly in our model. Incidentally, $\lambda_i p_1$ represents the portion of the total error rate attributable to the channels with gap-length distribution $p^i(k)$ and we could say that the percentage of channels of this type is λ'_i where $\lambda'_i p_{1,i} = \lambda_i p_1$.

To choose candidate distributions $p^i(k)$ we must first observe some principles, pertaining separately to the segments $i = 1,2,3$, which are noted in the next few paragraphs. We begin by examining the simplest

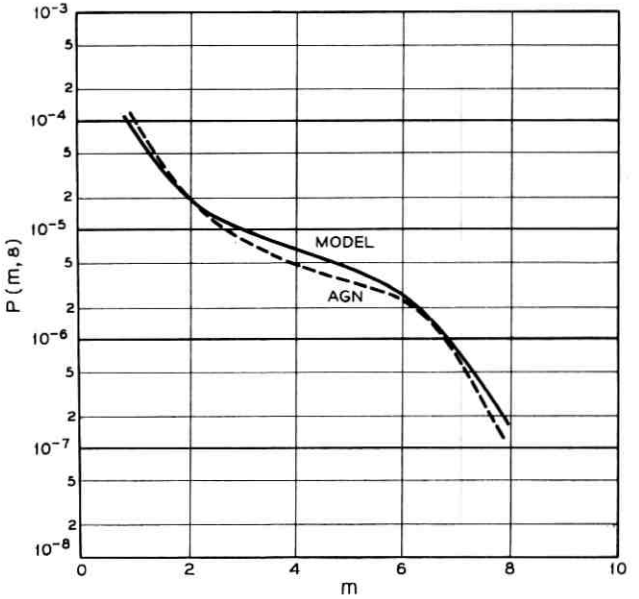


Fig. 4 — $P(m,8)$ for model of AGN data.

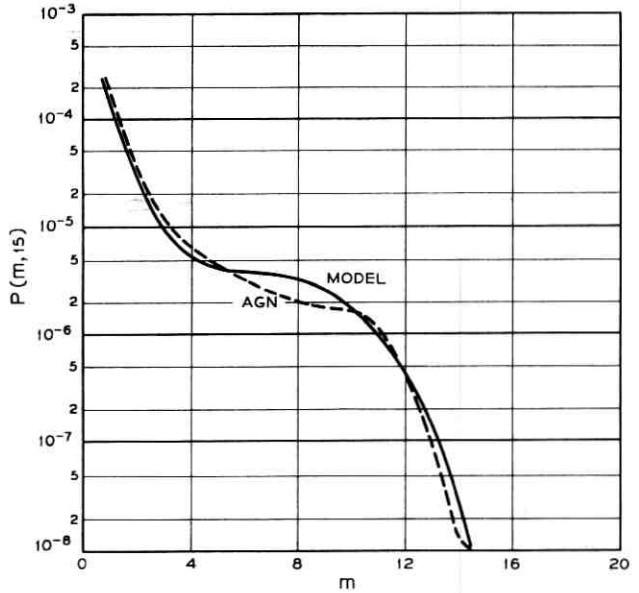


Fig. 5 — $P(m,15)$ for model of AGN data.

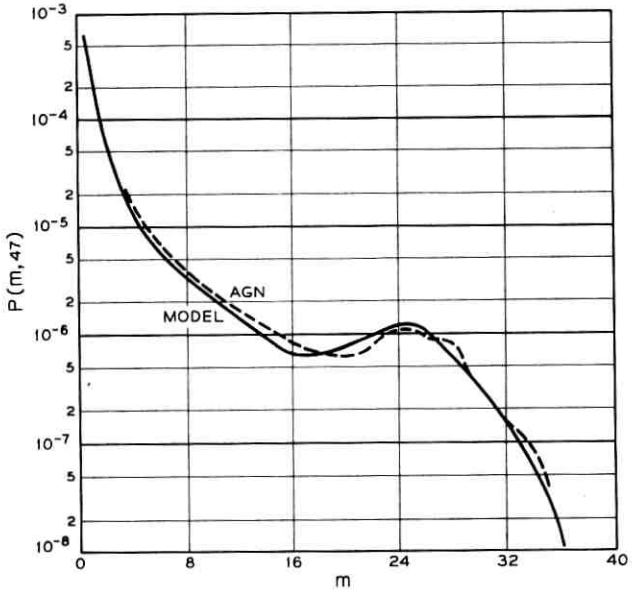


Fig. 6 — $P(m,47)$ for model of AGN data.

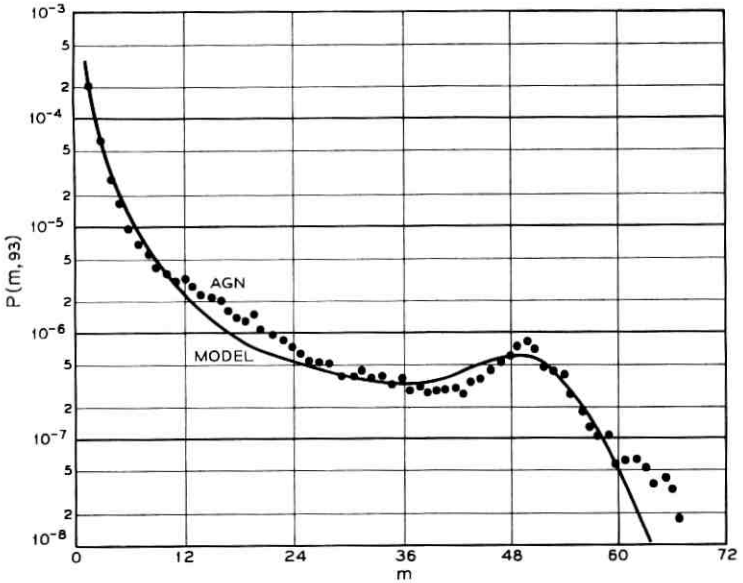


Fig. 7 — $P(m,93)$ for model of AGN data.

TABLE II — GAP-LENGTH DISTRIBUTIONS $p(k)$ AND WEIGHTING FACTORS FOR TW MODEL

Weighting Factor	λ	k					
		1	2	3	4	5	>5
Initial segment	57%	0.20	0.10	0.0	0.0	0.0	≈0
Intermediate segment	18%	0.35	0.25	0.15	0.05	0.0	≈0
Hump	25%	0.45	0.25	0.15	0.10	0.03	≈0

case analytically, namely that of a gap-length distribution $p(k)$ such that $p(1) = \alpha$ and $p(k) \approx 0$ for $k > 1$. Then the generating function $g(z)$ for $p(k)$ is essentially αz and

$$H_m(z) = p_1 z \left[\frac{1 - \alpha z}{1 - z}\right]^2 (\alpha z)^{m-1}.$$

Determining the coefficient of z^n in the above [using $(1 - z)^{-2} = 1 + 2z + 3z^2 + 4z^3 + \cdots$] we obtain

$$P(m,n) = p_1 \alpha^{m-1}[(n - m + 1) - 2\alpha(n - m) + \alpha^2(n - m - 1)]$$

which in a logarithmic plot as a function of m is almost a straight line.

A distribution of this type with α small, or a slight modification

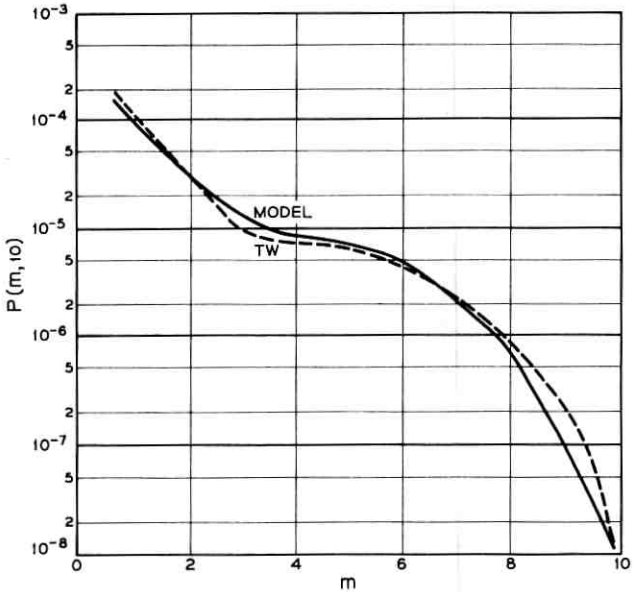


Fig. 8 — $P(m,10)$ for model of TW data.

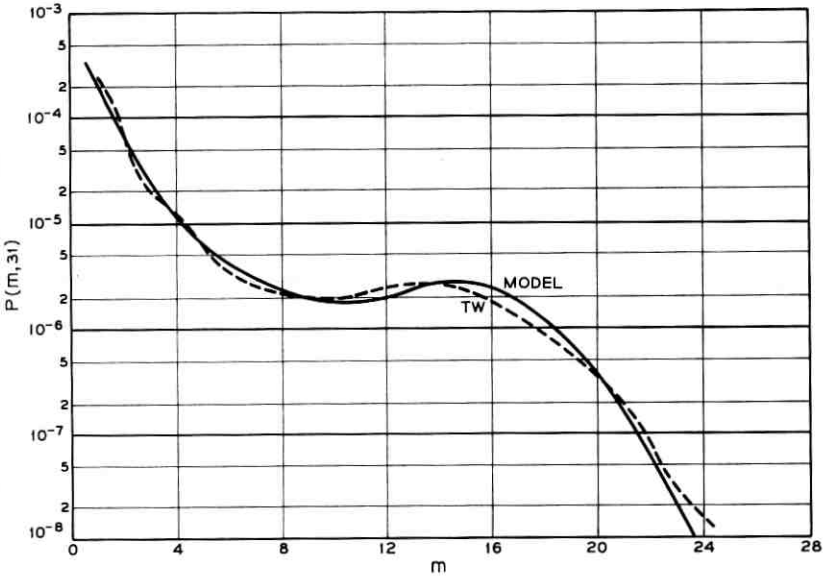


Fig. 9 — $P(m,31)$ for model of TW data.

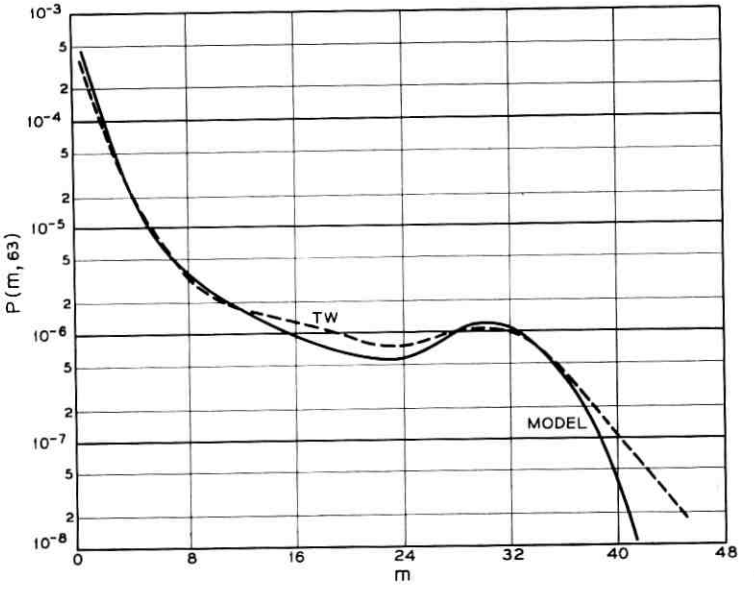


Fig. 10 — $P(m,63)$ for model of TW data.

thereof, may be useful in matching the initial segment of field test data as in the foregoing examples. In general, the initial segment of a $P(m,n)$ curve which has the same general shape as our two field test examples is determined by a gap-length distribution $p(k)$ having 5 to 40 per cent of the probability spread over the first three or four values of k. It is reasonable to assume also that $p(k)$ is monotonically decreasing and that $p(k) \approx 0$ for k much larger than 4.

For the intermediate segment of such a curve, the gap-length distribution may contain between 70 and 80 per cent of the probability in the first four or five terms. A definite hump begins to appear in the $P(m,n)$ curve when much over 90 per cent is contained in the same range. Amounts of 95 to 99 per cent produce humps as extreme as we note in our examples.

The exact way in which $p(k)$ is distributed over these first few terms can influence the shape of the $P(m,n)$ curve considerably. In general, large values of $p(1)$ cause the $P(m,n)$ to remain large for a more extended range of values of m. Beyond these few general remarks, the process of fitting remains a matter of trial and error. First we find gap-length distributions which give rise to $P(m,n)$ curves which approximate in shape the various segments of the curve we are trying to match. Then weighting factors are chosen so that when the various components are so weighted and added together, they yield numerical agreement with the desired curve.

Berger and Mandelbrot[2] and Sussman[3] have claimed that the error processes on the telephone network are indeed of the renewal type and they have some data[2] to support this view. We have not investigated this matter, but we do note that for the AGN data the composite gap-length distribution given in Table III is approximated reasonably well by that for our model. Our model, however, has $p(k) \approx 0$ for $k > 5$, whereas for the AGN data $p(k)\ 0 \not\approx$ for $k > 5$. These observations neither confirm or deny the existence of renewal processes on the telephone network. The accuracy with which our model approximates the field data is, however, indirect support for the notion that the error processes are at least not widely different from renewal processes.

The humps in the $P(m,n)$ curves at $m \approx n/2$ are rather remarkable. We have suspected that they are at least partially due to drop-outs and to the nature of the error-recording procedures in the field tests. For example, in the AGN tests during a drop-out, the error pattern recorded would coincide with the test word (Ref. 8, p. 1402) and would appear as repetitions of

$$1111011110010110000100000110110.$$

TABLE III — GAP-LENGTH DISTRIBUTIONS FOR AGN DATA AND MODEL

k =	1	2	3	4	5
AGN data	0.24	0.09	0.05	0.02	0.03
Model	0.20	0.14	0.04	0.01	0.03

When this pattern is present, the gap-length distribution induced would be that given in Table IV. This is essentially the distribution we have used to produce the hump in our model for the AGN data (cf. Table I). We found that it gave a slightly better fit than some other distributions we tried, but other distributions did yield rather good fittings. This is not sufficient evidence to conclude that the hump is entirely due to drop-outs but it does support the hypothesis that it is at least partially caused by them.

One other tentative interpretation of our models for the AGN and TW data lies in the gap-length distributions for their initial segments. The term $p(1)$ for the TW data is almost twice as large as that for the AGN data, which means that double errors would be more prevalent in the TW data. This is consistent with the occurrence of dibit errors for the four-phase data set employed in the TW tests.[6]

So many gap-length distributions seem to give reasonable approximations to the intermediate segment of the $P(m,n)$ curves of these field tests that it is difficult to ascribe much significance to them. They do, however, represent instances where errors have high probabilities of following other errors, thereby producing bursts. Very short drop-outs would offer one explanation, but there are no doubt others.

The accuracy of our approximations is certainly sufficient for many purposes, and in particular for the estimation of error rates for codes by the methods given in Ref. 5. An advantage of the models which is somewhat independent of accuracy is the following. When the block length n is very large there are two defects in the $P(m,n)$ values obtained directly from the field-test data. First, the plot of $P(m,n)$ becomes erratic due to the small sample size afforded by the field-test data, and, second, the computer processing time to obtain $P(m,n)$ becomes excessive. On the other hand, $P(m,n)$ computed from the model yields a

TABLE IV — GAP-LENGTH DISTRIBUTION FOR TEST WORD

k	1	2	3	4	5	>5
$p(k)$	0.563	0.250	0.062	0.0	0.125	0.0

smooth curve with very little time required in the computations. This appears to be a significant advantage of representing the telephone network by way of mathematical models.

V. ERROR RATES ON THE TELEPHONE NETWORK

The use of $P(m,n)$ in estimating error rates for error-detecting codes is described in Ref. 5. It has also been used for some error-correcting codes in Ref. 9. For error-correcting codes, the probability P_e of incorrect decoding may be conveniently and usefully bounded in some cases. For example, if the code is capable of correcting all $(m-1)$-fold (or fewer) errors, then certainly $P_e \leqq P(\geqq m,n)$ where

$$P(\geqq m,n) = \sum_{i=m}^{n} P(i,n)$$

and n is the code's block length.

In Ref. 5 we find another use for the probabilities $P(\geqq m,n)$. For an error-detecting code used with retransmissions for error correction, let P_u be the probability of an undetected error and P_r be the probability of retransmission. Then,

$$P_u \approx 2^{-c} P(\geqq d,n)$$

and

$$P_r \approx P(\geqq 1,n)$$

where d is the minimum distance of the code and c is the number of check bits in each code word. The above approximation for P_u is best if c is small, n is moderately large, and the check bits are not too trivial. It may be used in some other cases but with special caution if either extremely low error rates are desired or n is quite large.

Using the model for the AGN data, $P(\geqq m,n)$ vs n has been computed and displayed in Fig. 11 for $m = 1, \cdots, 10$. Fig. 11 is useful for the kind of estimates indicated above and for other considerations in error-control systems (e.g., synchronization).

VI. THE EFFECT OF INTERLEAVING ON ERROR RATES

Time division multiplexing (interleaving) has often been considered as a means of enhancing the error-control effectiveness of error-correcting codes. Its effect on the error statistics of a Gilbert burst-noise channel were noted in Ref. 5, p. 1987.

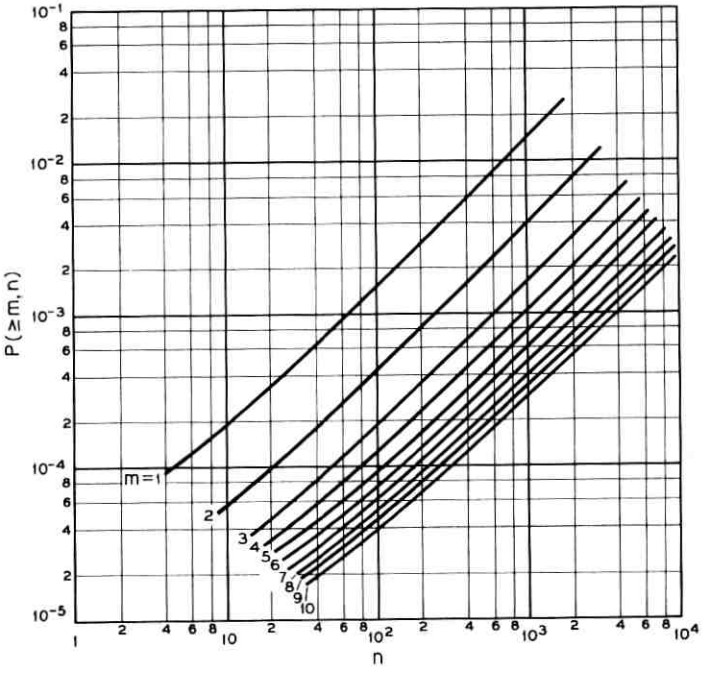


Fig. 11 — $P(\geq m,n)$ for $m = 1, \cdots, 10$.

With interleaving, the blocked bits from the data source are rearranged (by some delay and storage device) and put onto the line so that of two originally adjacent bits in a block, the second is the tth bit on line following the first. The number t will be called the "interleaving constant." When $t = 1$ there is no interleaving. We let n be the block length, and for error-rate considerations we are concerned with the probabilities $P_t(m,n)$ that m bit errors occur among the n bits of an interleaved data block. For $P_1(m,n)$ we write simply $P(m,n)$ as before. Thus, $P_t(m,n)$ is the probability of m errors among n bits which are equally spaced with $t - 1$ other bits between each two.

To obtain m errors among n bits spaced $t - 1$ bits apart, we must have r errors in the total block of tn bits where $m \leqq r \leqq (t - 1)\, n + m$. Given a total of r errors in tn bits, there are $\binom{r}{m}\binom{tn - r}{n - m}$ collections of n bits (from the total) containing m errors. There are, however, $\binom{tn}{n}$

possible collections of n bits. The probability that n bits selected from the tn bits contain m errors is hence

$$\binom{r}{m}\binom{tn-r}{n-m} \div \binom{tn}{n}$$

and, assuming that a sample of n regularly spaced bits is similar to a random sample of n bits, we may conclude that

$$P_t(m,n) \approx \sum_{r=m}^{(t-1)n+m} \frac{\binom{r}{m}\binom{tn-r}{n-m}}{\binom{tn}{n}} P(r,tn). \tag{1}$$

This formula could be used to approximate $P_t(m,n)$ from the $P(r,tn)$ values, except that $N = tn$ may be quite large and the computation of $P(r,N)$ then becomes infeasible even with the efficient programs used in connection with our mathematical model of the telephone network. Let us abandon this approach.

Now, $P(r,N)$ is expressed by

$$P(r,N) = \sum_{i=1}^{3} \lambda_i P^i(r,N)$$

where $P^i(r,N)$ is associated with the renewal channel determined by the gap-length distribution $p^i(j)$. If we interleave on one of these renewal channels, we effectively have another renewal channel determined by a modified gap-length distribution $p_t{}^i(j)$. Expressions for this will be given below. Thus,

$$P_t(m,n) = \sum_{i=1}^{3} \lambda_i P_t{}^i(m,n) \tag{2}$$

where $P_t{}^i(m,n)$ is obtained from $p_t{}^i(j)$ in the same way that $p^i(r,N)$ is obtained from $p^i(j)$ and involves only modest computations.

Using (2) we are then capable of computing the $P_t(m,n)$ values for the switched telephone network that are displayed and discussed below.

For a renewal channel the gap-length distribution $p(j)$ specifies the probability $Pr(0^{j-1}1 \mid 1)$ that, given a bit error, the next error is the jth following bit. (The superscript i is dropped from $p(j)$ here, since the three cases $i = 1$ to 3 are treated the same.) Let a be the autocorrelation function for the process so that $a(j) = \Pr(\alpha 1 \mid 1)$ where α is any binary word of length $j - 1$. Then

$$a(1) = p(1)$$

and

$$a(j) = p(j) + \sum_{s=1}^{j-1} p(s)a(j-s) \tag{3}$$

for $j > 1$.

Now, if we interleave with a constant t, error bits remain renewal events and $p_t(j)$ is the probability

$$\Pr\ (\alpha_1 0 \alpha_2 0 \cdots \alpha_{j-1} 0 \alpha_j 1 \mid 1)$$

that looking at every tth bit there are $j - 1$ correct bits preceding the next error (α_1, α_2, $\cdots$, α_j are arbitrary binary sequences of length $t - 1$).

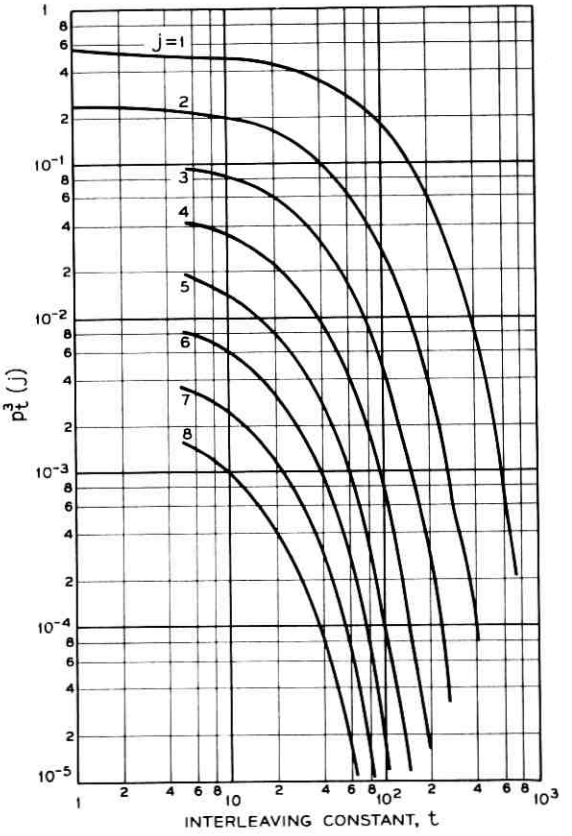


Fig. 12 — Gap-length distribution $p_t^3(j)$ vs t.

Thus,

$$p_t(1) = a(t)$$

and

$$p_t(j) = a(tj) - \sum_{s=1}^{j-1} a(ts)p_t(j - s) \tag{4}$$

for $j > 1$.

This last equation follows from the fact that the first term on the right is the probability that the tjth bit following an error is also in error, and the second term is the total probability that at least one of the tth, $(2t)$th, $\cdots$, $(j - 1)t$th bits is in error. Their difference then is $p_t(j)$.

We reintroduce the superscript i $(=1,2,3)$ on $p^i(j)$ to denote respectively the initial, intermediate and hump gap-length distributions given in Table I. The corresponding autocorrelation functions determined by (3) turn out to be well approximated (at multiples of 25) by the following formulae.

$$a^1(j\cdot 25) = 1.43 \times 10^{-13}\{2.295 \times 10^{-13}\}^{j-1}$$
$$a^2(j\cdot 25) = 0.990\{1.863 \times 10^{-2}\}^{j-1}$$
$$a^3(j\cdot 25) = 0.407\{0.763\}^{j-1}.$$

TABLE V — GAP-LENGTH DISTRIBUTIONS $p_t^i(j)$

t	j	$i = 1$	2	3
25	1	0	0.01	0.41
	2	0	0	0.14
	3	0	0	0.05
	4	0	0	0.02
	5	0	0	0.01
50	1	0	0.0002	0.31
	2	0	0	0.08
	3	0	0	0.02
	4	0	0	0.01
100	1	0	0	0.18
	2	0	0	0.03
	3	0	0	0.005
200	1	0	0	0.06
	2	0	0	0.004
	3	0	0	0.0002
400	1	0	0	0.008
	2	0	0	0.0001

Using (4), these give rise to the following gap-length distributions $p_t^i(j)$ (for $t = 25$, 50, 100, 200 and 400) shown in Table V (therein 0 indicates a number less than 10^{-4}).

To illustrate further the dependence on t we have plotted $p_t^3(j)$ ($j = 1$ to 8) in Fig. 12.

Using the gap-length distributions in Table V, we have computed $P_t(n,m)$ for $n = 15$, 31, 93 to obtain $P_t(\geqq m,n)$. These are summarized in Figs. 13 and 14, which give $P_t(\geqq m,31)$ vs t for $m = 1$ to 7 and $P_t(\geqq 3,n)$ vs n for the values of t in Table V.

The correction of all single and double errors is generally considered to be a reasonable procedure for capable codes of almost any block length. For such codes, the probability P_e of a decoding error cannot

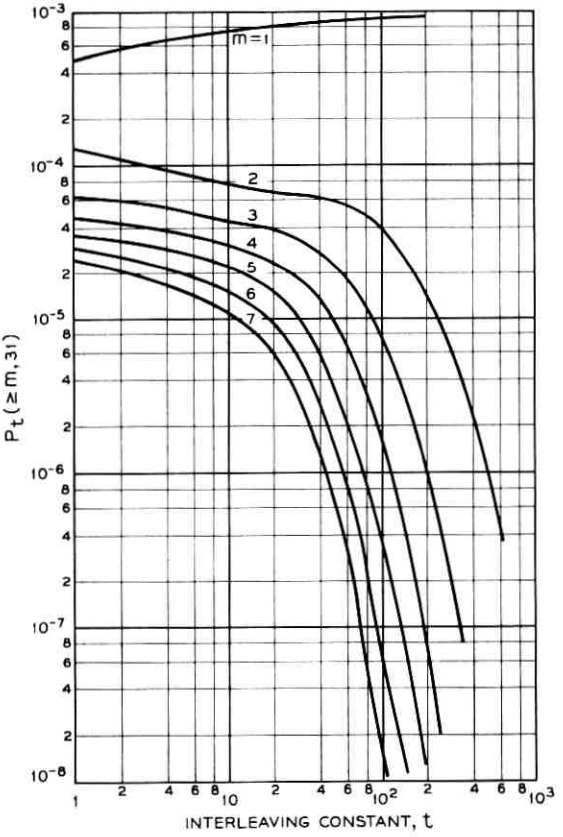


Fig. 13 — $P_t(\geqq m,31)$ vs t for $m = 1, \cdots, 7$.

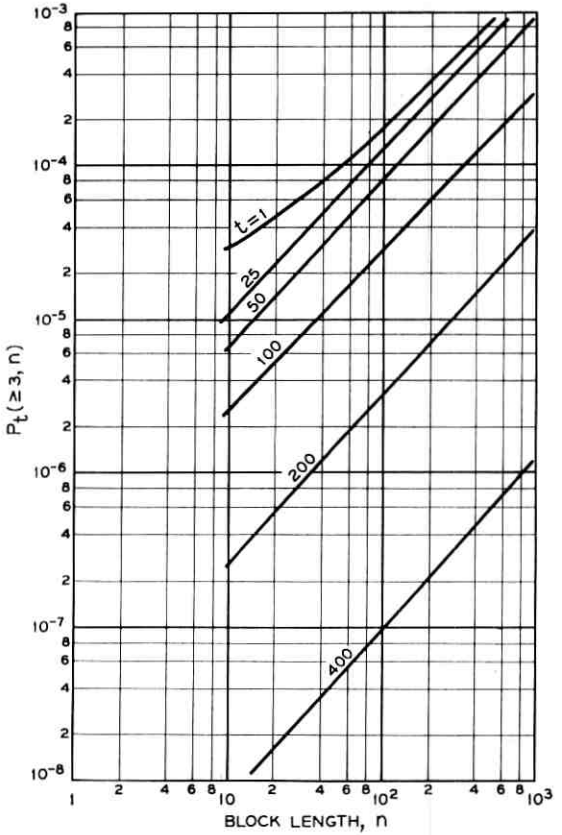


Fig. 14 — $P_t(\geqq 3,n)$ vs n.

exceed $P_t(\geqq 3,n)$, where n is the code's block length (and the interleaving constant is t).

Noting Fig. 13 and considering the Bose-Chaudhuri (31, 21) code,[10] which corrects all double errors, we conclude that interleaving with $t = 300$ would provide an error rate P_e of 10^{-7} or less. Such error rates are usually acceptable. The storage capacity required to obtain $t = 300$ for this block length is $31 \times 300 = 9{,}300$ bits.

VII. CONCLUSIONS

Representation of the telephone network by a combination of renewal-type channels may be accomplished by the specification of

slightly more than a dozen parameters. This has the advantages of being compact and convenient, and of admitting to extrapolation to large block lengths and giving accuracies which are more than adequate for most error-control evaluations.

Interleaving on the telephone network can significantly enhance the error-control effectiveness of error-correcting codes when the separation between adjacent bits of a code word is on the order of several hundred bits. The storage requirements for achieving such interleaving are excessive for general application, but, where computers or other special storage facilities are available, interleaving can provide an interesting trade-off between decoding complexity and simple storage.

REFERENCES

1. Gilbert, E. N., Capacity of a Burst-Noise Channel, B.S.T.J., **39,** Sept., 1960, p. 1253.
2. Berger, J. M., and Mandelbrot, B., A New Model for Error Clustering in Telephone Circuits, IBM J. Res. and Dev., **7,** July, 1963, p. 224.
3. Sussman, S. M., Analysis of the Pareto Model for Error Statistics on Telephone Circuits, IEEE Trans. Comm. Systems, June, 1963, p. 213.
4. Alexander, A. A., Gryb, R. M., and Nast, D. W. Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39,** May, 1960, p. 431.
5. Elliott, E. O., Estimates of Error Rates for Codes on Burst-Noise Channels, B.S.T.J., **42,** Sept., 1963, p. 1977.
6. Townsend, R. L., and Watts, R. N., Effectiveness of Error Control in Data Communications Over the Switched Telephone Network, B.S.T.J., **43,** Nov., 1964, p. 2611.
7. Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. I, 2nd ed., John Wiley and Sons, New York, 1959.
8. Morris, R., Further Analysis of Errors Reported in "Capabilities of the Telephone Network for Data Transmission," B.S.T.J., **41,** July, 1962, p. 1399.
9. MacWilliams, Jessie, Permutation Decoding of Systematic Codes, B.S.T.J., **43,** Jan., 1964, p. 485.
10. Peterson, W. W., *Error-Correcting Codes*, John Wiley and Sons, New York, 1961, p. 166.

Differential Detection of Binary FM

By R. R. ANDERSON, W. R. BENNETT,
J. R. DAVEY and J. SALZ

(Manuscript received August 27, 1964)

Detection of binary FM by multiplication of the received signal by itself delayed is analyzed. Error rates vs signal-to-noise ratio for additive Gaussian noise are calculated as a function of sampling time, differential delay at the receiver, and delay distortion in the channel. It is found that the differential detector can give better performance than the more conventional zero-crossing counter or frequency discriminator under conditions of severe delay distortion in the channel.

I. INTRODUCTION

It has been found possible to realize excellent practical performance in FM transmission of binary data by use of a detector in which the signal is multiplied by a delayed replica of itself. This method has been called "differential detection" on account of its resemblance to the scheme of that name in widespread use as a detector of phase-modulated waves. The name "product demodulation" has also been applied. We can regard the detector either as a particular kind of frequency discriminator or as a phase comparator operating on the phase changes inherent in an FM wave. The former concept is suitable for a steady-state analysis, while the latter is more convenient in the study of signal transitions.

II. THE DIFFERENTIAL DELAY PRODUCTOR AS A FREQUENCY DISCRIMINATOR

Viewed as a discriminator, the detector has a steady-state response function calculable by multiplying a sine wave $A \cos (\omega_c + \omega)t$ by the corresponding delayed wave $A \cos [(\omega_c + \omega)(t - \tau)]$, i.e.,

$$A \cos (\omega_c + \omega)t \cdot A \cos [(\omega_c + \omega)(t - \tau)]$$
$$= (A^2/2) \cos (\omega_c + \omega)\tau \qquad (1)$$
$$+ (A^2/2) \cos [2(\omega_c + \omega)t - (\omega_c + \omega)\tau].$$

When the double-frequency component is suppressed by a low-pass filter, we obtain the response

$$V_{lf} = (A^2/2) \cos (\omega_c + \omega)\tau. \tag{2}$$

Consider ω as the frequency deviation in radians/sec from the midband frequency ω_c. Then a discriminator characteristic can be realized by setting $\cos \omega_c\tau = 0$, $\sin \omega_c\tau = \pm 1$, giving

$$V_{lf} = \mp(A^2/2) \sin \omega\tau. \tag{3}$$

The resulting steady-state response is nearly proportional to frequency deviation over the range in which $\sin \omega\tau$ is approximately equal to $\omega\tau$.

A linear relationship is not necessary for binary FM detection, since only the sign of the deviation is significant. Unambiguous results in the noise-free case can be secured over a range of $\omega\tau$ from $-\pi$ to π. In particular, if $\omega = \pm\omega_b/2$ where ω_b is 2π times the bit rate, values of τ in the range zero to the reciprocal of the bit rate could be used, and a value equal to half the bit interval appears to be a good compromise. As τ is made small the linearity improves, and, as will be shown later, the performance approaches that of an ideal phase differentiator, which we shall refer to as a $d\varphi/dt$ detector. The latter type is of particular interest because its performance is closely approximated by either a zero-crossing counter or a frequency discriminator.

III. FM DETECTION BY DIFFERENTIAL PHASE COMPARISON

We illustrate the operation as a differential phase comparator by following a particular noise-free sequence through the detection process. The binary signal to be transmitted is shown in Fig. 1(a). It is assumed that this rectangular wave modulates the frequency of a carrier with a total shift between mark and space equal to the bit rate. This results in the phase change during a marking bit interval differing from that during a spacing bit interval by 360°. With respect to the mid-frequency as a reference, the variation of carrier phase versus time becomes ±180° per bit interval, as indicated by the solid triangular wave of Fig. 1(b). When the channel is shaped to give a raised-cosine pulse spectrum at the demodulator input, the phase-versus-time pattern becomes rounded at the transitions approximately as shown by the dotted waveform of Fig. 1(b). The received signal is passed through a network with an envelope delay of one-half bit interval and with a phase shift at midband frequency of 270°. The phase-versus-time pattern of the delayed signal is shown by the dashed-line wave of Fig. 1(b). For a long mark interval

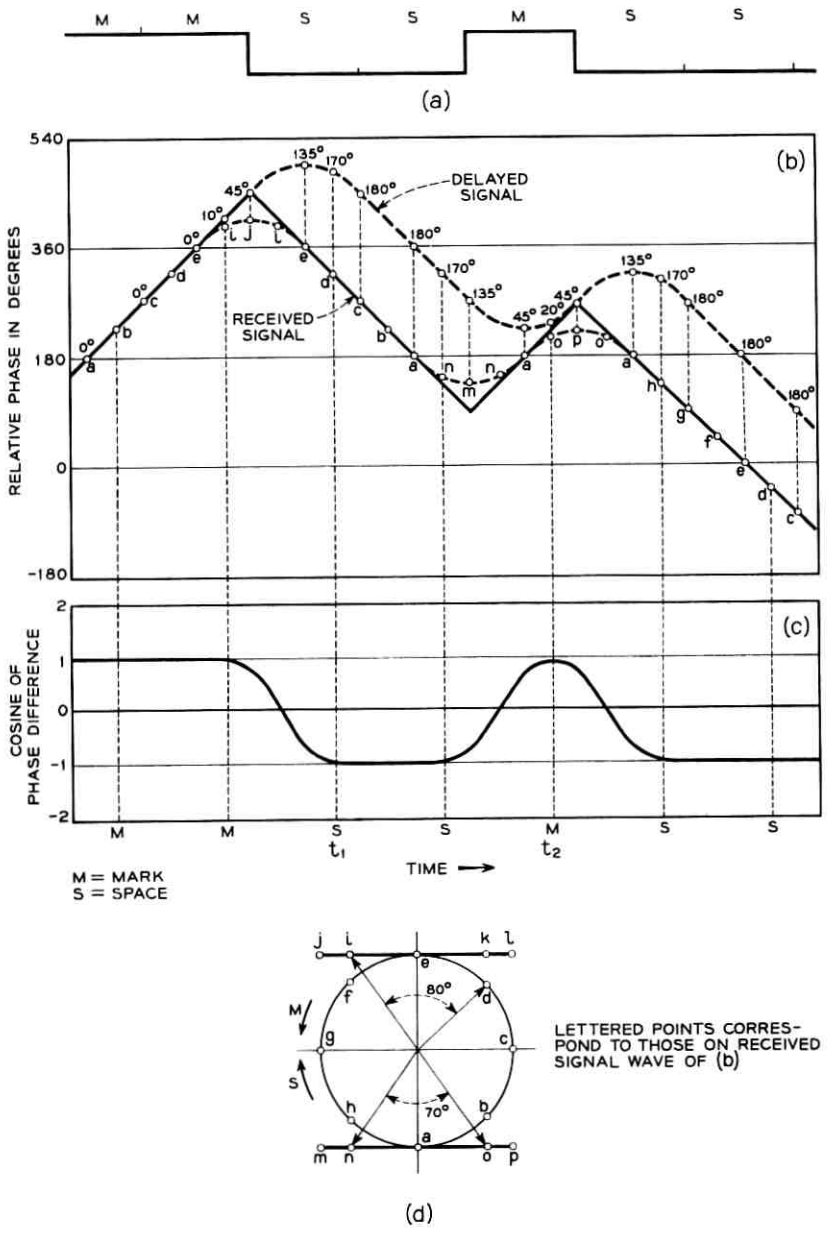


Fig. 1 — (a) Binary data, (b) phase of undelayed and delayed FM waves, (c) demodulated signal, (d) signal space diagram.

the two phase patterns are seen to be in phase, while for a long space interval they are 180° apart. The demodulation process consists of taking the product of these two waves with a switch-type modulator and filtering out the high-frequency component. If the amplitude modulation of the received signal is negligible, the demodulated signal becomes equal to the cosine of the phase difference between the delayed and undelayed versions of the signal.

The phase difference between the two signals at numerous points is indicated in Fig. 1(b). The shape of the demodulated wave, together with appropriate sampling times to recover the binary information, is shown in Fig. 1(c). It will be noted that the demodulated wave is delayed one-quarter bit interval from the instantaneous frequency of the received signal but is advanced one-quarter bit interval from the instantaneous frequency of the delayed signal.

A signal space diagram, such as described by J. R. Davey,[1] is given in Fig. 1(d) with lettered points corresponding to the lettered points on the phase pattern of the received signal. At a transition in the demodulated signal, the vectors representing the received signal and its delayed version are in phase at points such as (i), (k), (n), and (o). At a sampling instant having a transition on only one side, such as at t_1, the vectors representing the received signal and the half-bit delayed signal are at points such as (d) and (i). These form an angle of approximately 80° which, when the delayed signal is shifted 90°, becomes an angle of 10° or 170° depending on whether a mark or space is received. At a sampling instant having a transition on both sides, such as at t_2, the two vectors are at points such as (o) and (n). The angle is then about 70°, which, after the added 90° shift, results in an angle of 20° or 160°.

For a constant-amplitude signal the departures from the ideal 0° and 180° angles would decrease the amplitude of the demodulated signal at the sampling time. The diagram of Fig. 1(d) shows, however, that at least one of the vector amplitudes at these times is above steady state. If the increased signal amplitude appears at the linear input of the demodulating productor, it more than compensates for the phase error, thus tending toward an overshoot of the baseband signal. The low-pass filtering will of course determine the final extent, if any, of the overshoot. It is of interest to note that a linear productor would over-emphasize these amplitude variations. Consequently, it would be expected that a switched-type modulator would result in a more perfect eye pattern.*

* By "eye pattern" is meant the oscilloscope trace obtained by sweeping the detector output against a linear time base synchronized with the bit rate. A basic description of the properties of such patterns has been given by Brand and Carter.[2]

This does not mean, however, that the probability of decision error would be different. In fact, as will be shown later, limiting one of the two inputs to the multiplier, and hence obtaining in effect a switched-type modulator, does not change the error rate.

Fig. 2 shows a computer printout of the noise-free eye pattern corresponding to one-half bit delay when the demodulator follows a product law, the total frequency shift is equal to the bit rate, and full raised-cosine spectra apply. The origin is taken at the midpoint of the bit interval in the undelayed wave. Traces of like polarity are concurrent at this point, showing the absence of intersymbol interference at these particular instants. The peak responses of the detector are not reached until a time later by half the differential delay, and the individual peaks

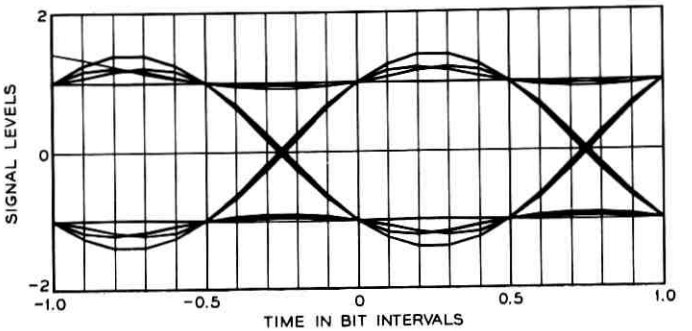


Fig. 2 — Eye pattern from product of undelayed and delayed FM waves with delay of one-half bit interval.

for different adjacent data sequences are spread over a range of values. The traces concur again at a time equal to the differential delay, and the same signal levels are observed as at the first point of concurrence. The decision-threshold (zero signal level) crossings are almost concurrent at a time preceding the origin by half the difference between the bit interval and the differential delay.

The time relations of Fig. 2 are peculiar to the choice of differential delay, ratio of frequency shift to bit rate, and spectral shaping. It will be shown later* for example, that if the differential delay is δ bit intervals, the first concurrence of traces is at the origin, the second one is at δ bit intervals later, the peaks occur at $\delta/2$, and the threshold crossings nearly at $-(1-\delta)/2$. If the raised-cosine shaping were changed to any other satisfying Nyquist's first criterion[5] for suppression of inter-

* See Figs. 9–12 and the discussion following equation (99).

symbol interference, all concurrences would be maintained except the threshold crossings, which could spread out more. Finally, if the relation between frequency shift and bit rate were changed, all concurrences would be destroyed, but this would not by itself imply a significant increase in error rate.

Eye patterns obtained when a limiter is inserted in one, but not both, of the two inputs are shown in Fig. 3. In Fig. 3(a), the undelayed signal is limited, giving in effect a switched modulator with the polarity of the undelayed signal switching the delayed signal. In Fig. 3(b), the delayed signal is limited, thereby interchanging the switching roles of Fig. 3(a). Comparing Fig. 3(a) with Fig. 2, we note that the concurrence of traces at 0.5 is destroyed and that the peaks of the responses are shifted to the left. In Fig. 3(b) the concurrence at 0 is destroyed and the response peaks are shifted to the right. In spite of these differences, which can be verified by fairly straightforward analysis, the probability of error at a specified sampling instant must be the same for all three cases, as will be shown in detail later.

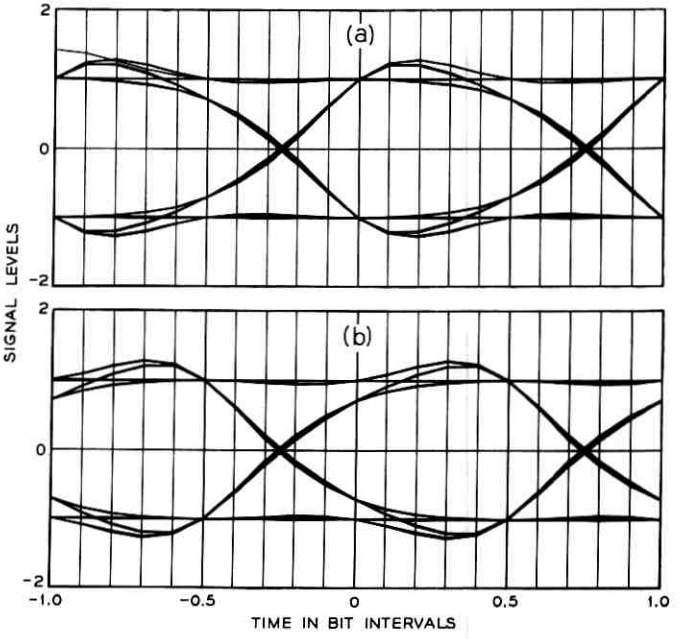


Fig. 3 — Eye patterns when one signal is limited before application to multiplier input. Delay of one-half bit interval. (a) Undelayed signal is limited; (b) delayed signal is limited.

The eye pattern for a $d\varphi/dt$ receiver, which is equivalent to the case of zero differential delay, is shown in Fig. 4. The traces concur at the peak response, which is at the origin, and the threshold crossings are half a bit interval away, as would be deduced by setting $\delta = 0$ in the discussion of the detector with differential delay. As δ approaches zero, we thus approach an equivalence with the performance of conventional frequency detectors.

The question of utility of eye patterns for nonlinear detection processes merits some further discussion. Although we cannot deduce error rates from them, we can at the least distinguish between "go" and "no go" conditions in the absence of noise. We can also use a given pattern as a basis for choosing the best sampling time. In making the choice,

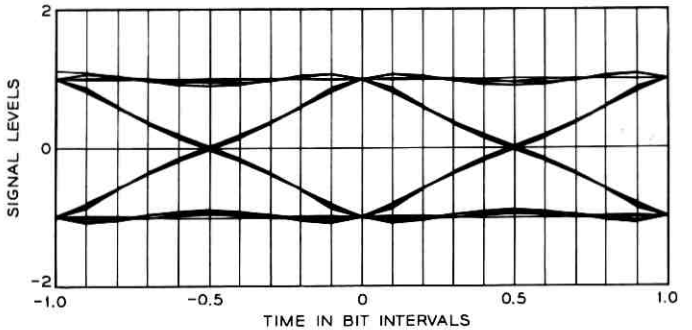


Fig. 4 — Eye pattern for $d\varphi/dt$ detector.

we consider both the horizontal and vertical margins, since relative immunity to timing jitter is important as well as the spread of sample values at the time of decision. We have verified in our calculations that the direction of change in the error rate with sampling time can be deduced by comparing eye openings for the two instants. Such deductions are not valid for comparing error rates corresponding to eye patterns for two different conditions.

Another use of the eye pattern is in laboratory diagnosis of system distortion. For such purposes we make use of established correspondence between the nature of the eye and various kinds of distortion in the particular system under test.

IV. THE EVALUATION PROBLEM

We are concerned here with the performance of the differential FM detector when the binary FM signal suffers transmission impairments.

We analyze particularly the effects of delay distortion and noise in the channel. A significant measure of performance is the curve of error rate vs signal-to-noise ratio at a specified bit rate for a channel with specified amplitude and delay variations with frequency. It is important that such a measure be determined over the range of transmission impairments encountered in actual channels. It is found that the principal virtue of the differential-delay scheme relative to the more familiar axis-crossing and frequency-discriminator types is an improved immunity to severe delay distortion.

Since there are many parameters which influence the performance of a data transmission system, the discussion would get out of hand if individual attention were given to all possible combinations of conditions. Fortunately, we can select representative regions of interest which are describable in terms of relatively few quantities. Our approach makes use of both direct analysis and digital computer back-up. In calculating the noise-free responses of the system under various conditions, we first establish formulas for a generalized data sequence. We then go to the computer to evaluate response functions vs time for a family of sequences of given length.

In theory it would be possible to calculate error rates by adding programmed noise in the computer calculation of response functions. A practical deterrent to such a procedure is the long computing time needed in the interesting cases of almost error-free transmission. It has been our experience that determination of error rates by computer simulation of additive noise is inferior to computer evaluation of analytic formulas for probability of error. Use of the latter procedure has been chiefly successful with additive Gaussian noise, but this does not necessarily imply serious limitation of utility. The premise that rank-order established on the basis of Gaussian noise holds for other kinds of interference has a good empirical foundation. A useful simplification from the Gaussian analysis is that the curves for error rate vs signal-to-noise ratio tend to be roughly parallel for different systems and to be characterized sufficiently well by their asymptotic slopes.

In terms of the normalized signal-to-noise ratio M, which is defined as the ratio of average signal power to the average white Gaussian noise power in a bandwidth equal to the bit rate, the asymptotic error rate is expressible[3] in the form $F(M) \exp(-\kappa M)$. The function $F(M)$ turns out to be of slight interest because of its relatively minor effect when M is large. For practical purposes the number κ determines the performance. For an ideal binary phase modulation system, κ has its maximum value of unity. The quantity $10 \log_{10}(1/\kappa)$ expresses noise impairment in db

relative to the ideal. It has been found possible to calculate this quantity directly without determining the entire error-rate curve.

The calculations we have made evaluate the effect of the following factors important in system design:

1. The sampling time relative to the signaling interval. The preferred sampling time is indicated by the eye patterns, but is more precisely established by error-rate calculations.

2. The length of the delay line. Equivalence of zero delay with a $d\varphi/dt$ detector establishes a reference at one end of the range in terms of a better-known system.

3. The data sequence. Results for the most and least vulnerable sequences are exhibited.

4. The delay distortion. Parabolic and linear variation with frequency are studied. The results are presented in terms of maximum delay variation expressed in bit intervals.

V. THE MODEL

A block diagram of the transmission system under study is shown in Fig. 5. The data source emits a sequence of binary symbols which for full information rate are independent of each other and have equal probability. The analysis can be generalized without analytical inconvenience by assigning a probability m_1 to one of the two binary symbols and $1 - m_1$ to the other. In conventional binary notation the symbols

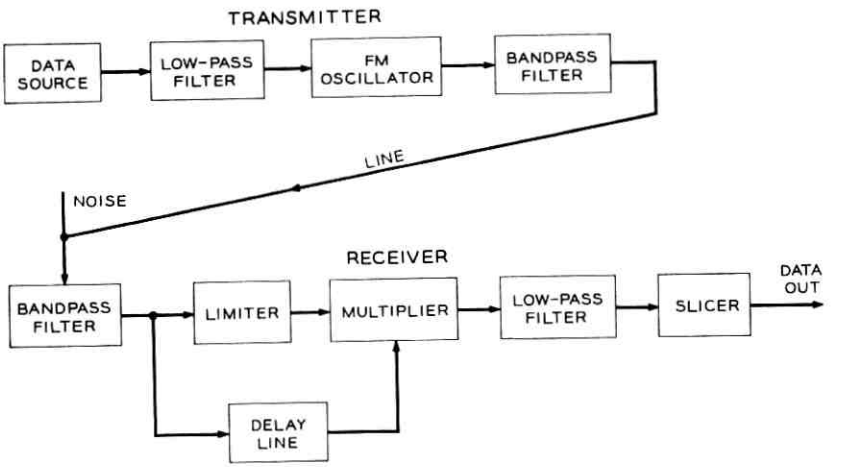


Fig. 5 — Binary FM transmission system with differential delay detection.

are 1 and 0. It is convenient to express binary frequency modulation of an oscillator in terms of positive and negative frequency deviations. The combination of data source and low-pass filter is accordingly defined by the shaped baseband data wave train

$$s(t) = \sum_{n=-\infty}^{\infty} b_n g(t - nT) \tag{4}$$

where

$$b_n = 2a_n - 1. \tag{5}$$

The values of a_n represent the data sequence in binary notation. The probability is m_1 that the typical a_n is unity, and $1 - m_1$ that it is zero. The value of b_n is $+1$ if a_n is unity, and -1 if a_n is zero. The function $g(t)$ represents a standard pulse emitted by the low-pass filter for a signal element centered at $t = 0$.

Ideally, the oscillator frequency follows the baseband signal wave $s(t)$. This would imply an output voltage from the FM oscillator specified by

$$V(t) = A \cos\left[\omega_c t + \theta_0 + \mu \int_{t_0}^{t} s(\lambda) d\lambda\right]. \tag{6}$$

Here, A is the carrier amplitude, ω_c is the frequency of the oscillator with no modulating signal applied, t_0 is an arbitrary reference time, θ_0 is the phase at $t = t_0$, and μ is a conversion factor relating frequency displacement to baseband signal voltage. The instantaneous angular frequency of the wave (6) is defined as the derivative of the argument of the cosine function. It can be written in the form $\omega_c + \omega_i$, where ω_i, the deviation from midband, is ideally expressed by

$$\omega_i = \mu s(t). \tag{7}$$

In the practical case, the transmitting bandpass filter restricts the frequency-modulated wave to the range of frequencies passed by the channel. The purpose of this filter is to prevent both waste of transmitted power in components which will not reach the receiver and contamination of the line at frequencies assigned to other channels. The result is a transformation of the voltage wave (6) to a band-limited form, which must depart in more or less degree from the ideal conditions of constant amplitude and of linear relationship between frequency and baseband signal. The line also inserts variations in amplitude- and phase-versus-frequency which cause further departures from the ideal. For our purposes it is sufficient to combine the line characteristics with those of the transmitting filter into a single composite

network function determining the wave presented to the receiving bandpass filter.

The receiving bandpass filter is necessary to exclude out-of-band noise and interference from the detector input. It also shapes the signal waveform and can include compensation for linear in-band distortion suffered in transmission. Two contradictory attributes are sought in the filter — a narrow band to reject noise and a wide band to supply a good signal wave to the detector. Previous work[3] has indicated a cosine filter as a near optimum.

The noise-free input to the detector will be written in the form

$$V_r(t) = P(t) \cos(\omega_c t + \theta) - Q(t) \sin(\omega_c t + \theta). \tag{8}$$

$P(t)$ and $Q(t)$ represent in-phase and quadrature signal modulation components respectively, which are associated with a carrier wave at the midband frequency ω_c with specified phase θ. Such a resolution can always be made, even though the details in actual examples may be burdensome. The added noise wave at the detector input is assumed to be Gaussian with zero mean and can likewise be written as

$$v(t) = x(t) \cos(\omega_c t + \theta) - y(t) \sin(\omega_c t + \theta). \tag{9}$$

If $v(t)$ represents Gaussian noise band-limited to $\pm 2\omega_c$, $x(t)$ and $y(t)$ are also Gaussian and are band-limited to $\pm\omega_c$. If the spectral density of $v(t)$ is $w_v(\omega)$, the spectral densities of $x(t)$ and $y(t)$ are given by[3]

$$w_x(\omega) = w_y(\omega) = w_v(\omega_c + \omega) + w_v(\omega_c - \omega), \qquad |\omega| < \omega_c. \tag{10}$$

In general, $x(t)$ and $y(t)$ are dependent, with cross-spectral density

$$w_{xy}(\omega) = j[w_v(\omega_c - \omega) - w_v(\omega_c + \omega)] \tag{11}$$

and cross-correlation function expressed in terms of $R_v(\tau)$, the autocorrelation function of $v(t)$, by

$$R_{xy}(\tau) = -2R_v(\tau) \sin \omega_c \tau. \tag{12}$$

The cross-correlation vanishes at $\tau = 0$, and hence the joint distribution of $x(t)$, $y(t)$ at any specified t is that of two independent Gaussian variables.

A convenient analytical model of the detector is a multiplier with delayed and undelayed waves applied as inputs and with a low-pass filter in the output to select the difference-frequency components of the product. In practical systems various departures from the basic model may offer a more convenient realization by physical circuits. The pure product law can be approximated by a switched modulator, an example of which is shown in Fig. 6(a). Here one of the two inputs operates a

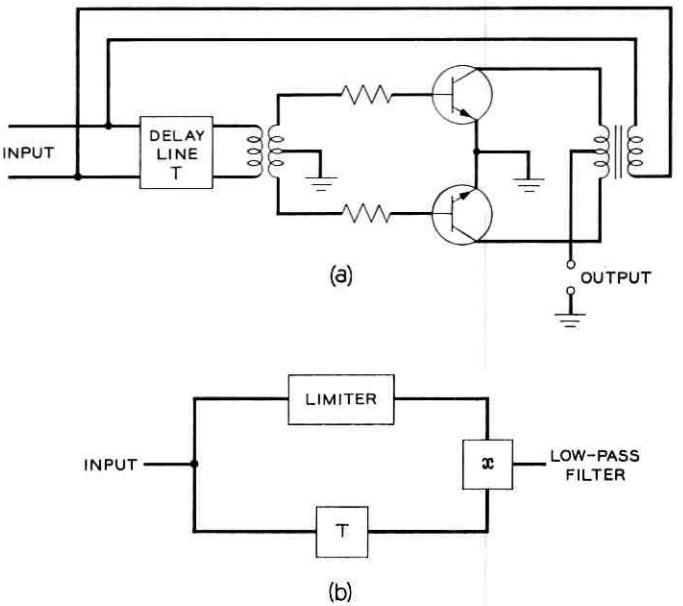


Fig. 6 — Differential detector with switched modulator.

two-transistor reversing switch in the path of the other input. The second input wave is in effect multiplied by a rectangular wave with axis crossings determined by the first input. The same result would be obtained from a strictly analog multiplier if the amplitude of one input were severely limited beforehand as shown in Fig. 6(b). We will base our analysis on a true productor with no limiter, but we show in Appendix A that limiting one input does not affect the results in the narrow-band case. In this detector the noise waves associated with the delayed and undelayed signal inputs are correlated. The amount of correlation depends on the value of the delay. The value of the delay relative to one bit interval will remain as a parameter to be optimized. However, we will require that the delay line should be designed to have a phase shift equal to an odd multiple of $\pi/2$ radians at the midband angular frequency ω_c.

VI. ANALYTICAL SOLUTION

By adding (8) and (9) we obtain the signal at the input of the detector

$$E(t) = x_1(t) \cos \omega_c t - y_1(t) \sin \omega_c t \tag{13}$$

where

$$\begin{aligned} x_1(t) &= x(t) + P(t) \\ y_1(t) &= y(t) + Q(t). \end{aligned} \tag{14}$$

The delayed signal is,

$$E(t - \tau) = x_1(t - \tau) \cos(\omega_c t - \omega_c \tau) - y_1(t - \tau) \sin(\omega_c t - \omega_c \tau) \tag{15}$$

where

$$0 < \tau < T.$$

In the case in which $\omega_c \tau = 270°$, we write for $E(t - \tau)$

$$E_d(t) = -x_{1d} \sin \omega_c t - y_{1d} \cos \omega_c t \tag{16}$$

where we set $f(t - \tau) \triangleq f_d$.

The low-frequency component $E_{lf}(t)$ of the product $E(t)E_d(t)$ is given by

$$2E_{lf} = y_1 x_{1d} - x_1 y_{1d} = \xi. \tag{17}$$

The performance of our system can be studied by analyzing the probability distribution of the quadratic form in (17). Since this is a binary system we only need the distribution at one point, namely "zero."

By a relabeling of the variables in which y_1, x_1, y_{1d}, and x_{1d} are replaced by x_1, x_2, x_3, and x_4, respectively, the calculation is reduced to the single problem of evaluating the probability that the quadratic form $x_1x_4 + x_2x_3$ is negative or positive, where x_1, x_2, x_3 and x_4 are Gaussian random variables with equal variances σ^2 and mean values given by:

$$\begin{aligned} E\{x_1\} &= Q(t), & E\{x_2\} &= -P(t) \\ E\{x_3\} &= Q(t - \tau), & E\{x_4\} &= P(t - \tau). \end{aligned} \tag{18}$$

We remark that all these average values, in general, will depend on the signal sequence.

A solution of this general problem in terms of an integral has been found[3] when the variables are independent. The present case is more complicated in that with an arbitrary delay τ, the noise samples become dependent. We must now include a nonzero covariance of x_1 and x_{1d} and of y_1 and y_{1d}. We point out that a solution based on uncorrelated noise samples would indicate a considerably poorer performance than found when the actual correlation is included. This is a case in which noise correlation is beneficial rather than harmful.

It has been found possible to apply the previous solution for the independent case by subjecting the four dependent Gaussian variables to a linear transformation. The new set of variables z_1, z_2, z_3, and z_4 becomes independent while preserving the invariance of the quadratic form on which decisions are based, i.e.,

$$z_1z_4 + z_2z_3 = x_1x_4 + x_2x_3 . \tag{19}$$

The nonzero covariances of the variables are:

$$\begin{aligned} \text{cov } (x_1 , x_3) &= \text{cov } (y,y_d) = r\sigma^2 \\ \text{cov } (x_2 , x_4) &= -\text{cov}(x,x_d) = -r\sigma^2. \end{aligned} \tag{20}$$

The value of r is the normalized autocorrelation function of the noise evaluated at lag time τ. The autocorrelation function is the Fourier transform of the spectral density. In the optimum receiver design the filter preceding the detector has a cosine amplitude-frequency response reaching zero at a frequency displacement from midband equal to the bit rate. Hence with white Gaussian noise on the line, the spectral density of the noise at the detector input is a squared- or raised-cosine function. Such a spectral function has the property that its Fourier transform, the autocorrelation function, decreases to zero when the lag time increases from zero to T. The solution we shall give is valid for any value of r within the permissible range from -1 to $+1$.

The transformation that satisfies (19) and at the same time diagonalizes the covariance matrix of the z variables is:

$$\begin{aligned} z_1 &= \tfrac{1}{2}(x_1 - x_2 + x_3 + x_4) \\ z_2 &= \tfrac{1}{2}(x_1 + x_2 + x_3 - x_4) \\ z_3 &= \tfrac{1}{2}(-x_1 + x_2 + x_3 + x_4) \\ z_4 &= \tfrac{1}{2}(x_1 + x_2 - x_3 + x_4). \end{aligned} \tag{21}$$

It can be verified that (19) is satisfied when this transformation is applied and also that if the correlation matrix of the x's is:

$$R_x = \sigma^2 \begin{bmatrix} 1 & 0 & r & 0 \\ 0 & 1 & 0 & -r \\ r & 0 & 1 & 0 \\ 0 & -r & 0 & 1 \end{bmatrix} \tag{22}$$

the correlation matrix of the z's becomes:

$$R_z = \sigma^2 \begin{bmatrix} 1+r & 0 & 0 & 0 \\ 0 & 1+r & 0 & 0 \\ 0 & 0 & 1-r & 0 \\ 0 & 0 & 0 & 1-r \end{bmatrix}. \tag{23}$$

We now recall the previously obtained result[3] for the four uncorrelated Gaussian variables z_1, z_2, z_3, and z_4, with the pair z_1 and z_2 having equal variances σ_0^2 and the pair z_3 and z_4 having equal variances σ_1^2:

$$\text{Prob}\,[(z_1z_4 + z_2z_3) < 0 \quad \text{when} \quad \bar{z}_1\bar{z}_4 + \bar{z}_2\bar{z}_3 > 0] = \Lambda(\rho,a,b)$$
$$= \frac{1}{2} - \frac{\rho}{2\sqrt{\pi}} \int_{-1}^{1} \exp(-\rho^2x^2)\ \text{erf}\,[a\rho(1-x^2)^{\frac{1}{2}} - bx]dx \tag{24}$$

where

$$\rho^2 = \frac{\bar{z}_1^2 + \bar{z}_2^2}{2\sigma_0^2} \tag{25}$$

$$a = \frac{\bar{z}_1\bar{z}_4 + \bar{z}_2\bar{z}_3}{2\sigma_1\sigma_2\rho^2} \tag{26}$$

$$b = \frac{\bar{z}_2\bar{z}_4 - \bar{z}_1\bar{z}_3}{2\sigma_1\sigma_2\rho}. \tag{27}$$

In our case

$$\sigma_0^2 = (1+r)\sigma^2, \qquad \sigma_1^2 = (1-r)\sigma^2. \tag{28}$$

It is shown in Appendix B that the integral in (24) can be simplified and reduced to a two-parameter form. We also show how the general asymptotic expression for large signal-to-noise ratio is derived.

The simplified version of (24) is

$$\Lambda(\rho,a,b) = \frac{1}{2\pi} \int_0^{\pi} \exp\left[\frac{-c^2}{1 + d^2\cos\varphi}\right] d\varphi \tag{29}$$

where

$$c^2 = \frac{2a^2\rho^2}{k^2 + a^2 + 1} \tag{30}$$

$$d^2 = \frac{[(k^2 + a^2 - 1)^2 + 4k^2]^{\frac{1}{2}}}{k^2 + a^2 + 1} \tag{31}$$

and $k = b/\rho$.

The signal-to-noise ratio appears only in the parameter ρ. The asymptotic expression for large signal-to-noise ratio is therefore obtained as the limiting form of (29) when c^2 becomes large. The result, as shown in Appendix B, is:

$$\Lambda(\rho,a,b) \equiv G(c,d) \sim \frac{1+d^2}{2cd\sqrt{2\pi}} \exp\left[\frac{-c^2}{1+d^2}\right]. \tag{32}$$

It is of particular interest in our problem to write the asymptotic expression in the form:

$$G(c,d) = \frac{\alpha}{\rho} e^{-\beta\rho^2} \tag{33}$$

The values of α and β are:

$$\begin{aligned} \alpha &= \frac{k^2+a^2+1+[(k^2+a^2-1)^2+4k^2]^{\frac{1}{2}}}{4\pi^{\frac{1}{2}}a[(k^2+a^2-1)^2+4k^2]^{\frac{1}{4}}} \\ \beta &= \frac{2a^2}{k^2+a^2+1+[(k^2+a^2-1)^2+4k^2]^{\frac{1}{2}}}. \end{aligned} \tag{34}$$

These are the general results we need.

VII. A REPRESENTATIVE EXAMPLE

A convenient analytical representation of a band-limited FM signal was first proposed by Sunde.[4] This representation approximates very closely the actual signals generated by practical data sets.[3] In the case of ideal transmission, we write for the signal at the input to the detector

$$E(t) = A \sin \omega_d t \sin \omega_c t - A s_1(t) \cos \omega_c t \tag{35}$$

where

$$s_1(t) = \sum_{n=-\infty}^{\infty} (-1)^n b_n g(t-nT) \tag{36}$$

$$T\omega_d = \pi, \qquad T \triangleq \text{bit interval.}$$

The standard pulse response $g(t)$ must satisfy Nyquist's first criterion, i.e.,

$$g(mT) = \delta_{m0}\,. \tag{37}$$

The value of b_n is $+1$ for mark and -1 for space.

As can be seen, the signal (35) can be synthesized by exciting a network having impulse response $g(t)$ by a series of $+$ or $-$ impulses occurring at integral multiples of the bit interval T. The sinusoidal com-

ponents of the signal are added to the response to the impulses. When delay distortion is present, the impulse response $g(t)$ is suitably modified by introducing a quadrature component.

The delayed version of (35) becomes:

$$E_d(t) = A \sin(\omega_d t - \omega_d T\delta) \sin(\omega_c t - \omega_c T\delta) - A s_1(t - T\delta) \cos(\omega_c t - \omega_c T\delta) \tag{38}$$

where $\delta = \tau/T$. When the noise components are added and the condition $\omega_c T\delta = 3\pi/2$ is satisfied, the low-frequency product is

$$\xi(t) = -[A \sin(\omega_d t - \pi\delta) + y(t - T\delta)][A s_1(t) + x(t)] + [A \sin \omega_d t + y(t)][A s_1(t - T\delta) + x(t - T\delta)]. \tag{39}$$

The function $\xi(t)$ represents the signal as it appears at the output of the ideal low-pass filter. To obtain the information, $\xi(t)$ must be sampled at integral multiples of the bit interval. The sample thus obtained at $t = T + \epsilon T$, $0 \leqq \epsilon \leqq 1$ is

$$\begin{aligned}\xi(T + T\epsilon) = &-\{A \sin[\pi(1 + \epsilon - \delta)] + y[T(1 + \epsilon - \delta)]\} \\ &\cdot\{A s_1[T(1 + \epsilon)] + x[T(1 + \epsilon)]\} \\ &+ \{A \sin[\pi(1 + \epsilon)] + y[T(1 + \epsilon)]\} \\ &\cdot\{A s_1[T(1 + \epsilon - \delta)] + x[T(1 + \epsilon - \delta)]\}.\end{aligned} \tag{40}$$

As is characteristic of nonlinear detection processes, the presence of noise introduces dependence on signal history. The memory can be minimized by using a pulse spectrum which satisfies the second as well as the first of Nyquist's criteria,[5] i.e., one which preserves the spacing of transition times as well as axis crossings. To illustrate let $\epsilon = 0$ and $\delta = \frac{1}{2}$. In this case (40) reduces to:

$$\xi(T) = -\{A + y(T/2)\}\{s_1(T) + x(T)\} + \{y(T)\}\{A s_1(T/2) + x(T/2)\} \tag{41}$$

where

$$s_1(t) = \sum_{n=-\infty}^{\infty} (-1)^n b_n g(t - nT) \tag{42}$$

$$s_1(T/2) = -\tfrac{1}{2}(b_1 - b_0), \qquad s_1(T) = -b_1 . \tag{43}$$

The memory is thus reduced to one previous symbol. It follows, therefore, that

$$\xi(T) = [x - A b_1][-y_d - A] + y[x_d - (A/2)(b_1 - b_0)]. \tag{44}$$

The probability of error is the weighted average of the conditional probabilities that the sample is negative when b_1 is $+1$ and that the sample is positive when $b_1 = -1$. Since the sample depends on the present and immediately preceding symbols, there are four different cases to consider: $b_1 = -b_0 = 1$; $b_1 = b_0 = 1$; $b_1 = -b_0 = -1$; and $b_1 = b_0 = -1$. If marking and spacing symbols are equally probable, each case has a probability of one-fourth. We now write for the average probability of error

$$P_e = \tfrac{1}{4} \sum_{n=1}^{4} \text{Prob } [(x_{1n}x_{4n} + x_{2n}x_{3n}) < 0] \tag{45}$$

where the variables x_{1n}, x_{2n}, x_{3n}, and x_{4n} are specified by Table I.

A physical interpretation of the detection process in this case can be obtained by regarding the quadratic form in (44) as the scalar product of two vectors, e.g.:

$$x_1x_4 + x_2x_3 = \mathbf{u} \cdot \mathbf{v} \tag{46}$$

where if $\mathbf{i},\mathbf{j}$ represent unit vectors along rectangular coordinate axes, the four possible pairs of vectors are:

$$\mathbf{u}_1 = \mathbf{i}(-A + x) + \mathbf{j}y$$
$$\mathbf{v}_1 = \mathbf{i}(-A - y_d) + \mathbf{j}(-A + x_d) \tag{47}$$

$$\mathbf{u}_2 = \mathbf{i}(-A + x) + \mathbf{j}y$$
$$\mathbf{v}_2 = \mathbf{i}(-A - y_d) + \mathbf{j}x_d \tag{48}$$

$$\mathbf{u}_3 = \mathbf{i}(A + x) - \mathbf{j}y$$
$$\mathbf{v}_3 = \mathbf{i}(A + y_d) + \mathbf{j}(A + x_d) \tag{49}$$

$$\mathbf{u}_4 = \mathbf{i}(A + x) - \mathbf{j}y$$
$$\mathbf{v}_4 = \mathbf{i}(A + y_d) + \mathbf{j}x_d\,. \tag{50}$$

Occurrence of error is synonymous with a negative value for the scalar product of any pair of vectors, and hence is also equivalent to an

TABLE I

n	x_{1n}	x_{2n}	x_{3n}	x_{4n}
1	$x - A$	y	$x_d - A$	$-y_d - A$
2	$x - A$	y	x_d	$-y_d - A$
3	$x + A$	$-y$	$x_d + A$	$y_d + A$
4	$x + A$	$-y$	x_d	$y_d + A$

angle greater than 90° between the vectors of a pair. We can use this representation to explain the beneficial effects of correlated noise samples. There is positive correlation between x and x_d as well as between y and y_d. This means that x_d is likely to have the same sign as x and similarly for y_d and y. Inspection of the vector components shows that such noise values have opposite effects on the size of the angle between the two vectors, as shown in Fig. 7, illustrating (49).

At this point we can obtain an explicit expression for the probability of error. From Table I, we note that the average values of the x's are given by Table II. From (21), the $\bar{z}$'s are given by Table III. It is seen that the cases of $n = 3$ and 4 differ from $n = 1$ and 2 respectively by a change of sign throughout. A change of sign in all the $\bar{z}$'s in (25)-(27)

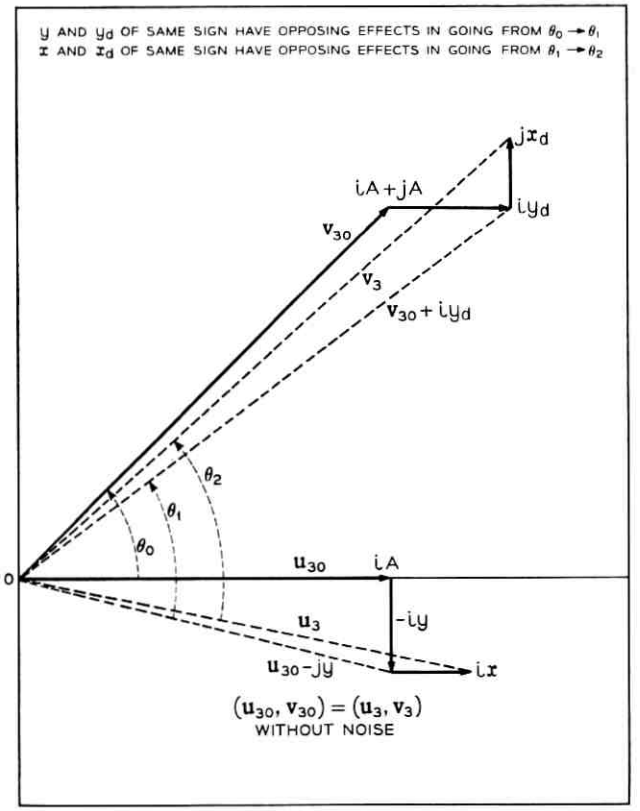


Fig. 7 — Effect of correlated noise samples on angle between vector inputs to detector.

TABLE II

n	$\bar{x}_{1n}$	$\bar{x}_{2n}$	$\bar{x}_{3n}$	$\bar{x}_{4n}$
1	$-A$	0	$-A$	$-A$
2	$-A$	0	0	$-A$
3	A	0	A	A
4	A	0	0	A

leaves the parameters a and b unchanged. Therefore, the four probabilities of (45) reduce to two distinct ones with the multiplying factor changed from $\frac{1}{4}$ to $\frac{1}{2}$.

The values of the parameters for the two cases are found to be

$$\rho_1^2 = \frac{5A^2}{4(1+r)\sigma^2} \qquad \rho_2^2 = \frac{A^2}{2(1+r)\sigma^2}$$

$$a_1 = \frac{2}{5}\sqrt{\frac{1+r}{1-r}} \qquad a_2 = \sqrt{\frac{1+r}{1-r}} \tag{51}$$

$$b_1 = -\frac{A}{2\sigma\sqrt{5(1-r)}} \qquad b_2 = 0.$$

As demonstrated in Ref. 3, the mean square of the signal wave on the line with random data is A^2 in the optimum case for white Gaussian noise. The noise input to the detector has a squared-cosine spectral density function, with σ^2 equal to the area. Hence σ^2 is one-half the mean-square value of the noise on the line in a band of width equal to twice the bit rate. In terms of the parameter M, defined as the ratio of average signal power on the line to average noise power in the bit-rate bandwidth, we have

$$M = A^2/\sigma^2. \tag{52}$$

Substituting the appropriate correlation value of $r = \frac{1}{2}$, we obtain

TABLE III

n	$\bar{z}_{1n}$	$\bar{z}_{2n}$	$\bar{z}_{3n}$	$\bar{z}_{4n}$
1	$-3A/2$	$-A/2$	$-A/2$	$-A/2$
2	$-A$	0	0	$-A$
3	$3A/2$	$A/2$	$A/2$	$A/2$
4	A	0	0	A

for the two sets of parameters

$$\begin{aligned} \rho_1^2 &= \frac{5}{6}M & \rho_2^2 &= \frac{M}{3} \\ a_1 &= \frac{2}{5}\sqrt{3} & a_2 &= \sqrt{3} \\ b_1 &= -\sqrt{\frac{M}{10}} & b_2 &= 0. \end{aligned} \tag{53}$$

The probability of error is given by

$$P_e = \tfrac{1}{2}\,\Lambda(\rho_1, a_1, b_1) + \tfrac{1}{2}\,\Lambda(\rho_2, a_2, b_2). \tag{54}$$

Using the computer program previously established for $\Lambda(\rho,a,b)$, S. Habib has obtained the uppermost curve of Fig. 8. Also shown are the ideal performance for coherent binary detection, the ideal performance for noncoherent binary detection, and an experimental curve obtained by E. R. Day.

It is also instructive to apply the asymptotic formulas given in (33). In this example, applying (53) to (34), we obtain

$$\begin{aligned} \alpha_1 &= \frac{3}{4}\sqrt{\frac{5}{3\pi}}, & \beta_1 &= \frac{2}{5} \\ \alpha_2 &= \frac{1}{2}\sqrt{\frac{3}{2\pi}}, & \beta_2 &= 1. \end{aligned} \tag{55}$$

Substituting these values in (30) we find

$$P_e \sim \frac{3}{4\sqrt{2\pi M}}\, e^{-M/3} + \frac{3}{8\sqrt{2\pi M}}\, e^{-M/3} \sim \frac{9}{8\sqrt{2\pi M}}\, e^{-M/3}. \tag{56}$$

In terms of the notation introduced in Section IV, $\kappa = \frac{1}{3}$. The asymptote is plotted in Fig. 8 and is found to be very close to the result of the exact calculation in the region of interest. Since the optimum coherent system has an error probability proportional to e^{-M}, the half-bit differential system in the limit requires 10 $\log_{10}$ 3 or 4.8 db more signal-to-noise ratio than the optimum for the same performance. Of this penalty, 3 db is accounted for by the steady-state informationless tones which make up half the average power of the FM signal. The remaining 1.8 db can be ascribed to the differential-detection scheme.

It is also possible to decode the message by providing a full-bit delay at the receiver. However, the decoding has to be performed on the transitions. In addition, the phase shift must be such that $\cos \omega_c T = 1$.

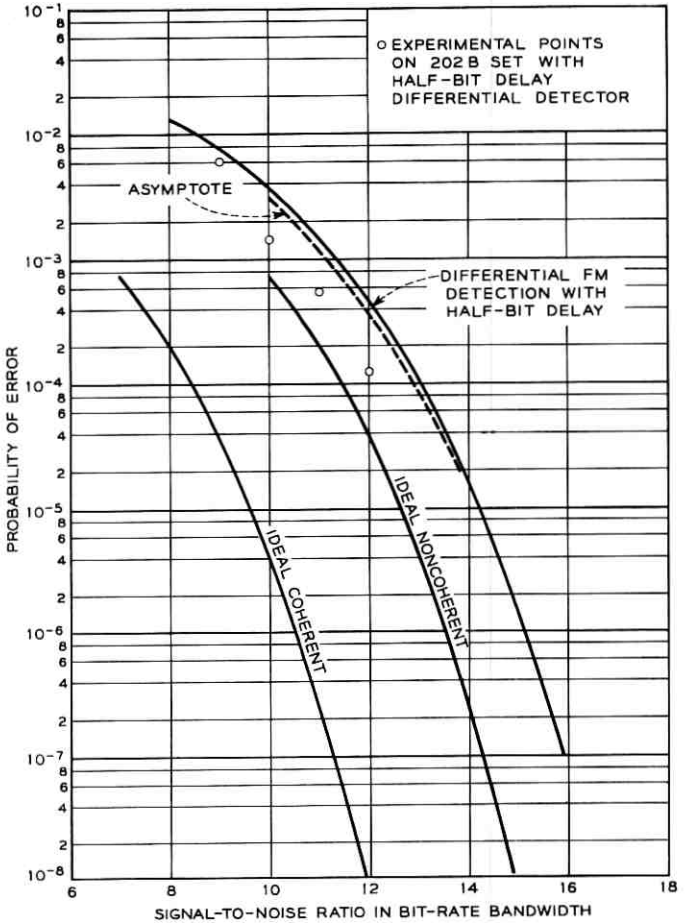


Fig. 8 — Error rates in binary data transmission.

As before, if we perform the indicated multiplication, reduce to a sum of single-frequency terms, and reject those that contain the midband frequency ω_c, we obtain the post-detection filter output, which we now designate as $V_{lf}(t)$. Again, to simplify the formulas we omit showing the functional dependence on t and use the subscript d to indicate values of functions at $t - T$. Then

$$2V_{lf} = x_1x_{1d} - x_1As_{1d} - x_{1d}As_1 + A^2s_1s_{1d} + A(y_1 - y_{1d}) \sin \omega_d t + y_1y_{1d} - A^2 \sin^2 \omega_d t. \tag{57}$$

Finally, we assume samples to be taken at multiples of T, which means that sin $\omega_d l$ becomes sin $\omega_d mT = 0$ and

$$s_1 = s_1(mT) = (-)^m b_m \tag{58}$$

$$s_{1d} - s_1[(m-1)T] = (-)^{m-1} b_{m-1} . \tag{59}$$

For the mth sample, then

$$2V_{lf} = x_1 x_{1d} + y_1 y_{1d} - (-)^m A(b_m x_{1d} - b_{m-1} x_1) - A^2 b_m b_{m-1} . \tag{60}$$

Let the binary information be coded in the transitions. Then, for example, a "1" is represented by $b_m = -b_{m-1}$ and a "0" by $b_m = b_{m-1}$. The two signaling conditions are then:

$$\begin{aligned} 2V_{lf1} &= A^2 + (-)^m b_m A(x_1 + x_{1d}) + x_1 x_{1d} + y_1 y_{1d} \\ &= [A + (-)^m b_m x_1][A + (-)^m b_m x_{1d}] + y_1 y_{1d} \end{aligned} \tag{61}$$

$$\begin{aligned} 2V_{lf0} &= -A^2 + (-)^m b_m A(x_1 - x_{1d}) + x_1 x_{1d} + y_1 y_{1d} \\ &= -[A - (-)^m b_m x_1][A + (-)^m b_m x_{1d}] + y_1 y_{1d} . \end{aligned} \tag{62}$$

The sampling interval T is typically large enough in this case to make the delayed noise samples independent of the direct samples. With this assumption we can regard x_1, y_1, x_{1d}, and y_{1d} as independent Gaussian variables. In the first condition, we set

$$\xi = A + (-)^m b_m x_1 \tag{63}$$

$$\xi_d = A + (-)^m b_m x_{1d} \tag{64}$$

$$2V_{lf} = \xi\xi_d + y_1 y_{1d} . \tag{65}$$

Then ξ, ξ_d, y_1, and y_{1d} are independent Gaussian variables with standard deviation σ, where σ is the rms noise voltage at the detector input, i.e., the rms value of either x_1 or y_1. The mean values of y_1 and y_{1d} are zero, and the mean values of ξ and ξ_d are A. In the second condition, we set

$$\xi = -A + (-)^m b_m x_1 \tag{66}$$

and obtain the same relations except that the mean value of ξ becomes $-A$ instead of A.

Correct decisions are made in the first condition if $2V_{lf}$ is positive, and in the second condition if $2V_{lf}$ is negative. Hence if we let

$$z = \xi\xi_d + y_1 y_{1d} \tag{67}$$

and designate $p_1(z)$ and $p_2(z)$ as the probability density functions of z

when ξ has the means A and $-A$ respectively, the probabilities of error in the two conditions become

$$P_1 = \int_{-\infty}^{0} p_1(z)dz \tag{68}$$

$$P_2 = \int_{0}^{\infty} p_2(z)dz. \tag{69}$$

This is the same problem solved in equations (54)–(58) of a previous paper[3], and the final result is found to be

$$P = P_1 = P_2 = \tfrac{1}{2}\, e^{-A^2/(2\sigma^2)} = \tfrac{1}{2}\, e^{-M/2}. \tag{70}$$

Equation (70) shows 3 db poorer performance than ordinary differentially coherent phase modulation. The one-bit delay differential system with transition coding thus suffers 3.9 db penalty relative to ideal coherent detection in the error probability range of 10^{-4}.

An interesting result is obtained when the delay line at the receiver is allowed to become small relative to the bit interval while still maintaining the condition that the phase shift at the carrier frequency ω_c equals 270°. We show in Appendix C that the performance in this case approaches the performance of an ideal phase differentiator.

VIII. RESULTS WITH NO DELAY DISTORTION

The first part of our numerical results deals with the performance of the differential FM detector as a function of the value of differential delay and the sampling instant in the absence of delay distortion on the line. As a preliminary, we show in Figs. 9–12 inclusive the computer print-outs of the eye patterns for differential-delay values of 0.2, 0.4,

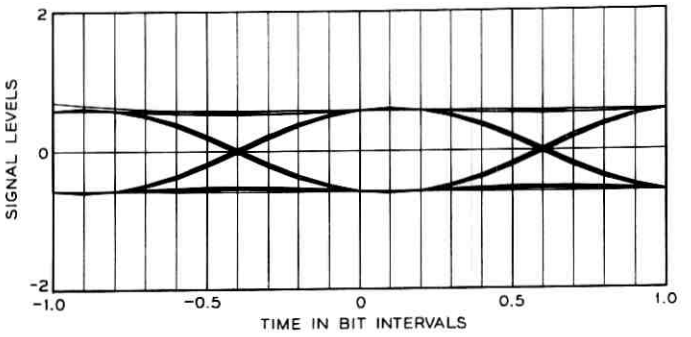


Fig. 9 — Eye pattern for differential delay of 0.2-bit interval.

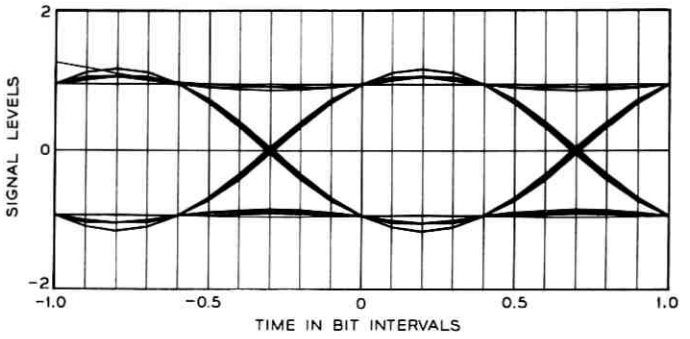


Fig. 10 — Eye pattern for differential delay of 0.4-bit interval.

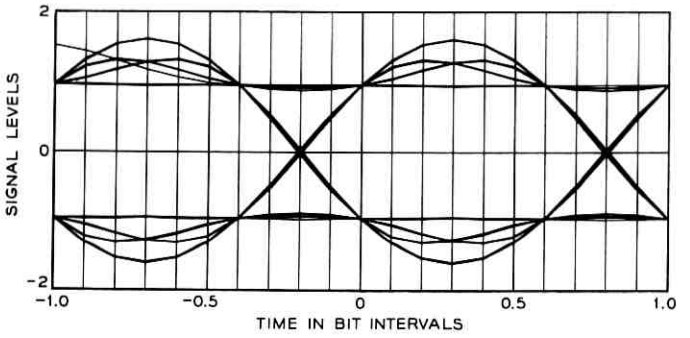


Fig. 11 — Eye pattern for differential delay of 0.6-bit interval.

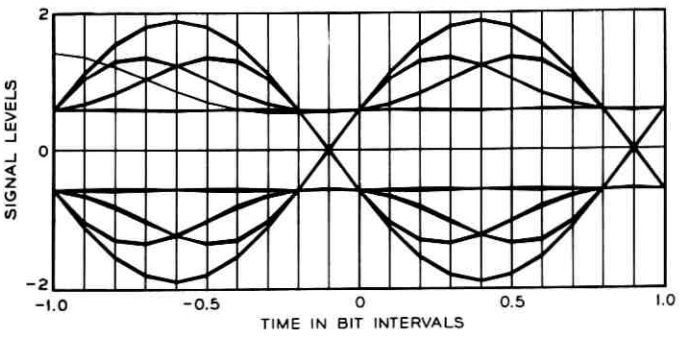


Fig. 12 — Eye pattern for differential delay of 0.8-bit interval.

0.6, and 0.8 bit interval. Some of the features found have been previously mentioned in Sec. III. As before stated, the patterns are not to be interpreted as furnishing quantitative evaluation of performance. For the latter, we rely entirely on error rates vs signal-to-noise ratio. The figures illustrate what would be seen at the detector output in the absence of noise for the various cases and indicate the preferred sampling time for each case.

For any pair of values of δ and ϵ in (40), ξ is a random variable possessing a probability density determined by the additive noise and previous signal history. We have conserved computation time without loss of essential information by concentrating attention on the asymptotic performance when the signal-to-noise ratio is large. The corresponding db impairment or degradation relative to optimum binary PM as expressed by $10 \log_{10} (1/\kappa)$ has been computed for various values of δ and ϵ and for all data sequences of length 5 bits. For a few representative cases, we have used the exact formula over a range of signal-to-noise ratios to indicate the degree of approximation given by the asymptotic formula for typical conditions of interest. A raised-cosine pulse spectrum on the line has been assumed. Details of the computations are given in Appendix D.

The IBM 7090 computer was programmed to evaluate the parameters developed in Appendix D for all 32 possible 5-bit sequences and for different values of δ and ϵ. Figs. 13–18 represent the degradation vs value of delay for different sampling instants across the bit. The sampling time is measured from the midpoint of the bit interval in the undelayed wave. The curves on each graph represent the various sets of sequences. It appears from the graphs that the 32 sequences tend to bunch into four distinct sets. We did not attempt to label or identify these sets, since the average performance is more closely determined by the sequences which suffer the most. This is because the relative probability of error for the sets varies exponentially with the degradation. It can be seen that as the value of delay exceeds half a bit the performance degrades rapidly.

We next examined the performance of the receiver for fixed delay line values and variable sampling instants. As shown in Appendix C, the $d\varphi/dt$ detector may be regarded as a limiting form of differential delay as the value of delay approaches zero. Fig. 19 shows the performance for the worst and best sequences as the sampling instant is varied across the bit. It is clear from this figure that the best sampling instant is in the middle of the bit, i.e., midway between transitions of the detector output wave. It turns out that the best sequence is the sequence of all marks or spaces, while the worst sequence is that of reversals.

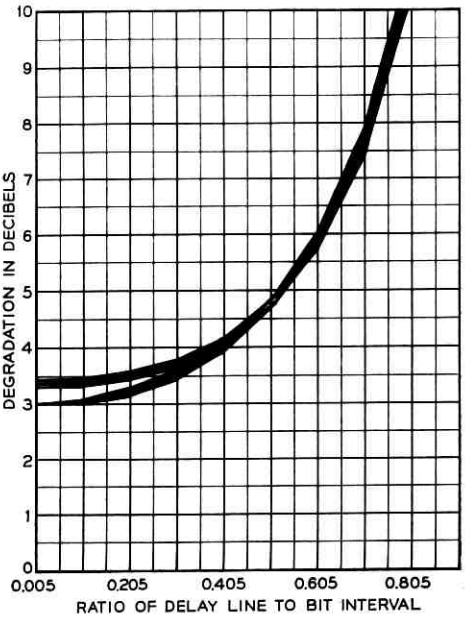


Fig. 13 — Db degradation vs differential delay with sampling at 0.005-bit interval.

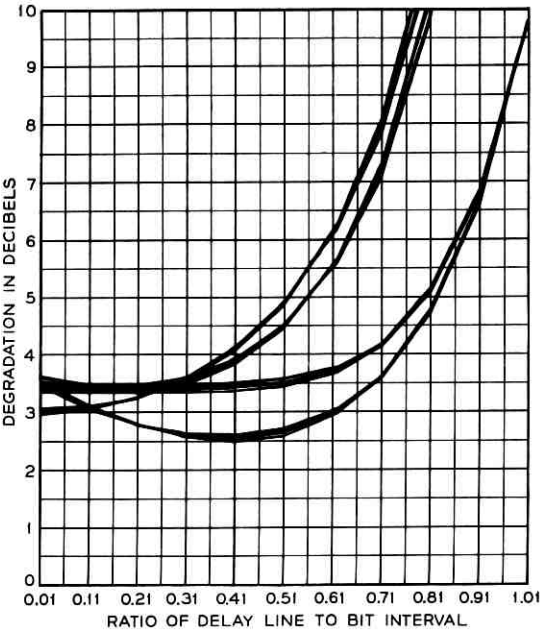


Fig. 14 — Db degradation vs differential delay with sampling at 0.11-bit interval.

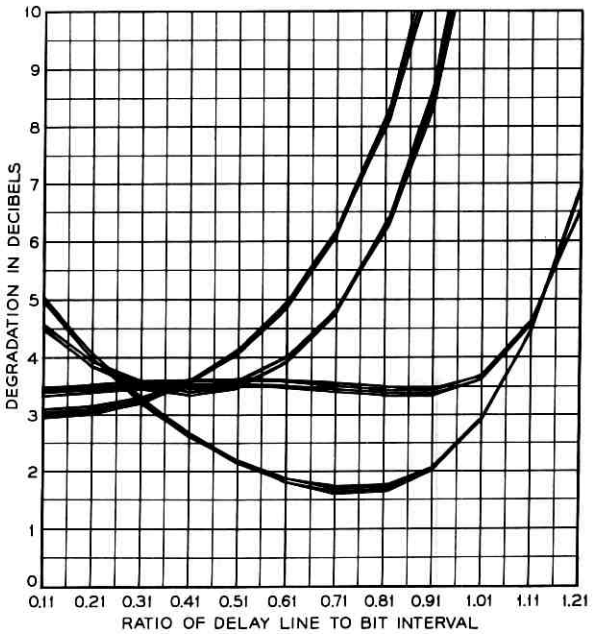


Fig. 15 — Db degradation vs differential delay with sampling at 0.21-bit interval.

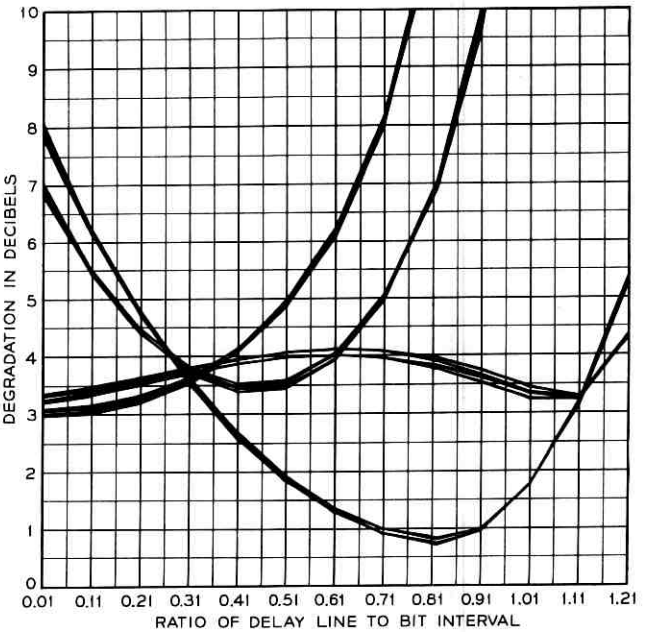


Fig. 16 — Db degradation vs differential delay with sampling at 0.31-bit interval.

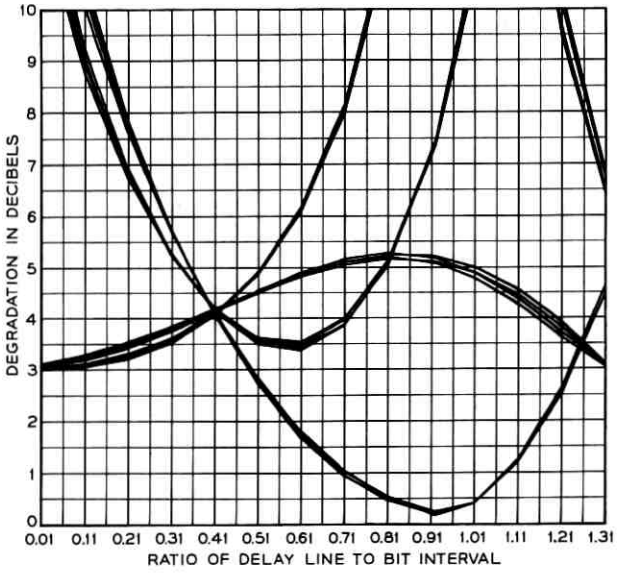


Fig. 17 — Db degradation vs differential delay with sampling at 0.41-bit interval.

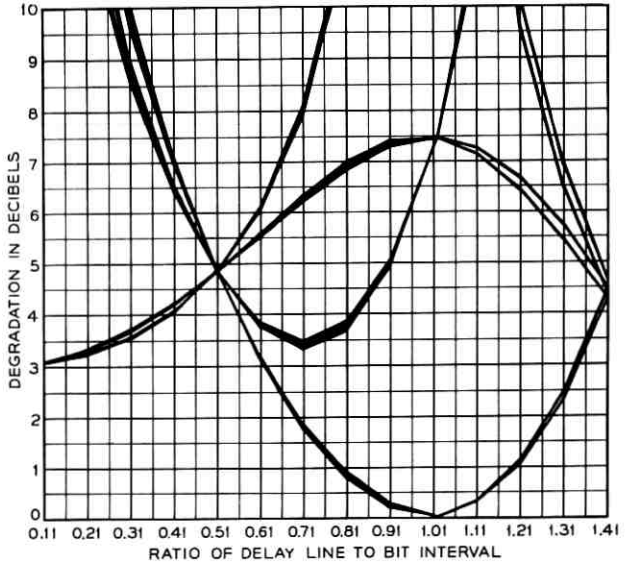


Fig. 18 — Db degradation vs differential delay with sampling at 0.51-bit interval.

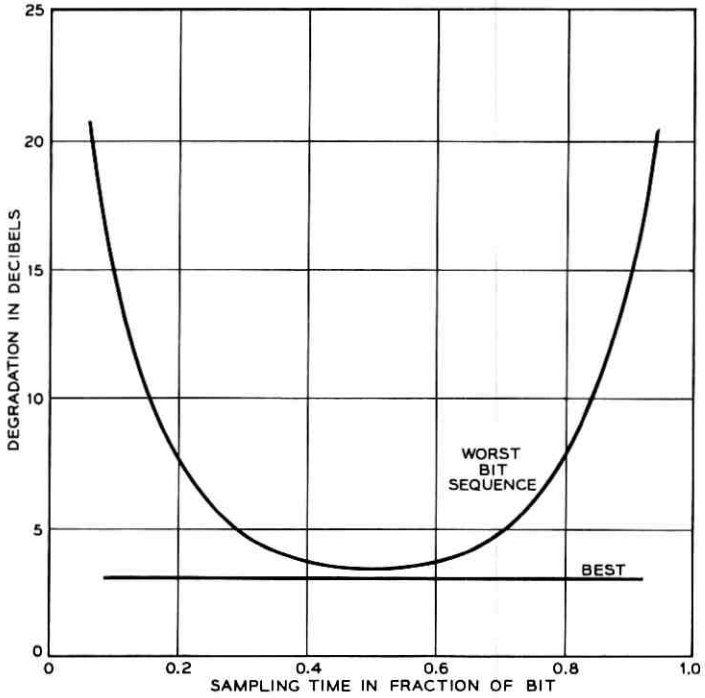


Fig. 19 — Db degradation vs sampling time for $d\varphi/dt$ detector with no delay distortion.

Fig. 20 exhibits the same information as Fig. 19 with the half-bit delay. In this case the worst sequence is a function of the sampling instant. At one-fourth bit and three-fourths bit sampling points referred to an origin at the transition points of the detector output, the intersymbol interference is zero, and at these instants the degradation is 4.8 db for all sequences. Nearer the center of the bit interval as seen from the detector output, the degradation is slightly greater. We nevertheless conclude that the best sampling instant is at this center, which we shall refer to as "mid-bit." By sampling at this time, we obtain a spread of half a bit interval tolerance to sampling jitter.

Similar curves are found when the receiver delay line is other than zero or a half-bit. In Fig. 21 we show the degradation of the worst sequence as a function of the receiver delay line. The zero and half-bit values are as shown on Figs. 19 and 20. The curve as shown applies to use of a delay line in which the phase shift is an odd multiple of $\pi/2$ at midband. We have previously shown that the signal can be recovered

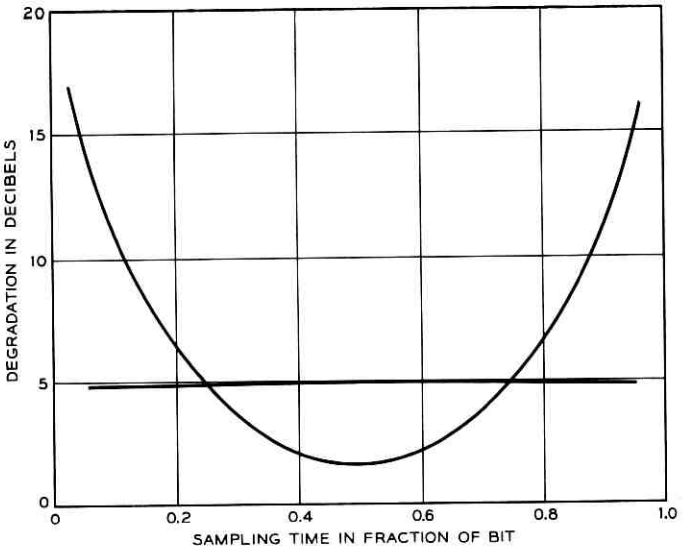


Fig. 20 — Db degradation vs sampling time for half-bit differential-delay detector with no delay distortion.

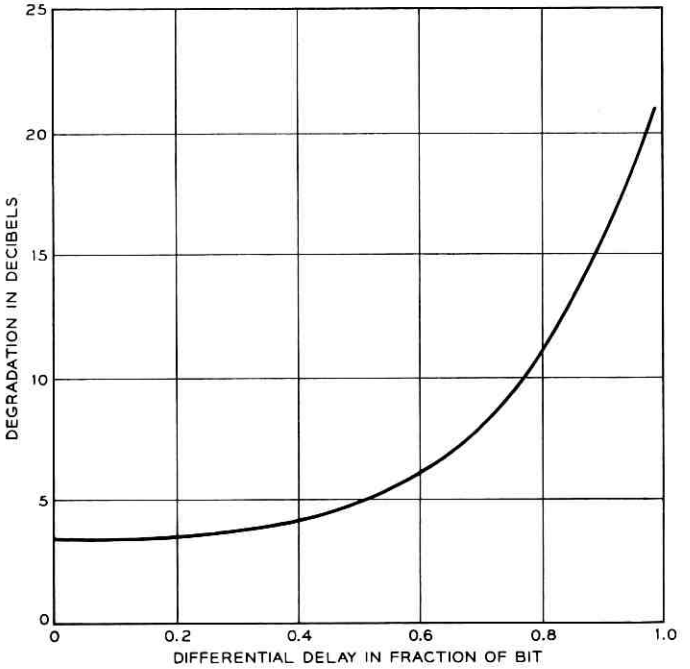


Fig. 21 — Db degradation vs differential delay for most vulnerable sequence sampled at mid-bit.

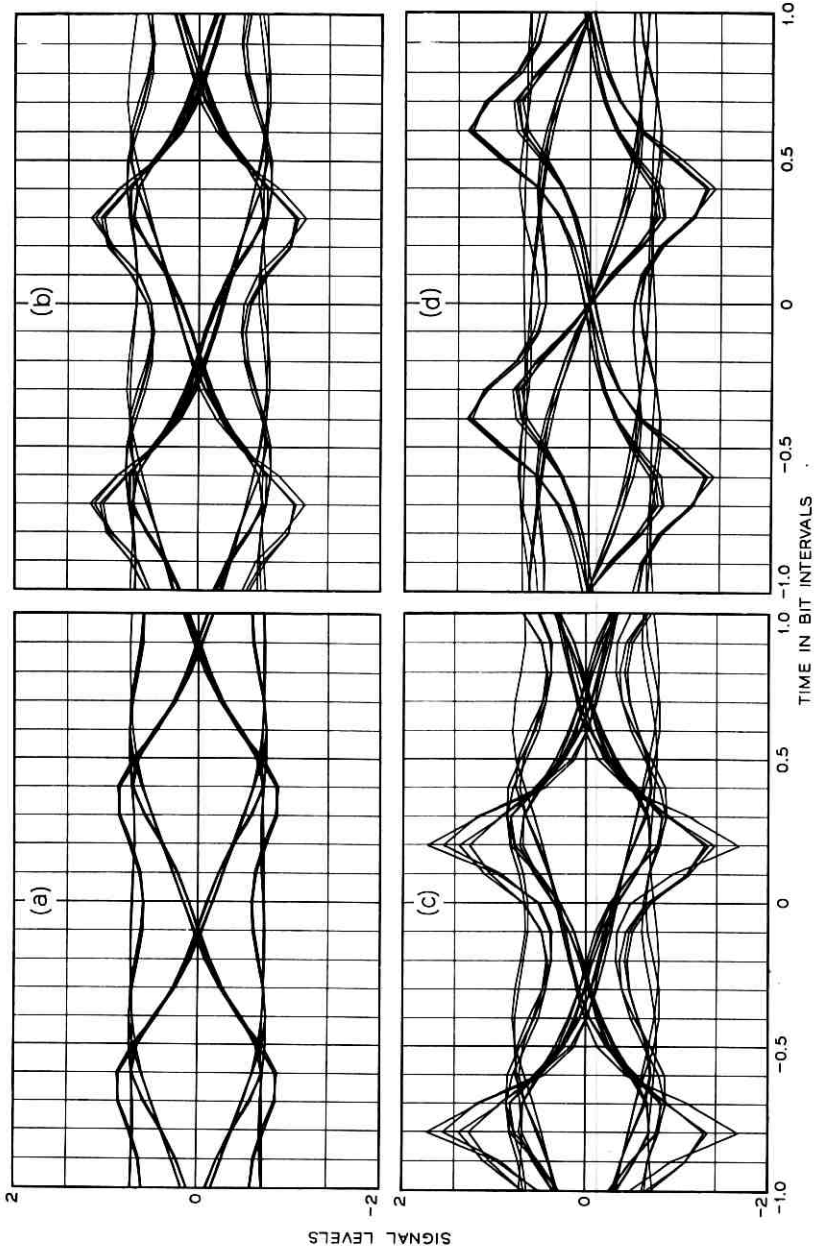


Fig. 22 — Eye patterns for $d\varphi/dt$ detector with delay distortion: (a) quadratic distortion of one bit interval; (b) quadratic, two bit intervals; (c) quadratic, three bit intervals; (d) linear, one bit interval.

with low degradation at a full-bit delay by differential encoding and use of a delay line with phase shift equal to a multiple of π at midband.

The above results indicate that, neglecting sampling jitter, least degradation results with least receiver delay, and the $d\varphi/dt$ receiver is best. Differences are small up to a half-bit. Addition of delay distortion alters this conclusion.

IX. EFFECTS OF DELAY DISTORTION

We preface our discussion by exhibiting sets of computed eye patterns for cases of linear and quadratic delay distortion in the channel. The amount of distortion is specified by the increment in delay measured in bit intervals between the center and edge of the transmission band. Results for the $d\varphi/dt$ detector are shown in Fig. 22 and for the half-bit delay differential detector in Fig. 23. The eye is found to close for smaller amounts of linear delay variation than for quadratic. As before, these results are not to be taken as quantitative measures of performance. The same numbers which determine the eye traces are also used in calculating error probabilities, but in a different way which precludes derivation of either final result directly from the other.

Fig. 24 shows the calculated degradation from the ideal for the $d\varphi/dt$ detector with various amounts of quadratic delay distortion measured in bit intervals. The results are plotted as a function of sampling time for the most vulnerable data sequence. There is an indication that the best sampling time is not at mid-bit when the delay distortion is large. However, the decreased tolerance to timing jitter would make such a shift undesirable. The effect of linear delay distortion, as exhibited in Fig. 25, is considerably worse. A comparison between effects of the two kinds of distortion is obtained in Fig. 26 by replotting the mid-bit sampling results of Figs. 24 and 25 as a function of delay distortion. As stated before, the $d\varphi/dt$ detector is equivalent to zero differential delay.

Figs. 27 and 28 present corresponding curves for the case of a half-bit differential delay. The shapes are similar to those for $d\varphi/dt$. There is slightly more tolerance to jitter, and the best sampling time is still at mid-bit.

In an effort to compare various receiver delay lines, previous data are cross-plotted in two ways. In Fig. 29, we in effect extend Fig. 21 to show how performance is affected by amount of differential delay when various amounts of specified linear and quadratic delay distortion are present in the channel. The curve for zero delay distortion replotted

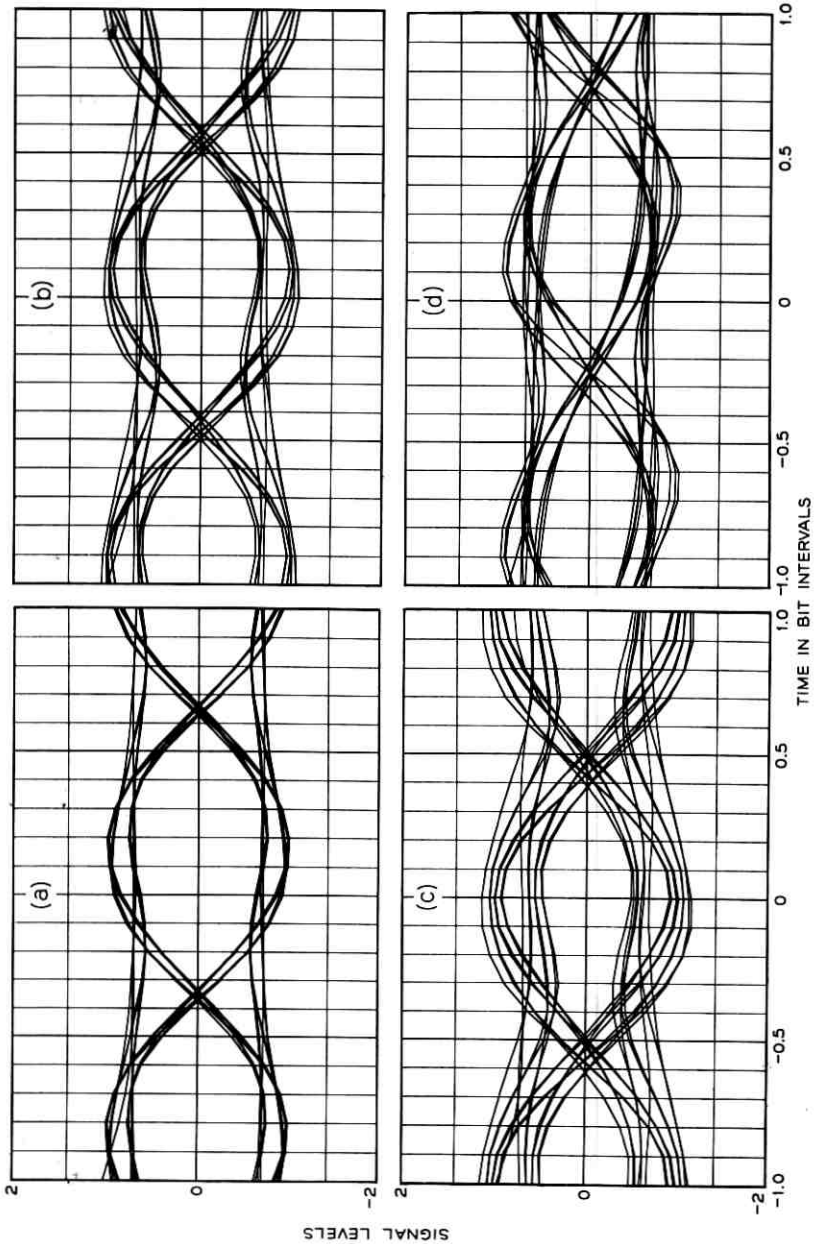


Fig. 23 — Eye patterns for half-bit differential delay detector with delay distortion: (a) quadratic distortion of one bit interval; (b) quadratic, two bit intervals; (c) quadratic, three bit intervals; (d) linear, one bit interval.

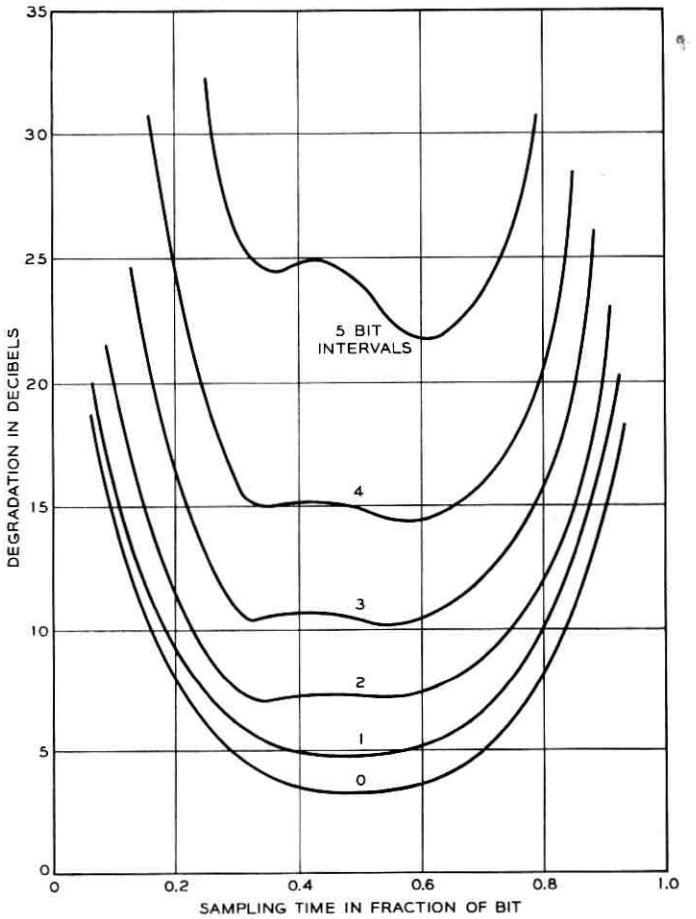


Fig. 24 — Db degradation of $d\varphi/dt$ detector vs sampling time for quadratic delay distortion.

from Fig. 21 shows the $d\varphi/dt$ receiver as best for a constant-delay channel. For quadratic delay distortion of two bit intervals or linear delay distortion of one bit interval, there is slightly less degradation at a half-bit of differential delay. For larger amounts of delay distortion the advantage from more receiver delay is greater.

The same conclusions can be drawn from Fig. 30, in which the abscissas are amounts of quadratic delay distortion and the curves are drawn for specified values of differential delay. The curves cross over from the condition of a preference for least differential delay with low delay

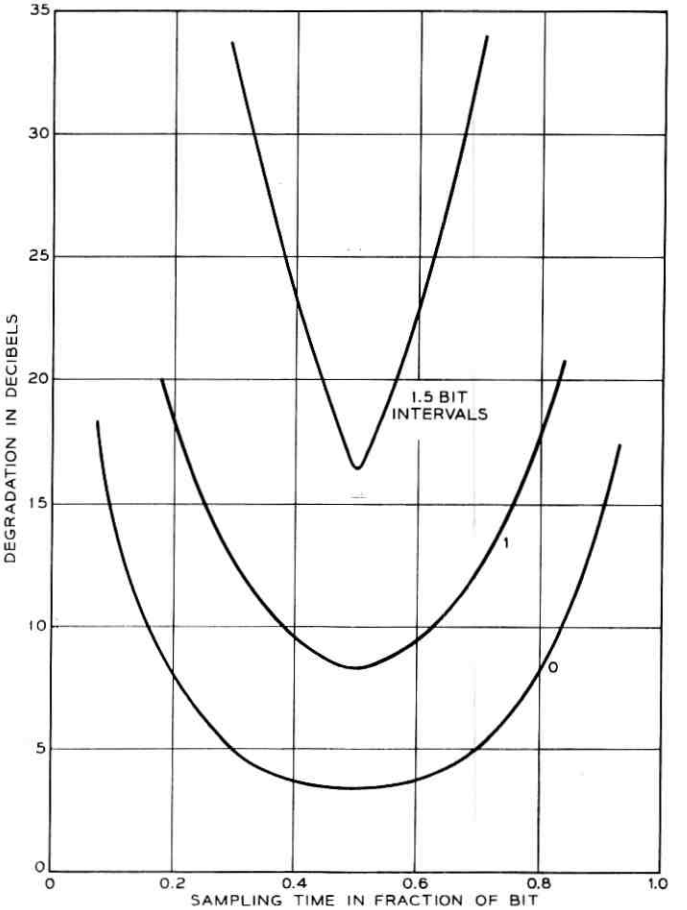


Fig. 25 — Db degradation of $d\varphi/dt$ detector vs sampling time for linear delay distortion.

distortion to a preference for greatest delay with high delay distortion. The exact choice of delay line depends on several factors. Say the maximum delay distortion is not greater than three bit intervals, then the half-bit delay is best over most of the range and only slightly worse over the rest. If the delay distortion is never more than one bit interval, a $d\varphi/dt$ detector is best, while if very high delay distortions are encountered, and fairly high degradation is permissible at lower values, a delay line of 0.7 bit interval would be better. It was also found that the longer delay lines provide more tolerance to jitter in the sampling time.

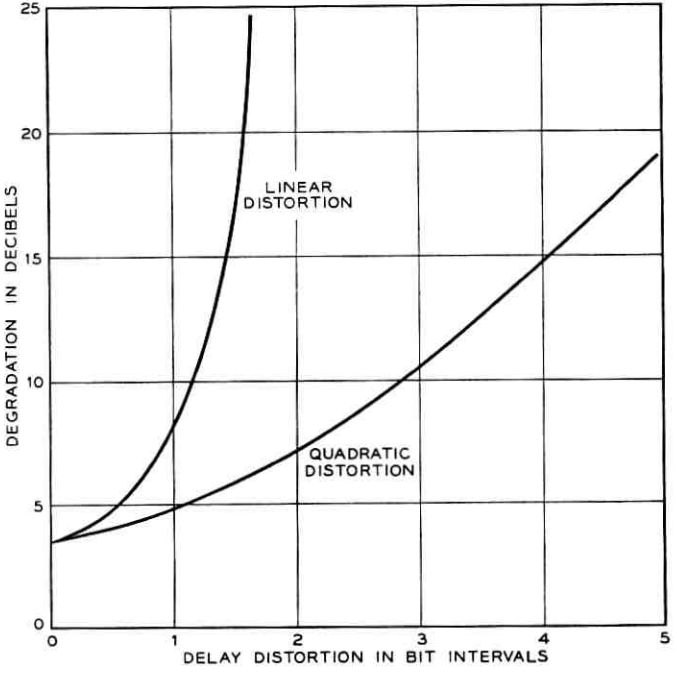


Fig. 26 — Db degradation of $d\varphi/dt$ detector vs linear and quadratic delay distortion.

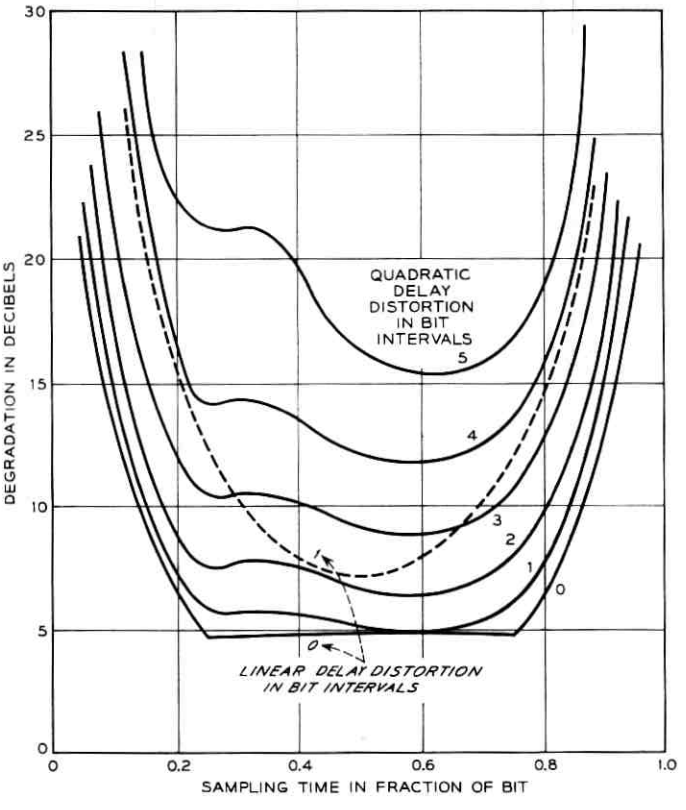


Fig. 27 — Db degradation with half-bit differential delay vs sampling time for quadratic delay distortion.

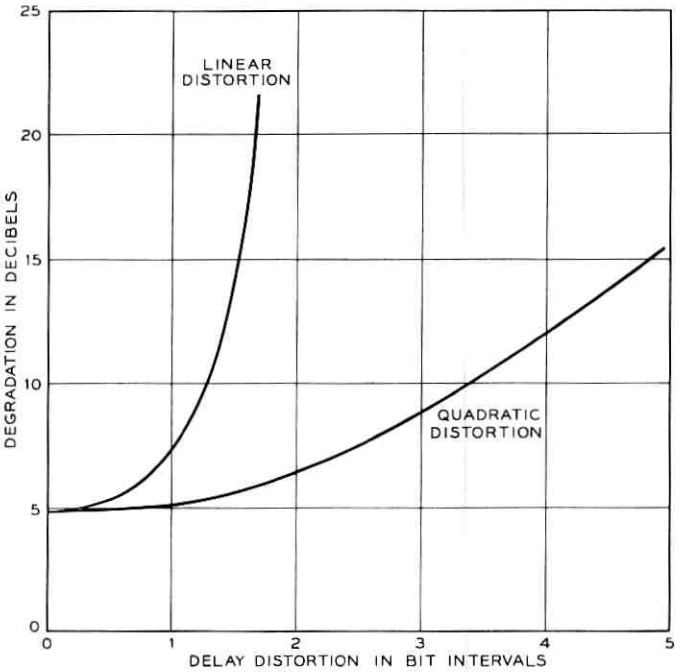


Fig. 28 — Db degradation with half-bit differential delay vs linear and quadratic delay distortion.

The asymptotic values of db impairment at large signal-to-noise ratio have been used throughout our discussion of Figs. 24–30 as a measure of performance for conditions of practical interest. As a check on the validity of this concept, complete curves of error probability vs signal-to-noise ratio have been computed from the exact formulas in representative cases. These curves are shown in Fig. 31 together with the asymptotic approximations. It appears that the latter are sufficient for most engineering applications.

APPENDIX A

Assume that a limiter is inserted in the undelayed input to the multiplier as shown in Fig. 6(b). The input to the limiter is then given by (13), which can also be written in the equivalent form

$$\begin{aligned} E(t) &= R(t) \cos [\omega_c t + \varphi(t)] \\ R(t) &= [x_1^2(t) + y_1^2(t)]^{\frac{1}{2}} \geqq 0 \\ \tan \varphi(t) &= y_1(t)/x_1(t). \end{aligned} \tag{71}$$

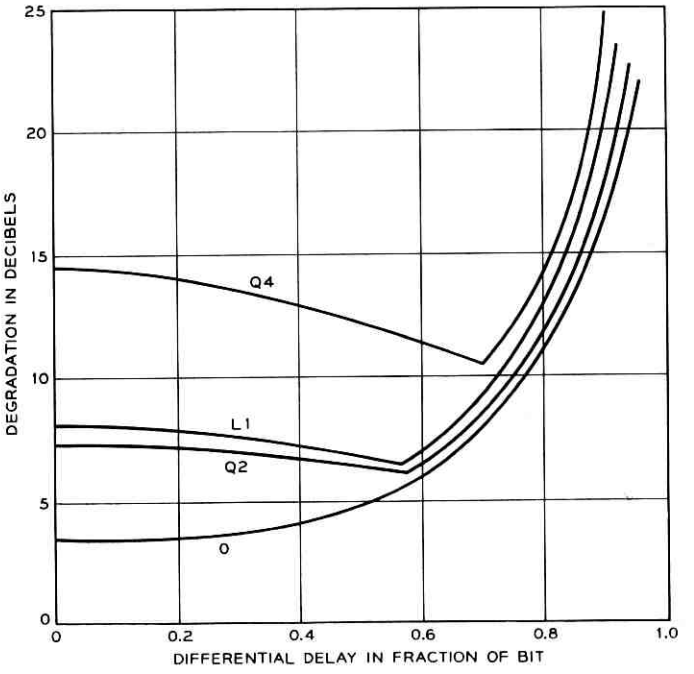


Fig. 29 — Db degradation vs differential delay for various amounts of linear and quadratic delay distortion. Curves are designated "L" for linear and "Q" for quadratic, followed by number of bit intervals of delay distortion.

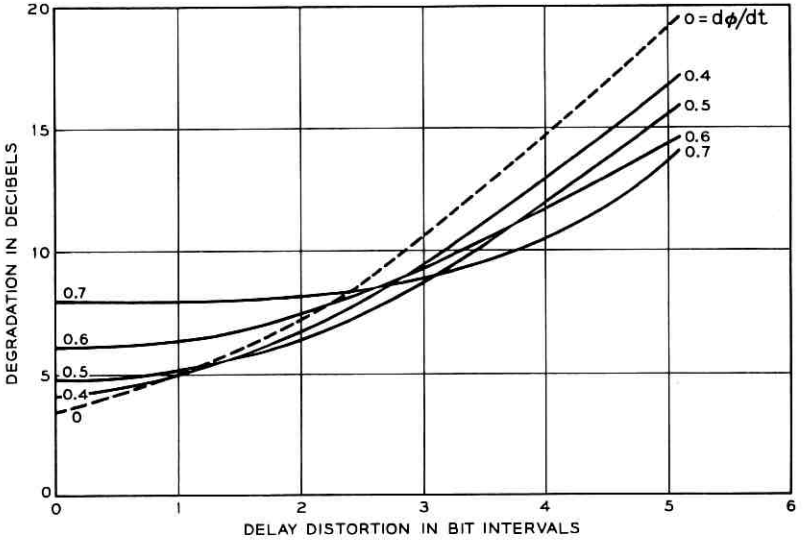


Fig. 30 — Db degradation vs quadratic delay distortion for various amounts of differential delay.

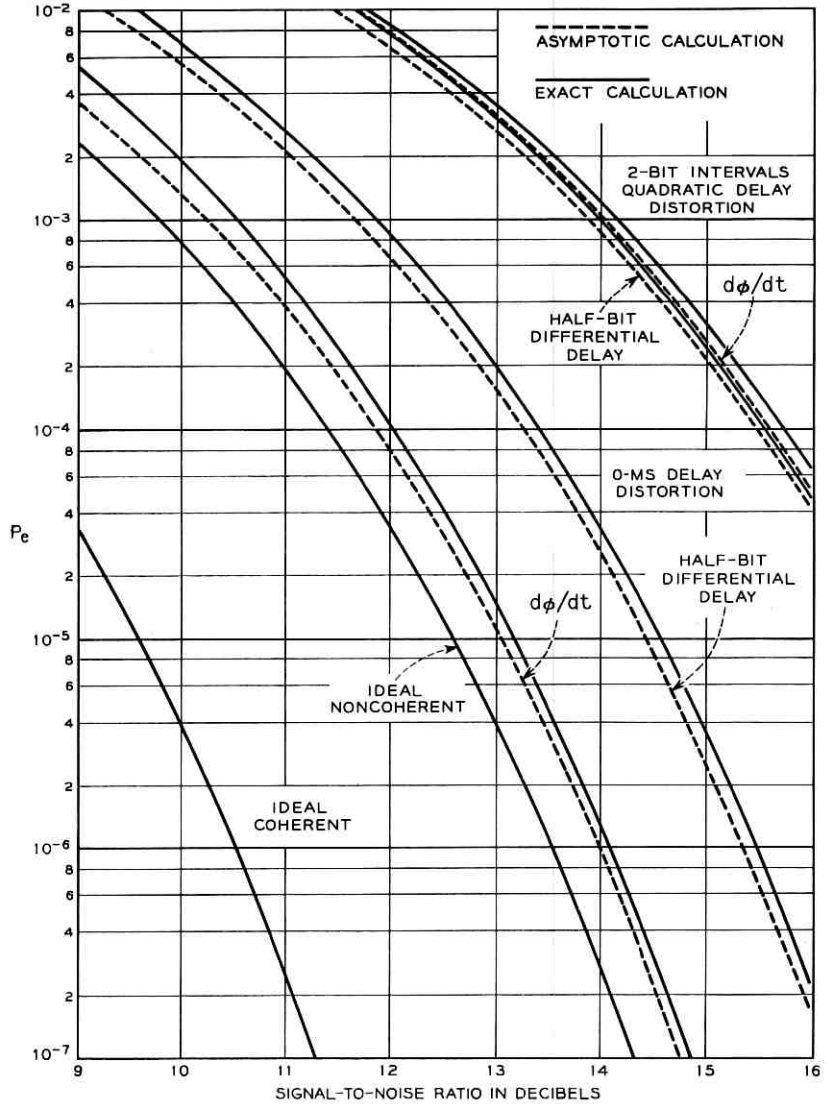


Fig. 31 — Comparison of results from exact and asymptotic formulas for error rates under representative conditions.

If the limiter is ideal, its output $E_L(t)$ is a positive constant E_0 when the value of the cosine function is positive and is a negative constant $-E_0$ when the cosine is negative. That is, if $z = \omega_c t + \varphi(t)$,

$$\begin{aligned} E_L(t) &= \begin{cases} E_0, & \cos z > 0 \\ -E_0, & \cos z < 0 \end{cases} \\ &= \frac{4E_0}{\pi} \sum_{n=0}^{\infty} \frac{(-)^n \cos (2n+1)z}{2n+1}. \end{aligned} \tag{72}$$

Multiplying $E_L(t)$ by the delayed signal and substituting $\cos \omega_c \tau = 0$, $\sin \omega_c \tau = -1$, gives:

$$\begin{aligned} E_L(t)E(t-\tau) = \frac{2E_0R(t-\tau)}{\pi} \sum_{n=0}^{\infty} \frac{(-)^n}{2n+1} \\ \cdot \{\sin [2n\omega_c t + (2n+1)\varphi(t) - \varphi(t-\tau)] \\ - \sin [2(n+1)\omega_c t + (2n+1)\varphi(t) \\ + \varphi(t-\tau)]\}. \end{aligned} \tag{73}$$

If it is possible by low-pass filtering to accept the term

$$\sin [\varphi(t) - \varphi(t-\tau)]$$

while rejecting the next higher-frequency terms $\sin [2\omega_c t + 3\varphi(t) - \varphi(t-\tau)]$ and $\sin [2\omega_c t + \varphi(t) + \varphi(t-\tau)]$, the filtered response becomes:

$$\begin{aligned} V_{lf}(t) &= \frac{2E_0R(t-\tau)}{\pi} \sin [\varphi(t) - \varphi(t-\tau)] \\ &= \frac{2E_0R(t-\tau)}{\pi} (\sin \varphi \cos \varphi_d - \cos \varphi \sin \varphi_d) \\ &= \frac{2E_0R(t-\tau)(y_1x_{1d} - x_1y_{1d})}{\pi(x_1^2 + y_1^2)^{\frac{1}{2}}(x_{1d}^2 + y_{1d}^2)^{\frac{1}{2}}}. \end{aligned} \tag{74}$$

The only term in (74) which can have both plus and minus signs is $y_1x_{1d} - x_1y_{1d}$, which is the same term found to be the basis of binary decisions in (17) for the case in which a pure product was taken. The switched modulator therefore gives the same error performance as the productor if there is sufficient frequency separation between the desired low-frequency output and the sidebands on $2\omega_c$.

An ideal limiter was assumed for simplicity in the argument just given, but the equivalence can be proved for a wide class of nonlinear devices in one of the two paths. It is sufficient that the output of the

device is of the same polarity as the input and can be expanded as a Fourier series in terms of the input values. We can then write

$$E_L(t) = F(R \cos z) = \frac{F_0}{2} + \sum_{n=1}^{\infty} F_n \cos nz \tag{75}$$

$$F_n = \frac{1}{\pi} \int_{-\pi/2}^{3\pi/2} F(R \cos z) \cos nz \, dz. \tag{76}$$

Note that R varies with time but cannot be negative.

In particular

$$F_1 = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} F(R \cos z) \cos z \, dz + \frac{1}{\pi} \int_{\pi/2}^{3\pi/2} F(R \cos z) \cos z \, dz. \tag{77}$$

In the first integral cos z is positive and hence $F(R \cos z)$ is positive. In the second integral cos z is negative, $F(R \cos z)$ is negative, and the product is positive. It follows that F_1 is positive.

If we repeat the calculation leading to (74) with the Fourier series (75) replacing (72), we find that $2E_0/\pi$ is replaced by F_1, and since F_1 is positive, the conclusion is the same. It will however be more difficult to isolate the desired low-frequency term when F_0 and F_2 are present, since these coefficients lead to components centered about ω_c. It is therefore preferable that the nonlinear function F have odd symmetry about the origin, in which case the even-order coefficients vanish.

The more general argument is useful in showing that the switched modulator does not have to be perfect. Note that the equivalence is destroyed if limiters are inserted in both paths. There would then be sidebands on harmonics in both inputs to the modulator, and the beats between these sidebands would generate additional components in the low-frequency band.

APPENDIX B

Simplification of the Error Probability Formula and Determination of Its Asymptotic Form

First, replace the parameter b by

$$k = \frac{b}{\rho} = \frac{\bar{z}_2\bar{z}_4 - \bar{z}_1\bar{z}_3}{2\sigma_1\sigma_2\rho^2}. \tag{78}$$

The expression to be evaluated is now

$$\Lambda_1(\rho,a,k) = \frac{1}{2} - \frac{\rho}{2\sqrt{\pi}} \int_{-1}^{1} e^{-\rho^2 x^2} \operatorname{erf} [\rho(a\sqrt{1-x^2} - kx)]dx. \quad (79)$$

Substitute $\rho x = x'$ and replace the error function by its definition as an integral. The result is

$$\Lambda_1(\rho,a,k) = \frac{1}{2} - \frac{1}{\pi} \int_{-\rho}^{\rho} \int_{0}^{a\sqrt{\rho^2-x^2}-kx} e^{-(x^2+y^2)}dx\,dy. \quad (80)$$

A graph of the region of integration in the xy-plane is shown in Fig. 32. From this figure, we deduce that by a change to polar coordinates, we obtain the equivalent expression

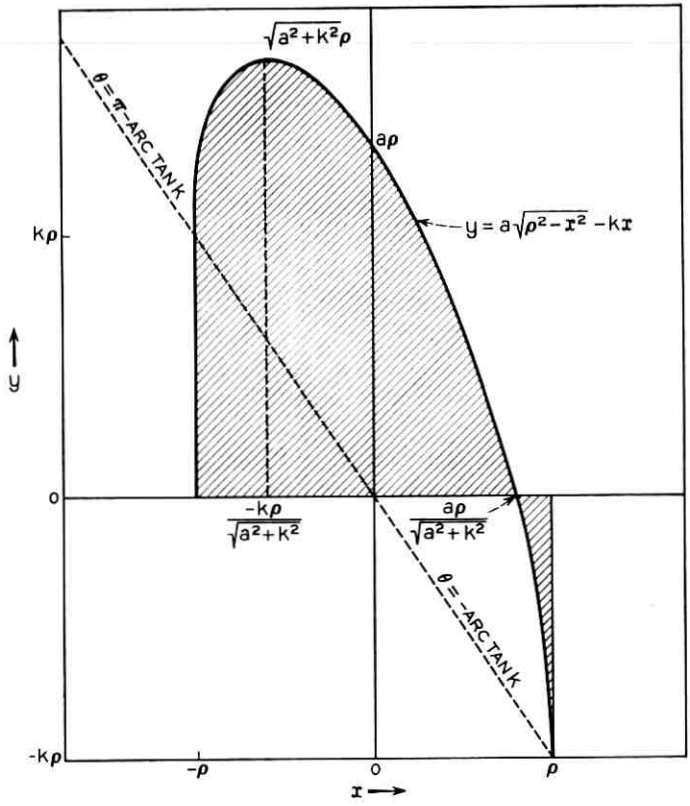


Fig. 32 — Region of integration for evaluation of double integral in (80).

$$\Lambda_1(\rho,a,k) = \frac{1}{2} - \frac{1}{\pi}\int_0^{\pi-\arctan k} d\theta \int_0^{f_1(\theta)} e^{-r^2} r\, dr - \frac{1}{\pi}\int_{\pi-\arctan k}^{\pi} d\theta \int_0^{-\rho/\cos\theta} e^{-r^2} r\, dr - \frac{1}{\pi}\int_0^{-\arctan k} d\theta \int_{f_1(\theta)}^{\rho/\cos\theta} e^{-r^2} r\, dr \tag{81}$$

where

$$f_1(\theta) = A\rho[(\sin\theta + k\cos\theta)^2 + a^2\cos^2\theta]^{-1/2}. \tag{82}$$

The integration with respect to r can be performed, and after some subsequent combining of terms, we find

$$\Lambda_1(\rho,a,k) = \frac{1}{2\pi}\int_0^{\pi} \exp\left[-\frac{a^2\rho^2}{(\sin\theta + k\cos\theta)^2 + a^2\cos^2\theta}\right] d\theta. \tag{83}$$

By the substitution $2\theta = \varphi$, (83) transforms to

$$\Lambda_1(\rho,a,k) = \frac{1}{4\pi}\int_0^{2\pi} \cdot \exp\left[-\frac{2a^2\rho^2}{k^2 + a^2 + 1 + (k^2 + a^2 - 1)\cos\varphi - 2k\sin\varphi}\right] d\varphi. \tag{84}$$

We then note that

$$(k^2 + a^2 - 1)\cos\varphi - 2k\sin\varphi = [(k^2 + a^2 - 1)^2 + 4k^2]^{\frac{1}{2}} \cos\left(\varphi + \arctan\frac{2k}{k^2 + a^2 - 1}\right). \tag{85}$$

Taking advantage of the fact that the range of integration is one complete period of the integrand in φ, we can replace the sum of φ and a constant angle by a new variable without changing the limits. Noting furthermore that the integrand then becomes an even function of the variable of integration, we obtain finally

$$\Lambda_1(\rho,a,k) = G(c,d) = \frac{1}{2\pi}\int_0^{\pi} \exp\left[-\frac{c^2}{1 + d^2\cos\varphi}\right] d\varphi \tag{86}$$

where c^2 and a^2 have the values given by (30) and (31).

Applying the method of steepest descents to the case in which c^2 is large, we write

$$G(c,d) = \frac{1}{2\pi}\int_0^{\pi} e^{-c^2\Phi(\varphi)}d\varphi \tag{87}$$

$$\begin{aligned}
\Phi(\varphi) &= (1 + d^2\cos\varphi)^{-1} \\
\Phi'(\varphi) &= d^2(1 + d^2\cos\varphi)^{-2}\sin\varphi \\
\Phi'(\varphi_0) &= 0 \text{ at } \varphi_0 = 0 \text{ or } \pi \\
\Phi''(\varphi_0) &= d^2(1 + d^2\cos\varphi_0)^{-2}\cos\varphi_0 > 0 \quad \text{for} \quad \varphi_0 = 0.
\end{aligned} \tag{88}$$

Hence, we set $\varphi_0 = 0$ and approximate $\Phi(\varphi)$ by

$$\begin{aligned}
\Phi(\varphi) &= \Phi(\varphi_0) + \tfrac{1}{2}\Phi''(\varphi_0)(\varphi - \varphi_0)^2 + \cdots \\
&= (1 + d^2)^{-1} + \tfrac{1}{2}d^2(1 + d^2)^{-2}\varphi^2 + \cdots .
\end{aligned} \tag{89}$$

If $c^2 \gg 1$,

$$\begin{aligned}
G(c,d) &\sim \frac{1}{2\pi}\exp\left(-\frac{c^2}{1 + d^2}\right)\int_0^{\pi}\exp\left[-\frac{c^2d^2\varphi^2}{2(1 + d^2)^2}\right]d\varphi \\
&\sim \frac{1 + d^2}{2cd\sqrt{2\pi}}\exp\left(-\frac{c^2}{1 + d^2}\right).
\end{aligned} \tag{90}$$

APPENDIX C

Limit of Performance as Delay Approaches Zero

As the delay τ is made small, the delayed variables in (17) can be expressed in terms of the undelayed ones by the linear approximations,

$$\begin{aligned}
x_{1d} &\approx x_1 - \tau\dot{x}_1 \\
y_{1d} &\approx y_1 - \tau\dot{y}_1 ,
\end{aligned} \tag{91}$$

where the dot signifies the derivative with respect to time. In the limit we then find

$$2E_{1f} \approx \tau(x_1\dot{y}_1 - y_1\dot{x}_1) . \tag{92}$$

The amplitude of the detected signal approaches zero as the delay is made small, but if τ is not actually zero, the lack of output can be compensated by linear amplification. Hence in the absence of imperfections other than additive noise in the channel, binary decisions are made on the basis of the sign of $x\dot{y}_1 - y_1\dot{x}_1$.

In the case of a $d\varphi/dt$ detector, binary decisions are made on the basis of the sign of

$$\frac{d\varphi}{dt} = \frac{d}{dt} \arctan \frac{y_1}{x_1} = \frac{x_1\dot{y}_1 - y_1\dot{x}_1}{x_1^2 + y_1^2}. \tag{93}$$

Since $x_1^2 + y_1^2$ cannot change sign, the decisions are actually made on the sign of $x_1\dot{y}_1 - y_1\dot{x}_1$. We conclude that the binary error rates of the two systems must approach equality as τ approaches zero.

APPENDIX D

Computational Details

For the pulse response (36) we use the impulse response corresponding to a raised-cosine spectrum, namely

$$g(t) = \frac{\sin\left(\frac{2\pi}{T}t\right)}{\pi\left(\frac{2t}{T}\right)\left[1 - \left(\frac{2t}{T}\right)^2\right]}. \tag{94}$$

The sampled value of the pulse train at $t = T + \epsilon T$ is then given by

$$s_1[T(1+\epsilon)] = \sum_n (-1)^n b_n \frac{\sin 2\pi[1 - \delta_2 - n]}{2\pi(1 - \delta_2 - n)[1 - 4(1 - \delta_2 - n)^2]} \equiv s_1^0(\delta_2) \tag{95}$$

and the delayed version after τ seconds by

$$s_1[T(1 + \epsilon - \delta)] = \sum_n (-1)^n b_n \frac{\sin 2\pi[1 - \delta_1 - n]}{2\pi(1 - \delta_1 - n)[1 - 4(1 - \delta_1 - n)^2]} = s_1^0(\delta_1) \tag{96}$$

where

$$\delta = \tau/T, \qquad \delta - \epsilon = \delta_1, \qquad \delta_1 - \delta = \delta_2. \tag{97}$$

In terms of the two new variables δ_1, δ_2, the noise correlation is given by

$$r(\delta) = \frac{\sin 2\pi(\delta_1 - \delta_2)}{2\pi(\delta_1 - \delta_2)[1 - 4(\delta_1 - \delta_2)^2]}. \tag{98}$$

The observed voltage or current ξ can now be written as a function of δ_1 and δ_2 in the following form:

$$\xi(\delta_1, \delta_2) = [As_1^0(\delta_2) + x][-A \sin \pi\delta_1 - y_d] + [A \sin \pi\delta_2 + y][As_1^0(\delta_1) + x_d]. \tag{99}$$

In the calculation of eye patterns, the noise samples x, x_d, y, and y_d are omitted. By dropping these terms in (99) and substituting successively $\epsilon = 0$ and $\epsilon = \delta$, we verify that the samples at these instants depend on only one value in the data sequence. This confirms the statements made in Section III in the discussion of eye patterns. When the noise samples are inserted, the entire previous signal history exerts its effect at all sampling instants, and in fact the preferred sampling instant from the standpoint of low error rate and tolerance to jitter in sampling time is not necessarily the one in which intersymbol interference vanishes in the absence of noise.

The probability of error for a particular sequence is:

$$\begin{aligned} P_e = {} & \tfrac{1}{2} \Pr \{[As_1^{+}(\delta_2) + x][-A \sin \pi\delta_1 - y_d] \\ & + [A \sin \pi\delta_2 + y][As_1^{+}(\delta_1) + x_d] < 0\} \\ & + \tfrac{1}{2} \Pr \{[As_1^{-}(\delta_2) + x][-A \sin \pi\delta_1 - y_d] \\ & + [A \sin \pi\delta_2 + y][As_1^{-}(\delta_1) + x_d] \geqq 0\} \end{aligned} \tag{100}$$

where

$$\begin{aligned} & s_1^{\pm}(u) \\ & = \sum_{n \neq 1} (-1)^n b_n \frac{\sin 2\pi[1 - \mu - n]}{2\pi(1 - \mu - n)[1 - 4(1 - \mu - n)^2] \pm r(\mu)} . \end{aligned} \tag{101}$$

Applying the transformation in (21) we obtain:

$$\begin{aligned} \bar{z}_1^{\pm} &= (A/2)[s_1^{\pm}(\delta_2) - \sin \pi\delta_2 + s_1^{\pm}(\delta_1) - \sin \pi\delta_1] \\ \bar{z}_2^{\pm} &= (A/2)[s_1^{\pm}(\delta_2) + \sin \pi\delta_2 + s_1^{\pm}(\delta_1) + \sin \pi\delta_1] \\ \bar{z}_3^{\pm} &= (A/2)[-s_1^{\pm}(\delta_2) + \sin \pi\delta_2 + s_1^{\pm}(\delta_1) - \sin \pi\delta_1] \\ \bar{z}_4^{\pm} &= (A/2)[s_1^{\pm}(\delta_2) + \sin \pi\delta_2 - s_1^{\pm}(\delta_1) - \sin \pi\delta_1] . \end{aligned} \tag{102}$$

The required parameters in (25)–(27) are

$$\rho^2 = \frac{M}{8A^2} \frac{\bar{z}_1^{\,2} + \bar{z}_2^{\,2}}{1 + r(\delta_1 - \delta_2)} \tag{103}$$

$$a = \frac{\bar{z}_1\bar{z}_4 + \bar{z}_2\bar{z}_3}{\bar{z}_1^{\,2} + \bar{z}_2^{\,2}} \sqrt{\frac{1 + r(\delta_1 - \delta_2)}{1 - r(\delta_1 - \delta_2)}} \tag{104}$$

$$k = \frac{\bar{z}_2\bar{z}_4 - \bar{z}_1\bar{z}_3}{\bar{z}_1^{\,2} + \bar{z}_2^{\,2}} \sqrt{\frac{1 + r(\delta_1 - \delta_2)}{1 - r(\delta_1 - \delta_2)}} . \tag{105}$$

As pointed out before, the asymptotic degradation is given by

$$\text{degradation} = 10 \log_{10} (1/\kappa)\text{db} \tag{106}$$

where

$$\kappa = \frac{k^2 + a^2 + 1 + [(k^2 + a^2 - 1)^2 + 4k^2]^{\frac{1}{2}}}{2a^2\rho_1^2} \tag{107}$$

and

$$\rho = M\rho_1 . \tag{108}$$

When delay distortion is present in the channel, a convolution is performed to evaluate the resulting in-phase and quadrature components of the noise-free input to the detector. It is convenient to combine the assumed delay distortion with the equivalent amplitude characteristic arising from signal pulse shaping, sending filter, transmission line, and receiving filter to form a single complex transmittance function

$$H(\omega) = A(\omega)e^{j\theta(\omega)} . \tag{109}$$

The impulse response of the medium is then

$$\begin{aligned} h(t) &= \int_{-\infty}^{\infty} H(\omega)e^{j\omega t}\, d\omega/2\pi \\ &= \int_{-\infty}^{\infty} A(\omega)e^{j[\theta(\omega)+\omega t]}\, d\omega/2\pi \\ &= \int_{0}^{\infty} A(\omega)\cos\,[\theta(\omega) + \omega t]\, d\omega/\pi . \end{aligned} \tag{110}$$

Let $A(\omega) = B(\omega - \omega_c) = B(\upsilon)$, $\theta(\omega) = \varphi(\omega - \omega_c) + \varphi_0 = \varphi(\upsilon) + \varphi_0$, and decompose (110) into the in-phase and quadrature components

$$\begin{aligned} h(t) &= \int_{-\omega_0}^{\infty} B(\upsilon)\cos\,[\varphi(\upsilon) + \upsilon t + \omega_c t + \varphi_0]\, d\upsilon/\pi \\ &= h_1(t)\cos\,\omega_c t - h_2(t)\sin\,\omega_c t, \end{aligned} \tag{111}$$

where

$$\begin{aligned} h_1(t) &= \int_{-\omega_0}^{\infty} B(\upsilon)\cos\,[\varphi(\upsilon) + \upsilon t + \varphi_0]\, d\upsilon/\pi \\ h_2(t) &= \int_{-\omega_0}^{\infty} B(\upsilon)\sin\,[\varphi(\upsilon) + \upsilon t + \varphi_0]\, d\upsilon/\pi . \end{aligned} \tag{112}$$

Since the medium is assumed to be linear, the signal input to the detector is the convolution of the input and the impulse response,

$$V_1(t) = \int_{-\infty}^{\infty} V(t - \tau)h(\tau)d\tau , \tag{113}$$

where $V(t)$ can be written in the form:

$$V(t) = P_0(t) \cos \omega_c t - Q_0(t) \sin \omega_c t. \tag{114}$$

Then from (111),

$$V_1(t) = \int_{-\infty}^{\infty} [P_0(t - \tau) \cos (\omega_c t - \omega_c \tau) - Q_0(t - \tau) \sin (\omega_c t - \omega_c \tau)][h_1(\tau) \cos \omega_c \tau - h_2(\tau) \sin \omega_c \tau] d\tau . \tag{115}$$

Dropping the double-frequency terms, we obtain for the input to the detector

$$V_r(t) = P(t) \cos \omega_c t - Q(t) \sin \omega_c t \tag{116}$$

where

$$\begin{aligned} P(t) &= \int_{-\infty}^{\infty} [P_0(t - \tau)h_1(\tau) - Q_0(t - \tau)h_2(\tau)] d\tau/2 \\ Q(t) &= \int_{-\infty}^{\infty} [P_0(t - \tau)h_2(\tau) + Q_0(t - \tau)h_1(\tau)] d\tau/2 . \end{aligned} \tag{117}$$

REFERENCES

1. Davey, J. R., Signal Space Diagrams, B.S.T.J., **43,** Nov., 1964, pp. 2973–2983.
2. Brand, S., and Carter, C. W., A 1,650-Bit-Per-Second Data System for Use over the Switched Telephone Network, A.I.E.E. Trans., Pt. I, Comm. and Elect., **80,** Jan., 1962, pp. 652–661.
3. 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.
4. Sunde, E. D., Ideal Pulses Transmitted by AM and FM, B.S.T.J., **38,** Nov., 1959, pp. 1357–1426.
5. Nyquist, H., Certain Topics in Telegraph Transmission, Trans., A.I.E.E **47,** pp. 617–644, Apr., 1928.

Nonlinear *RLC* Networks

By CHARLES A. DESOER* and JACOB KATZENELSON

(Manuscript received July 16, 1964)

This article considers the question of existence and uniqueness of the response of nonlinear time-varying RLC networks driven by independent voltage and current sources. It is proved that under certain conditions the response exists, is unique, and is defined by a set of ordinary differential equations satisfying some Lipschitz conditions. These conditions are of two types: (1) the network elements must have characteristics which satisfy suitable Lipschitz conditions and (2) the network must satisfy certain topological conditions. It should be noted that elements with nonmonotonic characteristics are allowed and that the element characteristics need to be continuous but not differentiable.

I. INTRODUCTION

This article considers the questions of existence and uniqueness of the response of nonlinear time-varying *RLC* networks. It is proved that under conditions imposed on the network elements and the network topology the response exists, is unique, and is defined by a set of ordinary differential equations satisfying some Lipschitz conditions. Thus, from the conditions imposed on the network it follows that the response of a network of this class is continuous whenever the sources applied to the network are continuous functions of time. In other words, for the class of networks under consideration, jump phenomena (of the type that occur in relaxation oscillators) are excluded.[1]

One motivation for studying this problem is the construction of nonlinear network models for physical devices and processes. The behavior of these models is often investigated by simulation studies performed on digital computers. It is clear that in order to get meaningful answers the existence and uniqueness of the model's response have to be assured. The simulation study requires the setting up of an appropriate set of differential equations and their integration. As networks of the class

* On leave of absence from the Department of Electrical Engineering, University of California, Berkeley, California.

considered here do not have jump phenomena, their equations can be integrated by some standard subroutines.

This article may be viewed as an extension to the *RLC* case of the articles by R. J. Duffin[2,3,4] and G. Birkhoff and J. B. Diaz[5] which were devoted to nonlinear resistive networks. We make heavy use of topological considerations and had to extend the techniques developed for the linear case by many people[6,7,8] P. R. Bryant in particular.[9,10] For further references see Ref. 16.

In the next section, we classify the network elements and exhibit the basic assumptions which hold for the remainder of the article. Some simple nonlinear circuits are also considered. Section III presents some standard reductions of sources and the definition of determinateness. Section IV deals with one-element-kind networks; its theorems are generalizations of Duffin's work and include some of his theorems as corollaries. The main result of the article is Theorem IV in Section V, which states the conditions under which a nonlinear *RLC* network is determinate. The conditions are of two types: (*i*) every characteristic has to satisfy suitable Lipschitz conditions and (*ii*) the network has to satisfy certain topological conditions. It has to be noted that, first, elements with nonmonotonic characteristics are allowed and, second, that each characteristic has to be representable by a function which is continuous but not necessarily differentiable. Finally, in Section VI we introduce a symbolic notation which allows us to write the differential equations for the nonlinear case in a manner which resembles that of the linear case.

II. ELEMENTS AND SIMPLE CIRCUITS

2.1 *Elements*

We assume that the reader has some familiarity with network theory, so that the basic concepts need not be defined.[11,12] A network may be considered as a set of points, called *nodes*, and a set of connecting *branches*. Each branch represents a physical *two-pole*. We assume that the voltage drop across each two-pole and the current through each two-pole can be measured at any time. The sign conventions are shown in Fig. 1: if, with respect to some arbitrary reference, the potential of A is larger (smaller) than the potential of B, then v is positive (negative); if the current actually flows in the direction of the arrow (opposite to the arrow) then i is positive (negative). Thus the product vi gives the power delivered by the outside world to the two-pole under consideration.

In most of the following, the branches consist of either a single source or a single *element* such as a resistor, an inductor or a capacitor. For each of these elements we shall adopt very broad definitions which we will narrow down in stating specific results. A two-pole is called a *resistor* if it is defined, for each t, by a set of ordered pairs (v, i), where v and i are finite numbers representing all the possible values, at time t, of the voltage and the current associated with the resistor. If the set of ordered pairs is independent of t, the resistor is said to be *time-invariant*. The set of (v, i) is called the *characteristic* of the resistor; for example, the characteristic of an *ideal diode* is given by

$$\{(0,i): 0 \leqq i < \infty\} \cup \{(v,0): -\infty < v \leqq 0\} .$$

A resistor is called *current-controlled* if, for all time and all currents in the interval $(-\infty, \infty)$, the voltage $v(t)$ is a function* of the current $i(t)$ and time t (we shall write $v(t) = \mathcal{R}(i(t),t)$), and the function $\mathcal{R}(i,t)$ is a piecewise-continuous function† of t for each fixed number i. A *voltage-*

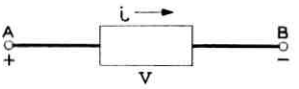


Fig. 1 — Sign conventions for two-pole.

controlled resistor is defined in the dual manner. For example, a voltage source is a current-controlled resistor and a current source is a voltage-controlled resistor. If a resistor is *current-controlled* and *time-invariant* then the characteristic can be represented by a function $v = \mathcal{R}(i)$. A resistor is called a *one-to-one resistor* if, for each t, the voltage is related to the current by a one-to-one mapping from $(-\infty, \infty)$ onto $(-\infty, \infty)$ which may depend on time.

A two-pole is called an *inductor* if it is defined, for each t, by a set of ordered pairs (φ, i) which represent the instantaneous flux and current associated with the inductor. The voltage across the inductor is given by $v = d\varphi/dt$. The *current-controlled* inductor, the *flux-controlled* inductor and the *one-to-one* inductor are defined as in the case of resistors. In the first two cases, if the elements are time-invariant, we shall write $\varphi = \mathcal{L}(i)$ and $i = \Gamma(\varphi)$, respectively.

* Unless specifically indicated, we follow modern usage: each function is single-valued; i.e., to each element of its domain it associates one and only one element of its range.

† A vector-valued function of time is said to be piecewise-continuous if it is continuous in every finite interval except at a finite number of points where it is discontinuous. At these points the function has a finite limit on the left as well as on the right.

A two-pole is called a *capacitor* if it is defined, for each t, by a set of ordered pairs (q,v) which represent the instantaneous charge and voltage associated with the capacitor. The current through the capacitor is given by $i = dq/dt$. The *charge-controlled capacitor*, the *voltage-controlled capacitor* and the *one-to-one capacitor* are defined as in the case of resistors. In the first two cases, if the elements are time invariant, we shall write $v = \mathfrak{D}(q)$ and $q = \mathfrak{C}(v)$, respectively.

Throughout the article we consider only elements whose characteristics can be represented, at all times, by a function defined on the interval $(-\infty, \infty)$. For example, Fig. 2(a), (b) and (c) represents the characteristics at time t of three time-varying resistors; we consider only resistors of the type shown in Fig. 2(a) and (b), since they are current- and

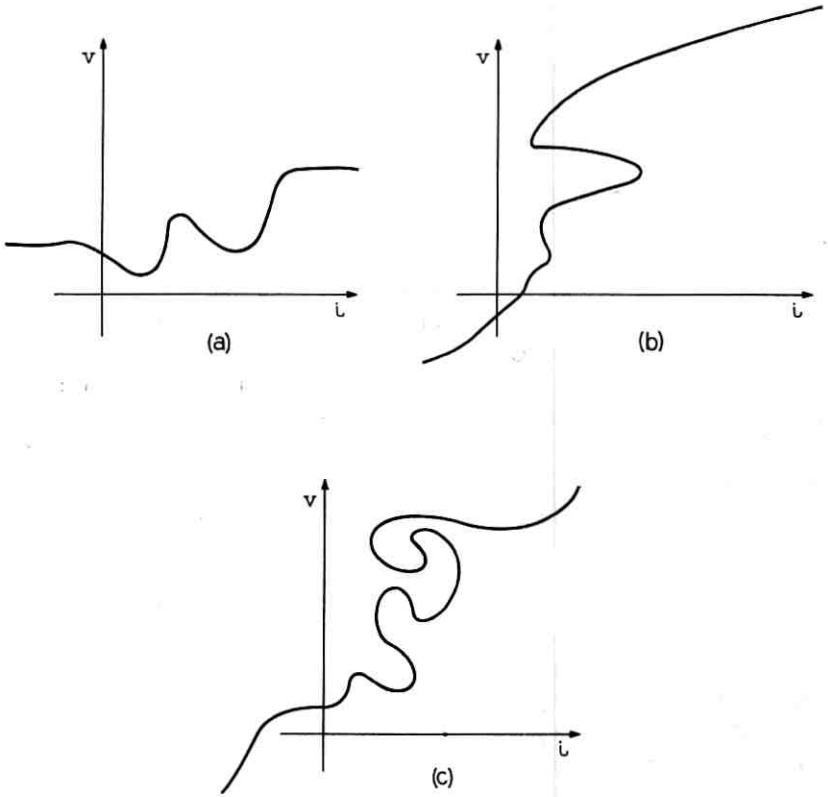


Fig. 2 — Characteristics at time t of three time-varying resistors: (a) and (b) are current- and voltage-controlled, respectively, while (c) is neither current- nor voltage-controlled.

voltage-controlled; these characteristics can be represented by

$$v(t) = \mathfrak{R}(i(t),t), \quad \text{and} \quad i(t) = \mathfrak{G}(v(t),t),$$

respectively. The characteristics of Fig. 2(c) cannot be represented in this way, and resistors of this type will not be considered.

Throughout the paper, whenever time-varying network elements are considered, it is assumed that the functions $\mathfrak{R}(\cdot,t)$, $\mathfrak{G}(\cdot,t)$, $\mathfrak{L}(\cdot,t)$, $\Gamma(\cdot,t)$, $\mathfrak{D}(\cdot,t)$, $\mathfrak{C}(\cdot,t)$ are piecewise-continuous functions of t for all fixed values of their first argument.

In addition to resistors, capacitors and inductors, our networks include voltage and current sources. Throughout this article we shall assume that the voltages of the voltage sources and currents of current sources are regulated functions of time.* For convenience we shall say that an element is continuous and monotonically increasing, when we mean that its characteristic is represented by a continuous monotonically increasing function which is defined on $(-\infty,\infty)$.

It is convenient to refer to functions like $\mathfrak{R}(\cdot,t)$ and $\mathfrak{D}(\cdot,t)$, which represent the characteristics of some elements, as the characteristics of the elements. This slight misuse of the concept of a function and a relation will be used only when there is no danger of confusion between the two.

2.2 *Two-Poles and Simple Connections of Two-Poles*

A two-pole is called *voltage-controlled* [current-controlled] if, for any initial time t_0 and for any initial state, the voltage $v(\cdot)$ [the current $i(\cdot)$] from t_0 on across its terminals determines uniquely the current $i(\cdot)$ through [the voltage $v(\cdot)$ across] the two-pole for $t \geqq t_0$.

A two-pole is said to be *one-to-one* if (a) it is both current-controlled and voltage-controlled and (b) it satisfies the following condition: for any initial state s_0, any initial time t_0, and any input current $i(\cdot)$, let $f(s_0,i)$ be the voltage appearing at the terminals; for any initial state s_0, any initial time t_0, and any input voltage $v(\cdot)$, let $g(s_0,v)$ be the current — then it is required that

$$g(s_0, f(s_0, i)) = i$$

for all initial states s_0 and all input currents $i(\cdot)$.

An immediate consequence of these definitions is that *any parallel connection of a finite number of voltage-controlled two-poles is voltage-controlled.*

* A vector-valued function of time is said to be regulated when, for all t, it has a limit on the left as well as a limit on the right.[13] A step function and a rectangular wave are regulated functions.

Consider the case where there are only two two-poles in the parallel connection. Let them be characterized by the functions

$$i_k = F_k(v, s_k(t_0)), \qquad (k = 1,2),$$

where v is the voltage across the parallel connection, i_k is the current through the kth two-pole, $s_k(t_0)$ is the state of the kth two-pole at time t_0. The v_k and i_k are real-valued functions defined on $[t_0, \infty)$. Kirchoff's current law implies that the current i through the parallel connection is given by

$$F_1(v, s_1(t_0)) + F_2(v, s_2(t_0));$$

hence, for fixed $(s_1(t_0), s_2(t_0))$, i is a function of v. This argument obviously extends, by induction, to the case where there are a finite number of two-poles.

A dual argument would show that *any series connection of a finite number of current-controlled two-poles is current-controlled.*

A parallel connection of current-controlled two-poles is not necessarily current-controlled. Refer to Fig. 3, which shows the characteristics of two current-controlled resistors. The dashed line shows the characteristic of the parallel connection: depending on the operating point there may be three distinct values of the voltage for the same input current. Dually,

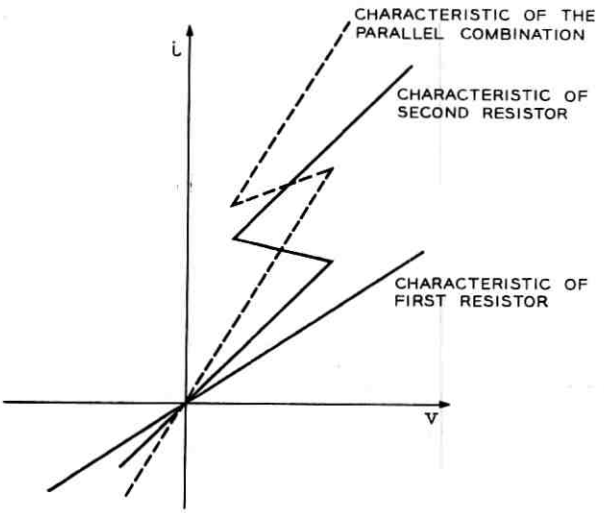


Fig. 3 — Parallel connection of two current-controlled resistors.

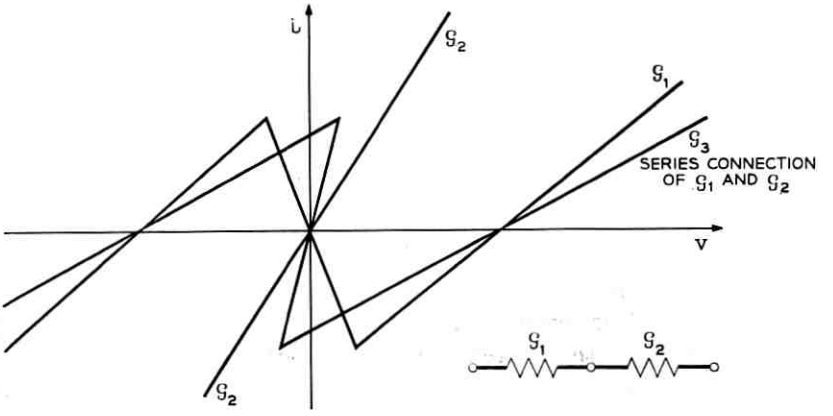


Fig. 4 — Series connection of two voltage-controlled resistors.

a series connection of voltage-controlled two-poles is not necessarily voltage-controlled.

To assume that each two-pole is one-to-one is not enough to cause both arbitrary parallel connections and arbitrary series connections to be one-to-one. Indeed, the well known characterization of continuous functions of bounded variation[14] implies that *any voltage-controlled resistor characteristic,* $i(t) = \mathfrak{R}(v(t),t)$, *that is continuous and of bounded variation in v can be obtained by connecting in parallel two one-to-one resistors whose characteristics are continuous and strictly monotonic.* (One resistor is monotonically increasing and the other is monotonically decreasing.) A dual statement holds for current-controlled resistors.

In fact, there are combined series and parallel connections of one-to-one two-poles that are neither voltage-controlled nor current-controlled. Refer to Fig. 4, which shows the series connection of $\mathcal{G}_1$ and $\mathcal{G}_2$. Fig. 5 shows how a voltage-controlled characteristic such as $\mathcal{G}_1$ may be obtained by connecting in parallel two one-to-one resistors. Putting the two resistors of characteristic $\mathcal{G}_1$ and $\mathcal{G}_2$ in series, we obtain (see Fig. 4) the characteristic $\mathcal{G}_3$, which is neither voltage-controlled nor current-controlled.

A (possibly time-varying) flux-controlled inductor is a voltage-controlled two-pole and, dually, a (possibly time-varying) charge-controlled capacitor is a current-controlled two-pole. If the inductor is flux-controlled, the current i is a function of the flux φ: $i(t) = \Gamma(\varphi(t),t)$. If $v(\cdot)$ is the voltage applied to the inductor and φ_0 is the flux through it at the initial time t_0, then by Lenz's law

$$\varphi(t) = \int_{t_0}^{t} v(t')dt' + \varphi_0$$

hence,

$$i(t) = \Gamma\left(\int_{t_0}^{t} v(t')dt' + \varphi_0, t\right) \quad \text{for all} \quad t.$$

2.3 *Examples of One-To-One Two-Poles*

We present here a set of sufficient conditions under which some elementary parallel or series connections of circuit elements constitute a two-pole which is either current-controlled, voltage-controlled or one-to-one. As the reader expects, quite specific assumptions will have to be

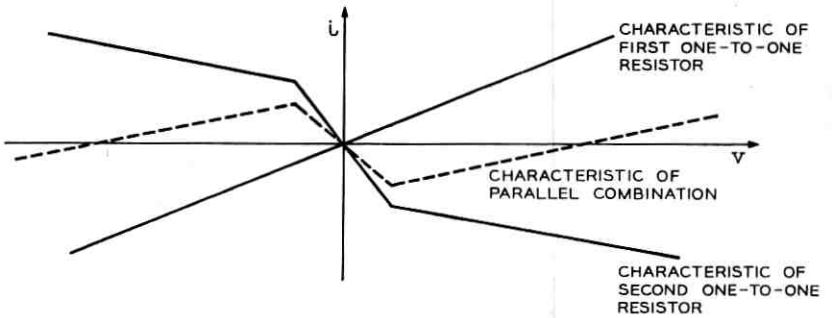


Fig. 5— Parallel connection of two one-to-one resistors.

made on the characteristics of the elements in order for the circuit to be a one-to-one two-pole.

The elements that we are going to consider are capacitors, resistors and inductors. Let us rank order these elements together with voltage sources and current sources in the following way: E,C,R,L,J. We shall say that a resistor is higher in rank than an inductor or a current source but lower in rank than a capacitor or a voltage source.

Until the end of this section, to simplify the discussion and without loss of generality, elements are assumed to be time-invariant.

Theorem: Consider the following circuits: the parallel RC, the parallel RL, the parallel LC and the parallel RLC circuit.

(A) If (a) the highest-ranked element is current-controlled (charge-controlled in the case of the capacitor),

(b) all other elements are voltage-controlled (flux-controlled in the case of the inductor), and

(*c*) *the characteristics of all elements satisfy a Lipschitz condition according to Table I,*

then the parallel circuit is current-controlled.

(*B*) *If, in addition* (*d*) *the highest ranked element is one-to-one, then each parallel circuit is one-to-one.*

Proof: We shall consider only the *RLC* circuit, since the proofs of the simpler cases follow in a similar way.

First let us prove that (*a*), (*b*), and (*c*) imply that the circuit is current-controlled. Let i_s be the source current. Then with the usual notation

$$i_s = \dot{q} + \mathcal{G}(v) + \Gamma(\varphi)$$

or, equivalently, using the fact that the capacitor is charge-controlled

$$\begin{cases} \dot{q} = -\mathcal{G}(\mathfrak{D}(q)) - \Gamma(\varphi) + i_s . & (1) \\ \dot{\varphi} = \mathfrak{D}(q) & (2) \end{cases}$$

By assumption (*c*) $\mathfrak{D}$, Γ and $\mathcal{G}$ satisfy Lipschitz conditions. Since the composite of two Lipschitzian functions is Lipschitzian, $\mathcal{G}(\mathfrak{D}(\cdot))$ is also Lipschitzian; therefore the system (1), (2) has a unique solution for each initial state and each current source. In this case the state is (q,φ). Thus the *RLC* is current-controlled.

Second we prove that (*a*), (*b*), (*c*), (*d*) imply that the *RLC* circuit is one-to-one. It is immediate that these assumptions imply that the *RLC* circuit is voltage-controlled. It remains to show that it is one-to-one.

Call $q_1(\cdot)$, $\varphi_1(\cdot)$, and $v_1(\cdot)$ the charge, flux and voltage resulting from the initial state (q_0, φ_0) at time t_0 and the input current i_s. The functions $q_1(\cdot)$ and $\varphi_1(\cdot)$ are the corresponding solutions of (1) and (2); $v_1(t) = \dot{\varphi}_1(t) = \mathfrak{D}(q_1(t))$. We have to show that, starting from the same initial state (q_0, φ_0) at time t_0, the input current resulting from the applied voltage v_1 is precisely i_s.

Let q_2, φ_2 and i_2 be the resulting charge, flux and input current. It is immediate that $v_1(t) = \dot{\varphi}_2(t) = \mathfrak{D}(q_2(t))$. Since $\varphi_2(t_0) = \varphi_0$, we have

TABLE I

Circuit	Highest-Ranked Element	Characteristics That Satisfy Lipschitz Conditions
RL	*R*	$\mathfrak{R}(i)$, $\Gamma(\varphi)$
RC	*C*	$\mathfrak{D}(q)$, $\mathcal{G}(v)$
LC	*C*	$\mathfrak{D}(q)$, $\Gamma(\varphi)$
RLC	*C*	$\mathfrak{D}(q)$, $\mathcal{G}(v)$, $\Gamma(\varphi)$

$\varphi_2 = \varphi_1$. Since the capacitor is one-to-one, q_2 is uniquely defined by the relation above in terms of v_1, hence $q_2 = q_1$. Finally, by Kirchhoff's current law

$$\begin{aligned} i_2 &= \dot{q}_2 + \mathcal{G}(v_2) + \Gamma(\varphi_2) \\ &= \dot{q}_1 + \mathcal{G}(\mathfrak{D}(q_1)) + \Gamma(\varphi_1). \end{aligned}$$

The last expression is precisely i_s by (1). Therefore $i_2 = i_s$. This concludes the proof that the parallel *RLC* circuit is a one-to-one two-pole. The dual case is covered by the following

Theorem: Consider the following circuits: the series RL, the series RC, the series LC and the series RLC circuit.

(A) If

(a) the lowest-ranked element is voltage-controlled (flux-controlled for inductor),

(b) all other elements are current-controlled (charge-controlled for capacitors), and

(c) the characteristics of all elements satisfy Lipschitz conditions according to Table II,

then the series circuit is voltage-controlled.

(B) If in addition (d) the lowest-ranked element is one-to-one, then each series circuit is one-to-one.

The proof is similar to that of the previous theorem and is therefore omitted.

III. REDUCTION OF THE NETWORK

Throughout the article we consider networks consisting of nonlinear time-varying resistors, capacitors, inductors (without mutual inductance) and independent sources. We shall label by $\mathfrak{N}$ the network under consideration. Usually, we consider each element and each source as constituting a branch of $\mathfrak{N}$. We denote by E,C,R,L,J the set of branches of $\mathfrak{N}$ which are voltage sources, capacitors, resistors, inductors and cur-

TABLE II

Circuit	Lowest-Ranked Element	Characteristics That Satisfy Lipschitz Conditions
RL	*L*	$\mathcal{R}(i)$, $\Gamma(\varphi)$
RC	*R*	$\mathfrak{D}(q)$, $\mathcal{G}(v)$
LC	*L*	$\mathfrak{D}(q)$, $\Gamma(\varphi)$
RLC	*L*	$\mathfrak{D}(q)$, $\mathcal{R}(i)$, $\Gamma(\varphi)$

rent sources, respectively. In our discussion, certain networks derived from $\mathfrak{N}$ will play an important role. In order to refer to them conveniently, let us define the following notations: Let "A" be a subset of the set of branches of $\mathfrak{N}$. Let us define[9]

$\mathfrak{N}_A$ to be the network derived from $\mathfrak{N}$ by removing all branches except the ones which are members of A,

$\mathfrak{N}_{(A)}$ to be the network derived from $\mathfrak{N}$ by replacing branches of set A by a short circuit, and

$\mathfrak{N}_{(A)\bullet}$ to be the network derived from $\mathfrak{N}$ by removing the branches of set A.

We shall use these notations as well as combinations of them. For example, $\mathfrak{N}_{(E)C}$ is the network derived from $\mathfrak{N}$ by first replacing branches of set E (the voltage sources) by short circuits and then removing all elements which do not belong to set C. Similarly, $\mathfrak{N}_{(E)(J)\bullet}$ is the network derived from $\mathfrak{N}$ by shorting all voltage sources and removing all current sources.

$S*\mathfrak{N}$ is defined to be the network derived from $\mathfrak{N}$ by separating it into the maximum number of separable subnetworks.

Throughout the article we assume that, first, for any cut set of current sources only, the source currents satisfy Kirchoff's current law, and, second, for any loop of voltage sources only, the source voltages satisfy Kirchoff's voltage law.

Without loss of generality we consider networks that are connected and nonseparable. This assumption does not exclude the possibility that $\mathfrak{N}_{(E)(J)\bullet}$ be both unconnected and/or separable. In the following we shall prove that without a loss of generality we can restrict the discussion to a network $\mathfrak{N}$ such that $\mathfrak{N}_{(E)(J)\bullet}$ is both connected and nonseparable. The proof consists of an algorithm which changes the configuration and reduces $\mathfrak{N}$ into a network $\mathfrak{N}'$ which has the following properties:

(*i*) $\mathfrak{N}'$ consists of connected subgraphs, $\mathfrak{N}_i'$, such that for each one of them, $\mathfrak{N}'_{i_{(E)(J)\bullet}}$ is connected and nonseparable.

(*ii*) For all branches of $\mathfrak{N}$ and $\mathfrak{N}'$ which are not sources, any set of branch currents and voltages is a solution of $\mathfrak{N}$ if and only if it is also a solution of $\mathfrak{N}'$ (when the latter is driven by the corresponding sources).

(*iii*) Current sources of $\mathfrak{N}'$ are linearly related to the current sources of $\mathfrak{N}$. The same is true for voltage sources.

The step-by-step reduction of the network $\mathfrak{N}$ to $\mathfrak{N}'$ is done as follows.

(1) From each loop which consists of voltage sources only, remove one voltage source.

(2) In each cut set which consists of current sources only, replace one of the current sources by a short circuit.

The resulting network is connected; it has a tree which includes *all* the voltage sources as tree branches and *all* the current sources as links.

(3) Each current source J whose fundamental loop includes more that one tree branch is removed from the network and is replaced by a set of current sources identical to J, each one placed in parallel with a tree branch of the fundamental loop.

(4) All current sources that are in parallel with voltage sources are removed.

(5) Any parallel connection of current sources is replaced by one equivalent current source.

(6) Consider each fundamental cut set defined by a voltage source. For each one of them insert in every link a voltage source equal to that which is in the tree branch and, finally, short circuit the tree branch voltage source.

(7) In each link, replace any series connection of voltage sources by one equivalent voltage source.

(8) Separate the network into the maximum number of connected, nonseparable subgraphs.

The resulting network is called $\mathfrak{N}'$. Property (*ii*) follows from the fact that all the steps of the above algorithm do not change the source contribution to any of the fundamental loop equations or the cut set equations. Property (*i*) follows from the fact that all current sources are links of $\mathfrak{N}'$ and all voltage sources are in a link. Property (*iii*) follows from steps (5) and (7). Finally, observe that $S*\mathfrak{N}_{(E)(J)\bullet}$ is identical with $\mathfrak{N}'_{(E)(J)\bullet}$.

It is well known that the state of the network is completely determined by all the voltages, fluxes, charges, and currents in the branches of the network. In the case of linear networks it is well known that certain proper subsets of these variables may be chosen as the state. For special classes of nonlinear *RLC* networks similar subsets will be indicated in the sequel.

We call a *solution* of an *RLC* network any set of voltages and currents of resistors, charges and voltages of capacitors, fluxes and currents of inductors which satisfy the Kirchhoff's laws and the branch characteristics. A network $\mathfrak{N}$ is said to be *determinate* if for any value of the initial state $\mathbf{s}_0$, given at any initial time t_0, and for any value of the sources $\mathbf{E}(\cdot)$, $\mathbf{J}(\cdot)$, there exists one and only one solution for $t \geqq t_0$ on some nonvanishing interval $[t_0, t_\alpha)$.

In the following section we shall describe a broad class of nonlinear *RLC* networks which are determinate.

IV. ONE-ELEMENT-KIND NETWORK

The purpose of this section is to establish a set of sufficient conditions under which a nonlinear (possibly time-varying) resistor network driven by a set of independent current sources and voltage sources has, for all possible inputs, one and only one set of branch voltages and branch currents that satisfy Kirchhoff's laws. Conditions under which the solution satisfies a Lipschitz condition with respect to the sources are also given.

The analysis of nonlinear resistor networks is almost identical with that of nonlinear capacitor networks or nonlinear inductor networks. Since the nonlinear resistors are the most flexible elements, we shall develop our analysis in terms of resistor networks.

Let us start by making three preliminary remarks:

(*i*) Given a resistor network together with an arbitrary distribution of current sources, it is always legitimate to assume that there are no cut sets of current sources only. (Dually, that there are no loops of voltage sources only.)

(*ii*) Any voltage source in series with a resistor may always be absorbed into a suitably redefined branch characteristic. Refer to Fig. 6, where v_1 and v_2 are the node voltages of nodes 1 and 2 referred to the same datum. Let the current through the resistor be given by its characteristic $g(v,t)$; since $g(v,t) = g(v_1 - v_2 - e,t)$ and since $e(\cdot)$ is a known function of time, we may introduce a new branch characteristic $g_{12}(\cdot,\cdot)$ specified at each instant of time by

$$g_{12}(v_1 - v_2,t) \triangleq g(v_1 - v_2 - e(t),t).$$

In other words, the voltage source e has been absorbed into the time dependence of g_{12}. A similar reasoning applies to a current-controlled resistor in series with a voltage source.

The dual case can be taken care of in the same manner: in this case, a current source which is in parallel with either a voltage-controlled or a current-controlled resistor can be absorbed into the branch.

Thus, without loss of generality, a network of nonlinear resistors and sources can be thought of as a network of nonlinear time-varying resist-

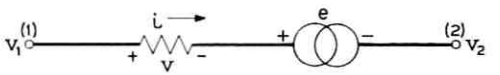


Fig. 6 — Voltage source in series with resistor.

ors with the understanding that the sources have been absorbed in the branch characteristics.

(*iii*) Thus when, as in Theorems I and II below, we consider a network of nonlinear time-varying resistors, we include the case of a network made up of time-varying resistors and of independent sources. There is no loss of generality in considering only connected networks, since it amounts to considering successively each separate part of an unconnected network.

We turn now to the statement of the main theorems.

Theorem I (Existence and Uniqueness): Consider a connected nonseparable network $\mathfrak{N}$ *of nonlinear (possibly time-varying) resistors. In case the resistor joining node* α *to node* β *is voltage-controlled, its characteristic is defined by the function* $g_{\alpha\beta}(\cdot,\cdot)$ *such that* $g_{\alpha\beta}(v_\alpha(t) - v_\beta(t),t)$ *is the current flowing through it at time* t *from node* α *to node* β*; here* v_α *and* v_β *are the node-to-datum voltages of nodes* α *and* β*. Similarly, if this resistor is current-controlled, its characteristic is defined by the function* $r_{\alpha\beta}(\cdot,\cdot)$ *such that* $r_{\alpha\beta}(i_{\alpha\beta}(t),t)$ *is the voltage difference between node* α *and node* β *at time* t*; here* $i_{\alpha\beta}$ *is the current through the resistor measured positively if it flows from* α *to* β*.*

If

(*a*) *there exists a tree* $\mathfrak{I}$ *such that all its tree branches are current-controlled and all its links are voltage-controlled,*

(*b*) *for all* α,β*, all* t *and all* x *in* $(-\infty,\infty)$

$g_{\alpha\beta}(x,t) = -g_{\beta\alpha}(-x,t)$ *if* (α,β) *is a link*

$r_{\alpha\beta}(x,t) = -r_{\alpha\beta}(-x,t)$ *if* (α,β) *is a tree branch*

(*c*) *for all links and all* t, $g_{\alpha\beta}(\cdot,t)$ *is a monotonically (not necessarily strictly) increasing continuous function defined on* $(-\infty,\infty)$*, and for all tree branches and all* t, $r_{\alpha\beta}(\cdot,t)$ *is a monotonically (not necessarily strictly) increasing continuous function defined on* $(-\infty,\infty)$*.*

Then,

for all current-sources i^s *connected between any pair of nodes and for all voltage sources* e^s *connected in series with network branches there exists one and only one set of branch voltages and branch currents that satisfy the Krichhoff laws and the branch characteristics.*

The conclusion of Theorem I can also be stated as follows: any network $\mathfrak{N}'$, formed from $\mathfrak{N}$ by inserting any set of voltage sources in series with any branch and any set of current sources between any node pair, is *determinate*.

Assumption (*b*) is a consequence of the physical meaning of the func-

tions g and r and of the sign conventions: from a physical point of view they do not restrict the nonlinear resistors in any way. The two corollaries that follow are special cases of Theorem I. Corollary I is an extension of Theorems 2 and 3 of Duffin,[3] and is implied by his 1948 paper.[4] Such an extension has been pointed out by I.W. Sandberg.[15]

Corollary 1: Consider a connected network of nonlinear voltage-controlled (possibly time-varying) resistors.

If

(a) for all branches and all t, $g_{\alpha\beta}(\cdot,t)$ is a monotonically (not necessarily strictly) increasing, continuous function defined on $(-\infty,\infty)$, and

(b) there exists a tree $\mathfrak{I}$ such that all its branches have $g_{\alpha\beta}$'s which are, for all t, monotonically increasing one-to-one mappings of $(-\infty,\infty)$ onto $(-\infty,\infty)$,

then the conclusion of Theorem I holds.

Proof: The conclusion follows directly from Theorem I since the tree branches have $g_{\alpha\beta}$'s that are, for all t, monotonically increasing one-to-one mappings of $(-\infty,\infty)$ onto $(-\infty,\infty)$; hence the tree branches are also current-controlled resistors satisfying assumption (c) of Theorem I.

Corollary 2: Consider a connected network of nonlinear, current-controlled (possibly time-varying) resistors.

If

(a) for all branches and all t, $r_{\alpha\beta}(\cdot,t)$ is a monotonically (not necessarily strictly) increasing, continuous function defined on $(-\infty,\infty)$, and

(b) there exists a tree $\mathfrak{I}$ such that its links have $r_{\beta\alpha}$'s which are, for all t, monotonically increasing one-to-one mappings of $(-\infty,\infty)$ onto $(-\infty,\infty)$, *then* the conclusion of Theorem I holds.

We consider now the extension of the Thévenin and Norton equivalent circuits to nonlinear resistive networks. If we pick an arbitrary node pair of such a network $\mathfrak{N}$, we may regard these nodes as the terminals of a two-terminal network: we shall call the characteristic of this two-terminal network *the input characteristic of* $\mathfrak{N}$ *at these two nodes.* Dually, if we pick a branch and insert two terminals in series with it, we obtain a two-terminal network: we shall call the characteristic of this two-terminal network *the branch-input characteristic of* $\mathfrak{N}$.

Theorem II (Thévenin and Norton equivalent circuits): Consider a network $\mathfrak{N}$ *satisfying the requirements of Theorem I together with the same kind of source distribution.*

Then

(*a*) *the input characteristic of* $\mathfrak{N}$ *at any node pair is that of a current-controlled resistor whose characteristic is a continuous, monotonically increasing function defined on* $(-\infty, \infty)$. *This characteristic may be represented by the Thévenin equivalent circuit of Fig. 7 (a): a series combination of a voltage source and a monotonically increasing current-controlled resistor whose characteristic passes through the origin.*

(*b*) *The branch-input characteristic of* $\mathfrak{N}$ *at any branch is that of a voltage-controlled resistor whose characteristic is a continuous, monotonically increasing function defined on* $(-\infty, \infty)$. *This characteristic may be represented by the Norton equivalent circuit of Fig. 7 (b): a parallel combination of a current source and a monotonically increasing voltage-controlled resistor whose characteristic passes through the origin.*

Let us consider some special cases of Theorem II.

Corollary 3: Consider a connected network of nonlinear (possibly time-varying) resistors satisfying assumptions (*a*), (*b*) and (*c*) of Theorem I.

(a) If, in addition, the characteristics of the tree branches of $\mathfrak{I}$ are strictly increasing, then the input characteristic at any node pair is that of a strictly increasing current-controlled resistor. If the characteristics of all tree branches of $\mathfrak{I}$ are continuous, monotonically increasing, one-to-one mappings of $(-\infty, \infty)$ onto $(-\infty, \infty)$ then so is the input characteristic at any node pair.

(b) If the characteristics of the links of the tree $\mathfrak{I}$ are strictly increasing, then the branch-input characteristic is that of a strictly increasing voltage-controlled resistor. If the characteristics of all links of $\mathfrak{I}$ are continuous, monotonically increasing, one-to-one mappings of $(-\infty, \infty)$ onto $(-\infty, \infty)$, then so is any branch-input characteristic.

Proof of Theorems I and II: The proof of these two theorems is divided

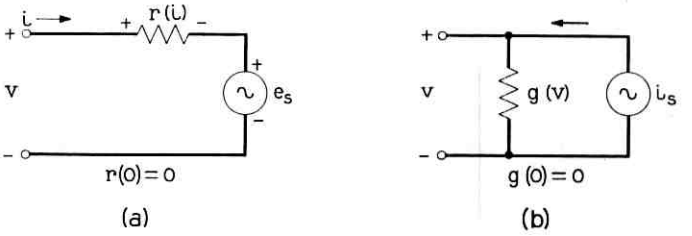


Fig. 7 — (a) Thévenin equivalent circuit: series combination of voltage source and a monotonically increasing current-controlled resistor whose characteristic passes through the origin. (b) Norton equivalent circuit: parallel combination of current source and monotonically increasing voltage-controlled resistor whose characteristic passes through the origin.

into two parts: in part one, we show that if Theorem I holds for a k-node network then Theorem II is true for a k-node network. In part two, we use this implication to prove Theorem I by induction.

The statement of the theorem allows time-varying resistors (and hence includes independent sources); however, in order to have as simple a notation as possible, we write down the proof as if all resistors were time-invariant.

Part One: We show that, for any integer $k \geqq 2$, if Theorem I holds for a k-node network then the input characteristic at any node pair is that of the Thévenin equivalent circuit specified in Theorem II (a). Let us connect the node pair under consideration to a current source i_s (see Fig. 8); this current source is viewed as an additional link, since it is a voltage-controlled resistor. By assumption, to each i_s there is one and only one set of branch currents and voltages that satisfy Kirchhoff's

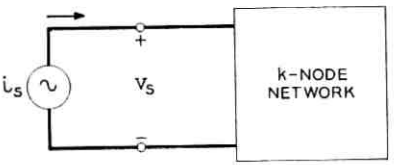


Fig. 8 — Node pair connected to current source.

laws and the branch characteristics. Consider two distinct values of i_s, namely, i_s and i_s'. Let the corresponding branch variables be $\mathbf{v},\mathbf{i}$ and $\mathbf{v}',\mathbf{i}'$. For each current-controlled branch define a number $\bar{r}$ (which depends on $\mathbf{i}$ and $\mathbf{i}'$) by the relation

$$v - v' \triangleq \Delta v = r(i) - r(i') \triangleq \bar{r} \cdot (i - i') = \bar{r} \cdot \Delta i.$$

Since all the current-controlled branches are monotonically increasing, $\bar{r} \geqq 0$. (If $\Delta i = 0$, $\bar{r}$ may be taken to be any nonnegative number.) Similarly, we define a $\bar{g}$ for each voltage-controlled resistor; again $\bar{g} \geqq 0$. The set of Δv's and Δi's together with Δv_s and Δi_s may be considered as a set of branch voltages and branch currents together with the source voltage and source current of a *linear* resistive network which is obtained by replacing each current-controlled resistor by a linear resistor of resistance $\bar{r}$, each voltage-controlled resistor by a linear resistor of conductance $\bar{g}$ and the current source by a current source Δi_s. Since the Δv's and Δi's satisfy Kirchhoff's laws, Tellegen's theorem[12] holds,

$$\Delta v_s \cdot \Delta i_s = \sum \Delta v \Delta i$$

where the sum is over all resistive branches.

Since all branches have monotonically increasing characteristics, this is a sum of nonnegative terms and $\Delta v_s \Delta i_s \geqq 0$. In other words, $\Delta i_s > 0$ implies that $\Delta v_s \geqq 0$: that is, the Thévenin equivalent circuit has a current-controlled monotonically increasing characteristic. The continuity of the characteristic follows from the following considerations: irrespective of the values of the $\bar{r}$'s and $\bar{g}$'s, the fact that assumption (c) of Theorem I requires them to be nonnegative implies that the current transfer ratio from the current source to any branch has a magnitude no larger than unity;[11] hence $\Delta i_s \to 0$ implies $\Delta i \to 0$ for all branches. Since the tree branches have continuous characteristics, it follows that, for them, $\Delta v \to 0$ and, by Kirchhoff's voltage law, the same holds for the links. Hence $\Delta i_s \to 0$ implies $\Delta v_s \to 0$, i.e., the current-controlled characteristic of the Thévenin equivalent circuit is continuous. The proof of part (b) of Theorem II follows exactly the dual of the above argument.

Part Two: Let us prove Theorem I for a two-node network (see Fig. 9). Let us plot on the (v,i) plane of Fig. 10 the characteristics of the current-controlled tree branch and that of the voltage-controlled link, taking into account the sign conventions defined on Fig. 9. By assumption, the functions g and r are both continuous and have $(-\infty, \infty)$ as domains; therefore their representative curves intersect at least at one point (v,i). We assert that it is the only one: indeed, suppose there were a second one, (v',i'); then the monotonicity of r and g imply, respectively

$$(v' - v)\,(i' - i) \geqq 0 \quad \text{and} \quad (v' - v)(i' - i) \leqq 0.$$

Hence

$$(v' - v)\,(i' - i) = 0.$$

Suppose $v' = v$; then since g is a function

$$i = g(-v) = g(-v') = i'.$$

Similarly, if $i' = i$, the fact that r is a function implies $v = v'$. Hence for all possible sources, there is one and only one set of branch voltages and currents that satisfies Kirchhoff's laws and the branch characteristics.

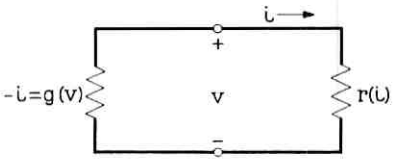


Fig. 9 — Two-node network.

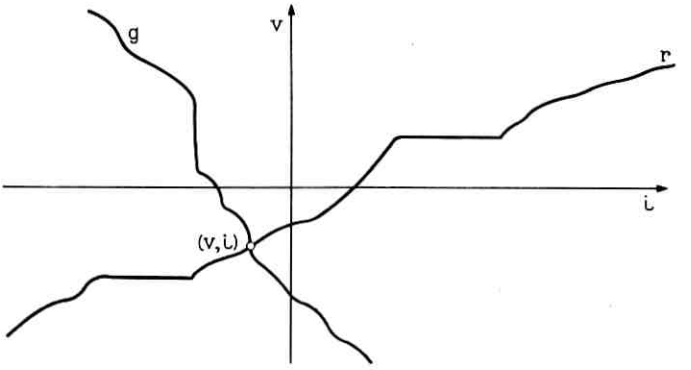


Fig. 10 — Characteristics of current-controlled tree branch and of voltage-controlled link as function of the tree branch current and voltage.

Thus Theorem I is established for a two-node network. The next step in the proof of Theorem I is to show that if it is true for an n-node network it is true for an $(n + 1)$-node network. Consider the n-node network shown in Fig. 11. We shall build out this network into an $(n + 1)$-node network.

Let us first connect the tree branch between node n and node $(n + 1)$, i.e., a current-controlled resistor. (There is no loss of generality in assuming that the numbering of the nodes is such that the branch $(n,n + 1)$ is a tree branch.) It is obvious that, for this network, the existence and uniqueness of the solution holds for all sources. Consequently, from part one of the proof, the input characteristic at any two nodes of this particular $(n + 1)$-node network has the equivalent circuits specified by Theorem II. The next step is to add a link, say between node k and node $(n + 1)$. Since the input characteristic at the node pair $(k,n + 1)$ is as specified in Theorem II (a), the voltage and current in the link are uniquely determined by the reasoning given for the case $n = 2$, and consequently the distribution of voltages and currents in all branches of the network is uniquely determined.

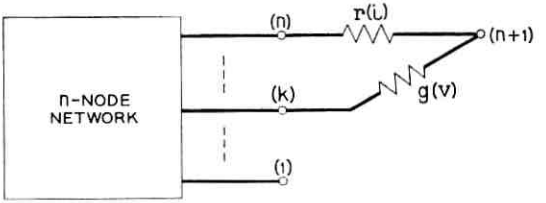


Fig. 11 — N-node network built out to $(n + 1)$-node network.

The process of constructing the $(n + 1)$-node network from the n-node network can be carried out step by step, adding a link at a time. Thus at the end of the process there is one and only one set of branch voltages and currents in the $(n + 1)$-node network that satisfies Kirchhoff's laws and the branch characteristics. Q.E.D.

For the purpose of solving the network differential equations of a general nonlinear *RLC* network it is important to know, for the resistive network case, under what conditions the function which maps the sources, $(\mathbf{E},\mathbf{J})$, into the branch voltages and currents, $(\mathbf{v},\mathbf{i})$, satisfies a Lipschitz condition. It is immediately clear that additional assumptions are required: consider Fig. 12, which shows the characteristic of a current-controlled resistor which fulfills the conditions of Theorem I. In the neighborhood of the operating point A, this resistor may appear to small signals either as an open circuit or a short circuit. Note that the same statement would apply if the resistor were voltage-controlled. It is obvious that under such conditions, the mapping $(\mathbf{E},\mathbf{J}) \rightarrow (\mathbf{v},\mathbf{i})$ will not satisfy a Lipschitz condition. As shown in the following theorem, only weak additional assumptions are required.

Theorem III: Consider a connected network of nonlinear (possibly time-varying) resistors which satisfies conditions (a), (b) and (c) of Theorem I. If, in addition, the following Lipschitz conditions are satisfied: there is a real-valued function $h(R,t)$, defined and positive for $R > 0$ and all t, such that

$$| g_{\alpha\beta}(x,t) - g_{\alpha\beta}(x',t) | \leqq h(R,t) | x - x' |$$

for all links of $\mathfrak{I}$, for all x, x' in $(-R,R)$ and all t and

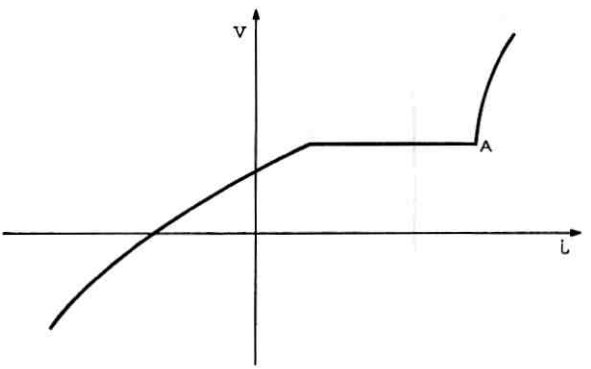


Fig. 12 — Characteristic of current-controlled resistor fulfilling conditions of Theorem I; note unbounded slope at point A.

$$| r_{\alpha\beta}(x,t) - r_{\alpha\beta}(x',t) | \leqq h(R,t) | x - x' |$$

for all branches of $\mathfrak{Z}$, *for all* x,x' *in* $(-R,R)$ *and all* t, *then the mapping which maps* $(\mathbf{E},\mathbf{J})$ *into* $(\mathbf{v},\mathbf{i})$ *satisfies a Lipschitz condition.**

Proof: Consider the effect of a change in the voltage sources $\mathbf{E}$ on the branch voltages $\mathbf{v}$ and branch currents $\mathbf{i}$. $\mathbf{E}$ is the vector whose ith component is the output voltage e_i of the source located in the ith branch. In the present $(n + 1)$ node network there are at most $n(n + 1)/2$ branches, hence $\mathbf{E}$ has at most that many components. Suppose that the change from $\mathbf{E}$ to $\mathbf{E} + \Delta\mathbf{E}$ is obtained by changing e_i to $e_i + \Delta e_i$ successively with $i = 1, 2, \cdots$. Call α,β the terminals of the first branch and call N_1 the remainder of the network (see Fig. 13). Since the input characteristic of N_1 is monotonically increasing, and since an increase of the voltage across the nonlinear resistance R increases the current through it or keeps it constant, the change in the input voltage of N_1,

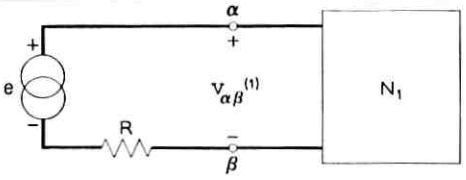


Fig. 13 — Nonlinear resistor R and voltage source in one branch of $(n + 1)$-node network; N_1 represents remainder of network.

$\Delta v_{\alpha\beta}{}^{(1)}$ due to the change of e_1 to $e_1 + \Delta e_1$ is such that $| \Delta e_1 | \geqq | \Delta v_{\alpha\beta}{}^{(1)} |$. (The superscript 1 indicates that only the source voltage in the first branch has been changed.) Call $\Delta v_k{}^{(1)}$ the corresponding change in the kth branch voltage. We assert that

$$| \Delta v_k{}^{(1)} | \leqq | \Delta v_{\alpha\beta}{}^{(1)} | \leqq | \Delta e_1 |.$$

For the particular change in the sources under consideration, we may define, as in the proof of Theorem I, for each tree branch a suitable $\bar{r}$ and for each link a suitable $\bar{g}$. Observe now that Δe_1 and the $\Delta v_k{}^{(1)}$ may be interpreted as being the source voltage and the resulting branch voltage of a linear network which has the same configuration as the given nonlinear network but in which each nonlinear resistor is replaced by $\bar{r}$ or $\bar{g}$ as required. By assumption (c) of Theorem I, all the $\bar{r}$'s and $\bar{g}$'s are nonnegative, hence all the voltage transfer ratios $| \Delta v_k{}^{(1)}/\Delta e^1 |$ of

* Incidentally, if the network $\mathfrak{N}$ was derived from another network $\mathfrak{N}_A$, by applying to $\mathfrak{N}_A$ the algorithm of Section III, then the mapping $(\mathbf{E},\mathbf{J}) \rightarrow (\mathbf{v},\mathbf{i})$ is one-to-one.

the linear network cannot[11] exceed 1 and the inequality asserted above follows. Thus, for all i's and k's,

$$| \Delta v_k^{(i)} | \leqq | \Delta e_i |.$$

Let Δv_k be the change in the voltage across the kth branch when $\mathbf{E}$ becomes $\mathbf{E} + \Delta\mathbf{E}$. Summing over i, using the triangle inequality, and defining the norm of a vector as the sum of the magnitude of its components, we get

$$| \Delta v_k | \leqq \| \Delta\mathbf{E} \|. \tag{3}$$

Since there are at most $n(n + 1)/2$ branches, we get finally

$$\| \Delta\mathbf{v} \| \leqq [n(n + 1)/2] \| \Delta\mathbf{E} \| \tag{4}$$

where $\Delta\mathbf{v}$ is the change in the branch voltages corresponding to the change of the voltage sources from $\mathbf{E}$ to $\mathbf{E} + \Delta\mathbf{E}$. We next bound the change in the branch currents. Applying (3) to a link and using the Lipschitz condition we find that

$$| \Delta i_k | \leqq h(R,t) \| \Delta\mathbf{E} \| \qquad \text{(for all links)}$$

and since there are at most $n(n - 1)/2$ links and the change in a tree branch current is equal to the change in the sum of currents of the links which belong to its fundamental cut set,

$$| \Delta i_k | \leqq h(R,t)[n(n - 1)/2] \| \Delta\mathbf{E} \| \qquad \text{(for all branches)}.$$

Thus

$$\| \Delta\mathbf{i} \| \leqq h(R,t)[n^2(n^2 - 1)/4] \| \Delta\mathbf{E} \|. \tag{5}$$

The effect of a change in the current sources from $\mathbf{J}$ to $\mathbf{J} + \Delta\mathbf{J}$ is obtained in a dual manner. Since the current transfer ratio may not exceed unity[11] we get

$$| \Delta i_k | \leqq \| \Delta\mathbf{J} \|$$

and

$$\| \Delta\mathbf{i} \| \leqq [n(n + 1)/2] \| \Delta\mathbf{J} \|. \tag{6}$$

This implies

$$| \Delta v_k | \leqq h(R,t) \| \Delta\mathbf{J} \| \qquad \text{(for all tree branches)}$$

and, by Kirchhoff's voltage law,

$$| \Delta v_k | \leqq h(R,t)n \| \Delta\mathbf{J} \| \qquad \text{(for all branches)}.$$

Finally

$$\| \Delta \mathbf{v} \| \leqq h(R,t)[n^2(n+1)/2] \| \Delta \mathbf{J} \|. \tag{7}$$

Using the usual product topology[17] for both the product spaces of voltage sources and current sources on the one hand and branch voltages and branch currents on the other, and invoking (4) to (7), we conclude that the mapping $(\mathbf{E},\mathbf{J}) \rightarrow (\mathbf{v},\mathbf{i})$ is Lipschitz.

V. NONLINEAR RLC NETWORKS

The previous section required all elements of the network to be of the same kind and to have a monotonically increasing characteristic. In this section both requirements are removed. In addition to independent sources, the network consists of nonlinear (possibly time-varying) resistors, capacitors and inductors and some of the elements are allowed to have characteristics with negative slope.

As a first step let us make one remark. Theorems I, II, and III would still hold if all resistors were montonically decreasing instead of monotonically increasing. In the more complicated situation considered here the same possible choice exists. For example, separable subnetworks of $\mathfrak{N}_{(E)C}$ which contain more than one capacitor could just as well contain monotonically decreasing capacitors. For simplicity, we shall assume that all monotonic elements are increasing.

In order to state the following theorem we need two definitions. A network (or subnetwork) is called a *self-loop* if it consists of a single branch whose end-points are identified: it consists of one branch and one node. A network (or subnetwork) is called an *open branch* if it consists of a single branch whose end-points are not identified: it consists of one branch and two nodes.

Theorem IV: Let $\mathfrak{N}$ *be a network of independent sources and nonlinear (possibly time-varying) resistors, capacitors and inductors (without mutual inductance) such that: capacitors of* $\mathfrak{N}$ *are either charge-controlled or monotonically increasing voltage-controlled; resistors are either voltage-controlled or current-controlled; inductors are either flux-controlled or monotonically increasing current-controlled. It is further assumed that* $\mathfrak{N}$ *and* $\mathfrak{N}_{(E)(J)}$ *are nonseparable and connected. The network* $\mathfrak{N}$ *is determinate if:*

(1) *The capacitor network* $S*\mathfrak{N}_{(E)C}$ *satisfies the following requirements:*

(*a*) *Open branches of* $S*\mathfrak{N}_{(E)C}$ *are charge-controlled and contain all charge-controlled capacitors which are not monotonically increasing.*

(*b*) *Each subnetwork of* $S*\mathfrak{N}_{(E)C}$ *which contains more than one element*

has a tree with monotonically increasing charge-controlled tree branches and monotonically increasing voltage-controlled links.

(c) *Self-loops of* $S*\mathfrak{N}_{(E)C}$ *are voltage-controlled.*

(2) *The resistive network* $S*\mathfrak{N}_{(EC)R}$ *satisfies the following requirements:*

(a) *Open branches are current-controlled and contain all current-controlled resistors which are not monotonically increasing.*

(b) *Each subnetwork which contains more than one element has a tree with monotonically increasing current-controlled tree branches and monotonically increasing voltage-controlled links.*

(c) *Self-loops are voltage-controlled and contain all voltage-controlled resistors which are not monotonically increasing.*

(3) *The inductive network* $S*\mathfrak{N}_{(ECR)L}$ *satisfies the following requirements:*

(a) *Open branches are current-controlled.*

(b) *Each subnetwork which contains more than one element has a tree with monotonically increasing current-controlled tree branches and monotonically increasing flux-controlled branches.*

(c) *Self-loops are flux-controlled and contain all flux-controlled inductors that are not monotonically increasing.*

(4) *In any finite interval, and for all time, the characteristics of the network resistors, capacitors and inductors satisfy a Lipschitz condition with respect to the following variables:*

tree branches: capacitors, with respect to q
resistors and inductors, with respect to i
links: capacitors and resistors, with respect to v
inductors, with respect to φ.

Remarks: Note that nonmonotonic voltage-controlled capacitors and current-controlled inductors were excluded from the discussion. Such capacitors and inductors may be included in the discussion provided they fall into the following trivial cases: each nonmonotonic voltage-controlled capacitor is in parallel with a voltage source and each nonmonotonic current-controlled inductor is in series with a current source. In such cases, $\mathfrak{N}_{(F)(J)*}$ is separable unless $\mathfrak{N}$ contains one element only.

The above conditions insure the existence and uniqueness[18] of the solution on some nonvanishing interval $[t_0, t_\alpha)$, where $t_\alpha > t_0$. The length of this interval cannot be specified without further assumptions on the Lipschitz constants $h(R,t)$. This is the well known problem of finite escape time. In particular, if for all branch characteristics the same Lipschitz constant can be used and holds over the whole domain of the characteristic, then the solution exists and is unique on $[t_0, \infty)$ for all regulated **E**'s and **J**'s.

Proof of Theorem IV: Let us denote the voltages and charges of the capacitive branches by $(\mathbf{e}_c, \mathbf{q}_c)$. Similarly, denote the voltages and currents of the resistive branches by $(\mathbf{e}_r, \mathbf{i}_r)$ and fluxes and currents of inductive branches by $(\varphi_L, \mathbf{i}_L)$. Voltage sources and current sources will denoted as usual by $\mathbf{E}$ and $\mathbf{J}$.

We assert that conditions (2) and (4) of Theorem IV imply, first, that the currents and the voltages of the resistive branches at time t, $(\mathbf{e}_r(t),\mathbf{i}_r(t))$ are uniquely determined by the values, *at the same time* t, of the capacitor voltages, the voltage sources, the inductor currents and the current sources, $(\mathbf{e}_c(t),\mathbf{E}(t),\mathbf{i}_L(t),\mathbf{J}(t))$; and, second, that the mappings

$$\mathbf{e}_r(t) = \mathbf{f}_{e_r}(\mathbf{e}_c(t),\mathbf{E}(t),\mathbf{i}_L(t),\mathbf{J}(t),t) \tag{8}$$

$$\mathbf{i}_r(t) = \mathbf{f}_{i_r}(\mathbf{e}_c(t),\mathbf{E}(t),\mathbf{i}_L(t),\mathbf{J}(t),t) \tag{9}$$

satisfy Lipschitz conditions.

Given any set of capacitor voltages $\mathbf{e}_c$ and inductor currents $\mathbf{i}_L$ such that Kirchoff's voltage law is satisfied in each loop formed by capacitors and voltage sources, and such that Kirchoff's current law is satisfied in each cut set formed by inductors and current sources, let us replace each capacitor with a voltage source whose voltage is equal to the voltage of the replaced capacitor and replace each inductor with a current source whose current is equal to the current of the replaced inductor. The network consists now of resistors, current sources and voltage sources only.

Let us use the algorithm of Section III to change the configuration of the sources and to separate the network into its separable parts. Let us denote the resulting network by $\mathfrak{N}^R$ and its sources by $(\mathbf{E}^R,\mathbf{J}^R)$.

The network $\mathfrak{N}^R$ has three sets of subnetworks: (a) connected nonseparable subnetworks which contain sources and two or more resistive branches, (b) subnetworks containing one resistive branch in parallel with a voltage source, and (c) subnetworks containing one resistive branch in parallel with a current source.

Consider the first set of subnetworks. Denote the branch voltages and branch currents of these subnetworks by $(\mathbf{e}_r, \mathbf{i}_r)_1$, and their sources by $(\mathbf{E}_1{}^R,\mathbf{J}_1{}^R)$. From conditions (2) and (4) of Theorem IV it follows that each subnetwork contains a tree whose resistive branches are monotonically increasing current-controlled resistors and whose links are monotonically increasing voltage-controlled resistors, and that all elements satisfy Lipschitz conditions. Therefore, from Theorems I and III of Section IV it follows that $(\mathbf{e}_r, \mathbf{i}_r)_1$ are uniquely denfied by $(\mathbf{E}_1{}^R,\mathbf{J}_1{}^R)$ and that the mapping $(\mathbf{e}_r(t),\mathbf{i}_r(t)) = \mathbf{f}_r(\mathbf{E}_1{}^R(t),\mathbf{J}_1{}^R(t),t)$ satisfies Lipschitz conditions.

The second set of networks corresponds to self-loops of $S*\mathfrak{N}_{(EC)R}$. In each subnetwork, the resistor is voltage-controlled, and hence the current through it is uniquely determined in terms of the voltage source. Since the characteristics satisfy a Lipschitz condition, the mapping from space $\mathbf{E}_2{}^R$ (the voltage sources) to $(\mathbf{e}_r, \mathbf{i}_r)_2$ (the branch voltages and currents in the subnetworks of the second set) satisfies Lipschitz conditions.

Subnetworks of the third set correspond to open branches of $S*\mathfrak{N}_{(EC)R}$. As each subnetwork contains only one resistor and a current source, the requirement that the branch be current-controlled is enough to insure uniqueness of $(\mathbf{e}_r, \mathbf{i}_r)_3$, the branch voltages and currents of this set, in terms of the corresponding sources $\mathbf{J}_3{}^R$. From condition (4) it follows that the mapping from $\mathbf{J}_3{}^R$ to $(\mathbf{e}_r, \mathbf{i}_r)_3$ satisfies the Lipschitz condition.

The voltages of sources $\mathbf{E}^R$ are linear combinations of the voltages $\mathbf{E}$ and $\mathbf{e}_c$, and the currents $\mathbf{J}^R$ are linear combinations of $\mathbf{J}$ and $\mathbf{i}_L$ (see Section IV). From this linearity property and from the properties of the above relations between the voltages and currents of the resistive branches and $(\mathbf{E}^R, \mathbf{J}^R)$ it follows that $(\mathbf{e}_r(t), \mathbf{i}_r(t))$ are uniquely defined by $(\mathbf{e}_c(t), \mathbf{E}(t), \mathbf{i}_L(t), \mathbf{J}(t))$ and that the mappings in (8) and (9) satisfy Lipschitz conditions.

Let us now consider the capacitors of the network $\mathfrak{N}$. Given any set of resistor currents $\mathbf{i}_r$ and inductor currents $\mathbf{i}_L$ such that Kirchoff's current law is satisfied in each cut set formed by resistors, inductors and current sources, let us replace the inductive and resistive branches of $\mathfrak{N}$ by current sources with currents equal to the corresponding currents $\mathbf{i}_L$, $\mathbf{i}_r$. The network consists now of sources and capacitors only. Let us use the algorithm of Section III to change the configuration of the sources and separate the network into its separable parts. The resulting network is denoted by $\mathfrak{N}^C$ and its sources by $\mathbf{E}^C, \mathbf{J}^C$. We are going to establish an analogy between $\mathfrak{N}^C$ and its sources by $\mathbf{E}^C, \mathbf{J}^C$. We are going to establish analogy between $\mathfrak{N}^C$ and $\mathfrak{N}^R$ and use the result just proved for $\mathfrak{N}^R$ to deduce a similar result for $\mathfrak{N}^C$.

$\mathfrak{N}^C$ consists of the three sets of subnetworks which were described in connections with $\mathfrak{N}^R$. Consider the second set of subnetworks of $\mathfrak{N}^C$, which consists of single capacitors in parallel with a voltage source. Except for the trivial case where $\mathfrak{N}$ consists only of a single capacitor in parallel with a voltage source, this set is empty, for otherwise $\mathfrak{N}_{(E)(J)*}$ would be separable. Condition (1) implies that each subnetwork of the first set has a tree, say τ_C, whose tree branches are monotonically increasing and charge-controlled, and whose links are monotonically increasing and voltage-controlled. For each subnetwork of this set, with each funda-

mental cut set of τ_C defined by a capacitive tree branch, we assign a variable q_i equal to the sum of the charges on all the capacitors of that cut set. For each subnetwork of the third set, we assign a q_i equal to the charge on the capacitor. $\mathbf{q}$ will denote the vector whose components are the q_i's.

The analogy between $\mathfrak{N}^C$ and $\mathfrak{N}^R$ is established in four steps:

(*i*) For $\mathfrak{N}^R$, the Kirchhoff current law applied to the ith cut set associated with a resistive tree branch reads

$$J_i^R = \sum_k i_{ki}$$

where i_{ki} is the current in the kth branch of the ith cut set. The i_{ki} are components of $\mathbf{i}_r$. For $\mathfrak{N}^C$, we have by definition of q_i,

$$q_i = \sum_k q_{ki}$$

where q_{ki} is the charge in the kth branch of the ith cut set. The q_{ki} are components of $\mathbf{q}$.

(*ii*) For both $\mathfrak{N}^R$ and $\mathfrak{N}^C$, the Kirchhoff voltage law holds.

(*iii*) Condition (1) imposes requirements on the topology and element characteristics of $\mathfrak{N}^C$ which are entirely similar to those imposed on $\mathfrak{N}^R$ by condition (2).

(*iv*) Finally, the elements of $\mathfrak{N}^R$ and $\mathfrak{N}^C$ satisfy analogous Lipschitz conditions by condition (4).

Therefore, the variables $(\mathbf{e}_c, \mathbf{q}_c)$ and $(\mathbf{E}^C, \mathbf{q})$ of $\mathfrak{N}^C$ are analogous to the variables $(\mathbf{e}_r, \mathbf{i}_r)$ and $(\mathbf{E}^R, \mathbf{J}^R)$ of $\mathfrak{N}^R$.

Remembering that $\mathbf{E}^C$ is linearly related to $\mathbf{E}$, we conclude that the voltages and charges of the capacitors at time t are uniquely determined by the values, *at the same time t*, of the voltage sources $\mathbf{E}(t)$ and $\mathbf{q}(t)$, and that the mapping

$$\mathbf{e}_c(t) = \mathbf{f}_{e_c}(\mathbf{E}(t), \mathbf{q}(t), t) \tag{10}$$

$$\mathbf{q}_c(t) = \mathbf{f}_{q_c}(\mathbf{E}(t), \mathbf{q}(t), t) \tag{11}$$

satisfies Lipschitz conditions.

Since q_i in any fundamental cut set is equal to the sum of the capacitor charges

$$\frac{dq_i(t)}{dt} = J_i^C(t)$$

where $J_i^C(t)$ is the contribution of the current sources to the ith cut set. As $\mathbf{J}^C$ is a linear combination of $\mathbf{i}_r$, $\mathbf{i}_L$, and $\mathbf{J}$ it follows that

$$\frac{d}{dt}\,\mathbf{q}(t) = \mathbf{f}_q(\mathbf{i}_r(t), \mathbf{i}_L(t), \mathbf{J}(t)) \tag{12}$$

where $\mathbf{f}_q$ is linear and does not depend explicitly on time.

Let us now consider the inductors of the network $\mathfrak{N}$. Given any set of resistor voltages $\mathbf{e}_r$ and capacitor voltages $\mathbf{e}_c$ such that Kirchoff's voltage law is satisfied in each loop formed by capacitors, resistors and voltage sources, let us replace the capacitor and resistor branches of $\mathfrak{N}$ with voltage sources equal to the corresponding voltages $\mathbf{e}_c$, $\mathbf{e}_r$. The network now consists of sources and inductors only. Let us again use the algorithm of Section III to change the configuration of the sources and separate the network into its separable parts. The resulting network is denoted by $\mathfrak{N}^L$ and its sources by $\mathbf{E}^L, \mathbf{J}^L$. As in the case of the resistive network, we are going to use the result previously proved for $\mathfrak{N}^R$ to deduce a similar result for $\mathfrak{N}^L$.

$\mathfrak{N}^L$ consists of the three sets of subnetworks which were described in connection with $\mathfrak{N}^R$. Consider the third set of subnetworks of $\mathfrak{N}^L$, which consists of single inductors in parallel with a current source. Except for the trivial case where $\mathfrak{N}$ consists only of a single inductor in parallel with a current source, this set is empty, for otherwise $\mathfrak{N}_{(E)(J)}$ would be separable. Condition (3) implies that each subnetwork of the first set has a tree, say τ_L , whose tree branches are monotonically increasing and current-controlled and whose links are monotonically increasing and flux-controlled. For each subnetwork of the first set, with each fundamental loop of τ_L defined by an inductive link, we assign a variable φ_i equal to the sum of the fluxes of all the inductors of that loop. For each subnetwork of the second set we assign a φ_i equal to the flux of the inductor. $\boldsymbol{\varphi}$ will denote the vector whose components are the φ_i's.

The analogy between $\mathfrak{N}^L$ and $\mathfrak{N}^R$ is established in four steps:

(*i*) For $\mathfrak{N}^R$ the Kirchhoff voltage law applied to the ith loop associated with a resistive link reads

$$E_i^R = \sum_k e_{ki}$$

where e_{ki} is the voltage across the kth branch of the ith loop. The e_{ki} are components of $\mathbf{e}_r$. For $\mathfrak{N}^L$ we have by definition of φ_i ,

$$\varphi_i = \sum_k \varphi_{ki}$$

where φ_{ki} is the flux in the kth branch of the ith loop. The φ_{ki} are components of $\boldsymbol{\varphi}_L$.

(*ii*) For both $\mathfrak{N}^R$ and $\mathfrak{N}^L$ the Kirchhoff current law holds.

(*iii*) Condition (3) imposes requirements on the topology and ele-

ment characteristics of $\mathfrak{N}^L$ which are entirely similar to those imposed on $\mathfrak{N}^R$ by condition (2).

(*iv*) Finally, the elements of $\mathfrak{N}^R$ and $\mathfrak{N}^L$ satisfy analogous Lipschitz conditions by condition (4).

Therefore, the variables $(\varphi_L, \mathbf{i}_L)$ and $(\varphi, \mathbf{J}^L)$ of $\mathfrak{N}^L$ are analogous to the variables $(\mathbf{e}_r, \mathbf{i}_r)$ and $(\mathbf{E}^R, \mathbf{J}^R)$.

As $\mathbf{J}^L$ is linearly related to $\mathbf{J}$, we conclude that fluxes and currents of the inductors at time t are uniquely determined by the values, at the same time t, of $(\varphi(t), \mathbf{J}(t))$, and that the mappings

$$\varphi_L(t) = \mathbf{f}_{\varphi_L}(\varphi(t), \mathbf{J}(t), t) \tag{13}$$

$$\mathbf{i}_L(t) = \mathbf{f}_{i_L}(\varphi(t), \mathbf{J}(t), t) \tag{14}$$

satisfy Lipschitz conditions.

Since φ_i in each loop is equal to the sum of the fluxes in the loop, it follows from the Kirchhoff voltage law that

$$\frac{d}{dt}\varphi_i(t) = E_i^L(t)$$

where $E_i^R(t)$ is the contribution of the voltage sources in the ith loop. As $\mathbf{E}^L$ is a linear combination of $\mathbf{e}_c$, $\mathbf{e}_r$, and $\mathbf{E}$, it follows that

$$\frac{d}{dt}\varphi(t) = \mathbf{f}_{\varphi}(\mathbf{e}_c(t), \mathbf{e}_r(t), \mathbf{E}(t)) \tag{15}$$

where $\mathbf{f}_{\varphi}$ is linear and does not depend explicitly on time.

Any solution of the network requires that (8), (9), (10), (11), (12), (13), (14) and (15) be satisfied simultaneously. It is shown in the following that these equations determine a unique solution.

In (12) and (15) substitute values of $\mathbf{e}_c$, $\mathbf{e}_r$, $\mathbf{i}_r$ and $\mathbf{i}_L$ from (10), (11), (9) and (14). The results are

$$\frac{d}{dt}\mathbf{q} = \mathbf{f}_q[\mathbf{f}_{i_r}(\mathbf{f}_{e_c}(\mathbf{E},\mathbf{q},t),\mathbf{E},\mathbf{f}_{i_L}(\varphi,\mathbf{J},t),\mathbf{J},t),\mathbf{f}_{i_L}(\varphi,\mathbf{J},t),\mathbf{J}] \tag{16}$$

$$\frac{d}{dt}\varphi = \mathbf{f}_{\varphi}[\mathbf{f}_{e_c}(\mathbf{E},\mathbf{q},t),\mathbf{f}_{e_r}(\mathbf{f}_{e_c}(\mathbf{E},\mathbf{q},t),\mathbf{E},\mathbf{f}_{i_L}(\varphi,\mathbf{J},t),\mathbf{J},t),\mathbf{E}]. \tag{17}$$

Since the right-hand sides of (16) and (17) are compositions of functions satisfying Lipschitz conditions, these equations may be rewritten as

$$\frac{d}{dt}\mathbf{q}(t) = \mathbf{F}_q(\mathbf{E}(t),\mathbf{q}(t),\varphi(t),\mathbf{J}(t),t) \tag{18}$$

$$\frac{d}{dt}\varphi(t) = \mathbf{F}_{\varphi}(\mathbf{E}(t),\mathbf{q}(t),\varphi(t),\mathbf{J}(t),t) \tag{19}$$

where $\mathbf{F}_q$ and $\mathbf{F}_\varphi$ satisfy Lipschitz conditions in $\mathbf{q}$ and φ. Therefore, for any $\mathbf{E}(\cdot)$ and $\mathbf{J}(\cdot)$ that are regulated functions of time[13] and for any initial values of φ and $\mathbf{q}$, the differential equations (18) and (19) determine uniquely $\varphi(\cdot)$ and $\mathbf{q}(\cdot)$, and the solutions are continuous functions of time.[18] In terms of $\mathbf{E}(\cdot)$, $\mathbf{J}(\cdot)$, $\varphi(\cdot)$ and $\mathbf{q}(\cdot)$, equations (8), (9), (10), (11), (13) and (14) determine uniquely the currents and voltages of the resistive branches, the voltages and charges of the capacitive branches, and the fluxes and currents of the inductive branches. Therefore the network $\mathfrak{N}$ is determinate. (Incidentally, the proof shows that the state of the network may be represented by $(\mathbf{q}, \varphi)$.)

It is worth indicating an immediate consequence of (18) and (19) and the other circuit relations.

Corollary: If the conditions of Theorem IV are satisfied, $\mathbf{E}$ and $\mathbf{J}$ are continuous functions of time, and all elements depend continuously on time, then $\mathbf{e}_c$, $\mathbf{q}_c$, $\mathbf{e}_r$, $\mathbf{i}_r$, φ_L, $\mathbf{i}_L$ are continuous functions of time; in other words, jump phenomena[1] are excluded.

Corollary: Let the network $\mathfrak{N}$ consist of independent sources, nonlinear (possibly time-dependent) monotonically increasing one-to-one resistors, capacitors and inductors. If the characteristics of all elements Lipschitz conditions as described in condition (4) of Theorem IV, the network $\mathfrak{N}$ is determinate.

Corollary: If branches of a network $\mathfrak{N}$ consist of: (a) voltage sources, current sources; (b) one-to-one monotonically increasing resistors, capacitors, and inductors whose characteristics satisfy condition (4) of Theorem IV; (c) one-to-one two-poles of the types described by the theorems of Section II and which satisfy conditions (*a*), (*b*), (*c*) and (*d*) of these theorems: *then* network $\mathfrak{N}$ is determinate.

Given a physical circuit or device, it may happen that a particular model of the circuit does not satisfy the conditions of Theorem IV. For example, this model $\mathfrak{N}'$ might be such that $S*\mathfrak{N}'_{(E)C}$ includes a parallel connection of two charge-controlled capacitors, $D_1(q)$ and $D_2(q)$, with only D_1 monotonically increasing. Under these conditions, it may happen that the current through the parallel combination does not determine uniquely the voltage across it. If, however, the model is changed (call it $\mathfrak{N}''$) and a resistor (or inductor) is inserted in series with D_2, then $S*\mathfrak{N}''_{(E)C}$ now includes an open branch D_2, condition (1) of Theorem IV is no longer violated, and $\mathfrak{N}''$ is determinate. Obviously, this idea may be used in the case of inductors and resistors.

Finally, let us conclude this section by a discussion which draws attention to some consequences of the conditions of Theorem IV. Some of the properties considered here will be used in the next section for writing in detail the network equations.

Let $\mathfrak{N}$ be a network which satisfies the conditions of Theorem IV. Denote by r,r_{1-1}, g the sets of resistors which are current-controlled (but not voltage-controlled), one-to-one, and voltage-controlled (and not current-controlled), respectively. Similarly, denote respectively by d,d_{1-1}, c the charge-controlled (and not voltage-controlled), one-to-one, and voltage-controlled capacitors, and by l,Γ_{1-1}, Γ the current-controlled (and not flux-controlled), one-to-one, and flux-controlled inductors.

Let us carry out the following operations:

(a) choose a forest of $\mathfrak{N}_E$

(b) choose a forest of $\mathfrak{N}_{(E)d}$

(c) choose a forest of $\mathfrak{N}_{(Ed)d_{1-1}}$

(d) choose a forest of $\mathfrak{N}_{(Edd_{1-1})r}$

(e) choose a forest of $\mathfrak{N}_{(Edd_{1-1}r)r_{1-1}}$

(f) choose a forest of $\mathfrak{N}_{(Edd_{1-1}rr_{1-1})l}$

(g) choose a forest of $\mathfrak{N}_{(Edd_{1-1}rr_{1-1}l)\Gamma_{1-1}}$.

Since the conditions of Theorem IV are satisfied by $\mathfrak{N}$, it follows that *the union of these forests forms a tree of* $\mathfrak{N}$ which we denote by τ. The construction of this tree is an extension of Bryant's procedure.[9]

This can be proved in the following way: From the conditions of Theorem IV it follows that the union of the forests chosen by (b) and (c) [by (d) and (e), by (f) and (g)] are forests of $\mathfrak{N}_{(E)C}$, [$\mathfrak{N}_{(EC)R}$, and $\mathfrak{N}_{(ECR)L}$, respectively]. Let us add the network's resistors to $\mathfrak{N}_{(ECR)L}$. This is done by splitting nodes and adding the new branches between them. Consider a node which was split, say, to three nodes and a resistor subnetwork connected between these nodes, It is clear that the subtree of this resistive subnetwork completes the forest of $\mathfrak{N}_{(ECR)L}$ for a forest of the $\mathfrak{N}_{(EC)RL}$. We can use the same argument to show that by adding the capacitors and voltage sources we get a forest of the network which includes all branches but the current sources. However, the current sources do not form any cut set and therefore are links of this forest. Thus τ, the union of these forests, is a tree of $\mathfrak{N}$.

From the construction of the tree, the conditions of Theorem IV, and the above discussion it follows that τ *contains all current and charge-controlled elements which are not one-to-one, and all voltage and flux-controlled elements which are not one-to-one are links of this tree.*

Consider the fundamental cut set of $\mathfrak{N}$ defined by an element of set d, a charge-controlled capacitor whose characteristic is not monotonically increasing. By assumption, this capacitor is an open branch of $S*\mathfrak{N}_{(E)C}$;

this cut set may not contain capacitors or voltage sources and therefore consists solely of resistors, inductors and current sources. Similar properties exist for fundamental loops defined by the various links. One can exhibit these properties by making a table in which links and tree branches are partitioned according to the types of their elements and properties of their characteristics; for each link and branch the table specifies the type of elements that are allowed to be in the corresponding loop or cut set. As the table is complicated, it is omitted and only some of the more interesting properties are listed below. Here we are going to make use of the rank-order of the elements, $ECRLJ$, defined in Section II:

(*i*) Tree branches with characteristics which are not monotonically increasing are the *highest* ranked elements in their own fundamental cut set. Thus, for example, a charged-controlled nonmonotonically increasing capacitor has a fundamental cut set which may include links which are resistors, inductors and current sources but no other capacitors.

(*ii*) Links with characteristics which are not monotonically increasing are the *lowest* ranked elements in their own fundamental loop. Thus, a fundamental loop defined by a nonmonotonically increasing resistor may have only capacitors or voltage sources in its tree branches.

VI. EQUATIONS FOR RLC NETWORKS

The purpose of this section is to write explicitly the equations of a nonlinear RLC circuit of the type considered in the previous section. Another purpose is to exhibit the similarities and differences between the equations that describe linear networks and those that describe the nonlinear networks under consideration.

To simplify the exposition consider the resistive network of Fig. 14. Call τ_1 the tree formed by the branches 1,2,3, and the voltage source E. If the network were linear, the fundamental cut set equations would read

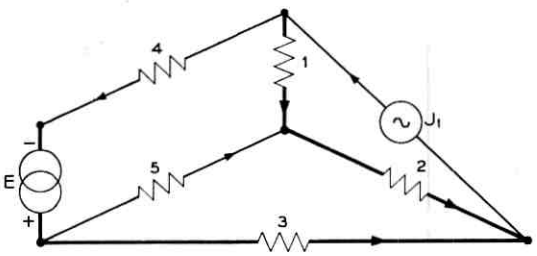


Fig. 14 — Resistive network.

$$
\begin{aligned}
g_1 e_1 + g_4(e_1 + e_2 - e_3 - E) &= J_1 \\
g_4(e_1 + e_2 - e_3 - E) + g_2 e_2 - g_5(-e_2 + e_3) &= J_1 \qquad (20) \\
-g_4(e_1 + e_2 - e_3 - E) + g_3 e_3 + g_5(-e_2 + e_3) &= 0
\end{aligned}
$$

where e_i is the voltage across the ith branch and g_i is the conductance of this branch. In the well known matrix form, the equations become

$$
\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 1 \end{bmatrix}
\begin{bmatrix} g_1 & 0 & 0 & 0 & 0 \\ 0 & g_2 & 0 & 0 & 0 \\ 0 & 0 & g_3 & 0 & 0 \\ 0 & 0 & 0 & g_4 & 0 \\ 0 & 0 & 0 & 0 & g_5 \end{bmatrix}
\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & -1 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix}
\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ E \end{bmatrix} = \begin{bmatrix} J_1 \\ J_1 \\ 0 \end{bmatrix} \qquad (21)
$$

or, more generally,

$$
\mathbf{\Delta}_{T(R),R}\mathbf{G}_R\mathbf{\Delta}'_{T(RE),R}\begin{bmatrix}\mathbf{e}\\ \mathbf{E}\end{bmatrix} + \mathbf{\Delta}_{T(R),L(J)}\mathbf{J} = \mathbf{0} \qquad (22)
$$

where $\mathbf{e}$, $\mathbf{E}$ and $\mathbf{J}$ are column vectors whose components are the tree branch voltages, $\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$, the voltage sources, $[E]$, and the current sources, $[J]$, of the network; $\mathbf{G}_R$ is the branch admittance matrix. The $\mathbf{\Delta}$'s are appropriate submatrices of the fundamental cut set matrix $\mathbf{Q}$. The first subscript of $\mathbf{\Delta}$ denotes the rows and the second subscript denotes the columns of $\mathbf{Q}$ whose intersection forms the submatrix. Thus $\mathbf{\Delta}_{T(RE),L(R)}$ is a submatrix formed by the intersection of rows corresponding to resistive and voltage source tree branches and columns corresponding to resistive links; $\mathbf{\Delta}_{T(R),R}$ is formed by the intersection of rows corresponding to resistive tree branches and columns corresponding to resistive branches. $\mathbf{\Delta}_{T(R),J}$ is defined similarly. The prime over a matrix indicates transposition. Now, let the resistors become monotonically increasing one-to-one nonlinear resistors. Without loss of generality we can assume these new resistors to be time invariant. Let $\bar{g}_1(\cdot)$, $\bar{g}_2(\cdot)$, $\bar{g}_3(\cdot)$, $\bar{g}_4(\cdot)$ and $\bar{g}_5(\cdot)$ be their characteristics.

The cut set equations are:

$$
\begin{aligned}
\bar{g}_1(e_1) + \bar{g}_4(e_1 + e_2 - e_3 - E) &= J_1 \\
\bar{g}_4(e_1 + e_2 - e_3 - E) + \bar{g}_2(e_2) - \bar{g}_5(e_3 - e_2) &= J_1 \qquad (23) \\
-\bar{g}_4(e_1 + e_2 - e_3 - E) + \bar{g}_5(e_3 - e_2) + \bar{g}_3(e_3) &= 0
\end{aligned}
$$

where, for example, $\bar{g}_1(e_1)$ is now the value of the function $\bar{g}_1$ evaluated at e_1.

The similarity between (20) and (23) suggests a shorthand notation

for writing the equations of nonlinear networks. By the product $\boldsymbol{A}*\mathbf{x}$ (where $\boldsymbol{A}$ is a diagonal matrix whose elements are functions $a_i(\cdot)$ and $\mathbf{x}$ is a column vector whose components are $x_1, x_2, \cdots, x_n$), we denote the column vector whose ith component is $a_i(x_i)$, that is, the ith diagonal element of $\boldsymbol{A}$ evaluated at the ith component of $\mathbf{x}$. With this symbolic notation the equations of the network of Fig. 14 can be written (for the nonlinear case) in a form analogous to (22).

$$\boldsymbol{\Delta}_{T(R),R}\boldsymbol{G}_R*\left(\boldsymbol{\Delta}'_{T(RE),R}\begin{bmatrix}\mathbf{e}\\ \mathbf{E}\end{bmatrix}\right)+\boldsymbol{\Delta}_{T(R),L(J)}\mathbf{J}=\mathbf{0}$$

where $\boldsymbol{G}_R$ is the diagonal matrix whose elements are the characteristics $\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_5$ and the $*$ operation must be interpreted as indicated above. $\boldsymbol{G}_R$ will be referred to the branch characteristic matrix. With this symbolic notation, cut set matrices, loop matrices and branch resistance matrices may be used to writing equations of nonlinear networks in the same way as for linear networks.

Let us now assume that the elements of the tree τ_1 of the network of Fig. 14 are monotonically increasing current-controlled but not voltage-controlled and the links are monotonically increasing voltage-controlled but not current-controlled. Since the tree branches are not voltage-controlled, the equations cannot be written in the form of (22). Let $\bar{r}_1(\cdot)$, $\bar{r}_2(\cdot)$ and $\bar{r}_3(\cdot)$ represent the characteristics of the tree branches and i_1, i_2 and i_3 be the currents of the corresponding tree branches. In terms of the tree branch voltages and currents the cut set equations become:

$$\begin{aligned} i_1+\bar{g}_4(e_1+e_2-e_3-E)&=J_1\\ i_2+\bar{g}_4(e_1+e_2+e_3-E)-\bar{g}_5(e_3-e_2)&=J_1\\ i_3-\bar{g}_4(e_1+e_2-e_3-E)+\bar{g}_5(e_2-e_3)&=0. \end{aligned} \tag{24}$$

The other set of equations is

$$\begin{aligned} e_1&=\bar{r}_1(i_1)\\ e_2&=\bar{r}_2(i_2)\\ e_3&=\bar{r}_3(i_3) \end{aligned} \tag{25}$$

or symbolically

$$\mathbf{i}_{T(R)}+\boldsymbol{\Delta}_{T(R),L(R)}\boldsymbol{G}_{L(R)}*\left(\boldsymbol{\Delta}'_{T(RE),L(R)}\begin{bmatrix}\mathbf{e}_{T(R)}\\ \mathbf{E}\end{bmatrix}\right)+\boldsymbol{\Delta}_{T(R),L(J)}\mathbf{J}=\mathbf{0}. \tag{26}$$

$$\mathbf{e}_{T(R)}=\boldsymbol{R}_{T(R)}*\mathbf{i}_{T(R)} \tag{27}$$

where $\mathbf{i}_{T(R)}$, $\mathbf{e}_{T(R)}$, $\mathbf{E}$ and $\mathbf{J}$ are the tree currents and voltages, and vol-

tage and current sources respectively. $\boldsymbol{G}$ and $\boldsymbol{R}$ are the link and branch characteristic matrices, and in our example

$$\boldsymbol{G}_{L(R)} = \begin{bmatrix} g_4 & 0 \\ 0 & \bar{g}_5 \end{bmatrix}, \qquad \boldsymbol{R}_{T(R)} = \begin{bmatrix} \mathfrak{R}_1 & 0 & 0 \\ 0 & \mathfrak{R}_2 & 0 \\ 0 & 0 & \mathfrak{R}_3 \end{bmatrix}.$$

The $\boldsymbol{\Delta}$'s are appropriate submatrices of the fundamental cut set matrix $\mathbf{Q}$. The first subscript denotes the rows and the second subscript denotes the columns of $\mathbf{Q}$ whose intersection forms the submatrix. Thus $\boldsymbol{\Delta}_{T(RE),L(R)}$ is a submatrix formed from the intersection of rows corresponding to resistive and voltage source tree branches and columns corresponding to resistive links. $\boldsymbol{\Delta}_{T(R),L(R)}$ and $\boldsymbol{\Delta}_{T(R),L(J)}$ are defined similarly. A comparison of (26), (27) and (22) shows that in the case of current-controlled tree branches and voltage-controlled links which are not one-to-one, we need *both* $\mathbf{i}_{T(R)}$ and $\mathbf{e}_{T(R)}$ for a straightforward writing of the cut set equations and the branch characteristic equations. Either $\mathbf{i}_{T(R)}$ or $\mathbf{e}_{T(R)}$ can be eliminated from the equations. The resulting equations are:

$$\mathbf{i}_{T(R)} + \boldsymbol{\Delta}_{T(R),L(R)}\boldsymbol{G}_{L(R)}*\left(\boldsymbol{\Delta}'_{T(RE),L(R)}\begin{bmatrix}\boldsymbol{R}_{T(R)}*\mathbf{i}_{T(R)} \\ \mathbf{E}\end{bmatrix}\right) + \boldsymbol{\Delta}_{T(R),L(J)}\mathbf{J} = \mathbf{0} \tag{28}$$

Or

$$\mathbf{e}_{T(R)} + \boldsymbol{R}_{T(R)}*\left\{\boldsymbol{\Delta}_{T(R),L(R)}\boldsymbol{G}_{L(R)}*\left(\boldsymbol{\Delta}'_{T(RE),L(R)}\begin{bmatrix}\mathbf{e}_{T(R)} \\ \mathbf{E}\end{bmatrix}\right)\right\} + \boldsymbol{R}_{T(R)}*(\boldsymbol{\Delta}_{T(R),L(J)}\mathbf{J} = \mathbf{0}. \tag{29}$$

Fundamental loop equations can be written in a similar way using both the voltages and currents of the links $\mathbf{i}_{L(R)}$ and $\mathbf{e}_{L(R)}$. The equations are

$$\mathbf{e}_{L(R)} + \mathbf{l}_{L(R),T(R)}\boldsymbol{R}_{T(R)}*\left(\mathbf{l}'_{L(RJ),T(R)}\begin{bmatrix}\mathbf{i}_{L(R)} \\ \mathbf{J}\end{bmatrix}\right) + \mathbf{l}_{L(R),T(E)}\mathbf{E} = \mathbf{0}$$

$$\mathbf{i}_{L(R)} = \boldsymbol{G}_{L(R)}*\mathbf{e}_{L(R)}$$

where the l's are appropriate submatrixes of the fundamental tie set matrix $\mathbf{B}$. Similarly to (26) and (27), either $\mathbf{e}_{L(R)}$ or $\mathbf{i}_{L(R)}$ can be eliminated.

We now write the equations for a general *RLC* network (which satisfies the requirements of Theorem IV) by performing the following steps†

† Other systems of variables are possible. For example, one can choose charges and fluxes as above and voltages of resistive links whose loop does not consist of capacitors and voltage sources only.

(*i*) A tree is chosen as explained in Section V.

(*ii*) Variables are chosen. We choose here the charges on the capacitive tree branches, $\mathbf{q}_D$, the currents of the resistive tree branches whose fundamental cut set *does not* consist of inductors and current sources only, $\mathbf{i}_R$, and the fluxes of the inductive links, φ_Γ.

The equations make use of the following characteristic branch matrixes: $\boldsymbol{C}$ and $\boldsymbol{D}$ are diagonal matrixes whose elements are the characteristics of voltage-controlled and charged-controlled capacitors, respectively; $\boldsymbol{G}$ and $\boldsymbol{R}$, those of voltage-controlled and current-controlled resistors, respectively; $\boldsymbol{L}$ and $\boldsymbol{\Gamma}$, those of current-controlled and flux-controlled inductors, respectively. Without loss of generality we can assume that the elements are time-invariant. The equations are

$$\frac{d}{dt}\left\{\mathbf{q}_D + \boldsymbol{\Delta}_{T(C),L(C)}\boldsymbol{C}_{L(C)}*\left(\boldsymbol{\Delta}'_{T(CE),L(C)}\begin{bmatrix}\boldsymbol{D}_{T(C)}*\mathbf{q}_D\\ \mathbf{E}\end{bmatrix}\right)\right\}$$
$$+\boldsymbol{\Delta}_{T(C),L(\bar{R})}\boldsymbol{G}_{L(\bar{R})}*\left(\boldsymbol{\Delta}'_{T(C\bar{R}E),L(\bar{R})}\begin{bmatrix}\boldsymbol{D}_{T(L)}*\mathbf{q}_D\\ \boldsymbol{R}_{T(\bar{R})}*\mathbf{i}_R\\ \mathbf{E}\end{bmatrix}\right) \tag{30}$$
$$+\boldsymbol{\Delta}_{T(C),L(L)}\boldsymbol{\Gamma}_{L(L)}*\varphi_\Gamma + \boldsymbol{\Delta}_{T(C),L(J)}\mathbf{J} = \mathbf{0}$$

$$\mathbf{i}_R + \boldsymbol{\Delta}_{T(\bar{R}),L(\bar{R})}\boldsymbol{G}_{L(\bar{R})}*\left(\boldsymbol{\Delta}'_{T(C\bar{R}E),L(\bar{R})}\begin{bmatrix}\boldsymbol{D}_{T(C)}*\mathbf{q}_D\\ \boldsymbol{R}_{T(\bar{R})}*\mathbf{i}_R\\ \mathbf{E}\end{bmatrix}\right) \tag{31}$$
$$+\boldsymbol{\Delta}_{T(\bar{R}),L(L)}\boldsymbol{\Gamma}_{L(L)}*\varphi_\Gamma + \boldsymbol{\Delta}_{T(\bar{R}),L(J)}\mathbf{J} = \mathbf{0}$$

$$\frac{d}{dt}\left\{\varphi_\Gamma + \mathbf{1}_{L(L),T(L)}\boldsymbol{L}_{T(L)}*\left(\mathbf{1}'_{L(LJ),T(L)}\begin{bmatrix}\boldsymbol{\Gamma}_{L(L)}*\varphi_\Gamma\\ \mathbf{J}\end{bmatrix}\right)\right\}$$
$$+\mathbf{1}_{L(L),T(R_1)}\boldsymbol{R}_{T(R_1)}*\left(\mathbf{1}'_{L(LJ),T(R_1)}\begin{bmatrix}\boldsymbol{\Gamma}_{L(L)}*\varphi_\Gamma\\ \mathbf{J}\end{bmatrix}\right) \tag{32}$$
$$+\mathbf{1}_{L(L),T(C)}\boldsymbol{D}_{T(C)}*\mathbf{q}_D + \mathbf{1}_{L(L),T(\bar{R})}\boldsymbol{R}_{T(\bar{R})}*\mathbf{i}_R$$
$$+\mathbf{1}_{L(L),T(E)}\mathbf{E} = \mathbf{0}$$

where R_1 is the set of resistive tree branches whose fundamental cut set contains inductive links and current sources only; and $\bar{R}$ is the set which contains all other resistive branches.

The terms in the brackets in (30) and (32) are equal to our state variables $\mathbf{q}$ and φ of Section V. One can write the equations in terms of these variables: the relations between $\mathbf{q}_D$ and $\mathbf{q}$ and φ_Γ and ϕ are given by

$$\mathbf{q}_D + \mathbf{\Delta}_{T(C),L(C)} \mathbf{C}_{L(C)} * \left(\mathbf{\Delta}'_{T(CE),L(C)} \begin{bmatrix} \mathbf{D}_{T(C)} * \mathbf{q}_D \\ \mathbf{E} \end{bmatrix} \right) = \mathbf{q}$$

$$\boldsymbol{\varphi}_\Gamma + \mathbf{1}_{L(L),T(L)} \mathbf{L}_{T(L)} * \left(\mathbf{1}'_{L(LJ),T(L)} \begin{bmatrix} \mathbf{\Gamma}_{L(L)} * \boldsymbol{\varphi}_\Gamma \\ \mathbf{J} \end{bmatrix} \right) = \boldsymbol{\varphi}.$$

In summary, the equations of the *RLC* nonlinear network are written in a way which is a generalization of the methods used in linear networks. However, great care must be taken of the fact that some characteristics are representable by functions which do not have inverses. This section indicated a method for tackling the problem. In this section, the equations are written in terms of three sets of variables: $\mathbf{q}_D$, the charges on the capacitive tree branches; $\boldsymbol{\varphi}_L$, the fluxes in the inductive links and $\mathbf{i}_R$, the currents in the resistive tree branches whose fundamental cut sets do not consist of only inductors and current sources. It is interesting to note that (except for the trivial case where $\mathfrak{N}$ consists of a single capacitor in parallel with a voltage source or a single inductor in parallel with a current source) the dimension of the state vector $(\mathbf{q},\boldsymbol{\varphi})$ used above is the same as in the linear case: [number of independent initial conditions] = [number of reactive elements] − [number of independent capacitor-only tie sets] − [number of independent inductor-only cut sets].[7,8,9]

REFERENCES

1. Minorsky, N., *Nonlinear Oscillations*, D. Van Nostrand, 1962, (part IV, especially Chap. 26) p. 614.
2. Duffin, R. J., Nonlinear Networks I, Bull. Am. Math. Soc. **52,** 1946, pp. 836–838.
3. Duffin, R. J., Nonlinear Networks II, Bull. Amer. Math. Soc. **53,** October, 1947, pp. 963–971.
4. Duffin R. J., Nonlinear Network IIb, Bull. Am. Math. Soc., **54,** 1948, pp. 119–127.
5. Birkhoff, G., and Diaz, J. B., Nonlinear Network Problems, Quart. Appl. Math., **13,** Jan., 1956, pp. 431–443.
6. Lock, K., *Coordinate Selection in Numerical Network Analysis*, Electrical Engineering Department, California Institute of Technology, Pasadena, California, 1962.
7. Bers, A., The Degree of Freedom in RLC Networks, IRE Trans. on Circuit Theory, **CT-6,** 1959, pp. 91–95.
8. Bryant, P. R., and Bers, A., The Degree of Freedom in RLC Networks, IRE Trans. on Circuit Theory, **CT-7,** 1960, p. 173.
9. Bryant, P. R., The Explicit Form of Bashkow's A Matrix, IRE Trans. on Circuit Theory, **CT-9,** Sept., 1962, pp. 303–306.
10. Bryant, P. R., The Order of Complexity of Electrical Networks, Proc. IEE, **106C,** June, 1959, pp. 174–188.
11. Weinberg, L., *Network Analysis and Synthesis*, McGraw-Hill, 1962.
12. Seshu, S., and Reed, M. B., *Linear Graphs and Electrical Networks*, Addison-Wesley, 1961.
13. Dieudonné, J., *Foundations of Modern Analysis*, Academic Press, 1960.
14. Apostol, T. M., *Mathematical Analysis*, Addison-Wesley, 1960.

15. Sandberg, I. W., private communication.
16. Branin, F. H., Jr., The Inverse of the Incidence Matrix of a Tree and the Formulation of the Algebraic First-Order Differential Equations of RLC Networks, IEEE Trans. on Circuit Theory, *CT-10*, Dec., 1963, p. 543–544.
17. Simmons, G. F., *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.
18. Coddington, E. A., and Levinson, N., *The Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

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 dept., and made several prototype magnetic-tape transports for storing digital data. He has recently conducted theoretical studies of data transmission systems by computer simulation. Member, AAAS, Sigma Xi and Tau Beta Pi.

William R. Bennett, B.S. in E.E., 1925, Oregon State University; M.A. in Physics, 1928, Ph.D., 1949, Columbia University; Bell Telephone Laboratories, 1925—. His early work was concerned with low-frequency transmission over wires and cables. He later became associated with the first coaxial carrier project and made basic studies on noise and distortion in broadband amplifiers. Time division multiplex and pulse code modulation were areas of subsequent major interest. He is now head of the data theory department in the data communications development laboratory. Fellow, IEEE; Member, American Physical Society, U.R.S.I., Sigma Xi, Tau Beta Pi and Eta Kappa Nu.

Paul T. Brady, B.E.E., 1958, Rensselaer Polytechnic Institute; M.S.E.E., 1960, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1961—. His work in human factors engineering has been concerned with studies of speech and voice-operated devices, especially as applied to satellite communication circuits.

James A. Cochran, B.S., 1956, M.S., 1957 and Ph.D., 1962 Stanford University; research mathematician Stanford Research Institute 1955–1958; research and teaching assistant Stanford University 1958–1961; Bell Telephone Laboratories 1962—. A member of the military research laboratory, he has been particularly concerned with antenna analysis and design and with various mathematical problems associated with the application of electromagnetic theory to microwave systems. Member, American Mathematical Society, Phi Beta Kappa and Sigma Xi.

JAMES R. DAVEY, B.S. in E.E., 1936, University of Michigan; Bell Telephone Laboratories, 1936—. He has been engaged in the design of telegraph and data transmission circuits for the following types of system; dc telegraph, multichannel AM and PM carrier telegraph, telegraph test and service boards, HF radio teletypewriter and a VHF ground-to-air data link. For the past several years he has been in charge of a department responsible for the development of various data terminals for use over telephone voice channels. Member, IEEE, Sigma Xi and Tau Beta Pi.

CHARLES A. DESOER, Sc.D., 1953, Massachusetts Institute of Technology; Bell Telephone Laboratories 1953–1958; University of California, Berkeley, 1958—. The academic year 1963–1964 was spent at Bell Telephone Laboratories. He has been concerned with the analysis and optimization of communication circuits, and the stability of control systems. He is the coauthor, with L. A. Zadeh, of the book *Linear System Theory*.

EDWIN O. ELLIOTT, A.B., 1949, M.A. 1951, Ph.D., 1959, University of California, Berkeley; Operations Evaluation Group of M.I.T., 1954–1958; Stanford Research Institute, 1958–1959; Assistant Professor of Mathematics, University of Nevada, Reno, 1959–1960; Bell Telephone Laboratories, 1960—. At Bell Laboratories he has been engaged in mathematical analysis of error-control methods for digital data communication systems and in the application of measure-theoretic techniques in the study of stochastic processes. He has also worked on problems in the congestion theory of traffic. Member, American Mathematical Society, Operations Research Society of America, Pi Mu Epsilon, Sigma Xi and Phi Beta Kappa.

JACOB KATZENELSON, B.Sc., 1957, and M.Sc., 1959, Technion, Israel Institute of Technology; Sc.D., 1962, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1962–September, 1964. He was engaged in studies of nonlinear networks and simulation of electronic circuits on digital computers. Mr. Katzenelson is now with the Electronic Systems Laboratory and Project Mac at M.I.T. Member, IEEE, Tau Beta Pi and Sigma Xi.

SAMUEL P. MORGAN, B.S., 1943, M.S., 1944, and Ph.D., 1947, California Institute of Technology; Bell Telephone Laboratories, 1947—. A research mathematician, Mr. Morgan has been particularly con-

cerned with the applications of electromagnetic theory to microwave and other problems. As Head, Mathematical Physics Department, he now supervises a research group in various fields of mathematical physics. Fellow, IEEE; member, American Physical Society, SIAM, Sigma Xi, Tau Beta Pi and AAAS.

J. Salz, B.S.E.E., 1955, M.S.E., 1956, Ph.D., 1961, University of Florida; The Martin Company, 1958–1960; Bell Telephone Laboratories, 1961—. He 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.

Er-Yung Yu, B. S., 1948, National Chiao-Tung University (China); M.S. 1957, Washington University; Ph.D. (Engineering Mechanics), 1960, Stanford University; Bell Telephone Laboratories, 1960—. Mr. Yu has been engaged in mechanics studies in problems of passive attitude control of satellites. He also participated in the Telstar satellite dynamics analysis and precession damper design. At present, he is working on the design of a feedback control system for laser frequency stabilization and on problems concerning the mechanical applications of lasers. Member, Sigma Xi and AIAA.

B.S.T.J. BRIEFS

A Note on a Special Class of One-Sided Distribution Sums

By R. D. BARNARD

(Manuscript received October 9, 1964)

I. INTRODUCTION

Occasionally encountered in the calculation of power spectra are limits of the form[1]

$$\lim_{N\to\infty} \sum_{n=0}^{N} [1 + F(n,N)]e^{inx},$$

where

$$F(n,N) \equiv \sum_{m=0}^{M} a_m(N)n^m \qquad M < \infty$$

$$a_m(N) = o(1) \qquad (N \to \infty) \qquad \forall m$$

$$x \in (-\infty,\infty), \qquad i = \sqrt{-1}.$$

It is shown here that these limits exist as distributions, or generalized functions,[2,3,4] and have several simple and useful representations. Specifically, we prove the following

Theorem:

$$\lim_{N\to\infty}^{(\mathrm{D})} \sum_{n=0}^{N} [1 + F(n,N)]e^{inx} = \frac{1}{2} + \pi \sum_{n=-\infty}^{\infty} \delta(x - 2\pi n) + \frac{i}{2}\cot\frac{x}{2}$$

$$= \lim_{\substack{\alpha\to 0 \\ \mathrm{Re}\ \alpha>0}}^{(\mathrm{D})} [1 - e^{-\alpha}e^{ix}]^{-1},$$

where $\lim^{(\mathrm{D})}$ and $\delta(\cdot)$ denote respectively a distribution limit[2,3] and the Dirac delta function.

II. ANALYSIS

Concerning notation, let C^∞ represent the space of infinitely differentiable scalar functions defined on the real line $(-\infty,\infty)$; C_d, the space of "rapidly decaying" test functions, viz., the linear vector space

$$C_d = \{\varphi \mid \varphi \in C^\infty, x^j\varphi^{(k)}(x) \to 0(|x| \to \infty)\forall j, k \geqq 0\};$$

and G, the space of generalized functions defined relative to the test functions of C_d. Finally, let Fg signify the generalized Fourier transform[2,3] of $g \in G$ with

$$F\varphi \equiv \int_{-\infty}^{\infty} \varphi(x) e^{-2\pi i y x} dx \qquad \varphi \in C_d .$$

The theorem under discussion is now established in terms of the following three lemmas:

Lemma I:

$$\lim_{N\to\infty}^{(D)} \sum_{n=0}^{N} [1 + F(n,N)] e^{inx} = \lim_{N\to\infty}^{(D)} \sum_{n=0}^{N} e^{inx} \equiv h(x) \in G.$$

Proof: Inasmuch as $Fh \in G$, then $h \in G$ and

$$\lim_{N} \int_{-\infty}^{\infty} a_m(N) \left[\sum_{n=0}^{N} e^{inx}\right] \varphi(x) dx = [\lim_{N} a_m(N)] \lim_{N} \int_{-\infty}^{\infty} \left[\sum_{n=0}^{N} e^{inx}\right] \varphi(x) dx$$
$$= 0 \quad \forall m.$$

Hence,

$$\lim_{N}^{(D)} \left[a_m(N) \sum_{n=0}^{N} e^{inx} \right] = 0 \quad \forall m,$$

and

$$\lim_{N\to\infty}^{(D)} \sum_{n=0}^{N} [1 + F(n,N)] e^{inx} = h(x) + \sum_{m=0}^{M} \left\{ \lim_{N}^{(D)} \left[a_m(N) \sum_{n=0}^{N} n^m e^{inx} \right] \right\}$$
$$= h(x) + \sum_{m=0}^{M} \left\{ (-i)^m \frac{d^m}{dx^m} \cdot \left[\lim_{N}^{(D)} a_m(N) \sum_{n=0}^{N} e^{inx} \right] \right\} = h(x).$$

Lemma II:

$$h(x) = \lim_{\substack{\alpha\to 0 \\ \text{Re } \alpha > 0}}^{(D)} [1 - e^{-\alpha} e^{ix}]^{-1}.$$

Proof: Setting

$$u(y) \equiv \begin{cases} 1, & y \geqq 0 \\ 0, & y < 0 \end{cases}$$

$$g_\alpha(y) \equiv \sum_{n=0}^{\infty} e^{-\alpha n} u\left(y - \frac{n}{2\pi}\right)$$

$$g_0(y) \equiv \lim_\alpha g_\alpha = \left[\sum_{n=0}^{\infty} u\left(y - \frac{n}{2\pi}\right)\right] \in G$$

and noting that

$$\begin{aligned} &\int_{-\infty}^{\infty} |\, g_0(y)\varphi(y)\,|\, dy < \infty \qquad \forall \varphi \in C_d \\ &|\, g_\alpha \,| \leqq g_0 \qquad \forall y \in (-\infty, \infty), \qquad \forall \operatorname{Re} \alpha > 0, \end{aligned} \tag{1}$$

one obtains by means of Lebesgue's dominated convergence theorem[5] the condition

$$\lim_\alpha \int_{-\infty}^{\infty} g_\alpha(y)\varphi(y)dy = \int_{-\infty}^{\infty} g_0(y)\varphi(y)dy \qquad \forall \varphi \in C_d. \tag{2}$$

Consequently,

$$\lim_\alpha{}^{(\mathrm{D})}\, g_\alpha = g_0\,, \tag{3}$$

and

$$h(x) = F^{-1}\cdot\sum_{n=0}^{\infty} \delta\left(y - \frac{n}{2\pi}\right) = F^{-1}\cdot\frac{d}{dy}\cdot g_0 = F^{-1}\cdot\frac{d}{dy}\cdot\lim_\alpha{}^{(\mathrm{D})}\, g_\alpha$$

$$= \lim_\alpha{}^{(\mathrm{D})}\cdot\sum_{n=0}^{\infty} e^{-\alpha n}e^{inx} = \lim_\alpha{}^{(\mathrm{D})}[1 - e^{-\alpha}e^{ix}]^{-1}.$$

Lemma III:

$$\lim_{\substack{\alpha\to 0 \\ \operatorname{Re}\alpha>0}}{}^{(\mathrm{D})}\, [1 - e^{-\alpha}e^{ix}]^{-1} = \frac{1}{2} + \pi\sum_{n=-\infty}^{\infty} \delta(x - 2\pi n) + \frac{i}{2}\cot\frac{x}{2}$$

Proof: From the definitions

$$C_\alpha(x) \equiv \frac{1}{2}\log\left[e^{-\alpha}\sin^2\frac{x}{2} + \left(\frac{1 - e^{-\alpha}}{2}\right)^2\right]$$

$$d_\alpha(x) \equiv \tan^{-1}\left[\frac{e^{-\alpha}\sin x}{1 - e^{-\alpha}\cos x}\right]$$

$$f_0(x) \equiv \pi\left[\sum_{n=0}^{\infty} u(x - 2\pi n) - \sum_{n=1}^{\infty} u(-x - 2\pi n)\right]$$

it is found that

$$| C_\alpha | \leqq \left| \log \left| \sin \frac{x}{2} \right| - \frac{1}{2} \right|$$

$$| d_\alpha | \leqq \frac{\pi}{2}$$

$$\lim_\alpha C_\alpha = \left[\log \left| \sin \frac{x}{2} \right| \right] \in G$$

$$\lim_\alpha d_\alpha = \left[f_0(x) - \frac{x}{2} \right] \in G$$

for all $\alpha \in (0,1)$ and almost all $x \in (-\infty, \infty)$. Therefore, as in (1), (2), and (3),

$$\lim_\alpha{}^{(D)} C_\alpha = \log \left| \sin \frac{x}{2} \right|$$

$$\lim_\alpha{}^{(D)} d_\alpha = f_0(x) - \frac{x}{2},$$

and

$$\begin{aligned} \lim_\alpha{}^{(D)} [1 - e^{-\alpha} e^{ix}]^{-1} &= \frac{d}{dx} \cdot \lim_\alpha{}^{(D)} [x + i \log (1 - e^{-\alpha} e^{ix})] \\ &= \frac{d}{dx} \cdot \lim_\alpha{}^{(D)} [x + d_\alpha(x) + i C_\alpha(x)] \\ &= \frac{1}{2} + \pi \sum_{n=-\infty}^{\infty} \delta(x - 2\pi n) + \frac{i}{2} \cot \frac{x}{2}. \end{aligned}$$

REFERENCES

1. Salz, J., The Spectral Density Function of Multilevel, Continuous-Phase FM, to be published.
2. Barnard, R. D., On the Spectral Properties of Single-Sideband Angle-Modulated Signals, B.S.T.J., **43,** Nov., 1964, p. 2811.
3. Temple, G., Generalized Functions, Proc. Roy. Soc. (London), **A228,** 1955, pp. 175–190.
4. Dunford, N., and Schwartz, J., *Linear Operators—Part II*, Interscience, New York, 1963.
5. Burkill, J. C., *The Lebesgue Integral*, Cambridge University Press, London, 1961, p. 41.